econometrics

MDPI

Article

Cross-Validation Model Averaging for Generalized Functional Linear Model

Haili Zhang [1,2] and Guohua Zou [3,*]

[1] University of Chinese Academy of Sciences, Beijing 100049, China; haili@amss.ac.cn
[2] Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China
[3] School of Mathematical Sciences, Capital Normal University, Beijing 100048, China
* Correspondence: ghzou@amss.ac.cn; Tel.: +86-15001122035

Received: 2 September 2019; Accepted: 18 February 2020; Published: 24 February 2020

Abstract: Functional data is a common and important type in econometrics and has been easier and easier to collect in the big data era. To improve estimation accuracy and reduce forecast risks with functional data, in this paper, we propose a novel cross-validation model averaging method for generalized functional linear model where the scalar response variable is related to a random function predictor by a link function. We establish asymptotic theoretical result on the optimality of the weights selected by our method when the true model is not in the candidate model set. Our simulations show that the proposed method often performs better than the commonly used model selection and averaging methods. We also apply the proposed method to Beijing second-hand house price data.

Keywords: generalized functional linear model; cross-validation; model averaging; asymptotic optimality

1. Introduction

In recent years, functional data have been increasingly popular in many scientific areas. A common question for the functional data is how to quantify the relationship between functional covariates and scalar responses. Functional linear model (FLM) and generalized functional linear model (GFLM) can take account of some associations between the response and the different points in the domain of the functional covariates, and therefore are two useful tools in many studies for functional data. These two models have now been widely used to solve practical problems, such as exploring the relationship between the growth and age in the life sciences, analyzing the weather data in different areas, recognizing the handwriting data, and conducting the diffusion tensor imaging studies. Functional data analysis usually represents functional covariates and coefficient functions by some linear combinations of a set of basis functions, such as a prespecified basis system like B-splines, Fourier and wavelet bases (James 2002), and data-adaptive basis functions from functional principal component analysis (FPCA) (Yao et al. 2005). We are concerned with the GFLM because it can estimate the flexible and nonlinear relationships between the functional covariates and scalar responses for many types of data such as binary response data, Poisson response data, and multivariate discrete response data. See, for example, James (2002), who expanded generalized linear models to generalized functional linear models with the functional principal component methodology and demonstrated that this approach can be performed for linear, logistic and censored regressions in simulations and real data analysis.

In econometrics, the relationship between time series and scalar response is often of interest. We can use GFLM instead of generalized linear model to handle the case where a time series with the dependence at different time points is used as the explanatory variables with dimension toward to infinity. On the other hand, prediction is often the main goal in econometric data analysis. Several

approaches have been proposed to select some important principal components in FPCA such as AIC, BIC, and leave-one-out cross-validation (Müller and Stadtmüller 2005). However, as we will demonstrate, the model selection alone, such as AIC, is not an optimal approach for the purpose of estimation and prediction. in one model selected by AIC or BIC may lead to the loss of information from other models. Different models often capture different data characteristics and therefore model averaging generally gets higher estimating or predicting accuracy, which has received extensive attention in recent years.

Model averaging has two research directions: Bayesian Model Averaging (BMA) and Frequentist Model Averaging (FMA). We will focus on the latter in this paper. A key problem with the FMA is the choice of weights assigned to different models. In this regard, various approaches have been developed. See, for example, smoothed AIC, smoothed BIC (Buckland et al. 1997), smoothed FIC (Hjort and Claeskens 2003; Claeskens and Carroll 2007; Zhang and Liang 2011; Zhang et al. 2012; Xu et al. 2014), Adaptive method (Yang 2001), MMA method (Hansen 2007; Wan et al. 2010), OPT method (Liang et al. 2011), JMA method (Hansen and Racine 2012; Zhang et al. 2013), and leave-subject-out cross-validation method (Gao et al. 2016), which apply to independent, or time series, or longitudinal data.

For functional data, some model averaging methods have been studied. Zhu et al. (2018) proposed a model averaging estimator based on Mallows' criterion for partial functional linear models whose response is a scalar and the predictors are a random vector and some functional variables. Zhang et al. (2018) proposed a Jackknife model averaging for fully functional linear models whose response and predictor are both functional processes. For generalized functional linear model designed for the case where the scalar response is nonlinearly dependent on functional explanatory variables, model averaging is a good alternative to model selection that may lead to instability in variable selection or coefficient estimation caused by randomness of the data collection and so on.

In this article, we consider model averaging methods for GFLM to capture the nonlinear characteristics hidden in the data and to reduce the prediction errors and risks. The contributions of this article are threefold: We first adopt FPCA to reduce the dimensions as it provides a parsimonious representation of functional data, and then present a novel model averaging procedure based on leave-one-out cross-validation criterion (CV). Second, we prove the consistency of parameter estimator under the misspecified model with some mild conditions. The dimension of the parameter can be divergent. Third, we establish the asymptotic optimality of our method in the squared loss sense for generalized linear model with a diverging number of parameters. Our work relaxes the condition that the expectations of estimators need to exist.

The rest of the article is organized as follows. In Section 2, we introduce our proposed model averaging method for GFLM. We then establish the asymptotic property of the proposed method in Section 3. Simulation studies and a real data example of second-hand house price in Beijing are presented in Section 4. Section 5 concludes. Proofs of theoretical results are provided in Appendix A and B.

2. Model Averaging for Generalized Functional Linear Model

2.1. The Generalized Functional Linear Model

The data we collected for the ith subject or experimental unit are $(\{X_i(t), t \in T\}, y_i), i = 1, \ldots, n$. We assume these data are generated independently. The predictor variable $X(t)$ $(t \in T)$ is a random curve corresponding to a square integrable stochastic process on a real interval T. The response variable is a real-valued random variable that may be continuous or discrete. For example, in a binary regression, one would have $y \in \{0, 1\}$.

Suppose that the given link function $g(\cdot)$ is a strictly monotone and twice continuously differentiable function with bounded derivatives and is thus invertible. This assumption is common in generalized linear model. See, for example, (Chen et al. 1999; Müller and Stadtmüller 2005; Ando and Li 2017). Moreover, we assume a variance function $\sigma^2(\cdot)$, which is strictly positive with upper

bound defined on the range of the link function. The generalized functional linear model or functional quasi-likelihood model is determined by a parameter function $\beta(\cdot)$, which is square integrable on its domain T, in addition to the link function $g(\cdot)$ and the variance function $\sigma^2(\cdot)$.

Given a real measure $d\omega$ on T, we define linear predictors

$$\eta_i = \alpha + \int \beta(t) X_i(t) d\omega(t), \qquad i = 1, \ldots n,$$

and conditional means $\mu_i = g(\eta_i)$, where $\mathbf{E}\left(y_i | X_i\left(t\right), t \in T\right) = \mu_i$, and $\mathbf{Var}\left(y_i | X_i(t), t \in T\right) = \sigma^2(\mu_i) = \tilde{\sigma}^2(\eta_i)$ with the function $\tilde{\sigma}^2(\eta_i) = \sigma^2(g(\eta_i))$. In a generalized functional linear model, the distribution of y_i would be specified with the exponential family. Thus, we should consider a functional quasi-likelihood model

$$y_i = g\left(\alpha + \int \beta(t) X_i(t) d\omega(t)\right) + e_i, \qquad i = 1, \ldots n, \tag{1}$$

where $\mathbf{E}\left(e_i | X_i(t), t \in T\right) = 0$ and $\mathbf{Var}\left(e_i | X_i(t), t \in T\right) = \sigma^2(\mu_i) = \tilde{\sigma}^2(\eta_i)$. Note that α is a constant, and the inclusion of an intercept allows us to require $\mathbf{E}\left(X_i(t)\right) = 0$ for all t. We assume the errors e_i are independent with the same variance. It is easy to obtain $\mathbf{E}(e_i) = 0$ and

$$\mathbf{Var}(e_i) = \mathbf{Var}\left\{\mathbf{E}(e_i | X_i(t), t \in T)\right\} + \mathbf{E}\{\mathbf{Var}(e_i | X_i(t), t \in T)\} = \mathbf{E}\left\{\tilde{\sigma}^2(\eta_i)\right\} = \sigma^2.$$

Following Müller and Stadtmüller (2005), we choose an orthonormal basis $\left\{\rho_j, i = 1, 2 \ldots\right\}$ of the function space $L^2(d\omega)$, that is $\int_T \rho_j(t)\rho_k(t)d\omega(t) = \delta_{jk}$, where $\delta_{jk} = 0$ for $j \neq k$ and $\delta_{jk} = 1$ for $j = k$. Then, we can expand the predictor process $X(t)$ and the parameter function $\beta(t)$ as

$$X(t) = \sum_{j=1}^{\infty} \varepsilon_j \rho_j(t),$$

and

$$\beta(t) = \sum_{j=1}^{\infty} \beta_j \rho_j(t),$$

[in the $L^2(d\omega)$ sense] with random variables ε_j and coefficients β_j given by $\varepsilon_j = \int X(t)\rho_j(t)d\omega(t)$ and $\beta_j = \int \beta(t)\rho_j(t)d\omega(t)$, respectively. By the previous assumptions that $X(t)$ and $\beta(t)$ are square integrable, we get $\sum_{j=1}^{\infty} \beta_j^2 < \infty$ and $\sum_{j=1}^{\infty} \mathbf{E}\varepsilon_j^2 < \infty$.

From the orthonormality of the basis function ρ_j and setting

$$\varepsilon_{i,j} = \int X_i(t)\rho_j(t)d\omega(t),$$

it follows immediately that

$$\eta_i = \alpha + \int \beta(t) X_i(t) d\omega(t) = \alpha + \sum_{j=1}^{\infty} \beta_j \varepsilon_{i,j}.$$

It will be convenient to work with standardized errors

$$e_i' = e_i / \sigma(\mu_i) = e_i / \tilde{\sigma}(\eta_i),$$

in which $\mathbf{E}(e_i' | X_i(t)) = 0$, $\mathbf{E}(e_i') = 0$, and $\mathbf{E}(e_i'^2) = 1$. Then, it will be sufficient to consider the following model,

$$y_i = g\left(\alpha + \sum_{j=1}^{\infty} \beta_j \varepsilon_{i,j}\right) + e_i' \tilde{\sigma}\left(\alpha + \sum_{j=1}^{\infty} \beta_j \varepsilon_{i,j}\right), \quad i = 1, \ldots, n, \tag{2}$$

where the function $g(\cdot)$ is known.

The number of parameter in model (2) is infinite. We address the difficulty caused by the infinite dimensionality of the predictors by approximating model (2) with a series of models where the number of predictors is truncated at $p = p_n$, and the dimension p_n can be a constant as large as possible with $p_n < n$. A heuristic truncation strategy is as follows. For the ith sample, a p-truncated linear predictor $\eta_{i,p}$ is

$$\eta_{i,p} = \alpha + \sum_{j=1}^{p} \beta_j \varepsilon_{i,j}.$$

The approximating model we use is

$$y_i = g\left(\alpha + \sum_{j=1}^{p} \beta_j \varepsilon_{i,j}\right) + e_i' \tilde{\sigma}\left(\alpha + \sum_{j=1}^{p} \beta_j \varepsilon_{i,j}\right), \quad i = 1, \ldots, n.$$

Now, we consider the estimation for generalized functional linear model. First, we use FPCA to get a set of orthogonal eigenfunctions as the basis functions in the space $L^2(d\omega)$. Then, we consider a series of candidate models. The number of candidate models is M. For the mth candidate model, we adopt the first p_m functional principal components to build the approximating model,

$$y_i = g\left(\alpha^{(m)} + \sum_{j=1}^{p_m} \beta_j^{(m)} \varepsilon_{i,j}\right) + e_i' \tilde{\sigma}\left(\alpha^{(m)} + \sum_{j=1}^{p_m} \beta_j^{(m)} \varepsilon_{i,j}\right), \quad i = 1, \ldots, n. \tag{3}$$

We assume that $p_1 < p_2 < \cdots < p_M$. That is, the candidate models are nested. Denote $\varepsilon_{i,0} = 1$ and $\beta_0^{(m)} = \alpha^{(m)}$, then we estimate the unknown parameter vector $\beta^{(m)} = (\beta_0^{(m)}, \beta_1^{(m)}, \ldots, \beta_{p_m}^{(m)})^T$ by solving the following estimating or score equation

$$U_{n,m}(\beta^{(m)}) = \frac{1}{n} \sum_{i=1}^{n} \left[y_i - g(\eta_{i,p_m})\right] \frac{g'(\eta_{i,p_m})}{\sigma^2(\mu_{i,p_m})} \varepsilon_{(i,p_m)} = 0, \tag{4}$$

where $\eta_{i,p_m} = \sum_{j=0}^{p_m} \beta_j^{(m)} \varepsilon_{i,j}$ and $\varepsilon_{(i,p_m)} = (\varepsilon_{i,0}, \ldots, \varepsilon_{i,p_m})^T$. Let $\hat{\beta}^{(m)}$ be the solution of the score equation $U_{n,m}(\beta^{(m)}) = 0$, i.e.,

$$U_{n,m}(\hat{\beta}^{(m)}) = \frac{1}{n} \sum_{i=1}^{n} [y_i - g(\hat{\eta}_{i,p_m})] \frac{g'(\hat{\eta}_{i,p_m})}{\sigma^2\left(g\left(\hat{\eta}_{i,p_m}\right)\right)} \varepsilon_{(i,p_m)} = 0. \tag{5}$$

2.2. Model Averaging Estimation

For each candidate model, we get the estimator of the unknown parameter vector by (4). Let

$$w \in H_n = \left\{ w \in [0,1]^M : \sum_{m=1}^{M} w_m = 1 \right\},$$

then we obtain the model averaging estimator of η_i:

$$\hat{\eta}_i(w) = \sum_{m=1}^{M} w_m \hat{\eta}_{i,p_m}, \tag{6}$$

where $\hat{\eta}_{i,p_m} = \sum_{j=0}^{p_m} \hat{\beta}_j^{(m)} \varepsilon_{ij}$. Thus, a model averaging estimator of the conditional mean μ_i is given by

$$\hat{\mu}_i(w) = g\left(\sum_{m=1}^{M} w_m \hat{\eta}_{i,p_m}\right). \tag{7}$$

Let $\tilde{\beta}_{-j}^{(m)}$ be the estimator of $\beta^{(m)}$ from (4) without the jth observation, that is,

$$U_{n,m,-j}(\beta^{(m)}) = \frac{1}{n-1}\sum_{i=1,i\neq j}^{n}\left[y_i - g(\eta_{i,p_m})\right]\frac{g'(\eta_{i,p_m})}{\sigma^2(\mu_{i,p_m})}\varepsilon_{(i,p_m)} = 0. \tag{8}$$

For the observation j, the leave-one-out truncated linear estimator of η_j under the mth model is

$$\tilde{\eta}_{j,p_m} = \tilde{\beta}_{-j}^{(m)T}\varepsilon_{(j,p_m)},$$

and the leave-one-out model averaging estimator of μ_j is

$$\tilde{\mu}_j(w) = g\left(\textstyle\sum_{m=1}^{M} w_m\tilde{\eta}_{j,p_m}\right).$$

Thus, we propose the following leave-one-out criterion for choosing weights in the model averaging estimator given by (7)

$$CV(w) = \sum_{i=1}^{n}(y_i - \tilde{\mu}_i(w))^2 = \sum_{i=1}^{n}\left[y_i - g\left(\textstyle\sum_{m=1}^{M} w_m\tilde{\eta}_{i,p_m}\right)\right]^2. \tag{9}$$

Let

$$\hat{w} = arg\min_{w\in H_n} CV(w)$$

be the weight vector from $CV(w)$ criterion. Then, plugging $\hat{w}$ into (7), we obtain the final model averaging estimator $\hat{\mu}_i(\hat{w}), i = 1, 2, \ldots, n$.

3. Asymptotic Property for Model Averaging Estimator

In this section, we will establish the optimal property of cross-validation model averaging for generalized functional linear model. We allow the dimension of each candidate model to be divergent as n tends to ∞.

Notations and Conditions

We denote the first and second derivatives of the function $g(\cdot)$ by $g^{'}(\ldots)$ and $g^{''}(\ldots)$, respectively, the diagonal matrix A with diagonal elements $a_1, a_2, \ldots, a_q$ by $A = diag(a_1, a_2, \ldots, a_q)$, the minimum singular value of matrix A by $\lambda_{\min}\{A\}$, and

$$\lambda_{n,m} = \lambda_{min}\left\{\frac{\varepsilon_n^{(m)T}\varepsilon_n^{(m)}}{n}\right\},$$

with

$$\varepsilon_n^{(m)} = \left(\varepsilon_{(1,p_m)}, \varepsilon_{(2,p_m)}, \ldots, \varepsilon_{(n,p_m)}\right)^T.$$

For any $\beta^{(m)} \in \mathbb{R}^{p_m+1}$, $n \in \mathbb{N}^+$, define

$$U_{n,m}^{(m)}\left(\beta^{(m)}\right) = \frac{1}{n}\sum_{i=1}^{n}\left[g\left(\beta^{(m)T}\varepsilon_{(i,p_m)}\right) - \mu_i\right]\frac{g^{'}\left(\beta^{(m)T}\varepsilon_{(i,p_m)}\right)}{\sigma^2\left(g\left(\beta^{(m)T}\varepsilon_{(i,p_m)}\right)\right)}\varepsilon_{(i,p_m)}. \tag{10}$$

We assume $|g'(\cdot)| \leqslant c < \infty$ and $|g''(\cdot)| \leqslant c_1 < \infty$, and $\sigma^2(\cdot)$ is strictly positive with bound $0 < d_1 \leqslant \sigma^2(\cdot) \leqslant d_2 < \infty$ and $\left|\sigma^{2'}(\cdot)\right| \leqslant d_3 < \infty$.

Consider the squared loss function

$$L_n(w) = \|\mu - \widehat{\mu}(w)\|^2,$$

where $\mu = (\mu_1, \mu_2, \ldots, \mu_n)^T$ and $\widehat{\mu}(w) = (\hat{\mu}_1(w), \hat{\mu}_2(w), \ldots, \hat{\mu}_n(w))^T$ are the two $n \times 1$ vectors, and $\|\cdot\|^2$ is Euclidean norm. Denote

$$R_n(w) = \sum_{i=1}^{n} \left[g(\eta_i) - g\left(\sum_{m=1}^{M} w_m \varepsilon_{(i,p_m)}^T \beta_\star^{(m)} \right) \right]^2,$$

and

$$\xi_n = \inf_{w \in H_n} R_n(w),$$

where $\beta_\star^{(m)}$ is the pseudo true parameter, which, like Flynn et al. (2013) and Lv and Liu (2014), is defined as the solution to the following score equation,

$$U_{n,m}^{(m)}\left(\beta^{(m)}\right) = 0,$$

and is a theoretical target under the mth candidate model with misspecification. We assume that such a solution is existent and $\left\|\beta_\star^{(m)}\right\|^2 / (p_m + 1) \leq C_b < \infty$. ξ_n represents the minimal bias between the true model and the final model generated by model averaging, which is an alternative to the risk based on $L_n(w)$. In this work, we do not require the expectation of $L_n(w)$ to exist, which is more relaxed than the common requirement on jackknife model averaging methods for generalized linear model. See, for example, Zhang et al. (2016) and Ando and Li (2017). In the following, we assume that $X_i(t), i = 1, 2, \ldots, n$ are non-random with $\sup_i |\eta_i| \leqslant C_\eta < \infty$.

Condition 1. *For some compact set Θ_m in $\mathbb{R}^{p_m+1}$,*

$$\lim_{n \to +\infty} \mathbf{P}\left\{0 \in U_{n,m}(\Theta_m)\right\} = 1$$

holds.

Condition 2. *(i) $\{e_i\}, i = 1, \ldots, n$ are mutually independent.*
(ii) $\mathbf{E}e_i = 0$.
(iii) $C_1 = \sup_i \mathbf{E}e_i^2 < \infty$.

Condition 3. $\sup_i \left\|\varepsilon_{(i,p_m)}\right\|^2 / (p_m + 1) \leqslant C_2 < \infty$.

Condition 4. $\sqrt{n}p^2/\xi_n \to 0$ *with* $p = \max_m p_m$ *and* $p^4/n = o(1)$.

Condition 5. $\sum_{i=1}^{n} \left(\tilde{\eta}_{i,p_m} - \hat{\eta}_{i,p_m}\right)^2 = O_p\left(p_m^4\right)$.

Condition 6. $\lambda_{min}\left\{\frac{\partial}{\partial \beta^{(m)}} U_{n,m}^{(m)}\left(\beta^{(m)}\right)\right\} \geqslant C_0 > 0$.

Condition 1 is a requirement for generalized model to guarantee the existence of solutions to (4). In general, the existence and consistency of roots obtained by solving (4) have to be checked, so we list Condition 1. The similar condition can be found in Balan and Schiopu-Kratina (2005). In the special case where the link function is $g(x) = x$, the solution of (4) is a generalized least squares estimator of $\beta^{(m)}$ and Condition 1 is easy to satisfy.

Condition 2 is common for generalized linear model. See, for instance, Chen et al. (1999) and Ando and Li (2017). The least squares estimator for linear regression models is strongly consistent under Condition 2. This condition is less restrictive than (A1) of Ando and Li (2017) for proving the optimality of the weight selection procedure.

Condition 3 is similar to (2.3) of Theorem 1 in Chen et al. (1999) and is due to the nonlinearity. A counterexample is given to show that $\hat{\beta}^{(m)}$ may not be consistent when Condition 3 (i) is dropped in Chen et al. (1999).

Condition 4 means that the speed of ξ_n tending to ∞ should be faster than that of $\sqrt{n}p^2$. This condition also implies that the true model is not in the candidate model set, which is a condition commonly used for optimal model averaging. It is easy to satisfy when the true model is an infinite dimensional model. This condition is an alternative to Condition C.3 of Zhang et al. (2016) and (A3) of Ando and Li (2017).

Condition 5 implies $n^{-1}\sum_{i=1}^{n}\left(\tilde{\eta}_{i,p_m}-\hat{\eta}_{i,p_m}\right)^2 = o_p(1)$ with $p_m^4/n = o(1)$. By Lemma A3 in the Appendix A and Condition 3, we have

$$\sum_{i=1}^{n}\left(\hat{\eta}_{i,p_m}-\varepsilon_{(i,p_m)}^T\beta_\star^{(m)}\right)^2 \leqslant \sum_{i=1}^{n}\left\|\varepsilon_{(i,p_m)}\right\|^2\left\|\hat{\beta}^{(m)}-\beta_\star^{(m)}\right\|^2 = O_p\left(p_m^4\right).$$

Then, with the following standard condition for the application of cross-validation,

$$\left|\frac{\sum_{i=1}^{n}\left(\tilde{\eta}_{i,p_m}-\varepsilon_{(i,p_m)}^T\beta_\star^{(m)}\right)^2}{\sum_{i=1}^{n}\left(\hat{\eta}_{i,p_m}-\varepsilon_{(i,p_m)}^T\beta_\star^{(m)}\right)^2}-1\right| = o_p(1),$$

which says that as n gets large, the difference between the ordinary and leave-one-out estimators of η_i under the mth candidate model gets small, it can be seen that

$$\sum_{i=1}^{n}\left(\tilde{\eta}_{i,p_m}-\hat{\eta}_{i,p_m}\right)^2 \leqslant 2\sum_{i=1}^{n}\left(\tilde{\eta}_{i,p_m}-\varepsilon_{(i,p_m)}^T\beta_\star^{(m)}\right)^2 + 2\sum_{i=1}^{n}\left(\hat{\eta}_{i,p_m}-\varepsilon_{(i,p_m)}^T\beta_\star^{(m)}\right)^2 = O_p\left(p_m^4\right),$$

which means Condition 5 is reasonable. For the one-parameter natural exponential family models, Ando and Li (2017) showed under some regularity conditions that $\sum_{i=1}^{n}\left(\tilde{\eta}_{i,p_m}-\hat{\eta}_{i,p_m}\right)^2 = O_p\left(p_m^2/n\right)$ satisfying our Condition 5. For the linear models where $g(x)=x$ and $\sigma^2(\cdot)=1$, $\sum_{i=1}^{n}\left(\tilde{\eta}_{i,p_m}-\hat{\eta}_{i,p_m}\right)^2 = O_P\left(p_m^2/n\right)$ under the assumption that $\varepsilon_{(i,p_m)}^T\left(\varepsilon_n^{(m)T}\varepsilon_n^{(m)}\right)^{-1}\varepsilon_{(i,p_m)} \leq cp_m/n$ for some constant $c<\infty$, which is commonly used to ensure the asymptotic optimality of cross-validation. See, for example, Condition (5.2) of Li (1987), Condition (5.2) of Andrews (1991), Condition (A.9) of Hansen and Racine (2012), Condition (C.2) of Zhang (2015), and Condition (C.3) of Zhao et al. (2018). In general, our Condition 5 is more relaxed than those in literature for the complex candidate models.

Condition 6 is to ensure that the pseduo true parameter $\beta_\star^{(m)}$ is unique. The consistency of the estimator of $\beta_\star^{(m)}$ can also be derived by this condition. See Lemma A3 in the Appendix A. In addition, the one-parameter natural exponential family considered in Theorem 1 of Ando and Li (2017) is an example with

$$\lambda_{min}\left\{\frac{\partial}{\partial\beta^{(m)}}U_{n,m}^{(m)}\left(\beta^{(m)}\right)\right\} = \lambda_{min}\left\{\frac{1}{n}\varepsilon_n^{(m)T}\Gamma\left(\beta^{(m)}\right)\varepsilon_n^{(m)}\right\},$$

where

$$\Gamma\left(\beta^{(m)}\right) = diag\left(g'\left(\varepsilon_{(1,p_m)}^T\beta^{(m)}\right), g'\left(\varepsilon_{(2,p_m)}^T\beta^{(m)}\right), \ldots, g'\left(\varepsilon_{(n,p_m)}^T\beta^{(m)}\right)\right).$$

By the commonly used assumption that $\lambda_{n,m} \geqslant c_0 > 0$ for some constant $c_0 < \infty$, and the assumption (4.3) in Ando and Li (2017), this example satisfies Condition 6.

Theorem 1. *Assume that Conditions 1–6 hold, then $\hat{w}$ is asymptotically optimal in the sense that*

$$\frac{L_n(\hat{w})}{\inf_{w\in H_n}L_n(w)} \xrightarrow{p} 1, \tag{11}$$

where $\xrightarrow{p}$ *means convergence in probability.*

Proof. See the Appendix B. □

Remark 1. *When the dimensions of the candidate models are fixed, condition 4 can be relaxed to* $n/\xi_n^2 \to 0$.

Remark 2. *It is easy to see that if we do not require that the weights sum to one, then we can use M instead of* 1 *as the upper bound of* $\sum_{m=1}^{M} w_m^2$ *in our proof. Thus, all the proofs are still valid for the fixed M. This implies that Theorem 1 remains true if we remove the constraint that the weights sum to one. In addition, as the candidate models are not necessarily nested in the proof, this theorem still holds when the candidate models are non-nested.*

4. Numerical Examples

4.1. Simulation I: Fixed Number of Candidate Models

In this section, we conduct simulation experiments to compare the finite sample performance of our model averaging methods and some commonly used model selection and model averaging methods. For model selection, we consider three methods: AIC, BIC, and FPCA. FPCA is an efficient and common method in functional data analysis, which determines the final model by the cumulative contributions of the functional principal components. For model averaging, we consider the following methods, S-AIC (smoothed AIC), S-BIC (smoothed BIC), and our cross-validation model averaging, which is denoted as CV1 if we restrict the sum of weights to be 1 as before, and CV2 if no constraint on the sum of weights is imposed.

The data generating process is as follows: the predictor variable is

$$X_i(t) = \sum_{j=1}^{J} \varepsilon_{i,j}\rho_j(t),$$

and the parameter function is

$$\beta(t) = \sum_{j=1}^{J} \beta_j\rho_j(t),$$

where $\rho_j(t)$ is a basis function with $t \in [0,1]$, and $j \geq 1$ and J is the number of the basis functions. Here, we use B-spline base and Fourier base. For B-spline base, we choose the order of the basis functions to be 2, and the number of the basis functions to be 20. As for Fourier base, we choose the number of the basis functions as 21 and the first basis to be a constant function.

In our simulation, the following four cases are considered.

Case 1 For $1 \leq j \leq 10$, β_j are generated from the standard normal distribution $N(0,1)$; for $10 < j \leq 20$, $\beta_j = 0$. The basis functions $\{\rho_j(t), t \in [0,1], 1 \leq j \leq 20\}$ are B-spline functions with parameters as mentioned above.

Case 2 For $1 \leq j \leq 20$, $\beta_j = j^{-2}$. The basis functions $\{\rho_j(t), t \in [0,1], 1 \leq j \leq 20\}$ are B-spline functions with parameters as mentioned above.

Case 3 For $1 \leq j \leq 11$, β_j are generated from the standard normal distribution $N(0,1)$; for $11 < j \leq 21$, $\beta_j = 0$. The basis functions $\{\rho_j(t), t \in [0,1], 1 \leq j \leq 21\}$ are Fourier functions with parameters as mentioned above.

Case 4 For $1 \leq j \leq 21$, $\beta_j = j^{-2}$. The basis functions $\{\rho_j(t), t \in [0,1], 1 \leq j \leq 20\}$ are Fourier functions with parameters as mentioned above.

We set the term $\varepsilon_{i,j}$ to be independently generated from $N(0, R^2/j^2)$, where $R = 1, 2, \ldots, 10$. The response variable y_i is generated from binomial distribution $Binomial(p(X_i(t)), 1)$ with the probability $p(X_i(t))$ being $g\left(\int_0^1 X_i(t)\beta(t)\,dt\right)$. We consider three types of link function $g(\cdot)$: logistic

link function $exp(\cdot)/(1+exp(\cdot))$, Probit link function, and Poisson link function. For the Poisson model, we only consider the simulations with $R = 1$ for Cases 1–4.

In the simulation, we use FPCA to obtain the nested candidate models. Each candidate model contains the first p_m principal components. The number of candidate models is 18 for Cases 1–2 and 19 for Cases 3–4. Then we adopt the weighted iterated least squares algorithm which is a common approach in generalized linear model to get the estimates for each model. For the weights, we use the 'fmincon' function in Matlab to get the solution of CV criterion.

The sample size is set as $n = 60, 200, 500$. We use the 80% data as the training data $\{Y_1, X_1\}$ with size n_1, and the remaining data as the test data $\{Y_2, X_2\}$ with size n_2. Then, we compare the prediction errors. We calculate the prediction accuracy ($\left\|\hat{Y}_2 - Y_2\right\|^2 / n_2$), fitting accuracy ($\left\|\hat{Y}_1 - Y_1\right\|^2 / n_1$), predictor coefficient prediction accuracy ($\left\|\hat{\eta}_{(2)} - \eta_{(2)}\right\|^2 / n_2$), and predictor coefficient fitting accuracy ($\left\|\hat{\eta}_{(1)} - \eta_{(1)}\right\|^2 / n_1$). We repeat this process 1000 times, and then obtain mean, median, and variance of these prediction errors for each method. To save space, we present only the results on the prediction accuracy. The results on the other type accuracies are available from the authors upon request. We only report the results for logistic link function due to space limitations. Other link function results are also available from the authors.

For Case 1, the prediction errors are summarized in Tables A1–A3. From Table A1, it is seen that with R varying from 1 to 10, the prediction errors are decreasing, because the difference of probability between the two groups (one group whose response is 1 and the other group whose response is 0) becomes larger. Our methods (CV1 and CV2 in the tables) always obtain the minimum error means (Mean in the tables), medians (Median in the tables), and variances (Var in the tables). However, there is no clear tendency between CV1 and CV2, which perform similarly in most of situations. When R is small, BIC is always better than AIC, and S-BIC is always better than S-AIC. This may be due to less parameters being useful for smaller R values, and in this case, a bigger penalty on the number of parameters in the model is preferred. Moreover, when the candidate models differ significantly, AIC or BIC performs similarly to S-AIC or S-BIC, respectively. As R becomes larger, the difference between AIC and BIC or S-AIC and S-BIC becomes smaller. FPCA is always superior to AIC, BIC, S-AIC, and S-BIC, and their differences become larger as R increases. Now, we turn to Tables A2 and A3. With the sample size n increasing from 60 to 200 and 500, we can see that the prediction errors decrease for each fixed R. The median and variance of prediction errors also become smaller. AIC and BIC behave increasingly similarly. CV1 and CV2 are still the best among all the methods, and followed by FPCA.

For Case 2, the prediction errors are given in Tables A4–A6. As shown earlier, CV1 and CV2 perform the best, and followed by FPCA. Likewise, S-AIC or S-BIC is better than AIC or BIC, respectively. For Table A4, with R varying from 1 to 10, the prediction errors are decreasing except FPCA method, which gets the minimum at $R = 7$ with a small fluctuation. CV1 and CV2 perform equally well for different R values and sample sizes. The difference between AIC and BIC becomes small with the sample size increasing. The similar phenomenon is observed for S-AIC and S-BIC.

For Case 3, the prediction errors are provided in Tables A7–A9. For $n = 60$ (Table A7), CV1 or CV2 is the best when R is between 1 and 5. However, when R is between 6 and 10, the two model selection methods—AIC and BIC—are the best. The similar conclusions can be found in Table A8 with $n = 200$ and Table A9 with $n = 500$, although in the latter case, CV1 actually performs the best for all of R values. The error rates of all methods become smaller with R increasing from 1 to 6 and then bigger with R varying from 7 to 10.

For Case 4, the prediction errors are presented in Tables A10–A12. For $n = 60$ in Table A10, CV1, CV2, and BIC are the best, and followed by AIC . In this design, S-AIC or S-BIC is not better than AIC or BIC. For $n = 200$ in Table A11, BIC is the best, and followed by AIC. For $n = 500$ in Table A12, CV1 always performs the best, and followed by BIC.

In summary, for out-of-sample prediction, our methods CV1 and CV2 perform the best in most of cases and have smaller variances and medians of errors. Furthermore, CV1 and CV2 often perform

equally well. This indicates that removing the restriction on the sum of weights may not lead to a better model averaging estimates.

4.2. Simulation II: Divergent Number of Candidate Models

We consider the situations where the number of candidate models tends to ∞ as the sample size increases. We set the sample size n to be 200, 400, and 1000, and the the number of candidate models to be $9n/100$ (So M=18,36, and 90 for the three sample sizes). The data generating process is as before: the predictor variable is $X_i(t) = \sum_{j=1}^{J} \varepsilon_{i,j}\rho_j(t)$, and the parameter function is $\beta(t) = \sum_{j=1}^{J} \beta_j\rho_j(t)$, where $\rho_j(t)$ is a 2-order B-spline basis function, $t \in [0,1], j \geq 1$, and $J = n/10$. For $1 \leq j \leq J$, $\beta_j = j^{-1/2}$. We set the term $\varepsilon_{i,j}$ to be independently generated from $N(0, R^2/j^{1/2})$, where $R = 1, 3, 7$. The response variable y_i is generated from binomial distribution with the logistic link.

The candidate models are nested. The algorithms used in the calculations are the same as that described in Section 4.1. For the simulation results, we report the errors of seven methods considered as Section 4.1. From Tables A13–A15, our methods—CV1 and CV2—perform the best in most of cases, and followed by FPCA, and SAIC. The difference between AIC and BIC, or SAIC and SBIC is decreasing with increasing R .

4.3. Application: Beijing Second-Hand House Price Data

We apply our method to the Beijing second-hand housing transaction price data, which is captured from the internet collected by the Guoxinda Group Corporation. Most of the data pass through the manual check. This data include the second-hand housing prices and the surrounding environment variables of the 2318 residential areas in Beijing. The second-hand housing prices data are monthly data from January 2015 to December 2017 for each residential area.

Our aim is to predict the increase level in house prices in next year. We are concerned about the relationship between price level to rise and the past housing price curves. We use the median of listing online prices of houses in a residential area as the house price for this residential area. We use the price curve of each residential area from January 2015 to December 2016 as a predictor variable. The response variable is a binary variable, which takes 1 if the rising ratio is high, and 0 otherwise. Here, we define the rising ratio for each district as the ratio of the average monthly price in 2017 to the average monthly price in 2016. The 25%, 50%, and 75% quantile ratios are 1.31, 1.37, and 1.44 , respectively. We focus on the residential areas whose housing prices are rising rapidly, and so if the ratio is higher than 75% quantile ratios of all residential areas, the response variable of this residential area takes 1 as its value, and 0 otherwise. Of the $n = 2318$ residential areas, 568 are rising fast, and 1750 are not.

For simplicity, we standardize all the price data. For each group, we plot the housing price trajectories in Figure 1. Failure to visually detect differences between the groups could result from overcrowding of these plots with too many curves, but when displaying fewer curves (lower panels of Figure 1), the same phenomenon remains. With a few exceptions, no clear visual differences between the two groups can be discerned. On the whole, the trajectories of per year from 2015 to 2016 are not much different. Therefore, the discrimination task at hand is difficult.

We randomly select 75% of all residential areas as the training set with size 1739, and the rest as the testing data with size 579. We use logistic link and B-spline functions to fit the house price curves. The number of the basis functions is 6, and the order of the B-spline basis functions is 2. Then, we adopt functional principal component analysis (Yao et al. 2005) to built the data-adaptive basis functions to reduce the dimension and deal with the correlations in house price time series.

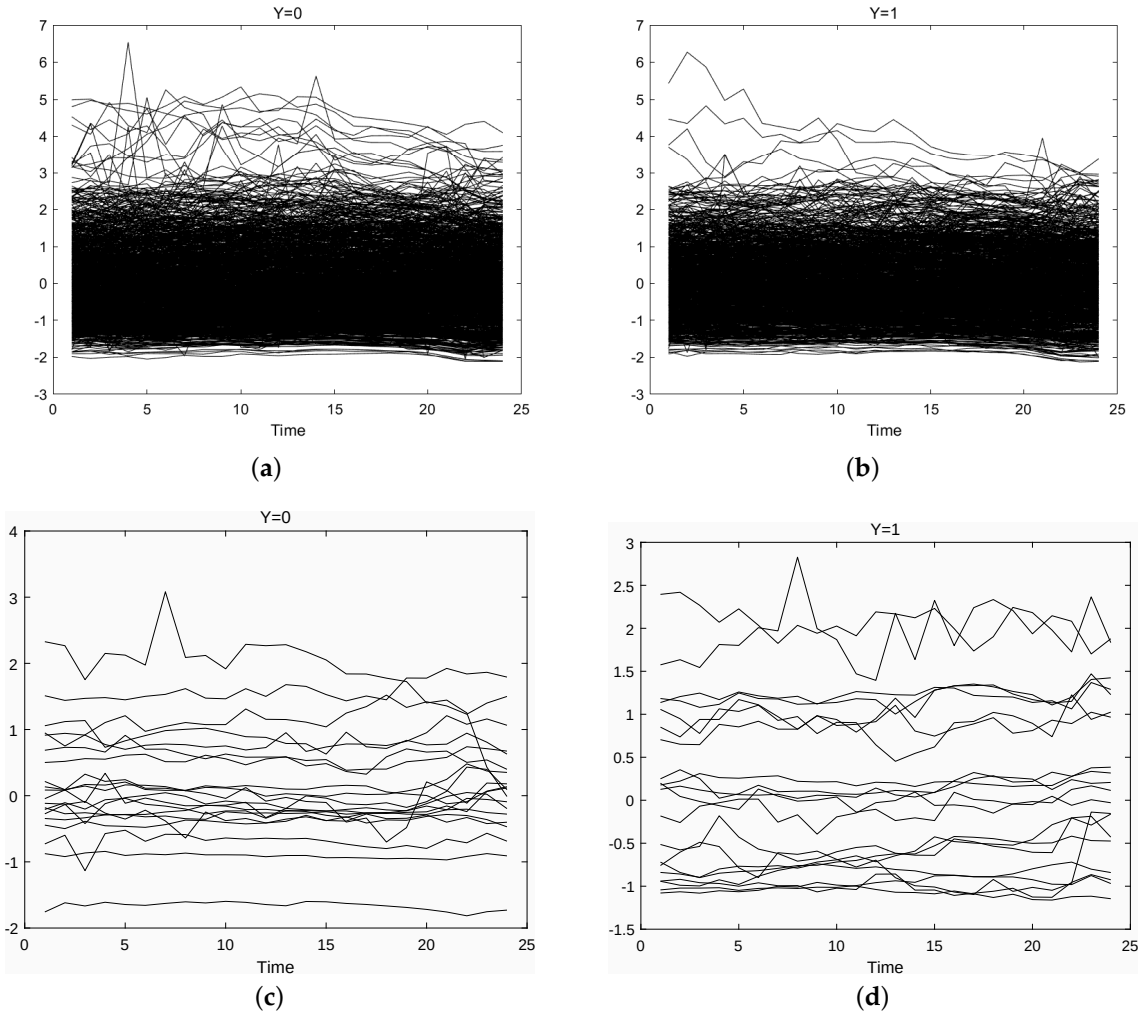

Figure 1. Predictor trajectories, corresponding to slightly smoothed monthly price curves. The low rising residential areas are in the upper left (**a**). The high rising residential areas are in upper right (**b**). Randomly selected profiles from the panels above are shown in the lower panels (**c**,**d**) for 20 districts.

We compare the out-of-sample prediction errors of the seven methods in Section 4. We repeat every method 20 times. The results are summarized in Tables 1 and 2. It can be observed from the tables that the error of CV1 or CV2 method is lower 10% on average than those of other methods, and overall, CV1 and CV2 behave similarly. As shown in the simulation above, this indicates the constraint that the sum of weights equals 1 makes sense in practical cases. AIC and BIC perform equally well, as both choose the largest model in most cases. We also find that FPCA is better than AIC or BIC. FPCA always selects the smallest model because the cumulative reliability of the first principal component is ~98%. Further, it is clear that the fitting error and prediction error of FPCA are similar. For the other methods, the fitting errors are always a little smaller than the prediction errors.

Table 1. Error of prediction.

Rounds	AIC	BIC	FPCA	S-AIC	S-BIC	CV1	CV2
1	0.301	0.301	0.275	0.301	0.301	0.221	0.221
2	0.292	0.292	0.247	0.292	0.292	0.178	0.176
3	0.290	0.290	0.242	0.290	0.290	0.187	0.187
4	0.280	0.280	0.233	0.280	0.280	0.176	0.174
5	0.276	0.276	0.233	0.276	0.276	0.147	0.149
6	0.316	0.316	0.233	0.316	0.316	0.188	0.188
7	0.269	0.269	0.244	0.269	0.269	0.164	0.164
8	0.294	0.294	0.225	0.294	0.294	0.174	0.174
9	0.316	0.316	0.235	0.316	0.316	0.187	0.187
10	0.282	0.282	0.242	0.282	0.282	0.174	0.173
11	0.292	0.292	0.240	0.292	0.292	0.162	0.162
12	0.285	0.285	0.261	0.285	0.285	0.188	0.188
13	0.282	0.282	0.219	0.282	0.282	0.150	0.149
14	0.264	0.264	0.280	0.264	0.264	0.188	0.188
15	0.282	0.282	0.247	0.282	0.282	0.187	0.187
16	0.295	0.295	0.269	0.295	0.295	0.185	0.185
17	0.328	0.328	0.252	0.328	0.328	0.204	0.202
18	0.301	0.301	0.245	0.301	0.301	0.187	0.187
19	0.278	0.278	0.209	0.278	0.278	0.150	0.150
20	0.311	0.311	0.249	0.311	0.311	0.183	0.183

Table 2. Error of fitting.

Rounds	AIC	BIC	FPCA	S-AIC	S-BIC	CV1	CV2
1	0.287	0.287	0.235	0.287	0.287	0.166	0.165
2	0.289	0.289	0.244	0.289	0.289	0.181	0.180
3	0.290	0.290	0.246	0.290	0.290	0.174	0.173
4	0.293	0.293	0.249	0.293	0.293	0.182	0.182
5	0.296	0.296	0.249	0.296	0.296	0.190	0.190
6	0.285	0.285	0.249	0.285	0.285	0.175	0.175
7	0.297	0.297	0.246	0.297	0.297	0.184	0.183
8	0.292	0.292	0.252	0.292	0.292	0.179	0.179
9	0.283	0.283	0.248	0.283	0.283	0.174	0.173
10	0.291	0.291	0.246	0.291	0.291	0.182	0.181
11	0.291	0.291	0.247	0.291	0.291	0.184	0.186
12	0.294	0.294	0.240	0.294	0.294	0.175	0.175
13	0.293	0.293	0.254	0.293	0.293	0.190	0.187
14	0.295	0.295	0.233	0.295	0.295	0.175	0.175
15	0.293	0.293	0.244	0.293	0.293	0.176	0.177
16	0.288	0.288	0.237	0.288	0.288	0.179	0.178
17	0.282	0.282	0.243	0.282	0.282	0.173	0.173
18	0.290	0.290	0.245	0.290	0.290	0.178	0.177
19	0.294	0.294	0.257	0.294	0.294	0.186	0.187
20	0.285	0.285	0.244	0.285	0.285	0.179	0.179

5. Concluding Remarks

In this paper, we proposed a model averaging approach under the framework of the generalized functional linear model. We showed that the weight chosen by the leave-one-out cross-validation method is asymptotically optimal in the sense of achieving the lowest possible squared error in a class

of model averaging estimators. It can be seen from the theoretical proof that our method is also valid for the non-nested candidate model set. Numerical analysis shows that for generalized functional linear model, cross-validation model averaging is a powerful tool for estimation and prediction. A further work is to develop model averaging inference procedures based on generalized functional linear model. In addition, how to combine other covariates into generalized functional linear model is also an interesting problem.

Author Contributions: H.Z. wrote the original draft. G.Z. reviewed and revised the whole paper. All authors have read and agreed to the published version of the manuscript.

Funding: Zou's work was partially supported by the Ministry of Science and Technology of China (Grant No. 2016YFB0502301) and the National Natural Science Foundation of China (Grant Nos. 11971323 and 11529101).

Acknowledgments: The authors thank the two referees for their constructive comments and suggestions that have substantially improved the original manuscript. The Beijing second-hand house price data is collected by the Guoxinda Group Corporation. This project was partially supported by the National Natural Science Foundation of China (Grant No.71571180).

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Lemmas and Proofs

The following Definition A1 and Lemma A1 can be found in Kahane (1968); Hoffmann (1974); Hoffmann and Pisier (1976); Zinn (1977); and Wu (1981).

Definition A1. *A linear map* $\nu : \mathbf{D} \to \mathbf{F}$ *is of type 2 if* $\sum_{i=1}^{n} \varepsilon_i \nu(s_i)$ *converges in* $\mathbf{F}$ *a.s. for all sequences* $\{s_i\} \subseteqq \mathbf{D}$ *such that* $\sum_{i=1}^{\infty} \|s_i\|_{\mathbf{D}}^2 < \infty$, *where* $\mathbf{D}$ *and* $\mathbf{F}$ *are Banach space,* $\{\varepsilon_i\}_{i=1}^{\infty}$ *are independent random variables such that* $\mathbf{P}(\varepsilon_i = 1) = \mathbf{P}(\varepsilon_i = -1) = 1/2$, *and a.s. means converges almost surely. A Banach space* $\mathbf{G}$ *is said to be type 2 if the identity map on* $\mathbf{G}$ *is type 2.*

Let (S, d) be a compact metric space and $\mathbf{C}(S)$ be the Banach space of real-valued continuous functions on S with the supremum norm

$$\|\nu\|_{\infty} = sup_{s \in S} |\nu(s)|,$$

for any $\nu \in \mathbf{C}(S)$. Denote a d-continuous metric ρ on S. Let $\mathbf{N}(S, \rho, \varepsilon)$ denote the minimal number of ρ-balls of radius less than or equal to ε which cover S, and set

$$\mathbf{H}(S, \rho, \varepsilon) = \log \mathbf{N}(S, \rho, \varepsilon).$$

We let

$$Lip(\rho) = \left\{ \nu \in \mathbf{C}(S) : \Lambda(\nu) = \sup_{s_1 \neq s_2 \in S} \frac{|\nu(s_1) - \nu(s_2)|}{\rho(s_1, s_2)} < \infty) \right\},$$

and for $\nu \in Lip(\rho)$, we define

$$\|\nu\|_{\rho} = \Lambda(\nu) + |\nu(s^*)|,$$

where s^* is some fixed point in S. In addition, assume that $\{\nu_j : j \geqslant 1\} \subseteq Lip(\rho)$ and $\{e_j : j \geqslant 1\}$ are independent real-valued random variables. Then, $\{\nu_j e_j\}$ are independent $Lip(\rho)$-valued random variables.

Lemma A1. *Let* (S, d) *denote a compact metric space. Suppose that* ρ *is a d-continuous metric on S with*

$$\int_0^{\delta} \mathbf{H}^{1/2}(S, \rho, u)\, du < \infty \quad \text{for some} \quad \delta > 0. \tag{A1}$$

Then we have $A < \infty$ *such that for all n,*

$$\mathbf{E}\left\|X_1 + X_2 + \cdots + X_n\right\|_\infty^2 \leqslant A\sum_{j=1}^{n}\mathbf{E}\left\|X_j\right\|_\rho^2, \tag{A2}$$

where $X_1, X_2, \ldots, X_n$ are independent $Lip(\rho)$-valued random variables with mean zeros.

Lemma A2. *For any $\beta^{(m)} \in \Theta_m$, define*

$$v_i\left(\beta^{(m)}\right) = \frac{g'\left(\varepsilon_{(i,p_m)}^T\beta^{(m)}\right)}{\sigma^2\left(g\left(\varepsilon_{(i,p_m)}^T\beta^{(m)}\right)\right)}.$$

Under Condition 3, we have

$$\sup_{\beta^{(m)}\in\Theta_m}\left\|\sum_{i\in[1,n]}\frac{v_i\left(\beta^{(m)}\right)}{p_m+1}\varepsilon_{(i,p_m)}e_i\right\| = O_p\left(\sqrt{np_m}\right). \tag{A3}$$

Proof of Lemma A2. First note that for any $l \in [0, p_m]$, we have

$$\begin{aligned}
&\Lambda\left(\frac{v_i}{p_m+1}\varepsilon_{i,l}\right)\\
&= \sup_{\beta_1^{(m)}\neq\beta_2^{(m)}\in\Theta_m}\frac{\left|v_i\left(\beta_1^{(m)}\right) - v_i\left(\beta_2^{(m)}\right)\right|\left|\varepsilon_{i,l}\right|}{(p_m+1)\times\rho\left(\beta_1^{(m)},\beta_2^{(m)}\right)}\\
&= \sup_{\beta_1^{(m)}\neq\beta_2^{(m)}\in\Theta_m}\frac{\left|v_i\left(\beta_1^{(m)}\right) - v_i\left(\beta_2^{(m)}\right)\right|}{\left|\varepsilon_{(i,p_m)}^T\beta_1^{(m)} - \varepsilon_{(i,p_m)}^T\beta_2^{(m)}\right|}\frac{\left|\varepsilon_{(i,p_m)}^T\beta_1^{(m)} - \varepsilon_{(i,p_m)}^T\beta_2^{(m)}\right|}{(p_m+1)\times\rho\left(\beta_1^{(m)},\beta_2^{(m)}\right)}\left|\varepsilon_{i,l}\right|\\
&= \sup_{\beta_1^{(m)}\neq\beta_2^{(m)}\in\Theta_m}\left\{\left|\frac{g''(\gamma_i)*\sigma^2(\gamma_i) - g'(\gamma_i)\sigma^{2'}(\gamma_i)}{\sigma^4(\gamma_i)}\right|\times\frac{\left|\varepsilon_{(i,p_m)}^T\beta_1^{(m)} - \varepsilon_{(i,p_m)}^T\beta_2^{(m)}\right|}{(p_m+1)\times\rho\left(\beta_1^{(m)},\beta_2^{(m)}\right)}\left|\varepsilon_{i,l}\right|\right\},
\end{aligned}$$

where the last step is by the mean-value theorem and γ_i is a point betweeen $\varepsilon_{(i,p_m)}^T\beta_1^{(m)}$ and $\varepsilon_{(i,p_m)}^T\beta_2^{(m)}$. From the assumptions that $g(\cdot)$ is a twice continuously differentiable function with bounded derivatives $|g'(\cdot)| \leqslant c < \infty$ and $|g''(\cdot)| \leqslant c_1 < \infty$, and $\sigma^2(\cdot)$ is strictly positive with bound $0 < d_1 \leqslant \sigma^2(\cdot) \leqslant d_2 < \infty$ and $\left|\sigma^{2'}(\cdot)\right| \leqslant d_3 < \infty$, we see that there is a constant $c' > 0$ such that $|v_i(\cdot)| \leqslant c' < \infty$, and

$$\begin{aligned}
&\Lambda\left(\frac{v_i}{p_m+1}\varepsilon_{i,l}\right)\\
&\leqslant \sup_{\beta_1^{(m)}\neq\beta_2^{(m)}\in\Theta_m}\left\{c'\times\frac{\left|\varepsilon_{(i,p_m)}^T\beta_1^{(m)} - \varepsilon_{(i,p_m)}^T\beta_2^{(m)}\right|}{(p_m+1)\times\rho\left(\beta_1^{(m)},\beta_2^{(m)}\right)}\left|\varepsilon_{i,l}\right|\right\}\\
&\leqslant \sup_{\beta_1^{(m)}\neq\beta_2^{(m)}\in\Theta_m}\left\{c'\times\frac{\left\|\varepsilon_{(i,p_m)}\right\|}{p_m+1}\left|\varepsilon_{i,l}\right|\right\}\\
&= c'\frac{\left\|\varepsilon_{(i,p_m)}\right\|}{p_m+1}\left|\varepsilon_{i,l}\right|,
\end{aligned}$$

where the second inequality is by Cauchy–Schwarz inequality. Therefore, we obtain

$$\begin{aligned}\left\|\frac{v_i}{p_m+1}\varepsilon_{i,l}\right\|_\rho &= \Lambda\left(\frac{v_i}{p_m+1}\varepsilon_{i,l}\right)+\left|\frac{v_i\left(\beta^{(m)*}\right)}{p_m+1}\varepsilon_{i,l}\right| \\ &\leqslant c'\frac{\left\|\varepsilon_{(i,p_m)}\right\|}{p_m+1}\left|\varepsilon_{i,l}\right|+c'\frac{1}{p_m+1}\left|\varepsilon_{i,l}\right| \\ &= c'\frac{\left\|\varepsilon_{(i,p_m)}\right\|+1}{p_m+1}\left|\varepsilon_{i,l}\right| \\ &< \infty.\end{aligned}$$

As Θ_m is a compact subset of $\mathbb{R}^{p_m+1}$, and $\rho(\beta_1^{(m)},\beta_2^{(m)})$ is the Euclidean metric in $\mathbb{R}^{p_m+1}$, (A1) is satisfied. Thus, by Lemma A1, there is a constant $A>0$ uniformly for all l such that for any $C>0$, we have

$$\begin{aligned}&\mathbf{P}\left\{\sup_{\beta^{(m)}\in\Theta_m}\left|\sum_{i\in[1,n]}\frac{v_i\left(\beta^{(m)}\right)}{p_m+1}\varepsilon_{i,l}e_i\right|^2>Cn\right\} \\ &=\mathbf{P}\left\{\left\|\sum_{i\in[1,n]}\frac{v_i}{p_m+1}\varepsilon_{i,l}e_i\right\|_\infty^2>Cn\right\} \\ &\leqslant\frac{1}{Cn}\mathbf{E}\left\|\sum_{i\in[1,n]}\frac{v_i}{p_m+1}\varepsilon_{i,l}e_i\right\|_\infty^2 \\ &\leqslant\frac{1}{Cn}A\left\{\sum_{i=1}^n\left\|\frac{v_i}{p_m+1}\varepsilon_{i,l}\right\|_\rho^2\right\}\sup_i\mathbf{E}e_i^2.\end{aligned}$$

Notice

$$\sup_{\beta^{(m)}\in\Theta_m}\left\|\sum_{i\in[1,n]}\frac{v_i\left(\beta^{(m)}\right)}{p_m+1}\varepsilon_{(i,p_m)}e_i\right\|^2\leqslant\sum_{l=0}^{p_m}\left\{\sup_{\beta^{(m)}\in\Theta_m}\left|\sum_{i\in[1,n]}\frac{v_i\left(\beta^{(m)}\right)}{p_m+1}\varepsilon_{i,l}e_i\right|^2\right\}.$$

Therefore, for any $\varepsilon>0$, letting $C=Ac'^2\left(\sqrt{C_2}+1\right)^2C_2C_1/\varepsilon$, we obtain

$$\begin{aligned}&\mathbf{P}\left\{\sup_{\beta^{(m)}\in\Theta_m}\left\|\sum_{i\in[1,n]}\frac{v_i\left(\beta^{(m)}\right)}{p_m+1}\varepsilon_{(i,p_m)}e_i\right\|^2>Cn\left(p_m+1\right)\right\} \\ &\leqslant\mathbf{P}\left\{\sum_{l=0}^{p_m}\left[\sup_{\beta^{(m)}\in\Theta_m}\left|\sum_{i\in[1,n]}\frac{v_i\left(\beta^{(m)}\right)}{p_m+1}\varepsilon_{i,l}e_i\right|^2\right]>Cn\left(p_m+1\right)\right\} \\ &\leqslant\sum_{l=0}^{p_m}\mathbf{P}\left\{\sup_{\beta^{(m)}\in\Theta_m}\left|\sum_{i\in[1,n]}\frac{v_i\left(\beta^{(m)}\right)}{p_m+1}\varepsilon_{i,l}e_i\right|^2>Cn\right\} \\ &\leqslant\sum_{l=0}^{p_m}\frac{1}{Cn}A\left\{\sum_{i=1}^n\left\|\frac{v_i}{p_m+1}\varepsilon_{i,l}\right\|_\rho^2\right\}\sup_i\mathbf{E}e_i^2 \\ &\leqslant\frac{Ac'^2}{Cn\left(p_m+1\right)^2}\sum_{i=1}^n\left\{\left[\left\|\varepsilon_{(i,p_m)}\right\|+1\right]^2\left\|\varepsilon_{(i,p_m)}\right\|^2\right\}\sup_i\mathbf{E}e_i^2 \\ &\leqslant\frac{Ac'^2\left(\sqrt{C_2}+1\right)^2C_2C_1}{C}=\varepsilon,\end{aligned}\tag{A4}$$

which implies (A3). □

Lemma A3. *Under Conditions 1–3 and 6, we have*

$$\left\|\hat{\beta}^{(m)}-\beta_{\star}^{(m)}\right\|^{2}=O_{p}\left(\frac{p_{m}^{3}}{n}\right), \tag{A5}$$

where $\hat{\beta}^{(m)}$ *belonging to* Θ_m *is the root of* (4).

Proof of Lemma A3. By the definition of $\beta_{\star}^{(m)}$ and Condition 6, then we have

$$\begin{aligned}\left\|U_{n,m}^{(m)}\left(\beta^{(m)}\right)\right\|^{2}&=\left\|U_{n,m}^{(m)}\left(\beta^{(m)}\right)-U_{n,m}^{(m)}\left(\beta_{\star}^{(m)}\right)\right\|^{2}\\&=\left\|\left.\frac{\partial U_{n,m}^{(m)}}{\partial\beta^{(m)}}\right|_{\beta^{(m)}=\overline{\beta}^{(m)}}\left(\beta^{(m)}-\beta_{\star}^{(m)}\right)\right\|^{2}\\&\geqslant C_{0}^{2}\left\|\beta^{(m)}-\beta_{\star}^{(m)}\right\|^{2},\end{aligned} \tag{A6}$$

where $\overline{\beta}^{(m)}$ is a point between $\beta^{(m)}$ and $\beta_{\star}^{(m)}$. Recalling that

$$\begin{aligned}U_{n,m}\left(\hat{\beta}^{(m)}\right)=&\frac{1}{n}\sum_{i=1}^{n}\left[y_{i}-g\left(\hat{\eta}_{i,p_m}\right)\right]\frac{g'\left(\hat{\eta}_{i,p_m}\right)}{\sigma^{2}\left(g\left(\hat{\eta}_{i,p_m}\right)\right)}\varepsilon_{(i,p_m)}\\=&\frac{1}{n}\sum_{i=1}^{n}\left[\mu_{i}+e_{i}-g\left(\hat{\eta}_{i,p_m}\right)\right]\frac{g'\left(\hat{\eta}_{i,p_m}\right)}{\sigma^{2}\left(g\left(\hat{\eta}_{i,p_m}\right)\right)}\varepsilon_{(i,p_m)}\\=&-U_{n,m}^{(m)}\left(\hat{\beta}^{(m)}\right)+\frac{1}{n}\sum_{i=1}^{n}e_{i}\frac{g'\left(\hat{\eta}_{i,p_m}\right)}{\sigma^{2}\left(g\left(\hat{\eta}_{i,p_m}\right)\right)}\varepsilon_{(i,p_m)}\\=&0,\end{aligned}$$

we obtain

$$\begin{aligned}U_{n,m}^{(m)}\left(\hat{\beta}^{(m)}\right)=&\frac{1}{n}\sum_{i=1}^{n}e_{i}\frac{g'\left(\hat{\eta}_{i,p_m}\right)}{\sigma^{2}\left(g\left(\hat{\eta}_{i,p_m}\right)\right)}\varepsilon_{(i,p_m)}\\=&\frac{1}{n}\varepsilon_{n}^{(m)T}V_{n}\left(\hat{\beta}^{(m)}\right)e,\end{aligned}$$

where $V_{n}\left(\hat{\beta}^{(m)}\right)=diag\left(\frac{g'\left(\hat{\eta}_{i,p_m}\right)}{\sigma^{2}\left(g\left(\hat{\eta}_{i,p_m}\right)\right)}\right)_{1\leqslant i\leqslant n}$. From (A6), we get

$$\left\{\left\|U_{n,m}^{(m)}\left(\beta^{(m)}\right)\right\|\leqslant C_{0}\delta\right\}\subseteq\left\{\left\|\beta^{(m)}-\beta_{\star}^{(m)}\right\|\leqslant\delta\right\},\quad\forall\delta>0. \tag{A7}$$

By Condition 1, for any $\kappa>0$, there is an N_1 such that for all $n>N_1$, we have

$$\mathbf{P}\left\{0\in U_{n,m}\left(\Theta_{m}\right)\right\}>1-\kappa.$$

From (A7), it can be seen that

$$\begin{aligned}&\left\{0\in U_{n,m}\left(\Theta_{m}\right)\right\}\\&=\left\{\text{There is a }\hat{\beta}_{*}^{(m)}\in\Theta_{m}\text{ such that }\frac{1}{n}\varepsilon_{n}^{(m)T}V_{n}\left(\hat{\beta}_{*}^{(m)}\right)e=U_{n,m}^{(m)}\left(\hat{\beta}_{*}^{(m)}\right)\right\}\\&\subset\left\{\sup_{\beta^{(m)}\in\Theta_{m}}\left\|\varepsilon_{n}^{(m)T}V_{n}\left(\beta^{(m)}\right)e\right\|>C_{0}n\delta\right\}\cup\left\{\left\|\hat{\beta}^{(m)}-\beta_{\star}^{(m)}\right\|\leqslant\delta\right\}.\end{aligned}$$

Then for any $C > 0$ and $n > N_1$, letting $\delta = C\sqrt{\frac{(p_m+1)^3}{n}}$, we have

$$\begin{aligned}
&\mathbf{P}\left\{\left\|\hat{\beta}^{(m)} - \beta_\star^{(m)}\right\| \leqslant \frac{C\left(p_m+1\right)^{3/2}}{\sqrt{n}}\right\} \\
&\geqslant 1-\kappa-\mathbf{P}\left\{\sup_{\beta^{(m)}\in\Theta_m}\left\|\varepsilon_n^{(m)T}V_n\left(\beta^{(m)}\right)e\right\| > C_0C\sqrt{n}\left(p_m+1\right)^{3/2}\right\} \\
&\geqslant 1-\kappa-\frac{Ac^2\left(\sqrt{C_2}+1\right)^2C_2C_1}{C_0^2C^2} \\
&\geqslant 1-\kappa-\frac{C'}{C_0^2C^2},
\end{aligned}$$

where $C' = Ac'^2\left(\sqrt{C_2}+1\right)^2C_2C_1$ and the second inequality is derived by (A4). As a result, for any $\kappa > 0$, we can select $C = \frac{\sqrt{C'}}{C_0\sqrt{\kappa}}$ such that

$$\mathbf{P}\left\{\left\|\hat{\beta}^{(m)} - \beta_\star^{(m)}\right\| > \frac{C\left(p_m+1\right)^{3/2}}{\sqrt{n}}\right\} < 2\kappa,$$

for sufficiently large n, thus

$$\left\|\hat{\beta}^{(m)} - \beta_\star^{(m)}\right\|^2 = O_p\left(\frac{p_m^3}{n}\right).$$

□

Lemma A4. *Under Conditions 1–4 and 6,*

$$\sup_{w\in H_n}\left|\frac{L_n(w)}{R_n(w)} - 1\right| = o_p(1). \tag{A8}$$

Proof of Lemma A4. Write $\Delta_i(w) = g(\eta_i) - g\left(\sum_{m=1}^M w_m\varepsilon_{(i,p_m)}^T\beta_\star^{(m)}\right)$. From the definition of $L_n(w)$, we have

$$\begin{aligned}
L_n(w) &= \sum_{i=1}^n\left[g(\eta_i) - g\left(\sum_{m=1}^M w_m\hat{\eta}_{i,p_m}\right)\right]^2 \\
&= \sum_{i=1}^n\left[g(\eta_i) - g\left(\sum_{m=1}^M w_m\varepsilon_{(i,p_m)}^T\beta_\star^{(m)}\right) + g\left(\sum_{m=1}^M w_m\varepsilon_{(i,p_m)}^T\beta_\star^{(m)}\right) - g\left(\sum_{m=1}^M w_m\hat{\eta}_{i,p_m}\right)\right]^2 \\
&= \sum_{i=1}^n\left\{\left[g(\eta_i) - g\left(\sum_{m=1}^M w_m\varepsilon_{(i,p_m)}^T\beta_\star^{(m)}\right)\right]^2 + \left[g\left(\sum_{m=1}^M w_m\varepsilon_{(i,p_m)}^T\beta_\star^{(m)}\right) - g\left(\sum_{m=1}^M w_m\hat{\eta}_{i,p_m}\right)\right]^2\right. \\
&\quad \left.+2\left[g(\eta_i) - g\left(\sum_{m=1}^M w_m\varepsilon_{(i,p_m)}^T\beta_\star^{(m)}\right)\right]\left[g\left(\sum_{m=1}^M w_m\varepsilon_{(i,p_m)}^T\beta_\star^{(m)}\right) - g\left(\sum_{m=1}^M w_m\hat{\eta}_{i,p_m}\right)\right]\right\} \\
&= \sum_{i=1}^n\Delta_i^2(w) + L_n^{(2)}(w) + \sum_{i=1}^n 2\Delta_i(w)\left[g\left(\sum_{m=1}^M w_m\varepsilon_{(i,p_m)}^T\beta_\star^{(m)}\right) - g\left(\sum_{m=1}^M w_m\hat{\eta}_{i,p_m}\right)\right] \\
&\triangleq R_n(w) + L_n^{(2)}(w) + L_n^{(3)}(w).
\end{aligned}$$

We note that

$$\frac{|L_n(w) - R_n(w)|}{R_n(w)} = \frac{\left|L_n^{(2)}(w) + L_n^{(3)}(w)\right|}{R_n(w)} \leq \sup_{w \in H_n} \left(\frac{|L_n^{(2)}(w)|}{R_n(w)} + 2\sqrt{\frac{|L_n^{(2)}(w)|}{R_n(w)}} \right).$$

Then, (A8) is valid if

$$\sup_{w \in H_n} \frac{\left|L_n^{(2)}(w)\right|}{R_n(w)} \xrightarrow{p} 0. \tag{A9}$$

Let η_{i*}^m be the point between $\varepsilon_{(i,p_m)}^T \beta_\star^{(m)}$ and $\hat{\eta}_{i,p_m}$, for fixed M,

$$\begin{aligned}
L_n^{(2)}(w) &= \sum_{i=1}^n \left[g\left(\sum_{m=1}^M w_m \varepsilon_{(i,p_m)}^T \beta_\star^{(m)} \right) - g\left(\sum_{m=1}^M w_m \hat{\eta}_{i,p_m} \right) \right]^2 \\
&= \sum_{i=1}^n \left[g'\left(\sum_{m=1}^M w_m \eta_{i*}^m \right) \left(\sum_{m=1}^M w_m \varepsilon_{(i,p_m)}^T \beta_\star^{(m)} - \sum_{m=1}^M w_m \hat{\eta}_{i,p_m} \right) \right]^2 \\
&= \sum_{i=1}^n \left\{ \left[g'\left(\sum_{m=1}^M w_m \eta_{i*}^m \right) \right]^2 \left[\sum_{m=1}^M w_m \left(\varepsilon_{(i,p_m)}^T \beta_\star^{(m)} - \hat{\eta}_{i,p_m} \right) \right]^2 \right\} \\
&\leqslant \sum_{i=1}^n \left[g'(w; \eta_{i*})^2 \sum_{m=1}^M \left(\varepsilon_{(i,p_m)}^T \beta_\star^{(m)} - \hat{\eta}_{i,p_m} \right)^2 \right] \\
&\leqslant c^2 \sum_{i=1}^n \sum_{m=1}^M \left[\varepsilon_{(i,p_m)}^T \left(\beta_\star^{(m)} - \hat{\beta}^{(m)} \right) \right]^2.
\end{aligned}$$

Then, by Lemma A3 and Condition 3, we have

$$\sup_{w \in H_n} L_n^{(2)}(w) \leqslant c^2 \sum_{i=1}^n \sum_{m=1}^M \left[\varepsilon_{(i,p_m)}^T \left(\beta_\star^{(m)} - \hat{\beta}^{(m)} \right) \right]^2 = O_p\Big(\sum_{m=1}^M p_m^4 \Big),$$

which, together with Condition 4, leads to (A9). □

Appendix B. Proof of Theorem 1

Let $\widetilde{\mu}(w) = (\tilde{\mu}_1(w), \tilde{\mu}_2(w), \ldots, \tilde{\mu}_n(w))^T$, and

$$\widetilde{L}_n(w) = \|\mu - \widetilde{\mu}(w)\|^2.$$

As in Li (1987) and Ando and Li (2014), we know that

$$\begin{aligned}
&CV(w) \\
&= \|e\|^2 + \widetilde{L}_n(w) + 2\langle e, \mu - \widetilde{\mu}(w) \rangle \\
&= \|e\|^2 + L_n(w) \left(\frac{\widetilde{L}_n(w)}{L_n(w)} + \frac{2\langle e, \mu - \widetilde{\mu}(w) \rangle}{L_n(w)} \right).
\end{aligned} \tag{A10}$$

As $\hat{w}$ minimizes $CV(w)$ over $w \in H_n$, it also minimizes $CV(w) - \|e\|^2$ over $w \in H_n$. Therefore, the claim

$$\frac{L_n(\widehat{w})}{\inf_{w \in H_n} L_n(w)} \xrightarrow{p} 1$$

is valid if

$$\sup_{w\in H_n}\left|\frac{\widetilde{L}_n(w)}{L_n(w)}-1\right|\xrightarrow{p}0 \tag{A11}$$

and

$$\sup_{w\in H_n}\left|\frac{\langle e,\mu-\widetilde{\mu}(w)\rangle}{L_n(w)}\right|\xrightarrow{p}0 \tag{A12}$$

hold. In fact, if we denote $w^*=arg\min_{w\in H_n}L_n(w)$, then

$$\frac{L_n(\widehat{w})}{\inf_{w\in\ H_n}L_n(w)}=\frac{L_n(\widehat{w})}{L_n(w^*)}\geqslant 1,$$

so we only need to prove

$$\frac{L_n(\widehat{w})}{L_n(w^*)}\leqslant 1+\delta_n,$$

where $\delta_n\geqslant 0$ for $n=1,2,\dots$, and $\delta_n\xrightarrow{p}0$. According to the definition of $\hat{w}$, we have $CV_n(\hat{w})\leqslant CV_n(w^*)$. Then, by (A10), we obtain

$$\|e\|^2+L_n(\hat{w})\left(\frac{\widetilde{L}_n(\hat{w})}{L_n(\hat{w})}+\frac{2\langle e,\mu-\widetilde{\mu}(\hat{w})\rangle}{L_n(\hat{w})}\right)\leqslant\|e\|^2+L_n(w^*)\left(\frac{\widetilde{L}_n(w^*)}{L_n(w^*)}+\frac{2\langle e,\mu-\widetilde{\mu}(w^*)\rangle}{L_n(w^*)}\right),$$

which is equivalent to

$$\frac{L_n(\hat{w})}{L_n(w^*)}\left(\frac{\widetilde{L}_n(\hat{w})}{L_n(\hat{w})}+\frac{2\langle e,\mu-\widetilde{\mu}(\hat{w})\rangle}{L_n(\hat{w})}\right)\leqslant\frac{\widetilde{L}_n(w^*)}{L_n(w^*)}+\frac{2\langle e,\mu-\widetilde{\mu}(w^*)\rangle}{L_n(w^*)}.$$

From (A11) and (A12), we have

$$\begin{aligned}&\frac{L_n(\hat{w})}{L_n(w^*)}\left(\frac{\widetilde{L}_n(\hat{w})}{L_n(\hat{w})}+\frac{2\langle e,\mu-\widetilde{\mu}(\hat{w})\rangle}{L_n(\hat{w})}\right)\\&\leqslant\frac{\widetilde{L}_n(w^*)}{L_n(w^*)}+\frac{2\langle e,\mu-\widetilde{\mu}(w^*)\rangle}{L_n(w^*)}\\&\leqslant\sup_{w\in H_n}\left|\frac{\widetilde{L}_n(w)}{L_n(w)}+\frac{2\langle e,\mu-\widetilde{\mu}(w)\rangle}{L_n(w)}\right|\\&\leqslant\sup_{w\in H_n}\left|\frac{\widetilde{L}_n(w)}{L_n(w)}-1\right|+1+\sup_{w\in H_n}\left|\frac{2\langle e,\mu-\widetilde{\mu}(w)\rangle}{L_n(w)}\right|,\end{aligned}$$

and

$$\begin{aligned}&\frac{L_n(\hat{w})}{L_n(w^*)}\left(\frac{\widetilde{L}_n(\hat{w})}{L_n(\hat{w})}+\frac{2\langle e,\mu-\widetilde{\mu}(\hat{w})\rangle}{L_n(\hat{w})}\right)\\&\geqslant\frac{L_n(\hat{w})}{L_n(w^*)}\left(1-\sup_{w\in H_n}\left|\frac{\widetilde{L}_n(w)}{L_n(w)}-1\right|-\sup_{w\in H_n}\left|\frac{2\langle e,\mu-\widetilde{\mu}(w)\rangle}{L_n(w)}\right|\right).\end{aligned}$$

Therefore,

$$\frac{1}{L_n(\hat{w})/L_n(w^*)}\geqslant\frac{1-\delta_n}{1+\delta_n}\to 1,$$

with $L_n(\hat{w})/L_n(w^*) \geqslant 1$, and

$$\delta_n = \sup_{w\in H_n}\left|\frac{\widetilde{L}_n(w)}{L_n(w)} - 1\right| + \sup_{w\in H_n}\left|\frac{2\langle e, \mu - \widetilde{\mu}(w)\rangle}{L_n(w)}\right|.$$

Thus, we obtain

$$\frac{L_n(\hat{w})}{L_n(w^*)} \xrightarrow{p} 1.$$

In the following, we prove (A11) and (A12).

Appendix B.1. Proof of (A11)

Notice that

$$\begin{aligned}
\left|\widetilde{L}_n(w) - L_n(w)\right| &= \left|\sum_{i=1}^n \tilde{\mu}_i(w)^2 - \sum_{i=1}^n \hat{\mu}_i(w)^2 + 2\sum_{i=1}^n \mu_i\left[\hat{\mu}_i(w) - \tilde{\mu}_i(w)\right]\right| \\
&= \left|\sum_{i=1}^n \left[\tilde{\mu}_i(w) - \hat{\mu}_i(w)\right]^2 - 2\sum_{i=1}^n \hat{\mu}_i(w)^2 + 2\sum_{i=1}^n \tilde{\mu}_i(w)\hat{\mu}_i(w) + 2\sum_{i=1}^n \mu_i\left[\hat{\mu}_i(w) - \tilde{\mu}_i(w)\right]\right| \\
&= \left|\sum_{i=1}^n \left[\tilde{\mu}_i(w) - \hat{\mu}_i(w)\right]^2 + 2\sum_{i=1}^n \left[\mu_i - \hat{\mu}_i(w)\right]\left[\hat{\mu}_i(w) - \tilde{\mu}_i(w)\right]\right| \\
&= \left|\left\|\widehat{\mu}(w) - \widetilde{\mu}(w)\right\|^2 + 2\left\langle \mu - \widehat{\mu}(w), \widehat{\mu}(w) - \widetilde{\mu}(w)\right\rangle\right| \\
&\leqslant \left\|\widehat{\mu}(w) - \widetilde{\mu}(w)\right\|^2 + 2\sqrt{L_n(w)}\left\|\widehat{\mu}(w) - \widetilde{\mu}(w)\right\|.
\end{aligned}$$

So,

$$\begin{aligned}
\left|\frac{\widetilde{L}_n(w)}{L_n(w)} - 1\right| &= \frac{\left|\widetilde{L}_n(w) - L_n(w)\right|}{L_n(w)} \\
&\leqslant \frac{\left\|\widehat{\mu}(w) - \widetilde{\mu}(w)\right\|^2 + 2\sqrt{L_n(w)}\left\|\widehat{\mu}(w) - \widetilde{\mu}(w)\right\|}{L_n(w)} \\
&= \frac{\left\|\widehat{\mu}(w) - \widetilde{\mu}(w)\right\|^2}{L_n(w)} + \frac{2\left\|\widehat{\mu}(w) - \widetilde{\mu}(w)\right\|}{\sqrt{L_n(w)}}.
\end{aligned}$$

Therefore, to prove (A11), it suffices to verify

$$\sup_{w\in H_n}\frac{\left\|\widehat{\mu}(w) - \widetilde{\mu}(w)\right\|^2}{L_n(w)} \xrightarrow{p} 0.$$

By Lemma A4, we need only to show

$$\sup_{w\in H_n}\frac{\left\|\widehat{\mu}(w) - \widetilde{\mu}(w)\right\|^2}{R_n(w)} \xrightarrow{p} 0. \tag{A13}$$

Let η^*_{i,p_m} be the point between $\tilde{\eta}_{i,p_m}$ and $\hat{\eta}_{i,p_m}$. Then, for any $\delta > 0$, we have

$$\begin{aligned}
&\mathbf{P}\left\{\sup_{w\in H_n}\frac{\left\|\widehat{\mu}(w) - \widetilde{\mu}(w)\right\|^2}{R_n(w)} > \delta\right\} \\
&\leqslant \mathbf{P}\left\{\sup_{w\in H_n}\left\|\widehat{\mu}(w) - \widetilde{\mu}(w)\right\|^2 > \delta\xi_n\right\}
\end{aligned}$$

$$
\begin{aligned}
&= \mathbf{P}\left\{\sup_{w\in H_n}\sum_{i=1}^{n}\left[\tilde{\mu}_i(w)-\hat{\mu}_i(w)\right]^2 > \delta\xi_n\right\}\\
&= \mathbf{P}\left\{\sup_{w\in H_n}\sum_{i=1}^{n}\left[g\left(\sum_{m=1}^{M} w_m\tilde{\eta}_{i,p_m}\right)-g\left(\sum_{m=1}^{M} w_m\hat{\eta}_{i,p_m}\right)\right]^2 > \delta\xi_n\right\}\\
&= \mathbf{P}\left\{\sup_{w\in H_n}\sum_{i=1}^{n}\left[g'\left(\eta^*_{i,p_m}\right)\left(\sum_{m=1}^{M} w_m\tilde{\eta}_{i,p_m}-\sum_{m=1}^{M} w_m\hat{\eta}_{i,p_m}\right)\right]^2 > \delta\xi_n\right\}\\
&\leqslant \mathbf{P}\left\{\max_{1\leqslant i\leqslant n, w\in H_n}\left|g'\left(\eta^*_{i,p_m}\right)\right|^2\sup_{w\in H_n}\sum_{i=1}^{n}\left[\sum_{m=1}^{M} w_m\left(\tilde{\eta}_{i,p_m}-\hat{\eta}_{i,p_m}\right)\right]^2 > \delta\xi_n\right\}\\
&\leqslant \mathbf{P}\left\{\max_{1\leqslant i\leqslant n, w\in H_n}\left|g'\left(\eta^*_{i,p_m}\right)\right|^2\sup_{w\in H_n}\sum_{i=1}^{n}\sum_{m=1}^{M}\left(\tilde{\eta}_{i,p_m}-\hat{\eta}_{i,p_m}\right)^2 > \delta\xi_n\right\}\\
&= \mathbf{P}\left\{\max_{1\leqslant i\leqslant n, w\in H_n}\left|g'\left(\eta^*_{i,p_m}\right)\right|^2\sum_{i=1}^{n}\sum_{m=1}^{M}\left(\tilde{\eta}_{i,p_m}-\hat{\eta}_{i,p_m}\right)^2 > \delta\xi_n\right\},
\end{aligned}
$$

which, together with the assumption that $g(\cdot)$ is a twice continuously differentiable function with bounded derivatives implying $\max_{1\leqslant i\leqslant n, w\in H_n}|g'(\eta^*_{i,p_m})|^2 \leqslant c^2 < \infty$, leads to

$$
\mathbf{P}\left\{\sup_{w\in H_n}\frac{\|\widehat{\mu}(w)-\widetilde{\mu}(w)\|^2}{R_n(w)} > \delta\right\} \leqslant \mathbf{P}\left\{c^2\sum_{i=1}^{n}\sum_{m=1}^{M}\left(\tilde{\eta}_{i,p_m}-\hat{\eta}_{i,p_m}\right)^2/\xi_n > \delta\right\}.
$$

Thus, to prove (A13), it suffices to show

$$
\sum_{i=1}^{n}\sum_{m=1}^{M}\left(\tilde{\eta}_{i,p_m}-\hat{\eta}_{i,p_m}\right)^2/\xi_n = o_p(1). \tag{A14}
$$

By Condition 5, for fixed M, we obtain

$$
\sum_{i=1}^{n}\sum_{m=1}^{M}\left(\tilde{\eta}_{i,p_m}-\hat{\eta}_{i,p_m}\right)^2 = O_p\left(\sum_{m=1}^{M} p_m^4\right),
$$

which, together with Condition 4, leads to (A14), and thus (A13) holds.

Appendix B.2. Proof of (A12)

As

$$
|\langle e, \mu-\widetilde{\mu}(w)\rangle| = \left|\sum_{i=1}^{n} e_i\left[g(\eta_i)-g\left(\sum_{m=1}^{M} w_m\tilde{\eta}_{i,p_m}\right)\right]\right|,
$$

it is sufficient to show

$$
\sup_{w\in H_n}\left|\sum_{i=1}^{n} e_i\left[g(\eta_i)-g\left(\sum_{m=1}^{M} w_m\tilde{\eta}_{i,p_m}\right)\right]\right|/R_n(w) \xrightarrow{p} 0.
$$

It is readily seen that

$$
\begin{aligned}
&\sup_{w\in H_n}\frac{\left|\sum_{i=1}^{n}e_i\left[g(\eta_i)-g\left(\sum_{m=1}^{M}w_m\tilde{\eta}_{i,p_m}\right)\right]\right|}{R_n(w)}\\
&\leqslant \sup_{w\in H_n}\frac{\left|\sum_{i=1}^{n}e_i\left[g(\eta_i)-g\left(\sum_{m=1}^{M}w_m\varepsilon_{(i,p_m)}^{T}\beta_{\star}^{(m)}\right)\right]\right|}{R_n(w)}\\
&\quad+\sup_{w\in H_n}\frac{\left|\sum_{i=1}^{n}e_i\left[g\left(\sum_{m=1}^{M}w_m\varepsilon_{(i,p_m)}^{T}\beta_{\star}^{(m)}\right)-g\left(\sum_{m=1}^{M}w_m\hat{\eta}_{i,p_m}\right)\right]\right|}{R_n(w)}\\
&\quad+\sup_{w\in H_n}\frac{\left|\sum_{i=1}^{n}e_i\left[g\left(\sum_{m=1}^{M}w_m\varepsilon_{(i,p_m)}^{T}\hat{\beta}^{(m)}\right)-g\left(\sum_{m=1}^{M}w_m\tilde{\eta}_{i,p_m}\right)\right]\right|}{R_n(w)}\\
&\triangleq \sup_{w\in H_n}A_n^{(1)}(w)+\sup_{w\in H_n}A_n^{(2)}(w)+\sup_{w\in H_n}A_n^{(3)}(w).
\end{aligned}
$$

Thus, we need only to prove

$$\sup_{w\in H_n}A_n^{(1)}(w)\xrightarrow{p}0, \tag{A15}$$

$$\sup_{w\in H_n}A_n^{(2)}(w)\xrightarrow{p}0, \tag{A16}$$

and

$$\sup_{w\in H_n}A_n^{(3)}(w)\xrightarrow{p}0. \tag{A17}$$

The proof of (A15) is similar to that of Wu (1981). We denote a metric

$$\rho(w,w')=\left\|w-w'\right\|,$$

which is on H_n. Let (H_n,ρ) be a compact metric space. Then $\mathbf{C}(H_n)$ is the Banach space of real-valued continuous functions on H_n with the supremum norm

$$\|\Delta\|_{\infty}=sup_{w\in H_n}\left|\Delta(w)\right|.$$

Let $\mathbf{N}(H_n,\rho,\varepsilon)$ denote the minimal number of ρ-balls of radius less than or equal to ε which cover H_n, and set

$$\mathbf{H}(H_n,\rho,\varepsilon)=\log\mathbf{N}(H_n,\rho,\varepsilon).$$

We let

$$Lip(\rho)=\left\{\Delta\in\mathbf{C}(H_n):\Lambda(\Delta)=\sup_{w\neq w'\in H_n}\frac{|\Delta(w)-\Delta(w')|}{\rho(w,w')}<\infty\right\},$$

and for $\Delta\in Lip(\rho)$, we define

$$\|\Delta\|_{\rho}=\Lambda(\Delta)+\left|\Delta(w^{*})\right|,$$

where w^{*} is some fixed point in H_n.

Recalling that $\Delta_i(w)=g(\eta_i)-g\left(\sum_{m=1}^{M}w_m\varepsilon_{(i,p_m)}^{T}\beta_{\star}^{(m)}\right)$, we have

$$
\begin{aligned}
&\Lambda\left(\frac{\Delta_i}{p+1}\right)\\
&=\sup_{w\neq w'\in H_n}\frac{|\Delta_i(w)-\Delta_i(w')|}{(p+1)\,\rho(w,w')}\\
&=\sup_{w\neq w'\in H_n}\left|g'(\gamma_{0,i})\right|\times\frac{\left|\sum_{m=1}^{M}w_m\varepsilon_{(i,p_m)}^{T}\beta_{\star}^{(m)}-\sum_{m=1}^{M}w_m'\varepsilon_{(i,p_m)}^{T}\beta_{\star}^{(m)}\right|}{(p+1)\,\rho(w,w')}
\end{aligned}
$$

$$\leqslant c \times \sup_{w \neq w' \in H_n} \frac{\left| \sum_{m=1}^{M} w_m \varepsilon_{(i,p_m)}^T \beta_\star^{(m)} - \sum_{m=1}^{M} w'_m \varepsilon_{(i,p_m)}^T \beta_\star^{(m)} \right|}{(p+1)\,\rho(w,w')}$$

$$\leqslant c \sqrt{\sum_{m=1}^{M} \frac{\left(\varepsilon_{(i,p_m)}^T \beta_\star^{(m)}\right)^2}{(p+1)^2}},$$

where $\gamma_{0,i}$ is a point between $\sum_{m=1}^{M} w_m \varepsilon_{(i,p_m)}^T \beta_\star^{(m)}$ and $\sum_{m=1}^{M} w'_m \varepsilon_{(i,p_m)}^T \beta_\star^{(m)}$. From the assumption $\left\|\beta_\star^{(m)}\right\|^2 / (p_m+1) \leqslant C_b < \infty$, and Condition 3, we obtain

$$\sup_i \Lambda(\Delta_i) \leqslant C_g < \infty. \tag{A18}$$

As for $\left|\frac{\Delta_i(w^*)}{p+1}\right|$, using Lagrange theorem, we have

$$\left|\frac{\Delta_i(w^*)}{p+1}\right| = \frac{1}{p+1}\left| g'(\zeta_i)\left(\eta_i - \sum_{m=1}^{M} w_m^* \varepsilon_{(i,p_m)}^T \beta_\star^{(m)}\right)\right|$$

$$\leqslant c \sqrt{\sum_{m=1}^{M} \frac{\left(\eta_i - \varepsilon_{(i,p_m)}^T \beta_\star^{(m)}\right)^2}{(p+1)^2}},$$

where ζ_i is a point between η_i and $\sum_{m=1}^{M} w_m^* \varepsilon_{(i,p_m)}^T \beta_\star^{(m)}$. Again, by Condition 3, $\left\|\beta_\star^{(m)}\right\|^2 / (p_m+1) \leqslant C_b < \infty$, and the assumption $\sup_i |\eta_i| \leq C_\eta < \infty$, we obtain

$$\sup_i |\Delta_i(w^*)| \leqslant \tilde{C} < \infty. \tag{A19}$$

For (A15), we have

$$\mathbf{P}\left\{\sup_{w \in H_n} A_n^{(1)}(w) > \delta\right\}$$

$$\leqslant \mathbf{P}\left\{\sup_{w \in H_n}\left|\sum_{i=1}^{n} e_i\left[g(\eta_i) - g\left(\sum_{m=1}^{M} w_m \varepsilon_{(i,p_m)}^T \beta_\star^{(m)}\right)\right]\right| > \delta\xi_n\right\}$$

$$= \mathbf{P}\left\{\sup_{w \in H_n}\left|\sum_{i=1}^{n} e_i \frac{\Delta_i(w)}{p+1}\right| > \delta\frac{\xi_n}{p+1}\right\}$$

$$\leqslant \frac{(p+1)^2\,\mathbf{E}\left[\sup_{w \in H_n}\left|\sum_{i=1}^{n} e_i \frac{\Delta_i(w)}{p+1}\right|\right]^2}{\delta^2 \xi_n^2}$$

$$= \frac{(p+1)^2\,\mathbf{E}\left\|\sum_{i=1}^{n} e_i \frac{\Delta_i}{p+1}\right\|_\infty^2}{\delta^2 \xi_n^2},$$

where $\delta > 0$ is an arbitrary constant. Since H_n is a compact subset of $\mathbb{R}^M$, and $\rho(w,w')$ is the Euclidean metric in $\mathbb{R}^M$, (A1) is satisfied. Therefore, by Lemma A1, we see that there is a constant $A < \infty$ such that for all n,

$$\begin{aligned}\mathbf{E}\left\|\sum_{i=1}^{n} e_i\frac{\Delta_i}{p+1}\right\|_{\infty}^2 &\leqslant A\sum_{i=1}^{n}\mathbf{E}\left\|e_i\frac{\Delta_i}{p+1}\right\|_{\rho}^2\\ &\leqslant A\sup_j \mathbf{E}e_j^2\sum_{i=1}^{n}\left[\Lambda\left(\frac{\Delta_i}{p+1}\right)+\left|\frac{\Delta_i(w^*)}{p+1}\right|\right]^2\\ &\leqslant 2A\sup_j \mathbf{E}e_j^2\sum_{i=1}^{n}\left(\Lambda^2\left(\frac{\Delta_i}{p+1}\right)+\left|\frac{\Delta_i(w^*)}{p+1}\right|^2\right)\\ &= O(n),\end{aligned}$$

where the last equality is because of (A18), (A19) and $\sup_j \mathbf{E}e_j^2 < \infty$. Therefore,

$$\mathbf{P}\left\{\sup_{w\in H_n} A_n^{(1)}(w) > \delta\right\} = O\left(\frac{(p+1)^2 n}{\xi_n^2}\right) \to 0,$$

and (A15) holds.

Denote $\tilde{\Delta}_i = g\left(\sum_{m=1}^{M} w_m \varepsilon_{(i,p_m)}^T \beta_\star^{(m)}\right) - g\left(\sum_{m=1}^{M} w_m \hat{\eta}_{i,p_m}\right)$. For (A16), we have

$$\begin{aligned}&\mathbf{P}\left\{\sup_{w\in H_n} A_n^{(2)}(w) > \delta\right\}\\ &\leqslant \mathbf{P}\left\{\sup_{w\in H_n}\left|\sum_{i=1}^{n} e_i\left[g\left(\sum_{m=1}^{M} w_m\varepsilon_{(i,p_m)}^T\beta_\star^{(m)}\right)-g\left(\sum_{m=1}^{M} w_m\hat{\eta}_{i,p_m}\right)\right]\right|^2 > \delta^2\xi_n^2\right\}\\ &\leqslant \mathbf{P}\left\{\sup_{w\in H_n}\left|\sum_{i=1}^{n} e_i^2\sum_{i=1}^{n}\tilde{\Delta}_i^2\right| > \delta^2\xi_n^2\right\}\\ &\leqslant \mathbf{P}\left\{\sup_{w\in H_n}\sum_{i=1}^{n}\tilde{\Delta}_i^2 > \delta^2\xi_n p^2/\sqrt{n}\right\}+\mathbf{P}\left\{\sum_{i=1}^{n} e_i^2 > \xi_n\sqrt{n}/p^2\right\}\\ &= \mathbf{P}\left\{\sup_{w\in H_n}\sum_{i=1}^{n}\left|g'\left(\sum_{m=1}^{M} w_m\tilde{\eta}_{i*}^m\right)\sum_{m=1}^{M}\left[w_m\left(\varepsilon_{(i,p_m)}^T\beta_\star^{(m)}-\hat{\eta}_{i,p_m}\right)\right]\right|^2 > \delta^2\xi_n p^2/\sqrt{n}\right\}\\ &\quad+\mathbf{P}\left\{\sum_{i=1}^{n} e_i^2 > \xi_n\sqrt{n}/p^2\right\}\\ &\leqslant \mathbf{P}\left\{\sup_{w\in H_n} c^2\sum_{i=1}^{n}\left|\sum_{m=1}^{M}\left[w_m\left(\varepsilon_{(i,p_m)}^T\beta_\star^{(m)}-\hat{\eta}_{i,p_m}\right)\right]\right|^2 > \delta^2\xi_n p^2/\sqrt{n}\right\}+\mathbf{P}\left\{\sum_{i=1}^{n} e_i^2 > \xi_n\sqrt{n}/p^2\right\}\\ &\leqslant \mathbf{P}\left\{c^2\sum_{i=1}^{n}\sum_{m=1}^{M}\left(\varepsilon_{(i,p_m)}^T\beta_\star^{(m)}-\hat{\eta}_{i,p_m}\right)^2 > \delta^2\xi_n p^2/\sqrt{n}\right\}+\mathbf{P}\left\{\sum_{i=1}^{n} e_i^2 > \xi_n\sqrt{n}/p^2\right\}\\ &\leqslant \mathbf{P}\left\{c^2\sum_{i=1}^{n}\sum_{m=1}^{M}\left(\varepsilon_{(i,p_m)}^T\beta_\star^{(m)}-\hat{\eta}_{i,p_m}\right)^2 > \delta^2\xi_n p^2/\sqrt{n}\right\}+\frac{p^2\sum_{i=1}^{n}\mathbf{E}e_i^2}{\xi_n\sqrt{n}}.\end{aligned}$$

From Lemma A3 and Condition 3 , we see that

$$\sum_{i=1}^{n}\sum_{m=1}^{M}\left(\varepsilon_{(i,p_m)}^T\beta_\star^{(m)}-\hat{\eta}_{i,p_m}\right)^2 = \sum_{i=1}^{n}\sum_{m=1}^{M}\left[\varepsilon_{(i,p_m)}^T\left(\beta_\star^{(m)}-\hat{\beta}^{(m)}\right)\right]^2 = O_p(\sum_{m=1}^{M} p_m^4).$$

Therefore, $\lim_{n\to+\infty}\mathbf{P}\left\{\sup_{w\in H_n} A_n^{(2)}(w) > \delta\right\} = 0$, that is, (A16) is valid.

Write $\overline{\Delta}_i = g\left(\sum_{m=1}^{M} w_m\varepsilon_{(i,p_m)}^T\hat{\beta}^{(m)}\right) - g\left(\sum_{m=1}^{M} w_m\tilde{\eta}_{i,p_m}\right)$. For (A17), we have

$$
\begin{aligned}
&\mathbf{P}\left\{\sup_{w\in H_n} A_n^{(3)}(w) > \delta\right\} \\
&\leqslant \mathbf{P}\left\{\sup_{w\in H_n}\left|\sum_{i=1}^{n} e_i\left[g\left(\sum_{m=1}^{M} w_m \varepsilon_{(i,p_m)}^T \hat{\beta}^{(m)}\right) - g\left(\sum_{m=1}^{M} w_m \tilde{\eta}_{i,p_m}\right)\right]\right|^2 > \delta^2\xi_n^2\right\} \\
&\leqslant \mathbf{P}\left\{\sup_{w\in H_n}\left|\sum_{i=1}^{n} e_i^2 \sum_{i=1}^{n}\overline{\Delta}_i^2\right| > \delta^2\xi_n^2\right\} \\
&\leqslant \mathbf{P}\left\{\sup_{w\in H_n}\sum_{i=1}^{n}\overline{\Delta}_i^2 > \delta^2\xi_n p^2/\sqrt{n}\right\} + \mathbf{P}\left\{\sum_{i=1}^{n} e_i^2 > \xi_n\sqrt{n}/p^2\right\} \\
&= \mathbf{P}\left\{\sup_{w\in H_n}\sum_{i=1}^{n}\left|g'\left(\sum_{m=1}^{M} w_m \eta_{i,p_m}^*\right)\sum_{m=1}^{M}\left[w_m\left(\varepsilon_{(i,p_m)}^T\hat{\beta}^{(m)} - \tilde{\eta}_{i,p_m}\right)\right]\right|^2 > \delta^2\xi_n p^2/\sqrt{n}\right\} \\
&\quad + \mathbf{P}\left\{\sum_{i=1}^{n} e_i^2 > \xi_n\sqrt{n}/p^2\right\} \\
&\leqslant \mathbf{P}\left\{\sup_{w\in H_n} c^2\sum_{i=1}^{n}\left|\sum_{m=1}^{M}\left[w_m\left(\varepsilon_{(i,p_m)}^T\hat{\beta}^{(m)} - \tilde{\eta}_{i,p_m}\right)\right]\right|^2 > \delta^2\xi_n p^2/\sqrt{n}\right\} + \mathbf{P}\left\{\sum_{i=1}^{n} e_i^2 > \xi_n\sqrt{n}/p^2\right\} \\
&\leqslant \mathbf{P}\left\{c^2\sum_{i=1}^{n}\sum_{m=1}^{M}\left(\varepsilon_{(i,p_m)}^T\hat{\beta}^{(m)} - \tilde{\eta}_{i,p_m}\right)^2 > \delta^2\xi_n p^2/\sqrt{n}\right\} + \mathbf{P}\left\{\sum_{i=1}^{n} e_i^2 > \xi_n\sqrt{n}/p^2\right\} \\
&\leqslant \mathbf{P}\left\{c^2\sum_{i=1}^{n}\sum_{m=1}^{M}\left(\varepsilon_{(i,p_m)}^T\hat{\beta}^{(m)} - \tilde{\eta}_{i,p_m}\right)^2 > \delta^2\xi_n p^2/\sqrt{n}\right\} + \frac{p^2\sum_{i=1}^{n}\mathbf{E}e_i^2}{\xi_n\sqrt{n}}.
\end{aligned}
$$

From Condition 5, we see that

$$
\sum_{i=1}^{n}\sum_{m=1}^{M}\left(\varepsilon_{(i,p_m)}^T\hat{\beta}^{(m)} - \tilde{\eta}_{i,p_m}\right)^2 = O_p\Big(\sum_{m=1}^{M} p_m^4\Big).
$$

Therefore, $\lim_{n\to+\infty}\mathbf{P}\left\{\sup_{w\in H_n} A_n^{(3)}(w) > \delta\right\} = 0$, that is, (A17) is valid.

Appendix C. Simulation Results in Section 4.1

Table A1. Prediction errors with n = 60 in Case 1.

R		AIC	BIC	FPCA	S-AIC	S-BIC	CV1	CV2
1	Mean	0.432	0.408	0.404	0.433	0.408	0.394	0.393
	Median	0.417	0.417	0.417	0.417	0.417	0.375	0.417
	Var	0.023	0.023	0.020	0.023	0.024	0.023	0.021
2	Mean	0.312	0.294	0.249	0.311	0.292	0.225	0.226
	Median	0.333	0.333	0.250	0.333	0.333	0.250	0.250
	Var	0.013	0.013	0.016	0.013	0.013	0.013	0.013
3	Mean	0.273	0.262	0.226	0.273	0.260	0.188	0.189
	Median	0.250	0.250	0.250	0.250	0.250	0.167	0.167
	Var	0.017	0.017	0.015	0.017	0.017	0.016	0.015
4	Mean	0.256	0.243	0.183	0.256	0.247	0.162	0.163
	Median	0.250	0.250	0.167	0.250	0.250	0.167	0.167
	Var	0.018	0.017	0.011	0.018	0.017	0.013	0.013
5	Mean	0.203	0.196	0.148	0.203	0.193	0.133	0.134
	Median	0.167	0.167	0.167	0.167	0.167	0.083	0.083
	Var	0.014	0.014	0.011	0.014	0.013	0.009	0.009

Table A1. *Cont.*

R		AIC	BIC	FPCA	S-AIC	S-BIC	CV1	CV2
6	Mean	0.234	0.233	0.135	0.234	0.233	0.117	0.115
	Median	0.250	0.250	0.125	0.250	0.250	0.083	0.083
	Var	0.016	0.016	0.010	0.016	0.016	0.010	0.010
7	Mean	0.214	0.213	0.149	0.214	0.214	0.118	0.117
	Median	0.208	0.208	0.167	0.208	0.250	0.083	0.083
	Var	0.014	0.015	0.010	0.014	0.015	0.009	0.008
8	Mean	0.213	0.209	0.134	0.213	0.210	0.104	0.103
	Median	0.250	0.167	0.125	0.250	0.167	0.083	0.083
	Var	0.012	0.012	0.009	0.012	0.012	0.008	0.008
9	Mean	0.196	0.196	0.128	0.196	0.196	0.096	0.099
	Median	0.167	0.167	0.083	0.167	0.167	0.083	0.083
	Var	0.014	0.014	0.012	0.014	0.015	0.008	0.008
10	Mean	0.209	0.208	0.126	0.209	0.206	0.088	0.087
	Median	0.167	0.167	0.083	0.167	0.167	0.083	0.083
	Var	0.016	0.016	0.009	0.016	0.016	0.006	0.006

Table A2. Prediction errors with n = 200 in Case 1.

R		AIC	BIC	FPCA	S-AIC	S-BIC	CV1	CV2
1	Mean	0.355	0.350	0.329	0.355	0.349	0.322	0.322
	Median	0.350	0.350	0.325	0.350	0.350	0.325	0.313
	Var	0.006	0.007	0.007	0.006	0.007	0.006	0.006
2	Mean	0.262	0.262	0.234	0.262	0.262	0.227	0.227
	Median	0.275	0.275	0.225	0.275	0.275	0.225	0.225
	Var	0.005	0.005	0.004	0.005	0.005	0.004	0.004
3	Mean	0.205	0.205	0.184	0.205	0.205	0.174	0.174
	Median	0.200	0.200	0.175	0.200	0.200	0.175	0.175
	Var	0.005	0.005	0.004	0.005	0.005	0.003	0.003
4	Mean	0.163	0.163	0.134	0.163	0.163	0.128	0.128
	Median	0.150	0.150	0.125	0.150	0.150	0.125	0.125
	Var	0.004	0.004	0.003	0.004	0.004	0.003	0.003
5	Mean	0.139	0.139	0.113	0.139	0.139	0.110	0.110
	Median	0.125	0.125	0.113	0.125	0.125	0.100	0.100
	Var	0.003	0.003	0.003	0.003	0.003	0.002	0.002
6	Mean	0.136	0.136	0.101	0.136	0.136	0.094	0.094
	Median	0.125	0.125	0.100	0.125	0.125	0.100	0.100
	Var	0.003	0.003	0.002	0.003	0.003	0.002	0.002
7	Mean	0.129	0.129	0.099	0.129	0.129	0.086	0.086
	Median	0.125	0.125	0.100	0.125	0.125	0.075	0.075
	Var	0.003	0.003	0.003	0.003	0.003	0.002	0.002
8	Mean	0.121	0.121	0.091	0.121	0.121	0.083	0.082
	Median	0.113	0.113	0.075	0.113	0.113	0.075	0.075
	Var	0.003	0.003	0.002	0.003	0.003	0.002	0.002
9	Mean	0.127	0.127	0.090	0.127	0.127	0.084	0.083
	Median	0.125	0.125	0.100	0.125	0.125	0.075	0.075
	Var	0.003	0.003	0.002	0.003	0.003	0.002	0.002
10	Mean	0.121	0.121	0.088	0.121	0.121	0.069	0.069
	Median	0.125	0.125	0.075	0.125	0.125	0.075	0.075
	Var	0.003	0.003	0.002	0.003	0.003	0.002	0.002

Table A3. Prediction errors with n = 500 in Case 1.

R		AIC	BIC	FPCA	S-AIC	S-BIC	CV1	CV2
1	Mean	0.349	0.349	0.332	0.349	0.349	0.330	0.330
	Median	0.345	0.345	0.330	0.345	0.345	0.330	0.330
	Var	0.002	0.002	0.002	0.002	0.002	0.002	0.002

Table A3. *Cont.*

R		AIC	BIC	FPCA	S-AIC	S-BIC	CV1	CV2
2	Mean	0.240	0.240	0.232	0.240	0.240	0.228	0.228
	Median	0.240	0.240	0.230	0.240	0.240	0.230	0.230
	Var	0.001	0.001	0.002	0.001	0.001	0.002	0.002
3	Mean	0.176	0.176	0.174	0.176	0.176	0.168	0.168
	Median	0.170	0.170	0.170	0.170	0.170	0.160	0.160
	Var	0.002	0.002	0.001	0.002	0.002	0.001	0.001
4	Mean	0.143	0.143	0.133	0.143	0.143	0.135	0.134
	Median	0.140	0.140	0.130	0.140	0.140	0.130	0.130
	Var	0.001	0.001	0.001	0.001	0.001	0.001	0.001
5	Mean	0.126	0.126	0.114	0.126	0.126	0.115	0.115
	Median	0.120	0.120	0.110	0.120	0.120	0.110	0.110
	Var	0.001	0.001	0.001	0.001	0.001	0.001	0.001
6	Mean	0.109	0.109	0.097	0.109	0.109	0.095	0.096
	Median	0.110	0.110	0.090	0.110	0.110	0.090	0.090
	Var	0.001	0.001	0.001	0.001	0.001	0.001	0.001
7	Mean	0.106	0.106	0.090	0.106	0.106	0.089	0.089
	Median	0.110	0.110	0.090	0.110	0.110	0.090	0.090
	Var	0.001	0.001	0.001	0.001	0.001	0.001	0.001
8	Mean	0.096	0.096	0.081	0.096	0.096	0.084	0.084
	Median	0.090	0.090	0.080	0.090	0.090	0.080	0.080
	Var	0.001	0.001	0.001	0.001	0.001	0.001	0.001
9	Mean	0.090	0.090	0.075	0.090	0.090	0.070	0.070
	Median	0.085	0.085	0.070	0.085	0.085	0.065	0.065
	Var	0.001	0.001	0.001	0.001	0.001	0.001	0.001
10	Mean	0.091	0.091	0.075	0.091	0.091	0.069	0.068
	Median	0.090	0.090	0.070	0.090	0.090	0.065	0.065
	Var	0.001	0.001	0.001	0.001	0.001	0.001	0.001

Table A4. Prediction errors with n = 60 in Case 2.

R		AIC	BIC	FPCA	S-AIC	S-BIC	CV1	CV2
1	Mean	0.362	0.346	0.359	0.359	0.342	0.351	0.354
	Median	0.333	0.333	0.333	0.333	0.333	0.333	0.333
	Var	0.021	0.021	0.021	0.021	0.021	0.021	0.022
2	Mean	0.315	0.251	0.262	0.300	0.245	0.245	0.248
	Median	0.333	0.250	0.250	0.250	0.250	0.250	0.250
	Var	0.020	0.016	0.016	0.019	0.015	0.015	0.016
3	Mean	0.269	0.193	0.208	0.257	0.188	0.185	0.184
	Median	0.250	0.167	0.167	0.250	0.167	0.167	0.167
	Var	0.016	0.014	0.014	0.015	0.013	0.012	0.013
4	Mean	0.258	0.174	0.176	0.252	0.167	0.163	0.164
	Median	0.250	0.167	0.167	0.250	0.167	0.167	0.167
	Var	0.018	0.013	0.012	0.017	0.013	0.012	0.012
5	Mean	0.244	0.145	0.169	0.239	0.137	0.138	0.135
	Median	0.250	0.167	0.167	0.250	0.167	0.083	0.083
	Var	0.017	0.010	0.013	0.017	0.010	0.011	0.011
6	Mean	0.234	0.142	0.150	0.227	0.131	0.122	0.119
	Median	0.250	0.167	0.167	0.250	0.083	0.083	0.083
	Var	0.018	0.010	0.012	0.017	0.010	0.009	0.009
7	Mean	0.214	0.127	0.142	0.205	0.118	0.113	0.110
	Median	0.167	0.083	0.167	0.167	0.083	0.083	0.083
	Var	0.016	0.011	0.012	0.016	0.010	0.009	0.009
8	Mean	0.230	0.120	0.156	0.223	0.110	0.105	0.107
	Median	0.250	0.083	0.167	0.167	0.083	0.083	0.083
	Var	0.018	0.010	0.014	0.017	0.009	0.009	0.010
9	Mean	0.204	0.121	0.160	0.192	0.108	0.100	0.099
	Median	0.167	0.083	0.167	0.167	0.083	0.083	0.083
	Var	0.017	0.009	0.016	0.016	0.009	0.008	0.008

Table A4. *Cont.*

R		AIC	BIC	FPCA	S-AIC	S-BIC	CV1	CV2
10	Mean	0.201	0.114	0.178	0.182	0.101	0.096	0.096
	Median	0.167	0.083	0.167	0.167	0.083	0.083	0.083
	Var	0.019	0.010	0.017	0.019	0.009	0.008	0.008

Table A5. Prediction errors with n = 200 in Case 2.

R		AIC	BIC	FPCA	S-AIC	S-BIC	CV1	CV2
1	Mean	0.369	0.336	0.349	0.369	0.336	0.342	0.341
	Median	0.375	0.325	0.350	0.375	0.325	0.350	0.338
	Var	0.007	0.007	0.006	0.006	0.007	0.006	0.006
2	Mean	0.265	0.253	0.239	0.265	0.248	0.233	0.233
	Median	0.275	0.250	0.250	0.275	0.250	0.225	0.225
	Var	0.006	0.005	0.005	0.006	0.005	0.005	0.005
3	Mean	0.204	0.204	0.184	0.204	0.203	0.175	0.175
	Median	0.200	0.200	0.175	0.200	0.200	0.175	0.175
	Var	0.004	0.004	0.003	0.004	0.004	0.003	0.003
4	Mean	0.175	0.175	0.147	0.175	0.175	0.143	0.142
	Median	0.175	0.175	0.150	0.175	0.175	0.150	0.125
	Var	0.004	0.004	0.004	0.004	0.004	0.004	0.003
5	Mean	0.157	0.157	0.130	0.157	0.157	0.118	0.118
	Median	0.150	0.150	0.125	0.150	0.150	0.125	0.125
	Var	0.004	0.004	0.003	0.004	0.004	0.003	0.003
6	Mean	0.148	0.148	0.120	0.148	0.148	0.108	0.107
	Median	0.150	0.150	0.125	0.150	0.150	0.100	0.100
	Var	0.004	0.004	0.003	0.004	0.004	0.002	0.002
7	Mean	0.150	0.150	0.116	0.150	0.150	0.092	0.091
	Median	0.150	0.150	0.113	0.150	0.150	0.100	0.100
	Var	0.003	0.003	0.003	0.003	0.004	0.002	0.002
8	Mean	0.162	0.161	0.125	0.162	0.161	0.091	0.092
	Median	0.150	0.150	0.125	0.150	0.150	0.088	0.100
	Var	0.005	0.005	0.004	0.005	0.005	0.002	0.002
9	Mean	0.173	0.167	0.130	0.173	0.165	0.086	0.087
	Median	0.175	0.175	0.125	0.175	0.150	0.075	0.075
	Var	0.004	0.004	0.004	0.004	0.004	0.002	0.002
10	Mean	0.192	0.172	0.147	0.192	0.167	0.088	0.090
	Median	0.200	0.175	0.150	0.200	0.150	0.075	0.075
	Var	0.006	0.005	0.005	0.006	0.005	0.002	0.002

Table A6. Prediction errors with n = 500 in Case 2.

R		AIC	BIC	FPCA	S-AIC	S-BIC	CV1	CV2
1	Mean	0.345	0.338	0.332	0.345	0.336	0.330	0.330
	Median	0.350	0.340	0.330	0.350	0.340	0.330	0.330
	Var	0.003	0.002	0.002	0.003	0.002	0.002	0.002
2	Mean	0.239	0.239	0.227	0.239	0.239	0.225	0.225
	Median	0.240	0.240	0.230	0.240	0.240	0.225	0.220
	Var	0.002	0.002	0.002	0.002	0.002	0.002	0.002
3	Mean	0.182	0.182	0.170	0.182	0.182	0.168	0.168
	Median	0.180	0.180	0.170	0.180	0.180	0.170	0.170
	Var	0.002	0.002	0.001	0.002	0.002	0.001	0.001
4	Mean	0.152	0.152	0.141	0.152	0.152	0.136	0.136
	Median	0.150	0.150	0.140	0.150	0.150	0.140	0.140
	Var	0.001	0.001	0.001	0.001	0.001	0.001	0.001
5	Mean	0.135	0.135	0.120	0.135	0.135	0.114	0.114
	Median	0.130	0.130	0.120	0.130	0.130	0.110	0.110
	Var	0.001	0.001	0.001	0.001	0.001	0.001	0.001

Table A6. *Cont.*

R		AIC	BIC	FPCA	S-AIC	S-BIC	CV1	CV2
6	Mean	0.129	0.129	0.110	0.129	0.129	0.100	0.101
	Median	0.130	0.130	0.110	0.130	0.130	0.100	0.100
	Var	0.001	0.001	0.001	0.001	0.001	0.001	0.001
7	Mean	0.128	0.128	0.107	0.128	0.128	0.092	0.093
	Median	0.130	0.130	0.100	0.130	0.130	0.090	0.090
	Var	0.002	0.002	0.001	0.002	0.002	0.001	0.001
8	Mean	0.134	0.134	0.109	0.134	0.134	0.086	0.087
	Median	0.130	0.130	0.110	0.130	0.130	0.080	0.080
	Var	0.002	0.002	0.001	0.002	0.002	0.001	0.001
9	Mean	0.147	0.147	0.117	0.147	0.147	0.086	0.088
	Median	0.140	0.140	0.110	0.140	0.140	0.090	0.090
	Var	0.002	0.002	0.002	0.002	0.002	0.001	0.001
10	Mean	0.171	0.171	0.135	0.171	0.171	0.093	0.096
	Median	0.170	0.170	0.135	0.170	0.170	0.090	0.090
	Var	0.003	0.003	0.002	0.003	0.003	0.001	0.001

Table A7. Prediction errors with n = 60 in Case 3.

R		AIC	BIC	FPCA	S-AIC	S-BIC	CV1	CV2
1	Mean	0.494	0.482	0.430	0.490	0.483	0.405	0.413
	Median	0.500	0.500	0.417	0.500	0.500	0.417	0.417
	Var	0.026	0.029	0.026	0.027	0.029	0.022	0.022
2	Mean	0.428	0.412	0.317	0.427	0.412	0.318	0.303
	Median	0.417	0.417	0.333	0.417	0.417	0.333	0.333
	Var	0.021	0.023	0.028	0.021	0.023	0.018	0.018
3	Mean	0.416	0.401	0.317	0.419	0.403	0.313	0.302
	Median	0.417	0.417	0.292	0.417	0.417	0.292	0.250
	Var	0.028	0.031	0.037	0.027	0.030	0.032	0.031
4	Mean	0.424	0.387	0.393	0.420	0.382	0.357	0.344
	Median	0.500	0.417	0.417	0.458	0.417	0.333	0.333
	Var	0.047	0.048	0.056	0.044	0.046	0.046	0.043
5	Mean	0.372	0.362	0.493	0.398	0.365	0.380	0.355
	Median	0.333	0.333	0.583	0.417	0.333	0.417	0.333
	Var	0.052	0.054	0.067	0.049	0.052	0.053	0.048
6	Mean	0.400	0.383	0.608	0.427	0.390	0.446	0.430
	Median	0.417	0.375	0.667	0.417	0.375	0.500	0.417
	Var	0.072	0.075	0.060	0.066	0.075	0.067	0.066
7	Mean	0.374	0.378	0.628	0.428	0.388	0.481	0.468
	Median	0.333	0.333	0.667	0.417	0.417	0.500	0.500
	Var	0.072	0.075	0.052	0.063	0.072	0.067	0.070
8	Mean	0.457	0.457	0.673	0.527	0.474	0.615	0.593
	Median	0.417	0.417	0.750	0.583	0.500	0.667	0.667
	Var	0.098	0.098	0.053	0.071	0.091	0.073	0.075
9	Mean	0.565	0.565	0.738	0.642	0.583	0.652	0.659
	Median	0.583	0.583	0.750	0.750	0.667	0.750	0.750
	Var	0.099	0.099	0.040	0.079	0.087	0.072	0.074
10	Mean	0.565	0.565	0.744	0.662	0.613	0.698	0.694
	Median	0.583	0.583	0.750	0.667	0.667	0.750	0.750
	Var	0.096	0.096	0.037	0.063	0.080	0.057	0.065

Table A8. Prediction errors with n = 200 in Case 3.

R		AIC	BIC	FPCA	S-AIC	S-BIC	CV1	CV2
1	Mean	0.406	0.403	0.366	0.406	0.401	0.341	0.342
	Median	0.400	0.400	0.350	0.400	0.400	0.325	0.325
	Var	0.006	0.007	0.007	0.006	0.007	0.007	0.007

Table A8. *Cont.*

R		AIC	BIC	FPCA	S-AIC	S-BIC	CV1	CV2
2	Mean	0.378	0.377	0.310	0.378	0.378	0.272	0.271
	Median	0.375	0.375	0.300	0.375	0.375	0.250	0.250
	Var	0.010	0.010	0.010	0.010	0.010	0.008	0.007
3	Mean	0.428	0.428	0.324	0.428	0.427	0.253	0.251
	Median	0.463	0.463	0.300	0.463	0.450	0.225	0.225
	Var	0.018	0.018	0.016	0.018	0.018	0.010	0.009
4	Mean	0.465	0.427	0.370	0.470	0.430	0.259	0.254
	Median	0.500	0.475	0.350	0.500	0.475	0.225	0.225
	Var	0.031	0.035	0.029	0.030	0.034	0.021	0.020
5	Mean	0.281	0.231	0.507	0.310	0.228	0.282	0.276
	Median	0.200	0.175	0.500	0.225	0.175	0.238	0.225
	Var	0.035	0.021	0.041	0.034	0.020	0.030	0.029
6	Mean	0.242	0.242	0.612	0.289	0.242	0.325	0.321
	Median	0.175	0.175	0.675	0.225	0.175	0.238	0.238
	Var	0.040	0.040	0.036	0.039	0.037	0.050	0.049
7	Mean	0.298	0.298	0.712	0.363	0.294	0.368	0.362
	Median	0.200	0.200	0.725	0.313	0.200	0.300	0.288
	Var	0.059	0.059	0.014	0.056	0.056	0.064	0.063
8	Mean	0.476	0.476	0.749	0.553	0.473	0.498	0.495
	Median	0.513	0.513	0.763	0.588	0.500	0.588	0.575
	Var	0.086	0.086	0.009	0.068	0.084	0.076	0.076
9	Mean	0.497	0.497	0.785	0.625	0.500	0.592	0.586
	Median	0.525	0.525	0.800	0.700	0.538	0.663	0.650
	Var	0.104	0.104	0.005	0.057	0.099	0.062	0.064
10	Mean	0.606	0.606	0.807	0.746	0.627	0.662	0.661
	Median	0.750	0.750	0.825	0.825	0.800	0.763	0.750
	Var	0.105	0.105	0.004	0.042	0.101	0.053	0.054

Table A9. Prediction errors with n = 500 in Case 3.

R		AIC	BIC	FPCA	S-AIC	S-BIC	CV1	CV2
1	Mean	0.394	0.394	0.360	0.394	0.394	0.338	0.338
	Median	0.390	0.390	0.355	0.390	0.390	0.340	0.340
	Var	0.004	0.004	0.003	0.004	0.004	0.003	0.003
2	Mean	0.345	0.345	0.280	0.345	0.345	0.241	0.244
	Median	0.340	0.340	0.275	0.340	0.340	0.240	0.240
	Var	0.005	0.005	0.003	0.005	0.005	0.002	0.002
3	Mean	0.426	0.426	0.286	0.426	0.426	0.190	0.200
	Median	0.430	0.430	0.270	0.430	0.430	0.190	0.200
	Var	0.008	0.008	0.008	0.008	0.008	0.002	0.002
4	Mean	0.524	0.490	0.390	0.526	0.490	0.170	0.190
	Median	0.550	0.540	0.400	0.550	0.540	0.160	0.180
	Var	0.017	0.025	0.018	0.015	0.025	0.002	0.003
5	Mean	0.225	0.199	0.535	0.241	0.198	0.168	0.170
	Median	0.160	0.160	0.560	0.180	0.160	0.150	0.160
	Var	0.028	0.018	0.017	0.027	0.018	0.006	0.006
6	Mean	0.186	0.183	0.665	0.225	0.184	0.183	0.183
	Median	0.140	0.140	0.680	0.180	0.140	0.140	0.150
	Var	0.014	0.013	0.009	0.014	0.012	0.013	0.011
7	Mean	0.251	0.251	0.735	0.322	0.252	0.251	0.253
	Median	0.170	0.170	0.740	0.260	0.170	0.170	0.190
	Var	0.033	0.033	0.004	0.028	0.031	0.033	0.028
8	Mean	0.376	0.376	0.776	0.511	0.379	0.376	0.383
	Median	0.335	0.335	0.780	0.520	0.335	0.335	0.385
	Var	0.065	0.065	0.002	0.048	0.062	0.065	0.057

Table A9. *Cont.*

R		AIC	BIC	FPCA	S-AIC	S-BIC	CV1	CV2
9	Mean	0.467	0.467	0.797	0.650	0.476	0.467	0.491
	Median	0.475	0.475	0.800	0.700	0.480	0.475	0.510
	Var	0.087	0.087	0.002	0.039	0.082	0.087	0.076
10	Mean	0.652	0.652	0.822	0.820	0.675	0.652	0.713
	Median	0.780	0.780	0.820	0.840	0.790	0.780	0.800
	Var	0.071	0.071	0.002	0.012	0.062	0.071	0.048

Table A10. Prediction errors with n = 60 in Case 4.

R		AIC	BIC	FPCA	S-AIC	S-BIC	CV1	CV2
1	Mean	0.389	0.378	0.417	0.396	0.381	0.381	0.387
	Median	0.417	0.333	0.417	0.417	0.417	0.417	0.417
	Var	0.024	0.024	0.023	0.023	0.023	0.022	0.024
2	Mean	0.286	0.268	0.363	0.299	0.269	0.268	0.268
	Median	0.250	0.250	0.333	0.250	0.250	0.250	0.250
	Var	0.022	0.021	0.029	0.022	0.022	0.021	0.021
3	Mean	0.230	0.219	0.382	0.259	0.228	0.219	0.219
	Median	0.167	0.167	0.333	0.250	0.167	0.167	0.167
	Var	0.024	0.023	0.040	0.024	0.023	0.023	0.023
4	Mean	0.186	0.181	0.460	0.242	0.199	0.181	0.181
	Median	0.167	0.167	0.417	0.167	0.167	0.167	0.167
	Var	0.022	0.022	0.048	0.031	0.024	0.022	0.022
5	Mean	0.195	0.194	0.545	0.284	0.216	0.194	0.194
	Median	0.167	0.167	0.583	0.250	0.167	0.167	0.167
	Var	0.029	0.030	0.054	0.046	0.034	0.030	0.030
6	Mean	0.213	0.211	0.642	0.374	0.256	0.211	0.211
	Median	0.167	0.167	0.667	0.333	0.167	0.167	0.167
	Var	0.042	0.042	0.045	0.062	0.049	0.042	0.042
7	Mean	0.208	0.210	0.680	0.424	0.268	0.210	0.210
	Median	0.167	0.167	0.750	0.417	0.167	0.167	0.167
	Var	0.052	0.053	0.037	0.068	0.060	0.053	0.053
8	Mean	0.228	0.228	0.727	0.513	0.310	0.228	0.228
	Median	0.167	0.167	0.750	0.500	0.250	0.167	0.167
	Var	0.059	0.059	0.025	0.067	0.071	0.059	0.059
9	Mean	0.259	0.258	0.730	0.572	0.366	0.258	0.258
	Median	0.167	0.167	0.750	0.583	0.250	0.167	0.167
	Var	0.084	0.084	0.030	0.069	0.091	0.084	0.084
10	Mean	0.303	0.303	0.761	0.665	0.455	0.303	0.303
	Median	0.167	0.167	0.750	0.750	0.417	0.167	0.167
	Var	0.099	0.099	0.020	0.047	0.096	0.099	0.099

Table A11. Prediction errors with n = 200 in Case 4.

R		AIC	BIC	FPCA	S-AIC	S-BIC	CV1	CV2
1	Mean	0.378	0.348	0.387	0.380	0.350	0.354	0.353
	Median	0.375	0.350	0.375	0.375	0.350	0.350	0.350
	Var	0.008	0.008	0.008	0.008	0.007	0.007	0.007
2	Mean	0.277	0.251	0.330	0.287	0.253	0.258	0.258
	Median	0.275	0.250	0.325	0.275	0.250	0.250	0.250
	Var	0.007	0.006	0.010	0.007	0.006	0.007	0.006
3	Mean	0.193	0.183	0.374	0.216	0.186	0.205	0.205
	Median	0.175	0.175	0.375	0.200	0.175	0.200	0.200
	Var	0.006	0.005	0.020	0.007	0.005	0.006	0.006
4	Mean	0.168	0.167	0.512	0.219	0.171	0.217	0.216
	Median	0.150	0.150	0.550	0.200	0.150	0.200	0.200
	Var	0.008	0.008	0.022	0.012	0.008	0.011	0.011

Table A11. *Cont.*

R		AIC	BIC	FPCA	S-AIC	S-BIC	CV1	CV2
5	Mean	0.141	0.141	0.613	0.237	0.152	0.237	0.237
	Median	0.125	0.125	0.650	0.200	0.125	0.200	0.200
	Var	0.008	0.008	0.020	0.019	0.009	0.018	0.018
6	Mean	0.132	0.132	0.700	0.294	0.146	0.292	0.291
	Median	0.100	0.100	0.700	0.250	0.125	0.250	0.250
	Var	0.011	0.011	0.010	0.030	0.013	0.029	0.029
7	Mean	0.138	0.138	0.742	0.392	0.161	0.381	0.377
	Median	0.100	0.100	0.750	0.375	0.125	0.375	0.375
	Var	0.014	0.014	0.007	0.039	0.017	0.033	0.033
8	Mean	0.154	0.154	0.769	0.512	0.193	0.490	0.487
	Median	0.100	0.100	0.775	0.550	0.125	0.500	0.500
	Var	0.023	0.023	0.004	0.042	0.028	0.039	0.039
9	Mean	0.175	0.175	0.788	0.624	0.232	0.583	0.580
	Median	0.100	0.100	0.800	0.675	0.125	0.625	0.625
	Var	0.038	0.038	0.005	0.032	0.046	0.035	0.035
10	Mean	0.192	0.192	0.800	0.695	0.282	0.654	0.653
	Median	0.100	0.100	0.800	0.725	0.175	0.688	0.675
	Var	0.049	0.049	0.004	0.024	0.063	0.029	0.029

Table A12. Prediction errors with n = 500 in Case 4.

R		AIC	BIC	FPCA	S-AIC	S-BIC	CV1	CV2
1	Mean	0.380	0.339	0.367	0.380	0.340	0.338	0.339
	Median	0.380	0.340	0.360	0.380	0.340	0.340	0.340
	Var	0.003	0.003	0.004	0.003	0.003	0.003	0.003
2	Mean	0.278	0.242	0.310	0.284	0.242	0.228	0.229
	Median	0.270	0.240	0.300	0.280	0.240	0.230	0.230
	Var	0.005	0.003	0.005	0.004	0.003	0.002	0.002
3	Mean	0.180	0.177	0.385	0.198	0.179	0.176	0.179
	Median	0.170	0.170	0.380	0.190	0.170	0.170	0.180
	Var	0.002	0.002	0.009	0.003	0.002	0.002	0.002
4	Mean	0.141	0.141	0.527	0.184	0.143	0.141	0.146
	Median	0.140	0.140	0.540	0.170	0.140	0.140	0.140
	Var	0.001	0.001	0.010	0.003	0.001	0.001	0.001
5	Mean	0.122	0.122	0.649	0.203	0.126	0.122	0.130
	Median	0.120	0.120	0.660	0.185	0.120	0.120	0.120
	Var	0.002	0.002	0.004	0.007	0.002	0.002	0.002
6	Mean	0.109	0.109	0.716	0.266	0.116	0.109	0.125
	Median	0.100	0.100	0.720	0.240	0.110	0.100	0.110
	Var	0.002	0.002	0.003	0.013	0.002	0.002	0.003
7	Mean	0.103	0.103	0.754	0.371	0.115	0.103	0.129
	Median	0.090	0.090	0.750	0.360	0.100	0.090	0.120
	Var	0.003	0.003	0.002	0.020	0.004	0.003	0.005
8	Mean	0.102	0.102	0.775	0.490	0.119	0.102	0.141
	Median	0.090	0.090	0.780	0.500	0.100	0.090	0.120
	Var	0.005	0.005	0.002	0.023	0.006	0.005	0.007
9	Mean	0.112	0.112	0.791	0.629	0.143	0.112	0.184
	Median	0.090	0.090	0.790	0.650	0.110	0.090	0.140
	Var	0.009	0.009	0.002	0.015	0.012	0.009	0.017
10	Mean	0.114	0.114	0.802	0.707	0.155	0.114	0.211
	Median	0.080	0.080	0.800	0.720	0.110	0.080	0.160
	Var	0.014	0.014	0.002	0.007	0.019	0.014	0.025

Appendix D. Simulation Results in Section 4.2

Table A13. Prediction errors with $R = 1$

N	R = 1	AIC	BIC	PCA	SAIC	SBIC	CV1	CV2
200	Mean	0.329	0.325	0.312	0.323	0.322	0.313	0.313
	Median	0.325	0.325	0.300	0.325	0.325	0.325	0.325
	Var	0.006	0.006	0.005	0.006	0.006	0.006	0.006
400	Mean	0.330	0.319	0.305	0.327	0.314	0.304	0.304
	Median	0.325	0.313	0.300	0.325	0.313	0.300	0.300
	Var	0.003	0.003	0.003	0.003	0.003	0.003	0.003
1000	Mean	0.332	0.326	0.304	0.330	0.326	0.305	0.304
	Median	0.330	0.320	0.303	0.330	0.320	0.303	0.300
	Var	0.001	0.001	0.001	0.001	0.001	0.001	0.001

Table A14. Prediction errors with $R = 3$

N	R = 3	AIC	BIC	PCA	SAIC	SBIC	CV1	CV2
200	Mean	0.173	0.173	0.168	0.173	0.173	0.162	0.162
	Median	0.175	0.175	0.175	0.175	0.175	0.175	0.163
	Var	0.003	0.003	0.002	0.003	0.003	0.002	0.002
400	Mean	0.172	0.172	0.163	0.172	0.171	0.167	0.169
	Median	0.175	0.175	0.163	0.175	0.175	0.175	0.175
	Var	0.001	0.001	0.002	0.001	0.001	0.001	0.002
1000	Mean	0.175	0.193	0.156	0.175	0.189	0.149	0.148
	Median	0.180	0.198	0.160	0.180	0.190	0.145	0.145
	Var	0.001	0.001	0.000	0.001	0.001	0.001	0.001

Table A15. Prediction errors with $R = 7$

N	R = 7	AIC	BIC	PCA	SAIC	SBIC	CV1	CV2
200	Mean	0.104	0.104	0.109	0.104	0.104	0.095	0.097
	Median	0.100	0.100	0.100	0.100	0.100	0.100	0.088
	Var	0.003	0.003	0.003	0.003	0.003	0.003	0.002
400	Mean	0.106	0.106	0.101	0.106	0.106	0.087	0.087
	Median	0.100	0.100	0.100	0.100	0.100	0.088	0.088
	Var	0.001	0.001	0.001	0.001	0.001	0.001	0.001
1000	Mean	0.109	0.109	0.103	0.109	0.109	0.084	0.083
	Median	0.110	0.110	0.105	0.110	0.110	0.085	0.080
	Var	0.001	0.001	0.000	0.001	0.001	0.000	0.000

References

Ando, Tomohiro, and Ker Chau Li. 2014. A model-averaging approach for high-dimensional regression. *Journal of the American Statistical Association* 109: 254–65. [CrossRef]

Ando, Tomohiro, and Ker Chau Li. 2017. A weight-relaxed model averaging approach for high-dimensional generalized linear models. *The Annals of Statistics* 45: 2654–79. [CrossRef]

Andrews, Donald W. K. 1991. Asymptotic optimality of generalized C_L, cross-validation, and generalized cross-validation in regression with heteroskedastic errors. *Journal of Econometrics* 47: 359–77. [CrossRef]

Balan, Raluca M., and Ioana Schiopu-Kratina. 2005. Asymptotic results with generalized estimating equations for longitudinal data. *The Annals of Statistics* 33: 522–41. [CrossRef]

Buckland, Steven T., Kenneth P. Burnham, and Nicole H. Augustin. 1997. Model selection: An integral part of inference. *Biometrics* 53: 603–18. [CrossRef]

Chen, Kani, Inchi Hu, and Zhiliang Ying. 1999. Strong consistency of maximum quasi-likelihood estimators in generalized linear models with fixed and adaptive designs. *The Annals of Statistics* 27: 1155–63. [CrossRef]

Claeskens, Gerda, and Raymond J. Carroll. 2007. An asymptotic theory for model selection inference in general semiparametric problems. *Biometrika* 94: 249–65. [CrossRef]

Flynn, Cheryl J., Clifford M. Hurvich, and Jeffrey S. Simonoff. 2013. Efficiency for regularization parameter selection in penalized likelihood estimation of misspecified models. *Journal of the American Statistical Association* 108: 1031–43. [CrossRef]

Gao, Yan, Xinyu Zhang, Shouyang Wang, and Guohua Zou. 2016. Model averaging based on leave-subject-out cross-validation. *Journal of Econometrics* 192: 139–51. [CrossRef]

Hansen, Bruce E. 2007. Least squares model averaging. *Econometrica* 75: 1175–89. [CrossRef]

Hansen, Bruce E., and Jeffrey S. Racine. 2012. Jacknife model averaging. *Journal of Econometrics* 167: 38–46. [CrossRef]

Hoffmann-Jørgensen, Jørgen. 1974. Sums of independent Banach space valued random variables. *Studia Mathematica* 52: 159–86. [CrossRef]

Hoffmann-Jørgensen, Jørgen, and Gilles Pisier. 1976. The law of large numbers and the central limit theorem in Banach spaces. *The Annals of Probability* 4: 587–99. [CrossRef]

Hjort, Nils L., and Gerda Claeskens. 2003. Frequentist model average estimators. *Journal of the American Statistical Association* 98: 879–99. [CrossRef]

James, Gareth M. 2002. Generalized linear models with functional predictors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64: 411–32. [CrossRef]

Kahane, Jean Pierrc. 1968. *Some Random Series of Functions*. Lexington: D. C. Heath.

Li, Ker Chau. 1987. Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set. *The Annals of Statistics* 15: 958–75. [CrossRef]

Liang, Hua, Guohua Zou, Alan T. K. Wan, and Xinyu Zhang. 2011. Optimal weight choice for frequentist model average estimators. *Journal of the American Statistical Association* 106: 1053–66. [CrossRef]

Lv, Jinchi, and Jun S. Liu. 2014. Model selection principles in misspecified models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76: 141–67. [CrossRef]

Müller, Hans Georg, and Ulrich Stadtmüller. 2005. Generalized functional linear models. *The Annals of Statistics* 33: 774–805. [CrossRef]

Wan, Alan T. K., Xinyu Zhang, and Guohua Zou. 2010. Least squares model averaging by Mallows criterion. *Journal of Econometrics* 156: 277–83. [CrossRef]

Wu, Chien-Fu. 1981. Asymptotic theory of nonlinear least squares estimation. *The Annals of Statistics* 9: 501–13. [CrossRef]

Xu, Ganggang, Suojin Wang, and Jianhua Z. Huang. 2014. Focused information criterion and model averaging based on weighted composite quantile regression. *Scandinavian Journal of Statistics* 41: 365–81. [CrossRef]

Yang, Yuhong. 2001. Adaptive regression by mixing. *Journal of the American Statistical Association* 96: 574–88. [CrossRef]

Yao, Fang, Müller Hans-Georg, and Wang Jane-Ling. 2005. Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association* 100: 577–90. [CrossRef]

Zhang, Xinyu, and Hua Liang. 2011. Focused information criterion and model averaging for generalized additive partial linear models. *The Annals of Statistics* 39: 174–200. [CrossRef]

Zhang, Xinyu, Alan T. K. Wan, and Sherry Z. Zhou. 2012. Focused information criteria model selection and model averaging in a Tobit model with a non-zero threshold. *Journal of Business and Economic Statistics* 30: 132–42. [CrossRef]

Zhang, Xinyu, Alan T. K. Wan, and Guohua Zou. 2013. Model averaging by jackknife criterion in models with dependent data. *Journal of Econometrics* 174: 82–94. [CrossRef]

Zhang, Xinyu. 2015. Consistency of model averaging estimators. *Economics Letters* 130: 120–23. [CrossRef]

Zhang, Xinyu, Dalei Yu, Guohua Zou, and Hua Liang. 2016. Optimal model averaging estimation for generalized linear models and generalized Linear mixed-effects models. *Journal of the American Statistical Association* 111: 1775–90. [CrossRef]

Zhang, Xinyu, Jeng-Min Chiou, and Yanyuan Ma. 2018. Functional prediction through averaging estimated functional linear regression models. *Biometrika* 105: 945–62. [CrossRef]

Zhao, Shangwei, Jun Liao, and Dalei Yu. 2018. Model averaging estimator in ridge regression and its large sample properties. *Statistical Papers*. [CrossRef]

Zhu, Rong, Guohua Zou, and Xinyu Zhang. 2018. Optimal model averaging estimation for partial functional linear models. *Journal of Systems Science and Mathematical Sciences* 38: 777–800.

Zinn, Joel. 1977. A note on the central limit theorem in Banach spaces. *The Annals of Probability* 5: 283–86. [CrossRef]