AStA Adv Stat Anal (2019) 103:475–501
https://doi.org/10.1007/s10182-018-00342-0

ORIGINAL PAPER

Estimation and variable selection for partial functional linear regression

Qingguo Tang[1] · Peng Jin[1]

Received: 5 May 2018 / Accepted: 22 November 2018 / Published online: 14 December 2018

Abstract We propose a new estimation procedure for estimating the unknown parameters and function in partial functional linear regression. The asymptotic distribution of the estimator of the vector of slope parameters is derived, and the global convergence rate of the estimator of unknown slope function is established under suitable norm. The convergence rate of the mean squared prediction error for the proposed estimators is also established. Based on the proposed estimation procedure, we further construct the penalized regression estimators and establish their variable selection consistency and oracle properties. Finite sample properties of our procedures are studied through Monte Carlo simulations. A real data example about the real estate data is used to illustrate our proposed methodology.

Keywords Partial functional linear regression · Functional principal component analysis · Variable selection · Asymptotic properties

1 Introduction

In the last two decades, there has been an increasing interest in regression models for functional variables as more and more data have arisen where the primary unit of observation can be viewed as a curve or in general a function, such as in biology, chemometrics, econometrics, geophysics, the medical sciences, meteorology and neurosciences. As a natural extension of the ordinary regression to the case where predictors include random functions and responses are scalars or functions, functional

✉ Qingguo Tang
tangqguo@163.com

1 School of Economics and Management, Nanjing University of Science and Technology, Nanjing 210094, China

linear regression analysis provides valuable insights into these problems. The effectively infinite-dimensional character of functional data analysis is a source of many of its differences from more conventional multivariate analysis. The functional linear model has been extensively studied and successfully applied; see Cardot et al. (2003), Ramsay and Silverman (2002, 2005), Cai and Hall (2006), Hall and Horowitz (2007), Reiss and Ogden (2010), Brunel and Roche (2015), Hsing and Eubank (2015), among many others.

It is frequently the case that a response is related to both a vector of finite length and a function-valued random variable as predictor variables. With a square integrable random function X on a compact set $\mathcal{T}$ in R and a d-dimensional vector of random variables $Z = (Z_1, \ldots, Z_d)^T$, we suppose that the scalar response Y is linearly related to predictor variables (X, Z) through the relationship

$$Y = \int_{\mathcal{T}} \gamma(t) X(t) \mathrm{d}t + Z^T \boldsymbol{\beta}_0 + \varepsilon, \tag{1.1}$$

where $\boldsymbol{\beta}_0$ is a $d \times 1$ vector of regression coefficients of Z, $\gamma(t)$ is a square integrable function on $\mathcal{T}$, and ε is a random error. Model (1.1) generalizes both the classical linear regression model and functional linear regression model which correspond to the cases $\gamma(t) = 0$ and $\boldsymbol{\beta}_0 = 0$, respectively. Moreover, this model includes the analysis of covariance model where the covariate is a random function, i.e., the model represents functional linear models between a scalar variable Y and a function-valued random variable X for each group simultaneously with the Z_k being scalar-valued indicator variables associated with subgroups. Zhang et al. (2007) proposed a two-stage functional mixed effects model to deal with measurement error and irregularly spaced time points and estimated the regression coefficient function using a two-stage nonparametric regression calibration method. Shin (2009) and Reiss and Ogden (2010) proposed the estimators of $\boldsymbol{\beta}_0$ and $\gamma(t)$ by generalizing the functional principal components estimation method in the functional linear regression and Shin and Lee (2012) considered a prediction of a scalar variable based on both a function-valued variable and a finite number of real-valued variables.

In this paper, we propose a new method for estimating the unknown parameters and function in model (1.1). Using functional principal component analysis, the unknown slope function is approximated by an average value which includes the unknown parameters. The estimators of the unknown parameters are obtained by solving a minimization problem. Although our method is obviously different from Shin (2009) and Shin and Lee (2012), we find that the estimators obtained by the two methods have the same behavior through simulation and further derivation. In fact, our estimators are more simple in expression and require less computation. Under conditions weaker than Shin (2009), we derive the asymptotic normality of the estimator of $\boldsymbol{\beta}_0$ and establish the global convergence rate of the estimator of the slope function $\gamma(t)$. Since our assumptions are weaker than that of Shin (2009), the asymptotic distribution of the estimator of $\boldsymbol{\beta}_0$ is different from that of Shin (2009) and Shin and Lee (2012). The proofs of our theorems are essentially different from Shin (2009). We establish the convergence rate of the mean squared prediction error for a predictor. Based on the proposed estimation procedure, we further propose a family of variable selection procedures via the penalized least squares using concave penalty functions. We show that the proposed penalized regression estimators have the variable selection consistency and oracle property of Fan and Li (2001).

Variable selection is particularly important when the true underlying model has a sparse representation. Identifying significant predictors will enhance the prediction performance of the fitted model. A penalty function generally facilitates variable selection in regression models. Various penalty functions have been used in the literature: the bridge regression (Frank and Friedman 1993), LASSO (Tibshirani 1996), SCAD (Fan and Li 2001), adaptive LASSO (Zou 2006), MCP (Zhang 2010), are well known. Liang and Li (2009) considered variable selection for partially linear models with measurement errors, Wang and Wang (2014) proposed adaptive Lasso estimators for ultrahigh-dimensional generalized linear models, and Aneirosa et al. (2015) investigated variable selection in partial linear regression with functional covariate. Fan et al. (2014) studied oracle optimality of folded concave penalized estimation.

The paper is organized as follows. Section 2 describes the estimation method and studies its asymptotic properties. Section 3 investigates an adaptive variable selection method and its asymptotic properties. Section 4 presents finite sample behaviors of the estimators. A real data example about the real estate data is given in Sect. 5. All proofs are relegated to "Appendix."

2 Estimation method and asymptotic results

Let Y be a real-valued random variable defined on a probability space $(\Omega, \mathcal{B}, P)$. Let Z be a d-dimensional vector of random variables with finite second moments, and let $\{X(t) : t \in \mathcal{T}\}$ be a zero-mean and second-order (i.e., $EX(t)^2 < \infty$ for all $t \in \mathcal{T}$) stochastic process defined on $(\Omega, \mathcal{B}, P)$ with sample paths in $L_2(\mathcal{T})$, the set of all square integrable functions on $\mathcal{T}$, where $\mathcal{T}$ is a bounded closed interval. ε is a random error with mean zero and is independent of (X, Z). Let $< \cdot, \cdot >$ and $\|\cdot\|$ represent, respectively, the $L_2(\mathcal{T})$ inner product and norm. Denote the covariance function of the process $X(t)$ by $K(s,t) = cov(X(s), X(t))$. We suppose that $K(s,t)$ is positive definite, in which case it admits a spectral decomposition in terms of strictly positive eigenvalues λ_j,

$$K(s,t) = \sum_{j=1}^{\infty} \lambda_j \phi_j(s)\phi_j(t), \quad s, t \in \mathcal{T}, \tag{2.1}$$

where (λ_j, ϕ_j) are (eigenvalue, eigenfunction) pairs for the linear operator with kernel K, the eigenvalues are ordered so that $\lambda_1 > \lambda_2 > \cdots$ and the functions $\phi_1, \phi_2, \ldots$ form an orthonormal basis for $L_2(\mathcal{T})$. This leads to the Karhunen–Loève representation

$$X(t) = \sum_{j=1}^{\infty} \xi_j \phi_j(t),$$

where the $\xi_j = \int_{\mathcal{T}} X(t)\phi_j(t)\mathrm{d}t$ are uncorrelated random variables with mean 0 and variance $E\xi_j^2 = \lambda_j$. Let $\gamma(t) = \sum_{j=1}^{\infty} \gamma_j \phi_j(t)$, then model (1.1) can be written as

$$Y = \sum_{j=1}^{\infty} \gamma_j \xi_j + Z^T \boldsymbol{\beta}_0 + \varepsilon. \tag{2.2}$$

By (2.2), we have

$$\gamma_j = E\{[Y - Z^T \boldsymbol{\beta}_0]\xi_j\}/\lambda_j. \tag{2.3}$$

Let $(X_i(t), Z_i, Y_i)$, $i = 1, \ldots, n$, be independent realizations of $(X(t), Z, Y)$ generated by the model (1.1). Empirical versions of K and of its spectral decomposition are

$$\hat{K}(s,t) = \frac{1}{n}\sum_{i=1}^{n} X_i(s)X_i(t) = \sum_{j=1}^{\infty} \hat{\lambda}_j \hat{\phi}_j(s)\hat{\phi}_j(t), \quad s, t \in \mathcal{T}.$$

Analogously to the case of K, $(\hat{\lambda}_j, \hat{\phi}_j)$ are (eigenvalue, eigenfunction) pairs for the linear operator with kernel $\hat{K}$, ordered such that $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq 0$. We take $(\hat{\lambda}_j, \hat{\phi}_j)$ and $\hat{\xi}_{ij} = \langle X_i, \hat{\phi}_j \rangle$ to be the estimators of (λ_j, ϕ_j) and $\xi_{ij} = \langle X_i, \phi_j \rangle$, respectively, and set

$$\tilde{\gamma}_j = \frac{1}{n\hat{\lambda}_j}\sum_{i=1}^{n} \left(Y_i - Z_i^T \boldsymbol{\beta}_0\right)\hat{\xi}_{ij}. \tag{2.4}$$

We use $\sum_{j=1}^{m} \tilde{\gamma}_j \hat{\xi}_j$ to approximate $\sum_{j=1}^{\infty} \gamma_j \xi_j$ in (2.2). Combining (2.2) and (2.4), we then solve the following minimization problem

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left\{ Y_i - \sum_{j=1}^{m} \frac{\hat{\xi}_{ij}}{n\hat{\lambda}_j} \sum_{l=1}^{n} \left(Y_l - Z_l^T \boldsymbol{\beta}\right)\hat{\xi}_{lj} - Z_i^T \boldsymbol{\beta} \right\}^2 \tag{2.5}$$

to obtain the estimator of $\boldsymbol{\beta}_0$. Define $\tilde{\xi}_{li} = \sum_{j=1}^{m} \frac{\hat{\xi}_{lj}\hat{\xi}_{ij}}{\hat{\lambda}_j}$, $\tilde{Y}_i = Y_i - \frac{1}{n}\sum_{l=1}^{n} Y_l \tilde{\xi}_{li}$ and $\tilde{Z}_i = Z_i - \frac{1}{n}\sum_{l=1}^{n} Z_l \tilde{\xi}_{li}$. Then, (2.5) can be written as

$$\min_{\beta} \sum_{i=1}^{n} \left(\tilde{Y}_i - \tilde{Z}_i^T \boldsymbol{\beta}\right)^2 \tag{2.6}$$

Let $\tilde{Y} = (\tilde{Y}_1, \ldots, \tilde{Y}_n)^T$ and $\tilde{Z} = (\tilde{Z}_1, \ldots, \tilde{Z}_n)^T$. Then the estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}_0$ is given by

$$\hat{\boldsymbol{\beta}} = (\tilde{Z}^T \tilde{Z})^{-1} \tilde{Z}^T \tilde{Y}. \tag{2.7}$$

The estimator of $\gamma(t)$ is given by $\hat{\gamma}(t) = \sum_{j=1}^{m} \hat{\gamma}_j \hat{\phi}_j(t)$ with

$$\hat{\gamma}_j = \frac{1}{n\hat{\lambda}_j} \sum_{i=1}^{n} \left(Y_i - Z_i^T \hat{\boldsymbol{\beta}} \right) \hat{\xi}_{ij}. \tag{2.8}$$

To implement our estimation method, we need to know how to choose m. The value for m can be selected by leave-one-curve-out cross-validation of the prediction error. Define CV function as

$$\mathrm{CV}(m) = \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{m} \hat{\gamma}_j^{-i} \hat{\xi}_{ij} - Z_i^T \hat{\boldsymbol{\beta}}^{-i} \right)^2,$$

where $\hat{\gamma}_j^{-i}$, $j = 1, \ldots, m$ and $\hat{\boldsymbol{\beta}}^{-i}$ are computed after removing (X_i, Z_i, Y_i). As an alternative to cross-validation, m can also be chosen by information criteria BIC. The BIC criteria as a function of m is given by

$$\mathrm{BIC}(m) = \log \left\{ \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{m} \hat{\gamma}_j \hat{\xi}_{ij} - Z_i^T \hat{\boldsymbol{\beta}} \right)^2 \right\} + \frac{\log n}{n}(m+1).$$

Large values of BIC indicate poor fits.

Remark 2.1 Noting that $\hat{\xi}_{ij} = \langle X_i, \hat{\phi}_j \rangle$, it can be easily shown that our estimators have the same performance as the estimators given in Shin (2009) and Shin and Lee (2012). However, our estimators are more simple in expression and require less computation.

In the following, we derive asymptotic normality of the estimator $\hat{\boldsymbol{\beta}}$ and the rate of convergence for the estimator $\hat{\gamma}(t)$. We make the following assumptions.

Assumption 1 X has finite fourth moment, in that $\int_{\mathcal{T}} E(X^4) < \infty$, and for each j, $E(\xi_j^4) < C_1 \lambda_j^2$ for some constant C_1.

Assumption 2 There exists a convex function φ defined on the interval [0, 1] such that $\varphi(0) = 0$ and $\lambda_j = \varphi(1/j)$ for $j \geq 1$.

Assumption 3 For Fourier coefficients γ_j, there exist constants $C_2 > 0$ and $\delta > 3/2$ such that $|\gamma_j| \leq C_2 j^{-\delta}$ for all $j \geq 1$.

Assumption 4 $m \to \infty$ and $n^{-1/2} m \lambda_m^{-1} \to 0$.

Assumption 5 $E(\|Z\|^4) < +\infty$.

Assumptions 1 and 3 are standard conditions for functional linear models; see, e.g., Cai and Hall (2006) and Hall and Horowitz (2007). Assumption 2 is slightly less restrictive than (3.2) of Hall and Horowitz (2007). Assumptions 4 can be easily verified and will be further discussed below.

Remark 2.2 Assumptions 2 and 4 are weaker than the assumptions for λ_j and m, respectively, in Shin (2009) and Shin and Lee (2012).

We first establish the asymptotic distribution of the estimator $\hat{\boldsymbol{\beta}}$. To derive the asymptotic normality of the estimator $\hat{\boldsymbol{\beta}}$, we need to adjust for the dependence of $Z = (Z_1, \ldots, Z_d)^T$ and $X(t)$, which is a common complication in semiparametric models. Let $\mathcal{G}$ denote the class of the random variables such that $G \in \mathcal{G}$ if $G = \sum_{j=1}^{\infty} g_j \xi_j$ and $|g_j| \le C_3 j^{-\delta}$ for all $j \ge 1$, where δ is defined in Assumption 3 and $C_3 > 0$ is a constant. Note that $\mathcal{G}$ is related to the first term on the right side of (2.2). Denote $G_r = \sum_{j=1}^{\infty} g_{rj} \xi_j$. Let

$$G_r^* = \operatorname{arginf}_{G_r \in \mathcal{G}} E\left[\left(Z_r - \sum_{j=1}^{\infty} g_{rj}\xi_j\right)^2\right].$$

Since

$$E\left[\left(Z_r - \sum_{j=1}^{\infty} g_{rj}\xi_j\right)^2\right] = E[(Z_r - E(Z_r|X))^2] + E\left[\left(E(Z_r|X) - \sum_{j=1}^{\infty} g_{rj}\xi_j\right)^2\right],$$

therefore,

$$G_r^* = \operatorname{arginf}_{G_r \in \mathcal{G}} E\left[\left(E(Z_r|X) - \sum_{j=1}^{\infty} g_{rj}\xi_j\right)^2\right].$$

Thus, G_r^* are the projections of $E(Z_r|X)$ onto the space $\mathcal{G}$. In other words, G_r^* is an element that belongs to $\mathcal{G}$ and it is the closest to $E(Z_r|X)$ among all the random variables in $\mathcal{G}$. Let $H_r = Z_r - G_r^*$ for $r = 1, \ldots, d$, and $H = (H_1, \ldots, H_d)^T$. We then have the following results.

Theorem 2.1 *Suppose that Assumptions 1–5 hold and $\Omega = E(HH^T)$ is invertible, then*

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \to_d N(0, \Omega^{-1}\sigma^2), \tag{2.9}$$

where $\to_d$ means convergence in distribution.

Remark 2.3 When the model is changed from functional linear model to partial functional linear model, to derive the asymptotic normality of the estimator $\hat{\boldsymbol{\beta}}$, it is key to handle the relation of the vector Z and $X(t)$. In our analysis, $Z_r, r = 1, \ldots, d$ are divided into two unrelated parts $G_r^* = \sum_{j=1}^{\infty} g_{rj}^* \xi_j$ and H_r. Consequently, (2.2) can be written as

$$Y = \sum_{j=1}^{\infty}\left(\gamma_j + \sum_{r=1}^{d} g_{rj}^* \beta_{0r}\right)\xi_j + H^T\boldsymbol{\beta}_0 + \varepsilon,$$

where $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0d})^T$. If $Z_r = \sum_{j=1}^{\infty} \tilde{g}_{rj}\xi_j + V_r$ and V_r is independent of $X(t)$, then $G_r^* = \sum_{j=1}^{\infty} \tilde{g}_{rj}\xi_j$ and $H_r = V_r$. If Z_r is independent of $X(t)$, then $G_r^* = 0$ and $H_r = Z_r$. If $E(Z_r|X(t)) = \sum_{j=1}^{\infty} \bar{g}_{rj}\xi_j$, then $G_r^* = \sum_{j=1}^{\infty} \bar{g}_{rj}\xi_j$ and $H_r = Z_r - G_r^*$. In Shin (2009) and Shin and Lee (2012), it is assumed that $E(Z_r|X(t)) = \sum_{j=1}^{\infty} \lambda_j^{-1} < K_{Z_k X}, \phi_j > \xi_j$, where $K_{Z_r X} = cov(Z_r, X)$ for $r = 1, \ldots, d$. In this case, $G_r^* = \sum_{j=1}^{\infty} \lambda_j^{-1} < K_{Z_k X}, \phi_j > \xi_j$ and $H_r = Z_r - G_r^*$, and the result of our Theorem 2.1 is the same as that of Theorem 3.1 in Shin (2009). Hence, Theorem 3.1 of Shin (2009) is a special case of our Theorem 2.1.

Next we establish the convergence rates of the estimators $\hat{\gamma}(t)$.

Theorem 2.2 *Assume that Assumptions 1–5 hold and that $n^{-1}m^2\lambda_m^{-1}\log m \to 0$. Then*

$$\int_{\mathcal{T}} \left\{\hat{\gamma}(t) - \gamma(t)\right\}^2 \mathrm{d}t = O_p\left(\frac{m}{n\lambda_m} + \frac{m}{n^2\lambda_m^2}\sum_{j=1}^{m}\frac{j^3\gamma_j^2}{\lambda_j^2} + \frac{1}{n\lambda_m}\sum_{j=1}^{m}\frac{\gamma_j^2}{\lambda_j} + m^{-2\delta+1}\right). \tag{2.10}$$

If $\lambda_j \sim j^{-\tau}, \tau > 1, m \sim n^{1/(\tau+2\delta)}, \delta > 2$ and $\delta > 1+\tau/2$, then $\sum_{j=1}^{m} j^3\gamma_j^2\lambda_j^{-2} \le C_4(\log m + m^{2\tau+4-2\delta})$ and $\sum_{j=1}^{m}\gamma_j^2\lambda_j^{-1} < +\infty$, where C_4 is a positive constant. We then have the following corollary.

Corollary 2.1 *Under Assumptions 1–5, if $\lambda_j \sim j^{-\tau}$, $\tau > 1$, $m \sim n^{1/(\tau+2\delta)}$ and $\delta > \min(2, 1+\tau/2)$, then it holds that*

$$\int_{\mathcal{T}} \left\{\hat{\gamma}(t) - \gamma(t)\right\}^2 \mathrm{d}t = O_p\left(n^{-(2\delta-1)/(\tau+2\delta)}\right). \tag{2.11}$$

The global convergence result (2.11) indicates that the estimator $\hat{\gamma}(t)$ attains the same convergence rate as those of the estimators of Hall and Horowitz (2007)*, which are optimal in the minimax sense.*

Let $\mathcal{S} = \{(Z_i, X_i, Y_i) : 1 \le i \le n\}$. In the following, for a new pair of predictor variables (Z_{n+1}, X_{n+1}) taking from the same population as the data and independent of the data, we shall derive the convergence rate of the mean squared prediction error (MSPE) given by

$$\begin{aligned}\mathrm{MSPE} = E\Bigg(\Bigg[\Bigg(&\int_{\mathcal{T}} \hat{\gamma}(t)X_{n+1}(t)\mathrm{d}t + Z_{n+1}^T\hat{\boldsymbol{\beta}}\Bigg) \\ &- \Bigg(\int_{\mathcal{T}} \gamma(t)X_{n+1}(t)\mathrm{d}t + Z_{n+1}^T\boldsymbol{\beta}_0\Bigg)\Bigg]^2 |\mathcal{S}\Bigg).\end{aligned}$$

Theorem 2.3 *Under Assumptions 1, 3 and 5, if $\lambda_j \sim j^{-\tau}$, $\tau > 1$, $m \sim n^{1/(\tau+2\delta)}$ and $\delta > \min(2, 1+\tau/2)$, then*

$$\mathrm{MSPE} = O_p(n^{-(\tau+2\delta-1)/(\tau+2\delta)}). \tag{2.12}$$

Remark 2.4 In practical application, $X(t)$ is only discretely observed. Without loss of generality, suppose $\mathcal{T} = [0, 1]$ and for each $i = 1, \ldots, n$, $X_i(t)$ is observed at n_i discrete points $0 = t_{i1} < \ldots < t_{in_i} = 1$. Typically, $\max_i \max_{1\le j\le n_i-1}(t_{i(j+1)} - t_{ij}) \to 0$ as $n \to \infty$ is also assumed. Based on the discrete observations, for each $i = 1, \ldots, n$, linear interpolation functions or spline interpolation functions can be used for the estimators of $X_i(t)$. For example, we can use the following linear interpolation function

$$\hat{X}_i(t) = X_i(t_{ij}) + \frac{(X_i(t_{i(j+1)}) - X_i(t_{ij}))}{t_{i(j+1)} - t_{ij}}(t - t_{ij}), \quad \text{for } t \in [t_{ij}, t_{i(j+1)}],\ j = 0, \ldots, n_i - 1$$

as the estimator of $X_i(t)$. It is necessary to point out that if $X_i(t), i = 1, \ldots, n$ are replaced by $\hat{X}_i(t), i = 1, \ldots, n$, the conclusions of Theorems 2.1–2.3 do not hold. We note that it is difficult to establish the related asymptotic properties by our current approach, and further research is expected.

3 Variable selection for partial functional linear model

In the variable selection problem, it is assumed that some components of $\boldsymbol{\beta}_0$ in model (1.1) are equal to zero. The goal is to identify and estimate the subset model. It has been argued that folded concave penalties are preferable to convex penalties such as the L_1-penalty in terms of both model-estimation accuracy and variable selection consistency (Lv and Fan 2009; Fan and Lv 2011). Let $p_{\nu_n}(|u|) = p_{a,\nu_n}(|u|)$ be general folded concave penalty functions defined on $u \in (-\infty, +\infty)$ satisfying

(a) The $p_{\nu_n}(u)$ are increasing and concave in $u \in [0, +\infty)$;
(b) The $p_{\nu_n}(u)$ are differentiable in $u \in (0, +\infty)$ with $p'_{\nu_n}(0) := p'_{\nu_n}(0+) \ge a_1\nu_n$, $p'_{\nu_n}(u) \ge a_1\nu_n$ for $u \in (0, a_2\nu_n]$, $p'_{\nu_n}(u) \le a_3\nu_n$ for $u \in [0, +\infty)$, and $p'_{\nu_n}(u) = 0$ for $u \in [a\nu_n, +\infty)$ with a prespecified constant $a > a_2$, where a_1, a_2 and a_3 are fixed positive constants.

The above family of general folded concave penalties contains several popular penalties including the SCAD penalty (Fan and Li 2001), the derivative of which is given by

$$p'_{\nu_n}(u) = \nu_n I_{\{u\le\nu_n\}} + \frac{(a\nu_n - u)_+}{a-1} I_{\{u>\nu_n\}} \quad \text{for some } a > 2,$$

and the MCP penalty (Zhang 2010), the derivative of which is given by

$$p'_{\nu_n}(u) = \left(\nu_n - \frac{u}{a}\right)_+ \quad \text{for some } a > 1.$$

It is easy to see that $a_1 = a_2 = a_3 = 1$ for the SCAD, and $a_1 = 1 - a^{-1}$, $a_2 = a_3 = 1$ for the MCP.

Based on the above analysis, we define a penalized least squares estimator of $\boldsymbol{\beta}_0$ as

$$\hat{\boldsymbol{\beta}}_{\text{PLS}} = \arg\min_{\boldsymbol{\beta}} (\tilde{Y} - \tilde{Z}\boldsymbol{\beta})^T (\tilde{Y} - \tilde{Z}\boldsymbol{\beta}) + n \sum_{k=1}^{d} p'_{\nu_n}(|\beta_k^{(0)}|)|\beta_k|, \tag{3.1}$$

where $\boldsymbol{\beta}^{(0)} = (\beta_1^{(0)}, \ldots, \beta_d^{(0)})^T$ is an initial estimator of $\boldsymbol{\beta}_0$. For example, $\boldsymbol{\beta}^{(0)}$ can be obtained from (2.7) in Sect. 2.

In the following, we show that the penalized least squares estimator defined by (3.1) has the oracle property (Fan and Li 2001). Without loss of generality, let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$, where $\boldsymbol{\beta}_1 \in \mathbf{R}^{d_1}$ and $\boldsymbol{\beta}_2 \in \mathbf{R}^{d-d_1}$. The vector of true parameters is denoted by $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{01}^T, \boldsymbol{\beta}_{02}^T)^T$ with each element of $\boldsymbol{\beta}_{01}$ being nonzero and $\boldsymbol{\beta}_{02} = 0$.

Theorem 3.1 *Suppose that the conditions of Theorem 2.1 hold. Let $p_{\nu_n}(\cdot)$ be general folded concave penalty functions satisfying assumptions (a) and (b) above and $\boldsymbol{\beta}^{(0)}$ be the estimator defined by (2.7). If $\nu_n \to 0$ and $\sqrt{n}\nu_n \to \infty$ as $n \to \infty$, then the penalized least squares estimator $\hat{\boldsymbol{\beta}}_{\text{PLS}} = (\hat{\boldsymbol{\beta}}_{\text{PLS1}}^T, \hat{\boldsymbol{\beta}}_{\text{PLS2}}^T)^T$ defined by (3.1) satisfies*

(1) Sparsity: $P(\hat{\boldsymbol{\beta}}_{\text{PLS2}} = 0) \to 1$.
(2) Asymptotic normality:

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\text{PLS1}} - \boldsymbol{\beta}_{01}) \to_d N(0, \Omega_1^{-1}\sigma^2), \tag{3.2}$$

where $\Omega_1 = E[(H_1, \ldots, H_{d_1})^T (H_1, \ldots, H_{d_1})]$.

Let

$$\hat{\gamma}_{\text{PLSj}} = \frac{1}{n\hat{\lambda}_j} \sum_{i=1}^{n} \left(Y_i - Z_i^T \hat{\boldsymbol{\beta}}_{\text{PLS}} \right) \hat{\xi}_{ij} \tag{3.3}$$

and $\hat{\gamma}_{\text{PLS}}(t) = \sum_{j=1}^{m} \hat{\gamma}_{PLSj} \hat{\phi}_j(t)$. We then have the following theorem.

Theorem 3.2 *(1) Under the assumptions of Theorems 3.1 and 2.2, the estimator $\hat{\gamma}_{\text{PLS}}(t)$ satisfies the conclusions of Theorem 2.2.*
(2) Under the assumptions of Theorems 3.1 and 2.3, the conclusions of Theorem 2.3 hold.

4 Simulation results

Since our estimators have the same performances as Shin (2009) and Shin and Lee (2012), in this section, we only investigate the finite sample performance of the penalized least squares estimators proposed in Sect. 3 by carrying out a Monte Carlo study. The data sets were generated from the following models

$$Y_i = \int_{\mathcal{T}} \gamma(t) X_i(t) \mathrm{d}t + Z_i^T \boldsymbol{\beta}_0 + \varepsilon_i, \tag{4.1}$$

Table 1 Results of Monte Carlo experiments for model (4.1)

	FP	FN	$\lvert\hat{\beta}_1 - \beta_1\rvert$	$\lvert\hat{\beta}_2 - \beta_2\rvert$	$\lvert\hat{\beta}_3 - \beta_3\rvert$
$n = 100$					
Setting 1	0.0200	0.0640	0.1288	0.1145	0.1012
Setting 2	0.0520	0.0780	0.1174	0.0890	0.0942
Setting 3	0.0360	0.0500	0.0924	0.0689	0.0693
Setting 4	0.0480	0.0580	0.1045	0.0742	0.0843
$n = 200$					
Setting 1	0.0040	0.0120	0.0952	0.0749	0.0744
Setting 2	0.0060	0.0280	0.0827	0.0719	0.0632
Setting 3	0.0040	0.0160	0.0637	0.0460	0.0478
Setting 4	0.0040	0.0140	0.0658	0.0529	0.0528

with $\mathcal{T} = [0, 1]$, $\boldsymbol{\beta}_0 = (2, 0, 1.5, 0, 0.3)^T$. We took $\gamma(t) = \sum_{j=1}^{50} \gamma_j \phi_j(t)$ and $X_i(t) = \sum_{j=1}^{50} \xi_{ij}\phi_j(t)$, where $\gamma_1 = 0.3$ and $\gamma_j = 4(-1)^{j+1} j^{-\delta}$, $j \geq 2$; $\phi_1(t) \equiv 1$ and $\phi_j(t) = 2^{1/2}\cos((j-1)\pi t)$, $j \geq 2$; the ξ_{ij}'s were independent and normal $N(0, \lambda_j)$. We let $Z_i = (Z_{i1}, \ldots, Z_{i5})^T$, when conditioning on ξ_{ij}, be a multivariate normal distribution with the mean vector $((1+\lambda_1)^{-1/2}\xi_{i1}, \ldots, (1+\lambda_5)^{-1/2}\xi_{i5})^T$ and the variance-covariance matrix $V = v_{kl}$ with $v_{kk} = (1+\lambda_k)^{-1}$ and $v_{kl} = 0.7((1+\lambda_k)(1+\lambda_l))^{-1/2}$ for $k, l = 1, \ldots, 5$, so that Z_i has a multivariate normal distribution with the zero-mean vector and the variance-covariance matrix whose diagonal elements are 1 and off-diagonal elements are v_{kl}. The errors ε_i were normally distributed with the mean 0 and the standard deviation 0.5. Similar to Shin and Lee (2012), we used 4 different sets of the eigenvalues, $\{\lambda_j\}$. In the two settings, $\lambda_j = j^{-\tau}$ and different values of τ are considered. In the other two settings, eigenvalues are "closely spaced" as in Hall and Horowitz (2007): $\lambda_1 = 1$, $\lambda_j = 0.2^2(1 - 0.0001j)^2$ if $2 \leq j \leq 4$, $\lambda_{5j+k} = 0.2^2\{(5j)^{-\tau/2} - 0.0001k\}^2$ for $j \geq 1$ and $0 \leq k \leq 4$.

1. Set $\tau = 1.1$ and $\delta = 2$ with the well-spaced eigenvalues.
2. Set $\tau = 1.1$ and $\delta = 2$ with the closely spaced eigenvalues.
3. Set $\tau = 3$ and $\delta = 2$ with the well-spaced eigenvalues.
4. Set $\tau = 3$ and $\delta = 2$ with the closely spaced eigenvalues.

All the results in this section are based on 500 replications. In all the simulated designs, we used the SCAD penalty function with $a = 3.7$. We set the sample size n to be 100 and 200, respectively. For each simulated data set, the penalized least squares estimators $\hat{\boldsymbol{\beta}}_{\text{PLS}}$ and $\hat{\gamma}_{\text{PLS}}(t)$ were computed by the procedure given in Sects. 2 and 3. The tuning parameter m is determined by BIC criterion as described in Sect. 2, and the tuning parameter ν_n in (3.1) is selected by the method given by Fan et al. (2014).

We measured the estimation accuracy for parametric estimators by the average l_1-losses: $|\hat{\beta}_1 - \beta_1|$, $|\hat{\beta}_3 - \beta_3|$, and $|\hat{\beta}_5 - \beta_5|$ over 500 replications. We also evaluated the selection accuracy by the average counts of false positive (FP) and false negative (FN) over the 500 replications; that is, the number of noise covariates included in the model and the number of signal covariates not included. Table 1 displays the simulation results

Table 2 Results of Monte Carlo experiments for model (4.1)

	$n = 100$			$n = 200$		
	Bias^2	Var	MISE	Bias^2	Var	MISE
Setting 1	0.1163	0.1608	0.2771	0.1100	0.0873	0.1973
Setting 2	0.1981	0.4146	0.6127	0.1901	0.3551	0.5452
Setting 3	0.1186	0.1185	0.2371	0.1215	0.0546	0.1760
Setting 4	0.1954	0.4159	0.6114	0.1817	0.3437	0.5254

Table 3 Results of Monte Carlo experiments under high-dimensional data

	FP	FN	$\|\hat{\beta}_1 - \beta_1\|$	$\|\hat{\beta}_2 - \beta_2\|$	$\|\hat{\beta}_3 - \beta_3\|$
$n = 100$					
Setting 1	0.0360	0.0840	0.1570	0.1217	0.1187
Setting 2	0.1560	0.1380	0.1336	0.1110	0.1166
Setting 3	0.0360	0.1580	0.1288	0.0912	0.1299
Setting 4	0.1060	0.0840	0.1134	0.0872	0.0975
$n = 200$					
Setting 1	0.0020	0.0140	0.0967	0.0795	0.0810
Setting 2	0.140	0.0440	0.0830	0.0769	0.0692
Setting 3	0.0020	0.0640	0.0875	0.0631	0.0911
Setting 4	0.0120	0.0140	0.0755	0.0557	0.0550

for model (4.1). We see from Table 1 that there is a general tendency for the average l_1-loss and PN and FN to decrease as n increases and there is a general tendency for the average l_1-loss to decrease as τ increases. Table 1 also shows that PNs and FNs for Settings 1 and 3 with the well-spaced eigenvalues are less than that for Settings 2 and 4 with the closely spaced eigenvalues, while PN and FN for the Setting 4 are less than that for the Setting 2.

Table 2 reports the integrated squared bias (Bias^2), integrated variance (Var) and mean integrated squared error (MISE) of the estimator $\hat{\gamma}(t)$ computed on a grid of 100 equally spaced points on $\mathcal{T}$. Table 2 shows that there is a general tendency for the MISE to decrease as τ increases. We also see from Table 2 that the MISEs for Settings 1 and 3 with the well-spaced eigenvalues are less than that for Settings 2 and 4 with the closely spaced eigenvalues.

In the following, we investigate the variable selection for high-dimensional data. In (4.1), let $Z_i = (Z_{i1}, \ldots, Z_{i30},)^T$, where $Z_{i1}, \ldots, Z_{i5}$ are taken the same as above, $Z_{i6}, \ldots, Z_{i30}$ are mutually independent and independent of $Z_{i1}, \ldots, Z_{i5}$ and $X_{ij} \sim N(0, 1)$ for $j = 6, \ldots, 30$, $\boldsymbol{\beta}_0 = (2, 0, 1.5, 0, 0.3, 0, \ldots, 0)^T$. The simulation results under this high-dimensional data are reported in Tables 3 and 4. We find that Tables 3 and 4 show conclusions similar to those in Tables 1 and 2. Comparing Table 3 with Table 1 and Table 4 with Table 2, we see that our penalized least squares estimators also behave well under the high-dimensional data.

Table 4 Results of Monte Carlo experiments under high-dimensional data

	$n = 100$			$n = 200$		
	Bias^2	Var	MISE	Bias^2	Var	MISE
Setting 1	0.1175	0.1676	0.2851	0.1112	0.0889	0.2001
Setting 2	0.2111	0.4712	0.6823	0.1956	0.3589	0.5545
Setting 3	0.1160	0.1337	0.2496	0.1196	0.0644	0.1839
Setting 4	0.1764	0.4358	0.6121	0.1868	0.3617	0.5485

5 A real data example

In this section, we analyze a real data set using the proposed methodology. For this purpose, we analyze the real estate data set which was collected from the statistical yearbooks of various cities, real estate market reports and statistical bulletins on national economic and social development in China. It includes the real estate data for 197 second-, third- and fourth-tier cities in China. In this data set, there are the average annual income of urban residents from 2000 to 2016, and the other data are based on 2016. Our purpose is to study the relationship between urban housing prices and their influencing factors. The response variable Y represents urban housing price. Since it takes many years of savings for the average resident to buy a house, we choose the average annual income of the residents as the functional covariate. Let $X_i^*(t)$ denote the average annual income of the residents of the ith city for the year t and $X_i(t) = X_i^*(t) - \bar{X}^*(t)$, where $\bar{X}^*(t) = \frac{1}{197}\sum_{i=1}^{197} X_i^*(t)$. The scalar covariates of primary interests include urban category (Z_2, Z_3), urban population (Z_4), urban GDP (Z_5), bank interest rate (Z_6), urban livability index (Z_7), urban comprehensive competitiveness (Z_8) and urban development index (Z_9). We note that among these variables the data of some variables such as Z_4 and Z_5 are very large, whereas those of some variables such as Z_6 are small. For this purpose, for each data of these variables, we first make the following modification: Let $\bar{z}_{i4}$, $i = 1, \ldots, 197$ be the observations of Z_4. Let $z_{i4} = \bar{z}_{i4}/\max \bar{z}_{i4}, i = 1, \ldots, 197$, so that the maximum of modified data of the variable Z_4 is 1. The data of the variables $Z_5, \ldots, Z_9$ are modified in a similar fashion. We construct the following partial functional linear model:

$$\log(Y_i) = \int_0^{17} \gamma(t)X_i(t)\mathrm{d}t + Z_{i1}\beta_{01} + \cdots + Z_{i9}\beta_{09} + \varepsilon_i, \tag{5.1}$$

where $Z_{i1} \equiv 1$, $Z_{i2} = 1$ and $Z_{i3} = 0$ stand for second-tier city, $Z_{i2} = 0$ and $Z_{i3} = 1$ stand for third-tier city, and $Z_{i2} = 0$ and $Z_{i3} = 0$ stand for fourth-tier city.

The estimators of unknown parameters and function in model (5.1) are computed by the method given in Sect. 2, and the tuning parameter m is determined by BIC criterion as described in Sect. 2. Table 5 exhibits the parametric estimators, and Fig. 1a shows the estimated curve of $\gamma(t)$ and its 95% confidence interval. We see from Table 5 that urban population, urban GDP, urban livability index , urban comprehensive competitiveness and urban development index have nonnegative effects, while bank interest rate has

Table 5 The parametric estimators for model (5.1)

β_{01}	β_{02}	β_{03}	β_{04}	β_{05}	β_{06}	β_{07}	β_{08}	β_{09}
8.2416	0.6821	0.2785	0.0102	0.2541	− 0.0006	0.0818	0.0036	0.2088

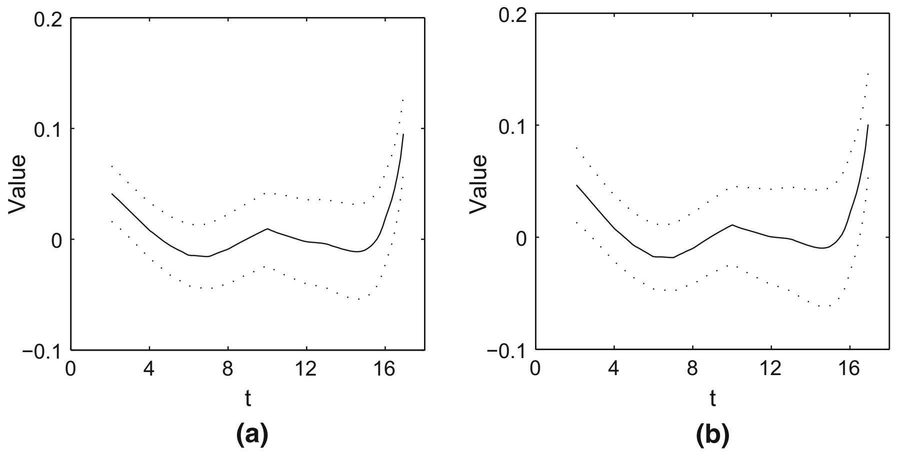

Fig. 1 The solid lines are the estimated curves of $\gamma(t)$, and the doted lines are their corresponding 95% point-wise confidence intervals. $\gamma(t)$ in (**a**) is computed by (2.8), and $\gamma(t)$ in (**b**) is computed by (3.3)

Table 6 The penalized least squares estimators of the parameters for model (5.1)

β_{01}	β_{02}	β_{03}	β_{04}	β_{05}	β_{06}	β_{07}	β_{08}	β_{09}
8.2424	0.6819	0.2785	0	0.2630	0	0.0804	0	0.2093

a negative effect. The fact that $\beta_{02} > \beta_{01} > 0$ in Table 5 indicates that the housing price for a third-tier city is larger than that for a fourth-tier city and the housing price for a second-tier city is larger than that for a third-tier city. We see from Fig. 1a that the estimated curve varies smoothly, but there is a rapid upward trend in the tail which shows that the effect of the average annual incomes of the residents on house prices varies greatly with different cities in recent years.

Table 6 exhibits the penalized least squares estimators of the parameters computed by the procedure given in Sect. 3, and Fig. 1b shows the estimated curve of $\gamma(t)$ computed by (3.3) and its 95% confidence interval. Table 6 shows that urban category, urban GDP, urban livability index and urban development index are important factors affecting house prices. Comparing Fig. 1b with Fig. 1a, we see that the difference between the two is not much.

To evaluate the prediction performance of our model and methods, we applied leave-one-out cross-validation to the data; i.e., when predicting the housing price for the ith city, we omit the data for this city when fitting the model. Figure 2 displays the boxplots for the absolute prediction errors $|\widehat{\log(y_j)} - \log(y_j)|,\ j = 1, \ldots, 197,$

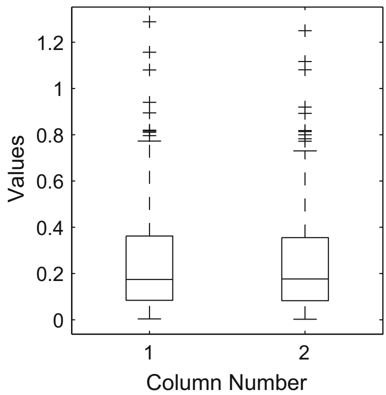

Fig. 2 Boxplots for the absolute prediction error $|\widehat{\log(y_j)} - \log(y_j)|$, $j = 1, \ldots, 197$, for two methods. Here 1 is the boxplot for the method given in Sect. 2 and 2 is the boxplot for the penalized method given in Sect. 3

for the method given in Sect. 2 and the penalized method given in Sect. 3. The mean values of these errors for the two methods are 0.2529 and 0.2521, respectively. These observations and Fig. 2 suggest that the penalized method is slightly better than the method given in Sect. 2.

Acknowledgements This work was supported by the National Social Science Foundation of China (16BTJ019), the Humanities and Social Science Foundation of Ministry of Education of China (14YJA910004) and Natural Science Foundation of Jiangsu Province of China (Grant No. BK20151481).

6 Appendix: Proofs

In this section, let $C > 0$ denote a generic constant of which the value may change from line to line. For a matrix $A = (a_{ij})$, set $\|A\|_\infty = \max_i \sum_j |a_{ij}|$ and $|A|_\infty = \max_{i,j} |a_{ij}|$. For a vector $v = (v_1, \ldots, v_k)^T$, set $\|v\|_\infty = \sum_{j=1}^k |v_j|$ and $|v|_\infty = \max_{1 \le j \le k} |v_j|$. Denote $W_l = \sum_{j=1}^\infty \gamma_j \xi_{lj}$, $\tilde{W}_i = W_i - \frac{1}{n}\sum_{l=1}^n W_l \tilde{\xi}_{li}$, $\tilde{\varepsilon}_i = \varepsilon_i - \frac{1}{n}\sum_{l=1}^n \varepsilon_l \tilde{\xi}_{li}$ and $\tilde{W} = (\tilde{W}_1, \ldots, \tilde{W}_n)^T$, $\tilde{\varepsilon} = (\tilde{\varepsilon}_1, \ldots, \tilde{\varepsilon}_n)^T$. Then

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = (\tilde{Z}^T \tilde{Z})^{-1} \tilde{Z}^T (\tilde{W} + \tilde{\varepsilon}). \tag{A.1}$$

Lemma A.1 *Suppose that Assumptions 1, 2, 4 and 5 hold, then it holds that*

$$\frac{1}{n}\tilde{Z}^T \tilde{Z} = \Omega + o_p(1).$$

Proof Let $\tilde{Z}_i = (\tilde{Z}_{i1}, \ldots, \tilde{Z}_{id})^T$. Set $\vec{\xi}_{li} = \sum_{j=1}^m \frac{\xi_{lj}\xi_{ij}}{\lambda_j}$, $\vec{Z}_{ir1} = Z_{ir} - \frac{1}{n}\sum_{l=1}^n Z_{lr}\vec{\xi}_{li}$ and $\vec{Z}_{ir2} = \frac{1}{n}\sum_{l=1}^n Z_{lr}(\tilde{\xi}_{li} - \vec{\xi}_{li})$. Then $\tilde{Z}_{ir} = \vec{Z}_{ir1} - \vec{Z}_{ir2}$ and

$$\frac{1}{n}\sum_{i=1}^{n}\tilde{Z}_{ir}\tilde{Z}_{iq} = \frac{1}{n}\sum_{i=1}^{n}(\vec{Z}_{ir1}\vec{Z}_{iq1} - \vec{Z}_{ir1}\vec{Z}_{iq2} - \vec{Z}_{ir2}\vec{Z}_{iq1} + \vec{Z}_{ir2}\vec{Z}_{iq2}), \quad r, q = 1, \ldots, d. \tag{A.2}$$

Let $\vec{Z}_{ir21} = \sum_{j=1}^{m} \frac{1}{\lambda_j}\left[\frac{1}{n}\sum_{l=1}^{n} Z_{lr}(\hat{\xi}_{lj} - \xi_{lj})\right]\xi_{ij}$, $\vec{Z}_{ir22} = \sum_{j=1}^{m}\left(\frac{1}{\hat{\lambda}_j} - \frac{1}{\lambda_j}\right)\left(\frac{1}{n}\sum_{l=1}^{n} Z_{lr}\hat{\xi}_{lj}\right)\xi_{ij}$ and $\vec{Z}_{ir23} = \sum_{j=1}^{m}\frac{1}{\hat{\lambda}_j}\left(\frac{1}{n}\sum_{l=1}^{n} Z_{lr}\hat{\xi}_{lj}\right)(\hat{\xi}_{ij} - \xi_{ij})$. We then have

$$|\vec{Z}_{ir2}\vec{Z}_{iq2}| \le \frac{3}{2}\left(\vec{Z}_{ir21}^2 + \vec{Z}_{ir22}^2 + \vec{Z}_{ir23}^2 + \vec{Z}_{iq21}^2 + \vec{Z}_{iq22}^2 + \vec{Z}_{iq23}^2\right). \tag{A.3}$$

Lemma 5.1 of Hall and Horowitz (2007) implies that

$$\hat{\xi}_{lj} - \xi_{lj} = \sum_{k\ne j}\frac{\xi_{lk}}{\hat{\lambda}_j - \lambda_k}\int \Delta\hat{\phi}_j\phi_k + \xi_{lj}\int(\hat{\phi}_j - \phi_j)\phi_j, \tag{A.4}$$

where $\Delta = \hat{K} - K$. We then obtain that

$$\begin{aligned}\left[\frac{1}{n}\sum_{l=1}^{n} Z_{lr}(\hat{\xi}_{lj} - \xi_{lj})\right]^2 &\le 2\left(\sum_{k\ne j}\frac{\vec{\xi}_{rk}}{\hat{\lambda}_j - \lambda_k}\int \Delta\hat{\phi}_j\phi_k\right)^2 + 2\left(\vec{\xi}_{rj}\int(\hat{\phi}_j - \phi_j)\phi_j\right)^2 \\ &\le 2\left[\sum_{k\ne j}\frac{\vec{\xi}_{rk}^2}{\left(\hat{\lambda}_j - \lambda_k\right)^2}\right]\left[\sum_{k=1}^{\infty}\left(\int \Delta\hat{\phi}_j\phi_k\right)^2\right] \\ &\quad + 2\vec{\xi}_{rj}^2\left(\int(\hat{\phi}_j - \phi_j)\phi_j\right)^2,\end{aligned}$$

where $\vec{\xi}_{rj} = \frac{1}{n}\sum_{l=1}^{n} Z_{lr}\xi_{lj}$. Lemma 1 of Cardot et al. (2007) implies that

$$|\lambda_j - \lambda_k| \ge \lambda_j - \lambda_{j+1} \ge \lambda_m - \lambda_{m+1} \ge \lambda_m/(m+1) \ge \lambda_m/(2m)$$

uniformly for $1 \le j \le m$. By (5.2) of Hall and Horowitz (2007), it holds that $\sup_{j\ge1}|\hat{\lambda}_j - \lambda_j| \le |\|\Delta\|| = O_p(n^{-1/2})$ and

$$\left(\int(\hat{\phi}_j - \phi_j)\phi_j\right)^2 \le \|\hat{\phi}_j - \phi_j\|^2 \le C\frac{|\|\Delta\||^2}{(\lambda_j - \lambda_{j+1})^2} \le C|\|\Delta\||^2\lambda_j^{-2}j^2, \tag{A.5}$$

where $|\|\Delta\|| = (\int_{\mathcal{T}}\int_{\mathcal{T}}\Delta^2(s,t)\mathrm{d}s\mathrm{d}t)^{1/2}$. Using Parseval's identity, we get that

$$\sum_{k=1}^{\infty}\left(\int \Delta\hat{\phi}_j\phi_k\right)^2 = \int\left(\int \Delta\hat{\phi}_j\right)^2 \le |\|\Delta\||^2 = O_p(n^{-1}).$$

Assumption 4 implies that $|\hat{\lambda}_j - \lambda_j| = o_p(\lambda_m/m)$. Consequently, $\sum_{k\neq j} \frac{\vec{\xi}_{rk}^2}{(\hat{\lambda}_j-\lambda_k)^2} = \sum_{k\neq j} \frac{\vec{\xi}_{rk}^2}{(\lambda_j-\lambda_k)^2}[1+o_p(1)]$, where $o_p(1)$ holds uniformly for $1 \leq j \leq m$. By arguments similar to those used in the proof of Lemma 2 of Cardot et al. (2007) and use the fact that $(\lambda_j - \lambda_k)^2 \geq (\lambda_k - \lambda_{k+1})^2$, we deduce that

$$\begin{aligned}\sum_{k\neq j} \frac{1}{(\lambda_j-\lambda_k)^2} E(\vec{\xi}_{rk}^2) &\leq C \sum_{k\neq j} \frac{1}{(\lambda_j-\lambda_k)^2} \left(n^{-1}\lambda_k + g_{rk}^2 \lambda_k^2\right) \\ &\leq C\left(n^{-1}\lambda_j^{-1} j^2 \log j + 1\right).\end{aligned}$$

Lemma 1 of Cardot et al. (2007) yields that

$$\sum_{j=1}^{m} \lambda_j^{-2} j^2 \log j \leq m^{-2}\lambda_m^{-2} \sum_{j=1}^{m} j^4 \log j \leq \lambda_m^{-2} m^3 \log m$$

and $\sum_{j=1}^m \lambda_j^{-1} \leq \lambda_m^{-1} m$. Therefore,

$$\sum_{j=1}^m \frac{1}{\lambda_j}\left[\frac{1}{n}\sum_{l=1}^n Z_{lr}(\hat{\xi}_{lj} - \xi_{lj})\right]^2 = O_p\left(n^{-2}\lambda_m^{-2} m^3 \log m + n^{-1}\lambda_m^{-1} m\right) \tag{A.6}$$

and

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} \vec{Z}_{ir21}^2 &\leq \left(\sum_{j=1}^{m} \frac{1}{\lambda_j}\left[\frac{1}{n}\sum_{l=1}^{n} Z_{lr}(\hat{\xi}_{lj}-\xi_{lj})\right]^2\right)\left(\sum_{j=1}^{m}\frac{1}{n\lambda_j}\sum_{i=1}^{n}\xi_{ij}^2\right) \\ &= O_p\left(n^{-2}\lambda_m^{-2} m^4 \log m + n^{-1}\lambda_m^{-1} m^2\right).\end{aligned} \tag{A.7}$$

Decomposing $\frac{1}{n}\sum_{l=1}^n Z_{lr}\hat{\xi}_{lj} = \vec{\xi}_{rj} + \frac{1}{n}\sum_{l=1}^n Z_{lr}(\hat{\xi}_{lj} - \xi_{lj})$ and using (A.6), we get

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} \vec{Z}_{ir22}^2 &\leq C \sum_{j=1}^{m} \frac{(\hat{\lambda}_j-\lambda_j)^2}{\lambda_j^3}\left(\frac{1}{n}\sum_{l=1}^{n} Z_{lr}\hat{\xi}_{lj}\right)^2 [1+o_p(1)]\left(\sum_{j=1}^{m}\frac{1}{n\lambda_j}\sum_{i=1}^{n}\xi_{ij}^2\right) \\ &= O_p\left(n^{-1}\lambda_m^{-1} m + n^{-3}\lambda_m^{-4} m^4 \log m + n^{-2}\lambda_m^{-3} m^2\right).\end{aligned} \tag{A.8}$$

By (A.10) of Tang (2015), it holds that

$$n\|\hat{\phi}_j - \phi_j\|^2/(j^2\log j) = O_p(1), \tag{A.9}$$

where $O_p(\cdot)$ holds uniformly for $1 \leq j \leq m$. Using (A.8) and (A.9), we obtain

$$\frac{1}{n}\sum_{i=1}^{n}\vec{Z}_{ir23}^{2} \le \left(\sum_{j=1}^{m}\frac{1}{\hat{\lambda}^{2}}\left(\frac{1}{n}\sum_{l=1}^{n}Z_{lr}\hat{\xi}_{lj}\right)^{2}\right)\left(\frac{1}{n}\sum_{i=1}^{n}\|X_i\|^2\right)\left(\sum_{j=1}^{m}\|\hat{\phi}_j-\phi_j\|^2\right)$$
$$= O_p\left(\left(n^{-2}\lambda_m^{-1}m^4 + n^{-1}m^3 + n^{-3}\lambda_m^{-3}m^6\log m + n^{-2}\lambda_m^{-2}m^4\right)\log m\right). \quad \text{(A.10)}$$

Hence, by (A.3), (A.7), (A.8), and (A.10) and Assumption 4, we conclude that

$$\frac{1}{n}\sum_{i=1}^{n}|\vec{Z}_{ir2}\vec{Z}_{iq2}| = O_p\left(n^{-2}\lambda_m^{-2}m^4\log m + n^{-1}\lambda_m^{-1}m^2\right) = o_p(1). \quad \text{(A.11)}$$

Define $\check{\xi}_{jr} = \frac{1}{n}\sum_{l=1}^{n}\lambda_j^{-1/2}\xi_{lj}Z_{lr}$. Since $E[\max_{1\le j\le m}(\check{\xi}_{jr} - E(\check{\xi}_{jr}))^2] \le \frac{1}{n}\sum_{j=1}^{m}\lambda_j^{-1}E(\xi_j Z_r)^2 \le Cn^{-1}$, we then have $\max_{1\le j\le m}|\check{\xi}_{jr} - E(\check{\xi}_{jr})| = O_p(n^{-1/2})$. Hence

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\vec{Z}_{ir1}\vec{Z}_{iq1}\\
&= \frac{1}{n}\sum_{i=1}^{n}Z_{ir}Z_{iq} - 2\sum_{j=1}^{m}\check{\xi}_{jr}\check{\xi}_{jq} + \sum_{j=1}^{m}\frac{\check{\xi}_{jr}\check{\xi}_{jq}}{n\lambda_j}\left(\sum_{i=1}^{n}\xi_{ij}^2\right) + \sum_{j\ne j'}\check{\xi}_{jr}\check{\xi}_{j'q}\bar{\xi}_{jj'}\\
&= \sum_{j=1}^{\infty}g_{rj}g_{qj}\lambda_j + E(H_rH_q) - 2\sum_{j=1}^{m}g_{rj}g_{qj}\lambda_j + \sum_{j=1}^{m}g_{rj}g_{qj}\lambda_j + o_p(1)\\
&= E(H_rH_q) + o_p(1),
\end{aligned} \quad \text{(A.12)}$$

where $\bar{\xi}_{jj'} = \frac{1}{n(\lambda_j\lambda_{j'})^{1/2}}\sum_{i=1}^{n}\xi_{ij}\xi_{ij'}$. Now Lemma A.1 follows from (A.2), (A.11), (A.12) and the fact that $\frac{1}{n}|\sum_{i=1}^{n}\vec{Z}_{ir1}\vec{Z}_{iq2}| \le \left(\frac{1}{n}\sum_{i=1}^{n}\vec{Z}_{ir1}^2\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n}\vec{Z}_{iq2}^2\right)^{1/2}$. □

Lemma A.2 *Under Assumptions 1–4, it holds that*

$$\sum_{j=1}^{m}\lambda_j\left[\gamma_j - \frac{1}{\hat{\lambda}_j}\left(\frac{1}{n}\sum_{l=1}^{n}W_l\hat{\xi}_{lj}\right)\right]^2 = O_p\left(n^{-1}\lambda_m^{-1}m\right).$$

Proof Set $S_1 = \sum_{j=1}^{m}\lambda_j\left[\gamma_j - \frac{1}{\lambda_j}\left(\frac{1}{n}\sum_{l=1}^{n}W_l\xi_{lj}\right)\right]^2$, $S_2 = \sum_{j=1}^{m}\frac{1}{\lambda_j}\left[\frac{1}{n}\sum_{l=1}^{n}W_l(\hat{\xi}_{lj}-\xi_{lj})\right]^2$ and $S_3 = \sum_{j=1}^{m}\lambda_j\left(\frac{1}{\hat{\lambda}_j}-\frac{1}{\lambda_j}\right)^2\left(\frac{1}{n}\sum_{l=1}^{n}W_l\hat{\xi}_{lj}\right)^2$. We have

$$\sum_{j=1}^{m} \lambda_j \left[\gamma_j - \frac{1}{\hat{\lambda}_j} \left(\frac{1}{n} \sum_{l=1}^{n} W_l \hat{\xi}_{lj} \right) \right]^2 \leq 3(S_1 + S_2 + S_3). \tag{A.13}$$

Since $E\left[\gamma_j - \frac{1}{\lambda_j}\left(\frac{1}{n}\sum_{l=1}^{n} W_l \xi_{lj}\right)\right] = 0$, then by Assumptions 1–3, we obtain that

$$E(S_1) = \sum_{j=1}^{m} \frac{1}{\lambda_j} Var\left(\frac{1}{n}\sum_{l=1}^{n} W_l \xi_{lj}\right) \leq \sum_{j=1}^{m} \frac{1}{n^2 \lambda_j} \sum_{l=1}^{n} E\left(W_l^2 \xi_{lj}^2\right) \leq Cm/n. \tag{A.14}$$

Similar to the proof of (A.6) and (A.8) and using Assumption 4, we deduce that

$$S_2 = O_p\left(n^{-2}\lambda_m^{-2} m^3 \log m + n^{-1}\lambda_m^{-1} m\right) = O_p(n^{-1}\lambda_m^{-1} m) \tag{A.15}$$

and

$$\begin{aligned} S_3 &\leq C \sum_{j=1}^{m} \frac{(\hat{\lambda}_j - \lambda_j)^2}{\lambda_j^3} \left(\bar{\zeta}_j^2 + \left[\frac{1}{n} \sum_{l=1}^{n} \zeta_l (\hat{\xi}_{lj} - \xi_{lj}) \right]^2 \right) [1 + o_p(1)] \\ &= O_p\left(n^{-1}\lambda_m^{-1} + n^{-3}\lambda_m^{-4} m^3 \log m + n^{-2}\lambda_m^{-3} m) = O_p(n^{-1}\lambda_m^{-1}\right). \end{aligned} \tag{A.16}$$

Now Lemma A.2 follows from (A.13)–(A.16). □

Lemma A.3 *Under Assumptions 1, 2, 4 and 5, it holds that*

$$\sum_{j=1}^{m} \lambda_j^{-1} \left(\sum_{i=1}^{n} \xi_{ij} \tilde{Z}_{ir} \right)^2 = O_p(nm + \lambda_m^{-2} m^4 \log m).$$

Proof Let $Z_{ir}^* = Z_{ir} - \sum_{j'=1}^{m} \frac{1}{\lambda_{j'}} \left(\frac{1}{n} \sum_{l=1}^{n} Z_{lr} \xi_{lj'} \right) \xi_{ij'}$. Observe that

$$\begin{aligned} \left(\sum_{i=1}^{n} \xi_{ij} \tilde{Z}_{ir} \right)^2 \leq\ & 4 \left(\sum_{i=1}^{n} \xi_{ij} Z_{ir}^* \right)^2 \\ & + 4 \left(\sum_{i=1}^{n} \xi_{ij} \sum_{j'=1}^{m} \frac{1}{\lambda_{j'}} \left[\frac{1}{n} \sum_{l=1}^{n} Z_{lr} \left(\hat{\xi}_{lj'} - \xi_{lj'} \right) \right] \xi_{ij'} \right)^2 \\ & + 4 \left(\sum_{i=1}^{n} \xi_{ij} \sum_{j'=1}^{m} \left(\frac{1}{\hat{\lambda}_{j'}} - \frac{1}{\lambda_{j'}} \right) \left[\frac{1}{n} \sum_{l=1}^{n} Z_{lr} \hat{\xi}_{lj'} \right] \xi_{ij'} \right)^2 \\ & + 4 \left(\sum_{i=1}^{n} \xi_{ij} \sum_{j'=1}^{m} \frac{1}{\hat{\lambda}_{j'}} \left[\frac{1}{n} \sum_{l=1}^{n} Z_{lr} \hat{\xi}_{lj'} \right] (\hat{\xi}_{ij'} - \xi_{ij'}) \right)^2 \end{aligned}$$

$$=: 4(T_{j1} + T_{j2} + T_{j3} + T_{j4}). \tag{A.17}$$

By direct computation and using Assumption 1, we get

$$E\left(\xi_{ij}^2 Z_{ir}^{*\,2}\right) \le 2E\left(\xi_{ij}^2 Z_{ir}^2\right) + 2E\left[\xi_{ij}^2\left(\sum_{j'=1}^{m}\frac{1}{\lambda_{j'}}\left(\frac{1}{n}\sum_{l=1}^{n} Z_{lr}\xi_{lj'}\right)\xi_{ij'}\right)^2\right]$$
$$\le C\left(\lambda_j + m\lambda_j/n^2 + (n-1)m\lambda_j/n^2 + m^2\lambda_j/n^2\right) \le C\lambda_j$$

and

$$\left|\sum_{i_1 \ne i_2} E\left(\xi_{i_1 j}\xi_{i_2 j} Z_{i_1 r}^* Z_{i_2 r}^*\right)\right| \le C[(n-1)(n+2)\lambda_j/n + (n-1)m\lambda_j/n] \le Cn\lambda_j.$$

Hence

$$E(T_{j1}) = \sum_{i=1}^{n} E\left(\xi_{ij}^2 Z_{ir}^{*\,2}\right) + \sum_{i_1 \ne i_2} E\left(\xi_{i_1 j}\xi_{i_2 j} Z_{i_1 r}^* Z_{i_2 r}^*\right) \le Cn\lambda_j. \tag{A.18}$$

Since $\sum_{j'=1}^{m}\frac{1}{\lambda_{j'}} E\left(\sum_{i=1}^{n}\xi_{ij}\xi_{ij'}\right)^2 \le Cn^2\lambda_j$, then by (A.6), we have

$$\sum_{j=1}^{m}\lambda_j^{-1}T_{j2} \le \left(\sum_{j'=1}^{m}\frac{1}{\lambda_{j'}}\left[\frac{1}{n}\sum_{l=1}^{n} Z_{lr}(\hat{\xi}_{lj'} - \xi_{lj'})\right]^2\right)$$
$$\times\left(\sum_{j=1}^{m}\lambda_j^{-1}\sum_{j'=1}^{m}\frac{1}{\lambda_{j'}}\left(\sum_{i=1}^{n}\xi_{ij}\xi_{ij'}\right)^2\right)$$
$$= O_p(n^{-2}\lambda_m^{-2}m^3\log m)O_p(n^2 m) = O_p(\lambda_m^{-2}m^4\log m). \tag{A.19}$$

Similar to the proof (A.8) and using Assumption 4, we deduce that

$$\sum_{j=1}^{m}\lambda_j^{-1}T_{j3} \le \left(\sum_{j'=1}^{m}\lambda_{j'}\left(\frac{1}{\hat{\lambda}_{j'}} - \frac{1}{\lambda_{j'}}\right)^2\left[\frac{1}{n}\sum_{l=1}^{n} Z_{lr}\hat{\xi}_{lj'}\right]^2\right)$$
$$\times\left(\sum_{j=1}^{m}\lambda_j^{-1}\sum_{j'=1}^{m}\frac{1}{\lambda_{j'}}\left(\sum_{i=1}^{n}\xi_{ij}\xi_{ij'}\right)^2\right)$$
$$= O_p\left(\lambda_m^{-2}m^2 + n^{-1}\lambda_m^{-4}m^4\log m\right) = o_p\left(\lambda_m^{-2}m^2\log m\right). \tag{A.20}$$

and

$$\sum_{j=1}^{m} \lambda_j^{-1} T_{j4} \le \left(\sum_{j'=1}^{m} \frac{1}{\lambda_{j'}^2} \left[\frac{1}{n} \sum_{l=1}^{n} Z_{lr} \hat{\xi}_{lj'} \right]^2 \right) [1 + o_p(1)]$$

$$\times \left(\sum_{j=1}^{m} \frac{1}{\lambda_j} \sum_{i=1}^{n} \xi_{ij}^2 \right) \left(\sum_{j'=1}^{m} \sum_{i=1}^{n} (\hat{\xi}_{ij'} - \xi_{ij'})^2 \right)$$

$$= O_p \left(n^{-1} \lambda_m^{-1} m^5 \log m + n^{-2} \lambda_m^{-3} m^7 (\log m)^2 \right) = o_p \left(\lambda_m^{-2} m^4 \log m \right). \tag{A.21}$$

Now Lemma A.3 follows from (A.17)–(A.21) and Assumption 4. □

Lemma A.4 *Under Assumptions 1–5, it holds that*

$$n^{-1/2} \left| \sum_{j=1}^{m} \frac{1}{\hat{\lambda}_j} \left(\frac{1}{n} \sum_{l=1}^{n} W_l \hat{\xi}_{lj} \right) \sum_{i=1}^{n} (\hat{\xi}_{ij} - \xi_{ij}) \tilde{Z}_{ir} \right| = o_p(1).$$

Proof Let $\breve{W}_j = \frac{1}{n} \sum_{l=1}^{n} W_l \hat{\xi}_{lj}$. Applying the Cauchy–Schwarz inequality, we get

$$\left(\sum_{j=1}^{m} \frac{1}{\hat{\lambda}_j} \breve{W}_j \sum_{i=1}^{n} (\hat{\xi}_{ij} - \xi_{ij}) \tilde{Z}_{ir} \right)^2 \le \left(\sum_{j=1}^{m} \frac{1}{\hat{\lambda}_j^2} \breve{W}_j^2 \right) \left(\sum_{j=1}^{m} \left(\sum_{i=1}^{n} (\hat{\xi}_{ij} - \xi_{ij}) \tilde{Z}_{ir} \right)^2 \right).$$

Using (A.4), (A.5), Assumption 4 and Parseval's identity and the arguments similar to those used to prove Lemma A.3, we deduce that

$$\sum_{j=1}^{m} \left(\sum_{i=1}^{n} (\hat{\xi}_{ij} - \xi_{ij}) \tilde{Z}_{ir} \right)^2$$

$$\le 2 \sum_{j=1}^{m} \left[\left(\sum_{k \ne j} (\hat{\lambda}_j - \lambda_k)^{-1} \int \Delta \hat{\phi}_j \phi_k \sum_{i=1}^{n} \xi_{ik} \tilde{Z}_{ir} \right)^2 \right.$$

$$\left. + \left(\sum_{i=1}^{n} \xi_{ij} \tilde{Z}_{ir} \right)^2 \left(\int (\hat{\phi}_j - \phi_j) \phi_j \right)^2 \right]$$

$$\le C \|\Delta\|^2 \sum_{j=1}^{m} \left[\sum_{k \ne j} \left(\hat{\lambda}_j - \lambda_k \right)^{-2} \left(\sum_{i=1}^{n} \xi_{ik} \tilde{Z}_{ir} \right)^2 + \lambda_j^{-2} j^2 \left(\sum_{i=1}^{n} \xi_{ij} \tilde{Z}_{ir} \right)^2 \right]$$

$$= O_p \left(\lambda_m^{-1} m^3 \log m + n^{-1} \lambda_m^{-3} m^6 \log m \right) = o_p(n).$$

Let $\vec{W}_j = \frac{1}{n}\sum_{l=1}^{n} W_l \xi_{lj}$. Decomposing $\frac{1}{n}\sum_{l=1}^{n} W_l \hat{\xi}_{lj} = \vec{W}_j + \frac{1}{n}\sum_{l=1}^{n} W_l(\hat{\xi}_{lj} - \xi_{lj})$ and using arguments similar to those used in the proof of (A.6) and using Assumption 4 , we obtain that

$$\sum_{j=1}^{m} \frac{1}{\hat{\lambda}_j^2} \breve{W}_j^2 = O_p(n^{-1}\lambda_m^{-1} m + 1 + n^{-2}\lambda_m^{-3} m^3 \log m + n^{-1}\lambda_m^{-2} m) = O_p(1).$$

This finishes the proof of Lemma A.4. □

Lemma A.5 *Under Assumptions 1–5, it holds that*

$$n^{-1/2}\left|\sum_{i=1}^{n} \tilde{W}_i \tilde{Z}_{ir}\right| = o_p(1).$$

Proof Observe that

$$\begin{aligned}
\sum_{i=1}^{n} \tilde{W}_i \tilde{Z}_{ir} &= \sum_{j=1}^{m}\left[\gamma_j - \frac{1}{\hat{\lambda}_j}\left(\frac{1}{n}\sum_{l=1}^{n} W_l \hat{\xi}_{lj}\right)\right]\sum_{i=1}^{n} \xi_{ij}\tilde{Z}_{ir} \\
&\quad - \sum_{j=1}^{m} \frac{1}{\hat{\lambda}_j}\left(\frac{1}{n}\sum_{l=1}^{n} W_l \hat{\xi}_{lj}\right)\sum_{i=1}^{n}\left(\hat{\xi}_{ij} - \xi_{ij}\right)\tilde{Z}_{ir} \\
&\quad + \sum_{j=m+1}^{\infty} \gamma_j \sum_{i=1}^{n} \xi_{ij}\tilde{Z}_{ir}
\end{aligned} \tag{A.22}$$

Lemmas A.2 and A.3 and Assumption 4 imply that

$$\begin{aligned}
& n^{-\frac{1}{2}}\left|\sum_{j=1}^{m}\left[\gamma_j - \frac{1}{\hat{\lambda}_j}\left(\frac{1}{n}\sum_{l=1}^{n} W_l \hat{\xi}_{lj}\right)\right]\sum_{i=1}^{n} \xi_{ij}\tilde{Z}_{ir}\right| \\
&\le n^{-\frac{1}{2}}\left(\sum_{j=1}^{m} \lambda_j\left[\gamma_j - \frac{1}{\hat{\lambda}_j}\left(\frac{1}{n}\sum_{l=1}^{n} W_l \hat{\xi}_{lj}\right)\right]^2\right)^{\frac{1}{2}}\left(\sum_{j=1}^{m} \lambda_j^{-1}(\sum_{i=1}^{n} \xi_{ij}\tilde{Z}_{ir})^2\right)^{\frac{1}{2}} \\
&= O_p\left(n^{-1/2}\lambda_m^{-1/2} m + n^{-1}\lambda_m^{-3/2} m^{5/2}(\log m)^{1/2}\right) = o_p(1).
\end{aligned} \tag{A.23}$$

By arguments similar to those used in the proof of Lemma A.3, we obtain that

$$\begin{aligned}
&\left(\sum_{j=m+1}^{\infty} \gamma_j \sum_{i=1}^{n} \xi_{ij}\tilde{Z}_{ir}\right)^2 \\
&\le \left(\sum_{j=m+1}^{\infty} \gamma_j^2\right)\left(\sum_{j=m+1}^{\infty}\left(\sum_{i=1}^{n} \xi_{ij}\tilde{Z}_{ir}\right)^2\right)
\end{aligned}$$

$$= O_p\left(nm^{-2\gamma+1} + \lambda_m^{-2} m^{-2\gamma+4} \log m\right) \sum_{j=m+1}^{\infty} \lambda_j$$
$$= o_p(n). \quad \text{(A.24)}$$

Now Lemma A.5 follows from (A.22)–(A.24) and Lemma A.4. □

Proof of Theorem 2.1 By arguments similar to those used to prove Lemmas A.4 and A.5, we deduce that $n^{-1/2} \sum_{i=1}^{n} \left(\frac{1}{n} \sum_{l=1}^{n} \varepsilon_l \tilde{\xi}_{li}\right) \tilde{Z}_{ir} = o_p(1)$. Hence

$$n^{-\frac{1}{2}} \sum_{i=1}^{n} \tilde{Z}_{ir} \tilde{\varepsilon}_i = n^{-\frac{1}{2}} \sum_{i=1}^{n} \tilde{Z}_{ir} \varepsilon_i + o_p(1).$$

We decompose $\sum_{i=1}^{n} \tilde{Z}_{ir} \varepsilon_i$ into three terms as

$$\sum_{i=1}^{n} \tilde{Z}_{ir} \varepsilon_i = \sum_{i=1}^{n} \varepsilon_i \left(Z_{ir} - \sum_{j=1}^{m} \frac{E(Z_{lr}\xi_j)}{\lambda_j} \xi_{ij}\right) - \sum_{i=1}^{n} \varepsilon_i \sum_{j=1}^{m} \frac{\xi_{ij}}{\lambda_j}$$
$$\left(\frac{1}{n} \sum_{l=1}^{n} Z_{lr} \xi_{lj} - E(Z_{lr}\xi_j)\right) - \sum_{i=1}^{n} \varepsilon_i \frac{1}{n} \sum_{l=1}^{n} Z_{lr} (\tilde{\xi}_{li} - \vec{\xi}_{li}).$$

Similar to the proof of Lemma A.4, we have $\sum_{i=1}^{n} \varepsilon_i \frac{1}{n} \sum_{l=1}^{n} Z_{lr} (\tilde{\xi}_{li} - \vec{\xi}_{li}) = o_p(n)$. Since

$$\sum_{i=1}^{n} \varepsilon_i \left(Z_{ir} - \sum_{j=1}^{m} \frac{E(Z_{lr}\xi_j)}{\lambda_j} \xi_{ij}\right) = \sum_{i=1}^{n} \varepsilon_i H_{ir} + \sum_{i=1}^{n} \varepsilon_i \sum_{j=m+1}^{\infty} g_{rj} \xi_{ij},$$

$\sum_{i=1}^{n} \varepsilon_i \sum_{j=1}^{m} \frac{\xi_{ij}}{\lambda_j} \left(\frac{1}{n} \sum_{l=1}^{n} Z_{lr} \xi_{lj} - E(Z_{lr}\xi_j)\right) = o_p(n)$ and $\sum_{i=1}^{n} \varepsilon_i \sum_{j=m+1}^{\infty} g_{kj} \xi_{ij} = o_p(n)$, it follows that

$$n^{-\frac{1}{2}} \sum_{i=1}^{n} \tilde{Z}_{ir} \tilde{\varepsilon}_i = n^{-\frac{1}{2}} \sum_{i=1}^{n} H_{ir} \varepsilon_i + o_p(1). \quad \text{(A.25)}$$

Now (2.9) follows from (A.1), Lemmas A.1 and A.5, (A.25) and the central limit theorem. The proof of Theorem 2.1 is finished. □

Lemma A.6 *Define* $\check{\gamma}_j = \frac{1}{\hat{\lambda}_j} E[(Y - Z^T \boldsymbol{\beta}_0)\xi_j]$. *Under the assumptions of Theorem 3.2, it holds that*

$$\sum_{j=1}^{m} (\hat{\gamma}_j - \check{\gamma}_j)^2 = O_p\left(n^{-1} m \lambda_m^{-1} + n^{-2} m \lambda_m^{-2} \sum_{j=1}^{m} \gamma_j^2 \lambda_j^{-2} j^3\right).$$

Proof Define $I_1 = \frac{1}{n}\sum_{i=1}^n (Y_i - Z_i^T \boldsymbol{\beta}_0)\xi_{ij} - \gamma_j\lambda_j$, $I_2 = \frac{1}{n}\sum_{i=1}^n (Y_i - Z_i^T \boldsymbol{\beta}_0)(\hat{\xi}_{ij} - \xi_{ij})$ and $I_3 = \frac{1}{n}\sum_{i=1}^n Z_i^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\hat{\xi}_{ij}$. Noting that $E[(Y - Z^T\boldsymbol{\beta}_0)\xi_j] = \gamma_j\lambda_j$, we have

$$\sum_{j=1}^m (\hat{\gamma}_j - \check{\gamma}_j)^2 \le 3\sum_{j=1}^m \lambda_j^{-2}\left(I_1^2 + I_2^2 + I_3^2\right)[1 + o_p(1)], \tag{A.26}$$

where $o_p(1)$ holds uniformly for $j = 1, \ldots, m$. Since $E(I_1) = 0$ and $E(I_1^2) \le \frac{1}{n}[\sum_{k=1}^\infty \gamma_k^2 E(\xi_k^2\xi_j^2) + \sigma^2\lambda_j] \le C\lambda_j/n$, we obtain that

$$\sum_{j=1}^m \lambda_j^{-2} I_1^2 = O_p\left(n^{-1}\sum_{j=1}^m \lambda_j^{-1}\right) = O_p(n^{-1}m\lambda_m^{-1}). \tag{A.27}$$

Let $M(t) = E[(Y_i - Z_i^T\boldsymbol{\beta}_0)X_i(t)] = \sum_{k=1}^\infty \gamma_k\lambda_k\phi_k(t)$. Then

$$I_2^2 \le 2\int_{\mathcal{T}}\left(\frac{1}{n}\sum_{i=1}^n (Y_i - Z_i^T\boldsymbol{\beta}_0)X_i(t) - M(t)\right)^2 \mathrm{d}t\,\|\hat{\phi}_j - \phi_j\|^2 + 2\left(\int_{\mathcal{T}} M(t)(\hat{\phi}_j(t) - \phi_j(t))\mathrm{d}t\right)^2.$$

Applying Assumption 1, it holds that

$$E\left(\int_{\mathcal{T}}\left(\frac{1}{n}\sum_{i=1}^n (Y_i - Z_i^T\boldsymbol{\beta}_0)X_i(t) - M(t)\right)^2 \mathrm{d}t\right) \le \frac{1}{n}\int_{\mathcal{T}} E[(Y_i - Z_i^T\boldsymbol{\beta}_0)^2 X_i^2(t)]\mathrm{d}t = O(n^{-1}).$$

From (A.9), we obtain $\sum_{j=1}^m \lambda_j^{-2}\|\hat{\phi}_j - \phi_j\|^2 = O_p(n^{-1}m^3\lambda_m^{-2}\log m)$. By arguments similar to those used in the proof of (5.15) of Hall and Horowitz (2007), it follows that

$$\sum_{j=1}^m \lambda_j^{-2}\left(\int_{\mathcal{T}} M(t)(\hat{\phi}_j(t) - \phi_j(t))\mathrm{d}t\right)^2 = O_p\left(\frac{m}{n\lambda_m} + \frac{m}{n^2\lambda_m^2}\sum_{j=1}^m \gamma_j^2\lambda_j^{-2}j^3 + \frac{m^3\log m}{n^2\lambda_m^2}\right).$$

Hence, using the assumption that $n^{-1}m^2\lambda_m^{-1}\log m \to 0$, we obtain

$$\sum_{j=1}^{m}\lambda_j^{-2}I_2^2 = O_p\left(n^{-1}m\lambda_m^{-1} + n^{-2}m\lambda_m^{-2}\sum_{j=1}^{m}\gamma_j^2\lambda_j^{-2}j^3\right). \tag{A.28}$$

Using Theorem 3.1, it holds that

$$\begin{aligned}\sum_{j=1}^{m}\lambda_j^{-2}I_3^2 &\le \left(\sum_{j=1}^{m}\frac{1}{n\lambda_j^2}\sum_{i=1}^{n}\hat{\xi}_{ij}^2\right)\left(\frac{1}{n}\sum_{i=1}^{n}[Z_i^T(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)]^2\right)\\ &= O_p\left(m\lambda_m^{-1} + n^{-1}m^3\lambda_m^{-2}\log m\right)O_p(n^{-1}) = O_p\left(n^{-1}m\lambda_m^{-1}\right).\end{aligned} \tag{A.29}$$

Now Lemma A.6 follows from combining (A.26)–(A.29). □

Proof of Theorem 2.2. Note that

$$\begin{aligned}\int_{\mathcal{T}}[\hat{\gamma}(t)-\gamma(t)]^2\mathrm{d}t \le C\Bigg(&\sum_{j=1}^{m}(\hat{\gamma}_j-\check{\gamma}_j)^2\\ &+\sum_{j=1}^{m}(\check{\gamma}_j-\gamma_j)^2 + m\sum_{j=1}^{m}\gamma_j^2\|\hat{\phi}_j-\phi_j\|^2 + \sum_{j=m+1}^{\infty}\gamma_j^2\Bigg)\end{aligned} \tag{A.30}$$

and

$$\sum_{j=1}^{m}(\check{\gamma}_j-\gamma_j)^2 = \sum_{j=1}^{m}\frac{(\hat{\lambda}_j-\lambda_j)^2}{\lambda_j^2}\gamma_j^2[1+o_p(1)] = O_p\left(n^{-1}\lambda_m^{-1}\sum_{j=1}^{m}\gamma_j^2\lambda_j^{-1}\right). \tag{A.31}$$

Assumption 3 implies that $m\sum_{j=1}^{m}\gamma_j^2\|\hat{\phi}_j-\phi_j\|^2 = O_p(mn^{-1}\sum_{j=1}^{m}\gamma_j^2 j^2\log j) = o_p(m/n)$ and $\sum_{j=m+1}^{\infty}\gamma_j^2 = O(m^{-2\gamma+1})$. Now (2.10) follows from Lemma A.6, (A.30) and (A.31). The proof of Theorem 2.2 is finished. □

Proof of Theorem 2.3. Observe that

$$\mathrm{MSPE} \le 2\{\|\hat{\gamma}-\gamma\|_K^2 + (\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)^T E(ZZ^T)(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\}, \tag{A.32}$$

where $\|\hat{\gamma}-\gamma\|_K^2 = \int_{\mathcal{T}}\int_{\mathcal{T}}K(s,t)[\hat{\gamma}(s)-\gamma(s)][\hat{\gamma}(t)-\gamma(t)]\mathrm{d}s\mathrm{d}t$. Under the assumptions of Theorem 2.3, using arguments similar to those used in the proof of Theorem 2 of Tang (2015), we deduce that $\|\hat{\gamma}-\gamma\|_K^2 = O_p(n^{-(\tau+2\delta-1)/(\tau+2\delta)})$. Now (2.12) follows from (A.32) and Theorem 2.1. The proof of Theorem 2.3 is finished. □

Lemma A.7 *Under the assumptions of Theorem 3.1, there exists a local minimizer* $\hat{\boldsymbol{\beta}}$ *of* (3.1) *such that* $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_p(n^{-1/2})$.

Proof Let

$$P_n(\boldsymbol{\beta}) = n \sum_{k=1}^{d} p'_{\nu_n}\left(\left|\beta_k^{(0)}\right|\right)|\beta_k|, \quad P_{n1}(\boldsymbol{\beta}) = n \sum_{k=1}^{d_1} p'_{\nu_n}\left(\left|\beta_k^{(0)}\right|\right)|\beta_k|$$

and $D_n(\boldsymbol{\beta}) = (\tilde{Y} - \tilde{Z}\boldsymbol{\beta})^T(\tilde{Y} - \tilde{Z}\boldsymbol{\beta}) + P_n(\boldsymbol{\beta})$. It suffices to prove that for any given $\varepsilon > 0$, there exists a constant C such that

$$P\left\{\sup_{\|u\|=C} D_n(\boldsymbol{\beta}_0 + n^{-1/2}u) > D_n(\boldsymbol{\beta}_0)\right\} \geq 1 - \varepsilon. \tag{A.33}$$

Note that

$$\begin{aligned} D_n(\boldsymbol{\beta}_0 + n^{-1/2}u) - D_n(\boldsymbol{\beta}_0) \geq & -2n^{-1/2}(\tilde{Y} - \tilde{Z}\boldsymbol{\beta}_0)^T\tilde{Z}u + n^{-1}u^T\tilde{Z}^T\tilde{Z}u \\ & + \left[P_{n1}\left(\boldsymbol{\beta}_{01} + n^{-1/2}u_1\right) - P_{n1}(\boldsymbol{\beta}_{01})\right] \end{aligned} \tag{A.34}$$

and

$$(\tilde{Y} - \tilde{Z}\boldsymbol{\beta}_0)^T\tilde{Z} = (\tilde{W} + \tilde{\varepsilon})^T\tilde{Z}.$$

By Lemma A.5, we have that $n^{-1/2}\tilde{W}^T\tilde{Z} = o_p(1)$. By (A.25), it follows that $n^{-1/2}\tilde{\varepsilon}^T\tilde{Z} = O_p(1)$. By Theorem 2.1, it holds that $\boldsymbol{\beta}^{(0)} \to_P \boldsymbol{\beta}_0$, and we then have $P\{P_{n1}(\boldsymbol{\beta}_{01} + n^{-1/2}u_1) - P_{n1}(\boldsymbol{\beta}_{01}) = 0\} \to 1$ as $n \to \infty$. Hence, for sufficiently large C, (A.33) follows from (A.34) and Lemma A.1 and the fact that Ω is positive definite. The proof of Lemma A.7 is complete. □

Proof of Theorem 3.1. We first prove that for any $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$ in the neighborhood $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| = O(n^{-1/2})$ for sufficiently large n and $\boldsymbol{\beta}_2 \neq \mathbf{0}$, with probability tending to 1, we have

$$D_n((\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)) > D_n((\boldsymbol{\beta}_1, \mathbf{0})). \tag{A.35}$$

Observe that

$$\begin{aligned} D_n\left((\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)\right) - D_n\left((\boldsymbol{\beta}_1, \mathbf{0})\right) = & -2\left(\tilde{W} - \tilde{Z}\left((\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{01})^T, \mathbf{0}^T\right)^T + \tilde{\varepsilon}\right)^T \tilde{Z}\left(\mathbf{0}^T, \boldsymbol{\beta}_2^T\right)^T \\ & + \left(\mathbf{0}^T, \boldsymbol{\beta}_2^T\right)\tilde{Z}^T\tilde{Z}\left(\mathbf{0}^T, \boldsymbol{\beta}_2^T\right)^T + n\sum_{k=d_1}^{d} p'_{\nu_n}\left(\left|\beta_k^{(0)}\right|\right)|\beta_k| \end{aligned}$$

By Lemma A.5, we have that $n^{-1/2}\tilde{W}^T\tilde{Z} = o_p(1)$. By (A.25), it follows that $n^{-1/2}\tilde{\varepsilon}^T\tilde{Z} = O_p(1)$. Hence, using Lemma A.1 and the fact that $\|\boldsymbol{\beta}_2\| = O(n^{-1/2})$

and $n^{1/2}\nu_n \to +\infty$ and the result of Theorem 2.1, we deduce that with probability tending to 1, it holds that

$$\begin{aligned} &D_n\left((\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)\right) - D_n\left((\boldsymbol{\beta}_1, \mathbf{0})\right) \\ &\quad = O_p\left(n^{1/2}\right)\sum_{k=d_1}^{d}|\beta_k| + n\sum_{k=d_1}^{d} p'_{\nu_n}\left(\left|\beta_k^{(0)}\right|\right)|\beta_k| \\ &\quad = n\nu_n \sum_{k=d_1}^{d}\left[O_p\left(\left(n^{1/2}\nu_n\right)^{-1}\right) + \nu_n^{-1} p'_{\nu_n}\left(\left|\beta_k^{(0)}\right|\right)\right]|\beta_k| > 0. \end{aligned}$$

By Lemma A.7 and (A.35), there exists a $\sqrt{n}$-consistent local minimizer $\check{\boldsymbol{\beta}} = (\check{\boldsymbol{\beta}}_1, \mathbf{0}^T)^T$ of (3.1). Note that

$$\begin{aligned} D_n((\hat{\boldsymbol{\beta}}_{\mathrm{PLS1}}, \hat{\boldsymbol{\beta}}_{\mathrm{PLS2}})) = D_n((\check{\boldsymbol{\beta}}_1, \mathbf{0})) - 2\sqrt{n}\Bigg[& n^{-1/2}(\tilde{Y} - \tilde{Z}\check{\boldsymbol{\beta}})^T \tilde{Z}(\hat{\boldsymbol{\theta}}_{\mathrm{PLS}} - \check{\boldsymbol{\beta}}) \\ & + n^{-1/2}(\hat{\boldsymbol{\theta}}_{\mathrm{PLS}} - \check{\boldsymbol{\beta}})^T \tilde{Z}^T \tilde{Z}(\hat{\boldsymbol{\theta}}_{\mathrm{PLS}} - \check{\boldsymbol{\beta}}) \\ & + \sqrt{n}\sum_{k=d_1+1}^{d} p'_{\nu_n}\left(\left|\beta_k^{(0)}\right|\right)|\hat{\beta}_{PLSk}|\Bigg], \end{aligned} \tag{A.36}$$

where $\hat{\boldsymbol{\beta}}_{\mathrm{PLS}} = (\hat{\beta}_{\mathrm{PLS1}}, \ldots, \hat{\beta}_{PLSd})^T$. Write $\tilde{Z} = (\tilde{\boldsymbol{Z}}_1, \tilde{\boldsymbol{Z}}_2)$. Since $\hat{\boldsymbol{\beta}}_{\mathrm{PLS}}$ is a minimizers of (3.1) and $\check{\boldsymbol{\beta}}$ is a local minimizer of (3.1), we then have that

$$(\tilde{Y} - \tilde{Z}\check{\boldsymbol{\beta}})^T \tilde{Z}(\hat{\boldsymbol{\theta}}_{\mathrm{PLS}} - \check{\boldsymbol{\beta}}) = (\tilde{W} + \tilde{\varepsilon})^T \tilde{\boldsymbol{Z}}_2 \hat{\boldsymbol{\theta}}_{\mathrm{PLS2}} + (\boldsymbol{\beta}_0 - \check{\boldsymbol{\beta}})\tilde{Z}^T \tilde{\boldsymbol{Z}}_2 \hat{\boldsymbol{\theta}}_{\mathrm{PLS2}}. \tag{A.37}$$

By Lemma A.5, we have that $n^{-1/2}\tilde{W}^T\tilde{\boldsymbol{Z}}_2 = o_p(1)$. By (A.25), it follows that $n^{-1/2}\tilde{\varepsilon}^T\tilde{\boldsymbol{Z}}_2 = O_p(1)$. The fact that $\boldsymbol{\beta}_0 - \check{\boldsymbol{\beta}} = O_p(n^{-1/2})$ and Lemma A.1 imply that $n^{-1/2}(\boldsymbol{\beta}_0 - \check{\boldsymbol{\beta}})\tilde{Z}^T\tilde{\boldsymbol{Z}}_2 = O_p(1)$. If $\hat{\boldsymbol{\beta}}_{\mathrm{PLS}} \neq \check{\boldsymbol{\beta}}$, under the assumptions of Theorem 3.1, then by (A.36) and (A.37), we have $D_n((\hat{\boldsymbol{\beta}}_{\mathrm{PLS1}}, \hat{\boldsymbol{\beta}}_{\mathrm{PLS2}})) > D_n((\check{\boldsymbol{\beta}}_1, \mathbf{0}))$. This is a contradiction to the fact that $\hat{\boldsymbol{\beta}}_{\mathrm{PLS}}$ is a minimizer of (3.1). So $\hat{\boldsymbol{\beta}}_{\mathrm{PLS2}} = 0$ and $\hat{\boldsymbol{\beta}}_{\mathrm{PLS1}} = \check{\boldsymbol{\beta}}_1$.

We now prove asymptotic normality part. Consider $D_n((\boldsymbol{\beta}_1, \mathbf{0}))$ as a function of $\boldsymbol{\beta}_1$. Noting that with probability tending 1, $\hat{\boldsymbol{\beta}}_{\mathrm{PLS1}}$ is the $\sqrt{n}$-consistent minimizer of $D_n((\boldsymbol{\beta}_1, \mathbf{0}))$ and satisfies

$$\frac{\partial D_n((\boldsymbol{\beta}_1, \mathbf{0}))}{\partial \boldsymbol{\beta}_1}\Bigg|_{\boldsymbol{\beta}_1 = \hat{\boldsymbol{\beta}}_{\mathrm{PLS1}}} = -2\tilde{\boldsymbol{Z}}_1^T(\tilde{Y} - \tilde{Z}\hat{\boldsymbol{\beta}}_{\mathrm{PLS}}) = 0$$

Hence

$$\hat{\boldsymbol{\beta}}_{\mathrm{PLS1}} - \boldsymbol{\beta}_{01} = \left(\tilde{\boldsymbol{Z}}_1^T\tilde{\boldsymbol{Z}}_1\right)^{-1}\tilde{\boldsymbol{Z}}_1^T\tilde{Y}.$$

By arguments similar to those used in the proof of (2.9), we can prove (3.2). The proof of Theorem 3.1 is finished. □

Proof of Theorem 3.2 Similar to the proofs of Theorems 2.2 and 2.3, we can complete the proof of Theorem 3.2. □

References

Aneirosa, G., Ferraty, F., Vieu, P.: Variable selection in partial linear regression with functional covariate. Statistics **49**, 1322–1347 (2015)

Brunel, E., Roche, A.: Penalized contrast estimation in functional linear models with circular data. Statistics **49**, 1298–1321 (2015)

Cai, T.T., Hall, P.: Prediction in functional linear regression. Ann. Stat. **34**, 2159–2179 (2006)

Cardot, H., Ferraty, F., Sarda, P.: Spline estimators for the functional linear model. Stat. Sin. **13**, 571–591 (2003)

Cardot, H., Mas, A., Sarda, P.: CLT in functional linear models. Probab. Theory Relat. Fields **138**, 325–361 (2007)

Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. **96**, 1348–1360 (2001)

Fan, J., Lv, J.: Non-concave penalized likelihood with np-dimensionality. IEEE Trans. Inf. Theory **57**, 5467–5484 (2011)

Fan, J., Xue, L., Zou, H.: Strong oracle optimality of folded concave penalized estimation. Ann. Stat. **42**, 819–849 (2014)

Frank, I., Friedman, J.: A statistical view of some chemometrics regression tools (with discussion). Technometrics **35**, 109–135 (1993)

Hall, P., Horowitz, J.L.: Methodology and convergence rates for functional linear regression. Ann. Stat. **35**, 70–91 (2007)

Hsing, T., Eubank, R.: Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators. Wiley, New York (2015)

Liang, H., Li, R.: Variable selection for partially linear models with measurement errors. J. Am. Stat. Assoc. **104**, 234–248 (2009)

Lv, J., Fan, J.: A unified approach to model selection and sparse recovery using regularized least squares. Ann. Stat. **37**, 3498–3528 (2009)

Ramsay, J.O., Silverman, B.W.: Applied Functional Data Analysis: Methods and Case Studies. Springer, New York (2002)

Ramsay, J.O., Silverman, B.W.: Functional Data Analysis. Springer, New York (2005)

Reiss, P.T., Ogden, R.T.: Functional generalized linear models with images as predictors. Biometrics **66**, 61–69 (2010)

Shin, H.: Partial functional linear regression. J. Stat. Plan. Inference **139**, 3405–3418 (2009)

Shin, H., Lee, M.H.: On prediction rate in partial functional linear regression. J. Multivar. Anal. **103**, 93–106 (2012)

Tang, Q.: Estimation for semi-functional linear regression. Statistics **49**, 1262–1278 (2015)

Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc. B **58**, 267–288 (1996)

Wang, M., Wang, X.: Adaptive Lasso estimators for ultrahigh dimensional generalized linear models. Stat. Prob. Lett. **89**, 41–50 (2014)

Zhang, C.H.: Nearly unbiased variable selection under mini-max concave penalty. Ann. Stat. **38**, 894–942 (2010)

Zhang, D., Lin, X., Sowers, M.F.: Two-stage functional mixed models for evaluating the effect of longitudinal covariate profiles on a scalar outcome. Biometrics **63**, 351–362 (2007)

Zou, H.: The adaptive lasso and its oracle properties. J. Am. Stat. Assoc. **101**, 1418–1429 (2006)