Methods for Analyzing Multivariate Longitudinal Data

Xin Lian

August 2022

Abstract

It is common to repeatedly measure the same subjects in clinical and social studies. In longitudinal studies, researchers are mostly interested in the evolution of observations across time on subject and population level. Such longitudinal data with inherent covariance within subjects across time violates assumptions of conventional linear models and requires specific methods. Sometimes the joint evolution between observations may also be of interest when more than one predictor or response is measured, which adds to the complexity of modeling and computation. Three commonly used families of methods will be introduced in this paper. To be specific, generalized linear mixed models for subject-specific analysis, marginal model for population-average analysis, and functional data modeling for stochastic process analysis will be emphasized. Univariate modeling and extensions to multivariate analysis of each method are discussed. A multivariate dataset from a laboratory study of patients with primary biliary cholangitis is used to show the application of a few of the methods presented.

1 Introduction

Longitudinal data arises from experiments and observations where attributes of the same subjects are measured repeatedly. The repeated measurement of responses are seen in clinical trials where the same set of participants are examined over a defined time period. For example, in studying the effect of treatments on HIV patients' immune suppression, CD4 cell count of each patient is measured daily for 40 weeks. Aside from the common interest of studying evolution

of responses and their association with covariates over time, longitudinal data analysis is different from conventional general linear regression and time series modeling because the correlation between responses for the same subject over time (serial correlation) is often significant and is of scientific interest.

There are a great number of resources on studying longitudinal data with a single univariate response. Extensions of conventional general linear regression models can be applied to describe subject-specific trajectories over time and offer inferences on both variances within or between subjects, for example in Cnaan (1997) [3] For large scale studies where the population mean structure is of greater interest, marginal models can be applied to lessen the computation burden. Examples with mixed-type predictors are illustrated in Albert (1999) [4] There are also responses that do not approximately follow an obvious probabilistic distribution, in which case the association between the predictors and the longitudinal response may be non-linear. Sometimes non-linear models are used given a known mechanistic structure of the longitudinal predictors and responses. Davidian (2003) [5] reviewed non-linear mixed models. More flexible but more complicated semi-parametric models are also used for multilevel longitudinal data. Model selection procedures of such models are discussed in Fan (2004) [6]. Another family of flexible non-parametric methods called functional data analysis (FDA) treats the data as functional data following a latent stochastic process. Rice (2004) [8] did a general review of developments in FDA.

A common scenario in clinical studies is simultaneous observation of multiple responses of the same subjects. For example, in examining the effect of antiretroviral treatment (ART) to patients with HIV, aside from monitoring the CD4 cell count reflecting the immune function of patients, researchers are also interested in monitoring HIV viral load as a direct indication of the effect of treatment. Interest in studying such data often lies in testing effects of covariates on different responses which requires fitting the distinct responses simultaneously. Sometimes, interest lies in detecting dependency or correlation between multiple longitudinal predictors or identifying homogeneous groups of subjects through clustering.

Despite the large variety of available methods and R packages for single response longitudinal data, applications of multivariate longitudinal data analysis are much more difficult to find. The extension from univariate model to multivariate model is not trivial because of the additional covariance required due to accounting for multiple responses at each time point from the same unit adds to the complexity of modeling and computation. As discussed in Verbeke (2014)

[10], when interest lies in the simultaneous evolution for multiple longitudinal responses, marginal studies on each univariate response often fail to give a sufficient inference to such covariance while direct joint analysis of all responses may be computationally infeasible.

This paper will introduce some of the most common methods that are applied to study univariate longitudinal data and their extensions to multivariate longitudinal data, often through dimension reduction. These methods and corresponding notations are introduced in Section 2. Applications of some methods discussed in Section 2 will be showcased on a longitudinal study of primary biliary cholangitis in Section 3. Section 4 briefly discusses some other classes of models that build upon the models introduced in earlier sections and suggests further developments.

2 Methods

There is a large variety of methods that are used to study longitudinal data. We are not going to exhaust the whole list, but will introduce some of the methods that can be applied to most general cases. In the following subsections, we first introduce generalized linear mixed models for interest in subject-specific study. Next, we introduce population-average models with a specific mean structure and covariance estimation approach called generalized estimating equations. Finally, a different class of models that treat the typical limited (i.e., sparse) observations as samples from infinite-dimension functional data and are based on non-parametric curve methods will also be introduced.

To start with, we introduce some notations for longitudinal data that are commonly used in different methods.

The methods discussed below are capable of dealing with unbalanced data with missing predictors or responses. For simplicity, we assume in the multivariate case that the predictors and responses of a given subject are all measured at each time of measurement but these time points may differ between subjects. Also, note that handling missing data, including imputation methods, is not the goal of this paper and thus is not discussed further.

In the sections below, for each subject $i \in \{1, 2, \ldots, n\}$ that is repeatedly measured, their longitudinal data with a single response generally follows the form:$(\boldsymbol{y}_i, \boldsymbol{X}_i)$ where

- $\boldsymbol{y}_i = [y_{i1}, \ldots, y_{im_i}]^T$ comprises all the responses across time for unit i with

dimension $(m_i \times 1)$

- Each y_{ij} comprises the response of unit i at time point j for $j \in \{1, 2, \ldots, m_i\}$ where m_i is the number of repeated measurements of subject i.
- $\boldsymbol{X}_i = [\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{im_i}]^T$ comprises all the P predictors of unit i across time with dimension $(m_i \times P)$
- Each $\boldsymbol{x}_{ij} = [x_{ij1}, \ldots, x_{ijP}]$ contains the predictors of unit i at time point j with dimension $(1 \times P)$ where x_{ijp} is the pth predictor of P predictors of unit i at time j for $p \in \{1, 2, \ldots, P\}$

In most models, we have regression coefficients $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_P]^T$ with dimension of $((P+1) \times 1)$ for the intercept and P predictors of $\boldsymbol{X}$, respectively.

For multiple response cases, an additional subscript k is introduced for each $\boldsymbol{y}_{ij}$ indicating the kth response; specifically, $\boldsymbol{y}_{ik} = [y_{i1k}, ..., y_{im_ik}]^T$ is the $(m_i \times 1)$ vector of observations for response k, $k \in \{1, 2, \ldots, K\}$. A set of regression coefficients $\boldsymbol{\beta}_k = [\beta_{k0}, \beta_{k1}, \ldots, \beta_{kP}]^T$ is specified for each response.

2.1 Subject-Specific Analysis using Generalized Linear Mixed Model

In many cases of clinical studies, we are interested in examining the trajectory of evolution of response with attributes of each subject over time. With assumptions of independence between subjects, the subject-specific trajectories can be described using a mixed-effects model. The mean structure is represented by shared "fixed effects" of covariates and intercept on response, while the subject-specific variation from the mean structure are modeled using iid zero-mean random variables as "random effects" of covariates, often time, and the intercept. Under single response setting popularized by Laird (1982) [11], the mixed-effects model can be parsimoniously constructed based on linear model with the form:

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z_i}\boldsymbol{b_i} + \boldsymbol{\varepsilon_i} \ ,$$

where $\boldsymbol{b_i}$ is called "random effects" and typically will have dimension no more than, and typically less than, $((P+1) \times 1)$ depending on design of the model. For random effects in this section, $\boldsymbol{b_i}$ has a dimension $(Q \times 1)$. In their most simple form, $\boldsymbol{b_i}$'s can be iid random variables that follow $N_Q(0, \boldsymbol{D})$ where $\boldsymbol{D}$ is a unknown covariance matrix with dimension of $(Q \times Q)$. $\boldsymbol{\varepsilon_i}$ is commonly

defined as iid random variables following $N(0, \boldsymbol{\Sigma_i})$, where $\boldsymbol{\Sigma_i} = \sigma^2 \boldsymbol{I_{m_i}}$, i.e., the conditional independence assumption, is often sufficient in this formation of a linear mixed model. A superior but more complicated setting of the covariance structure of random effects and error terms is discussed in Verbeke and Lesaffre (1997).[12] These effects represent the variations brought by different values of covariates on intercept and growth rate over time for each subjects' trajectory.

As in a conventional linear model, often the main interest is in the "fixed effects" $\boldsymbol{\beta}$, which is the vector of shared unknown constant parameters for all subjects with dimension of $((P+1) \times 1)$ with most general interest being in the fixed effects attached to the P covariates. According to the assumptions made above, the random variables $\boldsymbol{b_i}$ and $\boldsymbol{\varepsilon_i}$ are independent and follow zero-mean normal distributions. Therefore, the marginal distribution of $\boldsymbol{y_i}$ is also normal, with mean $E[\boldsymbol{y_i}] = \boldsymbol{X}_i\boldsymbol{\beta}$ and variance $Cov(\boldsymbol{y_i}) = \boldsymbol{Z_i D Z_i^T} + \boldsymbol{\Sigma_i}$.

Maximum likelihood estimation of $\boldsymbol{\beta}$ and $\boldsymbol{D}$ involves maximizing

$$l(\boldsymbol{\beta}, \boldsymbol{D}) \propto -\{\sum_{i=1}^{N} log|Cov(\boldsymbol{y_i})| + (\boldsymbol{y_i} - \boldsymbol{X_i\beta})^T Cov(\boldsymbol{y_i})^{-1}(\boldsymbol{y_i} - \boldsymbol{X_i\beta})\} \ ;$$

Unbiased estimates $\hat{\boldsymbol{\beta}}$, and $\hat{Cov}(\boldsymbol{y_i})$ can be obtained using restricted maximum likelihood (REML)[13]. The inference of $\boldsymbol{\beta}$ is not straightforward since $Cov(\boldsymbol{\beta})$ is not known. Based on REML inference on random effects, hypothesis tests on $\boldsymbol{\beta}$ based on Wald statistics and other methods are discussed in Kenward and Roger (1997).[14]

After estimation of $\hat{D}$, $\hat{\boldsymbol{b_i}}$ can be acquired as the expectation of random variable $\boldsymbol{b_i} \mid \boldsymbol{y_i}$ where

$$\hat{\boldsymbol{b_i}} = E[\boldsymbol{b_i} \mid \boldsymbol{y_i}] = \hat{D} Z_i^T \hat{Cov}(\boldsymbol{y_i})^{-1}(\boldsymbol{y_i} - \boldsymbol{X_i\beta}) \ .$$

This allows to describe subject-specific trajectories conditional on the random effects:

$$\hat{\boldsymbol{y_i}} = \boldsymbol{X}_i\hat{\boldsymbol{\beta}} + \boldsymbol{Z_i}\hat{\boldsymbol{b_i}} \ .$$

The prediction of responses can also be interpreted as a weighted average of population-mean and observed data:

$$\hat{\boldsymbol{y_i}} = (\boldsymbol{\Sigma_i} Cov(\boldsymbol{y_i})^{-1})\boldsymbol{X}_i\hat{\boldsymbol{\beta}} + (\boldsymbol{I}_{n_i} - \boldsymbol{\Sigma_i} Cov(\boldsymbol{y_i})^{-1})\boldsymbol{y_i} \ ,$$

where the decomposition of variance gives a desirable property of shrinkage

where subjects with less information converge to population average.

Generalized linear mixed effects model (GLMM) is an extension of this model as a unified approach to different data types and assumptions of distribution of random effects. Following conventions of GLM, instead of directly specifying a relation between response and predictors through link function, GLMM has the response conditional on random effects:

$$E[\boldsymbol{y_i}|\boldsymbol{b_i}] = \boldsymbol{\mu_i}, \quad g(\boldsymbol{\mu_i}) = \boldsymbol{X_i}\boldsymbol{\beta} + \boldsymbol{Z_i}\boldsymbol{b_i} \ ,$$

where $g()$ is the link function. For the conditional response that follows a normal distribution where $\boldsymbol{y_i}|\boldsymbol{b_i} \sim N(\boldsymbol{\mu_i}, \boldsymbol{\Sigma}^2_{\boldsymbol{m_i}})$, the link function is $g(\boldsymbol{\mu_i}) = \boldsymbol{\mu_i}$. Some other link functions for conditional responses following distributions of exponential family are $g(\boldsymbol{\mu_i}) = \log(\boldsymbol{\mu_i})$ for Poisson distribution and $g(\boldsymbol{\mu_i}) = \log(\frac{\boldsymbol{\mu_i}}{1-\boldsymbol{\mu_i}})$ for binomial distribution.

Like with linear mixed effect models, GLMMs have the advantage of tolerance with sparse design and unbalanced data. However, the estimation of such model is challenging since the likelihood function

$$L(\boldsymbol{y_1}, \ldots, \boldsymbol{y_n} \mid \boldsymbol{\beta}, \boldsymbol{D}) = \int \prod_i f(\boldsymbol{y_i} \mid \boldsymbol{b_i}) f(\boldsymbol{b_i}) d\boldsymbol{b_i} \ ,$$

which involves integration over $\boldsymbol{b_i}$, and in general the likelihood cannot be solved in closed form.

One widely used method is called adaptive Gaussian quadrature introduced in Rabe-Hesketh (2002) [15], where the integration is estimated by summation of Gaussian quadrature. This method can be computationally burdensome and limited to normal assumption of $\boldsymbol{b_i}$. Implementation of both linear mixed models and generalized linear mixed models are available in the R package *lme4*. [16]

Another popular approach is to estimate entities of interest through Bayesian techniques. Instead of finding estimates of $\boldsymbol{\beta}$, $\boldsymbol{b}$ and $\boldsymbol{D}$ by solving a maximum for $L(\boldsymbol{y_1}, \ldots, \boldsymbol{y_n} \mid (\boldsymbol{\beta}, \boldsymbol{b}), \boldsymbol{D})$, $\boldsymbol{\beta}$, $\boldsymbol{b}$ and $\boldsymbol{D}$ are treated as random variables that jointly follow a posterior given $\boldsymbol{Y}$. The posterior $\pi((\boldsymbol{\beta}, \boldsymbol{b}), \boldsymbol{D} \mid \boldsymbol{Y})$ does not have a closed form, thus Markov Chain Monte Carlo methods are used to draw samples from full conditionals $\pi((\boldsymbol{\beta}, \boldsymbol{b}) \mid \boldsymbol{D}, \boldsymbol{Y})$ and $\pi(\boldsymbol{D} \mid (\boldsymbol{\beta}, \boldsymbol{b}), \boldsymbol{Y})$. Detailed description of the algorithm can be found in Zhao (2006) [17]. Implementation of the method can be found in R package *MCMCglmm* [18].

For longitudinal data with multiple outcomes, GLMMs construct inter-outcome correlation using correlation of random effects between different outcomes. One

significant challenge in extending GLMMs to multivariate longitudinal data with multiple responses is the high number of parameters to be estimated and the cumbersome covariance structure of random effects that results in heavy calculation burden. Let $\boldsymbol{b_{ik}}$ be the vector of random effects specific for the kth response for unit i, then a joint density of $\boldsymbol{y_i}$ has the form

$$f(\boldsymbol{y_i}) = \int \dots \int f(\boldsymbol{y_{i1}}, \dots, \boldsymbol{y_{iK}} | \boldsymbol{b_{i1}}, \dots, \boldsymbol{b_{iK}}) f(\boldsymbol{b_{i1}}, \dots, \boldsymbol{b_{iK}}) d\boldsymbol{b_{i1}} \dots d\boldsymbol{b_{iK}} \ ,$$

It turns out that for a limited number of responses, MCMC method can simulate the posterior fairly well as it decomposes the posterior and evaluates several simpler conditional distributions. For maximum likelihood estimate, instead of drawing inference directly from full joint likelihood, a pairwise solution is proposed in Fieuws and Verbeke (2006) [19], where estimators of the joint model are approximated as means of estimators of all possible bivariate models. Instead of specifying the joint distribution of $f(\boldsymbol{b_{i1}}, \dots, \boldsymbol{b_{iK}})$, specifying and computing bivariate joint distributions such as bivariate normal are much easier. Asymptotic inference can be obtained according to composite likelihood methods.

2.2 Marginal Model

Another common aspect of studying longitudinal data is to focus on the population-averaged response. For example, in studying the overall effect of smoking on health over time on large population, the focus will be the marginal response of smokers against non-smokers instead of individual trajectory.

Marginal model only specifies the first two moments of the marginal response: a marginal mean and a single variance structure of $\boldsymbol{y}_i$. The marginal mean has the form: $\boldsymbol{\mu}_i = E[\boldsymbol{y}_i] = h(X_i\boldsymbol{\beta})$, where $h() = g^{-1}()$ is the link function for the mean response for unit i over time with the same dimension $(m_i \times 1)$ as $\boldsymbol{y}_i$. The variance structure is a function of the mean response usually in the form $Cov(\boldsymbol{y}_i) = \phi V(\boldsymbol{\mu}_i)$ where ϕ is called the dispersion parameter, and $V(.)$ is a variance function that may not necessarily correspond to the link function.

Maximum likelihood estimation of $\boldsymbol{\beta}$ then involves solving the score function

$$\sum_{i=1}^{n} (\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}})^T Cov^{-1}(\boldsymbol{y_i})(\boldsymbol{y_i} - \boldsymbol{\mu}_i) = 0 \ .$$

Unlike in subject-specific models where covariance between responses of general linear models can be estimated using maximum likelihood estimate on joint

distribution of all y_{ij}, likelihood function is not available in the marginal setting since it does not have a fully parameterized response y_{ij}. Given the nature of repeated measurements of longitudinal data that involves high correlation between y_{ij} for different measurements of the same unit, specifying the full likelihood function is not feasible. Additionally, an unstructured covariance matrix requires estimation of $m_i * (m_i + 1)/2$ parameters for m_i time points, which risks overparameterization and is often computationally unfeasible.

The generalized estimating equation (GEE) is an approach first proposed in Zeger and Liang (1986) [20] that uses a parsimonious structure for covariance without specifying the full likelihood. The covariance is constructed to be

$$Cov(\boldsymbol{y_i}) = \phi \boldsymbol{T}_i^{1/2} \boldsymbol{R}(\rho) \boldsymbol{T}_i^{-1/2} \ ,$$

where $\boldsymbol{T}_i$ is the diagonal matrix of variances and covariances of $\boldsymbol{y}_{ij}$, and $\boldsymbol{R}(\rho)$ is the *working correlation matrix* that estimates the unknown correlation structure. This construction is considered as the exact value of $Cov(\boldsymbol{y}_i)$ assuming a correct specification of $\boldsymbol{R}(\rho)$. Under this parameterization, we avoid specification of the distributional assumption of covariance and reduce the parameters to be estimated to ϕ and ρ where ρ is typically a scalar, though it can be a vector of correlation parameters in full generality.

Examples of widely used $R(\rho)$ including exchangeable correlation:

$$\begin{bmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \dots & \dots & 1 & \dots \\ \rho & \rho & \dots & 1 \end{bmatrix} \ ,$$

where all correlations are the same, and AR(1) which is generally more applicable to longitudinal data:

$$\begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^{m_i-1} \\ \rho & 1 & \rho & \dots & \rho^{m_i-2} \\ \rho^2 & \rho & 1 & \dots & \rho^{m_i-3} \\ \dots & \dots & \dots & 1 & \dots \\ \rho^{m_i-1} & \rho^{m_i-2} & \rho^{m_i-2} & \dots & 1 \end{bmatrix} \ ,$$

where correlation between repeated measurements on a subject decreases at constant exponential rate as they are farther apart in timing of measurement. This

specific form of auto-regressive correlation structure is most suitable for balanced data with evenly distanced measurements. Markov correlation structure, which generalizes AR1 structure to handle continuous distance, can be used for measurements with unevenly spaced measurements. Discussion on choice of correlation structure can be found in Cui and Qian (2007). [22]

We then have the approximated score functions as generalized estimation equations

$$\sum_{i=1}^{n}(\frac{\partial \boldsymbol{\mu_i}}{\partial \boldsymbol{\beta}})^t \boldsymbol{V}_i^{-1}(\boldsymbol{y_i}-\boldsymbol{\mu_i}) = 0 \ ,$$

where $\boldsymbol{V}_i = (\phi \boldsymbol{T}_i^{1/2}\boldsymbol{R}(\rho)\boldsymbol{T}_i^{-1/2})$ is the estimate of $Cov(\boldsymbol{y}_i)$. An estimate of $\hat{\boldsymbol{\beta}}$ can be achieved by solving the n equations using an extended version of Iteratively Reweighted Least Squares method. Implementation of the method is available in R package *geepack*. [21]

Model selection is of great significance for marginal model through GEE given the form of variance function and working correlation are assumed. Wang (2014) [24] gives detailed review of several quantitative criteria for model selection. One of the most important criteria and a desirable feature of GEE methods is to examine the form of working correlation, Fisher Information of the approximated score function can be used as the "Naive" variance of $\boldsymbol{\beta}$. A"Robust" variance of $\boldsymbol{\beta}$ can also be estimated using $Cov(\boldsymbol{y}_i)$ after the fitted model. The robust variance is indifferent to a misspecified working correlation thus a similar naive variance indicates a valid working correlation specification.

A straightforward extension to multiple response longitudinal data of the marginalized model can be done through stacking the $\boldsymbol{y}_{ik}$ response vectors. This joint analysis requires specification of correlation both within and between responses, which can be done through Kronecker product of marginal variances and working correlations. Both $\boldsymbol{R(\rho)}$ and $\boldsymbol{V}_i$ will then be matrix with dimension $((m_iK) \times (m_iK))$. $\boldsymbol{V}_i$ consists of diagonal blocks of $\{V_1, \ldots, V_k\}$.

However for $\boldsymbol{R(\rho)}$, since it now consists of correlation both within and between responses, correlation structure for single response model such as AR(1) is not appropriate here. More complicated structures, non-stationary M-dependent or unstructured for instance, may be specified for between responses correlation blocks. To avoid risk of misspecification, a formulation of working correlation matrix $\boldsymbol{R}$ using basis matrix derived from information matrix is proposed in Cho, H. (2016). [23] A description of construction of the Kronecker products and an implementation of extension of GEE is available in R package *mmm*.

[25]

2.3 Functional Data Analysis

Some parametric methods introduced above require specifications of distributional relationship between predictors and responses and calculation of likelihood or extension of quasi-likelihood estimates. When the distribution between predictor and response is difficult to specify, semi-parametric models may be used to offer more flexibility by adding latent variables. For example, change-point models in Cudeck and Klebe (2002) [7] use time-variant effects where β_p changes over time by introducing additional timepoint variables. More latent variable models are discussed in [26]

A major problem of these methods is risk of misspecification. Additionally, the complexity of the covariance and the difficulty in computation can be potentially overwhelming even for the parsimonious covariance structure in GEE method.

To offer flexibility without extra distributional assumptions, functional data analysis (FDA) treats responses and covariates as random functions following stochastic processes instead of having covariates and effects as scalars following a parametric distribution.

For data with univariate response, a function-on-function linear model of FDA has the form

$$\boldsymbol{y}_i(t) = \boldsymbol{x}_i(t) + \boldsymbol{\varepsilon}_i(t) \ ,$$

where $E[\boldsymbol{x}_i(t)] = \boldsymbol{\mu}(t)$, $E[\boldsymbol{\varepsilon}_i(t)] = 0$. Here, $\boldsymbol{y}_i(t)$, $\boldsymbol{x}_i(t)$, $\boldsymbol{\varepsilon}_i(t)$ are all random functions that are square integrable on L^2 space. The advantage of this setting is that it offers high degree of flexibility, however, the inference of the estimates is hard to achieve given the infinite dimension of the random functions.

A practically feasible estimate of the random functions can be achieved through Functional Principle Components Analysis (FPCA). According to the Karhunen–Loève theorem, random functions can be expanded into the form

$$\boldsymbol{x}_i(t) - \boldsymbol{\mu}(t) = \sum_{w=1}^{\infty} \xi_{iw} \boldsymbol{\nu}_w(t) \ ,$$

where $\boldsymbol{\nu}_w(t)$'s are eigenfunctions of covariance matrix of $\boldsymbol{x}_i(t)$ and random variables $\xi_{iw} = \int (\boldsymbol{x}_i(t) - \boldsymbol{\mu}(t)) \boldsymbol{\nu}_w(t) dt$ have the property $E[\xi_{iw}^2] = \lambda_{iw}$ which is the corresponding eigenvalue of $\boldsymbol{\nu}_w(t)$.

This transformation turns an infinite-dimensional random function into weighted sum of independent random variables that are estimable. The infinite number of random variables can be further reduced according to the uniform convergence property where $E[\boldsymbol{x}_i(t) - \boldsymbol{\mu}(t) - \sum_{w=1}^{W} \xi_{iw}\boldsymbol{\nu}_w(t)] \to 0$ as $W \to \infty$. Because eigenvalues λ_{iw}'s sum to total variation of x_i, they can be considered as weights of eigenfunctions. Therefore, the ν_w's corresponding to the greatest W eigenvalues that forms a basis are considered functional principle components and are used to approximate x_i. The approximation has the form

$$\hat{\boldsymbol{x}}_i(t) = \hat{\boldsymbol{\mu}}(t) + \sum_{w=1}^{W} \hat{\xi}_{iw}\hat{\boldsymbol{\nu}}_w(t) \ .$$

Therefore, the response

$$\boldsymbol{y}_i(t) \approx \hat{\boldsymbol{\mu}}(t) + \sum_{w=1}^{W} \hat{\xi}_{iw}\hat{\boldsymbol{\nu}}_w(t) + \boldsymbol{\varepsilon}_i(t) \ .$$

Estimates of $\hat{\xi}_{iw}$ and $\hat{\boldsymbol{\nu}}_w(t)$ can be directly calculated using conventional principle component analysis if $Var(x_i)$ is known or estimated by empirical observation. However, the covariance matrix is often computationally hard to estimate due to the high dimension.

To avoid an explicit covariance structure, analogous estimates of $\hat{\boldsymbol{\nu}}_w(t)$ use basis functions of $\boldsymbol{x}_i$ that are commonly independent locally smoothing functions of empirical observations. The most common forms of basis functions used include cubic splines, B-splines, sine/cosine or Fourier.

Estimates of $\hat{\xi}_{iw}$ can then be acquired following the procedure of principle component analysis by conditional expectation (PACE) proposed in Yao, Müller and Wang (2005) [27]

The number of W is selected subjectively. In an ideal situation, W should be very small, and the first W principle components correspond to over 90% of the total of the variance.

An extension to jointly analyze multiple dense-design functional data with potentially different variation is proposed in *MFPCA* [28] with R package available. It follows a two-step procedure, after separately fitting each of the single response models, the functional responses are constructed with re-weighted components according to principle components.

An application for multivariate data is to estimate the correlation of re-

sponses through the correlation of scores. Sharafoddini, Dubin, Lee (2021) [29] extracted the functional principle components' scores of a set of continuous markers and then implemented a clustering step that depended on different individual's functional PCs. Lim and Cheung (2020)[30] utilized *MFPCA* to compare performances of several clustering methods.

One particular disadvantage in interpretation of the model is that the functional principle components are mostly latent curves that are not observed values. To incorporate with the observed predictors, assuming that $\boldsymbol{X_i}$ and $\boldsymbol{y_i}$ have the same support and are correlated only on current time point, a function-on-function linear model called concurrent regression model has the form

$$\boldsymbol{y}_i(t) = \boldsymbol{\beta}_0(t) + \sum_{p=1}^{P} \boldsymbol{\beta}_p(t)\boldsymbol{x}_{ip}(t) + \boldsymbol{\varepsilon}_i(t) \ ,$$

where $\boldsymbol{\beta}_p(t)$ is the time-variant effect of the pth covariate and is a deterministic scalar value function. The uniqueness of such $\boldsymbol{\beta}(t)$ is assumed to exist via assumption of an invertible $E[XX^T]$.

A common family of approaches estimates $\boldsymbol{\beta}$ by linear smoothing. With the assumption of number U_p of basis functions, the expansion has the form

$$\hat{\boldsymbol{\beta}}_p(t) = \sum_{u_p=1}^{U_p} c_{u_p} B_{u_p}(t) \ ,$$

where $B(t)'s$ are predetermined basis functions, usually smoothing splines. $\hat{\boldsymbol{\beta}}_p(t)$'s can then be treated as smoothing spline estimators, and $\hat{c}_{u_p}$'s can be solved by minimizing the summation of L2-norm of $\{\boldsymbol{y}_i(t) - \sum_{p=1}^{P}\sum_{u_p=1}^{U_p} B_{u_p}(t)x_{ip}(t)\}$ for all n subjects with penalization on smoothness of $B(t)$'s. Both the number and the form of basis functions are subjectively selected. U can be selected via cross-validation and is suggested to be smaller than the number of time points. A detailed description can be found in Hoover (1998) [31], including discussion on hyper-parameter selection (bandwidth and smoothing parameter of $B(t)$). Variations of this method including two-stage estimation of varying coefficients using local polynomial smoothing and semi-parametric models with more efficient inference are discussed in Fan and Zhang (2008) [32] .

Though variations of smoothing spline estimate methods try to improve the efficiency, non-parametric local smoothing methods do not fully incorporate the within-subject variations and perform inconsistently for sparse designs. An in-

novative approach proposed by Sentürk and Müller (2010) [33] treats the varying coefficients $\beta(t)$ as functional data and estimates its value through estimates of auto- and cross-covariances of $Cov_{X,X}(s,t)$ and $Cov_{X,y}(s,t)$. For univariate $x(t)$, the corresponding effect $\beta(t)$ has its estimator with the least square form

$$\hat{\boldsymbol{\beta}}(t) = \frac{\hat{Cov}_{\boldsymbol{X},\boldsymbol{y}}(t,t)}{\hat{Cov}_{\boldsymbol{X},\boldsymbol{X}}(t,t)} \ .$$

The estimators of covariance structures are based on local smoothing on raw covariance constructed using functional principle components. Extension of this method to multivariate model with both functional and non-functional covariates is available and was introduced in Sentürk and Nguyen (2011) [34]. These estimators of $\beta(t)$ incorporate the underlying functional components of $X(t)$ and $y(t)$ and also all observations across entire time period, which helps eliminate bias and overcome sparseness. Discussion on the explicit form of covariance structures and implementation of this method are available in R package *FDA-PACE*[36].

For multiple responses, the variance structure of a varying coefficient model with functional responses and covariates is significantly more complicated. For functional-scalar model, the problem can be solved by decomposition of covariance of X and Y. Zhu (2012) [35] proposed a method that treats X as a scalar and introduces the subject-specific variations as a random function $\eta_{ij}(t)$. Local smoothing is applied to estimate $\eta_{ij}(t)$ and $\beta_j(t)$ ($j \in \{1,\ldots,P\}$ for P responses), while covariance for $\eta_{ij}(t)$ is estimated using FPCA. Auto- and cross variance of Y and X can then be estimated as a composition of Σ_η and variance of scalar X.

3 Analysis

In this section, we applied some of the methods mentioned in Section 2 to a real longitudinal dataset with multivariate responses.

Specifically, the methods are

- GLMM implemented using R package *MCMCglmm*
- Marginal model using GEE implemented from *mmm*
- FDA model using implementation from *fdapace*

The dataset analyzed here is a subset of original data from the Mayo Clinic trial in primary sclerosing cholangitis (PBC) conducted between 1974 and 1984. The subset data was studied in Komárek and Komárková (2013)[37] and consists of the patients that were not censored until day 910.

There are three binary responses: presence of edema(edema), presence of hepatomegaly or enlarged liver (hepato) and blood vessel malformations in the skin (spiders); and three continuous predictors: serum albumin (g/dl) (albumin), log of serum bilirunbin (mg/dl) (lbili) and platelet count (lplatelet). Figure 1 shows the trajectories of logarithm of platelet for all subjects, while Figure 2 shows one of the responses.

The measurements of each subject are roughly at 0, 6, 12 and 24 months after the observation start and are occasionally missing. This means we have a sparse design with unbalanced data, which can be handled by the selected methods with some restrictions.

The goal of the following analysis is to apply the selected methods to study the co-evolution between responses and predictors.

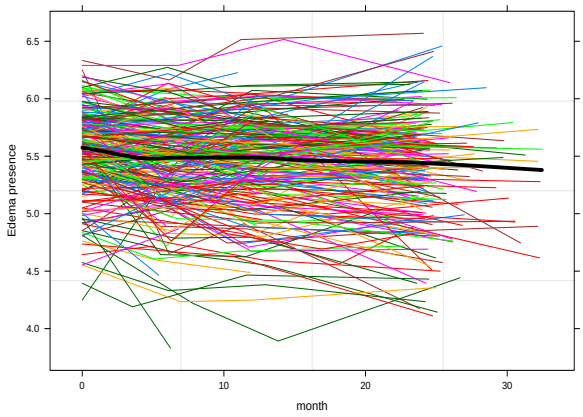

Figure 1: Log(platelet) vs month with loess curve superimposed

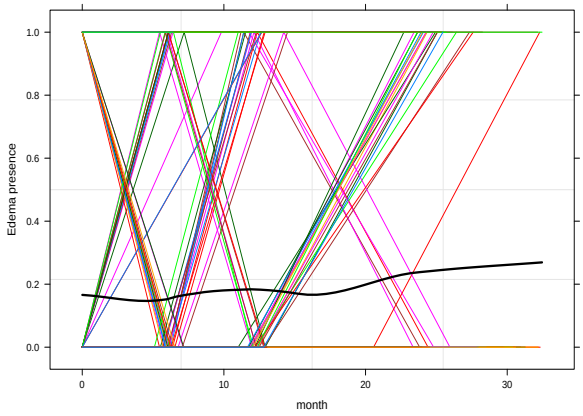

Figure 2: Edema vs month with loess curve superimposed

3.1 GLMM using *MCMCglmm*

We apply joint analysis of the three binary responses using a generalized linear model with all three predictors as fixed effects and intercept and time as random effect.

This method gives inference of within-subject variance, between-subject variance, and fixed effects through sampling of MCMC paths as in Figure 3. Inference of fixed effects is shown in Table 1, where we can see all covariates are significant except for albumin for the spider response where the confidence interval of the effect covers 0. Combined with inference of random effects, we can obtain prediction of the trajectory of any subject's responses, as shown for two

example participants for one response in Figure 4.

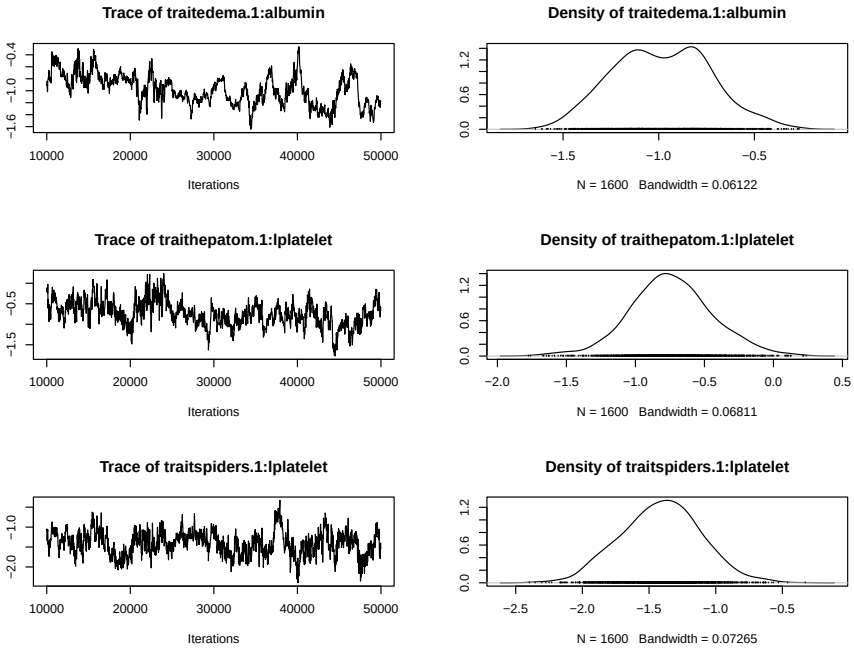

Figure 3: MCMC samples of some of the fixed effects

MCMCglmm fixed effect inference			
Fixed effects	post.mean	l-95% CI	u-95% CI
Response: hepatom			
lbili	1.428	0.968	1.936
albumin	-1.124	-1.844	-0.581
lplatelet	-0.740	-1.286	-0.074
Response: spiders			
lbili	1.600	1.004	2.175
albumin	-0.825	-1.669	0.000
lplatelet	-1.411	-1.989	-0.861
Response: edema			
lbili	0.655	0.175	1.086
albumin	-0.983	-1.463	-0.508
lplatelet	-1.245	-1.710	-0.749

Table 1: Estimation and inference of fixed effects corresponding to each of the responses in joint analysis

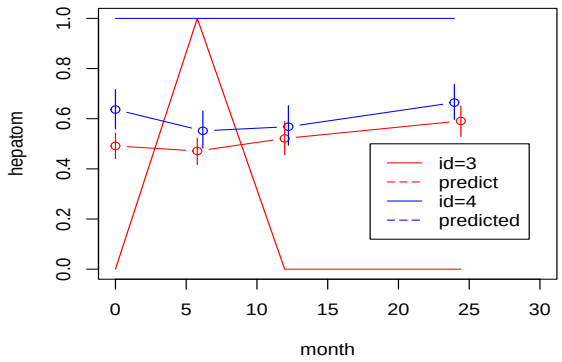

Figure 4: prediction of two subjects' trajectories of evolution of presence of hepatomegaly with 95% confidence interval

3.2 Marginal Model based on GEE using *mmm*

The marginal model jointly analyzes the three responses using all three predictors. The working correlation matrix has a dimension of (15×15) where $15 =$ (maximum of 5 measurements per subject) $\times$ (3 responses) and uses an exchangeable structure with $\rho = 0.176$ as in Table 2.

The estimates of the effects and both their naive and robust standard errors are shown in Table 3. Compared with results in GLMM model, estimates of effects have the same directions but are smaller in marginal model. Wald-statistic tests using both naive and robust standard error claim the effect of albumin for spiders response insignificant, as in the GLMM model.

Noticing that the naive standard error is quite different from robust standard error for some effects. In this case, the robust standard error is preferred. This indicates that the exchangeable correlation structure may not be appropriate to estimate the working correlation, and, seeing the four time points are not all equidistant, we might want to consider a continuous auto-regressive working correlation structure in a follow-up analysis.

	1	2	...	15
1	1	0.176	...	0.176
2	0.176	1	...	0.176
⋮	⋮	⋮	⋮	⋮
15	0.176	0.176	...	1

Table 2: Estimated exchangeable working correlation

GEE inference			
	Estimate	Naive S.E.	Robust S.E.
Response: hepatom			
Intercept	6.862	1.251	1.457
albumin	-0.886	0.186	0.194
lbili	0.748	0.096	0.114
lplatelet	-0.769	0.197	0.243
Response: spiders			
Intercept	0.966	1.213	1.546
albumin	-0.215	0.186	0.241
lbili	0.699	0.094	0.113
lplatelet	-0.285	0.196	0.238
Response: edema			
Intercept	4.145	1.328	1.752
albumin	-0.476	0.208	0.270
lbili	0.297	0.100	0.129
lplatelet	-0.753	0.214	0.276

Table 3: Estimation and inference of fixed effects corresponding to each of the responses in joint analysis

3.3 Functional Data Analysis using *fdapace*

In functional data analysis, the observed measurements are considered to follow latent stochastic processes with infinite dimensions. These processes can be estimated using functional principle component analysis implemented through PACE algorithm. The estimates of functional principle components of one of the predictors is shown in Figure 5. The scree plot indicates an ideal set of principle components as first two of the components explain over 90% of all variation of the data. The estimated mean function as a weighted average of eigenfunctions is more volatile in the latter half of the time period. This corresponds to the

observation in the design plot where the time of measurements is more sparse after 15 months.

The estimated principle components can then be used in a varying-coefficient concurrent regression model to estimate the predictors as functional data. As discussed in Seciton 2, the regression coefficients are also functions of time. The predicted mean function of one response and predicted mean function of fitted value of the response is shown in Figure 6 at 50 evenly-spaced time points over the range of all measurement time points. We can see the predicted mean function largely agrees with the estimated mean function at time points where number of observations is high. The prediction is very volatile around 20 and 30 where observations are few.

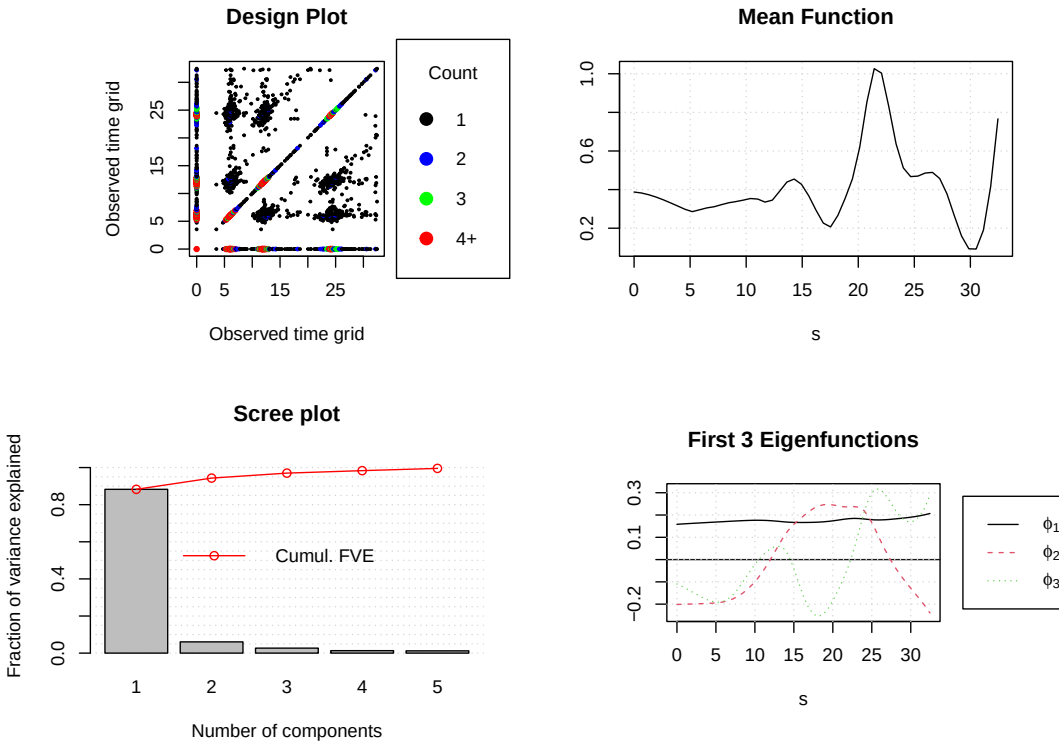

Figure 5: Functional principle component analysis of the continuous predictor Log(serum bilirunbin)

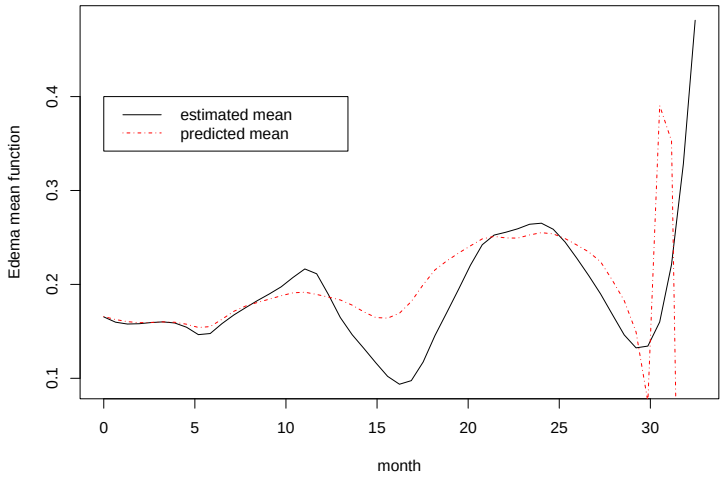

Figure 6: Estimated mean function of Edema using FPCA and Predicted mean function of Edema using varying-coefficient regression model based on functional predictors

4 Discussion

In this paper we introduced a non-exhausted list of methods that deal with longitudinal data. Three families of models are discussed in detail. The first family of models are generalized linear mixed effects models that are capable of describing subject-specific trajectories where random effects are specified. The next family of models are marginal models that focuses on the mean response of population without specifying subject-specific variations. Working correlation matrix is assumed to avoid dealing with an unstructured correlation matrix and to aid in estimation. For predictors with density of exponential family, we introduced the generalized estimating equations to estimate the effects of predictors. The last family of models introduced are functional data analysis models where observations of each subject is assumed to follow a stochastic process. These models require no distributional assumptions and are very flexible, but are computationally difficult with infinite dimensions. Functional principle components analysis is then applied to estimate the observations using weighted latent curves that explain most of the underlying variation.

We also briefly discussed the extension of each model to multivariate longitudinal data. The maximum likelihood estimate is computationally difficult to solve for subject-specific models when there are multiple responses because of the large dimension of covariance matrix. One approach to work around the likelihood function is to use Bayesian estimation to draw samples from full conditional distributions of each parameter. We applied this method using implementation from R package MCMCglmm to jointly analyze three binary responses in PBC data. The estimates and credible intervals of fixed effects are acquired from the MCMC samples. We can also predict the subject-specific trajectories of each response.

For marginal model, a straightforward extension to multiple responses uses Kronecker-product representation to jointly analyze three responses. An exchangeable correlation structure is used to describe correlations between responses.

In functional data analysis, latent stochastic process of each predictor is estimated using functional principle component analysis. In our analysis, an ideal estimate of a predictor is shown as the first two components explain more than 90% of the total variations of the predictor. A varying-coefficient model using implementation of functional concurrent regression is then applied in analysis of each response using functional principle components of predictors.

There are other widely used methods that are not covered in this paper. One method that is also popular particularly in clinical trials is joint model of survival responses and multivariate longitudinal predictors. The joint model fisrt estimates an underlying longitudinal trajectory for each predictor and uses the estimated values for covariate in survival models. Wulfsohn & Tsiatis (1997)[38] explored such model and argued its superiority over conventional survival models directly using observed covariate values. A review of joint models on multivariate outcomes can be found in Hickey (2016) [39]. Another method widely seen in social studies is structural equation model (SEM). This family of models focuses on the change processes where a construct is assumed with latent indicators of underlying states and paths between them. Little (2013) [40] did a systematic discussion on extending SEM to longitudinal data.

Further developments can be done when applying the methods covered in Section 3. All the methods can use hyperparameter tuning or model selection for better performance. For the estimation of fixed effects in GLMM using MCMC, the estimates may be improved with an appropriate prior because the MCMC sample paths seem to have poor mixing while the distribution of fixed effect of albumin for response edema seem to have multiple modes. For marginal model, the exchangeable correlation matrix may be an oversimplified estimate of the real underlying correlation structure, especially considering the correlation between responses. A more complicated correlation structure such as banded correlation matrix or continuous auto-regressive correlation matrix may be considered.

Another focus of joint analysis of multivariate longitudinal data is the correlation between outcomes. This was not comprehensively discussed in above sections due to the complexity of the problem and lack of implemented analysis methods in R. For GLMMs, the pairwise fitting [19] approach constructs an estimate of covariance matrix of inter-outcome responses. For marginal models using GEE, Rochon (1996)[2] proposed joint modelling in bivariate case using seemingly unrelated regression to construct an intra- and inter-outcomes covariance matrix for two separate GEE models. Another method proposed by Gray and Brookmeyer [1] explicitly estimates an overall treatment effect at each time point over different outcomes using odds ratios or correlation depending on data type. For functional data analysis, a joint analysis of dense-design multiple response longitudinal data is discussed in Happ & Greven (2018) [41].

References

[1] Gray, S. M., & Brookmeyer, R. (2000). *Multidimensional longitudinal data: estimating a treatment effect from continuous, discrete, or time-to-event response variables.* Journal of the American Statistical Association, 95(450), 396-406.

[2] Rochon, J. (1996). *Analyzing bivariate repeated measures for discrete and continuous outcome variables.* Biometrics, 740-750.

[3] Cnaan, A., Laird, N. M., & Slasor, P. (1997). *Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data.*, Statistics in medicine, 16(20), 2349-2380.

[4] PAUL S. ALBERT (1999), *Longitudinal Data Analysis (Repeated Measures) in Clinical Trials*, Statistics in Medicine, Vol. 18: 1707-1732

[5] Davidian, M., & Giltinan, D. M. (2003). *Nonlinear models for repeated measurement data: an overview and update.*, Journal of agricultural, biological, and environmental statistics, 8(4), 387-419.

[6] Fan, J., & Li, R. (2004)., *New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis.*, Journal of the American Statistical Association, 99(467), 710-723.

[7] Cudeck, R., & Klebe, K. J. (2002)., *Multiphase mixed-effects models for repeated measures data.*, Psychological methods, 7(1), 41.

[8] Rice, J. A. (2004), *Functional and longitudinal data analysis: perspectives on smoothing.*, Statistica Sinica, 631-647.

[9] Meyer, Peter M., (2007), *Characterizing daily urinary hormone profiles for women at midlife using functional data analysis*, American journal of epidemiology 165.8 : 936-945.

[10] Verbeke,Geert (2014), *The analysis of multivariate longitudinal data: A review*, Statistical methods in medical research 23.1 : 42-59.

[11] Laird, N. M., & Ware, J. H. (1982)., *Random-effects models for longitudinal data.*, Biometrics, 963-974.

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

[13] Harville, David A. (1977), *Maximum likelihood approaches to variance component estimation and to related problems*, Journal of the American statistical association 72.358 : 320-338.

[14] Kenward, M. G., & Roger, J. H. (1997)., *Small sample inference for fixed effects from restricted maximum likelihood.*, Biometrics, 983-997.

[15] Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2002). *Reliable estimation of generalized linear mixed models using adaptive quadrature.*, The Stata Journal, 2(1), 1-21.

[16] Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). *Fitting Linear Mixed-Effects Models Using lme4*, Journal of Statistical Software, 67(1), 1–48.

[17] Zhao, Y., Staudenmayer, J., Coull, B. A., & Wand, M. P. (2006). *General design Bayesian generalized linear mixed models.*, Statistical science, 35-51.

[18] Hadfield, J. D. (2010). *MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package.*, Journal of statistical software, 33, 1-22.

[19] Fieuws, S., & Verbeke, G. (2006). *Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles.*, Biometrics, 62(2), 424-431.

[20] Zeger, S. L., & Liang, K. Y. (1986)., *Longitudinal data analysis for discrete and continuous outcomes*, Biometrics, 121-130.

[21] Halekoh U, Højsgaard S, & Yan J (2006)., *The R Package geepack for Generalized Estimating Equations.*, Journal of Statistical Software, 15/2, 1–11.

[22] Cui, J. & Qian, G. (2007), *Selection of working correlation structure and best model in GEE analyses of longitudinal data.*, Communications in statistics—Simulation and computation, 36(5), 987-996.

[23] Cho, H. (2016). *The analysis of multivariate longitudinal data using multivariate marginal models.*, Journal of Multivariate Analysis, 143, 481-491.

[24] Wang, Ming (2014), *Generalized Estimating Equations in Longitudinal Data Analysis: A Review and Recent Developments*, Advances in Statistics 2014

[25] Asar, Özgür, & Özlem İlk.(2013) *mmm: an R package for analyzing multivariate longitudinal data with multivariate marginal models*, Computer Methods and Programs in Biomedicine 112.3: 649-654.

[26] Skrondal, A. & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models.*, Chapman and Hall/CRC.

[27] Yao, F., Müller, H. G., & Wang, J. L. (2005)., *Functional data analysis for sparse longitudinal data*, Journal of the American statistical association, 100(470), 577-590.

[28] Happ, C., & Greven, S. (2018). *Multivariate functional principal component analysis for data observed on different (dimensional) domains*, Journal of the American Statistical Association, 113(522), 649-659.

[29] Sharafoddini, Anis, Joel A. Dubin, & Joon Lee.(2021), *Identifying subpopulations of septic patients: A temporal data-driven approach*, Computers in Biology and Medicine 130 : 104182.

[30] Lim, Y., Cheung, Y. K., & Oh, H. S. (2020)., *A generalization of functional clustering for discrete multivariate longitudinal data.*, Statistical Methods in Medical Research, 29(11), 3205-3217.

[31] Hoover, D. R., Rice, J. A., Wu, C. O., & Yang, L. P. (1998)., *Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data*, Biometrika, 85(4), 809-822.

[32] Jianqing Fan & Wenyang Zhang (2008), *Statistical Methods with Varying Coefficient Models*, Stat Interface, 1(1): 179–195

[33] Şentürk, D., & Müller, H. G. (2010), *Functional varying coefficient models for longitudinal data*, Journal of the American Statistical Association, 105(491), 1256-1264.

[34] Sentürk, D. & Nguyen, D.V. (2011), *Varying Coefficient Models for Sparse Noise-contaminated Longitudinal Data*, Statistica Sinica 21(4), 1831-1856.

[35] HONGTU ZHU, RUNZE LI & LINGLONG KONG, *Multivariate Varying Coefficient Model for Functional Response*, The Annals of Statistics 2012, Vol. 40, No. 5, 2634–2666

[36] Gajardo A, Bhattacharjee S, Carroll C, Chen Y, Dai X, Fan J, Hadjipantelis P, Han K, Ji H, Zhu, C, Müller H, Wang J (2021). *fdapace: Functional Data Analysis and Empirical Dynamics*

[37] Komárek, A. & Komárková, L. (2013)., *Clustering for multivariate continuous and discrete longitudinal data.*, The Annals of Applied Statistics, 7(1), 177–200.

[38] Wulfsohn, M. S., & Tsiatis, A. A. (1997)., *A joint model for survival and longitudinal data measured with error.*, Biometrics, 330-339.

[39] Hickey, G. L., Philipson, P., Jorgensen, A. & Kolamunnage-Dona, R. (2016)., *Joint modelling of time-to-event and multivariate longitudinal outcomes: recent developments and issues.*, BMC medical research methodology, 16(1), 1-15.

[40] Little, T. D. (2013)., *Longitudinal structural equation modeling.*, Guilford press.

[41] Happ, C.,& Greven, S. (2018) , *Multivariate functional principal component analysis for data observed on different (dimensional) domains.*, Journal of the American Statistical Association, 113(522), 649-659.