Statistica Sinica Preprint No: SS-2023-0100	
Title	Estimation and Variable Selection under the Function-on-scalar Linear Model with Covariate Measurement Error
Manuscript ID	SS-2023-0100
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202023.0100
Complete List of Authors	Yifan Sun and Grace Y. Yi
Corresponding Authors	Grace Y. Yi
E-mails	gyi5@uwo.ca
Notice: Accepted version subject to English editing.	

Statistica Sinica

ESTIMATION AND VARIABLE SELECTION UNDER THE FUNCTION-ON-SCALAR LINEAR MODEL WITH COVARIATE MEASUREMENT ERROR

Yifan Sun[1], Grace Y. Yi[1,2,*]

[1]*Department of Statistical and Actuarial Sciences, University of Western Ontario, Canada*

[2]*Department of Computer Science, University of Western Ontario, Canada*

Abstract:

Function-on-scalar linear regression has been widely used to model the relationship between a functional response and multiple scalar covariates. Its utility is, however, challenged by the presence of measurement error, a ubiquitous feature in applications. Naively applying usual function-on-scalar linear regression to error-contaminated data often yields biased inference results. Further, estimation of the model parameters is complicated by the presence of inactive variables, especially when handling data with a large dimension. Building parsimonious and interpretable function-on-scalar linear regression models is in urgent demand to handle error-contaminated functional data. In this paper, we study this important problem and investigate the measurement error effects. We propose a debiased loss function, combined with a sparsity-inducing penalty function, to

*Corresponding author: Grace Y. Yi, gyi5@uwo.ca

simultaneously estimate functional coefficients and select salient predictors. An efficient computing algorithm is developed with tuning parameters determined by data-driven methods. Under mild conditions, the asymptotic properties of the proposed estimator are rigorously established, including estimation consistency, selection consistency, and the limiting distributions. The finite sample performance of the proposed method is assessed through extensive simulation studies, and the usage of the proposed method is illustrated by a real data application.

Key words and phrases: functional data analysis, function-on-scalar regression, measurement error, variable selection.

1. Introduction

Functional data analysis has attracted extensive attention in the last two decades (e.g., Ramsay and Silverman (2005); Horváth and Kokoszka (2012)). Typically, the function-on-scalar linear regression model (Ramsay and Silverman, 2005, Chapter 13) has been proven to be useful to describe the relationship between a functional response and multiple scalar covariates, and many methods have been proposed for inference about the coefficient functions of this model. To name a few, see Chiang et al. (2001), Ramsay and Silverman (2005, Section 13.4), Zhang and Chen (2007); Zhu et al. (2012), and the references therein.

Recently, research on variable selection has become increasingly inter-

esting, which is paramount in the era of big data. With data of a large dimension, usually only a small number of variables have the effects on explaining the change of the functional response and others have no explanatory effects. Excluding those inactive covariates is mandatory to build a parsimonious and interpretable model and conduct valid inference accordingly. Based on the widely-used methods developed for scalar response regression models, such as group LASSO (Yuan and Lin, 2006), adaptive LASSO (Zou, 2006), SCAD regularization (Fan and Li, 2001), and MCP regularization (Zhang, 2010), several methods have been proposed to handle variable selection for functional response models. Wang et al. (2007) proposed a group SCAD estimator and applied it to analyze gene expression data. Taking the within-subject correlation into consideration, Chen et al. (2016) introduced a group MCP procedure, combinded with the generalized least squares technique. Barber et al. (2017) and Fan and Reimherr (2017) extended group LASSO and group adaptive LASSO respectively to the high dimensional function-on-scalar model. Parodi and Reimherr (2018) developed a functional linear adaptive mixed estimation procedure to simultaneously obtain sparsity and smoothness for the coefficient functions. Cai et al. (2022) considered a robust selection approach with the exponential squared loss function employed.

While these methods can conduct variable selection for functional data analysis, they have a serious limitation for handling real data in applications. In reality, collected data commonly involve measurement error, a ubiquitous feature that has been a long standing concern in various fields, including medical research, epidemiological studies, nutritional studies, and cancer research (e.g., Yi and Cook (2005)). It is well known that naively ignoring measurement errors may lead to inconsistent estimators. A comprehensive account of measurement error can be found in Carroll et al. (2006), Yi (2017), and Yi et al. (2021), among others.

There have been only a few papers dealing with functional data with error-contaminated covariates. Zhu et al. (2019), Zhu et al. (2020), and Meng et al. (2021) investigated estimation under partially functional linear models, semi-functional partially linear models, and functional partially linear single index models, respectively, with additive measurement error models considered. Jiang et al. (2021) studied partially functional linear models where the variables are distorted by some multiplicative factors. However, the response variable for all those methods is merely a scalar but not a functional variable. Further, none of these methods discuss variable selection. Recently, Zhao et al. (2022) studied variable selection for a longitudinal varying coefficient model with a sparsely observed longitudinal

response and time-varying contaminated variables. They assumed that the components of measurement errors are uncorrelated and did not provide the asymptotic distribution for their proposed estimator.

In this paper, we consider this notable problem concerning functional data analysis for data with both measurement error and unimportant variables. Our contributions are multifaceted. First, we reveal the effects of covariate measurement error on the function-on-scalar linear regression analysis, and uncover an interesting connection of such effects with the usual ridge regression method for error-free settings. Secondly, we propose a method to conduct estimation and variable selection simultaneously with covariate measurement error effects fully accounted for. Our development encompasses both the least squares and generalized least squares loss functions. Theoretical properties are rigorously established for the proposed method. Finally, we develop an easy-to-implement computing algorithm and discuss two data-driven tuning parameter selection methods. The proposed procedure yields consistent estimators for the model coefficients and successfully detects all the active or inactive covariates in the presence of measurement error. To the best of our knowledge, this is the first work about regression models with the functional response and multiple covariates, which involve both measurement error and inactive variables. This research adds new

dimensions to the existing framework of classical function-on-linear models to accommodate error-contaminated data with irrelevant variables.

The rest of this article is organized as follows. Section 2 presents the basic model setup and a commonly used estimation method for ideal settings without measurement error. In Section 3, we investigate the measurement error effects and develop a valid method. A computational algorithm, together with tuning parameters selection methods, is presented in Section 4. Theoretical results are rigorously established in Section 5. In Section 6, we utilize our method to analyze a daily activity dataset, with the associated physiological, environmental and behavioral covariates included. Concluding remarks are placed in the final section. The technical details and the numerical studies are deferred to the online supplementary material, where we also include the extended development with both theoretical and numerical results.

2. Function-on-Scalar Model

2.1 Model and Notation

For subject i, let $Y_i(t)$ denote the functional response at time $t \in \mathcal{T}$ and let $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^\intercal$ represent an associated random vector of p-dimensional covariates, where $\mathcal{T} \subset \mathbb{R}$ is a compact set. Without loss of gen-

erality, $\boldsymbol{X}_i$ and $Y_i(t)$ are assumed to be centered so that the mean of $\boldsymbol{X}_i$ and of $Y_i(t)$ are both zero. Suppose that for $t \in \mathcal{T}$, $\{\{\boldsymbol{X}_i, Y_i(t)\} : i = 1, 2, \ldots, n\}$ is a random sample of n independent and identically distributed (i.i.d) random variables.

Consider the function-on-scalar regression model

$$Y_i(t) = \boldsymbol{\beta}(t)^\mathsf{T} \boldsymbol{X}_i + \varepsilon_i(t) \tag{2.1}$$

for $i = 1, 2, \ldots, n$ and $t \in \mathcal{T}$, where $\boldsymbol{\beta}(t) = (\beta_1(t), \ldots, \beta_p(t))^\mathsf{T}$ is the vector of unknown parameter functions, and $\varepsilon_i(t)$ is the random error process independent of $\boldsymbol{X}_i$ and satisfies that $\mathbb{E}\{\varepsilon_i(t)\} = 0$ for all $t \in \mathcal{T}$ and $\sup_{t\in\mathcal{T}} \mathbb{E}[\{\varepsilon_i(t)\}^2] < \infty$.

In practice, functional data $\{Y_i(t) : t \in \mathcal{T}\}$ are usually observed on a discrete grid. For ease of exposition, we assume that all the subjects in the sample are observed on a common grid $\boldsymbol{t} = \{t_1, \ldots, t_m\} \subset \mathcal{T}$, where the number m of grid points satisfies Assumption C3 in Section S1 of the supplementary material. Let $\boldsymbol{Y}_i(\boldsymbol{t}) = (Y_i(t_1), \ldots, Y_i(t_m))^\mathsf{T}$ and $\boldsymbol{\varepsilon}_i(\boldsymbol{t}) = (\varepsilon_i(t_1), \ldots, \varepsilon_i(t_m))^\mathsf{T}$. Then model (2.1) gives that

$$\boldsymbol{Y}_i(\boldsymbol{t}) = \boldsymbol{\beta}(\boldsymbol{t})^\mathsf{T} \boldsymbol{X}_i + \boldsymbol{\varepsilon}_i(\boldsymbol{t}), \tag{2.2}$$

where $\boldsymbol{\beta}(\boldsymbol{t}) = [\boldsymbol{\beta}(t_1), \cdots, \boldsymbol{\beta}(t_m)]$ is a $p \times m$ matrix of unknown quantities.

2.2 Least Squares Estimation

The least squares method may be considered to estimate the unknown coefficient functions $\boldsymbol{\beta}(t)$ due to the additive structure of model (2.1). However, owing to the infinite dimensionality, $\boldsymbol{\beta}(t)$ cannot be estimated without regularizations. To reduce the dimension of $\boldsymbol{\beta}(t)$, we employ the B-spline approximation (De Boor, 1978) as the basis to carry out our following development.

Let $\tau_1 < \tau_2 < \cdots < \tau_{M+1}$ denote $M+1$ knots on $\mathcal{T}$ with τ_1 and τ_{M+1} representing the two endpoints of $\mathcal{T}$, where M is chosen in light of both the sample size n and the number m of the observation points, as required in Assumption C in Section S1 of the supplementary material. Let $d+1$ denote the order of the B-spline function, where $d = 3$ is often considered in applications as well as in this paper. We then have the corresponding B-spline basis functions of order $d+1$ (De Boor, 1978, Chapter IX), denoted $\boldsymbol{\phi}(t) = (\phi_1(t), \ldots, \phi_{M+d}(t))^{\mathsf{T}}$, $t \in \mathcal{T}$.

For $j = 1, \ldots, p$ and $t \in \mathcal{T}$, we approximate the function $\beta_j(t)$ by a linear combination of $\boldsymbol{\phi}(t)$:

$$\beta_j(t) \approx \sum_{k=1}^{M+d} b_{jk}\phi_k(t) \triangleq \boldsymbol{b}_j^{\mathsf{T}}\boldsymbol{\phi}(t), \tag{2.3}$$

where $\boldsymbol{b}_j = (b_{j1}, \ldots, b_{j,M+d})^{\mathsf{T}}$ is the vector of coefficients to be determined.

With (2.3), the estimation of $\beta_j(t)$ now becomes the estimation of $\boldsymbol{b}_j$. Let $\boldsymbol{\Phi}(\boldsymbol{t}) = [\boldsymbol{\phi}(t_1), \cdots, \boldsymbol{\phi}(t_m)]^\mathsf{T}$ be the $m \times (M+d)$ matrix, and let $\boldsymbol{b} = (\boldsymbol{b}_1^\mathsf{T}, \cdots, \boldsymbol{b}_p^\mathsf{T})^\mathsf{T}$ be the $(M+d)p \times 1$ vector. Thus, the $m \times p$ matrix $\boldsymbol{\beta}(\boldsymbol{t})^\mathsf{T}$ in (2.2) can be approximated by $\boldsymbol{\Phi}(\boldsymbol{t})[\boldsymbol{b}_1, \cdots, \boldsymbol{b}_p]$. Neglecting the approximation errors, the model (2.2) can be rewritten as

$$\begin{aligned} \boldsymbol{Y}_i(\boldsymbol{t}) &\approx \boldsymbol{\Phi}(\boldsymbol{t})\big[\boldsymbol{b}_1, \cdots, \boldsymbol{b}_p\big]\boldsymbol{X}_i + \boldsymbol{\varepsilon}_i(\boldsymbol{t}) \\ &= (\boldsymbol{X}_i^\mathsf{T} \otimes \boldsymbol{\Phi}(\boldsymbol{t}))\boldsymbol{b} + \boldsymbol{\varepsilon}_i(\boldsymbol{t}), \end{aligned} \tag{2.4}$$

where $\otimes$ represents the Kronecker product. Let $\boldsymbol{Y}(\boldsymbol{t}) = (\boldsymbol{Y}_1^\mathsf{T}(\boldsymbol{t}), \ldots, \boldsymbol{Y}_n^\mathsf{T}(\boldsymbol{t}))^\mathsf{T}$, $\boldsymbol{\varepsilon}(\boldsymbol{t}) = (\boldsymbol{\varepsilon}_1(\boldsymbol{t})^\mathsf{T}, \ldots, \boldsymbol{\varepsilon}_n(\boldsymbol{t})^\mathsf{T})^\mathsf{T}$ and $\boldsymbol{X} = [\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n]^\mathsf{T}$. The equation (2.4) then yields

$$\boldsymbol{Y}(\boldsymbol{t}) \approx (\boldsymbol{X} \otimes \boldsymbol{\Phi}(\boldsymbol{t}))\boldsymbol{b} + \boldsymbol{\varepsilon}(\boldsymbol{t}). \tag{2.5}$$

The form of (2.5) is similar to the classical linear regression model, although the components in $\boldsymbol{\varepsilon}(\boldsymbol{t})$ may be correlated. The most direct way to estimate unknown $\boldsymbol{b}$ is the least squares method by minimizing the loss function

$$\tilde{L}_n(\boldsymbol{b}; \boldsymbol{Y}(\boldsymbol{t}), \boldsymbol{X}) \triangleq \frac{1}{2} \left\| \boldsymbol{Y}(\boldsymbol{t}) - (\boldsymbol{X} \otimes \boldsymbol{\Phi}(\boldsymbol{t}))\boldsymbol{b} \right\|^2, \tag{2.6}$$

where $\|\boldsymbol{a}\|$ represents the L_2 Euclidean norm $\sqrt{a_1^2 + \cdots + a_q^2}$ of the vector $\boldsymbol{a} \triangleq (a_1, \ldots, a_q)^\mathsf{T}$.

The formulation (2.6) focuses on expressing the differences between the responses $\boldsymbol{Y}(\boldsymbol{t})$ and their approximate mean $(\boldsymbol{X} \otimes \boldsymbol{\Phi}(\boldsymbol{t}))\boldsymbol{b}$, offering us a sim-

ple way to perform inference about model parameters without facilitating the correlation structure for the error terms $\boldsymbol{\varepsilon}(\boldsymbol{t})$. Incorporating possible dependence among the components of $\boldsymbol{\varepsilon}(\boldsymbol{t})$ can be done by adding a weight matrix to (2.6) to form a generalized least squares loss function. Its development is carried out in a manner similar to the development here. We defer the details to Section S5 of the supplementary material.

3. Measurement Error and Variable Selection

While (2.6) can be utilized to estimate $\boldsymbol{b}$, its validity hinges on the condition that $\boldsymbol{X}_i$ is precisely measured. This condition, however, is commonly violated in practice (Yi and Cook, 2005). Furthermore, some covariates in $\boldsymbol{X}_i$ may have no effect on the response. In this section, we consider data with these features.

3.1 Measurement Error Effects

In applications, the covariate vector $\boldsymbol{X}_i$ is often subject to measurement error, and we can only observe its surrogate version, denoted $\boldsymbol{X}_i^*$. To facilitate possible differences between $\boldsymbol{X}_i^*$ and $\boldsymbol{X}_i$, we consider the widely used classical additive measurement error model:

$$\boldsymbol{X}_i^* = \boldsymbol{X}_i + \boldsymbol{U}_i, \tag{3.7}$$

where the measurement error vector $\boldsymbol{U}_i \triangleq (U_{i1}, \ldots, U_{ip})^{\mathsf{T}}$ is independent of $\{Y_i(t) : t \in \mathcal{T}\}$ and $\boldsymbol{X}_i$. Further, $\boldsymbol{U}_1, \ldots, \boldsymbol{U}_n$ are assumed to be independent and identically distributed with mean vector $\mathbf{0}$ and covariance matrix, say, $\boldsymbol{\Sigma}$. To include the cases where some covariates in $\boldsymbol{X}_i$ are precisely measured, we allow matrix $\boldsymbol{\Sigma}$ to contain zero block submatrices and not to be strictly positive definite. To highlight the idea, we focus on the case where $\boldsymbol{\Sigma}$ is taken as known. Estimation of $\boldsymbol{\Sigma}$ is possible when repeated measurements of covariates or validation data are available (Carroll et al., 2006), as investigated in Section S3.7 of the supplementary material.

When the covariates are contaminated with measurement error, simply replacing $\boldsymbol{X}$ in loss function $\tilde{L}_n(\boldsymbol{b}; \boldsymbol{Y}(\boldsymbol{t}), \boldsymbol{X})$ in (2.6) with $\boldsymbol{X}^*$ results in biased estimator, where $\boldsymbol{X}^* = [\boldsymbol{X}_1^*, \ldots, \boldsymbol{X}_n^*]^{\mathsf{T}}$ which equals $\boldsymbol{X} + \boldsymbol{U}$ with $\boldsymbol{U} \triangleq [\boldsymbol{U}_1, \ldots, \boldsymbol{U}_n]^{\mathsf{T}}$. Indeed,

$$\begin{aligned}\tilde{L}_n(\boldsymbol{b}; \boldsymbol{Y}(\boldsymbol{t}), \boldsymbol{X}^*) &= \frac{1}{2} \left\| \boldsymbol{Y}(\boldsymbol{t}) - (\boldsymbol{X}^* \otimes \boldsymbol{\Phi}(\boldsymbol{t}))\boldsymbol{b} \right\|^2 \\ &= \tilde{L}_n(\boldsymbol{b}; \boldsymbol{Y}(\boldsymbol{t}), \boldsymbol{X}) + \frac{1}{2} \boldsymbol{b}^{\mathsf{T}} \left[(\boldsymbol{U}^{\mathsf{T}}\boldsymbol{U}) \otimes \left\{ \boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t}) \boldsymbol{\Phi}(\boldsymbol{t}) \right\} \right] \boldsymbol{b} \\ &\quad - \left\{ \boldsymbol{Y}(\boldsymbol{t}) - (\boldsymbol{X} \otimes \boldsymbol{\Phi}(\boldsymbol{t}))\boldsymbol{b} \right\}^{\mathsf{T}} (\boldsymbol{U} \otimes \boldsymbol{\Phi}(\boldsymbol{t}))\boldsymbol{b}, \end{aligned} \tag{3.8}$$

where the second step utilizes model (3.7). By the assumptions that $\boldsymbol{U}$ is independent of $\boldsymbol{Y}(\boldsymbol{t})$ and $\boldsymbol{X}$, $\mathbb{E}(\boldsymbol{U}) = \mathbf{0}$, and $\mathbb{E}(\boldsymbol{U}^{\mathsf{T}}\boldsymbol{U}) = n\boldsymbol{\Sigma}$, taking the expectation of (3.8) with respect to the joint distribution for the associated

random variables leads to

$$\mathbb{E}\left\{\tilde{L}_n(\boldsymbol{b};\boldsymbol{Y}(\boldsymbol{t}),\boldsymbol{X}^*)\right\} = \mathbb{E}\left\{\tilde{L}_n(\boldsymbol{b};\boldsymbol{Y}(\boldsymbol{t}),\boldsymbol{X})\right\} + \frac{n}{2}\boldsymbol{b}^\mathsf{T}\left[\boldsymbol{\Sigma}\otimes\left\{\boldsymbol{\Phi}^\mathsf{T}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})\right\}\right]\boldsymbol{b}. \tag{3.9}$$

Calling $\tilde{L}_n(\boldsymbol{b};\boldsymbol{Y}(\boldsymbol{t}),\boldsymbol{X}^*)$ the *naive loss* function, we see that the expectation of the *naive loss* function, called the *naive risk*, is larger than the expectation of the original loss function, $\mathbb{E}\left\{\tilde{L}_n(\boldsymbol{b};\boldsymbol{Y}(\boldsymbol{t}),\boldsymbol{X})\right\}$, i.e., the *true risk* derived from using the true covariate $\boldsymbol{X}$, unless at $\boldsymbol{b}=\boldsymbol{0}$. It also implies that directly applying (2.6) to $\boldsymbol{X}^*$, together with $\boldsymbol{Y}(\boldsymbol{t})$, to conduct inferences breaks down if the measurement error effects are naively ignored.

Motivated by (3.9), we construct a debiased loss function by subtracting the quadratic term in (3.9) to alleviate the measurement error effects:

$$L_n(\boldsymbol{b};\boldsymbol{Y}(\boldsymbol{t}),\boldsymbol{X}^*) \triangleq \tilde{L}_n(\boldsymbol{b};\boldsymbol{Y}(\boldsymbol{t}),\boldsymbol{X}^*) - \frac{n}{2}\boldsymbol{b}^\mathsf{T}\left[\boldsymbol{\Sigma}\otimes\left\{\boldsymbol{\Phi}^\mathsf{T}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})\right\}\right]\boldsymbol{b}, \tag{3.10}$$

which satisfies, by (3.9), $\mathbb{E}\{L_n(\boldsymbol{b};\boldsymbol{Y}(\boldsymbol{t}),\boldsymbol{X}^*)\} = \mathbb{E}\{\tilde{L}_n(\boldsymbol{b};\boldsymbol{Y}(\boldsymbol{t}),\boldsymbol{X})\}$. Then minimizing $L_n(\boldsymbol{b};\boldsymbol{Y}(\boldsymbol{t}),\boldsymbol{X}^*)$ with respect to $\boldsymbol{b}$ produces a consistent estimator of $\boldsymbol{b}$, provided regularity conditions.

This derivation has two implications. First, it indicates that in the presence of covariate measurement error, directly working with (2.6) by replacing the true covariates with their surrogate measurements does not necessarily ensure a consistent estimator. Secondly, the second term in

the right-hand-side of (3.10) resembles the L_2 or ridge regression penalty if taking the tuning parameter to be $-n\boldsymbol{\Sigma} \otimes \{\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})\}$ were allowed. In contrast to the usual ridge penalty without the negative sign, we call this term *the specialized ridge penalty*. Then the least squares estimator derived from (2.6) based on $\boldsymbol{X}$ can be equivalently regarded as a ridge regression estimator derived from using the surrogate measurement $\boldsymbol{X}^*$ with *the specialized ridge penalty* term. From now on we may use $L_n(\boldsymbol{b})$ to represent $L_n(\boldsymbol{b}; \boldsymbol{Y}(\boldsymbol{t}), \boldsymbol{X}^*)$ for simplicity from time to time.

3.2 Variable Selection

Suppose that $\boldsymbol{X}_i$ contains unimportant components for predicting the outcome and that only a small subset of the components in $\boldsymbol{X}_i$ is important in explaining $Y_i(t)$ in model (2.1). Without loss of generality, we assume that the first s covariates are active, or equivalently, $\beta_j(t) = 0$ if $j \in J_0$, where $J_0 = \{s+1, \ldots, p\}$, and the cardinality $|J_0|$ can be very close to p, suggesting the sparsity of model (2.1) is present. Our goal is to estimate $\boldsymbol{\beta}(t)$ and detect those inactive variables $\{X_{ij} : j \in J_0\}$ which correspond to the collection of all zero functions $\{\beta_j(t) : j \in J_0\}$.

Based on the loss function (3.10), we consider the penalized objective

function:

$$Q_n(\boldsymbol{b}) = L_n(\boldsymbol{b}; \boldsymbol{Y}(\boldsymbol{t}), \boldsymbol{X}^*) + nm \sum_{j=1}^{p} P_\lambda(\|\boldsymbol{b}_j\|), \tag{3.11}$$

where $P_\lambda(\cdot)$ is a penalty function with the tuning parameter $\lambda \geq 0$. Here, the penalty function $P_\lambda(\cdot)$ applies to each vector $\boldsymbol{b}_j$ collectively as a group, rather than to each individual element of $\boldsymbol{b}_j$; the inclusion m in the penalty term reflects that $\boldsymbol{Y}(\boldsymbol{t})$ in the loss function $L_n(\boldsymbol{b}; \boldsymbol{Y}(\boldsymbol{t}), \boldsymbol{X}^*)$ is m-dimensional. Thereby, an estimator of $\boldsymbol{b}$ is given by

$$\hat{\boldsymbol{b}} = \underset{\boldsymbol{b} \in \mathbb{R}^{p(M+d)}}{\operatorname{argmin}} \; Q_n(\boldsymbol{b}), \tag{3.12}$$

which is written as $\hat{\boldsymbol{b}} = (\hat{\boldsymbol{b}}_1^\mathsf{T}, \ldots, \hat{\boldsymbol{b}}_p^\mathsf{T})^\mathsf{T}$. We then take

$$\hat{\beta}_j(t) = \hat{\boldsymbol{b}}_j^\mathsf{T} \boldsymbol{\phi}(t) \tag{3.13}$$

as an estimator of $\beta_j(t)$ for $j = 1, \ldots, p$, and take $\hat{J}_0 = \{j : \hat{\boldsymbol{b}}_j = 0\}$ to be an estimate of J_0.

4. Computational Issues and Tuning Parameters Selection

In this section we develop an algorithm to minimize objective function (3.11) and propose methods to select tuning parameters. For ease of computation, we slightly modify the loss function $L_n(\boldsymbol{b})$ to obtain a strictly convex function.

First, $L_n(\boldsymbol{b})$ in (3.10) can be re-written as

$$L_n(\boldsymbol{b}) = \frac{1}{2}\boldsymbol{b}^{\mathsf{T}}\left[(\boldsymbol{X}^{*\mathsf{T}}\boldsymbol{X}^{*} - n\boldsymbol{\Sigma}) \otimes \{\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})\}\right]\boldsymbol{b} - \boldsymbol{b}^{\mathsf{T}}\left(\boldsymbol{X}^{*} \otimes \boldsymbol{\Phi}(\boldsymbol{t})\right)^{\mathsf{T}}\boldsymbol{Y}(\boldsymbol{t}) + \frac{1}{2}\boldsymbol{Y}(\boldsymbol{t})^{\mathsf{T}}\boldsymbol{Y}(\boldsymbol{t}), \tag{4.14}$$

which is a quadratic function of $\boldsymbol{b}$. Let $\boldsymbol{W} \triangleq \partial^2 L_n(\boldsymbol{b})/\partial\boldsymbol{b}\partial\boldsymbol{b}^{\mathsf{T}}$. It is easily seen that $\boldsymbol{W} = (\boldsymbol{X}^{*\mathsf{T}}\boldsymbol{X}^{*} - n\boldsymbol{\Sigma}) \otimes \{\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})\}$.

While $\boldsymbol{X}^{*\mathsf{T}}\boldsymbol{X}^{*}$ and $\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})$ are both positive definite, as assumed in Assumptions A2 and C3 in Section S1 of the supplementary material, unfortunately, $\boldsymbol{W}$ may be indefinite because the matrix $\boldsymbol{X}^{*\mathsf{T}}\boldsymbol{X}^{*} - n\boldsymbol{\Sigma}$ is not necessarily guaranteed to be positive definite for any sample, which is concretely shown in Section S3 of the supplementary material. Because of this, $L_n(\boldsymbol{b})$ may not be lower bounded, yielding the resultant estimator to be inconsistent. To overcome this issue, following the idea of Datta and Zou (2017), we replace the matrix $\boldsymbol{W}$ by its "nearest" positive definite matrix $\overline{\boldsymbol{W}}$, defined by the matrix projection. Specifically, for a small positive pre-determined threshold parameter τ, let

$$\mathscr{W}_\tau = \{\boldsymbol{W}_1 : \boldsymbol{W}_1 - \tau\boldsymbol{I} \text{ is semi-positive definite}\}.$$

Here and elsewhere, $\boldsymbol{I}$ represents an identity matrix whose dimension is indicated by the context. Define

$$\overline{\boldsymbol{W}} = \underset{\boldsymbol{W}_1 \in \mathscr{W}_\tau}{\operatorname{argmin}} \|\boldsymbol{W} - \boldsymbol{W}_1\|_{\max} \tag{4.15}$$

to be the matrix projection of $\boldsymbol{W}$, where $\|\cdot\|_{\max}$ denotes the maximum of the absolute values of all matrix entries. An algorithm for solving (4.15) is provided in Appendix of Datta and Zou (2017). The determination of $\overline{\boldsymbol{W}}$ is not very sensitive to the choice of τ, as shown in Section S3.5 of the online supplementary material.

Consequently, we consider a modified loss function for (4.14) by replacing $\boldsymbol{W}$ with its projected matrix $\overline{\boldsymbol{W}}$:

$$\bar{L}_n(\boldsymbol{b}) = \frac{1}{2}\boldsymbol{b}^{\mathsf{T}}\overline{\boldsymbol{W}}\boldsymbol{b} - \boldsymbol{b}^{\mathsf{T}}\left(\boldsymbol{X}^* \otimes \boldsymbol{\Phi}(\boldsymbol{t})\right)^{\mathsf{T}}\boldsymbol{Y}(\boldsymbol{t}) + \frac{1}{2}\boldsymbol{Y}(\boldsymbol{t})^{\mathsf{T}}\boldsymbol{Y}(\boldsymbol{t}),$$

which is quadratic in $\boldsymbol{b}$ with $\overline{\boldsymbol{W}}$ being positive definite, and thus has a unique global minimizer. It is immediate that solving the equation $\nabla_{\boldsymbol{b}}\bar{L}_n(\boldsymbol{b}) = \boldsymbol{0}$ yields the closed-form minimizer:

$$\tilde{\boldsymbol{b}} = \overline{\boldsymbol{W}}^{-1}\left(\boldsymbol{X}^* \otimes \boldsymbol{\Phi}(\boldsymbol{t})\right)^{\mathsf{T}}\boldsymbol{Y}(\boldsymbol{t}). \tag{4.16}$$

Because $\bar{L}_n(\boldsymbol{b})$ is a quadratic function of $\boldsymbol{b}$, the modified loss function $\bar{L}_n(\boldsymbol{b})$ can be written as

$$\bar{L}_n(\boldsymbol{b}) = \bar{L}_n(\tilde{\boldsymbol{b}}) + \frac{1}{2}(\boldsymbol{b} - \tilde{\boldsymbol{b}})^{\mathsf{T}}\overline{\boldsymbol{W}}(\boldsymbol{b} - \tilde{\boldsymbol{b}}). \tag{4.17}$$

Substituting $\bar{L}_n(\boldsymbol{b})$ for $L_n(\boldsymbol{b})$ in (3.11) yields a modified objective function for $Q_n(\boldsymbol{b})$:

$$\bar{Q}_n(\boldsymbol{b}) = \frac{1}{2}(\boldsymbol{b} - \tilde{\boldsymbol{b}})^{\mathsf{T}}\overline{\boldsymbol{W}}(\boldsymbol{b} - \tilde{\boldsymbol{b}}) + nm\sum_{j=1}^{p} P_\lambda(\|\boldsymbol{b}_j\|),$$

where the term unrelated to $\boldsymbol{b}$ is omitted. Since $\overline{\boldsymbol{W}}$ is positive definite, we have the unique Cholesky decomposition

$$\overline{\boldsymbol{W}} = \boldsymbol{V}^{\mathsf{T}}\boldsymbol{V}, \tag{4.18}$$

where $\boldsymbol{V}$ is an upper-triangular square matrix. Hence we can re-write $\bar{Q}_n(\boldsymbol{b})$ as

$$\bar{Q}_n(\boldsymbol{b}) = \frac{1}{2}\left\|\boldsymbol{V}\tilde{\boldsymbol{b}} - \boldsymbol{V}\boldsymbol{b}\right\|^2 + nm\sum_{j=1}^{p} P_\lambda(\|\boldsymbol{b}_j\|).$$

Rescaling $\bar{Q}_n(\boldsymbol{b})$ as $\tilde{Q}_n(\boldsymbol{b}) = \bar{Q}_n(\boldsymbol{b})/(nm)$ and re-writing it by including $(M+d)p$, we arrive at minimizing a standard penalized least squares function:

$$\begin{aligned}\tilde{Q}_n(\boldsymbol{b}) &= \frac{1}{2nm}\left\|\boldsymbol{V}\tilde{\boldsymbol{b}} - \boldsymbol{V}\boldsymbol{b}\right\|^2 + \sum_{j=1}^{p} P_\lambda(\|\boldsymbol{b}_j\|) \\ &= \frac{1}{2(M+d)p}\left\|\sqrt{\frac{(M+d)p}{nm}}\boldsymbol{V}\tilde{\boldsymbol{b}} - \sqrt{\frac{(M+d)p}{nm}}\boldsymbol{V}\boldsymbol{b}\right\|^2 + \sum_{j=1}^{p} P_\lambda(\|\boldsymbol{b}_j\|),\end{aligned} \tag{4.19}$$

where $\sqrt{\frac{(M+d)p}{nm}}\boldsymbol{V}\tilde{\boldsymbol{b}}$ is regarded as the pseudo-response with the "sample size" $(M+d)p$, the factor $\sqrt{\frac{(M+d)p}{nm}}\boldsymbol{V}$ before $\boldsymbol{b}$ is regarded as the pseudo design matrix, and $\boldsymbol{b}$ is taken as an unknown parameter with dimension $(M+d)p$. While any factor, rather than $(M+d)p$, can be included in (4.19), using $(M+d)p$ makes the resultant objective function (4.19) coincides with the standardized optimization problem considered by Breheny and Huang

(2015), thus allowing us to directly apply their algorithm. In particular, the function *grpreg* in R package *'grpreg'* is used, as done in our following simulation studies. Let $\hat{\boldsymbol{b}}_\tau$ denote the resultant estimator of $\boldsymbol{b}$.

Specifically, to apply *grpreg*, we need to calculate $\boldsymbol{V}$ and $\boldsymbol{V}\tilde{\boldsymbol{b}}$ to obtain the pseudo-design matrix and pseudo-response in (4.19), respectively. First, matrix $\boldsymbol{V}$ is calculated by the Cholesky decomposition of $\overline{\boldsymbol{W}}$, given by (4.18). Next, by (4.16) and (4.18), we determine $\boldsymbol{V}\tilde{\boldsymbol{b}}$ based on $\boldsymbol{V}^\top\boldsymbol{V}\tilde{\boldsymbol{b}} = (\boldsymbol{X}^* \otimes \boldsymbol{\Phi}(\boldsymbol{t}))^\top \boldsymbol{Y}(\boldsymbol{t})$. To find $\boldsymbol{V}\tilde{\boldsymbol{b}}$, it suffices to solve the linear systems of equations:

$$\boldsymbol{V}^\top\tilde{\boldsymbol{Y}} = (\boldsymbol{X}^* \otimes \boldsymbol{\Phi}(\boldsymbol{t}))^\top \boldsymbol{Y}(\boldsymbol{t}) \tag{4.20}$$

for the unknown vector $\tilde{\boldsymbol{Y}}$. Solving (4.20) can be efficiently done by forward substitution (Phillips and Taylor, 1996, Chapter 9) without computing the inverse of $\boldsymbol{V}$.

The minimization of $\tilde{Q}_n(\boldsymbol{b})$ in (4.19) basically depends on the choice of tuning parameters M and λ, which we suggest to be determined using the leave-one-out cross validation (CV) criterion. To be specific, for $i = 1, \ldots, n$, applying the formulation of (4.19) to the data with measurement for the ith subject removed and minimizing the resulting function with

respect to $\boldsymbol{b}$, we let $\hat{\boldsymbol{b}}_\tau^{[-i]}$ denote the resulting estimate of $\boldsymbol{b}$. Define

$$\mathrm{CV}(M,\lambda) = \frac{1}{n}\sum_{i=1}^{n}\Big\{ \Big\| \boldsymbol{Y}_i(\boldsymbol{t}) - (\boldsymbol{X}_i^{*\mathsf{T}} \otimes \boldsymbol{\Phi}(\boldsymbol{t}))\hat{\boldsymbol{b}}_\tau^{[-i]} \Big\|^2 - (\hat{\boldsymbol{b}}_\tau^{[-i]})^{\mathsf{T}} \left[\boldsymbol{\Sigma} \otimes \left\{\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})\right\}\right] \hat{\boldsymbol{b}}_\tau^{[-i]} \Big\}. \tag{4.21}$$

Except the factor n^{-1}, the CV function (4.21) resembles (3.10); a similar formulation was employed by Datta and Zou (2020) for high dimensional linear regression models.

The tuning parameters are determined sequentially by minimizing $\mathrm{CV}(M,\lambda)$ with respect to M and λ, which are evaluated over specified grids $\mathcal{M} \triangleq \{M_1,\ldots,M_D\}$ and $\boldsymbol{\Lambda} \triangleq \{\lambda_1,\ldots,\lambda_K\}$, respectively. To be concrete, for each fixed candidate of $M \in \mathcal{M}$, we first calculate $\hat{\lambda}_M = \mathrm{argmin}_{\lambda\in\boldsymbol{\Lambda}}\,\mathrm{CV}(M,\lambda)$, and then calculate $\hat{M} = \mathrm{argmin}_{M\in\mathcal{M}}\,\mathrm{CV}(M,\hat{\lambda}_M)$. Eventually, $(\hat{M},\,\hat{\lambda}_{\hat{M}})$ is taken as the value of (M,λ) when implementing (4.19).

Alternatively, we may consider the Bayesian information criterion (BIC), where we construct the BIC function as

$$\mathrm{BIC}(M,\lambda) = \log\left(\frac{1}{nm}\Big\|\boldsymbol{Y}(\boldsymbol{t}) - (\boldsymbol{X}^*\otimes\boldsymbol{\Phi}(\boldsymbol{t}))\hat{\boldsymbol{b}}_\tau\Big\|^2\right) + \frac{\log(nm)}{nm}\left(p - |\hat{J}_0|\right)(M+d), \tag{4.22}$$

with $|\hat{J}_0|$ representing the number of elements in $\hat{J}_0$. Then tuning parameters can be set as $(\hat{M},\,\hat{\lambda}) = \mathrm{argmin}_{(M,\,\lambda)\in\mathcal{M}\times\boldsymbol{\Lambda}}\,\mathrm{BIC}(M,\lambda)$. A similar criterion was used in Liang and Li (2009) to do variable selection for the partially linear model.

5. Theoretical Properties

In this section we develop asymptotic results for the proposed method. For any function $f(t)$, with $t \in \mathcal{T}$, let $\|f\|_{L_2} = \sqrt{\int_{\mathcal{T}} \{f(t)\}^2 \mathrm{d}t}$ and $\|f\|_{L_\infty} = \sup_{t\in\mathcal{T}} |f(t)|$ denote the L_2 and L_∞ norms of $f(t)$, respectively. Let $\rho_{\min}(\boldsymbol{A})$ denote the minimum eigenvalue of a square matrix $\boldsymbol{A}$. We present all regularity conditions and corresponding discussions to Section S1 in the supplementary material due to the space limit.

Our first theoretical result shows the existence of the minimizer of $Q_n(\boldsymbol{b})$ and its consistency. The asymptotic limit of $\hat{\boldsymbol{b}}$, denoted $\boldsymbol{b}_0$, is given by equation (S1.3) in the supplementary material.

Theorem 1. *Assume that the tuning parameter $\lambda \to 0$ as $n \to \infty$. Under Assumptions A-E in Section S1 of the supplementary material, there exists a local minimizer $\hat{\boldsymbol{b}}$ of $Q_n(\boldsymbol{b})$ such that*

$$\|\hat{\boldsymbol{b}} - \boldsymbol{b}_0\| = O_p(\sqrt{M/n}), \tag{5.23}$$

and hence, for $j = 1, \ldots, p$,

$$\|\hat{\beta}_j - \beta_j\|_{L_\infty} = O_p(\sqrt{M/n}) \text{ and } \|\hat{\beta}_j - \beta_j\|_{L_2} = O_p(\sqrt{1/n}). \tag{5.24}$$

Theorem 1 encompasses the results for scenarios without measurement error, where $\boldsymbol{\Sigma}$ in (3.7) is set to be $\boldsymbol{0}$. In this instance, the L_∞ and L_2

convergence rates in (5.24) are analogous to the results in Lin et al. (2017, Theorem 2) and Yu et al. (2021, Theorem 1), which are established for precisely measured data only. In the absence of measurement error, Chen et al. (2016) and Wang et al. (2007) established faster convergence rates for using spline methods than (5.24), but they ignored spline approximation error by treating the original function-on-scalar model as a parametric model. On the contrary, the L_2 convergence rate, established by Cai et al. (2022) for robust function-on-scalar linear regression, is slower than (5.24).

The next theorem presents the selection consistency and the point-wise limiting distribution. Define $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} = \mathrm{Cov}(\boldsymbol{\varepsilon}_i(\boldsymbol{t}))$, where $\boldsymbol{\varepsilon}_i(\boldsymbol{t})$ is the error vector given in the model (2.2). Let $\boldsymbol{X}_{Ii}$ denote the subvector of the active covariates in $\boldsymbol{X}_i$, i.e., $\boldsymbol{X}_{Ii} = (X_{i1}, \ldots, X_{is})^{\mathsf{T}}$, and let $\boldsymbol{U}_{Ii}$ denote the subvector of $\boldsymbol{U}_i$ in (3.7) corresponding to $\boldsymbol{X}_{Ii}$. Let $\boldsymbol{\Sigma}_{\boldsymbol{X}_I}$ represent the covariance matrix of $\boldsymbol{X}_{Ii}$, and let $\boldsymbol{\Sigma}_I$ represent the corresponding sub-matrix of $\boldsymbol{\Sigma}$ defined in (3.7), i.e., the covariance matrix of $\boldsymbol{U}_{Ii}$. Further, we use $\boldsymbol{\Gamma}$ to represent the $s^2 \times s^2$ matrix $\mathrm{Cov}\{\mathrm{vec}(\boldsymbol{U}_{Ii}\boldsymbol{U}_{Ii}^{\mathsf{T}})\}$, where $\mathrm{vec}(\boldsymbol{A})$ represents the column vector obtained by stacking the columns of matrix $\boldsymbol{A}$ under each other starting from the left. Define $\boldsymbol{B}_{0s} = [\boldsymbol{b}_{0,1}, \cdots, \boldsymbol{b}_{0,s}]$ to be the $(M+d) \times s$ matrix formed by the first s blocks of $\boldsymbol{b}_0$.

Theorem 2. *Suppose the conditions in Theorem 1 hold. Assume further*

that $\lambda\sqrt{n} \to \infty$ *as* $n \to \infty$.

(i) Then with probability tending to 1, the minimizer $\hat{\boldsymbol{b}}$ *in Theorem 1 satisfies* $\hat{\boldsymbol{b}}_j = \mathbf{0}$ *for all* $j \in J_0$.

(ii) Assume $\rho_{min}(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}) \geq C_5\xi_m$, *where* C_5 *is a positive constant and* $\{\xi_m : m = 1, 2, \dots\}$ *is a sequence of constants satisfying*

$$\frac{n\xi_m}{m} \to \infty \ \text{ and } \ \frac{M^{2q}\xi_m}{nm} \to \infty \ \text{ as } \ n \to \infty \ \text{ and } \ m \to \infty.$$

Then for any $t \in \mathcal{T}$ *and* $j \notin J_0$, *we have that*

$$\frac{\hat{\beta}_j(t) - \beta_j(t)}{\sqrt{\sigma_j^2(t)}} \xrightarrow{d} N(0,1) \quad \text{as } n \to \infty,$$

where

$$\begin{aligned}
&\sigma_j^2(t) \\
=&\frac{1}{n}\Big((\Omega_{jj} + \tilde{\Omega}_{jj})\boldsymbol{\phi}^{\mathsf{T}}(t)\{\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})\}^{-1}\{\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\boldsymbol{\Phi}(\boldsymbol{t})\}\{\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})\}^{-1}\boldsymbol{\phi}(t) \\
&+ \Omega_{jj}\boldsymbol{\phi}^{\mathsf{T}}(t)\boldsymbol{B}_{0s}\boldsymbol{\Sigma}_I\boldsymbol{B}_{0s}^{\mathsf{T}}\boldsymbol{\phi}(t) + \left[\boldsymbol{\Omega}_j^{\mathsf{T}} \otimes \{\boldsymbol{\phi}^{\mathsf{T}}(t)\boldsymbol{B}_{0s}\}\right]\boldsymbol{\Gamma}\left[\boldsymbol{\Omega}_j \otimes \{\boldsymbol{B}_{0s}^{\mathsf{T}}\boldsymbol{\phi}(t)\}\right]\Big),
\end{aligned} \tag{5.25}$$

with Ω_{jj}, $\boldsymbol{\Omega}_j$ *and* $\tilde{\Omega}_{jj}$ *being the* (j, j) *element of* $\boldsymbol{\Sigma}_{\boldsymbol{X}_I}^{-1}$, *the* j*th column of* $\boldsymbol{\Sigma}_{\boldsymbol{X}_I}^{-1}$ *and the* (j, j) *element of* $\boldsymbol{\Sigma}_{\boldsymbol{X}_I}^{-1}\boldsymbol{\Sigma}_I\boldsymbol{\Sigma}_{\boldsymbol{X}_I}^{-1}$, *respectively.*

Theorem 2(i) says that with probability tending to 1, $J_0 \subset \hat{J}_0$, which essentially implies $J_0 = \hat{J}_0$ by (5.23) in Theorem 1. Theorem 2(ii) establishes the asymptotic normal distribution for each estimator $\hat{\beta}_j(t)$ with $j \notin J_0$.

As in Theorem 1, Theorem 2 encompasses the case without measurement error, for which the second and third terms in (5.25) become zero, and then $\sigma_j^2(t)$ is simplified as the first term with $\tilde{\Omega}_{jj} = 0$. A similar formula was presented by Reiss et al. (2010) for the variance of $\hat{\boldsymbol{b}}$ without a proof. We now conclude this section with several remarks.

Remark 1. The validity of Theorems 1 and 2 requires conditions about the penalty function $P_\lambda(v)$ together with the tuning parameter λ. With the penalty function $P_\lambda(v)$ satisfying Condition E in Section S1 of the supplementary material, we require that $\lambda \to 0$ as $n \to \infty$ to ensure estimation consistency, as shown in Theorem 1, and that $\lambda\sqrt{n} \to \infty$ to achieve selection consistency, indicated by Theorem 2(i). These assumptions are also made by many others, including Fan and Li (2001) and Wang et al. (2007), for different contexts.

Remark 2. The formula (5.25) clearly shows the dependence of the limiting distribution of $\hat{\beta}_j(t)$ on the variability of the response model (2.1) and the measurement error (3.7), as well as on the B-spline approximation. If the random error process $\varepsilon_i(t)$ is white noise, the sequence ξ_m can be taken as constant 1 since $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} = \sigma_1^2 \boldsymbol{I}$, with a positive constant σ_1^2. When $\varepsilon_i(t)$ is a random function with well-defined continuous covariance function $C_\varepsilon(t, t') \triangleq \mathbb{E}\{\varepsilon_i(t)\varepsilon_i(t')\}$ for $t, t' \in \mathcal{T}$, then we can take ξ_m such that

$\xi_m \to 0$ as $m \to \infty$ if the eigenvalues of the covariance function $C_\varepsilon(\cdot,\cdot)$, denoted $\nu_1 \geq \cdots \geq \nu_k \geq \cdots$, decrease to 0 sufficiently fast as $k \to \infty$. For details about the eigenvalues decay rate of $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$, we refer the readers to Bunea and Xiao (2015, Section 5.1).

Remark 3. The order conditions for n, m and M, i.e., Assumption C1 in Section S1 of the supplementary material and the additional assumptions in Theorem 2(ii), are compatible. Those conditions can be met if, for example, we set ξ_m, m, and M to be of order m^{-k}, n^α, and n^β, respectively, for a constant k with $0 \leq k < q-1$ and positive constants α and β. Alternatively, setting $\frac{1}{k+1} > \alpha \geq \beta > \frac{1+\alpha(k+1)}{2q}$ enables Assumption C1 in Section S1 and the additional assumptions in Theorem 2(ii) to be compatible.

The preceding results are basically established by conditioning on pre-specified observation points $\boldsymbol{t} = \{t_1, \ldots, t_m\}$. When the observation times are taken as random variables, the results can still hold if the imposed conditions are modified, as discussed in Remark 2 in Section S1 of the supplementary material.

6. Real Data Analysis

The performance of the proposed method is evaluated by extensive simulation studies. For details, please refer to Section S3 in the supplementary

material. To illustrate the utility of the proposed method, here we analyze a real dataset.

Preventing obesity in childhood has received good attention, and one important task is to examine the association of children's daily physical activity with potential risk factors to help prevent obesity. Here we apply the function-on-scalar linear regression model to analyze the children's activity and obesity data (`http://jeffgoldsmith.com/IWAFDA/shortcourse_data.html`). In this study, 420 participants were recruited from 50 Head Start centers in New York between 2003 and 2005. All of them were asked to wear accelerometers which monitored the body activity intensity by summarizing the voltage signals in a period of 10 minutes, leading to 144 observations per day for each child. The experiment lasted 6 days and the original activity data were averaged across these days at each observation points. To use our setup to analyze the data, we re-scale the 24 hours time domain to be [0, 1] and centralize the activity data at each time points by subtracting the sample average from the individual measurements. For $i = 1, \ldots, n$ with $n = 420$, let $Y_i(t)$ denote the final response function for child i, where $t \in [0, 1]$. Figure 1 (a) plots all the functions $\{Y_i(t) : i = 1, \ldots .n\}$, showing that data in the morning time are more concentrated than at other times.

The covariates of interest include children's BMI Z-score, their triceps

and subscapular skinfold thicknesses, age, sex, study season, behavioural variables, presence of an asthma diagnosis, mother's birthplace, mother's work status, and the number of rooms at home (Rundle et al., 2009; Lovasi et al., 2011). More specifically, for subject i, let $\tilde{X}_{i,1}$ denote the BMI Z-score which is the measure of relative body mass index adjusted for child age and sex (De Onis et al., 1997, Section 5.3); let $\tilde{X}_{i,2}$ and $\tilde{X}_{i,3}$ respectively denote the triceps and subscapular skinfold thicknesses which are indicators of children's adiposity; and let $\tilde{X}_{i,4}$ represent the age (in years) at the recruitment. For subject i, sex is represented by a binary variable $\tilde{X}_{i,5}$, with value 1 or 0 indicating female or male; and study season, denoted $\tilde{X}_{i,6}$, shows whether measurements are obtained in warm months from May to September ($\tilde{X}_{i,6} = 1$) or cold months from October to April ($\tilde{X}_{i,6} = 0$). Let $\tilde{X}_{i,7}$ and $\tilde{X}_{i,8}$ denote two binary behavioural variables, which represent whether child i spent over 2 hours per day on watching TV ($\tilde{X}_{i,7}$ equals 1 if true, and 0 otherwise), and spent over 1 hour per day on playing video games ($\tilde{X}_{i,8}$ equals 1 if true, and 0 otherwise). The presence of an asthma diagnosis for subject i is represented by a binary variable $\tilde{X}_{i,9}$, with value 1 indicating diagnosed and 0 otherwise. Binary variable $\tilde{X}_{i,10}$ is used to represent mother's birthplace for child i, with value 1 indicating born in America and 0 otherwise. We use $\tilde{X}_{i,11}$ to characterize whether the mother

for child i has a job ($\tilde{X}_{i,11}$=1) or not ($\tilde{X}_{i,11}$=0). Let $\tilde{X}_{i,12}$ represent the number of rooms at home for child i. We center $\tilde{X}_{i,j}$ on its empirical mean for $j = 1, \ldots, 12$ and further divide by its empirical standard deviation for $j = 1, 2, 3$, and let $\{X_{i,j} : j = 1, \ldots, 12\}$ denote the resulting covariates.

To investigate how the activity profiles are associated with the covariates, we consider the model (2.1) with $p = 12$. It is known that the BMI Z-score $X_{i,1}$ are error-prone, and the triceps $X_{i,2}$ and subscapular skinfold thicknesses $X_{i,3}$ are measured by Lange calipers which are subject to reading error. That is, the actual measurement for $\{X_{i,j} : j = 1, 2, 3\}$ are the surrogate values, denoted $\{X^*_{i,j} : j = 1, 2, 3\}$, in the notation of Section 3.1. The measurement error model is given by model (3.7). In the absence of the information about the magnitude of measurement error degree, we conduct sensitivity analyses to explore possible measurement error effects on inference results.

To be specific, we consider the submatrix of the covariance matrix of the measurement error term $\boldsymbol{U}_i$ in (3.7) for the first three elements to be given by

$$\boldsymbol{\Sigma}_e = (0.2a)^2 \begin{pmatrix} 1 & -0.01c & 0.05c \\ -0.01c & 1 & -0.03c \\ 0.05c & -0.03c & 1 \end{pmatrix},$$

with $a = 0, 1, 2,$ or 3, and $c = 0, 5, 10,$ or 15 to reflect different degrees of measurement error; the marginal reliability ratio defined in Section S3 is 1, 0.96, 0.86 or 0.74 for $a = 0, 1, 2, 3$ respectively. Our choice of values for a and c include three special situations: (1) the case of no measurement error, reflected by $a = 0$; (2) settings with independently occurring measurement error for all the covariates, reflected by $c = 0$; and (3) the case where $\mathbf{\Sigma}_e$ is identical to the empirical covariance matrix of $\{X^*_{i,j} : j = 1, 2, 3\}$, given by $a = 5$ and $c = 1$. Here we report the results for $a = c = 1$ only and defer the results for other values of a and c to Section S4 of the supplementary material, where we also compare the performance of CV and BIC methods for the case with $a = c = 1$.

Let the degree of B-spline $d = 3$. We first apply the CV selection method to determine the tuning parameters. Two skinfold thicknesses covariates, $X_{i,2}$ and $X_{i,3}$, together with other three covariates, $X_{i,9}$, $X_{i,10}$ and $X_{i,11}$, are excluded in the model. With $a = c = 1$, we display in Figure 1 (b)-(h) the estimated coefficient functions and 90% bootstrap confidence regions for all active covariates, where those confidence regions are constructed as follows: 5000 bootstrap samples are repeatedly generated from the original sample using sampling with replacement. We then estimate coefficient functions for each bootstrap sample and pool these estimates

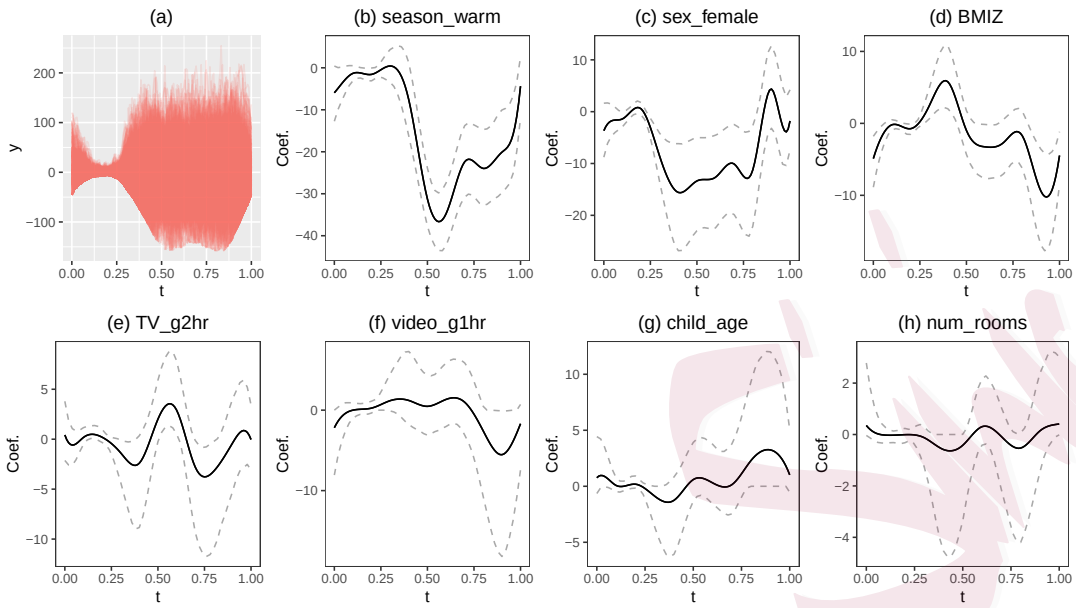

Figure 1: Centered activity profiles and coefficient function estimates.

together to obtain upper and lower 5% quantiles. The temporal effects of covariates are clearly seen. For example, BMI Z-score $X_{i,1}$ shows a positive effect on activity in the morning and a negative effect in the evening.

By contrast, we also analyze the data using the method of Chen et al. (2016) who did not consider the potential measurement error effects. As the number of the B-spline functions considered by Chen et al. (2016) must be pre-determined, we set it to be $M+d$, with M being selected by the proposed CV method and $d = 3$, to align with our setup. The results produced by the method of Chen et al. (2016) , not reported in details here, show that the norms of the estimated coefficient functions for mother's work status

$X_{i,10}$, the presence of an asthma diagnosis $X_{i,9}$, the number of rooms $X_{i,12}$, and two skinfold thicknesses $X_{i,2}$ and $X_{i,3}$ are much smaller than those for the rest of the covariates. These results are fairly comparable to the results produced by our method using CV, except that our estimated coefficient for the number of rooms is not penalized to 0, though it is small.

Finally, we comment that the specification of values for a and c is not unique; other values of a and c can be considered to assess the sensitivity of inference results to different magnitudes of measurement error, following the same procedure discussed here. While sensitivity analyses cannot reveal what the underlying truth is, such analyses help us understand the measurement error effects on affecting inference results, and thus enhance our interpretation of data analysis. The sensitivity analyses here hinges on the use of the measurement error model (3.7). When this model is inadequate to facilitate the measurement error process, the analysis results here do not necessarily uncover the impact of measurement error on inference truthfully.

7. Discussion

Function-on-scalar linear regression models have been proved to be useful to describe the relationship between a functional response and multiple

scalar covariates. Such models, however, cannot be directly applied to handle error-corrupted data. Naively ignoring the measurement error effects typically distorts inference results. This issue is further exacerbated by the presence of inactive variables. In this paper, we study the function-on-scalar linear regression model with additive covariate measurement error. Under the framework of B-spline approximation, we investigate the measurement error effects and reveal the connection of such effects with the ridge regression. We propose a debiased loss function, coupled with a sparsity-inducing penalty function, such as SCAD, to simultaneously estimate the coefficient functions and detect important predictors. Under mild conditions, estimation consistency, selection consistency, and the limiting distribution of the resultant estimators are rigorously established. We develop an efficient algorithm and tuning parameters selection methods to implement the proposed procedure. Numerical studies demonstrate the satisfactory performance of the proposed method, as opposed to the deleterious effects yielded from the naive method which disregards the measurement error effects.

Future extensions of this work may take the following directions. Here we consider the case where the dimension p of covariates is fixed. It is interesting to extend the development in this article to accommodate settings with a diverging p. Although the algorithm in Section 4 can still apply

to this case, establishing theoretical results requires extra care. In addition, the effect of using projected matrix (4.15) should be considered since the original matrix $\boldsymbol{W}$ is always not positive definite when $p > n$. In the current development, we propose to use cross validation or Bayesian information criteria to choose suitable values for the tuning parameters λ and M to work out the estimates of the model parameters. The performance of these two criteria is compared numerically. It is interesting to analytically compare how these methods may perform differently. Finally, it is interesting to generalize the proposed method to handle measurement error in both covariates and functional responses. Additional modeling of the measurement error process for the functional response is basically needed in such a circumstance.

Supplementary Material

The online supplementary material contains regularity conditions, technical proofs, simulation studies, and data analysis, along with the extended development with the least squares loss function (2.6) replaced by the generalized least squares loss function.

Acknowledgements

The authors thank an Associate Editor and the review team for their helpful comments on the initial version. The research was partially supported by the grants of the Discovery Grants Program and the Emerging Infectious Disease Modeling Program from the Natural Sciences and Engineering Research Council of Canada. Grace Y. Yi is Canada Research Chair in Data Science (Tier 1). Her research was undertaken, in part, thanks to funding from the Canada Research Chairs program.

References

Barber, R. F., M. Reimherr, and T. Schill (2017). The function-on-scalar lasso with applications to longitudinal gwas. *Electronic Journal of Statistics 11*(1), 1351–1389.

Breheny, P. and J. Huang (2015). Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Statistics and Computing 25*(2), 173–187.

Bunea, F. and L. Xiao (2015). On the sample covariance matrix estimator of reduced effective rank population matrices, with applications to fpca. *Bernoulli 21*(2), 1200–1230.

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

Carroll, R. J., D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu (2006). *Measurement Error*

in Nonlinear Models: A Modern Perspective. Chapman and Hall/CRC.

Chen, Y., J. Goldsmith, and R. T. Ogden (2016). Variable selection in function-on-scalar regression. *Stat 5*(1), 88–101.

Chiang, C.-T., J. A. Rice, and C. O. Wu (2001). Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variables. *Journal of the American Statistical Association 96*(454), 605–619.

Datta, A. and H. Zou (2017). Cocolasso for high-dimensional error-in-variables regression. *The Annals of Statistics 45*(6), 2400–2426.

Datta, A. and H. Zou (2020). A note on cross-validation for lasso under measurement errors. *Technometrics 62*(4), 549–556.

De Boor, C. (1978). *A Practical Guide to Splines*. Springer-Verlag New York.

De Onis, M., M. Blossner, W. H. Organization, et al. (1997). Who global database on child growth and malnutrition. Technical report, World Health Organization.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association 96*(456), 1348–1360.

Fan, Z. and M. Reimherr (2017). High-dimensional adaptive function-on-scalar regression. *Econometrics and Statistics 1*, 167–183.

Horváth, L. and P. Kokoszka (2012). *Inference for Functional Data with Applications*. Springer Science & Business Media.

REFERENCES

Jiang, Z., Z. Huang, and H. Zhu (2021). Estimation and inference for covariate adjusted partially functional linear regression models. *Statistics and Its Interface 14*(4), 359–371.

Liang, H. and R. Li (2009). Variable selection for partially linear models with measurement errors. *Journal of the American Statistical Association 104*(485), 234–248.

Lin, Z., J. Cao, L. Wang, and H. Wang (2017). Locally sparse estimator for functional linear regression models. *Journal of Computational and Graphical Statistics 26*(2), 306–318.

Lovasi, G. S., J. S. Jacobson, J. W. Quinn, K. M. Neckerman, M. N. Ashby-Thompson, and A. Rundle (2011). Is the environment near home and school associated with physical activity and adiposity of urban preschool children? *Journal of Urban Health 88*(6), 1143–1157.

Meng, S., Z. Huang, J. Zhang, and Z. Jiang (2021). Estimation on functional partially linear single index measurement error model. *Communications in Statistics-Theory and Methods*, 1–23.

Parodi, A. and M. Reimherr (2018). Simultaneous variable selection and smoothing for high-dimensional function-on-scalar regression. *Electronic Journal of Statistics 12*(2), 4602–4639.

Phillips, G. M. and P. J. Taylor (1996). *Theory and Applications of Numerical Analysis*. Elsevier.

Ramsay, J. and B. Silverman (2005). *Functional Data Analysis* (Second ed.). Springer-Verlag.

Reiss, P. T., L. Huang, and M. Mennes (2010). Fast function-on-scalar regression with penalized

basis expansions. *The International Journal of Biostatistics 6*(1).

Rundle, A., I. F. Goldstein, R. B. Mellins, M. Ashby-Thompson, L. Hoepner, and J. S. Jacobson (2009). Physical activity and asthma symptoms among new york city head start children. *Journal of Asthma 46*(8), 803–809.

Wang, L., G. Chen, and H. Li (2007). Group scad regression analysis for microarray time course gene expression data. *Bioinformatics 23*(12), 1486–1494.

Yi, G. Y. (2017). *Statistical Analysis with Measurement Error or Misclassification: Strategy, Method and Application*. Springer.

Yi, G. Y. and R. J. Cook (2005). Errors in the measurement of covariates. *Encyclopedia of Biostatistics 3*, 1741–1748.

Yi, G. Y., A. Delaigle, and P. Gustafson (2021). *Handbook of Measurement Error Models*. CRC Press.

Yu, S., G. Wang, L. Wang, and L. Yang (2021). Multivariate spline estimation and inference for image-on-scalar regression. *Statistica Sinica 31*(3), 1463–1487.

Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B 68*(1), 49–67.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics 38*(2), 894–942.

Zhang, J.-T. and J. Chen (2007). Statistical inferences for functional data. *The Annals of*

Statistics 35(3), 1052–1079.

Zhao, M., Y. Gao, and Y. Cui (2022). Variable selection for longitudinal varying coefficient errors-in-variables models. *Communications in Statistics-Theory and Methods 51*(11), 3713–3738.

Zhu, H., R. Li, and L. Kong (2012). Multivariate varying coefficient model for functional responses. *The Annals of Statistics 40*(5), 2634.

Zhu, H., R. Zhang, Z. Yu, H. Lian, and Y. Liu (2019). Estimation and testing for partially functional linear errors-in-variables models. *Journal of Multivariate Analysis 170*, 296–314.

Zhu, H., R. Zhang, and G. Zhu (2020). Estimation and inference in semi-functional partially linear measurement error models. *Journal of Systems Science and Complexity 33*(4), 1179–1199.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association 101*(476), 1418–1429.

first author affiliation

Department of Statistical and Actuarial Sciences, University of Western Ontario, Canada

E-mail: (first author email)

ysun2343@uwo.ca

second author affiliation

Department of Statistical and Actuarial Sciences, University of Western Ontario, Canada

REFERENCES

Department of Computer Science, University of Western Ontario, Canada

E-mail: (second author email)

gyi5@uwo.ca