Journal of Machine Learning Research 24 (2023) 1-41 Submitted 6/22; Published 7/23

# Functional L-Optimality Subsampling for Functional Generalized Linear Models with Massive Data

**Hua Liu**† LIUHUA_22@XJTU.EDU.CN
*School of Economics and Finance*
*Xi'an Jiaotong University*
*Xi'an, Shaanxi 710049, China*

**Jinhong You**† JOHNYOU07@163.COM
*School of Statistics and Management*
*Shanghai University of Finance and Economics*
*Shanghai 200433, China*

**Jiguo Cao**∗ JIGUO_CAO@SFU.CA
*Department of Statistics and Actuarial Science*
*Simon Fraser University*
*Burnaby, BC V5A 1S6, Canada*

**Editor:** Genevera Allen

## Abstract

Massive data bring the big challenges of memory and computation for analysis. These challenges can be tackled by taking subsamples from the full data as a surrogate. For functional data, it is common to collect multiple measurements over their domains, which require even more memory and computation time when the sample size is large. The computation would be much more intensive when statistical inference is required through bootstrap samples. Motivated by analyzing large-scale kidney transplant data, we propose an optimal subsampling method based on the functional L-optimality criterion for functional generalized linear models. To the best of our knowledge, this is the first attempt to propose a subsampling method for functional data analysis. The asymptotic properties of the resultant estimators are also established. The analysis results from extensive simulation studies and from the kidney transplant data show that the functional L-optimality subsampling (FLoS) method is much better than the uniform subsampling approach and can well approximate the results based on the full data while dramatically reducing the computation time and memory.

**Keywords:** Efficient computation algorithm; Functional data analysis; Kidney transplant; Large-scale data; Penalized B-spline

∗. Corresponding author.
†. These two authors contributed equally to this paper and shared the first authorship.

## 1. Introduction

Functional data are data collected at multiple times, spatial locations or any other continuum (Ramsay and Silverman, 2002; Morris, 2015; Wang et al., 2016). Our research is motivated by the analysis of a massive functional data set arising from studying organ transplantation. Organ transplantation is one of the great advances in modern medicine and is the ultimate effective means to treat diseases. It is necessary because it can replace those unusable organs of the recipients and make the recipients' life continue to be relayed and endless, but at the same time, the transplantation surgery is complex and expensive. The organ transplant data from the Organ Procurement Transplant Network/United Network for Organ Sharing (Optn/UNOS) as of September 2020 is a massive functional data, which is available at `https://optn.transplant.hrsa.gov/` with the permission of OPTN/UNOS. This data set collects the basic description (for example, age, race, gender, and height) of about 500,000 recipients of kidney transplants at the time of transplant since October 1, 1987, and their information during the followed-up period (for example, serum creatinine, recipient status and the follow-up time), which can be used to check whether the transplant was successful. The health status of the kidney can be measured by checking the estimated glomerular filtration rate (eGFR) with details shown in Section 6.1.

For this kind of massive functional data set, too much computing memory is required when using the full data for analysis, sometimes even exceeding the available computational resources. When the sample size of functional data is large, we also have to increase the computational efficiency. For instance, one important tool in functional data analysis (FDA) is the functional generalized linear model (James, 2002; Cardot and Sarda, 2005; Müller and Stadtmüller, 2005; Crainiceanu et al., 2009; McLean et al., 2014), which describes the relationship between the functional predictors and the scalar response from an exponential family distribution (for example, the Binomial distribution and Poisson distribution). An iterative optimization procedure is needed to estimate the functional generalized linear model. The corresponding computational complexity is $O(nK + \mathcal{I}nK^2)$ when using the penalized B-splines method (Cardot et al., 2003; Xiao, 2019), where $n$ is the sample size of functional data, $K$ is the number of knots, and $\mathcal{I}$ is the number of iterations. Usually, the number of knots, $K$, is chosen to be relatively large to capture the local features of the functional coefficients. We also need to select the optimal smoothing parameter by the Bayesian information criterion (BIC). Consequently, when the number of functional data is large, the computing time based on the full data may be too long.

To tackle the above computing challenges, an effective method is to take random subsamples from the massive data as a surrogate. The existing subsampling literature mainly focuses on models with only scalar variables. For instance, Ma et al. (2015) used the probabilities based on statistical leverage scores to randomly subsample data and established the asymptotic properties of the resultant estimators for linear regression. A method named information-based optimal subdata selection (IBOSS) proposed by Wang et al. (2019) selects subsample data deterministically without involving random sampling. Ai et al. (2021) investigated the optimal subsampling method under the A-optimality criterion (OSMAC) for generalized linear models. A Poisson subsampling method based

on the A-optimality or L-optimality criterion was used for maximum quasi-likelihood estimation in Yu et al. (2022). And Ma et al. (2022) studied the asymptotic normality and asymptotic unbiasedness of the leveraging sampling estimator. We refer the readers to Yao and Wang (2021) for a recent review of optimal subsampling methods of massive data when both the response and the predictors are scalar.

It is worth mentioning that there is little work on subsampling in the field of FDA. The simplest subsampling method to solve the computing challenges with functional predictors is to draw the sample uniformly at random, which will perform poorly when the leverage scores are non-uniform. Moreover, in order to make the B-spline approximation asymptotically unbiased, a relatively large $K$ is usually chosen. A roughness penalty is used to ensure the smoothness of the estimator, which results in the variance of the subsample estimator too complicated to be adapted by IBOSS in Wang et al. (2019) and Cheng et al. (2020). In addition, IBOSS is based on the order statistics of each scalar predictor variable. The functional predictor variable in the functional linear model is a curve and is difficult to be ordered. As a result, IBOSS is not suitable for the subsampling with functional predictors.

In this paper, we first estimate the functional coefficient using the subsampling data, and derive the asymptotic distribution of the general subsampling estimator. Then, we obtain the optimal subsampling probabilities by minimizing the asymptotic integrated mean squared errors (IMSE) and propose the functional L-optimality criterion. The derived optimal subsampling probability is not only related to the predictors but also the responses. In comparison, uniform subsampling treats every subject equally, which ignores the different information from subjects. Lastly, we attain the optimal subsampling estimator based on the optimal subdata drawn according to the optimal probability calculated above. Our proposed method is called the functional L-optimality subsampling (FLoS) method in this article. We establish the asymptotic results of the FLoS estimators for the functional generalized linear model. In addition, an R package `SubsamplingFunPredictors` has been developed for implementing the FLoS method. The R package and the R codes for the simulation studies can be downloaded at `https://github.com/caojiguo/FLoS`.

To the best of our knowledge, this is the first attempt to propose the subsampling method for functional data analysis. The FLoS method has several advantages. (1) The computing complexity for this method is $O(nK)$. It is significantly faster than $O(nK + \mathcal{I}nK^2)$ when using the full data. (2) The root integrated mean square errors (RIMSEs) of the estimators using the FLoS method are smaller than those using the uniform subsampling method. (3) We can calculate the subsampling probabilities on each subset independently. Therefore, distributed parallel computing can be adapted based on the FLoS method. (4) One by-product of the FLoS method is to make statistical inferences using multiple subsampling datasets in parallel computing, which has a more obvious advantage in reducing computing time.

The rest of this article is organized as follows. In Section 2, we briefly introduce the functional generalized linear model and the estimation based on the full data. Section 3 derives the optimal subsampling strategy and the optimal subsampling algorithm based on the functional L-optimality criteria for the estimator of the functional coefficient. The

asymptotic behaviours for the optimal subsampling estimator are also investigated in this section. The evaluation of the numerical performance of our proposed estimator via simulation studies is presented in Section 4. We also illustrate our method by analyzing a real data set in Section 5. Some conclusions and discussions are provided in Section 6.

## 2. Model and Full Data Estimation

This article studies optimal subsampling in regression with a functional predictor and a discrete scalar response. Let $Y$ be a scalar response, and $Z(t)$ be a process defined on a domain $[a, b]$, the form of the basic functional generalized linear model can be expressed as:

$$E(Y|X) = \psi\left(\alpha + \int_a^b Z(t)\beta(t)dt\right),$$

where $\alpha$ is the intercept, $\beta(t)$ is the unknown slope function, $\psi(\cdot)$ is a twice continuously differentiable function and the function $\psi^{-1}(\cdot)$ is called the link function.

We use the B-spline basis functions (de Boor, 1978) to approximate the functional coefficient $\beta(t)$. On the domain $[a, b]$, we define a knot sequence with $K$ interior knots $a = k_0 < k_1 < \ldots < k_K < k_{K+1} = b$. For $p$th degree B-spline basis, we define the additional knots: $k_{-p} = k_{-p+1} = \cdots = k_{-1} = k_0$, and $k_{K+1} = k_{K+2} = \cdots = k_{K+p+1}$. For $p \geq 1$, let $\mathcal{S}(p+1; k) = \{s(\cdot) \in \mathcal{C}^{p-1}[a, b] : s \text{ is a degree } p \text{ polynomial on each } [k_j, k_{j+1}]\}$ be the space of polynomial splines of degree $p$, where $\mathcal{C}^{p-1}([a, b])$ is the collection of all functions that have $p-1$ order bounded continuous derivatives on $[a, b]$. According to the definition of B-spline basis functions, the order is equal to $p+1$, and the total number of basis functions with degree $p$ and $K$ interior knots is $K + p + 1$. Denote the $p$th degree B-spline basis for $\mathcal{S}(p+1; k)$ as $\boldsymbol{N}_{p+1}(t) = (N_{j,p+1}(t) : -p \leq j \leq K)^T$ (Schumaker, 1981). And let $s_\beta(t) = \boldsymbol{N}_{p+1}^T(t)\boldsymbol{c}_\beta \in \mathcal{S}(p+1, k)$ be the approximation to the functional coefficient $\beta(t)$ (Claeskens et al., 2009), where $\boldsymbol{N}_{p+1}^T(t)$ is the transpose of $\boldsymbol{N}_{p+1}(t)$.

Suppose the full data $\mathcal{F}_n = \{(z_i(t), y_i), i = 1, \ldots, n; t \in [a, b]\}$ consists of $n$ independent and identically distributed (i.i.d) observations of $(Y, Z(t), t \in [a, b])$. Denote $\boldsymbol{x}_i(t) = (1, z_i(t))^T$, $\boldsymbol{X}(t) = (\boldsymbol{x}_1(t), \cdots, \boldsymbol{x}_n(t))^T$,

$$\boldsymbol{N}(t) = \begin{pmatrix} 1/(b-a) & 0 \\ \boldsymbol{0}_{K+p+1} & \boldsymbol{N}_{p+1}(t) \end{pmatrix},$$

$\boldsymbol{N}_i = \int_a^b \boldsymbol{N}(t)\boldsymbol{x}_i(t)dt$, $\boldsymbol{N} = \int_a^b \boldsymbol{X}(t)\boldsymbol{N}^T(t)dt = (\boldsymbol{N}_1, \ldots, \boldsymbol{N}_n)^T$, where $\boldsymbol{0}_{K+p+1}$ is a $(K + p + 1) \times 1$ vector with all elements equal to 0. Combining the maximum quasi-likelihood estimator in the generalized linear model (Chen et al., 1999; Yu et al., 2022) and the penalized B-splines, we can obtain the penalized quasi-likelihood estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{PQL}}(t) = (\widehat{\alpha}_{\mathrm{PQL}}, \widehat{\beta}_{\mathrm{PQL}}(t))^T = \boldsymbol{N}^T(t)\widehat{\boldsymbol{c}}_{\mathrm{PQL}}$ of the functional coefficient vector $\boldsymbol{\beta}(t) = (\alpha, \beta(t))^T$, where $\widehat{\boldsymbol{c}}_{\mathrm{PQL}} = (\widehat{\alpha}_{\mathrm{PQL}}, \widehat{\boldsymbol{c}}_{\beta,\mathrm{PQL}}^T)^T$ can be obatined by solving the following equation:

$$Q_{\mathrm{PQL}}(\boldsymbol{c}) = \sum_{i=1}^{n}\{y_i - \psi(\boldsymbol{N}_i^T\boldsymbol{c})\}\boldsymbol{N}_i - \lambda\boldsymbol{D}\boldsymbol{c} = 0, \tag{1}$$

with $\boldsymbol{c} = (\alpha, \boldsymbol{c}_\beta^T)^T$ and the non-negative smoothing parameter $\lambda$. In the above criterion, the first term is the quasi-likelihood function (Chen et al., 1999; Yu et al., 2022), and the second term is the roughness penalty that aims to enforce smoothness of $\hat{\beta}_{\mathrm{PQL}}(t)$, where $\boldsymbol{N}_{p+1}^{(q)}(t)$ is the $q$th order derivative of $\boldsymbol{N}_{p+1}(t)$ with $q \leq p$, $\boldsymbol{D}_q = \int_a^b \{\boldsymbol{N}_{p+1}^{(q)}(t)\}\{\boldsymbol{N}_{p+1}^{(q)}(t)\}^T dt$, and

$$\boldsymbol{D} = \begin{pmatrix} 0 & \mathbf{0}_{K+p+1}^T \\ \mathbf{0}_{K+p+1} & \boldsymbol{D}_q \end{pmatrix}.$$

## 3. The FLoS Method

In this section, we propose an optimal subsampling algorithm and then establish the asymptotic properties of the resultant estimators.

### 3.1 Subsample estimator

Denote the full data as $\mathcal{F}_n = \{(z_i(t), y_i), i = 1, \ldots, n; t \in [a, b]\}$. Let $\eta_i$ be the indicator variable, that is, $\eta_i = 1$ if $(z_i(t), y_i; t \in [a, b])$ is included in the subdata, and $\eta_i = 0$ otherwise. Thus, $\eta_i \sim$ Bernoulli$(p_i)$ with $\sum_{i=1}^n p_i = 1$. The subsample penalized quasi-likelihood estimator, denoted as $\widetilde{\boldsymbol{\beta}}_{\mathrm{PQL}}(t)$ is given by $\widetilde{\boldsymbol{\beta}}_{\mathrm{PQL}}(t) = \boldsymbol{N}^T(t)\widetilde{\boldsymbol{c}}_{\mathrm{PQL}}$, where $\widetilde{\boldsymbol{c}}_{\mathrm{PQL}}$ can be obtained through the equation

$$Q_{\mathrm{PQL}}^*(\boldsymbol{c}) := \sum_{i=1}^n R_i\{y_i - \psi(\boldsymbol{N}_i^T\boldsymbol{c})\}\boldsymbol{N}_i/(Lp_i) - \lambda\boldsymbol{D}\boldsymbol{c} = 0. \tag{2}$$

where $R_i = \sum_{l=1}^L \eta_{il}$ denotes the total number of times that the $i$-th observation is selected into the sample out of the $L$ sampling steps, $R_i \sim$ Binomial$(L, p_i)$, and the objective function is weighted by the sampling probabilities $p_i$.

In addition, let $\boldsymbol{\Psi} = \mathrm{Diag}(\dot{\psi}(\boldsymbol{N}_1^T\boldsymbol{c}), \cdots, \dot{\psi}(\boldsymbol{N}_n^T\boldsymbol{c}))$, $\boldsymbol{G}_{k,n}^{\psi} = \boldsymbol{N}^T\boldsymbol{\Psi}\boldsymbol{N}/n$ and $\boldsymbol{H}_{k,n}^{\psi} = \boldsymbol{G}_{k,n}^{\psi} + \lambda\boldsymbol{D}/n$, where $\dot{\psi}(\cdot)$ is the first order deriative of $\psi(\cdot)$. In the following theorem, we give the asymptotic normality of the subsample estimator.

**Theorem 1** *Under Assumptions 1-7, for any given $t$, as $L, n \to \infty$, we have* $\left\{\boldsymbol{N}^T(t)\right.$ $\left.\boldsymbol{H}_{k,n}^{\psi,-1}\boldsymbol{W}_p^{\psi}\ \boldsymbol{H}_{k,n}^{\psi,-1}\boldsymbol{N}(t)/L\right\}^{-1/2}(\widetilde{\boldsymbol{\beta}}_{\mathrm{PQL}}(t) - \boldsymbol{\beta}(t)) \to \mathbb{N}(\mathbf{0}_2, \boldsymbol{I}_2)$, *where*

$$\boldsymbol{W}_p^{\psi} = n^{-2}\sum_{i=1}^n E\left\{\left(y_i - \psi(\boldsymbol{N}_i^T\boldsymbol{c})\right)^2\right\}\boldsymbol{N}_i\boldsymbol{N}_i^T/p_i. \tag{3}$$

In (3), the term $\mathrm{E}\left\{\left(y_i - \psi(\boldsymbol{N}_i^T\boldsymbol{c})\right)^2\right\}$ is unknown, so the optimal subsampling probabilities are not directly implementable based the asymptotic covariance matrix of $\widetilde{\boldsymbol{\beta}}_{\mathrm{PQL}}(t)$. Next, we establish the asymptotic normality of estimator $\widetilde{\boldsymbol{\beta}}_{\mathrm{PQL}}(t)$ in approximating the full data estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{PQL}}(t)$ to obtain the optimal subsampling probabilities.

**Theorem 2** *Under Assumptions 1-7, for any given $t$, as $L, n \to \infty$, conditionally on $\mathcal{F}_n$ in probability, $\left\{\boldsymbol{N}^T(t)\boldsymbol{H}_{k,n}^{\psi,-1}\boldsymbol{V}_p^{\psi}\boldsymbol{H}_{k,n}^{\psi,-1}\boldsymbol{N}(t)/L\right\}^{-1/2}(\widetilde{\boldsymbol{\beta}}_{\mathrm{PQL}}(t)-\widehat{\boldsymbol{\beta}}_{\mathrm{PQL}}(t)) \to \mathbb{N}(\mathbf{0}_2, \boldsymbol{I}_2)$, where $\boldsymbol{V}_p^{\psi} = n^{-2}\sum_{i=1}^n\{y_i-\psi(\boldsymbol{N}_i^T\widehat{\boldsymbol{c}}_{\mathrm{PQL}})\}^2\boldsymbol{N}_i\boldsymbol{N}_i^T/p_i$.*

### 3.2 Optimal subsampling probabilities

For the functional coefficients, IMSE is a measure of the quality of their estimators. We want to find the optimal subsampling probabilities that minimize the IMSE of $\widetilde{\boldsymbol{\beta}}_{\mathrm{PQL}}$ in approximating $\widehat{\boldsymbol{\beta}}_{\mathrm{PQL}}$, where the IMSE is defined as follows

$$\mathrm{IMSE}(\widetilde{\boldsymbol{\beta}}_{\mathrm{PQL}}-\widehat{\boldsymbol{\beta}}_{\mathrm{PQL}}) = \int_a^b \boldsymbol{N}^T(t)(\boldsymbol{H}_{k,n}^{\psi})^{-1}\boldsymbol{V}_p^{\psi}(\boldsymbol{H}_{k,n}^{\psi})^{-1}\boldsymbol{N}(t)/L. \tag{4}$$

In (4), $L^{-1}(\boldsymbol{H}_{k,n}^{\psi})^{-1}\boldsymbol{V}_p^{\psi}(\boldsymbol{H}_{k,n}^{\psi})^{-1}$ is the asymptotic covariance matrix of $\widetilde{\boldsymbol{c}}_{\mathrm{PQL}}-\widehat{\boldsymbol{c}}_{\mathrm{PQL}}$, where $\boldsymbol{H}_{k,n}^{\psi}$ depends on the chosen smoothing parameter $\lambda$. In addition, from (4), we can see that only $\boldsymbol{V}_p^{\psi}$ depends on the sampling probability $p_i$ and the integral $\int_a^b \boldsymbol{N}^T(t)(\boldsymbol{H}_{k,n}^{\psi})^{-1}\boldsymbol{V}_p^{\psi}(\boldsymbol{H}_{k,n}^{\psi})^{-1}\boldsymbol{N}(t) \le \int_a^b \boldsymbol{N}^T(t)(\boldsymbol{H}_{k,n}^{\psi})^{-1}\boldsymbol{V}_{p^*}^{\psi}(\boldsymbol{H}_{k,n}^{\psi})^{-1}\boldsymbol{N}(t)$ if $\boldsymbol{V}_p^{\psi} \le \boldsymbol{V}_{p^*}^{\psi}$. We propose to obtain the optimal subsampling probability by minimizing $\boldsymbol{V}_p^{\psi}$. Several criteria exist for minimizing the matrix. Here we choose to minimize the trace of the matrix $\boldsymbol{V}_p^{\psi}$. Note that $L^{-1}\boldsymbol{V}_p^{\psi}$ is the asymptotic covariance matrix of $(\boldsymbol{H}_{k,n}^{\psi})^{-1}(\widetilde{\boldsymbol{c}}_{\mathrm{PQL}}-\widehat{\boldsymbol{c}}_{\mathrm{PQL}})$, where $(\boldsymbol{H}_{k,n}^{\psi})^{-1}(\widetilde{\boldsymbol{c}}_{\mathrm{PQL}}-\widehat{\boldsymbol{c}}_{\mathrm{PQL}})$ is a linear transformation of the estimator $\widetilde{\boldsymbol{c}}_{\mathrm{PQL}}-\widehat{\boldsymbol{c}}_{\mathrm{PQL}}$. Thus, minimizing $\mathrm{tr}(\boldsymbol{V}_p^{\psi})$ to obtain the optimal subsampling probability is termed the functional L-optimality criterion, which is the functional version of the L-optimality defined in Pukelsheim (2006) and Atkinson et al. (2007).

**Theorem 3** *If the subsampling probabilities $p_i, i = 1, \ldots, n$, are chosen as*

$$p_{\mathrm{PQL},i}^{\mathrm{FLoS}} = \frac{|y_i-\psi(\boldsymbol{N}_i^T\widehat{\boldsymbol{c}}_{\mathrm{PQL}})|\|\boldsymbol{N}_i\|_2}{\sum_{i=1}^n|y_i-\psi(\boldsymbol{N}_i^T\widehat{\boldsymbol{c}}_{\mathrm{PQL}})|\|\boldsymbol{N}_i\|_2}, \tag{5}$$

*then $\mathrm{tr}(\boldsymbol{V}_p^{\psi})$ attains its minimum, where $\|\boldsymbol{N}_i\|_2 = (\boldsymbol{N}_i^T\boldsymbol{N}_i)^{1/2}$.*

The subsampling probabilities (5) are related with the predictors and the response. Suppose the response $y_i \in \{0, 1\}, i = 1, \ldots, n$, we study the effect of the response on the subsampling probabilities. For the individuals with response $y_i = 1$, a smaller estimated probability $\psi(\boldsymbol{N}_i^T\widehat{\boldsymbol{c}}_{\mathrm{PQL}})$ using full data results in a larger subsampling probability $p_{\mathrm{PQL},i}^{\mathrm{FLoS}}$. On the other hand, for the samples with $y_i = 0$, the subsampling probability $p_{\mathrm{PQL},i}^{\mathrm{FLoS}}$ increases as the estimated probability $\psi(\boldsymbol{N}_i^T\widehat{\boldsymbol{c}}_{\mathrm{PQL}})$ increases. As a result, this subsampling method is more likely to select those samples that are more easily misclassified, which means this method improves the robustness of the subsample estimator.

Note that the corresponding computational complexity of $\widehat{\boldsymbol{c}}_{\mathrm{PQL}}$ in (5) using full data is $O(\mathcal{I}nK^2)$. Therefore, we need to replace $\widehat{\boldsymbol{c}}_{\mathrm{PQL}}$ by a pilot estimator, say $\widehat{\boldsymbol{c}}_{\mathrm{PQL}}^0$, which can be obtained by a uniform subsample with the sample size $L$. In addition, we need to

choose the smoothing parameter $\lambda$, the degree $p$ of the B-spline basis, and the number of knots $K$. In the penalized spline method, the choice of $K$ is not crucial (Cardot et al., 2003), as the roughness of the estimator is controlled by a roughness penalty, rather than the number of knots. Usually, in practice, we choose $p = 3$ and $K$ is chosen to be relatively large so that local features of $\beta(t)$ can be captured. Once $K$ and $p$ are fixed, we can select the smoothing parameter $\lambda$ by minimizing BIC. Using the full data to select the optimal $\lambda$ is computationally expensive. Therefore, we need to select the tuning parameter by BIC using the optimal subsample data. We give the practical subsampling procedure for the functional generalized linear model in Algorithm 1.

**Algorithm 1:** FLoS Algorithm for Functional Generalized Linear Model

**Input** : Data: $(\boldsymbol{N}_i, y_i; i = 1, \ldots, n)$, where $\boldsymbol{N}_i = (1, \int_a^b z_i(t)\boldsymbol{N}_{p+1}^T(t)dt)^T$.

- **Step 1:** Draw a subsample of size $L$ using the uniform sampling probabilities $p_i^0 = 1/n$, and use it to obtain the pilot estimator $\widehat{\boldsymbol{c}}_{\mathrm{PQL}}^0$ with $\lambda = 0$.
- **Step 2:** Applying $\widehat{\boldsymbol{c}}_{\mathrm{PQL}}^0$, we can get the approximate optimal subsampling probabilities $p_{\mathrm{PQL},i}^{\mathrm{FLoS},\widehat{\boldsymbol{c}}^0}$:

$$p_{\mathrm{PQL},i}^{\mathrm{FLoS},\widehat{\boldsymbol{c}}_{\mathrm{PQL}}^0} = \frac{|y_i - \psi(\boldsymbol{N}_i^T\widehat{\boldsymbol{c}}_{\mathrm{PQL}}^0)|\|\boldsymbol{N}_i\|_2}{\sum_{i=1}^n |y_i - \psi(\boldsymbol{N}_i^T\widehat{\boldsymbol{c}}_{\mathrm{PQL}}^0)|\|\boldsymbol{N}_i\|_2},$$

  Using the subsampling probabilities $p_{\mathrm{PQL},i}^{\mathrm{FLoS},\widehat{\boldsymbol{c}}_{\mathrm{PQL}}^0}$ to draw a random subsample with replacement of size $L$. Denote the subsample as $(\boldsymbol{N}_i^*, y_i^*)$, with associated subsampling probabilities $p_{\mathrm{PQL},i}^{*\mathrm{FLoS},\widehat{\boldsymbol{c}}_{\mathrm{PQL}}^0}$.
- **Step 3:** Given $\lambda$, we can obtain the estimate $\breve{\boldsymbol{c}}_{\mathrm{FLoS}}^{\mathrm{PQL}}(\lambda)$ through solving

$$Q_{\mathrm{PQL}}^{*\mathrm{FLoS}}(\boldsymbol{c}) = \sum_{i=1}^{L}(y_i^* - \psi(\boldsymbol{N}_i^{*T}\boldsymbol{c}))\boldsymbol{N}_i/(Lp_{\mathrm{PQL},i}^{*\mathrm{FLoS},\widehat{\boldsymbol{c}}_{\mathrm{PQL}}^0}) - \lambda\boldsymbol{D}\boldsymbol{c} = 0, \tag{6}$$

  thus, based on the optimal subsample data, we can use BIC to choose the optimal tuning parameter $\lambda$.

**Output** : Once we obtain the optimal $\lambda$, we can get the final estimator $\breve{\boldsymbol{\beta}}_{\mathrm{PQL}}^{\mathrm{FLoS}}(t) = \boldsymbol{N}^T(t)\breve{\boldsymbol{c}}_{\mathrm{FLoS}}^{\mathrm{PQL}}$.

In Algorithm 1, we need to use an iterative procedure, such as Newton's method introduced in Appendix C, to get the pilot estimator and solve Equation (6). In step 1 & 3, it takes $O(nK)$ computing complexity to calculate the matrix $\boldsymbol{N}$ and the subsampling probabilities. To get the pilot estimator $\widehat{\boldsymbol{c}}_{\mathrm{PQL}}^0$ in step 2, the computing complexity is $O(\mathcal{I}_0LK^2)$ where $\mathcal{I}_0$ is the number of iterations. In step 4, each iteration takes $O(LK^2)$

computing complexity and the whole procedure requires $O(\mathcal{I}_4 LK^2)$ with the number of iterations $\mathcal{I}_4$. Thus, when the full data size $n$ is very large, total computing complexity $O(nK + \mathcal{I}_0 LK^2 + \mathcal{I}_4 LK^2) \approx O(nK)$ is smaller than the total computing complexity based on full data $O(nK + \mathcal{I}nK^2) \approx O(\mathcal{I}nK^2)$. Thus, Algorithm 1 can reduce computing complexity dramatically. Algorithm 1 is also naturally suited for distributed storage and parallel computing. We can divide the full data into several subsets, simultaneously compute the $\boldsymbol{N}_i$ and optimal subsample probabilities on each subset. Combining the optimal subsample probabilities of each subset, we can get the indices of a random subsample in the full data and use these indices to extract the corresponding data on each subset. The asymptotic result of the estimator obtained from Algorithm 1 is presented Theorem 4.

**Theorem 4** *Under Assumptions 1-7, for any given $t$, as $L \to \infty$ and $n \to \infty$, conditionally on $\mathcal{F}_n$ in probability, $\left\{\boldsymbol{N}^T(t)(\boldsymbol{H}^{\psi}_{k,n})^{-1}\boldsymbol{V}^{\psi}_{\mathrm{FLoS}}(\boldsymbol{H}^{\psi}_{k,n})^{-1}\boldsymbol{N}(t)/L\right\}^{-1/2}(\breve{\boldsymbol{\beta}}^{\mathrm{FLoS}}_{\mathrm{PQL}}(t) - \widehat{\boldsymbol{\beta}}_{\mathrm{PQL}}(t)) \to \mathbb{N}(\boldsymbol{0}_2, \boldsymbol{I}_2)$, in distribution, where $\boldsymbol{V}^{\psi}_{\mathrm{FLoS}}$ has the minimum trace, and it has the explicit expression $\boldsymbol{V}^{\psi}_{\mathrm{FLoS}} = n^{-1}\sum_{i=1}^{n}|y_i - \psi(\boldsymbol{N}_i^T\widehat{\boldsymbol{c}}_{\mathrm{PQL}})|\boldsymbol{N}_i\boldsymbol{N}_i^T/\|\boldsymbol{N}_i\|_2 \times n^{-1}\sum_{i=1}^{n}|y_i - \psi(\boldsymbol{N}_i^T\widehat{\boldsymbol{c}}_{\mathrm{PQL}})|\|\boldsymbol{N}_i\|_2$.*

### 3.3 Extension

Our proposed method can be extended to the following functional generalized linear model with multiple functional predictors by several modifications:

$$E(Y|X) = \psi\left(\alpha + \sum_{m=1}^{M}\int_a^b Z_m(t)\beta_m(t)dt\right),$$

where $Y$ is the scalar response, $Z_m(t), m = 1, ..., M$ is functional predictor defined on domain $[a, b]$. Then, the penalized quasi-likelihood estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{PQL}}(t) = (\widehat{\alpha}_{\mathrm{PQL}}, \widehat{\beta}_{1,\mathrm{PQL}}(t), \cdots, \widehat{\beta}_{M,\mathrm{PQL}}(t))^T$ and $\widehat{\beta}_{m,\mathrm{PQL}}(t) = \boldsymbol{N}_{p+1}^T(t)\widehat{\boldsymbol{c}}_{m,\mathrm{PQL}}$, where $\widehat{\boldsymbol{c}}_{\mathrm{PQL}} = (\widehat{\alpha}_{\mathrm{PQL}}, \boldsymbol{c}_{1,\mathrm{PQL}}^T, \cdots, \boldsymbol{c}_{M,\mathrm{PQL}}^T)^T$ is the solution of the following equation:

$$Q_{\mathrm{PQL}}(\boldsymbol{c}) = \sum_{i=1}^{n}\left\{y_i - \psi\left(\boldsymbol{N}_i^T\boldsymbol{c}\right)\right\}\boldsymbol{N}_i - \sum_{m=1}^{M}\lambda_m\boldsymbol{D}_q\boldsymbol{c}_m = 0, \tag{7}$$

with $\boldsymbol{N}_i = \left(\alpha, \int_a^b z_{i1}(t)\boldsymbol{N}_{p+1}^T(t)dt, \cdots, \int_a^b z_{iM}(t)\boldsymbol{N}_{p+1}^T(t)dt\right)^T$. Let $\boldsymbol{Z}_m(t) = (z_{1m}(t), \cdots, z_{nm}(t))^T$, for each predictor $z_m(t)$, we compute a matrix $\boldsymbol{N}_m = \int_a^b \boldsymbol{Z}_m(t)\boldsymbol{N}_{p+1}^T(t)dt$. Denote $\boldsymbol{N} = (\boldsymbol{1}_n, \boldsymbol{N}_1, \cdots, \boldsymbol{N}_M)$ be the column catenation of the $n \times 1$ vector of 1s $\boldsymbol{1}_n$, $\boldsymbol{N}_1, \cdots, \boldsymbol{N}_M$ and corresponding set $\boldsymbol{D} = \mathrm{diag}(0, \boldsymbol{D}_q, \cdots, \boldsymbol{D}_q)$, where $\boldsymbol{D}$ is the matrix with 0 and $M$ blocks $\boldsymbol{D}_q$ in its main diagonal and zeros elsewhere. After replacing $\boldsymbol{N}$ and $\boldsymbol{D}_q$ by new defined $\boldsymbol{N}$ and $\boldsymbol{D}$, respectively, the estimations and algorithms described in Section 2 and Section 3.1 can be carried out to estimate $\beta_1(t), \ldots, \beta_M(t)$ simultaneously. It is worth mentioning that we can simultaneously compute all matrix $\boldsymbol{N}_1, \ldots, \boldsymbol{N}_M$.

## 4. Simulation Studies

We evaluate the performance of the proposed subsample estimators in terms of both estimations efficiency and computational efficiency in this section.

### 4.1 Simulation I

In this section, we evaluate the finite sample performance of the functional L-optimality subsampling method described in Algorithm 1 for estimating the functional logistic regression in comparison with the uniform subsampling method. We set the true functional coefficient $\beta(t) = 8 \times \sin(0.85\pi t)$. Denote the inverse logistic function as $\psi(\cdot) = \exp(\cdot)/(1+\exp(\cdot))$ and $p(x_i) = \psi(\int_0^1 x_i(t)\beta(t)dt)$. We generated responses $y(x_i) \sim$ Binomial$(1, p(x_i))$ as pseudo-Bernoulli random variables with probability $p(x_i)$. The functional predictor $x_i(t)$ is generated by $x_i(t) = \sum a_{ij}B_j(t)$, where $B_j(t)$ are cubic B-spline basis functions defined on 66 equally spaced knots in $[0, 1]$. We consider the following four different scenarios to generate the basis coefficients $a_{ij}$:

- **Scenario I.** The coefficients $a_{ij}$ are i.i.d from $N(0, 6)$. Figure 9 (a) in Appendix B shows that in the simulated data set under this scenario, the distribution of the probability $p(x_i)$ is symmetric about 0.5 and the number of 1's and the number of 0's in the responses are roughly equal.

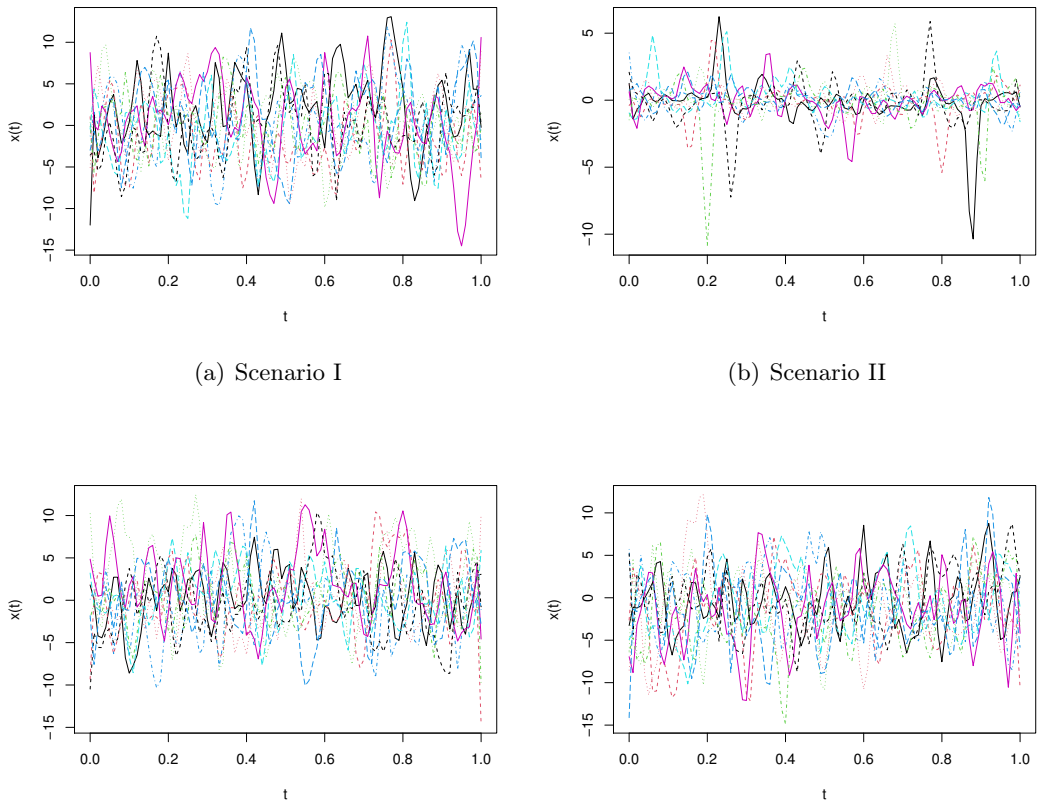
- **Scenario II.** We generate the coefficients $a_{ij}$ from the $t$ distribution with 2 degrees of freedom and zero mean, namely, $a_{ij} \overset{iid}{\sim} t_2$. For this scenario, Figure 9 (b) in Appendix B shows that the probability $p(x_i)$ is symmetric about 0.5 and is less uniform than those $p(x_i)$ of Scenario I. Similar to Scenario I, in the simulated data set under Scenario II, the number of 1's and the number of 0's in the responses are roughly equal.

- **Scenario III.** Similar with the setting in Wang et al. (2018), the coefficients $a_{ij}$ are iid from $N(0.3, 6)$. In this scenario, the distribution of probability $p(x_i)$ is skewed left and about 63.53% of responses are 1, which is shown in Figure 9 (c) in Appendix B. This data set illustrates imbalanced data.

- **Scenario IV.** The coefficients $a_{ij}$ are iid from $N(-0.8, 6)$. The data set generated under this scenario is an example of rare events data with about 17.85% of responses as 1, which is similar to the rare event data used in Wang et al. (2018). Figure 9 (d) in Appendix B shows that the distribution of probability $p(x_i)$ is skewed right.

Figure 1 displays a random subset of 10 curves for the functional predictor $x_i(t)$ under four scenarios when the sample size $n = 10^5$. It shows that the variation among the functional predictor $x_i(t)$ is the smallest when $a_{ij}$ is generated from a normal distribution, while the variation is the largest when $a_{ij}$ is generated from a $t$ distribution (Scenario II). It means that the data generated under Scenarios I, III, and IV is more uniform.

To evaluate the computational efficiency of the subsampling strategies, we record the CPU times (in seconds) of the two subsampling strategies and using the full data. This




(a) Scenario I

(b) Scenario II

(c) Scenario III

(d) Scenario IV

Figure 1: Ten examples of the simulated functional predictor $x_i(t)$ under four scenarios in Simulation I when the full sample size is $n = 10^5$.

paper uses the R programming language (enhanced R distribution Microsoft R 4.0.2) to implement each method. All computations are carried out on a computation platform with Intel Xeon 5 Cpu with 4 cores and 8G memory. Based on 300 replications, Table 1 displays the computation time for different combinations of the full data size $n$ and the subsample size $L$ under Scenario I. The results under the other three scenarios are similar and thus omitted. Table 1 shows that the functional L-optimality subsampling method is significantly faster than using the full data. The difference between the functional L-optimality subsampling method and the uniform subsampling method is small. In the implementation, we make the number of knots $K = \lceil 1.25 \times n^{1/4} \rceil$ according to Assumption 5, where $\lceil a \rceil$ means the least integer greater than or equal to $a$. When the full data size $n = 5 \times 10^6$, the size of the basis matrix of $\boldsymbol{N}_i$ is about 2.6GB and the computing time for using full data exceeds 18 minutes. Moreover, the basis matrix needs about 8.3GB memory under the full data size $n = 10^7$, which goes beyond the maximum memory of a general PC with 8G memory, so the estimation using the full data is not feasible. In this case, for the functional L-optimality subsampling method and the uniform subsampling method, we can take advantage of parallel computing to calculate the basis matrix $\boldsymbol{N}$ and the subsampling probability $p_{\mathrm{PQL},i}^{\mathrm{FLoS},\tilde{\boldsymbol{c}}^0}$. We then use the optimal subsampling data to estimate the functional generalized linear model.

Table 1: The computing time (in seconds) for estimating the functional generalized linear model in Simulation I using the functional L-optimality subsampling (FLoS) method and the uniform sampling (UNIS) method when the full data size $n = 10^5$, $10^6$, $5 \times 10^6$ and $n = 10^7$. When the full data size $n = 10^7$, the estimation is beyond the computing memory and fails when using the full data.

| $n$ | $L$ | FLoS | UNIS | FULL |
|---|---|---|---|---|
| $n = 10^5$ | 300 | 0.036 | 0.017 | 4.253 |
| | 3000 | 0.143 | 0.118 | |
| | 5000 | 0.225 | 0.194 | |
| $n = 10^6$ | 500 | 0.224 | 0.052 | 86.878 |
| | 5000 | 0.544 | 0.333 | |
| | 10000 | 0.862 | 0.630 | |
| $n = 5 \times 10^6$ | 1000 | 1.348 | 0.184 | 1107.457 |
| | 8000 | 2.390 | 1.161 | |
| | 20000 | 4.085 | 2.854 | |
| $n = 10^7$ | 3000 | 6.571 | 0.543 | **Fail** |
| | 20000 | 13.062 | 3.281 | |
| | 50000 | 24.163 | 7.835 | |

The simulation is repeated 300 times. The performances of the functional L-optimality subsampling method and the uniform subsampling method are measured by the root

IMSEs (RIMSEs) for the estimators:

$$\text{RIMSE} = \sqrt{\int (\breve{\beta}(t) - \beta(t))^2 dt}.$$

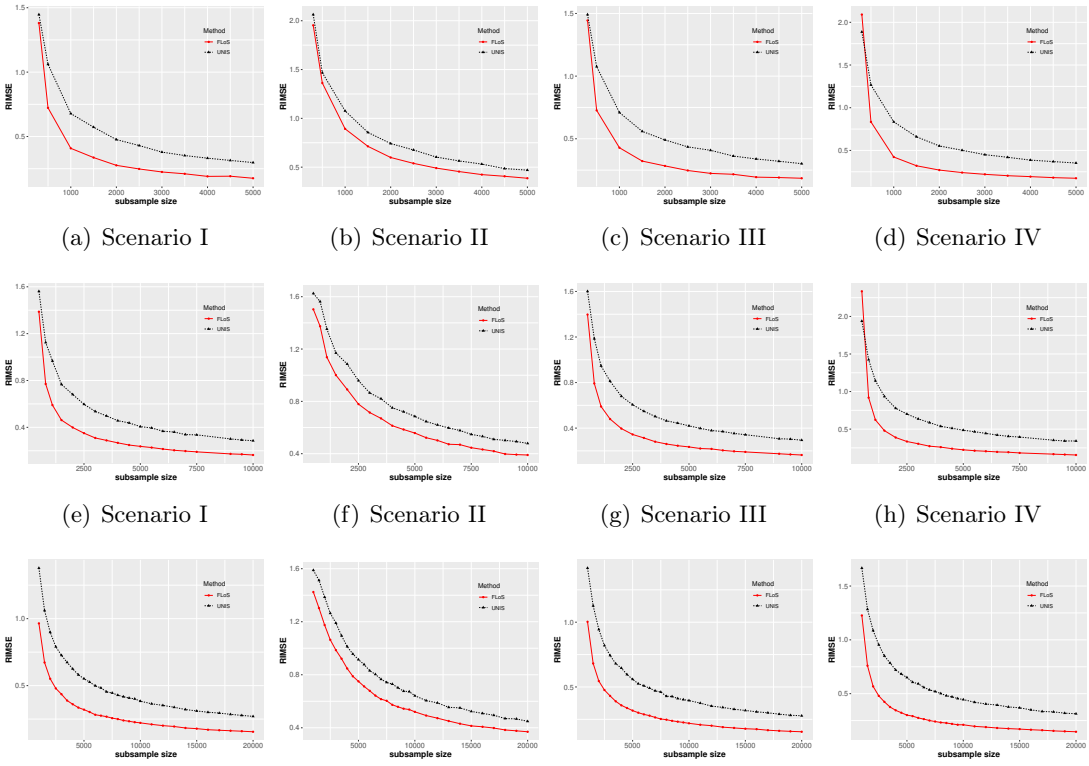
The first row of Figure 2 displays the mean of the RIMSEs against the subsample size $L = 300, 500, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000$ for full data size $n = 10^5$. When the full data size $n = 10^6$ and the subsample size $L = 500, 800, 1100, 1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000, 5500, 6000, 6500, 7000, 7500, 8000, 8500, 9000, 9500, 10000$, the figures of the RIMSEs for two subsample methods against the subsample size are resented in the second row of Figure 2. For full data size is $n = 5 \times 10^6$ and subsample size $L = 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000, 5500, 6000, 6500, 7000, 7500, 8000, 8500, 9000, 9500, 10000, 11000, 12000, 13000, 14000, 15000, 16000, 17000, 18000, 19000, 20000$, the bottom row of Figure 2 shows the RIMSEs against subsample size $L$.

Figure 2 shows that the functional L-optimality subsampling method outperforms the uniform subsampling approach for all scenarios under different full data sizes. The RIMSEs for both subsampling methods decrease and tend to stay stable as the subsample size increases. When the full data size is fixed, the more imbalanced the data, the greater the advantage of the functional L-optimality subsampling method over the uniform subsampling approach. Figure 3 shows that our method can still outperform the uniform subsampling approach when the proportion of 1's in the responses reaches around 5.88% ($a_{ij} \sim N(-2.5, 6)$) or even around 1.85% of the full data set with $n = 10^5$ ($a_{ij} \sim N(-3.5, 6)$). On the other hand, when the data is extremely rare data (for example, about 0.05% of 1's in the responses with $n = 10^5$, that is, $a_{ij} \sim N(-4.5, 6)$), neither subsampling methods nor the method using the full data work well. In Scenario II when the variation among functional predictors is larger, Figure 2 (b), (f) and (j) show that the functional L-optimality subsampling method also dominates the uniform subsampling approach.

To compare the performance of the two subsampling methods on the classification accuracy, Figure 4 displays proportions of correct classifications (PCC), which is defined as:

$$\text{PCC} = \frac{\#\{y_i = 1 \quad \text{and} \quad \psi(\boldsymbol{N}_i^T \widehat{\boldsymbol{c}}_{\text{PQL}}) > 0.5\} + \#\{y_i = 0 \quad \text{and} \quad \psi(\boldsymbol{N}_i^T \widehat{\boldsymbol{c}}_{\text{PQL}}) \leq 0.5\}}{n}. \tag{8}$$

Based on the PCC criterion, Figure 4 also shows that the PCC for the two methods increases and will stay stable as the subsample size $L$ increases. In addition, the functional L-optimality subsampling method performs better than the uniform subsampling approach in all four scenarios. Although the two methods do not perform very well for Scenario II, the functional L-optimality subsampling method is still significantly better than the uniform subsampling approach. In summary, regardless of whether the variation among the generated functional predictors is large or the responses are imbalanced, our proposed functional L-optimality subsampling method is better than the uniform subsampling approach.

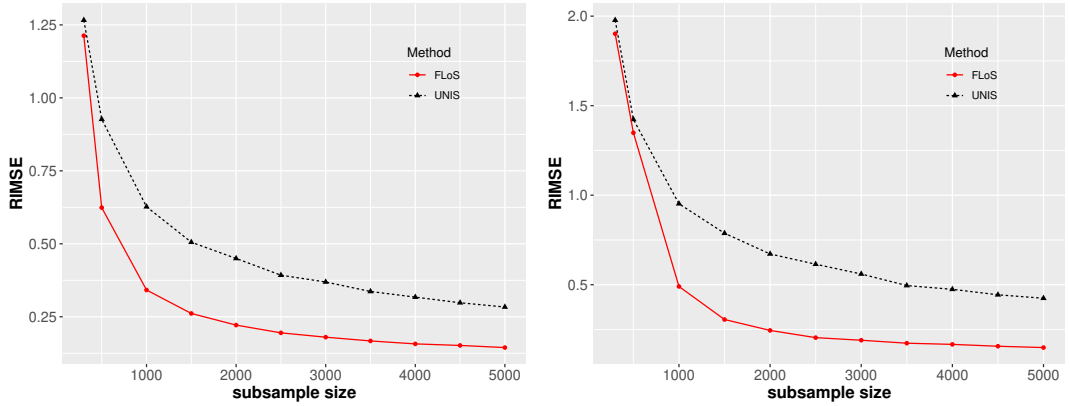



(a) Scenario I (b) Scenario II (c) Scenario III (d) Scenario IV

(e) Scenario I (f) Scenario II (g) Scenario III (h) Scenario IV

(i) Scenario I (j) Scenario II (k) Scenario III (l) Scenario IV

Figure 2: The average of the root integrated mean squared error (RIMSE) of the estimated functional coefficient $\breve{\beta}(t)$ in the functional logistic regression model (Simulation II) by using the functional L-optimality subsampling (FLoS) method and the uniform subsampling (UNIS) approach under four scenarios with various subsample sizes $L$ when the full data size $n = 10^5$ (Panels (a)-(d)), $10^6$ (Panels (e)-(h)), and $5 \times 10^6$ (Panels (i)-(l)).




(a) 5.88% of 1's in the responses

(b) 1.85% of 1's in the responses

Figure 3: The average of the root integrated mean squared error (RIMSE) of the estimated functional coefficient $\breve{\beta}(t)$ in the functional logistic regression model (Simulation II) by using the functional L-optimality subsampling (FLoS) method and the uniform subsampling (UNIS) approach from the rare event data with 5.88% or 1.85% of 1's in the response when the full data size $n = 10^5$.

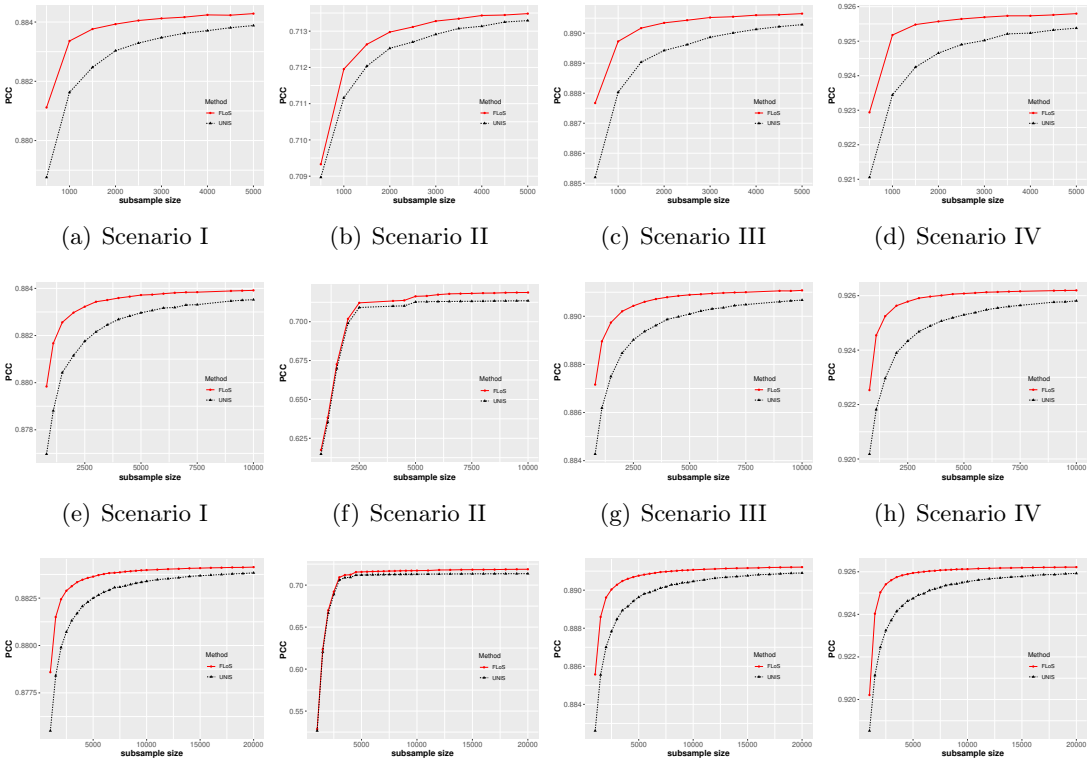
### 4.2 Simulation II

In this section, we evaluate the finite sample performance of the proposed subsampling method described in Algorithm 1 for estimating the functional Poisson regression in comparison with the uniform subsampling approach. We set the true functional coefficient $\beta(t) = \sin(0.5\pi t)$. Denote $\psi(\cdot) = \exp(\cdot)$ and $\lambda(x_i) = \psi(\int_0^1 x_i(t)\beta(t)dt)$, then we generated responses $y(x_i) \sim \text{Poisson}(\lambda(x_i))$ with the mean $\lambda(x_i)$. The simulation designs of the functional predictors $x_i(t)$ are the same as in Simulation I, except that we consider the following three different scenarios to generate the basis coefficients $a_{ij}$.

- **Scenario I.** The basis coefficients $a_{ij}$ are i.i.d from the standard normal distribution, namely, $a_{ij} \sim N(0, 1)$. Figures 10 (a) and (d) in Appendix B show that the distribution of the expected value $\lambda(x_i)$ ranges from 0.6 to 1.5 and is approximately symmetric about 1. About 70% of the responses are equal to 0 or 1.

- **Scenario II.** We generate the basis coefficients $a_{ij}$ from the $t$ distribution with 4 degrees of freedom and the variance is 1, namely, $a_{ij} \overset{iid}{\sim} t_4(0.5, 1)$. Figures 10 (b) and (e) in Appendix B shows that the $\lambda(x_i)$ varies from 0.6 to 1.5, and about 80% of responses lie between 0-2.

- **Scenario III.** We generate the basis coefficients $a_{ij}$ from the uniform distribution between 0 and 4, namely, $a_{ij} \overset{iid}{\sim} U(0, 4)$. Figures 10 (c) and (f) in Appendix B show that the expected value $\lambda(x_i)$ ranges from 2 to 6 and the distribution of responses is more uniform than in Scenario I and II.

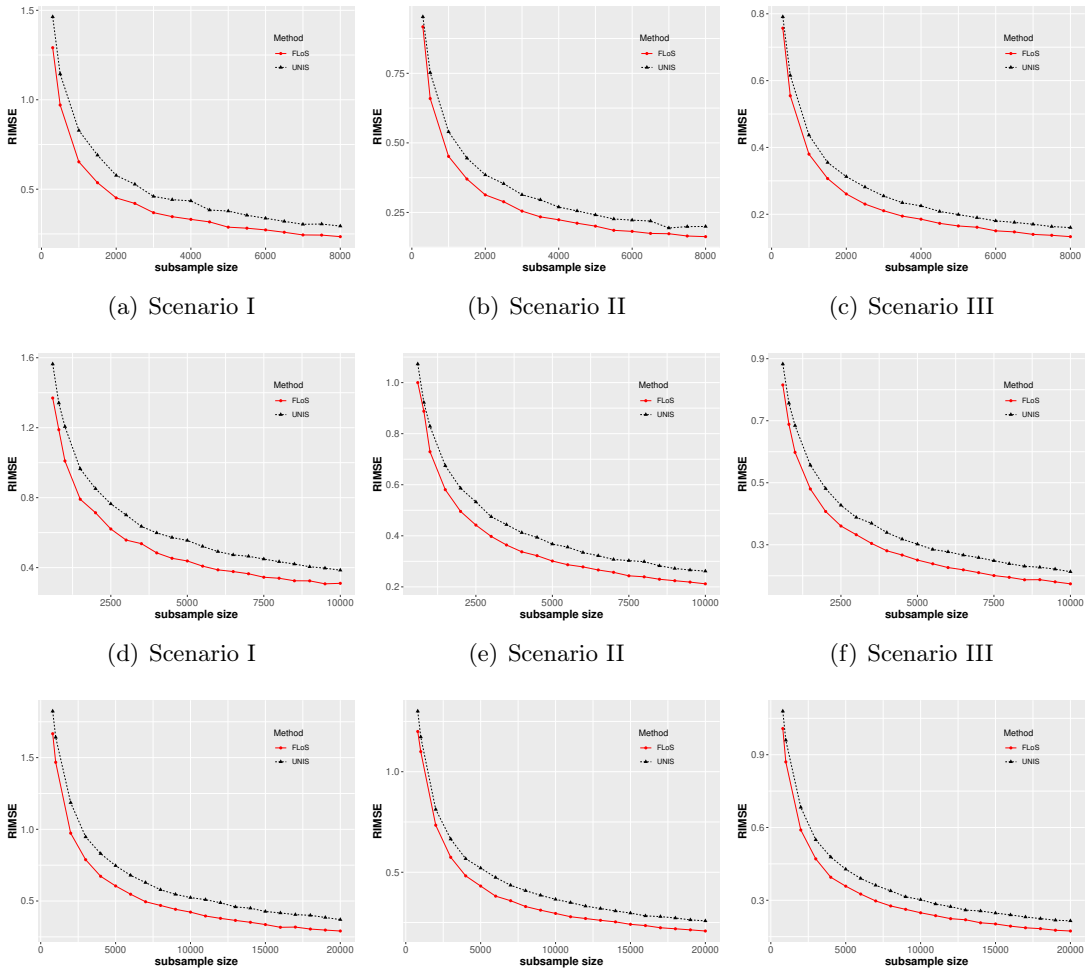



(a) Scenario I (b) Scenario II (c) Scenario III (d) Scenario IV

(e) Scenario I (f) Scenario II (g) Scenario III (h) Scenario IV

(i) Scenario I (j) Scenario II (k) Scenario III (l) Scenario IV

Figure 4: The average of the proportions of correct classifications (PCC) defined in (8) in the functional logistic regression model (Simulation II) by using the functional L-optimality subsampling (FLoS) method and the uniform subsampling (UNIS) approach under four scenarios with various subsample sizes $L$ when the full data size $n = 10^5$ (Panels (a)-(d)), $10^6$ (Panels (e)-(h)), and $5 \times 10^6$ (Panels (i)-(l)).




(a) Scenario I (b) Scenario II (c) Scenario III

(d) Scenario I (e) Scenario II (f) Scenario III

(g) Scenario I (h) Scenario II (i) Scenario III

Figure 5: The average of the root integrated mean squared error (RIMSE) of the estimated functional coefficient $\breve{\beta}(t)$ in the functional Poisson regression model by using the functional L-optimality subsampling (FLoS) method and the uniform subsampling (UNIS) approach under three scenarios with various subsample sizes $L$ when the full data size $n = 10^5$ (Panels (a)-(c)), $10^6$ (Panels (d)-(f)), and $5 \times 10^6$ (Panels (g)-(i)).

Same as in Simulation I, we choose $K = \lceil 1.25 * n^{1/4} \rceil$ in this section. Based on 300 replications, Figure 5 displays the mean of RIMSEs of the estimated functional coefficient $\breve{\beta}(t)$ in the functional Poisson regression model when using the functional L-optimality subsampling method and the uniform subsampling approach under three scenarios when the full data size $n = 10^5$, $10^6$, and $5 \times 10^6$. For full data size $n = 10^5$, we let subsample size $L = 300, 500, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000, 5500, 6000, 6500, 7000, 7500, 8000$. The subsample size $L = 600, 800, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000, 5500, 6000, 6500, 7000, 7500, 8000, 8500, 9000, 9500, 10000$ when the full data size is $n = 10^6$. And the subsample size $L = 800, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 11000, 12000, 13000, 14000, 15000, 16000, 17000, 18000, 19000, 20000$ for full data size $n = 5 \times 10^6$. Figure 5 shows that the functional L-optimality subsampling method outperforms the uniform subsampling approach for all three scenarios with different full data sizes. These numerical results are consistent with our theoretical results that the functional L-optimality subsampling method aims to minimize the IMSE of $\breve{\beta}(t)$ in approximating the estimator using the full data. Besides, when the full data size is fixed, the RIMSEs of $\breve{\beta}(t)$ using both methods become smaller and tend to stay stable as the subsample size $L$ increases.

## 5. Kidney Transplant Data

The kidneys are a pair of organs in the human body, whose primary function is to remove waste from the body through the production of urine and to regulate the chemical (electrolyte) composition of the blood. Renal failure means that the kidneys can no longer remove wastes and maintain electrolyte balance, which will threaten a human's life. Renal failure can be divided into acute renal failure and chronic renal failure. Regarding the treatment of chronic renal failure, one method is a kidney transplant. A successful kidney transplant can restore normal renal function to the patients and extend their survival time. After kidney transplantation, kidney transplant recipients still face a high probability of losing transplant function. It is also important to follow up the graft function and predict the patient's expected lifespan after a kidney transplant.

Creatinine is the waste product of creatine, which the muscles use to make energy. Typically, creatinine travels from the blood to the kidneys where it leaves the body in the urine. A high level of creatinine in the blood indicates that the kidney is not working correctly. On the other hand, only looking at how much creatinine is in the blood is not the best way to check how well the kidneys are working, because the level of creatinine in blood is related to age, race, gender, and body size. In other words, what's considered "normal" depends on these factors. The best way to know if kidneys are working properly is by looking at the glomerular filtration rate (GFR), which considers the creatinine level and the associated factors simultaneously (Keong et al., 2016; Dong et al., 2018, 2019, 2021; Shi et al., 2021; Dong et al., 2023). For adults (Age$\geq$ 19), we use the Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI, Levey et al. (2009)) equation to obtain the estimated glomerular filtration rate (eGFR, mL/min/1.73m$^2$). For children

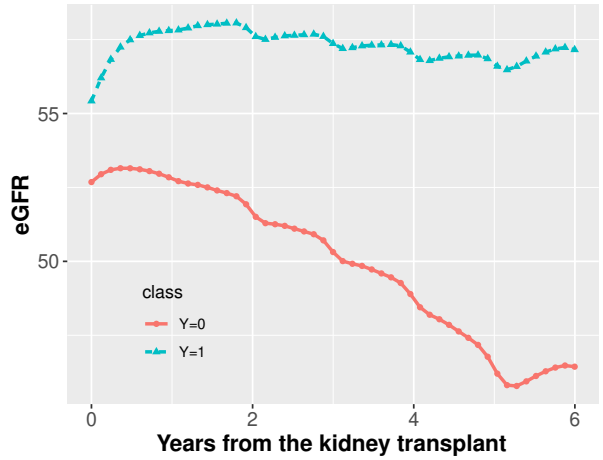



Figure 6: The mean eGFR curves for the group of recipients who die or need to be re-transplanted during the sixth to tenth year after the transplant ($Y = 0$) and the group of recipients who have lived for at least ten years after transplant ($Y = 1$).

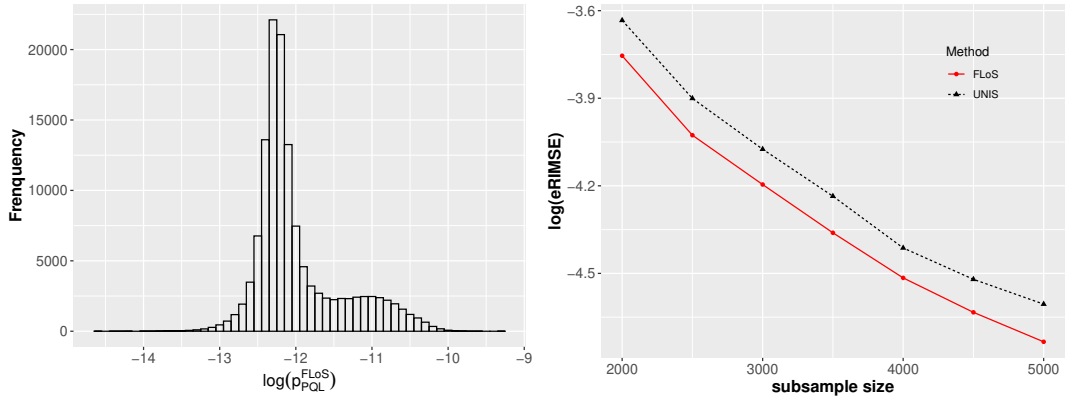
(Age$\leq$ 18), we use the Schwartz formula (Schwartz et al., 2009) to estimate the glomerular filtration rate.

Our objective is to predict whether kidney transplant recipients can survive over ten years based on their eGFR trajectories in the first six years after kidney transplant. The data resource used in this section is the kidney transplant data from the Optn/UNOS. After matching data and deleting missing data, there are $n = 130313$ recipients who have lived for at least six years after kidney transplant. We divide these recipients into two categories: the first category is the 30590 (23.3%) recipients who die or need to be re-transplanted during the sixth to tenth year after the transplant ($Y = 0$), and the other category is the 100713 (76.7%) recipients who have lived for at least ten years after transplant ($Y = 1$). Figure 6 display the mean eGFR trajectories for these two categories. It shows that the mean eGFR curve of $Y = 1$ is higher than that of $Y = 0$, which is consistent with the fact that a higher eGFR means a better renal function. For those recipients who have not lived for ten years after transplant, the eGFR shows a significant downward trend. On the contrary, the eGFR curve remains stable for those recipients who have lived for ten years after transplant.

We consider fitting a functional logistic regression model:

$$E(Y_i|\text{eGFR}_i) = \psi\left(\alpha + \int_0^6 \text{eGFR}_i(t) \times \beta(t)dt\right). \tag{9}$$

Figure 7 (a) displays the histogram of the log of the subsampling probabilities $p_{\text{PQL}}^{\text{FLoS}}$ in the the functional L-optimality subsampling method. It shows that the subsampling probabilities for different samples are very different. Because we do not know the true functional coefficient, we adopt the empirical root integrated mean square error (eRIMSE)




(a) (b)

Figure 7: (a) Histogram of the log of the subsampling probabilities $p_{\mathrm{PQL}}^{\mathrm{FLoS}}$ in the functional L-optimality subsampling method. (b) The logarithm of the empirical integrated mean square error (eIMSE) defined in (10) for the estimated functional coefficient using the functional L-optimality subsampling (FLoS) method and the uniform subsampling (UNIS) approach with different subsample sizes.

as the criterion for comparing two subsampling methods, which is defined as

$$\mathrm{eRIMSE} = S^{-1} \sum_{s=1}^{S} \sqrt{\int (\breve{\beta}^{(s)}(t) - \widehat{\beta}(t))^2 dt}, \tag{10}$$

where $\breve{\beta}^{(s)}(t)$ is the estimated functional coefficient using the s-th subsample data set, and $\widehat{\beta}(t)$ is the estimator using the full data. Figure 7 (b) displays the logarithm of the empirical integrated mean square error (eIMSE) defined in (10) for the estimated functional coefficient using both subsampling methods. It indicates that the functional L-optimality subsampling method has smaller eIMSEs than the uniform subsampling approach for all subsample sizes.

Figure 8 displays the estimated functional coefficient for the functional logistic regression model (9) by using the full data and using $L = 5000$ data subsampled with the functional L-optimality subsampling method. The two estimated functional coefficients are almost identical. Figure 8 also provides the corresponding 95% point-wise confidence interval for the functional coefficient based on 1000 subsampling datasets with the subsample size $L = 5000$ by using the functional L-optimality subsampling method. It shows that only the functional coefficient is significantly non-zero only from the fourth year after the transplant. Therefore, the information on eGFR during the 4th to the 5.5th year is more helpful to predict whether a recipient can live beyond ten years.

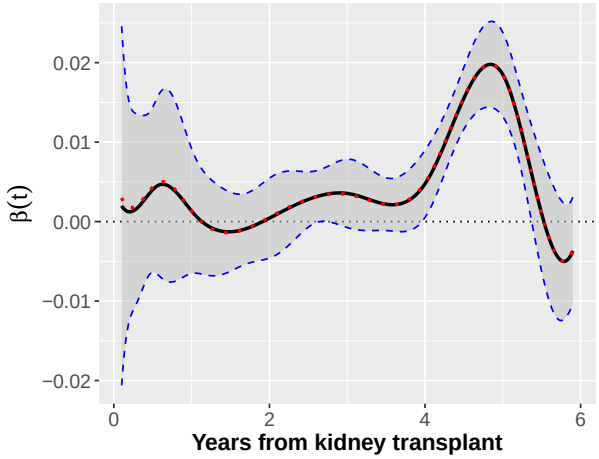



Figure 8: The red dotted line is the estimated $\beta(t)$ based on the full fata for predicting whether the recipient can live for at least 10 years after transplant based on the eGFR information in about the first six years. The black solid curve is the averaged estimated $\beta(t)$ using the FLoS method based on 1000 subsampling datasets with subsample size $L = 5000$. The blue dashed lines are 95% point-wise confidence limits on the curve based on 1000 subsampling datasets with the subsample size $L = 5000$.

## 6. Conclusions and Discussions

We propose the functional L-optimality subsampling method for estimating the functional generalized linear model to address the challenges brought when using extraordinary amounts of functional data. The asymptotic results of the subsample estimators have also been established. Several simulation studies show that our proposed method is computationally feasible and outperforms the uniform subsampling method for massive data. The proposed subsampling method is also demonstrated by analyzing the kidney transplant data. For the kidney transplant data, we find that the subsample estimators can well approximate the results obtained from the full data.

In this paper, we consider the subsampling for the scalar on function regressions. There are other functional regressions, such as function on scalar regressions (Zhu et al., 2012; Luo et al., 2016; Li et al., 2017; Cai et al., 2022) and function on function regressions (Sun et al., 2018; Cai et al., 2021a,b). For these two types of regressions, how to subsample is still an open problem. We will pursue these problems in our future research.

## Supplementary Materials

**R package**: An R package `SubsamplingFunPredictors` has been developed for implementing the proposed method. The R package and a demonstration are provided (SubsamplingFunPredictors_0.1.0.tar.gz, GNU zipped tar file) at `https://github.com/caojiguo/FLoS`.

**R code**: We provide the R codes at `https://github.com/caojiguo/FLoS`, which can be used to replicate the simulation studies included in the article.

## Acknowledgments

The authors sincerely thank the editor and three anonymous referees for their valuable comments, which lead to further improvement of this article. Dr Liu's research was supported by National Natural Science Foundation of China (NSFC) (No.12201487), the Project funded by China Postdoctoral Science Foundation (No.2022M722544), the Fundamental Research Funds for the Central Universities (SK2022044). Dr You's research was supported by the National Natural Science Foundation of China (NSFC) (No.11971291) and Innovative Research Team of Shanghai University of Finance and Economics. Dr Cao's research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery grant (RGPIN-2023-04057). The kidney transplant data set was supported in part by Health Resources and Services Administration contract 234-2005-370011C. The content about this data set is the responsibility of the authors alone and does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

## Appendix A. Assumptions

Before we present some assumptions used in the theorems, we first define some notations. If $0 < m < \infty$, $\mathcal{L}^m$ is defined as the space of functions $f(t)$ over the interval $[a, b]$ such that $\int_a^b |f(t)|^m dt < \infty$. With this convention, $\mathcal{L}^m$ is treated as a Banach space with the norm $\|f\|_m = (\int_a^b |f(t)|^m dt)^{1/m}$. When $m = 2$, we obtain the Hilbert space $\mathcal{L}^2$ with the inner product $\langle f, g\rangle = \int_a^b f(t)g(t)dt$ and the $\mathcal{L}_2$ norm $\|f\|_2$. And $\mathbb{R}^{m^*}$ is also a Hilbert space for a positive integer $m^*$. We also define $\langle \boldsymbol{u}, \boldsymbol{v}\rangle = \boldsymbol{u}^T\boldsymbol{v}$ as the inner product of vector $\boldsymbol{u}$ and $\boldsymbol{v}$.

**Assumption 1** *Let $\upsilon$ be a nonnegative integer, and $\kappa \in (0, 1]$ such that $d = \upsilon + \kappa \geq p+1$. We assume the unknown slope function $\beta(\cdot) \in \mathcal{H}^{(d)}([a, b])$, which is the class of function $f$ on $[a, b]$ whose $\upsilon$th deriative exists and satisfies a Lipschitz condition of order $\kappa$: $|f^{(\upsilon)}(t) - f^{(\upsilon)}(s)| \leq C_\upsilon |s - t|^\kappa$, for $s, t \in [a, b]$ and some constant $C_\upsilon > 0$.*

**Assumption 2** *For the functional predictor $Z(t)$, it holds that $\mathbb{E}(\|Z\|_4^4) < \infty$. In addition, the scalar response $Y$ satisfies that $\mathrm{E}(Y^4) < \infty$.*

**Assumption 3** *For the roughness penalty, we assume tuning parameter $\lambda$ satisfies that $\lambda = o(n^{1/2}K^{1/2-2q})$. Besides, we assume $q \leq p$.*

**Assumption 4** *Let $\delta_j = k_{j+1} - k_j$ and $\delta = \max_{0\leq j\leq K}(k_{j+1} - k_j)$. There exists a constant $M > 0$, such that*

$$\delta / \min_{0\leq j\leq K}(k_{j+1} - k_j) \leq M, \quad \max_{0\leq j\leq K-1} |\delta_{j+1} - \delta_j| = o(K^{-1}). \tag{11}$$

*In addition, let* $\boldsymbol{G}_{k,n}^{\psi} = \boldsymbol{N}^T \boldsymbol{\Psi} \boldsymbol{N}/n$ *and* $\boldsymbol{H}_{k,n}^{\psi} = (\boldsymbol{G}_{k,n}^{\psi} + \lambda/n\boldsymbol{D})$*. The smallest eigenvalue of* $\boldsymbol{G}_{k,n}^{\psi}$ *is greater than* $c_G/K$*, where* $c_G$ *is a positive constant.*

**Assumption 5** *The number of knots* $K = o(\sqrt{n})$ *and* $K = \omega(n^{1/(2d+1)})$*, where* $K = \omega(n^{1/(2d+1)})$ *means* $K/n^{1/(2d+1)} \to \infty$ *as* $n \to \infty$*.*

**Assumption 6** *We assume* $\max_{1 \le i \le n}(np_i)^{-1} = o_p(\sqrt{L})$ *and* $L = o(K^2)$*.*

**Assumption 7** *Let* $\dot{Q}_{\mathrm{PQL}}(\gamma, y)$ *be the first order derivative of* $Q_{\mathrm{PQL}}(\gamma, y)$ *with respect to* $\gamma$*. The function* $\dot{Q}_{\mathrm{PQL}}(\gamma, y) < 0$ *for* $\eta \in \mathbb{R}$ *and* $y$ *in the range of the response variable. The functions* $\psi(\cdot)$*, and the first order derivative of* $\psi(\cdot)$ *are continuous. There exist positive constants* $c_Q$ *and* $C_Q$ *such that* $c_Q \le \dot{Q}_{\mathrm{PQL}}(\gamma, y) \le C_Q$*. And for each* $\boldsymbol{z}$*,* $\mathrm{var}(Y|Z = z)$ *and* $\psi^{-1}(\int_a^b \alpha + z(t)\beta(t)dt)$ *are nonzero.*

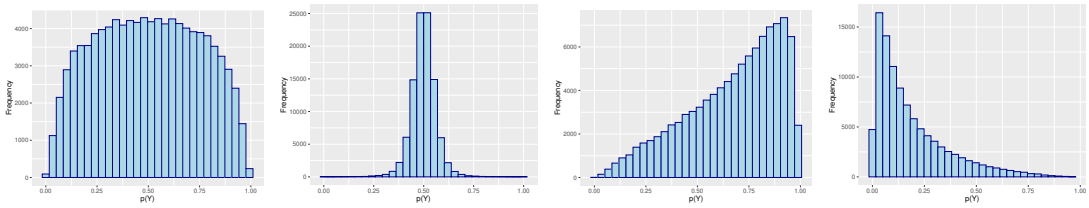
**Remark 5** *Assumption 1 is about the smoothness of the slope function, which has been widely used in the literature of nonparametric estimation (Liu et al., 2013; Kim and Wang, 2021; Yu et al., 2020). Assumption 2 gives some moment conditions on scalar response and functional predictor. Combing* $\|\boldsymbol{D}_q\|_\infty = O(K^{2q-1})$ *with Assumption 3, we can get* $\|\lambda \boldsymbol{D}\|_\infty = o(n^{1/2}K^{-1/2})$*. Thus, we can get* $\|\boldsymbol{H}_{k,n}^{\psi}\|_\infty = O(1/K)$*. Note that (11) in Assumption 4 implies that* $\delta \sim K^{-1}$*, that is,* $\delta$ *and* $K^{-1}$ *are rate-wise equivalent. The second condition in Assumption 4 implies that the functional predictor* $Z(t)$ *is away from zero in every small area of the domain* $[a, b]$*, which is reasonable to make the functional coefficient* $\beta(t)$ *estimable in the whole domain* $[a, b]$*. As mentioned in Ai et al. (2021), Assumption 6 restricts the weights in the estimation equation (2) and ensures the order of the extremely small subsampling probabilities. Besides, this assumption gives the order of the subsampling size L. Assumption 7 is a common assumption used under the quasi-likelihood framework (Carroll et al., 1997; Wang and Cao, 2018; Yu et al., 2020; Kim and Wang, 2021). And* $\dot{Q}_{\mathrm{PQL}}(\eta, y) < 0$ *ensures the uniqueness of the solution (7).*

## Appendix B. Additional Simulation Studies

This appendix includes the additional figures of the simulation studies in Section 4 and three additional simulation studies to show the good performances of the proposed functional L-optimality subsampling method.

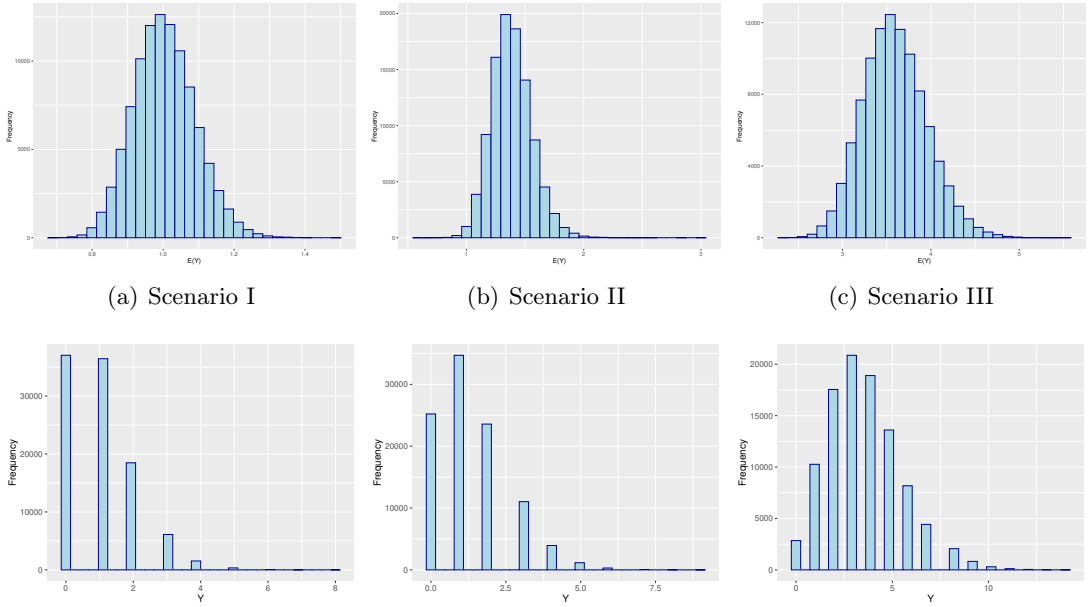
### B.1 Simulation I and II (Continued)

This section lists two additional plots (Figures 9 - 10) of Simulation I and II in the manuscript. Give one full data set with size $n = 10^5$ as an example, Figure 9 shows the histogram for $p(x_i)$ under four scenarios in Simulation I. Figure 10 shows the histogram for $\lambda(x_i)$ and the response $y_i$ under three scenarios in Simulation II based on one full data set with $n = 10^5$.




(a) Scenario I (b) Scenario II (c) Scenario III (d) Scenario IV

Figure 9: The histogram of a random example for $p(x_i)$ under four scenarios in Simulation I when the full data size is $n = 10^5$.




(a) Scenario I (b) Scenario II (c) Scenario III

(d) Scenario I (e) Scenario II (f) Scenario III

Figure 10: The histogram of a random example for $E(y_i) = \lambda(x_i)$ (Panels (a)-(c)) and $y_i$ (Panels (d)-(e)) under three scenarios in Simulation II when the full data size is $n = 10^5$.

## B.2 Simulation III

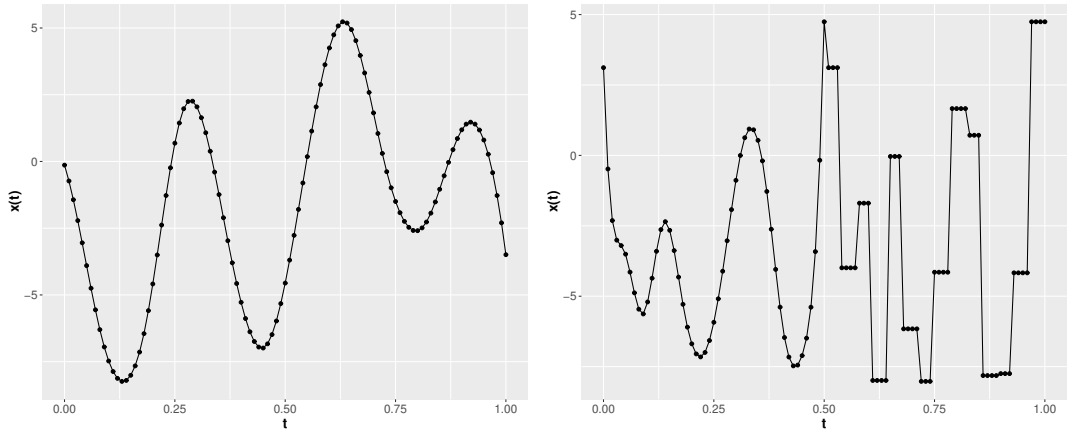
In this section, we want to study the finite sample performance of the functional L-optimality subsampling method described in Algorithm 1 for estimating the functional logistic regression when the true underlying smoothness of $X(t)$ across time is different. We set the true functional coefficient $\beta(t) = 1.8 \times \sin(0.85\pi t)$. Denote the inverse logistic function as $\psi(\cdot) = \exp(\cdot)/(1 + \exp(\cdot))$ and $p(x_i) = \psi(\int_0^1 x_i(t)\beta(t)dt)$, then we generated

responses $y(x_i) \sim$ Binomial$(1, p(x_i))$ as pseudo-Bernoulli r.v.s with probability $p(x_i)$. We consider the following two different scenarios to generate the functional predictor $x_i(t)$:

- **Scenario I.** The functional predictor $x_i(t)$ is generated by $x_i(t) = \sum a_{ij} B_j(t)$, where $B_j(t)$ are cubic B-spline basis functions defined on 12 equally spaced knots in $[0, 1]$, and the coefficients $a_{ij}$ are i.i.d from $N(0, 6)$.

- **Scenario II.** When $t \in [0, 0.5]$, $x_i(t)$ is generated by $x_i(t) = \sum a_{ij} B_j(t)$, where $B_j(t)$ are cubic B-spline basis functions defined on 12 equally spaced knots in $[0, 0.5]$ and the coefficients $a_{ij}$ are i.i.d from $N(0, 6)$. For $t \in (0.5, 1]$, $x_i(t) = \sum a_{ij} B_j(t)$, where $B_j(t)$ are constant B-spline basis functions defined on 15 equally spaced knots in $[0.5, 1]$ and $a_{ij} \overset{iid}{\sim} N(0, 6)$.

In Figure 11, we give one example of the functional predictor $x_i(t)$ under Scenario I and II, respectively. From Figure 11-(a), it is evident that the smoothness of $x_i(t)$ from Scenario I stays the same in the whole region. Figure 11-(b) displays an example of $x_i(t)$ curve from Scenario II. From this figure, we can see that $x_i(t)$ shows different levels of smoothness in the whole region, smooth when $0 < t < 0.5$, but rough when $0.5 < t < 1$. Let the full sample sizes $n = 10^5$ and the subsampling sizes $L = 200, 300, 500, 800, 1000, 1200, 1400, 1600, 1800, 2200, 2500, 2800, 3000$. Based on 300 replications, As shown in Figure 12, the improvement of FLoS relative to UNIS about RIMSE under scenario II is larger than that under scenario I. Then we can say that our proposed method FLoS has greater advantages when the smoothness of $X(t)$ across time is different.




(a) Scenario I (b) Scenario II

Figure 11: An example of the simulated functional predictor $x_i(t)$ under two scenarios in Simulation III when the full sample size is $n = 10^5$.

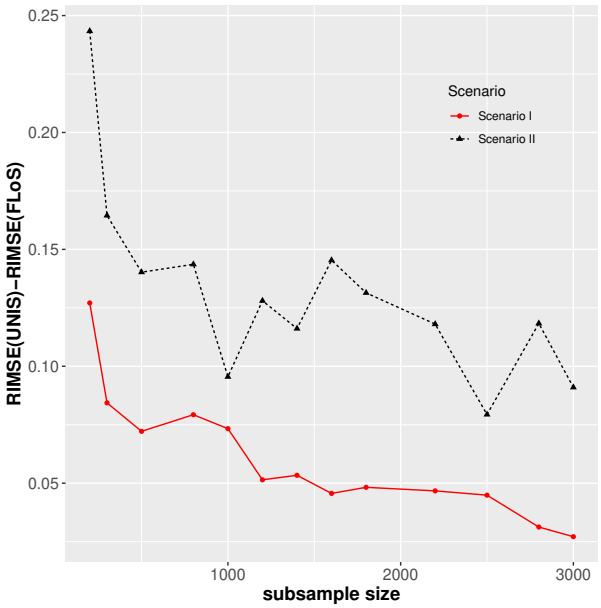


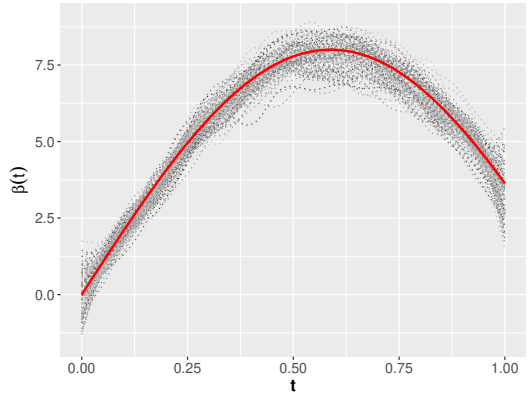

Figure 12: The comparison of the improvement in terms of the root integrated mean squared errors (RIMSE) using the functional L-optimality subsampling (FLoS) over using the uniform subsampling (UNIS) approach under two scenarios.

### B.3 Simulation IV

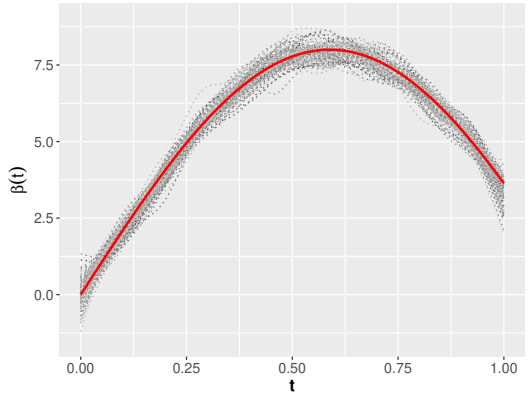
In this simulation study, we want to assess the variability of our proposed functional L-optimality subsampling method described in Algorithm 1. We set the true functional coefficient $\beta(t) = 8 \times \sin(0.85\pi t)$. Denote the inverse logistic function as $\psi(\cdot) = \exp(\cdot)/(1+\exp(\cdot))$ and $p(x_i) = \psi(\int_0^1 x_i(t)\beta(t)dt)$. We generate responses $y(x_i) \sim$ Binomial$(1, p(x_i))$ as pseudo-Bernoulli random variables with probability $p(x_i)$. The functional predictor $x_i(t)$ is generated by $x_i(t) = \sum a_{ij}B_j(t)$, where $B_j(t)$ are cubic B-spline basis functions defined on 66 equally spaced knots in $[0, 1]$ and the basis coefficients $a_{ij} \overset{iid}{\sim} N(0, 6)$.

Given one full data set with $n = 10^5$, Figure 13-(a), Figure 13-(b) and 13-(c) plot the true functional coefficient $\beta(t)$ (solid) and its 100 subsample estimators (dotted) based on the functional L-optimality subsampling method when the subsample size $L = 1000, 2000, 3000$ respectively. They show that the subsample estimators are close to the true curve and the gray area becomes narrow as the subsample size increases, especially for $0.1 < t < 0.4$. The estimators have less variability with the subsample size increasing. Moreover, to study the variability across the whole domain $[0, 1]$, we also give the boxplot (Figure 13-(d)) of the RIMSE based on 300 times subsampling procedure and subsample size $L = 300, 400, 600, 800, 1000, 1200, 1400, 1600, 1800, 2000, 2200, 2500, 2800, 3000$. Figure 13-(d) shows that the variability of the RIMSE decreases when the subsample size increases. In addition, to compare the variability of the subsample estimators using our proposed functional L-optimality subsampling method with that using the uniform

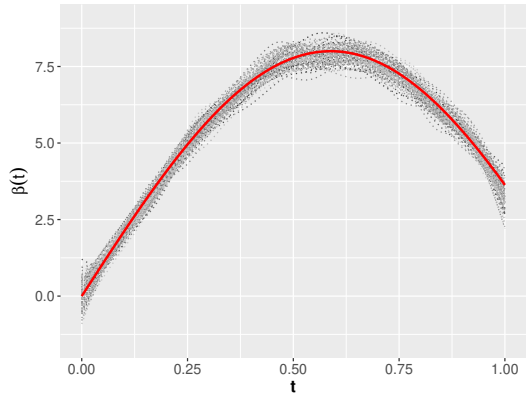
subsampling method, we present the standard deviation of the RIMSE of subsample estimators using two methods in Figure 14. Figure 14 displays that our method performs better than the uniform subsampling method in terms of the standard deviation.



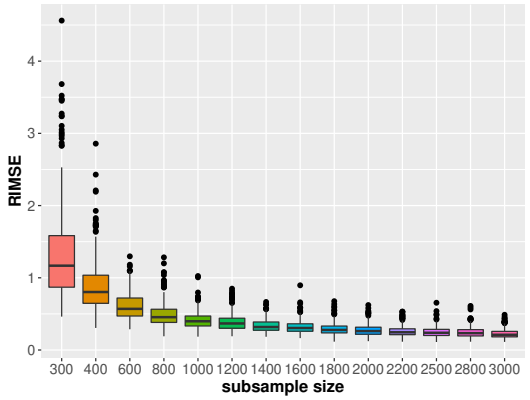

(a) $\breve{\beta}(t)$ when subsample size $L = 1000$




(b) $\breve{\beta}(t)$ when subsample size $L = 2000$




(c) $\breve{\beta}(t)$ when subsample size $L = 3000$




(d) Boxplot of the RIMSE of $\breve{\beta}(t)$

Figure 13: (a)-(c) The estimated functional coefficient $\breve{\beta}(t)$ (gray dotted curves) obtained by the functional L-optimality subsampling (FLoS) method when the subsample size $L = 1000, 2000, 3000$ and the full data size $n = 10^5$. The red solid curves are true $\beta(t)$. (d) Boxplot of the root integrated mean squared errors (RIMSE) of the subsampling estimators using the FLoS method under different subsample sizes when the full data size $n = 10^5$.

### B.4 Simulation V

To study the performance of the proposed functional L-optimality subsampling method described in Algorithm 1 when the functional predictor $x_i(t)$ is smooth, We set the true functional coefficient $\beta(t) = 1.8 \times \sin(0.85\pi t)$. Denote the inverse logistic function as

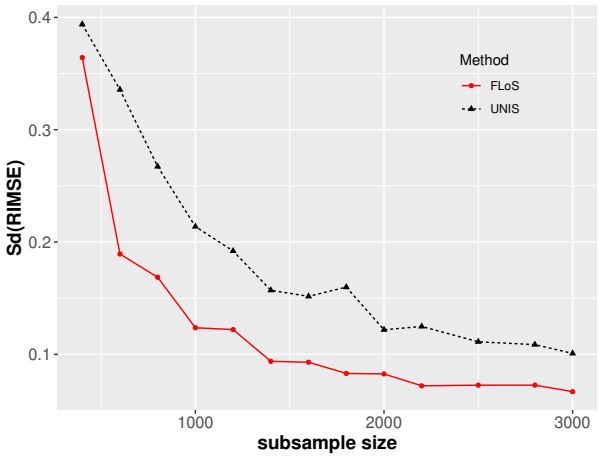



Figure 14: The standard deviation (sd) of the root integrated mean squared errors (RIMSE) of the estimated functional coefficient $\breve{\beta}(t)$ in the functional logistic regression model (Simulation IV) by using the functional L-optimality subsampling (FLoS) method and the uniform subsampling (UNIS) approach under with various subsample sizes L when the full data size $n = 10^5$.

$\psi(\cdot) = \exp(\cdot)/(1 + \exp(\cdot))$ and $p(x_i) = \psi(\int_0^1 x_i(t)\beta(t)dt)$, then we generated responses $y(x_i) \sim \text{Binomial}(1, p(x_i))$ as pseudo-Bernoulli r.v.s with probability $p(x_i)$. The functional predictor $x_i(t)$ is generated by $x_i(t) = \sum a_{ij}B_j(t)$, where $B_j(t)$ are cubic B-spline basis functions defined on 20 equally spaced knots in $[0, 1]$. The basis coefficients $a_{ij}$ are generated as $a_{i1} \overset{iid}{\sim} N(0, 6)$, $a_{ij} = a_{i,j-1} + \Lambda_{ij}$ and $\Lambda_{ij} \overset{iid}{\sim} N(0, 0.5)$ for $1 \leq i \leq n$ and $j \geq 2$.

Figure 15 displays one example of the simulated functional predictor $x_i(t)$, which shows that the functional predictor $x_i(t)$ is smooth. When the full sample size $n = 10^5$ and the subsample size $L = 300, 400, 600, 800, 1000, 1200, 1400, 1600, 1800, 2000, 2200, 2500, 2800, 3000$, we show the mean RIMSE of the estimators based on 300 replications in Figure 16-(a). And Figure 16-(b) displays the mean RIMSE of the estimators using two subsampling methods with $n = 10^6$ and subsample size $L = 500, 800, 1100, 1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000, 5500, 6000, 6500, 7000, 7500, 8000, 8500, 9000, 9500, 10000$. Figure 16-(c) and Figure 16-(d) also show the PCCs with different subsample size under $n = 10^5$ and $n = 10^6$, respectively. From this figure, it is clear that when the functional predictor $x_i(t)$ is smooth, the proposed method is superior to the uniform subsampling method, both for RIMSEs and PCCs.

## Appendix C. Estimation Steps for Functional Generalized Linear Model

In this section, we provide the estimation procedure for the functional generalized linear model to estimate the coefficient function $\boldsymbol{\beta}(\cdot)$ by solving the equation (1). Now we

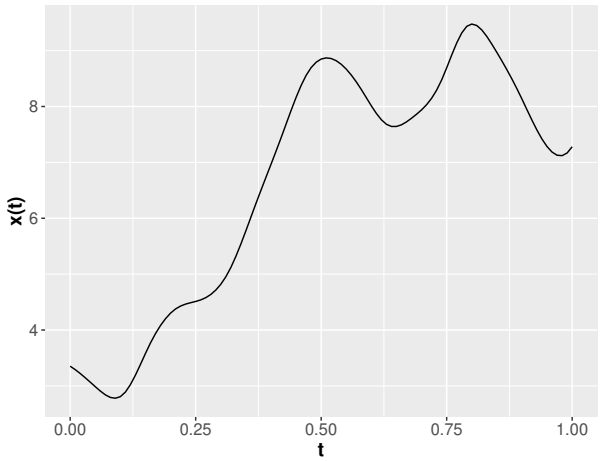



Figure 15: An example of the simulated functional predictor $x_i(t)$ under the setting of Simulation V when the full data size $n = 10^5$.

can apply the Newton–Raphson algorithm to iteratively solve it. The detailed estimation steps are given below:

**Step 0.** Obtain an initial estimate $\boldsymbol{c}_0$.

**Step 1.** At the $(k+1)$th iteration,

$$\boldsymbol{c}_{\mathrm{PQL}}^{(k+1)} = \boldsymbol{c}_{\mathrm{PQL}}^{(k)} + \dot{Q}_{\mathrm{PQL}}^{-1}(\boldsymbol{c}_{\mathrm{PQL}}^{(k)}) Q_{\mathrm{PQL}}(\boldsymbol{c}_{\mathrm{PQL}}^{(k)}). \tag{12}$$

**Step 2.** Repeat Step 1 until convergence is reached.

## Appendix D. Proof

This section includes the detailed proofs of the theoretical results. To prove our theorems in the manuscripts, we start by proving the following lemmas.

### D.1 Some Lemmas

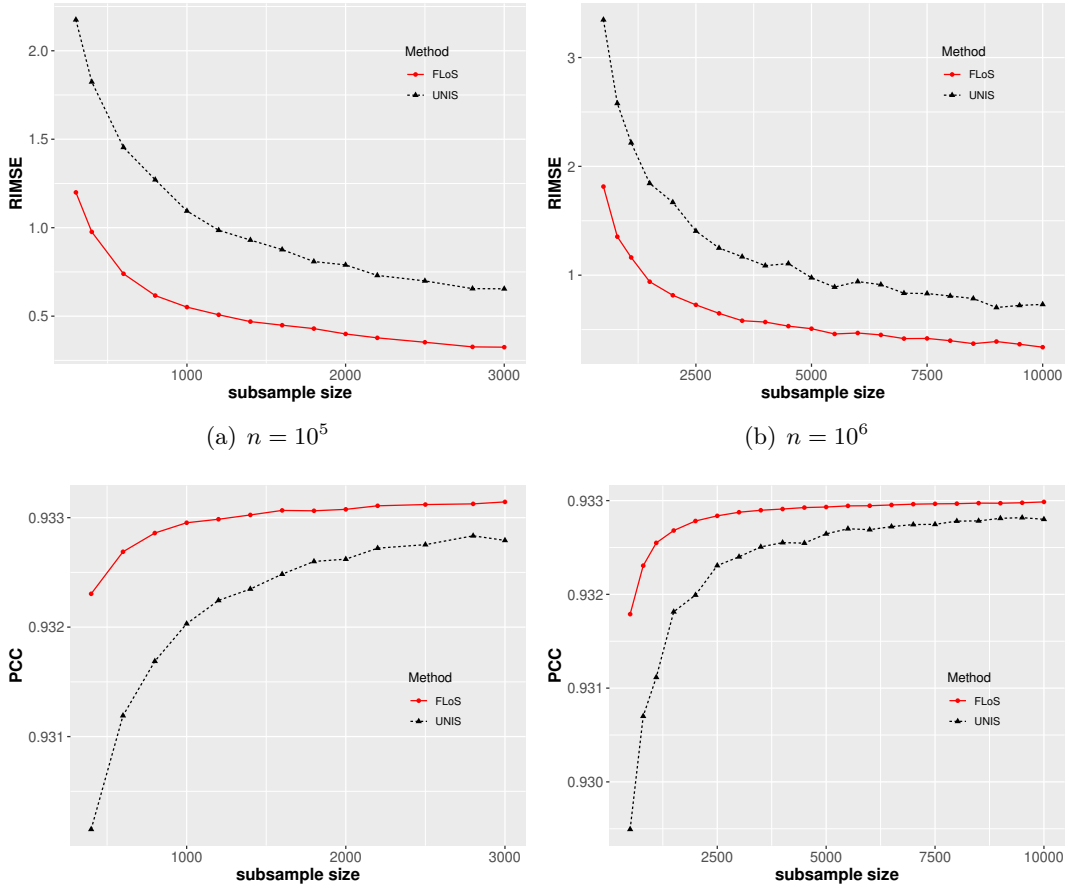
To prove our theorems in the manuscripts, we start by proving the following lemmas.

**Lemma 6** *Under Assumptions 1,* $s_\beta(t) - \beta(t) = b_a(t) + o(K^{-d})$.

**Proof** The proof of this lemma can be found in Barrow and Smith (1978). ■

**Lemma 7** *Under Assumption 2, 4, 5 and 7, (i) there exists constants* $C_G > c_G > 0$ *such that*

$$c_G K^{-1} \leq \rho_{min}(\boldsymbol{G}_{k,n}^{\psi}) \leq \rho_{max}(\boldsymbol{G}_{k,n}^{\psi}) \leq C_G K^{-1},$$




(a) $n = 10^5$ (b) $n = 10^6$

(c) $n = 10^5$ (d) $n = 10^6$

Figure 16: The average of the root integrated mean squared errors (RIMSE) (Panels (a)-(b)) and the proportions of correct classifications (PCC) defined in (8) (Panels (c)-(d)) in the functional logistic regression model (Simulation V) by using the functional L-optimality subsampling (FLoS) method and the uniform subsampling (UNIS) approach with various subsample sizes L when the full data size is $n = 10^5$ and $n = 10^6$, respectively.

*where $\rho_{min}$ and $\rho_{max}$ denote the smallest and largest eigenvalues of a matrix, respectively.*
*(ii) we can have $\|\boldsymbol{G}_{k,n}^{\psi}\|_{\infty} = O(K^{-1})$.*

**Proof** Using the result in (i), (ii) can be derived directly from Lemma 6.3 and Lemma 6.4 in (Zhou et al., 1998). We only give the proof of (i). For any non-zero vector $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_{K+p+1})^T$ with $\|\boldsymbol{\mu}\| = 1$, note that

$$\rho_{min}(\boldsymbol{G}_{k,n}^{\psi}) = \min \boldsymbol{\mu}^T \boldsymbol{G}_{k,n}^{\psi} \boldsymbol{\mu}, \quad \rho_{max}(\boldsymbol{G}_{k,n}^{\psi}) = \max \boldsymbol{\mu}^T \boldsymbol{G}_{k,n}^{\psi} \boldsymbol{\mu}.$$

According to the definition of $\boldsymbol{G}_{k,n}^{\psi}$, we get

$$\begin{aligned}
\boldsymbol{\mu}^T \boldsymbol{G}_{k,n}^{\psi} \boldsymbol{\mu} &= \frac{1}{n} \boldsymbol{\mu}^T \boldsymbol{N}^T \boldsymbol{\Psi} \boldsymbol{N} \boldsymbol{\mu} \\
&= \frac{1}{n} \sum_{i=1}^{n} \dot{\psi}(\boldsymbol{N}_i^T \boldsymbol{c}) \boldsymbol{\mu}^T \boldsymbol{N}_i \boldsymbol{N}_i^T \boldsymbol{\mu} \\
&= \frac{1}{n} \sum_{i=1}^{n} \dot{\psi}(\boldsymbol{N}_i^T \boldsymbol{c}) \left( \sum_{j=1}^{K+p+1} \mu_j N_{ij} \right)^2 .
\end{aligned}$$

For $|\sum_{j=1}^{K+p+1} \mu_j N_{ij}|$, by the property $\sum_{j=1}^{K+p+1} N_{p+1,j}(t) = 1$, we have

$$\begin{aligned}
\left| \sum_{j=1}^{K+p+1} \mu_j N_{ij} \right| &= \left| \int \sum_{j=1}^{K+p+1} \mu_j N_{p+1,j}(t) z_i(t) dt \right| \\
&\leq \int \left| \sum_{j=1}^{K+p+1} \mu_j N_{p+1,j}(t) z_i(t) \right| dt \\
&\leq \int \left\{ \left( \sum_{j=1}^{K+p+1} \mu_j^2 N_{p+1,j}(t) z_i^2(t) \right)^{1/2} \left( \sum_{j=1}^{K+p+1} N_{p+1,j}(t) \right)^{1/2} \right\} dt \\
&= \int \left\{ \left( \sum_{j=1}^{K+p+1} \mu_j^2 N_{p+1,j}(t) z_i^2(t) \right)^{1/2} \right\} dt \\
&\leq \int \left( \sum_{j=1}^{K+p+1} |\mu_j| |z_i(t)| N_{p+1,j}^{1/2}(t) \right) dt \\
&= \sum_{j=1}^{K+p+1} |\mu_j| \int |z_i(t)| N_{p+1,j}^{1/2}(t) dt \\
&\leq \sum_{j=1}^{K+p+1} |\mu_j| \left( \int z_i^2(t) dt \right)^{1/2} \left( \int N_{p+1,j}(t) dt \right)^{1/2} .
\end{aligned}$$

By the property $\int N_{p+1,j}(t)dt = O(K^{-1})$ and Assumption 2, one has $\left|\sum_{j=1}^{K+p+1} \mu_j N_{ij}\right| = O(K^{-1/2})$. Next, by Assumption 2, $\boldsymbol{\mu}^T \boldsymbol{G}_{k,n}^{\psi} \boldsymbol{\mu} \leq C_G K^{-1}$. Applying the second condition in Assumption 4, $c_G K^{-1} \leq \rho_{min}(\boldsymbol{G}_{k,n}^{\psi})$ holds as well. This completes the proof of (i). ■

**Lemma 8** *Under Assumptions 3 and 4, we can get (i) $\|\boldsymbol{D}_q\|_\infty = O(K^{2q-1})$; (ii) there are two positive constants $c_D < C_D$ such that $c_D K^{2q-1}\|\boldsymbol{\mu}\|_2^2 \leq \boldsymbol{\mu}^T \boldsymbol{D}_q \boldsymbol{\mu} \leq C_D K^{2q-1}\|\boldsymbol{\mu}\|_2^2$.*

**Proof** (i) Let $\boldsymbol{N}_{p+1-q}(t) = (N_{j,p+1-q}(t) : -p+q \leq j \leq K)^T$, from the derivative formula for B-spline functions (de Boor, 1978), we can get

$$s_\beta^{(q)}(t) = \boldsymbol{N}_{p+1-q}^T(t)\boldsymbol{c}^{(q)},$$

where $\boldsymbol{c}^{(q)}$ are defined recursively via

$$\begin{aligned} \boldsymbol{c}^{(1)} &= \boldsymbol{\nabla}_1 \boldsymbol{c}, \\ \boldsymbol{c}^{(q)} &= \boldsymbol{\nabla}_q \boldsymbol{c}^{(q-1)}, \end{aligned}$$

$$\boldsymbol{\nabla}_1 = (p+1-1) \times \begin{pmatrix} \frac{-1}{k_1 - k_{-p+1}} & \frac{1}{k_1 - k_{-p+1}} & 0 & \cdots & 0 \\ 0 & \frac{-1}{k_{-p+2} - k_{-p+3}} & \frac{1}{k_{-p+2} - k_{-p+3}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{-1}{k_{K+p} - k_K} & \frac{1}{k_{K+p} - k_K} \end{pmatrix},$$

$$\boldsymbol{\nabla}_q = (p+1-q) \times \begin{pmatrix} \frac{-1}{k_1 - k_{-p+q}} & \frac{1}{k_1 - k_{-p+q}} & 0 & \cdots & 0 \\ 0 & \frac{-1}{k_{-q+1} - k_{-p+2}} & \frac{1}{k_{-q+1} - k_{-p+2}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{-1}{k_{K+p+1-q} - k_K} & \frac{1}{k_{K+p+1-q} - k_K} \end{pmatrix},$$

$\boldsymbol{\nabla}_1$ is a $(K+p+1-1)\times(K+p+1)$ dimensional matrix and $\boldsymbol{\nabla}_q$ is a $(K+p+1-q)\times(K+p+1)$ dimensional matrix. So, we can rewrite the second penalty term in (1) as $\lambda \boldsymbol{c}^T \boldsymbol{\Delta}_q^T \boldsymbol{R} \boldsymbol{\Delta}_q \boldsymbol{c}$, where the matrix $\boldsymbol{R} = \int_a^b \boldsymbol{N}_{p+1-q}(t)\boldsymbol{N}_{p+1-q}^T(t)dt$ and $\boldsymbol{\Delta}_q = \boldsymbol{\nabla}_1 \ldots \boldsymbol{\nabla}_q$. Note that (A1) implies that $\delta \sim K^{-1}$, i.e., $\delta$ and $K^{-1}$ are rate-wise equivalent. Moreover, by definition $\|\Delta_l\|_\infty = O(K), l = 1, \ldots, q$, thus, $\|\boldsymbol{\Delta}_q\|_\infty = O(\delta^{-q}) = O(K^q)$. And, from the definition of B-spline, we can get $\|\boldsymbol{R}\|_\infty = O(K^{-1})$. Because $\boldsymbol{D}_q = \boldsymbol{\Delta}_q^T \boldsymbol{R} \boldsymbol{\Delta}_q$, the desired result holds. For (ii), it can be derived from Lemma 6.1 in Cardot et al. (2003) and the proof of it is omitted. ■

**Lemma 9** *Under Assumption 2-5 and 7, (i) there are some positive constants $c_H < C_H$ such that*

$$c_H K^{-1} \leq \rho_{min}(\boldsymbol{H}_{k,n}^{\psi}) \leq \rho_{max}(\boldsymbol{H}_{k,n}^{\psi}) \leq C_H K^{-1}.$$

*(ii)* $\|\boldsymbol{H}_{k,n}^{\psi}\|_{\infty} = O(K^{-1})$.

**Proof** Under Assumption 3, it can be directly derived from Lemma 7 and 8. ■

**Lemma 10** *Under Assumptions 2, 4, 6 and 7, as $n, L \to \infty$, in probability,*

$$\frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{Lp_i}\dot{\psi}(\boldsymbol{N}_i^T\boldsymbol{c})\boldsymbol{N}_i\boldsymbol{N}_i^T - \frac{1}{n}\sum_{i=1}^{n}\dot{\psi}(\boldsymbol{N}_i^T\boldsymbol{c})\boldsymbol{N}_i\boldsymbol{N}_i^T = o_{p|\mathcal{F}_n}(1). \tag{13}$$

**Proof** Direct calculation shows that conditionally on $\mathcal{F}_n$,

$$\mathrm{E}\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{Lp_i}\dot{\psi}(\boldsymbol{N}_i^T\boldsymbol{c})\boldsymbol{N}_i\boldsymbol{N}_i^T|\mathcal{F}_n\right\} = \frac{1}{n}\sum_{i=1}^{n}\dot{\psi}(\boldsymbol{N}_i^T\boldsymbol{c})\boldsymbol{N}_i\boldsymbol{N}_i^T.$$

Let $\boldsymbol{G}_{k,n}^{\psi,*} = n^{-1}\sum_{i=1}^{n}R_i\dot{\psi}(\boldsymbol{N}_i^T\boldsymbol{c})\boldsymbol{N}_i\boldsymbol{N}_i^T/(Lp_i)$, then $\boldsymbol{G}_{k,n}^{\psi,*} - \boldsymbol{G}_{k,n}^{\psi} = n^{-1}\sum_{i=1}^{n}(R_i - Lp_i)\dot{\psi}(\boldsymbol{N}_i^T\boldsymbol{c})\boldsymbol{N}_i\boldsymbol{N}_i^T/(Lp_i)$. For any component of $\boldsymbol{G}_{k,n}^{\psi,*} - \boldsymbol{G}_{k,n}^{\psi}$, we have

$$\begin{aligned}
&\mathrm{E}\left\{\boldsymbol{G}_{k,n}^{\psi,*j_1j_2} - \boldsymbol{G}_{k,n}^{\psi,j_1j_2}|\mathcal{F}_n\right\}^2\\
=&\frac{1}{n^2L}\sum_{i=1}^{n}\frac{1-p_i}{p_i}\left\{\dot{\psi}(\boldsymbol{N}_i^T\boldsymbol{c})\left(\boldsymbol{N}_i\boldsymbol{N}_i^T\right)^{j_1j_2}\right\}^2\\
\leq&\frac{1}{n^2L}\sum_{i=1}^{n}\frac{1}{p_i}\left\{\dot{\psi}(\boldsymbol{N}_i^T\boldsymbol{c})\left(\boldsymbol{N}_i\boldsymbol{N}_i^T\right)^{j_1j_2}\right\}^2\\
\leq&\max(\frac{1}{nLp_i})\frac{1}{n}\sum_{i=1}^{n}\left\{\dot{\psi}(\boldsymbol{N}_i^T\boldsymbol{c})\left(\boldsymbol{N}_i\boldsymbol{N}_i^T\right)^{j_1j_2}\right\}^2\\
=&o_p(\frac{1}{\sqrt{L}})O(K^{-2}) = o_p(1).
\end{aligned}$$

Thus, through Chebyshev's inequality, the desired conclusion follows. ■

**Lemma 11** *Under Assumptions 1 - 7, conditional on $\mathcal{F}_n$, as $n, L \to \infty$,*

$$\frac{\sqrt{L}}{n}\boldsymbol{V}_p^{\psi\,-1/2}Q_{\mathrm{PQL}}^{*}(\widehat{\boldsymbol{c}}_{\mathrm{PQL}}) \to N(0,1),$$

*in distribution.*

**Proof** Direct calculation shows that

$$\mathrm{E}\left\{\frac{1}{n}Q^*_{\mathrm{PQL}}(\widehat{\boldsymbol{c}}_{\mathrm{PQL}})|\mathcal{F}_n\right\} = \frac{1}{n}Q_{\mathrm{PQL}}(\widehat{\boldsymbol{c}}_{\mathrm{PQL}}) = 0,$$

and

$$\begin{aligned}
&\mathrm{Var}\left\{\frac{1}{n}Q^*_{\mathrm{PQL}}(\widehat{\boldsymbol{c}}_{\mathrm{PQL}})|\mathcal{F}_n\right\}\\
=&\frac{1}{n^2L}\sum_{i=1}^{n}\frac{1}{p_i}\left\{y_i-\psi(\boldsymbol{N}_i^T\widehat{\boldsymbol{c}}_{\mathrm{PQL}})\right\}^2\boldsymbol{N}_i\boldsymbol{N}_i^T-\frac{1}{n^2L}\sum_{i=1}^{n}\left\{y_i-\psi(\boldsymbol{N}_i^T\widehat{\boldsymbol{c}}_{\mathrm{PQL}})\right\}^2\boldsymbol{N}_i\boldsymbol{N}_i^T\\
=&\frac{1}{L}\boldsymbol{V}_p^{\psi}-o_p(1).
\end{aligned}$$

Next, we check the Lindeberg-Feller condition under the conditional distribution. Denote $\boldsymbol{\varpi}_i = \frac{1}{n}\left\{\frac{R_i}{Lp_i}\left\{y_i-\psi(\boldsymbol{N}_i^T\widehat{\boldsymbol{c}}_{\mathrm{PQL}})\right\}\boldsymbol{N}_i-\lambda\boldsymbol{D}\widehat{\boldsymbol{c}}_{\mathrm{PQL}}\right\}$. For any $\epsilon>0$,

$$\begin{aligned}
&\sum_{i=1}^{n}\mathrm{E}\left\{\|\boldsymbol{\varpi}_i\|^2\boldsymbol{I}(\|\boldsymbol{\varpi}_i\|>\epsilon)|\mathcal{F}_n\right\}\\
\le&\frac{1}{\epsilon}\sum_{i=1}^{n}\mathrm{E}\left\{\|\boldsymbol{\varpi}_i\|^3|\mathcal{F}_n\right\}\\
\le&\frac{1}{n^3}\sum_{i=1}^{n}\left\{\mathrm{E}(\frac{R_i^3|y_i-\psi(\boldsymbol{N}_i^T\widehat{\boldsymbol{c}}_{\mathrm{PQL}})|^3\|\boldsymbol{N}_i\|^3}{L^3p_i^3}|\mathcal{F}_n)+\|\lambda\boldsymbol{D}\widehat{\boldsymbol{c}}_{\mathrm{PQL}}\|^3\right\}\\
=&\frac{1}{n^3}\sum_{i=1}^{n}\frac{(L(L-1)(L-2)p_i^3+3L(L-1)p_i^2+Lp_i)|y_i-\psi(\boldsymbol{N}_i^T\widehat{\boldsymbol{c}}_{\mathrm{PQL}})|^3\|\boldsymbol{N}_i\|^3}{L^3p_i^3}+o(K^{-3})\\
=&o_p(1),
\end{aligned}$$

where the last equality holds by Assumptions 1 - 7. By Lindeberg-Feller central limit theorem, the desired result follows. ■

**Lemma 12** *Under Assumptions 1 - 7, as $n, L\to\infty$,*

$$\frac{\sqrt{L}}{n}\boldsymbol{W}_p^{\psi^{-1/2}}Q^*_{\mathrm{PQL}}(\boldsymbol{c})\to N(0,1),$$

*in distribution, where*

$$\boldsymbol{W}_p^{\psi}=\frac{1}{n^2}\sum_{i=1}^{n}\frac{E(y_i-\psi(\boldsymbol{N}_i^T\boldsymbol{c}))^2\boldsymbol{N}_i\boldsymbol{N}_i^T}{p_i}.$$

**Proof** Note that

$$\begin{aligned}
&\mathrm{E}\left\{\frac{1}{n}Q^*_{\mathrm{PQL}}(\boldsymbol{c})\right\} = \mathrm{E}\left\{\mathrm{E}\left\{\frac{1}{n}Q^*_{\mathrm{PQL}}(\boldsymbol{c})|\mathcal{F}_n\right\}\right\}\\
=&\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{N}_i(\psi(<\boldsymbol{z}_i,\boldsymbol{\beta}>)-\psi(\boldsymbol{N}_i^T\boldsymbol{c}))-\frac{1}{n}\lambda\boldsymbol{D}\boldsymbol{c}\\
=&o((nK)^{-1/2}),
\end{aligned}$$

and

$$\mathrm{Var}\left\{\frac{1}{n}Q^*(\boldsymbol{c})\right\} = \mathrm{E}\left\{\mathrm{Var}\left\{\frac{1}{n}Q^*(\boldsymbol{c})|\mathcal{F}_n\right\}\right\} + \mathrm{Var}\left\{\mathrm{E}\left\{\frac{1}{n}Q^*(\boldsymbol{c})|\mathcal{F}_n\right\}\right\}. \tag{14}$$

For the first term in (14), we can get

$$\begin{aligned}
&\mathrm{E}\left\{\mathrm{Var}\left\{\frac{1}{n}Q^*_{\mathrm{PQL}}(\boldsymbol{c})|\mathcal{F}_n\right\}\right\}\\
=&\mathrm{E}\left\{\frac{1}{n^2L}\sum_{i=1}^{n}\frac{(y_i-\psi(\boldsymbol{N}_i^T\boldsymbol{c}))^2\boldsymbol{N}_i\boldsymbol{N}_i^T}{p_i}-\frac{1}{n^2L}\sum_{i=1}^{n}(y_i-\psi(\boldsymbol{N}_i^T\boldsymbol{c}))^2\boldsymbol{N}_i\boldsymbol{N}_i^T\right\}\\
=&\frac{1}{n^2L}\sum_{i=1}^{n}\frac{\mathrm{E}(y_i-\psi(\boldsymbol{N}_i^T\boldsymbol{c}))^2\boldsymbol{N}_i\boldsymbol{N}_i^T}{p_i}-\frac{1}{n^2L}\sum_{i=1}^{n}\mathrm{E}(y_i-\psi(\boldsymbol{N}_i^T\boldsymbol{c}))^2\boldsymbol{N}_i\boldsymbol{N}_i^T\\
=&\frac{1}{L}\boldsymbol{W}_p^{\psi}-O(\frac{1}{nLK}).
\end{aligned}$$

Similarly, we deal with the second term in (14),

$$\begin{aligned}
\mathrm{Var}\left\{\mathrm{E}\left\{\frac{1}{n}Q^*_{\mathrm{PQL}}(\boldsymbol{c})|\mathcal{F}_n\right\}\right\} &= \mathrm{Var}\left\{\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{N}_i(y_i-\psi(\boldsymbol{N}_i^T\boldsymbol{c}))-\lambda\boldsymbol{D}\boldsymbol{c}\right\}\\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2\boldsymbol{N}_i\boldsymbol{N}_i^T\\
&= O(\frac{1}{nK}).
\end{aligned}$$

Under Assumption 1 - 7, $\frac{1}{L}\boldsymbol{W}_p^{\psi} = o(K^{-1}L^{-1/2})$. Consequently,

$$\mathrm{Var}\left\{\frac{1}{n}Q^*_{\mathrm{PQL}}(\boldsymbol{c})\right\} = \frac{1}{L}\boldsymbol{W}_p^{\psi}+O(\frac{1}{nK}).$$

Next, we check the Lindeberg-Feller condition, denote

$$\boldsymbol{\zeta}_i = \frac{1}{n}\left\{\frac{R_i}{Lp_i}\left\{y_i-\psi(\boldsymbol{N}_i^T\boldsymbol{c})\right\}\boldsymbol{N}_i-\lambda\boldsymbol{D}\boldsymbol{c}\right\}.$$

For any $\epsilon > 0$,

$$\begin{aligned}
&\sum_{i=1}^{n} \mathrm{E}\left\{\|\boldsymbol{\zeta}_i\|^2 \boldsymbol{I}(\|\boldsymbol{\zeta}_i\| > \epsilon)\right\} \\
\leq &\frac{1}{\epsilon}\sum_{i=1}^{n} \mathrm{E}\left\{\|\boldsymbol{\zeta}_i\|^3\right\} \\
\leq &\frac{1}{n^3}\sum_{i=1}^{n}\left\{\mathrm{E}(\frac{R_i^3|y_i - \psi(\boldsymbol{N}_i^T\boldsymbol{c})|^3\|\boldsymbol{N}_i\|^3}{L^3 p_i^3}) + \|\lambda \boldsymbol{D}\boldsymbol{c}\|^3\right\} \\
= &\frac{1}{n^3}\sum_{i=1}^{n}\frac{(L(L-1)(L-2)p_i^3 + 3L(L-1)p_i^2 + Lp_i)\mathrm{E}|y_i - \psi(\boldsymbol{N}_i^T\boldsymbol{c})|^3\|\boldsymbol{N}_i\|^3}{L^3 p_i^3} + o(K^{-3}) \\
= &o_p(1),
\end{aligned}$$

where the last equality holds by Assumptions 1 - 7. By Lindeberg-Feller's central limit theorem, the desired result follows. ■

### D.2 Proof of Main results

**Proof** ***of Theorem 1.*** By Taylor expansion, we can get

$$0 = \frac{1}{n}Q^*_{\mathrm{PQL}}(\widetilde{\boldsymbol{c}}_{\mathrm{PQL}}) = \frac{1}{n}Q^*_{\mathrm{PQL}}(\boldsymbol{c}) + \frac{1}{n}\dot{Q}^*_{\mathrm{PQL}}(\boldsymbol{c})(\widetilde{\boldsymbol{c}}_{\mathrm{PQL}} - \boldsymbol{c}) + o_p(1),$$

where $\dot{Q}^*_{\mathrm{PQL}}(\boldsymbol{c})$ is the first derivations of $Q^*_{\mathrm{PQL}}(\boldsymbol{c})$ about $\boldsymbol{c}$, and

$$\dot{Q}^*_{\mathrm{PQL}}(\boldsymbol{c}) = -\sum_{i=1}^{n} R_i/(Lp_i)\dot{\psi}(\boldsymbol{N}^T\boldsymbol{c})\boldsymbol{N}_i\boldsymbol{N}_i^T - \lambda\boldsymbol{D} = -n\boldsymbol{G}^{\psi,*}_{k,n} - \lambda\boldsymbol{D} \triangleq -n\boldsymbol{H}^{\psi,*}_{k,n}.$$

It is obviously that

$$\widetilde{\boldsymbol{c}}_{\mathrm{PQL}} - \boldsymbol{c} = \left\{-\frac{1}{n}\dot{Q}^*_{\mathrm{PQL}}(\boldsymbol{c})\right\}^{-1}\frac{1}{n}Q^*_{\mathrm{PQL}}(\boldsymbol{c}) + \left\{-\frac{1}{n}\dot{Q}^*_{\mathrm{PQL}}(\boldsymbol{c})\right\}^{-1}\cdot o_p(1),$$

and given $t$, we have

$$\begin{aligned}
\widetilde{\boldsymbol{\beta}}_{\mathrm{PQL}}(t) - \boldsymbol{\beta}(t) = &\boldsymbol{N}^T(t)\left\{-\frac{1}{n}\dot{Q}^*_{\mathrm{PQL}}(\boldsymbol{c})\right\}^{-1}\frac{1}{n}Q^*_{\mathrm{PQL}}(\boldsymbol{c}) + \boldsymbol{N}^T(t)\left\{-\frac{1}{n}\dot{Q}^*_{\mathrm{PQL}}(\boldsymbol{c})\right\}^{-1}\cdot o_p(1) \\
&+ s_\beta(t) - \beta(t) \\
= &\boldsymbol{N}^T(t)\left\{-\frac{1}{n}\dot{Q}^*_{\mathrm{PQL}}(\boldsymbol{c})\right\}^{-1}\frac{\boldsymbol{W}_p^{\psi 1/2}}{\sqrt{L}}\sqrt{L}\boldsymbol{W}_p^{\psi -1/2}\frac{1}{n}Q^*_{\mathrm{PQL}}(\boldsymbol{c}) \\
&+ \boldsymbol{N}^T(t)\left\{-\frac{1}{n}\dot{Q}^*_{\mathrm{PQL}}(\boldsymbol{c})\right\}^{-1}\cdot o_p(1) + s_\beta(t) - \beta(t).
\end{aligned} \tag{15}$$

From Lemma 10, we can get that $\boldsymbol{G}_{k,n}^{\psi,*} - \boldsymbol{G}_{k,n}^{\psi} \to 0$ and $\boldsymbol{H}_{k,n}^{\psi,*} - \boldsymbol{H}_{k,n}^{\psi} \to 0$ as $n, L \to \infty$. Thus, (15) can be written as

$$\begin{aligned}\widetilde{\boldsymbol{\beta}}_{\mathrm{PQL}}(t) - \boldsymbol{\beta}(t) = \boldsymbol{N}^T(t)\boldsymbol{H}_{k,n}^{\psi}\frac{\boldsymbol{W}_p^{\psi^{1/2}}}{\sqrt{L}}\sqrt{L}\boldsymbol{W}_p^{\psi^{-1/2}}\frac{1}{n}Q_{\mathrm{PQL}}^*(\boldsymbol{c}) + \boldsymbol{N}^T(t)\boldsymbol{H}_{k,n}^{\psi}\cdot o_p(1) \\ + s_\beta(t) - \beta(t).\end{aligned} \tag{16}$$

For Theorem 1, applying (16), as $n, L \to \infty$, it holds that

$$\begin{aligned}&(\boldsymbol{N}^T(t)(\boldsymbol{H}_{k,n}^{\psi})^{-1}\boldsymbol{W}_p^{\psi}(\boldsymbol{H}_{k,n}^{\psi})^{-1}\boldsymbol{N}(t))^{-1/2}\sqrt{L}(\widetilde{\boldsymbol{\beta}}_{\mathrm{PQL}}(t) - \boldsymbol{\beta}(t)) \\ =&(\boldsymbol{N}^T(t)(\boldsymbol{H}_{k,n}^{\psi})^{-1}\boldsymbol{W}_p^{\psi}(\boldsymbol{H}_{k,n}^{\psi})^{-1}\boldsymbol{N}(t))^{-1/2}\sqrt{L}\boldsymbol{N}^T(t)\boldsymbol{H}_{k,n}^{\psi}\frac{\boldsymbol{W}_p^{\psi^{1/2}}}{\sqrt{L}}\sqrt{L}\boldsymbol{W}_p^{\psi^{-1/2}}\frac{1}{n}Q_{\mathrm{PQL}}^*(\boldsymbol{c}) \\ &+ o_p(1).\end{aligned}$$

Therefore, from Assumption 6, Lemma 12 and Slutsky's theorem, we conclude that as $n, L \to \infty$, with probability approaching one,

$$(\boldsymbol{N}^T(t)(\boldsymbol{H}_{k,n}^{\psi})^{-1}\boldsymbol{W}_p^{\psi}(\boldsymbol{H}_{k,n}^{\psi})^{-1}\boldsymbol{N}(t))^{-1/2}\sqrt{L}(\widetilde{\boldsymbol{\beta}}_{\mathrm{PQL}}(t) - \boldsymbol{\beta}(t)) \to \mathbb{N}(\boldsymbol{0}_2, \boldsymbol{I}_2).$$

■

**Proof** ***of Theorem 2.*** With Lemma 10 and 11, and the similar reasoning as the proof for Theorem 1, we can prove Theorem 2. So, the details are omitted. ■

**Proof** ***of Theorem 3.*** Note that

$$\begin{aligned}\mathrm{tr}(\boldsymbol{V}_p^{\psi}) &= \mathrm{tr}\left[\frac{1}{n^2}\sum_{i=1}^{n}\frac{\{y_i - \psi(\boldsymbol{N}_i^T\widehat{\boldsymbol{c}}_{\mathrm{PQL}})\}^2\boldsymbol{N}_i\boldsymbol{N}_i^T}{p_i}\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{tr}\left[\frac{\{y_i - \psi(\boldsymbol{N}_i^T\widehat{\boldsymbol{c}}_{\mathrm{PQL}})\}^2\boldsymbol{N}_i\boldsymbol{N}_i^T}{p_i}\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\frac{\{y_i - \psi(\boldsymbol{N}_i^T\widehat{\boldsymbol{c}}_{\mathrm{PQL}})\}^2\|\boldsymbol{N}_i\|_2^2}{p_i} \\ &= \frac{1}{n^2}(\sum_{i=1}^{n}p_i)\sum_{i=1}^{n}\frac{\{y_i - \psi(\boldsymbol{N}_i^T\widehat{\boldsymbol{c}}_{\mathrm{PQL}})\}^2\|\boldsymbol{N}_i\|_2^2}{p_i} \\ &\geq \frac{1}{n^2}\left\{\sum_{i=1}^{n}|y_i - \psi(\boldsymbol{N}_i^T\widehat{\boldsymbol{c}}_{\mathrm{PQL}})|\|\boldsymbol{N}_i\|_2\right\}^2,\end{aligned}$$

where the last step is from the Cauchy-Schwarz inequality and the equality holds if and only if when $p_i \propto |y_i - \psi(\boldsymbol{N}_i^T\widehat{\boldsymbol{c}}_{\mathrm{PQL}})|\|\boldsymbol{N}_i\|_2$. And, to satisfy $\sum_{i=1}^{n}p_i = 1$, we let $p_i = \frac{|y_i - \psi(\boldsymbol{N}_i^T\widehat{\boldsymbol{c}}_{\mathrm{PQL}})|\|\boldsymbol{N}_i\|_2}{\sum_{i=1}^{n}|y_i - \psi(\boldsymbol{N}_i^T\widehat{\boldsymbol{c}}_{\mathrm{PQL}})|\|\boldsymbol{N}_i\|_2}$. ■

**Proof** ***of Theorem 4.*** Following the proof of Lemma 12, when $p_i = p_{\mathrm{PQL},i}^{\mathrm{FLoS},\widehat{\boldsymbol{c}}^0_{\mathrm{PQL}}}$, we have

$$\begin{aligned}
\boldsymbol{V}_p^{\psi} &= \frac{1}{n^2}\sum_{i=1}^{n}\frac{(y_i-\psi(\boldsymbol{N}_i^T\widehat{\boldsymbol{c}}_{\mathrm{PQL}}))^2\boldsymbol{N}_i\boldsymbol{N}_i^T}{p_i}\\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{\{y_i-\psi(\boldsymbol{N}_i^T\widehat{\boldsymbol{c}}_{\mathrm{PQL}})\}^2\boldsymbol{N}_i\boldsymbol{N}_i^T}{|y_i-\psi(\boldsymbol{N}_i^T\widehat{\boldsymbol{c}}^0_{\mathrm{PQL}})|\|\boldsymbol{N}_i\|_2}\times\frac{1}{n}\sum_{i=1}^{n}|y_i-\psi(\boldsymbol{N}_i^T\widehat{\boldsymbol{c}}^0_{\mathrm{PQL}})|\|\boldsymbol{N}_i\|_2\\
&\equiv \widetilde{\boldsymbol{V}}_1^{\psi}\times\widetilde{\boldsymbol{V}}_2^{\psi}.
\end{aligned}$$

Now we want to prove that $\widetilde{\boldsymbol{V}}_1^{\psi}-\boldsymbol{V}_1^{\psi}=o_p(1)$ and $\widetilde{\boldsymbol{V}}_2^{\psi}-\boldsymbol{V}_2^{\psi}=o_p(1)$, where $\boldsymbol{V}_1^{\psi}$ and $\boldsymbol{V}_2^{\psi}$ have the same expression of $\widetilde{\boldsymbol{V}}_1^{\psi}$ and $\widetilde{\boldsymbol{V}}_2^{\psi}$, respectively, except that $y_i-\psi(\boldsymbol{N}_i^T\widehat{\boldsymbol{c}}^0_{\mathrm{PQL}})$ is replaced by $y_i-\psi(\boldsymbol{N}_i^T\widehat{\boldsymbol{c}}_{\mathrm{PQL}})$. Combining $\widehat{\boldsymbol{c}}_{\mathrm{PQL}}-\boldsymbol{c}=o_p(1)$ with $\widehat{\boldsymbol{c}}^0_{\mathrm{PQL}}-\boldsymbol{c}=o_p(1)$, we can get $\widehat{\boldsymbol{c}}^0_{\mathrm{PQL}}-\widehat{\boldsymbol{c}}_{\mathrm{PQL}}=o_p(1)$, thus, $|\psi(\boldsymbol{N}_i^T\widehat{\boldsymbol{c}}^0_{\mathrm{PQL}})-\psi(\boldsymbol{N}_i^T\widehat{\boldsymbol{c}}_{\mathrm{PQL}})|=o_p(1)$ for each $i$. Therefore, $\mathrm{E}\left\{|\psi(\boldsymbol{N}_i^T\widehat{\boldsymbol{c}}^0_{\mathrm{PQL}})-\psi(\boldsymbol{N}_i^T\widehat{\boldsymbol{c}}_{\mathrm{PQL}})|\right\}\to 0$ and $\frac{|y_i-\psi(\boldsymbol{N}_i^T\widehat{\boldsymbol{c}}_{\mathrm{PQL}})|}{|y_i-\psi(\boldsymbol{N}_i^T\widehat{\boldsymbol{c}}^0_{\mathrm{PQL}})|}\le C_1$, where $C_1$ is a positive constant. Note that

$$\begin{aligned}
|\widetilde{\boldsymbol{V}}_1^{\psi}-\boldsymbol{V}_1^{\psi}| &= \frac{1}{n}\left|\sum_{i=1}^{n}\frac{\{y_i-\psi(\boldsymbol{N}_i^T\widehat{\boldsymbol{c}}_{\mathrm{PQL}})\}^2\boldsymbol{N}_i\boldsymbol{N}_i^T}{|y_i-\psi(\boldsymbol{N}_i^T\widehat{\boldsymbol{c}}^0_{\mathrm{PQL}})|\|\boldsymbol{N}_i\|_2}-\sum_{i=1}^{n}\frac{\{y_i-\psi(\boldsymbol{N}_i^T\widehat{\boldsymbol{c}}_{\mathrm{PQL}})\}\boldsymbol{N}_i\boldsymbol{N}_i^T}{\|\boldsymbol{N}_i\|_2}\right|\\
&\le \frac{1}{n}\sum_{i=1}^{n}\left|\frac{(y_i-\psi(\boldsymbol{N}_i^T\widehat{\boldsymbol{c}}_{\mathrm{PQL}}))(\psi(\boldsymbol{N}_i^T\widehat{\boldsymbol{c}}^0_{\mathrm{PQL}})-\psi(\boldsymbol{N}_i^T\widehat{\boldsymbol{c}}_{\mathrm{PQL}}))\boldsymbol{N}_i\boldsymbol{N}_i^T}{|y_i-\boldsymbol{N}_i^T\widehat{\boldsymbol{c}}^0|\|\boldsymbol{N}_i\|_2}\right|\\
&\le \frac{C_1}{n}\sum_{i=1}^{n}\left|\frac{(\psi(\boldsymbol{N}_i^T\widehat{\boldsymbol{c}}^0_{\mathrm{PQL}})-\psi(\boldsymbol{N}_i^T\widehat{\boldsymbol{c}}_{\mathrm{PQL}}))\boldsymbol{N}_i\boldsymbol{N}_i^T}{\|\boldsymbol{N}_i\|_2}\right|\\
&\le \frac{C_1}{n}\sum_{i=1}^{n}|(\psi(\boldsymbol{N}_i^T\widehat{\boldsymbol{c}}^0_{\mathrm{PQL}})-\psi(\boldsymbol{N}_i^T\widehat{\boldsymbol{c}}_{\mathrm{PQL}}))|\left|\frac{\boldsymbol{N}_i\boldsymbol{N}_i^T}{\|\boldsymbol{N}_i\|_2}\right|.
\end{aligned}$$

And

$$\begin{aligned}
|\widetilde{\boldsymbol{V}}_2^{\psi}-\boldsymbol{V}_2^{\psi}| &= \frac{1}{n}\left|\sum_{i=1}^{n}|y_i-\psi(\boldsymbol{N}_i^T\widehat{\boldsymbol{c}}^0_{\mathrm{PQL}})|\|\boldsymbol{N}_i\|_2-\sum_{i=1}^{n}|y_i-\psi(\boldsymbol{N}_i^T\widehat{\boldsymbol{c}}_{\mathrm{PQL}})|\|\boldsymbol{N}_i\|_2\right|\\
&\le \frac{1}{n}\sum_{i=1}^{n}\left|(\psi(\boldsymbol{N}_i^T\widehat{\boldsymbol{c}}^0_{\mathrm{PQL}})-\psi(\boldsymbol{N}_i^T\widehat{\boldsymbol{c}}_{\mathrm{PQL}}))\|\boldsymbol{N}_i\|_2\right|.
\end{aligned}$$

For any $\epsilon>0$, by Chebyshev's inquality

$$\begin{aligned}
&P\left\{\frac{1}{n}\sum_{i=1}^{n}|\psi(\boldsymbol{N}_i^T\widehat{\boldsymbol{c}}^0_{\mathrm{PQL}})-\psi(\boldsymbol{N}_i^T\widehat{\boldsymbol{c}}_{\mathrm{PQL}})|\left|\frac{\boldsymbol{N}_i\boldsymbol{N}_i^T}{\|\boldsymbol{N}_i\|_2}\right|>\epsilon\right\}\\
\le&\frac{1}{\epsilon n}\sum_{i=1}^{n}\mathrm{E}\left\{|\psi(\boldsymbol{N}_i^T\widehat{\boldsymbol{c}}^0_{\mathrm{PQL}})-\psi(\boldsymbol{N}_i^T\widehat{\boldsymbol{c}}_{\mathrm{PQL}})|\right\}\left|\frac{\boldsymbol{N}_i\boldsymbol{N}_i^T}{\|\boldsymbol{N}_i\|_2}\right|\\
\to&0.
\end{aligned}$$

and

$$\begin{aligned}
&P\left\{\frac{1}{n}\sum_{i=1}^{n}|\psi(\boldsymbol{N}_i^T\widehat{\boldsymbol{c}}_{\mathrm{PQL}}^0)-\psi(\boldsymbol{N}_i^T\widehat{\boldsymbol{c}}_{\mathrm{PQL}})\|\boldsymbol{N}_i\|_2>\epsilon\right\}\\
\leq&\frac{1}{\epsilon n}\sum_{i=1}^{n}\mathrm{E}\left\{|\psi(\boldsymbol{N}_i^T\widehat{\boldsymbol{c}}_{\mathrm{PQL}}^0)-\psi(\boldsymbol{N}_i^T\widehat{\boldsymbol{c}}_{\mathrm{PQL}})|\right\}\|\boldsymbol{N}_i\|_2\\
\to&0.
\end{aligned}$$

Thus, $\widetilde{\boldsymbol{V}}_1^{\psi}-\boldsymbol{V}_1=o_p(1)$, $\widetilde{\boldsymbol{V}}_2^{\psi}-\boldsymbol{V}_2^{\psi}=o_p(1)$ and $\widetilde{\boldsymbol{V}}_1^{\psi}\times\widetilde{\boldsymbol{V}}_2-\boldsymbol{V}_1^{\psi}\times\boldsymbol{V}_2^{\psi}=\widetilde{\boldsymbol{V}}_1^{\psi}\times\widetilde{\boldsymbol{V}}_2^{\psi}-\boldsymbol{V}_{\mathrm{FLoS}}^{\psi}=o_p(1)$.

Next, using a similar proof to the proof for Theorem 2, we can finish the proof of Theorem 4. ■

## References


Mingyao Ai, Jun Yu, Huiming Zhang, and Haiying Wang. Optimal subsampling algorithms for big data regressions. *Statistica Sinica*, 31(2):749–772, 2021.

Anthony Atkinson, Alexander Donev, and Randall Tobias. *Optimum Experimental Designs, with SAS*, volume 34. Oxford University Press, New York, 2007.

David L. Barrow and Philip W. Smith. Asymptotic properties of best $L_2[0,1]$ approximation by splines with variable knots. *Quarterly of applied mathematics*, 36(3):293–304, 1978.

Xiong Cai, Liugen Xue, and Jiguo Cao. Variable selection for multiple function-on-function linear regressionl. *Statistica Sinica*, 32(4):1–43, 2021a.

Xiong Cai, Liugen Xue, and Jiguo Cao. Robust penalized m-estimation for function-on-function linear regression. *Stat*, 10:e390, 2021b.

Xiong Cai, Liugen Xue, and Jiguo Cao. Robust estimation and variable selection for function-on-scalar regression. *Canadian Journal of Statistics*, 50(1):162–179, 2022.

Hervé Cardot and Pacal Sarda. Estimation in generalized linear models for functional data via penalized likelihood. *Journal of Multivariate Analysis*, 92(1):24–41, 2005.

Hervé Cardot, Frédéric Ferraty, and Pascal Sarda. Spline estimators for the functional linear model. *Statistica Sinica*, 13:571–591, 2003.

Raymond J. Carroll, Jianqing Fan, Irene Gijbels, and Matt P. Wand. Generalized partially linear single-index models. *Journal of the American Statistical Association*, 92(438): 477–489, 1997.

Kani Chen, Inchi Hu, and Zhiliang Ying. Strong consistency of maximum quasi-likelihood estimators in generalized linear models with fixed and adaptive designs. *The Annals of Statistics*, 27(4):1155–1163, 1999.

Qianshun Cheng, Haiying Wang, and Min Yang. Information-based optimal subdata selection for big data logistic regression. *Journal of Statistical Planning and Inference*, 209:112–122, 2020.

Gerda Claeskens, Tatyana Krivobokova, and Jean D. Opsomer. Asymptotic properties of penalized spline estimators. *Biometrika*, 96(3):529–544, 2009.

Ciprian M. Crainiceanu, Ana-Maria Staicu, and Chong-Zhi Di. Generalized multilevel functional regression. *Journal of the American Statistical Association*, 104(488): 1550–1561, 2009.

Carl de Boor. *A Practical Guide to Splines*. Springer, New York, 1978.

Jianghu (James) Dong, Liangliang Wang, Jiguo Cao, and Jagbir Gill. Functional principal component analysis of gfr curves after kidney transplant. *Statistical Methods in Medical Research*, 27(12):3785–3796, 2018.

Jianghu (James) Dong, Shijia Wang, Liangliang Wang, Jagbir Gill, and Jiguo Cao. Joint modelling for organ transplantation outcomes for patients with diabetes and the end-stage renal disease. *Statistical Methods in Medical Research*, 28(9):2724–2737, 2019.

Jianghu (James) Dong, Jiguo Cao, Jagbir Gill, Clifford Miles, and Troy Plumb. Functional joint models for chronic kidney disease in kidney transplant recipients. *Statistical Methods in Medical Research*, 30(8):1932–1943, 2021.

Jianghu (James) Dong, Haolun Shi, Liangliang Wang, Ying Zhang, and Jiguo Cao. Jointly modelling multiple transplant outcomes by a competing risk model via functional principal component analysis. *Journal of Applied Statistics*, 50(1):43–59, 2023.

Gareth M. James. Generalized linear models with functional predictors. *Journal of the Royal Statistical Society: Series B*, 64(3):411–432, 2002.

Farrah M. Keong, Yama A. Afshar, Stephen O. Pastan, Ritam Chowdhury, Jose N. Binongo, and Rachel E. Patzer. Decreasing estimated glomerular filtration rate is associated with increased risk of hospitalization after kidney transplantation. *Kidney International Reports*, 1(4):269–278, 2016.

Myungjin Kim and Li Wang. Generalized spatially varying coefficient models. *Journal of Computational and Graphical Statistics*, 30(1):1–10, 2021.

Andrew S. Levey, Lesley A. Stevens, Christopher H. Schmid, Yaping Zhang, Alejandro F. Castro III, Harold I. Feldman, John W. Kusek, Paul Eggers, Frederick Van Lente, Tom Greene, Josef Coresh, and CKD-EPI (Chronic Kidney Disease Epidemiology

Collaboration). A new equation to estimate glomerular filtration rate. *Annals of Internal Medicine*, 150(9):604–612, 2009.

Jialiang Li, Chao Huang, and Hongtu Zhu. A functional varying-coefficient single-index model for functional response data. *Journal of the American Statistical Association*, 112(519):1169–1181, 2017.

Rong Liu, Lijian Yang, and Wolfgang K. Härdle. Oracally efficient two-step estimation of generalized additive model. *Journal of the American Statistical Association*, 108 (502):619–631, 2013.

Xinchao Luo, Lixing Zhu, and Hongtu Zhu. Single-index varying coefficient model for functional responses. *Biometrics*, 72(4):1275–1284, 2016.

Ping Ma, Michael W. Mahoney, and Bin Yu. A statistical perspective on algorithmic leveraging. *The Journal of Machine Learning Research*, 16(1):861–911, 2015.

Ping Ma, Yongkai Chen, Xinlian Zhang, Xin Xing, Jingyi Ma, and Michael Mahoney. Asymptotic analysis of sampling estimators for randomized linear algebra algorithms. *The Journal of Machine Learning Research*, 23:1–45, 2022.

Mathew W. McLean, Giles Hooker, Ana-Maria Staicu, Fabian Scheipl, and David Ruppert. Functional generalized additive models. *Journal of Computational and Graphical Statistics*, 23(1):249–269, 2014.

Jeffrey S. Morris. Functional regression. *Annual Review of Statistics and Its Application*, 2(1):321–359, 2015.

Hans-Georg Müller and Ulrich Stadtmüller. Generalized functional linear models. *Annals of Statistics*, 33(2):774–805, 2005.

Friedrich Pukelsheim. *Optimal Design of Experiments*. Wiley, New York, 2006.

James O. Ramsay and Bernard W. Silverman. *Applied Functional Data Analysis*. Springer, New York, 2002.

Larry Schumaker. *Spline Functions: Basic Theory*. Wiley, New York, 1981.

George J. Schwartz, Alvaro Munoz, Michael F. Schneider, Robert H. Mak, Frederick Kaskel, Bradley A. Warady, and Susan L. Furth. New equations to estimate gfr in children with CKD. *Journal of the American Society of Nephrology*, 20(3):629–637, 2009.

Haolun Shi, Jianghu (James) Dong, Liangliang Wang, and Jiguo Cao. Functional principal component analysis for longitudinal data with informative dropout. *Statistics in Medicine*, 40:712—-724, 2021.

Xiaoxiao Sun, Pang Du, Xiao Wang, and Ping Ma. Optimal penalized function-on-function regression under a reproducing kernel hilbert space framework. *Journal of the American Statistical Association*, 113(524):1601–1611, 2018.

Haiying Wang, Rong Zhu, and Ping Ma. Optimal subsampling for large sample logistic regression. *Journal of the American Statistical Association*, 113(522):829–844, 2018.

Haiying Wang, Min Yang, and John Stufken. Information-based optimal subdata selection for big data linear regression. *Journal of the American Statistical Association*, 114 (525):393–405, 2019.

Jane-Ling Wang, Jeng-Min Chiou, and Hans-Georg Müller. Functional data analysis. *Annual Review of Statistics and Its Application*, 3:257–295, 2016.

Li Wang and Guanqun Cao. Efficient estimation for generalized partially linear single-index models. *Bernoulli*, 24(2):1101–1127, 2018.

Luo Xiao. Asymptotic theory of penalized splines. *Electronic Journal of Statistics*, 13(1): 747–794, 2019.

Yaqiong Yao and Haiying Wang. A review on optimal subsampling methods for massive datasets. *Journal of Data Science*, 19(1):1–22, 2021.

Jun Yu, Haiying Wang, Mingyao Ai, and Huiming Zhang. Optimal distributed subsampling for maximum quasi-likelihood estimators with massive data. *Journal of the American Statistical Association*, 117(537):265–276, 2022.

Shan Yu, Guannan Wang, Li Wang, Chenhui Liu, and Lijian Yang. Estimation and inference for generalized geoadditive models. *Journal of the American Statistical Association*, 115(530):761–774, 2020.

Shanggang Zhou, Xiaotong Shen, and Douglas A. Wolfe. Local asymptotics for regression splines and confidence regions. *The Annals of Statistics*, 26(5):1760–1782, 1998.

Hongtu Zhu, Runze Li, and Linglong Kong. Multivariate varying coefficient model for functional responses. *The Annals of Statistics*, 40(5):2634–2666, 2012.