University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2013 *Paper* 308

Vertically Shifted Mixture Models for Clustering Longitudinal Data by Shape

Brianna C. Heggeseth* Nicholas P. Jewell†

*University of California - Berkeley, brianna.c.heggeseth@williams.edu

†University of California - Berkeley, jewell@berkeley.edu

Vertically Shifted Mixture Models for Clustering Longitudinal Data by Shape

Brianna C. Heggeseth and Nicholas P. Jewell

Abstract

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 determine these distinct patterns is through cluster analysis, which seeks to separate objects (subjects, patients, observational units) into homogeneous groups. Many methods have been adapted for longitudinal data, but almost all of them fail to explicitly group trajectories according to distinct pattern shapes. To fulfill the need for clustering based explicitly on shape, we propose vertically shifting the data by subtracting the subject-specific mean directly removes the level prior to fitting a mixture modeling. This non-invertible transformation can result in singular covariance matrixes, which makes mixture model estimation difficult. Despite the challenges, this method outperforms existing clustering methods in a simulation study.

1 Introduction

A key advantage of a longitudinal study is its ability to measure individual change over time and to distinguish between "aging" affects and cohort variability. Typical longitudinal data analysis often involves modeling the conditional mean outcome as a function of time and explanatory variables while taking the inherent time dependence into account. This type of analysis provides a useful summary of relationships present within observed groups defined by categorical variables such as race/ethnicity or sex; however, if there are unmeasured subgroups in which the relationships differ, focusing only on the mean masks interesting and useful patterns. In these circumstances, it is more informative to partition the subjects into data-driven groups to estimate the relationships.

Cluster analysis methods have been adapted and developed specifically for approaching longitudinal data in this way, focusing on estimating trajectory groups in genomics and behavioral and cognitive development [18, 22, 29, 27]. Most of these methods involve fitting a finite mixture model to the outcome vectors. Other techniques cluster subjects via a partitioning method such as K-means [25, 19] or partitioning around mediods (PAM) [23] using on a dissimilarity measure developed specifically to take the time-ordered structure of longitudinal data into account.

These clustering methods are now being applied in other fields such as clinical medicine, public health, education, and economics. For example, researchers are attempting to discover distinct childhood growth patterns by clustering body mass index measurements over time with a view to determining which early life factors may contribute to following that pattern [31, 2]. The goal is to group children who grow similarly over time; however, none of the popular clustering methods explicitly group subjects based on the shape of the pattern while ignoring the trajectory level over time. Many authors have either been deceived or simply misunderstand this fact and interpret their clustering results as if the data were grouped by shape [31, 2, 30].

There has been little work and discussion on the best methods to utilize when shape, as distinct from level, is the feature of interest. The few proposals involve partitioning methods using dissimilarity measures based either on the derivative function or the Pearson correlation coefficient.

With longitudinal data, we do not directly observe the derivative function for each individual. Therefore, Möller-Levet et al. [28] and D'Urso [13] independently suggest estimating the derivative function via the difference quotient, calculating the slope of a linear interpolation between adjacent repeated measures. The dissimilarity between two individuals is measured as the squared Euclidean distance between the vectors of derivative estimates. As a result, individuals are compared on the shape of their underlying trend over time ignoring the original level of the data. Assume that the jth observed outcome for subject 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). This clustering

method essentially requires balanced data so the observations are fixed across individuals, $m_j = m$ and $t_{ij} = t_j$ for $j = 1, .., m$ and $i = 1, ..., n$. This ensures the estimated slopes are comparable between individuals. The forward difference quotient equals

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

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. Imagine two vectors that have the same underlying function but by chance, the noisy observations are on opposing sides of the function at every time point. Thus, the estimate derivatives are opposite signs for each interval and will be placed in separate groups despite the same underlying function. Now, 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 follow up, minimizes the variance but at the great expense of observing the rate of change during the follow up period. If time of observations are densely sampled, a functional approach smoothes out the noise using splines to estimate the function and then the derivative function. 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.

The Pearson correlation coefficient has been used to measure dissimilarity between two vectors based on the shape of the data [4, 14, 3]. In the context of functional clustering, Chiou and Li [3] suggest using the functional correlation as a similarity measure to cluster similar functions. Despite the wide use, there has been little to no discussion about how well the correlation does to discriminate between shapes. In the multivariate setting, a dissimilarity measure between two comparable vectors of equal length, $\mathbf{x} = (x_1, ..., x_m)$ and $\mathbf{y} = (y_1, ..., y_m)$, based on the Pearson correlation coefficient equals

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

where

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

and $\bar{x} = m^{-1}\sum_{j=1}^m x_j$ and $\bar{y} = m^{-1}\sum_{j=1}^m y_j$. Assume the data vectors are generated from the same model presented for the derivative-dissimilarity measure. Imagine the data is

observed with no noise, $\epsilon_{ij} = 0$ for $i = 1, ..., n$ and $j = 1, ..., m$. A vector with an underlying functional shape of $f(x)$ is perfectly positively correlated with a vector with functional shape $af(x)+b$ where $a, b \in \mathbb{R}$ and $a \geq 0$. As a result, two vectors that have the same shape but at different levels have a dissimilarity of zero since the correlation coefficient equals 1. However, two vectors that are multiples of each other are 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. Now, if the data is observed with error such that $\epsilon_{ij} \sim (0, \sigma_i^2)$, the dissimilarities usually increase especially when the noise is large relative to the functional shape.

Now, if two vectors are error-free observations of a constant, horizontal function, $f(x) = c$ where $c \in \mathbb{R}$, the correlation coefficient is undefined. When this type of data is observed with noise, the correlation is on average zero but variable when the number of observations is small and the magnitude of noise is large. In other words, the Pearson correlation coefficient cannot consistently detect that two horizontal vectors have the same shape. This methods fails to group by shape when the variance of the noise, σ_i^2, is large, the number of observations is small, and the data set includes horizontal patterns over time. These characteristics do not get discussed in the literature, but they have huge ramifications for clustering longitudinal data.

Both of these proposals implicitly remove the level while clustering, but they fail to detect shape when data is observed with high noise. Additionally, these dissimilarity measures based on comparing vectors are not conducive to use on irregularly sampled longitudinal data with inconsistent observation times. Also, the partitioning methods do not provide a way to estimate the impact of baseline factors on group membership while taking into account membership uncertainty.

In this paper, we propose vertically shifting each subject by subtracting its mean before fitting a multivariate mixture model in order to cluster longitudinal data by the shape over time. This method provides a probability framework and works with irregularly sampled longitudinal data. In Section 2, we discuss related work and then we present the model specification and provide details for the mean and covariance structure in Section 3. Implementation including parameter estimation is discussed in Section 4. We discuss challenges when modeling the transformed data in Section 5. Lastly, a simulation study demonstrates the merits of the proposed method in comparison to standard clustering methods and those that explicitly attempt to group based on shape in Section 6.

2 Related Work

A finite mixture model is a standard method for clustering multivariate data [16] and has been used for longitudinal applications [29, 22]. See [26] 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.

We 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 of the variable measures and dividing by the standard deviation so that each standardized variable is in comparable units and equally contributes 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 since the relationship between measurements within individuals is of interest. Therefore, any transformation performed should only be additive in nature to preserve the original 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.

A version of this idea has been implemented in the functional data analysis literature. For processes in a Hilbert space of square integrable functions with respect to the Lebesgue measure, dt, on the interval $\mathcal{T} = [0, T]$, Chiou and Li [3] propose using a mixture model and the Karhunen-Loève expansion for centered stochastic processes within their correlation-based clustering algorithm. The integral of the random function over interval $\mathcal{T}$ divided by T, the length of the interval, is subtracted to center the process; the resulting process integrates to zero. The integral of the process is the functional analogue to a mean vector in vector space; similarly, the resulting vector has mean zero after subtraction.

Although centering a process and a shifting a vector stems from the same idea, there are distinct consequences of subtracting the estimated level of a noisy curve observed at a finite number of points that do not arise when centering a smooth function. The term centering is used in the stochastic processes literature, but we use the term vertically shifting to refer to the procedure of subtracting the mean since it graphically describes the transformation of the noisy longitudinal data.

3 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}. We 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

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

where F is a probability distribution, $\mathbf{1}_{m_i}$ is a m-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 being in a particular shape group could depend on baseline covariates in the vector $\mathbf{w}_i$. 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

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

Applying 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, 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, we 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 model above would exactly follow 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 component given baseline covariates, $\mathbf{w}$. The full vector of parameters for the model is $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_K)$. 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.

3.1 Mean Structure

To focus 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 lower order polynomial basis such as $\{1, t, t^2\}$. However, this basis cannot capture complex shapes with drastic, local changes.

To allow for flexibility in the functional shape, the observation time interval, $[a, b]$, is broken up into smaller interval using L knots, $a < \tau_1 < \cdots < \tau_L < b$, and 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 [8, 34, 6, 9]. 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 [10, 5]. 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 $\boldsymbol{\mu}_k(t) = \sum_{j=1}^{L+p} \beta_j B_{j,p}(t)$. In the multivariate form, the mean vector at observation times $\mathbf{t}_i$ equals $\mathbf{x}_i\boldsymbol{\beta}_k$ where $\boldsymbol{\beta}_k = (\beta_{k,1}, ..., \beta_{k,L+p})$.

3.2 Covariance Structure

There are many potential assumptions to be made about the covariance of the random deviations from the mean. Here, we allow the covariances to differ between clusters. Since it is common for longitudinal data to have sparse, irregular time sampling, the covariance matrix needs structure to allow for parameter estimation as described by Jennrich and Schluchter [21] in their seminal paper. A common parameterization is conditional independence with constant variance where $\boldsymbol{\Sigma}_k = \sigma_k^2 \mathbf{I}_{m_i}$. 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 $\boldsymbol{\Sigma}_k = \sigma_k^2(\rho_k \mathbf{1}_{m_i}\mathbf{1}_{m_i}^T + (1-\rho_k)\mathbf{I}_{m_i})$ 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 between the two is the exponential correlation structure in which the dependence decays as the time between observations increases such that $[\boldsymbol{\Sigma}_k]_{jl} = \sigma_k^2 \exp(-|t_{ij} - t_{il}|/r_k)$ where $r_k > 0$ is the range of the dependence. If the range, r_k, is small, the correlation decays quickly, but if the 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 $[\boldsymbol{\Sigma}_k]_{jl} = \sigma^2 \rho_k^{|t_{ij} - t_{il}|}$ 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 if 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.

All of the covariance structures mentioned above are associated with weakly stationary processes with constant variance and correlation dependent only on the time lag between observations. If the variance or correlation function is non-constant but varying continuously, it could be potentially modeled as a function, but estimation is more difficult.

It is important to model the covariance structure correctly as misspecification can highly impact mixture model results in terms of parameter estimates and the final clustering if the groups are not well separated [20]. Transforming the data brings individuals with similar shapes closer but also brings others closer as well. In general, this decreases the separation between groups, which may force us to accurately model the correlation.

4 Implementation

Given a collection of independent observed outcome vectors from n individuals, $\mathbf{y}_1, ..., \mathbf{y}_n$. The first step is to calculate the mean for each subject, $\bar{y}_i$, and subtract the subject-specific mean from the observed outcome vector for $i = 1, ..., n$. This transformation results in independent vertically shifted vectors, $\mathbf{y}_1^*, ..., \mathbf{y}_n^*$.

Then, the order of the spline and the number and location of internal knots for the mean structure is chosen. The B-spline basis should be kept constant for all shape groups, so the simplest way to select the number of knots is through visual inspection of the full data set. If the most complex shape patterns is a lower order polynomial, no internal knots are necessary. However, if the most complex function has local activity, adding knots and increasing the order of the spline functions flexibly accommodates the twists and turns of the mean patterns. In choosing both the order of the polynomials and the number of knots, it is important to balance the number of mean parameters with the sample size. Every unit increase in the order or in the number of knots increases the number of parameters by K, the number of groups. In terms of location of the knots, one suggestion is to place knots at sample quantiles based on the sampling times of all the observations [33]. However, this strategy may not work well if the median time is not the point of deviation from a regular polynomial. If possible, it is best to place knots at local maxima, minima, and inflection points of the overall trends as to accommodate the differences from a polynomial function [15]. Once these are decided, the design matrices, $\mathbf{x}_i$, are calculated using widely available B-spline algorithms for $i = 1, ..., n$.

Parameters are estimated using maximum likelihood estimation via the EM algorithm. 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 3, 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 expectation-maximization (EM) algorithm [11]. This algorithm is applied in the framework where given $(\mathbf{t}_i, \mathbf{w}_i)$ each $\mathbf{y}_i^*$ is assumed to have stemmed from one of the mixture components and the indicator denoting its originating component is missing. The complete-data log likelihood is based on these indicator variables as well as the observed data $\{(\mathbf{y}_i^*, \mathbf{t}_i, \mathbf{w}_i)\}$. The expectation step (E-step) involves replacing the indicators by current values of their conditional expectation, which is the posterior probability of component membership, written as

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

for $i = 1, ..., n$ and $k = 1, ..., K$ using current estimates of the parameters. In the maximization step (M-step), the parameter estimates for the prior probabilities, linear mean, and covariance matrices are updated by maximizing the complete-data log likelihood using the posterior probabilities from the E-step in place of the indicator variables using numerical optimization. The E- and M-steps are alternated repeatedly until convergence. The EM algorithm guarantees convergence to a local maximum; global convergence may be attained through initializing the algorithm by randomly assigning individuals to initial components, running the algorithm multiple times and using the estimates associated with the highest log likelihood. Besides the parameter estimates, the algorithm returns the posterior probability estimates of component membership. These probabilities can be used to partition individuals into distinct clusters by selecting the cluster with the maximum posterior probability. However, unlike K-means, the posterior probability provides some measure of uncertainty in the group membership.

Estimation requires the number of clusters, K, to be known. In practice, this is not the case and 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 article, the criteria is the Bayesian Information Criterion (BIC) [35], 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. The BIC has been widely used for model selection with mixture models since the paper by Roeder and Wasserman [32]. In particular, the criteria has been to select the number of clusters [7, 17] with good results in practice. For regular models, the BIC was derived as an approximation to twice the log integrated likelihood using the Laplace method [36], but the necessary regularity conditions do not hold for mixture models in general [1]. However, Roeder and Wasserman [32] showed that the BIC leads to a consistent estimator of the mixture density, and Keribin [24] showed that the BIC is consistent for choosing the number of components in a mixture model.

5 Modeling Challenges

There are issues of identifiability with Gaussian mixture models that can be mitigated through some minor constraints [26]. In this section, we discuss some unique consequences of vertically shifting the data on the model and estimation.

5.1 Covariance of transformed data vectors

Let $\mathbf{Y} = (Y_1, ..., Y_m)$ be a random vector observed at times $\mathbf{t} = (t_1, ..., t_m)$ such that $\mathbf{Y} = \lambda \mathbf{1}_m + \boldsymbol{\mu} + \boldsymbol{\epsilon}$ such that $\lambda \sim F$, $\boldsymbol{\mu}$ is a vector of evaluations of a known deterministic 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.

Subtracting the mean of the elements of the vector by applying the transformation matrix $\mathbf{A} = \mathbf{I}_m - m^{-1}\mathbf{1}_m\mathbf{1}_m^T$ changes the correlation structure of the data. The covariance of the transformed random vector, $\mathbf{Y}^*$, equals

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

by the properties of covariance. One important property of this transformation is that it is non-invertible; once the mean is subtracted from the data, the original data cannot be recovered. This has tremendous consequences on the covariance matrix. Since $det(\mathbf{A}) = 0$, the determinant of $Cov(\mathbf{Y}^*)$ is always zero and the covariance matrix is singular. Intuitively, the matrix has to be singular because the removal of the mean constrains the data to sum to zero. In other words, the transformation projects the data onto the $(m-1)$-dimensional subspace orthogonal to the nullspace $\mathbf{1}$. However, this presents challenges when trying to model the covariance of the transformed data.

Rather than focusing on the covariance of the transformed vector, we work with the covariance of the transformed vectors after removing the relationship with time. Therefore, the covariance of the deviations of the transformed data from the known mean shape is

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

If the observation times are fixed, then $\boldsymbol{\mu}$ is not random and the covariance simplifies to $\mathbf{A}Cov(\boldsymbol{\epsilon})\mathbf{A}^T$. However, if the observation times are random, then $\boldsymbol{\mu}$ is a random vector and contributes variability. To better understand how best to model the transformed data, we explore this covariance matrix 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.

5.1.1 Fixed observation times

For the moment, assume that the observation times, $\mathbf{t}$, are fixed. Then

$$\begin{aligned} Cov(\mathbf{Y}^* - \boldsymbol{\mu}) &= \mathbf{A} Cov(\boldsymbol{\epsilon}) \mathbf{A}^T \\ &= \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T \end{aligned}$$

where $\boldsymbol{\Sigma}$ is the covariance of the original random errors. If the variance is constant over time, $\mathbf{V} = \sigma^2 I$, and the elements of the original vector are independent, $\mathbf{R}_i(\rho) = \mathbf{I}$, then the covariance of the deviations of the transformed data from the known mean shape is

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

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 m is large, the resulting correlation structure is approximately independent with variance σ^2.

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 of the deviations of the transformed data from the known mean shape is

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

where $a = \frac{-1}{m-1}$. This transformation maintains the exchangeable structure but with negative correlation on the off diagonal and decreased variance of $\sigma^2(1-\rho)\left(\frac{m-1}{m}\right)$. Again, if the number of observed data points is large, then the structure is approximately independent with variance $\sigma^2(1-\rho)$.

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,

$$\begin{aligned} Cov(\mathbf{Y}^* - \boldsymbol{\mu}) &= \sigma^2 \mathbf{A}\mathbf{R}(\rho)\mathbf{A}^T \\ &= \sigma^2 \left[\mathbf{R}(\rho) - m^{-1}\mathbf{1}\mathbf{1}^T\mathbf{R}(\rho) - m^{-1}\mathbf{R}(\rho)\mathbf{1}\mathbf{1}^T + m^{-2}\mathbf{1}\mathbf{1}^T\mathbf{R}(\rho)\mathbf{1}\mathbf{1}^T \right] \\ &= \sigma^2 \left[\mathbf{R}(\rho) - \text{ column mean vector } - \text{ row mean vector } + \text{ overall mean} \right]. \end{aligned}$$

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. For example, if $\sigma^2 = 1$, $\rho = 2$ and $\mathbf{t} = (1, 2, 3, 4)$, then the covariance matrix of the deviations of the transformed data from the known mean shape is

$$Cov(\mathbf{Y}^* - \boldsymbol{\mu}) = \begin{bmatrix} 0.499 & 0.009 & -0.229 & -0.278 \\ 0.009 & 0.307 & -0.087 & -0.229 \\ -0.229 & -0.087 & 0.307 & 0.009 \\ -0.278 & -0.229 & 0.009 & 0.499 \end{bmatrix}.$$

The variance and covariance changes over time with the covariance becoming negative as the time lag increases. If the number of observation times increase such that the observation period expands, the covariance of the transformed vector becomes close to the original covariance as column, row, and overall means decrease to zero when the number of pairs of measurements with large time lags increases. However, if the observation period remains fixed as the number of observations increases, the covariance after transformation continues to be non-stationary and has negative correlations.

We have calculated the covariance of the transformed random vector under three common covariance structures for the original data assume 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.

5.1.2 Random observation times

In practice, individuals in a longitudinal study are not typically observed at exactly the same times but rather at random times. When the times are random, the vector $\boldsymbol{\mu}$ is random because the elements are evaluations of the deterministic 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, $\boldsymbol{\Sigma}$, does not depend on time such as in the case of conditional independence or exchangeable, then the covariance simplifies to

$$\begin{aligned} Cov(\mathbf{Y}^* - \boldsymbol{\mu}) &= Cov((\mathbf{A} - \mathbf{I})\boldsymbol{\mu} + \mathbf{A}\boldsymbol{\epsilon}) \\ &= (\mathbf{A} - \mathbf{I})Cov(\boldsymbol{\mu})(\mathbf{A} - \mathbf{I})^T + \mathbf{A}Cov(\boldsymbol{\epsilon})\mathbf{A}^T \\ &= m^{-2}\mathbf{1}\mathbf{1}^T Cov(\boldsymbol{\mu})\mathbf{1}\mathbf{1}^T + \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T. \end{aligned}$$

Let the random times, $t_1, ..., t_m$ be independent with potentially different expected values and variances. The covariance of $\boldsymbol{\mu}$ equals a diagonal matrix with the jth diagonal entry

approximately equal to $Var(t_j)[\mu'(E(t_j))]^2$ by the delta method. Then, the covariance matrix of $\mathbf{Y}^*$ is the sum of two non-invertible matrices,

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

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

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

where the mean elements $\mu_{ij} = \mu(t_{ij})$. We repeat the simulation under different assumptions for the mean function, $\mu(t)$, and standard deviations of the observation times, σ_τ. Figure 1 and Figure 2 show the estimated autocorrelation functions of the deviations of the transformed data from the mean when $\mathbf{R}_i(\rho)$ is an exchangeable correlation matrix with correlation parameter $\rho = 0.5$ and when $\mathbf{R}_i(\rho)$ is an exponential correlation matrix with range parameter $\rho = 2$, respectively, under varying conditions for observations times and shape functions.

As the variance of observation times and the magnitude of the derivative mean function increases, the estimated correlation between deviations from the mean becomes more 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 [12], which may be appropriately flexible to approximate the covariance of the deviations from the mean of the transformed data.

5.1.3 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 covariance of transformed vector. As we saw above, the length of the vector, m, impacts the covariance of the transformed vector. Suppose the outcome vectors for a sample of individuals has 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.

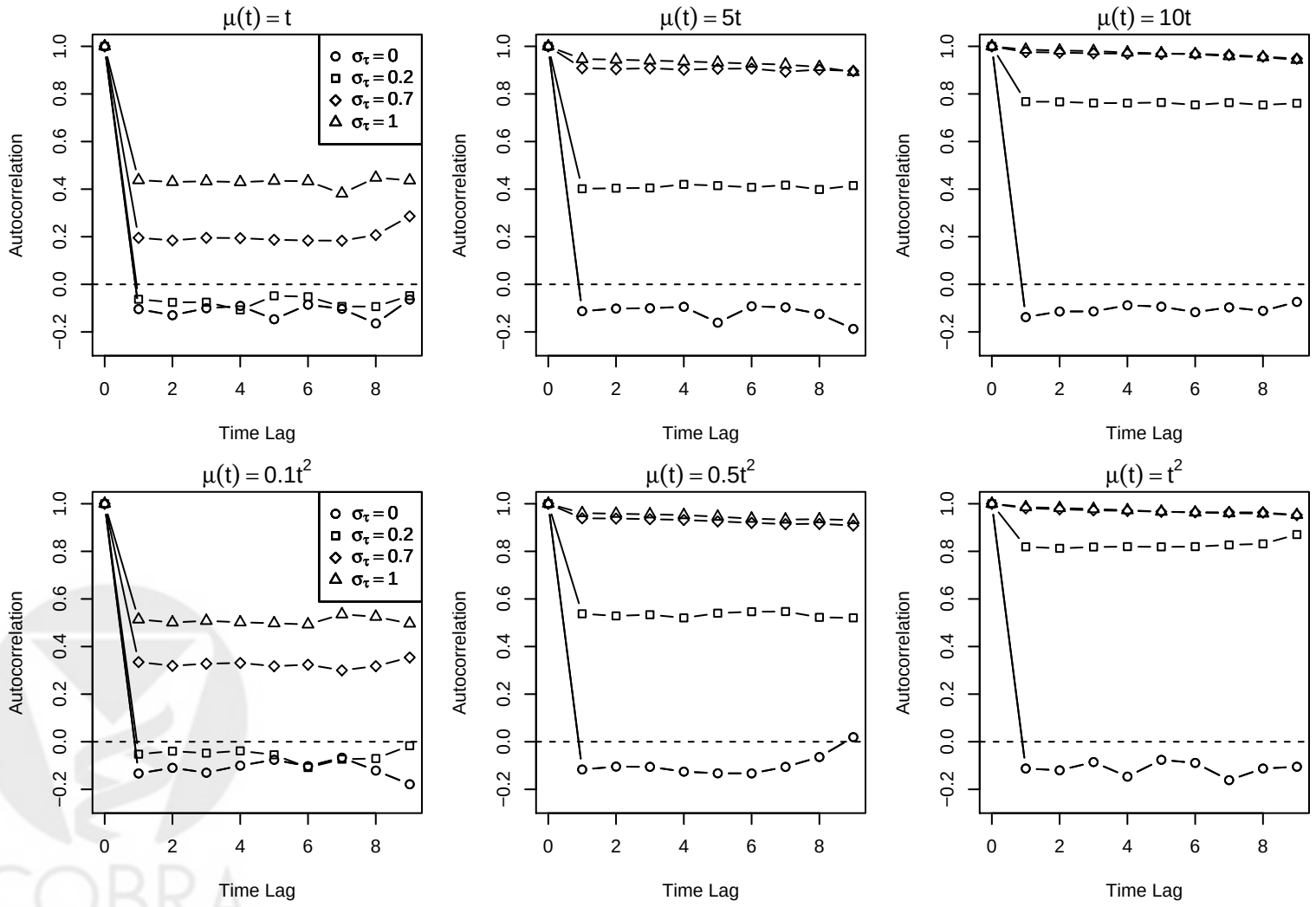

Figure 1: Estimated autocorrelation functions of the deviations from the mean from data generated with an exchangeable correlation error structure and random observation times under different mean functions, $\mu(t)$, and standard deviations of the random time perturbations, σ_τ.

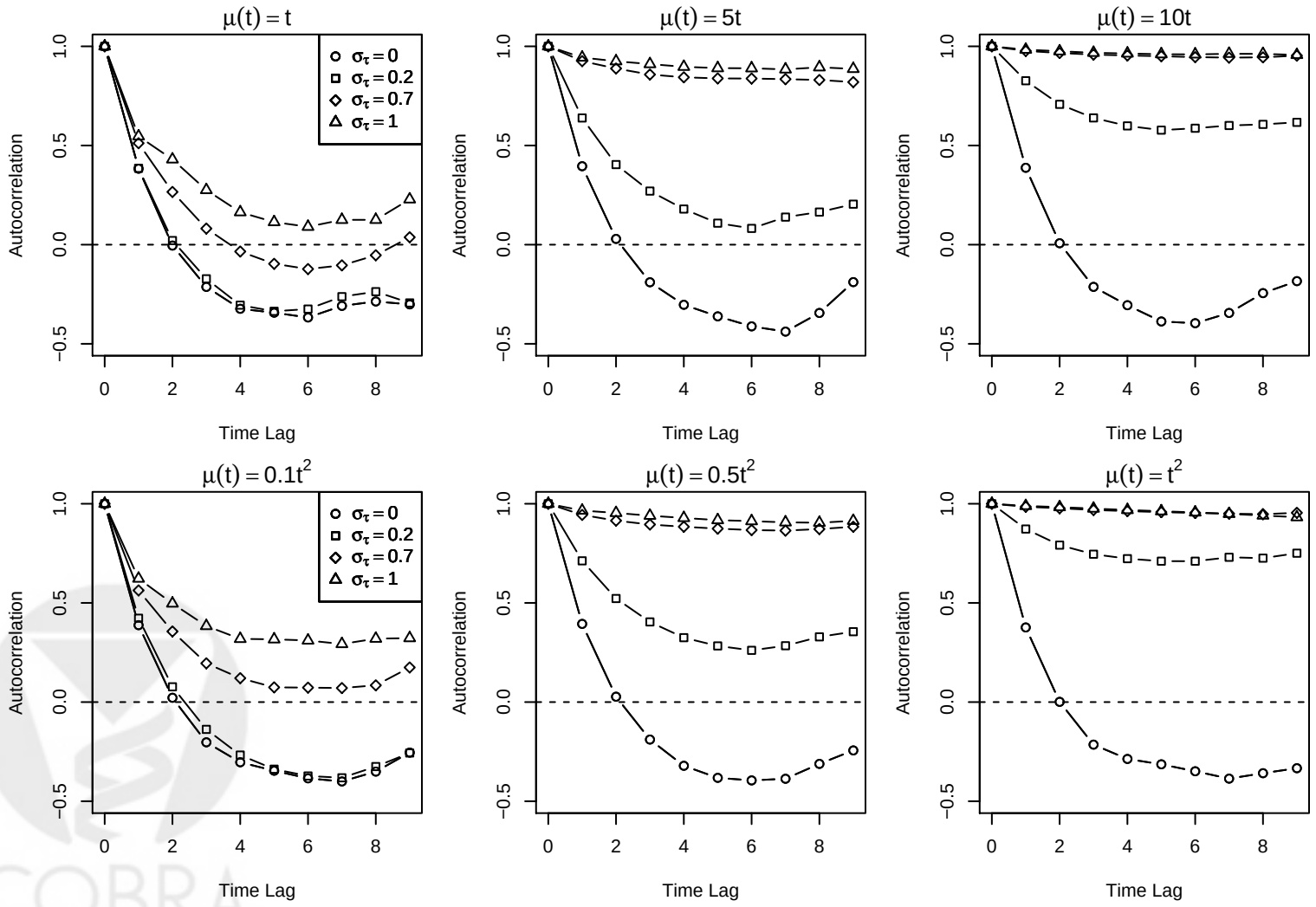

Figure 2: Estimated autocorrelation functions of the deviations from the mean from data generated with an exponential correlation error structure and random observation times under different mean functions, $\mu(t)$, and standard deviations of the random time perturbations, σ_τ.

6 Simulation

To compare the performance of the proposed clustering method with competing methods, we completed a simulation study with three trajectory shapes and three levels in the population. We restricted some shapes to particular levels to induce a relationship between level and shape. The five mean functions for generating the data with shapes at different levels are below:

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

Individuals follow these patterns with varying probabilities that depend on two 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 the independent simulated tosses of a fair coin such that $P(w_1 = 1) = P(w_1 = 0) = 0.50$ and $P(w_2 = 1) = P(w_2 = 0) = 0.50$.

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

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

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 coin toss, each shape group has about an equal frequency, marginally.

The second factor impacts the level, but I placed restrictions when creating the mean functions; therefore, level and shape are not independent of each other. Let L be a categorial random variable that indicates level group. $L = 1, 2, 3$ refers to the low, middle, and high group, respectively. Conditional on the shape group, all of the horizontal lines are in the middle level. For those in either the negative or positive slope groups, the chance of the high or low level equals

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

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

To summarize, individual trajectories are generated by first simulating two coin tosses to determine values for the two baseline binary factors. Then conditional on the factors, shape and level groups are randomly assigned by plugging the factors into the generalized logit functions above and drawing from multinomial distributions. Then, 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. Random noise is added to induce variability.

The random noise is made up of two components: individual-specific level perturbation and time-specific Gaussian measurement error. For individual i $(i = 1, ..., n)$ at the jth observation time $(j = 1, .., 5)$ with the lth mean function, the observed outcome equals

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

where σ_ϵ is the standard deviation of the measurement error and σ_λ is the standard deviation of the level perturbation. To create conditions with differing amount of overlap between groups, I let $\sigma_\lambda = 2, 3$ such that there are 4 or 6 standard deviation between the mean functions with the same shape. The magnitude of measurement error influences the signal to noise ratio and I let $\sigma_\epsilon = 0.5, 2$ to create two extreme conditions. There are four possible combinations of these two properties, representing the four conditions of the data-generating process in the simulation study.

For each condition, we generate a data set of $n = 500$ individuals using the process described above and apply five clustering methods: independence Gaussian mixture model, K-means on the difference quotients, PAM with a correlation-based dissimilarity measure, vertically shifting mixture models with independence, and vertically shifting mixture models with exponential correlation. For each method, we calculate the optimal K using the silhouette measure [23] for the partition methods and the BIC [35] for the models and estimate the misclassification rate when $K = 3$. The misclassification rate detects whether the method discovers the underlying shape structure. This is repeated $B = 500$ times such that we get 500 unique data sets under each condition on which we can apply each methods and summarize the results.

6.1 Results

Table 1 summarizes the simulations in terms of the number of groups and average misclassification rate. It is clear from this table that the standard Gaussian mixture model assuming independence does not select three groups as the optimal number of groups. The BIC with the independent mixture model consistently chooses five groups, the maximum we allow in the simulation, under all conditions. This method also does not perform well when forced to have $K = 3$. Only about 50% of the data is correctly specified in terms of the generating shape groups.

Of the established methods that are intended to group on shape, K-means on difference quotients selects the correct number of groups if the magnitude of the measurement error is small (Table 1). 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 25-27% misclassification, but the it does not consistently choose three groups. It only selects two and five groups.

Lastly, the method that prevailed amongst the competition is the vertically shifted mixture models. For every condition, the method chose three groups as the optimal number

99% of the time and when forced to $K = 3$, 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) of the individuals in terms of shape. The method worked well under both assumptions of independence and exponential correlation even though we know the true correlation structure is exchangeable with correlation -0.25. Therefore, in this case, the shapes are distinct enough that either correlation assumption worked well.

7 Conclusion

This method to cluster longitudinal data by the shape of the trajectory over time directly removes the level by subtraction individual-specific means prior to modeling. In contrast to partition methods based on observed vectors, this approach not only allows irregularly sampled data but may perform better when observation times are random, not fixed. The mixture model provides a probability framework to explicitly model the variability around an underlying smooth function. Therefore, the method does well even when the measurement error is large.

The standard method of mixture models applied to the original data values does not directly cluster based on shape. The methods created to focus on shape fail under fairly common circumstances. K-means applied to difference quotients works well when there is little measurement error, but fails to create distinct shape groups when there is moderate error. The correlation-based dissimilarity measure fails to group horizontal trajectories as similar. Therefore, it is hard to tell whether the high misclassification rate is due primarily to measurement error or the horizontal issues. Since it is common to have horizontal trajectories as well as substantial measurement error in longitudinal data, both of these methods are not recommended for wide use.

While the proposed method drastically out performs the competition, there are two main issues. First, subtracting the observed mean impacts the covariance in a way that makes it harder to model with familiar correlation structures. However, if the shapes drastically differ, using the the simple independence correlation structure may work well enough to detect the shape groups. Second, care needs to be taken when there is sparse and irregularly sampling.

References

[1] M. Aitkin and D. B. Rubin. "Estimation and Hypothesis Testing in Finite Mixture Models". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 47.1 (1985), pp. 67–75.

[2] M. A. Carter et al. "Trajectories of childhood weight gain: the relative importance of local environment versus individual social and early life factors". In: *PloS one* 7.10 (2012), e47065.

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

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

[5] M. G. Cox. "The Numerical Evaluation of B-Splines". In: *IMA Journal of Applied Mathematics* 10.2 (1972), pp. 134–149. DOI: 10.1093/imamat/10.2.134.

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

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

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

[9] C. De Boor. "Splines as linear combinations of B-splines. A survey." In: *Approximation Theory II*. New York: Academic Press, 1976, pp. 1–47.

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

[11] A. P. Dempster, N. M. Laird, and D. B. Rubin. "Maximum Likelihood from Incomplete Data via the EM Algorithm". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 39.1 (1977), pp. 1–38.

[12] P. Diggle et al. *Analysis of longitudinal data*. 2nd ed. New York: Oxford University Press, 2002.

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

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

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

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

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

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

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

[20] 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". In: *Statistics in Medicine to appear* (2013). DOI: 10.1002/sim.5729.

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

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

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

[24] C. Keribin. "Consistent Estimation of the Order of Mixture Models". In: *Sankhya: The Indian Journal of Statistics, Series A (1961-2002)* 62.1 (2000), pp. 49–66.

[25] 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.

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

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

[28] 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.

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

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

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

[32] K. Roeder and L. Wasserman. "Practical Bayesian Density Estimation Using Mixtures of Normals". In: *Journal of the American Statistical Association* 92.439 (1997), pp. 894–902. DOI: 10.1080/01621459.1997.10474044.

[33] D. Ruppert. "Selecting the Number of Knots for Penalized Splines". In: *Journal of Computational and Graphical Statistics* 11.4 (2002), pp. 735–757. DOI: 10.1198/106186002853.

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

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

[36] L. Tierney and J. B. Kadane. "Accurate approximations for posterior moments and marginal densities". In: *Journal of the American Statistical Association* 81.393 (1986), pp. 82–86.

σ_ϵ	σ_λ	$K=2$	$K=3$	$K=4$	$K=5$	MR
Independent Mixture						
0.50	2.00	0.00	0.00	10.00	490.00	0.41
2.00	2.00	0.00	0.00	24.00	476.00	0.39
0.50	3.00	0.00	0.00	1.00	499.00	0.45
2.00	3.00	0.00	0.00	5.00	495.00	0.45
K-means on Difference Quotients						
0.50	2.00	0.00	500.00	0.00	0.00	0.00
2.00	2.00	483.00	17.00	0.00	0.00	0.38
0.50	3.00	0.00	500.00	0.00	0.00	0.00
2.00	3.00	483.00	17.00	0.00	0.00	0.38
Correlation-based PAM						
0.50	2.00	403.00	0.00	0.00	97.00	0.25
2.00	2.00	500.00	0.00	0.00	0.00	0.27
0.50	3.00	403.00	0.00	0.00	97.00	0.25
2.00	3.00	500.00	0.00	0.00	0.00	0.27
Vertically Shifted Independent Mixture						
0.50	2.00	0.00	499.00	0.00	1.00	0.00
2.00	2.00	0.00	498.00	2.00	0.00	0.05
0.50	3.00	0.00	499.00	0.00	1.00	0.00
2.00	3.00	0.00	498.00	2.00	0.00	0.05
Vertically Shifted Exponential Mixture						
0.50	2.00	0.00	500.00	0.00	0.00	0.00
2.00	2.00	0.00	499.00	1.00	0.00	0.05
0.50	3.00	0.00	500.00	0.00	0.00	0.00
2.00	3.00	0.00	499.00	1.00	0.00	0.05

Table 1: Frequency table of the number of groups chosen and average misclassification rate (MR) ($K = 3$) for 500 replications of clustering methods applied to data generated under different values for σ_ϵ and σ_λ.