Statistica Sinica Preprint No: SS-2023-0067	
Title	Model Averaging Estimation for Partially Linear Functional Score Models
Manuscript ID	SS-2023-0067
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202023.0067

Complete List of Authors	Shishi Liu, Chunming Zhang, Hao Zhang, Rou Zhong and Jingxiao Zhang
Corresponding Authors	Jingxiao Zhang
E-mails	zhjxiaoruc@163.com

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0067

Model Averaging Estimation for Partially Linear Functional Score Models

Shishi Liu[1], Chunming Zhang[2], Hao Zhang[4], 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*

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: Model average, Mallows-type criterion, Functional data.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0067

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 et al. (1999, 2003); Yao et al. (2005); 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

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

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

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0067

regressions. For example, Müller and Yao (2008) and Fan et al. (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 et al. (2016) and Kong et al. (2016) developed partial functional linear regressions as

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

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 et al. (2019) and Tang et al. (2023) further augmented the modeling flexibility by introducing nonparametric

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0067

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 et al. (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

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

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 to process $X(t)$. 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 et al. (2014); Kong et al. (2016). Thereby, model uncertainty arises from the choice of functional space truncation. Besides, varying decisions on which scalar components to retain in the 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 et al. (1997); Clyde

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0067

and George (2004); 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 et al. (1997) and Hjort and Claeskens (2003) advocated information criterion-based weighting scheme, which assigns weights calculated from information criterion scores (such as AIC, BIC, etc.) for each candidate model. Hansen (2007) introduced a Mallow'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 et al. (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 et al., 2013; Cheng and Hansen, 2015; Gao et al., 2016; Zhang and Liu, 2023), which are particularly useful in

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0067

cases where deriving an unbiased estimator of squared risk is challenging. Moreover, Kullback-Leibler loss-based model averaging (Zhang et al., 2015; Fang et al., 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 et al. (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 et al. (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

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0067

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 et al. (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. Section 4 and Section 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 supplementary material.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0067

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 (iid) 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 et al. (2016); Zhu et al. (2018); Zhang et al. (2018); Wong et al. (2019) for examples. Hence, we follow this routine and establish our model as

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

where, albeit with a slight abuse of notation, $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$ now denotes the random error vector with conditional mean 0 and conditional variance matrix $\boldsymbol{\Omega} = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2)$ 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 $\mathcal{C}(s,t) = cov\{X(s), X(t)\}$. $\mathcal{T}$ is typically assumed to be a compact interval.

The classical FPCA takes eigen decomposition of the corresponding covariance operator as $(\mathcal{C}\psi_l)(t) = \lambda_l \psi_l(t)$, $l = 1, 2, \ldots$, where $\{\psi_1(t), \psi_2(t), \ldots\}$ is a set of (or-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0067

thonormal) eigenfunctions associated with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots > 0$. Then, $X(t)$ is projected onto the 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 k-th eigenfunction, and $var(\zeta_l) = \lambda_l$, $l = 1, \ldots, q$; refer to Rice and Silverman (1991); Hall et al. (2006).

Denote by $\boldsymbol{\zeta}_i = (\zeta_{i1}, \ldots, \zeta_{iq})$ 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 $\xi_{il} = \Phi\big(\zeta_{il}; \lambda_l^{-1/2}\big)$, where $\Phi(\cdot)$ is a continuously differentiable map from $\mathbb{R}$ to $[0, 1]$. And the resulting vector $\boldsymbol{\xi}_i = (\xi_{i1}, \ldots, \xi_{iq})^T$ represents the transformed FPC scores for $X_i(t)$. Similar to the process in Wong et al. (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 $\xi_{il} = \Phi\big(\zeta_{il}; \lambda_l^{-1/2}\big) = \Phi\big(\lambda_l^{-1/2}\zeta_{il}\big)$ 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

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0067

PLFS model comprises p_m regressors in $\mathbf{Z}_i$ and q_m regressors in $\boldsymbol{\xi}_i$,

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

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, $l = 1, \ldots, q_m$. Furthermore, let $\mathbf{K}_{(m)} = \big(K_{(m),ij}\big)$ be the $n \times n$ smoother matrix with elements

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

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

$$\widetilde{\boldsymbol{\theta}}_{(m)} = \big(\widetilde{\mathbf{Z}}_{(m)}^T \widetilde{\mathbf{Z}}_{(m)}\big)^{-1} \widetilde{\mathbf{Z}}_{(m)}^T (\mathbf{I} - \mathbf{K}_{(m)})\mathbf{Y},$$

$$\widetilde{\mathbf{f}}_{(m)}(\boldsymbol{\xi}_{(m)}) = \mathbf{K}_{(m)}(\mathbf{Y} - \mathbf{Z}_{(m)}\widetilde{\boldsymbol{\theta}}_{(m)}),$$

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

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0067

m-th candidate model is given by

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

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

Remark 1. In contrast to the exploration of model (1.2) with multiple functional predictors in Zhu et al. (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 $\boldsymbol{\zeta}_j$ and $\boldsymbol{\beta}_j$, to characterize the "features" in $X(t)$ and $\beta(t)$. This results in a reduced linear model as $Y = \mathbf{z}^T\boldsymbol{\theta} + \sum_{j=1}^{v}\boldsymbol{\zeta}_j^T\boldsymbol{\beta}_j + \varepsilon$. Then, they utilized the OLS estimator for each candidate model, whereas our work employs a kernel smoothing approach.

2.2 Weight choice criterion

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

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

Then the model averaging estimator of μ follows as

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

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0067

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

$$L_n(\boldsymbol{\omega}) = \|\widetilde{\boldsymbol{\mu}}(\boldsymbol{\omega}) - \boldsymbol{\mu}\|^2 = \|\mathbf{P}(\boldsymbol{\omega})\mathbf{Y} - \boldsymbol{\mu}\|^2,$$

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

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

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

It is evident that $\mathbb{E}\{C_n(\boldsymbol{\omega})|\mathbf{z}, X\} = R_n(\boldsymbol{\omega}) + tr(\boldsymbol{\Omega})$. Therefore, $C_n(\boldsymbol{\omega})$ serves as an unbiased estimator of the expected in-sample squared error loss plus a constant, similar to the Mallow's criterion proposed by Hansen (2007). As $tr(\boldsymbol{\Omega})$ is independent of $\boldsymbol{\omega}$, 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 $\widehat{\boldsymbol{\xi}}_{(m)}$. Specifically, suppose $X_i(t)$ is discretely measured with noise,

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

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 $\mathbf{Z}_i$. Now we focus on densely observed trajectories, allowing $X_i(t)$ to be effectively reconstructed from

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0067

$\{(t_{ij}, X_{ij}) : j = 1, \ldots, N_i\}$ using a smoother operator, see Kong et al. (2016); Wong et al. (2019). The reconstructed function is denoted by $\widetilde{X}_i(t)$. Then, the mean and covariance function of $X(t)$ are estimated by

$$\widehat{\nu}(t) = \frac{1}{n}\sum_{i=1}^{n} \widetilde{X}_i(t),$$
$$\widehat{\mathcal{C}}(s,t) = \frac{1}{n}\sum_{i=1}^{n} \big\{\widetilde{X}_i(s) - \widehat{\nu}(s)\big\}\big\{\widetilde{X}_i(t) - \widehat{\nu}(t)\big\}^T.$$

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

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

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

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

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

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

$$\widehat{\boldsymbol{\theta}}_{(m)} = \big(\widehat{\mathbf{Z}}_{(m)}^T\widehat{\mathbf{Z}}_{(m)}\big)^{-1}\widehat{\mathbf{Z}}_{(m)}^T(\mathbf{I} - \widehat{\mathbf{K}}_{(m)})\mathbf{Y},$$
$$\widehat{\mathbf{f}}_{(m)}(\widehat{\boldsymbol{\xi}}_{(m)}) = \widehat{\mathbf{K}}_{(m)}\big(\mathbf{Y} - \mathbf{Z}_{(m)}\widehat{\boldsymbol{\theta}}_{(m)}\big),$$

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

$$\widehat{\boldsymbol{\mu}}_{(m)} = \widehat{\mathbf{Z}}_{(m)}\big(\widehat{\mathbf{Z}}_{(m)}^T\widehat{\mathbf{Z}}_{(m)}\big)^{-1}\widehat{\mathbf{Z}}_{(m)}^T\big(\mathbf{I} - \widehat{\mathbf{K}}_{(m)}\big)\mathbf{Y} + \widehat{\mathbf{K}}_{(m)}\mathbf{Y} \equiv \widehat{\mathbf{P}}_{(m)}\mathbf{Y},$$
$$\widehat{\boldsymbol{\mu}}(\boldsymbol{\omega}) = \sum_{m=1}^{M} \omega_m\widehat{\boldsymbol{\mu}}_{(m)} = \sum_{m=1}^{M} \omega_m\widehat{\mathbf{P}}_{(m)}\mathbf{Y} = \widehat{\mathbf{P}}(\boldsymbol{\omega})\mathbf{Y},$$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0067

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

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

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

Let $\widetilde{\boldsymbol{\omega}} = \arg\min_{\boldsymbol{\omega}\in\mathcal{H}_n} \widehat{C}_n(\boldsymbol{\omega})$. 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 $M^* = \arg\max_{1\le m\le M}(p_m + q_m)$, leading to an estimator given by

$$\widehat{\boldsymbol{\Omega}} = \operatorname{diag}\big(\hat{\epsilon}^2_{(M^*),1}, \ldots, \hat{\epsilon}^2_{(M^*),n}\big), \tag{2.2}$$

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

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

$$\begin{aligned}
\widehat{\boldsymbol{\omega}} &= \arg\min_{\boldsymbol{\omega}} \widehat{C}_n(\boldsymbol{\omega})|_{\boldsymbol{\Omega}=\widehat{\boldsymbol{\Omega}}} \\
&= \arg\min_{\boldsymbol{\omega}} \|\mathbf{Y} - \widehat{\boldsymbol{\mu}}(\boldsymbol{\omega})\|^2 + 2tr\{\widehat{\mathbf{P}}(\boldsymbol{\omega})\widehat{\boldsymbol{\Omega}}\},
\end{aligned} \tag{2.3}$$

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

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0067

quadratic programming problem in the form of

$$\min_{\boldsymbol{\omega}} \widehat{C}_n(\boldsymbol{\omega})|_{\boldsymbol{\Omega}=\widehat{\boldsymbol{\Omega}}} = \min_{\boldsymbol{\omega}} \boldsymbol{\omega}^T \mathbf{H}^T \mathbf{H} \boldsymbol{\omega} + 2\boldsymbol{\omega}^T \mathbf{b}$$

$$\text{subject to} \quad \mathbf{1}^T \boldsymbol{\omega} = 1 \text{ and } \boldsymbol{\omega} \geq 0,$$

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

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0067

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 et al. (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 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

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0067

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

$$Y = \boldsymbol{\theta}^T \mathbf{z} + m_1(X_1) + \cdots + m_g(X_g) + \varepsilon,$$

$$Y = \boldsymbol{\theta}^T \mathbf{z} + \mathbf{f}_1(\boldsymbol{\xi}_1) + \cdots + \mathbf{f}_g(\boldsymbol{\xi}_g) + \varepsilon,$$

where $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_g$ are vectors of transformed FPC scores corresponding to $X_1(t)$, ..., $X_g(t)$, and $\mathbf{f}_1, \ldots, \mathbf{f}_g$ represent unknown functions for nonparametric modeling. The

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0067

approach of additive modeling for functional predictors, $m_1(X_1)+\cdots+m_g(X_g)$, rather than $m(X_1,\ldots,X_g)$, serves to alleviate the issue of the "curse of dimensionality" that often arises in nonparametric statistics. Regarding the estimation for $\mathbf{f}_1,\ldots,\mathbf{f}_g$, 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 $\widetilde{p} = \max_m p_m$, $\widetilde{q} = \max_m q_m$, and

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0067

$h = \min_m h_m$. 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

$$c_\lambda^{-1} l^{-\alpha} \leq \lambda_l \leq C_\lambda l^{-\alpha},$$
$$\lambda_l - \lambda_{l+1} \geq C_\lambda^{-1} l^{-1-\alpha}, \quad l = 1, 2, \ldots,$$

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 $\|X\| = \left\{ \int_{\mathcal{T}} X^2(t)dt \right\}^{1/2}$, 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); 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 et al. (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.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0067

(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 $m = 1, \ldots, M$.

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.), $i = 1, \ldots, n$.

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

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

(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 $m = 1, \ldots, M$.

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); Zhang et al. (2014) also. Assumption 3(b) is a common convergence condition in the literature, which requires

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0067

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 in Wan et al. (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. Remind that $\alpha > 1$ so that the order of $\widetilde{q}$ 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 $\widetilde{q}$ 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 $\eta_n^{-1}\widetilde{q}^2 = o_p(1)$ 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 $\widehat{\boldsymbol{\theta}}_{(m)}$ and $\widehat{\mathbf{f}}_{(m)}(\widehat{\xi}_{(m)})$. It further implies an intermediate result, as illustrated below in Lemma 1.

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

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

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

Lemma 1 demonstrates that the difference between $\mathbf{P}_{(m)}$ and its estimate $\widehat{\mathbf{P}}_{(m)}$ 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 et al. (2018), who did not account for the approximation effect of their estimated

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0067

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 $\lambda_{\max}\big(\mathbf{P}_{(m)} - \widehat{\mathbf{P}}_{(m)}\big) = o_p(1)$ as a high-level and technical condition, see Zhang et al. (2014); 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.

Now we provide the asymptotic optimality of the model averaging estimator when $\boldsymbol{\Omega}$ is known.

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

$$\frac{L_n(\widetilde{\boldsymbol{\omega}})}{\inf_{\boldsymbol{\omega}\in\mathcal{H}_n} L_n(\boldsymbol{\omega})} \to 1 \tag{3.1}$$

in probability as $n \to \infty$.

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

Following Liu and Okui (2013), we process $tr\{\widehat{\mathbf{P}}(\boldsymbol{\omega})\boldsymbol{\Omega}\}$ as one entity rather than considering $\boldsymbol{\Omega}$ in isolation, and estimate it by $\sum_{i=1}^{n} \hat{\epsilon}^2_{(M^*),i}\rho_{ii}(\boldsymbol{\omega})$ where $\rho_{ii}(\boldsymbol{\omega})$ is the i-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0067

th diagonal element of $\widehat{\mathbf{P}}(\boldsymbol{\omega})$. Denote $\rho_{ii}^{(m)}$ as the i-th diagonal element of $\widehat{\mathbf{P}}_{(m)}$. When $\boldsymbol{\Omega}$ is replaced by its estimate $\widehat{\boldsymbol{\Omega}}$ given in (2.2), provided that the following additional assumptions are satisfied, it can be shown that the model averaging estimator based on $\widehat{\boldsymbol{\omega}}$ shares the same asymptotic optimality as $\widetilde{\boldsymbol{\omega}}$ in Theorem 1.

Assumption 4. (a) There exists a constant c such that $\max_i \rho_{ii}^{(m)} \leq cn^{-1}|tr\big(\widehat{\mathbf{P}}_{(m)}\big)|$ (a.s.), uniformly for $m = 1, \ldots, M$.

(b) $tr\big(\mathbf{K}_{(m)}\big) = O(h^{-\widetilde{q}})$ (a.s.), uniformly for $m = 1, \ldots, M$.

(c) $\eta_n^{-1}\widetilde{p} = o_p(1)$ and $\eta_n^{-1}h^{-\widetilde{q}} = o_p(1)$.

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 et al. (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*

$$\frac{L_n(\widehat{\boldsymbol{\omega}})}{\inf_{\boldsymbol{\omega}\in\mathcal{H}_n} L_n(\boldsymbol{\omega})} \to 1 \tag{3.2}$$

in probability as $n \to \infty$.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0067

Theorem 2 shows that Theorem 1 remains valid when $\mathbf{\Omega}$ is replaced by $\widehat{\mathbf{\Omega}}$. Thus, the practically feasible $\widehat{\boldsymbol{\omega}}$ also enjoys the asymptotic optimality. The supplementary material provides the detailed proof for Theorems 1–2.

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 et al., 1997). For the m-th candidate model, AIC and BIC select the model with the smallest scores, defined as $AIC_m = \log(\widehat{\sigma}_m^2) + 2tr(\widehat{\mathbf{P}}_{(m)})/n$ and $BIC_m = \log(\widehat{\sigma}_m^2) + \log(n)tr(\widehat{\mathbf{P}}_{(m)})/n$, where $\widehat{\sigma}_m^2 = \|\mathbf{Y} - \widehat{\boldsymbol{\mu}}_{(m)}\|^2/n$. 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 et al. (2018).

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0067

The data is generated from the following PLFS model,

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

The vector $\boldsymbol{\xi}_i$ represents the transformed FPC scores derived from $\boldsymbol{\zeta}_i$, 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 $\psi_l(t) = \sqrt{2}\sin(\pi l t)$, $l = 1, \ldots, 4$, and ζ_{il} is i.i.d and simulated from $N(0, l^{-3/2})$, $i = 1, \ldots, n$. 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 $R^2 = var(\mu_i)/var(Y_i)$ ranges from 0.1 to 0.9, where $var(\mu_i)$ and $var(Y_i)$ 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

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0067

$X_i(t)$ is generated from

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

where $\psi_l(t) = \sqrt{2}\sin(\pi l t)$, $l = 1, \ldots, 10$, and ζ_{il} is i.i.d. and follows $N(0, l^{-3/2})$, $i = 1, \ldots, n$, $l = 2, \ldots, 10$. Consider another type of heteroskedasticity for independent random errors as $\varepsilon_i \sim N\big(0, \eta^2(Z_{i1}^2 + 0.01)\big)$. 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

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

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

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0067

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

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

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 1–2 depict the corresponding results for Examples 1–2. 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.

In Example 2, we consider a more intricate data generation process. The true models incorporate heteroscedastic random errors, a departure from Zhu et al. (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

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0067

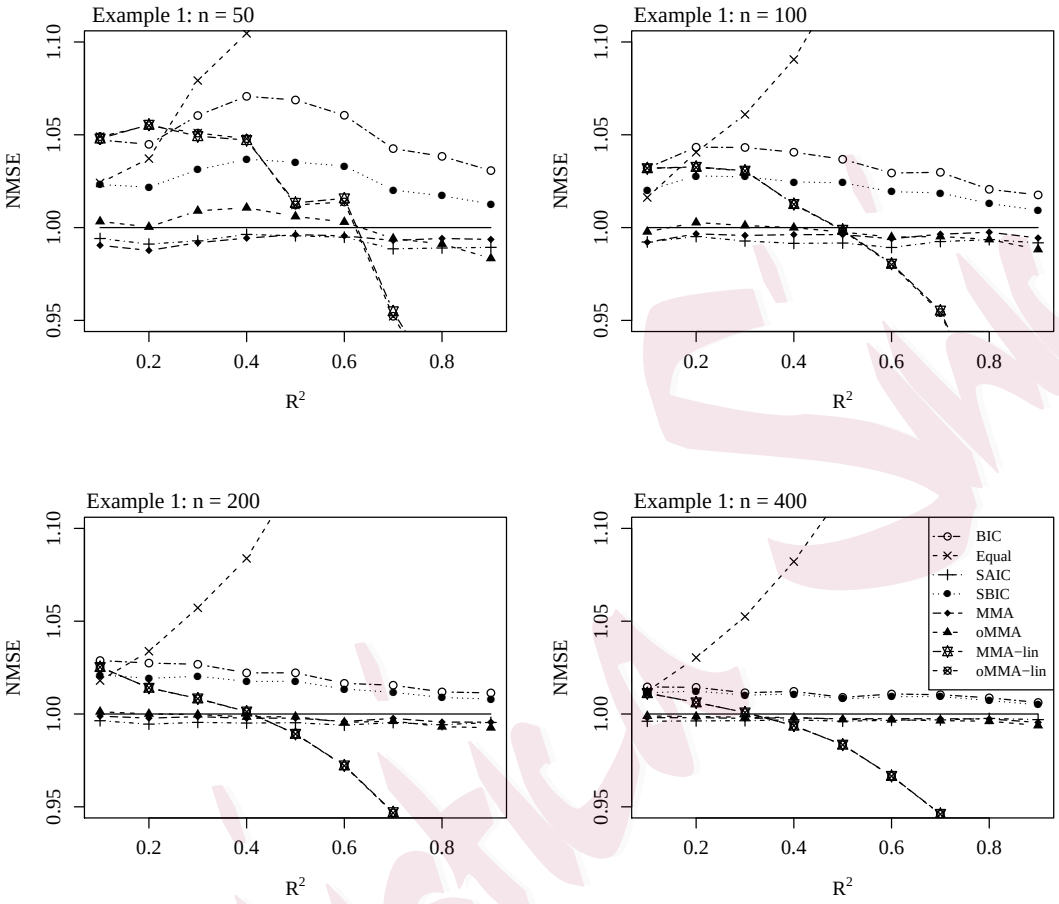

Figure 1: Normalized mean squared error (NMSE) comparisons for Example 1.

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 performs better than 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 knowl-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0067

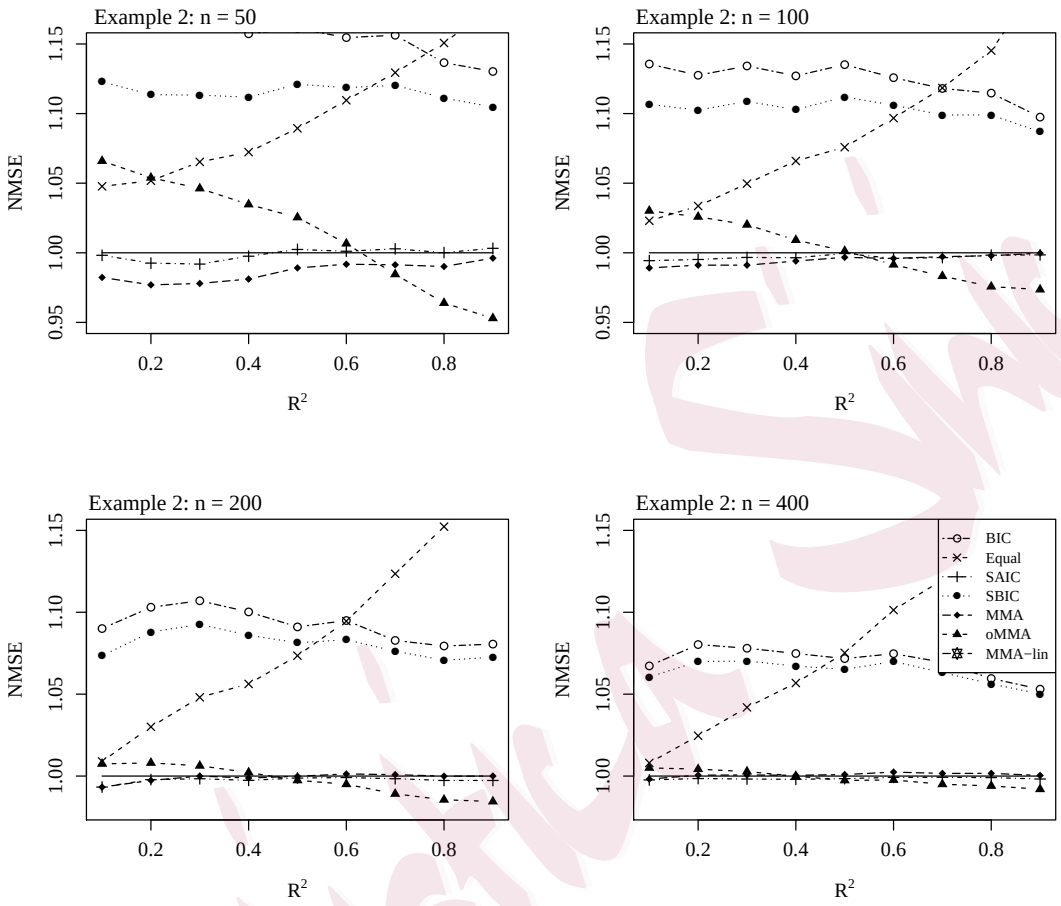

Figure 2: Normalized mean squared error (NMSE) comparisons for Example 2.

edge of the variance matrix $\mathbf{\Omega}$ 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 the oMMA's ability to effectively distinguish between signal (in candidate models) and approximation. In contrast, MMA employs an estimated variance matrix derived from the largest candidate model, which

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0067

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.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0067

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 7498 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 et al. (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 et al. (2017), we log-transformed all scalar variables except pH values and 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

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0067

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

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 $D = 1000$ runs, and for each repetition, we evaluated the mean squared prediction error (MSPE) using

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

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 $\widehat{\mu}_i^{(d)}$ 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

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0067

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	2.37E-29	3.93E-13	4.81E-38
t stat.	-13.93	-10.08	-16.50
t p.val	3.03E-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

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 (Racine and Parmeter, 2014), employing paired t-tests and Mann-Whitney-Wilcoxon (MWW) tests. These methods are designed to testing whether two competing approximate

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0067

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

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0067

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 et al. (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 Materials

The supplementary material contains additional simulations, additional details in real application, and the detailed proofs for Lemma 1, Theorems 1 and 2.

Acknowledgements

The authors greatly appreciate two referees and the associate editor for insightful comments. S. Liu's research was supported by the Zhejiang Provincial Department of

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0067

Education Scientific Research Project (No.Y202147117), and the Zhejiang Provincial Basic Public Welfare Research Program's Natural Science Foundation Exploration Project (No.LQ23A010017). C. Zhang's research was supported in part by the U.S. National Science Foundation grants DMS-2013486 and DMS-1712418, and provided by the University of Wisconsin-Madison Office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation. J. Zhang's research was supported by Public Health & Disease Control and Prevention, Fund for Building World-Class Universities (Disciplines) of Renmin University of China (to J.Z.) and the MOE Project of Key Research Institute of Humanities and Social Sciences (22JJD910001).

References

Ando, T. and K.-C. Li (2014). A model-averaging approach for high-dimensional regression. *Journal of the American Statistical Association 109*(505), 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*(2-3), 359–377.

Aneiros-Pérez, G. and P. Vieu (2006). Semi-functional partial linear regression. *Statistics & Probability Letters 76*(11), 1102–1110.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological) 57*(1), 289–300.

Buckland, S. T., K. P. Burnham, and N. H. Augustin (1997). Model selection: an integral part of inference.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0067

Biometrics 53(2), 603–618.

Cai, T. T. and P. Hall (2006). Prediction in functional linear regression. *The Annals of Statistics 34*(5), 2159–2179.

Cai, T. T. and M. Yuan (2012). Minimax and adaptive prediction for functional linear regression. *Journal of the American Statistical Association 107*(499), 1201–1216.

Cardot, H., F. Ferraty, and P. Sarda (1999). Functional linear model. *Statistics & Probability Letters 45*(1), 11–22.

Cardot, H., F. Ferraty, and P. Sarda (2003). Spline estimators for the functional linear model. *Statistica Sinica 13*, 571–591.

Cheng, X. and B. E. Hansen (2015). Forecasting with factor-augmented regression: A frequentist model averaging approach. *Journal of Econometrics 186*(2), 280–293.

Claeskens, G. and N. L. Hjort (2008). *Model selection and model averaging*. Cambridge University Press.

Clyde, M. and E. I. George (2004). Model uncertainty. *Statistical science 19*, 81–94.

Diebold, F. X. and R. S. Mariano (2002). Comparing predictive accuracy. *Journal of Business & economic statistics 20*(1), 134–144.

Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society: Series B (Methodological) 57*(1), 45–70.

Fan, J. and J. Lv (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society Series B: Statistical Methodology 70*(5), 849–911.

Fan, J. and R. Song (2010). Sure independence screening in generalized linear models with np-dimensionality. *The Annals of Statistics 38*(6), 3567–3604.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0067

Fan, Y., G. M. James, and P. Radchenko (2015). Functional additive regression. *The Annals of Statistics 43*(5), 2296–2325.

Fang, F., J. Li, and X. Xia (2022). Semiparametric model averaging prediction for dichotomous response. *Journal of Econometrics 229*(2), 219–245.

Gao, Y., X. Zhang, S. Wang, and G. Zou (2016). Model averaging based on leave-subject-out cross-validation. *Journal of Econometrics 192*(1), 139–151.

Hall, P., H.-G. Müller, and J.-L. Wang (2006). Properties of principal component methods for functional and longitudinal data analysis. *The annals of statistics 34*(3), 1493–1517.

Hansen, B. E. (2007). Least squares model averaging. *Econometrica 75*(4), 1175–1189.

Hansen, B. E. and J. S. Racine (2012). Jackknife model averaging. *Journal of Econometrics 167*(1), 38–46.

Hjort, N. L. and G. Claeskens (2003). Frequentist model average estimators. *Journal of the American Statistical Association 98*(464), 879–899.

Kong, D., K. Xue, F. Yao, and H. H. Zhang (2016). Partially functional linear regression in high dimensions. *Biometrika 103*(1), 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*(3), 958–975.

Liang, H., G. Zou, A. T. Wan, and X. Zhang (2011). Optimal weight choice for frequentist model average estimators. *Journal of the American Statistical Association 106*(495), 1053–1066.

Liu, Q. and R. Okui (2013). Heteroskedasticity-robust cp model averaging. *Econometrics Journal 16*, 463–472.

Müller, H.-G. and F. Yao (2008). Functional additive models. *Journal of the American Statistical Associa-*

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0067

tion 103(484), 1534–1544.

Racine, J. S. and C. F. Parmeter (2014, 02). Data-Driven Model Evaluation: A Test for Revealed Performance. In *The Oxford Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics*, pp. 308–345. Oxford University Press.

Ramsay, J. O. and B. W. Silverman (2005). *Functional data analysis (2nd ed.)*. New York: Springer.

Rice, J. A. and B. W. Silverman (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society: Series B (Methodological) 53*(1), 233–243.

Sila, A., G. Pokhariyal, and K. Shepherd (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*(3), 413–436.

Tang, Q., W. Tu, and L. Kong (2023). Estimation for partial functional partially linear additive model. *Computational Statistics & Data Analysis 177*, 107584.

Wan, A. T., X. Zhang, and G. Zou (2010). Least squares model averaging by mallows criterion. *Journal of Econometrics 156*(2), 277–283.

Wang, G., X.-N. Feng, and M. Chen (2016). Functional partial linear single-index model. *Scandinavian Journal of Statistics 43*(1), 261–274.

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.

Yao, F. and H.-G. Müller (2010). Functional quadratic regression. *Biometrika 97*(1), 49–64.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0067

Yao, F., H.-G. Müller, and J.-L. Wang (2005). Functional linear regression analysis for longitudinal data. *The Annals of Statistics 33*(6), 2873–2903.

Yu, D., L. Kong, and I. Mizera (2016). Partial functional linear quantile regression for neuroimaging data analysis. *Neurocomputing 195*, 74–87.

Zhang, H. and G. Zou (2020). Cross-validation model averaging for generalized functional linear model. *Econometrics 8*(1), 7.

Zhang, X., J.-M. Chiou, and Y. Ma (2018). Functional prediction through averaging estimated functional linear regression models. *Biometrika 105*(4), 945–962.

Zhang, X. and C.-A. Liu (2023). Model averaging prediction by k-fold cross-validation. *Journal of Econometrics 235*(1), 280–301.

Zhang, X., A. T. Wan, and S. Z. Zhou (2012). Focused information criteria, model selection, and model averaging in a tobit model with a nonzero threshold. *Journal of Business & Economic Statistics 30*(1), 132–142.

Zhang, X., A. T. Wan, and G. Zou (2013). Model averaging by jackknife criterion in models with dependent data. *Journal of Econometrics 174*(2), 82–94.

Zhang, X. and W. Wang (2019). Optimal model averaging estimation for partially linear models. *Statistica Sinica 29*, 693–718.

Zhang, X., D. Yu, G. Zou, and H. Liang (2016). Optimal model averaging estimation for generalized linear models and generalized linear mixed-effects models. *Journal of the American Statistical Association 111*(516), 1775–1790.

Zhang, X. and J. Yu (2018). Spatial weights matrix selection and model averaging for spatial autoregressive

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0067

models. *Journal of Econometrics 203*(1), 1–18.

Zhang, X., G. Zou, and R. J. Carroll (2015). Model averaging based on kullback-leibler distance. *Statistica Sinica 25*, 1583–1598.

Zhang, X., G. Zou, and H. Liang (2014). Model averaging and weight choice in linear mixed-effects models. *Biometrika 101*(1), 205–218.

Zhang, X., G. Zou, H. Liang, and R. J. Carroll (2020). Parsimonious model averaging with a diverging number of parameters. *Journal of the American Statistical Association 115*(530), 972–984.

Zhu, H., F. Yao, and H. H. Zhang (2014). Structured functional additive regression in reproducing kernel hilbert spaces. *Journal of the Royal Statistical Society. Series B, Statistical methodology 76*(3), 581–603.

Zhu, R., A. T. Wan, X. Zhang, and G. Zou (2019). A mallows-type model averaging estimator for the varying-coefficient partially linear model. *Journal of the American Statistical Association 114*(526), 882–892.

Zhu, R., G. Zou, and X. Zhang (2018). Optimal model averaging estimation for partial functional linear models. *Journal of Systems Science and Mathematical Sciences 38*, 777–800.

Zou, J., W. Wang, X. Zhang, and G. Zou (2022). Optimal model averaging for divergent-dimensional poisson regressions. *Econometric Reviews 41*(7), 775–805.