

MODEL AVERAGING ESTIMATION FOR PARTIALLY LINEAR FUNCTIONAL SCORE MODELS

Shishi Liu¹, Chunming Zhang², Hao Zhang³, Rou Zhong³
and Jingxiao Zhang^{*3}

¹*Hangzhou Dianzi University*, ²*University of Wisconsin*
and ³*Renmin University of China*

Abstract: The scalar-on-function regression is quite useful for modelling mixed-data in the context of scalar and functional variables. Under this class of regression, the paper aims at proposing a compelling alternative to model selection methods to address model selection uncertainty. The considered models characterize a scalar response using parametric effect of the scalar predictors and nonparametric effect of a functional predictor, and a model averaging estimation is developed based on Mallows-type criterion to assign weights for averaging. Further, the asymptotic optimality of the resulting estimator, in terms of achieving the smallest possible squared error loss, is established. Besides, simulation studies demonstrate its superiority to or comparability with some information criterion score-based model selection and averaging estimators. The proposed procedure is also applied to a mid-infrared spectra dataset for illustration.

Key words and phrases: Functional data, mallows-type criterion, model average.

1. Introduction

Functional data analysis (FDA) has received growing attention in recent decades due to its remarkable flexibility and widespread applicability in handling complex data, including variables defined on a continuum, such as time or space. Ramsay and Silverman (2005) offers a comprehensive introduction to FDA across various fields. One of the most extensively studied topics in FDA pertains to functional regression. A large body of research has been dedicated to developing regression models that incorporate functional predictors, with a predominant focus on functional parametric regression models, see Cardot, Ferraty and Sarda (1999, 2003), Yao, Müller and Wang (2005) and Cai and Hall (2006). These studies, falling into the category of scalar-on-function regression, assume specific forms of the regression model, such as functional linear model (FLM). The classic FLM is formulated to depict a linear relationship between a scalar response and a functional predictor, which is conventionally expressed as

$$Y = \beta_0 + \int_{\mathcal{T}} X(t)\beta(t)dt + \varepsilon, \quad (1.1)$$

*Corresponding author. E-mail: zhjxiaoruc@163.com

where Y represents the scalar response, $X(t)$ denotes the functional predictor defined on a continuum $\mathcal{T}$, β_0 and $\beta(t)$ represent the unknown coefficients, and ε is the error term.

Nevertheless, it is often noted that the above functional parametric regression may fall short in capturing potential non-linear associations between Y and $X(t)$, especially in complex data from fields like meteorology and biometrics. Consequently, statisticians have endeavored to develop nonparametric methodologies for functional regressions. For example, Müller and Yao (2008) and Fan, James and Radchenko (2015) introduced an additive form of the “features” for functional predictor(s) into models, derived from the basis expansion. This established the framework of functional additive model. Yao and Müller (2010) introduced the quadratic term of the functional predictor, and discussed the functional quadratic regression. These works illustrate both the necessity and capability of nonparametric modeling in effectively capturing non-linear characteristics in data.

The aforementioned models primarily concentrate on addressing functional predictors, potentially overlook the impact on the response variable from other available scalar predictors. Recently, researchers have extended their investigations to address a more common scenario in applications like neuroimaging and chemometrics studies, where both scalar and functional predictors coexist. It is referred to as mixed data or hybrid data in Ramsay and Silverman (2005). Many studies underscore the importance of employing a partially linear structure in handling functional regressions involving mixed data. For instance, Yu, Kong and Mizera (2016) and Kong et al. (2016) developed partial functional linear regressions as

$$Y = \theta_0 + \boldsymbol{\theta}^T \mathbf{z} + \int_{\mathcal{T}} X(t)\beta(t)dt + \varepsilon, \quad (1.2)$$

where θ_0 is the intercept, $\boldsymbol{\theta}$ is a vector of coefficients, $\mathbf{z}$ represents a vector of scalar predictors of interests. Compared to the FLM (1.1), these models integrate a linear combination of scalar predictors into the framework. Wong, Li and Zhu (2019) and Tang, Tu and Kong (2023) further augmented the modeling flexibility by introducing nonparametric effect of the functional predictor(s), which demonstrates improved fitting by considering both the partially linear structure and nonparametric modeling of functional predictor(s). Aneiros-Pérez and Vieu (2006) and Wang, Feng and Chen (2016) also incorporated a partially linear structure of scalar predictors into their respective functional nonparametric regression models. Inspired by these advancements, we aim to investigate a class of partially linear functional nonparametric regression models for mixed data as

$$Y = \theta_0 + \boldsymbol{\theta}^T \mathbf{z} + m\{X(t)\} + \varepsilon, \quad (1.3)$$

where $m(\cdot)$ is a nonparametric modeling function for $X(t)$. These models

leverage the capability of a partially linear framework to accommodate both scalar and functional predictors simultaneously, while also harnessing the flexibility of nonparametric modeling for the functional predictor.

An initial question arises: how should $X(t)$ be processed? A common practice in functional regressions is to project $X(t)$ onto a functional space with a finite functional basis. These projections associated with basis functions, considered to contain all information in $X(t)$, are then used for subsequent analysis; see Müller and Yao (2008), Zhu, Yao and Zhang (2014), and Kong et al. (2016). Thereby, model uncertainty arises from the choice of functional space truncation. In addition, varying decisions on which scalar components to retain in model (1.3) can also introduce model uncertainty. Hence, there is a level of uncertainty inherent in using selection techniques, a factor usually ignored by model selection approaches. To sum up, simply choosing one model may lead to inferior performance, as early noted by Draper (1995), Buckland, Burnham and Augustin (1997), Clyde and George (2004), and Claeskens and Hjort (2008).

To mitigate this issue, an alternative strategy to model selection is adopted in our work. Specifically, we employ a model averaging method for estimation, which combines multiple candidate models by assigning weights to each to address potential model uncertainty and deliver a more robust outcome. Model averaging paves an alternative way for tackling model uncertainty and has been extensively investigated in the scalar regression literature. For instance, Buckland, Burnham and Augustin (1997) and Hjort and Claeskens (2003) advocated an information-criterion-based weighting scheme, which assigns weights calculated from information criterion scores (such as AIC or BIC) for each candidate model. Hansen (2007) introduced a Mallows's criterion for weights optimization, demonstrating its asymptotic optimality.

This work has inspired the development of diverse Mallows-type model averaging procedures, which rely on an unbiased estimator of squared risk (up to a constant), in various model contexts. Zhang, Zou and Liang (2014) proposed a Mallows-type criterion for weights selection in linear mixed-effects models. They combined the conventional best linear unbiased estimators from each candidate model using the assigned weights. In a similar vein, Zhang and Wang (2019) and Zhu et al. (2019) also explored a Mallows-type criterion for weights selection in the context of partially linear models and varying-coefficient partially linear models. Besides, there are several other noteworthy model averaging methods, including jackknife model averaging (Hansen and Racine, 2012) and cross-validation model averaging (Zhang, Wan and Zou, 2013; Cheng and Hansen, 2015; Gao et al., 2016; Zhang and Liu, 2023), which are particularly useful in cases where deriving an unbiased estimator of squared risk is challenging. Moreover, Kullback–Leibler loss-based model averaging (Zhang, Zou and Carroll, 2015; Fang, Li and Xia, 2022; Zou et al., 2022) utilizes the divergence of the unconditional and conditional densities of Y given the fitted model, making it especially suitable for models with

generalized response variables. Collectively, these studies highlight the superior or comparable performance of model averaging methods to conventional model selection techniques in the context of scalar regressions.

It is worth noting that there have been growing developments in model averaging estimation for functional regressions. These works primarily employ the model averaging strategy to tackle model uncertainty arising from factors like functional space truncation, as previously discussed. For instance, Zhang, Chiou and Ma (2018) introduced a model averaging estimator for a linear model that includes a response and a predictor, both of which are of functional types. In addition, Zhang and Zou (2020) investigated a generalized functional linear model with a link function, incorporating a generalized scalar response and a functional predictor. And they developed a model averaging framework for this context. Moreover, Zhu, Zou and Zhang (2018) considered a mixed-data scenario and proposed an optimal model averaging method based on a Mallows-type criterion for model (1.2), where both scalar and functional predictors are parametrically included in the model. These procedures showcase their superiority over other competing selection methods within their frameworks, indicating the promising potential of model averaging for our model (1.3).

In summary, this article addresses the issue of mixed-data in functional regression by proposing a novel class of models that accommodate a partially linear structure for scalar predictors and a nonparametric effect of the functional predictor. By adopting this nonparametric approach, our work enhances modeling flexibility and facilitates the detection of potential non-linear effects on Y . This distinguishes our approach from that of Zhu, Zou and Zhang (2018), who also address the issue of mixed-data but employ a linear modeling method. Another special feature of our work, which differs from the existing functional regression literature, is that we focus on developing a model averaging approach for estimation. This idea of model averaging tackles model uncertainty stemming from selection procedures, thereby reducing the risk of selecting an inferior model by model selection methods. Furthermore, we propose a Mallows-type criterion for weights selection, based on an unbiased estimator of the squared risk. And we establish the asymptotic optimality of the resulting estimator in terms of achieving the lowest squared error loss, which also allows for non-nested and heteroscedastic candidate models.

The rest of this paper is organized as follows. Section 2 presents the model setup and the proposed model averaging estimator. The asymptotic optimality is established in Section 3. Sections 4 and 5 illustrate the simulation study and an empirical application. Section 6 concludes our work with a discussion. Additional simulations, additional details in real application, and all proofs are given in the online Supplementary Material.

2. Methodology

2.1. Model and estimator

Let Y be a scalar response variable associated with a scalar predictor vector $\mathbf{z}$ and a functional predictor X . Let $\{Y_i, \mathbf{Z}_i, X_i\}_{i=1}^n$ be independent identically distributed (i.i.d.) copies of $\{Y, \mathbf{z}, X\}$. We model the relationship between Y and $\{\mathbf{z}, X(\cdot)\}$ in the form of (1.3). As mentioned earlier, it is common practice to project X onto a functional space with a finite functional basis and utilize these projections to specify the effect of X . One of the most widely used bases is the eigen-basis derived from functional principal component analysis (FPCA). Refer to Müller and Yao (2008), Kong et al. (2016), Yu, Kong and Mizera (2016), Zhu, Zou and Zhang (2018), Zhang, Chiou and Ma (2018), and Wong, Li and Zhu (2019) for examples. Hence, we follow this routine and establish our model as

$$Y_i = \mu_i + \varepsilon_i = \mathbf{Z}_i^T \boldsymbol{\theta} + \mathbf{f}(\boldsymbol{\xi}_i) + \varepsilon_i, \quad (2.1)$$

where, albeit with a slight abuse of notation, $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^T$ now denotes the random error vector with conditional mean 0 and conditional variance matrix $\boldsymbol{\Omega} = \text{diag}(\sigma_1^2, \dots, \sigma_n^2)$ given $\{\mathbf{z}, X(\cdot)\}$. $\mathbf{f}(\cdot)$ is an unknown function for nonparametric modeling. The vector $\boldsymbol{\xi}_i$, which contains the information within each X_i , is derived through FPCA. Specifically, suppose $X(t)$, $t \in \mathcal{T}$ be a random function from Hilbert space $L^2(\mathcal{T})$ with mean function $\nu(t)$ and covariance function $\mathcal{C}(s, t) = \text{cov}\{X(s), X(t)\}$. $\mathcal{T}$ is typically assumed to be a compact interval.

The classical FPCA takes eigen decomposition of the corresponding covariance operator as $(\mathcal{C}\psi_l)(t) = \lambda_l \psi_l(t)$, $l = 1, 2, \dots$, where $\{\psi_1(t), \psi_2(t), \dots\}$ is a set of (orthonormal) eigenfunctions associated with eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots > 0$. Then, $X(t)$ is projected onto a finite set of eigenfunctions, resulting in a truncated Karhunen–Loève expansion

$$X(t) = \nu(t) + \sum_{l=1}^q \zeta_l \psi_l(t),$$

where $\zeta_l = \int_{\mathcal{T}} \{X(t) - \nu(t)\} \psi_l(t) dt$ represents the functional principal component (FPC) score associated with the l -th eigenfunction, and $\text{var}(\zeta_l) = \lambda_l$, $l = 1, \dots, q$; refer to Rice and Silverman (1991) and Hall, Müller and Wang (2006).

Denote by $\boldsymbol{\zeta}_i = (\zeta_{i1}, \dots, \zeta_{iq})$ the vector of FPC scores for $X_i(t)$. To avoid possible scale issues in nonparametric modeling, we apply a transformation to each ζ_{il} , yielding $\xi_{il} = \Phi(\zeta_{il}; \lambda_l^{-1/2})$, where $\Phi(\cdot)$ is a continuously differentiable map from $\mathbb{R}$ to $[0, 1]$. And the resulting vector $\boldsymbol{\xi}_i = (\xi_{i1}, \dots, \xi_{iq})^T$ represents the transformed FPC scores for $X_i(t)$. Similar to the process in Wong, Li and Zhu (2019), we can choose $\Phi(\cdot)$ to be a suitable cumulative distribution function (CDF) in practice. For instance, if ζ_{il} approximately follows Gaussian

distribution, we can specify $\Phi(\cdot)$ as the standard Gaussian distribution function, then $\xi_{il} = \Phi(\zeta_{il}; \lambda_l^{-1/2}) = \Phi(\lambda_l^{-1/2} \zeta_{il})$ will be almost uniform in $[0, 1]$.

Since transformed FPC scores are used in our model (2.1), we refer to it as the *partially linear functional score (PLFS) model*. Note that different functional space truncations, i.e., different choices for $\boldsymbol{\xi}$, as well as for $\mathbf{z}$, introduce model uncertainty. Hence, we investigate a model averaging procedure to address this issue. Assuming there are M candidate models approximating the true model. The m -th candidate PLFS model comprises p_m regressors in $\mathbf{Z}_i$ and q_m regressors in $\boldsymbol{\xi}_i$,

$$Y_i = \mu_{(m),i} + \varepsilon_{(m),i} = \mathbf{Z}_{(m),i}^T \boldsymbol{\theta}_{(m)} + \mathbf{f}_{(m)}(\boldsymbol{\xi}_{(m),i}) + \varepsilon_{(m),i},$$

where p_m and q_m are determined during the construction of the m -th candidate model, $\mathbf{Z}_{(m),i}$ is a $p_m \times 1$ vector, $\boldsymbol{\theta}_{(m)}$ is the corresponding unknown coefficients, $\boldsymbol{\xi}_{(m),i}$ is a $q_m \times 1$ vector, $\mathbf{f}_{(m)}$ is an unknown function mapping from $[0, 1]^{q_m}$ to $\mathbb{R}$, and $\varepsilon_{(m),i}$ contains both an approximation error of the m -th candidate model and a random error.

For each candidate model, the kernel smoothing method (Speckman, 1988) is employed in the estimation. Since multiple transformed FPC scores are to be processed, we adopt a product kernel function here denoted as $\mathcal{K}_{h_m}(\cdot) = \prod_{l=1}^{q_m} k_{h_{m,l}}(\cdot)$, where $k_{h_{m,l}}(\cdot)$ is a univariate kernel function and $h_{m,l}$ is a scalar bandwidth. We take $h_{m,l} = h_m$ for simplicity and clarity, $l = 1, \dots, q_m$. Furthermore, let $\mathbf{K}_{(m)} = (K_{(m),ij})$ be the $n \times n$ smoother matrix with elements

$$K_{(m),ij} = \frac{\mathcal{K}_{h_m}(\boldsymbol{\xi}_{(m),i} - \boldsymbol{\xi}_{(m),j})}{\sum_{j'=1}^n \mathcal{K}_{h_m}(\boldsymbol{\xi}_{(m),i} - \boldsymbol{\xi}_{(m),j'})}.$$

Then, the suggested kernel smoothing estimators for $\boldsymbol{\theta}_{(m)}$ and $\mathbf{f}_{(m)}(\boldsymbol{\xi}_{(m)})$ can be derived as follows,

$$\begin{aligned} \tilde{\boldsymbol{\theta}}_{(m)} &= \left(\tilde{\mathbf{Z}}_{(m)}^T \tilde{\mathbf{Z}}_{(m)} \right)^{-1} \tilde{\mathbf{Z}}_{(m)}^T (\mathbf{I} - \mathbf{K}_{(m)}) \mathbf{Y}, \\ \tilde{\mathbf{f}}_{(m)}(\boldsymbol{\xi}_{(m)}) &= \mathbf{K}_{(m)} \left(\mathbf{Y} - \mathbf{Z}_{(m)} \tilde{\boldsymbol{\theta}}_{(m)} \right), \end{aligned}$$

where $\mathbf{Y} = (Y_1, \dots, Y_n)^T$, $\mathbf{Z}_{(m)}$ is an $n \times p_m$ matrix, and $\tilde{\mathbf{Z}}_{(m)} = (\mathbf{I} - \mathbf{K}_{(m)})\mathbf{Z}_{(m)}$. Obviously, $\tilde{\boldsymbol{\theta}}_{(m)}$ is actually a least square estimate and $\tilde{\mathbf{f}}_{(m)}$ is a Nadaraya–Watson (local constant) estimator. Therefore, the estimation of $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)^T$ under the m -th candidate model is given by

$$\begin{aligned} \tilde{\boldsymbol{\mu}}_{(m)} &= \mathbf{Z}_{(m)} \tilde{\boldsymbol{\theta}}_{(m)} + \tilde{\mathbf{f}}_{(m)}(\boldsymbol{\xi}_{(m)}) \\ &= \tilde{\mathbf{Z}}_{(m)} \left(\tilde{\mathbf{Z}}_{(m)}^T \tilde{\mathbf{Z}}_{(m)} \right)^{-1} \tilde{\mathbf{Z}}_{(m)}^T (\mathbf{I} - \mathbf{K}_{(m)}) \mathbf{Y} + \mathbf{K}_{(m)} \mathbf{Y} \equiv \mathbf{P}_{(m)} \mathbf{Y}. \end{aligned}$$

Let $\tilde{\mathbf{P}}_{(m)} \equiv \tilde{\mathbf{Z}}_{(m)} (\tilde{\mathbf{Z}}_{(m)}^T \tilde{\mathbf{Z}}_{(m)})^{-1} \tilde{\mathbf{Z}}_{(m)}^T$, which is idempotent, and $\mathbf{P}_{(m)} \equiv \tilde{\mathbf{P}}_{(m)} (\mathbf{I} - \mathbf{K}_{(m)}) + \mathbf{K}_{(m)}$.

Remark 1. In contrast to the exploration of model (1.2) with multiple functional predictors in Zhu, Zou and Zhang (2018), where both the effects of scalar and functional predictors are assumed to be linear, our model (2.1) enables the detection of potential non-linear association in data through nonparametric modeling. They employed the projections obtained from FPCA of $X_j(t)$ and the corresponding coefficient function $\beta_j(t)$, denoted by vectors ζ_j and η_j , to characterize the “features” in $X(t)$ and $\beta(t)$. This results in a reduced linear model as $Y = \mathbf{z}^T \boldsymbol{\theta} + \sum_{j=1}^v \zeta_j^T \eta_j + \varepsilon$. Then, they utilized the OLS estimator for each candidate model, whereas our work employs a kernel smoothing approach.

2.2. Weight choice criterion

Let $\boldsymbol{\omega} = (\omega_1, \dots, \omega_M)^T$ be a weight vector in the unit simplex of $\mathbb{R}^M$,

$$\mathcal{H}_n = \left\{ \boldsymbol{\omega} \in [0, 1]^M : \sum_{m=1}^M \omega_m = 1 \right\}.$$

Then, the model averaging estimator of μ follows as

$$\tilde{\boldsymbol{\mu}}(\boldsymbol{\omega}) = \sum_{m=1}^M \omega_m \tilde{\boldsymbol{\mu}}_{(m)} = \sum_{m=1}^M \omega_m \mathbf{P}_{(m)} \mathbf{Y} = \mathbf{P}(\boldsymbol{\omega}) \mathbf{Y},$$

where $\mathbf{P}(\boldsymbol{\omega}) = \sum_{m=1}^M \omega_m \mathbf{P}_{(m)}$. Define the square error loss function and the corresponding conditional risk function as

$$\begin{aligned} L_n(\boldsymbol{\omega}) &= \|\tilde{\boldsymbol{\mu}}(\boldsymbol{\omega}) - \boldsymbol{\mu}\|^2 = \|\mathbf{P}(\boldsymbol{\omega}) \mathbf{Y} - \boldsymbol{\mu}\|^2, \\ R_n(\boldsymbol{\omega}) &= \mathbb{E}\{L_n(\boldsymbol{\omega}) | \mathbf{z}, X\} = \|(\mathbf{P}(\boldsymbol{\omega}) - \mathbf{I})\boldsymbol{\mu}\|^2 + \text{tr}\{\mathbf{P}^T(\boldsymbol{\omega})\mathbf{P}(\boldsymbol{\omega})\boldsymbol{\Omega}\}, \end{aligned}$$

where $\|\cdot\|$ denotes the L^2 norm of a vector, and $\text{tr}(A)$ represents the trace of matrix A . So, we can select the optimal weights based on the following Mallows-type criterion

$$C_n(\boldsymbol{\omega}) = \|\mathbf{Y} - \tilde{\boldsymbol{\mu}}(\boldsymbol{\omega})\|^2 + 2\text{tr}\{\mathbf{P}(\boldsymbol{\omega})\boldsymbol{\Omega}\}.$$

It is evident that $\mathbb{E}\{C_n(\boldsymbol{\omega}) | \mathbf{z}, X\} = R_n(\boldsymbol{\omega}) + \text{tr}(\boldsymbol{\Omega})$. Therefore, $C_n(\boldsymbol{\omega})$ serves as an unbiased estimator of the expected in-sample squared error loss plus a constant, similar to the Mallows's criterion proposed by Hansen (2007). As $\text{tr}(\boldsymbol{\Omega})$ is independent of $\boldsymbol{\omega}$, the optimal weights can be obtained by minimizing $C_n(\boldsymbol{\omega})$ given that $\boldsymbol{\Omega}$ is known.

However, obtaining complete curves of $X(t)$ is often infeasible in real-world measurements, rendering the FPC scores $\boldsymbol{\zeta}$ and the corresponding transformed FPC scores $\boldsymbol{\xi}$ unobservable. Consequently, the above-mentioned procedure cannot be directly implemented. To ensure practical applicability, we substitute the original $\boldsymbol{\xi}_{(m)}$ with its estimator $\hat{\boldsymbol{\xi}}_{(m)}$. Specifically, suppose $X_i(t)$ is discretely measured with noise,

$$X_{ij} = X_i(t_{ij}) + e_{ij}, \quad i = 1, \dots, n, \quad j = 1, \dots, N_i,$$

where e_{ij} 's are independent measurement errors with mean 0 and variance σ_e^2 . Additionally, the errors e_{ij} are also independent of X_i and $\mathbf{Z}_i$. We now focus on densely observed trajectories, allowing $X_i(t)$ to be effectively reconstructed from $\{(t_{ij}, X_{ij}) : j = 1, \dots, N_i\}$ using a smoother operator, see Kong et al. (2016) and Wong, Li and Zhu (2019). The reconstructed function is denoted by $\tilde{X}_i(t)$. Then, the mean and covariance function of $X(t)$ are estimated by

$$\begin{aligned}\hat{\nu}(t) &= \frac{1}{n} \sum_{i=1}^n \tilde{X}_i(t), \\ \hat{\mathcal{C}}(s, t) &= \frac{1}{n} \sum_{i=1}^n \left\{ \tilde{X}_i(s) - \hat{\nu}(s) \right\} \left\{ \tilde{X}_i(t) - \hat{\nu}(t) \right\}^T.\end{aligned}$$

The spectral decomposition $\hat{\mathcal{C}}(s, t) = \sum_{l=1}^{n-1} \hat{\lambda}_l \hat{\psi}_l(s) \hat{\psi}_l(t)$ yields sample eigenvalues $\{\hat{\lambda}_l\}$ and eigenfunctions $\{\hat{\psi}_l\}$. The estimates for FPC scores are subsequently obtained by

$$\hat{\zeta}_{il} = \int_{\mathcal{T}} \left\{ \tilde{X}_i(t) - \hat{\nu}(t) \right\} \hat{\psi}_l(t) dt, \quad \hat{\xi}_{il} = \Phi \left(\hat{\zeta}_{il}; \hat{\lambda}_l^{-1/2} \right).$$

Once we obtain $\hat{\boldsymbol{\xi}}_{(m)}$, the original quantities listed above have their practical substitutes.

The smoother matrix is now denoted as $\hat{\mathbf{K}}_{(m)}$ with elements

$$\hat{K}_{(m),ij} = \frac{\mathcal{K}_{h_m}(\hat{\boldsymbol{\xi}}_{(m),i} - \hat{\boldsymbol{\xi}}_{(m),j})}{\sum_{j'=1}^n \mathcal{K}_{h_m}(\hat{\boldsymbol{\xi}}_{(m),i} - \hat{\boldsymbol{\xi}}_{(m),j'})}.$$

Then, the final kernel smoothing estimators for $\boldsymbol{\theta}_{(m)}$ and $\mathbf{f}_{(m)}$ are given by

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{(m)} &= \left(\hat{\mathbf{Z}}_{(m)}^T \hat{\mathbf{Z}}_{(m)} \right)^{-1} \hat{\mathbf{Z}}_{(m)}^T \left(\mathbf{I} - \hat{\mathbf{K}}_{(m)} \right) \mathbf{Y}, \\ \hat{\mathbf{f}}_{(m)} \left(\hat{\boldsymbol{\xi}}_{(m)} \right) &= \hat{\mathbf{K}}_{(m)} \left(\mathbf{Y} - \mathbf{Z}_{(m)} \hat{\boldsymbol{\theta}}_{(m)} \right),\end{aligned}$$

where $\hat{\mathbf{Z}}_{(m)} = (\mathbf{I} - \hat{\mathbf{K}}_{(m)}) \mathbf{Z}_{(m)}$. Furthermore, the m -th estimator and the model averaging estimator for $\boldsymbol{\mu}$ are

$$\begin{aligned}\hat{\boldsymbol{\mu}}_{(m)} &= \hat{\mathbf{Z}}_{(m)} \left(\hat{\mathbf{Z}}_{(m)}^T \hat{\mathbf{Z}}_{(m)} \right)^{-1} \hat{\mathbf{Z}}_{(m)}^T \left(\mathbf{I} - \hat{\mathbf{K}}_{(m)} \right) \mathbf{Y} + \hat{\mathbf{K}}_{(m)} \mathbf{Y} \equiv \hat{\mathbf{P}}_{(m)} \mathbf{Y}, \\ \hat{\boldsymbol{\mu}}(\boldsymbol{\omega}) &= \sum_{m=1}^M \omega_m \hat{\boldsymbol{\mu}}_{(m)} = \sum_{m=1}^M \omega_m \hat{\mathbf{P}}_{(m)} \mathbf{Y} = \hat{\mathbf{P}}(\boldsymbol{\omega}) \mathbf{Y},\end{aligned}$$

where $\hat{\mathbf{P}}(\boldsymbol{\omega}) = \sum_{m=1}^M \omega_m \hat{\mathbf{P}}_{(m)}$. Denote $\bar{\mathbf{P}}_{(m)} \equiv \hat{\mathbf{Z}}_{(m)} (\hat{\mathbf{Z}}_{(m)}^T \hat{\mathbf{Z}}_{(m)})^{-1} \hat{\mathbf{Z}}_{(m)}^T$, which is still idempotent, and $\hat{\mathbf{P}}_{(m)} \equiv \bar{\mathbf{P}}_{(m)} (\mathbf{I} - \hat{\mathbf{K}}_{(m)}) + \hat{\mathbf{K}}_{(m)}$.

Additionally, the modified loss, the conditional risk, and the Mallows-type criterion are transformed into

$$\begin{aligned}\widehat{L}_n(\boldsymbol{\omega}) &= \|\widehat{\boldsymbol{\mu}}(\boldsymbol{\omega}) - \boldsymbol{\mu}\|^2 = \|\widehat{\mathbf{P}}(\boldsymbol{\omega})\mathbf{Y} - \boldsymbol{\mu}\|^2, \\ \widehat{R}_n(\boldsymbol{\omega}) &= \left\| \left\{ \widehat{\mathbf{P}}(\boldsymbol{\omega}) - \mathbf{I} \right\} \boldsymbol{\mu} \right\|^2 + \text{tr} \left\{ \widehat{\mathbf{P}}^T(\boldsymbol{\omega}) \widehat{\mathbf{P}}(\boldsymbol{\omega}) \boldsymbol{\Omega} \right\}, \\ \widehat{C}_n(\boldsymbol{\omega}) &= \|\mathbf{Y} - \widehat{\boldsymbol{\mu}}(\boldsymbol{\omega})\|^2 + 2\text{tr} \left\{ \widehat{\mathbf{P}}(\boldsymbol{\omega}) \boldsymbol{\Omega} \right\}.\end{aligned}$$

Let $\widetilde{\boldsymbol{\omega}} = \operatorname{argmin}_{\boldsymbol{\omega} \in \mathcal{H}_n} \widehat{C}_n(\boldsymbol{\omega})$. Note that the covariance matrix $\boldsymbol{\Omega}$ is typically unknown in practice. Hence, we should estimate $\boldsymbol{\Omega}$ to obtain a computationally feasible criterion. Following Hansen (2007), we estimate $\boldsymbol{\Omega}$ based on the largest candidate model indexed by $M^* = \operatorname{argmax}_{1 \leq m \leq M} (p_m + q_m)$, leading to an estimator given by

$$\widehat{\boldsymbol{\Omega}} = \operatorname{diag} \left(\hat{\epsilon}_{(M^*),1}^2, \dots, \hat{\epsilon}_{(M^*),n}^2 \right), \quad (2.2)$$

where $(\hat{\epsilon}_{(M^*),1}, \dots, \hat{\epsilon}_{(M^*),n})^T = \mathbf{Y} - \widehat{\boldsymbol{\mu}}_{(M^*)}$.

With $\boldsymbol{\Omega}$ replaced by $\widehat{\boldsymbol{\Omega}}$, we select the optimal weights as follows,

$$\begin{aligned}\widehat{\boldsymbol{\omega}} &= \operatorname{argmin}_{\boldsymbol{\omega}} \widehat{C}_n(\boldsymbol{\omega})|_{\boldsymbol{\Omega}=\widehat{\boldsymbol{\Omega}}} \\ &= \operatorname{argmin}_{\boldsymbol{\omega}} \|\mathbf{Y} - \widehat{\boldsymbol{\mu}}(\boldsymbol{\omega})\|^2 + 2\text{tr} \left\{ \widehat{\mathbf{P}}(\boldsymbol{\omega}) \widehat{\boldsymbol{\Omega}} \right\},\end{aligned} \quad (2.3)$$

which can be treated as a feasible counterpart of $\widehat{C}_n(\boldsymbol{\omega})$. Let $\mathbf{H} = (\mathbf{Y} - \widehat{\boldsymbol{\mu}}_{(1)}, \dots, \mathbf{Y} - \widehat{\boldsymbol{\mu}}_{(M)})$ and $\mathbf{b} = \{\text{tr}(\widehat{\mathbf{P}}_{(1)} \widehat{\boldsymbol{\Omega}}), \dots, \text{tr}(\widehat{\mathbf{P}}_{(M)} \widehat{\boldsymbol{\Omega}})\}^T$. It is clear that (2.3) is a standard quadratic programming problem in the form of

$$\begin{aligned}\min_{\boldsymbol{\omega}} \widehat{C}_n(\boldsymbol{\omega})|_{\boldsymbol{\Omega}=\widehat{\boldsymbol{\Omega}}} &= \min_{\boldsymbol{\omega}} \boldsymbol{\omega}^T \mathbf{H}^T \mathbf{H} \boldsymbol{\omega} + 2\boldsymbol{\omega}^T \mathbf{b} \\ \text{subject to } \mathbf{1}^T \boldsymbol{\omega} &= 1 \text{ and } \boldsymbol{\omega} \geq 0,\end{aligned}$$

where $\mathbf{1}$ is a vector with all entries equal to 1. The problem can be efficiently optimized by the R package *quadprog*.

Remark 2. Our work exhibits several distinctions when compared to model averaging approaches to scalar regressions with a partially linear structure. Notably, Zhang and Wang (2019) as well as Zhu et al. (2019) investigated Mallows-type model averaging estimators for partially linear models and varying-coefficient partially linear models, respectively. Both our studies and theirs employed the profile least squares estimation method with kernel smoothing techniques. However, their treatment of scalar variables does not entail the subsequent challenges associated with handling functional predictors, a pivotal focus in our own research. Regarding the source of model uncertainty, while they deliberated on both which predictors are included in candidate models and which are allocated to linear or non-linear parts, our natural division between scalar and

functional predictors avoids this kind of uncertainty. We concentrate solely on determining which components are retained in the models.

2.3. Implementation details

2.3.1. Model preparation

A key problem in implementing the proposed model averaging method lies in the preparation of candidate models. Depending on the context, the candidate models can be prepared in various ways. In specialized fields like economics and finance, candidate models are often postulated based on different theories for prediction. This entails using prior knowledge of the model setup. Without these expert theories, all possible specifications of predictors can be considered when preparing candidate models.

Typically, with p variables in $\mathbf{z}$ and q scores in $\boldsymbol{\xi}$, we have a total number of $M = (2^p - 1) \cdot (2^q - 1)$ candidate models, requiring that at least one component from each part is included. However, when p or q is large, estimating and averaging all possible candidate models is computationally infeasible. Hence, a model screening step prior to model averaging is more desirable. For example, a backward elimination procedure before model averaging was employed by Zhang, Wan and Zhou (2012), but it may still be computationally burdensome when p is large. In addition, the well-known screening procedure based on marginal correlations between the predictors and the response (Fan and Lv, 2008; Fan and Song, 2010) was performed to prepare candidate models in Ando and Li (2014). This ordering-based model screening strategy was also adopted by Zhang et al. (2016). Moreover, Zhang et al. (2016) advocated a top m model screening approach, which use penalized regression with various tuning parameters to screen out some candidate models. More recently, Zhang et al. (2020) utilized the order of entering the solution path of penalized regression to sequence predictors, and then prepared candidate models in a nested manner. Zou et al. (2022) ordered the covariates first based on their marginal correlations with the dependent variable, and constructed the candidate model by including one extra covariate at each time based on the ordering.

In our context, when both p and q are small, we consider all possible candidate models with diverse specifications in $\mathbf{z}$ and $\boldsymbol{\xi}$, ensuring the inclusion of at least one component in each part. This leads to a total of $M = (2^p - 1) \cdot (2^q - 1)$ candidate models. In cases where p or q is large, we may adopt a pre-screening strategy that combines ordering screening and threshold screening. Specifically, we arrange the scalar variables based on their marginal correlation with Y , and screen out the most informative FPC scores, which collectively account for a certain proportion of explained variance in FPCA. Subsequently, we construct candidate models in a nested fashion by incorporating the first few scalar predictors and transformed FPC scores in the ordering. Denote by p_{sc} and q_{sc} the numbers of screened out scalar predictors and transformed FPC scores.

Finally, we have $p_{sc} \times q_{sc}$ candidate models in all, which significantly eases the computational load.

2.3.2. Multiple functional predictors

If more than one functional predictor is available, denoted as $X_1(t), \dots, X_g(t)$, we may extend our model (1.3) and (2.1) to an additive form:

$$\begin{aligned} Y &= \boldsymbol{\theta}^T \mathbf{z} + m_1(X_1) + \dots + m_g(X_g) + \varepsilon, \\ Y &= \boldsymbol{\theta}^T \mathbf{z} + \mathbf{f}_1(\boldsymbol{\xi}_1) + \dots + \mathbf{f}_g(\boldsymbol{\xi}_g) + \varepsilon, \end{aligned}$$

where $\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_g$ are vectors of transformed FPC scores corresponding to $X_1(t), \dots, X_g(t)$, and $\mathbf{f}_1, \dots, \mathbf{f}_g$ represent unknown functions for nonparametric modeling. The approach of additive modeling for functional predictors, $m_1(X_1) + \dots + m_g(X_g)$, rather than $m(X_1, \dots, X_g)$, serves to alleviate the issue of the “curse of dimensionality” that often arises in nonparametric statistics. Regarding the estimation for $\mathbf{f}_1, \dots, \mathbf{f}_g$, we can also perform profile kernel smoothing method outlined above. The advantage of this method is its computational efficiency, as the estimator can be computed without iteration. However, as discussed in Speckman (1988), it may produce different estimators depending on the order of $\mathbf{f}_j$, due to its hierarchical nature. Consequently, in model averaging context, a more efficient and stable estimator needs to be explored, and we leave this for future work.

As for the kernel function employed in smoothing, we also utilize a product kernel function $\mathcal{K}_{j,h_j}(u) = \prod_{l=1}^{q_j} k_{j,h_j}(u)$ to address each $\mathbf{f}_j(\boldsymbol{\xi}_j)$. Here, $k_{j,h_j}(\cdot)$ represents a univariate kernel function and h_j is a scalar bandwidth. For small q_j , second-order kernel functions are commonly used in the field. If q_j increases, the application of higher-order kernel functions, such as fourth or sixth-order, can help reduce smoothing bias, albeit at the cost of increased variance. In our development, we advocate performing a pre-screening procedure before model averaging when many $\{\xi_{j,l}\}$ are available. This helps alleviate estimation bias for the nonparametric part.

3. Asymptotic Optimality

Define $\eta_n = \inf_{\boldsymbol{\omega}} R_n(\boldsymbol{\omega})$, and let $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ denote the largest and smallest singular value of a matrix, respectively. Let $\boldsymbol{\omega}_m^0$ be a weight vector in which the m -th component is 1 and the others are 0. Let $\tilde{p} = \max_m p_m$, $\tilde{q} = \max_m q_m$, and $h = \min_m h_m$. We focus on $X(t)$, which is densely observed with noise, and consider a fixed M here. The following assumptions are required for the model averaging estimator to achieve asymptotic optimality.

Assumption 1.

(a) *The eigenvalue sequence $\{\lambda_l\}$ of $X(t)$ satisfies*

$$\begin{aligned} c_\lambda^{-1} l^{-\alpha} &\leq \lambda_l \leq C_\lambda l^{-\alpha}, \\ \lambda_l - \lambda_{l+1} &\geq C_\lambda^{-1} l^{-1-\alpha}, \quad l = 1, 2, \dots, \end{aligned}$$

where c_λ and C_λ represent generic positive constants, and $\alpha > 1$ to ensure $\sum_{l=1}^{\infty} \lambda_l < \infty$.

(b) $\mathbb{E}(\|X\|^4) < \infty$ *where $\|X\| = \{\int_{\mathcal{T}} X^2(t) dt\}^{1/2}$, and there exists a constant $C_\zeta > 0$ such that $\mathbb{E}(\zeta_l^2 \zeta_j^2) \leq C_\zeta \lambda_l \lambda_j$ and $\mathbb{E}(\zeta_l^2 - \lambda_l)^2 < C_\zeta \lambda_l^2$, $\forall l \neq j$.*

Assumption 1 imposes mild restrictions on $X(t)$, which are widely used in functional regressions, see Cai and Hall (2006) and Cai and Yuan (2012). Assumption 1(a) assumes that the eigenvalues decay at a polynomial rate, which is a relatively slow rate, allowing $X(t)$ to be flexibly modelled as a L^2 process. Moreover, it requires that the spacings among eigenvalues not be too small to ensure the identifiability and consistency of sample eigenvalues and eigenfunctions. Assumption 1(b), same as Assumption 2 in Wong, Li and Zhu (2019), places a weak moment restriction on $X(t)$, which is easily satisfied when $X(t)$ is a Gaussian process.

Assumption 2.

(a) *The kernel function $k(\cdot)$ is a bounded symmetric density with compact support and continuously bounded first derivative function.*

(b) $\max_i \sum_{j=1}^n |K_{(m),ij}| = O(1)$ (a.s.), $\max_j \sum_{i=1}^n |K_{(m),ij}| = O(1)$ (a.s.), *uniformly for $m = 1, \dots, M$.*

Assumption 2 imposes specific requirements on the kernel estimation method. Assumption 2(a) is commonly used for kernel functions and is met by various types, including second-order uniform, Epanechnikov, and quartic kernels. Assumption 2(b) bounds the elements of the smoother matrix, which has been discussed in Speckman (1988) and is similar to condition 1 in Zhang and Wang (2019).

Assumption 3.

(a) *For some integer $G \geq 1$, $\max_i \mathbb{E}(\varepsilon_i^{4G} | \mathbf{Z}_i, X_i) < \infty$ (a.s.), $i = 1, \dots, n$.*

(b) $M \eta_n^{-2G} \sum_{m=1}^M \{R_n(\boldsymbol{\omega}_m^0)\}^G = o_p(1)$.

(c) $\tilde{q} = O(n^{1/(2+2\alpha)})$ *where α relates to the decay rate of eigenvalues $\{\lambda_l\}$ in Assumption 1, $n^{1/2} \eta_n^{-1} \tilde{q} = o_p(1)$, $\eta_n^{-1} \tilde{q}^2 = o_p(1)$.*

(d) $\|\boldsymbol{\mu}\|^2/n = O(1)$, a.s.

- (e) *The smallest nonzero singular value of the kernel smoother $\mathbf{K}_{(m)}$, i.e., the square root of nonzero eigenvalue of $\mathbf{K}_{(m)}^T \mathbf{K}_{(m)}$, is bounded away from 0 (a.s.), for all $m = 1, \dots, M$.*

Parts (a), (b), and (d) of Assumption 3 are standard conditions for model averaging. Assumption 3(a) constrains the conditional moment of random errors, which is easily satisfied by Gaussian error; see Hansen (2007) and Zhang, Zou and Liang (2014). Assumption 3(b) is a common convergence condition in the literature, which requires that η_n goes to infinity rapidly enough, implying that there is no simple approximating model with zero bias. It holds in scenarios where all candidate models are misspecified, as indicated by Wan, Zhang and Zou (2010), Zhang and Wang (2019), Zhu et al. (2019), and others. Assumption 3(c) regulates the growth rate of the number of FPC scores, which guarantees effective estimation accuracy. Recall that $\alpha > 1$, so that the order of $\tilde{q}$ is smaller than $n^{-1/4}$, which may also help alleviate fitting issues in kernel smoothing. The remaining two restrictions in Assumption 3(c) are technical conditions that limit the dimensionality of the nonparametric part. Specifically, they ensure that the diverging rate of $\tilde{q}$ is constrained by the diverging rate of η_n as n approaches infinity. If η_n goes to infinity at a rate greater than $n^{1/2}$, then $\eta_n^{-1} \tilde{q}^2 = o_p(1)$ is implied directly. Assumption 3(d) concerns the sum of μ_i^2 and is commonly used in regression contexts, see Liang et al. (2011). Assumption 3(e) serves as a technical condition ensuring the stability of estimation for $\hat{\boldsymbol{\theta}}_{(m)}$ and $\hat{\mathbf{f}}_{(m)}(\hat{\xi}_{(m)})$. It further implies an intermediate result, as illustrated below in Lemma 1.

Lemma 1. *Under Assumptions 1, 2, 3(c) and 3(e), we have*

$$\lambda_{\max}(\mathbf{P}_{(m)} - \hat{\mathbf{P}}_{(m)}) = O_p(n^{-1/2} q_m),$$

for all $m = 1, \dots, M$.

Lemma 1 demonstrates that the difference between $\mathbf{P}_{(m)}$ and its estimate $\hat{\mathbf{P}}_{(m)}$ diminishes as n approaches infinity. This suggests that the perturbation stemming from our estimated FPC scores can be bounded, provided certain conditions met. Unlike Zhu, Zou and Zhang (2018), who did not account for the approximation effect of their estimated FPC scores and directly derived asymptotic optimality assuming that true $\mathbf{P}_{(m)}$ and $\mathbf{P}(\boldsymbol{\omega})$ are obtained based on the estimated FPC scores, our approach considers the estimation influence. Other works, addressing estimation influences in different model setups such as the estimation of the variance of random coefficients in mixed-effects models and the auto-regression coefficient in spatial auto-regressive models, often employ $\lambda_{\max}(\mathbf{P}_{(m)} - \hat{\mathbf{P}}_{(m)}) = o_p(1)$ as a high-level and technical condition, see Zhang, Zou and Liang (2014) and Zhang and Yu (2018). Here, we choose less intricate but equally effective assumptions, which include mild restrictions on $X(t)$ and the kernel method, as well as several order requirements on the growth rate of the

number of FPC scores and a standard condition regarding the smallest singular value of the matrix $\mathbf{K}_{(m)}$. We derive this lemma as an intermediate result instead of listing it as a condition.

We now provide the asymptotic optimality of the model averaging estimator when $\mathbf{\Omega}$ is known.

Theorem 1. *Under Assumptions 1–3, it holds that*

$$\frac{L_n(\tilde{\omega})}{\inf_{\omega \in \mathcal{H}_n} L_n(\omega)} \rightarrow 1 \quad (3.1)$$

in probability as $n \rightarrow \infty$.

Theorem 1 illustrates the asymptotic optimality of $\tilde{\omega}$ in the sense that the squared loss based on the weight vector $\tilde{\omega}$ is asymptotically identical to that obtained using the infeasible optimal weight vector if $\mathbf{\Omega}$ is known.

Following Liu and Okui (2013), we process $\text{tr}\{\hat{\mathbf{P}}(\omega)\mathbf{\Omega}\}$ as one entity rather than considering $\mathbf{\Omega}$ in isolation, and estimate it by $\sum_{i=1}^n \hat{\epsilon}_{(M^*),i}^2 \rho_{ii}(\omega)$, where $\rho_{ii}(\omega)$ is the i -th diagonal element of $\hat{\mathbf{P}}(\omega)$. Denote $\rho_{ii}^{(m)}$ as the i -th diagonal element of $\hat{\mathbf{P}}_{(m)}$. When $\mathbf{\Omega}$ is replaced by its estimate $\hat{\mathbf{\Omega}}$ given in (2.2), provided that the following additional assumptions are satisfied, it can be shown that the model averaging estimator based on $\hat{\omega}$ shares the same asymptotic optimality as $\tilde{\omega}$ in Theorem 1.

Assumption 4.

- (a) *There exists a constant c such that $\max_i \rho_{ii}^{(m)} \leq cn^{-1} |\text{tr}(\hat{\mathbf{P}}_{(m)})|$ (a.s.), uniformly for $m = 1, \dots, M$.*
- (b) *$\text{tr}(\mathbf{K}_{(m)}) = O(h^{-\tilde{q}})$ (a.s.), uniformly for $m = 1, \dots, M$.*
- (c) *$\eta_n^{-1} \tilde{p} = o_p(1)$ and $\eta_n^{-1} h^{-\tilde{q}} = o_p(1)$.*

Assumption 4(a), same as Condition 5 in Zhang and Wang (2019) and Condition (C.6) in Zhu et al. (2019), means that there should not be any dominant or strongly influential subjects as shown in Li (1987) and Andrews (1991). Assumption 4(b) is typically used in kernel smoothing technique; see Condition (h) of Speckman (1988) and Condition 4 of Zhang and Wang (2019). Assumption 4(c), similar to Condition (C.9) of Zhu et al. (2019) and Condition 3 of Zhang, Chiou and Ma (2018), places additional restrictions on the growth rate of the dimensionality of scalar predictors and FPC scores.

Theorem 2. *Under Assumptions 1–4, we have that*

$$\frac{L_n(\hat{\omega})}{\inf_{\omega \in \mathcal{H}_n} L_n(\omega)} \rightarrow 1 \quad (3.2)$$

in probability as $n \rightarrow \infty$.

Theorem 2 shows that Theorem 1 remains valid when $\mathbf{\Omega}$ is replaced by $\hat{\mathbf{\Omega}}$. Thus, the practically feasible $\hat{\mathbf{\omega}}$ also enjoys the asymptotic optimality. The Supplementary Material provides the detailed proof for Theorems 1 and 2.

4. Simulation Study

In this section, we compare the finite sample performance of the proposed Mallows-type model averaging (MMA) estimator to several popular model selection and averaging estimators, including AIC, BIC, equally weighting, SAIC and SBIC (Buckland, Burnham and Augustin, 1997). For the m -th candidate model, AIC and BIC select the model with the smallest scores, defined as $AIC_m = \log(\hat{\sigma}_m^2) + 2tr(\hat{\mathbf{P}}_{(m)})/n$ and $BIC_m = \log(\hat{\sigma}_m^2) + \log(n)tr(\hat{\mathbf{P}}_{(m)})/n$, where $\hat{\sigma}_m^2 = \|\mathbf{Y} - \hat{\boldsymbol{\mu}}_{(m)}\|^2/n$. Equally weighting simply assigns uniform weights of $1/M$ to each candidate model. SAIC and SBIC assign weights to the m -th candidate as $\omega_m^{AIC} = \exp(-AIC_m/2)/\sum_{m=1}^M \exp(-AIC_m/2)$ and $\omega_m^{BIC} = \exp(-BIC_m/2)/\sum_{m=1}^M \exp(-BIC_m/2)$ respectively. Additionally, we also compare to the oracle MMA (oMMA) estimator, assuming the variance matrix $\mathbf{\Omega}$ of the random error vector ε is known. Furthermore, both the MMA and oracle MMA estimators derived from model (1.2) (MMA-lin and oMMA-lin) are also performed to illustrate the benefit of nonparametric modeling in detecting potential non-linear effect of $X(t)$. This helps to further illustrate the modeling difference between our approach and that of Zhu, Zou and Zhang (2018).

The data is generated from the following PLFS model,

$$Y_i = \mu_i + \varepsilon_i = \sum_{j=1}^{M_0} \theta_j Z_{ij} + \mathbf{f}(\boldsymbol{\xi}_i) + \varepsilon_i, \quad i = 1, \dots, n. \quad (4.1)$$

The vector $\boldsymbol{\xi}_i$ represents the transformed FPC scores derived from ζ_i , where ζ_{il} is independently generated from $N(0, \lambda_l)$. Here, the standard Gaussian CDF, denoted as $\Phi(\cdot)$, is employed as the transformation. We now consider the following examples. Two additional examples are provided in the Supplementary Material.

Example 1. To illustrate the effectiveness of nonparametric modeling of the functional predictor, we perform a comparison under linear effect of $X_i(t)$. $M_0 = 4$ and $\boldsymbol{\theta} = (1.5, 0.7, 0.2, -0.4)^T$. $\mathbf{Z}_i$ is a 4×1 vector that follows a multivariate normal distribution with zero means and a variance-covariance matrix $\Sigma = (0.5^{|a-b|})_{4 \times 4}$. The functional predictor $X_i(t)$ is obtained by

$$X_i(t) = \sum_{l=1}^4 \zeta_{il} \psi_l(t), \quad t \in [0, 1],$$

where $\psi_l(t) = \sqrt{2} \sin(\pi l t)$, $l = 1, \dots, 4$, and ζ_{il} is i.i.d. and simulated from $N(0, l^{-3/2})$, $i = 1, \dots, n$. The random error term ε_i is i.i.d. and follows $N(0, \eta^2)$.

η controls the signal-to-noise ratio and we vary it such that $R^2 = \text{var}(\mu_i)/\text{var}(Y_i)$ ranges from 0.1 to 0.9, where $\text{var}(\mu_i)$ and $\text{var}(Y_i)$ denote the variances of μ_i and Y_i , respectively. In this scenario, data is generated from the following process,

$$Y_i = \sum_{j=1}^4 \theta_j Z_{ij} + \int_0^1 X_i(t) \beta(t) dt + \varepsilon_i,$$

where the coefficient function $\beta(t) = 1 + \log(1+t)$, $t \in [0, 1]$.

Example 2. $M_0 = 50$ and $\theta_j = j^{-2/3}$. Consider $\mathbf{z}$ and $X(t)$ being correlated. Simulate $(\mathbf{Z}_i, \zeta_{i1}) \sim MN(0, \Sigma)$, where $\Sigma = (0.5^{|a-b|})_{51 \times 51}$. The functional predictor $X_i(t)$ is generated from

$$X_i(t) = \sum_{l=1}^{10} \zeta_{il} \psi_l(t), \quad t \in [0, 1],$$

where $\psi_l(t) = \sqrt{2} \sin(\pi l t)$, $l = 1, \dots, 10$, and ζ_{il} is i.i.d. and follows $N(0, l^{-3/2})$, $i = 1, \dots, n$, $l = 2, \dots, 10$. Consider another type of heteroskedasticity for independent random errors as $\varepsilon_i \sim N\{0, \eta^2(Z_{i1}^2 + 0.01)\}$. Varying η such that R^2 ranging from 0.1 to 0.9. And the non-linear effect of $X_i(t)$ is introduced by

$$\mathbf{f}(\boldsymbol{\xi}_i) = \exp\left(\sum_{l=1}^{10} \frac{\xi_{il}}{l}\right),$$

where $\xi_{il} = \Phi(\lambda_l^{-1/2} \zeta_{il})$.

Assume that $X(t)$ is observed at 100 equally-spaced grids on the defined interval with measurement error. Denote the i -th observation of X at time point t_j by $X_{ij} = X_i(t_j) + e_{ij}$, where measurement errors e_{ij} 's are independent $N(0, 0.2)$ variables. The sample size is set as $n = 50, 100, 200$ and 400.

In Example 1, we omit z_4 and ξ_4 in preparing candidate models, so all candidate models are misspecified. With different specifications of which elements in $\{z_1, z_2, z_3\}$ and $\{\xi_1, \xi_2, \xi_3\}$ are included in the model, we have a total number of $M = 49$ candidate models for Example 1. As for Example 2, we first conduct pre-screening and then construct candidate models in a nested way. That is, order scalar predictors according to their marginal correlations to the response Y , and screen out the first $p_{sc} = \lceil n^{1/3} \rceil$ variables, where $\lceil x \rceil$ denotes the smallest integer larger than x . Besides, the first q_{sc} FPC scores which account for as least 85% of the cumulative variance explained proportion are picked out in order. Then, we set each candidate model include the first a scalar predictors and the first b transformed FPC scores in the ordering. As a result, we have $p_{sc} \times q_{sc}$ candidate models in all.

We employ the Epanechnikov kernel $k(u) = 3(1 - u^2)I(|u| \leq 1)/4$ for illustration, with the bandwidth h_m set to $n^{-1/(1+q_m)}$ based on the rule-of-thumb

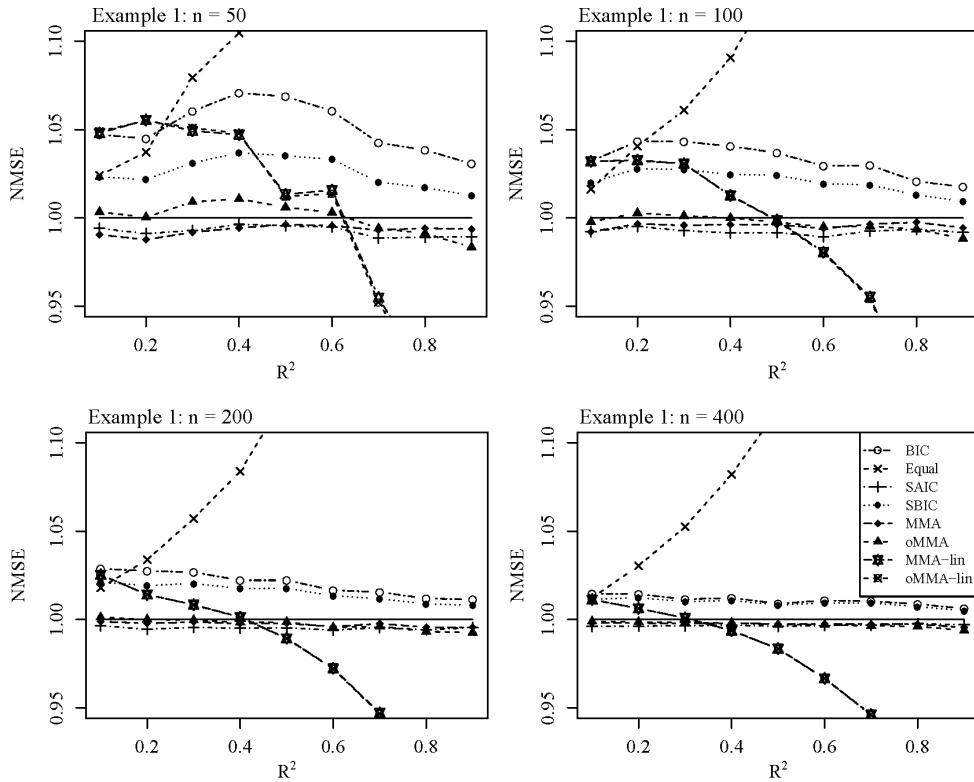


Figure 1. Normalized mean squared error (NMSE) comparisons for Example 1.

method, $m = 1, \dots, M$. The mean squared error (MSE) of each method is presented,

$$MSE = \frac{1}{nD} \sum_{d=1}^D \left\| \hat{\boldsymbol{\mu}}^{(d)} - \boldsymbol{\mu}^{(d)} \right\|^2,$$

where $D = 200$ denotes the number of repetitions and d represents the d -th trial. For easy comparison, all MSE's are normalized by dividing by the MSE of AIC model selection estimator. Thus, a normalized MSE (NMSE) smaller than 1 indicates the corresponding estimator is superior to AIC estimator, and vice versa.

Figures 1 and 2 depict the corresponding results for Examples 1 and 2. In Example 1, we investigate the performances of all methods under a partially linear functional linear model. As shown in Figure 1, both the MMA-lin and oMMA-lin outperform the other methods, particularly for medium and large R^2 values, underscoring the effectiveness and efficiency of linear modeling when linear effects are prevalent in the data.

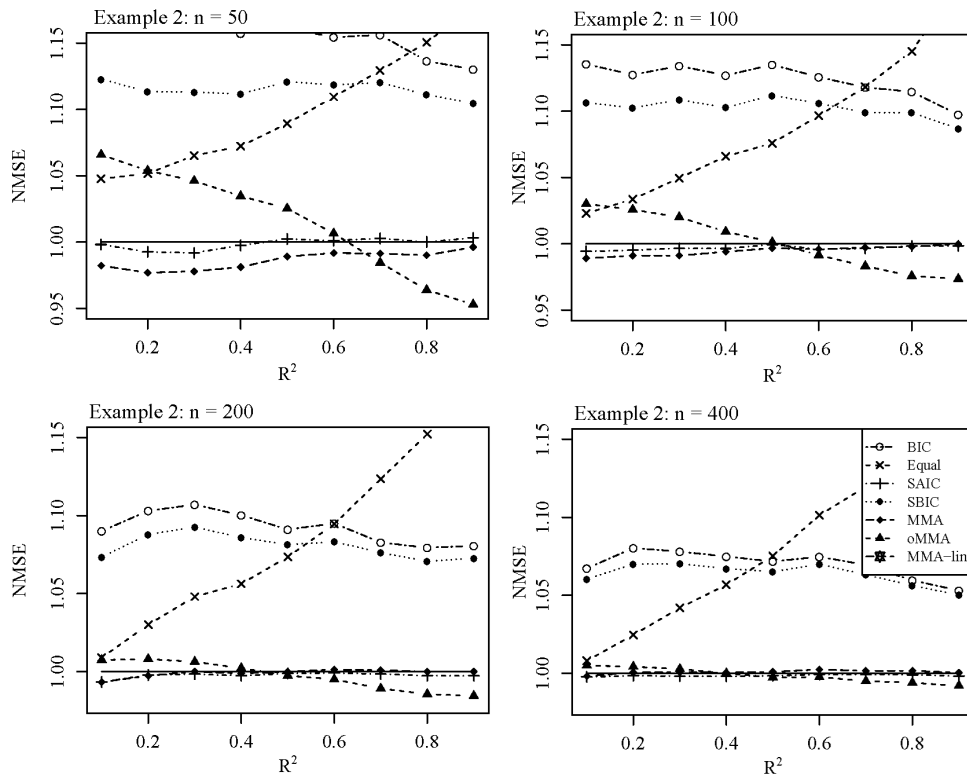


Figure 2. Normalized mean squared error (NMSE) comparisons for Example 2.

In Example 2, we consider a more intricate data generation process. The true models incorporate heteroscedastic random errors, a departure from Zhu, Zou and Zhang (2018) which primarily addresses homoscedastic error term. Hence, we omit the comparison to the oMMA-lin method here. In Example 2, both the MMA and oMMA estimators exhibit a clear advantage over other estimates, for small and large values of R^2 , respectively. However, MMA-lin performs worse in Example 2, with its lines entirely falling outside each subfigure. This could be attributed to the much stronger non-linear effects in Example 2, where linear modeling encounters challenges in dealing with this type of effects.

The oMMA outperforms MMA for large R^2 values, whereas MMA excels for smaller R^2 values. This discrepancy can be attributed to the signal-to-noise ratio inherent in the dataset. Specifically, when the signal predominates, the oracle knowledge of the variance matrix Ω significantly enhances performance. Conversely, when noise prevails, the oMMA estimator can capture the information regarding the error terms but may struggle to precisely quantify the magnitude of approximation error. This limitation could potentially compromise oMMA's ability to effectively distinguish between the signal (in candidate models) and the approximation. In contrast, MMA employs an estimated variance matrix derived

from the largest candidate model, which mitigates this issue. Additionally, when sample size n is limited, MMA exhibits an advantage over AIC and SAIC methods. With increasing n , these estimators tend to behave similarly in most cases for R^2 . On the other hand, BIC and SBIC yield less satisfactory results in our framework, as their parsimonious nature does not align with our objective of minimizing MSE. Furthermore, equally weighting performs worse in each case, suggesting caution in its use especially when dealing with a strong signal.

To summarize, both the proposed MMA and oMMA estimators demonstrate competitive performance when compared to the other estimators across most scenarios. Their superior performance can be attributed, in part, to the fact that their optimality does not hinge on the correct specification of candidate models. This means that the true model needs not be included in the set of candidate models. In addition, equally weighting method generally yields poorer outcomes in most cases, especially as R^2 increases. Both BIC and SBIC display poor performance in our simulations, likely due to their parsimonious nature not aligning with our objective. It is noteworthy that the model averaging estimators, SAIC and SBIC, consistently outperform their model selection counterparts, AIC and BIC. The differences between AIC and SAIC, as well as BIC and SBIC, tend to diminish as R^2 increases. This underscores the potential capability of model averaging approach in real-world data analysis. Furthermore, the comparisons between MMA-type and MMA-lin-type estimators highlight the necessity of nonparametric modeling in certain scenarios.

5. Application to Real Data

In this section, we illustrate the application of our proposed method using a dataset on soil properties analysis. The dataset includes mid-infrared spectroscopy measurements (functional predictor X) of 160 soil samples collected in western Kenya. These measurements were obtained within the spectrum range of 7,498 to 600 cm^{-1} . Additionally, the dataset contains reference measurements, including the total carbon percentage (Y), as well as 19 other covariates ($\mathbf{z}$), such as exchangeable aluminium ($mg \cdot kg^{-1}$), boron concentration ($mg \cdot kg^{-1}$), and the exchangeable calcium-to-magnesium ratio. Further details can be found in Sila, Pokhariyal and Shepherd (2017). With the inclusion of numerous variables, a large amount of model uncertainty arises, as the empirical results may vary across different model specifications. Consequently, model averaging is likely to gain an advantage in such situations.

We retained a sample of 158 observations after removing all incomplete records. To assess the performance of our proposed procedure, we randomly allocated 90% of the records (143) as the training set and constructed the test set using the remaining data points (15). Following Sila, Pokhariyal and Shepherd (2017), we log-transformed all scalar variables except pH values and

Table 1. Average MSPEs, median MSPEs, and their standard errors on the test set

	AIC	BIC	Equal	SAIC	SBIC	MMA	MMA-lin
Mean	0.339	0.339	0.377	0.332	0.335	0.329	0.347
Median	0.329	0.323	0.359	0.320	0.323	0.316	0.332
Ste	0.119	0.121	0.145	0.117	0.116	0.117	0.126

then standardized them. For $X(t)$, we utilized the leading 3 FPC scores for modeling, which collectively account for at least 90% of the explained variance in the functional data from the training set and do not exceed the order $\lceil n^{1/(2+\alpha)} \rceil$, as listed in Assumption 3(c). Given the relatively large number of scalar predictors, we initially screened out $\mathbf{z}$ whose absolute marginal correlation exceeds 0.1, resulting in 13 remaining covariates (see details in the Supplementary Material). We then adopted a nested approach for preparing candidate models. Each candidate model includes one extra component from both the parametric and nonparametric parts. We conducted $D = 1,000$ runs, and for each repetition, we evaluated the mean squared prediction error (MSPE) using

$$MSPE^{(d)} = \frac{1}{n_{test}} \sum_{i \in TEST_d} \left(Y_i^{(d)} - \hat{\mu}_i^{(d)} \right)^2, \quad d = 1, \dots, D,$$

where n_{test} denotes the sample size of test set, $Y_i^{(d)}$ represents the i -th response of the test set $TEST_d$, and $\hat{\mu}_i^{(d)}$ signifies the prediction for $Y_i^{(d)}$.

The boxplots and the empirical cumulative distribution functions of $\{MSPE^{(d)}\}$ across D runs are displayed in the Supplementary Material. Table 1 presents the average and median MSPEs, as well as their standard errors across D repetitions. It is worth noting that MMA yields the smallest average and median MSPEs, which demonstrates its enhanced predictive capability compared to other model averaging and selection estimators. SAIC follows as the second-best performer with slightly larger average and median MSPEs, but still inferior to MMA. AIC exhibits a bit larger average MSPE than SAIC and MMA. Additionally, BIC and SBIC show similar performances to AIC and SAIC on this dataset, while the MSPEs of SBIC are slightly larger than SAIC. The equally weighting method delivers significantly larger MSPEs, reinforcing the need for cautious application in practice. Furthermore, MMA-lin exhibits poor MSPE on this dataset, suggesting that nonparametric modeling proves more effective than linear modeling for this data. Notably, the average and median MSPEs of model averaging estimators are mostly smaller than those of their model selection counterparts. This underscores that, when prediction performance is of primary interest, model averaging stands as a favorable alternative to model selection.

To further highlight the superiority of MMA, we conducted a Diebold–Mariano (DM) test (Diebold and Mariano, 2002) and a data-driven approach

Table 2. Results of test statistics and adjusted p-values.

	MMA/AIC	MMA/BIC	MMA/Equal
DM stat.	-11.66	-7.356	-13.55
DM p.val	42.37E-29	3.93E-13	44.81E-38
t stat.	-13.93	-10.08	-16.50
t p.val	43.03E-40	8.37E-23	8.30E-54
	MMA/SAIC	MMA/SBIC	MMA/MMA-lin
DM stat.	-4.110	-5.128	-6.293
DM p.val	2.14E-05	2.11E-07	3.48E-10
t stat.	-6.597	-7.258	-8.713
t p.val	3.40E-11	4.72E-13	9.12E-18

(Racine and Parmeter, 2014), employing paired t -tests and Mann–Whitney–Wilcoxon (MWW) tests. These methods are designed to test whether two competing approximate models are equivalent in terms of their prediction performances. Table 2 provides the test statistics for the first two tests, along with their corresponding p-values for one-sided tests. We include the corresponding results of MWW test in the Supplementary Material to save space. A negative value of test statistics indicates that the respective method is less accurate than MMA. Notably, all p-values have been adjusted using the Benjamini & Hochberg method (Benjamini and Hochberg, 1995) to account for multiple comparisons. In the context of one-sided test employed here, the alternative hypothesis suggests that the other method is less accurate than MMA.

All test statistics in Table 2 exhibit negative values, strongly indicating the superior prediction accuracy of MMA. Furthermore, the adjusted p-values for the one-sided tests approach zero, which demonstrate that MMA outperforms its competing approaches with high confidence. Collectively, these findings affirm that the proposed MMA procedure yields competitive outcomes in comparison to alternative methods.

6. Conclusion and Discussion

We have introduced a Mallows-type model averaging approach for PLFS models, which address the model uncertainty in the mixed-data scenario involving both scalar and functional predictors. The theoretical analysis has verified the asymptotic optimality of MMA estimator in the context of densely observed functional predictors with measurement error. Furthermore, our extensive numerical study has revealed that the performance of the proposed estimator surpasses that of classical competing model selection and averaging methods in various cases, particularly in scenarios characterized by a large degree of model uncertainty.

Several aspects deserve future research. Firstly, while we have advocated an additive modeling approach in scenarios involving multiple functional predictors,

further investigation is warranted to ascertain the most effective and efficient means for conducting model averaging in such cases. Additionally, the asymptotic optimality is derived under the assumption that all candidate models are misspecified. Recent work, as exemplified by Zhang et al. (2020), Zou et al. (2022), and Fang, Li and Xia (2022), has explored consistent properties in situations where correct models exist within the model space. Investigating the consistency of the model averaging approach in such contexts presents a promising yet challenging direction for future research. Lastly, while our present study focused on a scalar response, it is crucial to note that responses of binary, censored, and functional nature are prevalent in practical applications. Thus, extending the proposed methodology to accommodate these diverse response types is of great importance. This endeavor would broaden the applicability and impact of our approach in real-world settings.

Supplementary Material

The online Supplementary Material contains additional simulations, additional details in real application, and the detailed proofs for Lemma 1, Theorems 1 and 2.

Acknowledgments

The authors greatly appreciate the two referees and the associate editor for their insightful comments. The authors' departmental and research center affiliations are acknowledged as follows. Shishi Liu is affiliated with the School of Economics, Hangzhou Dianzi University. Chunming Zhang is affiliated with the Department of Statistics, University of Wisconsin. Hao Zhang is affiliated with the School of Statistics, Renmin University of China. Rou Zhong and Jingxiao Zhang are affiliated with the Center for Applied Statistics and the School of Statistics, Renmin University of China.

Shishi Liu's research was supported by the Zhejiang Provincial Department of Education Scientific Research Project (Y202147117), and the Zhejiang Provincial Natural Science Foundation of China (LQ23A010017). Chunming Zhang's research was supported in part by U.S. National Science Foundation grants DMS-2013486 and DMS-1712418, and by the University of Wisconsin-Madison Office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation. Jingxiao Zhang's research was supported by the Public Health & Disease Control and Prevention, the Fund for Building World-Class Universities (Disciplines) of Renmin University of China, and the MOE Project of Key Research Institute of Humanities and Social Sciences (22JJD910001).

References

- Ando, T. and Li, K.-C. (2014). A model-averaging approach for high-dimensional regression. *Journal of the American Statistical Association* **109**, 254–265.
- Andrews, D. W. (1991). Asymptotic optimality of generalized CL, cross-validation, and generalized cross-validation in regression with heteroskedastic errors. *Journal of Econometrics* **47**, 359–377.
- Aneiros-Pérez, G. and Vieu, P. (2006). Semi-functional partial linear regression. *Statistics & Probability Letters* **76**, 1102–1110.
- Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 289–300.
- Buckland, S. T., Burnham, K. P. and Augustin, N. H. (1997). Model selection: An integral part of inference. *Biometrics* **53**, 603–618.
- Cai, T. T. and Hall, P. (2006). Prediction in functional linear regression. *The Annals of Statistics* **34**, 2159–2179.
- Cai, T. T. and Yuan, M. (2012). Minimax and adaptive prediction for functional linear regression. *Journal of the American Statistical Association* **107**, 1201–1216.
- Cardot, H., Ferraty, F. and Sarda, P. (1999). Functional linear model. *Statistics & Probability Letters* **45**, 11–22.
- Cardot, H., Ferraty, F. and Sarda, P. (2003). Spline estimators for the functional linear model. *Statistica Sinica* **13**, 571–591.
- Cheng, X. and Hansen, B. E. (2015). Forecasting with factor-augmented regression: A frequentist model averaging approach. *Journal of Econometrics* **186**, 280–293.
- Claeskens, G. and Hjort, N. L. (2008). *Model Selection and Model Averaging*. Cambridge University Press.
- Clyde, M. and George, E. I. (2004). Model uncertainty. *Statistical Science* **19**, 81–94.
- Diebold, F. X. and Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & Economic Statistics* **20**, 134–144.
- Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 45–70.
- Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **70**, 849–911.
- Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *The Annals of Statistics* **38**, 3567–3604.
- Fan, Y., James, G. M. and Radchenko, P. (2015). Functional additive regression. *The Annals of Statistics* **43**, 2296–2325.
- Fang, F., Li, J. and Xia, X. (2022). Semiparametric model averaging prediction for dichotomous response. *Journal of Econometrics* **229**, 219–245.
- Gao, Y., Zhang, X., Wang, S. and Zou, G. (2016). Model averaging based on leave-subject-out cross-validation. *Journal of Econometrics* **192**, 139–151.
- Hall, P., Müller, H.-G. and Wang, J.-L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *The Annals of Statistics* **34**, 1493–1517.
- Hansen, B. E. (2007). Least squares model averaging. *Econometrica* **75**, 1175–1189.
- Hansen, B. E. and Racine, J. S. (2012). Jackknife model averaging. *Journal of Econometrics* **167**, 38–46.
- Hjort, N. L. and Claeskens, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association* **98**, 879–899.

- Kong, D., Xue, K., Yao, F. and Zhang, H. H. (2016). Partially functional linear regression in high dimensions. *Biometrika* **103**, 147–159.
- Li, K.-C. (1987). Asymptotic optimality for CP, CL, cross-validation and generalized cross-validation: Discrete index set. *The Annals of Statistics* **15**, 958–975.
- Liang, H., Zou, G., Wan, A. T. and Zhang, X. (2011). Optimal weight choice for frequentist model average estimators. *Journal of the American Statistical Association* **106**, 1053–1066.
- Liu, Q. and Okui, R. (2013). Heteroskedasticity-robust CP model averaging. *Econometrics Journal* **16**, 463–472.
- Müller, H.-G. and Yao, F. (2008). Functional additive models. *Journal of the American Statistical Association* **103**, 1534–1544.
- Racine, J. S. and Parmeter, C. F. (2014). Data-driven model evaluation: A test for revealed performance. In *The Oxford Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics*, 308–345. Oxford University Press.
- Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. 2nd Edition. Springer, New York.
- Rice, J. A. and Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society. Series B (Methodological)* **53**, 233–243.
- Sila, A., Pokhariyal, G. and Shepherd, K. (2017). Evaluating regression-kriging for mid-infrared spectroscopy prediction of soil properties in western Kenya-East Africa. *Geoderma Regional* **10**, 39–47.
- Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society. Series B (Methodological)* **50**, 413–436.
- Tang, Q., Tu, W. and Kong, L. (2023). Estimation for partial functional partially linear additive model. *Computational Statistics & Data Analysis* **177**, 107584.
- Wan, A. T., Zhang, X. and Zou, G. (2010). Least squares model averaging by mallows criterion. *Journal of Econometrics* **156**, 277–283.
- Wang, G., Feng, X.-N. and Chen, M. (2016). Functional partial linear single-index model. *Scandinavian Journal of Statistics* **43**, 261–274.
- Wong, R. K., Li, Y. and Zhu, Z. (2019). Partially linear functional additive models for multivariate functional data. *Journal of the American Statistical Association* **114**, 406–418.
- Yao, F. and Müller, H.-G. (2010). Functional quadratic regression. *Biometrika* **97**, 49–64.
- Yao, F., Müller, H.-G. and Wang, J.-L. (2005). Functional linear regression analysis for longitudinal data. *The Annals of Statistics* **33**, 2873–2903.
- Yu, D., Kong, L. and Mizera, I. (2016). Partial functional linear quantile regression for neuroimaging data analysis. *Neurocomputing* **195**, 74–87.
- Zhang, H. and Zou, G. (2020). Cross-validation model averaging for generalized functional linear model. *Econometrics* **8**, 7.
- Zhang, X., Chiou, J.-M. and Ma, Y. (2018). Functional prediction through averaging estimated functional linear regression models. *Biometrika* **105**, 945–962.
- Zhang, X. and Liu, C.-A. (2023). Model averaging prediction by K-fold cross-validation. *Journal of Econometrics* **235**, 280–301.
- Zhang, X., Wan, A. T. and Zhou, S. Z. (2012). Focused information criteria, model selection, and model averaging in a tobit model with a nonzero threshold. *Journal of Business & Economic Statistics* **30**, 132–142.
- Zhang, X., Wan, A. T. and Zou, G. (2013). Model averaging by jackknife criterion in models with dependent data. *Journal of Econometrics* **174**, 82–94.

- Zhang, X. and Wang, W. (2019). Optimal model averaging estimation for partially linear models. *Statistica Sinica* **29**, 693–718.
- Zhang, X., Yu, D., Zou, G. and Liang, H. (2016). Optimal model averaging estimation for generalized linear models and generalized linear mixed-effects models. *Journal of the American Statistical Association* **111**, 1775–1790.
- Zhang, X. and Yu, J. (2018). Spatial weights matrix selection and model averaging for spatial autoregressive models. *Journal of Econometrics* **203**, 1–18.
- Zhang, X., Zou, G. and Carroll, R. J. (2015). Model averaging based on Kullback–Leibler distance. *Statistica Sinica* **25**, 1583–1598.
- Zhang, X., Zou, G. and Liang, H. (2014). Model averaging and weight choice in linear mixed-effects models. *Biometrika* **101**, 205–218.
- Zhang, X., Zou, G., Liang, H. and Carroll, R. J. (2020). Parsimonious model averaging with a diverging number of parameters. *Journal of the American Statistical Association* **115**, 972–984.
- Zhu, H., Yao, F. and Zhang, H. H. (2014). Structured functional additive regression in reproducing kernel Hilbert spaces. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **76**, 581–603.
- Zhu, R., Wan, A. T., Zhang, X. and Zou, G. (2019). A mallows-type model averaging estimator for the varying-coefficient partially linear model. *Journal of the American Statistical Association* **114**, 882–892.
- Zhu, R., Zou, G. and Zhang, X. (2018). Optimal model averaging estimation for partial functional linear models. *Journal of Systems Science and Mathematical Sciences* **38**, 777–800.
- Zou, J., Wang, W., Zhang, X. and Zou, G. (2022). Optimal model averaging for divergent-dimensional poisson regressions. *Econometric Reviews* **41**, 775–805.

(Received February 2023; accepted May 2024)

Supplementary Material: Model Averaging for Partially Linear Functional Score Models

Shishi Liu¹, Chunming Zhang², Hao Zhang⁴, Rou Zhong^{3,4} and Jingxiao Zhang^{3,4,*}

¹*School of Economics, Hangzhou Dianzi University*

²*Department of Statistics, University of Wisconsin*

³*Center for Applied Statistics, Renmin University of China*

⁴*School of Statistics, Renmin University of China*

Supplementary Material

This supplementary file contains additional simulations, additional details in real application, and all technical proofs. Here we omit the notation *a.s.* (almost surely) for convenience.

S1 Additional Simulations

Example 3. $M_0 = 4$ and $\boldsymbol{\theta} = (1.5, 0.7, 0.2, -0.4)^T$. $\mathbf{Z}_i$ is a 4×1 vector that follows a multivariate normal distribution with zero means and a variance-covariance matrix $\Sigma = (0.5^{|a-b|})_{4 \times 4}$. The functional predictor $X_i(t)$ is obtained by

$$X_i(t) = \sum_{l=1}^4 \zeta_{il} \psi_l(t), \quad t \in [0, 1],$$

where $\psi_l(t) = \sqrt{2} \sin(\pi lt)$, $l = 1, \dots, 4$, and ζ_{il} is i.i.d and simulated from $N(0, l^{-3/2})$, $i = 1, \dots, n$. The random error term ε_i is i.i.d. and follows $N(0, \eta^2)$.

η controls the signal-to-noise ratio and we vary it such that $R^2 = \text{var}(\mu_i)/\text{var}(Y_i)$ ranges from 0.1 to 0.9, where $\text{var}(\mu_i)$ and $\text{var}(Y_i)$ denote the variances of μ_i and Y_i , respectively. And the non-linear effect of $X_i(t)$ is introduced by

$$\mathbf{f}(\boldsymbol{\xi}_i) = \exp \left\{ \sum_{l=1}^4 \xi_{il}/l \right\},$$

where $\xi_{il} = \Phi(\lambda_l^{-1/2}\zeta_{il})$. All settings are identical to that of Example 1 except for the non-linear effect of $X(t)$.

Example 4. $M_0 = 50$ and $\theta_j = j^{-1/2}$. The case where $\mathbf{z}$ and $X(t)$ being correlated is considered. Simulate $(\mathbf{Z}_i, \zeta_{i1}) \sim MN(0, \Sigma)$, where $\Sigma = (0.5^{|a-b|})_{51 \times 51}$. The functional predictor $X_i(t)$ is simulated by

$$X_i(t) = \sum_{l=1}^5 \zeta_{il} \psi_l(t), \quad t \in [0, 10],$$

where $\psi_l(t) = \cos(\pi lt/5)/\sqrt{5}$, $l = 1, \dots, 5$, and ζ_{il} is i.i.d. and follows $N(0, l^{-2})$, $i = 1, \dots, n$, $l = 2, \dots, 5$. The independent ε_i 's are heteroscedastic as $\varepsilon_i \sim N(0, \eta^2(u_i^2 + 0.01))$, where u_i follows $U[-1, 1]$. Still varying η such that R^2 varies between 0.1 and 0.9. And the non-linear effect of $X_i(t)$ is generated from

$$\mathbf{f}(\boldsymbol{\xi}_i) = 2\left(\xi_{i1}\xi_{i2} - \frac{1}{2}\right) + \left(\xi_{i3} - \frac{1}{2}\right)^2 - \frac{1}{12} + \frac{1}{4}\left(\xi_{i4} - \frac{1}{2}\right) + \frac{1}{5}\left(\xi_{i5} - \frac{1}{2}\right),$$

where $\xi_{il} = \Phi(\lambda_l^{-1/2}\zeta_{il})$.

Identical to Example 1, we still omit z_4 and ξ_4 in preparing candidate models in Example 3, so all candidate models are misspecified. With different specifications of which elements in $\{z_1, z_2, z_3\}$ and $\{\xi_1, \xi_2, \xi_3\}$ are included in the model, we have a total number of $M = (2^3 - 1)(2^3 - 1) = 49$ candidate models for Ex-

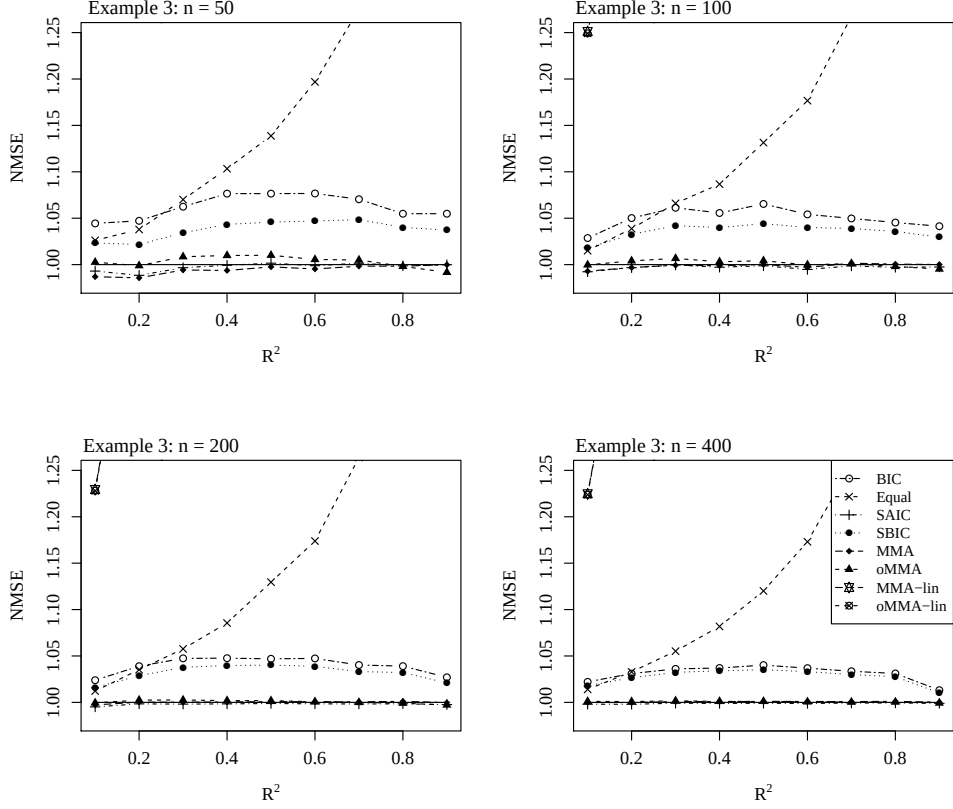


Figure S1: Normalized mean squared error (NMSE) comparisons for Example 3.

amples 1 and 2. As for Example 4, similar to that of Example 2, a pre-screening is conducted first, and candidate models are then constructed in a nested way.

In Example 3, both MMA-type and AIC-type estimators in Figure S1 demonstrate superiority over other competing estimators. Particularly, when sample size n is limited, MMA exhibits an advantage over AIC and SAIC methods. With increasing n , these estimators tend to behave similarly in most cases for R^2 . This underscores the effectiveness of MMA approach. Additionally, it is observed that the oracle MMA (oMMA) is less effective for small values of R^2 and n , which

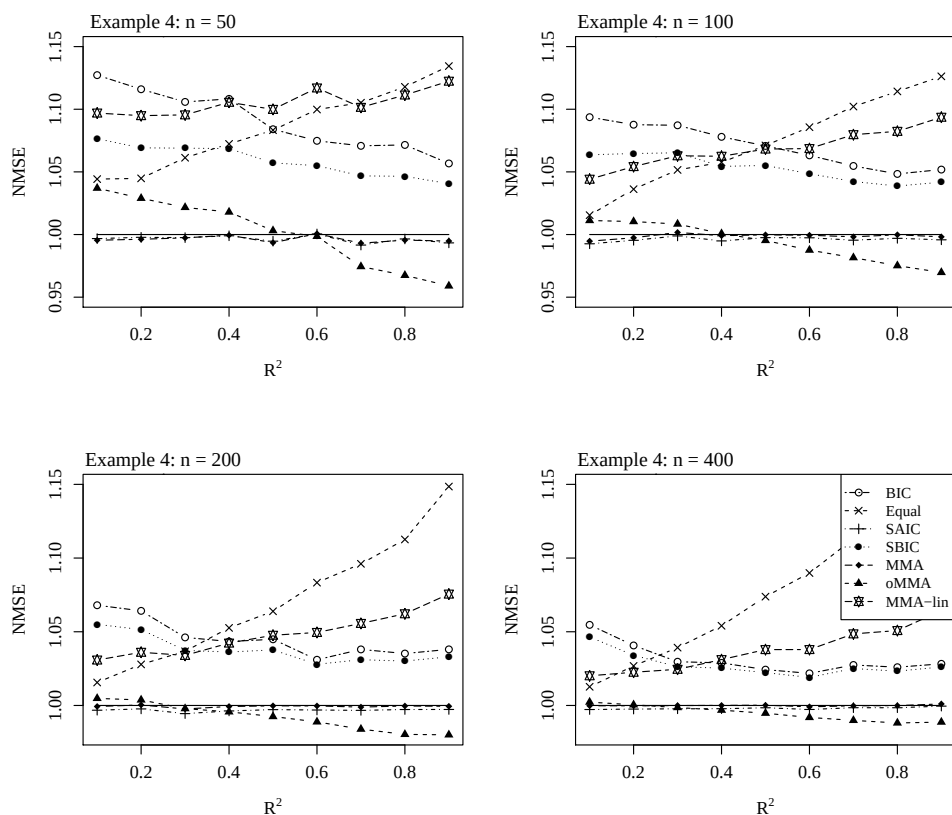


Figure S2: Normalized mean squared error (NMSE) comparisons for Example 4.

can be attributed to the signal-to-noise ratio in data, as discussed earlier in the manuscript. Also, Figure S1 illustrates that SAIC and SBIC outperform their model selection counterparts, AIC and BIC. The differences tend to decrease as R^2 or n grows. Finally, it is worth noting that the lines for MMA-lin and oracle MMA-lin (oMMA-lin) mostly fall outside the subfigures; only two points of each can be seen at the top left of the subfigures. This implies that partially linear functional linear models are inefficient in detecting non-linear effects, highlighting the necessity of exploring nonparametric modeling. Similar trends to those

in Figure 1 can also be observed in Figure S1.

In Example 4, as depicted in Figure S2, the oracle MMA (o-MMA) demonstrates a clear edge over other competing methods for medium and large R^2 values. Additionally, MMA outperforms the others for small R^2 and n values. As n increases, the differences between MMA and AIC-type methods diminish. On the other hand, MMA-lin deteriorates as R^2 increases, coinciding with the strengthening signal of the non-linear component. Notably, equally weighting still performs the poorest for most large R^2 values. Similar trends to those in Figure 2 can also be observed in Figure S2.

S2 Details in Real Application

We excluded one covariate, phosphorous ($mg \cdot kg^{-1}$), from the original set of 19 covariates due to its incompleteness. The sample marginal correlations of remaining covariates with our response variable, the total carbon percentage, are presented in Table 1, ranging from -0.3508 to 0.6523. To prepare candidate model, we firstly screened out scalar variables whose absolute marginal correlations exceed 0.1, and then adopted the nested fashion described in the manuscript.

Through $D = 1000$ repetitions, we finally obtained D mean squared prediction error (MSPE) values. Figures S3 and S4 present boxplots and empirical cumulative distribution functions of MSPEs for each approach. It is evident from Figure S4 that a visual stochastic dominance relationship exists between

Table 1: Sample marginal Correlations to the total carbon percentage (Y).

	marginal correlation
Exchangeable Magnesium ($cmol_c \cdot kg^{-1}$)	0.6523
Exchangeable Bases ($cmol_c \cdot kg^{-1}$)	0.6124
Exchangeable Potassium ($cmol_c \cdot kg^{-1}$)	0.6093
Exchangeable Calcium ($cmol_c \cdot kg^{-1}$)	0.5698
Boron Concentration ($mg \cdot kg^{-1}$)	0.5022
Soil Electrical Conductivity ($dS \cdot m^{-1}$)	0.4172
Sulphur ($mg \cdot kg^{-1}$)	0.3890
Soil pH in Water	0.2202
Iron Concentration ($mg \cdot kg^{-1}$)	0.2143
Copper Concentration ($mg \cdot kg^{-1}$)	0.1963
Zinc ($mg \cdot kg^{-1}$)	0.1733
Exchangeable Calcium-to-Magnesium Ratio	0.0826
Exchangeable Manganese ($mg \cdot kg^{-1}$)	0.0749
Exchangeable Acidity ($cmol_c \cdot kg^{-1}$)	-0.0147
Exchangeable Aluminium ($mg \cdot kg^{-1}$)	-0.0533
Exchangeable Sodium ($cmol_c \cdot kg^{-1}$)	-0.0991
Exchangeable Sodium Ratio (%)	-0.3088
Exchangeable Sodium Percentage (%)	-0.3508

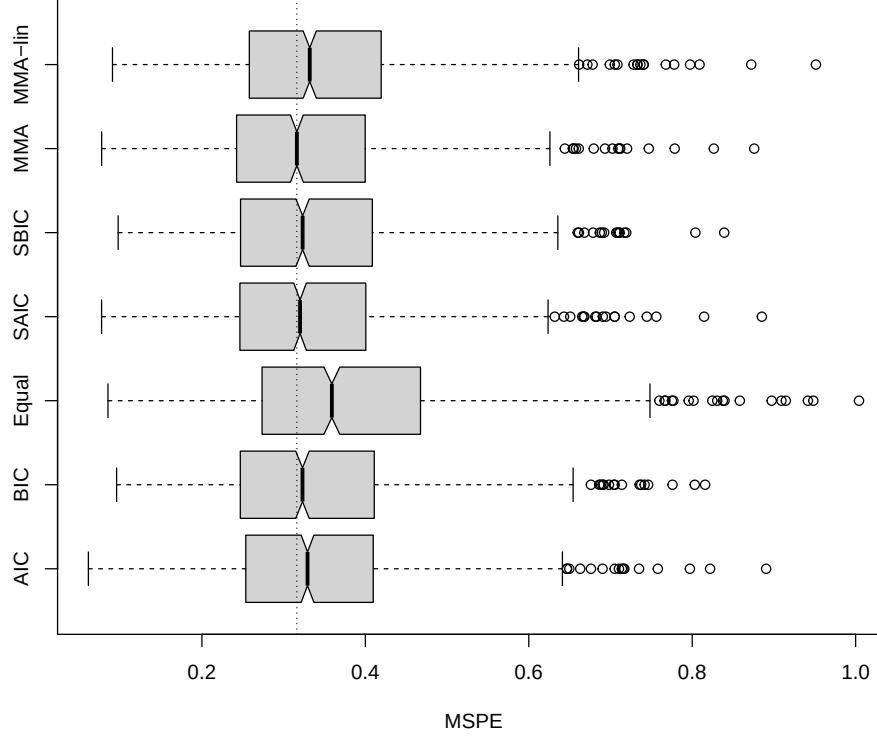


Figure S3: Boxplots of MSPE across $D = 1000$ repetitions for the soil dataset. The vertical dashed line represents the median MSPE value of MMA.

MMA and other competitors, indicating the preference for MMA on this dataset. Moreover, Figure S4 reveals that linear modeling may be underspecified, and the equally weighting scheme may be inefficient in practice.

Additionally, the remaining results for the data-driven method using paired Mann-Whitney-Wilcoxon (MWW) test are listed in Table 2. The alternative hypothesis indicates that the other method is less accurate than MMA. A small statistic value (relative to $D(D + 1)/2 \approx 250, 250$) of the MWW test suggests

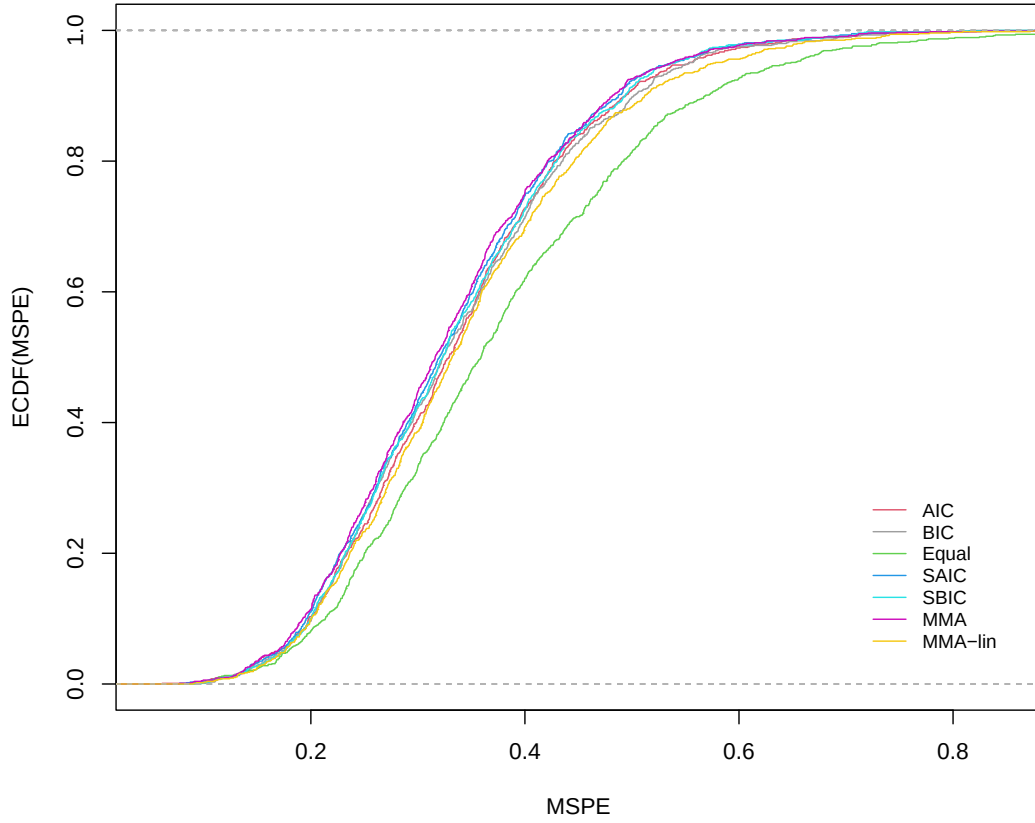


Figure S4: ECDFs of MSPE based on $D = 1000$ values for the soil dataset.

that the respective method is less accurate than MMA. The p-values have also been adjusted using the BH method, as employed in the manuscript. Again, both the MWW statistics and corresponding adjusted p-values in Table 2 demonstrate that MMA outperforms other competitors in terms of prediction accuracy for this dataset.

Table 2: Results of test statistics and adjusted p-values for the data-driven approach using paired MWW test.

	MMA/AIC	MMA/BIC	MMA/Equal	MMA/SAIC	MMA/SBIC	MMA/MMA-lin
mww stat.	1.31E+05	1.53E+05	1.16E+05	2.03E+05	1.79E+05	1.89E+05
mww p.val	4.63E-39	1.09E-26	1.12E-48	9.17E-08	4.54E-15	9.61E-12

S3 Some Lemmas

The estimation error of the transformed FPC score is of order $O_p(n^{-1/2}l)$ as shown in [4], and we list the result here while omitting the detailed proof.

Lemma S3.1. *Suppose the transformation function $\Phi(\cdot)$ has bounded derivative.*

Under Assumption 1, there exists a constant $C > 0$ such that $\mathbb{E}(\widehat{\xi}_{il} - \xi_{il})^2 \leq Cl^2/n$ uniformly for $l \leq J_n$, where $J_n = \lfloor (2C_\lambda O_p(1))^{-1/(1+\alpha)} n^{1/(2+2\alpha)} \rfloor$.

Lemma S3.2. *Under Assumptions 1, 2 and 3(c), we have $\lambda_{\max}(\mathbf{K}_{(m)}) = O(1)$, $\lambda_{\max}(\widehat{\mathbf{K}}_{(m)}) = O_p(1)$, and $\lambda_{\max}(\mathbf{P}_{(m)}) = O(1)$, $\lambda_{\max}(\widehat{\mathbf{P}}_{(m)}) = O_p(1)$, for all $m = 1, \dots, M$.*

Proof. For any square matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ (see [1]), we have

$$\lambda_{\max}(\mathbf{M}_1 \mathbf{M}_2) \leq \lambda_{\max}(\mathbf{M}_1) \lambda_{\max}(\mathbf{M}_2), \quad (\text{S3.1})$$

$$\text{and } \lambda_{\max}(\mathbf{M}_1 + \mathbf{M}_2) \leq \lambda_{\max}(\mathbf{M}_1) + \lambda_{\max}(\mathbf{M}_2).$$

These two inequalities will be frequently used in the following proofs.

By an inequality of Reisz (see [2]), we obtain that

$$\lambda_{\max}^2(\mathbf{K}_{(m)}) \leq \max_i \sum_{j=1}^n |\mathbf{K}_{(m),ij}| \cdot \max_j \sum_{i=1}^n |\mathbf{K}_{(m),ij}|,$$

which implies that $\lambda_{\max}(\mathbf{K}_{(m)}) = O(1)$. Hence,

$$\begin{aligned}\lambda_{\max}(\mathbf{P}_{(m)}) &= \lambda_{\max}(\tilde{\mathbf{P}}_{(m)})(1 + \lambda_{\max}(\mathbf{K}_{(m)})) + \lambda_{\max}(\mathbf{K}_{(m)}) \\ &= (1 + \lambda_{\max}(\mathbf{K}_{(m)})) + \lambda_{\max}(\mathbf{K}_{(m)}) = O(1).\end{aligned}$$

From Lemma S3.1, we obtain that

$$\begin{aligned}\hat{\xi}_{il} - \xi_{il} &= O_p(n^{-1/2}l), \\ \hat{\xi}_{il} - \hat{\xi}_{jl} &= \xi_{il} - \xi_{jl} + O_p(n^{-1/2}l), \quad \text{for } k \leq J_n.\end{aligned}$$

Applying Taylor series expansion and Assumption 2,

$$\begin{aligned}\hat{\mathbf{K}}_{(m),ij} &= \frac{\mathcal{K}(\hat{\boldsymbol{\xi}}_{(m),i} - \hat{\boldsymbol{\xi}}_{(m),j})}{\sum_{j'=1}^n \mathcal{K}(\hat{\boldsymbol{\xi}}_{(m),i} - \hat{\boldsymbol{\xi}}_{(m),j'})} \\ &= \left\{ \mathcal{K}(\boldsymbol{\xi}_{(m),i} - \boldsymbol{\xi}_{(m),j}) + \sum_{l=1}^{q_m} k'(\xi_{il} - \xi_{jl}) \prod_{u \neq l} k(\xi_{iu} - \xi_{ju})(\hat{\xi}_{il} - \xi_{il} + \xi_{jl} - \hat{\xi}_{jl}) \right. \\ &\quad \left. + o_p(n^{-\frac{1}{2}}q_m) \right\} / \left\{ \sum_{j'=1}^n [\mathcal{K}(\boldsymbol{\xi}_{(m),i} - \boldsymbol{\xi}_{(m),j'}) + O_p(n^{-\frac{3}{2}}q_m)] \right\} \\ &= \frac{\mathcal{K}(\boldsymbol{\xi}_{(m),i} - \boldsymbol{\xi}_{(m),j})}{\sum_{j'=1}^n \mathcal{K}(\boldsymbol{\xi}_{(m),i} - \boldsymbol{\xi}_{(m),j'})} + \left\{ \sum_{l=1}^{q_m} k'(\xi_{il} - \xi_{jl}) \prod_{u \neq l} k(\xi_{iu} - \xi_{ju})(\hat{\xi}_{il} - \xi_{il} + \xi_{jl} - \hat{\xi}_{jl}) \right\} \\ &\quad / \left\{ \sum_{j'=1}^n \mathcal{K}(\boldsymbol{\xi}_{(m),i} - \boldsymbol{\xi}_{(m),j'}) \right\} + o_p(n^{-\frac{1}{2}}q_m) \\ &= \mathbf{K}_{(m),ij} + O_p(n^{-\frac{3}{2}}q_m), \quad q_m \leq J_n,\end{aligned}$$

i.e., $\hat{\mathbf{K}}_{(m),ij} = \mathbf{K}_{(m),ij} + O_p(n^{-\frac{3}{2}}q_m)$. Note that q_m is no larger than J_n and it is common for kernel smoothing to restrict the dimension of $\boldsymbol{\xi}$ to handle the curse of dimensionality. By Assumptions 2 and 3, we can show that

$$\begin{aligned}\max_i \sum_{j=1}^n |\hat{\mathbf{K}}_{(m),ij}| &= \max_i \sum_{j=1}^n |\mathbf{K}_{(m),ij}| + O_p(n^{-\frac{1}{2}}q_m) = O_p(1), \\ \max_j \sum_{i=1}^n |\hat{\mathbf{K}}_{(m),ij}| &= \max_j \sum_{i=1}^n |\mathbf{K}_{(m),ij}| + O_p(n^{-\frac{1}{2}}q_m) = O_p(1),\end{aligned}$$

uniformly for $m = 1, \dots, M$. Similarly,

$$\lambda_{\max}^2(\widehat{\mathbf{K}}_{(m)}) \leq \max_i \sum_{j=1}^n |\widehat{\mathbf{K}}_{(m),ij}| \max_j \sum_{i=1}^n |\widehat{\mathbf{K}}_{(m),ij}| = O_p(1),$$

$$\begin{aligned} \lambda_{\max}(\widehat{\mathbf{P}}_{(m)}) &= \lambda_{\max}(\overline{\mathbf{P}}_{(m)})(1 + \lambda_{\max}(\widehat{\mathbf{K}}_{(m)})) + \lambda_{\max}(\widehat{\mathbf{K}}_{(m)}) \\ &= (1 + \lambda_{\max}(\widehat{\mathbf{K}}_{(m)})) + \lambda_{\max}(\widehat{\mathbf{K}}_{(m)}) = O_p(1). \end{aligned}$$

In addition,

$$\max_i \sum_{j=1}^n |\widehat{\mathbf{K}}_{(m),ij} - \mathbf{K}_{(m),ij}| = O_p(n^{-\frac{1}{2}}q_m),$$

$$\max_j \sum_{i=1}^n |\widehat{\mathbf{K}}_{(m),ij} - \mathbf{K}_{(m),ij}| = O_p(n^{-\frac{1}{2}}q_m),$$

$$\lambda_{\max}^2(\widehat{\mathbf{K}}_{(m)} - \mathbf{K}_{(m)}) \leq \max_i \sum_{j=1}^n |\widehat{\mathbf{K}}_{(m),ij} - \mathbf{K}_{(m),ij}| \max_j \sum_{i=1}^n |\widehat{\mathbf{K}}_{(m),ij} - \mathbf{K}_{(m),ij}|,$$

which leads to

$$\lambda_{\max}(\widehat{\mathbf{K}}_{(m)} - \mathbf{K}_{(m)}) = O_p(n^{-\frac{1}{2}}q_m). \tag{S3.2}$$

□

Lemma S3.3 corresponds to Lemma 1 in the paper.

Lemma S3.3. *Under Assumptions 1, 2 and 3(c)(e), we have*

$$\lambda_{\max}(\mathbf{P}_{(m)} - \widehat{\mathbf{P}}_{(m)}) = O_p(n^{-\frac{1}{2}}q_m),$$

for all $m = 1, \dots, M$.

Proof. Take a decomposition as

$$\begin{aligned} \mathbf{P}_{(m)} - \widehat{\mathbf{P}}_{(m)} &= (\widetilde{\mathbf{P}}_{(m)} - \overline{\mathbf{P}}_{(m)}) + (\mathbf{K}_{(m)} - \widehat{\mathbf{K}}_{(m)}) + \widetilde{\mathbf{P}}_{(m)}(\widehat{\mathbf{K}}_{(m)} - \mathbf{K}_{(m)}) \\ &\quad + (\overline{\mathbf{P}}_{(m)} - \widetilde{\mathbf{P}}_{(m)})\mathbf{K}_{(m)} + (\widetilde{\mathbf{P}}_{(m)} - \overline{\mathbf{P}}_{(m)})(\mathbf{K}_{(m)} - \widehat{\mathbf{K}}_{(m)}). \end{aligned} \quad (\text{S3.3})$$

Recalling Eq.(S3.1), it suffices to determine the order of $\lambda_{\max}(\widetilde{\mathbf{P}}_{(m)} - \overline{\mathbf{P}}_{(m)})$ and $\lambda_{\max}(\mathbf{K}_{(m)} - \widehat{\mathbf{K}}_{(m)})$. We have already quantified $\lambda_{\max}(\mathbf{K}_{(m)} - \widehat{\mathbf{K}}_{(m)})$ in Eq.(S3.2). Remind that $\widetilde{\mathbf{P}}_{(m)}$ and $\overline{\mathbf{P}}_{(m)}$ are projection matrices related to $\widetilde{\mathbf{Z}}_{(m)} = (\mathbf{I} - \mathbf{K}_{(m)})\mathbf{Z}_{(m)}$ and $\widehat{\mathbf{Z}}_{(m)} = (\mathbf{I} - \widehat{\mathbf{K}}_{(m)})\mathbf{Z}_{(m)}$, i.e. $\widetilde{\mathbf{P}}_{(m)} = \widetilde{\mathbf{Z}}_{(m)}(\widetilde{\mathbf{Z}}_{(m)}^T \widetilde{\mathbf{Z}}_{(m)})^{-1} \widetilde{\mathbf{Z}}_{(m)}^T$ and $\overline{\mathbf{P}}_{(m)} = \widehat{\mathbf{Z}}_{(m)}(\widehat{\mathbf{Z}}_{(m)}^T \widehat{\mathbf{Z}}_{(m)})^{-1} \widehat{\mathbf{Z}}_{(m)}^T$. We simplify the notations as $\widetilde{\mathbf{P}}_{(m)} = \mathbf{H}_{(\mathbf{I} - \mathbf{K}_{(m)})\mathbf{Z}_{(m)}}$ and $\overline{\mathbf{P}}_{(m)} = \mathbf{H}_{(\mathbf{I} - \widehat{\mathbf{K}}_{(m)})\mathbf{Z}_{(m)}}$, where $\mathbf{H}_A$ represents the (normalised) projection operator generated from A . It is obviously observed that the deviation between two projection operators $\widetilde{\mathbf{P}}_{(m)}$ and $\overline{\mathbf{P}}_{(m)}$ derives from the difference between $\mathbf{K}_{(m)}$ and $\widehat{\mathbf{K}}_{(m)}$, so using Assumption 3(e) it holds that

$$\begin{aligned} \|\widetilde{\mathbf{P}}_{(m)} - \overline{\mathbf{P}}_{(m)}\| &\leq \|\mathbf{H}_{\mathbf{I} - \mathbf{K}_{(m)}} - \mathbf{H}_{\mathbf{I} - \widehat{\mathbf{K}}_{(m)}}\| = \|\mathbf{H}_{\mathbf{K}_{(m)}} - \mathbf{H}_{\widehat{\mathbf{K}}_{(m)}}\| \\ &\leq C(\|\mathbf{K}_{(m)} - \widehat{\mathbf{K}}_{(m)}\|)(\lambda_{\max}(\mathbf{K}_{(m)}) + \lambda_{\max}(\widehat{\mathbf{K}}_{(m)})) + 2\lambda_{\max}(\mathbf{K}_{(m)})\lambda_{\max}(\widehat{\mathbf{K}}_{(m)}) = O_p(n^{-\frac{1}{2}}q_m), \end{aligned} \quad (\text{S3.4})$$

where C is related to the smallest nonzero singular value of $\mathbf{K}_{(m)}$.

Finally according to Eq.(S3.3), combining Eqs.(S3.1), (S3.2) and (S3.4), we have

$$\lambda_{\max}(\mathbf{P}_{(m)} - \widehat{\mathbf{P}}_{(m)}) = O_p(n^{-\frac{1}{2}}q_m),$$

which completes the proof. $\square$

S4 Proof of Theorem 1

Proof. Firstly, it follows from Lemma S3.2 that

$$\begin{aligned} \sup_{\boldsymbol{\omega}} \lambda_{\max}(\widehat{\mathbf{P}}(\boldsymbol{\omega})) &= \sup_{\boldsymbol{\omega}} \lambda_{\max}\left(\sum_{m=1}^M \omega_m \widehat{\mathbf{P}}_{(m)}\right) \leq \sup_{\boldsymbol{\omega}} \sum_{m=1}^M \omega_m \lambda_{\max}(\widehat{\mathbf{P}}_{(m)}) \\ &\leq \max_{1 \leq m \leq M} \lambda_{\max}(\widehat{\mathbf{P}}_{(m)}) = O_p(1), \end{aligned} \quad (\text{S4.5})$$

and similarly,

$$\sup_{\boldsymbol{\omega}} \lambda_{\max}(\mathbf{P}(\boldsymbol{\omega})) = O_p(1). \quad (\text{S4.6})$$

Let $\widehat{\mathbf{A}}(\boldsymbol{\omega}) = \mathbf{I} - \widehat{\mathbf{P}}(\boldsymbol{\omega})$ and $\mathbf{A}(\boldsymbol{\omega}) = \mathbf{I} - \mathbf{P}(\boldsymbol{\omega})$. From the definition of $L_n(\boldsymbol{\omega})$,

$\widehat{C}_n(\boldsymbol{\omega})$ and $R_n(\boldsymbol{\omega})$, we have

$$\begin{aligned} \widehat{C}_n(\boldsymbol{\omega}) &= L_n(\boldsymbol{\omega}) + \|\varepsilon\|^2 - 2\boldsymbol{\mu}^T(\mathbf{P}(\boldsymbol{\omega}) - \widehat{\mathbf{P}}(\boldsymbol{\omega}))\boldsymbol{\mu} - 2\varepsilon^T(\mathbf{P}(\boldsymbol{\omega}) - \widehat{\mathbf{P}}(\boldsymbol{\omega}))\varepsilon - 2\varepsilon^T(\mathbf{P}(\boldsymbol{\omega}) - \widehat{\mathbf{P}}(\boldsymbol{\omega}))\boldsymbol{\mu} \\ &\quad - 2\varepsilon^T(\mathbf{P}(\boldsymbol{\omega}) - \widehat{\mathbf{P}}(\boldsymbol{\omega}))^T\boldsymbol{\mu} - 2\varepsilon^T(\widehat{\mathbf{P}}(\boldsymbol{\omega}) - \mathbf{P}(\boldsymbol{\omega}))^T\mathbf{P}(\boldsymbol{\omega})\boldsymbol{\mu} + 2\varepsilon^T\mathbf{P}^T(\boldsymbol{\omega})(\widehat{\mathbf{P}}(\boldsymbol{\omega}) - \mathbf{P}(\boldsymbol{\omega}))\boldsymbol{\mu} \\ &\quad + \boldsymbol{\mu}^T(\mathbf{P}(\boldsymbol{\omega}) + \widehat{\mathbf{P}}(\boldsymbol{\omega}))^T(\mathbf{P}(\boldsymbol{\omega}) - \widehat{\mathbf{P}}(\boldsymbol{\omega}))\boldsymbol{\mu} + \varepsilon^T(\mathbf{P}(\boldsymbol{\omega}) + \widehat{\mathbf{P}}(\boldsymbol{\omega}))^T(\mathbf{P}(\boldsymbol{\omega}) - \widehat{\mathbf{P}}(\boldsymbol{\omega}))\varepsilon \\ &\quad + \varepsilon^T(\mathbf{P}(\boldsymbol{\omega}) + \widehat{\mathbf{P}}(\boldsymbol{\omega}))^T(\mathbf{P}(\boldsymbol{\omega}) - \widehat{\mathbf{P}}(\boldsymbol{\omega}))\boldsymbol{\mu} + 2\varepsilon^T\mathbf{A}(\boldsymbol{\omega})\boldsymbol{\mu} \\ &\quad - 2[\varepsilon^T\mathbf{P}(\boldsymbol{\omega})\varepsilon - \text{tr}(\mathbf{P}(\boldsymbol{\omega})\boldsymbol{\Omega})] - 2[\text{tr}(\mathbf{P}(\boldsymbol{\omega})\boldsymbol{\Omega}) - \text{tr}(\widehat{\mathbf{P}}(\boldsymbol{\omega})\boldsymbol{\Omega})], \end{aligned}$$

and

$$L_n(\boldsymbol{\omega}) - R_n(\boldsymbol{\omega}) = \varepsilon^T\mathbf{P}^T(\boldsymbol{\omega})\mathbf{P}(\boldsymbol{\omega})\varepsilon - \text{tr}(\mathbf{P}^T(\boldsymbol{\omega})\mathbf{P}(\boldsymbol{\omega})\boldsymbol{\Omega}) - 2\varepsilon^T\mathbf{P}^T(\boldsymbol{\omega})\mathbf{A}(\boldsymbol{\omega})\boldsymbol{\mu}.$$

So similar to the proof of Theorem 2.1 of [1], in order to prove Eq.(4) in Theorem

1, we need only to verify that

$$\sup_{\boldsymbol{\omega}} \frac{|\varepsilon^T\mathbf{A}(\boldsymbol{\omega})\boldsymbol{\mu}|}{R_n(\boldsymbol{\omega})} = o_p(1), \quad (\text{S4.7})$$

$$\sup_{\boldsymbol{\omega}} \frac{|\varepsilon^T\mathbf{P}(\boldsymbol{\omega})\varepsilon - \text{tr}(\mathbf{P}(\boldsymbol{\omega})\boldsymbol{\Omega})|}{R_n(\boldsymbol{\omega})} = o_p(1), \quad (\text{S4.8})$$

$$\sup_{\omega} \frac{|\varepsilon^T \mathbf{P}^T(\omega) \mathbf{P}(\omega) \varepsilon - \text{tr}(\mathbf{P}^T(\omega) \mathbf{P}(\omega) \mathbf{\Omega})|}{R_n(\omega)} = o_p(1), \quad (\text{S4.9})$$

$$\sup_{\omega} \frac{|\boldsymbol{\mu}^T \mathbf{A}^T(\omega) \mathbf{P}(\omega) \varepsilon|}{R_n(\omega)} = o_p(1), \quad (\text{S4.10})$$

$$\sup_{\omega} \frac{|\varepsilon^T (\mathbf{P}(\omega) - \hat{\mathbf{P}}(\omega))^T \boldsymbol{\mu}|}{R_n(\omega)} = o_p(1), \quad (\text{S4.11})$$

$$\sup_{\omega} \frac{|\varepsilon^T (\mathbf{P}(\omega) - \hat{\mathbf{P}}(\omega)) \boldsymbol{\mu}|}{R_n(\omega)} = o_p(1), \quad (\text{S4.12})$$

$$\sup_{\omega} \frac{|\varepsilon^T (\hat{\mathbf{P}}(\omega) - \mathbf{P}(\omega))^T \mathbf{P}(\omega) \boldsymbol{\mu}|}{R_n(\omega)} = o_p(1), \quad (\text{S4.13})$$

$$\sup_{\omega} \frac{|\varepsilon^T \mathbf{P}^T(\omega) (\hat{\mathbf{P}}(\omega) - \mathbf{P}(\omega)) \boldsymbol{\mu}|}{R_n(\omega)} = o_p(1), \quad (\text{S4.14})$$

$$\sup_{\omega} \frac{|\varepsilon^T (\mathbf{P}(\omega) + \hat{\mathbf{P}}(\omega))^T (\mathbf{P}(\omega) - \hat{\mathbf{P}}(\omega)) \boldsymbol{\mu}|}{R_n(\omega)} = o_p(1), \quad (\text{S4.15})$$

$$\sup_{\omega} \frac{|\boldsymbol{\mu}^T (\mathbf{P}(\omega) - \hat{\mathbf{P}}(\omega)) \boldsymbol{\mu}|}{R_n(\omega)} = o_p(1), \quad (\text{S4.16})$$

$$\sup_{\omega} \frac{|\boldsymbol{\mu}^T (\mathbf{P}(\omega) + \hat{\mathbf{P}}(\omega))^T (\mathbf{P}(\omega) - \hat{\mathbf{P}}(\omega)) \boldsymbol{\mu}|}{R_n(\omega)} = o_p(1), \quad (\text{S4.17})$$

$$\sup_{\omega} \frac{|\varepsilon^T (\mathbf{P}(\omega) - \hat{\mathbf{P}}(\omega)) \varepsilon|}{R_n(\omega)} = o_p(1), \quad (\text{S4.18})$$

$$\sup_{\omega} \frac{|\varepsilon^T (\mathbf{P}(\omega) + \hat{\mathbf{P}}(\omega))^T (\mathbf{P}(\omega) - \hat{\mathbf{P}}(\omega)) \varepsilon|}{R_n(\omega)} = o_p(1), \quad (\text{S4.19})$$

and

$$\sup_{\omega} \frac{|\text{tr}(\mathbf{P}(\omega) \mathbf{\Omega}) - \text{tr}(\hat{\mathbf{P}}(\omega) \mathbf{\Omega})|}{R_n(\omega)} = o_p(1). \quad (\text{S4.20})$$

Note that Eqs.(S4.7)–(S4.10) do not include any $\hat{\cdot}$ terms. From Eq.(S4.6) and Assumption 3, Eqs.(S4.7)–(S4.10) can be shown by using the same steps as in the proof of Theorem 1 of [3].

For proving Eq.(S4.19), by (S3.1), it is seen that

$$\begin{aligned}
& \sup_{\boldsymbol{\omega}} \frac{|\varepsilon^T (\mathbf{P}(\boldsymbol{\omega}) + \hat{\mathbf{P}}(\boldsymbol{\omega}))^T (\mathbf{P}(\boldsymbol{\omega}) - \hat{\mathbf{P}}(\boldsymbol{\omega})) \varepsilon|}{R_n(\boldsymbol{\omega})} \\
& \leq \eta_n^{-1} \frac{1}{2} \sup_{\boldsymbol{\omega}} \left| \varepsilon^T \left[(\mathbf{P}(\boldsymbol{\omega}) + \hat{\mathbf{P}}(\boldsymbol{\omega}))^T (\mathbf{P}(\boldsymbol{\omega}) - \hat{\mathbf{P}}(\boldsymbol{\omega})) + (\mathbf{P}(\boldsymbol{\omega}) - \hat{\mathbf{P}}(\boldsymbol{\omega}))^T (\mathbf{P}(\boldsymbol{\omega}) + \hat{\mathbf{P}}(\boldsymbol{\omega})) \right] \varepsilon \right| \\
& \leq \eta_n^{-1} \frac{1}{2} \|\varepsilon\|^2 \cdot \sup_{\boldsymbol{\omega}} \lambda_{\max} \left[(\mathbf{P}(\boldsymbol{\omega}) + \hat{\mathbf{P}}(\boldsymbol{\omega}))^T (\mathbf{P}(\boldsymbol{\omega}) - \hat{\mathbf{P}}(\boldsymbol{\omega})) + (\mathbf{P}(\boldsymbol{\omega}) - \hat{\mathbf{P}}(\boldsymbol{\omega}))^T (\mathbf{P}(\boldsymbol{\omega}) + \hat{\mathbf{P}}(\boldsymbol{\omega})) \right] \\
& \leq \eta_n^{-1} \|\varepsilon\|^2 \cdot \sup_{\boldsymbol{\omega}} \lambda_{\max}(\mathbf{P}(\boldsymbol{\omega}) + \hat{\mathbf{P}}(\boldsymbol{\omega})) \lambda_{\max}(\mathbf{P}(\boldsymbol{\omega}) - \hat{\mathbf{P}}(\boldsymbol{\omega})) \\
& \leq \eta_n^{-1} \|\varepsilon\|^2 \cdot \sup_{\boldsymbol{\omega}} [\lambda_{\max}(\mathbf{P}(\boldsymbol{\omega})) + \lambda_{\max}(\hat{\mathbf{P}}(\boldsymbol{\omega}))] \cdot \sum_{m=1}^M \omega_m \lambda_{\max}(\mathbf{P}_{(m)} - \hat{\mathbf{P}}_{(m)}) \\
& \leq n \eta_n^{-1} \cdot \frac{\|\varepsilon\|^2}{n} \cdot \sup_{\boldsymbol{\omega}} [\lambda_{\max}(\mathbf{P}(\boldsymbol{\omega})) + \lambda_{\max}(\hat{\mathbf{P}}(\boldsymbol{\omega}))] \cdot \max_{1 \leq m \leq M} \lambda_{\max}(\mathbf{P}_{(m)} - \hat{\mathbf{P}}_{(m)}) \\
& = o_p(1),
\end{aligned}$$

where the last step is from Eqs.(S4.5)–(S4.6), Assumption 3 and Lemma S3.3.

By Lemma S3.3 and Assumption 3, we can prove Eqs.(S4.15)–(S4.18) in a similar way.

For Eq.(S4.12),

$$\begin{aligned}
 & \sup_{\omega} \frac{|\varepsilon^T (\mathbf{P}(\omega) - \widehat{\mathbf{P}}(\omega)) \boldsymbol{\mu}|}{R_n(\omega)} \\
 & \leq \eta_n^{-1} \|\boldsymbol{\mu}\| \cdot \sup_{\omega} \|(\mathbf{P}(\omega) - \widehat{\mathbf{P}}(\omega))^T \varepsilon\| \\
 & \leq \eta_n^{-1} \cdot \|\boldsymbol{\mu}\| \cdot \sup_{\omega} \lambda_{\max}(\mathbf{P}(\omega) - \widehat{\mathbf{P}}(\omega)) \cdot \|\varepsilon\| \\
 & \leq n\eta_n^{-1} \cdot \frac{\|\boldsymbol{\mu}\|}{\sqrt{n}} \frac{\|\varepsilon\|}{\sqrt{n}} \cdot \max_{1 \leq m \leq M} \lambda_{\max}(\mathbf{P}_{(m)} - \widehat{\mathbf{P}}_{(m)}) \\
 & = o_p(1),
 \end{aligned}$$

where the last step is from Lemma S3.3 and Assumption 3. Similarly, we can verify Eqs.(S4.11), (S4.13)–(S4.14) by Lemma S3.3, Assumption 3 and Eqs.(S4.5)–(S4.6).

Now we consider the last Eq.(S4.20). Note that $\boldsymbol{\Omega}$ is a diagonal matrix,

$$\begin{aligned}
 & \sup_{\omega} \frac{|tr(\mathbf{P}(\omega)\boldsymbol{\Omega}) - tr(\widehat{\mathbf{P}}(\omega)\boldsymbol{\Omega})|}{R_n(\omega)} \\
 & = \sup_{\omega} \frac{|tr[(\mathbf{P}(\omega) - \widehat{\mathbf{P}}(\omega))\boldsymbol{\Omega}]|}{R_n(\omega)} \\
 & \leq \eta_n^{-1} \sup_{\omega} |tr(\mathbf{P}(\omega) - \widehat{\mathbf{P}}(\omega))| \lambda_{\max}(\boldsymbol{\Omega}) \\
 & \leq n\eta_n^{-1} \max_{1 \leq m \leq M} \lambda_{\max}(\mathbf{P}_{(m)} - \widehat{\mathbf{P}}_{(m)}) \lambda_{\max}(\boldsymbol{\Omega}) \\
 & = o_p(1),
 \end{aligned}$$

where the last step is from Lemma S3.3 and Assumption 3. This completes the proof of Theorem 1. $\square$

S5 Proof of Theorem 2

Proof. Note that

$$\widehat{C}_n(\boldsymbol{\omega})|_{\boldsymbol{\Omega}=\widehat{\boldsymbol{\Omega}}} = \widehat{C}_n(\boldsymbol{\omega}) + 2tr(\widehat{\mathbf{P}}(\boldsymbol{\omega})\widehat{\boldsymbol{\Omega}}) - 2tr(\widehat{\mathbf{P}}(\boldsymbol{\omega})\boldsymbol{\Omega}).$$

From the result of Theorem 1, to prove Eq.(5) in Theorem 2, it suffices to prove that

$$\sup_{\boldsymbol{\omega}} \frac{|tr(\widehat{\mathbf{P}}(\boldsymbol{\omega})\widehat{\boldsymbol{\Omega}}) - tr(\widehat{\mathbf{P}}(\boldsymbol{\omega})\boldsymbol{\Omega})|}{R_n(\boldsymbol{\omega})} = o_p(1). \quad (\text{S5.21})$$

Let $\mathbf{Q}_{(m)} = \text{diag}(\rho_{11}^{(m)}, \dots, \rho_{nn}^{(m)})$ and $\mathbf{Q}(\boldsymbol{\omega}) = \sum_{m=1}^M \omega_m \mathbf{Q}_{(m)}$. To prove Eq.(S5.21), we decompose the left-hand side of Eq.(S5.21) into four parts as follows.

$$\begin{aligned} & \sup_{\boldsymbol{\omega}} \frac{|tr(\widehat{\mathbf{P}}(\boldsymbol{\omega})\widehat{\boldsymbol{\Omega}}) - tr(\widehat{\mathbf{P}}(\boldsymbol{\omega})\boldsymbol{\Omega})|}{R_n(\boldsymbol{\omega})} \\ &= \sup_{\boldsymbol{\omega}} \frac{|(\mathbf{Y} - \widehat{\mathbf{P}}_{(M^*)}\mathbf{Y})^T \widehat{\mathbf{Q}}(\boldsymbol{\omega})(\mathbf{Y} - \widehat{\mathbf{P}}_{(M^*)}\mathbf{Y}) - tr(\widehat{\mathbf{Q}}(\boldsymbol{\omega})\boldsymbol{\Omega})|}{R_n(\boldsymbol{\omega})} \\ &= \sup_{\boldsymbol{\omega}} \frac{|(\boldsymbol{\mu} + \varepsilon)^T (\mathbf{I} - \widehat{\mathbf{P}}_{(M^*)})^T \widehat{\mathbf{Q}}(\boldsymbol{\omega})(\mathbf{I} - \widehat{\mathbf{P}}_{(M^*)})(\boldsymbol{\mu} + \varepsilon) - tr(\widehat{\mathbf{Q}}(\boldsymbol{\omega})\boldsymbol{\Omega})|}{R_n(\boldsymbol{\omega})} \\ &\leq \sup_{\boldsymbol{\omega}} \frac{|\boldsymbol{\mu}^T (\mathbf{I} - \widehat{\mathbf{P}}_{(M^*)})^T \widehat{\mathbf{Q}}(\boldsymbol{\omega})(\mathbf{I} - \widehat{\mathbf{P}}_{(M^*)})\boldsymbol{\mu}|}{R_n(\boldsymbol{\omega})} + \sup_{\boldsymbol{\omega}} \frac{2|\varepsilon^T (\mathbf{I} - \widehat{\mathbf{P}}_{(M^*)})^T \widehat{\mathbf{Q}}(\boldsymbol{\omega})(\mathbf{I} - \widehat{\mathbf{P}}_{(M^*)})\boldsymbol{\mu}|}{R_n(\boldsymbol{\omega})} \\ &\quad + \sup_{\boldsymbol{\omega}} \frac{|\varepsilon^T (\mathbf{I} - \widehat{\mathbf{P}}_{(M^*)})^T \widehat{\mathbf{Q}}(\boldsymbol{\omega})(\mathbf{I} - \widehat{\mathbf{P}}_{(M^*)})\varepsilon|}{R_n(\boldsymbol{\omega})} + \sup_{\boldsymbol{\omega}} \frac{|tr(\widehat{\mathbf{Q}}(\boldsymbol{\omega})\boldsymbol{\Omega})|}{R_n(\boldsymbol{\omega})} \\ &\equiv \Xi_1 + \Xi_2 + \Xi_3 + \Xi_4. \end{aligned}$$

Now define $\rho = \max_{1 \leq m \leq M} \max_{1 \leq i \leq n} |\rho_{ii}^{(m)}|$. From Assumption 4 and Lemma

S3.2, we have $\max_{1 \leq m \leq M} |tr(\hat{\mathbf{K}}_{(m)})| = \max_{1 \leq m \leq M} |tr(\mathbf{K}_{(m)})| + O_p(n^{-\frac{1}{2}}\tilde{q})$ and

$$\begin{aligned}
 \rho &\leq cn^{-1} \max_{1 \leq m \leq M} |tr(\hat{\mathbf{P}}_{(m)})| \\
 &\leq cn^{-1} \max_{1 \leq m \leq M} |tr(\bar{\mathbf{P}}_{(m)})| + cn^{-1} \max_{1 \leq m \leq M} |tr(\bar{\mathbf{P}}_{(m)}\hat{\mathbf{K}}_{(m)})| + cn^{-1} \max_{1 \leq m \leq M} |tr(\hat{\mathbf{K}}_{(m)})| \\
 &\leq cn^{-1} \max_{1 \leq m \leq M} rank(\bar{\mathbf{P}}_{(m)}) + cn^{-1} \frac{1}{2} \max_{1 \leq m \leq M} [\lambda_{\max}(\bar{\mathbf{P}}_{(m)}\hat{\mathbf{K}}_{(m)} + \hat{\mathbf{K}}_{(m)}^T \bar{\mathbf{P}}_{(m)}) \\
 &\quad \cdot rank(\bar{\mathbf{P}}_{(m)}\hat{\mathbf{K}}_{(m)} + \hat{\mathbf{K}}_{(m)}^T \bar{\mathbf{P}}_{(m)})] + cn^{-1} \max_{1 \leq m \leq M} |tr(\hat{\mathbf{K}}_{(m)})| \\
 &\leq cn^{-1}\tilde{p} + cn^{-1} \cdot 2\tilde{p} \cdot \lambda_{\max}(\bar{\mathbf{P}}_{(m)})\lambda_{\max}(\hat{\mathbf{K}}_{(m)}) + cn^{-1} \max_{1 \leq m \leq M} |tr(\hat{\mathbf{K}}_{(m)})| \\
 &= cn^{-1}\tilde{p} + cn^{-1}\tilde{p} \cdot O_p(1) + cn^{-1} \cdot O_p(h^{-\tilde{q}} + n^{-\frac{1}{2}}\tilde{q}) \\
 &= O_p(n^{-1}\tilde{p} + n^{-1}h^{-\tilde{q}} + n^{-\frac{3}{2}}\tilde{q}).
 \end{aligned} \tag{S5.22}$$

It follows from Lemma S3.2, Assumptions 3–4, and Eqs.(S4.5) and (S5.22) that

$$\begin{aligned}
 \Xi_1 &\leq \eta_n^{-1} \sup_{\omega} \lambda_{\max}(\hat{\mathbf{Q}}(\omega)) \cdot \|(\mathbf{I} - \hat{\mathbf{P}}_{(M^*)})\boldsymbol{\mu}\|^2 \\
 &\leq \eta_n^{-1} \rho \cdot \|(\mathbf{I} - \hat{\mathbf{P}}_{(M^*)})\boldsymbol{\mu}\|^2 \\
 &\leq \eta_n^{-1} \rho \cdot [1 + \lambda_{\max}(\hat{\mathbf{P}}_{(M^*)})]^2 \cdot \|\boldsymbol{\mu}\|^2 \\
 &= \eta_n^{-1} \cdot O_p(n^{-1}\tilde{p} + n^{-1}h^{-\tilde{q}} + n^{-\frac{3}{2}}\tilde{q}) \cdot O_p(1) \cdot O_p(n) \\
 &= O_p(\eta_n^{-1}\tilde{p} + \eta_n^{-1}h^{-\tilde{q}} + n^{-\frac{1}{2}}\eta_n^{-1}\tilde{q}).
 \end{aligned}$$

Using Lemma S3.2, Assumptions 3–4, and Eqs.(S4.5) and (S5.22), we obtain

that

$$\begin{aligned}
\Xi_2 &\leq 2\eta_n^{-1} \cdot \|(\mathbf{I} - \widehat{\mathbf{P}}_{(M^*)})\boldsymbol{\mu}\| \cdot \sup_{\boldsymbol{\omega}} \|\widehat{\mathbf{Q}}(\boldsymbol{\omega})(\mathbf{I} - \widehat{\mathbf{P}}_{(M^*)})\varepsilon\| \\
&\leq 2\eta_n^{-1} \cdot \|(\mathbf{I} - \widehat{\mathbf{P}}_{(M^*)})\boldsymbol{\mu}\| \cdot \sup_{\boldsymbol{\omega}} \lambda_{\max}(\widehat{\mathbf{Q}}(\boldsymbol{\omega})) \cdot \|(\mathbf{I} - \widehat{\mathbf{P}}_{(M^*)})\varepsilon\| \\
&\leq 2\eta_n^{-1} \cdot \|(\mathbf{I} - \widehat{\mathbf{P}}_{(M^*)})\boldsymbol{\mu}\| \cdot \rho \cdot \|(\mathbf{I} - \widehat{\mathbf{P}}_{(M^*)})\varepsilon\| \\
&\leq 2\eta_n^{-1} \cdot (1 + \lambda_{\max}(\widehat{\mathbf{P}}_{(M^*)})) \cdot \|\boldsymbol{\mu}\| \cdot \rho \cdot (1 + \lambda_{\max}(\widehat{\mathbf{P}}_{(M^*)})) \cdot \|\varepsilon\| \\
&= 2\eta_n^{-1} \cdot O_p(1) \cdot O_p(n^{\frac{1}{2}}) \cdot O_p(n^{-1}\tilde{p} + n^{-1}h^{-\tilde{q}} + n^{-\frac{3}{2}}\tilde{q}) \cdot O_p(1) \cdot O_p(n^{\frac{1}{2}}) \\
&= O_p(\eta_n^{-1}\tilde{p} + \eta_n^{-1}h^{-\tilde{q}} + n^{-\frac{1}{2}}\eta_n^{-1}\tilde{q}).
\end{aligned}$$

Using Lemma S3.2, Assumptions 3–4, and Eqs.(S4.5) and (S5.22), we have

$$\begin{aligned}
\Xi_3 &\leq \eta_n^{-1} \cdot \sup_{\boldsymbol{\omega}} \lambda_{\max}(\widehat{\mathbf{Q}}(\boldsymbol{\omega})) \cdot \|(\mathbf{I} - \widehat{\mathbf{P}}_{(M^*)})\varepsilon\|^2 \\
&\leq \eta_n^{-1} \cdot \rho [1 + \lambda_{\max}(\widehat{\mathbf{P}}_{(M^*)})]^2 \cdot \|\varepsilon\|^2 \\
&= \eta_n^{-1} \cdot O_p(n^{-1}\tilde{p} + n^{-1}h^{-\tilde{q}} + n^{-\frac{3}{2}}\tilde{q}) \cdot O_p(1) \cdot O_p(n) \\
&= O_p(\eta_n^{-1}\tilde{p} + \eta_n^{-1}h^{-\tilde{q}}).
\end{aligned}$$

Using Assumptions 3–4, and Eqs.(S4.5) and (S5.22), we have

$$\begin{aligned}
\Xi_4 &\leq \eta_n^{-1} \cdot n \sup_{\boldsymbol{\omega}} \lambda_{\max}(\widehat{\mathbf{Q}}(\boldsymbol{\omega})) \cdot \lambda_{\max}(\boldsymbol{\Omega}) \\
&\leq \eta_n^{-1} \cdot n \cdot \rho \cdot \lambda_{\max}(\boldsymbol{\Omega}) \\
&= \eta_n^{-1} \cdot n \cdot O_p(n^{-1}\tilde{p} + n^{-1}h^{-\tilde{q}} + n^{-\frac{3}{2}}\tilde{q}) \cdot O_p(1) \\
&= O_p(\eta_n^{-1}\tilde{p} + \eta_n^{-1}h^{-\tilde{q}} + n^{-\frac{1}{2}}\eta_n^{-1}\tilde{q}).
\end{aligned}$$

Finally, it follows from Assumptions 3–4 that $\Xi_1 = o_p(1)$, $\Xi_2 = o_p(1)$, $\Xi_3 = o_p(1)$ and $\Xi_4 = o_p(1)$. Therefore, we have verified Eq.(S5.21) and this completes the proof.

□

Remark 1. In our current development, we focus on the setting of densely observed functional data with noise. Through our theoretical derivation, we find that the estimation error from FPCA mainly influences asymptotic optimality via the convergence rate of $\lambda_{\max}(\mathbf{P}_{(m)} - \hat{\mathbf{P}}_{(m)})$. When $\{X_i(t)\}$ are sparsely or irregularly observed, similar to cases encountered in longitudinal studies, we can similarly justify the optimality for this sparse and irregular setting. This is contingent on FPCA for sparsely or irregularly observed data yielding appropriate convergence rates for the estimator of transformed FPC scores $\{\hat{\xi}_{il} - \xi_{il}\}$.

Bibliography

- [1] Li, K.-C. (1987). Asymptotic optimality for cp, cl, cross-validation and generalized cross-validation: discrete index set. *The Annals of Statistics* 15(3), 958–975.
- [2] Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society: Series B (Methodological)* 50(3), 413–436.
- [3] Wan, A. T., X. Zhang, and G. Zou (2010). Least squares model averaging by mallows criterion. *Journal of Econometrics* 156(2), 277–283.
- [4] Wong, R. K., Y. Li, and Z. Zhu (2019). Partially linear functional additive

models for multivariate functional data. *Journal of the American Statistical Association* 114(525), 406–418.