A functional multiple imputation approach to incomplete longitudinal data

Yulei He,[a][*][†] Recai Yucel[b] and Trivellore E. Raghunathan[c]

In designed longitudinal studies, information from the same set of subjects are collected repeatedly over time. The longitudinal measurements are often subject to missing data which impose an analytic challenge. We propose a functional multiple imputation approach modeling longitudinal response profiles as smooth curves of time under a functional mixed effects model. We develop a Gibbs sampling algorithm to draw model parameters and imputations for missing values, using a blocking technique for an increased computational efficiency. In an illustrative example, we apply a multiple imputation analysis to data from the Panel Study of Income Dynamics and the Child Development Supplement to investigate the gradient effect of family income on children's health status. Our simulation study demonstrates that this approach performs well under varying modeling assumptions on the time trajectory functions and missingness patterns. Copyright © 2011 John Wiley & Sons, Ltd.

Keywords: cubic smoothing spline; functional data analysis; Gibbs sampling; missing data; multiple imputation; Panel Study of Income Dynamics

1. Introduction

In designed longitudinal studies, information from the same set of subjects are measured repeatedly over time. The aim of such studies is to frequently estimate the mean or individual response at a certain time, to relate time-invariant or time-dependent covariates to repeatedly measure response variables, or to relate the response variables to each other. Missing data often occur in longitudinal measurements because subjects may miss visits during the study, because some variables may not be measured at particular visits, or because subjects may drop out prematurely. The absence of complete data is a serious impediment to longitudinal data analysis.

The commonly used complete-case or available-case analysis (AC) is based on observed data without any imputation or other adjustments. They are well known to be inefficient and may give biased results if data are not missing completely at random [1]. Other naive missing data approaches include the mean imputation method, which imputes a missing value at a specific time by the mean value of the observed sample at that time, and the last observed value carried forward method, which imputes a missing value at a specific time by the observed value immediately before that time. These two imputation methods, although simple to implement, do not adequately reflect the temporal pattern as well as the relationship among response variables and therefore may produce misleading results.

More principled approaches include the likelihood-based methods, inverse probability weighted estimation equations and multiple imputation. See Discussion (Section 6) for a general comment on the connection and differences among three approaches. In this paper, we focus on the multiple imputation approach to incomplete longitudinal data where the repeated measurements are subject to missing values. In multiple imputation, missing data are 'filled-in' by several plausible sets of values to generate

[a] *Department of Health Care Policy, Harvard Medical School, Boston, MA 02115, U.S.A.*
[b] *Department of Epidemiology and Biostatistics, University of Albany School of Public Health, Rensselaer, NY 12144, U.S.A.*
[c] *Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI 48109, U.S.A.*
[*] *Correspondence to: Yulei He, Department of Health Care Policy, Harvard Medical School, Boston, MA 02115, U.S.A.*
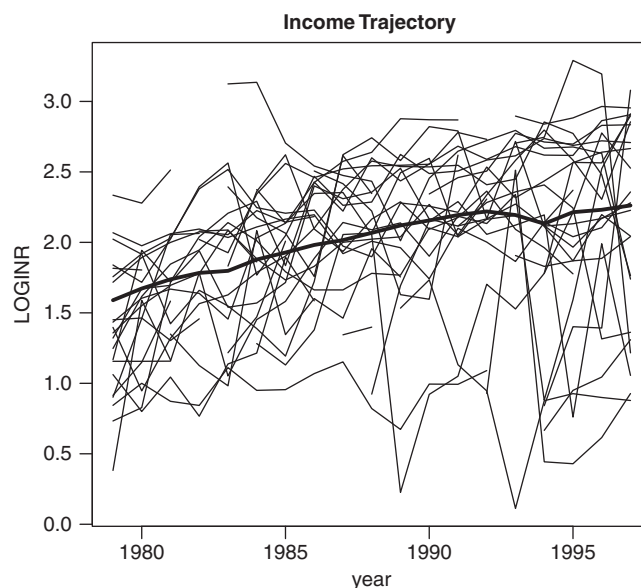[†] *E-mail: he@hcp.med.harvard.edu*

Figure 1. Log-transformed income to needs ratio data (LOGINR) from 1979 to 1997 of PSID; thick line: average profile of a sample of $n=2172$ families; thin lines: income series from 25 randomly selected families.

completed datasets. The analysis results obtained from each completed dataset are combined in a way to account for imputation uncertainty and form a final inference [2].

Multiple imputation is very popular among practitioners because of its great flexibility. The analysts, who generally do not have special statistical expertise required by alternative missing data methods (e.g. likelihood-based methods), can readily analyze the completed datasets using existing statistical procedures or routines. In addition, the analysis procedures applied to the imputed data can be incompatible (uncongenial) with the statistical model used to create imputations and still lead to practically adequate multiple imputation inferences in many cases [3], allowing more room for applying a variety of imputation analyses. Further, the increasing availability of imputation software makes it relatively easy to generate imputations under some well-developed models [4].

We assume missingness at random (MAR) for the incomplete longitudinal data, that is, the probability of missingness is not related to unobserved values conditioning on observed data [2]. Under MAR, multiple imputation approaches have been developed based on the linear mixed model and its multivariate extensions for longitudinal data [5, 6], and can be implemented using the library 'pan' in R package (www.r-project.org). In these methods, however, the growth curves of the response variables are limited to linear/quadratic or other simple parametric functions of time. This assumption may not be general enough to characterize the temporal trends exhibited by some real data. For example, Figure 1 plots the (log-transformed) annual records of family income to needs ratio over a 19-year period from the Panel Study of Income Dynamics (PSID) (Section 4). The mean trajectory (thick line) appears to be close to linear over time, while the individual series (thin line) appear to show more curvature. The main focus of the analysis is to relate family income status at different time spans to children's health status. Given the incomplete longitudinal data, it is of practical interest to develop a multiple imputation approach which appropriately incorporates the temporal patterns as well as covariate information. Certainly, a less restrictive assumption on the time functions might be more desired than imposing some parametric assumptions, which might be incorrect.

In our motivating example from PSID, one plausible strategy is to treat the income series as smooth curves and estimate the missing values using nonparametric methods. This is in lieu of methods used in functional data analysis [7] in which the basic unit of analysis is a curve. Some of the common applications of functional data analysis involve curve-type longitudinal data such as height growth curves, hormone profiles and CD4 counts over time [8–10]. For instance, the authors of [11] proposed a class of functional mixed effect models that can be viewed as extensions of linear mixed models for the curve data, estimating the response profiles using cubic smoothing splines. Methods in functional data analysis are also highly relevant to semiparametric or nonlinear regression [12] and smoothing techniques.

Despite the vibrant research on functional data analysis, little attention has been paid to its application on imputation and missing data analysis. In this paper, we explore this area by developing a multiple imputation approach for missing longitudinal responses using the functional mixed models. Our goal is to create imputations that retain the correct temporal trends using a data-driven strategy. The rest of this paper is organized as follows. Section 2 describes the functional mixed model. Section 3 develops a Gibbs sampling algorithm for drawing missing values and model parameters. Section 4 describes the example which motivates this study and illustrates the application. In Section 5, we conduct some simulation studies to evaluate the performance of the proposed approach. Finally, Section 6 concludes this paper with a discussion.

2. Functional mixed model

2.1. General setup

Let y_{ij} be the response of subject i at time t_{ij} ($i=1,\ldots,m$, $j=1,\ldots,n_i$), and we assume y_{ij} can be subject to missing values. The functional mixed model [11] takes the following form

$$y_{ij} = X_{ij}\beta(t_{ij}) + Z_{ij}\alpha_i(t_{ij}) + \varepsilon_{ij}, \varepsilon_{ij} \sim \mathrm{N}(0, \sigma^2), \tag{1}$$

where $X_{ij}=(x_{ij1},\ldots,x_{ijp})(1\times p)$ and $Z_{ij}=(z_{ij1},\ldots,z_{ijq})(1\times q)$ are design matrices, $\beta(t)=(\beta_1(t),\ldots,\beta_p(t))^{\mathrm{T}}$ is a $p\times 1$ vector of fixed functions, $\alpha_i(t)=(\alpha_{i1}(t),\ldots,\alpha_{iq}(t))^{\mathrm{T}}$ is a $q\times 1$ vector of random functions that are modelled as realizations of Gaussian processes $A(t)=(a_1(t),\ldots,a_q(t))^{\mathrm{T}}$(collection of processes) with zero means; and ε_{ij} is the measurement error. For simplicity, here we assume both X's and Z's are fully observed. The problem of handling missing X's and Z's is briefly addressed in Section 6.

The functional mixed model (1) can be regarded as a generalization of the linear mixed model by treating both the fixed and random effects as smooth curves of t. Similar to the interpretation of the linear mixed model, $\beta(t)$ can be interpreted as population-average profiles, $Z_{ij}\alpha_i(t)$ as the ith subject-specific deviation from the population-average curves, and hence $X_{ij}\beta(t)+Z_{ij}\alpha_i(t)$ as the ith subject-specific function. As pointed out by [8], a variety of longitudinal curve models can be viewed as special cases of the functional mixed model. For example, if both $\beta(t)$'s and $\alpha_i(t)$'s are modeled as vector of polynomials, model (1) is reduced to the random coefficients model [13].

2.2. Estimation using cubic smoothing splines

Guo [11] proposed an estimation procedure for model (1) using cubic smoothing splines. This approach exploits the connection between cubic smoothing splines and linear mixed model. This feature can be traced back to [14], which showed that the estimate of a smoothing spline can be obtained through the posterior estimate of a Gaussian stochastic process. Following Wahba's [14] Bayesian approach, $\beta_k(t)$'s and $\alpha_{il}(t)$'s are modeled as

$$\beta_k(t) = B_{k1} + B_{k2}t + \lambda_{bk}^{-1/2}\int_0^t W_k(s)\,\mathrm{d}s, \quad k=1,\ldots,p, \tag{2}$$

$$\alpha_{il}(t) = A_{il1} + A_{il2}t + \lambda_{al}^{-1/2}\int_0^t W_{il}(s)\,\mathrm{d}s, \quad i=1,\ldots,m, \quad l=1,\ldots,q, \tag{3}$$

where B_{k1} and B_{k2} have diffuse priors and are treated as fixed in the terminology of linear mixed model, i.e. $((B_{k1}, B_{k2})^{\mathrm{T}} \sim \mathrm{N}(\mathbf{0}, \tau\mathbf{I}_2)$ with $\tau\to\infty$), A_{il1} and A_{il2} are random intercepts and slopes, i.e. $(A_{il1}, A_{il2})^{\mathrm{T}} \sim i.i.d.\mathrm{N}(\mathbf{0}, \mathbf{\Omega}_l)$ for $i=1,\ldots,m$ and $l=1,\ldots,q$, $W_k(s)$'s and $W_{il}(s)$'s are independent Weiner processes, and λ_{bk}'s and λ_{al}'s are smoothing parameters controlling the balance between smoothness and 'goodness of fit' for $\beta_k(t)$'s and $\alpha_{il}(t)$'s, respectively. Therefore, $\lambda_{bk}^{-1/2}\int_0^t W_k(s)\mathrm{d}s$ and $\lambda_{al}^{-1/2}\int_0^t W_{il}(s)\mathrm{d}s$ model the nonlinear curvature of $\beta_k(t)$ and $\alpha_{il}(t)$'s, respectively.

For simplicity of notations, we assume that each subject has equal number of time points, i.e. $n_1=\cdots=n_m=n$, and $\mathbf{t}=(t_1,\ldots,t_n)^{\mathrm{T}}$. More generally for unbalanced designs, $\mathbf{t}$ is the collection of distinctive time points across different subjects. As shown by Green and Zhang et al. [15, 16], the

estimation of $\beta_k(t)$'s and $\alpha_{il}(t)$'s at the design points $\mathbf{t}$ can be obtained through the following linear mixed models:

$$\beta_k(\mathbf{t}) = \mathbf{T}\zeta_k + \mathbf{B}\mathbf{a}_k, \zeta_k = (B_{k1}, B_{k2})^\mathrm{T}, \mathbf{a}_k \sim \mathrm{N}\left(\mathbf{0}, \frac{1}{\lambda_{bk}}\mathbf{I}_{n-2}\right), \tag{4}$$

$$\alpha_{il}(\mathbf{t}) = \mathbf{T}\zeta_{il} + \mathbf{B}\mathbf{a}_{il}, \zeta_{il} = (A_{il1}, A_{il2})^\mathrm{T} \sim \mathrm{N}(\mathbf{0}, \mathbf{\Omega}_l), \mathbf{a}_{il} \sim \mathrm{N}\left(\mathbf{0}, \frac{1}{\lambda_{al}}\mathbf{I}_{n-2}\right), \tag{5}$$

where $\beta_k(\mathbf{t}) = (\beta_k(t_1), \ldots, \beta_k(t_n))^\mathrm{T}$, $\alpha_{il}(\mathbf{t}) = (\alpha_{il}(t_1), \ldots, \alpha_{il}(t_n))^\mathrm{T}$, $\mathbf{T} = \{\mathbf{1}, \mathbf{t}\}$, and $\mathbf{B}$ is an $n \times (n-2)$ matrix which satisfies $\mathbf{B}\mathbf{B}^\mathrm{T} = \mathbf{R}^1$, where $\mathbf{R}^1$ is the integrated Weiner covariance matrix evaluated at $\mathbf{t}$ [14]. In mixed models (4) and (5), ζ_k models the fixed intercept and slope for function $\beta_k(t)$, ζ_{il} models the random intercept and slope for function $\alpha_{il}(t)$ of subject i. Both $\mathbf{a}_k$ and $\mathbf{a}_{il}$ are $(n-2) \times 1$ vector of random effects so that $\mathbf{B}\mathbf{a}_k$ and $\mathbf{B}\mathbf{a}_{il}$ model the departure of $\beta_k(t)$ and $\alpha_{il}(t)$ from the straight lines, respectively. The inverse of smoothing parameters, $1/\lambda_{bk}$ and $1/\lambda_{al}$, are the variance components of the mixed models (4) and (5), respectively. When λ's $\rightarrow \infty$, the estimated functions are close to linear curves. On the other hand, when λ's $\rightarrow 0$, the estimated functions tend to interpolate all the data points. Note that parameterizations of the linear mixed model format of cubic smoothing splines may appear slightly different in different literature. For example, the associated random effects are modeled as vectors of dimension n in [11]. However, it can be easily shown that the estimates of smooth functions remain the same under those different parameterizations.

Substituting equations (4) and (5) into model (1), we obtain a linear mixed model representation of the functional mixed model

$$\mathbf{Y}_i = \widetilde{\mathbf{X}}_i^\mathbf{T}\widetilde{\boldsymbol{\beta}} + \widetilde{\mathbf{X}}_i^\mathbf{B}\widetilde{\mathbf{b}} + \widetilde{\mathbf{Z}}_i^\mathbf{T}\widetilde{\boldsymbol{\beta}}_i + \widetilde{\mathbf{Z}}_i^\mathbf{B}\widetilde{\mathbf{b}}_i + \epsilon_i, \epsilon_i \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I}_n), \quad i = 1, \ldots, m, \tag{6}$$

where

$$\mathbf{Y}_i = (y_{i1}, \ldots, y_{in})^\mathrm{T},$$

$$\widetilde{\mathbf{X}}_i^\mathbf{T} = (\overline{\mathbf{X}}_{i1}\mathbf{T}, \ldots, \overline{\mathbf{X}}_{ip}\mathbf{T}), \widetilde{\mathbf{X}}_i^\mathbf{B} = (\overline{\mathbf{X}}_{i1}\mathbf{B}, \ldots, \overline{\mathbf{X}}_{ip}\mathbf{B}) \quad \text{with } \overline{\mathbf{X}}_{ik} = \mathrm{diag}\{x_{i1k}, \ldots, x_{ijk}, \ldots, x_{\mathrm{ink}}\},$$

$$\widetilde{\mathbf{Z}}_i^\mathbf{T} = (\overline{\mathbf{Z}}_{i1}\mathbf{T}, \ldots, \overline{\mathbf{Z}}_{iq}\mathbf{T}), \widetilde{\mathbf{Z}}_i^\mathbf{B} = (\overline{\mathbf{Z}}_{i1}\mathbf{B}, \ldots, \overline{\mathbf{Z}}_{iq}\mathbf{B}) \quad \text{with } \overline{\mathbf{Z}}_{il} = \mathrm{diag}\{z_{i1l}, \ldots, z_{ijl}, \ldots, z_{inl}\},$$

$$\widetilde{\boldsymbol{\beta}} = (\zeta_1^\mathrm{T}, \ldots, \zeta_p^\mathrm{T})^\mathrm{T},$$

$$\widetilde{\mathbf{b}} = (\mathbf{a}_1^\mathrm{T}, \ldots, \mathbf{a}_p^\mathrm{T})^\mathrm{T} \sim \mathrm{N}(\mathbf{0}, \mathrm{diag}\{1/\lambda_{b1}, \ldots, 1/\lambda_{bp}\} \otimes \mathbf{I}_{n-2}),$$

$$\widetilde{\boldsymbol{\beta}}_i = (\zeta_{i1}^\mathrm{T}, \ldots, \zeta_{iq}^\mathrm{T})^\mathrm{T} \sim \mathrm{N}(\mathbf{0}, \mathrm{diag}\{\mathbf{\Omega}_1, \ldots, \mathbf{\Omega}q\} \otimes \mathbf{I}_2), \text{ and}$$

$$\widetilde{\mathbf{b}}_i = (\mathbf{a}_{i1}^\mathrm{T}, \ldots, \mathbf{a}_{iq}^\mathrm{T})^\mathrm{T} \sim \mathrm{N}(\mathbf{0}, \mathrm{diag}\{1/\lambda_{a1}, \ldots, 1/\lambda_{aq}\} \otimes \mathbf{I}_{n-2}).$$

Under the linear mixed model format (6), the restrictive maximum likelihood estimation approaches have been proposed [11, 16–20], and the inverse of the smoothing parameters are estimated as variance components. The connection between the cubic smoothing spline and the linear mixed model, therefore, allows fitting of the functional mixed model using existing software, such as SAS PROC MIXED (www.sas.com).

Other than cubic smoothing splines, functional mixed models can also be estimated using kernel smoothers [21]. When data exhibit many local features like peaks or jumps, a more general wavelet-based approach can be used [22, 23]. In this paper, we focus on using the cubic smoothing spline approach to functional mixed models mainly because it provides a continuum of models from a trend linear in time to treating time as a factor (obtained as the smoothing parameter λ's tend to ∞ and 0, respectively). Smoothing splines are also relatively easier to formulate and program and thus might be more accessible to practitioners. However, alternative computational methods under the functional mixed model can also be considered for imputation.

3. Gibbs sampling algorithm

We use the Gibbs sampling technique [24] to obtain the draws of missing values in y_{ij}'s and model parameters from their posterior distributions under model (6). To complete a fully Bayesian

model specification, we now specify the prior distributions of the linear mixed model (6). Denoting IG() as inverse gamma distribution and IW() as inverse Wishart distribution, we impose conjugate prior distributions for the parameters as $\zeta_k \sim N(\mathbf{0}, \boldsymbol{\Sigma}_k)$, $1/\lambda_{bk} \sim IG(\gamma_{bk}/2, \mu_{bk}/2)$, $\boldsymbol{\Omega}_l \sim IW(d_l, c_l R_l)$, $1/\lambda_{al} \sim IG(\gamma_{al}/2, \mu_{al}/2)$, and $\sigma^2 \sim IG(\gamma_0/2, \mu_0/2)$ for $k = 1, \ldots, p$ and $l = 1, \ldots, q$. Values of the hyper-parameters of these prior distributions can be chosen based upon prior knowledge or to give diffuse priors for a more objective inference. We recommend to run the Gibbs sampling with different priors to ensure that the results are not sensitive to the choice.

Because model (6) has multiple random effects with large dimensions, and some of them (e.g. $\mathbf{b}$ and $\mathbf{b}_i$) might be highly correlated *a posteriori*, the default Gibbs sampling algorithm which updates one parameter at a time can lead to slow convergence. We adopt the strategy of de-conditioning [25] by blocking the parameters into groups and then updating them simultaneously to achieve a better convergence. We rewrite model (6) as

$$\mathbf{Y}_i = \sum_{k=1}^{p} \widetilde{\widetilde{\mathbf{X}}}_{ik} \mathbf{U}_k + \sum_{l=1}^{q} \widetilde{\widetilde{\mathbf{Z}}}_{il} \mathbf{U}_{il} + \epsilon_i, \quad i = 1, \ldots, m, \tag{7}$$

where $\widetilde{\widetilde{\mathbf{X}}}_{ik} = (\overline{\mathbf{X}}_{ik}\mathbf{T}, \overline{\mathbf{X}}_{ik}\mathbf{B})$, $\widetilde{\widetilde{\mathbf{Z}}}_{il} = (\overline{\mathbf{Z}}_{il}\mathbf{T}, \overline{\mathbf{Z}}_{il}\mathbf{B})$, $\mathbf{U}_k = (\zeta_k^T, \mathbf{a}_k^T)^T$, and $\mathbf{U}_{il} = (\zeta_{il}^T, \mathbf{a}_{il}^T)^T$. The key of the algorithm is to group fixed and random effect coefficients into blocks as $\mathbf{U}_k$'s and $\mathbf{U}_{il}$'s, which are the coefficients corresponding to the population-average profile and individual profile, and draw them iteratively from their conditional distributions given the remaining model components. The variance components can then be drawn based on draws of $\mathbf{U}_k$'s and $\mathbf{U}_{il}$'s.

With a little abuse of notations, let $[X|Y]$ denote the conditional distribution of X given Y, and denote variance components as the collection of λ's, Ω's and σ^2. From equation (7), it is easy to obtain the following distributions:

(a) the conditional distribution of $\mathbf{Y}_i$ given $\bigcup\{\mathbf{U}_k\}_{k=1,\ldots,p}$ (i.e. the collection of $\mathbf{U}_k$'s) and variance components:

$$\left[\mathbf{Y}_i \,\Big|\, \bigcup\{\mathbf{U}_k\}_{k=1,\ldots,p}, \text{variance components}\right] \sim N\left(\sum_{k=1}^{p} \widetilde{\widetilde{\mathbf{X}}}_{ik}\mathbf{U}_k, \mathbf{V}_i\right),$$

where $\mathbf{V}_i = \sum_{l=1}^{q} \widetilde{\widetilde{\mathbf{Z}}}_{il}\text{diag}\{\boldsymbol{\Omega}_l, 1/\lambda_{al} \otimes \mathbf{I}_{n-2}\}\widetilde{\widetilde{\mathbf{Z}}}_{il}^T + \sigma^2\mathbf{I}_n$,

(b) the conditional distribution of $\mathbf{U}_k$ given variance components

$$[\mathbf{U}_k|\text{variance components}] \sim N(\mathbf{0}, \mathbf{W}_k), \quad \text{where } \mathbf{W}_k = \text{diag}\{\boldsymbol{\Sigma}_k, 1/\lambda_{bk} \otimes \mathbf{I}_{n-2}\}$$

(c) the conditional distribution of $\mathbf{Y}_i - \sum_{k=1}^{p} \widetilde{\widetilde{\mathbf{X}}}_{ik}\mathbf{U}_k$ given $\bigcup\{\mathbf{U}_{il}\}_{l=1,\ldots,q}$ and variance components:

$$\left[\mathbf{Y}_i - \sum_{k=1}^{p} \widetilde{\widetilde{\mathbf{X}}}_{ik}\mathbf{U}_k \,\Big|\, \bigcup\{\mathbf{U}_{il}\}_{l=1,\ldots,q}, \text{variance components}\right] \sim N\left(\sum_{l=1}^{q} \widetilde{\widetilde{\mathbf{Z}}}_{il}\mathbf{U}_{il}, \sigma^2\mathbf{I}_n\right),$$

and

(d) the conditional distribution of $\mathbf{U}_{il}$ given variance components:

$$[\mathbf{U}_{il}|\text{variance components}] \sim N(\mathbf{0}, \boldsymbol{\Lambda}_l), \quad \text{where } \boldsymbol{\Lambda}_l = \text{diag}\{\boldsymbol{\Omega}_l, 1/\lambda_{al} \otimes \mathbf{I}_{n-2}\}$$

From distributions in (a) and (b), it is easy to show that $\mathbf{U}_k$'s ($k = 1, \ldots, p$) can be drawn from:
Step 1:

$$\left[\mathbf{U}_k \,\Big|\, \bigcup\{\mathbf{Y}_i\}_{i=1,\ldots,m}, \bigcup\{\mathbf{U}_s\}_{s=1,\ldots,p,\ s\neq k}, \text{variance components}\right]$$

$$\sim N\left(\left(\sum_{i=1}^{m} \widetilde{\widetilde{\mathbf{X}}}_{ik}^T\mathbf{V}_i^{-1}\widetilde{\widetilde{\mathbf{X}}}_{ik} + \mathbf{W}_k^{-1}\right)^{-1}\left(\sum_{i=1}^{m} \widetilde{\widetilde{\mathbf{X}}}_{ik}^T\mathbf{V}_i^{-1}\left(\mathbf{Y}_i - \sum_{s\neq k} \widetilde{\widetilde{\mathbf{X}}}_{is}\mathbf{U}_s\right)\right), \left(\sum_{i=1}^{m} \widetilde{\widetilde{\mathbf{X}}}_{ik}^T\mathbf{V}_i^{-1}\widetilde{\widetilde{\mathbf{X}}}_{ik} + \mathbf{W}_k^{-1}\right)^{-1}\right).$$

Distributions in (c) and (d) imply that $\mathbf{U}_{il}$'s $(i=1,\ldots,m, l=1,\ldots,q)$ can be drawn from:
Step 2:

$$\left[\mathbf{U}_{il}|\mathbf{Y}_i, \bigcup\{\mathbf{U}_k\}_{k=1,\ldots,p}, \bigcup\{\mathbf{U}_{it}\}_{t=1,\ldots,q,t\neq l}, \text{variance components}\right]$$

$$\sim \mathrm{N}((\widetilde{\widetilde{\mathbf{Z}}}_{il}^{\mathrm{T}}\widetilde{\widetilde{\mathbf{Z}}}_{il}/\sigma^2+\Lambda_l^{-1})^{-1}\left(\widetilde{\widetilde{\mathbf{Z}}}_{il}^{\mathrm{T}}\left(\mathbf{Y}_i-\sum_{k=1}^{p}\widetilde{\widetilde{\mathbf{X}}}_{ik}\mathbf{U}_k-\sum_{t\neq l}\widetilde{\widetilde{\mathbf{Z}}}_{it}\mathbf{U}_{it}\right)\right), \quad (\widetilde{\widetilde{\mathbf{Z}}}_{il}^{\mathrm{T}}\widetilde{\widetilde{\mathbf{Z}}}_{il}/\sigma^2+\Lambda_l^{-1})^{-1}).$$

Once $\mathbf{U}_k$'s and $\mathbf{U}_{il}$'s are sampled, the variance components can be sampled from:
Step 3:

$$[1/\lambda_{bk}|\mathbf{a}_k]\sim \mathrm{IG}\left(\frac{\gamma_{bk}+n-2}{2}, \frac{\mu_{bk}+\mathbf{a}_k^{\mathrm{T}}\mathbf{a}_k}{2}\right), \quad k=1,\ldots,p,$$

$$\left[\mathbf{\Omega}_l|\bigcup\{\zeta_{il}\}_{i=1,\ldots,m}\right]\sim \mathrm{IW}\left(m+d_l, \left(\sum_{i=1}^{m}\zeta_{il}\zeta_{il}^{\mathrm{T}}+(c_l\mathbf{R}_l)^{-1}\right)^{-1}\right), \quad l=1,\ldots,q,$$

$$\left[1/\lambda_{al}|\bigcup\{\mathbf{a}_{il}\}_{i=1,\ldots,m}\right]\sim \mathrm{IG}\left(\frac{\gamma_{al}+(n-2)m}{2}, \frac{\mu_{al}+\sum_{i=1}^{m}\mathbf{a}_{il}^{\mathrm{T}}\mathbf{a}_{il}}{2}\right), \quad l=1,\ldots,q,$$

$$\left[\sigma^2|\bigcup\{\mathbf{Y}_i\}_{i=1,\ldots,m}\bigcup\{\mathbf{U}_k\}_{k=1,\ldots,p}, \bigcup\{\mathbf{U}_{il}\}_{i=1,\ldots,m, l=1,\ldots,q}\right]$$

$$\sim \mathrm{IG}\left(\frac{\gamma_0+nm}{2}, \frac{\mu_0+\sum_{i=1}^{m}(\mathbf{Y}_i-\sum_{k=1}^{p}\widetilde{\widetilde{\mathbf{X}}}_{ik}\mathbf{U}_k-\sum_{l=1}^{q}\widetilde{\widetilde{\mathbf{Z}}}_{il}\mathbf{U}_{il})^{\mathrm{T}}(\mathbf{Y}_i-\sum_{k=1}^{p}\widetilde{\widetilde{\mathbf{X}}}_{ik}\mathbf{U}_k-\sum_{l=1}^{q}\widetilde{\widetilde{\mathbf{Z}}}_{il}\mathbf{U}_{il})}{2}\right).$$

To draw the missing values in $\mathbf{Y}_i$'s, suppose $\mathbf{Y}_i=(\mathbf{Y}_{i,\mathrm{mis}}, \mathbf{Y}_{i,\mathrm{obs}})$ and denote $n_{i,\mathrm{mis}}$ as the number of missing values in $\mathbf{Y}_i$, the posterior distribution of the missing values given the observed values and parameters is
Step 4:

$$[\mathbf{Y}_{i,\mathrm{mis}}|\mathbf{Y}_{i,\mathrm{obs}}, \text{parameters}]\sim \mathrm{N}\left(\sum_{k=1}^{p}\widetilde{\widetilde{\mathbf{X}}}_{ik,\mathrm{mis}}\mathbf{U}_k+\sum_{l=1}^{q}\widetilde{\widetilde{\mathbf{Z}}}_{il,\mathrm{mis}}\mathbf{U}_{il}, \sigma^2\mathbf{I}_{n_{i,\mathrm{mis}}}\right),$$

for $i=1,\ldots,m$, where $\widetilde{\widetilde{\mathbf{X}}}_{ik,\mathrm{mis}}$ and $\widetilde{\widetilde{\mathbf{Z}}}_{ik,\mathrm{mis}}$ denote the submatrix of $\widetilde{\widetilde{\mathbf{X}}}_{ik}$ and $\widetilde{\widetilde{\mathbf{Z}}}_{ik}$ corresponding to the missing items in the ith subject.

Given the starting values of model parameters, the above steps 1–4 define a cycle of the Gibbs sampler. Executing this cycle repeatedly creates sequences of draws of parameters and missing data. Time series plots and Gelman–Rubin statistics [26] can be used to determine whether the Gibbs chain converges. After the Gibbs sampler converges, we can select M (e.g. $M=5$) draws of the missing values from one Gibbs sampler sequence, with long lag times between each draw to avoid autocorrelation, to obtain M completed data sets. Or we can carry out M independent Gibbs sampling chains and choose the last draw of missing values from each chain upon convergence. Although the fully Bayesian inferences of the functional mixed model parameters can be obtained from the posterior draws of the parameters, using imputed data can be advantageous for analyses that do not directly conform to the functional mixed model (Section 4). In this work, we code the Gibbs sampling imputation algorithm using a matrix-based language GAUSS (www.aptech.com). Some sample code are attached in Appendix,[‡] and please contact the authors for further information.

4. Application

4.1. Survey background

The PSID is a longitudinal survey of a national representative sample of U.S. families. Originated in 1968, the PSID interviewed and re-interviewed individuals from sampled families every year until

1996. Since 1997, the interviewing has taken place every other year. The panel is based on a complex survey design for households. It follows adults through the full life course and also includes families when children of the original sample grow up and establish separate households or when marriage goes separate ways. The PSID emphasizes the dynamic aspects of economic and demographic behavior, but it contains a wide range of measures, including health, sociological, and psychological ones. In 1997, the PSID supplemented its main data collection with additional data on 0- to 12-year-old children, their parents, and other care-givers. The Child Development Supplement (CDS) cohort studies a broad array of developmental outcomes such as physical health, emotional well-being, cognitive abilities, achievements, and social relationships with family and peers. More information about the PSID and CDS can be obtained from the web site http://psidonline.isr.umich.edu.

It is well known that there exists positive association between health and income in adulthood. Research using the data from the PSID, CDS, and other public-use databases has shown that such association has antecedent in childhood [27, 28], although mixed results were obtained regarding the role of timing of household income in affecting a child's health. One possibility is that the effect of income depends on the age of the child when the income was received [27]. Another possibility is that investment decisions are made on long-run average income, in which case the timing is not important [28]. In this example, we illustrate the proposed imputation approach in analyzing a subset of the PSID and CDS data to compare the effect of household income on children's health development at different time spans. However, we do not attempt to present a comprehensive and rigid study for the subjective matters, which can be found in the aforementioned literature.

4.2. Multiple imputation

We limit our multiple imputation analysis to a sample of 2172 families included in the PSID cohort from 1979 to 1997, and augment it with the data from their children collected in the CDS. For simplicity, we only include the data from the eldest child in each family and ignore the sampling design in our analysis. The longitudinal response variables are annual record of family income to needs ratio from 1979 to 1997, which is defined as the ratio of household income to the poverty threshold for the corresponding family size. The average missingness rate for the income data is approximately 15.1 per cent per year, with a declining trend at the end. The missingness pattern is rather arbitrary (i.e. nonmonotone). The children's health status (HEALTH: 1=Excellent–very good, 82.4 per cent; 0=Good–fair–poor, 17.6 per cent) was collected in the CDS cohort. In addition, the subset data include some fully observed auxiliary variables which may be associated with income or health status. These covariates include: children's gender (GENDER: 1=Male, 51.8 per cent; 0=Female, 48.2 per cent); children's age (AGE: mean=6.8 years, std=3.7 years); household head's race (RACE: 1=White, 51.8 per cent; 0=Non-white, 48.2 per cent); and household head's education level (EDU: 1=More than high school, 43.1 per cent; 0=High school or less, 56.9 per cent).

We apply a log-transformation to accommodate the skewness of income data and carry out the imputation on the transformed scale. We use the functional mixed model (1) to characterize the trends of log-transformed income series. For illustrative purposes, we ignore the possible autocorrelation of the measurement error terms. The covariate effects of HEALTH, RACE, and EDU are treated as fixed smooth functions as we hypothesize families with different such characteristics may have different income trajectory. Individual family's deviation from population-average profile are treated as random smooth curves.

We use independent $N(0, 10^3)$ priors for all fixed effects, independent IW $(10^{-3}, 10^{-3}I_2)$ priors for covariance matrices, and independent IG $(10^{-3}, 10^{-3})$ priors for the inverse of smoothing parameters and error variance. These hyper-parameters reflect vague prior information on the parameters. To ensure reliable results, we also examine the sensitivity of the estimation results to the values of hyper-parameters of the conjugate priors. We find that various reasonable values of the hyper-parameters lead to essentially the same results because of the large sample size. The Gibbs chains achieve convergence after burn-in period of 2500 iterations, assessed by examining the time series plots, sample autocorrelation plots, and Gelman–Rubin statistics from multiple chains. We run five independent Gibbs sampler sequences of 5000 iterations and use the missing values drawn from the last iteration of each sequence to obtain five completed data sets (that is, $M=5$). We find that the multiple imputation analysis results essentially remain unchanged when we increase the number of imputations, as the relative efficiency [2] of five imputations for the analyses of interest (Section 4.3) exceed 95 per cent.

Table I. Functional mixed model parameter estimates.

Parameter	Posterior Mean	95 per cent credible interval
$\lambda_{\text{HEALTH}=0}$	693	(209, 1703)
$\lambda_{\text{HEALTH}=1}$	782	(245, 1799)
λ_{SUBJECT}	341	(311, 372)
σ^2	9.77×10^{-2}	$(9.58 \times 10^{-2}, 9.96 \times 10^{-2})$
$\delta_{0,\text{HEALTH}}$	5.73×10^{-3}	$(-5.54 \times 10^{-2}, 6.59 \times 10^{-2})$
$\delta_{1,\text{HEALTH}}$	8.70×10^{-3}	$(3.99 \times 10^{-3}, 1.33 \times 10^{-2})$

Notes: $\delta_{0,\text{HEALTH}}$ is the estimate of intercept differences of population-average income profile between two groups of families classified by HEALTH (HEALTH=1 vs HEALTH=0). $\delta_{1,\text{HEALTH}}$ is the estimate of slope differences of population-average income profile between two groups of families classified by HEALTH.

Table I lists the posterior means and 95 per cent posterior confidence intervals for some of the model parameters. The large scales of the smoothing parameters suggest that the log-transformed income series are rather close to straight lines, although the individual series are more noisy (less smooth) than the population average profiles. Families whose child is in better health status has significantly larger slope ($\delta_{1,\text{HEALTH}} = 8.70 \times 10^{-3}$), indicating a faster growth on their income. Families with higher education also had a faster income growth; there was no difference on income growth between families with different races, although white families had apparently higher income at the baseline (results not shown).

Figure 2 plots the observed values, multiply imputed values, as well as the estimated family-specific log-transformed income trajectories from randomly selected three families. The family-specific curve estimates are the posterior means of the smooth functions from the Gibbs samples, and they are similar across independent Gibbs chains. There exist some curvature for the estimated trends but the deviation from linearity is not strong. The imputed values ('+' in Figure 2) for each family are scattered the estimated trend, and they are different across multiple Gibbs chains, incorporating the uncertainty of imputation.

4.3. Post-imputation analyses

We investigate the effect of family income on children's health status on three periods: (1) six years prior to birth; (2) the first three years since birth (i.e. critical developmental period) [27]; and (3) the child's entire life until 1997. We conduct the following analyses using the multiply imputed datasets:

(A) Estimate the differences of the average income to needs ratio during the three periods (1)–(3) between two groups of families classified by HEALTH;
(B) Fit three logistic regression models corresponding to three periods, and in each model the average income to needs ratio during the period and other covariates are used to predict HEALTH;
(C) Fit a mixed model assuming both fixed and random linear trends for the log-transformed income to needs ratio from 1979 to 1997, and estimate the slope difference between two groups of families classified by HEALTH, controlling for other covariates.

These analyses represent typical ones applied to longitudinal data, that is, the univariate descriptive statistics (A), multivariate regression ignoring correlation among repeated measurements (B), and mixed-model fitting (C).

The estimates from each completed dataset are combined using the multiple imputation combining rules [2]. In addition to the functional multiple imputation approach (FUN), we also apply alternative missing data methods. These approaches include the AC analysis procedure, the last observation carried forward method (LOCF), the mean imputation method (MEAN), and the multiple imputation approach based on a random effects model assuming linear growth curves of log-income (LIN, implemented using R library 'pan' with $M = 5$). Note that AC discards incomplete cases in analyses, while LOCF, MEAN, and LIN fill in these missing cases for each subject repeatedly over time and thus create completed data for analyses.

Table II shows the estimates of interest and their associated 95 per cent confidence intervals (inside the parenthesis). The results are somewhat similar across the methods. The income difference estimates suggest that children with better health status are more likely to be in higher income family. In addition, the logistic regression analysis results imply that the family income status at different time spans
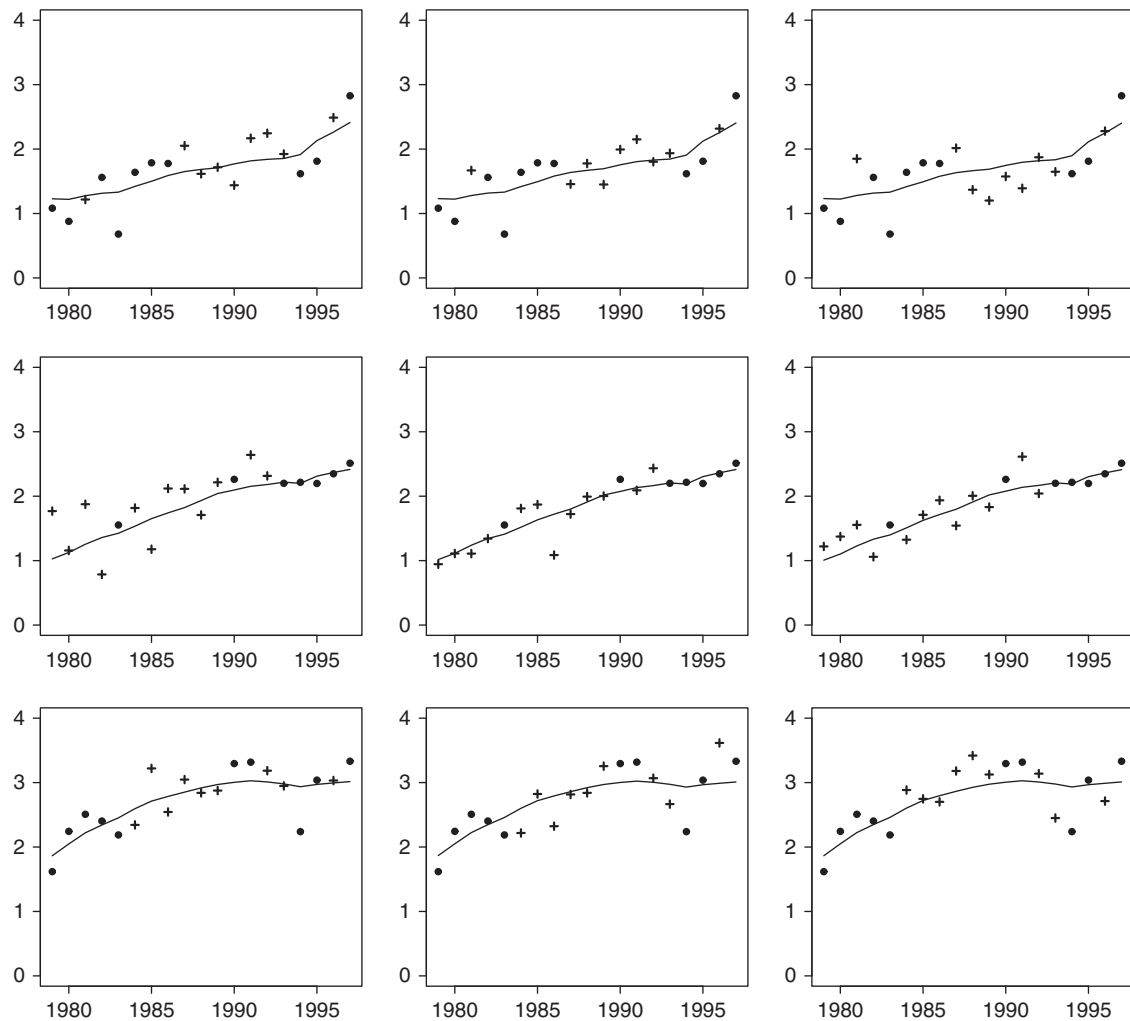
Figure 2. Trajectory plots of the observed and imputed data of log-transformed income to needs ratio from selected subjects; row (from top to bottom): three subjects; column (from left to right): three imputations per subject; concrete line: family-specific curve estimates; dots: observed values; plus: imputed values.

have significant and rather similar effects on children's health development. The mixed model analysis results further quantifies the difference in income growth for the two groups over the study period. Our illustrative example indicates that, rather than the income at specific age period, the long run average income determines health investment and health status, which is consistent with findings from [28].

5. Simulation studies

In this section we use Monte Carlo simulations to systematically evaluate the performance of the proposed approach and compare it with those alternatives.

5.1. Design

5.1.1. Complete-data generation. We consider the functional mixed model

$$y_{ij}|x_i = \beta_x(t_{ij}) + \alpha_i(t_{ij}) + \varepsilon_{ij}, \quad \varepsilon_{it} \sim N(0, \sigma^2),$$

$$x_i = 1 + \text{Bernoulli}(0.5),$$

and allow $\beta_x(t)$ and $\alpha_i(t)$'s to be various functions of t including linear, polynomial, and trigonometric curves.

Table II. Analysis results relating family income to children's health.

	DIFF (6 yrs prior)	DIFF (Critical)	DIFF (entire)
AC	2.488 (1.686, 3.291)	3.240 (2.003, 4.476)	3.865 (2.730, 4.999)
LOCF	2.375 (1.621, 2.130)	3.275 (2.144, 4.406)	3.888 (2.768, 5.007)
MEAN	2.046 (1.339, 2.753)	2.956 (1.878, 4.034)	3.646 (2.559, 4.734)
LIN	2.263 (1.513, 3.012)	3.217 (2.079, 4.355)	3.888 (2.761, 5.016)
FUN	2.338 (1.561, 3.114)	3.196 (2.044, 4.348)	3.886 (2.756, 5.016)
	OR (6 yrs prior)	OR (Critical)	OR (entire)
AC	1.378 (1.055, 1.802)	1.303 (1.042, 1.629)	1.445 (1.185, 1.762)
LOCF	1.418 (1.082, 1.857)	1.402 (1.113, 1.767)	1.491 (1.214, 1.832)
MEAN	1.360 (1.017, 1.817)	1.391 (1.090, 1.774)	1.480 (1.197, 1.831)
LIN	1.316 (1.008, 1.718)	1.363 (1.082, 1.717)	1.468 (1.198, 1.798)
FUN	1.343 (1.019, 1.770)	1.331 (1.057, 1.675)	1.443 (1.194, 1.787)
	SLOPE $\times 10^3$		
AC	8.46 (3.81, 13.10)		
LOCF	8.72 (4.18, 13.27)		
MEAN	9.05 (4.72, 13.38)		
LIN	8.40 (3.73, 13.08)		
FUN	8.49 (3.90, 13.07)		

Notes: DIFF denotes the difference in the average income to needs ratio in the analysis (A) from Section 4.3. OR denotes the odds ratio estimates associated with the average income to needs ratio (1 unit change=10) in the analysis (B) from Section 4.3. SLOPE denotes the slope difference estimated in the analysis (C) from Section 4.3.

Study I: We first consider the case that $\beta_1(t) \neq \beta_2(t)$ so that two groups have different distributions of longitudinal profiles. The estimands of interest (Section 5.1.3) focus on the differences between two groups. The data-generating functions are as follows:

scenario I.1. linear curves: $t = 1, 2, \ldots, 20,$

$$\beta_x(t) = \begin{cases} 3+6t & \text{if } x_i = 1, \\ 6+5t & \text{if } x_i = 2, \end{cases}$$

$$\alpha_i(t) = a_{0i} + a_{1i}t,$$

$$(a_{0i}, a_{1i})^{\mathrm{T}} \sim \mathrm{N}\left(\mathbf{0}, \begin{pmatrix} 100 & -0.5 \\ -0.5 & 0.36 \end{pmatrix}\right), \quad \sigma^2 = 27;$$

scenario I.2. polynomial curves: $t = 0, 0.2, 0.4, \ldots, 3.8,$

$$\beta_x(t) = \begin{cases} 16-12t+7t^2-9t^3+6t^4-t^5 & \text{if } x_i = 1, \\ 14+12t-7t^2+8t^3-4t^4+0.7t^5 & \text{if } x_i = 2, \end{cases}$$

$$\alpha_i(t) = a_{0i} + a_{1i}t + a_{2i}t^2 + a_{3i}t^3 + a_{4i}t^4,$$

$$(a_{0i}, a_{1i}, a_{2i}, a_{3i}, a_{4i})^{\mathrm{T}} \sim \mathrm{N}(\mathbf{0}, \mathrm{diag}(20, 2, 0.1, 0.01, 0.001)), \sigma^2 = 1;$$

scenario I.3. trigonometric curves: $t = 1/21, 2/21, \ldots, 20/21,$

$$\beta_x(t) = \begin{cases} 8\sin(4\pi t)+6\cos(4\pi t) & \text{if } x_i = 1, \\ -\sin(4\pi t)-2\cos(4\pi t) & \text{if } x_i = 2, \end{cases}$$

$$\alpha_i(t) = a_{0i} + a_{1i}\sin(4\pi t) + a_{2i}\cos(4\pi t),$$

$$(a_{0i}, a_{1i}, a_{2i})^t \sim \mathrm{N}(\mathbf{0}, \mathrm{diag}(9, 4, 4)), \sigma^2 = 1.$$

Study II: We also consider the case where $\beta_1(t) = \beta_2(t)$ so that two groups have the same distributions of longitudinal profiles. The corresponding estimand of interest is the type-I error rate of falsely rejecting the null hypothesis that these two groups differ. The data-generating models are as follows:

scenario II.1. linear curves: $t = 1, 2, \ldots, 20$,

$$\beta_1(t) = \beta_2(t) = 3 + 6t,$$

$$\alpha_i(t) = a_{0i} + a_{1i}t,$$

$$(a_{0i}, a_{1i})^{\mathrm{T}} \sim \mathrm{N}\left(\mathbf{0}, \begin{pmatrix} 100 & -0.5 \\ -0.5 & 0.36 \end{pmatrix}\right), \quad \sigma^2 = 27;$$

scenario II.2. polynomial curves: $t = 0, 0.2, 0.4, \ldots, 3.8$,

$$\beta_1(t) = \beta_2(t) = 16 - 12t + 7t^2 - 9t^3 + 6t^4 - t^5,$$

$$\alpha_i(t) = a_{0i} + a_{1i}t + a_{2i}t^2 + a_{3i}t^3 + a_{4i}t^4,$$

$$(a_{0i}, a_{1i}, a_{2i}, a_{3i}, a_{4i})^{\mathrm{T}} \sim \mathrm{N}(\mathbf{0}, \mathrm{diag}(20, 2, 0.1, 0.01, 0.001)), \sigma^2 = 1;$$

scenario II.3. trigonometric curves: $t = 1/21, 2/21, \ldots, 20/21$,

$$\beta_1(t) = \beta_2(t) = -\sin(4\pi t) - 2\cos(4\pi t),$$

$$\alpha_i(t) = a_{0i} + a_{1i}\sin(4\pi t) + a_{2i}\cos(4\pi t),$$

$$(a_{0i}, a_{1i}, a_{2i})^t \sim \mathrm{N}(\mathbf{0}, \mathrm{diag}(1, 1, 4)), \sigma^2 = 1.$$

In every scenarios considered, 500 complete datasets are generated, and each set contains 100 subjects ($m = 100$) with longitudinal Y's measured over 20 time points (waves).

5.1.2. Creating missing data. We assume that the covariate X is fully observed and impose both monotone and nonmonotone missingness patterns for Y's. In a monotone pattern, subjects never come back after they drop out of the study. In a nonmonotone pattern, the missingness is rather intermittent so a subject can have observed response at a later wave even if he/she missed earlier ones [1]. In principle under MAR, the missingness of a case at wave j can be dependent on observed values at any wave before j (i.e. wave $1, 2, \ldots, j-1$). However, we surmise that in most real life scenarios, the probability of missingness at wave j is affected more by earlier observations closer to wave j such as at wave $j-1$ or $j-2$ and affected less by those closer to the starting time (i.e. baseline). Therefore, we consider two scenarios for modeling the dependency of missingness in the simulation.

In a simple scenario, the missingness at wave j is dependent on the most recently observed value. To create a monotone pattern, the Y's are fully observed for the first five waves, while for the later waves ($j \geqslant 6$), the missing values are produced using the following equation:

$$Pr(y_{ij} \text{ is missing}|y_{i1}, \ldots, y_{i,j-1}) = \begin{cases} \Phi(\lambda_0 + \lambda_1 x_i + \lambda_2 y_{i,j-1} + \lambda_3 x_i y_{i,j-1}) & \text{if } y_{i,j-1} \text{ is observed,} \\ 1 & \text{if } y_{i,j-1} \text{ is missing,} \end{cases}$$

where Φ is the cumulative probability distribution function for the standard normal distribution. To create a nonmonotone pattern, the Y's are fully observed at the baseline (the first wave), while for the later waves ($j \geqslant 2$), the missing values are produced using the following equation:

$$Pr(y_{ij} \text{ is missing}|y_{i1}, \ldots, y_{i,j-1}) = \Phi(\lambda_0 + \lambda_1 x_i + \lambda_2 y_{ij*} + \lambda_3 x_i y_{ij*}),$$

where y_{ij*} denotes the most recently observed value before wave j. Unlike the monotone pattern, y_{ij*} could be values at a wave earlier than $j-1$. Note that the interaction between the group indicator and observed values in the missingness equations allows the pattern of dependency differs between two groups. With certain choices of λ's, for example, we can produce scenarios where in one group, the missingness tends to occur for subject whose observed values are larger, while it is less so or even in opposite direction in another group. It might be expected that throwing away incomplete cases in these scenarios or not incorporating the trend appropriately would lead to bias for the estimates of the difference in the two groups.

In a more general scenario, we let the probability of missingness at wave j depends on observed values from the two recent waves. For the monotone missingness, the Y's are fully observed for the first five waves, while for the later waves ($j \geqslant 6$), the missing values are produced using the following equation:

$$Pr(y_{ij} \text{ is missing } | y_{i1}, \ldots, y_{i,j-1})$$

$$= \begin{cases} \Phi(\lambda_0 + \lambda_1 x_i + \lambda_2 y_{i,j-1} + \lambda_3 y_{i,j-2} + \lambda_4 x_i y_{i,j-1} + \lambda_5 x_i y_{i,j-2}) & \text{if } y_{i,j-1} \text{ is observed,} \\ 1 & \text{if } y_{i,j-1} \text{ is missing.} \end{cases}$$

For the nonmonotone missingness, the Y's are fully observed on the first two waves, while for the later waves ($j \geqslant 3$), the missing values are produced using the following equation:

$$Pr(y_{ij} \text{ is missing } | y_{i1}, \ldots, y_{i,j-1}) = \Phi(\lambda_0 + \lambda_1 x_i + \lambda_2 y_{ij*} + \lambda_3 y_{ij**} + \lambda_4 x_i y_{ij*} + \lambda_5 x_i y_{ij**}),$$

where y_{ij*} and y_{ij**} denote the observed values from two recent waves before wave j, and they can be at waves earlier than $j-1$ and $j-2$ for some cases.

The coefficients λ's are chosen to yield the average missingness rate per wave close to 20 per cent in all scenarios considered. For example, Figure 3 shows one simulated incomplete dataset from various curves in the simple scenario. We could include more waves of observed values (e.g. $j-1, j-2, j-3, \ldots$) in the above equations for even more general missingness mechanisms. But this would increase the complexity of the missingness probability equations and we are less capable of controlling average missingness rates in the simulation.

5.1.3. Estimands of interest. In Study I, we consider three estimands that distinguish the two groups (i.e. $x = 1$ vs $x = 2$):

(i) The between-group difference for the mean of responses from wave 12 till wave 16, i.e.

$$\bar{y}_{12-16,x=1} - \bar{y}_{12-16,x=2} = \frac{\sum_{i=1}^{100} I_{\{x_i=1\}} \sum_{j=12}^{j=16} y_{ij}}{5 \sum_{i=1}^{100} I_{\{x_i=1\}}} - \frac{\sum_{i=1}^{100} I_{\{x_i=2\}} \sum_{j=12}^{j=16} y_{ij}}{5 \sum_{i=1}^{100} I_{\{x_i=2\}}}.$$

(ii) Run a logistic regression for the group indicator X using the average of Y's from wave 12 till wave 16 as the predictor, and estimate the associated coefficient;

(iii) Fit with the data-generating model (Section 5.1.1), which are mixed models for longitudinal y_{ij}'s, and estimate the mean difference in selected coefficients between two groups. Specifically, in scenarios I.1 and I.2, the estimand of interest is the difference in the coefficient for the linear term t, and in scenario I.3, the estimand of interest is the difference in the coefficient associated with $\sin(4\pi t)$.

Matched with analysis (A)–(C) in Section 4.3, estimands (i) and (ii) focus on the 'local' difference between the two groups and estimand (iii) is pertaining to the 'global' difference. These estimands also represent typical analyses applied to longitudinal data, that is, the univariate descriptive statistics (i), multivariate regression ignoring correlations within repeated measurements (ii), and mixed-model fitting (iii).

In Study II, we apply the same analyses as in Study I and test the null hypothesis that the quantity of interest is 0, which indicates no difference between two groups. We obtain the average rate of rejecting the hypothesis (i.e. Type-I error) over 500 simulations. Since the two groups have the same distribution of longitudinal profiles, we would expect the estimated type-I error would be close to the nominal level 5 per cent for the method working properly.

5.1.4. Methods and evaluation criteria. For all scenarios considered, we apply FUN to the incomplete datasets. Same prior distributions as those used in the real data example are adopted in the Gibbs sampling. We use the last draws of five independent Gibbs chains with 3000 iterations for multiple imputation, with the relative efficiency exceeding 95 per cent for estimands of interest. Other missing data methods including AC, LOCF, MEAN, and LIN (with $M=5$) are also applied. In scenarios 2 and 3, we also apply the linear mixed-model imputation assuming a fifth-order polynomial for the mean response and a fourth-order polynomial for the random effects using R library 'pan'(POLY).
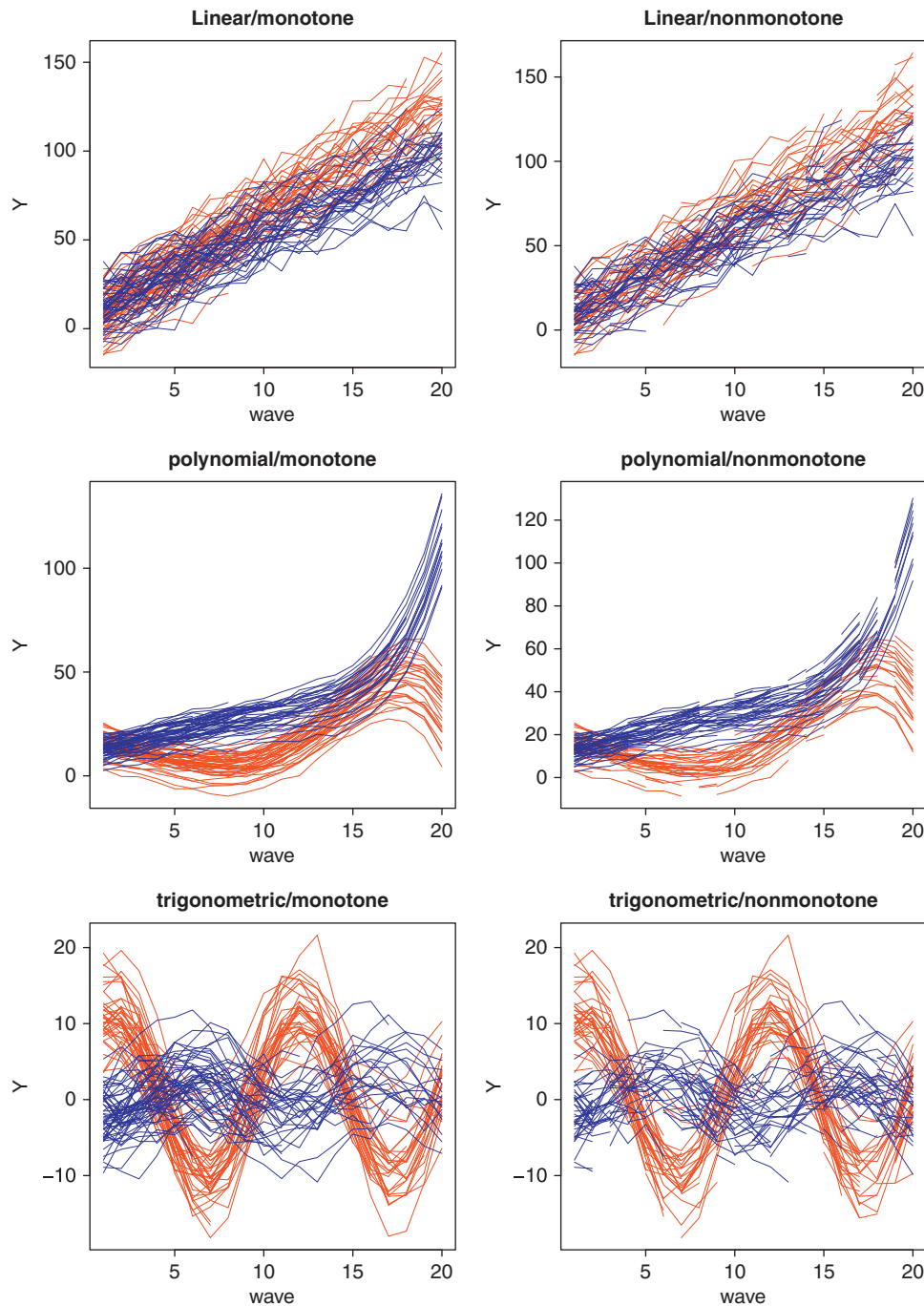
Figure 3. Simulated longitudinal data with missing values. Red (Gray on print version):
$x = 1$; Blue (Black on print version): $x = 2$.

In both studies, the inferences made on the data before deletion (BD) are taken as the gold standard.
We evaluate the performance of the methods using the following four criteria [29]:

Relative bias (RBIAS = |Bias/True| × 100 per cent) (Study I): it measures the accuracy of the point
estimator, and a reasonable upper limit for the relative bias can be taken as 5 per cent.

Root of the relative mean squared error (RRMSE = $\sqrt{\text{MSE(Method)}/\text{MSE(BD)}}$) [30] (Study I): the
root of the mean squared error (RMSE) has been extensively used in the literature of simulation studies
because it is an integrated measure of bias and variance and has the same units as the quantity being
estimated. RRMSE is a standardized version of the RMSE, measuring the increase in RMSE relative
to that of the BD analysis.

Table III. Results from simulation study I: missingness dependent on the observed values from the most recent wave.

Function	Missingness pattern	Method	Mean difference			Logit regression			Mixed-model		
			|RBIAS|	RRMSE	COV	|RBIAS|	RRMSE	COV	|RBIAS|	RRMSE	COV
Linear	Monotone	AC	**5.13**	1.14	930	4.55	1.13	908	0.04	1.09	950
		MEAN	29.22	1.48	**654**	**8.06**	1.28	932	**45.46**	3.79	**70**
		LOCF	**32.26**	1.89	**844**	**64.99**	2.21	**84**	**47.42**	4.54	**708**
		LIN	0.04	1.03	964	0.55	1.03	916	0.06	1.12	952
		FUN	0.15	1.03	948	0.51	1.03	918	0.10	1.14	950
Linear	Nonmonotone	AC	0.11	1.03	950	**5.74**	.95	918	0.02	1.02	956
		MEAN	**26.39**	1.34	**670**	**30.14**	1.70	902	**30.00**	2.54	**160**
		LOCF	**8.67**	1.05	928	**8.32**	0.99	908	2.34	1.05	954
		LIN	0.14	1.00	952	0.29	0.99	934	0.02	1.02	956
		FUN	0.01	1.00	956	0.19	0.99	928	0.01	1.02	958
Polynomial	Monotone	AC	4.52	1.27	942	**7.28**	1.27	**888**	0.05	1.04	934
		MEAN	**34.10**	2.90	**122**	**7.43**	1.54	952	**50.63**	11.12	**22**
		LOCF	**13.36**	1.81	**880**	**53.02**	2.61	**24**	**35.58**	7.69	**2**
		LIN	**33.92**	3.01	**528**	**8.80**	1.25	990	**40.22**	8.81	**810**
		POLY	0.05	1.07	954	0.68	1.03	942	0.03	1.07	934
		FUN	0.31	1.06	946	4.14	1.11	952	0.95	1.14	938
Polynomial	Nonmonotone	AC	1.38	1.12	942	**11.80**	1.10	**816**	0.04	1.09	942
		MEAN	**28.81**	2.45	**132**	**36.20**	2.31	**818**	**60.56**	13.70	**398**
		LOCF	**11.13**	1.33	**860**	**17.32**	1.45	956	**37.43**	8.48	**454**
		LIN	**28.53**	2.56	**504**	**10.35**	1.23	984	**34.05**	8.23	972
		POLY	0.06	1.00	956	0.83	0.99	940	0.20	1.11	950
		FUN	0.04	1.00	956	0.67	0.99	940	1.25	1.16	946
Trigonometric	Monotone	AC	**22.82**	1.59	904	**16.18**	1.41	966	0.02	1.03	958
		MEAN	34.84	1.53	**578**	**15.59**	1.51	960	**25.73**	6.15	**0**
		LOCF	**89.01**	3.69	**418**	**94.34**	3.13	**6**	**13.55**	3.36	**160**
		LIN	**74.27**	3.07	**514**	**81.32**	2.77	**158**	**16.55**	4.05	**128**
		POLY	**24.47**	1.41	**882**	**24.49**	1.39	948	3.78	1.47	**818**
		FUN	2.12	1.15	972	**6.91**	1.05	962	0.09	1.25	904
Trigonometric	Nonmonotone	AC	**5.57**	1.12	940	**9.01**	0.99	914	0.01	1.01	954
		MEAN	**31.31**	1.45	**678**	2.83	1.22	948	**24.23**	5.76	**0**
		LOCF	**35.86**	1.70	**728**	**28.21**	1.43	914	**7.40**	2.01	**614**
		LIN	**30.10**	1.55	**860**	**35.88**	1.48	**734**	**17.94**	4.31	**22**
		POLY	**9.97**	1.09	938	**12.91**	1.16	964	1.02	1.07	**888**
		FUN	2.19	1.03	940	2.69	1.06	942	0.26	1.04	904

Notes: Mean difference: between-group difference in the average of responses from wave 12 till wave 16. Logit regression: a logistic regression for the group indicator using the average of responses from wave 12 till wave 16 as the predictor. Mixed model: fit with the data-generating model for longitudinal responses and estimate the mean difference in selected coefficients between two groups. These analyses correspond to estimands (i)–(iii) in Section 5.1.3. Highlighted numbers indicate that the absolute relative bias are larger than 5 per cent and coverage rates are lower than 90 per cent.

Coverage rate of the 95 per cent confidence interval estimates across 500 replicates (COV) (Study I): if a method is working well, the actual coverage should be close to the nominal rate (95 per cent). The performance of a method can be regarded to be poor if its coverage drops below 90 per cent.

Type-I error rate when the two groups have the same distribution of repeated outcomes (Study II). If a method is working well, the actual Type-I error rate should be close to the nominal level (5 per cent). The performance of a method can be regarded to be poor if its Type-I error rate exceeds 10 per cent.

5.2. Results

Tables III and IV include the results from study I, where curves from two groups have different distributions. Biases and coverage rates that are beyond acceptable limits are shown with bold characters. As expected, the naive missing data methods, such as AC, MEAN, and LOCF, can perform badly when estimating the mean difference and logistic regression across all three types of curves. When fitting the random coefficient models, however, AC is performing well because the estimation is based on

Table IV. Results from simulation study I: missingness dependent on the observed values from the two recent waves.

Function	Missingness pattern	Method	Mean difference			Logit regression			Mixed-model		
			|RBIAS|	RRMSE	COV	|RBIAS|	RRMSE	COV	|RBIAS|	RRMSE	COV
Linear	Monotone	AC	**25.52**	1.61	**828**	**29.59**	1.66	924	0.43	1.12	962
		MEAN	**26.42**	1.36	**662**	**42.09**	2.10	908	**45.85**	3.82	**9**
		LOCF	**36.31**	2.15	**826**	**72.62**	2.44	**28**	**23.80**	3.48	900
		LIN	0.76	1.05	940	2.60	1.09	934	0.71	1.15	962
		FUN	0.11	1.05	952	0.79	1.05	920	0.12	1.17	964
Linear	Nonmonotone	AC	**5.26**	1.05	948	0.01	0.98	918	0.06	1.04	946
		MEAN	**31.32**	1.47	**472**	**74.16**	3.04	**674**	**51.12**	4.17	**0**
		LOCF	**7.24**	0.96	932	**25.34**	1.42	918	**18.01**	1.75	**690**
		LIN	0.04	1.00	952	0.29	1.01	924	0.00	1.04	944
		FUN	0.16	1.01	956	0.36	1.01	930	0.02	1.05	944
Polynomial	Monotone	AC	**8.11**	1.42	910	**13.45**	1.24	**830**	0.17	1.17	946
		MEAN	**39.73**	3.32	**38**	2.07	1.30	938	**43.80**	9.73	**100**
		LOCF	**28.85**	2.82	**678**	**43.89**	2.19	**112**	**29.38**	6.39	**12**
		LIN	**43.94**	3.91	**390**	0.21	0.97	974	**28.16**	6.40	984
		POLY	0.27	1.04	948	1.39	1.00	932	0.00	1.04	944
		FUN	0.47	1.04	940	3.54	1.06	952	0.64	1.06	946
Polynomial	Nonmonotone	AC	**8.00**	1.26	908	**15.24**	1.17	**764**	0.01	1.13	950
		MEAN	**42.71**	3.52	**0**	**21.60**	1.84	902	**64.95**	14.83	**318**
		LOCF	**13.82**	1.46	**780**	3.30	1.02	920	**24.66**	6.05	**756**
		LIN	**31.08**	2.77	**464**	**8.51**	1.18	978	**17.79**	4.78	1000
		POLY	0.12	1.01	956	0.76	0.99	938	0.27	1.16	950
		FUN	0.10	1.01	954	0.54	0.99	938	3.71	1.56	918
Trigonometric	Monotone	AC	**76.2**	3.10	**316**	**95.54**	3.52	**566**	0.01	1.04	950
		MEAN	**17.11**	1.05	**890**	**106.69**	3.89	**472**	**24.94**	6.00	**0**
		LOCF	**74.30**	3.12	**478**	**86.18**	2.88	**2**	**7.58**	2.08	**604**
		LIN	**18.61**	1.33	946	**30.00**	1.38	**838**	**12.63**	3.16	**312**
		POLY	**17.16**	1.22	930	**22.84**	1.38	970	1.29	1.14	912
		FUN	1.57	1.12	968	2.80	1.07	962	0.68	1.15	922
Trigonometric	Nonmonotone	AC	**17.27**	1.29	912	**30.52**	1.34	**732**	0.03	1.01	948
		MEAN	**60.12**	2.39	**168**	**32.04**	1.63	**804**	**18.20**	4.38	**6**
		LOCF	**56.86**	2.42	**534**	**24.07**	1.27	944	3.67	1.30	**890**
		LIN	**56.57**	2.36	**522**	**58.33**	2.09	**460**	**16.27**	3.94	**3**
		POLY	**7.82**	1.08	944	**7.64**	1.08	960	2.26	1.16	**884**
		FUN	4.21	1.04	942	0.10	1.07	952	1.29	1.07	912

Notes: Mean difference: between-group difference in the average of responses from wave 12 till wave 16. Logit regression: a logistic regression for the group indicator using the average of responses from wave 12 till wave 16 as the predictor. Mixed model: fit with the data-generating model for longitudinal responses and estimate the mean difference in selected coefficients between two groups. These analyses correspond to estimands (i)–(iii) in Section 5.1.3. Highlighted numbers indicate that the absolute relative bias are larger than 5 per cent and coverage rates are lower than 90 per cent.

the observed-data likelihood when only the outcome is missing [31]. For linear response functions, LIN performs well in all estimands because it is based on the actual data-generating model, while for nonlinear response functions, either polynomial or trigonometric, LIN clearly breaks down with large biases and low coverage rates. As expected, POLY is performing well under the polynomial curves because the imputation model matches with the data-generating model. But with trigonometric curves, POLY can yield some bias and have low coverage rates. This indicates that correctly modeling the temporal trend is important in imputing the longitudinal response. Overall, FUN yields satisfactory performance across all scenarios and comes out as the top method. The results are similar across two missingness mechanisms tested.

Tables V and VI present the results from study II, where the curves from the two groups are generated from the same distribution. Type-I error rates that are beyond the acceptable limit are shown with bold characters. The overall pattern is similar to that from the simulation study I. For the mean difference and logit regression coefficient, the *ad hoc* approaches including AC, MEAN, and LOCF can have

Table V. Results from simulation study II: missingness dependent on the observed values from the most recent wave.

Function	Missingness pattern	Method	Mean difference Type-I error	Logit regression Type-I error	Mixed-model Type-I error
Linear	Monotone	AC	**73.6**	**71.4**	5.4
		MEAN	**31.2**	**25.4**	**26.4**
		LOCF	**50.4**	**49.6**	**97.4**
		LIN	5.2	3.8	4.2
		FUN	4.6	4.4	4.2
Linear	Nonmonotone	AC	**14.2**	**13.0**	4.4
		MEAN	**16.0**	**12.4**	**12.6**
		LOCF	**60.8**	**57.4**	**100.0**
		LIN	5.0	3.8	4.4
		FUN	4.8	4.2	3.0
Polynomial	Monotone	AC	**13.2**	**10.8**	5.8
		MEAN	8.6	8.0	4.0
		LOCF	**18.0**	**15.2**	**85.0**
		LIN	6.6	5.6	1.8
		POLY	4.4	4.2	5.0
		FUN	4.6	4.4	5.2
Polynomial	Nonmonotone	AC	**27.8**	**26.8**	6.4
		MEAN	7.4	6.6	1.6
		LOCF	**28.4**	**27.4**	**24.4**
		LIN	**19.0**	**15.8**	**95.4**
		POLY	4.4	4.0	5.8
		FUN	4.4	4.2	6.6
Trigonometric	Monotone	AC	5.4	3.6	4.4
		MEAN	**14.6**	**11.0**	6.0
		LOCF	**58.0**	**50.2**	**17.4**
		LIN	3.4	2.0	2.8
		POLY	3.6	2.6	8.8
		FUN	3.0	3.6	6.6
Trigonometric	Nonmonotone	AC	**11.2**	9.6	4.6
		MEAN	**10.4**	9.6	5.6
		LOCF	**49.0**	**48.6**	7.6
		LIN	7.6	5.4	4.0
		POLY	5.2	3.8	**11.6**
		FUN	5.2	4.2	9.6

Notes: Mean difference: between-group difference in the average of responses from wave 12 till wave 16. Logit regression: a logistic regression for the group indicator using the average of responses from wave 12 till wave 16 as the predictor. Mixed model: fit with the data-generating model for longitudinal responses and estimate the mean difference in selected coefficients between two groups. These analyses correspond to estimands (i)–(iii) in Section 5.1.3. Note: Highlighted numbers indicate that the Type-I error is larger than 10 per cent.

large type-I error rates across all three types of curves. For the mixed model coefficient estimates, AC has reasonable type-I error rates across scenarios because the estimation is based on the observed-data likelihood when only the outcome is missing [31]. When the curves are linear and polynomial, the error rates from LIN or POLY are around the nominal level, respectively. This is expected because the imputation model matches the true data-generating model under the respective situation. Overall, the error rate from FUN is close to the nominal level irrespective of the type of curves tested, suggesting its good performance and robustness.

6. Discussion

We have proposed a multiple imputation approach in which a functional mixed model is used to create imputations for incomplete repeated measurements. This approach avoids relying on parametric assumptions on time functions by using nonparametric cubic smoothing splines. The draw of missing

Table VI. Results from simulation study II: missingness dependent on the observed values from the two recent waves.

Function	Missingness pattern	Method	Mean difference Type-I error	Logit regression Type-I error	Mixed-model Type-I error
Linear	Monotone	AC	**71.6**	**68.4**	4.8
		MEAN	**30.2**	**27.0**	**260**
		LOCF	**32.4**	**31.8**	**90.8**
		LIN	5.6	4.6	4.2
		FUN	6.4	6.0	3.0
Linear	Nonmonotone	AC	9.2	8.6	5.8
		MEAN	**12.6**	**10.6**	**12.0**
		LOCF	**26.8**	**24.6**	**96.2**
		LIN	5.2	4.0	6.0
		FUN	4.8	3.8	4.2
Polynomial	Monotone	AC	**21.2**	**18.2**	5.8
		MEAN	**11.0**	9.8	5.6
		LOCF	**10.8**	9.4	**95.2**
		LIN	6.6	3.6	**12.2**
		POLY	3.8	3.2	5.6
		FUN	4.4	3.8	5.0
Polynomial	Nonmonotone	AC	**28.6**	**27.0**	7.0
		MEAN	**10.0**	**8.8**	2.0
		LOCF	**16.6**	**15.2**	**61.0**
		LIN	**22.0**	**19.0**	**72.4**
		POLY	4.4	4.2	6.0
		FUN	4.6	4.4	6.8
Trigonometric	Monotone	AC	6.6	4.6	3.4
		MEAN	**15.6**	**12.2**	5.2
		LOCF	**98.0**	**96.2**	**37.4**
		LIN	3.8	3.2	2.2
		POLY	6.8	7.2	6.2
		FUN	2.2	2.8	4.6
Trigonometric	Nonmonotone	AC	6.6	4.8	5.0
		MEAN	8.0	7.2	5.0
		LOCF	**44.0**	**44.8**	6.6
		LIN	3.8	2.8	3.0
		POLY	3.8	3.6	7.0
		FUN	5.0	3.8	8.2

Notes: Mean difference: between-group difference in the average of responses from wave 12 till wave 16. Logit regression: a logistic regression for the group indicator using the average of responses from wave 12 till wave 16 as the predictor. Mixed model: fit with the data-generating model for longitudinal responses and estimate the mean difference in selected coefficients between two groups. These analyses correspond to estimands (i)–(iii) in Section 5.1.3. Highlighted numbers indicate that the Type-I error is larger than 10 per cent.

data at the designed time points can be obtained using Gibbs sampling algorithm on the linear mixed model format of the functional mixed model. To enhance the practical use of this approach, imputation routines such as R library 'pan', which contains the Gibbs sampling algorithm for linear mixed models, can be modified so that practitioners can implement it without programming their own algorithm. Our simulation study has shown that this data-driven imputation strategy might be less sensitive to the functional forms of the longitudinal trends than methods relying on simple parametric functions such as linear or quadratic curves.

Several immediate extensions of the functional mixed model can be considered. First, if missing data occur in both outcome (Y's) and covariates (X's and Z's) in the functional mixed model (1), distributions for X's and Z's can be specified to jointly impute the missing data. Second, a multivariate extension of the functional mixed model can be formulated. In such extension, the correlations among the random functions of different response variables can be introduced through the covariance of the Gaussian processes of cubic smoothing splines. Third, similar to the extension from linear mixed model to generalized linear mixed model, a class of generalized functional mixed model can be proposed for nonnormal outcomes.

Our study has several limitations. First, all data-generating models used in the simulation study follow the functional mixed model (1), which assumes a linear relationship between the covariates and (fixed or random) coefficient functions of time. On the other hand, many types of longitudinal data can be modeled using nonlinear mixed models [32], such as in pharmacokinetic and pharmacodynamic analysis. Of potential interest is to assess the performance of the proposed approach under some classic types of nonlinear mixed models.

Second, we limit our study to ignorable missingness. In many practical cases, the incompleteness are often nonignorable, that is, the probability of missingness can depend on unobserved values [2]. Nonignorable missing data problems in longitudinal studies is rather challenging and the solution often involves modeling the missingness mechanism under some hypothetical assumptions [33]. On the other hand, models for nonignorably missing data are often specified by modifying a baseline ignorable model, for example, by hypothesizing a scale shift relative to ignorable imputations or a 'not missing at random' selection model. The proposed method can then be used as the baseline model, which is a starting point for developing a nonignorable model.

Third, we only consider the cubic smoothing splines, which is only one of the many smoothing methods developed. Cubic smoothing splines might be applicable to cases where data are rather 'smooth', lacking many peaks or jumps. In our illustrative example, we use a log transformation to smooth the income data before applying the functional modeling. When data exhibit many local features like spikes or changing points which cannot be well accommodated by transformation, more general wavelet-based approach can be used [22, 23] to estimate the functional mixed model. In addition, double smoothing estimators of the local linear regression [34] exhibit certain advantages over cubic smoothing splines when data are sparse. Therefore, it is of further interest to adapt these smoothing and functional data analysis techniques to missing data imputation and investigate their properties. On the other hand, functional data analysis methods are usually complicated, and it requires additional research to suit them for practical use by developing user-friendly software routines. Our future research plan pertains to extending 'pan' so that it can carry out imputation using functional mixed models as proposed in this paper. Therefore, practitioners can implement it without writing their own code.

This paper is focused on developing an imputation approach to incomplete longitudinal data. For general missing data problems, the other two common approaches are the likelihood-based methods and inverse probability weight (IPW) methods. Raghunathan [35] provided an overview of three approaches for general practitioners, while more technical aspects can be found in [36] and references thereof. Here we briefly comment on their connections and differences in general as well as in relevance to our context.

Let $Y_{\text{com}} = (Y_{\text{obs}}, Y_{\text{mis}})$ refer to the collection of observed data and missing data, and θ refer to the model parameters. Under MAR, the likelihood-based approaches obtain the maximum likelihood estimates of θ by integrating out the missing values (e.g. using EM algorithm) in the observed-data likelihood $L(\theta|Y_{\text{obs}}) \propto \int P(Y_{\text{obs}}, Y_{\text{mis}}|\theta) \mathrm{d}Y_{\text{mis}}$, where $P(Y_{\text{obs}}, Y_{\text{mis}}|\theta)$ is the imposed complete-data model. Based on the same complete-data model, multiple imputations of the missing values can be drawn from their posterior predictive distributions $P(Y_{\text{mis}}|Y_{\text{obs}})$ under some assumed prior distributions for θ. If the estimand of interest Q from the multiple imputation analysis is a function of θ, then multiple imputation provides approximate Bayesian inference for Q. Therefore, with a large sample and a diffuse prior distribution, multiple imputation and likelihood-based methods produce similar answers [37]. The advantage of multiple imputation over likelihood-based methods mainly lies on the flexibility of the former. Unlike the likelihood-based methods, multiple imputation separates the inferences into two phases: the imputation phase, in which the imputations are created, and the analysis phase, in which the completed-data inferences are obtained and combined. Because the phases are distinct, imputation and analysis maybe carried out on different occasions and by different persons. In addition, a variety of analyses can be applied to imputed datasets, not limited to the original form of θ.

In our problem, we expect that the likelihood analysis results would be similar to those under the multiple imputation. This is most obvious when the estimand of interest Q is the mixed-model slope coefficient θ in the simulation study. We note that AC is the maximum likelihood estimator. The corresponding results are similar to those of multiple imputation when the complete-data model matches with the data-generating model (i.e. LIN under linear curves and POLY under polynomial curves) and under the proposed functional mixed model. For the other two estimands, the difference in the mean and logit regression coefficient, the functional relationship between them and the mixed model coefficient exists but is less obvious. Since they are difficult to directly infer using the likelihood-based methods, multiple imputation analysis under an appropriate model appears to be a more desirable approach.

The IPW approach is based on modeling the missingness mechanism to obtain the estimated probability of being observed or missing. For the observed cases, estimation equations are applied yet weighted by the inverse of being respondent to obtain the consistent estimates. The idea originated from [38] in sample surveys, and developed by Robins *et al.* (e.g. [39, 40]) for more general settings. IPW may exhibit some robustness property, as it does not depend on the knowledge or assumption of the distribution of the observed data like likelihood-based methods/multiple imputation do. But the generic form of IPW may be inefficient since it only uses data from complete cases. In addition, problem can arise due to the fact that the resulting estimates might be unstable because certain subsets of the study sample might have small response probabilities so their inverse (weights) would be large [1]. To overcome the inefficiency of IPW, Robins and colleagues (e.g. [41]) proposed improved IPW estimates which are theoretically more efficient under the MAR assumption. A further development, the so-called doubly robust, or doubly protected estimators, are robust under certain conditions to misspecification of the model for the probability of responses. As yet, the technical nature of these methods and the lack of available generally applicable software have restricted their accessibility and uptake by the wider research community.

The application of IPW in our test scenarios is complicated by the long data series and nonmonotone missingness pattern. We believe it is beyond the scope of our paper. From our limited experiences, most of IPW applications for longitudinal data have been limited to the case of monotone missingness with a few time points. With monotone missingness, the response probability models can be decomposed into a series of logistic regressions of modeling response at wave j on the observed data at wave $j-1$, $j-2, \ldots$. The response weight at wave j is therefore the product of the response weights estimated from these logistic regressions. In a rather long series as in our example, we would expect that the weights for observed cases at later waves could get rather large and thus lead to a significant loss of efficiency and unstable results. With nonmonotone missingness, the model of probability of response at wave j is more difficult to specify, as the probability of being respondent can be associated with the observed data at both later time such as wave $j+1$ and earlier time. However, unlike the monotone missingness, subject with an observed value at wave j may have a missing value at wave $j-1$ and this further complicates the modeling. Another challenge lies in the variance estimation of IPW, which usually requires computationally intensive resampling methods such as Jackknife and bootstrap. There are a few literature comparing the IPW (and its double robust extensions) with multiple imputation under cross-sectional data [42], survival data [43], and longitudinal dropout [5, 22]. The results show that IPW tends to be less efficient and its validity relies on the correctness of the probability model for missingness, while the double robust extension of IPW shows edges over multiple imputation if the complete-data model is misspecified while the missingness probability model is correctly specified in the former approach. Therefore, extending IPW and its double robust extensions to more complicated longitudinal designs with missing data and comparing it with multiple imputation approach is of important interest.

References

1. Little RJA, Rubin DB. *Statistical Analysis of Missing Data*. Wiley: New York, 2002.
2. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. Wiley: New York, 1987.
3. Meng XL. Multiple-imputation inferences with uncongenial sources of input. *Statistical Science* 1994; **9**:538–558.
4. Harel O, Zhou XH. Multiple imputation: review of theory, implementation, and software. *Statistics in Medicine* 2007; **26**:3057–3077.
5. Liu M, Taylor JMG, Belin TR. Multiple imputation and posterior simulation for multivariate missing data in longitudinal studies. *Biometrics* 2000; **56**:1157–1163.
6. Schafer JL, Yucel RM. Computational strategies for multivariate linear mixed-effects models with missing values. *Journal of Computational and Graphical Statistics* 2002; **11**:437–457.
7. Ramsay JO, Silverman BW. *Applied Functional Data Analysis*. Springer: Berlin, 2005.
8. Guo W. Functional data analysis in longitudinal studies using smoothing splines. *Statistical Methods in Medical Research* 2004; **13**:49–62.
9. Rice JA. Functional and longitudinal data analysis: perspectives on smoothing. *Statistica Sinica* 2004; **14**:631–647.
10. Zhao X, Marron JS, Wells MT. The functional data analysis view of longitudinal data. *Statistica Sinica* 2004; **14**:789–808.
11. Guo W. Functional mixed effect models. *Biometrics* 2002; **58**:121–128.
12. Ruppert D, Wand MP, Carroll RJ. *Semiparametric Regression*. Cambridge University Press: Cambridge, 2003.
13. Longford NT. *Random Coefficient Models*. Oxford University Press: U.K., 1994.
14. Wahba G. Improper priors, spline smoothing and the problem of guarding against model errors in regression. *Journal of the Royal Statistical Society*, *Series B* 1978; **40**:364–372.

15. Green PJ. Penalized likelihood for general semi-parametric regression models. *International Statistical Review* 1987; **55**:245–260.
16. Zhang D, Lin X, Raz J, Sowers M. Semiparametric stochastic mixed models for longitudinal data. *Journal of the American Statistical Association* 1998; **93**:710–719.
17. Brumback BA, Rice JA. Smoothing spline models for the analysis of nested and crossed samples of curves (with Discussion). *Journal of the American Statistical Association* 1998; **93**:961–994.
18. Wang Y. Mixed effects smoothing spline analysis of variance. *Journal of the Royal Statistical Society*, *Series B* 1998; **60**:159–174.
19. Wang Y. Smoothing spline models with correlated random errors. *Journal of the American Statistical Association* 1998; **93**:341–348.
20. Verbyla P, Cullis BR, Kenward MG, Welham SJ. The analysis of designed experiments and longitudinal data by using smoothing splines (with Discussion). *Applied Statistics* 1999; **48**:269–311.
21. Wand MP, Jones MC. *Kernel Smoothing*. Chapman and Hall: London, 1995.
22. Morris JS, Carroll RJ. Wavelet-Based functional mixed models. *Journal of the Royal Statistical Society*, *Series B* 2006; **68**:179–199.
23. Antoniadis A, Spapatinas T. Estimation and inference in functional mixed-effects models. *Computational Statistics and Data Analysis* 2007; **51**:4793–4813.
24. Gelfand AE, Smith AFM. Sampling-based approaches to calculate marginal densities. *Journal of the American Statistical Association* 1990; **85**:398–409.
25. Chib S, Carlin BP. On MCMC sampling in hierarchical longitudinal models. *Statistics and Computing* 1999; **9**:17–26.
26. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences (with Discussion). *Statistical Science* 1992; **7**:457–511.
27. Brooks-Gunn J, Duncan GJ, Maritato N. Poor families, poor outcomes: the well-being of children and youth. In *Consequences of Growing up Poor*, Duncan GJ, Brooks-Gunn J (eds). Russell Sage Foundation: New York, 1997; 1–17.
28. Case A, Lubtosky D, Paxson C. Economic status and health in childhood: the origins of the gradient. *American Economic Review* 2002; **92**(5):1308–1334.
29. Demirtas H, Freels SA, Yucel RM. The plausibility of multivariate normality assumption when multiply imputing non-gaussian continuous outcomes: a simulation assessment. *Journal of Statistical Computation and Simulation* 2008; **78**:2069–2084.
30. Little RJA, An H. Robust likelihood-based analysis of multivariate data with missing values. *Statistica Sinica* 2004; **14**:949–968.
31. Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods* 2001; **6**:330–351.
32. Davidian M, Giltinan DM. *Nonlinear Models for Repeated Measurement Data*. Chapman and Hall/CRC: London/Boca Raton, 1995.
33. Daniels MJ, Hogan JW. *Missing Data in Longitudinal Studies*: *Strategies for Bayesian Modeling and Sensitivity Analysis*. Chapman and Hall/CRC: London/Boca Raton, 2008.
34. He H, Huang LS. Double-smoothing for bias reduction in local linear regression. *Journal of Statistical Planning and Inference* 2009; **139**:1056–1072.
35. Raghunathan TE. What do we do with missing data? Some options for analysis of incomplete data. *Annual Review of Public Health* 2004; **25**:99–117.
36. Molenberghs G, Kenward MG. *Missing Data in Clinical Studies*. Wiley: West Sussex, 2007.
37. Schafer JL. Multiple imputation in multivariate problems when the imputation and analysis model differ. *Statistica Neerlandica* 2003; **57**:19–35.
38. Horvitz DG, Thomospn DJ. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 1952; **47**:663–685.
39. Robins JM, Rotnitzky A. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association* 1995; **90**:122–129.
40. Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association* 1995; **90**:106–121.
41. Sharfstein DO, Rotnitzky A, Robins JM. Adjusting for nonignorable drop-out using semiparametric nonresponse models (with Discussion). *Journal of the American Statistical Association* 1999; **94**:1096–1146.
42. Carpenter JR, Kenward MG, Vansteelandt S. A comparison of multiple imputation and double robust estimation for analyses with missing data. *Journal of the Royal Statistical Society*, *Series A* 2006; **169**:571–584.
43. Qi L, Wang YF, He Y. A comparison of multiple imputation and fully augmented weighted estimators for Cox regression with missing covariate. *Statistics in Medicine* 2010; **29**:2592–2604.