Analysis of multivariate non-gaussian functional data: a semiparametric latent process approach

Jiakun Jiang^{a,b}, Huazhen Lin^{b,*}, Qingzhi Zhong^{b} , and Yi Li^{c}
^{a} Center for Statistics and Data Science, Beijing Normal University at Zhuhai, China.
^{b}Center of Statistical Research and School of Statistics, Southwestern University of Finance and Economics, Chengdu, Sichuan, China.
^{c}Department of Biostatistics, University of Michigan, Ann Arbor, USA.

Abstract

Commonly assumed for multivariate functional regression models are normality and structural dependence, which, however, may not hold in practice. To relax these restrictions, we propose a new semiparametric transformation latent process functional regression model for multivariate functional data. Our model does not require normality assumptions or any specific dependence structures among multivariate response curves or intra-individual variability across time. We propose a combined likelihood- and estimating equation-based method to estimate parameters, transformation functions and covariance structures. We establish theoretical properties, including $\sqrt{n}$−consistency and asymptotic normality, for the proposed estimators. The utility of the method is illustrated via extensive simulations and analyses of an elderly cognitive evolution dataset, which yield a better fit than the other competing methods and some interesting findings.

Keywords: Functional regression analysis, Latent process, Normal transformation model, Semi-parametric
AMS subject classifications: 62H25

1. Introduction

Multivariate functional data are commonly collected in psychological studies to measure unobservable outcomes, e.g. cognitive functions, with a series of measurements in response to a battery of tests administered over time. In this case, the latent cognitive process, viewed as a common cognitive factor across all of the psychometric tests, may better predict dementia and cognitive decline ([14, 43, 47]). It would be of substantial interest to focus on this latent process by describing its evolution as well as the impactful factors. However, analyses of this type of data are hampered by (a) unobservable latent cognitive evolution process; (b) unknown dependency among multivariate longitudinal or functional data; (c) unknown links between the latent process and the multivariate functional data; and (d) non-Gaussian multivariate response curves as reflected by the psychometric test results; see Figures 2-4 in the Supplementary Material. The goal of the paper is to propose a functional regression model which allowing non-Gaussian response curves and without the specifications of link functions and covariance structures.

Limited work has been done for modeling multivariate functional responses. Dunson ([11]) proposed a dynamic latent variable model (DLVM) in which each response is related to a latent variable through a generalized linear model, and the serial dependency is accounted by a linear transition structure model which stipulates that the latent traits linearly depend on their chistorical values and covariates. The articles [4, 9, 22] extended the DLVM to accommodate categorical data, survival data and mixed-type data. However, as [13] alluded to, with a discrete-time formulation, DLVM can only fit regularly balanced time series and may have limited usage for analyzing data collected at irregular and possibly subject-specific time points. An extension to these settings is difficult because the number of parameters increases with the number of subject-specific time points. Moreover, DLVM requires a Gaussian assumption on

[*] Corresponding author. Email address: linhz@swufe.edu.cn

the functional responses as well as a specified link between the responses and the latent process. In addition, [8] proposed a multivariate functional linear regression model in which both the response and predictor variables contain multivariate random trajectories; see [2, 16] for a comprehensive review.

To address some of these limitations, [35, 65] proposed functional linear mixed models; [20] related functional binary or count data to a latent Gaussian process; [50, 52, 56] applied a generalized Gaussian process regression model for non-Gaussian functional data; [25]developed a Gaussian latent process threshold model for longitudinal ordinal data; [43, 44] proposed a transformation latent process model for multivariate non-Gaussian longitudinal data. However, all of the methods require specifications of the covariance structures of the latent process and the links between the latent process and the functional responses, which may not be desirable as results are sensitive to these misspecifications.

We propose a semiparametric transformation latent process functional regression model for multivariate non-Gaussian functional data. By not specifying link functions and covariance structures, our model does not require normality assumptions or any specific dependence structures among multivariate response curves or intra-individual variability across time, and provides a convenient means to model the dependency among the multivariate non-Gaussian functional responses, and to explore the biologic processes governing the cognitive impairment. Furthermore, our model allows measurements to be taken at irregular and possibly subject-specific time points. With the added flexibility, our model has smaller out-of-sample prediction errors than the existing methods as shown in our analysis of the Cognitive Decline data, which motivated the proposed method; see Section 5 and Table 4.

The remainder is organized as follows. The proposed method and a two-stage estimation procedure are introduced in Section 2. The uniformly consistent and asymptotically normal properties are derived in Section 3. Section 4 contains simulation results and Section 5 presents an application to an elderly cognitive evolution study. Section 6 concludes the paper with concluding remarks. Technical proofs and related results are relegated to the Appendix and Supplementary Materials.

2. Model and Estimation

2.1. Model

Suppose that there are n independent subjects with observations $(\mathbf{Y}_i(t_{ij}), \mathbf{X}_{ij}, \mathbf{Z}_{ij}), i \in \{1, 2, \dots, n\},\ j \in \{1, 2, \dots, n_i\}$, where $\mathbf{Y}_i(t_{ij}) = (Y_{i1}(t_{ij}), \dots, Y_{ip}(t_{ij}))^\top$ are continuous outcomes of a p-dimensional vector measured at individual-specific time points t_{ij}, $\mathbf{X}_{ij}$ and $\mathbf{Z}_{ij}$ are respectively p_1 and p_2 dimensional covariates, representing two different sets of features. For example, $\mathbf{X}_{ij}$ are confounders and Z_{ij} are covariates of interest covariate, such as treatment assignment or dosage of regimen. Without loss of generality, we assume that the time points t_{ij}'s are a random sample from a certain population with a bounded support, say, $[0, 1]$. Also to facilitate large sample property derivations, we assume a random covariate design, that is, $(\mathbf{X}_{ij}, \mathbf{Z}_{ij})$ are random variables, jointly following an unspecified distribution. Denote by $\eta_i = (\eta_i(t))_{t\geq 0}$ the latent process (e.g. cognitive ability) for individual i with $i \in \{1, \dots, n\}$, and our goal is to describe its evolution over time and evaluate how the covariates, $\mathbf{Z}_{ij}$, may impact it. As opposed to the existing models and by relaxing the normality assumptions on $\mathbf{Y}_i(t)$, we consider a nonparametric transformation that will transform $\mathbf{Y}_i(t)$ into a normal variable before linking it to η_i. To illustrate the idea, we note that, for a continuous variable Y with a distribution function F, it is possible to find a transformation to "normalize" it. Specifically, let Φ be the standard normal distribution function and take $H(\cdot) = \Phi^{-1}(F(\cdot))$. Then, $H(Y)$ has a standard normal distribution. This motivates us to find a monotonic transformation $H_m(\cdot)$ for $Y_{im}(t)$ such that the observed outcomes $Y_{im}(t), m = 1, \dots, p$ depend on the unobserved latent variables $\eta_i(t)$ through a normal transformation linear model,

$$H_m(Y_{im}(t_{ij})) = \mathbf{X}_{ij}^\top \boldsymbol{\beta}_m + \lambda_m \eta_i(t_{ij}) + \epsilon_{imj},\ m = 1, \dots, p, \tag{1}$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_p)^\top$ is a matrix of regression coefficients, $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_p)^\top$ is a vector of factor loadings and $\boldsymbol{\epsilon}_{ij} = (\epsilon_{i1j}, \dots, \epsilon_{ipj})^\top$ is distributed as $N(0, \boldsymbol{\Sigma}_e)$ with $\boldsymbol{\Sigma}_e = \operatorname{diag}(\sigma_1^2, \dots, \sigma_p^2)$. For identifiability, we let $\lambda_1 = 1$ and $H_m(E_{nm}) = c > 0$, with $E_{nm} = \frac{1}{N} \sum_{i=1}^n \sum_{j=1}^{n_i} Y_{im}(t_{ij})$, where $N = \sum_{i=1}^n n_i$. It is noteworthy that our transformation function is to apply to the entire random process $Y_{im}(t)$, rather than *only* to its realizations on individual and discrete time points, an idea stemmed from [21] which considered a common Box-Cox transformation for longitudinal data; also see [17, 33].

In addition, the dependence among the multiple functional responses is reflected through the shared latent process $\eta_i(\cdot)$, e.g. a common cognitive factor across multiple response curves. We posit the following model for $\eta_i(t)$:

$$\eta_i(t_{ij}) = \mathbf{Z}_{ij}^\top \boldsymbol{\alpha} + \delta_i(t_{ij}),\ j \in \{1, \ldots, n_i\}, \tag{2}$$

where $\mathbf{Z}_{ij}$ is the of interest covariate vector with p_2-dimension, such as treatment assignment or dosage of regimen, $\boldsymbol{\alpha}$ is used to evaluate the impact of the covariates on the latent process, $\delta_i(t)$ is a Gaussian process with unknown mean $\mu(t)$ and unknown covariance structure $C(s,t) = Cov(\delta_i(s), \delta_i(t))$, and is independent of the error process $\boldsymbol{\epsilon}_i(t)$ with $\boldsymbol{\epsilon}_i(t_{ij}) = \boldsymbol{\epsilon}_{ij}$, Our models can accommodate multivariate functional data consisting of mixtures of count, ordinal and continuous variables by linking discrete outcomes to continuous latent variables as in [12, 37].

However, for irregular time points, the dimension of parameters may diverge as $n \to \infty$, making it difficult to apply the DLVM model described in [11] for discrete times. [43, 44]and [25] tackled the problem by specifying both the mean and covariance structures of the latent process, which were restricted in practice.

To deal with these challenges, we propose to draw inference based on (1) and (2). Applying the Karhunen-Loeve expansion ([3]) yields

$$\delta_i(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_{ik}\phi_k(t), \tag{3}$$

where $\mu(t) = E\{\delta_i(t)\}$ is the overall mean function; $\phi_k(t)$ is the k-th orthonomal eigenfunction of the covariance function $C(s,t) = Cov(\delta_i(s), \delta_i(t))$, satisfying $\int \phi_k(t)\phi_j(t)dt = 1$ if $j = k$, and 0 otherwise; the ξ_{ik} are the functional principal component scores for the stochastic process $\delta_i(t)$ with $E(\xi_{ik}) = 0$, $\mathrm{var}(\xi_{ik}) = \rho_k$ and $\mathrm{cov}(\xi_{ij}, \xi_{ik}) = 0$ if $j \neq k$. Here, ρ_k is the eigenvalue corresponding to the eigenfunction $\phi_k(\cdot)$. For identifiability, we require $\phi_k(0) > 0$, because $\xi_{ik}\phi_k(t) = \{-\xi_{ik}\}\{-\phi_k(t)\}$. Since $\sup_{t\in[0,1]} E[\sum_{k=1}^{\infty} \xi_{ik}\phi_k(t) - \sum_{k=1}^{K_n} \xi_{ik}\phi_k(t)]^2 \to 0$ as $K_n \to \infty$ ([59]), then we suppose

$$\delta_i(t) = \mu(t) + \sum_{k=1}^{K_n} \xi_{ik}\phi_k(t),\ \text{with } K_n \to \infty. \tag{4}$$

When K_n does not depend on n, model (4) has been extensively considered in the literature of functional principle component analysis (FPCA) ([20, 26, 40, 41, 59]). For more flexibility, we allow $K_n \to \infty$ as $n \to \infty$ ([19, 27, 32]).

2.2. *Estimation*

Let $\boldsymbol{\xi}_i = (\xi_{i1}, \ldots, \xi_{iK_n})^\top$, $\mathbf{H}(\mathbf{Y}_i(t)) = (H_1(Y_{i1}(t)), \ldots, H_p(Y_{ip}(t)))^\top$ and $\mathbf{H}(\mathbf{Y}_i) = (\mathbf{H}(\mathbf{Y}_i(t_{i1})), \ldots, \mathbf{H}(\mathbf{Y}_i(t_{i,n_i})))$. Substituting (2) and (4) into (1) gives that $\mathbf{H}(\mathbf{Y}_i(t_{ij})) = \boldsymbol{\beta}\mathbf{X}_{ij} + \boldsymbol{\lambda}\mathbf{Z}_{ij}^\top \alpha + \boldsymbol{\lambda}\mu(t_{ij}) + \boldsymbol{\lambda}\boldsymbol{\xi}_i^\top \boldsymbol{\Phi}(t_{ij}) + \boldsymbol{\epsilon}_{ij}$ subject to $\int_t \boldsymbol{\Phi}(t)\boldsymbol{\Phi}(t)^\top dt = \mathbf{I}$, where $\boldsymbol{\Phi}(t) = (\phi_1(t), \ldots, \phi_{K_n}(t))^\top$, $\mathbf{I}$ is an identity matrix throughout the paper and may have different dimensions in different places. Then given $(\mathbf{X}_i, \mathbf{Z}_i, \mathbf{t}_i)$,

$$\mathrm{Vec}\,(\mathbf{H}\,(\mathbf{Y}_i)) \sim N\left(\mathrm{Vec}\left(\boldsymbol{\beta}\mathbf{X}_i^\top + \boldsymbol{\lambda}\alpha^\top \mathbf{Z}_i^\top + \boldsymbol{\lambda}\mu(\mathbf{t}_i)^\top\right), \Gamma_i\right), \tag{5}$$

where $\mathrm{Vec}\,(A) = (\mathbf{a}_1^\top, \ldots, \mathbf{a}_p^\top)^\top$ coerces $A = (\mathbf{a}_1, \ldots, \mathbf{a}_p)$ into a vector, $\mathbf{X}_i = (\mathbf{X}_{i1}, \ldots, \mathbf{X}_{i,n_i})^\top$, $\mathbf{Z}_i = (\mathbf{Z}_{i1}, \ldots, \mathbf{Z}_{i,n_i})^\top$, $\mu(\mathbf{t}_i) = (\mu(t_{i1}), \ldots, \mu(t_{i,n_i}))^\top$,

$$\Gamma_i = \left\{(\boldsymbol{\Phi}(t_{i1}), \ldots, \boldsymbol{\Phi}(t_{i,n_i}))^\top \boldsymbol{\Lambda} (\boldsymbol{\Phi}(t_{i1}), \ldots, \boldsymbol{\Phi}(t_{i,n_i}))\right\} \otimes (\boldsymbol{\lambda}\boldsymbol{\lambda}^\top) + \mathbf{I} \otimes \boldsymbol{\Sigma}_e,$$

$\boldsymbol{\Lambda} = \mathrm{diag}(\rho_1, \ldots, \rho_{K_n})$. Suppose $\boldsymbol{\varsigma} = (\sigma_1^2, \ldots, \sigma_p^2)^\top$ and $\boldsymbol{\tau} = (\rho_1, \ldots, \rho_K)^\top$, then all of the finite parameters can be denoted as $\boldsymbol{\theta} = (\mathrm{Vec}(\boldsymbol{\beta})^\top, \boldsymbol{\lambda}^\top, \boldsymbol{\alpha}^\top, \boldsymbol{\varsigma}^\top)^\top \in A$, where A are bounded closed sets in R^{d_0} and $d_0 = (p_1 + 2)p + p_2 - 1$. Hence $\Theta = (\boldsymbol{\theta}^\top, \boldsymbol{\tau}^\top, \mu, \boldsymbol{\Phi}^\top)^\top$ are all of parameters and functions to be estimated. By (5), the observed likelihood is

$$\begin{aligned}\mathbb{L}_n(\Theta; \mathbf{H}) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{pn_i/2}|\Gamma_i|^{1/2}} \exp\Big[-\frac{1}{2}\left\{\mathrm{Vec}\,(\mathbf{H}\,(\mathbf{Y}_i)) - \mathrm{Vec}\left(\boldsymbol{\beta}\mathbf{X}_i^\top + \boldsymbol{\lambda}\alpha^\top \mathbf{Z}_i^\top + \boldsymbol{\lambda}\mu(\mathbf{t}_i)^\top\right)\right\}^\top \\ \times \Gamma_i^{-1} \left\{\mathrm{Vec}\,(\mathbf{H}(\mathbf{Y}_i)) - \mathrm{Vec}\left(\boldsymbol{\beta}\mathbf{X}_i^\top + \boldsymbol{\lambda}\alpha^\top \mathbf{Z}_i^\top + \boldsymbol{\lambda}\mu(\mathbf{t}_i)^\top\right)\right\}\Big],\end{aligned} \tag{6}$$

where $\mathbf{H} = (H_1, H_2, \ldots, H_p)$.

To estimate $\mu(t), \phi_1(t), \ldots, \phi_{K_n}(t)$, we propose to use B-spline smoothing ([6, 24, 48]). Let

$$\mathcal{G} = \{g(\cdot) : |g^{(l)}(v_1) - g^{(l)}(v_2)| \le c_0 |v_1 - v_2|^s, \;\; 0 \le v_1, v_2 \le 1\}, \tag{7}$$

where l and s are nonnegative integers, $r = l + s \ge 2$, and $c_0 > 0$ is a constant. Assuming that $\mu \in \mathcal{G}$ and $\phi_k \in \mathcal{G}$ for $k \in \{1, \ldots, K_n\}$, we approximate $\mu(t)$ and $\phi_k(t)$ by $\mu_n(t) = \boldsymbol{\vartheta}_\mu^\top B_n(t)$ and $\phi_{nk}(t) = \boldsymbol{\vartheta}_k^\top B_n(t)$ for $k \in \{1, \ldots, K_n\}$, respectively, where $B_n(\cdot) = \{b_1(\cdot), \ldots, b_{q_n}(\cdot)\}^\top$ is a set of B-spline basis functions of order $l+1$ with knots $0 = t_0 < t_1 < \ldots < t_{M_n} = 1$, satisfying $\max(t_j - t_{j-1} : j = 1, \ldots, M_n) = O(n^{-v})$ for a constant $v \in (0, 0.5)$ and $q_n = M_n + l$, $\boldsymbol{\vartheta}_\mu$ and $\boldsymbol{\vartheta}_k$ are the B-spline coefficients of q_n-dimension corresponding to $\mu(t)$ and $\phi_k(t)$, respectively. Denote by $\mathcal{G}_n = \left\{\boldsymbol{\zeta}^\top B_n(t) : \boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_{q_n})^\top \in R^{q_n}, \max_{1 \le i \le q_n} |\zeta_i| \le c_0, t \in [0,1]\right\}$. By Corollary 6.21 in [48], for any $g \in \mathcal{G}$, there exists a $g_n \in \mathcal{G}_n$ such that $\|g_n - g\|_\infty = O(n^{-rv})$. We hence estimate $\Theta_n = (\boldsymbol{\theta}^\top, \boldsymbol{\tau}^\top, \boldsymbol{\vartheta}_\mu^\top, \boldsymbol{\vartheta}_k^\top, k = 1, \ldots, K_n)^\top$ and $\mathbf{H}$ by

$$(\hat{\Theta}_n, \hat{\mathbf{H}}_n) = \operatorname{argmax}_{\Theta_n, \mathbf{H}} \log L_n(\Theta_n; \mathbf{H}), \tag{8}$$

where $L_n(\Theta_n; \mathbf{H})$ is $\mathbb{L}_n(\Theta; \mathbf{H})$ with $\mu(t)$ and $\phi_k(t)$ replaced by $\mu_n(t) = \boldsymbol{\vartheta}_\mu^\top B_n(t)$ and $\phi_{nk}(t) = \boldsymbol{\vartheta}_k^\top B_n(t)$, respectively, for $k = 1, \ldots, K_n$.

As the likelihood function $L_n(\Theta_n; \mathbf{H})$ involves the infinite dimensional functions ($\mathbf{H}$) and infinite parameters (Θ_n), a direct maximization is not feasible. We resort to an iterative two-stage approach. First, we use a series of estimating equations to estimate $\mathbf{H}$ given Θ_n. Then we estimate Θ_n by maximizing $L_n(\Theta_n; \mathbf{H})$ with $\mathbf{H}$ replaced by its estimate. We repeat the procedure until convergence.

2.2.1. Estimating Θ_n given $\mathbf{H}$

Given $\mathbf{H}$, maximizing $\log L_n(\Theta_n; \mathbf{H})$ with respect to Θ_n is computationally expensive, as we have to compute Γ_i^{-1}, $i = 1, \ldots, n$, at each iterative step. On the other hand, if $\boldsymbol{\xi}_i$ were observed, the joint log-likelihood would be

$$\begin{aligned} \mathcal{L}_n(\Theta_n; \mathbf{H}) = \sum_{i=1}^n \Bigg(& -\frac{1}{2} \log |\boldsymbol{\Lambda}| - \frac{1}{2} \boldsymbol{\xi}_i^\top \boldsymbol{\Lambda}^{-1} \boldsymbol{\xi}_i - \frac{n_i}{2} \sum_{m=1}^p \log(\sigma_m^2) \\ & - \frac{1}{2} \sum_{j=1}^{n_i} \sum_{m=1}^p \frac{\left\{H_m(Y_{imj}) - \boldsymbol{\beta}_m^\top \mathbf{X}_{ij} - \lambda_m \mathbf{Z}_{ij}^\top \boldsymbol{\alpha} - \lambda_m \boldsymbol{\vartheta}_\mu^\top B_n(t_{ij}) - \lambda_m \boldsymbol{\xi}_i^\top \boldsymbol{\vartheta} B_n(t_{ij})\right\}^2}{\sigma_m^2} \Bigg), \end{aligned}$$

where $Y_{imj} = Y_{im}(t_{ij})$ and $\boldsymbol{\vartheta} = (\boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_{K_n})^\top$. This is a much easier objective function to work with, which motivates us to treat the $\boldsymbol{\xi}_i$ as missing data and invoke the EM algorithm ([10]). Specifically, with $\mathcal{O}_i = \{\mathbf{H}(\mathbf{Y}_i), \mathbf{X}_i, \mathbf{Z}_i, \mathbf{t}_i\}$, differentiating $E\{\mathcal{L}_n(\Theta_n; \mathbf{H}) | \mathcal{O}_i, i = 1, \ldots, n\}$, with respect to Θ_n and setting the derivatives to zero lead to the following estimation equations for $m = 1, \ldots, p$ and $k = 1, \ldots, K_n$:

$$\sigma_m^2 = \frac{1}{N} \sum_{i=1}^n \sum_{j=1}^{n_i} E\left(\left[\tilde{Y}_{imj} - \boldsymbol{\beta}_m^\top \mathbf{X}_{ij} - \lambda_m W_{ij}(\Theta_n)\right]^2 | \mathcal{O}_i\right), \; \rho_k = \frac{1}{n} \sum_{i=1}^n E\left[\xi_{ik}^2 | \mathcal{O}_i\right],$$

$$\boldsymbol{\beta}_m = \left(\sum_{i=1}^n \sum_{j=1}^{n_i} \mathbf{X}_{ij} \mathbf{X}_{ij}^\top\right)^{-1} \sum_{i=1}^n \sum_{j=1}^{n_i} \mathbf{X}_{ij} E\left(\left\{\tilde{Y}_{imj} - \lambda_m W_{ij}(\Theta_n)\right\} | \mathcal{O}_i\right), \; \lambda_m = \frac{\sum_{i=1}^n \sum_{j=1}^{n_i} \left(\tilde{Y}_{imj} - \boldsymbol{\beta}_m^\top \mathbf{X}_{ij}\right) E\left\{W_{ij}(\Theta_n) | \mathcal{O}_i\right\}}{\sum_{i=1}^n \sum_{j=1}^{n_i} E\left\{W_{ij}(\Theta_n)^2 | \mathcal{O}_i\right\}},$$

$$\boldsymbol{\alpha} = \left(\sum_{i=1}^n \sum_{j=1}^{n_i} \sum_{m=1}^p \frac{\lambda_m^2}{\sigma_m^2} \mathbf{Z}_{ij} \mathbf{Z}_{ij}^\top\right)^{-1} \times \sum_{i=1}^n \sum_{j=1}^{n_i} \sum_{m=1}^p \frac{\lambda_m \mathbf{Z}_{ij} E\left[\left\{\tilde{Y}_{imj} - \boldsymbol{\beta}_m^\top \mathbf{X}_{ij} - \lambda_m \boldsymbol{\vartheta}_\mu^\top B_n(t_{ij}) - \lambda_m \boldsymbol{\xi}_i^\top \boldsymbol{\vartheta} B_n(t_{ij})\right\} | \mathcal{O}_i\right]}{\sigma_m^2},$$

$$\boldsymbol{\vartheta}_\mu = \left\{\sum_{i=1}^n \sum_{j=1}^{n_i} \sum_{m=1}^p \frac{\lambda_m^2}{\sigma_m^2} B_n(t_{ij}) B_n(t_{ij})^\top\right\}^{-1} \times \sum_{i=1}^n \sum_{j=1}^{n_i} \sum_{m=1}^p \frac{\lambda_m}{\sigma_m^2} B_n(t_{ij}) E\left[\left\{\tilde{Y}_{imj} - \boldsymbol{\beta}_m^\top \mathbf{X}_{ij} - \lambda_m \mathbf{Z}_{ij}^\top \boldsymbol{\alpha} - \lambda_m \boldsymbol{\xi}_i^\top \boldsymbol{\vartheta} B_n(t_{ij})\right\} | \mathcal{O}_i\right],$$

$$\boldsymbol{\vartheta}_k = \left[\sum_{i=1}^{n}\sum_{j=1}^{n_i}\sum_{m=1}^{p} \frac{\lambda_m^2 B_n(t_{ij}) B_n(t_{ij})^\top E\left\{\xi_{ik}^2 | O_i\right\}}{\sigma_m^2}\right]^{-1} \times$$
$$\sum_{i=1}^{n}\sum_{j=1}^{n_i}\sum_{m=1}^{p} \frac{\lambda_m B_n(t_{ij}) E\left[\left\{\tilde{Y}_{imj} - \boldsymbol{\beta}_m^\top \mathbf{X}_{ij} - \lambda_m \mathbf{Z}_{ij}^\top \boldsymbol{\alpha} - \lambda_m \boldsymbol{\vartheta}_\mu^\top B_n(t_{ij}) - \lambda_m \sum_{r\neq k} \xi_{ir} \boldsymbol{\vartheta}_r^\top B_n(t_{ij})\right\} \xi_{ik} | O_i\right]}{\sigma_m^2},$$

where $\tilde{Y}_{imj} = H_m(Y_{imj})$ and $W_{ij}(\Theta_n) = \mathbf{Z}_{ij}^\top \boldsymbol{\alpha} + \boldsymbol{\vartheta}_\mu^\top B_n(t_{ij}) + \boldsymbol{\xi}_i^\top \boldsymbol{\vartheta} B_n(t_{ij})$. The right side of those equations involve $E(\boldsymbol{\xi}_i | O_i)$ and $E(\boldsymbol{\xi}_i^{\otimes 2} | O_i)$ which are given in the following paragraph. At each iteration, each Θ_n on the right side of the equations is replaced by its updated values, and the procedure is with closed-form expressions.

To estimate Θ_n using the above equations, we need to compute the conditional mean and conditional variance matrices of $\boldsymbol{\xi}_i$ given O_i, that is, $E(\boldsymbol{\xi}_i^{\otimes r} | O_i)$ with $r = 1, 2$, where $a^{\otimes 2} = aa^\top$ and $a^{\otimes 1} = a$. Denote $\boldsymbol{\epsilon}_i = (\boldsymbol{\epsilon}_{i1}^\top, \boldsymbol{\epsilon}_{i2}^\top, \ldots, \boldsymbol{\epsilon}_{i,n_i}^\top)^\top$, $\boldsymbol{\Psi}_1(\mathbf{t}_i)^\top = (\boldsymbol{\vartheta} B_n(\mathbf{t}_i)) \otimes \boldsymbol{\lambda}^\top$, $B_n(\mathbf{t}_i) = (B_n(t_{i1}), \ldots, B_n(t_{i,n_i}))$, and $\boldsymbol{\Psi}(\mathbf{t}_i) = \begin{pmatrix} \boldsymbol{\Psi}_1(\mathbf{t}_i) & \mathbf{I} \\ \mathbf{I} & 0 \end{pmatrix}$. Then $\begin{pmatrix} \mathrm{Vec}\left(\mathbf{H}(\mathbf{Y}_i) - \boldsymbol{\beta} \mathbf{X}_i^\top - \lambda \alpha^\top \mathbf{Z}_i^\top - \lambda \boldsymbol{\vartheta}_\mu^\top B_i\right) \\ \boldsymbol{\xi}_i \end{pmatrix} = \boldsymbol{\Psi}(\mathbf{t}_i) \begin{pmatrix} \boldsymbol{\xi}_i \\ \boldsymbol{\epsilon}_i \end{pmatrix} \sim N(0, \boldsymbol{\Psi}(\mathbf{t}_i) \mathbf{D} \boldsymbol{\Psi}(\mathbf{t}_i)^\top)$, where $\mathbf{D} = \mathrm{diag}\,(\boldsymbol{\Lambda}, \mathbf{D}_2)$ and $\mathbf{D}_2 = \mathbf{I} \otimes \boldsymbol{\Sigma}_e$. Hence, given $O_i = \{\mathbf{H}(\mathbf{Y}_i), \mathbf{X}_i, \mathbf{Z}_i, \mathbf{t}_i\}$, $\boldsymbol{\xi}_i$ is a multivariate normal vector with a mean vector

$$E(\boldsymbol{\xi}_i | O_i) = \boldsymbol{\Lambda} \boldsymbol{\Psi}_1(\mathbf{t}_i)^\top \left\{\boldsymbol{\Psi}_1(\mathbf{t}_i) \boldsymbol{\Lambda} \boldsymbol{\Psi}_1(\mathbf{t}_i)^\top + \mathbf{D}_2\right\}^{-1} \mathrm{Vec}\left(\mathbf{H}(\mathbf{Y}_i) - \boldsymbol{\beta} \mathbf{X}_i^\top - \lambda \alpha^\top \mathbf{Z}_i^\top - \lambda \boldsymbol{\vartheta}_\mu^\top B_i\right),$$

and a covariance matrix $Var(\boldsymbol{\xi}_i | O_i) = \boldsymbol{\Lambda} - \boldsymbol{\Lambda} \boldsymbol{\Psi}_1(\mathbf{t}_i)^\top \left\{\boldsymbol{\Psi}_1(\mathbf{t}_i) \boldsymbol{\Lambda} \boldsymbol{\Psi}_1(\mathbf{t}_i)^\top + \mathbf{D}_2\right\}^{-1} \boldsymbol{\Psi}_1(\mathbf{t}_i) \boldsymbol{\Lambda}$, and then $E(\boldsymbol{\xi}_i^{\otimes 2} | O_i) = Var(\boldsymbol{\xi}_i | O_i) + E^2(\boldsymbol{\xi}_i | O_i)$.

2.2.2. Estimating $\mathbf{H}$ *given* Θ_n

Denote $V_{ij}(\Theta_n) = \mathbf{X}_{ij}^\top \beta_m + \lambda_m \mathbf{Z}_{ij}^\top \alpha + \lambda_m \boldsymbol{\vartheta}_\mu^\top B_n(t_{ij})$ and $\sigma_{ij}^2(\Theta_n) = \lambda_m^2 B_n(t_{ij})^\top \boldsymbol{\vartheta}^\top \Lambda \boldsymbol{\vartheta} B_n(t_{ij}) + \sigma_m^2$. For any given y, we have $Pr\{Y_{imj} \le y | \mathbf{X}_{ij}, \mathbf{Z}_{ij}, t_{ij}\} = Pr\{H_m(Y_{imj}) \le H_m(y) | \mathbf{X}_{ij}, \mathbf{Z}_{ij}, t_{ij}\} = \Phi\left(\frac{H_m(y) - V_{ij}(\Theta_n)}{\sigma_{ij}(\Theta_n)}\right)$, where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal random variable. We estimate $H_m(y)$ using the following estimating equations,

$$\sum_{i=1}^{n}\sum_{j=1}^{n_i}\left[I\left(Y_{imj} \le y\right) - \Phi\left(\frac{H_m(y) - V_{ij}(\Theta_n)}{\sigma_{ij}(\Theta_n)}\right)\right] = 0, \tag{9}$$

for any given y in the support of $Y_{imj}, m = 1, \ldots, p$. Specifically, let $v_{m1}, \ldots, v_{mS_m}$ denote the distinct points of Y_{imj}, $i = 1, 2, \ldots, n, j = 1, 2, \ldots, n_i$. Given $y = v_{ms}$, we estimate $\theta = H_m(y)$ by solving the equations defined by (9),

$$\sum_{i=1}^{n}\sum_{j=1}^{n_i}\left[I\left(Y_{imj} \le y\right) - \Phi\left(\frac{\theta - V_{ij}(\Theta_n)}{\sigma_{ij}(\Theta_n)}\right)\right] = 0, \tag{10}$$

with respect to θ. Equation (9) entails $\hat{H}_m(y)$ to be a nondecreasing step function with jumps only at Y_{imj}. Varying y among $\{v_{m1}, \ldots, v_{mS_m}\}$ and repeating the estimation procedure for each y, we obtain the curve estimator of $H_m(\cdot)$ for each $m = 1, \ldots, p$.

Remark 1. Computation for $H_m(\cdot)$ is feasible. With the closed-form estimator for Θ_n at each step, the proposed method can be straightforwardly implementable. Unlike a traditional nonparametric approach ([23]), our approach does not involve nonparametric smoothing and thus does not need to choose smoothing parameters.

Remark 2. Denote the resulting estimators by $\hat{\boldsymbol{\theta}}$, $\hat{\Theta}_n$ and $\hat{\mathbf{H}}_n$ for $\boldsymbol{\theta}$, Θ_n and $\mathbf{H}$, respectively. Then we estimate the mean function $\mu(t)$ and covariance function $C(s,t)$ by $\hat{\mu}(t) = \hat{\boldsymbol{\vartheta}}_\mu^\top B_n(t)$ and $\hat{C}(s,t) = \hat{\boldsymbol{\Phi}}(t)^\top \hat{\boldsymbol{\Lambda}} \hat{\boldsymbol{\Phi}}(s) = B_n(t)^\top \hat{\boldsymbol{\vartheta}}^\top \hat{\boldsymbol{\Lambda}} \hat{\boldsymbol{\vartheta}} B_n(s)$, respectively. We estimate $C(s,t) = \boldsymbol{\Phi}(t)^\top \boldsymbol{\Lambda} \boldsymbol{\Phi}(s)$ through the estimation of the univariate function $\phi_k(t)$. This differs from the existing FPCA methods, which estimate the univariate function $\phi_k(\cdot)$ via the bivariate function $C(s,t)$, and may incur efficiency loss.

In practice, we select the rank K_n of FPCAs and the number of knots M_n by minimizing the BIC criterion defined in (11) in the data example section. Our simulation studies show that the proposed BIC method performs well.

3. Large sample properties

We establish the asymptotic properties for $\hat{\boldsymbol{\theta}}$, $\hat{\mu}(t)$, $\hat{C}(s,t)$ and $\hat{\mathbf{H}}_n$. Let $\|v\|$ denote the Euclidean norm for a vector v and the L_2 norm $\|f\|_2^2 = \int_0^1 f^2(t)dt$ for any function $f(\cdot)$. Let $\boldsymbol{\Theta} = (\boldsymbol{\theta}^\top, \mu, C)^\top$ and its estimator $\hat{\boldsymbol{\Theta}} = (\hat{\boldsymbol{\theta}}^\top, \hat{\mu}, \hat{C})^\top$. Define the distance between $\boldsymbol{\Theta}_1 = (\boldsymbol{\theta}_1^\top, \mu_1, C_1)^\top$ and $\boldsymbol{\Theta}_2 = (\boldsymbol{\theta}_2^\top, \mu_2, C_2)^\top$ as $d(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2) = \left(\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|^2 + \|\mu_1 - \mu_2\|_2^2 + \|C_1 - C_2\|_2^2\right)^{1/2}$. We use the subscript "0" for true value. That is, $\boldsymbol{\theta}_0$ is the true value of $\boldsymbol{\theta}$. We list the following regularity conditions.
(A1) $\mathbf{X}_i$ and $\mathbf{Z}_i$ are bounded with a compact support.
(A2) $\boldsymbol{\theta}_0$ is an interior point of the bounded set A and $\mu_0 \in \mathcal{G}, \phi_{k0} \in \mathcal{G}, k \in \{1, \dots, K_n\}$.
(A3) There exists $[\underline{y}_m, \bar{y}_m]$ such that $\frac{1}{N}\sum_{i=1}^n \sum_{j=1}^{n_i} I(Y_{imj} \notin [\underline{y}_m, \bar{y}_m]) = o_p(n^{-1/2}), m \in \{1, 2, \dots, p\}$.
(A4) The $Y_{ij}(t)$ is continuous response. The transformation function $H_m(y)$ is strictly increasing, and its first derivative is continuous over $y \in [\underline{y}_m, \bar{y}_m]$.
(A5) $\Delta/\delta \leq c_0$ uniformly in n, where $\delta = \min\limits_{1\leq j\leq M_n} |t_j - t_{j-1}|$, $\Delta = \max\limits_{1\leq j\leq M_n} |t_j - t_{j-1}| = O(n^{-\upsilon})$, and $t_j, j \in \{1, \dots, M_n\}$ are knots of B-spline.
(A6) $\int_0^1 \mathrm{Var}(\delta_i(t))dt < \infty$.
(A7) $K_n = O(n^e)$ with $0 \leq e < \min(1-\upsilon, 2r\upsilon)$.
(A8) $\max_i n_i < \infty$.

Conditions (A1) and (A2) are commonly assumed in the semiparametric literature, for example [5] for generalized partially linear single-index model, [63] for local linear simultaneous confidence corridor of sparse functional data, [30] for variance function partially linear single-index model, and [6] for varying coefficient transformation model. Condition (A1) is a technical issue which can be relaxed to their high-order moments being bounded. Condition (A3) is used to avoid the tail problem, which is also required by [31] for transformation model. Condition (A4) is commonly used for the transformation function ([23, 64]). Condition (A5) is used for spline analysis ([34], and Condition (A6) is used to avoid unbounded covariances ([18]). In practice, K_n is small and hence Condition (A7) is easily satisfied. The following theorems stipulate the consistency and asymptotic normality of the proposed estimators, and the proofs are deferred to the Supplementary materials.

Theorem 1. *Under Conditions (A1)-(A7), we have*

$$P\left(\sup_{y\in[\underline{y}_m, \bar{y}_m]} |\hat{H}_{mn}(y) - H_{m0}(y)| \to 0\right) \to 1,\ d(\hat{\boldsymbol{\Theta}}, \boldsymbol{\Theta}_0) = O_p\left(n^{-\min(\frac{1-\upsilon-e}{2}, r\upsilon - \frac{e}{2})}\right),\ m \in \{1, 2, \dots, p, \}$$

where r is a smooth parameter defined in (7) and $0 < \upsilon < 0.5$ *is given for determining the density of the knots in the spline basis* $B_n(\cdot)$.

Remark 3. The convergence rate of $\hat{\boldsymbol{\Theta}}$ is determined by the number of eigenfunctions K_n, smoothness of unknown function and number of spline knots. When the number of spline knots q_n increases, the approximation error (bias) decreases, but the stochastic error (variance) increases. The larger K_n is and consequently the larger the number of estimated functions is, the slower convergence rate is. Particularly, when K_n is constant and $e = 0$, then $d(\hat{\boldsymbol{\Theta}}, \boldsymbol{\Theta}_0) = O_p\left(n^{-\min(\frac{1-\upsilon}{2}, r\upsilon)}\right)$ and the choice of $\upsilon = \frac{1}{2r+1}$ yields the optimal rate of convergence $n^{\frac{r}{1+2r}}$ for the non-parametric function ([51]).

Theorem 2. *Assume that Conditions (A1)-(A7) hold with* $r \geq 2$ *and* $\frac{1}{4r} < v < \frac{1}{2}$. *Then, we have* $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(0, I^{-1}(\boldsymbol{\theta}_0))$, *where* $I(\boldsymbol{\theta}_0)$ *is defined at the end of the proof of Theorem 2 in the Appendix.*

4. Simulation

We assess the finite-sample performance of the proposed method via Monte Carlo simulations. We investigate the robustness of the method and whether the robustness is gained at the expense of efficiency. We also investigate the sensitivity of our method to the normality assumption on the transformed response. For robustness, we consider the latent class mixed models (lcmm) method ([43, 45]), where both the transformation function and covariance structure are specified, and the method without transformation (WOT), which is performed by our proposed method by setting

$H_m(y) = y$. To assess the efficiency, we modify the proposed method by allowing the transformation to be correctly specified and the covariance structure to be estimated by the proposed method. We term this approach the correct transformation (CT) approach.

We assess the performance of the estimator in bias, standard deviation (sd), and root mean squared error (RMSE), which are defined by

$$bias = \left[\frac{1}{n_{grid}} \sum_{i=1}^{n_{grid}} \{E\hat{g}(x_i) - g(x_i)\}^2\right]^{1/2}, \quad sd = \left[\frac{1}{n_{grid}} \sum_{i=1}^{n_{grid}} E\{\hat{g}(x_i) - E\hat{g}(x_i)\}^2\right]^{1/2},$$

and RMSE $= \left[bias^2 + sd^2\right]^{1/2}$, where x_i $(i = 1, \ldots, n_{grid})$ are the grid points in which the function $g(\cdot)$ is estimated and $E\hat{g}(x_i)$ is approximated by its sample mean based on N simulated data sets. In the following examples, we use the cubic B-spline approximation with the number of knots M_n and the rank K_n selected by maximizing

$$\text{BIC}(M_n, K_n) \quad = \quad \log\left\{L_n(\Theta_n; \mathbf{H})\right\} - \frac{1}{2}\log(Np)\text{df}(\Theta_n), \tag{11}$$

where df(Θ_n) is the number of parameters in Θ_n. The knots placed at the quantiles of the observation times. We set $n_{grid} = 100$.

We simulate $N = 200$ datasets, each with $(n, p) = (200, 3)$. The observation number of each subject is sampled from a discrete uniform distribution on $\{1, \ldots, 10\}$, and the locations of the measurements are randomly chosen from $U(0, 1)$. The covariates X_{i1} and X_{i2} are generated from an independent standard normal distribution $N(0, 1)$. The covariates Z_i are generated from the Bernoulli distribution with a success probability $1/2$. Finally, we set $\mu(t) = t^2 - t, \lambda_1 = \lambda_2 = \lambda_3 = 1, \boldsymbol{\beta}_1 = (\beta_{11}, \beta_{12})^\top = (0.25, 0.25)^\top, \boldsymbol{\beta}_2 = (\beta_{21}, \beta_{22})^\top = (0.25, 0.25)^\top, \boldsymbol{\beta}_3 = (\beta_{31}, \beta_{32})^\top = (-0.5, -0.5)^\top, \alpha = 1$, and $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 1/2$. We consider the following six examples.

Example 1. In models (1) and (2), we set $H_1(t) = H_2(t) = H_3(t) = 5\log(t)$ and $\eta(t) = \mathbf{Z}^T\boldsymbol{\alpha} + \mu(t) + \sum_{k=1}^{2} \xi_k\phi_k(t)$, with $\phi_1(t) = \sqrt{2}sin(\pi t), \phi_2(t) = \sqrt{2}sin(2\pi t), \xi_1 \sim N(0, \rho_1), \xi_2 \sim N(0, \rho_2), \rho_1 = 1$ and $\rho_2 = 1/4$.

Example 2. The example is the same as Example 1, except that $H_1(t) = H_2(t)= H_3(t) = 10(\sqrt{t} - 1)$. Neither the transformation function nor the covariance structure satisfies the requirements of [43] in Examples 1 and 2.

Example 3. The example is the same as Example 1, except that $H_1(t) = H_2(t) = H_3(t) = 5t$. Here, the transformation function satisfies the requirement of [43], but the covariance structure does not follow that of [43].

Example 4. The data are generated in the same way as in Example 1, except that $H_1(t) = H_2(t) = H_3(t) = (\frac{1}{5}t)^3$ and that the covariance structure follows [43]. That is, $\eta_i(t) = \mathbf{Z}^T\boldsymbol{\alpha} + \mu(t) + \mathbf{f}(t)^T u_i + \omega_i(t)$, where $\mathbf{f}(t) = (1, t, t^2)$, u_i is a multivariate normal random variable with mean zero and identity covariance matrix and $\omega_i(t)$ is a Brownian motion with mean zero and covariance $\text{cov}(\omega_i(t), \omega_i(s)) = \min(t, s)$. Therefore, the covariance structure does not satisfy the requirement of our method, but follows the requirement of [43]. This setup is to examine the robustness of our method to the covariance structure.

Example 5. The setting is the same as in Example 1, except that $K_n = 4$, eigen-functions $\phi_1(t) = \sqrt{2}sin(\pi t), \phi_2(t) = \sqrt{2}sin(2\pi t), \phi_3(t) = \sqrt{2}sin(3\pi t), \phi_4(t) = \sqrt{2}sin(4\pi t)$ and $\rho_1 = 1, \rho_2 = 0.75, \rho_3 = 0.5, \rho_4 = 0.25$. The setting is used to examine the performance of the proposed method for relatively large K_n.

Example 6. Data are generated in the same way as in Example 2, except that the random error ϵ_{ikj} is generated from a mixed distribution with each component being the centralized and scaled gamma distribution $\sigma \times (Gamma(\tau, 1) - \tau)/\sqrt{\tau}$, $\sigma = \sqrt{2}/2$. We take τ =1, 3, 10 and 50. As our method assumes a Gaussian distribution on the transformed response, this example examines the sensitivity of our method to this normality assumption.

Simulation studies with the same setting except of bounded covariates $X_{i1}, X_{i2} \sim U[-2, 2]$ are also implemented. We apply the proposed method, CT, WOT and lcmm to analyze each simulated dataset. Table 1 and Table 1 in the Supplementary Material present the bias, empirical sd, and RMSE of the estimates for the parameters, mean function $\mu(\cdot)$ and covariance function $C(\cdot, \cdot)$ for Examples 1 and 2. The lcmm method specifies the transformation function by a β-cumulative distribution function and the covariance function by the sum of a quadratic polynomial and Brownian motions. Because the covariance and transformation functions in Examples 1 and 2 do not satisfy the requirement of [43], the lcmm method is numerically unstable, as it fails in 88 and 62 out of the 200 runs in Examples 1 and 2, respectively. WOT fails in 53 out of the 200 runs in each of Examples 1 and 2. The summaries are based on the convergent cases. Using a useful rule of thumb that unbiasedness means a standardized bias (bias as a percentage

Table 1: The bias, empirical sd, and RMSE of the estimates for the parameters, mean function $\mu(\cdot)$ and covariance function $C(\cdot,\cdot)$ for Example 1. The lcmm method [43] specifies the transformation function by a β-cumulative distribution function and the covariance function by the sum of a quadratic polynomial and Brownian motions. Because the covariance and transformation functions in Example 1 do not satisfy the requirement of [43], the lcmm method is numerically unstable, as it fails in 88 out of the 200 runs in Example 1. WOT fails in 53 out of the 200 runs. The summaries are based on the convergent cases.

	lcmm			proposed			CT			WOT		
	bias	sd	RMSE	bias	sd	RMSE	bias	sd	RMSE	bias	sd	RMSE
β_{11}	0.617	0.371	0.721	0.001	0.033	0.033	0.001	0.032	0.032	0.193	0.008	0.193
β_{12}	0.622	0.381	0.730	0.006	0.034	0.035	0.003	0.033	0.033	0.195	0.008	0.195
β_{21}	0.626	0.370	0.728	0.000	0.034	0.034	0.002	0.034	0.034	0.192	0.008	0.193
β_{22}	0.630	0.380	0.736	0.004	0.037	0.038	0.001	0.036	0.036	0.194	0.008	0.195
β_{31}	1.244	0.733	1.444	0.011	0.038	0.040	0.003	0.036	0.036	.389	0.010	0.389
β_{32}	1.253	0.753	1.462	0.005	0.037	0.037	0.000	0.035	0.035	0.386	0.009	0.386
σ_1^2	6.617	6.575	9.328	0.000	0.028	0.028	0.004	0.023	0.023	0.468	0.012	0.468
σ_2^2	6.567	6.346	9.132	0.004	0.032	0.033	0.000	0.024	0.024	0.467	0.012	0.468
σ_3^2	6.673	6.759	9.498	0.005	0.035	0.035	0.006	0.024	0.024	0.465	0.012	0.466
α	2.477	1.461	2.876	0.003	0.057	0.058	0.004	0.055	0.055	0.7770	0.013	0.777
λ_2	*	*	*	0.001	0.025	0.025	0.000	0.023	0.023	0.000	0.006	0.006
λ_3	*	*	*	0.012	0.028	0.030	0.001	0.023	0.023	0.006	0.008	0.010
$\mu(\cdot)$	0.167	0.106	0.198	0.011	0.088	0.089	0.003	0.076	0.076	1.163	0.021	1.163
$C(\cdot,\cdot)$	13.28	15.77	20.62	0.037	0.130	0.135	0.017	0.132	0.133	0.754	0.012	0.754

of the SD) not exceeding 40% ([38]), we may infer that the proposed estimator and the CT method are unbiased. In addition, these two methods have relatively smaller and comparable SDs. In contrast, the lcmm method and the WOT method with identity transformation are biased and the lcmm method is seriously inefficient. This indicates that misspecifications of the transformation function and covariance structures may lead to large biases and variations. As the proposed method is comparable to the CT estimator in mean squared errors, it is robust with little loss of efficiency. Similar patterns are observed in Table 2 in the Supplementary Materials for the results of Examples 3 and 4, which further confirms that misspecification of either the transformation function or covariance structure may lead to large biases and variations. When increasing K_n to 4, Table 3 in the Supplementary Materials indicates that the proposed method still performs well.

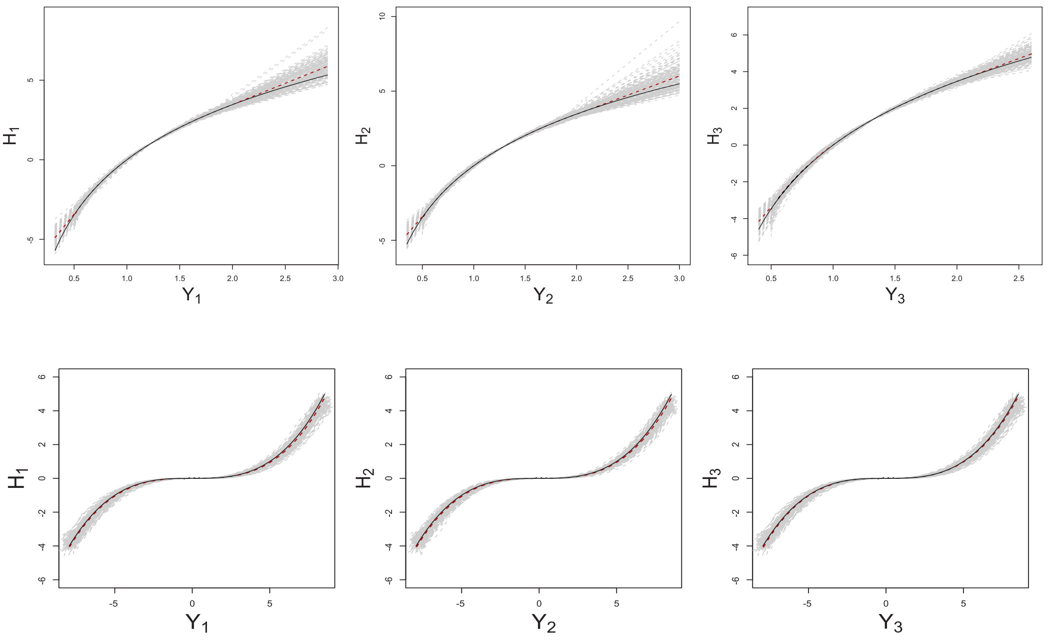

Fig. 1: The average estimates of transformation functions over 200 replications using the proposed method (solid-true curves; red dotted-proposed estimator; dashed-95% empirical point-wise confidence limit of proposed estimator) for Examples 1 and 4 in first and second rows, respectively.

Table 2 presents the results for Example 6. When $\tau \geq 10$, both skewness (SK) and excess kurtosis (EK) are less than one and the proposed estimator is nearly unbiased. When both skewness and excess kurtosis are around 1 to 2, the proposed estimator incurs slight biases. Only in the extreme cases, where the skewness and excess kurtosis are both larger than 2 and the error distribution is severely non-normal, is the proposed estimator moderately biased. Table 2 hints at our method's robustness to the normality assumption on the error.

Fig.1 displays the averaged estimates of the transformation functions based on 200 replicates, along with the 95% pointwise confidence intervals, for Examples 1 and 4. The averaged estimates coincide with the truth. Similar results are obtained from the other examples and omitted here.

In Examples 1 to 5, we generate $\mathbf{X}_i = (X_{i1}, X_{i2})^\top$ from unbounded normal distribution. We also perform simulation studies with the same setting except of bounded covariates $X_{i1}, X_{i2} \sim U[-2, 2]$, similar results with those in Table 1 and Fig.1 can be obtained, suggesting that the proposed method works well for both bounded and unbounded covariates.

We select the number of the interior knots M_n and the rank K_n of FPCA by maximizing the BIC in (11). To check the performance of (11), we vary M_n and K_n over the grids and calculate the corresponding BIC for Example 1. Fig.1 in the Supplementary Materials, which depicts BIC values over (M_n, K_n), shows that the largest BIC is achieved when $M_n = 3$ and $K_n = 2$, which is the truth. Fig.1 in the Supplementary Materials also shows that the optimal K_n and M_n are nearly independent, meaning K_n and M_n can be separately selected. Furthermore, the proposed method does not

Table 2: The bias, empirical sd, and RMSE of the estimates for the parameters, mean function $\mu(\cdot)$ and covariance function $C(\cdot,\cdot)$ for Example 6 to examines the sensitivity of our method to normality assumption. The random error ϵ_{ikj} of the transformed response is generated from a mixed distribution with each component being the centralized and scaled gamma distribution $\sigma \times (Gamma(\tau, 1) - \tau)/\sqrt{\tau}$, $\sigma = \sqrt{2}/2$ with τ =1, 3, 10 and 50. SK and EK denote skewness and excess kurtosis respectively.

τ	1			3			10			50		
SK	2			1.15			0.63			0.28		
EK	6			2			0.6			0.12		
	bias	sd	RMSE	bias	sd	RMSE	bias	sd	RMSE	bias	sd	RMSE
β_{11}	0.007	0.036	0.035	0.005	0.033	0.034	0.001	0.035	0.035	0.001	0.035	0.035
β_{12}	0.004	0.034	0.035	0.004	0.034	0.034	0.002	0.036	0.036	0.005	0.036	0.037
β_{21}	0.005	0.034	0.034	0.001	0.034	0.035	0.000	0.033	0.033	0.003	0.033	0.033
β_{22}	0.004	0.034	0.035	0.002	0.036	0.036	0.004	0.035	0.035	0.007	0.036	0.037
β_{31}	0.003	0.036	0.036	0.003	0.034	0.035	0.000	0.035	0.035	0.009	0.037	0.038
β_{32}	0.002	0.037	0.037	0.001	0.035	0.035	0.003	0.037	0.037	0.000	0.039	0.039
σ_1^2	0.005	0.043	0.044	0.005	0.038	0.038	0.006	0.032	0.033	0.001	0.032	0.032
σ_2^2	0.007	0.039	0.039	0.008	0.037	0.038	0.001	0.029	0.030	0.003	0.031	0.031
σ_3^2	0.006	0.046	0.046	0.008	0.033	0.034	0.006	0.033	0.034	0.000	0.031	0.031
α	0.039	0.064	0.075	0.017	0.065	0.067	0.000	0.062	0.062	0.006	0.066	0.066
λ_2	0.005	0.026	0.027	0.003	0.027	0.027	0.001	0.027	0.027	0.003	0.030	0.030
λ_3	0.009	0.030	0.032	0.006	0.028	0.028	0.001	0.026	0.026	0.003	0.030	0.030
$\mu(\cdot)$	0.079	0.096	0.125	0.041	0.089	0.098	0.017	0.088	0.089	0.008	0.088	0.088
$C(\cdot,\cdot)$	0.032	0.130	0.134	0.030	0.127	0.131	0.036	0.129	0.134	0.038	0.127	0.133

seem sensitive to the choice of M_n.

5. Analysis of Cognitive Decline among the Elderly

As cognitive decline develops subtly among the elderly, the progress, though slow, may severely impede the quality of life of affected individuals ([36, 60]). The education level, as a well-established risk factor for some neurological disorders, such as Alzheimer's disease, plays an uncertain role in cognitive decline among the elderly ([58]), which has sparked much research interest; see [1, 39, 61, 62], among many others. However, as most studies did not fully utilize the multifaceted measurements for the cognitive capacity, the results were often conflicting.

By accounting for the multivariate functional measurements of cognitive curves, we examined the effect of education on cognitive decline using the French prospective cohort study (PAQUID), a prospective cohort study (Dordogne & Gironde) that aimed to explore functional and cerebral aging ([29]). Included in the study were subjects over 65 years of age at initial visit, and were tested at years 1, 3, 5, 8, 10 and 13 after the initial visit. At each visit, a battery of psychometric tests, including the Mini-Mental State Examination (MMSE), Benton Visual Retention Test (BVRT) and Isaacs Set Test (IST), were administered, followed by a screening procedure for dementia. MMSE (Y_1) is a global cognitive test evaluating various dimensions of cognition (memory, calculation, orientation in space and time, language and word recognition), ranging from 0 to 30. BVRT (Y_2) evaluates the immediate visual memory, and ranges from 0 to 15. IST (Y_3) evaluates the semantic verbal fluency, processing speed and memory, and ranges from 0 to 40. For each test, lower values indicate more severe impairment. The covariates of interest include gender (X) and educational level (Z, 1=with primary schooling and 0=otherwise). The analyzable samples consist of 490 subjects after removing the missing data. To illustrate the distributions of MMSE, IST and BVRT given X, Z and t, we discretized the continuous t (age) into four intervals (65-75, 75-80, 80-85, 85-100) to ensure roughly equal sizes within each interval, and obtained 16 groups with various combinations of age, gender ($X = 0, 1$) and educational level ($Z = 0, 1$). We plotted the histograms of response data (MMSE, BVRT, IST) for each group in Fig.2-4 in the Supplementary Materials, none of which seemed to be normally distributed.

We applied the proposed method to investigate how the cognitive capacity declines over time, and whether and how education affects the declining process. We used models (1) and (2), where the multivariate response curves included three functional outcomes, namely, Y_1, Y_2 and Y_3, measured at various times, t_{ij} (e.g. the jth observation time for subject i), and the covariates included X and Z. For improved stability, we rescaled t_{ij} into $[0, 1]$ at the data pre-processing stage.

We adopted the cubic B-spline approximation. The number of the interior knots $M_n = 4$ and the rank $K_n = 1$ of FPCA were selected by maximizing $\text{BIC}(M_n, K_n)$. K_n is chosen to be 1 because the first eigen-component explains more than 95% of the variance of the multivariate response processes. The SD was calculated based 200 bootstrap samples, in which each subject was treated as a resampling unit to preserve the dependence structure of the data. The choice of 200 was determined by monitoring the stability of the SD.

The estimated mean function, eigenfunction and transformation functions based on the proposed method are given in Fig.2, along with the corresponding 95% point-wise confidence limits (the dashed lines) over 200 bootstrap replications. The estimates of β_k, σ_k^2 for $k \in \{1, 2, 3\}, \alpha$ and ρ_1 along with the estimated standard errors are presented in Table 3. Fig.2 shows a decreasing mean function, implying that cognitive capacity declined with age in general. The eigenfunction clearly decreased until 93 with a large standard error after 93, suggesting that the cognition capacity declines at least until 93 but with an unclear pattern after 93, likely because of very few patients surviving past 93.

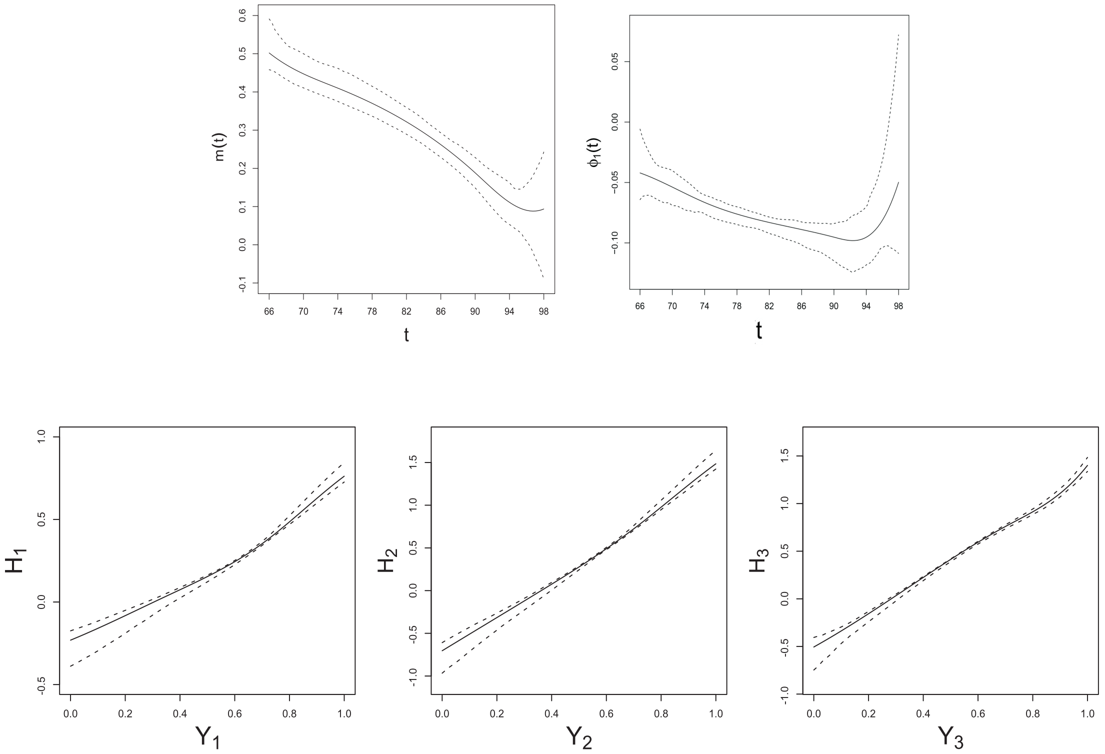

Fig. 2: Estimates (solid) and 95% confident limit (dashed) of mean function $m(t)$, first eigen-function $\phi_1(t)$, and transformation functions H_1, H_2 and H_3 for three responses MMSE(Y_1), BVRT(Y_2) and IST(Y_3) respectively in the Cognitive Decline data. The 95% point-wise confidence limits (the dashed lines) are obtained based on 200 bootstrap replications.

The coefficient estimates for β_k for $k \in \{1, 2, 3\}$ in Table 3 show that men and women did not have a significant

Table 3: The estimator (Est.), sd, and $p-$valus of parameters for cognitive decline data. β_1,β_2 and β_3 are the coefficients of gender (X), and σ_1, σ_2 and σ_3 are standard deviations of random errors for three responses MMSE(Y_1), BVRT(Y_2) and IST(Y_3) respectively. α is the coefficient of educational level (Z). ρ_1 is the first eigenvalue for the latent process $\delta_i(t)$.

		β_1	β_2	β_3	α	σ_1^2	σ_2^2	σ_3^2	ρ_1
Proposed	Est.	0.002	0.052	-0.054	0.182	0.046	0.148	0.100	2.600
	sd	0.014	0.028	0.024	0.016	0.006	0.018	0.009	0.195
	p value	0.886	0.063	0.024	0.000	0.000	0.000	0.000	0.000
lcmm	Est.	-0.009	0.014	-0.005	0.111	*	*	*	*
	sd	0.005	0.005	0.006	0.008	*	*	*	*
	p value	0.071	0.005	0.404	0.000	*	*	*	*

difference in MMSE (Y_1) and BVRT (Y_2), while women tended to perform better than men in the verbal fluency test (IST, Y_3). In contrast, the lcmm method does not detect the effect of gender on the verbal fluency test. The result of the proposed method is in agreement with the previous results that women tended to perform better in verbal skills ([46, 57]). The estimate of α in Table 3 suggests that subjects with primary schooling had significantly better cognitive capacity than those without, which agrees with the previous results that highly educated subjects tended to perform better in cognitive tests ([28, 53]). Fig.2 reveals that the estimated transformation functions for Y_1 and Y_2 were more non-linear than that for Y_3. This is not surprising as the distribution of Y_1 or Y_2 is more non-normal like compared to that of Y_3.

Finally, to evaluate the usefulness of our proposed transformation model, we compared the out-of-sample prediction error of the methods with data-driven transformation and without transformation (WOT), and the lcmm method. Specially, we randomly divided the data into the training and testing subsets, with ratios of 1:2, 1:1 and 2:1, respectively. We used the three methods to fit the training data sets. For each subject in the testing data sets, we predicted the Y_1, Y_2 and Y_3 by the fitted model obtained from the training data sets, and computed the mean square prediction error (MSPE) of Y_1, Y_2,Y_3, respectively, and their sum ($MSPE_{sum}$) in each of the testing data sets. The root-prediction error is displayed in Table 4, which suggests that the proposed method has much smaller prediction errors than the lcmm and the method without transformation.

Table 4: The out-of-sample mean square prediction error (MSPE) and their summation ($MSPE_{sum}$) of the methods with data-driven transformation (Proposed) and without transformation (WOT), and the lcmm for the Cognitive Decline data. We consider three randomly dividing with the ratios of training and testing subsets being 1:2, 1:1 and 2:1, respectively.

Training set rate	Proposed				lcmm				WOT			
	Y_1	Y_2	Y_3	$MSPE_{sum}$	Y_1	Y_2	Y_3	$MSPE_{sum}$	Y_1	Y_2	Y_3	$MSPE_{sum}$
1/3	0.084	0.009	0.006	0.085	0.636	0.038	0.000	0.637	0.220	0.245	0.275	0.429
1/2	0.095	0.014	0.006	0.096	0.449	0.054	0.000	0.452	0.219	0.242	0.271	0.424
2/3	0.111	0.036	0.007	0.117	0.194	0.104	0.000	0.220	0.215	0.244	0.271	0.424

6. Concluding remarks

We have proposed a new semiparametric transformation latent process regression model, in which we allow the distribution of the responses, the dependent structure of multiple response curves, and the pattern of intraindividual variability to be unspecified. Our model renders more flexibility and can be applicable to non-normal data as we have demonstrated. To overcome the difficulty of inferring $(\mathbf{H}, \Theta_n)$ simultaneously, we propose a convenient iterative algorithm that combines a simple one-dimension Newton-Raphson for $\mathbf{H}$ and a closed form expression of Θ_n at each step. Therefore, our inferential procedure is feasible and implementable. Finally, we have established the large sample properties of the resulting estimators, based on which confidence intervals can be constructed.

It will be of interest to extend the methods to accommodate discrete response variables. We envision that the related theory and implementation may be nontrivial. In practice, the dimension of covariates may be large. To handle the high dimensionality of covariates, we may need to resort to penalization approaches. Suitable penalty functions and regularity conditions need to be rigorously constructed, which merit further studies.

Acknowledgements

The research was supported by National Natural Science Foundation of China (No.11931014 and 11829101) and Fundamental Research Funds for the Central Universities (No. JBK1806002) of China.

Appendix

Denote $Pf = \int f(x)dP(x),\ P_nf = \frac{1}{n}\sum_{i=1}^n f(x_i)$,

$$\begin{aligned}\mathcal{G}_n &= \left\{\boldsymbol{\zeta}^\top B_n(t) : \boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_{q_n})^\top \in R^{q_n}, \max_{1\le i\le q_n} |\zeta_i| \le c_0, t \in [0,1]\right\},\\ \boldsymbol{\Omega}_n &= \{(\overrightarrow{\boldsymbol{\beta}}^\top, \boldsymbol{\lambda}^\top, \boldsymbol{\alpha}^\top, \boldsymbol{\varsigma}^\top)^\top \in A, \boldsymbol{\tau} \in B \in R^K, \mu \in \mathcal{G}_n, \boldsymbol{\Phi} \in \prod_{k=1}^K \mathcal{G}_n\} = A \times B \times \mathcal{G}_n \times \prod_{k=1}^K \mathcal{G}_n,\end{aligned}$$

$$\ell_i(\Theta_n; \mathbf{H}) = -\frac{1}{2}\log(|\Gamma_i|) - \frac{1}{2}\left\{\overrightarrow{\mathbf{H}(\mathbf{Y}_i)} - \overrightarrow{\boldsymbol{\beta}\mathbf{X}_i^\top + \lambda\alpha^\top\mathbf{Z}_i^\top + \lambda\mu(\mathbf{t}_i)^\top}\right\}^\top \Gamma_i^{-1}\left\{\overrightarrow{\mathbf{H}(\mathbf{Y}_i)} - \overrightarrow{\boldsymbol{\beta}\mathbf{X}_i^\top + \lambda\alpha^\top\mathbf{Z}_i^\top + \lambda\mu(\mathbf{t}_i)^\top}\right\},$$

$$Q_n(\Theta_n; \mathbf{H}) = \sum_{i=1}^n \ell_i(\Theta_n; \mathbf{H}),\ V_{ij,m}(\Theta_n) = \mathbf{X}_{ij}^\top\boldsymbol{\beta}_m + \lambda_m\mathbf{Z}_{ij}^\top\alpha + \lambda_m\mu(t_{ij}),\ W_{ij,m}(\Theta_n) = \lambda_m^2\boldsymbol{\Phi}^\top(t_{ij})\boldsymbol{\Lambda}\boldsymbol{\Phi}(t_{ij}) + \sigma_m^2,$$

$$\psi_{imj}(w; y, \Theta_n) = I\left(Y_{imj} \le y\right) - \Phi\left(\frac{w - V_{ij,m}(\Theta_n)}{\sqrt{W_{ij,m}(\Theta_n)}}\right),\ \Psi_{mn}(w; y, \Theta_n) = \frac{1}{N}\sum_{i=1}^n\sum_{j=1}^{n_i}\psi_{imj}(w; y, \Theta_n),$$

$$\Psi_m(w; y, \Theta_n) = E\left\{\Phi\left(\frac{H_{m0}(y) - V_{ij,m}(\Theta_0)}{\sqrt{W_{ij,m}(\Theta_0)}}\right) - \Phi\left(\frac{w - V_{ij,m}(\Theta_n)}{\sqrt{W_{ij,m}(\Theta_n)}}\right)\right\}.$$

$\hat{H}_{mn}(y; \Theta_n)$ is the estimator of $H_m(y)$ given Θ_n and is the solution of (10) with respect to $H_m(y)$. Define $\hat{\mathbf{H}}_n(y; \Theta_n) = (\hat{H}_{1n}(y; \Theta_n), \hat{H}_{2n}(y; \Theta_n), \ldots, \hat{H}_{pn}(y; \Theta_n))^\top$.

We first give several lemmas. We define the covering number of class $\mathcal{L}_n = \{\ell(\Theta_n; \hat{\mathbf{H}}_n(\cdot; \Theta_n)) : \Theta_n \in \boldsymbol{\Omega}_n\}$. In particular, for any $\epsilon > 0$, define the covering number $N(\epsilon, \mathcal{L}_n, L_1(P_n))$ as the smallest value of κ for which there exist $\{\Theta_{n,j} \in \boldsymbol{\Omega}_n, j = 1, \ldots, \kappa\}$ such that

$$\min_{j\in\{1,\ldots,\kappa\}} \frac{1}{n}\sum_{i=1}^n |\ell_i(\Theta_n; \hat{\mathbf{H}}_n(\cdot; \Theta_n)) - \ell_i(\Theta_{n,j}; \hat{\mathbf{H}}_n(\cdot; \Theta_{n,j}))| < \epsilon,$$

for all $\Theta_n \in \boldsymbol{\Omega}_n$. If no such κ exists, define $N(\epsilon, \mathcal{L}_n, L_1(P_n)) = \infty$. We remark that this theory relies on the complicated modern empirical process theory ([55]). To establish the asymptotic normality, we also employ the Riesz representation theorem. First, we give some lemmas and their proofs.

Lemma 1. *The covering number of the class $\boldsymbol{\Omega}_n$ satisfies*

$$N(\epsilon, \boldsymbol{\Omega}_n, L_2) \le c_0 M_n^{Kq_n}\epsilon^{-\{d_0+(K+1)q_n\}},$$

where $M_n = O(n^v)$.

Proof: The result follows by applying Lemma 2.5 and Corollary 2.6 in [15]. □

Lemma 2. *Under conditions (A1)-(A5), $\hat{H}_{mn}(y; \Theta_n), m = 1, \ldots, p$ defined by (10) satisfies*

$$\sup_{\Theta_n\in\boldsymbol{\Omega}_n, y\in[\underline{y}_m, \bar{y}_m]} |\hat{H}_{mn}(y; \Theta_n) - H_m(y; \Theta_n)| \to 0, \text{ for given } y \in [\underline{y}_m, \bar{y}_m] \text{ and } \Theta_n \in \boldsymbol{\Omega}_n$$

where $H_m(y; \Theta_n)$ satisfies

$$\Psi_m(H_m(y; \Theta_n); y, \Theta_n) = 0. \tag{12}$$

Proof: It follows from the law of large numbers and the monotonicity of $H_{m0}(y)$ that for given $\varrho \geq 0$, $y \in [\underline{y}_m, \bar{y}_m]$, $\Theta_n \in \mathbf{\Omega}_n$,

$$\frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\left[I\left(Y_{imj} \leq y\right) - \Phi\Big(\frac{H_{m0}(y) - V_{ij,m}(\Theta_n)}{\sqrt{W_{ij,m}(\Theta_n)}} - \varrho\Big)\right] \to E\Big\{\Phi\Big(\frac{H_{m0}(y) - V_{ij,m}(\Theta_0)}{\sqrt{W_{ij,m}(\Theta_0)}}\Big) - \Phi\Big(\frac{H_{m0}(y) - V_{ij,m}(\Theta_n)}{\sqrt{W_{ij,m}(\Theta_n)}} - \varrho\Big)\Big\}, \quad (13)$$

almost surely as $n \to \infty$, where $N = \sum_{i=1}^{n} n_i$. Furthermore, by Lemma 1 and Theorem 19.4 of [54], $\mathbf{\Omega}_n$ is a P-Glivenko-Cantelli class. This, coupled with the fact that $\Phi\Big(\frac{H_{m0}(y) - V_{ij,m}(\Theta_n)}{\sqrt{W_{ij,m}(\Theta_n)}} - \varrho\Big)$ is a continuous and bounded function of $\Theta_n \in \mathbf{\Omega}_n$, and the indicator function class $\{I\left(Y_{imj} \leq y\right)\}$ is of VC class, implies that (13) also holds uniformly over $y \in [\underline{y}_m, \bar{y}_m]$ and $\Theta_n \in \mathbf{\Omega}_n$ by [15].

On the other hand, it follows from (13) that for large ϱ,

$$\frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\left[I\left(Y_{imj} \leq y\right) - \Phi\Big(\frac{H_{m0}(y) - V_{ij,m}(\Theta_n)}{\sqrt{W_{ij,m}(\Theta_n)}} - \varrho\Big)\right] > 0, \quad (14)$$

$$\frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\left[I\left(Y_{imj} \leq y\right) - \Phi\Big(\frac{H_{m0}(y) - V_{ij,m}(\Theta_n)}{\sqrt{W_{ij,m}(\Theta_n)}} + \varrho\Big)\right] < 0. \quad (15)$$

This, together with the monotonicity and continuity of Φ, implies that there exists a unique $\hat{H}_{mn}(y;\Theta_n)$ such that

$$\frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\left[I\left(Y_{imj} \leq y\right) - \Phi\Big(\frac{\hat{H}_{mn}(y;\Theta_n) - V_{ij,m}(\Theta_n)}{\sqrt{W_{ij,m}(\Theta_n)}}\Big)\right] = 0, \quad (16)$$

for given y and $\Theta_n \in \mathbf{\Omega}_n$. Similarly, there is a unique $H_m(y;\Theta_n)$ satisfying (12) for given y and Θ_n. In addition, note that

$$\begin{aligned}\Psi_{mn}(\hat{H}_{mn}(y;\Theta_n); y, \Theta_n) &= \{\Psi_{mn}(\hat{H}_{mn}(y;\Theta_n); y, \Theta_n) - \Psi_m(\hat{H}_{mn}(y;\Theta_n); y, \Theta_n)\} \\ &+\{\Psi_m(H_m(y;\Theta_n); y, \Theta_n) - \Psi_{mn}(H_m(y;\Theta_n); y, \Theta_n)\} + \{\Psi_m(\hat{H}_{mn}(y;\Theta_n); y, \Theta_n) - \Psi_m(H_m(y;\Theta_n); y, \Theta_n)\} \\ &+\Psi_{mn}(H_m(y;\Theta_n); y, \Theta_n). \quad (17)\end{aligned}$$

By Lemma 1 and the uniform strong law of large numbers, we have

$$\Psi_{mn}(H_m(y;\Theta_n); y, \Theta_n) \to \Psi_m(H_m(y;\Theta_n); y, \Theta_n) = 0,$$

almost surely uniformly in $y \in [\underline{y}_m, \bar{y}_m]$ and $\Theta_n \in \mathbf{\Omega}_n$. Then (17) and conditions (A1)-(A4) imply that

$$0 = \|\Psi_{mn}(\hat{H}_{mn}(y;\Theta_n); y, \Theta_n)\| \geq C_m\|\hat{H}_{mn}(y;\Theta_n) - H_m(y;\Theta_n)\| - \xi_n, \quad (18)$$

where $C_m > 0$ which does not depend on y, and

$$\xi_n = \sup_{\Theta_n \in \mathbf{\Omega}_n, y \in [\underline{y}_m, \bar{y}_m]} \|\Psi_{mn}(H_m(y;\Theta_n); y, \Theta_n)\| \to 0.$$

Hence, (18) implies that $\hat{H}_{mn}(y;\Theta_n)$ converges to $H_m(y;\Theta_n)$ uniformly in $y \in [\underline{y}_m, \bar{y}_m]$ and $\Theta_n \in \mathbf{\Omega}_n$. □

Lemma 3. *Under Conditions (A1)-(A5), the covering number of the class $\mathcal{L}_n$ satisfies*

$$N(\epsilon, \mathcal{L}_n, L_1(P_n)) \leq c_0 M_n^{Kq_n} \epsilon^{-[d_0+(K+1)q_n]}.$$

Proof: For any $\Theta^{(1)} = (\overrightarrow{\boldsymbol{\beta}}_1^\top, \boldsymbol{\lambda}_1^\top, \boldsymbol{\alpha}_1^\top, \boldsymbol{\varsigma}_1^\top, \boldsymbol{\tau}_1^\top, \mu_1, \mathbf{\Phi}_1^\top)^\top \in \mathbf{\Omega}_n$, $\Theta^{(2)} = (\overrightarrow{\boldsymbol{\beta}}_2^\top, \boldsymbol{\lambda}_2^\top, \boldsymbol{\alpha}_2^\top, \boldsymbol{\varsigma}_2^\top, \boldsymbol{\tau}_2^\top, \mu_2, \mathbf{\Phi}_2^\top)^\top \in \mathbf{\Omega}_n$ and $\mathbf{\Phi}_j = (\phi_{j,1}, \ldots, \phi_{j,K})$, $j \in \{1, 2\}$, we have

$$\begin{aligned}P_n\ell(\Theta^{(1)}; \hat{\mathbf{H}}_n(\ldots;\Theta^{(1)})) - P_n\ell(\Theta^{(2)}; \hat{\mathbf{H}}_n(\,\cdot\,;\Theta^{(2)})) &= P_n\{\ell(\Theta^{(1)}; \mathbf{H}(\,\cdot\,;\Theta^{(1)})) - \ell(\Theta^{(2)}; \mathbf{H}(\,\cdot\,;\Theta^{(2)}))\} \\ +P_n\{\ell(\Theta^{(1)}; \hat{\mathbf{H}}_n(\,\cdot\,;\Theta^{(1)})) - \ell(\Theta^{(1)}; \mathbf{H}(\,\cdot\,;\Theta^{(1)}))\} &- P_n\{\ell(\Theta^{(2)}; \hat{\mathbf{H}}_n(\,\cdot\,;\Theta^{(2)})) - \ell(\Theta^{(2)}; \mathbf{H}(\,\cdot\,;\Theta^{(2)}))\}. \quad (19)\end{aligned}$$

By the uniform convergence of $\hat{H}_{mn}(y;\Theta_n)$ to $H_m(y;\Theta_n), m \in \{1,2,\ldots,p\}$ described in Lemma 2, we have

$$\begin{aligned} &P_n\{\ell(\Theta^{(1)};\hat{\mathbf{H}}_n(\,\cdot\,;\Theta^{(1)})) - \ell(\Theta^{(1)};\mathbf{H}(\,\cdot\,;\Theta^{(1)}))\} = o_p(1),\\ &P_n\{\ell(\Theta^{(2)};\hat{\mathbf{H}}_n(\,\cdot\,;\Theta^{(2)})) - \ell(\Theta^{(2)};\mathbf{H}(\,\cdot\,;\Theta^{(2)}))\} = o_p(1). \end{aligned} \tag{20}$$

Using the Taylor expansion, we obtain

$$\begin{aligned} |\ell(\Theta^{(1)};\mathbf{H}(\,\cdot\,;\Theta^{(1)})) - \ell(\Theta^{(2)};\mathbf{H}(\,\cdot\,;\Theta^{(2)}))| \le c_0\Big(&\|\vec{\boldsymbol{\beta}}_1 - \vec{\boldsymbol{\beta}}_2\| + \|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\| + \|\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2\| + \|\boldsymbol{\varsigma}_1 - \boldsymbol{\varsigma}_2\|\\ &+\|\boldsymbol{\tau}_1 - \boldsymbol{\tau}_2\| + \sum_{k=0}^{K}\|\phi_{1,k} - \phi_{2,k}\|_\infty\Big). \end{aligned} \tag{21}$$

where $\phi_{1,0} = \mu_1, \phi_{2,0} = \mu_2$. Denote $\boldsymbol{\zeta}_{j,k} = (\zeta_{jk1},\ldots,\zeta_{jkq_n})$ to be the spline coefficients of $\phi_{j,k}$ where $j \in \{1,2\}$ and $k \in \{0,1,\ldots,K\}$, respectively. We have

$$\|\phi_{1,k} - \phi_{2,k}\|_\infty \le \max_{1\le i\le q_n}|\zeta_{1ki} - \zeta_{2ki}| := \|\boldsymbol{\zeta}_{1,k} - \boldsymbol{\zeta}_{2,k}\|_\infty. \tag{22}$$

Combining (21) and (22), we obtain

$$\begin{aligned} &|\ell(\Theta^{(1)};\mathbf{H}(\,\cdot\,;\Theta^{(1)})) - \ell(\Theta^{(2)};\mathbf{H}(\,\cdot\,;\Theta^{(2)}))|\\ &\le c_0\left(\|\vec{\boldsymbol{\beta}}_1 - \vec{\boldsymbol{\beta}}_2\| + \|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\| + \|\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2\| + \|\boldsymbol{\varsigma}_1 - \boldsymbol{\varsigma}_2\| + \|\boldsymbol{\tau}_1 - \boldsymbol{\tau}_2\| + \sum_{k=0}^{K}\|\boldsymbol{\alpha}_{1,k} - \boldsymbol{\alpha}_{2,k}\|_\infty\right). \end{aligned} \tag{23}$$

Combining (19), (20), (23) and Lemma 1 and mimicking the calculation on page 94 of [55], we have $N(\epsilon,\mathcal{L}_n,L_1(P_n)) \le c_0 M_n^{Kq_n}\epsilon^{-[d_0+(K+1)q_n]}$. □

Lemma 4. *Under Conditions (A1)-(A5), we have*

$$\sup_{\Theta_n\in\boldsymbol{\Omega}_n}|P_n\ell(\Theta_n;\hat{\mathbf{H}}_n(\,\cdot\,;\Theta_n)) - P\ell(\Theta_n;\mathbf{H}(\,\cdot\,;\Theta_n))| \to 0 \textit{ almost surely,}$$

Proof: By lemma 2, we have $\sup_{\Theta_n\in\boldsymbol{\Omega}_n}|P_n\{\ell(\Theta_n;\hat{\mathbf{H}}_n(\,\cdot\,;\Theta_n)) - \ell(\Theta_n;\mathbf{H}(\,\cdot\,;\Theta_n))\}| \to 0$. Then we need to prove $\sup_{\Theta_n\in\boldsymbol{\Omega}_n}|P_n\ell(\Theta_n;\mathbf{H}(\,\cdot\,;\Theta_n)) - P\ell(\Theta_n;\mathbf{H}(\,\cdot\,;\Theta_n))| \to 0$. Note that $|\ell(\Theta_n;\mathbf{H}(\,\cdot\,;\Theta_n))|$ is bounded under Conditions (A1)-(A4). So, without loss of generality, we assume $\sup_{\Theta_n\in\boldsymbol{\Omega}_n}|\ell(\Theta_n;\mathbf{H}(\,\cdot\,;\Theta_n))| \le 1$. Then $P\ell^2(\Theta_n;\mathbf{H}(\,\cdot\,;\Theta_n)) \le P(\sup_{\Theta_n\in\boldsymbol{\Omega}_n}|\ell(\Theta_n;\mathbf{H}(\,\cdot\,;\Theta_n))|)^2 \le 1$. Let $\alpha_n = n^{-1/2+\phi_1}(\log n)^{1/2}$ with $(\upsilon + e)/2 < \phi_1 < 1/2$. Obviously $\{\alpha_n\}$ is a non-increasing sequence of positive numbers. Also for a given $\epsilon > 0$, let $\epsilon_n = \epsilon\alpha_n$. Then for sufficiently large n and any $\Theta_n \in \boldsymbol{\Omega}_n$, we have

$$\frac{\mathrm{var}(P_n\ell(\Theta_n;\mathbf{H}(\,\cdot\,;\Theta_n)))}{(4\epsilon_n)^2} \le \frac{(1/n)P\ell^2(\Theta_n;\mathbf{H}(\,\cdot\,;\Theta_n))}{16\epsilon^2\alpha_n^2} \le \frac{1}{16n\epsilon^2\alpha_n^2} \le \frac{1}{16\epsilon^2 n^{2\phi_1}\log n} \le \frac{1}{2}.$$

Let $O = \{O_1,\ldots,O_n\}$ represent the observed data. Let P_n^o denote the signed measure that places mass $\pm\frac{1}{n}$ at each of $\{O_1,\ldots,O_n\}$, with the random ± signs independent of the O_i. Then, by [42] and $\mathrm{var}(P_n\ell(\Theta_n;\mathbf{H}(\,\cdot\,;\Theta_n)))/(4\epsilon_n)^2 \le 1/2$, the following symmetrization inequality holds

$$P(\sup_{\Theta_n\in\boldsymbol{\Omega}_n}|P_n\ell(\Theta_n;\mathbf{H}(\,\cdot\,;\Theta_n)) - P\ell(\Theta_n;\mathbf{H}(\,\cdot\,;\Theta_n))| > 8\epsilon_n) \le 4P(\sup_{\Theta\in\boldsymbol{\Omega}_n}|P_n^o\ell(\Theta_n;\mathbf{H}(\,\cdot\,;\Theta_n))| > 2\epsilon_n). \tag{24}$$

Given O, select $(\Theta_n^{(1)},\ldots,\Theta_n^{(\kappa)})$, where $\kappa = N(\epsilon_n/2,\mathcal{L}_n,L_1(P_n))$, such that

$$\min_{j\in\{1,\ldots,\kappa\}}P_n|\ell(\Theta_n;\mathbf{H}(\,\cdot\,;\Theta_n)) - \ell(\Theta_n^{(j)};\mathbf{H}(y;\Theta_n^{(j)}))| \le \frac{\epsilon_n}{2},$$

for all $\Theta_n \in \boldsymbol{\Omega}_n$. For each $\Theta_n \in \boldsymbol{\Omega}_n$, denote

$$\Theta_n^* = \arg\min_{\Theta_n^{(j)}}P_n|\ell(\Theta_n;\mathbf{H}(\,\cdot\,;\Theta_n)) - \ell(\Theta_n^{(j)};\mathbf{H}(y;\Theta_n^{(j)}))|.$$

Note that

$$|P_n^o(\ell(\Theta_n;\mathbf{H}(\,\cdot\,;\Theta_n)) - \ell(\Theta_n^*;\mathbf{H}(y;\Theta_n^*)))| = \Big|\frac{1}{n}\sum_{i=1}^{n} \pm(\ell_i(\Theta_n;\mathbf{H}(\,\cdot\,;\Theta_n)) - \ell_i(\Theta_n^*;\mathbf{H}(y;\Theta_n^*)))\Big|$$

$$\leq \frac{1}{n}\sum_{i=1}^{n}\Big|(\ell_i(\Theta_n;\mathbf{H}(\,\cdot\,;\Theta_n)) - \ell_i(\Theta_n^*;\mathbf{H}(y;\Theta_n^*)))\Big| = P_n\Big|\ell(\Theta_n;\mathbf{H}(\,\cdot\,;\Theta_n)) - \ell(\Theta_n^*;\mathbf{H}(y;\Theta_n^*))\Big|. \tag{25}$$

Then, by the definition of Θ_n^* and (25), we have

$$\begin{aligned}
&P(\sup_{\Theta_n\in\mathbf{\Omega}_n} |P_n^o\ell(\Theta_n;\mathbf{H}(\,\cdot\,;\Theta_n))| > 2\epsilon_n|\mathcal{O}) \\
&\leq \quad P(\sup_{\Theta_n\in\mathbf{\Omega}_n} [|P_n^o\ell(\Theta_n^*;\mathbf{H}(y;\Theta_n^*))| + P_n^o|\ell(\Theta_n;\mathbf{H}(\,\cdot\,;\Theta_n)) - \ell(\Theta_n^*;\mathbf{H}(y;\Theta_n^*))|] > 2\epsilon_n|\mathcal{O}) \\
&\leq \quad P(\max_j |P_n^o\ell(\Theta_n^{(j)};\mathbf{H}(y;\Theta_n^{(j)}))| > \frac{3\epsilon_n}{2}|\mathcal{O}) \leq N(\epsilon_n/2,\mathcal{L}_n,L_1(P_n))\max_j P(|P_n^o\ell(\Theta_n^{(j)};\mathbf{H}(y;\Theta_n^{(j)}))| > \frac{3\epsilon_n}{2}|\mathcal{O}).
\end{aligned} \tag{26}$$

According to the definition of $N(\epsilon_n/2,\mathcal{L}_n,L_1(P_n))$, for each $\Theta_n^{(j)}$, there exists $\check{\Theta}_n^{(j)}$ such that $P_n|\ell(\check{\Theta}_n^{(j)};\mathbf{H}(y;\check{\Theta}_n^{(j)})) - \ell(\Theta_n^{(j)};\mathbf{H}(y;\Theta_n^{(j)}))| \leq \frac{\epsilon_n}{2}$. Therefore, we obtain

$$\begin{aligned}
&P(|P_n^o\ell(\Theta_n^{(j)};\mathbf{H}(y;\Theta_n^{(j)}))| > \frac{3\epsilon_n}{2}|\mathcal{O}) \\
&\leq \quad P([|P_n^o\ell(\check{\Theta}_n{}^{(j)};\mathbf{H}(y;\check{\Theta}_n^{(j)}))| + P_n|\ell(\check{\Theta}_n{}^{(j)};\mathbf{H}(y;\check{\Theta}_n^{(j)})) - \ell(\Theta_n^{(j)};\mathbf{H}(y;\Theta_n^{(j)}))|] > \frac{3\epsilon_n}{2}|\mathcal{O}) \\
&\leq \quad P(|P_n^o\ell(\check{\Theta}_n^{(j)};\mathbf{H}(y;\check{\Theta}_n^{(j)}))| > \epsilon_n|\mathcal{O}).
\end{aligned} \tag{27}$$

By Hoeffding's inequality, we have

$$\begin{aligned}
P(|P_n^o\ell(\check{\Theta}_n^{(j)};\mathbf{H}(y;\check{\Theta}_n^{(j)}))| > \epsilon_n|\mathcal{O}) \quad &= \quad P(|\sum_{i=1}^{n} \pm\ell_i(\check{\Theta}_n^{(j)};\mathbf{H}(y;\check{\Theta}_n^{(j)}))| > n\epsilon_n|\mathcal{O}) \\
&\leq \quad 2\exp\Big[-2(n\epsilon_n)^2/\sum_{i=1}^{n}(2\ell_i(\check{\Theta}_n^{(j)};\mathbf{H}(y;\check{\Theta}_n^{(j)})))^2\Big] \leq 2\exp(-n\epsilon_n^2/2).
\end{aligned} \tag{28}$$

The last inequality of (28) holds because $|\ell_i(\check{\Theta}_n^{(j)};\mathbf{H}(y;\check{\Theta}_n^{(j)}))| \leq 1$. Combining (26), (27), (28) and Lemma 3, we obtain

$$P(\sup_{\Theta_n\in\mathbf{\Omega}_n} |P_n^o\ell(\Theta_n;\mathbf{H}(\,\cdot\,;\Theta_n))| > 2\epsilon_n|\mathcal{O}) \quad \leq \quad 2N(\epsilon_n/2,\mathcal{L}_n,L_1(P_n))\exp(-n\epsilon_n^2/2) \leq c_0M_n^{Kq_n}(\epsilon_n/2)^{-[d_0+(K+1)q_n]}\exp(-n\epsilon_n^2/2).$$

Note that the right-hand side does not depend on $\mathcal{O}$. Then by taking expectations over $\mathcal{O}$, we have

$$P(\sup_{\Theta_n\in\mathbf{\Omega}_n} |P_n^o\ell(\Theta_n;\mathbf{H}(\,\cdot\,;\Theta_n))| > 2\epsilon_n) \leq c_0M_n^{Kq_n}(\epsilon_n/2)^{-[d_0+(K+1)q_n]}\exp(-n\epsilon_n^2/2). \tag{29}$$

By (24) and (29), we obtain

$$\begin{aligned}
&P(\sup_{\Theta_n\in\mathbf{\Omega}_n} |P_n\ell(\Theta_n;\mathbf{H}(\,\cdot\,;\Theta_n)) - P\ell(\Theta_n;\mathbf{H}(\,\cdot\,;\Theta_n))| > 8\epsilon_n) \leq 4P(\sup_{\Theta_n\in\mathbf{\Omega}_n} |P_n^o\ell(\Theta_n;\mathbf{H}(\,\cdot\,;\Theta_n))| > 2\epsilon_n) \\
&\leq \quad c_0M_n^{Kq_n}(\epsilon_n/2)^{-[d_0+(K+1)q_n]}\exp(-n\epsilon_n^2/2) \leq c_0\exp(-c_0\epsilon^2n^{2\phi_1}\log n).
\end{aligned}$$

The last inequality follows as $(\upsilon + e)/2 < \phi_1 < 1/2$. Hence

$$\sum_{n=1}^{\infty} P(\sup_{\Theta_n\in\mathbf{\Omega}_n} |P_n\ell(\Theta_n;\mathbf{H}(\,\cdot\,;\Theta_n)) - P\ell(\Theta_n;\mathbf{H}(\,\cdot\,;\Theta_n))| > 8\epsilon_n) < \infty.$$

By the Borel-Cantelli lemma, we have

$$\sup_{\Theta_n\in\mathbf{\Omega}_n} |P_n\ell(\Theta_n;\mathbf{H}(\,\cdot\,;\Theta_n)) - P\ell(\Theta_n;\mathbf{H}(\,\cdot\,;\Theta_n))| \to 0,$$

almost surely, which completes the proof of Lemma 4. □

Now we are ready to prove Theorems 1-2.

Proof of Theorem 1. The proof has two steps. The first step proves the consistency of $\hat{\Theta}_n$ and $\hat{\mathbf{H}}_n(y;\hat{\Theta}_n)$. The second step deals with the convergent rate of $\hat{\Theta}_n$.

Step 1 (consistency). Under Condition (A2) and by Corollary 6.21 of [48], there exist $\mu_{n0} = \mathbf{a}_{n0}^{\top}B_n(t)$ and $\phi_{nk0} = \mathbf{b}_{nk0}^{\top}B_n(t)$ such that

$$\sup_{t\in[0,1]} |\mu_{n0}(t) - \mu_0(t)| = O(q_n^{-r}) \textit{ and } \sup_{t\in[0,1]} |\phi_{nk0}(t) - \phi_{k0}(t)| = O(q_n^{-r}),$$

where $\mu_0(t)$ and $\phi_{k0}(t)$ denote the true functions of $\mu(t)$and $\phi_k(t)$ where $k \in \{1,\dots,K\}$. Denote $C_0(s,t) = \sum_{k=1}^{K} \rho_{k0}\phi_{k0}(t)\phi_{k0}(s)$, $C_{n0}(s,t) = \sum_{k=1}^{K} \rho_{k0}\phi_{kn0}(t)\phi_{kn0}(s)$, and $\hat{C}(s,t) = \sum_{k=1}^{K} \hat{\rho}_k\hat{\phi}_k(t)\hat{\phi}_k(s)$. Then

$$\sup_{(t,s)\in[0,1]^2} |C_0(s,t) - C_{n0}(s,t)| \le \sum_{k=1}^{K} \rho_{k0}\sup|\phi_{k0}(t)\phi_{k0}(s) - \phi_{kn0}(t)\phi_{kn0}(s)|$$
$$\le \sum_{k=1}^{K} \rho_{k0}\{\sup|\phi_{k0}(s)||\phi_{k0}(t) - \phi_{kn0}(t)| + \sup|\phi_{kn0}(t)||\phi_{k0}(s) - \phi_{kn0}(s)|\} = O(K^{1/2}q_n^{-r}).$$

Let $\boldsymbol{\Theta}_{n0} = (\boldsymbol{\theta}_0, \mu_{n0}, C_{n0})^{\top}$ and $\boldsymbol{\Theta}_0 = (\boldsymbol{\theta}_0, \mu_0, C_0)^{\top}$. Then,

$$d(\boldsymbol{\Theta}_{n0}, \boldsymbol{\Theta}_0) = O(n^{-r\upsilon}). \tag{30}$$

Let $M(\Theta_n;\hat{\mathbf{H}}_n(y;\Theta_n)) = -\ell(\Theta_n;\hat{\mathbf{H}}_n(\,\cdot\,;\Theta_n))$ and $K_\epsilon = \{\Theta_n : d(\Theta_n,\Theta_{n0}) \ge \epsilon, \Theta_n \in \boldsymbol{\Omega}_n\}$ for $\epsilon > 0$ and

$$\zeta_{1n} = \sup_{\Theta_n\in\boldsymbol{\Omega}_n} |P_nM(\Theta_n;\hat{\mathbf{H}}_n(\,\cdot\,;\Theta_n)) - PM(\Theta_n;\hat{\mathbf{H}}_n(\,\cdot\,;\Theta_n))|,$$
$$\zeta_{2n} = P_nM(\Theta_{n0};\hat{\mathbf{H}}_n(\,\cdot\,;\Theta_{n0})) - PM(\Theta_{n0};\hat{\mathbf{H}}_n(\,\cdot\,;\Theta_{n0})).$$

Then one can show that

$$\inf_{K_\epsilon} PM(\Theta_n;\hat{\mathbf{H}}_n(\,\cdot\,;\Theta_n)) = \inf_{K_\epsilon}\{PM(\Theta_n;\hat{\mathbf{H}}_n(\,\cdot\,;\Theta_n)) - P_nM(\Theta_n;\hat{\mathbf{H}}_n(\,\cdot\,;\Theta_n)) + P_nM(\Theta_n;\hat{\mathbf{H}}_n(\,\cdot\,;\Theta_n))\}$$
$$\le \zeta_{1n} + \inf_{K_\epsilon} P_nM(\Theta_n;\hat{\mathbf{H}}_n(\,\cdot\,;\Theta_n)). \tag{31}$$

If $\hat{\Theta}_n \in K_\epsilon$, then we have

$$\inf_{K_\epsilon} P_nM(\Theta_n;\hat{\mathbf{H}}_n(\,\cdot\,;\Theta_n)) = P_nM(\hat{\Theta}_n;\hat{\mathbf{H}}_n(\,\cdot\,;\Theta_n)) \le P_nM(\Theta_{n0};\hat{\mathbf{H}}_n(\,\cdot\,;\Theta_{n0})) = \zeta_{2n} + PM(\Theta_{n0};\hat{\mathbf{H}}_n(\,\cdot\,;\Theta_{n0})). \tag{32}$$

Let $\delta_\epsilon = \inf_{K_\epsilon} PM(\Theta_n;\hat{\mathbf{H}}_n(\,\cdot\,;\Theta_n)) - PM(\Theta_{n0};\hat{\mathbf{H}}_n(\,\cdot\,;\Theta_{n0}))$. One can verify $\delta_\epsilon > 0$ under the conditions in Theorem 1 when n is large enough. By (31) and (32), we have $\inf_{K_\epsilon} PM(\Theta_n;\hat{\mathbf{H}}_n(\,\cdot\,;\Theta_n)) \le \zeta_{1n} + \zeta_{2n} + PM(\Theta_{n0};\hat{\mathbf{H}}_n(\,\cdot\,;\Theta_{n0})) = \zeta_n + PM(\Theta_{n0};\hat{\mathbf{H}}_n(\,\cdot\,;\Theta_{n0}))$, with $\zeta_n = \zeta_{1n} + \zeta_{2n}$, and hence $\zeta_n \ge \delta_\epsilon$ by the definition of δ_ϵ. Since $\{\hat{\Theta}_n \in K_\epsilon\} \subseteq \{\zeta_n \ge \delta_\epsilon\}$, then $\bigcup_{i=1}^{\infty}\bigcap_{n=i}^{\infty}\{\hat{\Theta}_n \in K_\epsilon\} \subseteq \bigcup_{i=1}^{\infty}\bigcap_{n=i}^{\infty}\{\zeta_n \ge \delta_\epsilon\}$. By Lemma 2 and the strong law of large numbers, we have both $\zeta_{1n} \to 0$ and $\zeta_{2n} \to 0$ almost surely. Therefore, $\bigcup_{i=1}^{\infty}\bigcap_{n=i}^{\infty}\{\zeta_n \ge \delta_\epsilon\}$ is null set when n is large enough, which proves that $d(\hat{\Theta}_n, \Theta_{n0}) \to 0$ almost surely as $n \to \infty$. Since

$$\hat{C}(s,t) - C_{n0}(s,t) = \sum_{k=1}^{K} \hat{\rho}_k\hat{\phi}_k(t)\hat{\phi}_k(s) - \hat{\rho}_k\phi_{kn0}(t)\phi_{kn0}(s) + \hat{\rho}_k\phi_{kn0}(t)\phi_{kn0}(s) - \rho_{k0}\phi_{kn0}(t)\phi_{kn0}(s)$$
$$= \sum_{k=1}^{K} \hat{\rho}_k\Big[\hat{\phi}_k(t)\{\hat{\phi}_k(s) - \phi_{kn0}(s)\} + \phi_{kn0}(s)\{\hat{\phi}_k(t) - \phi_{kn0}(t)\}\Big] + \{\hat{\rho}_k - \rho_{k0}\}\phi_{kn0}(t)\phi_{kn0}(s). \tag{33}$$

Then $\hat{C}(s,t) \to C_{n0}(s,t)$ almost surely. Combining with (30), we have $d(\hat{\boldsymbol{\Theta}}_n, \boldsymbol{\Theta}_0) \to 0$ almost surely. Further, with Lemma 2, we have $\hat{\mathbf{H}}_{mn}(y) \equiv \hat{\mathbf{H}}_{mn}(y;\hat{\Theta}_n) \to \mathbf{H}_m(y;\Theta_0) \equiv H_{m0}(y)$ uniformly in $y \in [\underline{y}_m, \overline{y}_m]$ for $m = 1,2,\dots,p$.

Step 2 (convergence rate). We establish the convergence rate of $\hat{\Theta}_n$ by using Theorem 3.4.1 of [55]. For any $\eta > 0$, define

$$\mathcal{F}_\eta = \{\ell(\Theta_n; \hat{\mathbf{H}}_n(\cdot\,;\Theta_n)) - \ell(\Theta_{n0}; \hat{\mathbf{H}}_n(\cdot\,;\Theta_{n0})) : \Theta_n \in \mathbf{\Omega}_n, \eta/2 \leq d(\Theta_n, \Theta_{n0}) \leq \eta\}.$$

For Θ in the neighborhood of Θ_0, the compactness of the parameter space implies that $P\{\ell(\Theta_0; H_0) - \ell(\Theta; \hat{\mathbf{H}}_n(\cdot\,;\Theta))\} \asymp d^2(\Theta, \Theta_0)$. Hence

$$P(\ell(\Theta_{n0}; \hat{\mathbf{H}}_n(\cdot\,;\Theta_{n0})) - \ell(\Theta_0; \mathbf{H}_0)) \asymp d^2(\Theta_{n0}, \Theta_0) \leq c_0 n^{-2r\upsilon}. \tag{34}$$

For large n, by (34) we have

$$P(\ell(\Theta_n; \hat{\mathbf{H}}_n(\cdot\,;\Theta_n)) - \ell(\Theta_{n0}; \hat{\mathbf{H}}_n(\cdot\,;\Theta_{n0}))) \quad \leq \quad c_0\eta^2 + c_0 n^{-2r\upsilon} = O_p(\eta^2),$$

for any $\ell(\Theta_n; \hat{\mathbf{H}}_n(\cdot\,;\Theta_n)) - \ell(\Theta_{n0}; \hat{\mathbf{H}}_n(\cdot\,;\Theta_{n0})) \in \mathcal{F}_\eta$. Following [49], we can establish that for $0 < \varepsilon < \eta$, $\log N_{[]}(\varepsilon, \mathcal{F}_\eta, L_2(P)) \leq c_0 K q_n \log(\eta/\varepsilon)$. Under Conditions (A1)-(A5), $\mathcal{F}_\eta$ is uniformly bounded. Therefore, by Lemma 3.4.2 of [55], we obtain

$$E_P \|n^{1/2}(P_n - P)\|_{\mathcal{F}_\eta} \leq c_0 J_{[]}(\eta, \mathcal{F}_\eta, L_2(P)) \left\{1 + \frac{J_{[]}(\eta, \mathcal{F}_\eta, L_2(P))}{\eta^2\sqrt{n}}\right\},$$

where $J_{[]}(\eta, \mathcal{F}_\eta, L_2(P)) = \int_0^\eta \{1 + \log N_{[]}(\varepsilon, \mathcal{F}_\eta, L_2(P))\}^{\frac{1}{2}} d\varepsilon \leq c_0\sqrt{Kq_n}\eta$. Denote $\phi_n(\eta) = \sqrt{Kq_n}\eta + Kq_n/\sqrt{n}$. It follows that $\phi_n(\eta)/\eta$ is decreasing in η, and $r_n^2\phi_n(1/r_n) = r_n\sqrt{Kq_n} + r_n^2 Kq_n/n^{1/2} \leq c_0 n^{1/2}$, where $r_n = (Kq_n)^{-1/2}n^{1/2} = n^{(1-\upsilon-e)/2}$. Noting that $P_n(\ell(\Theta_n; \hat{H}_n(\cdot\,;\Theta_n)) - \ell(\Theta_{n0}; \hat{H}_n(\cdot\,;\Theta_{n0}))) \geq 0$ and $d(\hat{\Theta}_n, \Theta_{n0}) \leq d(\hat{\Theta}_n, \Theta_0) + d(\Theta_0, \Theta_{n0}) \to 0$ in probability, by applying Theorem 3.4.1 of [55], we have $n^{(1-\upsilon-e)/2}d(\hat{\Theta}_n, \Theta_{n0}) = O_P(1)$. According to (33), we have $n^{(1-\upsilon-e)/2}d(\hat{\mathbf{\Theta}}_n, \mathbf{\Theta}_{n0}) = O_P(1)$. This together with $d(\mathbf{\Theta}_0, \mathbf{\Theta}_{n0}) = O(n^{-r\upsilon+e/2})$ yields that $d(\hat{\mathbf{\Theta}}_n, \mathbf{\Theta}_0) = O_P(n^{-(1-\upsilon-e)/2} + n^{-r\upsilon+e/2}) = O(n^{-\min(\frac{1-\upsilon-e}{2}, r\upsilon-e/2)})$. □

Proof of Theorem 2. Let $\mathbf{\Omega} - \mathbf{\Theta}_0$ be $\mathbf{\Omega}$ excluding $\mathbf{\Theta}_0$. Let $\tilde{\Omega}$ denote the linear span of $\mathbf{\Omega} - \mathbf{\Theta}_0$ and define the Fisher inner product on the space $\tilde{\Omega}$ as $< v, \check{v} >= P\{\dot{\ell}(\Theta_0; \mathbf{H}(\cdot\,;\Theta_0))[v]\dot{\ell}(\Theta_0; \mathbf{H}(\cdot\,;\Theta_0))[\check{v}]\}$ for $v, \check{v} \in \tilde{\Omega}$, the Fisher norm as $\|v\| =< v, v >$, where $\dot{\ell}(\Theta_0; \mathbf{H}(\cdot\,;\Theta_0))[v] = \left.\frac{d\ell(\Theta_0 + sv; \mathbf{H}(\cdot\,;\Theta_0))}{ds}\right|_{s=0}$ is the first order directional derivative of $\ell(\Theta_0; \mathbf{H}(\cdot\,;\Theta_0))$ at the direction $v \in \tilde{\Omega}$ (evaluated at Θ_0). Also let $\bar{\tilde{\Omega}}$ be the closed linear span of $\tilde{\Omega}$ under the Fisher norm. Then $(\bar{\tilde{\Omega}}, \|\cdot\|)$ is a Hilbert space. For a vector of d_0-dimension $b = (b_1^\top, b_2^\top, b_3^\top, b_4^\top)^\top$ with $\|b\| \leq 1$ and for any $v \in \tilde{\Omega}$, define a smooth functional of Θ as $h(\Theta) = b^\top\boldsymbol{\theta} = b_1^\top\overrightarrow{\boldsymbol{\beta}} + b_2^\top\boldsymbol{\lambda} + b_3^\top\boldsymbol{\alpha} + b_4^\top\boldsymbol{\varsigma}$ and $\dot{h}(\Theta_0)[v] = \left.\frac{dh(\Theta_0 + sv)}{ds}\right|_{s=0}$, where $\boldsymbol{\theta} = (\overrightarrow{\boldsymbol{\beta}}^\top, \boldsymbol{\lambda}^\top, \boldsymbol{\alpha}^\top, \boldsymbol{\varsigma}^\top)^\top$, whenever the right hand-side limit is well defined. According to the Riesz representation theorem, there exists a $v^* \in \bar{\tilde{\Omega}}$ such that $\dot{h}(\Theta_0)[v] =< v, v^* >$ for all $v \in \bar{\tilde{\Omega}}$ and $\|v^*\| = \|\dot{h}(\Theta_0)\|$. Note $h(\Theta) - h(\Theta_0) = \dot{h}(\Theta_0)(\Theta - \Theta_0)$. Thus, according to the Cramér-Wold device, to prove Theorem 2, it suffices to show that

$$\sqrt{n} < \hat{\Theta}_n - \Theta_0, v^* > \xrightarrow{d} N(0, b^\top I^{-1}(\boldsymbol{\theta}_0)b), \tag{35}$$

due to $b^\top\{(\hat{\overrightarrow{\boldsymbol{\beta}}}_n^\top, \hat{\boldsymbol{\lambda}}_n^\top, \hat{\boldsymbol{\alpha}}_n^\top, \hat{\boldsymbol{\varsigma}}_n^\top, \hat{\boldsymbol{\tau}}_n^\top)^\top - (\overrightarrow{\boldsymbol{\beta}}_0^\top, \boldsymbol{\lambda}_0^\top, \boldsymbol{\alpha}_0^\top, \boldsymbol{\varsigma}_0^\top, \boldsymbol{\tau}_0^\top)^\top\} = h(\hat{\Theta}_n) - h(\Theta_0) = \dot{h}(\Theta_0)(\hat{\Theta}_n - \Theta_0) =< \hat{\Theta}_n - \Theta_0, v^* >$. In fact, (35) holds when $\sqrt{n} < \hat{\Theta}_n - \Theta_0, v^* > \xrightarrow{d} N(0, \|v^*\|^2)$ and $\|v^*\|^2 = b^\top I^{-1}(\boldsymbol{\theta}_0)b$.

We will take two steps to prove (35). First, we prove $\sqrt{n} < \hat{\Theta}_n - \Theta_0, v^* > \to_d N(0, \|v^*\|^2)$. According to the result of Corollary 6.21 in [48], there exists a $\Pi_n v^* \in \mathbf{\Omega}_n - \mathbf{\Theta}_0$ such that $\|\Pi_n v^* - v^*\| = O(n^{-r\upsilon})$. In addition, under the assumptions $r \geq 2$ and $1/2 > v > 1/4r$, we have $\delta_n\|\Pi_n v^* - v^*\| = o(n^{-1/2})$ where $\delta_n = n^{-\min\{(1-\upsilon-e)/2, r\upsilon\}}$. For any $\Theta \in \{\Theta \in \mathbf{\Omega} : d(\Theta, \Theta_0) = O(\delta_n)\}$, define the first and second order directional derivative at the directions $v, \check{v}$ as

$$\dot{\ell}(\Theta; \mathbf{H}(\cdot\,;\Theta))[v] = \frac{d\ell(\Theta + sv, O)}{ds}|_{s=0},$$

$$\ddot{\ell}(\Theta; \mathbf{H}(\cdot\,;\Theta))[v, \check{v}] = \left.\frac{d^2\ell(\Theta + sv + \check{s}\check{v}, O)}{d\check{s}ds}\right|_{s=0,\check{s}=0} = \left.\frac{d\dot{\ell}(\Theta + \check{s}\check{v}, O)[\check{v}]}{d\check{s}}\right|_{\check{s}=0}.$$

Define $r(\Theta - \Theta_0; \mathbf{H}(\cdot\,;\Theta)) = \ell(\Theta; \mathbf{H}(\cdot\,;\Theta)) - \ell(\Theta_0; \mathbf{H}(\cdot\,;\Theta_0)) - \dot{\ell}(\Theta_0; \mathbf{H}(\cdot\,;\Theta_0))(\Theta - \Theta_0)$ and let $\varepsilon_n = o(n^{-1/2})$. Then, by the definition of $\hat{\Theta}_n$ and $P\dot{\ell}(\Theta_0; H(\cdot\,;\Theta_0))[\prod_n v^*] = 0$, we have

$$\begin{aligned}
0 &\le P_n\{\ell(\hat{\Theta}_n; \mathbf{H}(\cdot\,;\hat{\Theta}_n)) - \ell(\hat{\Theta}_n \pm \varepsilon_n \Pi_n v^*; \mathbf{H}(\cdot\,;\hat{\Theta}_n))\} = P_n\{r(\hat{\Theta}_n - \Theta_0; \mathbf{H}(\cdot\,;\hat{\Theta}_n)) + \dot{\ell}(\Theta_0; \mathbf{H}(\cdot\,;\Theta_0))(\hat{\Theta}_n - \Theta_0) \\
&\quad - r(\hat{\Theta}_n + \varepsilon_n \Pi_n v^* - \Theta_0; \mathbf{H}(\cdot\,;\hat{\Theta}_n)) - \dot{\ell}(\Theta_0; \mathbf{H}(\cdot\,;\Theta_0))(\hat{\Theta}_n + \varepsilon_n \Pi_n v^* - \Theta_0)\} \\
&= \mp \varepsilon_n P_n \dot{\ell}(\Theta_0; \mathbf{H}(\cdot\,;\Theta_0))[v^*] \mp \varepsilon_n P_n \dot{\ell}(\Theta_0; \mathbf{H}(\cdot\,;\Theta_0))[\Pi_n v^* - v^*] + (P_n - P)\{r(\hat{\Theta}_n - \Theta_0; \mathbf{H}(\cdot\,;\hat{\Theta}_n)) \\
&\quad - r(\hat{\Theta}_n \pm \varepsilon_n \Pi_n v^* - \Theta_0; \mathbf{H}(\cdot\,;\hat{\Theta}_n))\} + P\{r(\hat{\Theta}_n - \Theta_0; \mathbf{H}(\cdot\,;\hat{\Theta}_n)) - r(\hat{\Theta}_n \pm \varepsilon_n \Pi_n v^* - \Theta_0; \mathbf{H}(\cdot\,;\hat{\Theta}_n))\} \\
&= \mp \varepsilon_n P_n \dot{\ell}(\Theta_0; \mathbf{H}(\cdot\,;\Theta_0))[v^*] \mp I_1 + I_2 + I_3. \qquad (36)
\end{aligned}$$

We will investigate the asymptotic behavior of I_1, I_2 and I_3. For I_1, it follows from Conditions (A1)-(A5), the Chebyshev inequality, and $||\Pi_n v^* - v^*|| = o(1)$ that

$$I_1 = \varepsilon_n \times o_p(n^{-1/2}). \qquad (37)$$

For I_2, due to the mean value theorem, we obtain that

$$\begin{aligned}
I_2 &= (P_n - P)\{\ell(\hat{\Theta}_n; \mathbf{H}(\cdot\,;\hat{\Theta}_n)) - \ell(\hat{\Theta}_n \pm \varepsilon_n \Pi_n v^*; \mathbf{H}(\cdot\,;\hat{\Theta}_n)) \pm \varepsilon_n \ell(\Theta_0; H(\cdot\,;\Theta_0))[\varepsilon_n \Pi_n v^*]\} \\
&= \mp \varepsilon_n (P_n - P)[\{\dot{\ell}(\tilde{\Theta}; \mathbf{H}(\cdot\,;\tilde{\Theta})) - \dot{\ell}(\Theta_0; \mathbf{H}(\cdot\,;\Theta_0))\}[\Pi_n v^*]], \qquad (38)
\end{aligned}$$

where $\tilde{\Theta}$ lies between $\hat{\Theta}_n$ and $\hat{\Theta}_n \pm \varepsilon_n \Pi_n v^*$. By Theorem 2.8.3 of [55], we know that $\{\dot{\ell}(\Theta; \mathbf{H}(\cdot\,;\Theta))[\Pi_n v^*] : ||\Theta - \Theta_0|| = O_p(\delta_n)\}$ is a Donsker class. Hence by Theorem 2.11.23 of [55], we get $I_2 = \varepsilon_n \times o_p(n^{-1/2})$. Since

$$\begin{aligned}
&P(r[\Theta - \Theta_0; \mathbf{H}(\cdot\,;\Theta)]) = P\{\ell(\Theta; \mathbf{H}(\cdot\,;\Theta)) - \ell(\Theta_0; \mathbf{H}(\cdot\,;\Theta_0)) - \dot{\ell}(\Theta_0; \mathbf{H}(\cdot\,;\Theta_0))[\Theta - \Theta_0]\} \\
&\quad = \frac{1}{2}P\{\ddot{\ell}(\tilde{\Theta}; \mathbf{H}(\cdot\,;\tilde{\Theta}))[\Theta - \Theta_0, \Theta - \Theta_0] - \ddot{\ell}(\Theta_0; \mathbf{H}(\cdot\,;\Theta_0))[\Theta - \Theta_0, \Theta - \Theta_0]\} + \frac{1}{2}P\ddot{\ell}(\Theta_0; \mathbf{H}(\cdot\,;\Theta_0))[\Theta - \Theta_0, \Theta - \Theta_0] \\
&\quad = \frac{1}{2}P\ddot{\ell}(\Theta_0; \mathbf{H}(\cdot\,;\Theta_0))[\Theta - \Theta_0, \Theta - \Theta_0] + \varepsilon_n \times o_p(n^{-1/2}),
\end{aligned}$$

where $\tilde{\Theta}$ is between Θ and Θ_0 and the last equation follows from the Taylor expansion, conditions A(1)-A(5) and $r \ge 2$, $1/2 > \upsilon > 1/4r$. Therefore

$$\begin{aligned}
I_3 &= -\frac{1}{2}\{||\hat{\Theta}_n - \Theta_0||^2 - ||\hat{\Theta}_n \pm \varepsilon_n \Pi_n v^* - \Theta_0||^2\} + \varepsilon_n \times o_p(n^{-1/2}) = \pm \varepsilon_n < \hat{\Theta}_n - \Theta_0, \Pi_n v^* > + \frac{1}{2}||\varepsilon_n \Pi_n v^*||^2 + \varepsilon_n \times o_p(n^{-1/2}) \\
&= \pm \varepsilon_n < \hat{\Theta}_n - \Theta_0, \Pi_n v^* - v^* + v^* > + \frac{1}{2}||\varepsilon_n \Pi_n v^*||^2 + \varepsilon_n \times o_p(n^{-1/2}) \\
&= \pm \varepsilon_n < \hat{\Theta}_n - \Theta_0, v^* > + \frac{1}{2}||\varepsilon_n \Pi_n v^*||^2 + \varepsilon_n \times o_p(n^{-1/2}) = \pm \varepsilon_n < \hat{\Theta}_n - \Theta_0, v^* > + \varepsilon_n \times o_p(n^{-1/2}), \qquad (39)
\end{aligned}$$

where the last equality holds because $\delta_n ||\Pi_n v^* - v^*|| = o(n^{-1/2})$, Cauchy-Schwarz inequality, and $||\Pi_n v^*||^2 \to ||v^*||$. By (36) - (39), combined with $P\dot{\ell}(\Theta_0; \mathbf{H})[v^*] = 0$, we can establish that

$$\begin{aligned}
0 &\le P_n\{\ell(\hat{\Theta}_n; \mathbf{H}(\cdot\,;\hat{\Theta}_n)) - \ell(\hat{\Theta}_n \pm \varepsilon_n \Pi_n v^*; \mathbf{H}(\cdot\,;\hat{\Theta}_n))\} \\
&= \mp \varepsilon_n P_n \dot{\ell}(\Theta_0; \mathbf{H}(\cdot\,;\Theta_0))[v^*] \pm \varepsilon_n < \hat{\Theta}_n - \Theta_0, v^* > + \varepsilon_n \times o_p(n^{-1/2}) \\
&= \mp \varepsilon_n (P_n - P)\dot{\ell}(\Theta_0; \mathbf{H}(\cdot\,;\Theta_0))[v^*] \pm \varepsilon_n < \hat{\Theta}_n - \Theta_0, v^* > + \varepsilon_n \times o_p(n^{-1/2}).
\end{aligned}$$

Therefore, we obtain $\mp \sqrt{n}(P_n - P)\dot{\ell}(\Theta_0; \mathbf{H})[v^*] \pm \sqrt{n} < \hat{\Theta}_n - \Theta_0, v^* > + o_p(1) \ge 0$. Further, with the central limit theorem, we have $\sqrt{n} < \hat{\Theta}_n - \Theta_0, v^* >= \sqrt{n}(P_n - P)\dot{\ell}(\Theta_0; \mathbf{H})[v^*] + o_p(1) \to N(0, ||v^*||^2)$ and $||v^*||^2 = ||\dot{\ell}(\Theta_0; \mathbf{H})[v^*]||^2$.

Now we calculate $||v^*||$. Rewrite $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{d_0})$. For each component $\boldsymbol{\theta}_q$ where $q = 1, 2, \ldots, d_0$, let $\psi_q^* = (b_{1q}^*, b_{2q}^*, \ldots, b_{(K+1)q+K}^*)$ be the minimizer of $E\{\ell_{\boldsymbol{\theta}} \cdot e_q - \ell_{b_1}[b_{1q}] - \tilde{\ell}_{b_2}[b_{2q}] - \ldots - \ell_{b_{(K+1)q+K}}[b_{(K+1)q+K}]\}^2$ with respect to $\psi_q = (b_{1q}, b_{2q}, \ldots, b_{(K+1)q+K})$, where $\ell_{\boldsymbol{\theta}} = (\ell_{\vec{\boldsymbol{\beta}}}^\top, \ell_{\boldsymbol{\lambda}}^\top, \ell_{\boldsymbol{\alpha}}^\top, \ell_{\boldsymbol{\varsigma}}^\top)^\top$, $\ell_{\vec{\boldsymbol{\beta}}} = (\ell_{\beta_1}^\top, \ldots, \ell_{\beta_{pp_1}}^\top)^\top$, $\ell_{\boldsymbol{\lambda}} = (\ell_{\lambda_2}, \ldots, \ell_{\lambda_p})^\top$, $\ell_{\boldsymbol{\alpha}} = (\ell_{\alpha_1}^\top, \ldots, \ell_{\alpha_{p_2}}^\top)^\top$, $\ell_{\boldsymbol{\varsigma}} = (\ell_{\varsigma_1}^\top, \ldots, \ell_{\varsigma_p}^\top)^\top$, $\ell_{\beta_k} = \frac{\partial \ell}{\partial \beta_k}$, $\ell_{\lambda_k} = \frac{\partial \ell}{\partial \lambda_k}$, $\ell_{\alpha_k} = \frac{\partial \ell}{\partial \alpha_k}$, $\ell_{\varsigma_k} = \frac{\partial \ell}{\partial \varsigma_k}$, and e_q is a d_0 dimensional vector of zeros except with the $q-$th element equal to 1.

Define a vector $S_{\boldsymbol{\theta}}$ of dimension d_0, with the q-th element as $\ell_{\boldsymbol{\theta}} \cdot e_q - \ell_{b_1}[b_{1q}^*] - \ell_{b_2}[b_{2q}^*] - \ldots - \ell_{b_{s_0}}[b_{(K+1)q+K}^*]$, and then $I(\boldsymbol{\theta}) = E(S_{\boldsymbol{\theta}} S_{\boldsymbol{\theta}}^\top)$, $I(\boldsymbol{\theta}_0) = E(S_{\boldsymbol{\theta}_0} S_{\boldsymbol{\theta}_0}^\top)$. Further, following [7], we obtain

$$||v^*|| = ||\dot{h}(\Theta_0)|| = \sup_{v \in \bar{V}: ||v|| > 0} \frac{|\dot{h}(\Theta_0)|^2}{v^2} = b^\top [E(S_{\boldsymbol{\theta}_0} S_{\boldsymbol{\theta}_0}^\top)]^{-1} b = b^\top I^{-1}(\boldsymbol{\theta}_0) b.$$

Supplementary material

It contain Fig.1-3 for the histograms of MMSE, BVRT and IST in each group of the Cognitive Decline data, Tables 1-2 for the simulation results of Examples 3-5, and the R code of our simulation.

[1] D. Alley, K. Suthers, E. Crimmins, Education and cognitive decline in older Americans: results from the ahead sample, Research on Aging 29 (2007) 73–94.

[2] G. Aneiros, R. Cao, R. Fraiman, C. Genest, P. Vieu, Recent advances in functional data analysis and high-dimensional statistics, Journal of Multivariate Analysis 170 (2019) 3–9.

[3] R. B. Ash, M. F. Gardner, Topics in Stochastic Processes: Probability and Mathematical Statistics: A Series of Monographs and Textbooks, volume 27, Academic Press, 2014.

[4] B. Cai, D. B. Dunson, J. B. Stanford, Dynamic model for multivariate markers of fecundability, Biometrics 66 (2010) 905–913.

[5] R. J. Carroll, J. Fan, I. Gijbels, M. P. Wand, Generalized partially linear single-index models, Journal of the American Statistical Association 92 (1997) 477–489.

[6] K. Chen, X. Tong, Varying coefficient transformation models with censored data, Biometrika 97 (2010) 969–976.

[7] X. Chen, Y. Fan, V. Tsyrennikov, Efficient estimation of semiparametric multivariate copula models, Journal of the American Statistical Association 101 (2006) 1228–1240.

[8] J.-M. Chiou, Y.-F. Yang, Y.-T. Chen, Multivariate functional linear regression and prediction, Journal of Multivariate Analysis 146 (2016) 301–312.

[9] M. J. Daniels, S.-L. T. Normand, Longitudinal profiling of health care units based on continuous and discrete patient outcomes, Biostatistics 7 (2006) 1–15.

[10] A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society: Series B (Methodological) 39 (1977) 1–22.

[11] D. B. Dunson, Dynamic latent trait models for multidimensional longitudinal data, Journal of the American Statistical Association 98 (2003) 555–563.

[12] D. B. Dunson, Bayesian dynamic modeling of latent trait distributions, Biostatistics 7 (2006) 551–568.

[13] D. B. Dunson, Empirical bayes density regression, Statistica Sinica 17 (2007) 481–504.

[14] C. Fabrigoule, I. Rouch, A. Taberly, L. Letenneur, D. Commenges, J.-M. Mazaux, J.-M. Orgogozo, J.-F. Dartigues, Cognitive process in preclinical phase of dementia, Brain 121 (1998) 135–141.

[15] S. Van de Geer, Empirical Processes in M-estimation, volume 6, Cambridge University Press, 2000.

[16] A. Goia, P. Vieu, An introduction to recent advances in high/infinite dimensional statistics, Journal of Multivariate Analysis 146 (2016) 1–6.

[17] M. J. Gurka, L. J. Edwards, K. E. Muller, L. L. Kupper, Extending the Box–Cox transformation to the linear mixed model, Journal of the Royal Statistical Society: Series A (Statistics in Society) 169 (2006) 273–288.

[18] P. Hall, J. L. Horowitz, Methodology and convergence rates for functional linear regression, Annals of Statistics 35 (2007) 70–91.

[19] P. Hall, M. Hosseini-Nasab, On properties of functional principal components analysis, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68 (2006) 109–126.

[20] P. Hall, H.-G. Müller, F. Yao, Modelling sparse generalized longitudinal observations with latent gaussian processes, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70 (2008) 703–723.

[21] D. J. Hand, M. J. Crowder, Practical Longitudinal Data Analysis, volume 34, CRC Press, London, 1996.

[22] A. H. Herring, J. Yang, Bayesian modeling of multiple episode occurrence and severity with a terminating event, Biometrics 63 (2007) 381–388.

[23] J. L. Horowitz, Semiparametric estimation of a regression model with an unknown transformation of the dependent variable, Econometrica: Journal of the Econometric Society (1996) 103–137.

[24] J. Z. Huang, Local asymptotics for polynomial spline regression, Annals of Statistics 31 (2003) 1600–1635.

[25] H. Jacqmin-Gadda, C. Proust-Lima, H. Amiéva, Semi-parametric latent process model for longitudinal ordinal data: Application to cognitive decline, Statistics in Medicine 29 (2010) 2723–2731.

[26] G. M. James, T. J. Hastie, C. A. Sugar, Principal component models for sparse functional data, Biometrika 87 (2000) 587–602.

[27] D. Kong, K. Xue, F. Yao, H. H. Zhang, Partially functional linear regression in high dimensions, Biometrika 103 (2016) 147–159.

[28] L. J. Launer, M. A. Dinkgreve, C. Jonker, C. Hooijer, J. Lindeboom, Are age and education independent correlates of the mini-mental state exam performance of community-dwelling elderly?, Journal of Gerontology 48 (1993) 271–277.

[29] L. Letenneur, D. Commenges, J.-F. Dartigues, P. Barberger-Gateau, Incidence of dementia and alzheimer's disease in elderly community residents of south-western france, International Journal of Epidemiology 23 (1994) 1256–1261.

[30] H. Lian, H. Liang, R. J. Carroll, Variance function partially linear single-index models, Journal of the Royal Statistical Society. Series B, Statistical methodology 77 (2015) 171–194.

[31] H. Lin, X. Zhou, G. Li, A direct semiparametric receiver operating characteristic curve regression with unknown link and baseline functions, Statistica Sinica 22 (2012) 1427–1456.

[32] Z. Lin, H.-G. Müller, F. Yao, Mixture inner product spaces and their application to functional data analysis, The Annals of Statistics 46 (2018) 370–400.

[33] S. R. Lipsitz, J. Ibrahim, G. Molenberghs, Using a box–cox transformation in the analysis of longitudinal data with incomplete responses, Journal of the Royal Statistical Society: Series C (Applied Statistics) 49 (2000) 287–296.

[34] M. Lu, Y. Zhang, J. Huang, Semiparametric estimation methods for panel count data using monotone b-splines, Journal of the American Statistical Association 104 (2009) 1060–1070.

[35] J. S. Morris, R. J. Carroll, Wavelet-based functional mixed models, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68 (2006) 179–199.

[36] M. C. Morris, D. A. Evans, L. E. Hebert, J. L. Bienias, Methodological issues in the study of cognitive decline, American Journal of Epidemiology 149 (1999) 789–793.

[37] B. Muthén, A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators, Psychometrika 49 (1984) 115–132.

[38] M. K. Olsen, J. L. Schafer, A two-part random-effects model for semicontinuous longitudinal data, Journal of the American Statistical Association 96 (2001) 730–745.

[39] J. M. Parisi, G. W. Rebok, Q.-L. Xue, L. P. Fried, T. E. Seeman, E. K. Tanner, T. L. Gruenewald, K. D. Frick, M. C. Carlson, The role of education and intellectual activity on cognition, Journal of Aging Research (2012).

[40] D. Paul, J. Peng, Consistency of restricted maximum likelihood estimators of principal components, The Annals of Statistics 37 (2009) 1229–1271.

[41] J. Peng, D. Paul, A geometric approach to maximum likelihood estimation of the functional principal components from sparse longitudinal data, Journal of Computational and Graphical Statistics 18 (2009) 995–1015.

[42] D. Pollard, Convergence of Stochastic Processes, Springer, New York, 1984.

[43] C. Proust, H. Jacqmin-Gadda, J. M. Taylor, J. Ganiayre, D. Commenges, A nonlinear model with latent process for cognitive evolution using multivariate longitudinal data, Biometrics 62 (2006) 1014–1024.

[44] C. Proust-Lima, H. Amieva, H. Jacqmin-Gadda, Analysis of multivariate mixed longitudinal data: a flexible latent process approach, British Journal of Mathematical and Statistical Psychology 66 (2013) 470–487.

[45] C. Proust-Lima, V. Philipps, B. Liquet, Estimation of extended mixed models using latent classes and latent processes: the r package lcmm, Journal of Statistical Software 78 (2017) 1–56.

[46] M. Reite, C. M. Cullum, J. Stocker, P. Teale, E. Kozora, Neuropsychological test performance and meg-based brain lateralization: sex differences, Brain Research Bulletin 32 (1993) 325–328.

[47] T. A. Salthouse, H. E. Hancock, E. J. Meinz, D. Z. Hambrick, Interrelations of age, visual acuity, and cognitive functioning, The Journals of Gerontology Series B: Psychological Sciences and Social Sciences 51 (1996) 317–330.

[48] L. Schumaker, Spline functions: basic theory, Cambridge University Press, 2007.

[49] X. Shen, W. H. Wong, Convergence rate of sieve estimates, The Annals of Statistics (1994) 580–615.

[50] J. Shi, T. Choi, Gaussian Process Regression Analysis for Functional Data, CRC Press, London, 2011.

[51] C. J. Stone, Optimal rates of convergence for nonparametric estimators, The Annals of Statistics (1980) 1348–1360.

[52] C. L. Sung, Y. Hung, W. Rittase, C. Zhu, C. J. Wu, A generalized gaussian process model for computer experiments with binary time series, Journal of the American Statistical Association 115 (2020) 945–956.

[53] T. N. Tombaugh, N. J. McIntyre, The mini-mental state examination: a comprehensive review, Journal of the American Geriatrics Society 40 (1992) 922–935.

[54] A. W. Van der Vaart, Asymptotic Statistics, volume 3, Cambridge University Press, Cambridge, 2000.

[55] A. W. Van der Vaart, . A. Wellner, Weak Convergence and Empirical Processes: With Applications to Statistics, Springer, New York, 1996.

[56] B. Wang, J. Shi, Generalized Gaussian process regression model for non-gaussian functional data, Journal of the American Statistical Association 109 (2014) 1123–1133.

[57] W. Wiederholt, D. Cahn, N. M. Butters, D. P. Salmon, D. Kritz-Silverstein, E. Barrett-Connor, Effects of age, gender and education on selected neuropsychological tests in an elderly community cohort, Journal of the American Geriatrics Society 41 (1993) 639–647.

[58] R. Wilson, L. Hebert, P. Scherr, L. Barnes, C. M. De Leon, D. Evans, Educational attainment and cognitive decline in old age, Neurology 72 (2009) 460–465.

[59] F. Yao, H.-G. Müller, J.-L. Wang, Functional data analysis for sparse longitudinal data, Journal of the American Statistical Association 100 (2005) 577–590.

[60] J. A. Yesavage, J. O. Brooks, On the importance of longitudinal research in Alzheimer's disease, Journal of the American Geriatrics Society (1991).

[61] L. B. Zahodne, M. M. Glymour, C. Sparks, D. Bontempo, R. A. Dixon, S. W. MacDonald, J. J. Manly, Education does not slow cognitive decline with aging: 12-year evidence from the victoria longitudinal study, Journal of the International Neuropsychological Society 17 (2011) 1039–1046.

[62] L. B. Zahodne, Y. Stern, J. J. Manly, Differing effects of education on cognitive decline in diverse elders with low versus high educational attainment, Neuropsychology 29 (2015) 649–657.

[63] S. Zheng, L. Yang, W. K. Härdle, A smooth simultaneous confidence corridor for the mean of sparse functional data, Journal of the American Statistical Association 109 (2014) 661–673.

[64] X. Zhou, H. Lin, E. Johnson, Non-parametric heteroscedastic transformation regression models for skewed data with an application to health care costs, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70 (2008) 1029–1047.

[65] H. Zhu, J. S. Morris, F. Wei, D. D. Cox, Multivariate functional response regression, with application to fluorescence spectroscopy in a cervical pre-cancer study, Computational Statistics & Data Analysis 111 (2017) 88–101.