Longitudinal Cluster Analysis with Applications to Growth Trajectories

by

Brianna Christine Heggeseth

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Nicholas Jewell, Chair
Assistant Professor Cari Kaufman
Associate Professor Alan Hubbard

Spring 2013

Longitudinal Cluster Analysis with Applications to Growth Trajectories

Abstract

Longitudinal Cluster Analysis with Applications to Growth Trajectories

by

Brianna Christine Heggeseth

Doctor of Philosophy in Statistics

University of California, Berkeley

Professor Nicholas Jewell, Chair

Longitudinal studies play a prominent role in health, social, and behavioral sciences as well as in the biological sciences, economics, and marketing. By following subjects over time, temporal changes in an outcome of interest can be directly observed and studied. An important question concerns the existence of distinct trajectory patterns. One way to discover potential patterns in the data is through cluster analysis, which seeks to separate objects (individuals, subjects, patients, observational units) into homogeneous groups. There are many ways to cluster multivariate data. Most methods can be categorized into one of two approaches: nonparametric and model-based methods. The first approach makes no assumptions about how the data were generated and produces a sequence of clustering results indexed by the number of clusters $k = 2, 3, ...$ and the choice of dissimilarity measure. The later approach assumes data vectors are generated from a finite mixture of distributions. The bulk of the available clustering algorithms are intended for use on data vectors with exchangeable, independent elements and are not appropriate to be directly applied to repeated measures with inherent dependence.

Multivariate Gaussian mixtures are a class of models that provide a flexible parametric approach for the representation of heterogeneous multivariate outcomes. When the outcome is a vector of repeated measurements taken on the same subject, there is often inherent dependence between observations. However, a common covariance assumption is conditional independence—that is, given the mixture component label, the outcomes for subjects are independent. In Chapter 2, I study, through asymptotic bias calculations and simulation, the impact of covariance misspecification in multivariate Gaussian mixtures. Although maximum likelihood estimators of regression and prior probability parameters are not consistent under misspecification, they have little asymptotic bias when mixture components are well separated or if the assumed correlation is close to the truth even when the covariance is misspecified. I also present a robust standard error estimator and show that it outperforms conventional estimators in simulations and can provide evidence that the model is misspecified.

The main goal of a longitudinal study is to observed individual change over time; therefore, observed trajectories have two prominent features: level and shape of change over time. These features are typically associated with baseline characteristics of the individual. Grouping by shape and level separately provides an opportunity to detect and estimate these relationships. Although many nonparametric and model-based methods have been adapted for longitudinal data, most fail to explicitly group individuals according to the shape of their repeated measure trajectory. Some methods are thought to group by shape, but the dissimilarity between trajectories is not defined in terms of any one specific feature of the data. Rather, the methods are based on the entire vector and cluster trajectories by the level because it tends to dominate the variability between data vectors. These methods discover shape groups only if level and shape are correlated.

To fulfill the need for clustering based explicitly on shape, I propose three methods Chapter 4 that are adaptations of available algorithms. One approach is to use a dissimilarity measure based on estimated derivatives of functions underlying the trajectories. One challenge for this approach is estimating the derivatives with minimal bias and variance. The second approach explicitly models the variability in the level within a group of similarly shaped trajectories using a mixture model resulting in a multilayer mixture model. One difficulty with this method comes in choosing the number of shape clusters. Lastly, vertically shifting the data by subtracting the subject-specific mean directly removes the level prior to modeling. This non-invertible transformation can result in singular covariance matrixes, which makes parameter estimation difficult. In theory, all of these methods should cluster based on shape, but each method has shortfalls. I compare these methods with existing clustering methods in a simulation study in Chapter 5 and find that the vertical shifted mixture model outperforms the existing and other proposed methods.

A subset of the clustering methods are then compared on a real data set of childhood growth trajectories from the Center for the Health Assessment of Mothers and Children of Salinas (CHAMACOS) study in Chapter 6. Vertically shifting the data prior to fitting a mixture model results in groups based on the shape of their growth over time in contrast to the standard mixture model assuming either conditional independence or a more general correlation. The group means do not drastically change between methods for this data set, but group membership differs enough to impact inference about the relationship between baseline covariates and distinct groups.

To Paul, my partner in life.

Contents

List of Figures

List of Tables

Acknowledgments

This dissertation would never have come to fruition without the support of many individuals, and it is with pleasure that I acknowledge them.

I would like to express my utmost gratitude to my advisor, Nicholas Jewell, for giving me the freedom to explore unfamiliar territory and guidance when I needed it the most. He has taught me how to ask good questions and concisely express my ideas. I am indebted to him for helping me establish my career by providing opportunities for collaboration and professional development. Many thanks to Alan Hubbard and Cari Kaufman for their expertise and support as members of my dissertation committee. I want to thank Brenda Eskenazi and Kim Harley for access to data from the CHAMACOS cohort study and their contributed expertise regarding child development. I further thank and acknowledge Bin Yu, Deborah Nolan, and Barbara Abrams for their unending support and encouragement throughout my time at Berkeley. I strive to be the inspirational role model these three women have been to me.

I want to thank all of the faculty in the statistics department for providing me with a strong technical foundation on which I complete my research and the staff for providing administrative and computing support.

Thanks to *Statistics in Medicine* for accepting a subset of the material presented in Chapters 1 and 2 for inclusion in the publication titled, 'The Impact of Covariance Misspecification in Multivariate Gaussian Mixtures on Estimation and Inference: An Application to Longitudinal Modeling' to appear soon [51].

Many fellow graduate students have helped me stay sane through these challenging years. Their support and care helped me overcome stumbling blocks and continue my work. I greatly value their friendship and I deeply appreciate the time we have had together.

Most importantly, none of this would have been possible without the love and patience of my family. My parents told me to set my sights high and pushed me to reach my full potential. My friend, partner, and husband, Paul, never gives up on me and brings joy to my life each and every day.

Chapter 1

Introduction

Longitudinal studies play a prominent role in health, social, and behavioral sciences as well as in the biological sciences, economics, and marketing. By following subjects over time, temporal changes in an outcome of interest can be directly observed and studied. This dissertation studies the task of clustering longitudinal data. Statisticians, machine learning communities, and applied researchers have investigated this problem for other highly structured multivariate data sets such as time series and functional data, and it is important to distinguish between these different types of data. While all of these data structures are ordered by a variable such as time, they have different characteristics that call for different methods. Time series are sequences of data points measured uniformly over time from random processes on only a few units. Examples include monthly unemployment figures, yearly global temperatures, and hourly stock prices over time [128]. Functional data are assumed to provide information about smooth curves or functions that have been measured on a densely sampled grid. The data include many curves, one for each observational unit, and may be observed with substantial measurement error. Data such as hip angles over walking cycles and height during adolescence that have underlying smooth curves are categorized as functional data [111, 112].

In this dissertation, I focus on longitudinal data which includes sporadic repeated measurements of the same outcome observed on many subjects at sparse, potentially irregular times over a long period of time. Long-term ongoing longitudinal studies include the National Longitudinal Survey of Youth, which has attempted to interview individuals annually and then biannually for the past 30 years, and the Framingham Heart Study, which has interviewed three generations of subjects biannually over the past 60 years. These long-term studies have observed each individual as many as 30 times. The majority of longitudinal studies only observe individuals 5 to 15 times during the follow-up period due to limited resources. For the purposes of this dissertation, I often refer to a longitudinal time-ordered vector of outcomes as a trajectory as a reminder that the data vectors are ordered and represent a path of development over time.

An important question concerns the existence of distinct trajectory patterns. Cluster analysis seeks to separate objects (individuals, subjects, patients, observational units) into

homogeneous clusters. There are many ways to cluster multivariate data and there are a multitude of methods and algorithms available to use. Most methods can be categorized into one of two approaches: nonparametric and model-based methods. The first approach makes no assumptions about how the data were generated and produces a sequence of clustering results indexed by the number of clusters $k = 2, 3, ...$ and the choice of dissimilarity measure. The later approach assumes data vectors are generated from a finite mixture of distributions. The clustering results include the conditional probability of group membership and estimates of the parameters for the assumed data generation distribution [87, 37]. Both of these approaches are widely used and the pros and cons of the approaches have been discussed in depth [82, 35].

The bulk of the available clustering methods are intended for use on data vectors with exchangeable, independent elements and are not appropriate to be directly applied to repeated measures with inherent dependence [36]. Recent work to cluster longitudinal data has focused on extending and adjusting the standard clustering methods. Dissimilarity measures have been designed to be sensitive to permutations in the time order by incorporating the relationships between neighboring observations [15] or by calculating dissimilarity between coefficients of a functional projection rather than the raw vectors [125, 132, 1, 131, 54]. In the model-based approach, one can incorporate a functional basis in the mean structure [99, 41] and explicitly model the temporal correlation structure [97, 38, 89] to incorporate the longitudinal nature into the model.

1.1 Longitudinal data

In classical univariate statistics, each individual or subject gives rise to a single measurement, termed the outcome. In multivariate statistics, that one measurement is replaced by a vector of multiple outcome measurements. In longitudinal studies, a vector of measurements is observed for each subject, but the vector represents one outcome measured repeatedly at a sequence of observation times. Therefore, the data have characteristics of both multivariate and time series data.

Longitudinal data differ from multivariate data in that the time ordering imparts dependence among measurements that is not present in a typical multivariate data set. They differ from classical time series data because the data consists of a large number of independent trajectories that are sparsely and potentially irregularly sampled over time, rather than a few random processes that are uniformly sampled over time. These properties call for methods specific to longitudinal data [26]. Due to the sparsity, it is necessary to borrow strength between individuals by modeling the mean outcome as a function of explanatory variables. When there are distinct patterns in the data, averaging can be done within the groups.

In this dissertation, I let Y_{ij} represent an outcome random variable and $\mathbf{x}_{ij}$ be a design vector of length p based on explanatory variables observed at time t_{ij} for individual $i = 1, ..., n$. I let $\mathbf{w}_i$ be a design vector of length q based on baseline variables observed at or before t_{i1}. The set of repeated outcomes for individual i together in a vector is denoted as

$\mathbf{Y}_i = (Y_{i1}, ..., Y_{im_i})$ with mean $E(\mathbf{Y}_i) = \boldsymbol{\mu}_i$ and $m_i \times m_i$ covariance matrix $Cov(\mathbf{Y}_i) = \boldsymbol{\Sigma}_i$, where the element in the jth row and lth column of $\boldsymbol{\Sigma}_i$ is the covariance between Y_{ij} and Y_{il}. The covariance matrix can be written as $\boldsymbol{\Sigma}_i = \mathbf{V}_i^{1/2}\mathbf{R}_i\mathbf{V}_i^{1/2}$ where $\mathbf{R}_i$ is a $m_i \times m_i$ correlation matrix of $\mathbf{Y}_i$ and $\mathbf{V}_i$ is a $m_i \times m_i$ diagonal matrix with variances as the non-zero elements. In this dissertation, individuals are assumed independent of each other; therefore, $\mathbf{Y}_i$ and $\mathbf{Y}_{i'}$ are independent for $i \neq i'$. In general, I use capital letters to represent random variables or matrices, bold font for vectors and matrices, and small letters for specific observations. Hence, I denote $\mathbf{y}_i$ as the observed outcome vector for individual i.

1.2 Cluster analysis

One of our most natural abilities is grouping and sorting objects into categories. We simplify our world by recognizing similarities, grouping similar objects, and naming the groups. For example, when children first learn colors, different shades and hues of blue are grouped together as blue. Similarly, golden retrievers, poodles, and terriers are all dogs. These group labels allow us to summarize a large amount of information in a way that can be easily understood. Identifying and labeling groups of similar items is also important when summarizing a large data set so that information can be retrieved more efficiently by providing a concise description of patterns of similarities and differences in the data.

Cluster analysis is one tool to explore groups within a data set and it has found its way into various scientific disciplines. From mathematics and statistics to biology and genetics to economics and market research, each field uses its own terminology to describe the grouping process. It is called numerical taxonomy in biology, unsupervised pattern recognition in the machine learning literature, and data segmentation in market research, but the problem is same: to find groups of similar objects.

The problem is well defined, but the solution is not. There is no agreed upon criteria that determines that one grouping is better than another. In general, most statisticians agree that clusters should be formed by maximizing the similarity within groups and maximizing the dissimilarity between groups. Some believe that there are naturally occurring groups that need to be discovered, but others remark that cluster analysis can always find clusters even if there is no true underlying group structure. However, Bonner [8] suggests that the user is the ultimate judge of the meaning and value of the clusters. Clusters should not be judged on whether they are true or false but rather in their usefulness. No method works well on every data set to find meaningful groups. For this reason, there is a wealth of clustering algorithms in the literature.

In this dissertation, the general goal of clustering is to maximize similarity within groups understanding that the analysis is highly dependent on how the user determines individuals to be similar. In this way, clusters are meaningful to the user while maintaining the basic mathematical philosophy of clustering. I overview two approaches to clustering as they provide a foundation for the methods discussed in this dissertation.

1.3 Nonparametric clustering methods

The first clustering approach focuses on explicitly defining similarity between subjects rather than making assumptions about how the data were generated. The three key ingredients to these methods are the dissimilarity measure, the clustering algorithm, and the number of clusters.

1.3.1 Dissimilarity measures

There are many terms used for a measure to quantify the distinctiveness of a pair of subjects. Metric, distance, dissimilarity, and similarity are all related concepts. A metric is any function, d, that satisfies the following three mathematical properties,

$$\begin{aligned} (i) \quad & d(\mathbf{x}, \mathbf{y}) \geq 0 \text{ with equality if and only if } \mathbf{x} = \mathbf{y} \\ (ii) \quad & d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x}) \\ (iii) \quad & d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z}) \end{aligned}$$

for vectors $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^m$. Some people use the term distance function to describe a function that satisfies the first two conditions but not necessarily the third condition, known as the triangle inequality. Therefore, all metrics are also distance functions by this definition.

The most commonly used metric on Euclidean space is the Minkowski distance of order p defined as

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_p = \left(\sum_{j=1}^{m} |x_j - y_j|^p \right)^{1/p}$$

where $\mathbf{x} = (x_1, ..., x_m)$ and $\mathbf{y} = (y_1, ..., y_m)$. The metric equals the Euclidean distance when $p = 2$,

$$||\mathbf{x} - \mathbf{y}||_2 = \sqrt{\sum_{j=1} (x_j - y_j)^2},$$

and the Manhattan distance when $p = 1$,

$$||\mathbf{x} - \mathbf{y}||_1 = \sum_{j=1} |x_j - y_j|.$$

As p increases to infinity, the distance converges to the Chebyshev distance,

$$||\mathbf{x} - \mathbf{y}||_\infty = \max_j |x_j - y_j|.$$

Other distance measures are based on variants of the correlation coefficient. These do not satisfy the triangle inequality and are not considered metrics. One popular distance measure based on the Pearson correlation coefficient is defined as

$$d(\mathbf{x}, \mathbf{y}) = 1 - \frac{(\mathbf{x} - \bar{x})^T (\mathbf{y} - \bar{y})}{\|\mathbf{x} - \bar{x}\|_2 \|\mathbf{y} - \bar{y}\|_2}$$

where $\bar{x}$ represents the average of the elements in $\mathbf{x}$ and similarly for $\bar{y}$. The codomain of this distance function is $[0, 2]$. Perfect negative linear correlation results in the largest distance, perfect positive linear correlation results in the smallest, and no linear correlation is in the middle. The Pearson correlation coefficient can be replaced with its uncentered version to calculate the cosine-angle distance, defined as

$$d(\mathbf{x}, \mathbf{y}) = 1 - \frac{\mathbf{x}^T \mathbf{y}}{\|\mathbf{x}\|_2 \|\mathbf{y}\|_2}.$$

When vectors are exactly the same, the angle between the vectors is 0 and the distance equals 0. On the other hand, the maximum value of 2 is attained when two vectors are 180 degrees apart. Both of these distance measures are invariant to scaling such that $d(\mathbf{x}, a \cdot \mathbf{x}) = 0$ for $a > 0$ and the Pearson correlation distance is additionally invariant to shifts such that $d(\mathbf{x}, a \cdot \mathbf{x} + b) = 0$ for $a > 0$ and $a, b \in \mathbb{R}$.

It is also important to note that the distance functions introduced thus far require vectors of equal length and are invariant to permutation of the order of the elements in the vector.

More generally, a similarity function satisfies the following three conditions,

(i) $S(\mathbf{x}, \mathbf{y}) \geq 0$

(ii) $S(\mathbf{x}, \mathbf{y}) = S(\mathbf{y}, \mathbf{x})$

(iii) $S(\mathbf{x}, \mathbf{y})$ increases in a monotone fashion as $\mathbf{x}$ and $\mathbf{y}$ are more similar

where objects $\mathbf{x}$ and $\mathbf{y}$ need not be vectors in Euclidean space. A dissimilarity function is defined in the same way except that the function increases as the objects are more dissimilar. This general definition allows the function to be flexible enough to take into account how one deems two objects to be similar or dissimilar. In many cases the term distance and dissimilarity is used interchangeably if the objects are numeric vectors. The choice of the dissimilarity measure needs to be made after careful study as it can have a dramatic effect on the final clustering results.

1.3.2 Algorithms

Clustering algorithms aim to optimize a criterion based on the chosen dissimilarity measure. Most clustering algorithms can be classified into two group according to their general search strategies: hierarchical and partitioning algorithms. Hierarchical methods involve constructing a tree of clusters in which the root node is a cluster containing all objects and the leaves are clusters each containing one object. The tree can be constructed in a divisive manner (i.e., top down by recursively dividing groups) or agglomerative manner (i.e., bottom up by recursively combining groups). In order to combine groups, there are different ways of measuring the distance between groups of objects: single-linkage, complete-linkage, average-linkage. This approach should be used when there is a priori scientific knowledge of a hierarchical structure in the data.

Partition methods aim to map objects into $k \geq 2$ disjoint groups by maximizing a criterion without any larger hierarchical structure. Two popular methods include K-means [81, 49] and partitioning around medoids (PAM) [65].

K-means is one of the simplest unsupervised learning algorithms used to solve the well-known clustering problem. Given a set of vectors $(\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n)$ where $\mathbf{y}_i \in \mathbb{R}^m$ for all $i = 1, ..., n$, the K-means clustering algorithm aims to partition the n vectors into K sets $\{C_1, ..., C_K\}$ so as to minimize the sum of squared Euclidean distance to the assigned cluster centroids denoted as

$$\min_{\{C_k\}_{k=1}^K} \sum_{k=1}^{K} \sum_{i:\mathbf{y}_i \in C_k} \|\mathbf{y}_i - \boldsymbol{\mu}_k\|_2^2$$

where $\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_K$ are the centroids or mean vectors of the K clusters. To find the sets that minimize the criterion, the algorithm starts by strategically or randomly choosing K data vectors as the centroids. The general K-means algorithm proceeds by alternating between two steps:

- Assignment step: Assign each observation to the cluster with the closest centroid. The kth set equals

$$C_k = \{\mathbf{y}_i : \|\mathbf{y}_i - \boldsymbol{\mu}_j\|_2^2 \geq \|\mathbf{y}_i - \boldsymbol{\mu}_k\|_2^2 \ \ \forall \ \ 1 \leq j \leq k, \ \ i = 1, ..., n\}$$

  such that every $\mathbf{y}_i$ is in one and only one set.

- Update step: Calculate the centroid of each new set,

$$\boldsymbol{\mu}_k = \frac{1}{|C_k|} \sum_{i:\mathbf{y}_i \in \mathbf{C}_k} \mathbf{y}_i.$$

The algorithm continues to iterate until the sets no longer change. Depending on the initial partition, the algorithm is expected to converge to local optima; therefore, completing multiple random starts offers the best way to obtain a global optimum. The K-means algorithm works best when data clusters are about equal in size and shape. The algorithm attempts to find spherical clusters since the K-means algorithm is based on squared Euclidean distance. If the groups are not dense spherical densities with many spurious outliers, the centroid may not be representative of any of the cluster members.

The partitioning around medoids (PAM) algorithm attempts to overcome some of the issues with K-means. The algorithm operates on a user-provided dissimilarity matrix rather than squared Euclidean distance. It is robust to outliers since the medoid or middle vector for each group is selected from the observed data vectors rather than based on mean calculations. To find K sets of vectors to minimize the sum of dissimilarity of the vectors with their respective medoids, K individuals are first randomly chosen as medoids denoted as $\boldsymbol{\nu}_1, ..., \boldsymbol{\nu}_K$. The PAM algorithm alternates between the following two steps:

- Build step: Assign each observational unit to the closest medoid with respect to the given dissimilarity matrix.
- Swap step: For each $k = 1, ..., K$, swap the medoid, $\boldsymbol{\nu}_k$, with each non-medoid observation and compute the sum of the dissimilarities of the vectors with their closest medoid. Select the configuration with the smallest sum of dissimilarities.

The algorithm continues until there is no change in the medoids. Although it has a longer run time, this building and swapping procedure can return a smaller sum of dissimilarity in contrast to a simple assignment and update procedure in the K-means algorithm.

1.3.3 Choosing the number of components

A major challenge in clustering is to determine the optimal number of clusters. For the partition methods, a maximum number of clusters is chosen $K_{max} < n$ and the algorithm is run for each value $K = 2, 3, ..., K_{max}$. Then, some take a direct approach to choose the optimal K by optimizing functions of within and between cluster dissimilarity [90] or the average silhouette [65]. Others take a testing approach by comparing a cluster summary with its expectation under an appropriate null distribution using the Gap statistic [134] or the CLEST approach [28].

The average silhouette has the advantage of working well with any cluster routine and dissimilarity measure. For each vector i, the silhouette $s(i)$ is defined as

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

where $a(i)$ is the average dissimilarity of the ith vector to all other vectors in its cluster, A, $d(i, C)$ is average dissimilarity of the ith vector to all vectors in cluster C and $b(i) = \min_{C \neq A} d(i, C)$. The overall average silhouette width $\bar{s}$ is the average of $s(i)$ over all observational units in the data set. The chosen clustering algorithm is run and the overall average silhouette is calculated for each possible value of K. Then the optimal K is chosen to maximize the average silhouette while minimizing the within-group dissimilarities in comparison to the between-group dissimilarities.

1.4 Model-based clustering methods

For a chosen number of clusters, partitioning and hierarchical algorithms divide vectors into non-overlapping, non-empty sets. Each subject is in one and only one cluster. This type of clustering is referred to as hard clustering in contrast to soft or fuzzy clustering where each subject can be in multiple groups to varying degrees. This distinction largely impacts subjects that are near the edge between two groups in terms of the chosen dissimilarity measure. With hard clustering, those subjects are forced in one group. With soft clustering,

those subjects on the edge contribute to multiple clusters to varying degrees and the clustering results include association levels that indicate the uncertainty in group membership. Probabilistic model-based methods allow for soft clustering with the possibility of formal inference [39]. Assuming a probability distribution for the data provides a framework in which to estimate the probability of group membership as well as to estimate parameters and make inferences on the relationship between covariates and group membership. The main model used for clustering is the finite mixture model.

1.4.1 Finite mixture model

In a general finite mixture, the density for a random variable $\mathbf{y}$ takes the form

$$f(\mathbf{y}|\boldsymbol{\theta}) = \pi_1 f_1(\mathbf{y}|\boldsymbol{\theta}_1) + \cdots + \pi_K f_K(\mathbf{y}|\boldsymbol{\theta}_K)$$

where f_k and $\boldsymbol{\theta}_k$ are the density and parameters of the kth component and π_k is the probability an observation belongs to the kth component ($\pi_k > 0$; $\sum_k^K \pi_k = 1$). The full parameter vector $\boldsymbol{\theta}$ includes the prior probabilities π_k and the component parameters $\boldsymbol{\theta}_k$. In many situations, the component densities are assume multivariate Gaussian parameterized by a mean vector $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$,

$$f_k(\mathbf{y}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{\exp\left(-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\mathbf{y}-\boldsymbol{\mu}_k)\right)}{\sqrt{\det(2\pi\boldsymbol{\Sigma}_k)}}.$$

The Gaussian distribution can be reparameterized to accommodate longitudinal data and to reduce the number of parameters. If the mean outcome values are thought to depend on explanatory variables, the mean vectors are replaced with $\boldsymbol{\mu}_k = \mathbf{x}\boldsymbol{\beta}_k$ such that $\mathbf{x}$ is a design matrix based on those variables that impact the mean. The covariance matrix can be simplified by assuming a structure such as independence ($\boldsymbol{\Sigma}_k = \sigma_k^2\mathbf{I}$), compound symmetry ($\boldsymbol{\Sigma}_k = \sigma_k^2(\rho_k\mathbf{1}\mathbf{1}^T + (1-\rho_k)\mathbf{I})$), or exponential covariance ($[\boldsymbol{\Sigma}_k]_{jl} = \sigma_k^2\exp(-|t_{ij} - t_{il}|/r_k)$). On the other hand, the covariance structure can be parameterized through the eigenvalue decomposition of the form

$$\boldsymbol{\Sigma}_k = \lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$$

where $\mathbf{D}_k$ is the orthogonal matrix of eigenvectors, $\mathbf{A}_k$ is a diagonal matrix whose elements are proportional to the eigenvalues, and λ_k is a proportional constant [5]. The mixture model can also be extended to include more complexity through further parameterizations. If baseline factors are thought to influence group membership, the prior group probabilities can be parameterized using a generalized logit model

$$\pi_k(\mathbf{w}, \boldsymbol{\gamma}) = \frac{\exp(\mathbf{w}^T\boldsymbol{\gamma}_k)}{\sum_{j=1}^K \exp(\mathbf{w}^T\boldsymbol{\gamma}_j)}$$

where $\boldsymbol{\gamma}_K = \mathbf{0}$ and $\mathbf{w}$ is a design vector based on baseline factors.

1.4.2 Expectation-maximization algorithm

Under the assumption that $\mathbf{y} = \{\mathbf{y}_1, ..., \mathbf{y}_n\}$ are independent realizations from the mixture distribution $f(\mathbf{y}|\boldsymbol{\theta})$ defined above, the log-likelihood function for the full parameter vector $\boldsymbol{\theta}$ is given by

$$\log L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(\mathbf{y}_i|\boldsymbol{\theta}).$$

The maximum likelihood estimate of $\boldsymbol{\theta}$ is obtained by finding an appropriate root of the score equation, $\partial \log L(\boldsymbol{\theta})/\partial\boldsymbol{\theta} = \mathbf{0}$. Solutions of this equation corresponding to local maxima can be found iteratively through the expectation-maximization (EM) algorithm [25], which was originally presented in the context of an incomplete data problem. In the clustering case, $\mathbf{y}_i$ is assumed to have stemmed from one of the components and the label denoting its originating component is missing. Let z_{ik} equal 1 if $\mathbf{y}_i$ was generated from component k and 0 otherwise. The component label vectors $\mathbf{z}_i = (z_{i1}, ..., z_{iK})$ for $i = 1, .., n$ are assumed to be realizations of random vectors $\mathbf{Z}_i = (Z_{i1}, ..., Z_{iK})$ that follow a categorical distribution with probabilities $\pi_1, ..., \pi_K$. The complete data log-likelihood function, based on these component labels and the observed data $\mathbf{y}$, is given by

$$\log L_c(\boldsymbol{\theta}) = \sum_{k=1}^{K}\sum_{i=1}^{n} z_{ik} \left[\log \pi_k + \log f_k(\mathbf{y}_i|\boldsymbol{\theta}_k)\right].$$

The EM algorithm treats the z_{ik} as missing data and iteratively imputes the missing values and then estimates the parameters. The expectation step (E-step) involves taking the conditional expectation of the complete log-likelihood given the data and the current value of the parameter estimates. Since the complete log-likelihood is linear in the unobserved data z_{ik}, this step involves calculating the conditional expectation of Z_{ik} given the observed data. On the tth iteration, the expectation is

$$\begin{aligned} E_{\boldsymbol{\theta}^{(t-1)}}(Z_{ik}|\mathbf{y}) &= P_{\boldsymbol{\theta}^{(t-1)}}(Z_{ik}|\mathbf{y}) \\ &= \alpha_{ik}^{(t)} \end{aligned}$$

using the estimates of the parameters from the previous iteration. The quantity $\alpha_{ik}^{(t)}$ is the posterior probability that the ith individual belongs to component k, written as

$$\alpha_{ik}^{(t)} = \pi_k^{(t-1)} f_k(\mathbf{y}_i|\boldsymbol{\theta}_k^{(t-1)}) / \sum_{j=1}^{K} \pi_j^{(t-1)} f_j(\mathbf{y}_i|\boldsymbol{\theta}_j^{(t-1)})$$

for $i = 1, ..., n$ and $k = 1, ..., K$.

In the maximization step (M-step), the parameter estimates for the prior probabilities and component parameters are updated by maximizing the conditional expectation of the complete-data log-likelihood from the E-step. On the tth iteration, the updated estimates of

the prior probabilities $\pi_k^{(t)}$ are calculated independently of the estimates of the component parameters $\boldsymbol{\theta}_k^{(t)}$. The prior probability estimates equal

$$\pi_k^{(t)} = n^{-1} \sum_{i=1}^{n} \alpha_{ik}^{(t)}.$$

If the generalized logit function is used for the prior probabilities to include dependence on baseline covariates such that $\pi_k = \pi_k(\mathbf{w}_i, \boldsymbol{\gamma}) = \frac{\exp(\mathbf{w}_i^T \boldsymbol{\gamma}_k)}{\sum_{j=1}^{K} \exp(\mathbf{w}_i^T \boldsymbol{\gamma}_j)}$, then the updated value of $\boldsymbol{\gamma}^{(t)} = (\boldsymbol{\gamma}_1^{(t)}, ..., \boldsymbol{\gamma}_{K-1}^{(t)})$ is an appropriate root of

$$\sum_{k=1}^{K} \sum_{i=1}^{n} \alpha_{ik}^{(t)} \partial \log \pi_k(\mathbf{w}_i, \boldsymbol{\gamma}) / \partial \boldsymbol{\gamma} = \mathbf{0}$$

calculated using a numerical optimization routine. The component parameters are updated by finding an appropriate root of

$$\sum_{i=1}^{n} \alpha_{ik}^{(t)} \partial \log f_k(\mathbf{y}_i | \boldsymbol{\theta}_k) / \partial \boldsymbol{\theta}_k = \mathbf{0}$$

at the tth iteration for each component $k = 1, ..., K$. Closed-form solutions to this equation exist for many Gaussian mixture models.

The E-step and M-step are alternated repeatedly until convergence. Dempster, Laird, and Rubin [25] showed the incomplete likelihood increases monotonically at each iteration of this algorithm. Hence, the EM algorithm guarantees convergence to a local maximum; global convergence may be attained by running the algorithm multiple times randomly assigning individuals to initial components and using the estimates associated with the one with the highest log-likelihood. For more details about the EM algorithm, see McLachlan and Krishnan [88].

When the algorithm has converged, the values of the parameters and posterior probabilities from the last iteration are taken as the final estimates. The component estimates characterize the mean and variances within the clusters and the posterior probabilities provide the strength of group association for the soft clustering. This clustering can be translated into a hard clustering by choosing the group label that maximizes the posterior probability, $\arg\max_k \alpha_{ik}$, for each subject. Then, one measure of uncertainty in the group labels is $(1 - \max_k \alpha_{ik})$. A variant of the EM algorithm called the classification EM [12], in which the posterior probabilities are converted to group indicators before the M-step, is equivalent to the K-means algorithm for a Gaussian mixture with $\Sigma_k = \sigma \mathbf{I}$. For further discussion about other estimation approaches besides maximum likelihood via the EM algorithm, see McLachlan and Peel [87].

1.4.3 Estimation issues

Although an estimation tool exists, there are potential issues of parameter identifiability with mixture models. Frühwirth-Schnatter [40] distinguished between three types of nonidentifiability: invariance to relabeling of components, potential overfitting, and nonidentifiability due to the family of component distribution and the covariate design matrix. The first two issues are resolved through constraints such as $\boldsymbol{\theta}_k \neq \boldsymbol{\theta}_{k'}$ for all $k, k' = 1, ..., K$, $k \neq k'$. The last concern is often solved by assuming Gaussian components since finite mixtures of multivariate Gaussians are identifiable [133, 150]. However, Hennig [52] suggested that the introduction of a regression structure to a Gaussian mixture requires a full rank design matrix as well as a rich covariate domain for regression parameters to be identifiable. On the other hand, prior probabilities parameters from a generalized logit are identifiable by setting the parameters of one component to zero such as $\boldsymbol{\gamma}_K = \mathbf{0}$ [61].

Besides identifiability, there are other known issues with finite mixture models. McLachlan and Peel [87] noted that the sample size must be quite large for asymptotic theory to accurately describe the finite sampling properties of the estimator. Also, when component variances are allowed to vary between components, the mixture likelihood function is unbounded, and each observation gives rise to a singularity on the boundary of the parameter space [21, 66]. However, Kiefer [67] outlined theory that guarantees that there exists a particular local maximizer of the mixture likelihood function that is consistent, efficient, and asymptotically normal if the mixture is not overfit. To avoid issues of singularities and spurious local modes in the EM algorithm, Hathaway [50] considered constrained maximum likelihood estimation for multivariate Gaussian mixtures based on the following constraint on the smallest eigenvalue of the matrix $\boldsymbol{\Sigma}_h \boldsymbol{\Sigma}_j^{-1}$, denoted as $\lambda_{\min}(\boldsymbol{\Sigma}_h \boldsymbol{\Sigma}_j^{-1})$,

$$\min_{1 \leq h \neq j \leq K} \lambda_{\min}(\boldsymbol{\Sigma}_h \boldsymbol{\Sigma}_j^{-1}) \geq c > 0$$

for some positive constant $c \in [0, 1]$ to ensure a global maximizer.

1.4.4 Choosing the number of components

Just as with partitioning methods, it is difficult to choose the number of mixture components and the problem has not been completely resolved [87]. A natural approach is to test the hypothesis that $K = K_0$ against the alternative that $K = K_1$ with a likelihood ratio test. However, the regularity conditions do not hold for the test statistic to have its usual asymptotic null distribution. This problem has been considered by many authors (see McLachlan and Peel [87] for a summary) who have suggested modified test statistics [148], approximate null distributions [77], and resampling methods to empirically estimate the null distribution [85].

Another approach is to use information theoretic methods to select a model. A set of candidate models are compared by measuring the information lost when the model is used to approximate the true reality. The model with the lowest information loss and lowest

information criterion is chosen as the best model. Two popular information criterion are the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC).

The information loss can be described in term of the Kullback-Leibler (KL) information [69] of the true distribution with respect to the fitted model. If $g(\mathbf{y})$ is the true density, the KL information of $g(\mathbf{y})$ with respect to the estimated model $f(\mathbf{y}|\hat{\boldsymbol{\theta}})$ is

$$\int \log[g(\mathbf{y})/f(\mathbf{y}|\hat{\boldsymbol{\theta}})]g(\mathbf{y})d\mathbf{y} = \int \log[g(\mathbf{y})]g(\mathbf{y})d\mathbf{y} - \int \log[f(\mathbf{y}|\hat{\boldsymbol{\theta}})]g(\mathbf{y})d\mathbf{y}.$$

Since the first term does not involve the model, only the second term is relevant. When comparing models, this integral can be estimated using the empirical distribution, which places equal mass $1/n$ at each observation $\mathbf{y}_i$,

$$\begin{aligned}\int \log[f(\mathbf{y}|\hat{\boldsymbol{\theta}})]g(\mathbf{y})d\mathbf{y} &\approx \frac{1}{n}\sum_{i=1}^{n} \log f(\mathbf{y}_i|\hat{\boldsymbol{\theta}}) \\ &= \frac{1}{n}\log L(\hat{\boldsymbol{\theta}}).\end{aligned}$$

This is a biased estimate of the expected log density. Akaike [3, 2] showed that the bias is asymptotically equal to d, where d is equal to the number of parameters in the model. Thus, he proposed choosing a model that minimizes

$$AIC = -2\log L(\hat{\boldsymbol{\theta}}) + 2d.$$

The AIC criterion is often used to choose the number of components in a mixture model. However, many authors have observed that it is inconsistent [68] and tends to overestimate the number of components in a mixture [130, 13].

The BIC was originally derived in a Bayesian framework as an approximation to the integrated likelihood [124], which is used in Bayes factors to compare two models. For further information about Bayes factors, see Kass and Raftery [64]. The BIC is calculated as

$$BIC = -2\log L(\hat{\boldsymbol{\theta}}) + d\log n$$

where d is the number of parameters in the model and n is the sample size. The criterion can be justified in a frequentist framework, but the regularity conditions breakdown for mixture models. Nevertheless, Fraley and Raftery [37] noted considerable theoretical and practical evidence to support the use for clustering. Unlike the AIC, the BIC has been shown not to asymptotically underestimate the true number of components [72]. In the context of nonparametric density estimation, Roeder and Wasserman [116] showed that the density estimate chosen with the BIC is consistent. However, if the component densities assumptions do not hold, Biernacki, Celeux, and Govaert [6] found that it tends to fit too many components.

To overcome the shortcomings of the BIC for the mixture model, Biernacki, Celeux, and Govaert [6] proposed the ICL-BIC based on the integrated complete likelihood for a mixture

with K components,

$$ICL - BIC = -2 \log L(\hat{\boldsymbol{\theta}}) + d \log n - 2 \sum_{k=1}^{K} \sum_{i=1}^{n} \hat{z}_{ik} \log \alpha_{ik}$$

where α_{ik} is the posterior probability of individual i belonging to component k, $\hat{z}_{ik}$ is the maximum a posteriori estimate of group membership, d is the total number of parameters, and n is the sample size. The last term is an entropy term based on the strength of the group classification. If the components of the mixture are well separated, then the entropy will be close to zero, but when the components are largely overlapping, the entropy is large and acts as a penalty for having poorly separated clusters. While this criteria has an additional penalty, it is not used as frequently as BIC in practice.

When comparing information criterion between competing models, large differences correspond to strong evidence for one model over the other. Kass and Raftery [64] provide standard conventions for calibrating differences in BIC which can guide the final choice of models. When the differences are small, the smallest number of components that fit the data is chosen for the sake of parsimony.

1.5 Thesis outline

This dissertation is organized in the following manner. In Chapter 2, I develop a series of simulations to study the impact of misspecifying the covariance structure in finite mixture models for longitudinal data and present finite-sample and asymptotic bias results. In Chapter 3, I motivate the rest of the dissertation by discussing the general clustering methodology for longitudinal data and illustrate how these methods do not cluster based on the shape of change over time despite the natural tendency to interpret the results using shape terminology. Chapter 4 presents three proposed methods to cluster based solely on the shape and compares them to standard methods through a simulation study in Chapter 5. Chapter 6 applies these methods to longitudinal growth data of children. In Chapter 7, I overview the results of the dissertation, discuss the contributions, and present areas of future work.

Chapter 2

Covariance misspecification in mixture models

In this chapter, I focus on finite mixture models, the model-based method of clustering longitudinal data. Data are assumed to be generated from a mixture of Gaussian distributions. While no model is perfect, misspecifying the model with simplifying assumptions about the shape of the distributions can cause bias in the estimates. I investigate the impact of incorrectly assuming repeated measures of the same outcome over time are independent.

2.1 Introduction

Multivariate Gaussian mixtures are a class of models that provide a flexible parametric approach for the representation of heterogeneous multivariate outcomes potentially originating from distinct subgroups in the population. An overview of finite mixture models is available in Chapter 1 and in many texts [35, 135, 86, 87, 40]. We can estimate covariate effects on the outcome as well as group membership probabilities by extending mixture models to include a regression structure for both the mean and prior probabilities. See Cruz-Mesía, Quintana, and Marshall [18] for a review of finite mixture models with a regression mean structure and Wedel [143] for a history of concomitant variable models that use baseline variables to explain variation in subgroups. These extensions are used in several medical applications [121] including epidemiology, genomics, and pharmacology in addition to other fields including astronomy, biology, economics, and speech recognition. When the multivariate outcome is a vector of repeated measures taken over time, these methods are identified as group-based trajectory modeling [99, 100] or latent-class growth analysis [95, 94]. See Pickles and Croudace [107] and references within for a review of mixture methods applied to longitudinal data. The use of mixture models for multivariate data is increasing due to computational advances that have made maximum likelihood (ML) parameter estimation possible, via the EM algorithm, through model-specific packages such as Proc Traj in SAS [62], FlexMix [71] and mclust [38] in R, and software such as Mplus [98] and Latent Gold

[139].

Despite the increased use of these models, the sensitivity of estimated regression coefficients to model assumptions has only been explored to a limited degree. In a multivariate mixture model, one must specify the component distribution, the form of the mean, the structure of the covariance matrix, and the number of components; therefore, there are many ways to misspecify the model. For example, in practice, the number of components is unknown and model selection procedures based on the Bayesian information criterion are often employed. However, if the specified covariance structure is too restrictive relative to the truth, the estimated number of components will typically be greater than the true number because more components are needed to model the extra variability. The literature in estimating the number of components is vast [103] and continues to debate this unresolved issue. Owing to the potential complexity of mixture models, simplifying assumptions are made to reduce the dimension of the parameter space, to make estimation possible, and for computational convenience. In particular, many researchers assume Gaussian components and/or restrict the components to have equal variance, both of which are known to result in asymptotic bias if the assumptions are not met [44, 76]. In this chapter, I assume that the number of components, mean structure, and distribution are known and focus on other indeterminacies such as the covariance matrix.

In terms of the covariance matrix, correlation functions [26], eigenvalue and Cholesky decompositions [5, 89], as well as mixed effects structures [97] are used to impose structure and parsimony. Additionally, one common assumption is conditional independence—given the mixture component label, the outcomes for a subject are assumed independent [105, 96]. Of the available software that estimate regression effects for the mean and prior probabilities, most of them make this simplifying assumption. This restriction is convenient when the data are unbalanced or if the sample size is small to make estimation of the covariance parameters more stable. Despite the wealth of proposed covariance models, there has been little work done in the area of mixture models with misspecified covariance structures, and the conditional independence assumption is unlikely to hold in many multivariate data settings, specifically in longitudinal applications. If the mixture consists of one component, the work carried out by Liang and Zeger [75] suggests that regression estimates are asymptotically unbiased. However, these properties do not hold with additional components since estimation includes prior probabilities as well as component parameters.

Here, I investigate the impact of covariance misspecification on ML estimation of parameters and standard errors in multivariate Gaussian mixture models. In particular, our focus is on the assumption of conditional independence for the covariance structure; therefore, we assume the number of components, the distribution, and the mean structure is known. This chapter is organized as follows. Section 2.2 presents the model specification. Section 2.3 describes the estimation procedure, issues, and asymptotic properties of the parameter estimators based on the seminal results of White [146]. In Section 2.4, I present a series of simulations of a simple misspecified example to compare asymptotic and finite-sample bias of parameter and standard error estimates under varying levels of dependence and separation between components. In Section 2.5, I apply these ideas to body mass index data

from a national longitudinal study to demonstrate the effects of misspecification on potential inferences made in practice.

2.2 Model specification

In a finite multivariate mixture, the density of a random vector $\mathbf{y}$ takes the form

$$f(\mathbf{y}) = \pi_1 f_1(\mathbf{y}) + \cdots + \pi_K f_K(\mathbf{y})$$

where $\pi_k > 0$ for $k = 1, ..., K$ and $\sum_k^K \pi_k = 1$. The parameters π_k are prior probabilities, and the functions $f_1, ..., f_K$ are component densities, assumed multivariate Gaussian here.

I extend the general model to allow other factors to affect the mean as well as the prior probabilities. Let $\mathbf{y} \in \mathbb{R}^m$ be a random vector whose distribution, conditional on regression covariates, $\mathbf{x}$, and concomitant variables, $\mathbf{w}$, is a mixture of K Gaussian densities with prior probabilities $\pi_1(\mathbf{w}, \boldsymbol{\gamma}), ..., \pi_K(\mathbf{w}, \boldsymbol{\gamma})$. That is, the conditional mixture density for $\mathbf{y}$ is defined by

$$f(\mathbf{y}|\mathbf{x}, \mathbf{w}, \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k(\mathbf{w}, \boldsymbol{\gamma}) f_k(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}_k) \tag{2.1}$$

where $f_k(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}_k)$ denotes the m-variate Gaussian probability density function with mean $\mathbf{x}\boldsymbol{\beta}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$, $\boldsymbol{\theta}_k$ includes both $\boldsymbol{\beta}_k$ and $\boldsymbol{\Sigma}_k$, $\mathbf{x}$ is a $m \times p$ matrix, and $\mathbf{w}$ is a vector of length q. The regression covariates include measures that affect the mean, whereas the concomitant variables influence the prior probabilities. This general structure allows the possibility that some baseline variables could be in both $\mathbf{x}$ and $\mathbf{w}$.

I parameterize the prior probabilities using the generalized logit model with the form

$$\pi_k(\mathbf{w}, \boldsymbol{\gamma}) = \frac{\exp(\mathbf{w}^T \boldsymbol{\gamma}_k)}{\sum_{j=1}^{K} \exp(\mathbf{w}^T \boldsymbol{\gamma}_j)}$$

for $k = 1, ..., K$ where $\boldsymbol{\gamma}_k \in \mathbb{R}^q$, $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^T, ..., \boldsymbol{\gamma}_K^T)$ where $\boldsymbol{\gamma}_K = \mathbf{0}$.

Throughout this chapter, I generally assume conditional independence with constant variance within a component where $\boldsymbol{\Sigma}_k = \sigma_k^2 \mathbf{I}_m$ as the *proposed estimation model*, but it is straightforward to extend the covariance model to include standard longitudinal correlation structures such as exchangeable or exponential. Therefore, the vector of all unknown parameters $\boldsymbol{\theta}$ consists of the prior probabilities parameters $\boldsymbol{\gamma}_k$ and the component regression and variance parameters $\boldsymbol{\theta}_k^T = (\boldsymbol{\beta}_k^T, \sigma_k^2)$ for $k = 1, ..., K$ and could include correlation parameters.

2.3 Parameter estimation

2.3.1 EM algorithm

Under the assumption that $\mathbf{y}_1, ..., \mathbf{y}_n$ are independent realizations from the mixture distribution, $f(\mathbf{y}|\mathbf{x}, \mathbf{w}, \boldsymbol{\theta})$, defined in equation (2.1), the log-likelihood function for the parameter

vector $\boldsymbol{\theta}$ is given by

$$\log L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(\mathbf{y}_i|\mathbf{x}_i, \mathbf{w}_i, \boldsymbol{\theta}).$$

The ML estimate of $\boldsymbol{\theta}$ is obtained by finding an appropriate root of the score equation, $\partial \log L(\boldsymbol{\theta})/\partial\boldsymbol{\theta} = \mathbf{0}$. Solutions of this equation corresponding to local maxima can be found iteratively through the EM algorithm [25], which is thoroughly described in Section 1.4.2.

2.3.2 Asymptotic properties of estimates

If the true underlying data-generating distribution is a member of the specified model class, then ML estimation via the EM algorithm gives parameter estimates that are consistent [140, 70]. However, if the specified model does not contain the true underlying mixture, then the ML estimators potentially have asymptotic bias [44, 76]. Here, I am interested in the impact of misspecifying the covariance matrix structure on parameter estimation and inference.

General theoretical results for ML estimators are given by White [146]. Our investigation is a special case where the covariance matrices of mixture components are incorrectly specified but the mean structure and distribution are known. Let $f(\mathbf{y}|\boldsymbol{\theta})$ be the assumed estimation model, $g(\mathbf{y})$ be the true density, and C be a compact subset of the parameter space. It follows that the ML estimator $\hat{\boldsymbol{\theta}}_n$ is consistent for the parameter vector $\boldsymbol{\theta}^*$ that minimizes the Kullback-Leibler information $\int \log[g(\mathbf{y})/f(\mathbf{y}|\boldsymbol{\theta})]g(\mathbf{y})d\mathbf{y} = \int \log[g(\mathbf{y})]g(\mathbf{y})d\mathbf{y} - \int \log[f(\mathbf{y}|\boldsymbol{\theta})]g(\mathbf{y})d\mathbf{y}$, under some regularity conditions [146], which is equivalent to maximizing $\int \log[f(\mathbf{y}|\boldsymbol{\theta})]g(\mathbf{y})d\mathbf{y}$ with respect to $\boldsymbol{\theta}$.

In the case of mixture densities, this integral is mathematically intractable. Lo [76] used a modified EM algorithm for univariate data that maximized $\int \log[f(\mathbf{y}|\boldsymbol{\theta})]g(\mathbf{y})d\mathbf{y}$ with respect to $\boldsymbol{\theta}$ in order to estimate $\boldsymbol{\theta}^*$. This procedure could be adapted to bivariate data, but for outcome vectors of larger dimension, this procedure is not as useful. It is known that for outcome vectors $\{\mathbf{y}_i\}_{i=1,\ldots,n}$ generated from the true density, under suitable regularity conditions [59],

$$\sup_{\boldsymbol{\theta}\in C} \frac{1}{n} \sum_{i=1}^{n} \log f(\mathbf{y}_i|\boldsymbol{\theta}) \xrightarrow{a.s.} \sup_{\boldsymbol{\theta}\in C} \int_{-\infty}^{\infty} \log f(\mathbf{y}|\boldsymbol{\theta}) g(\mathbf{y}) d\mathbf{y} \quad \text{as } n \to \infty$$

Therefore, to investigate asymptotic bias under a misspecified covariance structure when $g(\mathbf{y})$ is known, I numerically approximate $\boldsymbol{\theta}^*$ using the EM algorithm on a large sample from $g(\mathbf{y})$ of size $n = 100,000$.

In addition to consistency, White [146] also showed that $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) \to N(0, C(\boldsymbol{\theta}^*))$, where the asymptotic covariance matrix is $C(\boldsymbol{\theta}^*) = A(\boldsymbol{\theta}^*)^{-1}B(\boldsymbol{\theta}^*)A(\boldsymbol{\theta}^*)^{-1}$, with

$$A(\boldsymbol{\theta}^*) = \left\{ E\left(\frac{\partial^2 \log f(\mathbf{y}_i|\boldsymbol{\theta}^*)}{\partial\theta_j \partial\theta_l} \right) \right\}, \quad B(\boldsymbol{\theta}^*) = \left\{ E\left(\frac{\partial \log f(\mathbf{y}_i|\boldsymbol{\theta}^*)}{\partial\theta_j} \cdot \frac{\partial \log f(\mathbf{y}_i|\boldsymbol{\theta}^*)}{\partial\theta_l} \right) \right\}.$$

Moreover, $C_n(\hat{\boldsymbol{\theta}}_n) = A_n(\hat{\boldsymbol{\theta}}_n)^{-1}B_n(\hat{\boldsymbol{\theta}}_n)A_n(\hat{\boldsymbol{\theta}}_n)^{-1} \overset{a.s.}{\to} C(\boldsymbol{\theta}^*)$, with

$$A_n(\hat{\boldsymbol{\theta}}_n) = \left\{ \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \log f(\mathbf{y}_i|\hat{\boldsymbol{\theta}}_n)}{\partial\theta_j \partial\theta_l} \right\}, \quad B_n(\hat{\boldsymbol{\theta}}_n) = \left\{ \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \log f(\mathbf{y}_i|\hat{\boldsymbol{\theta}}_n)}{\partial\theta_j} \cdot \frac{\partial \log f(\mathbf{y}_i|\hat{\boldsymbol{\theta}}_n)}{\partial\theta_l} \right\}.$$

Following a similar procedure as Boldea and Magnus [7], I derive the score vector and Hessian needed to calculate A_n and B_n for a multivariate Gaussian mixture model as specified in this chapter. Derivations are found in Appendix A.

If the model is correctly specified, then both $-A_n(\hat{\boldsymbol{\theta}}_n)^{-1}$ and $B_n(\hat{\boldsymbol{\theta}}_n)^{-1}$ are consistent estimators of $C(\boldsymbol{\theta}^*)$ [146], and two possible variance-covariance estimates for the parameter estimator are

$$\widehat{\mathrm{Cov}}_1(\hat{\boldsymbol{\theta}}_n) = \widehat{\mathbf{W}}_1 = -(nA_n(\hat{\boldsymbol{\theta}}_n))^{-1}$$

and

$$\widehat{\mathrm{Cov}}_2(\hat{\boldsymbol{\theta}}_n) = \widehat{\mathbf{W}}_2 = (nB_n(\hat{\boldsymbol{\theta}}_n))^{-1}$$

On the other hand, $A_n(\hat{\boldsymbol{\theta}}_n)^{-1}B_n(\hat{\boldsymbol{\theta}}_n)A_n(\hat{\boldsymbol{\theta}}_n)^{-1}$ provides a consistent estimator of $C(\boldsymbol{\theta}^*)$ *despite any misspecification*. Therefore, a third and *robust* variance estimate of the parameter estimator is given by

$$\widehat{\mathrm{Cov}}_3(\hat{\boldsymbol{\theta}}_n) = \widehat{\mathbf{W}}_3 = n^{-1}A_n(\hat{\boldsymbol{\theta}}_n)^{-1}B_n(\hat{\boldsymbol{\theta}}_n)A_n(\hat{\boldsymbol{\theta}}_n)^{-1}.$$

I refer to calculated standard error estimates corresponding to these three indexed variance-covariance estimates throughout the rest of the chapter.

2.4 Simulation study

I carry out two series of simulations to examine the behavior of the ML estimators in terms of bias under misspecification of the covariance structure for finite samples from a multivariate mixture. Specifically, I am mainly interested in the impact of dependence in the true error structure on bias in parameter and standard error estimates when the conditional independence is assumed incorrectly and how this is affected by the (i) level of dependence and (ii) the separation between mixture components. Secondly, I investigate the behavior of the bias when the estimation structure gets closer to the true correlation structure by comparing the bias under three correlation structures.

In all of the simulations, data sets with sample size n are generated from an m-variate Gaussian mixture model with parameters $(\boldsymbol{\gamma}_k, \beta_k, \sigma_k^2, \mathbf{R}_k)$ for $k = 1, ..., K$ where $\mathbf{R}_k$ is the true correlation structure as follows:

- Fix K.
- For each subject, $i = 1, ..., n$,

- Fix $\mathbf{x}_i = \mathbf{1}_m$ and $\mathbf{w}_i = 1$.
- Construct matrices $\mathbf{A}_k$ such that $\mathbf{A}_k\mathbf{A}_k^T = \mathbf{R}_k$ for $k = 1, .., K$ using the Cholesky decomposition.
- Randomly assign group membership, h_i, by drawing a value from the categorical distribution defined by $P(h = k) = \pi_k(\mathbf{w}_i, \boldsymbol{\gamma})$ for $k = 1, ..., K$.
- Draw m standard normal random values $\mathbf{e}_i$ and let

$$\mathbf{y}_i = \mathbf{x}_i\beta_{h_i} + \sigma_{h_i}\mathbf{A}_{h_i}\mathbf{e}_i$$

Thus, $\mathbf{y}_i \sim N(\mathbf{x}_i\beta_{h_i}, \sigma^2_{h_i}\mathbf{R}_{h_i})$. I then estimate the parameters and standard errors, $\widehat{SE}_1$, $\widehat{SE}_2$, $\widehat{SE}_3$, using constrained maximum likelihood via the EM algorithm [50] doing five random initializations, on the basis of a multivariate mixture model with a specified correlation structure and known design matrix.

For simplicity, I focus on an example of two Gaussian components ($K = 2$) with constant mean vectors (i.e. no relationship between covariates and $\mathbf{y}$), one component with independent errors, the other with some level of dependency in the errors. For the first series, the latter dependence is based on an exchangeable correlation structure where all outcomes in an observational unit are equally correlated, which is mathematically equivalent to a random intercept model if the correlation is positive.

To investigate the influence of the level of dependence, I set the vector length to $m = 5$, equal prior probabilities ($\gamma_1 = 0$; baseline variables have no effect), mean of the components to $\beta_1 = 1$ and $\beta_2 = 3$, and the variance of the components to $\sigma_1^2 = 0.25$ and $\sigma_2^2 = 1$. The errors are independent ($\mathbf{R}_1 = \mathbf{I}_m$) in component one and we let the level of dependence vary with $\rho = 0, 0.5, 0.99$ within the exchangeable structure ($\mathbf{R}_2 = \rho(\mathbf{1}_m\mathbf{1}_m^T - \mathbf{I}_m) + \mathbf{I}_m$ where $\mathbf{1}_m$ is a m-length vector of 1's) for component two. I present the bias of parameter estimates and the three standard error estimates under these conditions.

Then, I consider the separation between two component distributions using the concept of c-separation [20]. Two Gaussian distributions, $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, in $\mathbb{R}^m$ are c-separated if $||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||_2 \geq c\sqrt{m \cdot \max(\lambda_{max}(\boldsymbol{\Sigma}_1), \lambda_{max}(\boldsymbol{\Sigma}_2))}$ where $\lambda_{max}(\boldsymbol{\Sigma})$ is the largest eigenvalue of $\boldsymbol{\Sigma}$. Dasgupta [20] notes that two Gaussians are almost completely non-overlapping when $c = 2$. This inequality can be rearranged to establish a measure of separation,

$$S = ||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||_2 / \sqrt{m \cdot \max(\lambda_{max}(\boldsymbol{\Sigma}_1), \lambda_{max}(\boldsymbol{\Sigma}_2))},$$

which is a standardized Euclidean distance between mean vectors. In this simulation, I calculate the value of S for data-generating component densities as a measure of the separation between the two components and if $S > 2$, the components do not overlap and are well separated. For this series of simulations, I again use a strong level of dependence ($\rho = 0.99$) in the exchangeable structure, a vector length of $m = 5$, but vary the mean and variance of the second component ($\beta_2 = 3, 5$ and $\sigma_2^2 = 0.25, 1, 4$) to invoke different degrees of separation between components.

I perform 1000 replications of each simulation for sample sizes $n = 100, 500, 1000$. I approximate the true standard error with the standard deviation of the replicates. To estimate the asymptotic bias of the model parameters ($n = \infty$), I complete one replication with $n = 100,000$.

The two prong simulation described above focuses on the impact of using a conditional independence estimation model under different levels of dependence and separation in the data-generating components. In practice, I can choose correlation structures other than conditional independence. To explore the bias under difference covariance assumptions, I run a short simulation adjusting the data-generating model from above to use an exponential correlation structure, such that the dependence decreases as the time lag increases, rather than the constant dependence from the exchangeable structure. Therefore, for component two, the correlation between two measurements within a subject that are observed d time units apart is $\exp(-d/r)$ where r, the range parameter, determines how quickly the correlation decays to zero. This structure is general enough so that if r is close to zero, the correlation matrix is close to conditional independence and if r is very large, the structure is close to exchangeable correlation with strong dependence.

For this simulation, I continue using the two Gaussian components ($K = 2$) with constant mean vectors ($\beta_1 = 1$ and $\beta_2 = 3$) of length $m = 5$ with observations at times $t = 1, 2, 3, 4$ and 5, one component with independent errors, and the second component with a moderate level of dependence that decays exponentially ($r = 3$). I estimate the prior probability, mean, and variance parameters assuming different correlation structures: conditional independence, exchangeable, and exponential correlation. I estimate and compare the finite-sample bias of parameters and standard errors by letting $n = 500$ with 1000 replications. Additionally, I compare the conventional estimate $\widehat{\mathbf{W}}_1$ of the covariance matrix of $\widehat{\boldsymbol{\theta}}$ and the robust estimate $\widehat{\mathbf{W}}_3$. If the estimation model is close to the true structure, the matrices should be similar and $\mathbf{Q} = \widehat{\mathbf{W}}_1^{-1}\widehat{\mathbf{W}}_3$ should be close to the identity matrix. I calculate $RJ = tr(\mathbf{Q})/\nu$ where ν is the length of $\widehat{\boldsymbol{\theta}}$, which has been termed the RJ criteria and should be close to 1 if the estimation model is close to the truth [127, 117].

2.4.1 Results

Table 2.1 lists bias estimates for the dependence-varying simulation study. The estimates range from close to zero when $\rho = 0$ to magnitudes of upwards of 0.3 when $\rho = 0.99$. It is clear from this table that stronger dependence in the errors results in greater finite-sample and asymptotic bias when estimating under the conditional independence assumption. Additionally, the magnitude of bias seems to reach the asymptotic levels at sample sizes of $n = 500$, but it is important to note that the estimates for the asymptotic bias, based on one replication with $n = 100,000$, are only numerically accurate to two decimal places for γ_1, σ_1^2, and σ_2^2. We see this numerical inaccuracy when $\rho = 0$ since the asymptotic bias should be zero when the conditional independence assumption is met. In terms of standard error estimates, the bias increases with increased dependence with values ranging from 0.001 when $\rho = 0$ to 0.111 when $\rho = 0.99$. We see a divergence between the three variance estimators

with $\widehat{SE}_3$, the robust estimator, consistently having the least bias (Table 2.2). When the model is correctly specified with $\rho = 0$, the three estimators are similar as supported by asymptotic theory.

			Bias Estimates		
n	$\widehat{\gamma}_1$	$\widehat{\beta}_1$	$\widehat{\sigma}_1^2$	$\widehat{\beta}_2$	$\widehat{\sigma}_2^2$
$\rho = 0.00$					
100	-0.004 (0.006)	0.000 (0.001)	-0.000 (0.001)	-0.002 (0.002)	-0.006 (0.003)
500	0.005 (0.003)	-0.000 (0.000)	-0.000 (0.000)	0.001 (0.001)	-0.000 (0.001)
1000	0.004 (0.002)	-0.000 (0.000)	-0.000 (0.000)	0.001 (0.001)	-0.000 (0.001)
∞	-0.001	0.000	0.001	-0.000	0.002
$\rho = 0.50$					
100	0.125 (0.006)	0.031 (0.001)	0.024 (0.001)	0.106 (0.003)	-0.136 (0.003)
500	0.125 (0.003)	0.028 (0.001)	0.024 (0.000)	0.101 (0.002)	-0.124 (0.001)
1000	0.125 (0.002)	0.028 (0.000)	0.024 (0.000)	0.098 (0.001)	-0.125 (0.001)
∞	0.115	0.027	0.024	0.095	-0.125
$\rho = 0.99$					
100	0.370 (0.007)	0.087 (0.002)	0.033 (0.001)	0.327 (0.005)	-0.410 (0.005)
500	0.346 (0.003)	0.078 (0.001)	0.030 (0.001)	0.310 (0.002)	-0.388 (0.002)
1000	0.350 (0.002)	0.079 (0.001)	0.030 (0.000)	0.309 (0.001)	-0.385 (0.001)
∞	0.353	0.082	0.031	0.315	-0.383

Table 2.1: Bias estimates (SE) of maximum likelihood parameter estimators when the covariance structure of a two-component Gaussian mixture is assumed to be conditionally independent based on 1000 replications under each mixture distribution with $m = 5$, $\gamma_1 = \gamma_2 = 0$, $\beta_1 = 1$, $\mathbf{R}_1 = \mathbf{I}_m$, $\sigma_1^2 = 0.25$, $\beta_2 = 3$, $\mathbf{R}_2 = \mathbf{R}(\rho)$ and $\sigma_2^2 = 1$ where $\mathbf{R}(\rho)$ is the exchangeable correlation matrix with parameter ρ. Asymptotic estimates ($n = \infty$) are based on one replication with $n = 100,000$. Values equal to zero represent values less than 0.001.

Bias Estimates

ρ	n	$\widehat{SE}_1(\widehat{\gamma}_1)$	$\widehat{SE}_2(\widehat{\gamma}_1)$	$\widehat{SE}_3(\widehat{\gamma}_1)$	$\widehat{SE}_1(\widehat{\beta}_1)$	$\widehat{SE}_2(\widehat{\beta}_1)$	$\widehat{SE}_3(\widehat{\beta}_1)$	$\widehat{SE}_1(\widehat{\sigma}_1^2)$	$\widehat{SE}_2(\widehat{\sigma}_1^2)$	$\widehat{SE}_3(\widehat{\sigma}_1^2)$
0.00	100	0.007	0.007	0.007	0.000	0.001	-0.000	-0.001	0.001	-0.001
	500	0.002	0.002	0.002	-0.000	0.000	-0.000	0.000	0.000	0.000
	1000	0.001	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.000
0.50	100	0.004	0.004	0.004	-0.006	-0.008	-0.002	-0.007	-0.009	-0.004
	500	0.002	0.002	0.002	-0.003	-0.004	-0.001	-0.002	-0.004	-0.001
	1000	-0.001	-0.001	-0.001	-0.002	-0.002	-0.000	-0.001	-0.002	-0.000
0.99	100	-0.003	-0.003	-0.002	-0.026	-0.035	-0.010	-0.013	-0.016	-0.008
	500	-0.005	-0.005	-0.004	-0.011	-0.015	-0.003	-0.005	-0.007	-0.002
	1000	-0.003	-0.003	-0.002	-0.007	-0.010	-0.001	-0.003	-0.005	-0.002

ρ	n	$\widehat{SE}_1(\widehat{\beta}_2)$	$\widehat{SE}_2(\widehat{\beta}_2)$	$\widehat{SE}_3(\widehat{\beta}_2)$	$\widehat{SE}_1(\widehat{\sigma}_2^2)$	$\widehat{SE}_2(\widehat{\sigma}_2^2)$	$\widehat{SE}_3(\widehat{\sigma}_2^2)$
0.00	100	0.001	0.003	0.001	-0.008	-0.004	-0.010
	500	0.001	0.001	0.001	0.001	0.001	0.000
	1000	0.000	0.000	0.000	0.001	0.001	0.001
0.50	100	-0.046	-0.070	-0.005	-0.019	-0.028	-0.004
	500	-0.021	-0.032	-0.002	-0.010	-0.016	-0.001
	1000	-0.014	-0.021	-0.000	-0.009	-0.013	-0.002
0.99	100	-0.092	-0.122	-0.017	-0.085	-0.111	-0.022
	500	-0.040	-0.054	-0.006	-0.035	-0.049	-0.003
	1000	-0.027	-0.037	-0.002	-0.026	-0.036	-0.004

Table 2.2: Bias estimates of the three standard error estimators (SE_1, SE_2, SE_3) when the covariance structure of a two-component Gaussian mixture is assumed to be conditionally independent based on 1000 replications under each mixture distribution with $m = 5$, $\gamma_1 = \gamma_2 = 0$, $\beta_1 = 1$, $\mathbf{R}_1 = \mathbf{I}_m$, $\sigma_1^2 = 0.25$, $\beta_2 = 3$, $\mathbf{R}_2 = \mathbf{R}(\rho)$ and $\sigma_2^2 = 1$ where $\mathbf{R}(\rho)$ is the exchangeable correlation matrix with parameter ρ. Approximate standard error is based on the estimated standard deviation of the simulation distribution. Values equal to zero represent values less than 0.001.

Figure 2.1 shows that the relationship between the level of component separation and the magnitude of bias is complex. As in the previous simulation, sample sizes of $n = 500$ and larger produce similar bias estimates so we only present the asymptotic results. When the level of separation is high, $S > 2$, then the magnitude of the bias is small, but when there is some overlap, $S < 2$, there is not a clear, consistent relationship between the value of S and the magnitude of the estimated bias for all parameters. That is, for two sets of parameter values, such as $(\beta_2 = 3, \sigma_2^2 = 0.25)$ and $(\beta_2 = 5, \sigma_2^2 = 1)$, that have the same level of separation, $S = 2.836$, the magnitude of the bias for all parameter estimates is drastically different for the two settings. However, in general, the bias decreases as the level of separation increases for a fixed mean parameter. The only exception is that the estimator for the first component mean $(\widehat{\beta}_1)$ has increased bias when $S = 1.418$ as compared to $S = 0.709$, but the bias then decreases when $S = 2.836$. It appears that when there is high overlap between two components, there is a point at which the bias peaks and then starts to decrease as σ_2 increases even though the amount of overlap continues to increase. Lastly, similar to the parameter estimates, the greater amount of separation results in less bias in the standard errors with biases as large as 1.0 unit in the situation with the most overlap and as little as less than 0.001 when components are well separated. Again, the robust estimator again has the lowest bias.

The simulations based on dependence and separation demonstrate the finite-sample and asymptotic bias in the ML estimators when the covariance structure is misspecified as conditional independence and the mixture components overlap. However, if two components are well separated, the misspecification of the dependence in the errors does not result in large biases and thus any finite-sample bias could be removed potentially conventional techniques such as bootstrapping with careful tracking of component labels [45]. Additionally, when there is no covariance misspecification or when components are well separated, all of the standard error estimates are similar and have little bias. However, when there is misspecification in the dependence structure, the estimates basely solely on the Hessian matrix or the score vector understate the true variability while the robust estimate has little bias. In cases where the true level of dependence is high, the bias in the Hessian estimator, $\widehat{SE}_1$, and the robust version, $\widehat{SE}_3$, can differ by as much as a relative factor of 2. In simulations not shown, using unequal prior probabilities result in similar conclusions. When the component proportions are unbalanced, the magnitude of bias increases when a majority of observations units originate from the misspecified component (here component 2).

Figure 2.2 shows the absolute bias of parameter estimates under the three different covariance assumptions when the data was generating with the exponential correlation structure for component 2. As expected, when the model is correctly specified, there is very little bias. I note that assuming the exchangeable structure, while incorrect, results in less bias than assuming conditional independence. As the RJ criteria gets closer to 1 from 1.97 to 1.02 to 0.99 using independence, exchangeable, and then exponential, the model structure gets closer to the true structure resulting in little bias in the parameter estimates.

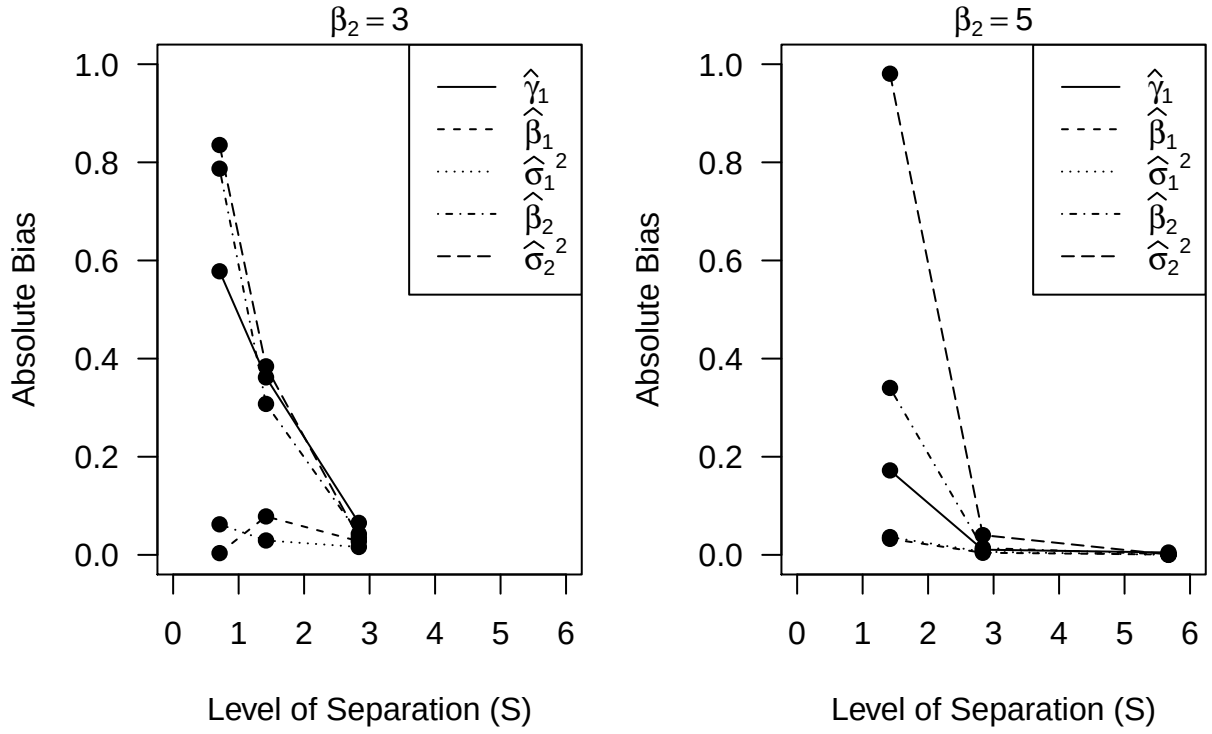

Figure 2.1: Asymptotic bias estimates of maximum likelihood parameter estimators when the covariance structure of a two-component Gaussian mixture is assumed to be conditionally independent based on one replication with $n = 100,000$ under each mixture distribution with $m = 5$, $\gamma_1 = \gamma_2 = 0$, $\beta_1 = 1$, $\mathbf{R}_1 = \mathbf{I}_m$, $\sigma_1^2 = 0.25$, $\mathbf{R}_2 = \mathbf{R}(\rho)$, and $\rho = 0.99$ where $\mathbf{R}(\rho)$ is the exchangeable correlation matrix with parameter ρ. The level of separation (S) is calculated using the true mixture distribution. For $\beta_2 = 3$, variance parameters, $\sigma_2^2 = 0.25, 1, 4$ result in $S = 2.836, 1.418, 0.709$, respectively. For $\beta_2 = 5$, variance parameters, $\sigma_2^2 = 0.25, 1, 4$ result in $S = 5.671, 2.836, 1.418$, respectively. Values of $S \geq 2$ indicate almost completely separated components.

2.5 Body mass index data example

To look at the behavior of the parameter and standard error estimates in practice, we use data from the 1979 National Longitudinal Survey of Youth (NLSY79). The NLSY79 is a nationally representative sample of 12,686 young American men and women aged 14-22 years in 1979. The cohort, interviewed annually from 1979 to 1994 and biennially thereafter, has provided health and economic data for a total of 23 interviews (until 2008). In particular, the available body weight data for the 1979 cohort span a twenty-five year period [105]. We study body mass index (BMI) over time as it is an important longitudinal measure for public health and elucidating obesity development. Self-reported weight was collected in 17

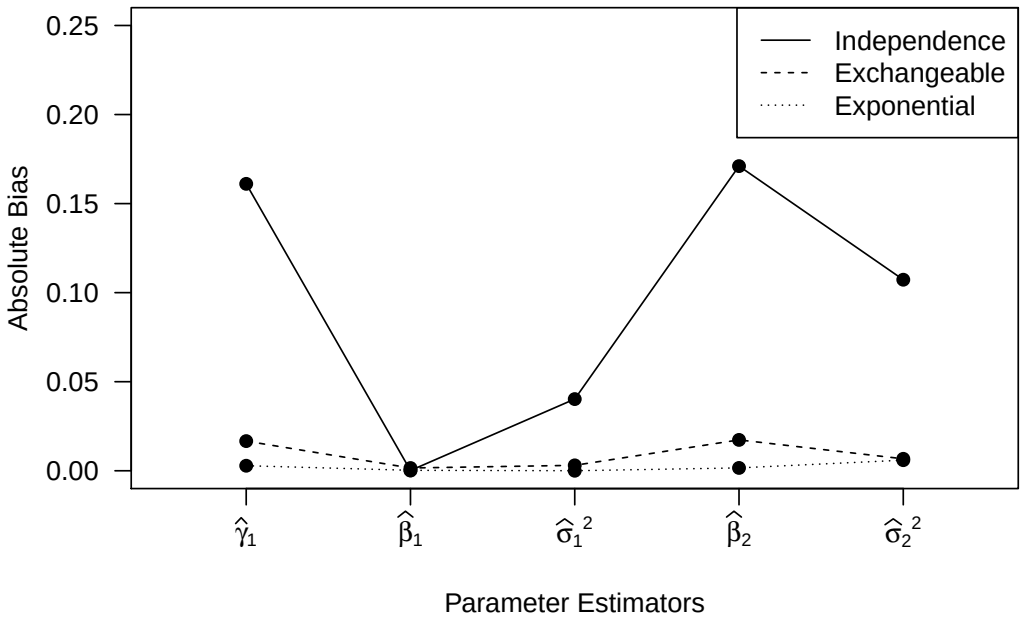

Figure 2.2: Bias estimates of maximum likelihood parameter estimators when the covariance structure of a two-component Gaussian mixture is assumed to be conditionally independent, exchangeable, and exponential structure based on 1000 replications under each mixture distribution with $n = 500$, $m = 5$, $\gamma_1 = \gamma_2 = 0$, $\beta_1 = 1$, $\mathbf{R}_1 = \mathbf{I}_m$, $\sigma_1^2 = 0.25$, $\beta_2 = 3$, $\mathbf{R}_2 = \mathbf{R}(r)$ and $\sigma_2^2 = 2$ where $\mathbf{R}(r)$ is the exponential correlation matrix based on the observation times $(1, 2, 3, 4, 5)$ and $r = 3$. Mean values of the RJ criteria are $RJ = 1.97, 1.02, 0.99$ for the three covariance assumptions, respectively.

interviews and height in five of those. BMI [weight (kg)/height^2 (m^2)] was calculated for each interview based on the weight and the average height.

For the purposes of this chapter, the complex sampling structure is ignored and we randomly sample 500 subjects who were at least 18 years of age in 1981 and reported all 17 weight measurements. Of this sample, 51% are female, 54% are non-Hispanic/non-Black, 29.2% Black and 16.8% Hispanic. To model the BMI outcomes, I allow a quadratic relationship between mean BMI and age and include sex as a baseline concomitant variable. Therefore, for $i = 1, ..., 500$, we assume that the observed data were generated according to

$$BMI_i = \beta_{k0} + \beta_{k1} \cdot (AGE_i - 18) + \beta_{k2} \cdot (AGE_i - 18)^2 + \epsilon_i$$

with probability

$$\pi_k(\mathbf{w}, \boldsymbol{\gamma}) = \frac{e^{\gamma_{k0} + I(MALE_i)\gamma_{k1}}}{\sum_j e^{\gamma_{j0} + I(MALE_i)\gamma_{j1}}}$$

where $\epsilon_i \sim N(0, \sigma_k^2 \mathbf{R}_{ik})$ for $k = 1, ..., 4$. The choice of four groups is based on previous research [105]. Using the EM algorithm with five random initializations, we estimate parameters and standard errors and present the estimates that produced the highest log-likelihood. For the sake of comparison, we complete the estimation assuming conditional independence ($\mathbf{R}_{ik} = \mathbf{I}_m$), and under an exchangeable ($\mathbf{R}_{ik} = \rho_k(\mathbf{1}_m \mathbf{1}_m^T - \mathbf{I}_m) + \mathbf{I}_m$) and exponential ($[\mathbf{R}_{ik}]_{jl} = \exp(-d_{ijl}/r_k)$ where d_{ijl} is the Euclidean distance between the ages at the jth and lth interviews for subject i) correlation model.

2.5.1 Results

Parameter and standard errors are estimated for a four-component multivariate Gaussian mixture model assuming conditional independence, exchangeable, and exponential correlation (Table 2.3). The regression parameter estimates are used to calculate the mean curves for the four groups under all three covariance assumptions, and we see that the mean curves differ between the models mainly in terms of the innermost curves (Figure 2.3). Under exchangeable correlation, one of the middle curves represents little BMI increase over time in contrast to the other groups. Under the exponential correlation assumption, the two lowest groups have a similar pattern over time, but the dependence differs between these groups with the range parameters estimated as $r_1 = 2.973$ and $r_2 = 23.579$ indicating that component 2 has more long range dependence between the BMI outcomes than component 1. Our simulation results suggest the magnitude of bias in the parameter estimator depends on how close the assumed correlation structure is to the truth and the overlap between components. I note there are no well-separated components and we see bias in the mean estimates by comparing the three covariance assumptions.

Given that the repeated outcome is BMI, we expect some dependence in the error structure within individuals. I consider the level of dependence in errors by plotting the estimated autocorrelation function by calculating the empirical variogram of the residuals from the conditional independence model [26] for each estimated component by randomly assigned each individual to a component using posterior probabilities [141]. The estimated autocorrelation function of the residuals shows strong dependence between residuals within 5 to 10 years and the correlation decreases with increasing time lags (Figure 2.4). This correlation structure is therefore neither consistent with conditional independence nor exchangeable correlation, but rather decreases to zero which is more consistent with the exponential correlation structure. We see that the robust standard estimators are almost twice those of estimates using the standard estimators under conditional independence, and the RJ criteria, which compares the conventional and robust estimates of the covariance matrix the parameters, suggests that the exponential correlation structure is the one closest to the truth.

In this data example, we see the influence of the covariance structure on the estimates, especially in terms of the regression parameters. With the simulation results and the RJ criteria, we expect the exponential correlation model fits the data the best out of the three structures. However, I note that I fixed the number of components to be four for the sake of consistency and this may not be the optimal number of components. In practice, this

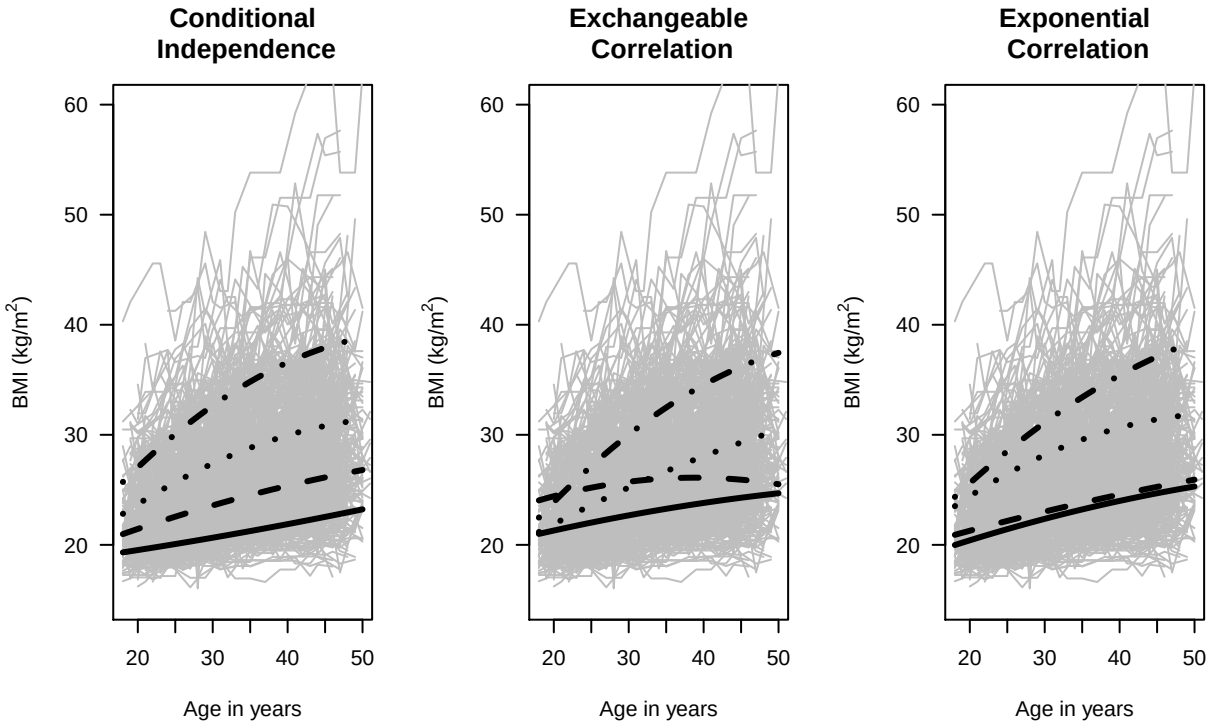

Figure 2.3: Random sample of 500 body mass index (BMI) trajectories from NLSY79 and mean curves for the four components estimated using a Gaussian mixture model specified with a quadratic mean under the covariance assumptions: conditional independence, exchangeable, and exponential correlation. The labeled are consistent with the tables in the text: component 1 (solid), component 2 (dashed), component 3 (dotted), and component 4 (dashed-dot). Additionally, the RJ criteria was calculated for each covariance assumption: $RJ = 7.34, 3.02, 2.22$ under conditional independence, exchangeable, and exponential, respectively.

value is estimated from the data as mentioned earlier. This data application highlights the impact of covariance misspecification as well as the fact that the mean structure may not be the only aspect differing between individuals; the level of dependence and variability also distinguish groups of individuals.

2.6 Discussion

I have shown that covariance misspecification of a two-component Gaussian mixture may produce very little bias in regression and prior probability parameter estimates when the components are well separated. This is well aligned with Lo's univariate findings [76]. However, when there is some overlap in the component distributions, assuming the wrong correlation

structure can produce asymptotically biased parameter estimates, the magnitude dependent on the level of separation and how close the structure is to the truth. With misspecified mixture models, the potential for biased prior probabilities and regression parameters estimates differs from the one-component models for which general estimating equations [75] produce unbiased estimates despite dependence present in the errors. Depending on the context and precision of the estimates, the bias may or may not have practical significance, but it is important to note that the ML estimators are inconsistent under covariance misspecification and there may be substantial bias when the components are not well separated.

In addition to potential biases in the parameter estimates, the simulations provide evidence that conventional standard errors estimates that are based solely on the score equation, or the Hessian, can be extremely biased and underestimate the true variability of the estimates when the covariance structure is misspecified. Therefore, standard errors should be robustly estimated using White's estimator that sandwiches the two conventional estimators. I use the exact formula for this estimator since the numerical approximations to the Hessian matrix and score vector are not by-products of the EM algorithm. To my knowledge, very few software programs automatically use a robust standard error estimator, but it should be implemented in every mixture model software as the default variance estimator and presented along with standard estimators to allow for comparisons, calculation of the RJ criteria, and the detection of misspecification bias.

Given our results, I recommend three things when estimating parameters in a mixture model. First, count the number of subjects whose maximum posterior probability is less than 0.95. If this count is non-zero, this indirectly indicates that the component distribution are not well separated, suggesting that specifying the correct correlation structure is important. Second, if the components are not well separated, fit the mixture model using several correlation structures such as conditional independence, exchangeable, and exponential correlation. For each model, calculate the RJ criteria based on the conventional and robust estimated variance-covariance matrix. Compare the parameter estimates to see if they change under the different assumptions and assess the RJ criteria values to see which structure results in a value closest to one. Choose the most parsimonious model that has an RJ criteria value close to one. Third, if none of these three structure fulfills this requirement, consider a more complex, potentially non-stationary covariance matrix as well as other sources of misspecification such as an incorrect number of components, assumed distribution, or an inflexible mean structure.

This simulation study is limited, but the results likely apply to more complex mean structures and a larger number of components. In future studies, the impact of bias should be explored for more than two components with all components potentially having a misspecified covariance structure and for non-stationary covariance structures. Additionally, mixture models as specified in this chapter group individuals with similar trajectories over time; in the next chapter, I discuss methods that distinguish between the shape of the trajectory and the vertical level of the curve when grouping individuals together.

	Independence			**Exchangeable**			**Exponential**		
	Estimate	$\widehat{SE}_1$	$\widehat{SE}_3$	Estimate	$\widehat{SE}_1$	$\widehat{SE}_3$	Estimate	$\widehat{SE}_1$	$\widehat{SE}_3$
γ_{10}	0.83	0.20	0.46	0.12	0.18	0.26	0.42	0.23	0.44
γ_{11}	-1.80	0.38	0.57	0.97	0.30	0.43	-0.89	0.73	2.28
β_{10}	19.32	0.14	0.18	20.99	0.21	0.22	19.98	0.36	0.47
β_{11}	0.10	0.02	0.03	0.16	0.01	0.02	0.22	0.05	0.06
β_{12}	0.00	0.00	0.00	-0.00	0.00	0.00	-0.00	0.00	0.00
σ_1^2	2.97	0.11	0.28	6.33	0.72	1.00	5.72	0.43	0.98
ρ_1	-	-	-	0.85	0.02	0.02	-	-	-
r_1	-	-	-	-	-	-	2.97	0.21	0.28
γ_{20}	0.60	0.21	0.44	-0.45	0.22	0.27	0.72	0.22	0.52
γ_{21}	0.44	0.28	0.39	0.85	0.34	0.46	1.42	0.51	1.69
β_{20}	20.98	0.12	0.31	24.04	0.52	0.69	20.91	0.24	0.39
β_{21}	0.24	0.02	0.02	0.20	0.03	0.07	0.19	0.02	0.03
β_{22}	-0.00	0.00	0.00	-0.00	0.00	0.00	-0.00	0.00	0.00
σ_2^2	3.15	0.10	0.37	21.11	2.93	4.72	8.15	0.79	2.46
ρ_2	-	-	-	0.81	0.03	0.02	-	-	-
r_2	-	-	-	-	-	-	23.58	2.02	4.21
γ_{30}	0.28	0.22	0.48	0.13	0.19	0.28	0.17	0.25	0.53
γ_{31}	0.53	0.30	0.40	1.07	0.30	0.42	1.74	0.51	1.58
β_{30}	22.82	0.21	0.72	21.16	0.28	0.37	23.52	0.39	0.56
β_{31}	0.45	0.03	0.05	0.37	0.02	0.03	0.46	0.05	0.09
β_{32}	-0.01	0.00	0.00	-0.00	0.00	0.00	-0.01	0.00	0.00
σ_3^2	6.42	0.26	1.16	11.24	1.23	2.23	13.05	1.00	2.85
ρ_3	-	-	-	0.82	0.02	0.02	-	-	-
r_3	-	-	-	-	-	-	10.22	0.68	0.83
γ_{41}	0	-	-	0	-	-	0	-	-
γ_{40}	0	-	-	0	-	-	0	-	-
β_{40}	25.71	0.48	0.67	22.47	0.57	0.48	24.36	1.05	0.83
β_{41}	0.68	0.07	0.11	0.72	0.04	0.08	0.64	0.14	0.10
β_{42}	-0.01	0.00	0.00	-0.01	0.00	0.00	-0.01	0.00	0.00
σ_4^2	26.86	1.15	7.89	32.98	3.48	6.93	44.79	4.09	13.76
ρ_4	-	-	-	0.68	0.03	0.04	-	-	-
r_4	-	-	-	-	-	-	8.06	0.65	0.85

Table 2.3: Parameter and standard error estimates ($\widehat{SE}_1, \widehat{SE}_3$) for a random sample of 500 from NLSY79 assuming a four-component mixture model with quadratic mean and the following correlation structures: conditional independence, exchangeable, and exponential correlation. Values equal to zero represent values less than 0.01. Additionally, the RJ criteria was calculated each covariance assumption: $RJ = 7.34, 3.02, 2.22$ under conditional independence, exchangeable, and exponential, respectively.

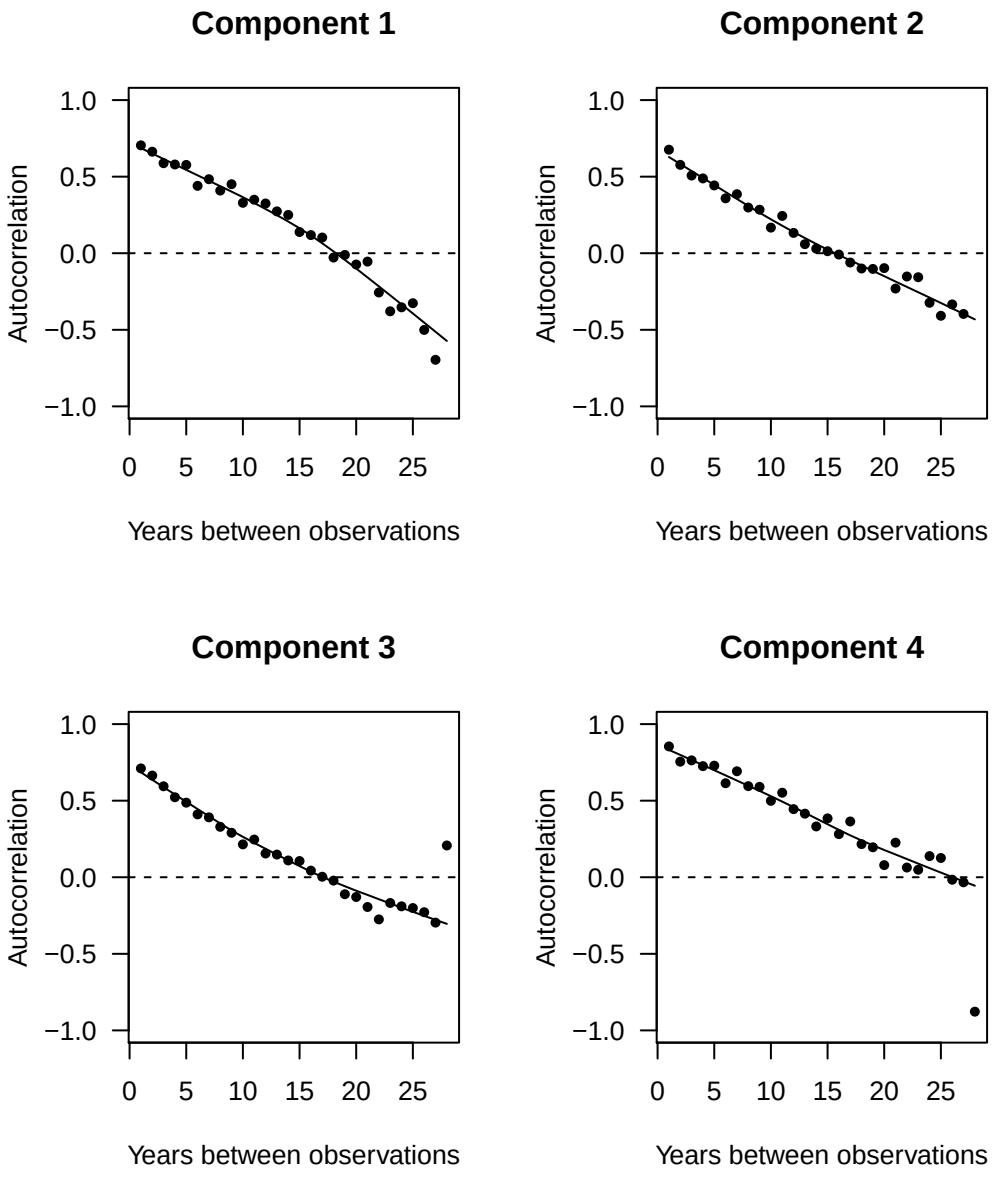

Figure 2.4: Smoothed sample autocorrelation of component residuals of estimated Gaussian mixture model specified with a quadratic mean and conditional independence with a random sample of 500 body mass index trajectories from NLSY79 randomly assigned to components based on estimated posterior probabilities.

Chapter 3

Shape-based clustering

Longitudinal data is collected with the intent of studying the change over time. One side effect of collecting this type of data is the inherent dependence between the repeated measures. The last chapter provided evidence that this nuisance cannot be ignored when clustering data using a finite mixture model. In this chapter, I discuss clustering methods with the research goals of longitudinal studies in mind.

3.1 Motivation

There is no one way to cluster data. There are many characteristics of quantitative data on which to base a grouping. This is not only true for data sets. We constantly group products, people, and projects based on different qualities and characteristics. However, we tend to forget this subjective choice when faced with a data set filled with continuous measurements. Many automatically use squared Euclidean distance to determine similarity. This may work adequately when the data are organized into vectors of independent explanatory factors all of which are equally important features of the object or subject. However, when the vector has time-ordered structure, a dissimilarity measure that is based only the element-wise differences between vectors ignores this important structural property of the data.

There has been some work in adapting and developing clustering methods for longitudinal data in the fields of behavior science and genomics [122, 43, 62, 98, 89]. Most of these methods involve fitting a finite mixture model with adjustments to the mean or covariance to make it more adaptable to longitudinal data. Others cluster subjects with a partition algorithm using a dissimilarity measure developed specifically to take the time-ordered structure of longitudinal data into account. Applying cluster methods to longitudinal data has gained popularity in other areas of application. For example, there is a growing literature about clustering distinct childhood body mass index growth patterns over time with a view to determining life factors that contribute to patterns [109, 10].

The main goal of a longitudinal study is to study individual change over time. However,

none of the standard clustering methods adapted for longitudinal data explicitly group subjects based on the shape of the pattern over time. Although they claim to group similar patterns, the similarity is not defined in terms of any one specific feature of the data but rather just in terms of the difference in the vector elements. In the time series literature, Wang, Smith, and Hyndman [142] suggested estimating and using global features of time series processes such as the trend, seasonality, autocorrelation, and kurtosis to define similarity when clustering. Some of these features apply to longitudinal data. The trend over time in particular is of considerable interest. We go a step further and break trend into two components: level and shape.

In many circumstances, these two characteristics provide distinct information that can elicit different actions. For example, when investing in stocks, the magnitude of the share price provides information about the value and financial practices of a company. While that information is important, the feature that may drive a decision to buy or sell shares of a stock is the historical change in price over time. Similarly, a high average body mass index triggers concerns about health, but knowing how that value is changing over time indicates whether the problem is improving or worsening and if an intervention may be necessary. When studying gender inequality in salaries, it is important to differentiate between discrimination in starting salaries and systematic discrimination over time. I distinguish between these two features of a trajectory when thinking about clustering as they provide different information.

Despite the interest in the shape of the change over time, too many researchers apply a standard clustering method without thought on the interpretation of their results. Not only are the groups driven by the level rather than the shape of the trajectories, but the means within a given cluster may not be representative of any specific individual's trajectory since it may be an average of trajectories with different shapes at the same level. It is common for authors applying these standard clustering methods to make incorrect conclusions about shape [147, 93, 9, 109, 84].

Imagine measuring alcohol consumption over time for a sample of young adults. This cohort may contain individuals who never drink as well as those that have moderate or heavy drinking habits. Besides different levels of alcohol consumption, there may be a variety of behavior patterns over time such as escalation, reduction, or stabilization. All of these patterns can occur at each level of alcohol consumption. Figure 3.1 illustrates a possible scenario of four individuals, two of who drink heavily and two of who are low to moderate drinkers. Within both levels, one of the individuals has escalating behavior and the other is reducing alcohol consumption over time. We assume linear change for simplicity. The shape of the curve, the slope in this case, is independent of the consumption level. If K-means is applied to the observed vectors, two clusters are discovered that are determined by the level of consumption with the two heavy drinkers grouped together and the two moderate consumers in another group. Consequently, the mean trajectory for each group is a horizontal line, which disguises the fact that the alcohol consumption is not stable for any of the individuals. Knowing what factors impact the alcohol consumption level is important for public health, but the knowledge of what type of people have escalating or reducing behavior in a population informs what and when interventions need to be implemented.

This is a trivial, highly simplified example, but it illustrates the type of results that occur in practice [84].

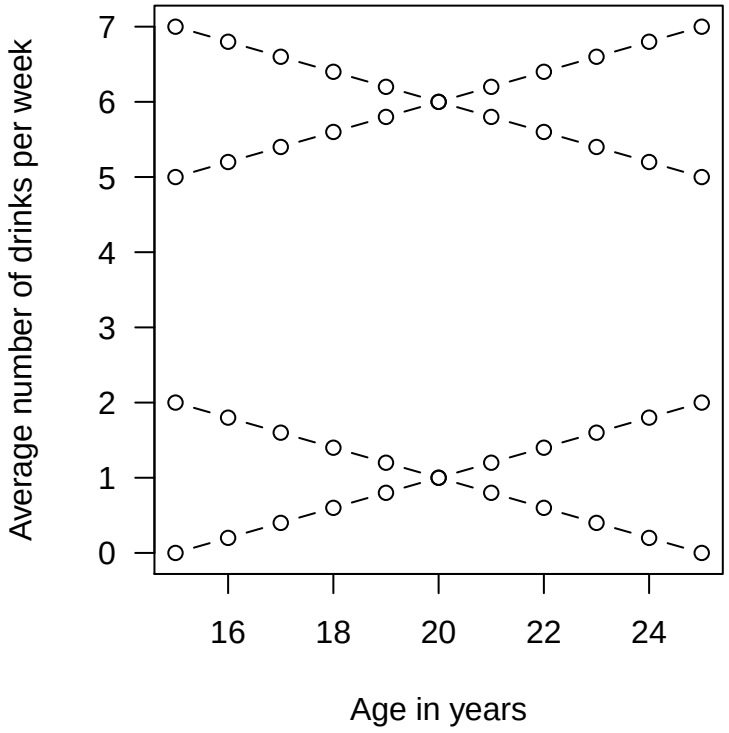

Figure 3.1: Graph of linear trajectories representing the hypothetical alcohol consumption of four individuals.

There are factors that influence the level and perhaps other factors that affect the shape of the patterns and it is important to separate these two relationships so as to not muddy the interpretations. If the goal is to detect and compare groups based on their temporal change over time, we need methods that answer the following research questions: Are there distinct shape patterns in the data? How many patterns are there? Are there baseline factors that impact the shape of an individual's trajectory? In the next section, I discuss two popular standard clustering methods in the context of clustering based on shape and highlight the situations in which these methods fail to address such research questions.

3.2 Limitations of standard clustering methods

Two standard methods for clustering multivariate data include partition methods based on a dissimilarity measure (e.g. Euclidean distance) and model-based methods such as finite mixture models. These methods were introduced in the first chapter of this thesis and I now illustrate the limitations of these methods in answering research questions about shape.

I assume that there are n subjects such that for subject i, we observe m_i repeated measures of an outcome of interest. I denote the vector of measured outcomes as $\mathbf{y}_i = (y_{i1}, ..., y_{im_i})$ for $i = 1, ..., n$ and $\mathbf{w}_i \in \mathbb{R}^q$ as the design vector based on baseline variables that may impact group membership. Lastly, the corresponding time of the data collection for subject i is $\mathbf{t}_i = (t_{i1}, ..., t_{im_i})$.

3.2.1 Partitioning algorithms

If the outcome is measured at the same times for every subject such that $m_i = m$ and $\mathbf{t}_i = \mathbf{t}$ for all $i = 1, ..., n$, longitudinal data fit into a typical multivariate data framework where applying a partitioning method like the K-means algorithm [81, 49] to the observed vector is typically appropriate. See Chapter 1 for details of the iterative algorithm.

I now illustrate how this algorithm fails to address all three of the questions previously posed. Using the simple alcohol consumption example, let $n = 4$, and suppose the average number of drinks per week is collected over a ten year period, $\mathbf{t} = (15, 16, ..., 24, 25)$. Let the observed outcome vectors be linear in trend as shown in Figure 3.1. Note that there are two shape groups: increasing and decreasing. In this circumstance, the level and the shape are not strongly associated since there are different levels within each shape group. The K-means algorithm groups on the basis of squared Euclidean distance between individuals. The distances between these four subjects are listed in Table 3.1.

	Low Decreasing	High Decreasing	Low Increasing	High Increasing
Low Decreasing	0	275.0	17.6	292.6
High Decreasing	275.0	0	292.6	17.6
Low Increasing	17.6	292.6	0	275.0
High Increasing	292.6	17.6	275.0	0

Table 3.1: Squared Euclidean distance matrix for the hypothetical alcohol consumption vectors of four individuals: high level but slowly decreasing, high level but slowly increasing, low level but slowly decreasing, and low level but slowly increasing.

The level dominates such that the subjects with the same level (high or low) are clearly the most similar using this dissimilarity measure. The distance between those with same shape have a distance of about fifteen times that of distances within levels. Clustering using Euclidean distance thus detects distinct levels, and if the distinct shapes do not coincide with the distinct levels, the K-means algorithm fails at grouping individuals based on shape and detecting the number of trend patterns. The mean curves over time for the two groups end up being horizontal even though the alcohol consumption is not stable over time for any individual.

Imagine if the shape of the trajectory were correlated with the vertical level where all of the heavy drinkers slowly decrease the number of drinks per week and the light drinkers increase over the years; then, K-means algorithm results in clusters based on level and thus

on shape in this case (Figure 3.2). The algorithm detects the shape groups only when the trajectories with similar shapes also have similar levels, which may be the case in some applications but is not true for many data sets. I have focused on the K-means algorithm, but these conclusions hold for the PAM algorithm when Euclidean or squared Euclidean distance is used.

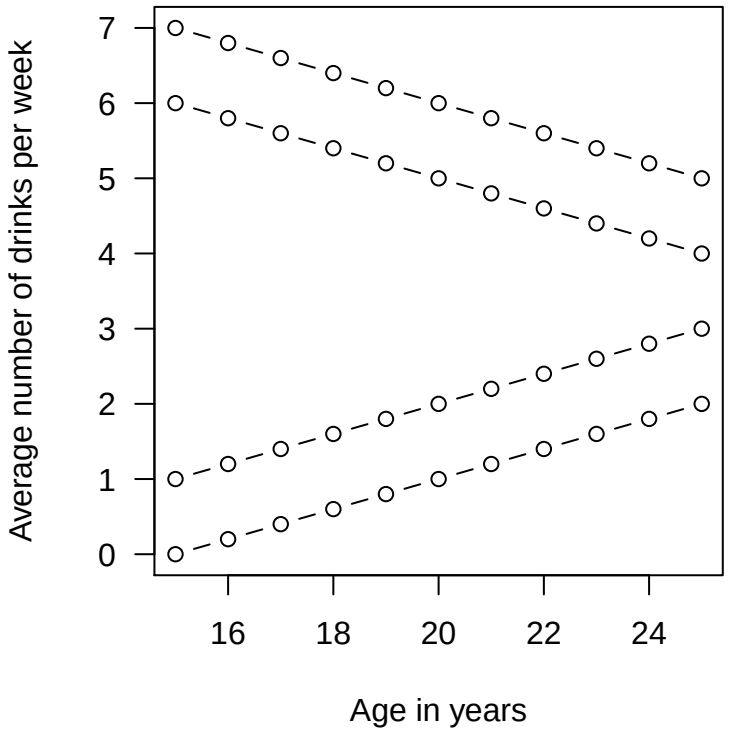

Figure 3.2: Graph of linear trajectories representing the hypothetical alcohol consumption of four individuals where shape and level are strongly associated.

Lastly, if baseline factors are related to the cluster membership, the only option is to complete an analysis after the clustering algorithm is complete. One technique is to fit a multinomial logistic regression model for the group labels with baseline factors as the explanatory variables. This analysis automatically assumes that the group labels are known and fixed; it does not take into account any variability in the algorithm or the uncertainty in the group memberships. Therefore, any inference should be done with caution as the standard errors are calculated conditional on cluster labels and do not take into account other sources of uncertainty.

3.2.2 Finite mixture models

In contrast to partition methods, finite mixture models provide a probability framework in which to take into account the uncertainty of group memberships and simultaneously estimate the relationship between baseline factors and groups. Additionally, the model is sufficiently flexible to accommodate irregular sampling which frequently occurs in longitudinal studies. A finite mixture model is a weighted sum of component distributions. I assume a parametric form for these distributions and that parameters differ between components.

In general, I assume there are K latent groups that occur in the population with frequencies $\pi_1, ..., \pi_K$. If subject i is a member of the kth group, the observed data vector for that subject is given by

$$\mathbf{y}_i = \boldsymbol{\mu}_k + \boldsymbol{\epsilon}_i \quad \boldsymbol{\epsilon}_i \sim N(0, \boldsymbol{\Sigma}_k).$$

Maximum likelihood estimation via the EM algorithm is used to estimate the model parameters and the posterior probabilities of whether a subject belongs to each component. These probabilities provide a way to soft cluster subjects to groups as a subject 'belongs' to every cluster with some probability. Subjects can then be hard clustered through assignment based on the maximum posterior probability.

If the outcome vector is repeated measures over time, there are some necessary adjustments that need to be made to the mixture model. The outcome vectors may not have equal length and the measurements may not be observed at the same times between individuals. Therefore, structure must be imposed on the mean vector and the covariance matrix. A regression structure can be used by assuming a linear model for the mean, $\boldsymbol{\mu}_k = \mathbf{x}\boldsymbol{\beta}_k$. The design matrices based on explanatory variables $\mathbf{x}$ must include time components that can model the shape of the curve. In terms of the covariance structure, most software packages restrict the structure to conditional independence, which can be problematic as shown in Chapter 2.

If conditional independence is assumed for the correlation structure, then the likelihood function is based on squared Euclidean distances between observations and group means scaled by the estimated variance. Akin to the K-means algorithm, maximizing the likelihood with the original data vectors generally fails to group individuals by shape if shape and level are weakly dependent. If the level of the outcome vector is largely determined by a random intercept such that $\mathbf{y}_i = \lambda_i \mathbf{1} + \mathbf{x}_i\boldsymbol{\beta}_k + \boldsymbol{\epsilon}_i$ where $\lambda_i \sim N(0, \tau_k^2)$ and $\boldsymbol{\epsilon}_i \sim N(0, \sigma_k^2 I)$ for subjects in the kth group, then it is possible to model the variability of the level through the correlation structure. In general, if the distribution of the level within a shape group is known and can be correctly modeled within each shape group, then including that into the model can produce the shape groups. In practice, such distributions are not known and clustering results are sensitive to incorrect assumptions.

It only takes a simple simulation to show that the standard finite mixture fails to group individuals by shape, but Nagin's book leads people to believe that a finite mixture model assuming conditional independence "lends itself to analyzing questions that are framed in terms of the shape of the developmental course of the outcome of interest" and "focuses on identification of different trajectory shapes and on examining how the prevalence of the

shape and the shape itself relate to predictors" [100]. This shows a gap in the understanding present in the literature.

3.3 Methods extended for shape

There has been work done in developing and adapting clustering methods for longitudinal data, but there has been much less discussion in the literature about the best methods to utilize when shape of pattern over time is the feature of interest. In this dissertation, the term shape refers to the shape of the underlying functional pattern over time.

There are many ways to describe the shape of a function. The first derivative of the function with respect to time provides the instantaneous rate of change at a point in time. The sign of the derivative indicates whether the function is increasing or decreasing at a point. Local and global minima and maxima of the function are determined by where the first derivative is zero. Lastly, the second derivative provides information about the growth of the first derivative and thus whether the function is concave up or down at any point.

These are all aspects of the shape and they may be important in different scientific settings. For example, when clustering genes, the location of the peaks in expression levels may be of interest so focusing on local maxima would be useful when clustering [80]. In other circumstances, it is only necessary to know whether the function is increasing or decreasing so that only the sign of the first derivative is needed [106].

For this thesis, I define shape as the pattern of the function after disregarding the vertical level. There are a few ways to compare the shape of two function using this definition. One way is to compare the first derivatives as differentiation removes the level and uniquely describes the shape of the remaining curve. However, it is difficult to estimate the derivative of a discretized, noisy version of the function. Another popular approach involves calculating the Pearson correlation coefficient between two vectors of discretized versions of the functions. These two ideas have been implemented in a clustering framework for longitudinal data.

3.3.1 Derivative-based dissimilarity

One approach to clustering on the basis of shape is to compare derivative functions. The first derivative of a function describes the rate of change while ignoring the intercept or level of the original function. With longitudinal data, we do not directly observe the derivative function for each individual. Assume that the jth observed outcome for individual i at time t_{ij} is a realization of the model

$$y_{ij} = f_i(t_{ij}) + \epsilon_{ij}$$

for $i = 1, ..., n$ and $j = 1, ..., m_i$ where $\epsilon_{ij} \overset{iid}{\sim} (0, \sigma_i^2)$, f_i is a subject-specific differentiable, continuous function, and m_i is relatively small (usually around 5 to 10). When the observation times are consistent across individuals with $m_j = m$ and $t_{ij} = t_j$ for $j = 1, .., m$ and $i = 1, ..., n$, Möller-Levet et al. [92] and D'Urso [29] independently suggested estimating the

derivative of the underlying smooth function via the difference quotient by calculating the slope of a linear interpolation between adjacent repeated measures,

$$\hat{f}_i^{'}(t_j) = (y_{i\ j+1}) - y_{ij})/(t_{j+1} - t_j).$$

By the mean value theorem, this is an unbiased estimate of the true derivative $f_i^{'}(\tau)$ at a point $\tau \in [t_j, t_{j+1}]$ such that

$$E(\hat{f}_i^{'}(t_j)) = f_i(t_{j+1}) - f_i(t_j))/(t_{j+1} - t_j) = f_i^{'}(\tau).$$

However, the estimate is highly variable if σ_i^2 is large since

$$\text{Var}(\hat{f}_i^{'}(\tau)) = 2\sigma_i^2/(t_{j+1} - t_j)^2.$$

Large variability in the estimates impacts the cluster analysis if enough estimates are far from the true derivative. The dissimilarity between two individuals is measured as the squared Euclidean distance between the vectors of difference quotients. High variability can lead to high dissimilarity for individuals with similarly underlying shape. One way to minimize the variance is to maximize the time between observations. Observing only two observations, one at baseline and another at the end of the follow-up period, minimizes the variance but at the expense of observing the rate of change during the follow-up period. If the times of observation are densely sampled, a functional approach smoothes out the noise using splines to estimate the function and then the derivative function [132]. In either circumstance, the derivatives are independently estimated for each individual and there is no direct way to borrow strength between individuals to better estimate the derivative even if some individuals are thought to have a common shape.

Our main concern is how well these methods cluster based on shape; therefore, it is important to determine the behavior of the Euclidean distance using these estimated derivatives and how well it can distinguish between two noisy curves in terms of their shape over time. Assume there are three subjects ($n = 3$) observed ten times ($m = 10$) uniformly at intervals of Δ, $\mathbf{t}_1 = \mathbf{t}_2 = \mathbf{t}_3 = (\Delta, 2 \cdot \Delta, ..., 10 \cdot \Delta)$. Two subjects have the same horizontal shape over time and the third has a positive slope, $f_1(t) = f_2(t) = 0$, $f_3(t) = a * t$, and the variability of the noise is the same amongst the subjects, $\sigma_1 = \sigma_2 = \sigma_3 = \sigma$. The outcomes for individual i at the jth observation time equals

$$y_{ij} = f_i(t_j) + \epsilon_{ij} \quad \epsilon_{ij} \sim N(0, \sigma^2)$$

for $i = 1, 2, 3$ and $j = 1, ..., 10$. The vectors of difference quotients for the three subjects are denoted as $\mathbf{dy}_1$, $\mathbf{dy}_2$, $\mathbf{dy}_3$. I am interested in the probability of the event that the distance between $\mathbf{dy}_1$ and $\mathbf{dy}_2$, which are both generated from the same horizontal shape, is less than the distance between $\mathbf{dy}_2$ and $\mathbf{dy}_3$, which are generated from different shapes. Given values for σ, Δ, a, I empirically estimate

$$P(\|\mathbf{dy}_1 - \mathbf{dy}_2\|_2^2 < \|\mathbf{dy}_2 - \mathbf{dy}_3\|_2^2)$$

by generating data for the three subjects 5000 times. I complete this simulation under many conditions specified by a combination of $a = 0.25, 1, 5$, $\Delta = 0.5, 1, 1.5, 2, 2.5, 3$, and $\sigma = 0.5, 1, 1.5, 2, 2.5, 3$. The estimated probabilities are plotted against the ratio, σ/Δ, for different slopes of $f_3(t)$ (Figure 3.3).

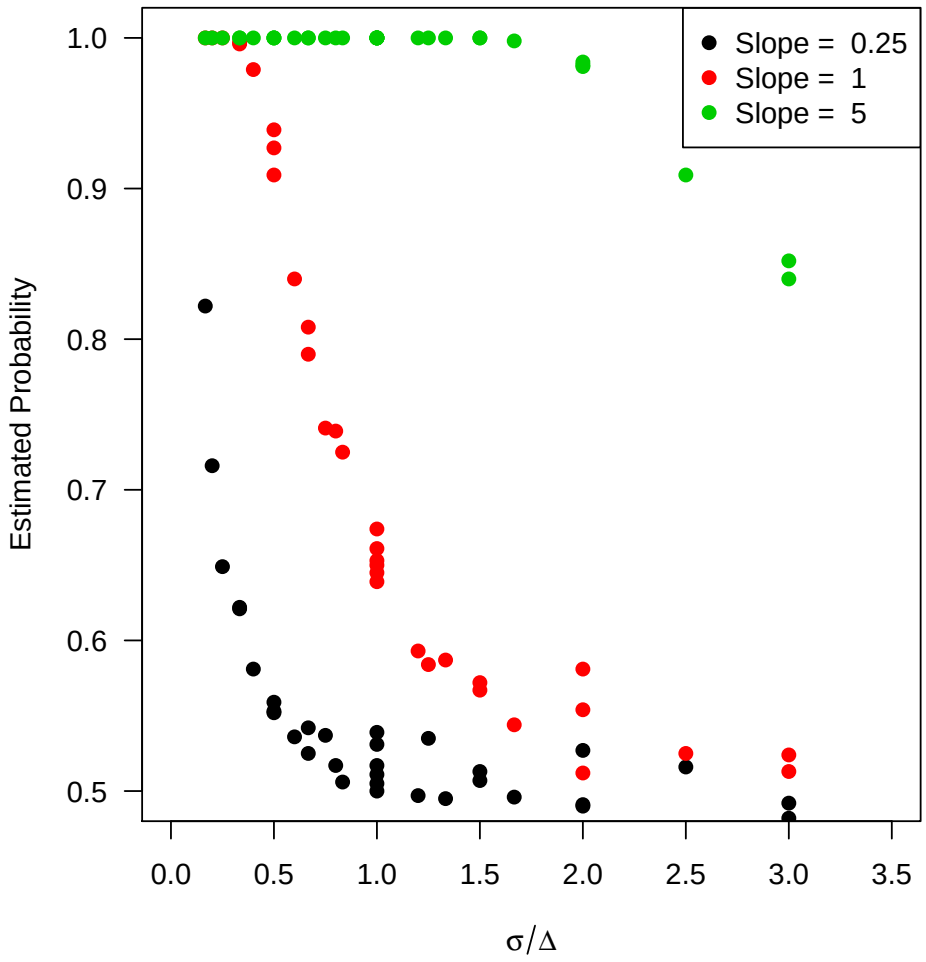

Figure 3.3: Empirical probability that the distance between $\mathbf{dy}_1$ and $\mathbf{dy}_2$, which share the same underlying horizontal function, is smaller than the distance between $\mathbf{dy}_1$ and $\mathbf{dy}_3$ for different ratios of standard deviation to length of time lags (σ/Δ) and slopes of the linear function underlying $\mathbf{y}_3$ based on 5000 replications.

I note that as the ratio of variance to time interval increases, the probability of the two horizontal observed vectors being closer than the two vectors from different shapes decreases to 0.50. Therefore, if the variance is large in comparison to the sampling increments, it is a toss up as to whether the Euclidean distance between the difference quotients correctly determines which vectors were generated from the same shape. On the other hand, if the variance is small relative to the time between samples, the procedure finds the two horizon-

tal lines more similar than two lines with different slopes most of the time. As the slope increases and thus the difference between shapes increases, the method correctly determines similarity. There is a trade-off between bias and variance so individuals need to be observed at intervals relative to the measurement error and degree to which the shapes differ without compromising the ability to measure the shape. Otherwise, this procedure fails at grouping individuals with similar patterns over time.

3.3.2 Correlation-based dissimilarity

Another approach is to calculate the Pearson correlation coefficient which is invariant to trajectory level [15, 30, 14]. Despite the wide use, there has been little discussion about how well the correlation does to discriminate between shapes in this context. In the multivariate setting, one dissimilarity measure based on the Pearson correlation coefficient between two comparable vectors is

$$d_{Cor}(\mathbf{x}, \mathbf{y}) = 1 - Cor(\mathbf{x}, \mathbf{y})$$

where

$$Cor(\mathbf{x}, \mathbf{y}) = \frac{\sum_{j=1}^{m}(x_j - \bar{x})(y_j - \bar{y})}{\sqrt{\sum_{j=1}^{m}(x_j - \bar{x})^2}\sqrt{\sum_{j=1}^{m}(y_j - \bar{y})^2}}.$$

This measure takes values between 0 and 2 with extremes attained only when there is perfect linear positive or negative correlation. If there is zero noise, a vector with an underlying non-constant functional shape of $f(t)$ is perfectly positively correlated with a vector with functional shape of $a \cdot f(t) + b$ where $a > 0$ and $b \in \mathbb{R}$. As a result, two vectors that have the same shape over time but at different levels have a dissimilarity of zero since the correlation coefficient equals 1. However, two vectors with underlying functions that are scalar multiples of each other are also placed in the same cluster despite having drastically different shapes. While this method succeeds in grouping individuals with the same shape at different levels, it fails in the sense that is also groups individuals with different shapes. If the function is observed with error such that $\epsilon_{ij} \sim (0, \sigma_i^2)$, the magnitude of the correlation is generally deflated such that the dissimilarity moves closer to the middle value of 1.

The correlation between two constant functions is undefined. With noise, the expected dissimilarity value is 1 with variation dependent on the number of observations. In other words, the Pearson correlation coefficient does not consistently detect two horizontal curves to have the same shape especially with a small number of observations, which is typical in longitudinal data. This dissimilarity measure fails to detect similar underlying shapes when σ_i^2 is large, the number of observations is low, and the data set includes constant or stable patterns over time.

The related cosine-angle dissimilarity measure has also been used to determine the similarity of two vector profiles [30]. The measure is the uncentered version of the Pearson

correlation dissimilarity,

$$d_{Cos}(\mathbf{x}, \mathbf{y}) = 1 - \frac{\sum_{j=1}^{m} x_j y_j}{\sqrt{\sum_{j=1}^{m} x_j^2}\sqrt{\sum_{j=1}^{m} y_j^2}}.$$

Since the mean is not subtracted from the vector elements, it is defined for two vectors based on constant functions. However, the measure is invariant to scaling transformations of the vectors and not invariant to vertical shifts, which are undesirable properties for a dissimilarity measure based on our definition of shape similarity.

3.4 Discussion

The standard clustering methods generally fail to answer the three research questions posed at the beginning of this chapter: Are there distinct shape patterns in the data? How many patterns are there? Are there baseline factors that impact the shape of an individual's trajectory? There have been some attempt to adjust the input vector to group by shape, but these methods based on difference quotients and the Pearson correlation coefficient generally fail when data observed equals a smooth function plus moderate variability. Additionally, both methods require data to be collected at fixed time points for all of the subjects. Ideally, we want a method that exploits all of the properties of longitudinal data, which includes the possibility of sporadic, irregularly sampled data.

In the next chapter, I propose three extensions of existing method that attempt to answer the research questions above for real longitudinal data sets. Before I present the proposed methodology, it is worth discussing two common data-generating models that allow a group structure based on shape and individuals can differ in the vertical level within the group. In a functional data approach [111], we assume that the jth observation for the ith individual at time t_{ij} is

$$y_{ij} = f_i(t_{ij}) + \epsilon_{ij}$$

for $i = 1, ..., n$ and $j = 1, ..., m_i$ where $\epsilon_{ij} \overset{iid}{\sim} (0, \sigma_i^2)$ and f_i is a random function sampled from a Hilbert space of square integrable functions on a real interval.

In a typical longitudinal data analysis approach [26], we assume that the jth observation for the ith individual at time t_{ij} in group k is

$$y_{ij} = \lambda_i + \mu_k(t_{ij}) + \epsilon_{ij}$$

for $i = 1, ..., n$ and $j = 1, ..., m_i$ where λ_i adjusts the vertical level of an individual's curve, $\mu_k(t)$ is a continuous mean function of time for cluster k, and ϵ_{ij} is random error that may be correlated within subject i but independent between subjects. If λ_i is assumed random, then the model is considered a random intercept model within clusters.

In the next chapter, I present three new methods adapted from related literature that address clustering based on shape over time and then compare them with the methods discussed in this chapter.

Chapter 4

Proposed shape-based methods

In this chapter, I present three new clustering techniques that attempt to answer the three research questions presented in the last chapter: Are there distinct shape patterns in the longitudinal data? How many patterns are there? Are there baseline factors that impact the shape of an individual's trajectory? For each method, I discuss related work and then introduce necessary background, notation, and the model specification. I describe the implementation process and any foreseen issues and limitations. I then discuss the advantages and disadvantages of each. In the next chapter, a simulation study compares the proposed methods with those presented earlier in this thesis in addressing research questions about shape.

4.1 Derivative spline coefficients partitioning

The first method is based on the idea that differentiation removes the level from a function and provides information about the shape. Unlike the difference quotient method described in Chapter 3, the proposed method smoothes out noise prior to calculating the derivative and does not require the data to be regularly sampled for all subjects. Using techniques from functional data analysis [111], outcome data for each individual are projected onto a functional basis to estimate a smooth function of time. The estimated function is differentiated to get an indirect estimate of the derivative function. The dissimilarity between individuals is based on these estimates. By first removing the level, this method theoretically clusters individuals with similarly shaped trajectories.

4.1.1 Related work

In Chapter 3, I introduced the difference quotient dissimilarity measure separately suggested by D'Urso [29] and Möller-Levet et al. [92]. Möller-Levet et al. referred to this measure as the short time-series distance and developed it to identify similar shapes in microarray data. In contrast, D'Urso took a physics view of the data and referred to the difference

quotient measure as the longitudinal-velocity dissimilarity measure. Additionally, he used a difference quotient in velocities to calculate a measure inspired by acceleration and then combined the cross-sectional and evolutive information of the trajectories into one dissimilarity measure.

Similar to D'Urso, Zerbe presented three distance measures for growth curves based on position, velocity, and acceleration [151, 122]. Rather than using difference quotients, he suggested estimating the velocity by first fitting a polynomial of degree d to each individual's growth curve using least squares. For individual i, he let $\mathbf{y}_i$ be the vector of observed outcomes at times $\mathbf{t}_i$. Then, the estimated curve is $\hat{f}_i(t) = (1 \;\; t \;\; t^2 \;\; ... \;\; t^d)\, \hat{\boldsymbol{\beta}}_i$ where

$$\hat{\boldsymbol{\beta}}_i = (\mathbf{X}_i^T \mathbf{X}_i)^{-1} \mathbf{X}_i^T \mathbf{y}_i$$

and $\mathbf{X}_i$ is the within-individual design matrix determined by the polynomial function evaluated at the observation times of individual i. The estimated derivative is equal to $\hat{f}_i'(t) = [0 \;\; 1 \;\; 2t \;\; ... \;\; dt^{d-1}]\, \hat{\boldsymbol{\beta}}_i$. In other words, he suggested projecting the data onto a polynomial basis of degree d, differentiating the basis, and calculating the estimated derivative function, which can be represented by a polynomial basis of degree $d-1$. The dissimilarity between the ith and jth individuals based on the velocity equals

$$d(\mathbf{y}_i, \mathbf{y}_j) = \left[\int_{\mathcal{T}} [\hat{f}_i'(t) - \hat{f}_j'(t)]^2 dt \right]^{1/2}$$

for a chosen time interval, $\mathcal{T}$. This integral is easy to evaluate since the estimated derivatives functions are represented using a polynomial basis.

Tarpey and Kinateder [132], at the end of their functional clustering paper, briefly mentioned a few suggestions to cluster data after getting 'rid of dominating variability in the intercept.' One of their proposals was to cluster individuals based on the derivatives of the estimated functions. In personal correspondence with one of the authors, the details of the implementation were clarified. After projecting individuals' data onto a finite Fourier basis, they differentiated the Fourier basis functions and used the K-means algorithm on the coefficients of the derivative functions. Thus, the estimated function for individual i is $\hat{f}_i(t) = (1 \;\; \sin(2\pi t) \;\; \cos(2\pi t) \;\; ... \;\; \sin(2\pi w t) \;\; \cos(2\pi w t))\, \hat{\boldsymbol{\beta}}_i$ where

$$\hat{\boldsymbol{\beta}}_i = (\mathbf{X}_i^T \mathbf{X}_i)^{-1} \mathbf{X}_i^T \mathbf{y}_i$$

and $\mathbf{X}_i$ is the within-individual design matrix determined by the Fourier expansion evaluated at the observation times of individual i. Since $\frac{d}{dt}\cos(2\pi w t) = -2\pi w \sin(2\pi w t)$ and $\frac{d}{dt}\sin(2\pi w t) = 2\pi w \cos(2\pi w t)$, the estimated derivative function for individual i is repre-

sented using the same Fourier expansion with new coefficients

$$\hat{\boldsymbol{\alpha}}_i = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & -2\pi & \cdots & 0 & 0 \\ 0 & 2\pi & 0 & \ddots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \ddots & 0 & -2\pi w \\ 0 & 0 & 0 & \cdots & 2\pi w & 0 \end{pmatrix} \hat{\boldsymbol{\beta}}_i$$

that are transforms of the original coefficients. The dissimilarity between the ith and jth individuals based on the derivative coefficients equals

$$d(\mathbf{y}_i, \mathbf{y}_j) = (\hat{\boldsymbol{\alpha}}_i - \hat{\boldsymbol{\alpha}}_j)^T(\hat{\boldsymbol{\alpha}}_i - \hat{\boldsymbol{\alpha}}_j).$$

These two approaches use the same general procedure. Individual trajectories are smoothed by projecting the data onto a chosen basis, the derivative of the estimated function is calculated by differentiating the basis functions, and the dissimilarity between individuals is based on the estimated derivative functions. Now with any smoothing procedure, there is a fundamental bias-variance tradeoff. In this case, including higher ordered polynomial terms or more sine and cosine functions in the basis decreases the bias but increases the variance. The basis functions need to be selected so as to strike a balance between the two. The type of basis also affects the differentiation process. For example, polynomial and Fourier bases are computationally convenient in that differentiation results in a basis of the same type.

Despite the similarities, these two methods differ in how the dissimilarity measure is defined between two individuals. Zerbe used the L_2 distance between two derivative functions; Tarpey and Kinateder calculated the squared Euclidean distance between the basis coefficient vectors for the estimated derivative function. This reflects the diversity in the functional cluster analysis literature; some use the L_2 distance on functions [54] while others calculate the Euclidean distance between the linear coefficients of the basis function [1, 125, 132].

Tarpey [131] reconciled these two dissimilarity measures by showing that clustering functional data using the L_2 metric on function space can be achieved by running K-means on a suitable linear transformation of the basis coefficients. If $y(t)$ is a functional realization represented as $y(t) = \sum_j \beta_u u_j(t)$and $\mu(t)$ is a functional cluster mean represented as $\mu(t) = \sum_j \gamma_j u_j(t)$, then the squared L^2 distance between them on interval $\mathcal{T}$ is

$$\begin{aligned} \int_{\mathcal{T}} (y(t) - \mu(t))^2 dt &= \int_{\mathcal{T}} (\sum_j (\beta_j - \gamma_j) u_j(t))^2 dt \\ &= \sum_j \sum_l (\beta_j - \gamma_j)(\beta_l - \gamma_l) \int_{\mathcal{T}} u_j(t) u_l(t) dt \\ &= (\boldsymbol{\beta} - \boldsymbol{\gamma})^T \mathbf{W} (\boldsymbol{\beta} - \boldsymbol{\gamma}) \\ &= (\mathbf{W}^{1/2}(\boldsymbol{\beta} - \boldsymbol{\gamma}))^T (\mathbf{W}^{1/2}(\boldsymbol{\beta} - \boldsymbol{\gamma})) \end{aligned}$$

where $\mathbf{W}_{jl} = \int_{\mathcal{T}} u_j(t)u_l(t)dt$. Therefore, clustering with the L^2 distance is equivalent to plugging transformed coefficients, $\mathbf{W}^{1/2}\boldsymbol{\beta}$, into the K-means algorithm. Consequently, when the functions are represented using an orthogonal basis such as the Fourier expansion, K-means on the coefficients is equivalent to the L^2 implementation.

Both functional bases presented thus far are restrictive. The Fourier basis only works well when the data is periodic in nature and a polynomial basis does not provide a general structure to represent complex functions with few parameters. To allow for flexibility in the functional shape, the observation time interval, $[a, b]$, can be broken up into smaller interval using L internal knots, $a < \tau_1 < \cdots < \tau_L < b$, so that polynomials of order p are fit in each subinterval. This piecewise polynomial can be expressed as a linear combination of truncated power functions and polynomials of order p. In other words, $\{1, t, t^2, ..., t^{p-1}, (t - \tau_1)_+^{p-1}, ..., (t - \tau_L)_+^{p-1}\}$ is a basis for a piecewise polynomial with knots at $\tau_1, ..., \tau_L$. However, the normal equations associated with the truncated power basis are highly ill-conditioned.

A better conditioned basis for the same function space is the B-spline basis [22, 123, 19, 23], which extends the advantages of polynomials to include greater flexibility [1]. To my knowledge, no other study has focused on clustering longitudinal data using the coefficients of the B-spline derivative estimate.

In the following sections, I introduce B-spline functions and demonstrate how they can be used to estimate derivative functions. Then, the implementation and practical decisions that need to be made are presented and discussed.

4.1.2 B-spline functions

I fit a p-order spline function to each subject i in order to estimate its underlying smooth, f_i. For the sake of being self-contained, I include some background on splines. Let $t \in [a, b]$ where $a, b \in \mathbb{R}$ and $\xi_0 = a < \xi_1 < \cdots < \xi_L < b = \xi_{L+1}$ be a subdivision of the interval $[a, b]$ by L distinct points, termed internal knots. The knot sequence is augmented by adding replicates at the beginning and end, $\tau = [\tau_1, ..., \tau_{L+2p}]$ for $p \in \mathbb{N}$, such that

$$\begin{aligned} \tau_1 = \tau_2 = \cdots = \tau_p &= \xi_0 \\ \tau_{j+p} &= \xi_j, \qquad j = 1, ..., L \\ \xi_{L+1} = \tau_{L+p+1} = \tau_{L+p+2} &= \cdots = \tau_{L+2p} \end{aligned}$$

The spline function, $f(t)$, is a polynomial of order p on every interval $[\tau_{j-1}, \tau_j]$ and has $p - 2$ continuous derivatives on the interval (a, b). The set of spline functions of order p for a fixed sequence of knots, $\tau = [\tau_1, ..., \tau_{L+2p}]$, is a linear space of functions with $L + p$ free parameters. A useful basis $B_{1,p}(t), ..., B_{L+p,p}(t)$ for this linear space is given by Schoenberg's B-splines

[19, 23] defined as

$$B_{j,1}(t) = \begin{cases} 1 \text{ if } \tau_j \leq t < \tau_{j+1} \\ 0 \text{ otherwise} \end{cases}$$

$$B_{j,l}(t) = \frac{t - \tau_j}{\tau_{j+l-1} - \tau_j} B_{j,l-1}(t) + \frac{\tau_{j+l} - t}{\tau_{j+l} - \tau_{j+1}} B_{j+1,l-1}(t)$$

where $l = 2, ..., p$ and $j = 1, ..., L + 2p - l$. If I adopt the convention that $B_{j,1}(t) = 0$ for all $t \in \mathbb{R}$ if $\tau_j = \tau_{j+1}$, then by induction $B_{j,l}(t) = 0$ if $\tau_j = \tau_{j+1} = \cdots = \tau_{j+l}$. Hence, $B_{1,l}(t) = 0$ for $t \in \mathbb{R}$ and $l < p$ on the defined knot sequence. The B-spline function of order p is defined by

$$f(t) = \sum_{j=1}^{L+p} \beta_j B_{j,p}(t).$$

Let $\mathbf{y}_i = (y_{i1}, ..., y_{im_i})$ denote a vector of repeated observations for individual i observed at times $\mathbf{t}_i = (t_{i1}, ..., t_{im_i})$ for $i = 1, ..., n$. I assume that $y_{ij} = f_i(t_{ij}) + \epsilon_{ij}$ such that $E(\epsilon_{ij}) = 0$ for all $j = 1, ..., m_i$ and $i = 1, ..., n$. To estimate f_i, I fix the order of the B-spline to p and the internal knots and then estimate the coefficients, $\boldsymbol{\beta}_i = (\beta_{i,1}, ..., \beta_{i,(L+p)})$ using least squares. The estimated function is $\hat{f}_i(t) = \sum_{j=1}^{L+p} \hat{\beta}_{i,j} B_{j,p}(t)$ where

$$\hat{\boldsymbol{\beta}}_i = (\mathbf{X}_i^T \mathbf{X}_i)^{-1} \mathbf{X}_i \mathbf{y}_i$$

and $\mathbf{X}_i$ is the within-individual design matrix based on the B-spline basis functions.

To estimate $f_i'(t)$, the estimated function $\hat{f}_i(t)$ is differentiated with respect to t such that

$$\hat{f}_i'(t) = \sum_{j=1}^{L+p} \hat{\beta}_{i,j} B_{j,p}'(t).$$

Prochazkova [108] showed that this can be simplified to

$$\hat{f}_i'(t) = \sum_{j=1}^{L+p} \hat{\beta}_{i,j} \left[\frac{p-1}{\tau_{j+p-1} - \tau_j} B_{j,p-1}(t) - \frac{p-1}{\tau_{j+p} - \tau_{j+1}} B_{j+1,p-1}(t) \right].$$

However, this can be written in terms of a B-spline basis of one order lower,

$$\hat{f}_i'(t) = \sum_{j=1}^{L+p-1} (\hat{\beta}_{i,j+1} - \hat{\beta}_{i,j}) \frac{p-1}{\tau_{j+p} - \tau_{j+1}} B_{j+1,p-1}(t).$$

Adjusting the knot sequence to only have $p - 1$ replicates at the beginning and end results in

$$\hat{f}_i'(t) = \sum_{j=1}^{L+p-1} \hat{\alpha}_{i,j} B_{j,p-1}(t)$$

where $\hat{\alpha}_{i,j} = (\hat{\beta}_{i,j+1} - \hat{\beta}_{i,j}) \frac{p-1}{\tau_{j+p} - \tau_{j+1}}$. These derivative coefficients, $\hat{\boldsymbol{\alpha}}_i = [\hat{\alpha}_{i,1}, ..., \hat{\alpha}_{i,(L+p-1)}]$, can be used to cluster trajectories with similar shape with the K-means algorithm [81, 49].

4.1.3 Implementation

To use this method in practice, decisions about the B-spline basis need to be made. The order of the polynomials must be selected. This together with the number of internal knots impacts the flexibility of the B-spline function. Cubic B-splines of order four have been shown to have good mathematical properties and are used frequently in practice [58]. However, depending on the number of data points observed per subject, it may be necessary to use a quadratic polynomial ($p = 3$) due to the restriction that the sum of the order and the number of internal knots must be less than or equal to the minimum number of observation points per subject for estimation to be possible.

The number of internal knots plays a role in the flexibility of the class of spline functions and should be enough to fit the features in the data. As with the order, the main limiting factor in choosing L is the number of data points per subject. The longitudinal data sets considered in this thesis have about five to ten data points per subjects. It is important not to over fit the individual curves so for data with limited observation, it may only be possible to have at most one internal knot. If there are many repeated measures, model selection information criteria or cross-validation can be used to choose L [114]. The location of the internal knots is another issue of discussion. There are some suggested data-driven ways to select knot location [152], but Ruppert [120] supports fixing the knots at sample quantiles. I generally follow this suggestion and adjust them as necessary to the areas with the most functional activity.

Once the order of the polynomials and number and location of knots are chosen, then smooth functions and their derivatives are estimated using least squares separately for every subject. The coefficients from the estimated derivative functions become the input vectors to the K-means algorithm (see Chapter 1 for more details). This algorithm converges, but there is no guarantee that it will find the grouping that globally minimizes the objective function. Therefore, in practice, the algorithm is run multiple times with different random initializations and the clustering that minimizes the objective function is chosen. I use 25 random starts.

The number of clusters must be fixed in order to run the K-means algorithm. However, K is unknown and of interest to researchers in practice. While no perfect mathematical criterion exists, a number of heuristics (see [134] and discussion therein) are available for choosing K. For this thesis, K is chosen for partition methods so as to maximize the overall average silhouette width [118]. See Chapter 1 for technical details of the silhouette width. By maximizing the overall average silhouette width, the dissimilarity between clusters is maximized and the dissimilarity within clusters is minimized resulting in distinct groups. In this case, the dissimilarity between two individuals is defined as squared Euclidean distance between derivative coefficients.

K-means is a partitioning algorithm; therefore, by definition every subject is hard clustered into one of the groups. There is no stated uncertainty in the group memberships even if a subject is between two clusters. In order to estimate the relationship between baseline variables and shape group membership, I assume the cluster labels are known. Given the

subject grouping labels from the partition, $\{c_i\}$ such that $c_i \in \{1, 2, ..., K\}$ for all $i = 1, ...n$, I fit a multinomial logistic regression model, which is an extension of the logistic regression model, using the group labels as the outcome and baseline factors as explanatory variables such that

$$P(c_i = k|\mathbf{w}_i) = \frac{\exp(\mathbf{w}_i^T \boldsymbol{\gamma}_k)}{\sum_{j=1}^{K} \exp(\mathbf{w}_i^T \boldsymbol{\gamma}_j)}$$

for all $k = 1, ..., K$ and $i = 1, ..., n$ with $\boldsymbol{\gamma}_K = 0$ where $\mathbf{w}_i$ is the design vector based on baseline factors. The parameters $\boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_{K-1}$ are estimated via maximum likelihood estimation. However, it is important to note that the estimated standard errors from a Hessian calculation do not include any group membership uncertainty as the hard clustering labels are assumed known for the estimation process.

4.2 Multilayer mixture model

The second method attempts to make up for the lack of uncertainty in groups memberships by using a model-based approach. Traditional clustering methods fail when the level and shape are weakly dependent resulting in individuals with the same shaped trajectory at different levels. As mentioned previously, if the levels within a shape group are normally distributed, using an exchangeable correlation matrix in a finite mixture model takes that variability into account. However, if the levels do not satisfy those assumptions, a Gaussian mixture could be used to model the non-Gaussian distribution of levels within a shape group.

4.2.1 Related work

Model-based methods provide the probability framework that dissimilarity-based methods lack. One benefit is the ability to simultaneously take into account uncertainty about of the parameter estimates and cluster membership. As mentioned in Chapter 1, the main model-based method is the finite mixture model. However, a standard mixture of Gaussians fit to the original data does not distinguish between the level and shape and clusters subjects based on the dominant source of variability. Another limitation of Gaussian mixtures for grouping data is that it cannot handle clusters that are non-normal. For this reason, the mixture model has been extended to include multiple layers such that each cluster is allowed to be a mixture [74]. For example, an obvious model for a data set with two groups each with bimodal densities would be a multilayer mixture of two clusters each with two components (see Figure 4.1 for a diagram of the model structure). A variation of this idea has been used to cluster non-normal groups by fitting a mixture with many components and then systematically combining components to make more meaningful clusters [53].

In this thesis, the goal is to have clusters be meaningful in terms of distinguishing between shape. Therefore, the idea of multilayer mixtures can be used to model K non-normal shape clusters composed of individuals with the same mean shape at different levels.

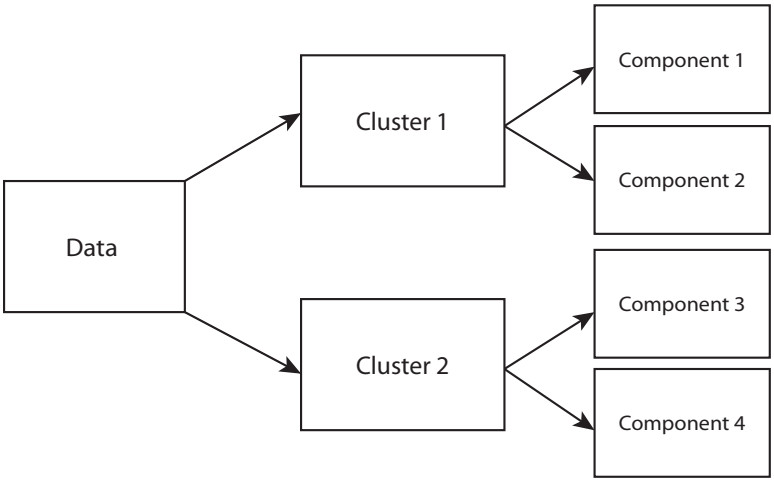

Figure 4.1: Diagram of multilayer mixture model showing that each cluster is composed of potentially more than one component.

4.2.2 Model specification

In a multilayer mixture model, K is the number of clusters and J_k is the number of components in the kth cluster ($k = 1, ..., K$) such that $J = \sum_{k=1}^{K} J_k$ is the total number of components in the entire model. Let j uniquely index all of the components such that $j = 1, ..., J$. For ease of explanation, let $c(j) : \{1, ..., J\} \rightarrow \{1, 2, ..., K\}$ be a cluster assigning function that specifies the cluster to which a component belongs. Let $f(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be the probability density function of a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Then, the probability density function for cluster k is

$$f_k(\mathbf{y}) = \sum_{j:c(j)=k} \pi_{j|c(j)} f(\mathbf{y}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$$

where $\pi_{j|c(j)}$ is the probability of being in component j given a subject is in cluster $c(j)$ and $\sum_{j:c(j)=k} \pi_{j|c(j)} = 1$ for all $k = 1, 2, ..., K$. Given baseline factors, let the probability of cluster k be $\pi_k(\mathbf{w}, \boldsymbol{\gamma}) = \exp(\mathbf{w}^T \boldsymbol{\gamma}_k) / \sum_{l=1}^{K} \exp(\mathbf{z}^T \boldsymbol{\gamma}_l)$ such that $\boldsymbol{\gamma}_K = 0$ and $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_K)$. The probability density function for the multilayer mixture is written as

$$g(\mathbf{y}|\mathbf{w}) = \sum_{k=1}^{K} \pi_k(\mathbf{w}, \boldsymbol{\gamma}) f_k(\mathbf{y}) = \sum_{k=1}^{K} \pi_k(\mathbf{w}, \boldsymbol{\gamma}) \sum_{j:c(j)=k} \pi_{j|c(j)} f(\mathbf{y}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j).$$

Since every component belongs to one and only one cluster, the above equation reduces to a regular mixture model if $\bar{\pi}_j(\mathbf{w}, \boldsymbol{\gamma}) = \pi_{c(j)}(\mathbf{w}, \boldsymbol{\gamma})\pi_{j|c(j)}$ with the density written as

$$g(\mathbf{y}|\mathbf{w}) = \sum_{j=1}^{J} \bar{\pi}_j(\mathbf{w}, \gamma) f(\mathbf{y}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j).$$

To adjust this model to satisfy the goal of clustering on shape, individuals in the clusters are assumed to have the same shaped trajectory over time and the varying levels are modeled by a mixture of components. The component mean structure includes a regression parameterization to model the smooth underlying group mean shape function over time plus a component-specific level. The design matrix for the regression can easily include B-spline basis functions presented in Section 4.1.2. Then, the regression parameters are constrained to be the same within shape clusters and the intercepts are allowed to differ in each level component. Hence, $\boldsymbol{\mu}_j = \lambda_j \mathbf{1} + \mathbf{x}\boldsymbol{\beta}_{c(j)}$, where $\mathbf{x}$ is a design matrix of B-spline basis functions excluding the first basis function to allow for estimation of intercept terms for each component. This allows the clusters to be based on shape while the components can have different levels (see Figure 4.2 for a diagram of the model structure). Additionally, I assume conditional independence within components, $\boldsymbol{\Sigma}_j = \sigma_j^2 \mathbf{I}$ for all $j = 1, ..., J$, for simplicity and to allow for irregularly sampled longitudinal data.

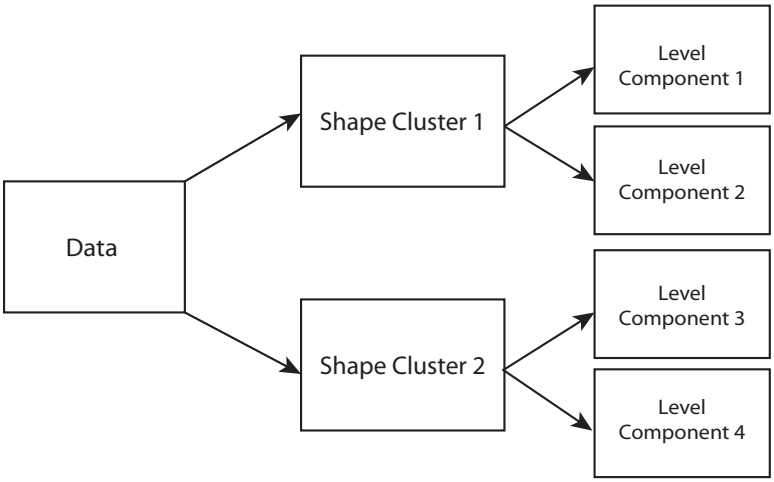

Figure 4.2: Diagram of multilayer mixture model showing that each shape cluster is composed of potentially more than one level component.

4.2.3 Implementation

Let $\mathbf{y}_i = (y_{i1}, ..., y_{im_i})$ denote a vector of repeated observations for individual i observed at times $\mathbf{t}_i = (t_{i1}, ..., t_{im_i})$ for $i = 1, ..., n$. For each individual i, B-spline basis functions are evaluated at the observation times to create design matrices $\mathbf{x}_i$ for the mean regression and baseline covariates are combined in design vectors $\mathbf{w}_i$ for the group membership regression. Now, B-splines are used so the mean structure is flexible enough to accommodate complex shapes. The number and location of internal knots are dealt with in the same manner as in the first proposed method.

Denote the true shape cluster identity of individual i by η_i where $\eta_i \in \{1, ..., K\}$. Then, the parameters of the multilayer mixture model $\boldsymbol{\eta} = (\eta_1, ..., \eta_n)$ and $\boldsymbol{\theta} = \{\boldsymbol{\gamma}_k, \pi_{j|c(j)}, \lambda_j, \boldsymbol{\beta}_k, \sigma_j^2;\ j = 1, ..., J, k = 1, ..., K\}$ are estimated by maximizing the classification log-likelihood function,

$$\begin{aligned} L(\boldsymbol{\theta}, \boldsymbol{\eta}) &= \sum_{i=1}^{n} \log \pi_{\eta_i}(\mathbf{w}_i, \boldsymbol{\gamma}) f_{\eta_i}(\mathbf{y}_i|\mathbf{x}_i) \\ &= \sum_{i=1}^{n} \log \left[\pi_{\eta_i}(\mathbf{w}_i, \boldsymbol{\gamma}) \sum_{j:c(j)=\eta_i} \pi_{j|c(j)} f(\mathbf{y}_i|\lambda_j \mathbf{1} + \mathbf{x}_i \boldsymbol{\beta}_{\eta_i}, \sigma_j^2 \mathbf{I}) \right] \end{aligned} \tag{4.1}$$

using a modified EM algorithm called the classification expectation maximization (CEM) algorithm [12, 87]. The modification involves adding a classification step between the expectation step and maximization step where individuals are assigned to shape clusters.

In order to start the iterative algorithm, the individuals are initially assigned into shape clusters and component groups. Initialization can involve randomly partitioning individuals into components or strategically partitioning individuals into shape clusters using a computationally fast procedure such as the first proposed method of this thesis and then randomly partition the individuals into components.

Let $\boldsymbol{\theta}^{(t)}$ and $\boldsymbol{\eta}^{(t)}$ be the current estimates of the parameters at the tth iteration of the algorithm. The CEM algorithm updates these estimates as follows:

1. Expectation step: For each individual i, compute the posterior probability of being in shape cluster k

$$\alpha_{i,k} = \pi_k(\mathbf{w}_i, \boldsymbol{\gamma}) f_k(\mathbf{y}_i|\mathbf{x}_i) / \sum_{j=1}^{K} \pi_j(\mathbf{w}_i, \boldsymbol{\gamma}) f_j(\mathbf{y}_i|\mathbf{x}_i)$$

   for $i = 1, ..., n$ and $k = 1, ..., K$.

2. Classification step: Hard classify subjects to shape clusters according to $\eta_i^{(t+1)} = \arg\max_k \alpha_{i,k}$.

3. Maximization step: For each shape cluster, use maximum likelihood estimation to update parameter vector $\boldsymbol{\theta}$ by embedding an EM procedure initialized with the current parameter values.

This algorithm increases the classification log-likelihood at each iteration. The statement is below and the proof is in Appendix B.

Theorem 1. *The classification likelihood $L(\boldsymbol{\theta}, \boldsymbol{\eta})$ defined in equation (4.1) is non-decreasing after each update of $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$ by the CEM algorithm. That is, $L(\boldsymbol{\theta}^{(t+1)}, \boldsymbol{\eta}^{(t+1)}) \geq L(\boldsymbol{\theta}^{(t)}, \boldsymbol{\eta}^{(t)})$ for all t.*

In the maximization step of the algorithm, parameters are estimated by maximizing the likelihood for each shape cluster. However, the shape parameters are constrained to be equal for all components within the cluster. To estimate both the component and cluster-specific parameters simultaneously, this thesis uses computational methods suggested by Grün and Leisch [46]. For shape cluster k, the outcome vector for subject i is temporarily replaced by a vector of $\mathbf{y}_i$ repeated J_k times. The new design matrix is set equal to $(\mathbf{I}_{J_k} \otimes \mathbf{1}_{m_i}, \mathbf{1}_{J_k} \otimes \mathbf{x}_i)$ where $\otimes$ refers to the Kronecker product. Lastly, the covariance matrix structure is block diagonal with $\sigma_j^2 \mathbf{I}_{m_i}$ in each block for j that satisfy $c(j) = k$. The likelihood function is maximized with respect to the parameters for cluster k and its components using profile likelihoods and a standard numerical optimization routine.

With this complex mixture structure, it is necessary to select the number of shape clusters as well as the number of components for each cluster k. It is recommended to fix K and then use model selection criteria to select J_k for each $k = 1, ..., K$. As suggested by Li [74], the BIC is used to select J_k even though the regularity conditions do not hold as the criteria has been shown to be a useful informal guide in practice.

Robust standard errors for the parameter estimates can be found following the same procedure as Boldea and Magnus [7].

4.3 Vertical shifted mixture model

The goal of cluster analysis is to partition a collection of individuals into homogeneous groups. In this thesis, similar individuals are those with the same outcome shape pattern over time. If two curves only differ by vertical shifts, they are placed in the same group. The first proposed method uses derivatives to implicitly remove the level and the second method directly models the variability in the level with an additional layer of mixture models. In this method, I consider subtracting the subject-specific mean from the outcome measurements to remove the level. A finite mixture of densities with a mean shape curve and a covariance function is fit to the vertically shifted data.

4.3.1 Recent work

A finite mixture model is a standard method for clustering multivariate data [36] and has been used for longitudinal applications [98, 62]. See Chapter 1 for an extensive summary of finite mixture models. However, for longitudinal data, the models are commonly used for the observed data without much regard to the goal of clustering by shape.

I suggest removing the level by subtracting out the mean outcome level prior to modeling. Subtracting the mean is not a novel idea in statistics or even cluster analysis. In fact, experimental data such as gene expression microarrays are often normalized to compensate for variability in the measurement device between samples. In cluster analysis of multivariate data, it is recommended that each variable is standardized by subtracting the mean and dividing by the standard deviation within each variable so the variables are in comparable units and equally contribute to the grouping process. This is not recommended for the longitudinal setting where each variable is a repeated measurement at a different time point.

To compare shapes, we want to maintain the original scale of the data since the relationship between measurements within individuals is of interest. A translation via centering [14] or vertically shifting preserves the shape of the data over time. In general, pre-processing the data can provide a path to answering the research question but any transformation of the data should be carefully studied for potential unintended consequences.

4.3.2 Model specification

Let $\mathbf{y}_i = (y_{i1}, ..., y_{im_i})$ denote an outcome vector of repeated observations for individual i, $i = 1, ..., n$. The vector of corresponding times of observation for individual i is denoted as $\mathbf{t}_i = (t_{i1}, ..., t_{im_i})$ and $\mathbf{w}_i$ is a q-length design vector based on time-fixed factors that are typically collected at or before time t_{i1}. I assume that there are K mean shape functions $\mu_k(t)$ in the population such that the outcome vector for individual i in shape group k is realization of

$$\mathbf{y}_i = \lambda_i \mathbf{1}_{m_i} + \boldsymbol{\mu}_i + \boldsymbol{\epsilon}_i, \quad \lambda_i \sim F_\lambda, \quad \boldsymbol{\epsilon}_i \sim N(0, \boldsymbol{\Sigma}_k)$$

where F_λ is a probability distribution, $\mathbf{1}_{m_i}$ is an m_i-length vector of 1's, and $\mu_{ij} = \mu_k(t_{ij})$ is the jth element of a m_i-length vector of mean values evaluated at the observation times, $\mathbf{t}_i$. The outcome vector is determined by a mean shape function, a random intercept, and potentially correlated random errors. The probability of having a particular shape could depend on baseline covariates. Let $\bar{y}_i = m_i^{-1} \sum_{j=1}^{m_i} y_{ij} = \lambda_i + \bar{\mu}_i + \bar{\epsilon}_i$ be the mean of the outcome measurements for individual i. This measure of the vertical level of the data vector can be removed by applying a linear transformation, $\mathbf{A}_i = \mathbf{I}_{m_i} - m_i^{-1} \mathbf{1}_{m_i} \mathbf{1}_{m_i}^T$, to the vector of observations. The vertically shifted vector for individual i equals

$$\begin{aligned}
\mathbf{y}_i^* &= \mathbf{A}_i \mathbf{y}_i \\
&= \mathbf{A}_i(\lambda_i \mathbf{1}_{m_i} + \boldsymbol{\mu}_{ik} + \boldsymbol{\epsilon}_i) \\
&= \mathbf{A}_i(\boldsymbol{\mu}_i + \boldsymbol{\epsilon}_i) \\
&= \boldsymbol{\mu}_i - \bar{\mu}_i + \boldsymbol{\epsilon}_i - \bar{\epsilon}_i.
\end{aligned}$$

Multiplying the symmetric matrix $\mathbf{A}_i$ to the vector $\mathbf{y}_i$ subtracts the individual mean $\bar{y}_i$ from each element $\mathbf{y}_i$. This results in the removal of the random intercept λ_i, leaving the mean function evaluated at the observation times plus random error shifted by a random constant $\bar{\mu}_i + \bar{\epsilon}_i$. Clearly, we do not have to worry about F_λ, the distribution of the random intercept, or any other time-fixed factors that only impact the level of the outcome.

Once the level is removed, I assume the vertically shifted data $\mathbf{y}_i^*$ follow a Gaussian mixture of K groups with mean shape functions and random errors. If the observation times are fixed, vertically shifted data generated from the specified model follows this Gaussian mixture. Thus, conditional on observation times $\mathbf{t}$ and baseline covariates $\mathbf{w}$, $\mathbf{y}^*$ is assumed to be a realization from a finite mixture model with density

$$f(\mathbf{y}^*|\mathbf{t}, \mathbf{w}, \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k(\mathbf{w}, \boldsymbol{\gamma}) f_k(\mathbf{y}^*|\mathbf{t}, \boldsymbol{\theta}_k)$$

where $\pi_k(\mathbf{w}, \boldsymbol{\gamma})$ is the prior probability of being in the kth shape component given baseline covariates, $\mathbf{w}$. To allow baseline covariates to impact the probability of having a certain shape pattern over time, the prior probabilities are parameterized using the generalized logit function of the form

$$\pi_k(\mathbf{w}, \boldsymbol{\gamma}) = \frac{\exp(\mathbf{w}^T \boldsymbol{\gamma}_k)}{\sum_{j=1}^{K} \exp(\mathbf{w}^T \boldsymbol{\gamma}_j)}$$

for $k = 1, ..., K$ where $\boldsymbol{\gamma}_k \in \mathbb{R}^q$, $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_K)$, and $\boldsymbol{\gamma}_K = \mathbf{0}$. For continuous outcome vectors, the component densities $f_k(\mathbf{y}^*|\mathbf{t}, \boldsymbol{\theta}_k)$ are multivariate Gaussian densities with mean and covariance dependent on time. The full vector of parameters for the model is $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_K)$.

Mean Structure

Only focusing on shape, the mean is modeled as a smooth function of time represented by a chosen functional basis. If the shape is periodic in nature, a Fourier basis is appropriate. Another common basis is a polynomial basis such as a quadratic or cubic basis. However, this type of basis cannot capture complex shapes with local changes with only a few parameters like a B-spline basis.

A B-spline function of order p with L internal knots, $\tau_1, ..., \tau_L$, is defined by a linear combination of coefficients and B-spline basis functions

$$\mu(t) = \sum_{j=1}^{L+p} \beta_j B_{j,p}(t)$$

where the basis functions $B_{j,p}(t)$ are defined iteratively [24, 17] (see Section 4.1.2 for more details). Values from the pth order B-spline basis functions taken at observation times $\mathbf{t}_i$ can be used in a design matrix $\mathbf{x}_i$ to linearly model the mean vector. Thus, the mean of the kth shape cluster is approximated by the linear function $\mu_k(t) = \sum_{j=1}^{L+p} \beta_{k,j} B_{j,p}(t)$. In the multivariate form, the mean vector at observation times $\mathbf{t}_i$ equals $\mathbf{x}_i \boldsymbol{\beta}_k$ where $\boldsymbol{\beta}_k = (\beta_{k,1}, ..., \beta_{k,L+p})$.

Covariance Structure

There are many potential assumptions to be made about the covariance matrix. Here, we allow the covariances to differ between clusters. Since it is common for longitudinal data to have sparse, irregular time sampling, we need to impose structure on the covariance matrix to allow for parameter estimation as described by Jennrich and Schluchter [60] in their seminal paper. A common structure is conditional independence with constant variance where $\mathbf{\Sigma}_k = \sigma_k^2 \mathbf{I}_{m_i}$. This is typically an unrealistic assumption for longitudinal data since there is inherent dependence between repeated measures on the same unit. Compound symmetry, which is also known as exchangeable correlation, is a popular correlation structure in longitudinal analysis where all repeated measures are equally correlated. This is typically paired with constant variance such that $\mathbf{\Sigma}_k = \sigma_k^2(\rho_k \mathbf{1}_{m_i}\mathbf{1}_{m_i}^T + (1-\rho_k)\mathbf{I}_{m_i})$ where $-1 \leq \rho_k \leq 1$ is the correlation between any two distinct measurements within an individual. This dependence structure describes the resulting correlation matrix of a random intercept model.

Another structure that provides a compromise is the exponential correlation structure in which the dependence decays as the time between observations increases such that the element in the jth row and lth column of $\mathbf{\Sigma}_k$ is $\sigma_k^2 \exp(-|t_{ij} - t_{il}|/r_k)$ where $r_k > 0$ is the range of the dependence. If the range r_k is small, the correlation decays quickly, but if r_k is large, there is long range dependence between measurements within an individual. This structure is similar to the correlation matrix generated from a continuous autoregressive model of order one such that the element in the jth row and lth column of $\mathbf{\Sigma}_k$ is $\sigma^2 \rho_k^{|t_{ij} - t_{ill}|}$ where ρ_k is the correlation for measurements observed one unit of time apart. If $\rho_k = \exp(-1/r_k)$, then the two parameterization result in the same structure as long as the correlation between two measures is constrained to be positive. This is a reasonable assumption for longitudinal data in the original form but it many not be acceptable for the transformed data as discussed later. It is important to model the covariance structure correctly as removing the vertical level increases the potential overlap between shape components (see Chapter 2).

4.3.3 Implementation

Given a collection of independent observed outcome vectors from n individuals, $\mathbf{y}_1, ..., \mathbf{y}_n$, we remove the level by subtracting the subject-specific mean from the observed outcome vector, $\mathbf{y}_i^* = \mathbf{y}_i - \bar{y}_i$ for $i = 1, ..., n$. After a visual inspection of the data, the order of the spline and the number and location of internal knots for the B-spline mean structure is chosen. Adding knots and increasing the order of the spline functions flexibly accommodates the twists and turns of the mean patterns but also increases the number of parameters. The knots can be placed at local maxima, minima, and inflection points of the overall trends [34] or at sample quantiles based on the sampling times of all the observations [120]. Design matrices $\mathbf{x}_i$ are calculated using widely available B-spline algorithms for $i = 1, ..., n$.

Under the assumption that $\mathbf{y}_1^*, ..., \mathbf{y}_n^*$ are independent realizations from the mixture distribution $f(\mathbf{y}^*|\mathbf{t}, \mathbf{w}, \boldsymbol{\theta})$ defined in Section 4.3.2, the log-likelihood function for the parameter

vector $\boldsymbol{\theta}$ is given by

$$\log L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(\mathbf{y}_i^* | \mathbf{t}_i, \mathbf{w}_i, \boldsymbol{\theta}).$$

The maximum likelihood estimate of $\boldsymbol{\theta}$ is obtained by finding an appropriate root of the score equation, $\partial \log L(\boldsymbol{\theta})/\partial \boldsymbol{\theta} = \mathbf{0}$. Solutions of this equation corresponding to local maxima can be found iteratively through the EM algorithm [25] (see Section 1.4.2 for technical details about the EM algorithm).

Estimation requires the number of clusters to be known. In practice, this is not the case and the value of K is chosen. The most popular way to choose K is by setting a maximum value such that $K_{max} < n$, fitting the model under all values of $K = 2, ..., K_{max}$, and choosing the value that optimizes a chosen criteria. In this thesis, I use is the Bayesian Information Criterion (BIC) [124], defined as

$$BIC = -2 \log L(\hat{\boldsymbol{\theta}}) - d \log(n)$$

where d is the length of $\boldsymbol{\theta}$, the number of parameters in the mixture model, and $L(\boldsymbol{\theta})$ is the likelihood function for the parameter vector.

There are issues of identifiability with Gaussian mixture models that can be mitigated through some minor constraints [87]. Next, we explore some unique consequences of vertically shifting the data on modeling and estimation.

Covariance of vertically shifted data

Let $\mathbf{Y} = (Y_1, ..., Y_m)$ be a random vector observed at times $\mathbf{t} = (t_1, ..., t_m)$ such that $\mathbf{Y} = \lambda \mathbf{1}_m + \boldsymbol{\mu} + \boldsymbol{\epsilon}$ where $\lambda \sim F_\lambda$, $\boldsymbol{\mu}$ is a vector of evaluations of a function $\mu(t)$ at times $\mathbf{t}$, and $\boldsymbol{\epsilon} \sim (0, \boldsymbol{\Sigma})$. Let $\boldsymbol{\Sigma} = \mathbf{V}^{1/2}\mathbf{R}(\rho)\mathbf{V}^{1/2}$ where $\mathbf{R}(\rho)$ is an $m \times m$ correlation matrix based on the parameter ρ and potentially the associated observation times, and $\mathbf{V}$ is a $m \times m$ matrix with variances along the diagonal.

One important property of this transformation is that it is not invertible; once the mean is subtracted from the data, the original data cannot be recovered. This has an impact on the correlation structure of the data. The covariance of the transformed random vector after removing the mean vector equals

$$\begin{aligned} Cov(\mathbf{Y}^* - \boldsymbol{\mu}) &= Cov(\mathbf{A}(\lambda \mathbf{1}_m + \boldsymbol{\mu} + \boldsymbol{\epsilon}) - \boldsymbol{\mu}) \\ &= Cov((\mathbf{A} - \mathbf{I}_m)\boldsymbol{\mu} + \mathbf{A}\boldsymbol{\epsilon}). \end{aligned}$$

If the observation times are fixed, then $\boldsymbol{\mu}$ is not random and the covariance matrix $\mathbf{A}Cov(\boldsymbol{\epsilon})\mathbf{A}^T$ is singular since $det(\mathbf{A}) = 0$. However, if the observation times are random, then $\boldsymbol{\mu}$ is a random vector and contributes to the overall variability. To better understand how to model the transformed data, we explore the covariance when the observation times are fixed and random. From this point on, $\mathbf{I}_m$ will be written as $\mathbf{I}$ and $\mathbf{1}_m$ as $\mathbf{1}$ for simplification.

Fixed observation times

If the observation times $\mathbf{t}$ are fixed, then

$$Cov(\mathbf{Y}^* - \boldsymbol{\mu}) = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T$$

where $\boldsymbol{\Sigma}$ is the covariance of the original random errors. If the variance is constant over time, $\mathbf{V} = \sigma^2\mathbf{I}$, and the elements of the original vector are independent, $\mathbf{R}(\rho) = \mathbf{I}$, then the covariance can be written as

$$Cov(\mathbf{Y}^* - \boldsymbol{\mu}) = \sigma^2 \left(\frac{m-1}{m}\right)(a\mathbf{1}\mathbf{1}^T + (1-a)\mathbf{I})$$

where $a = \frac{-1}{m-1}$. Therefore, if the observation times are fixed and the data has independent errors, subtracting the estimated mean induces negative exchangeable correction between the observations of magnitude $\frac{-1}{m-1}$. Additionally, the variance decreases to $\sigma^2\frac{m-1}{m}$.

If the errors in the original data have constant variance, $\mathbf{V} = \sigma^2\mathbf{I}$, and are exchangeable with $\mathbf{R}(\rho) = \rho\mathbf{1}\mathbf{1}^T + (1-\rho)\mathbf{I}$, then the covariance is written as

$$Cov(\mathbf{Y}^* - \boldsymbol{\mu}) = \sigma^2(1-\rho)\left(\frac{m-1}{m}\right)(a\mathbf{1}\mathbf{1}^T + (1-a)\mathbf{I})$$

where $a = \frac{-1}{m-1}$. This transformation maintains the exchangeable structure but with negative correlation on the off diagonal and decreased variance of $\sigma^2(1-\rho)\left(\frac{m-1}{m}\right)$.

On the other hand, if the original correlation is exponential such that the correlation decreases as time lags increases, $Cor(Y_j, Y_l) = \exp(-|t_j - t_l|/\rho)$, the resulting covariance after transformation is not a recognizable structure. In fact, the covariance can no longer be written as a function of time lags. The covariance matrix is a linear combination of the original correlation matrix, column and row means, and the overall mean correlation,

$$Cov(\mathbf{Y}^* - \boldsymbol{\mu}) = \sigma^2\left[\mathbf{R}(\rho) - m^{-1}\mathbf{1}\mathbf{1}^T\mathbf{R}(\rho) - m^{-1}\mathbf{R}(\rho)\mathbf{1}\mathbf{1}^T + m^{-2}\mathbf{1}\mathbf{1}^T\mathbf{R}(\rho)\mathbf{1}\mathbf{1}^T\right]$$

This non-stationary covariance matrix includes negative correlations when the mean of the correlations within each column and within each row are positive and substantial.

We have calculated the covariance of the transformed random vector under three common covariance structures for the original data assuming fixed observation times. All of these covariance matrices are not invertible since $det(\mathbf{A}) = 0$. In particular, if prior to transformation, the errors are independent or exchangeable, the correlation of the resulting transformed data is exchangeable equal to $\frac{-1}{m-1}$. This particular value has significant meaning as it is the lower bound for correlation in an exchangeable matrix. This means that the true parameter value of the correlation for the transformed vector is on the boundary of the parameter space. Therefore, even if the true structure is known, estimating parameters for the true model is difficult. Conditional independence or the exponential structure may be an adequate approximation to regularize the estimation, especially if m is moderately large.

Random observation times

In practice, individuals in a longitudinal study are not typically observed at exactly the same times but rather at sporadic times. When the times are random, $\boldsymbol{\mu}$ is random because the elements are evaluations of the function $\mu(t)$ at random times. Therefore, the transformed vector has variability due to the random times in addition to the errors.

If the covariance of the original errors does not depend on time, then the covariance simplifies to

$$Cov(\mathbf{Y}^* - \boldsymbol{\mu}) = m^{-2}\left(\sum_{j=1}^{m} Var(t_j)[\mu'(E(t_j))]^2\right)\mathbf{1}\mathbf{1}^T + \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T$$

using the delta method assuming the random times $t_1, ..., t_m$ are independent. This matrix is the sum of two non-invertible matrices, which need not be non-invertible. In fact, if the variance of the times and/or the derivative of the deterministic function $\mu(t)$ is large, the positive magnitude of the first matrix may be large enough to counteract negative correlations in the second matrix.

If the original covariance is dependent on the random times, the mean vector and error structure both depend on the time. We explore the impact of transforming the data through empirical simulations. Let the observation times equal random perturbations around specified goal times such that $\mathbf{t}_i = \mathbf{T} + \boldsymbol{\tau}_i$ where $\boldsymbol{\tau}_i \sim N(0, \sigma_\tau^2\mathbf{I})$ and $\mathbf{T} = (1, 2, ..., 9, 10)$. We generate $n = 500$ realizations of the model,

$$\mathbf{y}_i = \boldsymbol{\mu}_i + \boldsymbol{\epsilon}_i \quad \text{where } \boldsymbol{\epsilon}_i \sim N(0, \mathbf{R}_i(\rho))$$

where the vector elements $\mu_{ij} = \mu(t_{ij})$. We repeat the simulation under different assumptions for the mean function and standard deviations of the observation times. Figure 4.3 shows the estimated autocorrelation functions of the deviations of the transformed data from the mean when $\mathbf{R}_i(\rho)$ is an exponential correlation matrix with range parameter $\rho = 2$ under varying conditions for observations times and shape functions. As the variance of the observation times and the magnitude of the derivative mean function increases, the estimated correlation in the transformed space becomes positive. Thus, variability in the observations times can result in covariance structures that are no longer singular.

In practice, if the data are regularly or very close to regularly sample, negative correlations are problematic for estimation and an independence or exponential correlation structure may be the best option. If the data are irregularly sampled, one potential covariance model is an additive model that combines a random intercept with the exponential correlation [26], which may be appropriately flexible to approximate the covariance of the deviations from the mean of the transformed data.

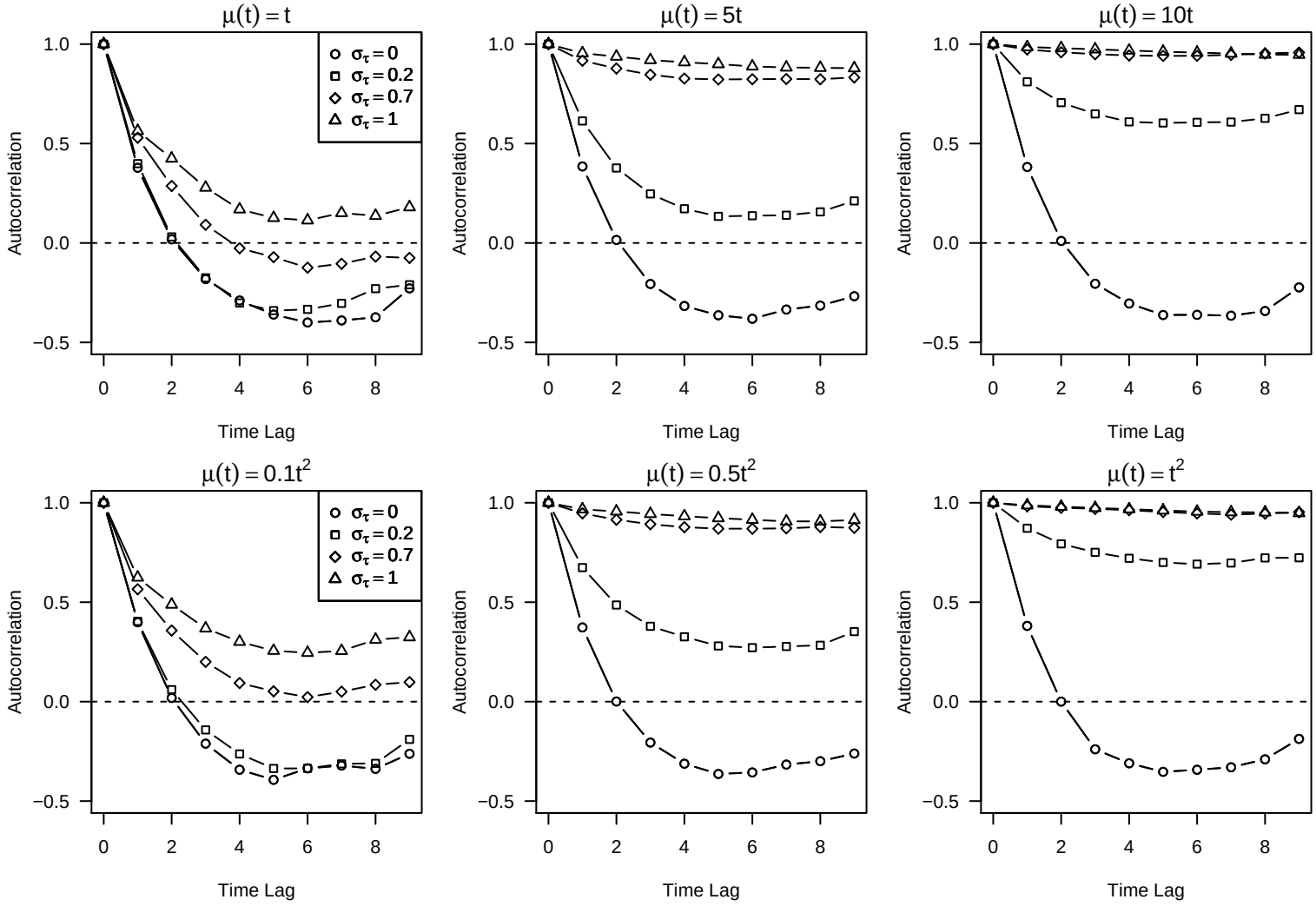

Figure 4.3: Estimated autocorrelation functions of the deviations from the mean from data generated with an exponential correlation error structure and random observation times under different mean functions, $\mu(t)$, and standard deviations of the random time perturbations, σ_τ.

Unbalanced observation times

In addition to the issues of fixed versus random sampling, having an unequal number of observations per subject can impact the estimation of the vertically shifted model. As we saw above, the length of the vector impacts the covariance of the transformed vector. Suppose the outcome vectors for a sample of individuals have the same mean shape and covariance over time, but each individual is observed a different number of times because they were unavailable for an interview or two. Transforming the vectors by subtracting means based on a variety of number of observations induces a different covariance structure for each individual based on the length of outcome vector. If there is quite a bit of variability in the number of observations, it may impact clustering to assume they share the same covariance structure during the estimation/clustering process. However, if the number of observation times is large for all subjects and the observation period is long, then the covariance matrices should be similar.

Additionally, if the unbalanced nature of the data is due to lost to follow-up during a longitudinal study, clustering based on the shape should be done with caution. If the general shape of the curve during the observation period is not measured adequately by the number of observations, it does not make sense to try and cluster those individuals with the rest who have more fully observed curves.

4.4 Discussion

In this chapter, I described three different approaches to the problem of clustering irregularly sampled longitudinal data by shape. The first method focuses on clustering derivative functions. Projecting the data onto a B-spline basis removes noise and provides an estimate of the smooth underlying function. The derivative can then be calculated from the estimated function. This should be an improvement over difference quotients, which simply linearly interpolates the data points with no regard to error. One difficulty with this method is choosing the correct order and knots for the basis so as to not over fit the data when there are only a few data points. While this may improve upon other methods, this approach has limitations. There is no direct way to borrow strength between individuals when estimating their derivative function even if shape is hypothesized to be a common factor. Also, partitioning methods do not lend themselves to analysis of baseline factors since by definition, the algorithm does not provide any uncertainty estimates in group membership.

Rather than ignoring the level, the second approach attempts to directly model the variability in the level by assuming that for each shape group, the distribution for the level can be approximated with a Gaussian mixture model. Assuming a multilayered mixture model provides a probability framework to take into account uncertainty while estimating the relationship between baseline factors and group membership. However, this model requires a large number of individuals in each shape group for the mixture to model the distribution well. It is not clear how robust this method is for small sample and group sizes.

Lastly, the third approach directly removes the level by subtracting individual-specific means prior to modeling. This allows individuals to be compared without making specific assumptions about the distribution of the level while providing the probability framework. There are two difficulties with this method. First, subtracting an observed mean impacts the covariance in a way that makes it harder to model with a known correlation structure. Second, care needs to be taken when there is sparse and irregularly sampling.

I compare these methods in practice with a simulation study in the next chapter.

Chapter 5

Simulations

In this chapter, I compare the three proposed methods presented in the last chapter with competing clustering methods in a simulation study. The goal of this study is to examine the performance of these methods when applied to data of individuals with the same shape trajectory at a variety of levels. No method works well in all situations; therefore, the data are generated under a variety of conditions to test the methods on their robustness to model misspecification, noise, and group overlap. The ability to detect and create homogenous shape groups indicates good performance. The outcomes of interest are the data-driven number of groups, misclassification with reference to the generating shape groups, and the parameter estimates for the cluster mean shape and baseline factors.

First, I describe the data-generating process used throughout the simulation study. Then, I discuss the specific details of the simulation implementation in terms of the variety of data conditions, methods used, and the simulation outcomes of interest. Lastly, I present the results and compare the performance of the methods.

5.1 Data-generating process

When deciding upon the data-generating process, it is important to choose a simple example that addresses the goals of the study and can generalize to other situations. This simulation study is based on a population with three trajectory shapes and three outcome levels. Since complex functions can be locally approximated with linear functions, the three shapes used in the simulation are straight lines with different slopes—positive, negative, and zero. We restrict some shapes to particular levels to induce a relationship between level and

shape. The five mean functions for generating data with three shapes at different levels are

$$\begin{aligned}
\mu_1(t) &= -1 - t && \text{(negative slope, low level)} \\
\mu_2(t) &= 11 - t && \text{(negative slope, high level)} \\
\mu_3(t) &= 0 && \text{(zero slope, middle level)} \\
\mu_4(t) &= -11 + t && \text{(positive slope, low level)} \\
\mu_5(t) &= 1 + t && \text{(positive slope, high level)}
\end{aligned}$$

Individuals follow these mean patterns with varying probabilities that depend on two baseline factors. The first factor w_1 impacts the shape and w_2 impacts the level. Individuals are randomly assigned values of these two binary factors with independent simulated tosses of a fair coin.

Let S be a categorical random variable that indicates the shape/slope group such that $S = 1, 2, 3$ refers to the negative, zero, and positive slope groups, respectively. Conditional on the baseline factors, the probability of being in the kth shape group is

$$P(S = k|w_1) = \frac{\exp(\gamma_{0k} + \gamma_{1k} w_1)}{\sum_{l=1}^{3} \exp(\gamma_{0l} + \gamma_{1l} w_1)}$$

for $k = 1, 2, 3$ where $\gamma_{01} = 2, \gamma_{11} = -4, \gamma_{02} = 1.5, \gamma_{12} = -2, \gamma_{03} = \gamma_{13} = 0$ and $w_1 \in \{0, 1\}$. Since the value of w_1 is determined by a fair coin toss, each shape group has about an equal probability, marginally.

The second factor impacts the level. Let L be a categorial random variable that indicates level group such that $L = 1, 2, 3$ refers to the low, middle, and high group, respectively. Only those who follow the constant function have middle level outcome measures. For those in either the negative or positive slope groups, the chance of the high or low level is

$$P(L = k|S = 1 \text{ or } S = 3, w_2) = \frac{\exp(\zeta_{0k} + \zeta_{1k} w_2)}{\sum_{l \in \{1,3\}} \exp(\zeta_{0l} + \zeta_{1l} w_2)}$$

for $k = 1, 3$ and $w_2 \in \{0, 1\}$ where $\zeta_{01} = 0, \zeta_{11} = 0, \zeta_{03} = -3, \zeta_{13} = 6$. Again, each outcome level group has about the same marginal probability.

For each individual, the chosen mean function is evaluated at five equidistant observation times $t = 1, 3.25, 5.5, 7.75, 10$ that span the period 1 to 10 units and random noise is added to induce variability. The random noise is made up of two independent ingredients: individual-specific level perturbation and time-specific Gaussian measurement error. For individual i $(i = 1, ..., n)$ following the lth mean function, the observed outcome at the jth observation time $(j = 1, .., 5)$ is a realization of

$$y_{ij} = \lambda_i + \mu_l(t_j) + \epsilon_{ij} \quad \text{where} \quad \epsilon_{ij} \overset{iid}{\sim} N(0, \sigma_\epsilon^2), \lambda_i \overset{iid}{\sim} F_\lambda(0, \sigma_\lambda^2)$$

where σ_ϵ is the standard deviation of the measurement error, σ_λ is the standard deviation of the level perturbation, and F_λ is the probability distribution of the level perturbation.

This imposes an exchangeable correlation structure within an individual's outcome vector conditional on the mean function. The distribution of the level perturbation as well as the standard deviations of the random noise are values that vary in the simulation to create different conditions on which to test the clustering methods. The values are chosen to alter the overlap of the groups and the signal to noise ratio and are discussed in the next section.

5.2 Implementation

In this section, I describe the random noise conditions, overview the clustering methods included in this study, and discuss the outcomes of interest used to compare the methods in the simulation.

5.2.1 Simulation conditions

The first ingredient of the random noise is the level perturbation. To test the sensitivity of the methods to distributional assumptions, I use the uniform distribution and the Gaussian distribution to generate the level perturbations. To create conditions with differing amount of overlap between groups, I let $\sigma_\lambda = 2$ or 3 such that the vertical shift between the mean functions with the same shape is about 4 or 6 standard deviations. The second ingredient is the measurement error, the magnitude of which influences the signal to noise ratio. I let $\sigma_\epsilon = 0.5$ or 2 to create two extreme conditions, strong and weak signals relative to noise. The eight possible combinations represent the conditions of the data-generating process in the simulation study.

5.2.2 Clustering methods

For each condition, I generate a data set of $n = 500$ individuals using the process described above and apply each method to cluster the data into $K = 2, 3, 4$, and 5 groups. Each method produces different output making direct comparisons difficult.

Partition methods produce group labels for each individual while model-based methods result in posterior probability estimates for each individual and group. Therefore, I translate the estimated probabilities into group labels by placing individuals into the group that has the highest posterior probability.

Besides grouping results, the group representatives are in different formats. The K-means algorithm results in a mean vector within each group, the interpretation of which depends on the input vectors. On the other hand, I model the mean with spline basis functions in the model-based clustering methods. The model output includes the estimated coefficients for the group-specific mean functions. To make the group representatives from the partition and model-based methods comparable, I transform all of the output into coefficients of a differentiated B-spline function which describes the shape of the curves. I project the data vectors onto a common spline basis, if not already in that form, differentiate the spline

functions, and rearrange terms to calculate the coefficients of the spline function one order lower.

The common spline used throughout the simulation study is a quadratic B-spline with no internal knots and boundary knots at 0.5 and 10.5. A quadratic B-spline basis without internal knots is equivalent to a Bernstein polynomial basis [79] which generalizes the standard polynomial basis; in general, I advocate using a B-spline basis that provides flexibility and is computationally stable. As mentioned above, this function is differentiated and the coefficients are transformed in order to get the coefficients of the derivative B-spline function for final method comparisons. See Section 4.1.2 for more details on derivatives of B-spline functions.

For every clustering method, I estimate the relationship between the two baseline factors, w_1 and w_2, and the group membership. This estimation procedure differs between the methods as it may simultaneously occur while estimating group parameters or after the clustering algorithm is complete. I now overview the clustering methods used in this study.

K-means

The K-means algorithm is a general clustering procedure for numerical vectors. K groups are determined through an iterative assignment and update process to minimize the within-group sum of squares distance to the group mean vectors. Note that the objective function indirectly imposes a spherical shape on the groups in the vector space.

To choose the optimal value of K in this simulation study, I use the silhouette measure, which measures the compactness of clusters, for all applications of the K-means algorithm. See Section 1.3 for more details.

To estimate the relationship between baseline factors and group membership produced from the K-means algorithm, the groups labels are taken as known and become the outcomes for a multinomial logistic regression with the observed baseline factors as explanatory variables. Let c_i be the group label produced by the clustering algorithm and w_{i1}, w_{i2} be the baseline factors for individual i, then the model assumes that

$$\log \frac{P(c_i = k)}{P(c_i = K)} = \gamma_{0k} + \gamma_{1k} w_{i1} + \gamma_{2k} w_{i2}$$

for $k = 1, .., K-1$. Estimates for the coefficients $\boldsymbol{\gamma} = (\gamma_{01}, \gamma_{11}, \gamma_{21}, ..., \gamma_{0\,K-1}, \gamma_{1\,K-1}, \gamma_{2\,K-1})$ are calculated via maximum likelihood estimation as long as the number of individuals in each group is adequately large.

K-means on original data

The K-means algorithm applied to the 500 original data vectors results in groups based on both the level and shape. Both characteristics can play a role since this method is based on squared Euclidean distance of the original outcome measures. In order to compare the estimated group representatives to those of the following methods, the K group mean vectors

are projected onto the common B-spline basis and the coefficients are then transformed into derivative coefficients.

K-means on spline coefficients

To focus on the individual trends without noise, the original data vectors are projected onto the common B-spline basis prior to clustering. This process in effect smoothes each individual trajectory into a quadratic function. The vector of basis coefficients for each individual are used as the input for the K-means algorithm. This technique originates in the functional data analysis literature [112] and has been used as a way to decrease the dimension of the data prior to clustering.

K-means on derivative spline coefficients

This is the first proposed method of this thesis. A slight adjustment to the previous method allows the clustering to be based on the derivative function, which describes the shape, rather than the original function. Before the algorithm is applied, the spline coefficients are transformed into coefficients of the derivative function written as a B-spline function of one lower order.

K-means on difference quotients

Another way of clustering on the basis of derivatives is to calculate difference quotients by taking the pair-wise differences within the ordered data vector and divide by the time between observations,

$$\Delta_{ij} = \frac{y_{ij+1} - y_{ij}}{t_{j+1} - t_j}$$

for $i = 1, ..., n$ and $j = 1, ..., 4$. The vector of difference quotients, $\Delta_i = (\Delta_{i1}, ..., \Delta_{i4})$, is used as the input for the K-means algorithm.

PAM with correlation measure

Partitioning around medoids (PAM) is a variant of the K-means algorithm. Rather than focusing on the distance to the group mean or centroid, this method uses the middle data vector or medoid as the representative of the group, which results in a more robust procedure less sensitive to outliers. The algorithm is general enough to allow for other measures of dissimilarity beyond squared Euclidean distance. For this study, the chosen dissimilarity measure is a linear transformation of the Pearson's correlation coefficient, which measures the linear association between two vectors and is believed to detect similarity in the shape of longitudinal data [15, 30, 14]. The dissimilarity between two vectors is defined as

$$d(\mathbf{y}_1, \mathbf{y}_2) = 1 - Cor(\mathbf{y}_1, \mathbf{y}_2)$$

where $Cor(\mathbf{y}_1, \mathbf{y}_2)$ is defined in Section 1.3. Perfect positive linear correlation results in a dissimilarity of 0 and perfect negative correlation results in a value of 2. If a vector is an additive shift of another vector, they are deemed to be similar. This is one desired property for a clustering method in this thesis. However, the same goes for multiplicative transformations of data vectors which change the shape of the trajectory. Therefore, this measure cannot detect magnitude differences in the slope of trend lines and cannot distinguish between a horizontal line and one with a positive or negative slope.

To choose the number of groups and estimate the relationship between baseline factors and group membership, the same procedures as the K-means algorithm are used.

Finite mixture model

In contrast to the partition methods mentioned above, using a probability model allows for uncertainty in the model parameters and group membership. The data are fit to the model using the EM algorithm in order to maximize the likelihood function (see Section 1.4.2 for more details). Rather than being separated into K groups, mixture models allow individuals to contribute to every group to some degrees. For example, if two groups are not well separated, then an individual in the overlap of the groups could contribute equally to the estimation of both groups' parameters.

For this simulation study, I generally assume the K components of the finite mixture model are multivariate Gaussian with a component-specific mean function that can be expressed as a quadratic B-spline with boundary knots at 0.5 and 10.5. The estimated coefficients for the mean parameters are then transformed into the coefficients for the derivative spline function for the final comparison. Additionally, the prior probabilities of the components are parameterized using a generalized logit function such that the logit is equal to a linear combination of the two baseline factors. For individual i,

$$\log \frac{P(c_i = k)}{P(c_i = K)} = \log \frac{\pi_k(w_{i1}, w_{i2}, \boldsymbol{\gamma})}{\pi_K(w_{i1}, w_{i2}, \boldsymbol{\gamma})} = \gamma_{0k} + \gamma_{1k} w_{i1} + \gamma_{2k} w_{i2}.$$

The parameters, $\boldsymbol{\gamma} = (\gamma_{01}, ..., \gamma_{2\,K-1})$ are estimated simultaneously with the mean and covariance parameters. The number of components or groups is selected by minimizing the model selection criteria, BIC, over model fits with $K = 2, 3, 4, 5$ (see Chapter 1 for more details).

Finite mixture model with independence

For the standard model fit to the original data, the covariance structure is assumed to be independent such that the variance is constant over time but differs between components. Therefore, the only covariance parameters in the model are the component-specific variances.

Finite mixture model with exponential correlation

As discussed in Chapter 2, assuming independence for longitudinal data when there is inherent dependence between repeated measures can result in bias and potentially misleading clustering results. Therefore, we use an exponential correlation structure that maintains the stationarity of the covariance structure but allows the correlation to decay as the time between observations increases. The number of covariance parameters increases to two per component: variance and range of dependence.

Multilayer mixture model

This is the second proposed method of this thesis. In generalizing the standard finite mixture model to allow shape groups to have different level groups, I assume that each shape component of the mixture model is a mixture of level components assumed to be multivariate Gaussian. In order to minimize the number of parameters to be estimated, I assume independence between repeated measures. For this two-layered model, the prior probabilities for the shape components, rather than those of the level components, are parameterized to relate baseline factors to group membership. In addition to specifying the number of shape components, the number of level components within each shape component need to be set prior to estimating the parameters. It is difficult to choose both of these values in an automatic way; therefore, we fix the number of shape components to be $K = 3$ since in a simulation it is known a priori. Then we use the BIC to select the number of level components within the shape components [74].

Vertically shifted independence mixture model

This is the third proposed method of this thesis. Rather than trying to explicitly model the level, I remove the level prior to fitting the model by subtracting the individual-specific mean. Before the model is fit to the data, the trajectories are vertically shifted so they lie in the same range. The hope is that shape drives the group development once the level is removed. As mentioned in Chapter 4, subtracting the mean impacts the covariance structure and can result in negative correlation between repeated measures, which is hard to model with a legitimate correlation function. In this simulation, the correlation structure of the original data for this data-generating process is exchangeable and observation times are fixed. Therefore, we know that the correlation of the transformed errors is exchangeable with magnitude $\frac{-1}{4}$. The true correlation is on the boundary of the parameter space; therefore, I use the independence correlation structure to model the leftover correlation after removing the mean. The same estimation and modeling procedures as the standard finite mixture model apply.

Vertically shifted exponential mixture model

Rather than assuming independence for the transformed vector, I use the exponential correlation structure to model the error structure after removing the mean.

5.2.3 Outcomes of interest

For each method and condition, I generate a data set of $n = 500$ individuals and calculate the following outcomes of interest. This is repeated so that we get 500 unique data sets under each condition on which we can apply each method and summarize the results.

Number of groups

For each replication, I use the method-specific procedure of choosing the final data-driven number of groups and record the final value of K for each data set. Note that this K is not chosen for the multilayer mixture as the number of shape groups is fixed to be equal to three.

Misclassification rate

To detect whether the method discovers the underlying shape structure, I fix $K = 3$ and compare the cluster memberships labels to the true shape group membership using a contingency table. The cluster label columns of the contingency table are reordered such that the trace of the 3 by 3 inner matrix is maximized to match the clusters to the true shape groups. The sum of the off-diagonal elements of the newly permuted matrix represents the number of misclassified individuals. This value is divided by the sample size to get the rate. The best method results in zero misclassifications. To summarize the replications, I present the mean misclassification rate of all 500 replications.

Adjusted Rand Index

Another way to measure the similarity between the produced clusters and the true shape clusters is the Adjusted Rand Index [57, 90]. This measure assesses the agreement of two partitions. In our case, the true shape groups form the reference partition and I compare it to the cluster partition produced by the method when $K = 3$.

Given a set of n objects suppose $U = \{u_1, ..., u_K\}$ and $V = \{v_1, ..., v_{K'}\}$ represent two different partitions of the objects such that $\cup_{i=1}^{K} u_i = \cup_{j=1}^{K'} v_j$ and $u_i \cap u_{i'} = \emptyset = v_j \cap v_{j'}$ for $1 \leq i \neq i' \leq K$ and $1 \leq j \neq j' \leq K'$. The amount of overlap between the two partitions can be written in the form of a contingency table with elements n_{ij} denoting the number of objects that are common to groups u_i and v_j and $n_{i\cdot}$ and $n_{\cdot j}$ denoting row and column sums, respectively. Let a be the number of pairs of objects that are placed in the same group in U and in the same cluster in V, b be the number of pairs of objects in the same group in U but not in the same cluster in V, c be the number of pairs of objects in the same cluster

in V but not in the same group in U, and d be the number of pairs of objects in different groups and different clusters in both partitions. The quantities a and d are interpreted as agreements and b and c as disagreements. The Rand Index [113] equals $\frac{a+d}{a+b+c+d} = \frac{a+d}{\binom{n}{2}}$. The index can yield a value between 0 and 1, where 1 indicates that the clusters are exactly the same.

One issue with the Rand Index is that the expected value of the index does not take on a constant value and thus it is hard to evaluate raw indices. Hubert and Arabie [57] suggested using the generalized hypergeometric distribution for the null model when the partitions are picked at random. This leads to the Adjusted Rand Index, which is of the form $\frac{\text{Index}-\text{Expected Index}}{\text{Max Index}-\text{Expected Index}}$. The index is bounded above by 1 and takes on the value 0 when the index equals its expected value. In the contingency table notation, the Adjusted Rand Index can be simplified to

$$\frac{\sum_{ij}\binom{n_{ij}}{2} - \sum_i \binom{n_{i\cdot}}{2}\sum_i \binom{n_{\cdot j}}{2}/\binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{n_{i\cdot}}{2} + \sum_j \binom{n_{\cdot j}}{2}\right] - \sum_i \binom{n_{i\cdot}}{2}\sum_i \binom{n_{\cdot j}}{2}/\binom{n}{2}}$$

A method that performs well has an Adjusted Rand Index value close to 1. To summarize the replications, I present the mean Adjusted Rand Index for the 500 replications.

Mean parameters

As mentioned in the brief descriptions of the methods used in the simulation, the format for the group representative differs between each method. Six of the ten methods use the B-spline basis to model the mean as a function and one method explicitly uses the derivative of the B-spline function. It is not possible to map a derivative function back to the original function. Therefore, I map the clustering representatives when $K = 3$ to the derivative B-spline function so that I can compare the shape of estimated representatives within each group. For reference, the true mean curves are projected onto the B-spline basis and then differentiated.

Just as with the misclassification rate, it is necessary to match the group labels to the shape groups of the true generating distribution. This is completed by permuting the columns until the trace of the 3 by 3 inner matrix of the contingency table between the true shape labels and the clustering group labels is maximized. Once the groups are matched to the truth, the mean squared error (MSE) is estimated for each derivative coefficient using the 500 replications.

Baseline factors

Once groups are discovered, the natural question is to ask who are in the groups and how else do they differ? For each condition, method, and replication, the coefficients of the generalized logit $\{\gamma_{k0}, \gamma_{k1}, \gamma_{k2}\}$ for $k = 1, ..., K-1$ are estimated either simultaneously in the model setting or a posteriori for the partition methods when $K = 3$. To summarize the

estimates from all of the replications, I plot the density of the values and compare them to the true generating values. The clustering is unstable if the density plot is multimodal.

5.3 Results

Tables 5.1, 5.2, and 5.3 summarize the simulations in terms of the chosen number of groups, average misclassification rate, and average Adjusted Rand Index. It is clear from these tables that the three standard methods—the K-means algorithm applied to original vectors, K-means on spline coefficients, and the independent Gaussian mixture model—do not select three groups as the optimal number of groups (Table 5.1). For almost all the conditions, the K-means algorithm on the original data obscures any shape differences in the data by vertically splitting the individuals in half: a high and a low group. If the noise in the data is smoothed out prior to clustering by applying K-means to the B-spline coefficients, five groups is favored except when high variance in the level results in overlap between groups and the algorithm selects two groups. The BIC consistently chooses five groups for the independent mixture model. It recognizes the five mean functions but no similarities in shape or level. All three of these methods also do not perform well when forced to have the true number of shape groups, $K = 3$. Only about 50-60% of the data is correctly specified in terms of the generating shape groups. Small average values for the Adjusted Rand Index confirm that groups from these methods do not match the shape groups.

If the correlation structure of the finite mixture model is generalized to exponential correlation, there is more variability in the number of groups chosen. When the time-specific measurement error is small ($\sigma_\epsilon = 0.5$), the correct number of groups is chosen about 20% of the time (Table 5.1). Under these same circumstances, there is perfect alignment with the generating shape groups when the method is forced to created three groups. However, when the magnitude of the noise is large ($\sigma_\epsilon = 2$), three is never chosen as the optimal number of groups and the shape groups are not discovered when $K = 3$. So even though this method does not directly cluster based on shape, having a general enough correlation structure allows the method to pick up the correct groups when the number of group is known and the measurement error is small. However, it does not choose the correct number of groups in the majority of simulations.

Of the established methods that are intended to group on shape, K-means on difference quotients picks the correct number of groups only if the magnitude of the measurement error is small relative to the strength of the signal (Table 5.2). If the variability around the individual mean is large, the method chooses two groups and misclassifies about 38% of the individuals when forced to have three groups. Using the correlation dissimilarity measure with the PAM algorithm gives slightly better results with only 24-27% misclassification, but it never chooses three groups.

Now, I compare the three proposed methods (Table 5.3). Using K-means with the derivative spline coefficients is a slight improvement over the difference quotients such that the true number of groups is chosen more frequently even under conditions of higher measurement

error. However, this method did not decrease the misclassification rate or increase the Adjusted Rand Index under those same conditions.

The multilayer mixture model is successful in correctly classifying individuals into shape groups when the level components are well separated ($\sigma_\epsilon = 0.5, \sigma_\lambda = 2$). The method seems to be robust to some misspecification as the results do not degrade when the level is generated by the uniform distribution in comparison with the assumed Gaussian model. However, if the distribution of the level is not distinctly bimodal, the misclassification rate is quite high.

Lastly, the method that prevails amongst the competition is the vertically shifted mixture model. In this case, either correlation assumption works well. For every condition, the method chose three groups as the optimal partition at least 99% of the time and when forced to have $K = 3$, the method discovered the shape groups with little misclassification. Only when the measurement error is large ($\sigma_\epsilon = 2$) did the method misclassify 5% (about 25 individuals) in terms of shape.

To compare the group representatives from all methods with the true mean shapes used to generate the data, we transformed the estimates when $K = 3$ to coefficients of the derivative spline as previously described. There are two coefficients per group and I denote them as α_{11}, α_{21} for group 1, α_{12}, α_{22} for group 2, and α_{13}, α_{23} for group 3. Tables 5.4, 5.5, and 5.6 present the mean squared errors for the derivative spline coefficients under the simulation conditions for all clustering methods. When the clustering algorithm finds the shape groups correctly, the MSE is very close to zero for all coefficients. This occurs for the vertically shifted mixture under all conditions and both derivative-based methods and the exponential mixture model when the measurement error is small. Additionally, the coefficients in the multilayer mixture model have a small MSE only when the level groups are well separated.

In general, the standard methods find one group with a horizontal shape on average indicated by values close to zero for α_{21} and α_{22}. Large MSE values (close to or greater than 1) indicate clustering instability in that the method finds fundamentally different groups for every randomly generated data set resulting in great variation in the estimates (Table 5.5 and 5.6). This occurs for a variety of reasons. In the case of the derivative spline coefficients, overfitting the five observations results in high variability. For the correlation-based method, the inconsistent dissimilarity between noisy horizontal vectors causes extra variability.

I have determined that some of the methods discover the shape groups under conditions of small measurement error by comparing the group labels to the true generating labels comparing the estimates of the mean shapes with known generating values. Lastly, I want to know how the choice of method impacts the estimation of the relationship between baseline factors and group membership. For simplification, we present density plots of the estimated coefficients for group 1 and 2 for the factors w_1 and w_2 for only three methods—K-means on the original data, K-means on the difference quotients, and vertically shifted exponential mixture models—after setting $K = 3$. As a reminder, the first baseline factor w_1 was used to generate shape groups and the second w_2 was used to generate level groups.

It is clear that the K-means on the original data picks up the relationship between the group membership and the second baseline factor w_2 (Figure 5.1). The density plots with dashed lines for the coefficients of w_2 are concentrated at non-zero values while the solid line

densities for the coefficients of w_1 revolve around zero. For the two uniform conditions with $\sigma_\lambda = 2$, the density plots for the coefficients of w_1 are bimodal, which indicates some variability in the groups chosen. Overall, K-means picks up the level groups since the coefficients for w_2 are significantly non-zero.

On the other hand, the K-means on the difference quotients finds a relationship with w_1 (Figure 5.2). The average coefficient values for w_1 are non-zero while the coefficients for w_2 revolve around zero. However, when $\sigma_\epsilon = 2$, one of the density plots is bimodal indicating variability which was indicated by a higher misclassification rate.

Lastly, with the vertically shifting exponential mixture model (Figure 5.3), there is a significant relationship between group membership and w_1, the factor that was used to generate shape groups. The second factor is estimated to have no relationship with the final grouping.

F_λ	σ_ϵ	σ_λ	$K=2$	$K=3$	$K=4$	$K=5$	MR	ARI
K-means on Original Data								
Uniform	0.50	2	0	0	0	500	0.42	0.25
Uniform	2.00	2	492	0	0	8	0.42	0.25
Uniform	0.50	3	500	0	0	0	0.50	0.09
Uniform	2.00	3	500	0	0	0	0.51	0.09
Gaussian	0.50	2	0	0	0	500	0.38	0.33
Gaussian	2.00	2	273	0	0	227	0.39	0.29
Gaussian	0.50	3	412	0	0	88	0.46	0.16
Gaussian	2.00	3	500	0	0	0	0.47	0.14
K-means on Splines Coefficients								
Uniform	0.50	2	0	0	0	500	0.46	0.19
Uniform	2.00	2	31	0	1	468	0.47	0.18
Uniform	0.50	3	189	0	10	301	0.48	0.17
Uniform	2.00	3	500	0	0	0	0.48	0.15
Gaussian	0.50	2	0	0	0	500	0.44	0.22
Gaussian	2.00	2	0	0	0	500	0.45	0.20
Gaussian	0.50	3	1	0	0	499	0.46	0.18
Gaussian	2.00	3	490	0	0	10	0.47	0.16
Independent Mixture								
Uniform	0.50	2	0	0	2	498	0.44	0.22
Uniform	2.00	2	0	0	2	498	0.40	0.28
Uniform	0.50	3	0	0	1	499	0.49	0.12
Uniform	2.00	3	0	0	2	498	0.49	0.12
Gaussian	0.50	2	0	0	10	490	0.41	0.25
Gaussian	2.00	2	0	0	24	476	0.39	0.29
Gaussian	0.50	3	0	0	1	499	0.45	0.17
Gaussian	2.00	3	0	0	5	495	0.45	0.17
Exponential Mixture								
Uniform	0.50	2	0	65	171	264	0.00	1.00
Uniform	2.00	2	0	0	23	477	0.40	0.40
Uniform	0.50	3	0	113	157	230	0.00	1.00
Uniform	2.00	3	0	0	37	463	0.41	0.35
Gaussian	0.50	2	0	114	136	250	0.00	1.00
Gaussian	2.00	2	0	0	34	466	0.41	0.38
Gaussian	0.50	3	0	141	121	238	0.00	1.00
Gaussian	2.00	3	0	1	40	459	0.41	0.35

Table 5.1: The number of times each value of K was chosen and the average misclassification rate (MR) and average Adjusted Rand Index (ARI) when $K = 3$ for 500 replications of standard clustering methods applied to data generated under different conditions for the F_λ and the standard deviation of ϵ and λ.

F_λ	σ_ϵ	σ_λ	$K=2$	$K=3$	$K=4$	$K=5$	MR	ARI
K-means on Difference Quotients								
Uniform	0.50	2	0	500	0	0	0.00	1.00
Uniform	2.00	2	480	20	0	0	0.38	0.31
Uniform	0.50	3	0	500	0	0	0.00	1.00
Uniform	2.00	3	480	20	0	0	0.38	0.31
Gaussian	0.50	2	0	500	0	0	0.00	1.00
Gaussian	2.00	2	483	17	0	0	0.38	0.31
Gaussian	0.50	3	0	500	0	0	0.00	1.00
Gaussian	2.00	3	483	17	0	0	0.38	0.31
Correlation-based PAM								
Uniform	0.50	2	380	0	0	120	0.24	0.49
Uniform	2.00	2	500	0	0	0	0.27	0.46
Uniform	0.50	3	380	0	0	120	0.24	0.49
Uniform	2.00	3	500	0	0	0	0.27	0.46
Gaussian	0.50	2	403	0	0	97	0.25	0.49
Gaussian	2.00	2	500	0	0	0	0.27	0.45
Gaussian	0.50	3	403	0	0	97	0.25	0.49
Gaussian	2.00	3	500	0	0	0	0.27	0.45

Table 5.2: The number of times each value of K was chosen and the average misclassification rate (MR) and average Adjusted Rand Index (ARI) when $K = 3$ for 500 replications of clustering methods intended to group by shape applied to data generated under different conditions for the F_λ and the standard deviation of ϵ and λ.

F_λ	σ_ϵ	σ_λ	$K=2$	$K=3$	$K=4$	$K=5$	MR	ARI
K-means on Derivative Splines Coefficients								
Uniform	0.50	2	0	500	0	0	0.00	1.00
Uniform	2.00	2	15	189	164	132	0.39	0.27
Uniform	0.50	3	0	500	0	0	0.00	1.00
Uniform	2.00	3	15	189	164	132	0.39	0.27
Gaussian	0.50	2	0	500	0	0	0.00	1.00
Gaussian	2.00	2	18	191	171	120	0.39	0.27
Gaussian	0.50	3	0	500	0	0	0.00	1.00
Gaussian	2.00	3	18	191	171	120	0.39	0.27
Multilayer Mixture								
Uniform	0.50	2					0.07	0.82
Uniform	2.00	2					0.29	0.35
Uniform	0.50	3					0.33	0.32
Uniform	2.00	3					0.46	0.12
Gaussian	0.50	2					0.09	0.75
Gaussian	2.00	2					0.30	0.34
Gaussian	0.50	3					0.31	0.34
Gaussian	2.00	3					0.41	0.17
Vertically Shifted Independent Mixture								
Uniform	0.50	2	0	499	1	0	0.00	1.00
Uniform	2.00	2	0	499	1	0	0.05	0.87
Uniform	0.50	3	0	499	1	0	0.00	1.00
Uniform	2.00	3	0	499	1	0	0.05	0.87
Gaussian	0.50	2	0	499	0	1	0.00	1.00
Gaussian	2.00	2	0	498	2	0	0.05	0.87
Gaussian	0.50	3	0	499	0	1	0.00	1.00
Gaussian	2.00	3	0	498	2	0	0.05	0.87
Vertically Shifted Exponential Mixture								
Uniform	0.50	2	0	500	0	0	0.00	1.00
Uniform	2.00	2	0	500	0	0	0.05	0.87
Uniform	0.50	3	0	500	0	0	0.00	1.00
Uniform	2.00	3	0	500	0	0	0.05	0.87
Gaussian	0.50	2	0	500	0	0	0.00	1.00
Gaussian	2.00	2	0	499	1	0	0.05	0.87
Gaussian	0.50	3	0	500	0	0	0.00	1.00
Gaussian	2.00	3	0	499	1	0	0.05	0.87

Table 5.3: The number of times each value of K was chosen and the average misclassification rate (MR) and average Adjusted Rand Index (ARI) when $K = 3$ for 500 replications of proposed clustering methods applied to data generated under different conditions for the F_λ and the standard deviation of ϵ and λ.

F_λ	σ_ϵ	σ_λ	α_{11}	α_{21}	α_{12}	α_{22}	α_{13}	α_{23}
K-means on Original Data								
Uniform	0.50	2	0.71	0.71	0.09	0.09	0.72	0.72
Uniform	2.00	2	0.78	0.77	0.06	0.06	0.80	0.80
Uniform	0.50	3	0.87	0.88	0.01	0.01	0.88	0.88
Uniform	2.00	3	0.88	0.89	0.01	0.01	0.89	0.88
Gaussian	0.50	2	0.92	0.92	0.01	0.01	0.91	0.92
Gaussian	2.00	2	0.90	0.91	0.02	0.02	0.89	0.90
Gaussian	0.50	3	0.89	0.89	0.01	0.01	0.89	0.90
Gaussian	2.00	3	0.89	0.89	0.01	0.01	0.88	0.91
K-means on Splines Coefficients								
Uniform	0.50	2	0.50	0.50	0.21	0.22	0.53	0.53
Uniform	2.00	2	0.49	0.55	0.23	0.29	0.53	0.58
Uniform	0.50	3	0.52	0.53	0.23	0.24	0.59	0.59
Uniform	2.00	3	0.56	0.63	0.26	0.29	0.60	0.69
Gaussian	0.50	2	0.49	0.49	0.19	0.20	0.56	0.56
Gaussian	2.00	2	0.47	0.55	0.21	0.29	0.54	0.62
Gaussian	0.50	3	0.68	0.69	0.14	0.14	0.69	0.70
Gaussian	2.00	3	0.64	0.76	0.16	0.20	0.64	0.76
Independent Mixture								
Uniform	0.50	2	0.87	0.87	0.01	0.01	0.87	0.87
Uniform	2.00	2	0.85	0.86	0.02	0.02	0.87	0.86
Uniform	0.50	3	0.85	0.85	0.02	0.02	0.86	0.86
Uniform	2.00	3	0.84	0.83	0.02	0.02	0.86	0.85
Gaussian	0.50	2	0.88	0.88	0.01	0.01	0.88	0.88
Gaussian	2.00	2	0.86	0.86	0.02	0.02	0.86	0.86
Gaussian	0.50	3	0.91	0.91	0.01	0.01	0.90	0.90
Gaussian	2.00	3	0.90	0.88	0.01	0.01	0.89	0.90
Exponential Mixture								
Uniform	0.50	2	0.00	0.00	0.00	0.00	0.00	0.00
Uniform	2.00	2	0.28	0.30	0.38	0.37	0.26	0.27
Uniform	0.50	3	0.00	0.00	0.00	0.00	0.00	0.00
Uniform	2.00	3	0.12	0.13	0.39	0.37	0.10	0.09
Gaussian	0.50	2	0.00	0.00	0.00	0.00	0.00	0.00
Gaussian	2.00	2	0.26	0.25	0.41	0.40	0.24	0.24
Gaussian	0.50	3	0.00	0.00	0.00	0.00	0.00	0.00
Gaussian	2.00	3	0.12	0.12	0.39	0.37	0.11	0.11

Table 5.4: Mean squared error for derivative spline coefficients when $K = 3$ for 500 replications of standard clustering methods applied to data generated under different conditions for the distribution of λ and the standard deviation of ϵ and λ.

F_λ	σ_ϵ	σ_λ	α_{11}	α_{21}	α_{12}	α_{22}	α_{13}	α_{23}
K-means on Difference Quotients								
Uniform	0.50	2	0.00	0.00	0.00	0.00	0.00	0.00
Uniform	2.00	2	0.10	0.09	0.20	0.17	0.10	0.10
Uniform	0.50	3	0.00	0.00	0.00	0.00	0.00	0.00
Uniform	2.00	3	0.10	0.09	0.20	0.17	0.10	0.10
Gaussian	0.50	2	0.00	0.00	0.00	0.00	0.00	0.00
Gaussian	2.00	2	0.10	0.09	0.20	0.18	0.10	0.10
Gaussian	0.50	3	0.00	0.00	0.00	0.00	0.00	0.00
Gaussian	2.00	3	0.10	0.09	0.20	0.18	0.10	0.10
PAM with Correlation Measure								
Uniform	0.50	2	0.91	0.91	0.46	0.46	1.88	1.85
Uniform	2.00	2	1.22	1.17	0.64	0.66	2.14	2.02
Uniform	0.50	3	0.91	0.91	0.46	0.46	1.88	1.85
Uniform	2.00	3	1.22	1.17	0.64	0.66	2.14	2.02
Gaussian	0.50	2	0.83	0.87	0.47	0.46	1.77	1.77
Gaussian	2.00	2	1.20	1.31	0.67	0.64	1.95	2.12
Gaussian	0.50	3	0.83	0.87	0.47	0.46	1.77	1.77
Gaussian	2.00	3	1.20	1.31	0.67	0.64	1.95	2.12

Table 5.5: Mean squared error for derivative spline coefficients when $K = 3$ for 500 replications of clustering methods intended to group by shape applied to data generated under different conditions for the distribution of λ and the standard deviation of ϵ and λ.

F_λ	σ_ϵ	σ_λ	α_{11}	α_{21}	α_{12}	α_{22}	α_{13}	α_{23}
K-means on Derivative Splines Coefficients								
Uniform	0.50	2	0.00	0.00	0.00	0.00	0.00	0.00
Uniform	2.00	2	0.50	0.49	1.21	1.29	0.41	0.51
Uniform	0.50	3	0.00	0.00	0.00	0.00	0.00	0.00
Uniform	2.00	3	0.50	0.49	1.21	1.29	0.41	0.51
Gaussian	0.50	2	0.00	0.00	0.00	0.00	0.00	0.00
Gaussian	2.00	2	0.46	0.45	1.26	1.27	0.48	0.52
Gaussian	0.50	3	0.00	0.00	0.00	0.00	0.00	0.00
Gaussian	2.00	3	0.46	0.45	1.26	1.27	0.48	0.52
Multilayer Mixture								
Uniform	0.50	2	0.02	0.02	0.01	0.01	0.03	0.03
Uniform	2.00	2	0.12	0.12	0.01	0.01	0.12	0.11
Uniform	0.50	3	0.33	0.33	0.03	0.03	0.33	0.32
Uniform	2.00	3	0.40	0.40	0.02	0.02	0.42	0.40
Gaussian	0.50	2	0.01	0.01	0.00	0.00	0.01	0.01
Gaussian	2.00	2	0.13	0.13	0.01	0.01	0.13	0.13
Gaussian	0.50	3	0.31	0.32	0.02	0.02	0.29	0.29
Gaussian	2.00	3	0.35	0.34	0.01	0.01	0.34	0.35
Vertically Shifted Independent Mixture								
Uniform	0.50	2	0.00	0.00	0.00	0.00	0.00	0.00
Uniform	2.00	2	0.01	0.01	0.01	0.01	0.01	0.01
Uniform	0.50	3	0.00	0.00	0.00	0.00	0.00	0.00
Uniform	2.00	3	0.01	0.01	0.01	0.01	0.01	0.01
Gaussian	0.50	2	0.00	0.00	0.00	0.00	0.00	0.00
Gaussian	2.00	2	0.01	0.01	0.01	0.01	0.01	0.01
Gaussian	0.50	3	0.00	0.00	0.00	0.00	0.00	0.00
Gaussian	2.00	3	0.01	0.01	0.01	0.01	0.01	0.01
Vertically Shifted Exponential Mixture								
Uniform	0.50	2	0.00	0.00	0.00	0.00	0.00	0.00
Uniform	2.00	2	0.01	0.01	0.01	0.01	0.01	0.01
Uniform	0.50	3	0.00	0.00	0.00	0.00	0.00	0.00
Uniform	2.00	3	0.01	0.01	0.01	0.01	0.01	0.01
Gaussian	0.50	2	0.00	0.00	0.00	0.00	0.00	0.00
Gaussian	2.00	2	0.01	0.01	0.01	0.01	0.01	0.01
Gaussian	0.50	3	0.00	0.00	0.00	0.00	0.00	0.00
Gaussian	2.00	3	0.01	0.01	0.01	0.01	0.01	0.01

Table 5.6: Mean squared error for derivative spline coefficients when $K = 3$ for 500 replications of proposed clustering methods applied to data generated under different conditions for the distribution of λ and the standard deviation of ϵ and λ.

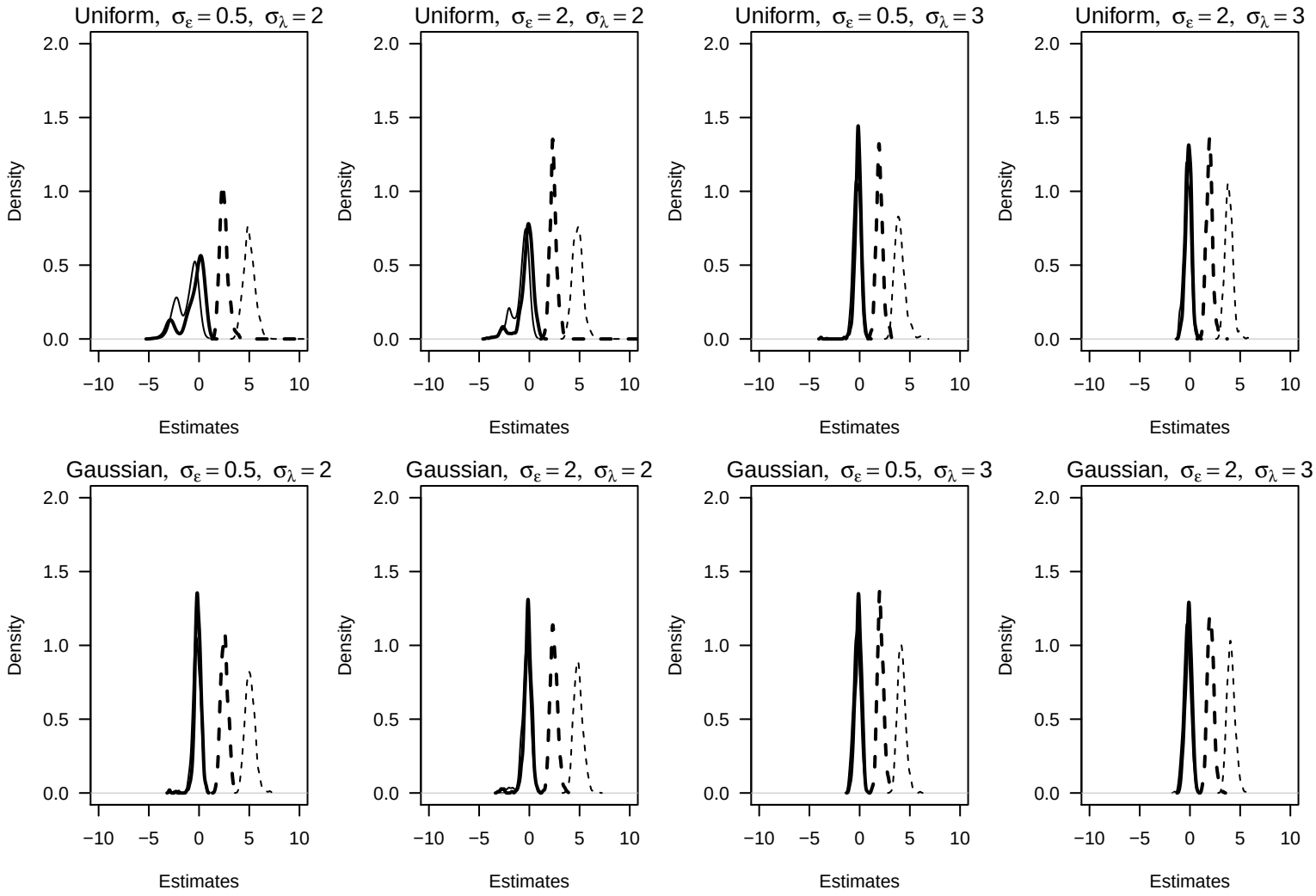

Figure 5.1: Density plots of estimated multinomial coefficients from analysis after running K-means on original data under simulation conditions. Coefficients for group 3 are fixed equal to 0. Solid: w_1, dashed: w_2. Thin: group 1, thick: group 2.

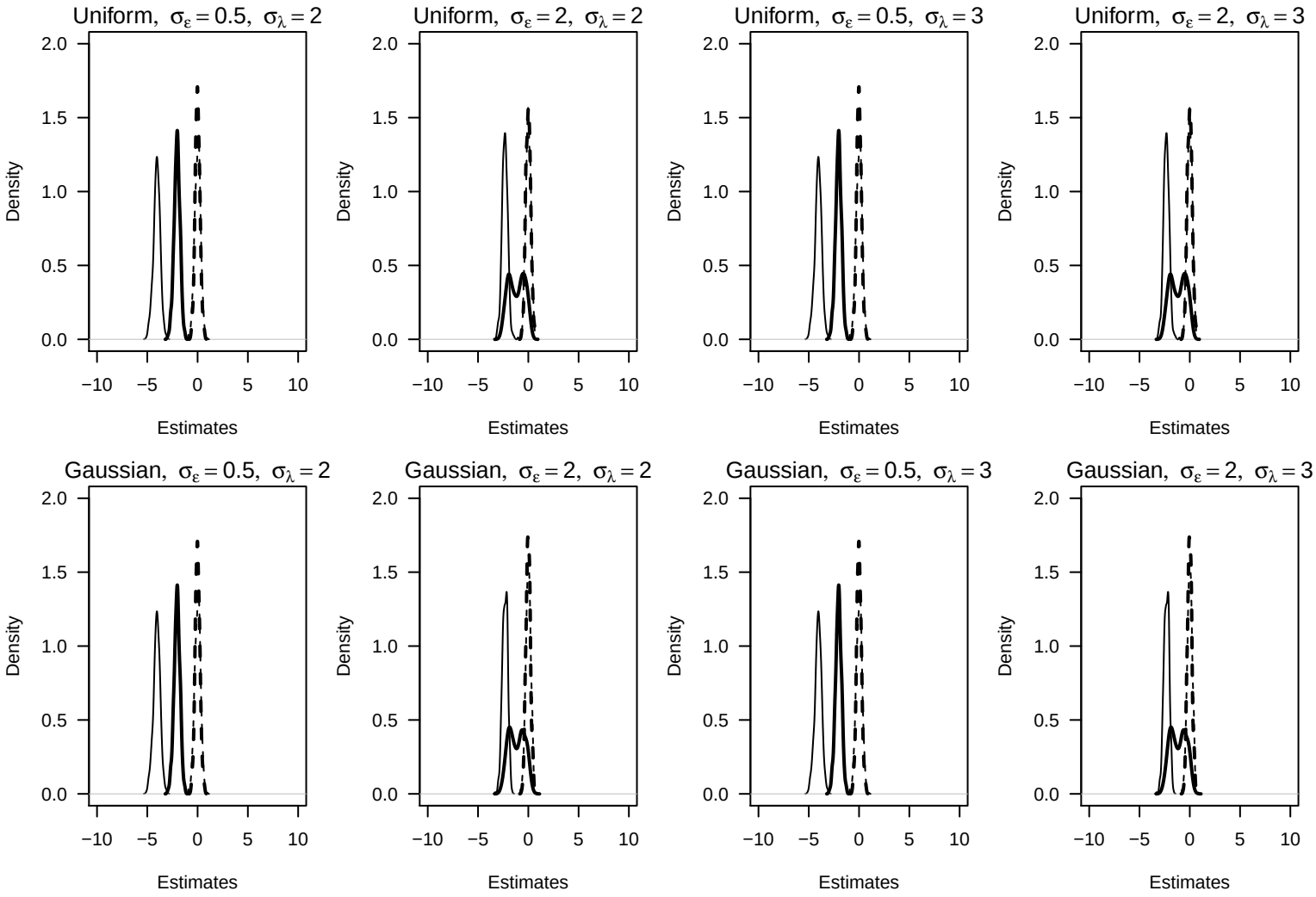

Figure 5.2: Density plots of estimated multinomial coefficients from analysis after running K-means on difference quotients under simulation conditions. Coefficients for group 3 are fixed equal to 0. Solid: w_1, dashed: w_2. Thin: group 1, thick: group 2.

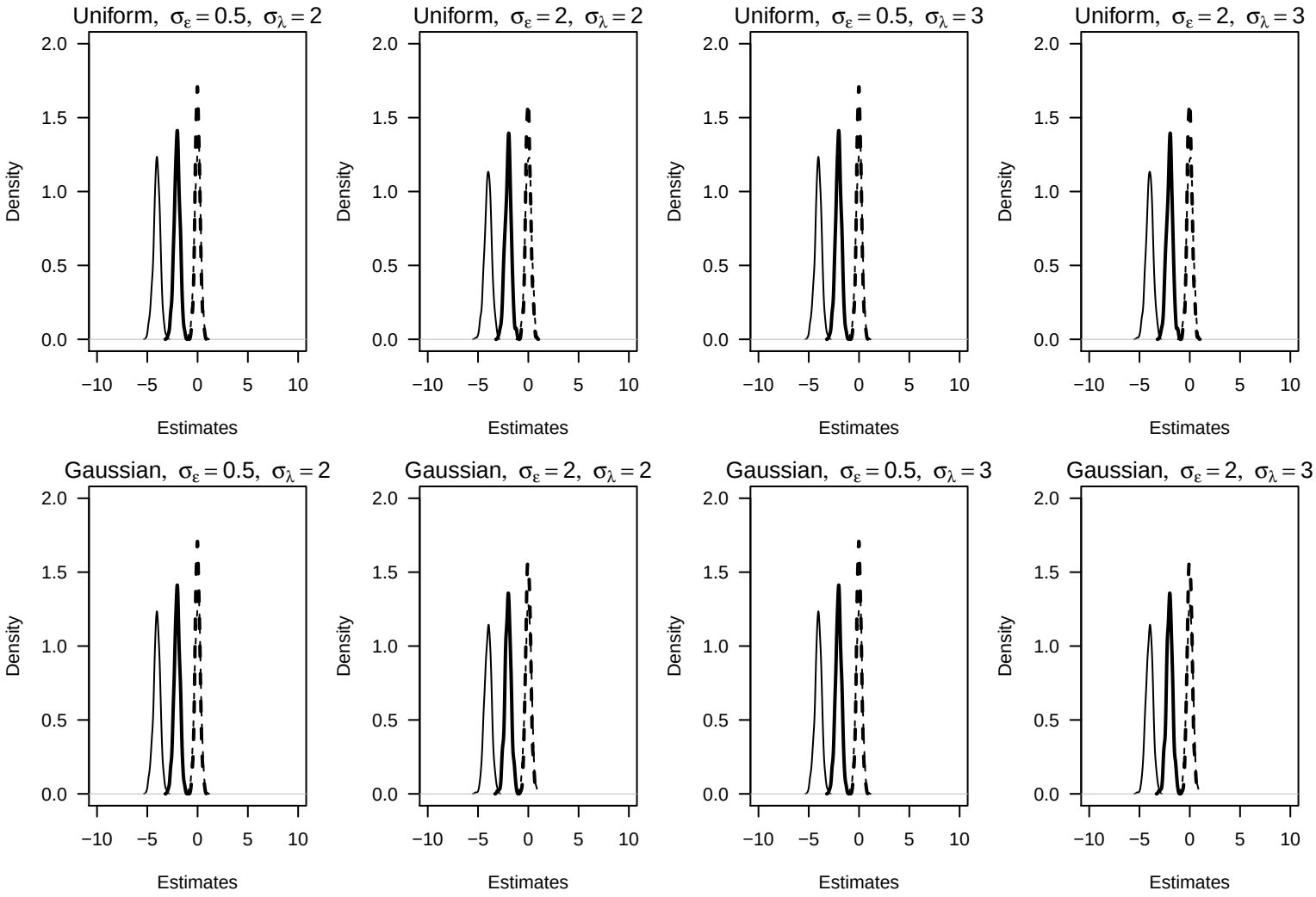

Figure 5.3: Density plots of simultaneously estimated multinomial coefficients from the vertically shifted exponential mixture model under simulation conditions. Coefficients for group 3 are fixed equal to 0. Solid: w_1, dashed: w_2. Thin: group 1, thick: group 2.

5.4 Discussion

This study provides evidence against standard clustering methods and insight into the strengths and weaknesses of the proposed methods to cluster by shape. K-means on original data finds groups based mainly on level. K-means on spline coefficients generalizes the original K-means to work with irregularly sampled data, but does not improve performance when the number of observations is small. The finite mixture model with independence also fails to detect shapes when shape and level are not perfectly correlated. Surprisingly, the mixture model with exponential correlation was able to detect the three shape groups when measurement error was small, but it fails to select the correct number of groups.

Of the two methods previously thought to detect shape, PAM with the correlation measure failed on all accounts in this simulation. Not only did it choose the incorrect number of shape groups, it could not find the shape groups when forced to have the correct number of shape groups. Quotient differences, on the other hand, worked well when measurement error was small. It was able to choose the correct number of groups and find the shape groups.

The K-means on the derivative spline coefficients is an attempt to use the derivative for clustering while accommodating irregularly sampled data. For data with small measurement error, this method performed well. However, under larger measurement error, it performed worse than difference quotients. This primarily was due to the small number of observation times over which the data are smoothed for each individual. I imagine the performance would improve with more observations over time. The multilayer mixture model worked well when the data fit into the shape group, level component framework such that the components are well separated. Unfortunately, this is not a realistic for most data sets. Vertically shifting the data prior to fitting a mixture model outperformed the rest of the methods. Measurement error and overlap had little impact on the results. It consistency found the three shape groups even when the covariance was not modeled perfectly.

These results are limited as I focused on only one generating framework. The shapes are straight lines with very few observations. While this is restrictive, it does not drastically deviate from real data sets with long range trends. An interesting extension of this simulation would be to change the slopes of the lines to further explore how the separation impacts the proposed models. I kept the data generating correlation simple by including a random intercept on top of independent errors. One could try a more complex structure to detect the impact of covariance misspecification like in Chapter 2. Despite these limitations, the conclusions should generalize to similar situations.

Chapter 6

Data application

6.1 Introduction

Obesity has become one of the most burdensome public health issues that faces the United States today [138]. This epidemic impacts children and adults alike. A recent study estimates that about 32% of children in the U.S. are overweight or obese [102]. This alarming statistic has been a call to action for public officials and parents to change how we live and eat. However, encouraging healthy diets and physical activity through increased education and community involvement has seen limited success [47, 78]. It is generally accepted that genetic, metabolic, and environmental factors in addition to behavioral factors play a role in regulating a child's weight [138]. Yet, the exact determinants of childhood obesity are poorly understood.

As it is difficult to change the behavior of individuals, recent research has focused on other potential causes and venues for intervention. One hypothesis is that exposure to hormone-altering chemicals prenatally and during early postnatal development disrupt and reprogram the metabolic system to favor weight gain [137]. A number of rodent studies suggest that prenatal exposure to endocrine disrupters such as bisphenol A (BPA) is associated with increased body weight after adjusting for exercise and amount and type of food [119, 4, 55, 56, 91, 129, 144, 149]. A small number of cross-sectional human studies suggest a positive association between BPA and obesity in children and adults [11, 126, 136]. However, only one study has investigated human prenatal BPA exposure and they found a significant negative association between prenatal exposure and body mass index at age 9 but only in girls [48].

To fully understand the potential impact of prenatal chemical exposure on childhood growth development, it is necessary to observe the height and weight of children during the time between infancy and adolescence. Quality longitudinal growth studies are becoming more common, but most data analyses do not fully take advantage of the longitudinal nature of the data and only look at two cross-sectional points to determine change over time. In the past few years, researchers have noted the importance of the entire growth trajectory of individuals with a view to determine if there are distinct growth trajectory patterns [109, 10,

73, 42]. In this case, we determine whether growth trajectory patterns are associated with early factors such as prenatal BPA exposure.

6.1.1 Body mass index trajectories

Quantifying and evaluating children's growth involves measuring both height and weight. Body mass index (BMI), calculated as weight (kg)/height^2 (m^2), is used as an indicator of total body fat for most individuals [115]. BMI is also used to clinically classify individuals as overweight or obese. The Center for Disease Control (CDC) publishes clinical charts that include BMI-for-age reference growth curves in terms of percentiles and z-scores using cross-sectional data from National Health and Nutrition Examination Survey (NHANES). The z-scores are not intended for longitudinal analysis as the CDC performs different normalizing transformations for each age in months to resolve the skewed sex and age-specific distributions for BMI. Since the reference data is cross-sectional, the charts may not reflect typical age-related patterns of BMI change. Therefore, we concentrate on the original BMI trajectories rather than of BMI z-scores over time.

One way to detect alterations in growth patterns is to apply cluster analysis methods to childhood BMI trajectory data. Then, additional analysis can be completed to estimate the relationship between growth patterns and early life factors such as exposure to BPA as well as other important factors such as maternal BMI, maternal smoking, maternal gestational weight gain, maternal age, birth weight, and duration of breastfeeding which have been suggested to be potential predictors of increased risk of a high-rising BMI growth as well as the early and late onset of obesity [109, 10, 73].

6.1.2 Clustering

The clustering methods used in many longitudinal applications including growth trajectories are typically based on a finite mixture model, which is a probabilistic model for representing groups within the overall population that occur with different frequencies but are not known *a priori*. In the simplest form, the mixture density for the random outcome vector $\mathbf{y}$ is a weighted sum of K group densities written as

$$f(\mathbf{y}|\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{y}|\boldsymbol{\theta}_k)$$

where π_k is the probability of belonging to the kth group and $\pi_k > 0$ for all $k = 1, .., K$ and $\sum_{k=1}^{K} \pi_k = 1$. The probability densities f_k are typically assumed to be multivariate Gaussian for continuous outcomes such as BMI. With longitudinal data, the main goal is to study the change over time so the mean within each group is modeled conditional on the observations times $\mathbf{t}$. This model is the basis for many popular methods and software that have rapidly become commonplace in the growth trajectory literature.

The Proc Traj add-on package for SAS [62], used in the BMI literature [109, 10], fits group-based trajectory models. These models are finite mixture models such that for a continuous outcome measure, the group distributions are assumed to be multivariate Gaussian with a polynomial mean function and the repeated measures within an individual are assumed independent conditional on the group membership. Thus, the model assumes that the outcome data $\mathbf{y}_i = (y_{i1}, ..., y_{im_i})$ observed at times $\mathbf{t}_i = (t_{i1}, t_{i2}, ..., t_{im_i})$ for subject i who belongs to group k is generated according to

$$\mathbf{y}_i = \beta_{0k} + \beta_{1k}\mathbf{t}_i + \beta_{2k}\mathbf{t}_i^2 + \beta_{3k}\mathbf{t}_i^3 + \boldsymbol{\epsilon}_i$$

where $\boldsymbol{\epsilon}_i \sim N(0, \sigma^2 I)$ for a cubic mean model. Additionally, the group probabilities can be modeled using a generalized logit function to allow time-stable covariates determine group membership (setting $\boldsymbol{\gamma}_K = 0$ for identifiability),

$$\pi_k = \frac{\exp(\boldsymbol{\gamma}_k^T \mathbf{w}_i)}{\sum_{l=1}^{K} \exp(\boldsymbol{\gamma}_l^T \mathbf{w}_i)}$$

where $\mathbf{w}_i$ is a design vector based on time-stable covariates.

Another software program, Mplus [98], commonly used in the literature [73, 42], fits a generalization of this mixture model, termed a growth mixture model, which allows random effects in the mean structure to account for within-group variation that is ignored in the Proc Traj model specification. Assuming a distribution for the cluster-specific slope and intercept coefficients attempts to model within-individual dependence that is inherent in repeated measures. These random effects can be incorporated into the distribution of the errors. Thus, the observed data for subject i is assumed to be generated by

$$\mathbf{y}_i = \beta_{0k} + \beta_{1k}\mathbf{t}_i + \beta_{2k}\mathbf{t}_i^2 + \beta_{3k}\mathbf{t}_i^3 + \boldsymbol{\epsilon}_i$$

where $\boldsymbol{\epsilon}_i \sim N(0, \boldsymbol{\Sigma}_{ik})$ with $\boldsymbol{\Sigma}_{ik} = \boldsymbol{\Lambda}_i \boldsymbol{\Psi}_k \boldsymbol{\Lambda}_i^T + \boldsymbol{\Theta}_k$, where $\boldsymbol{\Theta}_k$ is the covariance matrix of the random errors, $\boldsymbol{\Psi}_K$ is the covariance matrix for the random effects, and

$$\boldsymbol{\Lambda}_i = \begin{pmatrix} 1 & t_{i1} & t_{i1}^2 & t_{i1}^3 \\ 1 & \vdots & \vdots & \vdots \\ 1 & t_{im_i} & t_{im_i}^2 & t_{im_i}^3 \end{pmatrix}$$

This model is equivalent to the Proc Traj model when $\boldsymbol{\Psi}_k = \mathbf{0}$ and $\boldsymbol{\Theta}_k = \sigma^2 \mathbf{I}$ for all $k = 1, ..., K$. It is important to note that the random effects indirectly specify a complex, non-stationary covariance structure. This is problematic since misspecifying the covariance structure in finite multivariate Gaussian mixture models can cause bias in the mean estimates and result in an incorrect number of chosen groups as seen in Chapter 2. Special attention should be paid to modeling the dependence inherent in repeated measures of an outcome over time. I recommend fitting a model using a well-known stationary correlation structure such as exchangeable or exponential correlation before fitting a complex non-stationary model via random effects.

These model-based methods are typically understood to group individuals such that members of the same group share a similar pattern of change over time [42]. This suggests that two individuals with similar patterns of change but different vertical levels would be grouped together. However, model parameters and subsequently group membership are estimated by maximizing a likelihood function, which is based on Euclidean distance between observed data and group means if normality is assumed (see Chapter 3). This process results in groups being primarily determined by the level despite subtle differences in the shape especially if the level and shape of a trajectory are independent or weakly dependent. Therefore, these methods do not directly group trajectories on shape and resulting groups may include growth trajectories of the same level but different shapes over time. This translates into estimated group memberships and means not accurately representing shape groups present in the data.

Due to this misunderstanding, researchers present the results of a mixture model analysis by discussing the shape of the estimated mean growth pattern for each group [109, 10]. However, the mean only represents the average shape of all trajectories in the group; perhaps no one individual follows the path of the mean especially when the groups are not homogeneous in terms of shape. Additionally, the estimated relationships between risk factors and the resulting groups are deceitful when the group descriptions are inaccurate and misunderstood. Care needs to be taken when making conclusions based on these methods especially in describing the shape of the trajectories within each group.

The goal of this research is to group individuals on the basis of the shape of their growth trajectory, so it is appropriate to compare the new methods proposed in Chapter 4 with those used in practice on a real data set. The vertically shifted mixture model is based on the same foundation as those described above. The main difference is that rather than fitting a model to the original data, the input for the model is the transformed or vertically shifted data vector. Each individual's mean BMI is subtracted from all of their measured BMI values. In effect, each individual is normalized to have mean zero and the level is removed without eliminating the variability at any time point. By removing the mean prior to fitting a multivariate Gaussian mixture model, we allow the clustering method to focus directly on the shape rather than the level. Thus, the resulting groups can be honestly summarized by describing the shape of the mean trajectory of the group and estimated associations with risk factors can be interpreted accurately in terms of shape groups. I illustrate the difference between the methods with longitudinal BMI data on a sample of children living in the Salinas Valley, CA.

6.2 Data

The Center for the Health Assessment of Mothers and Children of Salinas (CHAMACOS) Study is a longitudinal birth cohort study designed to assess the health effects of pesticides and other environmental exposures on the growth and development of low-income children living in the agricultural Salinas Valley, CA [31, 33]. Of 601 pregnant women enrolled in the

study in 1999-2000, a total of 527 mostly Latino mother-child singleton pairs were followed through a live-birth delivery and 327 pairs continued to be followed through the 9-year interview. Baseline maternal characteristics were measured at the start of the study and maternal urine and blood samples were taken twice during pregnancy and then again shortly after delivery to measure levels of pesticide and chemical exposure. Child height and weight were measured without jackets and shoes by trained staff at interviews that occurred at birth and when the child was approximately 1, 2, 3 $\frac{1}{2}$, 5, 7, and 9 years of age. BMI was calculated as weight (kg) divided by height squared (m^2) for ages 2 and over. The exact age of the child was also recorded due to variability in the interview times. For this thesis, we limit our analysis to 303 children who have at least four recorded BMI measures. Details of the study are published elsewhere [32]. All study activities were approved by the Committee for the Protection of Human Subjects at the University of California, Berkeley.

For illustrative purposes, we limit the discussion to two baseline risk factors, maternal pre-pregnancy BMI and BPA exposure. One of the strongest known predictors of a child's BMI is maternal pre-pregnancy BMI. At the start of the study, the maternal height was measured and used together with self-reported pre-pregnancy weight to calculate the BMI. During the first and second half of pregnancy, maternal BPA exposure was measured via urine samples. See Harley et al. [48] for details of how the concentrations were measured.

Characterizing BMI growth trajectories of children and distinguishing distinct patterns are of great importance to public health officials. We use this data set to illustrate the differences between the clustering methods and how they work in practice with real data.

6.3 Cluster methods

I use the standard and novel methods described in this thesis and compare the clustering groups and the inference on baseline factors that may be related to cluster membership. Specifically, we fit a multivariate Gaussian mixture with conditional independence like Proc Traj (Model 1), a multivariate Gaussian mixture with exponential correlation (Model 2), and the vertically shifted mixture model proposed in Chapter 4 (Model 3). There is some variability in the observation times, but the growth is rather gradual so the independence assumption may be the best approximation for the covariance of the transformed data in Model 3.

All of these model-based methods require a mean structure. I use a B-spline basis to model the mean function. Based on the visual inspection of the data, I determine that quadratic basis functions with one internal knot should be sufficient to model the complexity given the limited number of time points. The internal knot is place at the median age of 5 years.

To compare the methods in terms of baseline factor relationships, I fit the model with two different factors: maternal pre-pregnancy BMI, which is known to impact the child's BMI, and the log base two transformed maternal BPA exposure during pregnancy, which is

hypothesized to impact growth trajectories. The relationships are estimated separately from each other but simultaneously with the other model parameters.

To estimate parameters for all of the models, I use maximum likelihood estimation via the EM algorithm. The algorithm is run five times for each model with random starts and the final estimate are from the fit with the highest likelihood. For each model, the number of groups is chosen by fitting the models for $K = 2, 3, 4, 5, 6$ and choosing the value of K that minimizes the BIC. To make inferences about the parameters, I estimate the robust sandwich standard error for the parameter estimators [146]. Odds ratios for the baseline factors are calculated by exponentiating the coefficients and are presented with exponentiated confidence intervals.

The cluster groups are visualized by plotting individual BMI trajectories colored according to the group assignment made by maximizing the posterior probability. The group means are represented in the adjoining panel. For the vertically shifted model, the group means are shifted such that the mean BMI at the age 2 equals the average BMI of the individuals in the particular group at the age 2 interview.

6.4 Results

Figure 6.1 shows the clustering results for the three mixture models with maternal pre-pregnancy BMI as a baseline factor that impacts group membership. The chosen number of groups for each model is $K = 5$, $K = 4$, and $K = 5$ for Model 1, 2, and 3, respectively. Model 1 as compared to Model 2 requires more groups to model the variability within the BMI trajectories due to the limited correlation structure. Neither Model 1 nor Model 2 group individuals based on the shape of their trajectories, but it is hard to see this without comparing the plot of groups trajectories to those from Model 3. The plot from the last model brings attention to individuals with high BMI levels at age two with relatively stable trajectories over time and individuals with moderate BMI levels whose BMI drastically increased over the follow-up period. Both of these types of individuals are categorized with trajectories with very different shapes in the other two models.

It is interesting to note that the mean shapes do not drastically change between the three models even though group membership changes (right side of Figure 6.1). Thus, the number of membership changes is relatively small in this case so as to not to change the shape of the means for this data set. Even though it is tempting to compare the starting levels of the mean curves, it is important to realize that the level was removed to focus on the shape in Model 3 and the artificial levels for presentation are based on the average BMI at age 2 within the groups. Thus in Model 3, group 1 includes more individuals with lower BMI at age 2 in comparison to the other models.

Table 6.1 presents the estimated odds ratios comparing group membership odds for a one unit increase in maternal pre-pregnancy BMI. The reference group has the highest numerical label with the most stable shape over time. There seems to be a consistent monotone trend in the odds ratios for pre-pregnancy BMI as the derivative of the trajectory shape increases, but

the inference depends on the model used. We see that it highly predicts group memberships in Model 1 as the individuals are divided mainly by BMI level. However, generalizing the correlation structure in Model 2 increases the standard error enough to decrease the strength of the evidence. While the point estimates are similar for Model 3, the standard errors are slightly smaller but not enough to provide strong evidence that maternal pre-pregnancy BMI also impacts the growth pattern over time.

	Odds Ratio	95% CI	$P > \|z\|$
Model 1: Group 1	1.32	(1.2, 1.45)	0
Model 1: Group 2	1.24	(1.14, 1.35)	0
Model 1: Group 3	1.15	(1.04, 1.27)	0.006
Model 1: Group 4	1.11	(1.02, 1.21)	0.017
Model 1: Group 5 (ref.)	-	-	-
Model 2: Group 1	1.3	(0.85, 1.98)	0.232
Model 2: Group 2	1.23	(0.88, 1.73)	0.23
Model 2: Group 3	1.11	(0.88, 1.39)	0.387
Model 2: Group 4 (ref.)	-	-	-
Model 3: Group 1	1.28	(0.92, 1.79)	0.136
Model 3: Group 2	1.24	(0.93, 1.67)	0.143
Model 3: Group 3	1.18	(0.89, 1.55)	0.247
Model 3: Group 4	1.04	(0.85, 1.28)	0.7
Model 3: Group 5 (ref.)	-	-	-

Table 6.1: Odds ratio estimates and confidence intervals for a one unit increase in maternal pre-pregnancy BMI comparing each group to the reference group for the three mixture models (Model 1: independence, Model 2: exponential, Model 3: vertically shifted).

Figure 6.2 shows the clustering results for the three mixture models with maternal BPA exposure during pregnancy as a baseline factor that impacts group membership. The chosen number of groups for each model is $K = 5$, $K = 4$, and $K = 5$ for Model 1, 2, and 3, respectively. The clustering looks very similar to those from the models with pre-pregnancy BMI. The main differences occur with Model 3. The group means slightly differ in shape from those with maternal BMI. This is natural since all of the parameters include group membership are simultaneously estimated. Therefore, baseline factors have a minor influence on the group membership for a trajectory in the overlap between two groups. This change of group can be enough to change the mean curve.

Table 6.2 gives the estimated odds ratios comparing group membership odds for a one unit increase in log base 2 transformed maternal BPA exposure. The reference group is again the last group with the most stable trajectory shape. In generalizing the correlation structure (Model 2 vs. Model 1), the estimated relationship between BPA and group membership changes. Specifically, the odds ratio estimates for group 2 switches from 0.9 in Model 1 to 1.24 in Model 2. This suggests that increased BPA increases the probability of being in

group 2 in comparison to group 4. This magnitude change is maintained by the vertically shifted model (Model 3), and the standard errors are smaller relative to Model 2 to provide more evidence that there is small effect of BPA exposure on the shape of the growth curve over time.

	Odds Ratio	95% CI	$P > \|z\|$
Model 1: Group 1	0.84	(0.58, 1.21)	0.345
Model 1: Group 2	0.9	(0.65, 1.25)	0.542
Model 1: Group 3	1.04	(0.75, 1.46)	0.804
Model 1: Group 4	0.92	(0.65, 1.29)	0.618
Model 1: Group 5 (ref.)	-	-	-
Model 2: Group 1	0.79	(0.56, 1.11)	0.181
Model 2: Group 2	1.24	(0.84, 1.83)	0.285
Model 2: Group 3	1.01	(0.73, 1.39)	0.956
Model 2: Group 4 (ref.)	-	-	-
Model 3: Group 1	0.8	(0.49, 1.29)	0.356
Model 3: Group 2	1.33	(0.95, 1.87)	0.095
Model 3: Group 3	1.01	(0.71, 1.43)	0.948
Model 3: Group 4	0.88	(0.6, 1.29)	0.52
Model 3: Group 5 (ref.)	-	-	-

Table 6.2: Odds ratio estimates and confidence intervals for a one unit increase in maternal BPA exposure during pregnancy in log base 2 units comparing each group to the reference group for the three mixture models (Model 1: independence, Model 2: exponential, Model 3: vertically shifted).

6.5 Discussion

To explore the population in terms of their heterogeneous growth patterns over time, we used multivariate finite mixture models. Rather than averaging out all of the interesting growth patterns, mixture models allow for a finite number of relationships. Standard mixture models group these relationships based on the feature that dominate the variability. If this feature is the level of the trajectory, the estimation of the mixture model may not highlight interesting patterns over time. Therefore, we use a simple adjustment to remove the vertical level and only focus on how the children's growth changes over time.

For this data set, the mean functions do not drastically change between the standard model and the vertically shifted model, but the group memberships change. Moving individuals between groups impacts the interpretation of baseline factors and their association with the probability of being in a particular group. The well-studied association of pre-pregnancy BMI and child's BMI level [145] is supported with this data set. However, this data set does

not provide strong evidence for an association with the shape of the growth pattern. This is an important distinction to make as genetic factors may impact early childhood BMI but not necessarily determine the growth trajectory over time. While this data set does not provide strong evidence to suggest that prenatal exposure to BPA is associated with the level or shape of the growth pattern, the shift in the estimates between Model 1 and Model 3 for the second group of late gainers highlights the importance of investigating the factors that impact level and those that impact shape separately.

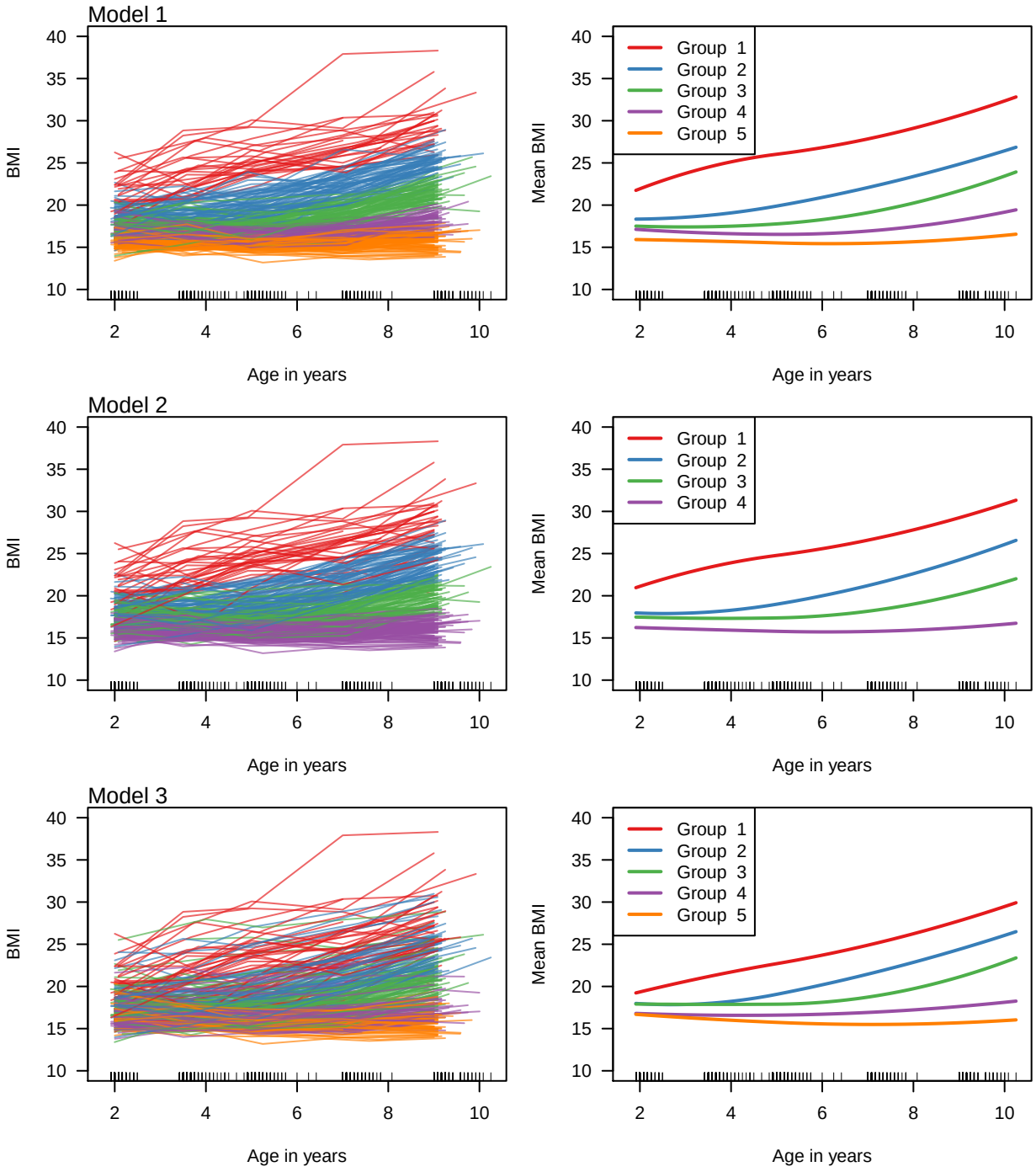

Figure 6.1: Clustered BMI trajectories colored according to the group assignment made by maximizing the posterior probability and group mean functions for three mixture models (Model 1: independence, Model 2: exponential, Model 3: vertically shifted) with maternal pre-pregnancy BMI as the baseline factor.

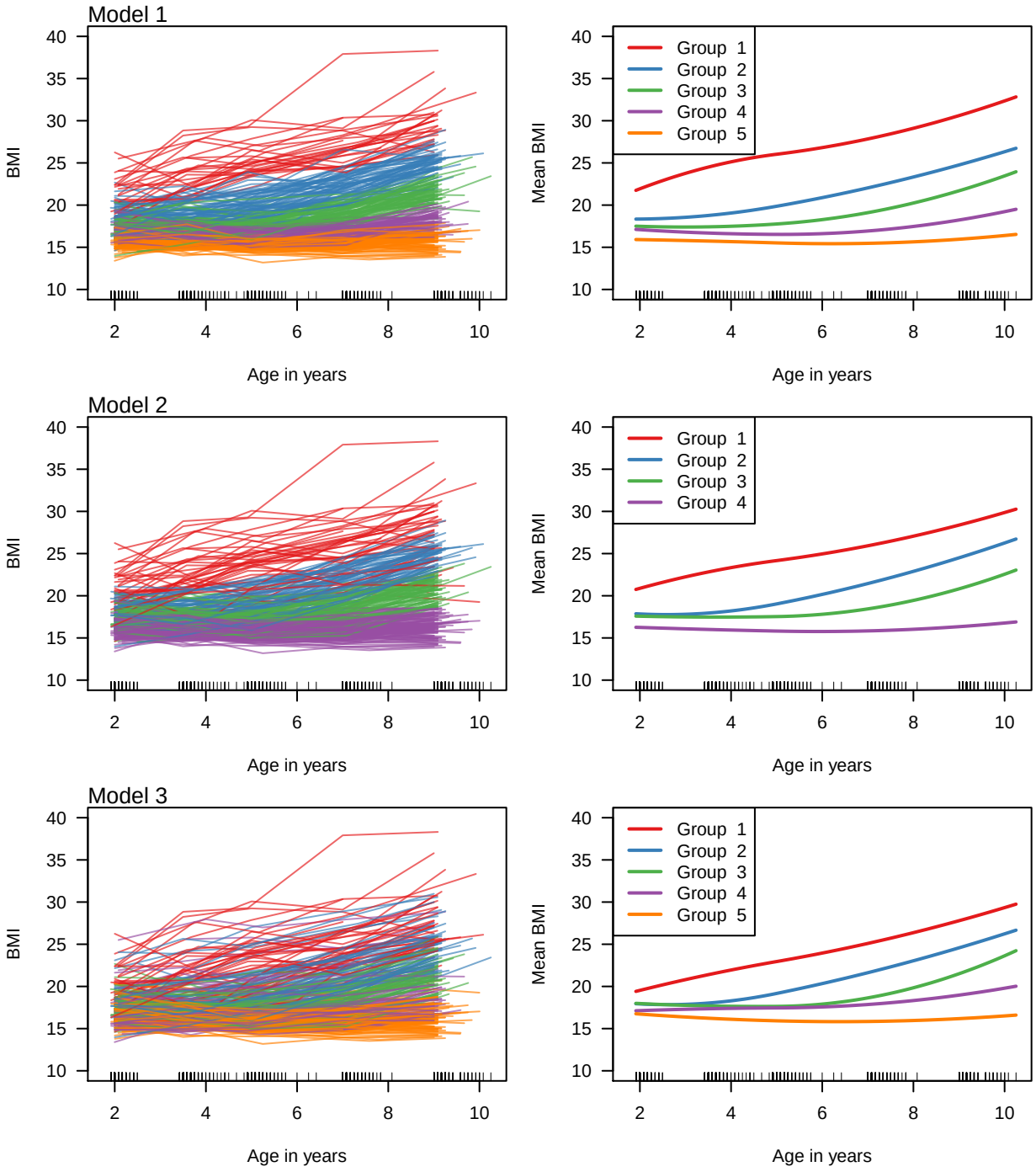

Figure 6.2: Clustered BMI trajectories colored according to the group assignment made by maximizing the posterior probability and group mean functions for three mixture models (Model 1: independence, Model 2: exponential, Model 3: vertically shifted) with maternal BPA exposure during pregnancy as the baseline factor.

Chapter 7

Conclusions

Longitudinal data sets include measurements repeated at potentially irregularly intervals over time on a relatively large number of subjects. This highly structured data requires methods that account for the time-ordering and dependence between measurements on the same subject. I focus on cluster analysis methods, propose new methods to group longitudinal data by the pattern of outcome change over time, and compare the performance of standard clustering methods with the proposed in answering research questions about change over time.

7.1 Contributions

This thesis makes several original contributions. The first major contribution is the study of covariance misspecification in mixture models in Chapter 2. When components of a mixture model overlap, it is important to correctly model the within-vector dependence to avoid asymptotic and finite sample bias in the parameter estimates when the number of components is known. Naively assuming conditional independence may bias the clustering results and lead to incorrect conclusions. This is contrast to most longitudinal data analysis in which the mean can be estimated in an unbiased manner without modeling the inherent dependence and robust standard error estimate are used for inference. Although not studied in this thesis, covariance misspecification also impacts the choice of the optimal number of components in a mixture model when it is not known a priori. If the dependence structure is more complex than the assumed model, more components are needed to model the variability and dependence. This has practical importance when fitting a mixture model to longitudinal data as there is inherent dependence within subjects. This work adds to and is consistent with the limited literature on misspecifying mixture in that well-separated components are more robust to misspecification.

When clustering longitudinal data, it is important to consider how two trajectories are deemed similar. Longitudinal studies are typically carried out to directly study the change over time. The second contribution of this thesis is raising awareness that most clustering

algorithms do not group individuals based on how trajectories change over time. Rather, most methods, including the standard finite mixture model, group individuals on the feature that dominates the variability, which is often the overall outcome level. The few methods that have been suggested to cluster based on shape are limited as they only succeed in certain circumstances.

The third contribution is adapting and extending standard methods to cluster based on shape while overcoming the shortfalls of the current methodology. The three proposed methods approach the problem from different angles. All of the methods have issues and challenges, but vertically shifted mixture models outperforms them when in clustering short noisy trajectories.

Lastly, this thesis studies growth trajectories by clustering children growth trajectories into distinct groups and estimating the relationship between baseline factors and group membership. The proposed vertically shifted mixture model is fit to CHAMACOS BMI outcome data and the results are juxtaposed with those from standard clustering methods that are currently used in the literature. In this data set, level and shape are moderately related so the group mean curves are similar between clustering methods, but the groups differ enough in composition to impact the direction and inference of the relationship with baseline factors. Maternal pre-pregnancy BMI is significantly associated with growth trajectory groups primarily determined by level but not necessarily with those based on shape. On the other hand, this study provides some modest evidence to suggest that prenatal exposure to BPA may impact the rate of growth but not necessarily the level of BMI.

7.2 Limitations

While this thesis attempts to be thorough in its study of these clustering methods, it cannot be exhaustive. The study on covariance misspecification primarily focuses on wrongly assuming conditional independence when the true data-generating dependence for one of the groups is exchangeable with constant variance. There are infinitely many simulation possibilities for generating and estimating covariance models. Additionally, the trends used to generate the data were limited to horizontal lines over time knowing that more complex shapes translate into more well-separated components. To fully understand the impact of misspecification, one could study different mean shapes, more components, non-stationary generating covariance structures and varying vector lengths. However, the study highlighted key issues of assuming independence for longitudinal data, choosing a covariance model, and the impact of component separation on clustering. More suggestions for more simulation studies are listed at the end of Chapter 2.

The context in which I discuss and compare different clustering methods is a longitudinal data set with only five to ten repeated outcome measurements sparsely observed over time, which is common in the field of public health. Some of the methods discussed may work better on longer trajectories. In that case, if the data are sampled on a dense grid, more analysis options are available in the functional data analysis literature. While the restricted

scope may be considered a shortfall of this thesis, I believe it is worth focusing our attention on this type of data set since it is a common structure in practice.

In terms of the proposed methods, the vertically shifted mixture model excels in practice even though there are theoretic issues. By subtracting the mean, the transformed vector lies on a subspace of one dimension smaller and thus the covariance matrix of the vertically shifted vector is always singular no matter the original dependence structure. However, we model the covariance of the deviations from the mean for the transformed vectors. If the observation times are the same for every individual, this matrix is singular and regularization of the covariance matrix is necessary for estimation. When this is the case, the vertically shifted model may not be able to accurately detect subtle shape differences due to the misspecification in the covariance matrix. If the data are observed at random times, the transformation may have a valid covariance matrix and modeling may be more accurate.

Many believe that the gold standard for longitudinal studies is to measure individuals at evenly spaced, regular times. However, there is no scientific reasoning for this uniform temporal design [16]. Rather, the observation times should be based on the expected trajectory shape; more observations are needed during times of rapid or non-linear change. Work in this thesis suggests a random temporal design may be preferable when vertically shifting the data, but it is also important to have enough data points throughout the follow-up period to thoroughly observe individual trajectories. Therefore, a temporal design should be chosen based on the expected shape plus some systematic or random variability in the actual measurements times.

7.3 Future work

In addition to answering questions, the work in this thesis leads to many new questions. Mixture models are sensitive to assumptions and incorrectly modeling the covariance structure can lead to bias. However, it is hard to know whether there is bias when modeling with real data. For most statistical models, resampling methods can be used to estimate the magnitude of the bias; however, implementation for mixture models is not straightforward. Grün and Leisch [45] provided suggestions for bootstrapping finite mixture models as a diagnostic tool but did not discuss the issue of bias. This is an area needing more work as practitioners need some indication that the results are potentially misleading due to bias.

While this thesis has focused on longitudinal data with few time observations, some of the clustering methods discussed in this thesis may perform better when there is more data points per person. Future work should include increasing the number of observations when comparing the three proposed methods. Additionally, as the number of observations increases, the line that distinguishes longitudinal data from functional data is blurred. It is worth determining the point at which functional clustering algorithms may be more applicable than those proposed in this thesis. It is also worth investigating the possibility of extending these ideas of shape to non-continuous outcome data [62] as well as multivariate longitudinal trajectories [63, 29].

I have limited the discussion to only include basis functions of time in the design matrix for the mean regression, but the model is general enough to allow other covariates. However, including time-fixed covariates that only impact the vertical level is futile when clustering based on shape. If these covariates are thought to impact the shape, one can include them in the generalized logit design vector so as to estimate the association with shape group membership. On the other hand, time-varying covariates can be used to model the mean trajectory shape. However, the interpretation of the estimated coefficients in the context of clustering is not straightforward. Nagin et al. [101] suggested using time-varying binary covariates to model abrupt or sharp changes in individual trajectories caused by life course turning points. However, if these binary covariates change states multiple times during the follow-up period, a hidden Markov model or switching regression models [110] may be more appropriate. Future work needs to consider the interpretation of shape clusters when continuous time-varying covariates are introduced into the mean structure as groups are formed not only on the basis of the relationship with time but also with these new covariates.

One of the largest contributions of this thesis is separating shape and level into two distinct characteristics of longitudinal trajectories. Research questions usually focus on one or the other so it is important to separate these features so as to not muddle the interpretations when doing clustering analysis. One suggestion for systematically studying both features is to complete a pseudo two-part model in which two mechanisms impact the outcome values. The standard two-part model views a semicontinuous univariate outcome response as the result of two processes, one determining whether the outcome is 0 and the other determining the actual value if it is non-zero. This idea was first introduced in the econometrics literature to describe health expenditures using a pair of regression equations, one for the probability of expenditure and one for the conditional mean of expenditure [27, 83]. Two-part models have been applied to longitudinal data [104], but we do not want to split our response into zeros and non-zeros. Rather, we have a time-ordered outcome response that we believe resulted from two processes, one determining the level and the other determining the pattern of change over time. In many longitudinal applications, researchers are interested in both shape and level and want to study them simultaneously.

Bibliography

[1] C. Abraham et al. "Unsupervised curve clustering using B-splines". *Scandinavian Journal of Statistics* 30.3 (2003), pp. 581–595.

[2] H. Akaike. "A new look at the statistical model identification". *IEEE Transactions on Automatic Control* 19.6 (1974), pp. 716–723. DOI: 10.1109/TAC.1974.1100705.

[3] H. Akaike. "Information theory and an extension of the maximum likelihood principle". In: *Proceedings of 2nd International Symposium on Information Theory*. Tsahkadsor, Armenian SSR, 1973, pp. 267–281.

[4] B. T. Akingbemi et al. "Inhibition of testicular steroidogenesis by the xenoestrogen bisphenol A is associated with reduced pituitary luteinizing hormone secretion and decreased steroidogenic enzyme gene expression in rat leydig cells". *Endocrinology* 145.2 (2004), pp. 592–603. DOI: 10.1210/en.2003-1174.

[5] J. D. Banfield and A. E. Raftery. "Model-based Gaussian and non-Gaussian clustering". *Biometrics* 49.3 (1993), pp. 803–821.

[6] C. Biernacki, G. Celeux, and G. Govaert. "Assessing a mixture model for clustering with the integrated completed likelihood". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22.7 (2000), pp. 719–725. DOI: 10.1109/34.865189.

[7] O. Boldea and J. R. Magnus. "Maximum likelihood estimation of the multivariate normal mixture model". *Journal of the American Statistical Association* 104.488 (2009), pp. 1539–1549. DOI: 10.1198/jasa.2009.tm08273.

[8] R. E. Bonner. "On some clustering techniques". *IBM Journal of Research and Development* 8.1 (1964), pp. 22–32. DOI: 10.1147/rd.81.0022.

[9] J. M. Broadbent, W. M. Thomson, and R. Poulton. "Trajectory patterns of dental caries experience in the permanent dentition to the fourth decade of life". *Journal of Dental Research* 87.1 (2008), pp. 69–72. DOI: 10.1177/154405910808700112.

[10] M. A. Carter et al. "Trajectories of childhood weight gain: The relative importance of local environment versus individual social and early life factors". *PLoS ONE* 7.10 (2012), e47065.

[11] J. L. Carwile and K. B. Michels. "Urinary bisphenol A and obesity: NHANES 2003-2006". *Environmental Research* 111.6 (2011), pp. 825–830.

[12] G. Celeux and G. Govaert. "A classification EM algorithm for clustering and two stochastic versions". *Computational Statistics & Data Analysis* 14.3 (1992), pp. 315–332. DOI: 10.1016/0167-9473(92)90042-E.

[13] G. Celeux and G. Soromenho. "An entropy criterion for assessing the number of clusters in a mixture model". *Journal of Classification* 13.2 (1996), pp. 195–212. DOI: 10.1007/BF01246098.

[14] J.-M. Chiou and P.-L. Li. "Correlation-based functional clustering via subspace projection". *Journal of the American Statistical Association* 103.484 (2008), pp. 1684–1692. DOI: 10.1198/016214508000000814.

[15] A. Chouakria and P. Nagabhushan. "Adaptive dissimilarity index for measuring time series proximity". *Advances in Data Analysis and Classification* 1.1 (2007), pp. 5–21.

[16] L. M. Collins. "Analysis of longitudinal data: The integration of theoretical model, temporal design, and statistical model". *Annual Review of Psychology* 57 (2006), pp. 505–528.

[17] M. G. Cox. "The numerical evaluation of B-splines". *IMA Journal of Applied Mathematics* 10.2 (1972), pp. 134–149. DOI: 10.1093/imamat/10.2.134.

[18] R. De la Cruz-Mesía, F. A. Quintana, and G. Marshall. "Model-based clustering for longitudinal data". *Computational Statistics & Data Analysis* 52.3 (2008), pp. 1441–1457. DOI: 10.1016/j.csda.2007.04.005.

[19] H. B. Curry and I. J. Schoenberg. "On Pólya frequency functions IV: the fundamental spline functions and their limits". *Journal d'Analyse Mathématique* 17.1 (1966), pp. 71–107.

[20] S. Dasgupta. "Learning mixtures of Gaussians". In: *Proceedings of the IEEE Symposium on Foundations of Computer Science*. New York, NY, 1999, pp. 634–644.

[21] N. E. Day. "Estimating the components of a mixture of normal distributions". *Biometrika* 56.3 (1969), pp. 463–474. DOI: 10.1093/biomet/56.3.463.

[22] C. De Boor. *A Practical Guide to Splines*. New York: SpringerVerlag, 1978.

[23] C. De Boor. "Splines as linear combinations of B-splines. A survey." In: *Approximation Theory II*. Ed. by G. G. Lorentz, C. K. Chui, and L. L. Schumaker. New York: Academic Press, 1976, pp. 1–47.

[24] C. De Boor. "On calculating with B-splines". *Journal of Approximation Theory* 6.1 (1972), pp. 50–62.

[25] A. P. Dempster, N. M. Laird, and D. B. Rubin. "Maximum likelihood from incomplete data via the EM algorithm". *Journal of the Royal Statistical Society. Series B (Methodological)* 39.1 (1977), pp. 1–38.

[26] P. Diggle et al. *Analysis of Longitudinal Data*. 2nd ed. New York: Oxford University Press, 2002.

[27] N. Duan et al. "A comparison of alternative models for the demand for medical care". *Journal of Business & Economic Statistics* 1.2 (1983), pp. 115–126.

[28] S. Dudoit and J. Fridlyand. "A prediction-based resampling method for estimating the number of clusters in a dataset". *Genome Biology* 3.7 (2002), pp. 1–21. DOI: 10.1186/gb-2002-3-7-research0036.

[29] P. D'Urso. "Dissimilarity measures for time trajectories". *Statistical Methods & Applications* 9.1 (2000), pp. 53–83.

[30] M. B. Eisen et al. "Cluster analysis and display of genome-wide expression patterns". *Proceedings of the National Academy of Sciences* 95.25 (1998), pp. 14863–14868.

[31] B. Eskenazi et al. "Association of in utero organophosphate pesticide exposure and fetal growth and length of gestation in an agricultural population". *Environmental Health Perspectives* 112.10 (2004), pp. 1116–1124.

[32] B. Eskenazi et al. "CHAMACOS, a longitudinal birth cohort study: Lessons from the fields". *Journal of Children's Health* 1.1 (2003), pp. 3–27.

[33] B. Eskenazi et al. "Methodologic and logistic issues in conducting longitudinal birth cohort studies: Lessons learned from the Centers for Children's Environmental Health and Disease Prevention Research". *Environmental Health Perspectives* 113.10 (2005), pp. 1419–1429.

[34] R. L. Eubank. *Nonparametric Regression and Spline Smoothing*. New York, NY: Marcel Dekker, 1999.

[35] B. S. Everitt and D. J. Hand. *Finite Mixture Distributions*. London: Chapman & Hall, 1981.

[36] B. S. Everitt et al. *Cluster Analysis*. 5th ed. London: John Wiley & Sons, 2011.

[37] C. Fraley and A. E. Raftery. "How many clusters? Which clustering method? Answers via model-based cluster analysis". *The Computer Journal* 41.8 (1998), pp. 578–588. DOI: 10.1093/comjnl/41.8.578.

[38] C. Fraley and A. E. Raftery. "MCLUST: Software for model-based cluster analysis". *Journal of Classification* 16.2 (1999), pp. 297–306.

[39] C. Fraley and A. E. Raftery. "Model-based clustering, discriminant analysis, and density estimation". *Journal of the American Statistical Association* 97.458 (2002), pp. 611–631. DOI: 10.1198/016214502760047131.

[40] S. Frühwirth-Schnatter. *Finite Mixture and Markov Switching Models*. New York: Springer, 2006.

[41] S. Gaffney and P. Smyth. "Trajectory clustering with mixtures of regression models". In: *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Diego, 1999, pp. 63–72. DOI: 10.1145/312129.312198.

[42] F. L. Garden et al. "Body mass index (BMI) trajectories from birth to 11.5 years: Relation to early life food intake". *Nutrients* 4.10 (2012), pp. 1382–1398. DOI: 10.3390/nu4101382.

[43] C. Genolini and B. Falissard. "KmL: K-means for longitudinal data". *Computational Statistics* 25.2 (2010), pp. 317–328.

[44] G. Gray. "Bias in misspecified mixtures". *Biometrics* 50.2 (1994), pp. 457–470.

[45] B. Grün and F. Leisch. "Bootstrapping finite mixture models". In: *Proceedings of the 16th Symposium on Computational Statistics*. Ed. by A. J. Prague, Czech Republic, 2004, pp. 1115–1122.

[46] B. Grün and F. Leisch. "FlexMix version 2: Finite mixtures with concomitant variables and varying and constant parameters". *Journal of Statistical Software* 28.4 (2008), pp. 1–35.

[47] W. Hardeman et al. "Interventions to prevent weight gain: A systematic review of psychological models and behaviour change methods". *International Journal of Obesity and Related Metabolic Disorders* 24.2 (2000), pp. 131–143.

[48] K. G. Harley et al. "Prenatal and postnatal bisphenol A exposure and body mass index in childhood in the CHAMACOS cohort". *Environmental Health Perspectives* (2013), to appear.

[49] J. A. Hartigan and M. A. Wong. "Algorithm AS 136: A k-means clustering algorithm". *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28.1 (1979), pp. 100–108.

[50] R. J. Hathaway. "A constrained formulation of maximum-likelihood estimation for normal mixture distributions". *The Annals of Statistics* 13.2 (1985), pp. 795–800. DOI: 10.1214/aos/1176349557.

[51] B. C. Heggeseth and N. P. Jewell. "The impact of covariance misspecification in multivariate Gaussian mixtures on estimation and inference: An application to longitudinal modeling". *Statistics in Medicine* (2013), to appear. DOI: 10.1002/sim.5729.

[52] C. Hennig. "Identifiability of models for clusterwise linear regression". *Journal of Classification* 17.2 (2000), pp. 273–296. DOI: 10.1007/s003570000022.

[53] C. Hennig. "Methods for merging Gaussian mixture components". *Advances in Data Analysis and Classification* 4.1 (2010), pp. 3–34.

[54] D. B. Hitchcock, J. G. Booth, and G. Casella. "The effect of pre-smoothing functional data on cluster analysis". *Journal of Statistical Computation and Simulation* 77.12 (2007), pp. 1043–1055.

[55] M. Hiyama et al. "Bisphenol-A (BPA) affects reproductive formation across generations in mice". *Journal of Veterinary Medical Science* 73.9 (2011), pp. 1211–1215.

[56] K. L. Howdeshell et al. "Exposure to bisphenol A advances puberty". *Nature* 401.6755 (1999), pp. 763–764.

[57] L. Hubert and P. Arabie. "Comparing partitions". *Journal of Classification* 2.1 (1985), pp. 193–218.

[58] G. M. James and C. A. Sugar. "Clustering for sparsely sampled functional data". *Journal of the American Statistical Association* 98.462 (2003), pp. 397–408.

[59] R. I. Jennrich. "Asymptotic properties of non-linear least squares estimators". *The Annals of Mathematical Statistics* 40.2 (1969), pp. 633–643. DOI: 10.1214/aoms/1177697731.

[60] R. I. Jennrich and M. D. Schluchter. "Unbalanced repeated-measures models with structured covariance matrices". *Biometrics* 42.4 (1986), pp. 805–820.

[61] W. Jiang and M. A. Tanner. "On the identifiability of mixtures-of-experts". *Neural Networks* 12.9 (1999), pp. 1253–1258. DOI: 10.1016/S0893-6080(99)00066-0.

[62] B. L. Jones, D. S. Nagin, and K. Roeder. "A SAS procedure based on mixture models for estimating developmental trajectories". *Sociological Methods & Research* 29.3 (2001), pp. 374–393.

[63] B. L. Jones and D. S. Nagin. "Advances in group-based trajectory modeling and an SAS procedure for estimating them". *Sociological Methods & Research* 35.4 (2007), pp. 542–571. DOI: 10.1177/0049124106292364.

[64] R. E. Kass and A. E. Raftery. "Bayes factors". *Journal of the American Statistical Association* 90.430 (1995), pp. 773–795. DOI: 10.1080/01621459.1995.10476572.

[65] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken: John Wiley & Sons, 1990.

[66] J. Kiefer and J. Wolfowitz. "Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters". *The Annals of Mathematical Statistics* 27.4 (1956), pp. 887–906. DOI: 10.1214/aoms/1177728066.

[67] N. M. Kiefer. "Discrete parameter variation: Efficient estimation of a switching regression model". *Econometrica* 46.2 (1978), pp. 427–434.

[68] A. B. Koehler and E. S. Murphree. "A comparison of the Akaike and Schwarz criteria for selecting model order". *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 37.2 (1988), pp. 187–195.

[69] S. Kullback and R. A. Leibler. "On information and sufficiency". *The Annals of Mathematical Statistics* 22.1 (1951), pp. 79–86.

[70] L. M. Le Cam. "On some asymptotic properties of maximum likelihood estimates and related Bayes estimates". *University of California Publications in Statistics* 1.11 (1953), pp. 277–330.

[71] F. Leisch. "FlexMix: A general framework for finite mixture models and latent class regression in R". *Journal of Statistical Software* 11.8 (2004), pp. 1–18.

[72] B. G. Leroux. "Consistent estimation of a mixing distribution". *The Annals of Statistics* 20.3 (1992), pp. 1350–1360.

[73] C. Li et al. "Developmental trajectories of overweight during childhood: Role of early life factors". *Obesity (Silver Spring)* 15.3 (2007), pp. 760–771.

[74] J. Li. "Clustering based on a multilayer mixture model". *Journal of Computational and Graphical Statistics* 14.3 (2005), pp. 547–568.

[75] K.-Y. Liang and S. L. Zeger. "Longitudinal data analysis using generalized linear models". *Biometrika* 73.1 (1986), pp. 13–22. DOI: 10.1093/biomet/73.1.13.

[76] Y. Lo. "Bias from misspecification of the component variances in a normal mixture". *Computational Statistics & Data Analysis* 55.9 (2011), pp. 2739–2747. DOI: 10.1016/j.csda.2011.04.007.

[77] Y. Lo, N. R. Mendell, and D. B. Rubin. "Testing the number of components in a normal mixture". *Biometrika* 88.3 (2001), pp. 767–778. DOI: 10.1093/biomet/88.3.767.

[78] C. B. Lombard, A. A. Deeks, and H. J. Teede. "A systematic review of interventions aimed at the prevention of weight gain in adults". *Public Health Nutrition* 12.11 (2009), pp. 2236–2246.

[79] G. G. Lorentz. *Bernstein Polynomials*. Toronto: University of Toronto Press, 1953.

[80] Y. Luan and H. Li. "Clustering of time-course gene expression data using a mixed-effects model with B-splines". *Bioinformatics* 19.4 (2003), pp. 474–482. DOI: 10.1093/bioinformatics/btg014.

[81] J. MacQueen. "Some methods for classification and analysis of multivariate observations". In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. Ed. by L. M. Le Cam and J. Neyman. Vol. 1. Berkeley, 1967, pp. 281–297.

[82] J. Magidson and J Vermunt. "Latent class models for clustering: A comparison with k-means". *Canadian Journal of Marketing Research* 20.1 (2002), pp. 36–43.

[83] W. G. Manning et al. "A two-part model of the demand for medical care: Preliminary results from the health insurance study". In: *Health, Economics, and Health Economics*. Ed. by J. van der Gaag and M. Perlman. Amsterdam: North-Holland, 1981.

[84] S. I. McCoy et al. "A trajectory analysis of alcohol and marijuana use among Latino adolescents in San Francisco, California". *Journal of Adolescent Health* 47.6 (2010), pp. 564–574.

[85] G. J. McLachlan. "On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture". *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 36.3 (1987), pp. 318–324.

[86] G. J. McLachlan and K. E. Basford. *Mixture Models: Inference and Applications to Clustering*. New York: Marcel Dekker, 1988.

[87] G. J. McLachlan and D. Peel. *Finite Mixture Models*. New York: Wiley, 2000.

[88] G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. New York: Wiley, 1997.

[89] P. D. McNicholas and T. B. Murphy. "Model-based clustering of longitudinal data". *Canadian Journal of Statistics* 38.1 (2010), pp. 153–168. DOI: 10.1002/cjs.10047.

[90] G. W. Milligan and M. C. Cooper. "An examination of procedures for determining the number of clusters in a data set". *Psychometrika* 50.2 (1985), pp. 159–179. DOI: 10.1007/BF02294245.

[91] J. Miyawaki et al. "Perinatal and postnatal exposure to bisphenol A increases adipose tissue mass and serum cholesterol level in mice". *Journal of Atherosclerosis and Thrombosis* 14.5 (2007), pp. 245–252.

[92] C. Möller-Levet et al. "Fuzzy clustering of short time-series and unevenly distributed sampling points". In: *Proceedings of the Fifth International Conference on Intelligent Data Analysis*. Ed. by M. R. Berthold et al. Berlin, 2003, pp. 330–340.

[93] S. Mulvaney et al. "Trajectories of symptoms and impairment for pediatric patients with functional abdominal pain: a 5-year longitudinal study". *Journal of the American Academy of Child & Adolescent Psychiatry* 45.6 (2006), pp. 737–744. DOI: 10.1097/10.chi.0000214192.57993.06.

[94] B. Muthén. "Second-generation structural equation modeling with a combination of categorical and continuous latent variables: New opportunities for latent class–latent growth modeling". In: *New Methods for the Analysis of Change*. Ed. by L. M. Collins and A. G. Sayers. Washington, D.C.: American Psychological Association, 2001, pp. 291–322. DOI: 10.1037/10409-010.

[95] B. Muthén and L. Muthén. "Integrating person-centered and variable-centered analyses: Growth mixture modeling with latent trajectory classes". *Alcoholism: Clinical and Experimental Research* 24.6 (2000), pp. 882–891. DOI: 10.1111/j.1530-0277.2000.tb02070.x.

[96] B. Muthén et al. "General approaches to analysis of course: Applying growth mixture modeling to randomized trials of depression medication". In: *Causality and Psychopathology: Finding the Determinants of Disorders and their Cures*. Ed. by P. Shrout, K. Keyes, and K. Ornstein. New York: Oxford University Press, 2008, pp. 159–178.

[97] B. Muthén and K. Shedden. "Finite mixture modeling with mixture outcomes using the EM algorithm". *Biometrics* 55.2 (1999), pp. 463–469. DOI: 10.1111/j.0006-341X.1999.00463.x.

[98] L. K. Muthén and B. O. Muthén. *Mplus User's Guide*. Los Angeles, 1998-2010.

[99] D. S. Nagin. "Analyzing developmental trajectories: A semiparametric, group-based approach". *Psychological Methods* 4.2 (1999), pp. 139–157. DOI: 10.1037/1082-989X.4.2.139.

[100] D. S. Nagin. *Group-Based Modeling of Development*. Cambridge: Harvard University Press, 2005.

[101] D. S. Nagin et al. "Life course turning points: The effect of grade retention on physical aggression". *Developmental Psychopathology* 15.2 (2003), pp. 343–361.

[102] C. L. Ogden, M. D. Carroll, and K. M. Flegal. "High body mass index for age among US children and adolescents, 2003-2006". *Journal of the American Medical Association* 299.20 (2008), pp. 2401–2405. DOI: 10.1001/jama.299.20.2401.

[103] A. Oliveira-Brochado and F. V. Martins. *Assessing the number of components in mixture models: A review*. FEP Working Papers 194. Universidade do Porto, Faculdade de Economia do Porto, 2005.

[104] M. K. Olsen and J. L. Schafer. "A two-part random-effects model for semicontinuous longitudinal data". *Journal of the American Statistical Association* 96.454 (2001), pp. 730–745.

[105] T. Ostbye, R. Malhotra, and L. R. Landerman. "Body mass trajectories through adulthood: Results from the National Longitudinal Survey of Youth 1979 Cohort (1981-2006)". *International Journal of Epidemiology* 40.1 (2011), pp. 240–250. DOI: 10.1093/ije/dyq142.

[106] T. L. Phang et al. "Trajectory clustering: A non-parametric method for grouping gene expression time courses, with applications to mammary development". In: *Proceedings of the 8th Pacific Symposium on Biocomputing*. Lihue, Hawaii, 2003, pp. 351–362.

[107] A. Pickles and T. Croudace. "Latent mixture models for multivariate and longitudinal outcomes". *Statistical Methods in Medical Research* 19.3 (2010), pp. 271–289. DOI: 10.1177/0962280209105016.

[108] J. Prochazkova. "Derivative of B-Spline function". In: *Proceedings of the 25th Conference on Geometry and Computer Graphics*. Prague, Czech Republic, 2005.

[109] L. E. Pryor et al. "Developmental trajectories of body mass index in early childhood and their risk factors: An 8-year longitudinal study". *Archives of Pediatrics & Adolescent Medicine* 165.10 (2011), pp. 906–912. DOI: 10.1001/archpediatrics.2011.153.

[110] R. E. Quandt. "A new approach to estimating switching regressions". *Journal of the American Statistical Association* 67.338 (1972), pp. 306–310.

[111] J. O. Ramsay and B. W. Silverman. *Applied Functional Data Analysis: Methods and Case Studies*. New York: Springer, 2002.

[112] J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. 2nd ed. New York: Springer, 2005.

[113] W. M. Rand. "Objective criteria for the evaluation of clustering methods". *Journal of the American Statistical Association* 66.336 (1971), pp. 846–850.

[114] J. A. Rice and C. O. Wu. "Nonparametric mixed effects models for unequally sampled noisy curves". *Biometrics* 57.1 (2001), pp. 253–259. DOI: 10.1111/j.0006-341X.2001.00253.x.

[115] A. F. Roche et al. "Grading body fatness from limited anthropometric data". *The American Journal of Clinical Nutrition* 34.12 (1981), pp. 2831–2838.

[116] K. Roeder and L. Wasserman. "Practical Bayesian density estimation using mixtures of normals". *Journal of the American Statistical Association* 92.439 (1997), pp. 894–902. DOI: 10.1080/01621459.1997.10474044.

[117] A. Rotnitzky and N. P. Jewell. "Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data". *Biometrika* 77.3 (1990), pp. 485–497. DOI: 10.1093/biomet/77.3.485.

[118] P. J. Rousseeuw. "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis". *Journal of Computational and Applied Mathematics* 20 (1987), pp. 53–65.

[119] B. S. Rubin et al. "Perinatal exposure to low doses of bisphenol A affects body weight, patterns of estrous cyclicity, and plasma LH levels". *Environmental Health Perspectives* 109.7 (2001), pp. 675–680.

[120] D. Ruppert. "Selecting the number of knots for penalized splines". *Journal of Computational and Graphical Statistics* 11.4 (2002), pp. 735–757. DOI: 10.1198/106186002853.

[121] P. Schlattmann. *Medical Applications of Finite Mixture Models*. Springer, 2009.

[122] E. D. Schneiderman, S. M. Willis, and C. J. Kowalski. "Clustering on the basis of longitudinal data". *Computers in Biology and Medicine* 23.5 (1993), pp. 399–406. DOI: 10.1016/0010-4825(93)90137-P.

[123] L. L. Schumaker. *Spline Functions: Basic Theory*. New York: John Wiley & Sons, 1981.

[124] G. Schwarz. "Estimating the dimension of a model". *The Annals of Statistics* 6.2 (1978), pp. 461–464.

[125] N. Serban and L. Wasserman. "CATS: Cluster analysis by transformation and smoothing". *Journal of the American Statistical Association* 100.471 (2005), pp. 990–999.

[126] A. Shankar, S. Teppala, and C. Sabanayagam. "Urinary bisphenol A levels and measures of obesity: Results from the national health and nutrition examination survey 2003-2008". *ISRN Endocrinology* 2012 (2012), p. 965243.

[127] J. Shults et al. "A comparison of several approaches for choosing between working correlation structures in generalized estimating equation analysis of longitudinal binary data". *Statistics in Medicine* 28.18 (2009), pp. 2338–2355. DOI: 10.1002/sim.3622.

[128] R. H. Shumway and D. S. Stoffer. *Time Series Analysis and Its Applications*. 3rd ed. New York: Springer, 2011.

[129] E. Somm et al. "Perinatal exposure to bisphenol A alters early adipogenesis in the rat". *Environmental Health Perspectives* 117.10 (2009), pp. 1549–1555.

[130] G. Soromenho. "Comparing approaches for testing the number of components in a finite mixture model". *Computational Statistics* 9.4 (1994), pp. 65–82.

[131] T. Tarpey. "Linear transformations and the k-means clustering algorithm". *The American Statistician* 61.1 (2007), pp. 34–40.

[132] T. Tarpey and K. K. J. Kinateder. "Clustering functional data". *Journal of Classification* 20.1 (2003), pp. 93–114.

[133] H. Teicher. "Identifiability of finite mixtures". *The Annals of Mathematical Statistics* 34.4 (1963), pp. 1265–1269. DOI: 10.1214/aoms/1177703862.

[134] R. Tibshirani, G. Walther, and T. Hastie. "Estimating the number of clusters in a data set via the gap statistic". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63.2 (2001), pp. 411–423. DOI: 10.1111/1467-9868.00293.

[135] D. M. Titterington, A. F. M. Smith, and U. E. Makov. *Statistical Analysis of Finite Mixture Distributions*. New York: Wiley, 1985.

[136] L. Trasande, T. M. Attina, and J. Blustein. "Association between urinary bisphenol A concentration and obesity prevalence in children and adolescents". *Journal of the American Medical Association* 308.11 (2012), pp. 1113–1121.

[137] R. S. Tuma. "Environmental chemicals—not just overeating—may cause obesity". *Journal of the National Cancer Institute* 99.11 (2007), p. 835. DOI: 10.1093/jnci/djk230.

[138] U.S. Department of Health and Human Services. *The Surgeon General's Call to Action to Prevent and Decrease Overweight and Obesity*. Rockville, MD: U. S. Department of Health and Human Services, Public Health Services, Office of the Surgeon General, 2001.

[139] J. K. Vermunt and J. Magidson. *Latent GOLD 4.0 User's Guide*. Belmont, Massachusetts, 2005.

[140] A. Wald. "Note on the consistency of the maximum likelihood estimate". *The Annals of Mathematical Statistics* 20.4 (1949), pp. 595–601. DOI: 10.1214/aoms/1177729952.

[141] C.-P. Wang, C. H. Brown, and K. Bandeen-Roche. "Residual diagnostics for growth mixture models". *Journal of the American Statistical Association* 100.471 (2005), pp. 1054–1076. DOI: 10.1198/016214505000000501.

[142] X. Wang, K. Smith, and R. Hyndman. "Characteristic-based clustering for time series data". *Data Mining and Knowledge Discovery* 13.3 (2006), pp. 335–364. DOI: 10.1007/s10618-005-0039-x.

[143] M. Wedel. "Concomitant variables in finite mixture models". *Statistica Neerlandica* 56.3 (2002), pp. 362–375. DOI: 10.1111/1467-9574.t01-1-00072.

[144] J. Wei et al. "Perinatal exposure to bisphenol A at reference dose predisposes offspring to metabolic syndrome in adult rats on a high-fat diet". *Endocrinology* 152.8 (2011), pp. 3049–3061.

[145] R. C. Whitaker et al. "Predicting obesity in young adulthood from childhood and parental obesity". *New England Journal of Medicine* 337.13 (1997), pp. 869–873.

[146] H. White. "Maximum likelihood estimation of misspecified models". *Econometrica* 50.1 (1982), pp. 1–25.

[147] M. Windle and M. Wiesner. "Trajectories of marijuana use from adolescence to young adulthood: Predictors and outcomes". *Development and Psychopathology* 16.4 (2004), pp. 1007–1027. DOI: 10.1017/S0954579404040118.

[148] J. H. Wolfe. *A Monte Carlo study of sampling distribution of the likelihood ratio for mixtures of multinormal distributions*. Tech. rep. San Diego: U. S. Naval Personnel and Training Research Laboratory.

[149] X. Xu et al. "Changed preference for sweet taste in adulthood induced by perinatal exposure to bisphenol A—A probable link to overweight and obesity". *Neurotoxicology and Teratology* 33.4 (2011), pp. 458 –463. DOI: 10.1016/j.ntt.2011.06.002.

[150] S. J. Yakowitz and J. D. Spragins. "On the identifiability of finite mixtures". *The Annals of Mathematical Statistics* 39.1 (1968), pp. 209–214. DOI: 10.1214/aoms/1177698520.

[151] G. O. Zerbe. "A new nonparametric technique for constructing percentiles and normal ranges for growth curves determined from longitudinal data." *Growth* 43.4 (1979), pp. 263–272.

[152] S. Zhou and X. Shen. "Spatially adaptive regression splines and accurate knot selection schemes". *Journal of the American Statistical Association* 96.453 (2001), pp. 247–259. DOI: 10.1198/016214501750332820.

Appendix A

Standard errors derivations

Let $\mathbf{y}_1, ..., \mathbf{y}_n$ be a sample of independent m-variate random vectors drawn from mixture model

$$f(\mathbf{y}|\mathbf{x}, \mathbf{w}, \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k(\mathbf{w}, \boldsymbol{\gamma}) f_k(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}_k)$$

where $f_k(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}_k)$ is the density for a multivariate Gaussian with mean $\mathbf{x}\boldsymbol{\beta}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$. The log-likelihood is written as

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(\mathbf{y}_i|\mathbf{x}_i, \mathbf{w}_i, \boldsymbol{\theta})$$

where the complete parameter vector is $\boldsymbol{\theta} = (\boldsymbol{\gamma}_1^T, ..., \boldsymbol{\gamma}_{K-1}^T, \boldsymbol{\theta}_1^T, ..., \boldsymbol{\theta}_K^T)^T$ and $\boldsymbol{\theta}_k$ includes mean and covariance parameters. The score vector is defined by $\mathbf{q}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \mathbf{q}_i(\boldsymbol{\theta})$, where

$$\mathbf{q}_i(\boldsymbol{\theta}) = \frac{\partial \log f(\mathbf{y}_i|\mathbf{x}_i, \mathbf{w}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = ((\mathbf{q}_i^{\gamma_1})^T, ..., (\mathbf{q}_i^{\gamma_{K-1}})^T, (\mathbf{q}_i^{\theta_1})^T, ..., (\mathbf{q}_i^{\theta_K})^T)^T$$

and the Hessian matrix defined by $\mathbf{Q}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \mathbf{Q}_i(\boldsymbol{\theta})$, where

$$\mathbf{Q}_i(\boldsymbol{\theta}) = \frac{\partial^2 \log f(\mathbf{y}_i|\mathbf{x}_i, \mathbf{w}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} = \begin{pmatrix} \mathbf{Q}_i^{\gamma_1\gamma_1} & \cdots & \mathbf{Q}_i^{\gamma_1\gamma_{K-1}} & \mathbf{Q}_i^{\gamma_1\theta_1} & \cdots & \mathbf{Q}_i^{\gamma_1\theta_K} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{Q}_i^{\gamma_{K-1}\gamma_1} & \cdots & \mathbf{Q}_i^{\gamma_{K-1}\gamma_{K-1}} & \mathbf{Q}_i^{\gamma_{K-1}\theta_1} & \cdots & \mathbf{Q}_i^{\gamma_{K-1}\theta_K} \\ \mathbf{Q}_i^{\theta_1\gamma_1} & \cdots & \mathbf{Q}_i^{\theta_1\gamma_{K-1}} & \mathbf{Q}_i^{\theta_1\theta_1} & \cdots & \mathbf{Q}_i^{\theta_1\theta_K} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{Q}_i^{\theta_K\gamma_1} & \cdots & \mathbf{Q}_i^{\theta_K\gamma_{K-1}} & \mathbf{Q}_i^{\theta_K\theta_1} & \cdots & \mathbf{Q}_i^{\theta_K\theta_K} \end{pmatrix}.$$

Also, we define

$$\phi_{ik} = \pi_k(\mathbf{w}_i, \boldsymbol{\gamma}) f_k(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\theta}_k), \;\; \alpha_{ik} = \frac{\phi_{ik}}{\sum_k \phi_{ik}} \tag{A.1}$$

$$\mathbf{b}_{ik} = \boldsymbol{\Sigma}_{ik}^{-1}(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\beta}_k), \;\; \mathbf{B}_{ik} = \boldsymbol{\Sigma}_{ik}^{-1} - \mathbf{b}_{ik}\mathbf{b}_{ik}^T, \tag{A.2}$$

Theorem 2. *For the multivariate Gaussian mixture model, the contribution of the ith observation to the score vector with respect to the parameters, $\boldsymbol{\gamma}_k$ (k=1,...,K-1) and $\boldsymbol{\theta}_k$ $(k = 1, .., K)$, is given by*

$$\mathbf{q}_i^{\gamma_k} = (\alpha_{ik} - \pi_k(\mathbf{w}_i, \boldsymbol{\gamma}))\mathbf{w}_i, \quad \mathbf{q}_i^{\theta_k} = \alpha_{ik}\mathbf{c}_{ik}.$$

The contribution of the ith observation to the Hessian matrix is given by

$$\begin{aligned}
\mathbf{Q}_i^{\gamma_k\gamma_k} &= [\alpha_{ik}(1-\alpha_{ik}) - \pi_k(\mathbf{w}_i, \boldsymbol{\gamma})(1-\pi_k(\mathbf{w}_i, \boldsymbol{\gamma}))]\mathbf{w}_i\mathbf{w}_i^T \\
\mathbf{Q}_i^{\gamma_k\gamma_l} &= -[\alpha_{ik}\alpha_{il} - \pi_k(\mathbf{w}_i, \boldsymbol{\gamma})\pi_l(\mathbf{w}_i, \boldsymbol{\gamma})]\mathbf{w}_i\mathbf{w}_i^T
\end{aligned}$$

$$\begin{aligned}
\mathbf{Q}_i^{\gamma_k\theta_k} = \alpha_{ik}(1-\alpha_{ik})\mathbf{w}_i\mathbf{c}_{ik}^T, \quad \mathbf{Q}_i^{\gamma_k\theta_l} &= -\alpha_{ik}\alpha_{il}\mathbf{w}_i\mathbf{c}_{il}^T \\
\mathbf{Q}_i^{\theta_k\theta_k} = -(\alpha_{ik}\mathbf{C}_{ik} - \alpha_{ik}(1-\alpha_{ik})\mathbf{c}_{ik}\mathbf{c}_{ik}^T), \quad \mathbf{Q}_i^{\theta_k\theta_l} &= -\alpha_{ik}\alpha_{il}\mathbf{c}_{ik}\mathbf{c}_{il}^T \quad (k \neq l).
\end{aligned}$$

where $\mathbf{c}_{ik}$ and $\mathbf{C}_{ik}$ are defined as follows.

If $\boldsymbol{\Sigma}_{ik} = \sigma_k^2\mathbf{I}$, then $\boldsymbol{\theta}_k = (\boldsymbol{\beta}_k, \sigma_k^2)$ and

$$\mathbf{c}_{ik} = \begin{pmatrix} \mathbf{x}_i^T\mathbf{b}_{ik} \\ -\frac{1}{2}\operatorname{tr}(\mathbf{B}_{ik}) \end{pmatrix}, \quad \mathbf{C}_{ik} = \begin{pmatrix} \frac{1}{\sigma_k^2}\mathbf{x}_i^T\mathbf{x}_i & \frac{1}{\sigma_k^2}\mathbf{x}_i^T\mathbf{b}_{ik} \\ \frac{1}{\sigma_k^2}\mathbf{b}_{ik}^T\mathbf{x}_i & \frac{1}{2\sigma_k^2}\operatorname{tr}(\boldsymbol{\Lambda}_{ik}) \end{pmatrix}$$

where $\boldsymbol{\Lambda}_{ik} = \boldsymbol{\Sigma}_{ik}^{-1} - 2\mathbf{B}_{ik}$.

If $\boldsymbol{\Sigma}_{ik} = \sigma_k^2\mathbf{R}_{ik}$, then $\boldsymbol{\theta}_k = (\boldsymbol{\beta}_k, \sigma_k^2, \rho_k)$ and

$$\mathbf{c}_{ik} = \begin{pmatrix} \mathbf{x}_i^T\mathbf{b}_{ik} \\ -\frac{1}{2}\operatorname{tr}(\mathbf{B}_{ik}\mathbf{R}_{ik}) \\ -\frac{\sigma_k^2}{2}\operatorname{tr}(\mathbf{B}_{ik}\mathbf{R}_{ik}') \end{pmatrix},$$

$$\mathbf{C}_{ik} = \begin{pmatrix} \frac{1}{\sigma_k^2}\mathbf{x}_i^T\mathbf{R}_{ik}^{-1}\mathbf{x}_i & \frac{1}{\sigma_k^2}\mathbf{x}_i^T\mathbf{b}_{ik} & \mathbf{x}_i^T\mathbf{R}_{ik}^{-1}\mathbf{R}_{ik}'\mathbf{b}_{ik} \\ \frac{1}{\sigma_k^2}\mathbf{b}_{ik}^T\mathbf{x}_i & \frac{1}{2\sigma_k^2}\operatorname{tr}(\boldsymbol{\Lambda}_{ik}\mathbf{R}_{ik}) & \frac{1}{2}\operatorname{tr}(\boldsymbol{\Lambda}_{ik}\mathbf{R}_{ik}') \\ \mathbf{b}_{ik}^T\mathbf{R}_{ik}'\mathbf{R}_{ik}^{-1}\mathbf{x}_i & \frac{1}{2}\operatorname{tr}(\boldsymbol{\Lambda}_{ik}\mathbf{R}_{ik}') & \frac{\sigma_k^2}{2}\operatorname{tr}(\boldsymbol{\Lambda}_{ik}\mathbf{R}_{ik}'\mathbf{R}_{ik}^{-1}\mathbf{R}_{ik}') \end{pmatrix}$$

where $\boldsymbol{\Lambda}_{ik} = \boldsymbol{\Sigma}_{ik}^{-1} - 2\mathbf{B}_{ik}$, $\mathbf{R}_{ik}$ is a correlation matrix based on the parameter ρ_k, and $\mathbf{R}_{ik}'$ is the derivative of $\mathbf{R}_{ik}$ with respect to ρ_k.

If $\boldsymbol{\Sigma}_{ik} = \nu_k^2\mathbf{J} + \sigma_k^2\mathbf{R}_{ik}$, then $\boldsymbol{\theta}_k = (\boldsymbol{\beta}_k, \sigma_k^2, \rho_k, \nu_k^2)$ and

$$\mathbf{c}_{ik} = \begin{pmatrix} \mathbf{x}_i^T\mathbf{b}_{ik} \\ -\frac{1}{2}\operatorname{tr}(\mathbf{B}_{ik}\mathbf{R}_{ik}) \\ -\frac{\sigma_k^2}{2}\operatorname{tr}(\mathbf{B}_{ik}\mathbf{R}_{ik}') \\ -\frac{1}{2}\operatorname{tr}(\mathbf{B}_{ik}\mathbf{J}) \end{pmatrix}, \quad \mathbf{C}_{ik} = \begin{pmatrix} \mathbf{x}_i^T\boldsymbol{\Sigma}_{ik}^{-1}\mathbf{x}_i & (\mathbf{C}_{ik}^1)^T \\ \mathbf{C}_{ik}^1 & \mathbf{C}_{ik}^2 \end{pmatrix}$$

with

$$\mathbf{C}^1_{ik} = \begin{pmatrix} \mathbf{b}^T_{ik}\mathbf{R}_{ik}\boldsymbol{\Sigma}^{-1}_{ik}\mathbf{x}_i \\ \sigma^2_k\mathbf{b}^T_{ik}\mathbf{R}'_{ik}\boldsymbol{\Sigma}^{-1}_{ik}\mathbf{x}_i \\ \mathbf{b}^T_{ik}\mathbf{J}\boldsymbol{\Sigma}^{-1}_{ik}\mathbf{x}_i \end{pmatrix},$$

$$\mathbf{C}^2_{ik} = \begin{pmatrix} \frac{1}{2}\operatorname{tr}(\boldsymbol{\Lambda}_{ik}\mathbf{R}_{ik}\boldsymbol{\Sigma}^{-1}_{ik}\mathbf{R}_{ik}) & \frac{\sigma^2_k}{2}\operatorname{tr}(\boldsymbol{\Lambda}_{ik}\mathbf{R}_{ik}\boldsymbol{\Sigma}^{-1}_{ik}\mathbf{R}'_{ik}) & \frac{1}{2}\operatorname{tr}(\boldsymbol{\Lambda}_{ik}\mathbf{R}_{ik}\boldsymbol{\Sigma}^{-1}_{ik}\mathbf{J}) \\ \frac{\sigma^2_k}{2}\operatorname{tr}(\boldsymbol{\Lambda}_{ik}\mathbf{R}'_{ik}\boldsymbol{\Sigma}^{-1}_{ik}\mathbf{R}_{ik}) & \frac{(\sigma^2_k)^2}{2}\operatorname{tr}(\boldsymbol{\Lambda}_{ik}\mathbf{R}'_{ik}\boldsymbol{\Sigma}^{-1}_{ik}\mathbf{R}'_{ik}) & \frac{\sigma^2_k}{2}\operatorname{tr}(\boldsymbol{\Lambda}_{ik}\mathbf{R}'_{ik}\boldsymbol{\Sigma}^{-1}_{ik}\mathbf{J}) \\ \frac{1}{2}\operatorname{tr}(\boldsymbol{\Lambda}_{ik}\mathbf{J}\boldsymbol{\Sigma}^{-1}_{ik}\mathbf{R}_{ik}) & \frac{\sigma^2_k}{2}\operatorname{tr}(\boldsymbol{\Lambda}_{ik}\mathbf{J}\boldsymbol{\Sigma}^{-1}_{ik}\mathbf{R}'_{ik}) & \frac{1}{2}\operatorname{tr}(\boldsymbol{\Lambda}_{ik}\mathbf{J}\boldsymbol{\Sigma}^{-1}_{ik}\mathbf{J}) \end{pmatrix}$$

where $\mathbf{J}$ *is an* $m \times m$ *matrix of 1's,* $\boldsymbol{\Lambda}_{ik} = \boldsymbol{\Sigma}^{-1}_{ik} - 2\mathbf{B}_{ik}$, $\mathbf{R}_{ik}$ *is a correlation matrix based on the parameter* ρ_k*, and* $\mathbf{R}'_{ik}$ *is the derivative of* $\mathbf{R}_{ik}$ *with respect to* ρ_k*.*

Proof. I follow a procedure similar to Boldea and Magnus [7]. Let $g(\mathbf{w}) = \sum_{j=1}^K \exp(\mathbf{w}^T\boldsymbol{\gamma}_j)$. Then $\pi_k(\mathbf{w}, \boldsymbol{\gamma}) = \frac{\exp(\mathbf{w}^T\boldsymbol{\gamma}_k)}{g(\mathbf{w})}$. Taking the logarithm of both sides, we get

$$\log \pi_k(\mathbf{w}, \boldsymbol{\gamma}) = \mathbf{w}^T\boldsymbol{\gamma}_k - \log g(\mathbf{w}) \tag{A.3}$$

Note that $\partial \log g(\mathbf{w})/\partial\boldsymbol{\gamma}_j = \pi_j(\mathbf{w}, \boldsymbol{\gamma})\mathbf{w}$. Then,

$$\begin{aligned} \frac{d\log\pi_k(\mathbf{w},\boldsymbol{\gamma})}{d\boldsymbol{\gamma}_j} &= I(j=k)\mathbf{w} - \pi_j(\mathbf{w},\boldsymbol{\gamma})\mathbf{w} \quad \text{for } j = 1, ..., K-1 \\ \frac{d^2\log\pi_k(\mathbf{w},\boldsymbol{\gamma})}{d\boldsymbol{\gamma}_j d\boldsymbol{\gamma}_l^T} &= [-I(j=l)\pi_j(\mathbf{w},\boldsymbol{\gamma}) + \pi_j(\mathbf{w},\boldsymbol{\gamma})\pi_l(\mathbf{w},\boldsymbol{\gamma})]\mathbf{w}\mathbf{w}^T \quad \text{for } j, l = 1, ..., K-1 \end{aligned} \tag{A.4}$$

Let ϕ_{ik} and α_{ik} be defined as in (A.1). Then since $f(\mathbf{y}_i|\mathbf{x}_i, \mathbf{w}_i, \boldsymbol{\theta}) = \sum_{k=1}^K \phi_{ik}$, we obtain

$$d\log f(\mathbf{y}_i|\mathbf{x}_i,\mathbf{w}_i,\boldsymbol{\theta}) = \frac{df(\mathbf{y}_i|\mathbf{x}_i,\mathbf{w}_i,\boldsymbol{\theta})}{f(\mathbf{y}_i|\mathbf{x}_i,\mathbf{w}_i,\boldsymbol{\theta})} = \sum_{k=1}^K \frac{d\phi_{ik}}{\sum_{j=1}^K \phi_{ij}} = \sum_{k=1}^K \alpha_{ik} d\log\phi_{ik} \tag{A.5}$$

and

$$\begin{aligned} d^2\log f(\mathbf{y}_i|\mathbf{x}_i,\mathbf{w}_i,\boldsymbol{\theta}) &= \left(\frac{d^2 f(\mathbf{y}_i|\mathbf{x}_i,\mathbf{w}_i,\boldsymbol{\theta})}{f(\mathbf{y}_i|\mathbf{x}_i,\mathbf{w}_i,\boldsymbol{\theta})} - \left(\frac{df(\mathbf{y}_i|\mathbf{x}_i,\mathbf{w}_i,\boldsymbol{\theta})}{f(\mathbf{y}_i|\mathbf{x}_i,\mathbf{w}_i,\boldsymbol{\theta})}\right)^2\right) \\ &= \left(\frac{\sum_{k=1}^K d^2\phi_{ik}}{\sum_{k=1}^K \phi_{ik}} - \left(\frac{\sum_{k=1}^K d\phi_{ik}}{\sum_{k=1}^K \phi_{ik}}\right)^2\right) \\ &= \left(\sum_{k=1}^K \alpha_{ik}(d^2\log\phi_{ik} + (d\log\phi_{ik})^2) - (\sum_{k=1}^K \alpha_{ik} d\log\phi_{ik})^2\right) \end{aligned} \tag{A.6}$$

To evaluate these, we first need the first- and second-order derivatives of $\log \phi_{ik}$. Since,

$$\log f_k(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\theta}_k) = -\frac{m}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}_{ik}^{-1}| - (\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\beta}_k)^T\boldsymbol{\Sigma}_{ik}^{-1}(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\beta}_k)$$

we find

$$\begin{aligned} d\log f_k(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\theta}_k) &= -\frac{1}{2}d\log|\boldsymbol{\Sigma}_{ik}| + (\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\beta}_k)^T\boldsymbol{\Sigma}_{ik}^{-1}d(\mathbf{x}_i\boldsymbol{\beta}_k) \\ &\quad - \frac{1}{2}(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\beta}_k)^T d\boldsymbol{\Sigma}_{ik}^{-1}(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\beta}_k) \\ &= -\frac{1}{2}\operatorname{tr}(\boldsymbol{\Sigma}_{ik}^{-1}d\boldsymbol{\Sigma}_{ik}) + (\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\beta}_k)^T\boldsymbol{\Sigma}_{ik}^{-1}d(\mathbf{x}_i\boldsymbol{\beta}_k) \\ &\quad + \frac{1}{2}(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\beta}_k)^T\boldsymbol{\Sigma}_{ik}^{-1}d\boldsymbol{\Sigma}_{ik}\boldsymbol{\Sigma}_{ik}^{-1}(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\beta}_k) \end{aligned}$$

and

$$\begin{aligned} d^2\log f_k(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\theta}_k) &= -\frac{1}{2}\operatorname{tr}(d\boldsymbol{\Sigma}_{ik}^{-1}d\boldsymbol{\Sigma}_{ik}) - d(\mathbf{x}_i\boldsymbol{\beta}_k)^T\boldsymbol{\Sigma}_{ik}^{-1}d(\mathbf{x}_i\boldsymbol{\beta}_k) + (\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\beta}_k)^T d\boldsymbol{\Sigma}_{ik}^{-1}d(\mathbf{x}_i\boldsymbol{\beta}_k) \\ &\quad - (\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\beta}_k)^T\boldsymbol{\Sigma}_{ik}^{-1}d\boldsymbol{\Sigma}_{ik}\boldsymbol{\Sigma}_{ik}^{-1}d(\mathbf{x}_i\boldsymbol{\beta}_k) \\ &\quad - (\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\beta}_k)^T\boldsymbol{\Sigma}_{ik}^{-1}d\boldsymbol{\Sigma}_{ik}\boldsymbol{\Sigma}_{ik}^{-1}d\boldsymbol{\Sigma}_{ik}\boldsymbol{\Sigma}_{ik}^{-1}(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\beta}_k) \\ &= \frac{1}{2}\operatorname{tr}(\boldsymbol{\Sigma}_{ik}^{-1}d\boldsymbol{\Sigma}_{ik}\boldsymbol{\Sigma}_{ik}^{-1}d\boldsymbol{\Sigma}_{ik}) - d(\mathbf{x}_i\boldsymbol{\beta}_k)^T\boldsymbol{\Sigma}_{ik}^{-1}d(\mathbf{x}_i\boldsymbol{\beta}_k) \\ &\quad - 2(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\beta}_k)^T\boldsymbol{\Sigma}_{ik}^{-1}d\boldsymbol{\Sigma}_{ik}\boldsymbol{\Sigma}_{ik}^{-1}d(\mathbf{x}_i\boldsymbol{\beta}_k) \\ &\quad - (\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\beta}_k)^T\boldsymbol{\Sigma}_{ik}^{-1}d\boldsymbol{\Sigma}_{ik}\boldsymbol{\Sigma}_{ik}^{-1}d\boldsymbol{\Sigma}_{ik}\boldsymbol{\Sigma}_{ik}^{-1}(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\beta}_k) \end{aligned}$$

and hence, using (A.4) and the definitions (A.1)-(A.2),

$$\begin{aligned} d\log\phi_{ik} &= d\log\pi_k(\mathbf{w}_i, \boldsymbol{\gamma}) + (\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\beta}_k)^T\boldsymbol{\Sigma}_{ik}^{-1}d(\mathbf{x}_i\boldsymbol{\beta}_k) - \frac{1}{2}\operatorname{tr}(\boldsymbol{\Sigma}_{ik}^{-1}d\boldsymbol{\Sigma}_{ik}) \\ &\quad + \frac{1}{2}(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\beta}_k)^T\boldsymbol{\Sigma}_{ik}^{-1}d\boldsymbol{\Sigma}_{ik}\boldsymbol{\Sigma}_{ik}^{-1}(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\beta}_k) \\ &= \mathbf{w}_i d\boldsymbol{\gamma}_k - \pi_k(\mathbf{w}_i, \boldsymbol{\gamma})\mathbf{w}_i d\boldsymbol{\gamma}_\cdot + \mathbf{b}_{ik}^T d(\mathbf{x}_i\boldsymbol{\beta}_k) - \frac{1}{2}\operatorname{tr}(\mathbf{B}_{ik}d\boldsymbol{\Sigma}_{ik}) \end{aligned}$$

and

$$\begin{aligned} d^2\log\phi_{ik} &= d^2\log\pi_k(\mathbf{w}_i, \boldsymbol{\gamma}) - d(\mathbf{x}_i\boldsymbol{\beta}_k)^T\boldsymbol{\Sigma}_{ik}^{-1}d(\mathbf{x}_i\boldsymbol{\beta}_k) - 2(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\beta}_k)^T\boldsymbol{\Sigma}_{ik}^{-1}d\boldsymbol{\Sigma}_{ik}\boldsymbol{\Sigma}_{ik}^{-1}d(\mathbf{x}_i\boldsymbol{\beta}_k) \\ &\quad - (\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\beta}_k)^T\boldsymbol{\Sigma}_{ik}^{-1}d\boldsymbol{\Sigma}_{ik}\boldsymbol{\Sigma}_{ik}^{-1}d\boldsymbol{\Sigma}_{ik}\boldsymbol{\Sigma}_{ik}^{-1}(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\beta}_k) + \frac{1}{2}\operatorname{tr}(\boldsymbol{\Sigma}_{ik}^{-1}d\boldsymbol{\Sigma}_{ik}\boldsymbol{\Sigma}_{ik}^{-1}d\boldsymbol{\Sigma}_{ik}) \\ &= -d\boldsymbol{\gamma}_\cdot^T\pi_\cdot(\mathbf{w}_i, \boldsymbol{\gamma})\mathbf{w}_i\mathbf{w}_i^T d\boldsymbol{\gamma}_\cdot + d\boldsymbol{\gamma}_\cdot^T\pi_\cdot(\mathbf{w}_i, \boldsymbol{\gamma})\pi_\cdot(\mathbf{w}_i, \boldsymbol{\gamma})\mathbf{w}_i\mathbf{w}_i^T d\boldsymbol{\gamma}_\cdot \\ &\quad - d(\mathbf{x}_i\boldsymbol{\beta}_k)^T\boldsymbol{\Sigma}_{ik}^{-1}d(\mathbf{x}_i\boldsymbol{\beta}_k) - 2\mathbf{b}_{ik}^T(d\boldsymbol{\Sigma}_{ik})\boldsymbol{\Sigma}_{ik}d(\mathbf{x}_i\boldsymbol{\beta}_k) \\ &\quad - \frac{1}{2}\operatorname{tr}(\boldsymbol{\Sigma}_{ik}^{-1} - 2\mathbf{B}_{ik})d\boldsymbol{\Sigma}_{ik}\boldsymbol{\Sigma}_{ik}^{-1}d\boldsymbol{\Sigma}_{ik} \end{aligned}$$

For specific structures of $\mathbf{\Sigma}_{ik}$, we substitute in derivative of $\mathbf{\Sigma}_{ik}$.

If $\mathbf{\Sigma}_{ik} = \sigma_k^2\mathbf{I}$, then $\boldsymbol{\theta}_k^T = (\boldsymbol{\beta}_k^T, \sigma_k^2)$ and

$$d\mathbf{\Sigma}_{ik} = d\sigma_k^2\mathbf{I}.$$

If $\mathbf{\Sigma}_{ik} = \sigma_k^2\mathbf{R}_{ik}$, then $\boldsymbol{\theta}_k^T = (\boldsymbol{\beta}_k^T, \sigma_k^2, \rho_k)$ and

$$d\mathbf{\Sigma}_{ik} = d\sigma_k^2\mathbf{R}_{ik} + \sigma_k^2\mathbf{R}_{ik}^{'}d\rho_k.$$

If $\mathbf{\Sigma}_{ik} = \nu_k^2\mathbf{J} + \sigma_k^2\mathbf{R}_{ik}$, then $\boldsymbol{\theta}_k^T = (\boldsymbol{\beta}_k^T, \sigma_k^2, \rho_k)$ and

$$d\mathbf{\Sigma}_{ik} = d\nu_k^2\mathbf{J} + d\sigma_k^2\mathbf{R}_{ik} + \sigma_k^2\mathbf{R}_{ik}^{'}d\rho_k.$$

Using these calculations, we can rewrite the previously equations with the definitions of $\mathbf{c}_{ik}$ and $\mathbf{C}_{ik}$ in the the theorem,

$$d\log\phi_{ik} = \mathbf{w}_i d\boldsymbol{\gamma}_k - \pi_k(\mathbf{w}_i, \boldsymbol{\gamma})\mathbf{w}_i d\boldsymbol{\gamma}_\cdot + \mathbf{c}_{ik}^T d\boldsymbol{\theta}_k \tag{A.7}$$

and

$$d^2\log\phi_{ik} = -d\boldsymbol{\gamma}_\cdot^T\pi_\cdot(\mathbf{w}_i, \boldsymbol{\gamma})\mathbf{w}_i\mathbf{w}_i^T d\boldsymbol{\gamma}_\cdot + d\boldsymbol{\gamma}_\cdot^T\pi_\cdot(\mathbf{w}_i, \boldsymbol{\gamma})\pi_{\cdot\cdot}(\mathbf{w}_i, \boldsymbol{\gamma})\mathbf{w}_i\mathbf{w}_i^T d\boldsymbol{\gamma}_{\cdot\cdot} - d\boldsymbol{\theta}_k^T\mathbf{C}_{ik}d\boldsymbol{\theta}_k \tag{A.8}$$

Inserting (A.7) in (A.5), and (A.8) and (A.7) in (A.6) completes the proof. □

By definition, the conventional variance-covariance estimates are
$A_n(\hat{\boldsymbol{\theta}}_n) = n^{-1}\sum_{i=1}^n \mathbf{Q}_i(\hat{\boldsymbol{\theta}}_n)$ and $B_n(\hat{\boldsymbol{\theta}}_n) = n^{-1}\sum_{i=1}^n \mathbf{q}_i(\hat{\boldsymbol{\theta}}_n)\mathbf{q}_i^T(\hat{\boldsymbol{\theta}}_n)$.

Appendix B

Ascending property of the CEM algorithm

Theorem 1. *The classification likelihood $L(\boldsymbol{\theta}, \boldsymbol{\eta})$ defined in Section 4.2 is non-decreasing after each update of $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$ by the CEM algorithm. That is, $L(\boldsymbol{\theta}^{(t+1)}, \boldsymbol{\eta}^{(t+1)}) \geq L(\boldsymbol{\theta}^{(t)}, \boldsymbol{\eta}^{(t)})$ for all t.*

Proof. We now prove the ascending property of the CEM algorithm used to estimate parameters in the multilayer mixture model, which is stated in Theorem 1 in Chapter 4. We need to show that

$$L(\boldsymbol{\theta}^{(t+1)}, \boldsymbol{\eta}^{(t+1)}) \geq L(\boldsymbol{\theta}^{(t)}, \boldsymbol{\eta}^{(t)}). \tag{B.1}$$

The CEM algorithm alternatives between optimizing $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$, keeping the other fixed. After the classification step, the parameter vector is updated from $(\boldsymbol{\theta}^{(t)}, \boldsymbol{\eta}^{(t)})$ to $(\boldsymbol{\theta}^{(t)}, \boldsymbol{\eta}^{(t+1)})$. Then the maximization step updates $(\boldsymbol{\theta}^{(t)}, \boldsymbol{\eta}^{(t+1)})$ to $(\boldsymbol{\theta}^{(t+1)}, \boldsymbol{\eta}^{(t+1)})$. To prove B.1, it suffices to show that

$$\begin{aligned} L(\boldsymbol{\theta}^{(t)}, \boldsymbol{\eta}^{(t+1)}) &\geq L(\boldsymbol{\theta}^{(t)}, \boldsymbol{\eta}^{(t)}) \\ L(\boldsymbol{\theta}^{(t+1)}, \boldsymbol{\eta}^{(t+1)}) &\geq L(\boldsymbol{\theta}^{(t)}, \boldsymbol{\eta}^{(t+1)}). \end{aligned}$$

Note that

$$\begin{aligned}
L(\boldsymbol{\theta}^{(t)}, \boldsymbol{\eta}) &= \sum_{i=1}^{n} \log \left[\pi_{\eta_i}(\mathbf{w}_i, \boldsymbol{\gamma}^{(t)}) \sum_{j:c(j)=\eta_i} \pi_{j|c(j)}^{(t)} f(\mathbf{y}_i | \lambda_j^{(t)} \mathbf{1} + \mathbf{x}_i \boldsymbol{\beta}_{\eta_i}^{(t)}, (\sigma_j^2)^{(t)} \mathbf{I}) \right] \\
&\leq \sum_{i=1}^{n} \max_{\eta_i} \log \left[\pi_{\eta_i}(\mathbf{w}_i, \boldsymbol{\gamma}^{(t)}) \sum_{j:c(j)=\eta_i} \pi_{j|c(j)}^{(t)} f(\mathbf{y}_i | \lambda_j^{(t)} \mathbf{1} + \mathbf{x}_i \boldsymbol{\beta}_{\eta_i}^{(t)}, (\sigma_j^2)^{(t)} \mathbf{I}) \right] \\
&= \sum_{i=1}^{n} \log \left[\pi_{\eta_i^{(t+1)}}(\mathbf{w}_i, \boldsymbol{\gamma}^{(t)}) \sum_{j:c(j)=\eta_i^{(t+1)}} \pi_{j|c(j)}^{(t)} f(\mathbf{y}_i | \lambda_j^{(t)} \mathbf{1} + \mathbf{x}_i \boldsymbol{\beta}_{\eta_i^{(t+1)}}^{(t)}, (\sigma_j^2)^{(t)} \mathbf{I}) \right] \\
&= L(\boldsymbol{\theta}^{(t)}, \boldsymbol{\eta}^{(t+1)})
\end{aligned}$$

for all $\boldsymbol{\eta}$. This is true because in the algorithm,

$$\eta_i^{(t+1)} = \arg\max_{\eta_i} \pi_{\eta_i}(\mathbf{w}_i, \boldsymbol{\gamma}^{(t)}) \sum_{j:c(j)=\eta_i} \pi_{j|c(j)}^{(t)} f(\mathbf{y}_i \lambda_j^{(t)} \mathbf{1} + \mathbf{x}_i \boldsymbol{\beta}_{\eta_i}^{(t)}, (\sigma_j^2)^{(t)} \mathbf{I}).$$

Thus, we have $L(\boldsymbol{\theta}^{(t)}, \boldsymbol{\eta}^{(t+1)}) \geq L(\boldsymbol{\theta}^{(t)}, \boldsymbol{\eta}^{(t)})$. To prove the other part of the inequality, note that

$$\begin{aligned}
L(\boldsymbol{\theta}, \boldsymbol{\eta}^{(t+1)}) &= \sum_{i=1}^{n} \log \left[\pi_{\eta_i^{(t+1)}}(\mathbf{w}_i, \boldsymbol{\gamma}) \sum_{j:c(j)=\eta_i^{(t+1)}} \pi_{j|c(j)} f(\mathbf{y}_i | \lambda_j \mathbf{1} + \mathbf{x}_i \boldsymbol{\beta}_{\eta_i^{(t+1)}}, \sigma_j^2 \mathbf{I}) \right] \\
&= \sum_{k=1}^{K} \sum_{i=1}^{n} I(\eta_i^{(t+1)} = k) \log \left[\pi_k(\mathbf{w}_i, \boldsymbol{\gamma}) \sum_{j:c(j)=k} \pi_{j|c(j)} f(\mathbf{y}_i | \lambda_j \mathbf{1} + \mathbf{x}_i \boldsymbol{\beta}_k, \sigma_j^2 \mathbf{I}) \right] \\
&= \sum_{k=1}^{K} \sum_{i=1}^{n} I(\eta_i^{(t+1)} = k) \log \pi_k(\mathbf{w}_i, \boldsymbol{\gamma}) \\
&\quad + \sum_{k=1}^{K} \sum_{i=1}^{n} I(\eta_i^{(t+1)} = k) \log \sum_{j:c(j)=k} \pi_{j|c(j)} f(\mathbf{y}_i | \lambda_j \mathbf{1} + \mathbf{x}_i \boldsymbol{\beta}_k, \sigma_j^2 \mathbf{I}) \\
&\leq \max_{\boldsymbol{\gamma}} \sum_{k=1}^{K} \sum_{i=1}^{n} I(\eta_i^{(t+1)} = k) \log \pi_k(\mathbf{w}_i, \boldsymbol{\gamma}) \\
&\quad + \sum_{k=1}^{K} \max_{\pi_{j|c(j)}, \boldsymbol{\beta}_k, \lambda_j, \sigma_j \text{ for } j:c(j)=k} \sum_{i=1}^{n} I(\eta_i^{(t+1)} = k) \log \sum_{j:c(j)=k} \pi_{j|c(j)} f(\mathbf{y}_i | \lambda_j \mathbf{1} + \mathbf{x}_i \boldsymbol{\beta}_k, \sigma_j^2 \mathbf{I}) \\
&= L(\boldsymbol{\theta}^{(t+1)}, \boldsymbol{\eta}^{(t+1)})
\end{aligned}$$

for all $\boldsymbol{\theta}$. This suggests that $L(\boldsymbol{\theta}, \boldsymbol{\eta}^{(t+1)})$ can be maximized by first using numerical optimization to find $\boldsymbol{\gamma}$ that maximizes the first term, and then the rest of the parameters can be

optimized separately for the mixture model within each cluster using the samples assigned to the cluster by $\boldsymbol{\eta}^{(t+1)}$. The maximization step in the CEM computes $\boldsymbol{\theta}^{(t+1)}$ exactly in this fashion. Therefore, we have $L(\boldsymbol{\theta}^{(t)}, \boldsymbol{\eta}^{(t+1)}) \leq L(\boldsymbol{\theta}^{(t+1)}, \boldsymbol{\eta}^{(t+1)})$. □