Biometrika (2020), **103**, 1, *pp*. 1–13
Printed in Great Britain Advance Access publication on 31 July 2018

Functional principal component analysis with informative observation times

BY PEIJUN SANG
Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario N2L3G1, Canada
peijun.sang@uwaterloo.ca

DEHAN KONG
Department of Statistical Sciences, University of Toronto, Toronto, Ontario M5G 1X6, Canada
dehan.kong@utoronto.ca

AND SHU YANG
Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695, U.S.A.
syang24@ncsu.edu

SUMMARY

Functional principal component analysis has been shown to be invaluable for revealing variation modes of longitudinal outcomes, which serves as important building blocks for forecasting and model building. Decades of research have advanced methods for functional principal component analysis often assuming independence between the observation times and longitudinal outcomes. Yet such assumptions are fragile in real-world settings where observation times may be driven by outcome-related reasons. Rather than ignoring the informative observation time process, we explicitly model the observational times by a general counting process dependent on time-varying prognostic factors. Identification of the mean, covariance function, and functional principal components ensues via inverse intensity weighting. We propose using weighted penalized splines for estimation and establish consistency and convergence rates for the weighted estimators. Simulation studies demonstrate that the proposed estimators are substantially more accurate than the existing ones in the presence of a correlation between the observation time process and the longitudinal outcome process. We further examine the finite-sample performance of the proposed method using the Acute Infection and Early Disease Research Program study.

Some key words: Functional data analysis; Informative sampling; Missing at random.

1. INTRODUCTION

Longitudinal data have been extensively studied in the literature of statistics. Our research is motivated by the investigation of the disease progression in HIV-positive patients. Highly active antiretroviral therapy (HAART) has been shown to be an effective treatment for HIV in improving the immunological function and delaying the progression to AIDS (Hecht et al., 2006). Our goal is to study the mean trend and variation mode of CD4 counts, an indicator

of immune function, over time after treatment initiation, which is of major importance: first, it depicts a whole picture of how the disease evolves over time and thus provides new insights into the treatment mechanism. Second, it enables the prediction of disease progression and helps patients manage the disease better. Third, such information can also be used to design optimal treatment regimes for better clinical outcomes (Guo et al., 2021).

Parametric random effect models (Laird & Ware, 1982) and generalized estimating equations models (Liang & Zeger, 1986) are commonly adopted to fit longitudinal data; see Diggle et al. (2002) for a comprehensive overview. Though the latter does not need to specify the parametric distribution of the longitudinal response, it imposes a specific form of the mean response. To better understand the complexity of real-world data, Figure 2(a) shows the trajectories of CD4 counts at the follow-up visits from five randomly selected patients in the motivating application, where the number and timing of the visits differ from one to the next. Extracting useful information from such data has become a challenging statistical problem.

Functional data analysis offers a nonparametric means to modeling longitudinal data at irregularly spaced times. Repeated measurements of a longitudinal response from a subject are regarded as sparsely sampled from a continuous random function subject to measurement errors. Moreover, the underlying true random function is typically modeled in a nonparametric manner, thus avoiding model misspecification suffered from the two aforementioned approaches. To estimate the mean and covariance functions of the underlying continuous function from sparse observations, existing approaches usually assume that the observation times are independent of the longitudinal responses and then apply nonparametric smoothing techniques such as kernel smoothing to the aggregated observations from all subjects; see Yao et al. (2005), Li & Hsing (2010) and Zhang & Wang (2016) for instance.

Yet the independence assumption of the observation times and responses is restrictive in practice; e.g., patients with deteriorative health conditions may be more likely to visit the health care facilities (Phelan et al., 2017). Without addressing the informative observation time process, the study results can be biased and misleading (Lin et al., 2004; Sun et al., 2021). Xu et al. (2024) considered using a marked point process to model the informative visit times in longitudinal studies. But their work assumes that both the longitudinal outcome process and the latent process used to define the intensity function of the point process are Gaussian, which may not hold in practice. To address the same issue, Weaver et al. (2023) assumed that both the intensity function of the point process and the longitudinal outcome process depend on a positive latent factor. This assumption is slightly restrictive and can hardly be verified since it implies the dependence between observation times and the longitudinal outcome can be completely explained through this single latent factor. In this article, we propose to model the observation time process by a general counting process with an intensity function depending on time-dependent confounders. But it should be noted that the time-dependent confounders can be just functions of the observed outcome themselves. To account for the effect of the observation time process when estimating the mean function, we leverage the inverse of the intensity function at each observation time point as its weight and then apply penalized B-spline functions to the aggregated observations. This idea is further extended to estimating the covariance function with the tensor product of B-spline bases, weighted by a product of the inverse of the intensity functions at the two time points, to correct the selection bias of the pairs of observations. Variation modes can thus be visualized through an eigen-decomposition on the estimated covariance function, which is referred to as functional principal component analysis.

The proposed functional principal component analysis accounts for the dependence between the response process and the observation time process via inverse intensity function weighting. This fills an important gap in the literature as traditional approaches often assume that response

observations are independent of the observation times, which however is likely to be violated in real-world studies. Moreover, we establish consistency and convergence rates of our proposed estimator when estimating the mean function, covariance function, and functional principal components of a random function. Numerical studies demonstrate that in contrast to the traditional approaches, our approach can yield consistent estimates when the response observation times are indeed correlated with the underlying response process.

2. Basic Setup

2.1. *Functional principal component analysis and observation time process*

Suppose that X is a random function defined on a compact set $\mathbb{I} \subset \mathbb{R}$. Let $L^2(\mathbb{I})$ denote the collection of measurable square-integrable functions on $\mathbb{I}$. Furthermore, we assume that $\int_{\mathbb{I}} \mathbb{E}\{X^2(t)\}\mathrm{d}t < \infty$. Let $\mu(t) = \mathbb{E}\{X(t)\}$ and $C(s,t) = \mathrm{cov}\{X(s), X(t)\}$ denote the mean function and the covariance function of X, respectively. Then we can define the covariance operator $C : L^2(\mathbb{I}) \to L^2(\mathbb{I})$ that satisfies $(Cf)(t) = \int_{\mathbb{I}} C(s,t)f(s)\mathrm{d}s$ for any $f \in L^2(\mathbb{I})$. It follows from Mercer's theorem that there exists an orthonormal basis $(\varphi_j)_j$ of $L^2(\mathbb{I})$ and a sequence of nonnegative decreasing eigenvalues $(\kappa_j)_j$ such that $C(s,t) = \sum_{j=1}^{\infty} \kappa_j \varphi_j(s)\varphi_j(t)$. The eigenfunctions of C, φ_j's are also referred to as functional principal components of X. In fact, X admits the following Karhunen-Loève expansion, $X(t) = \mu(t) + \sum_{j=1}^{\infty} \zeta_j \varphi_j(t)$, where $\zeta_j = \int_{\mathbb{I}}\{X(t) - \mu(t)\}\varphi_j(t)\mathrm{d}t$ is called the jth functional principal component score of X and satisfies $\mathbb{E}(\zeta_j\zeta_k) = \delta_{jk}\kappa_j$, where $\delta_{jk} = 1$ if $j = k$ and 0 otherwise. The expansion is useful to approximate an infinite-dimensional random function because approximating $X(t)$ by $\mu(t) + \sum_{j=1}^{p} \zeta_j\varphi_j(t)$ yields the minimal mean squared error when using an arbitrary orthonormal system consisting of p functions for any $p \in \mathbb{N}^+$. Additionally, functional principal component analysis enables us to understand variation modes of this random function, as it displays the greatest variations along with the directions of principal components.

In practice, a fully observed trajectory of a random function may not be accessible due to various practical hurdles and is only observed at sparsely and irregularly spaced time points. To describe the irregularly-spaced observation time process for observing $X_i(t)$, let the set of visit times be $0 \le t_{i1} < \ldots < t_{im_i} \le \tau$, where m_i is the total number of observations, and τ denotes the predetermined study end time. Therefore, the domain of the random function $X(t)$ is $\mathbb{I} = [0, \tau]$. In stark contrast to the regular time setting, the observed time points are allowed to vary from one subject to another. Let $X_{ij} = X_i(t_{ij}) + \epsilon_{ij}$ denote the noisy observation of the ith random function at time t_{ij}, where ϵ_{ij} is the measurement error. Our primary interest is to perform functional principal component analysis from observations $\{X_{ij} : j = 1, \ldots, m_i, i = 1, \ldots, n\}$. Yao et al. (2005) and Li & Hsing (2010) address this problem under the assumption that the observed time points are independently and identically distributed, and X_{ij}, t_{ij} and m_i are independent of each other for subject i. However, in practice, whether or not there exists an observation at one particular time point often depends on the response process. Therefore, analysis of such data requires assumptions on the mechanism for the observation time process.

2.2. *Informative observation times*

Let $N_i(t)$ be the general counting process for the observation times; that is, $N_i(t) = \sum_{j=1}^{\infty} I(t_{ij} \le t)$ for $t \in [0, \tau]$. We use overline to denote the history; e.g., $\overline{X}_i(t) = \{X_i(u) : 0 \le u \le t\}$ is the history of the stochastic process X until time t for the ith subject. It is possible that the dependence between the longitudinal outcome and the observation times can be explained by merely using functions of the observed outcomes. Next we focus on a more complicated sce-

nario, where an auxiliary process is also involved in inducing the dependence of the outcome and observation times. In addition to the response process, we also observe a covariate process $Z_i(t)$ that is related $X_i(t)$ and $N_i(t)$, which can be multivariate, time-independent, or time-varying. Let $\overline{X}_i^{\rm obs}(t) = \{X_i(s) : \mathrm{d}N_i(s) = 1, 0 \le s \le t\}$ and $\overline{N}_i(t) = \{N_i(s) : 0 \le s \le t\}$ be the history of observed variables and observation times through t, respectively. We denote the observed history of variables for subject i at time t as $\overline{O}_i(t) = \{\overline{X}_i^{\rm obs}(t-), \overline{N}_i(t-), \overline{Z}_i^{\rm obs}(t-)\}$, where $t-$ indicates the time up to but excluding t. We use $\mathcal{F}_{it}^*$ to denote the filtration generated by $\overline{O}_i(t)$ and $X_i(t)$, and $E\{\mathrm{d}N_i(t) \mid \mathcal{F}_{it}^*\}$ denotes the estimated number of observation made in $[t, t+\mathrm{d}t)$, given the observed history up to t for the ith subject. Let $\lambda\{t \mid \overline{O}_i(t)\} = \mathbb{E}\{\mathrm{d}N_i(t) \mid \overline{O}_i(t)\}/\mathrm{d}t$ denote the conditional intensity function of $N_i(t)$ given the observed history up to t, but not including t. For the above notations, we suppress i to denote their population counterparts. In practice, the irregular observation times can be due to a number of reasons that may be related to subjects' responses, in which case, we say that the observation times are informative. In this case, ignoring the observation time process leads to biased results for the response variable. Similar to the missing data literature, we require a further assumption to identify the mean and variance functions of $X(t)$ under an informative observation time process.

Assumption 1. (i) $\mathbb{E}\{\mathrm{d}N_i(t) \mid \overline{O}_i(t), X_i(t)\} = \mathbb{E}\{\mathrm{d}N_i(t) \mid \overline{O}_i(t)\}$, and (ii) $\lambda\{t \mid \overline{O}_i(t)\} > 0$ almost surely for $i = 1, \ldots, n$.

Assumption 1(i) implies that the observed history collects all prognostic variables that affect the observation time process. It is plausible when $\overline{O}_i(t)$ includes the past observed responses $\overline{X}^{\rm obs}(t-)$, historical observation pattern $\overline{N}(t-)$, and past observed important auxiliary confounder process $\overline{Z}^{\rm obs}(t-)$ that is related to both observation time and response. Assumption 1(ii) suggests that all subjects have a positive probability of visiting at any time t. Assumption 1 is key toward identification, see §2.3; however, it is not verifiable based on the observed data and thus requires careful consultation of subject matter knowledge.

2.3. *Identification via inverse intensity function weighting*

We show that Assumption 1 leads to the identification of $\mu(t)$, $C(t,s)$, and $\varphi_j(t)$ by providing a brief outline of the proof below, while a detailed proof is deferred to §S.1 in the supplementary material. First, by the law of total expectation, we have for any $t \le \tau$,

$$\mathbb{E}[X(t)\lambda^{-1}\{t \mid \overline{O}(t)\}\mathrm{d}N(t)] = \mathbb{E}[X(t)\lambda^{-1}\{t \mid \overline{O}(t)\}\mathbb{E}\{\mathrm{d}N(t) \mid \mathcal{F}_t^*\}] = \mu(t)\mathrm{d}t. \quad (1)$$

Weighting by $\lambda^{-1}\{t \mid \overline{O}(t)\}$ serves to create a pseudo-population in which the observation time process is no longer associated with $X(t)$ as if the observed responses were sampled completely at random. Thus, $\mu(t)$ is identifiable.

Next, assuming $s < t$ and by the double use of the law of total expectation, we have

$$\begin{aligned}
&\mathbb{E}[\{X(t)-\mu(t)\}\{X(s)-\mu(s)\}\lambda^{-1}\{s \mid \overline{O}(s)\}\mathrm{d}N(s)\lambda^{-1}\{t \mid \overline{O}(t)\}\mathrm{d}N(t)] \\
&= \mathbb{E}[\{X(t)-\mu(t)\}\{X(s)-\mu(s)\}\lambda^{-1}\{s \mid \overline{O}(s)\}\mathrm{d}N(s)\lambda^{-1}\{t \mid \overline{O}(t)\}\mathbb{E}\{\mathrm{d}N(t) \mid \mathcal{F}_t^*\}] \\
&= \mathbb{E}[\{X(t)-\mu(t)\}\{X(s)-\mu(s)\}\lambda^{-1}\{s \mid \overline{O}(s)\}\mathrm{d}N(s)\mathrm{d}t] \\
&= \mathbb{E}[\{X(t)-\mu(t)\}\{X(s)-\mu(s)\}\lambda^{-1}\{s \mid \overline{O}(s)\}\mathbb{E}\{\mathrm{d}N(s) \mid X(t), \mathcal{F}_s^*\}\mathrm{d}t] \\
&= \mathbb{E}[\{X(t)-\mu(t)\}\{X(s)-\mu(s)\}\mathrm{d}t\mathrm{d}s] = C(t,s)\mathrm{d}t\mathrm{d}s. \quad (2)
\end{aligned}$$

Weighting by $\lambda^{-1}\{s \mid \overline{O}(s)\}\lambda^{-1}\{t \mid \overline{O}(t)\}$ serves to create a pseudo-population in which the observation time process is no longer associated with $\{X(t)-\mu(t)\}\{X(s)-\mu(s)\}$. Hence, $C(t,s)$ is identifiable.

3. ESTIMATION

In practice, the intensity function for the observation time process is unknown and requires modeling and estimation. Following Lin et al. (2004) and Yang et al. (2018, 2020), we assume the intensity function follows a proportional intensity function $\lambda\{t \mid \overline{O}(t)\} = \lambda_0(t)\exp[g\{\overline{O}(t)\}^{\mathrm{T}}\beta]$, where $g(\cdot)$ is a pre-specified multivariate function of $\overline{O}(t)$. Let $\theta = \{\lambda_0(t), \beta\}$. Under Assumption 1, the estimator of θ, denoted by $\widehat{\theta} = \{\widehat{\lambda}_0(t), \widehat{\beta}\}$, can be obtained from the standard software.

We treat the estimated intensity function, $\{\widehat{\lambda}_0(t_{ij})\}^{-1}\exp[-g\{\overline{O}_i(t_{ij})\}^{\mathrm{T}}\widehat{\beta}]$, as the sampling weight of X_{ij}. To estimate the mean function, because we cannot accurately recover each trajectory of X_i from sparse and noisy observations, we propose using weighted penalized splines to borrow information from aggregated observations from all subjects. In particular, let $0 = \xi_0 < \xi_1 \le \cdots \le \xi_K < \xi_{K+1} = \tau$ be a sequence of knots. The number of interior knots $K = K_n = n^{\eta}$ with $0 < \eta < 0.5$ being a positive integer such that $\max_{1\le k\le K+1}|\xi_k - \xi_{k-1}| = O(n^{-\eta})$. Let $\mathcal{S}_n$ be the space of polynomial splines of order $l \ge 1$ consisting of functions h satisfying: (i) in each subinterval, h is a polynomial of degree $l-1$; and (ii) for $l \ge 2$ and $0 \le l' \le l-2$, h is l' times continuously differentiable on $[0,\tau]$. Let $\{B_j(\cdot), 1 \le j \le q_n\}$, $q_n = K_n + l$, be the normalized B-spline basis functions of $\mathcal{S}_n$. Then for any $h \in \mathcal{S}_n$, there exists $\gamma = (\gamma_1, \ldots, \gamma_{q_n})^{\mathrm{T}} \in \mathbb{R}^{q_n}$ such that $h(t) = \sum_{j=1}^{q_n}\gamma_j B_j(t) = \gamma^{\mathrm{T}}B(t)$ for $t \in [0,\tau]$. To account for the effect of the observation time process on estimating the mean function, we define the weight

$$w_{ij}(\mu) = \{\widehat{\lambda}_0(t_{ij})\}^{-1}\exp[-\widehat{\beta}^{\mathrm{T}}g\{\overline{O}_i(t_{ij})\}] \tag{3}$$

for the jth observation of the ith subject, where $j = 1, \ldots, m_i$ and $i = 1, \ldots, n$. Let m be a positive integer, smaller than l. Suppose the penalty term in the penalized splines is

$$\int_0^{\tau}\gamma^{\mathrm{T}}\left\{B^{(m)}(t)\right\}^{\otimes 2}\gamma\,\mathrm{d}t,$$

where $a^{\otimes 2} = aa^{\mathrm{T}}$ for any matrix or column vector a. Consequently, the penalty matrix is $Q_\mu = \int_0^{\tau}\{B^{(m)}(t)\}^{\otimes 2}\,\mathrm{d}t$. We then estimate the mean function by $\widehat{\mu}(t) = B^{\mathrm{T}}(t)\widehat{\gamma}_\mu$ with

$$\widehat{\gamma}_\mu = \underset{\gamma\in\mathbb{R}^{q_n}}{\arg\min}\left[\sum_{i=1}^{n}\sum_{j=1}^{m_i}\{X_{ij} - B(t_{ij})^{\mathrm{T}}\gamma\}^2 w_{ij}(\mu) + \frac{\lambda_\mu}{2}\gamma^{\mathrm{T}}Q_\mu\gamma\right], \tag{4}$$

where $\lambda_\mu > 0$ is a tuning parameter controlling the roughness of the estimated mean function.

Next, we present an estimator of the covariance function. Let $G_i(t_{ij}, t_{il}) = \{X_{ij} - \widehat{\mu}(t_{ij})\}\{X_{il} - \widehat{\mu}(t_{il})\}$ be the raw estimate of the covariance function evaluated at (t_{ij}, t_{il}). By (2), we introduce

$$w_{ijl}(C) = \{\widehat{\lambda}_0(t_{ij})\widehat{\lambda}_0(t_{il})\}^{-1}\exp\left(-\widehat{\beta}^{\mathrm{T}}[g\{\overline{O}_i(t_{ij})\} + g\{\overline{O}_i(t_{il})\}]\right), \tag{5}$$

where $j, l = 1, \ldots, m_i$ and $i = 1, \ldots, n$, to account for the effect of the observation time process on estimating the covariance function. We use the tensor product of $B_j(t)$'s to estimate this bivariate covariance function. More specifically, $C(t,s)$ is approximated by $\sum_{1\le j_1\le j_2\le q_n}\eta_{j_1j_2}B_{j_1}(t)B_{j_2}(s)$. To ensure that $C(t,s) = C(s,t)$, we require $\Xi = (\eta_{j_1j_2})$ to be a $q_n \times q_n$ symmetric matrix. Denote $D(t,s) = B(t)\otimes B(s)$, which is a vector of length q_n^2, and $\eta = \mathrm{vec}(\Xi)$. Then the estimated covariance function is $\widehat{C}(t,s) = \sum_{j_1,j_2=1}^{q_n}\widehat{\eta}_{j_1j_2}B_{j_1}(t)B_{j_2}(s)$,

where $\widehat{\Xi}$ is obtained by solving the minimization problem:

$$\widehat{\Xi} = \underset{\Xi=\Xi^{\mathrm{T}}}{\arg\min} \left[\sum_{i=1}^{n} \sum_{1\le j\neq l\le m_i} \left\{ G_i(t_{ij}, t_{il}) - \sum_{j_1,j_2=1}^{q_n} \eta_{j_1 j_2} B(t_{ij}) B(t_{il}) \right\}^2 w_{ijl}(C) + \frac{\lambda_C}{2} \eta^{\mathrm{T}} Q_C \eta \right]. \tag{6}$$

Here Q_C is a $q_n^2 \times q_n^2$ penalty matrix with (j_1, j_2)th entry being

$$\int_0^\tau \int_0^\tau \left\{ \sum_{i+j=m} \binom{m}{i} \partial^i D_{j_1}(t,s) \partial^j D_{j_2}(t,s) \right\} \mathrm{d}t\mathrm{d}s$$

(Lai & Wang, 2013), and $\lambda_C > 0$ is a tuning parameter that controls the trade-off between fidelity to the data and plausibility of $C(t,s)$. The functional principal components are then estimated by solving

$$\int_0^\tau \widehat{C}(s,t)\varphi_j(s)\mathrm{d}s = \widehat{\kappa}_j \varphi_j(t) \tag{7}$$

subject to $\int_0^\tau \varphi_j^2(t)\mathrm{d}t = 1$ and $\int_0^\tau \varphi_j(t)\varphi_k(t)\mathrm{d}t = 0$ when $j \neq k$. Detailed steps for solving these equations can be found in Chapter 8.4 of Ramsay & Silverman (2005).

In the following numerical implementations, we take $m = 2$. The generalized cross-validation is used to choose the tuning parameters λ_μ, λ_C and the number of basis function q_n. In particular, let Y denote the vector of length $N = \sum_{i=1}^n m_i$ consisting of observations $X_{ij}, j = 1, \ldots, m_i, i = 1, \ldots n$. Let $W = \mathrm{diag}\{w_{ij}(\mu), j = 1, \ldots, m_i, i = 1, \ldots, n\}$ and $Y_w = W^{1/2}Y$. According to Chapter 3 of Gu (2013), the generalized cross-validation score for (4) is

$$V_\mu(\lambda_\mu) = \frac{N^{-1} Y_w^{\mathrm{T}} \{I - A_w(\lambda_\mu)\}^2 Y_w}{[N^{-1}\mathrm{tr}\{I - A_w(\lambda_\mu)\}]^2},$$

where I is an $N \times N$ identity matrix and $A_w(\lambda_\mu)$ is the so-called smoothing matrix satisfying $\widehat{Y}_w = W^{1/2}\widehat{Y} = A_w(\lambda_\mu)Y_w$. An explicit form of $A_w(\lambda_\mu)$ can be found in (3.12) of Gu (2013). We select λ_μ by minimizing $V_\mu(\lambda_\mu)$. The smoothing parameter λ_C for the covariance function estimate defined in (6) is chosen in a similar manner. Moreover, the number of basis functions q_n is selected by gradually increasing its value in a grid until it leads to a significant decrease in the generalized cross-validation score; see §S.2.2 of the supplementary material for details.

4. THEORETICAL PROPERTIES

4.1. *Large sample properties of the mean function estimator*

For $d \in N^+$, let $C^d([0,\tau])$ denote the class of functions with continuous dth derivatives over $[0,\tau]$. Without loss of generality, we assume $\tau = 1$ in the theoretical analysis. Below, we present the regularity assumptions for deriving the large sample properties for proposed mean and covariance function estimators, as well as the estimated functional principal components.

Assumption 2. The true mean function of X, $\mu(\cdot)$, belongs to $C^d([0,\tau])$ for some $d \geq \max(2, m)$.

Assumption 3. The knots are equally spaced in $\mathcal{S}_n$. The order of the spline functions satisfies $l \geq d$ and $l > m$.

Remark 1. Assumptions 2 and 3 ensure that there exists a spline function $\tilde{\mu}(\cdot) = B^{\mathrm{T}}(\cdot)\tilde{\gamma} \in \mathcal{S}_n$ such that $\|\mu - \tilde{\mu}\|_\infty = O(q_n^{-d})$. The equal-spaced knots assumption is used for the convenience of deriving the decay rate; see Proposition 4.2 of Xiao (2019). Proof of this result is similar to that of Lemma 1 of Smith & Barrow (1979) and is omitted here. Without the equal-spaced knots assumption, deriving the decay rate of the eigenvalues of the relevant penalty matrix would be more challenging. This is left for future research.

Assumption 4. Therefore exists some constant $\delta > 2$ such that $\mathbb{E}(\|X\|_\infty^\delta) < \infty$.

Assumption 5. The random errors ϵ_{ij}'s are independent and identically distributed with mean 0 and $\mathbb{E}(\|\epsilon\|_\infty^\delta) < \infty$, where δ is defined in Assumption 4 and ϵ denotes the random process of the error.

Establishing the uniform convergence rate on the estimated mean function entails strong moment conditions on X and ϵ as in Assumptions 4 and 5. Similar assumptions are considered in Li & Hsing (2010) and Zhang & Wang (2016).

Assumption 6. In the intensity function $\lambda(t) = \lambda_0(t)\exp[g\{\overline{O}(t)\}^{\mathrm{T}}\beta_0]$, $\lambda_0(t)$ belongs to $C^p([0,\tau])$ for some $p \geq d$ and is strictly positive, and $g\{\overline{O}(t)\}$ is almost surely bounded over $[0,\tau]$.

This assumption specifies a smoothness property for the baseline intensity to ensure that a desirable convergence rate can be achieved when replacing the true intensity function with the estimated one in (3). This assumption is commonly adopted in a semiparametric Cox model (Cox, 1972) for modeling the intensity function for a counting process. We can estimate β by the partial likelihood approach, the cumulative baseline intensity function $\Lambda_0(t) = \int_0^t \lambda_0(s)\mathrm{d}s$ by Breslow's estimator, and further $\lambda_0(t)$ by a kernel-smoothed estimator defined in (S18). More details can be found in §S.2 in the supplementary material. Under Assumption 6, according to Andersen et al. (1993), $\hat{\beta}$ is $\sqrt{n}$-consistent, and $\hat{\lambda}_0(t)$ is a consistent estimator of λ with rate $n^{-p/(2p+1)}$, if the bandwidth h_n satisfies $h_n \asymp n^{-1/(2p+1)}$ and the kernel K is of order $\lfloor p \rfloor$, which denotes the greatest integer strictly less than p (Tsybakov, 2009, p. 5). Under this assumption, the convergence rate of $\hat{\lambda}(t)$ is no slower than the uniform convergence rate given in Theorem 1. Consequently, this semiparametric estimate of $\lambda(t)$ will not affect the uniform convergence rate of the proposed mean and/or covariance function. The following theorem establishes the uniform convergence rate for the proposed mean function estimator.

THEOREM 1. *Assume Assumptions 1–6 hold. Then the estimated mean function* $\widehat{\mu}(t) = B(t)^{\mathrm{T}}\widehat{\gamma}_\mu$*, where* $\widehat{\gamma}_\mu$ *is defined in* (4)*, satisfies*

$$\sup_{t\in[0,\tau]} |\widehat{\mu}(t) - \mu(t)| = O_P\left\{ q_n^{-d} + \lambda_\mu q_n^m + \left(\frac{q_n \log n}{n}\right)^{\frac{1}{2}} \right\}, \tag{8}$$

provided that $\lambda_\mu q_n^{2m} = O(1), q_n^\delta = O\{(n/\log n)^{\delta-2}\}$ *and* $\log n/n = o(q_n^{-4})$.

Remark 2. If $q_n \asymp (n/\log n)^{1/(1+2d)}$ and $\lambda_\mu = o\{q_n^{-(d+m)}\}$, the uniform convergence rate of $\widehat{\mu}$ is $O_P\{(n/\log n)^{-d/(1+2d)}\}$. Our mean function estimator achieves the optimal convergence rate $\{n/\log(n)\}^{-d/(2d+1)}$, established in Stone (1982) for independent and identically distributed data and in Li & Hsing (2010) for sparse functional data with the assumption that the observational times are independent of the functional data.

To derive the convergence rate for the proposed covariance function estimator, we further need the following assumption.

Assumption 7. The true covariance function of X, $C(\cdot,\cdot)$, belongs to $C^d([0,\tau]^2)$.

Similarly to the mean function, by the result on p.149 of De Boor (1978), Assumption 7 leads the existence of $\tilde{\eta} \in \mathbb{R}^{q_n^2}$ such that

$$\sup_{(t,s)\in[0,\tau]^2} |C(t,s) - D^{\mathrm{T}}(t,s)\tilde{\eta}| = O(q_n^{-d}).$$

The following theorem establishes the convergence rate for the proposed covariance function estimator.

THEOREM 2. *Assume Assumptions 1–7 hold with some $\delta > 4$ in Assumptions 4 and 5. The estimated covariance function $\widehat{C}(t,s) = D(t,s)^{\mathrm{T}}\widehat{\eta}$, where $\widehat{\eta} = \mathrm{vec}(\widehat{\Xi})$ defined in* (6)*, satisfies*

$$\sup_{(t,s)\in[0,\tau]^2} |\widehat{C}(t,s) - C(t,s)| = O_P\left\{ q_n^{-d} + \lambda_C q_n^m + \left(\frac{q_n^2 \log n}{n}\right)^{\frac{1}{2}} \right\},$$

provided that $\lambda_C q_n^{2m} = O(1), q_n^{\delta} = O\{(n/\log n)^{(\delta-2)/2}\}$ and $\log n/n = o(q_n^{-4})$.

Remark 3. If $q_n \asymp (n/\log n)^{1/(2d+2)}$ and $\lambda_C = O(q_n^{-m-d})$, then the uniform convergence rate of $\hat{C}$ is $O_P\{(n/\log n)^{-d/(2d+2)}\}$. In other words, the uniform convergence rate of the covariance function estimator is the same as the optimal rate established in Stone (1982) for independent and identically distributed data and in Li & Hsing (2010) for sparse functional data with the assumption that the observational times are independent of the functional data.

COROLLARY 1. *Under the same assumptions of Theorem 2, $q_n \asymp n^{1/(4d+2)}$ and $\lambda_C = O(q_n^{-m-d})$, for $1 \le j \le j_0$ satisfying $\kappa_1 > \cdots > \kappa_{j_0} > \kappa_{j_0+1} \ge 0$, we have*

$$|\widehat{\kappa}_j - \kappa_j| = O_P(n^{-\frac{1}{2}}) \quad \textit{and} \quad \left\{\int_0^\tau |\widehat{\varphi}_j(t) - \varphi_j(t)|^2 \mathrm{d}t\right\}^{\frac{1}{2}} = O_P(n^{-\frac{d}{2d+1}}),$$

where $\widehat{\kappa}_j$ and $\widehat{\varphi}_j$ denote the jth eigenvalue and eigenfunction of $\widehat{C}(t,s)$, respectively.

This conclusion is similar to Theorem 1 of Hall et al. (2006), which is a refined result of Theorem 2 of Yao et al. (2005).

5. SIMULATION STUDIES

5.1. Simulation design

The simulated response process $\{X_i(t) : i = 1,\ldots,n\}$ is generated by $X_i(t) = \sin(t + 1/2) + \sum_{k=1}^{50} \nu_k \zeta_{ik} \varphi_k(t) + \epsilon_i(t)$ for $t \in [0,\tau]$ with $\tau = 3$, where $\nu_k = (-1)^{k+1}(k+1)^{-1}$, ζ_{ik}'s are independently following a uniform distribution over $[-\sqrt{3},\sqrt{3}]$ and $\varphi_k(t) = \sqrt{2/3}\cos(k\pi t)$ for $k \ge 1$ and $\epsilon_i(t)$'s are independently normally distributed across both i and t, with mean 0 and variance 0.01. We consider the following design for observation times. The observation times of $X_i(\cdot)$ are generated sequentially by a general counting process with the intensity function $\lambda\{t \mid \overline{X}_i^{\mathrm{obs}}(t-)\} = \exp\{2X_i^{\mathrm{obs}}(t-)\}$. This design leads to sparse observations of $X_i(t)$ with an average of 11.5 observations on each trajectory. We vary the sample size from $n = 100$ to $n = 200$.

We compare the proposed estimators with the unweighted functional principal component analysis (Yao et al., 2005) without adjusting for the informative observation time process. For a fair comparison, we use the penalized spline for smoothing instead of the original local linear smoother proposed by Yao et al. (2005). For the proposed estimators, we consider both cases

when the true intensity function is known or estimated to examine the impact of intensity function estimation on subsequent analysis. We report the mean integrated squared errors for the estimated mean function, covariance function, and first functional principal component, defined as $\int_0^\tau \{\widehat{\mu}(t) - \mu(t)\}^2 \mathrm{d}t$, $\int_0^\tau \int_0^\tau \{\widehat{C}(s,t) - C(s,t)\}^2 \mathrm{d}s\mathrm{d}t$, and $\int_0^\tau \{\widehat{\phi}_1(t) - \phi_1(t)\}^2 \mathrm{d}t$, respectively. We also report the bias and the standard error of the estimated first eigenvalue, denoted by $\widehat{\lambda}_1$.

Table 1: Mean integrated squared errors ($\times 0.01$) for the estimated mean function, covariance function, and first functional principal component. The actual numerical values are the ones displayed in the table multiplied by 0.01. Standard deviations are presented in the bracket.

n	$\widehat{\mu}(t)$			$\widehat{C}(s,t)$			$\widehat{\varphi}_1(t)$			$\widehat{\kappa}_1$		
	UW	TW	EW	UW	TW	EW	UW	TW	EW	UW	TW	EW
100	6.41	1.13	1.10	3.18	2.63	2.40	14.4	9.8	9.2	4.46	1.11	1.09
	(2.00)	(.67)	(.63)	(1.55)	(1.04)	(.85)	(8.4)	(5.3)	(4.6)	(4.73)	(3.96)	(3.81)
200	6.29	.83	.82	2.94	1.91	1.75	12.6	7.1	6.6	4.68	.50	.56
	(1.45)	(.46)	(.42)	(1.94)	(.56)	(.49)	(7.9)	(3.7)	(3.2)	(4.02)	(3.00)	(2.91)

UW denotes the unweighted method, TW denotes the proposed method assuming that the true intensity function is known, and EW denotes the proposed method where the intensity function is estimated.

Table 1 summarizes mean integrated squared errors over 200 Monte Carlo runs. Figure 1 plots the average of the estimated mean functions and the first functional principal components across 200 Monte Carlo replicates. The unweighted method shows clear biases in estimating the mean function and the first principal component, while our proposed weighted method can reduce the biases. Interestingly, the proposed estimators with estimated weights improve the counterparts with true weights in terms of the mean integrated squared errors of $\widehat{C}(s,t)$ and $\widehat{\varphi}_1(t)$; see Table 1. This phenomenon is similar to the inverse propensity weighting estimator of the average treatment effect, where one can achieve better efficiency by using the estimated propensity score instead of using the true score.

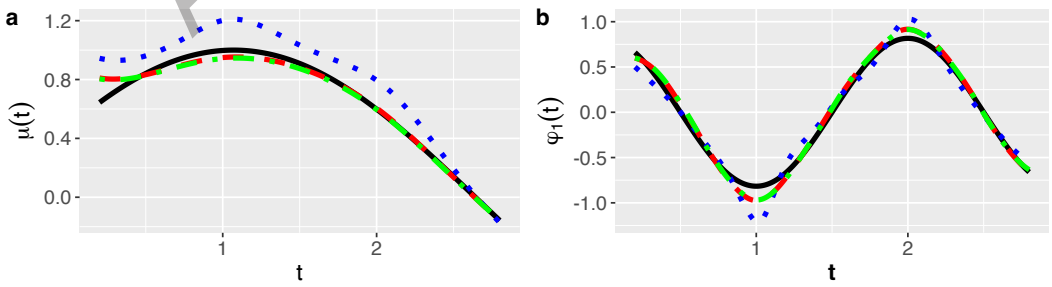

Fig. 1: Simulation results of the average of estimated mean function (Panel a) and the average of estimated first functional principal component (Panel b) across 200 Monte Carlo replicates. In both panels, the black solid line denotes the true function, while the red dashed, green dash-dotted, and blue dotted lines denote the estimates from the proposed method with estimated weights, the proposed method with true weights, and the unweighted method, respectively.

In addition, we consider various designs in §S.3 of the supplementary material:
1. $\lambda(t)$ depends on both an auxiliary process Z and the past history of X, and the true process X also depends on Z. Z can either be a null set or be a multivariate random vector or a stochastic

process.

2. The baseline intensity function $\lambda_0(t)$ can be set to be a constant or a linear function.

3. The observational time is independent of the response process.

For all these settings, our proposed method performs similarly to the comparison shown earlier; see §S.3 for details.

6. APPLICATION

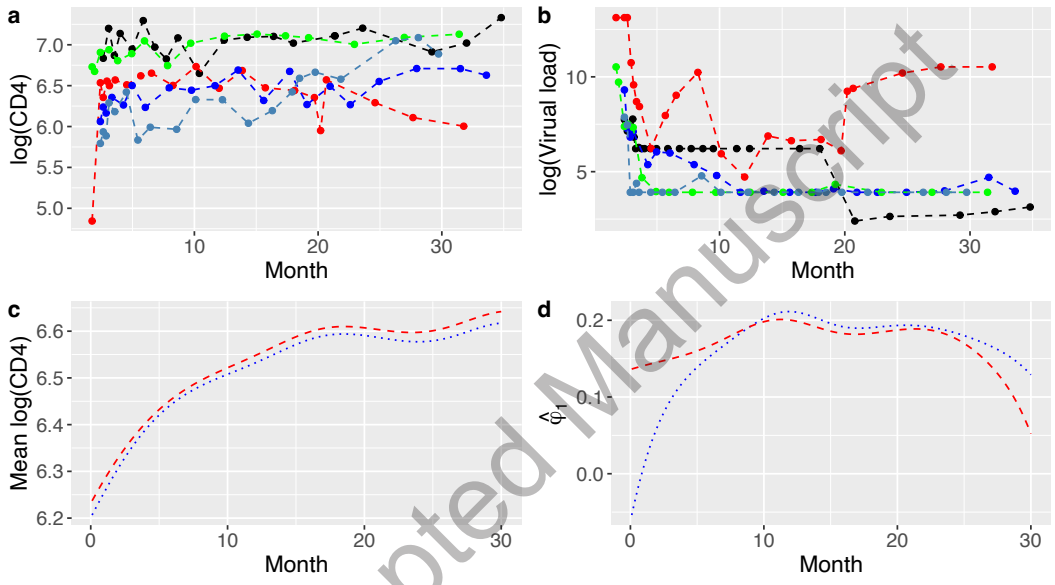

Fig. 2: (a) & (b) Trajectories of log CD4 counts and log viral load observed at irregularly spaced follow-up visits from 5 randomly selected patients. (c) & (d) Estimated mean function and the first eigenfunction of log CD4 counts from the unweighted method and the proposed weighted method. Here red dashed and blue dotted lines represent the estimates from the weighted and unweighted methods, respectively.

Most existing studies focused only on the treatment effect of highly active antiretroviral therapy on a clinical endpoint at a fixed time point, e.g. CD4 counts at two years after treatment initiation (Yang, 2021). On the contrary, our goal is to study the mean trend and variation mode of CD4 counts. The observational AIEDRP (Acute Infection and Early Disease Research Program) Core 01 study was established by Hecht et al. (2006). It established a cohort of newly infected HIV patients. The patients were protocolized to visit the physicians for outcome assessment such as CD4 count and viral load at weeks 2, 4, and 12, and then every 12 weeks thereafter, through week 96. In our analysis, we include 72 patients from the AIEDRP program who initiated HAART between 52 and 92 days after HIV diagnosis. These patients also had more than 2 visits during the study follow-up. The outcome of interest is log CD4 count, lower values meaning worse immunological function. A unique challenge arises due to substantial variability in the follow-up visit times at which patient outcomes were assessed. Figure 2(a) and (b) show the trajectories of log CD4 counts and log viral load at the follow-up visits from 5 randomly selected patients, respectively. The number and timing of visits differ from one patient to the next, resulting in irregularly spaced observations. Moreover, such irregular visit times can be due to obstacles that may be related to patients' health status and thus informative about the outcome of interest.

We apply the proposed method to estimate the mean trend and variation mode of log CD4 counts over time. To address the irregularly spaced and informative observation times, we model the intensity of visit times by a Cox proportional intensity function adjusting for log CD4 counts and log viral load at the closest past visit. The fitted result for the intensity function, presented in §S.4 of the supplementary material, shows that patients with lower CD4 counts and higher viral load are more likely to visit. Figure 2(c) displays the estimated mean functions of the log CD4 counts from unweighted and weighted analyses. The unweighted estimator shows persistently lower means than the weighted estimator over time. This is in line with the fitted result of the intensity function which suggests that the worse outcomes are more likely to be assessed and thus the unweighted estimator is biased downward. Figure 2(d) displays the estimated first eigenfunction which depicts the dominant mode of variation of CD4 counts. The weighted and unweighted analyses tend to agree on the variation mode after 10 months; however, there exist great discrepancies between them before 10 months. The weighted analysis uncovers the phase transitions of CD4 counts following treatment initiation: an immediate dramatic change, followed by a plateau between 10 months and 20 months, and a rebound after 20 months. Such transitions are reasonable because antiretroviral therapy promptly reduces the amount of HIV and helps recover the immune system and produce more CD4 cells, while drug resistance can be developed in extended long treatment uptake and affects CD4 counts to change.

For the sensitivity analysis of the intensity function for the observation times of CD4 counts, we fit another intensity function with $g(\overline{O}(t))$ taken as the log viral load and its square. This new intensity function leads to a similar estimate of the mean function and the first eigenfunction of log CD4 counts. More details can be found in §S.4 of the supplementary material.

7. Discussion

To handle the informative observation time process, we describe identifying assumptions that are tantamount to the missingness at random assumption; that is, the unobserved outcomes are unrelated to the probabilities of observations so long as controlling for observed information. Our weighting strategy can be readily extended to other functional principal component analyses, such as the Principal Analysis by Conditional Expectation proposed by Yao et al. (2005). Empirical results in §S.3.1 and §S.4 demonstrate similar performance to the proposed approach, while theoretical comparisons will be explored in future research. More robust and efficient estimation than weighting-alone estimators can be developed by using the augmentation of the conditional mean functions (Coulombe & Yang, 2024), which will be another interesting future research topic. In practice, if a prognostic variable that is related to the observation time process is not captured in the data, the observed information is not sufficient to explain away the dependence between the longitudinal outcomes and the observational time process, leading to observations not at random or missingness not at random (Pullenayegum & Lim, 2016; Sun et al., 2021). Because such assumptions are untestable, sensitivity analysis methodology is critically important for assessing the robustness of the study conclusion against violation of assumptions; however, no such methodology has been developed previously. In the future, we will develop a sensitivity analysis toolkit following (Robins et al., 1999; Yang & Lok, 2017; Smith et al., 2022) for functional data with irregular observation times.

Acknowledgement

The authors thank the editor, associate editor, and three referees for their constructive comments, which have substantially improved the paper. Sang and Kong's research are partially

supported by the Natural Science and Engineering Research Council of Canada. Yang's research is partially supported by the U.S. National Institutes of Health and National Science Foundation.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes the proofs of the identification results and theorems, additional details of numerical implementations, and additional results of numerical studies in the main manuscript. The code to implement the method can be found at https://github.com/spj1125/FPCA.

REFERENCES

ANDERSEN, P. K., BORGAN, O., GILL, R. D. & KEIDING, N. (1993). *Statistical Models Based on CountingProcesses*. Springer, New York.

COULOMBE, J. & YANG, S. (2024). Multiply robust estimation of marginal structural models in observational studies subject to covariate-driven observations. *Biometrics* .

COX, D. R. (1972). Regression models and life-tables. *J. R. Stat. Soc. Ser. B* **34**, 187–202.

DE BOOR, C. (1978). *A Practical Guide to Splines*. Springer, New York.

DIGGLE, P., DIGGLE, P. J., HEAGERTY, P., LIANG, K.-Y. & ZEGER, S. (2002). *Analysis of Longitudinal Data*. Oxford University Press.

GU, C. (2013). *Smoothing Spline ANOVA Models 2nd edition*. Springer, New York.

GUO, W., ZHOU, X.-H. & MA, S. (2021). Estimation of optimal individualized treatment rules using a covariate-specific treatment effect curve with high-dimensional covariates. *J. Amer. Statist. Assoc.* **116**, 309–321.

HALL, P., MÜLLER, H.-G. & WANG, J.-L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *Ann. Stat.* **34**, 1493–1517.

HECHT, F. M., WANG, L., COLLIER, A., LITTLE, S., MARKOWITZ, M., MARGOLICK, J., KILBY, J. M., DAAR, E., CONWAY, B. & NETWORK, A. (2006). A multicenter observational study of the potential benefits of initiating combination antiretroviral therapy during acute HIV infection. *J. Infect. Dis.* **194**, 725–733.

LAI, M.-J. & WANG, L. (2013). Bivariate penalized splines for regression. *Stat. Sin.* **23**, 1399–1417.

LAIRD, N. M. & WARE, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963–974.

LI, Y. & HSING, T. (2010). Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *Ann. Stat.* **38**, 3321–3351.

LIANG, K.-Y. & ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

LIN, H., SCHARFSTEIN, D. O. & ROSENHECK, R. A. (2004). Analysis of longitudinal data with irregular, outcome-dependent follow-up. *J. R. Stat. Soc. Ser. B* **66**, 791–813.

PHELAN, M., BHAVSAR, N. A. & GOLDSTEIN, B. A. (2017). Illustrating informed presence bias in electronic health records data: how patient interactions with a health system can impact inference. *eGEMs* **5**.

PULLENAYEGUM, E. M. & LIM, L. S. (2016). Longitudinal data subject to irregular observation: A review of methods with a focus on visit processes, assumptions, and study design. *Stat Methods Med Res* **25**, 2992–3014.

RAMSAY, J. O. & SILVERMAN, B. W. (2005). *Functional Data Analysis 2nd edition*. Springer, New York.

ROBINS, J. M., ROTNITZKY, A. & SCHARFSTEIN, D. O. (1999). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials*. Springer, New York, pp. 1–94.

SMITH, B., YANG, S., APTER, A. J. & SCHARFSTEIN, D. O. (2022). Trials with irregular and informative assessment times: a sensitivity analysis approach. *arXiv preprint arXiv:2204.11979* .

SMITH, P. & BARROW, D. (1979). Efficient L2 approximation by splines. *Numer Math* **33**, 101–114.

STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Stat.* **10**, 1040–1053.

SUN, Y., MCCULLOCH, C. E., MARR, K. A. & HUANG, C.-Y. (2021). Recurrent events analysis with data collected at informative clinical visits in electronic health records. *J. Amer. Statist. Assoc.* **116**, 594–604.

TSYBAKOV, A. B. (2009). *Introductin to Nonparametric estimators*. Springer, New York.

WEAVER, C., XIAO, L. & LU, W. (2023). Functional data analysis for longitudinal data with informative observation times. *Biometrics* **79**, 722–733.

XIAO, L. (2019). Asymptotic theory of penalized splines. *Electron. J. Stat* **13**, 747–794.

XU, G., ZHANG, J., LI, Y. & GUAN, Y. (2024). Bias-correction and test for mark-point dependence with replicated marked point processes. *J. Amer. Statist. Assoc.* **119**, 217–231.

YANG, S. (2021). Semiparametric estimation of structural nested mean models with irregularly spaced longitudinal observations. *Biometrics (to appear)* .

YANG, S. & LOK, J. J. (2017). Sensitivity analysis for unmeasured confounding in coarse structural nested mean models. *Stat. Sin.* **28**, 1703–1723.

YANG, S., PIEPER, K. & COOLS, F. (2020). Semiparametric estimation of structural failure time model in continuous-time processes. *Biometrika* **107**, 123–136.

YANG, S., TSIATIS, A. A. & BLAZING, M. (2018). Modeling survival distribution as a function of time to treatment discontinuation: A dynamic treatment regime approach. *Biometrics* **74**, 900–909.

YAO, F., MÜLLER, H.-G. & WANG, J.-L. (2005). Functional data analysis for sparse longitudinal data. *J. Amer. Statist. Assoc.* **100**, 577–590.

ZHANG, X. & WANG, J.-L. (2016). From sparse to dense functional data and beyond. *Ann. Stat.* **44**, 2281–2321.

[*Received on 2 January 2017. Editorial decision on 1 April 2017*]