Multivariate longitudinal data in presence of missing data

A comparison between mixed-effects models and multivariate functional principal component analysis

Arnau García

May 16, 2025

The main goal in this report is to conduct a comparison between mixed-effects models (MM) [5] and multivariate functional principal component analysis (MFPCA) [3] to model multivariate longitudinal data in presence of missing data. Firstly, a theoretical analysis on the relationship of MMs as well as MFPCA with missingness mechanisms will be carried out. Then, an empirical simulation study will be performed in order to assess how both methods respond to different missing data scenarios.

MMs are widely used in the literature, since they effectively address missing data challenges. Nonetheless, the usage of random effects implies a high computational cost, which in the situation where we have several longitudinal markers increases even more. These computational costs mean that in practice, when one has a large number of longitudinal biomarkers, the problem becomes prohibitive.

On the other hand, functional principal component analysis (FPCA) [15] is a non-parametric approach that is flexible and well suited to model sparsely sampled longitudinal data at a potentially lower computational cost. In order to deal with multiple longitudinal outcomes, we will use the multivariate functional principal component analysis presented in [3]. This is because the latter is a flexible approach, and is the chosen one in all the references consulted in the literature [7] [6].

In health studies, FPCA is very appealing since it allows to describe and summarize temporal trajectories in a parsimonious way using only a few principal components and individual contributions to these components. However, to be suitable for the analysis of longitudinal data in health studies, FPCA-based methods should handle the challenges of missing data. The main unsolved challenge is precisely the behavior of FPCA techniques with missingness

mechanisms. No theoretical analysis has been carried out. A numerical assessment by means of a simulation study has been done recently in [12]. Nonetheless, neither a theoretical nor an empirical analysis has been made in the case of multivariate longitudinal data.

Let us assume that we have n individuals, as well as L different longitudinal biomarkers for each patient in the study, that we denote as $\boldsymbol{y}_i = (\boldsymbol{y}_{1i}, \ldots, \boldsymbol{y}_{li}, \ldots, \boldsymbol{y}_{Li})^\top$ for the i-th individual. Moreover, $\boldsymbol{y}_{li}$ is the $n_{li} \times 1$ longitudinal response vector for the i-th subject and the l-th longitudinal response, with element $y_{li}(t_{ij})$ being the value of the longitudinal outcome taken at time point t_{ij}, $j = 1, \ldots, n_{li}$ (n_{li} is the number of measurements for the i-th subject, l-th longitudinal response).

Furthermore, we assume the L longitudinal outcomes to be continuous. Although mixed-effects models are able to handle different types of data (binary or count data) using generalized linear mixed model theory, this is not the case for FPCA-based methods. The latter can only deal with continuous outcomes.

Missing data mechanisms

Some definitions have to be introduced to work with missing data in a proper formal framework. We use [10] as a benchmark to define the different quantities in the multivariate case. We define the missing data indicator as

$$R_{li}(t_{ij}) = \begin{cases} 1 \text{ if } y_{li}(t_{ij}) \text{ is observed} \\ 0 \text{ otherwise.} \end{cases}$$

Therefore, we obtain a partition of the complete response vector $\boldsymbol{y}_i = (\boldsymbol{y}_{1i}, \ldots, \boldsymbol{y}_{Li})^\top$ into two subvectors, the observed data subvector $\boldsymbol{y}_i^o$ containing those $y_{li}(t_{ij})$ for which $R_{li}(t_{ij}) = 1$, and the missing data subvector $\boldsymbol{y}_i^m$ containing the remaining components. The vector $\boldsymbol{R}_i = (\boldsymbol{R}_{1i}, \ldots, \boldsymbol{R}_{li}, \ldots, \boldsymbol{R}_{Li})^\top$, where $\boldsymbol{R}_{li} = (R_{li}(t_{i1}), \ldots, R_{li}(t_{in_{li}}))^\top$, and the process generating $\boldsymbol{R}_i$ are referred to as the missing data process.

When missingness is restricted to dropout or attrition, the missing data indicator can be replaced by the vector $\boldsymbol{R}_i^d$ defined as $\boldsymbol{R}_i^d = (R_{1i}^d, \ldots, R_{li}^d, \ldots, R_{Li}^d)^\top$, where

$$R_{li}^d = 1 + \sum_{j=1}^{n_{li}} R_{li}(t_{ij}).$$

For an incomplete sequence, R_{li}^d denotes the occasion at which dropout occurs, whereas for a complete sequence, $R_{li}^d = n_{li} + 1$, for each longitudinal outcome $l = 1, \ldots, L$. In both cases R_{li}^d equals one plus the length of the observed measurement sequence, whether this is complete or incomplete.

The appropriateness of different methods of analysis of incomplete longitudinal data is determined by the missing data mechanism. The missing data mechanism can be thought of as the probability model describing the relation between the missing data $\boldsymbol{R}_i$ and response data $\boldsymbol{y}_i$ processes. A taxonomy of missing data mechanisms based on the conditional density of the missingness process $\boldsymbol{R}_i$ given the complete response vector $\boldsymbol{y}_i = (\boldsymbol{y}_i^o, \boldsymbol{y}_i^m)$:

$$p(\boldsymbol{R}_i|\boldsymbol{y}_i^o, \boldsymbol{y}_i^m),$$

is developed in [8]. The three types of missing mechanisms (presented in [8] or [9]) are:

- **Missing completely at random (MCAR):**

Data is called missing completely at random (MCAR) when the probability that the biomarkers are missing does not depend on the biomarkers at all. That is, longitudinal data are MCAR when $\boldsymbol{R}_i$ is independent of both $\boldsymbol{y}_i^o$ and $\boldsymbol{y}_i^m$,

$$p(\boldsymbol{R}_i|\boldsymbol{y}_i^o, \boldsymbol{y}_i^m) = p(\boldsymbol{R}_i).$$

An example of MCAR longitudinal data is encountered in health surveys in which subjects go in and out of the study after providing a predetermined number of repeated measurements. Since the number and timing of the measurements is determined by design, the probability of obtaining a response is unrelated to the actual measurements.

- **Missing at random (MAR):**

MAR assumes that the probability of missingness depends on the set of observed responses, but is unrelated to the outcomes that should have been obtained. That is, longitudinal data are MAR when $\boldsymbol{R}_i$ is conditionally independent of $\boldsymbol{y}_i^m$ given $\boldsymbol{y}_i^o$,

$$p(\boldsymbol{R}_i|\boldsymbol{y}_i^o, \boldsymbol{y}_i^m) = p(\boldsymbol{R}_i|\boldsymbol{y}_i^o).$$

An important characteristic of MAR is that the predictive distribution of the missing longitudinal responses $\boldsymbol{y}_i^m$ given the observed data $\boldsymbol{y}_i^o$ and $\boldsymbol{R}_i$, does not depend on the missingness process. We can easily see this with

$$\begin{aligned} p(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o, \boldsymbol{R}_i) &= \frac{p(\boldsymbol{y}_i^m, \boldsymbol{y}_i^o, \boldsymbol{R}_i)}{p(\boldsymbol{y}_i^o, \boldsymbol{R}_i)} = \frac{p(\boldsymbol{R}_i|\boldsymbol{y}_i^o, \boldsymbol{y}_i^m)p(\boldsymbol{y}_i^o, \boldsymbol{y}_i^m)}{p(\boldsymbol{R}_i|\boldsymbol{y}_i^o)p(\boldsymbol{y}_i^o)} \\ &= \frac{p(\boldsymbol{R}_i|\boldsymbol{y}_i^o)p(\boldsymbol{y}_i^o, \boldsymbol{y}_i^m)}{p(\boldsymbol{R}_i|\boldsymbol{y}_i^o)p(\boldsymbol{y}_i^o)} = \frac{p(\boldsymbol{y}_i^o, \boldsymbol{y}_i^m)}{p(\boldsymbol{y}_i^o)} = p(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o). \end{aligned}$$

An example of MAR longitudinal data arises when a study protocol requires patients whose response value exceeds a specific medically relevant threshold to be removed from the study. In this case, missingness is under the control of the investigator and is related to the observed components of $\boldsymbol{y}_i$ only.

- **Missing not at random (MNAR):**

MNAR assumes that the probability that longitudinal responses are missing depends on a subset of the responses we would have observed. Specifically, the distribution of $\boldsymbol{R}_i$ depends on at least some elements of $\boldsymbol{y}_i^m$, even if we condition on $\boldsymbol{y}_i^o$,

$$p(\boldsymbol{R}_i|\boldsymbol{y}_i^o, \boldsymbol{y}_i^m) = p(\boldsymbol{R}_i|\boldsymbol{y_i^m}) \quad \text{or} \quad p(\boldsymbol{R}_i|\boldsymbol{y}_i^o, \boldsymbol{y}_i^m) = p(\boldsymbol{R}_i|\boldsymbol{y}_i^o, \boldsymbol{y_i^m}).$$

An example of MNAR longitudinal data arises in pain studies in which patients may ask for rescue medication when their pain levels exceed the threshold they can tolerate. However, contrary to MAR, the predictive distribution of $\boldsymbol{y}_i^m$ conditional on $\boldsymbol{y}_i^o$ is not the same as in the target population, but rather depends on both $\boldsymbol{y}_i^o$ and on $p(\boldsymbol{R}_i|\boldsymbol{y}_i)$. Thus, the model assumed for the missingness process is crucial and must be included in the analysis

Furthermore, it is important to remark that in the definitions above we have implicitly assumed that the probability of missingness may depend on covariates. Note also that so far we have not worked with parameters. All the section has been carried out without the use of parameters, in all generality so that it can be used for parametric and non-parametric models.

Finally, when working under the parametric paradigm, it is important to keep in mind that all the classification and concepts presented above are equally valid working from the frequentist approach as from the Bayesian approach. A discussion on analysis of problems involving missing data under the Bayesian framework is presented in [1]. Nonetheless, the main difference when working under the Bayesian approach is that Bayesian inference draws no distinction between missing data and parameters. Both are uncertain, and they have a joint posterior distribution, conditional on observed data. The practical distinction arises when setting up the joint model for observed data, unobserved data, and parameters.

Mixed-effects models

The reader is addressed to [13] to see a review of several type of parametric models used to conduct multivariate longitudinal data analysis. Among all the possibilities, in this work we choose the random effects model, because of its flexibility and good properties. In the random effects model, the association between the outcomes is generated by allowing the random effects themselves to be correlated. Moreover, another reason for using these models is our interest in further application in joint models for longitudinal and time to event data [10].

Since we assumed continuous longitudinal outcomes, we focus on normally distributed longitudinal outcomes. Then, we use a linear mixed-effects model to describe the subject-specific longitudinal trajectories. Specifically, we have

$$y_{li}(t) = m_{li}(t) + \varepsilon_{li}(t) = \boldsymbol{x}_{li}^{\top}(t)\boldsymbol{\beta}_l + \boldsymbol{z}_{li}^{\top}(t)\boldsymbol{b}_{li} + \varepsilon_{li}(t), \quad l = 1, \ldots, L, \tag{1}$$

where $y_{li}(t)$ denotes the value of the lth longitudinal outcome for the ith subject at time point t, $\boldsymbol{x}_{li}(t)$ and $\boldsymbol{z}_{li}(t)$ denote the time-dependent design vectors for the fixed-effects $\boldsymbol{\beta}$ and for the random effects $\boldsymbol{b}_{li}$, respectively. Finally, $\varepsilon_{li}(t)$ are the corresponding error terms for the l-th longitudinal outcome, that are assumed to be normally distributed with mean 0 and variance σ^2, as well as independent of the random effects, and $Cov(\varepsilon_i(t), \varepsilon_i(t')) = 0$ for $t \neq t'$. To account for the association between the multiple longitudinal outcomes, we link their corresponding random effects. More specifically, the complete vector of random effects $\boldsymbol{b}_i = (\boldsymbol{b}_{1i}^{\top}, \boldsymbol{b}_{2i}^{\top}, \ldots, \boldsymbol{b}_{Li}^{\top})^{\top}$ is assumed to follow a multivariate normal distribution with mean zero and variance-covariance matrix $\boldsymbol{D}$.

We assume the repeated measurements within each longitudinal process to be conditionally independent on the random effects. In addition, we assume conditional independence between the L different longitudinal outcomes, on the random effects. That is, we let random effects account for the correlation between measurements within a subject, as well as between the different longitudinal outcomes. Under these assumptions, the likelihood contribution for the i-th subject conditional on the parameters and random effects takes the form:

$$\begin{aligned} p\left(\boldsymbol{y}_i \mid \boldsymbol{b}_i, \boldsymbol{\theta}\right) &= \big[\prod_{l=1}^{L}\prod_{j=1}^{n_{li}} p\left(y_{li}(t_{ij})|\boldsymbol{b}_{li}, \boldsymbol{\theta}\right)\big] p(\boldsymbol{b}_i|\boldsymbol{\theta}) \\ &\propto \left(\sigma^2\right)^{-\frac{1}{2}\sum_{l=1}^{L} n_{li}} \exp\left\{-\sum_{l=1}^{L}\sum_{j=1}^{n_{li}} \frac{\left(y_{li}(t_{ij}) - \boldsymbol{x}_{li}(t_{ij})^{\top}\boldsymbol{\beta}_l - \boldsymbol{z}_{li}(t_{ij})^{\top}\boldsymbol{b}_{li}\right)^2}{2\sigma^2}\right\} \\ &\quad \times \det(\boldsymbol{D})^{-1/2} \exp\left(-\boldsymbol{b}_i^{\top}\boldsymbol{D}^{-1}\boldsymbol{b}_i/2\right), \end{aligned} \tag{2}$$

where $\boldsymbol{\theta}$ denotes the vector of all parameters involved in the model. In the first line of (2) after the proportionality symbol, the reader can see the contribution of the longitudinal outcomes taking into account the different assumptions we have made. In the second line, we have the contribution of the random effects to the likelihood.

If working under the frequentist approach, estimates of the parameters $\boldsymbol{\theta}$ are derived by maximizing the log-likelihood:

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log p(\boldsymbol{y}_i; \boldsymbol{\theta}) = \sum_{i=1}^{n} \log \int p(\boldsymbol{y}_i, \boldsymbol{b}_i; \boldsymbol{\theta}) d\boldsymbol{b}_i = \sum_{i=1}^{n} \log \int p(\boldsymbol{y}_i | \boldsymbol{b}_i, \boldsymbol{\theta}_y) p(\boldsymbol{b}_i | \boldsymbol{\theta}_b) d\boldsymbol{b}_i,$$

where $\boldsymbol{\theta}_y$ and $\boldsymbol{\theta}_b$ are denoting the parameters for the longitudinal outcomes and the parameters of the random-effects, respectively. When the model is linear in the random effects, the integral has an analytical solution. In other cases, numerical integration is required.

On the other hand, if working under the Bayesian paradigm, first, prior distributions for all the parameters involved in the model must be specified. Then, taking into account that the posterior distribution is a compromise between the prior distribution and the likelihood, we derive posterior inferences using, for instance, a Markov chain Monte Carlo algorithm (MCMC). Regarding the prior distributions, we can take standard non-informative priors, such as normal priors for all the regression coefficients, and inverse-gamma priors for the variance-related parameters.

Although MM provides us with a flexible and well-defined framework for modeling longitudinal data, we must not lose sight of the fact that we are working with a parametric model and that when L grows, so does the number of parameters and the complexity of our model. In addition, we must sometimes make assumptions about the variance structure of the random-effects, to avoid working with a prohibitive number of covariance parameters.

An essential feature of the MM, key in this work, is its robustness to MAR data, which ensures an unbiased estimator under MAR [11]. Nevertheless, when the missing data are suspected to be MNAR, a joint model of the missingness mechanism is required. When working under the MNAR case, the model requires a correct specification of the missing data mechanism to yield robust estimates.

Multivariate functional principal component analysis

To read about all the theoretical concepts, as well as go through all the asymptotic properties and its proofs, the reader is addressed to [3]. In order to present in this work the MFPCA, and how estimation is conducted, we take profit of the relationship between univariate and multivariate FPCA for finite Karhunen-Loève decomposition, stated and proven in [3]. Moreover, we follow [6], as well as [7], to carry out the explanations in this section. It is important to take into account that MFPCA is a non-parametric approach. Our goal in this section is to explain the methods behind MFPCA, and then try to discuss how missing data mechanisms can affect to these methods.

To model continuous longitudinal outcomes, we assume that the observed data $\boldsymbol{y}_{li}$ is a noisy measurement of the latent outcome process $X_{li}(t)$, where time $t \in \mathcal{T} = [0, \tau]$ and $\tau = max\{T_i^* : i = 1, \ldots, n\}$. This is, we end up with $y_{li}(t_{ij}) = X_{li}(t_{ij}) + \varepsilon_{li}(t_{ij})$, where $\varepsilon_{li}(t_{ij})$ are independent errors centered and with variance $\sigma^2_{\varepsilon_l}$.

Let $\boldsymbol{X}_i(\boldsymbol{t}) = \{X_{li}(t_l)\}_{l=1,\ldots,L}$ be the multivariate longitudinal processes of the ith subject, where $\boldsymbol{t} = (t_1, \ldots, t_L) \in \boldsymbol{\mathcal{T}} = \mathcal{T}_1 \times \cdots \mathcal{T}_L$, and $\mathcal{T}_1, \ldots, \mathcal{T}_L$ are, possibly different, closed intervals. Let us explain how to model $\boldsymbol{X}_i(t)$ by means of MFPCA. A two-step procedure will be used. In the first step, we let $\mu_l(t)$ be the unknown smoothed mean function of $X_{li}(t)$ (the longitudinal process of the ith subject and lth longitudinal outcome). Moreover, we let $\Sigma_l(t, t') = Cov\left(X_{li}(t), X_{li}(t')\right)$ be the covariance function that models the correlation of outcome l's trajectory between time points t and t'. By the Mercer's theorem (see the Mercer's Theorem), the spectral decomposition of the covariance function is

$$\Sigma_l(t, t') = Cov\left(X_{li}(t), X_{li}(t')\right) = \sum_{r=1}^{\infty} \lambda_{lr}\phi_{lr}(t)\phi_{lr}(t'),$$

where $\{\lambda_{lr}\}_{r=1,\ldots,\infty}$ are non-increasing eigenvalues and $\{\phi_{lr}(t)\}_{r=1,\ldots,\infty}$ are the corresponding orthonormal eigenfunctions. Furthermore, the Karhunen Loève (see KL theorem) expansion of the stochastic process $X_{li}(t)$ is given by

$$X_{li}(t) = \mu_l(t) + \sum_{r=1}^{\infty} \xi_{ilr}\phi_{lr}(t), \tag{3}$$

where the so-called FPC scores $\{\xi_{ilr}\}_{r=1,\ldots,\infty}$ are uncorrelated random variables with mean zero and variance λ_{lr}. It is, we have that $E(\xi_{ilr}) = 0$, for all $r \in \mathbb{N}$, as well as $E(\xi_{ilr_1}\xi_{ilr_2}) = \mathbf{1}(r_1 = r_2)\lambda_{lr_1}$, for all $r_1, r_2 \in \mathbb{N}$. Moreover, we can think in $\phi_{lr}(t)$ as a changing pattern of the longitudinal process, and the random variable ξ_{ilr} is describing how strongly the data from the ith subject follow this pattern. We assume that the longitudinal process can be approximated by the first R_l eigenfunctions. Thus, the truncated version of the individual trajectory, $X_{li}(t) \approx \mu_l(t) + \sum_{r=1}^{R_l} \xi_{ilr}\phi_{lr}(t)$, will be the quantity used. A prespecified percentage of variance explained (PVE) will be used to determine the number of components R_l. Say PVE $= 80\%, 90\%, 95\%$.

In our framework, observed outcomes of $\boldsymbol{y}_{li}$ are available only at discrete random times t_{ij}, as well as missing for $t > T_i^*$. The FPCA is conducted via principal analysis by conditional estimation (PACE) algorithm [15]. The PACE method has been proven to be versatile and powerful when applied to sparse and irregularly measured longitudinal data contaminated with measurements errors. The PACE algorithm is applied to all the observed data for the lth outcome, generating the following estimated quantities: mean function $\hat{\mu}_l(t)$, error variance $\hat{\sigma}_{\varepsilon_l}$, covariance function $\hat{\Sigma}_l(t, t')$, eigenvalues $\hat{\lambda}_{lr}$, and eigenfunctions $\hat{\phi}_{lr}(t)$. Mean, covariance, and eigenfunctions are assumed to be smooth. Then, local linear smoothers are used for function and surface estimation, fitting local lines in one dimension and local planes in two dimensions by weighted least squares. Firstly, mean function is estimated based on the pooled data from all individuals. Hereinafter, covariance function is estimated by means

of surface estimation. Let $\hat{G}_l(t,t')$ be a smooth surface estimate of $\Sigma(t,t')$. Finally, the estimates of eigenfunctions and eigenvalues correspond to the solutions $\hat{\lambda}_{lr}$ and $\hat{\phi}_{lr}(t)$ of the eigenequations,

$$\int_{\mathcal{T}_l} \hat{G}_l(t,t')\hat{\phi}_{lr}(t)dt = \hat{\lambda}_{lr}\hat{\phi}_{lr}(t'),$$

where the functions $\hat{\phi}_{lr}(t)$ are constraint to $\int_{\mathcal{T}_l} \hat{\phi}_{lr}(t)^2dt = 1$ and $\int_{\mathcal{T}_l} \hat{\phi}_{lr_1}(t)\hat{\phi}_{lr_2}(t)dt = 0$ for $r_2 < r_1$. Estimation of eigenfunctions is done by discretizing the smoothed covariance.

Using the quantities above, we can estimate the FPC scores of subject i as follows

$$\hat{\xi}_{ilr} = E\left(\xi_{ilr}|\boldsymbol{y}_{li}\right) = \hat{\lambda}_{lr}(\hat{\boldsymbol{\phi}}_{ilr})^\top \hat{\Sigma}^{-1}_{\boldsymbol{y}_{li}}(\boldsymbol{y}_{li} - \hat{\boldsymbol{\mu}}_{li}) \tag{4}$$

where we have the vectors $\hat{\boldsymbol{\mu}}_{li} = (\hat{\mu}_l(t_{11}), \ldots, \hat{\mu}_l(t_{1n_{li}}))$, $\hat{\boldsymbol{\phi}}_{ilr} = (\hat{\phi}_{lr}(t_{11}), \ldots, \hat{\phi}_{lr}(t_{1n_{li}}))$ and $\hat{\Sigma}_{\boldsymbol{y}_{li}}$ is a $n_{li} \times n_{li}$ matrix with entries

$$\left(\hat{\Sigma}_{\boldsymbol{y}_{li}}\right)_{j,j'} = \hat{\Sigma}_l(t_{ij}, t_{ij'}) + \mathbf{1}(j = j')\hat{\sigma}^2_{\varepsilon_l}.$$

We apply the PACE algorithm on all L longitudinal outcomes and estimate the eigenfunctions $\hat{\boldsymbol{\phi}}(t) = (\hat{\phi}_{l1}(t), \ldots, \hat{\phi}_{lR_l}(t))$ and the FPC scores $\hat{\boldsymbol{\xi}}_{li} = (\hat{\xi}_{il1}, \ldots, \hat{\xi}_{ilR_l})$, where R_l is a proper truncation for the lth longitudinal outcome. Let us denote the vector of estimated FPC scores across all longitudinal outcomes for the ith subject as $\hat{\boldsymbol{\xi}}_i = (\hat{\boldsymbol{\xi}}_{1i}, \ldots, \hat{\boldsymbol{\xi}}_{li})$, the length of the previous vector is denoted as $R_+ = \sum_{l=1}^{L} R_l$.

Since each longitudinal outcome is associated with the disease progression, there may exist nonnegligible correlation among the FPC scores derived from multiple longitudinal outcomes. In the second step of the procedure used, MFPCA implicitly models the correlations among outcomes by means of the correlations among the FPC scores.

Now, we consider the $R_+ \times R_+$ matrix $\boldsymbol{Z}$, consisting of $R_j \times R_k$ blocks $\boldsymbol{Z^{(jk)}}$ with entries

$$Z^{(jk)}_{mn} = Cov\left(\xi_{ijm}, \xi_{ikn}\right), \quad m = 1, \ldots, R_j,\ n = 1, \ldots, R_k,\ j, k = 1 \ldots, L.$$

Let V be an $n \times R_+$ matrix such that whose ith row is $\hat{\boldsymbol{\xi}}_i^\top$. That is, the i-th row is $(\hat{\xi}_{i11}, \ldots, \hat{\xi}_{i1R_1}, \ldots, \hat{\xi}_{iL1}, \ldots, \hat{\xi}_{iLR_L})$. Then, an estimate of the block matrix $\boldsymbol{Z}$ is given by $\hat{Z} = (n-1)^{-1}V^\top V$. A matrix eigenanalysis is performed on $\hat{Z}$, resulting into estimated eigenvalues $\{\hat{\nu}_k\}_{k=1,\ldots,R_+}$ and orthonormal eigenvectors $\{\hat{\boldsymbol{c}}_k\}_{k=1,\ldots,R_+}$. Estimates for the multivariate eigenfunctions for the lth outcome are given by

$$\hat{\Psi}_{lk}(t) = \sum_{r=1}^{R_l} [\hat{\boldsymbol{c}}_k]_r^{(l)} \hat{\phi}_{lr}(t),$$

where $[\hat{\boldsymbol{c}}_k]_r^{(l)}$ is denoting the lth block of the orthonormal eigenvector $\hat{\boldsymbol{c}}_k$. The set of multivariate eigenfunctions $\hat{\boldsymbol{\Psi}}_k = \{\hat{\Psi}_{lk}(t)\}_{k=1,\dots,R_+}$ characterizes the kth changing pattern of the multivariate longitudinal processes $\boldsymbol{X}_i(t)$. Estimates for the MFPC scores of subject i can be computed as

$$\hat{\rho}_{ik} = \sum_{l=1}^{L} \sum_{r=1}^{R_l} [\hat{\boldsymbol{c}}_k]_r^{(l)} \hat{\xi}_{ilr} = V_{i\cdot} \hat{c}_k, \tag{5}$$

where $V_{i\cdot}$ denotes the i-th row of the matrix V. Lastly, the lth longitudinal outcome, $X_{li}(t)$, can be sufficiently approximated by selecting the first $R^* \leq R_+$ scores and eigenfunctions based on PVE, and computing

$$E(y_{li}(t)) = \hat{X}_{li}(t) \approx \hat{\mu}_l(t) + \sum_{k=1}^{R^*} \hat{\rho}_{ik} \hat{\Psi}_{lk}(t). \tag{6}$$

In the two-step procedure explained above, we have taken profit of the relationship between univariate and multivariate FPCA, stated and proven in [3] as said before. In particular, in the first step we are applying univariate FPCA for each longitudinal outcome. In the second step, we are using the aforementioned relationship to estimate the scores as well as the other components involved in multivariate FPCA.

Implementation of MFPCA can be found in the R package `MFPCA` [4]. An important observation is that the current implementation requires the longitudinal data to lie on a grid with fixed intervals. Nonetheless, missing data is allowed. A consequence of using a fixed grid, is that we lose flexibility compared with our initial framework. When using FPCA-based methods in this work, the observed longitudinal data $y_{li}(t_{ij})$ is first rounded to the nearest time corresponding to the fixed grid. A $n \times J \times L$ input matrix will be used to deal with the longitudinal data, where J is the grid length needed to accommodate the largest observation time.

Having shown how MFPCA can be used to approximate longitudinal data, now we want to show how weights can be used in order to account for missing data. Different articles in the literature have attempted to address the challenge of missing data with FPCA by making use of weights [14]. Nonetheless, when one looks at the application of FPCA in the literature, one rarely sees the use of these weights in practice. We think that the proper using of weights can be useful to account for missing data.

We can use weights in two levels when working under the MFPCA framework. In the first level, focusing in each longitudinal outcome separately, we can use as subject-specific weights $w_{li} = n_{li}^{-1}$, when estimating the mean as well as the covariance function via local linear smoothers. That is, we use the inverse of the number of observation points as weights. Notice that these weights would be used in the first step of the two step procedure shown above. Thus, this weighting is used when doing univariate FPCA for each longitudinal outcome.

In the second level, when performing the second step of the procedure, we can use weights to account for different domains, ranges or amounts of variation between the L different longitudinal outcomes. These weights are a generalization of the framework exposed above, and are explained in [3]. In this extension, MFPCA is based on a weighted scalar product. Given weights $w_1, \ldots, w_L > 0$, the matrix eigenanalysis is now performed in $\hat{Z}_w = (n-1)^{-1} D V^\top V D$, where $D = diag(w_1^{1/2}, \ldots, w_1^{1/2}, \ldots, w_L^{1/2}, \ldots, w_L^{1/2})$ is a $R_+ \times R_+$ matrix in which we have the weights $w_l^{1/2}$ repeated R_l times in the diagonal for each $l = 1, \ldots, L$. Thus, we have indeed, $\sum_{l=1}^{L} R_l = R_+$ elements in the diagonal. In the above framework, in which we did not take into account the weights, $D = Id$ (the $R_+ \times R_+$ identity matrix). If $\hat{\boldsymbol{c}}_k$ and $\hat{\nu}_k$ are the eigenvectors and eigenvalues, respectively, obtained from the new matrix eigenanalysis. Then, the estimated orthonormal eigenfunctions $\hat{\Psi}_{lk}$ and associated scores $\hat{\rho}_{ik}$ can be computed as

$$\hat{\Psi}_{lk}(t_l) = \left(w_l \hat{\nu}_k \hat{\boldsymbol{c}}_k^\top \hat{Z}_w \hat{\boldsymbol{c}}_k\right)^{-1/2} \sum_{l=1}^{L} \sum_{r_1=1}^{R_l} \sum_{r_2=1}^{R_m} [\hat{Z}_w]_{r_1 r_2}^{(lm)} [\hat{\boldsymbol{c}}_k]_{r_2}^{(m)},$$

$$\hat{\rho}_{ik} = (\hat{\nu})^{1/2} \left(\hat{\boldsymbol{c}}_k^\top \hat{Z}_w \hat{\boldsymbol{c}}_k\right)^{-1/2} V_{i\cdot} D \hat{\boldsymbol{c}}_k.$$

Observe that this extension allow as to use weights accounting for the size of the intervals in which we observe our longitudinal data. Taking into account the difference between these intervals can allow us to make adjustments that will have a positive impact on our MFPCA. Moreover, if we suspect that some longitudinal outcome will give different results from the rest due to possible missing data, we can deal with this through these weights.

Simulation study

A simulation study has been executed to evaluate the performances of MFPCA to analyze longitudinal data in presence of missing data. In this section we aim to 1) perform a comparison of with predicted values via MFPCA with those obtained by mixed models, 2) evaluate the robustness of MFPCA estimations to dropout. A similar simulation study has been carried out in [12]. However, in [12] only one longitudinal outcome is taken into account. Therefore, this work can be seen as a generalization to the multivariate case.

- **Comparison between MFPCA and MM**

Given that in this work we are interested in dynamic predictions, it is important to assess whether predictions derived using MFPCA may be affected by the presence of missing data. Moreover, we want to compare the predictive power of MFPCA with that of MM (considered the gold standard).

Six different missing data scenarios have been considered. For all these scenarios the following simulation specifications are common. We consider $L = 3$ longitudinal processes, all of them are generated using linear mixed models. Training and testing data have been independently simulated, both containing 200 subjects. The visit process was generated from a fixed grid with a unit of distance between visits that go from $t = 0$ to $t = 9$, with a random uniform noise around the theoretical visits, except for $t = 0$ which is fixed for all subjects. This is, for the ith individual we have the observation times (t_{ij}): $(0, 1+u(-1/2, 1/2), 2+u(-1/2, 1/2), \ldots, 9+u(-1/2, 1/2))$. Where $u(a, b)$ are numbers generated from a continuous uniform distribution on the interval (a, b). This uniform noise add some irregularity, making the observation times specific for each individual, which we believe reproduces a longitudinal study situation in a realistic way. Two binary variables (say sex and treatment) were simulated and added as a baselines covariates. Let us show the model specification for the three longitudinal outcomes simulated. The linear mixed models used are:

$$\begin{aligned}
y_{1i}(t_{ij}) &= m_{1i}(t_{ij}) + \varepsilon_{1i}(t_{ij}) \\
&= (\beta_0^1 + b_{0i}^1) + (\beta_1^1 + b_{1i})t_{ij} + \beta_2^1 \texttt{sex}_i + \beta_3^1 \texttt{sex}_i t_{ij} + \varepsilon_{1i}(t_{ij}), \\
y_{2i}(t_{ij}) &= m_{2i}(t_{ij}) + \varepsilon_{2i}(t_{ij}) \\
&= (\beta_0^2 + b_{0i}^2) + (\beta_1^2 + b_{1i})\sqrt{t_{ij}} + \varepsilon_{2i}(t_{ij}), \\
y_{3i}(t_{ij}) &= m_{i3}(t_{ij}) + \varepsilon_{3i}(t_{ij}) \\
&= (\beta_0^3 + b_{0i}^3) + (\beta_1^3 + b_{1i}^3)t_{ij} + \beta_2^3 \texttt{treatment}_i + \varepsilon_{3i}(t_{ij}).
\end{aligned}$$

Where the super-indices in the parameters denote the number of the longitudinal model, $\texttt{sex}_i$ denotes a binary sex indicator and $\texttt{treatment}_i$ denotes a binary treatment indicator. The random effects were assumed to follow a multivariate normal distribution with mean zero and the following covariance matrix:

$$D = \Sigma\Sigma^t,$$

where Σ is a 6×6 matrix.

Then, dropout has been simulated under different scenarios. We considered dropout with two different intensities: 30% and 50%. Where 30% (idem 50%) of intensity means 30% of longitudinal data is missing because of dropout. The six missing data scenarios used are the following:

- MCAR: ith subject drops out at t_{ij} with a probability determined by a logistic model with t_{ij} as predictor.
- Fixed MAR: ith subject drops out at t_{ij} if $y_{1i}(t_{i(j-1)}) > \nu_1$ or $y_{2i}(t_{i(j-1)}) > \nu_2$ or $y_{3i}(t_{i(j-1)}) < \nu_3$.
- Threshold MAR: ith subject drops out at t_{ij} with a probabilities determined by three logistic models each one with the indicators $y_{1i}(t_{i(j-1)}) > \nu_1$, $y_{2i}(t_{i(j-1)}) > \nu_2$, $y_{3i}(t_{i(j-1)}) < \nu_3$ as predictors.
- Increasing MAR: ith subject drops out at t_{ij} with a probabilities determined by three logistic models each one with $y_{1i}(t_{i(j-1)})$, $y_{2i}(t_{i(j-1)})$, $y_{3i}(t_{i(j-1)})$ as predictors.
- Threshold MNAR: ith subject drops out at t_{ij} with a probabilities determined by three logistic models each one with the indicators $y_{1i}(t_{ij}) > \nu_1$, $y_{2i}(t_{ij}) > \nu_2$, $y_{3i}(t_{ij}) < \nu_3$ as predictors.
- Increasing MNAR: ith subject drops out at t_{ij} with a probabilities determined by three logistic models each one with $y_{1i}(t_{ij})$, $y_{2i}(t_{ij})$, $y_{3i}(t_{ij})$ as predictors.

The generated longitudinal data of 100 randomly selected individuals are shown in Figure 1. In the left part of Figure 1, the complete longitudinal profiles of those 100 individuals are shown. Observe that we have $L = 3$ different longitudinal outcomes. In the right panel of the graphic we have the incomplete longitudinal profiles. This is, longitudinal profiles under different missing data scenarios.

Predictions for the longitudinal values in both, training and testing data, have been computed from the estimations of MFPCA and MM on the training set. Predictions of interest are those for the missing longitudinal values due dropout (true values are known from the simulation procedure). To compute predictions we have used the true MM (the one used to generate the data) as well as MFPCA. By doing this, we are comparing predictions derived with MFPCA with predictions with the best-case scenario, which is the true MM model. Regarding the MFPCA, 80% of PVE has been used in those scenarios with 30% of dropout intensity, and 65% of PVE with 50% of dropout intensity.

To fit the MM, Bayesian models have been used. Specifically, we have used the `rstanarm` R package [2] (function `stan_mvmer`). There are several reasons why we have used the Bayesian approach to fit these models. First, because it is in line with the rest of the work. And, in addition, because this type of models are not easy to fit under the frequentist paradigm, and there are no packages that do it for the multivariate case. Nonetheless, fitting Bayesian models increases the computational time. Regarding MFPCA, we have used `MFPCA` R package [4]. In order to derive predictions with this package, the observation times must lie in a fixed grid. Thus, we have had to create a fixed grid of points from 0 to 9 with a distance of 0.5 between points, and all observation times have been approximated to the nearest point on

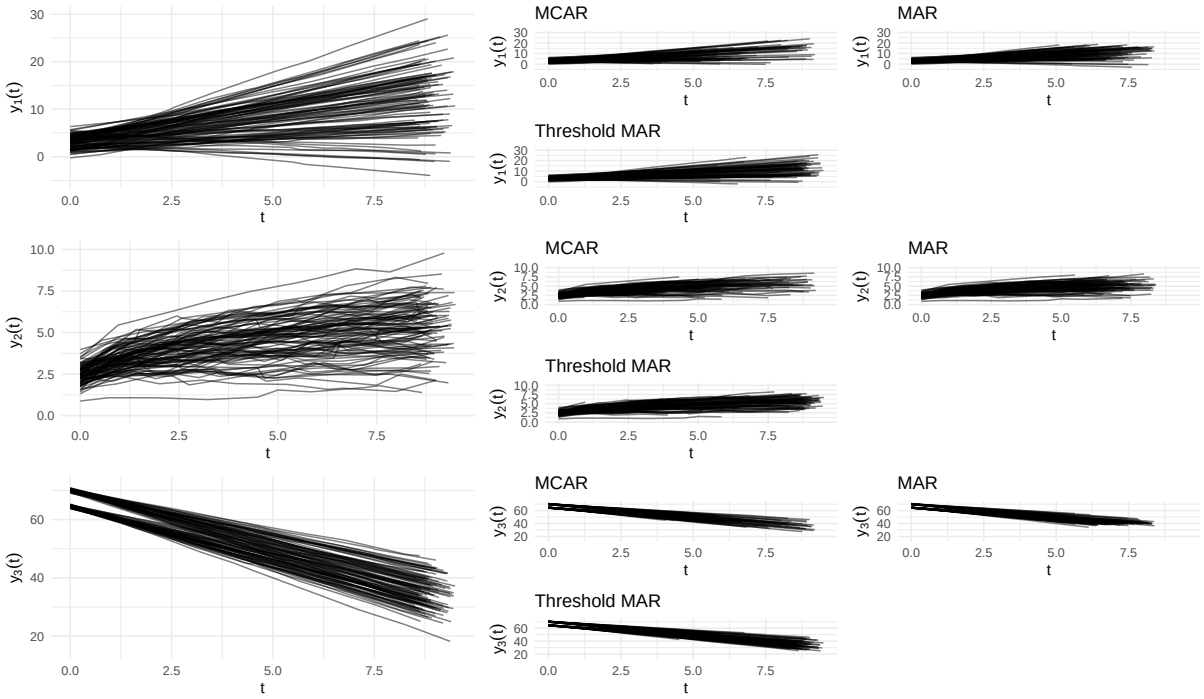

Figure 1: Generated individual longitudinal data, for 100 randomly selected subjects, by means of a multivariate Bayesian MM. In the left panel three plots with the complete longitudinal profiles. In the right panel longitudinal profiles after dropout for the different scenarios (30% of dropout intensity).

the grid.

For each scenario, the procedure was replicated on 100 different simulated data sets. The performance were evaluated by means of the root mean square error (RMSE). Hence, RMSE computed on the missing data of the training and testing set will be reported. Results are shown in Figure 2, as well as Figure 3.

Focusing first on the results for those scenarios with a 30% dropout rate, shown in Figure 2, we can see that boxplots for the RMSEs of MFPCA predictions present much more variability than boxpolots for the data-generating MM. However, we can see a similar trend regarding MFPCA-related boxplots across the different missing data scenarios simulated. This is a sign that the MFPCA predictions are robust under MAR. In addition, we can observe that predictions via MFPCA are similar to those predictions computed by means of the data-generating MM, since RMSEs seems to be similar. Moreover, while the data-generating MM is affected by over-fitting and RMSEs are bigger in testing data than in training data, it seems that for MFPCA-based predictions account fairly good for over-fitting. Having said that, we have to keep in mind that we are comparing MFPCA-based predictions with those

predictions derived from the data-generating mixed model. Thus, of course, we expect the MM to rank systematically as the best one for every missing data scenario. Nonetheless, our main interest in this section is to assess how good the predictions based on MFPCA can be, and how these predictions behave under different missing data scenarios. Certainly, in practice, the true model is unknown, and so we expect MFPCA-based predictions to be even better when comparing with some candidate MM. Another important comment is that MFPCA-based methods are non-parametric, and we do not need to design the appropriate model expression to obtain these results. The only thing we have to choose is the PVE, so that we can calculate the appropriate number of scores with the data at hand.

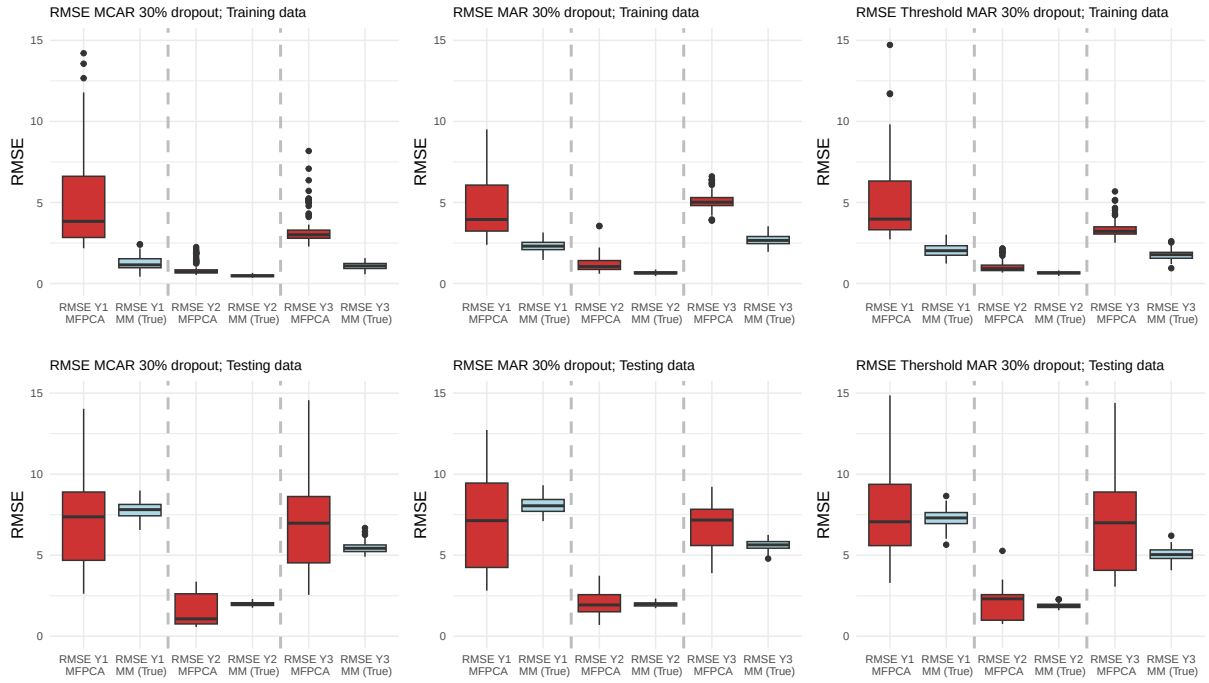

Figure 2: RMSEs computing on missing training data (three graphics at the top) and testing data (three graphics at the bottom). All scenarios with 30% rate of dropout.

Regarding those scenarios with 50% of dropout, results are reported in Figure 3. We can observe in this case, contrary to what we would expect, that the variability of the RMSEs with respect to MFPCA-based predictions is lower compared to that of the scenarios with 30% dropout. Moreover, we again observe a similar trend in the results between the different missing data scenarios. This seems to indicate that the predictions calculated via MFPCA are robust under MAR scenario. In addition to this, the conclusions we can draw are similar to those already mentioned above.

- **Further comments and things to do:**

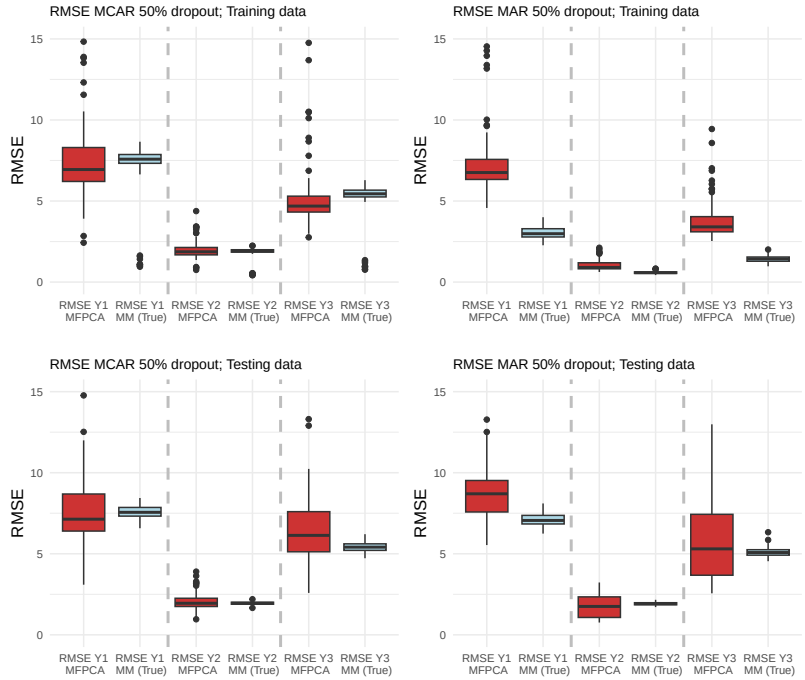

Figure 3: RMSEs computing on missing training data (three graphics at the top) and testing data (three graphics at the bottom). All scenarios with 50% rate of dropout.

There are several items that are pending in this work. Throughout the work some parts are written in red color. This is because there are pending things to do related to that. Let us comment it briefly:

1. Finish the simulations: we have to finish the simulations of all the missing data scenarios. Since we are using Bayesian multivariate MM, simulations are computationally expensive, and take some time.

2. Also a very important part of the simulation study we have planned is to evaluate the robustness of MFPCA estimations to dropout. To do this we will compute the principal components via MFPCA under different missing data scenarios. These simulations should not take so much time, since we would not fit any Bayesian multivariate MM, but only MFPCA (which are less computationally expensive).

3. Another item to be explored is the usage of weights to account for missing data, and see whether these weights can improve predictions based on MFPCA.

References

[1] Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.

[2] Ben Goodrich, Jonah Gabry, Imad Ali, and Sam Brilleman. rstanarm: Bayesian applied regression modeling via Stan., 2024. R package version 2.32.1.

[3] Clara Happ and Sonja Greven. Multivariate functional principal component analysis for data observed on different (dimensional) domains. *Journal of the American Statistical Association*, 113(522):649–659, 2018.

[4] Clara Happ-Kurz. *MFPCA: Multivariate Functional Principal Component Analysis for Data Observed on Different Dimensional Domains*, 2022. R package version 1.3-10.

[5] Nan M Laird and James H Ware. Random-effects models for longitudinal data. *Biometrics*, pages 963–974, 1982.

[6] Kan Li and Sheng Luo. Dynamic prediction of alzheimer's disease progression using features of multiple longitudinal outcomes and time-to-event data. *Statistics in medicine*, 38(24):4804–4818, 2019.

[7] Jeffrey Lin and Sheng Luo. Deep learning for the dynamic prediction of multivariate longitudinal and survival data. *Statistics in medicine*, 41(15):2894–2907, 2022.

[8] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

[9] Geert Molenberghs and Michael Kenward. *Missing data in clinical studies*. John Wiley & Sons, 2007.

[10] Dimitris Rizopoulos. *Joint models for longitudinal and time-to-event data: With applications in R*. CRC press, 2012.

[11] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[12] Corentin Ségalas, Catherine Helmer, Robin Genuer, and Cécile Proust-Lima. Functional principal component analysis as an alternative to mixed-effect models for describing sparse repeated measures in presence of missing data. *Statistics in Medicine*, 43(26):4899–4912, 2024.

[13] Geert Verbeke, Steffen Fieuws, Geert Molenberghs, and Marie Davidian. The analysis of multivariate longitudinal data: a review. *Statistical methods in medical research*, 23(1):42–59, 2014.

[14] Caleb Weaver, Luo Xiao, and Wenbin Lu. Functional data analysis for longitudinal data with informative observation times. *Biometrics*, 79(2):722–733, 2023.

[15] Fang Yao, Hans-Georg Müller, and Jane-Ling Wang. Functional data analysis for sparse longitudinal data. *Journal of the American statistical association*, 100(470):577–590, 2005.