Hindawi Publishing Corporation
Abstract and Applied Analysis
Volume 2013, Article ID 687151, 11 pages
http://dx.doi.org/10.1155/2013/687151

Research Article

Remodeling and Estimation for Sparse Partially Linear Regression Models

Yunhui Zeng,[1,2] Xiuli Wang,[3] and Lu Lin[1]

[1] *Shandong University Qilu Securities Institute for Financial Studies and School of Mathematical Science, Shandong University, Jinan 250100, China*
[2] *Supercomputing Center, Shandong Computer Science Center, Jinan 250014, China*
[3] *College of Mathematics Science, Shandong Normal University, Jinan 250014, China*

Correspondence should be addressed to Lu Lin; linlu@sdu.edu.cn

Received 11 October 2012; Accepted 14 December 2012

Academic Editor: Xiaodi Li

When the dimension of covariates in the regression model is high, one usually uses a submodel as a working model that contains significant variables. But it may be highly biased and the resulting estimator of the parameter of interest may be very poor when the coefficients of removed variables are not exactly zero. In this paper, based on the selected submodel, we introduce a two-stage remodeling method to get the consistent estimator for the parameter of interest. More precisely, in the first stage, by a multistep adjustment, we reconstruct an unbiased model based on the correlation information between the covariates; in the second stage, we further reduce the adjusted model by a semiparametric variable selection method and get a new estimator of the parameter of interest simultaneously. Its convergence rate and asymptotic normality are also obtained. The simulation results further illustrate that the new estimator outperforms those obtained by the submodel and the full model in the sense of mean square errors of point estimation and mean square prediction errors of model prediction.

1. Introduction

Consider the following partially linear regression model:

$$Y = \beta^T X + \gamma^T Z + f(U) + \varepsilon, \tag{1}$$

where Y is a scalar response variable, X and Z are, respectively, p-dimensional and q-dimensional continuous-valued covariates with p being finite and $p \ll q$, β is the parameter vector of interest and γ is the nuisance parameter vector which is supposed to be sparse in the sense that $\|\gamma\|_2$ is small, $f(\cdot)$ is an unknown function satisfying $Ef(U) = 0$ for identification, ε is the random error satisfying $E(\varepsilon \mid X, Z, U) = 0$. For simplicity, we assume that U is univariate. Let (Y_i, X_i, Z_i, U_i), $i = 1, \ldots, n$, be i.i.d. observations of (Y, X, Z, U) obtained from the above model.

A feature of the model is that the parametric part contains both the parameter vector of interest and nuisance parameter vector. The reason for this coefficient separation is as follows. In practice we often use such a model to distinguish the main treatment variables of interest from the state variables. For instance, in a clinical trial, X consists of treatment variables and can be easily controlled, Z is a vector of many clinical variables, such as patient ages and body weights. The variables in Z may have an impact on Y but are not of primary interest and the effects may be small. In order to make up for potentially nonnegligible effects on the response Y, the nuisance covariate Z are introduced into model (1); see Shen et al. [1]. Model (1) contains all relevant covariates and in this paper we call it full model.

The purpose of this paper is to estimate β, the parameter of interest, when $\gamma^T Z$ is removed from the model. The main idea is remodeling based on the following working model:

$$Y = \beta^T X + f(U) + \eta. \tag{2}$$

As is known, $E(\eta \mid X = x, U = u)$ is a nonzero function if $\gamma^T E(Z \mid X, U) \neq 0$, which relies on two elements, one is $E(Z \mid X, U)$, related with the correlation between the covariates of Z and (X, U), the other is γ, determined by the nuisance parameter in the removed part. Thus the least squares estimator based on model (2) may be inconsistent. In the following, we will make use of the above two elements. Specifically, in the first stage, we shall construct a remodeled model by a multistep-adjustment to correct the submodel bias based on the correlation information between the covariates. This adjustment is motivated by Gai et al. [2]. In the paper, they proposed a nonparametric adjustment by adding a univariate nonparametric estimation to the working model (2), and it can dramatically reduce the bias of the working model. But this only holds in a subset of the covariates, although the subset may be fairly large. In order to obtain a globally unbiased working model for linear regression model, Zeng et al. [3] adjusted the working model by multiple steps. Because only those variables in Z correlated with (X, U) may have impact on estimation of β, in each step a univariate nonparametric part was added to the working model and consequently a globally unbiased working model was obtained.

However, when many components of Z are correlated with (X, U), the number of nonparametric functions added in the above working model is large. Such a model is improper in practice. Thus, in the second stage, we further simplify the above adjusted model by a semiparametric variable selection procedure proposed by Zhao and Xue [4]. Their method can select significant parametric and nonparametric components simultaneously under sparsity condition for semiparametric varying coefficient partially linear models. The relevant papers include Fan and Li [5], Wang et al. [6, 7], among others. After two-stage remodeling, the final model is conditionally unbiased. Based on this model, the estimation and model prediction are significantly improved.

The rest of this paper is organized as follows. In Section 2, a multistep adjustment and remodeled models are firstly proposed, then the models are further simplified via the semiparametric SCAD variable selection procedure. A new estimator of the parameter of interest based on the simplified model is derived, its convergence rate and asymptotic normality are also obtained. Simulations are given in Section 3. A short conclusion and some remarks are contained in Section 4. Some regular conditions and theoretical proofs are presented in the appendix.

2. New Estimator for the Parameter of Interest

In this paper, we suppose that covariate Z has zero mean, p is finite and $p \ll q$, $E(\varepsilon \mid X, Z, U) = 0$ and $\mathrm{Var}(\varepsilon \mid X, Z, U) = \sigma^2$. We also assume that covariates X and U and parameter β are prespecified, so that the submodel (2) is a fixed model.

2.1. Multistep-Adjustment by Correlation. In this subsection, we first adjust the submodel to be conditionally unbiased by a multistep-adjustment.

When Z is normally distributed, the principal component analysis (PCA) method will be used. Let Σ_Z be the covariance matrix of Z, then there exists an orthogonal $q \times q$ matrix Q such that $Q\Sigma_Z Q^T = \Lambda$, where Λ is the diagonal matrix $\mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_q)$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_q \geq 0$ being eigenvalues of Σ_Z. Denote $Q^T = (\tau_1, \tau_2, \ldots, \tau_q)$ and $\widetilde{Z}^{(j)} = \tau_j^T Z$.

When Z is centered but nonnormally distributed, we shall apply independent component analysis (ICA) method. Assume that Z is generated by a nonlinear combination of independent components $\widetilde{Z}^{(i)}$, that is $Z = F(\widetilde{Z})$, where $F(\cdot)$ is an unknown nonlinear mapping from R^q to R^q, $\widetilde{Z}$ is an unknown random vector with independent components. By imposing some constraints on the nonlinear mixing mapping F or the independent components $\widetilde{Z}^{(i)}$, the independent components $\widetilde{Z}^{(i)}$ can be properly estimated. See Simas Filho and Seixas [8] for an overview of the main statistical principles and some algorithms for estimating the independent components. For simplicity, in this paper we suppose that $Z = (Z^{(1)}, \ldots, Z^{(q)})^T$ with $Z^{(l)} = \Sigma_{j=1}^q F_{lj}(\widetilde{Z}^{(j)})$, $l = 1, \ldots, q$, and $F_{lj}(\cdot)$ are scalar functions.

In the above two cases, $\widetilde{Z}^{(j)}$'s are independent of each other. Set K_0 to be the size of set $M_0 = \{j : E(\widetilde{Z}^{(j)} \mid X, U) \neq 0,\ 1 \leq j \leq q\}$. Without loss of generality, let $M_0 = \{1, \ldots, K_0\}$.

We construct the following adjusted model:

$$Y = \beta^T X + \sum_{j=1}^{K_0} g_j\left(\widetilde{Z}^{(j)}\right) + f(U) + \zeta_{K_0}, \tag{3}$$

where $g_j(\widetilde{Z}^{(j)}) = E(Y - \beta^T X - f(U) \mid \widetilde{Z}^{(j)}) = \gamma^T E(Z \mid \widetilde{Z}^{(j)})$, $j = 1, \ldots, K_0$ and $\zeta_{K_0} = Y - \beta^T X - g_1(\widetilde{Z}^{(1)}) - \cdots - g_{K_0}(\widetilde{Z}^{(K_0)}) - f(U)$. The model (3) is based on Z's population and depends on the distributions of X, U and Z. It is easy to see that model (3) is conditionally unbiased, that is, $E(\zeta_{K_0} \mid X, U, \widetilde{Z}^{(j)}, 1 \leq j \leq K_0) = 0$.

The adjusted model (3) is an additive partially linear model, in which $\beta^T X$ is the parametric part, $f(U)$ and $g_j(\widetilde{Z}^{(j)})$, $j = 1, \ldots, K_0$, are the nonparametric parts and ζ_{K_0} is the random error. Compared with the submodel (2), the nonparametric parts $g_j(\widetilde{Z}^{(j)})$, $j = 1, \ldots, K_0$, may be regarded as bias-corrected terms for the random error η. For centered Z, $E(g_j(\widetilde{Z}^{(j)})) = 0$, $j = 1, \ldots, K_0$, the nonparametric components $g_1(\widetilde{Z}^{(1)}), \ldots, g_{K_0}(\widetilde{Z}^{(K_0)})$ can be properly identified. In fact, centered Z can be relaxed to any Z such that satisfies $\gamma^T E(Z) = 0$.

When Z is centered and normally distributed, the nonparametric parts $g_j(\widetilde{Z}^{(j)}) = \alpha_j \tau_j^T Z = \alpha_j \widetilde{Z}^{(j)}$, $j = 1, \ldots, K_0$. So the multistep adjusted model (3) is really a partially linear model

$$Y = \beta^T X + \alpha^T \widetilde{Z}_{K_0} + f(U) + \zeta_{K_0} \tag{4}$$

with $\alpha = (\alpha_1, \ldots, \alpha_{K_0})^T$ and $\widetilde{Z}_{K_0} = (\widetilde{Z}^{(1)}, \ldots, \widetilde{Z}^{(K_0)})^T$. Specially, when $f(U) \equiv 0$, the full model is a linear model, the multistep adjusted model is also a linear model

$$Y = \beta^T X + \alpha^T \widetilde{Z}_{K_0} + \zeta_{K_0}. \tag{5}$$

But when the variables in Z are not jointly normal, the nonparametric parts g_j can be highly nonlinear, which are similar to the results of marginal regression; see Fan et al. [9].

2.2. Model Simplification. When the most of the features in the full model are correlated, then K_0 is very large and even is close to q. In this case, the adjusted model (3) is improper in practice, so we shall use the group SCAD regression procedure, proposed by Wang et al. [6], and the semiparametric variable selection procedure, proposed by Zhao and Xue [4], to further simplify the model.

Let $s = |\mathscr{M}_*|$ with $\mathscr{M}_* = \{1 \le j \le K_0 : E(g_j(\widetilde{Z}^{(j)}))^2 > 0\}$, and assume that the model (3) is sparse, that is, s is small. We define the semiparametric penalized least squares as

$$F(\beta, g(\cdot), f(\cdot)) = \sum_{i=1}^{n}\left\{Y_i - \beta^T X_i - \sum_{j=1}^{K_0} g_j\left(\widetilde{Z}_i^{(j)}\right) - f(U_i)\right\}^2 + n\sum_{j=1}^{K_0} p_{\lambda_j}\left(\left\|g_j\left(\widetilde{Z}^{(j)}\right)\right\|\right), \qquad (6)$$

where $\|g_j(\widetilde{Z}^{(j)})\| = (E(g_j(\widetilde{Z}^{(j)}))^2)^{1/2}$, and $p_\lambda(\cdot)$ is the SCAD penalty function with λ being a tuning parameter defined as

$$p'_\lambda(w) = \lambda\left\{I(w \le \lambda) + \frac{(a\lambda - w)_+}{(a-1)\lambda} I(w > \lambda)\right\}, \qquad (7)$$

with $a > 2$, $w > 0$ and $p_\lambda(0) = 0$. In (6), $g(\cdot)$ denotes the set $\{g_j(\widetilde{Z}^{(j)}),\ j = 1, \ldots, K_0\}$. Because g_j are nonparametric functions, thus they cannot be directly applied for minimization. Here we will replace $f(\cdot)$ and $g(\cdot)$ by basis function approximations. For $1 \le j \le K_0$, let $\{\Psi_{jk},\ k = 1, \ldots, L\}$ be orthogonal basis functions satisfying

$$E\left(\Psi_{jk}\Psi_{jl}\right) \equiv \int_{\text{supp}} \Psi_{jk}\left(\widetilde{Z}^{(j)}\right)\Psi_{jl}\left(\widetilde{Z}^{(j)}\right) r_j\left(\widetilde{Z}^{(j)}\right) dZ = \delta_{kl} = \begin{cases} 0, & k \ne l; \\ 1, & k = l, \end{cases} \qquad (8)$$

where $r_j(\widetilde{Z}^{(j)})$ is the density function of $\widetilde{Z}^{(j)}$. Similarly, let $\{\Psi_{0k},\ k = 1, \ldots, L\}$ be orthogonal basis functions satisfying the above condition which is only replaced by the support and density function of U. Denote $\Psi_j(\widetilde{Z}^{(j)}) = (\Psi_{j1}(\widetilde{Z}^{(j)}), \ldots, \Psi_{jL}(\widetilde{Z}^{(j)}))^T$, $\Psi_0(U) = (\Psi_{01}(U), \ldots, \Psi_{0L}(U))^T$. Then $g_j(\widetilde{Z}^{(j)})$ and $f(U)$ can be approximated by

$$g_j\left(\widetilde{Z}^{(j)}\right) \approx \theta_j^T \Psi_j\left(\widetilde{Z}^{(j)}\right), \qquad f(U) \approx \nu^T \Psi_0(U). \qquad (9)$$

Denote $\|\theta_j\|_2 = (\theta_j^T\theta_j)^{1/2}$, invoking that $E(\Psi_j(\widetilde{Z}^{(j)}) \cdot \Psi_j^T(\widetilde{Z}^{(j)})) = \mathbf{I}_L$ the identity matrix, we get

$$F(\beta, \theta, \nu) = \sum_{i=1}^{n}\left\{Y_i - \beta^T X_i - \theta^T \Psi_{\mathbf{i}} - \nu^T \Psi_{\mathbf{0i}}\right\}^2 + n\sum_{j=1}^{K_0} p_{\lambda_j}\left(\left\|\theta_j\right\|_2\right), \qquad (10)$$

where $\theta = (\theta_1^T, \ldots, \theta_{K_0}^T)^T$, $\Psi_{\mathbf{i}} \equiv \Psi(\widetilde{Z}_i) = \text{Vec}(\Psi_1(\widetilde{Z}_i^{(1)}), \ldots, \Psi_{K_0}(\widetilde{Z}_i^{(K_0)}))$, $\Psi_{\mathbf{0i}} \equiv \Psi_0(U_i)$.

Denote by $\widehat{\beta}$, $\widehat{\theta} = (\widehat{\theta}_1^T, \ldots, \widehat{\theta}_{K_0}^T)^T$ and $\widehat{\nu}$ the least squares estimators based on the penalized function (10), that is $(\widehat{\beta}, \widehat{\theta}, \widehat{\nu}) = \arg\min_{\beta \in R^p, \theta_j \in R^L, \nu \in R^L} F(\beta, \theta, \nu)$. Let $\widehat{g}_j \equiv \widehat{g}_j(\widetilde{Z}^{(j)}) = \widehat{\theta}_j^T \Psi_j(\widetilde{Z}^{(j)})$ and $\widehat{f} \equiv \widehat{f}(U) = \widehat{\nu}^T \Psi_0(U)$, then $\widehat{g}_j$ is an estimator of $g_j(\widetilde{Z}^{(j)})$, $\widehat{f}$ is an estimator of $f(U)$.

Let $\widehat{\mathscr{M}_n} = \{1 \le j \le K_0 : \widehat{\theta}_j \ne 0\}$ and $K_n = \widehat{|\mathscr{M}_n|}$. For simplicity, we assume that $\mathscr{M}_* = \{1, 2, \ldots, s\}$ and $\widehat{\mathscr{M}_n} = \{1, 2, \ldots, K_n\}$. So we get the following simplified working model

$$Y = \beta^T X + \sum_{j=1}^{K_n} g_j\left(\widetilde{Z}^{(j)}\right) + f(U) + \zeta_{K_n}, \qquad (11)$$

where $g_j(\widetilde{Z}^{(j)}) = E(\gamma^T Z \mid \widetilde{Z}^{(j)})$, $j = 1, \ldots, K_n$ and $\zeta_{K_n} = Y - \beta^T X - g_1(\widetilde{Z}^{(1)}) - \cdots - g_{K_n}(\widetilde{Z}^{(K_n)}) - f(U)$. Under the assumption of sparsity, the model (11) contains all of significant nonparametric functions and fully utilizes both the correlation information of covariates and the model sparsity on nuisance covariate.

If Z is centered and normally distributed with covariance matrix $\Sigma_Z = I_q$ the identity matrix, then $\tau_j = e_j$, $j = 1, \ldots, q$, where e_j denotes the unit vector with 1 at position j, and α is sparse with $\alpha_j = \gamma^T \tau_j = \gamma_j$. So the model (4) is sparse. For model (5), the special case of model (4), we can apply the SCAD penalty method proposed by Fan and Li [5] to select variables in $\widetilde{Z}_{K_0}$ and estimate parameters α and β simultaneously. The selected covariate and the corresponding parameter are denoted by $\widetilde{Z}_{K_n}$ and α_{K_n}, the resulting parameter estimators are denoted by $\widehat{\alpha_{K_n}}$ and $\widehat{\beta}$, respectively. Finally, we can use the simplified model

$$Y = \beta^T X + \alpha_{K_n}^T \widetilde{Z}_{K_n} + \zeta_{K_n} \qquad (12)$$

for model prediction. Under the condition of sparsity, its model size is much smaller than those of the multistep adjusted model (5) and the full model (1).

2.3. Asymptotic Property of Point Estimator. Let β_0, θ_0, ν_0, and $g_{j0}(\cdot)$, $f_0(\cdot)$ be the true values of β, θ, ν, and $g_j(\cdot)$, $f(\cdot)$, respectively, in model (3). Without loss of generality, we assume that $g_{j0}(\widetilde{Z}^{(j)}) = 0$, $j = s+1, \ldots, K_0$, and $g_{j0}(\widetilde{Z}^{(j)})$, $j = 1, \ldots, s$, are all nonzero components.

We suppose that $g_j(\widetilde{Z}^{(j)})$, $j = 1, \ldots, K_0$ can be expressed as $\sum_{k=1}^{\infty} \theta_{jk} \Psi_{jk}(\widetilde{Z}^{(j)})$ and $f(U)$ can be expressed as $\sum_{k=1}^{\infty} \nu_k \Psi_{0k}(U)$, θ_j and ν belong to the Sobolev ellipsoid $S(r, M) = \{\theta : \sum_{k=1}^{\infty} \theta_k^2 k^{2r} \leq M,\ M > 0, r > 0\}$.

The following theorem gives the consistency of the penalized SCAD estimators.

Theorem 1. *Suppose that the regularity conditions (C1)–(C5) in the appendix hold and the number of terms* $L = O_p(n^{1/(2r+1)})$. *Then,*

(i) $\|\widehat{\beta} - \beta_0\| = O_p(n^{-r/(2r+1)} + a_n)$,

(ii) $\|\widehat{g}_j(\cdot) - g_{j0}(\cdot)\| = O_p(n^{-r/(2r+1)} + a_n)$, $j = 1, \ldots, K_0$,

(iii) $\|\widehat{f}(\cdot) - f_0(\cdot)\| = O_p(n^{-r/(2r+1)} + a_n)$,

where $a_n = \max_j\{|p'_{\lambda_j}(\|\theta_{j0}\|_2)| : \theta_{j0} \neq 0\}$.

From the last paragraph of Section 2.2 we know that, for linear regression model and normally distributed Z, the multistep adjusted model (5) is a linear model. By orthogonal basis functions, such as power series, we have $r = \infty$, then $\|\widehat{\beta} - \beta_0\| = O_p(n^{-1/2} + a_n)$, implying the estimator $\widehat{\beta}$ has the same convergence rate as that of the SCAD estimator in Fan and Li [5].

Theorem 2. *Suppose that the regularity conditions (C1)–(C6) in the appendix hold and the number of terms* $L = O_p(n^{1/(2r+1)})$. *Let* $\lambda_{\max} = \max_j\{\lambda_j\}$ *and* $\lambda_{\min} = \min_j\{\lambda_j\}$. *If* $\lambda_{\max} \to 0$ *and* $n^{r/(2r+1)}\lambda_{\min} \to \infty$ *as* $n \to \infty$, *then, with probability tending to 1,* $\widehat{g}_j(\cdot) \equiv 0$, $j = s+1, \ldots, K_0$.

Remark 3. By Remark 1 of Fan and Li [5], we have that, if $\lambda_{\max} \to 0$ as $n \to \infty$, then $a_n \to 0$. Hence from Theorems 1 and 2, by choosing proper tuning parameters, the variable selection method is consistent and the estimators of nonparametric components achieve the optimal convergence rate as if the subset of true zero coefficients was already known; see Stone [10].

Let $\theta^* = (\theta_1^T, \ldots, \theta_s^T)^T$ be the nonzero components of θ, corresponding covariates are denoted by Ψ_i^*, $i = 1, \ldots, n$. In addition, let

$$\Sigma = \frac{1}{\sigma_{K_0}^2} \left\{E\left(XX^T\right) - E\left(X\Psi^{*T}\right) E^{-1}\left(\Psi^*\Psi^{*T}\right) E\left(\Psi^* X^T\right) \right.$$
$$\left. -E\left(X\breve{\Psi}_0^T\right) E^{-1}\left(\breve{\Psi}_0\breve{\Psi}_0^T\right) E\left(\breve{\Psi}_0 X^T\right)\right\}, \tag{13}$$

where $\sigma_{K_0}^2 = \operatorname{Var}(\zeta_{K_0 i})$ for homoscedastic case, $\breve{\Psi}_0 = \Psi_0 - E(\Psi_0\Psi^{*T})E^{-1}(\Psi^*\Psi^{*T})\Psi^*$.

Theorem 4. *Suppose that the regularity conditions (C1)–(C6) in the appendix hold and the number of terms* $L = O_p(n^{1/(2r+1)})$. *If* Σ *is invertible, then*

$$\sqrt{n}\left(\widehat{\beta} - \beta_0\right) \xrightarrow{\mathscr{D}} N\left(0, \Sigma^{-1}\right), \tag{14}$$

where "$\xrightarrow{\mathscr{D}}$" *denotes the convergence in distribution.*

Remark 5. From Theorems 1 and 4, it can be found that the penalized estimators have the oracle property. Furthermore, the estimator of the parameter of interest has the same asymptotic distribution as that based on the correct submodel.

2.4. Some Issues on Implementation. In the adjusted model (4), τ_j, $j = 1, \ldots, K_0$ are used. When the population distribution is not available, they need to be approximated by estimators. When Z is normally distributed and eigenvalues r_j, $j = 1, \ldots, q$ of the covariance matrix Σ_Z are different from each other, then $\sqrt{n}(u_j - \tau_j)$ is asymptotically $N(\mathbf{0}, V_j)$ with $V_j = \sum_{l \neq j}^{q}(r_j r_l/(r_j - r_l)^2)\tau_l\tau_l^T$, where u_j is the jth eigenvector of $S = (1/(n-1))\sum_{i=1}^{n}(Z_i - \overline{Z})(Z_i - \overline{Z})^T$ with $\overline{Z} = (1/n)\sum_{i=1}^{n} Z_i$; see Anderson [11]. For the case when the population size is large and comparable with the sample size, if the covariance matrix is sparse, we can use the method in Rütimann and Bühlmann [12] or Cai and Liu [13] to estimate the covariance matrix. So we can use u_j to approximate τ_j. When τ_j in model (4) are replaced by these consistent estimators, one can see that the approximation error can be neglected without changing the asymptotic property.

The nonparametric parts $g_l(\widetilde{Z}^{(l)})$ in the adjusted model depend on the univariate variable $\widetilde{Z}^{(l)}$, for $l = 1, \ldots, K_0$. So it needs to choose the steps K_0 firstly. In real implementation, we compute all the q multiple correlation coefficients of $\widetilde{Z}^{(l)}$ $(l = 1, \ldots, q)$ with X and U. Then we choose the components $R = \{\widetilde{Z}^{(l)}: |\operatorname{mcorr}(\widetilde{Z}^{(l)}, (X, U))| \geq \delta,\ l = 1, \ldots, q\}$ for given small number $\delta > 0$, where $\operatorname{mcorr}(u, V)$ denotes the multicorrelation coefficient between u and V and can be approximated by its sample form; see Anderson [11].

There are some tuning parameters needing to choose in order to implement the two-stage remodeling procedure. Fan and Li [5] showed that the SCAD penalty with $a = 3.7$ performs well in a variety of situations. Hence, we use their suggestion throughout this paper. We still need to choose the positive integer L for basis functions and the tuning parameter λ_j of the penalty functions. Similar to the adaptive lasso of Zou [14], we suggest taking $\lambda_j = \lambda/\|\widehat{\theta}_j^{(0)}\|_2$, where $\widehat{\theta}_j^{(0)}$ is initial estimator of θ_j by using ordinary least squares method based on the first term in (10). So the two remaining parameters L and λ can be selected simultaneously using the leave-one-out CV or GCV method; see Zhao and Xue [4] for more details.

3. Simulation Studies

In this section, we investigate the behavior of the newly proposed method by simulation studies.

3.1. Linear Model with Normally Distributed Covariates. The dimensions of the full model (1) and the submodel (2) are chosen to be 100 and 5, respectively. We set $\beta = (0.5, 3.5, 2.5, 1.5, 4.0)^T$ and $\gamma = (\gamma_1, \gamma_2, \mathbf{0}_{55}^T)^T$, where $\gamma_2 \sim \text{Unif}[-0.5, 0.5]^{30}$, a 30-dimensional uniform distribution on $[-0.5, 0.5]^{30}$, and γ_1 is chosen in the following ways:

Case (I). $\gamma_1 \sim \text{Unif}[0.5, 1.0]^{10}$.

Case (II). $\gamma_1 = (1.0, 1.0, 1.0, 1.5, 1.5, 1.5, 2.0, 2.0, 2.0, 2.0)$.

We assume that $(X^T, Z^T)^T \sim N((\mathbf{1}_5^T, \mathbf{0}_{40}^T, \mathbf{1}_{55}^T)^T, \Sigma\Sigma^T)$, where

$$\Sigma = \left(\sigma_{ij}\right), \qquad \sigma_{ij} = \sigma_{ji} = \begin{cases} 1.0, & j = i,\ i = 1, \ldots, p + q; \\ c, & j = i + p,\ i = 1, 3, \ldots, q; \\ 0, & \text{otherwise}, \end{cases} \tag{15}$$

with $c = 0.5$ or $c = 0.8$. The error term ε is assumed to be normally distributed as $N(0, 0.3^2)$.

Here we denote the submodel (2) as model (I), the multistep adjusted linear model (5) as model (II), the two-stage model (12) as model (III), and the full model (1) as model (IV). We compare mean square errors (MSEs) of the new two-stage estimator $\widehat{\beta}_{\text{TS}}$ based on model (III) with the estimator $\widehat{\beta}_S$ based on model (I), the multistep estimator $\widehat{\beta}_M$ based on model (II), the SCAD estimator $\widehat{\beta}_{\text{SCAD}}$ and the least squares estimator $\widehat{\beta}_F$ based on model (IV). We also compare mean square prediction errors (MSPEs) of the above mentioned models with corresponding estimators.

The data are simulated from the full model (1) with sample size $n = 100$ and simulation times $m = 1000$. We use the sample-based PCA approximations to substitute τ_j's. The parameter a in the SCAD penalty function is set to be 3.7 and λ is selected by leave-one-out CV method.

Table 1 reports the MSEs of point estimators on the parameter β and the MSPEs of model predictions. From the table, we have the following findings: (1) $\widehat{\beta}_F$ has the largest MSEs and $\widehat{\beta}_S$ takes the second place, nearly all the new estimator $\widehat{\beta}_{\text{TS}}$ has the smallest MSEs. (2) When $c = 0.5$, the MSEs of $\widehat{\beta}_{\text{SCAD}}$ are smaller than those of $\widehat{\beta}_M$, while when $c = 0.8$ they are larger than those of $\widehat{\beta}_M$. These show that if the correlation between the covariates is strong, the MSEs of $\widehat{\beta}_{\text{SCAD}}$ are larger than those of $\widehat{\beta}_M$, the multistep-adjustment is necessary, so the estimations and model predictions based on two-stage model are significantly improved. (3) In case (I) and (II) the simulation results have the similar performance. (4) Similar to the trend of the MSEs of the five estimators, the MSPE of the two-stage adjusted model is the smallest among the mentioned five models.

In summary, Table 1 indicates that the two-stage adjusted linear model (12) performs much better than the full model, and better than the submodel, the SCAD-penalized model and the multistep adjusted model.

3.2. Partially Linear Model with Nonnormally Distributed Covariates. The dimensions of the linear part in the full model (1) and the submodel (2) are chosen to be 50 and 5, respectively. We set $\beta = (0.5, 3.5, 2.5, 1.5, 4.0)^T$, $\gamma = (\gamma_1, \gamma_2, \mathbf{0}_{25}^T)^T$, $f(u) = u^2 * \sin(3u)$, where

$$\gamma_1 = (0.5, 0.1, 0.8, 0.2, 0.5, 0.2, 0.6, 0.5, 0.1, 0.9),$$

$\gamma_2 \sim \text{Unif}[-0.3, 0.3]^{10}$, a 10-dimensional uniform distribution on $[-0.3, 0.3]$.

We assume that the covariates are distributed in the following two ways.

Case (I). $(X^T, Z^T, U)^T \sim t(\mathbf{0}_{51}^T, \Sigma\Sigma^T)$, a 51-dimensional student distribution with degree of freedom df = 5, where

$$\Sigma = \left(\sigma_{ij}\right),$$
$$\sigma_{ij} = \sigma_{ji} = \begin{cases} 1.0, & j = i,\ i = 1, \ldots, p + q + 1; \\ 0.95, & j = i + p,\ i = 1, 2, \ldots, q + 1; \\ 0.9, & j = i + p - 2,\ i = 1, 2, \ldots, q + 3; \\ 0, & \text{otherwise}. \end{cases} \tag{16}$$

Case (II). $X = (1/(1 + c))(W_1 + cV)$, $Z = (Z_1^T, Z_2^T, Z_3^T, Z_4^T)^T$ with $Z_1 = (1/(1 + c))(W_2 + cV)$, $Z_2 = W_3$, $Z_3 = (1/(1 + c))(W_4 + cV)$, $Z_4 = W_5$, $U = W_5^{(1)}$, where $W_1, W_2, W_3, W_4 \sim \text{Unif}[-1.0, 1.0]^5$, $W_5 \sim \text{Unif}[-1.0, 1.0]^{30}$, $V \sim \text{Unif}[-1.0, 1.0]^5$, uniform distributions on $[-1.0, 1.0]$ and constant $c = 0.1$. All W_1, W_2, W_3, W_4, W_5, and V are independent.

The error term ε is assumed to be normally distributed as $N(0, 0.3^2)$.

Here we denote the submodel (2) as model $(\text{I})'$, the multistep adjusted additive partially linear model (3) as model $(\text{II})'$, the two-stage model (11) as model $(\text{III})'$ and the full model (1) as model $(\text{IV})'$. We compare mean square errors (MSEs) of the new two-stage estimator $\widehat{\beta}_{\text{TS}}$ based on model $(\text{III})'$ with the estimator $\widehat{\beta}_S$ based on model $(\text{I})'$, the estimator $\widehat{\beta}_M$ based on model $(\text{II})'$ and the least squares estimator $\widehat{\beta}_F$ based on model $(\text{IV})'$. We also compare the mean average square errors (MASEs) of the nonparametric estimators of $f(\cdot)$ and the mean square prediction errors (MSPEs) of different models with corresponding estimators.

The data are simulated from the full model (1) with sample size $n = 100$ and simulation times $m = 500$. We use the sample-based approximations of ICA, see Hyvärinen and Oja [15]. The parameter a in the SCAD penalty function is set to be 3.7, the number L and the parameter λ is selected by GCV method. We use the standard Fourier orthogonal basis as the basis functions.

Table 2 reports the MSEs of point estimators on the parameter β, the MASEs of $f(\cdot)$ and the MSPEs of model predictions. From the table, we have the following results: (1) $\widehat{\beta}_F$ has the largest MSEs, its MSEs are much larger than the MSEs of the other estimators, and the new estimator $\widehat{\beta}_{\text{TS}}$ always has the smallest MSEs. (2) The MASEs of $f(\cdot)$ have

Table 1: MSEs on the parameter β and MSPEs of the two-stage adjusted linear model (12) compared with the submodel, the SCAD-penalized model, the multistep adjusted model and the full model.

No.	Item	$\widehat{\beta}_S$	$\widehat{\beta}_{\mathrm{SCAD}}$	$\widehat{\beta}_M$	$\widehat{\beta}_{\mathrm{TS}}$	$\widehat{\beta}_F$
Case (I) $c = 0.5$	MSEs	0.3079	0.0457	0.0660	0.0571	1.6105×10^3
		0.1763	0.0206	0.0346	0.0176	1.0940×10^3
		0.1396	0.0481	0.0631	0.0461	4.2049×10^3
		0.1870	0.0196	0.0349	0.0186	5.0183×10^3
		0.1131	0.0517	0.0609	0.0430	6.2615×10^3
	MSPEs	3.4780	1.1896	1.6512	1.0679	3.0499×10^2
Case (I) $c = 0.8$	MSEs	0.1568	0.6191	0.0934	0.0826	1.2494×10^3
		0.6239	0.1060	0.0090	0.0083	1.0456×10^2
		0.8829	0.8173	0.0895	0.1039	2.6368×10^2
		0.5882	0.0919	0.0107	0.0100	7.6452×10^1
		1.0799	0.9829	0.0961	0.0929	1.1610×10^3
	MSPEs	4.7930	2.6700	0.8354	0.7771	1.3223×10^2
Case (II) $c = 0.5$	MSEs	0.4272	0.0660	0.0849	0.0557	4.3002×10^2
		0.6371	0.0318	0.0499	0.0295	3.7893×10^3
		0.4560	0.0715	0.0927	0.0588	1.2784×10^3
		0.5926	0.0306	0.0491	0.0287	6.7354×10^3
		0.9052	0.0734	0.0874	0.0583	2.5047×10^2
	MSPEs	6.8634	1.5096	2.0780	1.2077	5.0464×10^3
Case (II) $c = 0.8$	MSEs	0.6764	0.4263	0.1212	0.0960	1.3904×10^3
		0.9721	0.1060	0.0107	0.0102	4.0743×10^2
		0.6242	0.4756	0.1146	0.1003	1.0498×10^3
		1.0282	0.0954	0.0112	0.0098	5.6031×10^2
		1.3420	0.5474	0.1341	0.1124	9.9632×10^2
	MSPEs	7.9928	2.1165	0.9514	0.8469	2.3110×10^2

Table 2: MSEs on the parameter β, MASEs of $f(\cdot)$ and MSPEs of the two-stage adjusted model (11) compared with the submodel, multistep adjusted model and the full model.

No.	Item	$\widehat{\beta}_S$	$\widehat{\beta}_M$	$\widehat{\beta}_{\mathrm{TS}}$	$\widehat{\beta}_F$
Case (I)	MSEs	0.4352	5.0403	0.3267	2.9753×10^1
		0.6859	1.2820×10^1	0.3328	1.4593×10^1
		1.1152	8.1542	0.3723	1.4391×10^1
		1.8489	7.2055	1.3194	2.4036×10^1
		3.3079	1.6144×10^1	1.9989	4.8575×10^1
	MASEs	3.0887	5.9814	3.0175	3.0633
	MSPEs	4.6047	7.0331×10^1	3.5536	3.9648
Case (II)	MSEs	0.0377	0.6144	0.0191	—[1]
		0.0449	1.0876	0.0305	—
		0.0332	3.7510	0.0246	—
		0.0396	0.4324	0.0238	—
		0.0512	1.1995	0.0335	—
	MASEs	0.4722	0.5220	0.4126	0.4380
	MSPEs	0.9221	9.3068	0.8053	—

[1] "—" denotes the algorithm collapsed and returned no value.

similar trend to the MSEs of the four estimators, while the differences are not very noticeable. (3) Similar to the MSEs of the estimators, the MSPEs of the two-stage adjusted model are the smallest among the four models. (4) In Case (II), the simulation results of models $(\mathrm{I})'$, $(\mathrm{II})'$ and $(\mathrm{III})'$, perform a little better than those in Case (I) because of the correlation structure among the covariates.

In summary, Table 2 indicates that the two-stage adjusted model (11) performs much better than the full model and the multistep adjusted model, and better than the submodel.

4. Some Remarks

In this paper, the main objective is to consistently estimate the parameter of interest β. When estimating the parameter of interest, its bias is mainly determined by the relevant variables, and its variance may be impacted by other variables. Because variable selection much relies on the sparsity of the parameter, when we directly consider the partially linear model, some irrelevant variables with nonzero coefficients may be selected in the final model. This may affect the estimation of the parameter β on its efficiency and stability. Thus based on the prespecified submodel, a two-stage remodeling method is proposed. In the new remodeling procedure, the correlation among the covariates (X, Z) and the sparsity of the regression structure are fully used. So the final model is sufficiently simplified and conditionally unbiased. Based on the simplified model, the estimation and model prediction are significantly improved. Generally, after the first stage the adjusted model is an additive partially linear model. Therefore, the remodeling method can be applied to partially linear regression model with linear regression model as a special case.

From the remodeling procedure, we can see that it can be directly applied to additive partially linear model, in which the nonparametric function $f(U)$ has componentwise additive form. As for general partially linear model with multivariate nonparametric function, we should resort to multivariate nonparametric estimation method. If the dimension of covariate U is high, it may be faced with "the curse of dimensionality".

In the procedure of model simplification, orthogonal series estimation method is used. This is only for technical convenience, because the semiparametric penalized least squares (6) can be easily transformed into parametric penalized least squares (10) and then the theoretic results are obtained. Although other nonparametric methods such as kernel and spline can be used without any essential difficulty, they can not directly achieve this goal. Compared with kernel method, it is somewhat difficult for series method to establish the asymptotic normality result for the nonparametric component $f(U)$ under primitive conditions.

Appendix

A. Some Conditions and Proofs

A.1. Regularity Conditions (C1)–(C6).

(C1) $(\widetilde{Z}, U)$ has finite nondegenerate compact support, denoted as $\mathrm{supp}(\widetilde{Z}, U)$.

(C2) The density function $r_j(t)$ of $\widetilde{Z}(j)$ and $r_0(t)$ of U satisfies $0 < L_1 \le r_j(t) \le L_2 < \infty$ on its support for $0 \le j \le K_0$ for some constants L_1 and L_2, and it is continuously differentiable.

(C3) $G(\widetilde{Z}, U) = E(XX^T \mid \widetilde{Z}, U)$ and $E(\zeta_{K_0}^2 \mid \widetilde{Z}, U)$ are continuous. For given $\widetilde{Z}$ and u, $G(\widetilde{Z}, u)$ is positive definite, and its eigenvalues are bounded.

(C4) $\sup_{(\widetilde{Z},u)\in \mathrm{supp}(\widetilde{Z},U)} E(\|X\|^3 \mid \widetilde{Z} = \widetilde{Z}, U = u) < \infty$, $Ef(U) = 0$, the first two derivatives of $f(\cdot)$ are Lipschitz continuous of order one.

(C5) $b_n = \max_j\{p''_{\lambda_j}(\|\theta_{j0}\|_2) : \theta_{j0} \neq 0\} \to 0$ as $n \to \infty$.

(C6) $\liminf_{n\to\infty} \liminf_{\|\theta_{j0}\|_2 \to 0} \lambda_j^{-1} p'_{\lambda_j}(\|\theta_{j0}\|_2) > 0$ for $j = s+1, \ldots, K_0$ where s satisfies $\gamma^T E(E(Z \mid \widetilde{Z}^{(j)})E(Z^T \mid \widetilde{Z}^{(j)}))\gamma > 0$ for $1 \le j \le s$; $\gamma^T E(E(Z \mid \widetilde{Z}^{(j)})E(Z^T \mid \widetilde{Z}^{(j)}))\gamma = 0$ for $s < j \le K_0$.

Conditions (C1)–(C3) are some regular constraints on the covariates and condition (C4) is some constraints on the regression structure as those in Härdle et al. [16]. Conditions (C5)-(C6) are assumptions on the penalty function which are similar to those used in Fan and Li [5] and Wang et al. [7].

A.2. Proof for Theorem 1. Let $\delta = n^{-r/(2r+1)} + a_n$, $\beta = \beta_0 + \delta T_1$, $\theta = \theta_0 + \delta T_2$, $\nu = \nu_0 + \delta T_3$ and $T = (T_1^T, T_2^T, T_3^T)^T$. Firstly, we shall prove that, $\forall \epsilon > 0$, $\exists C > 0$, $P\{\inf_{\|T\|=C} F(\beta, \theta, \nu) > F(\beta_0, \theta_0, \nu_0)\} \ge 1 - \epsilon$.

Denote $D(\beta, \theta, \nu) = L^{-1}\{F(\beta, \theta, \nu) - F(\beta_0, \theta_0, \nu_0)\}$, then we have

$$\begin{aligned} & D(\beta, \theta, \nu) \\ & = \frac{1}{L}\sum_{i=1}^{n}\Big[\left(T_1^T X_i + T_2^T \Psi(\widetilde{Z}_i) + T_3^T \Psi_0(U_i)\right) \\ & \qquad \times (-2\delta Y_i) + 2\delta\left(\beta_0^T X_i + \theta_0^T \Psi(\widetilde{Z}_i) + \nu_0^T \Psi_0(U_i)\right) \\ & \qquad \times \left(T_1^T X_i + T_2^T \Psi(\widetilde{Z}_i) + T_3^T \Psi_0(U_i)\right) \\ & \qquad + \delta^2 \left(T_1^T X_i + T_2^T \Psi(\widetilde{Z}_i) \right. \\ & \qquad \left. + T_3^T \Psi_0(U_i)\right)^2\Big] \\ & \quad + \frac{n}{L}\sum_{j=1}^{K_0}\left(p_{\lambda_j}(\|\theta_j\|_2) - p_{\lambda_j}(\|\theta_{j0}\|_2)\right) \\ & \ge -\frac{2\delta}{L}\sum_{i=1}^{n}\left(\zeta_{K_0 i} + R(\widetilde{Z}_i, U_i)\right) \\ & \qquad \times \left(T_1^T X_i + T_2^T \Psi(\widetilde{Z}_i) + T_3^T \Psi_0(U_i)\right) \\ & \quad + \frac{\delta^2}{L}\sum_{i=1}^{n}\left(T_1^T X_i + T_2^T \Psi(\widetilde{Z}_i) + T_3^T \Psi_0(U_i)\right)^2 \\ & \quad + \frac{n}{L}\sum_{j=1}^{s}\left(p_{\lambda_j}(\|\theta_j\|_2) - p_{\lambda_j}(\|\theta_{j0}\|_2)\right) \\ & \equiv I_1 + I_2 + I_3, \end{aligned} \tag{A.1}$$

where $R(\widetilde{Z}_i, U_i) = \sum_{j=1}^{K_0} R_j(\widetilde{Z}_i) + R_0(U_i)$ with $R_j(\widetilde{Z}_i) = g_j(\widetilde{Z}_i^{(j)}) - \theta_j^T \Psi_j(\widetilde{Z}_i^{(j)})$, $j = 1, \ldots, K_0$ and $R_0(U_i) = f(U_i) - \nu^T \Psi_0(U_i)$.

By the conditions (C1) and (C2), the maximal squared bias of $g_j(\widetilde{Z}^{(j)})$ is equal to

$$\left\|R_j\left(\widetilde{Z}^{(j)}\right)\right\|^2 = \sum_{k=L+1}^{\infty} \theta_{jk}^2 \le \sum_{k=L+1}^{\infty} \theta_{jk}^2 \left(\frac{k}{L}\right)^{2r} \le ML^{-2r}, \quad \text{(A.2)}$$

so $\|R_j(\widetilde{Z}^{(j)})\| = O(L^{-r})$. Similarly, $\|R_0(U)\| = O(L^{-r})$. Then,

$$\begin{aligned}&\sum_{i=1}^{n} R\left(\widetilde{Z}_i, U_i\right)\left(T_1^T X_i + T_2^T \Psi\left(\widetilde{Z}_i\right) + T_3^T \Psi_0\left(U_i\right)\right)\\ &\quad = O_p\left(nK_0 L^{-r}\|T\|\right).\end{aligned} \quad \text{(A.3)}$$

Noticing that $E(\zeta_{K_0} \mid X, \widetilde{Z}, U) = 0$, by Zhao and Xue [4], we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \zeta_{K_0 i}\left(T_1^T X_i + T_2^T \Psi\left(\widetilde{Z}_i\right) + T_3^T \Psi_0\left(U_i\right)\right) = O_p\left(\|T\|\right). \quad \text{(A.4)}$$

So

$$\begin{aligned}I_1 &= -\frac{2\delta}{L}\left[O_p\left(nK_0 L^{-r}\|T\|\right) + O_p\left(\sqrt{n}\|T\|\right)\right]\\ &= O_p\left(1 + n^{r/(2r+1)} a_n\right)\|T\|.\end{aligned} \quad \text{(A.5)}$$

Similarly, we have

$$\begin{aligned}0 < I_2 &= O_p\left(nL^{-1}\delta^2\|T\|^2\right)\\ &= O_p\left(1 + 2n^{r/(2r+1)} a_n + n^{2r/(2r+1)} a_n^2\right)\|T\|^2.\end{aligned} \quad \text{(A.6)}$$

By properly choosing a sufficiently large C, I_2 dominates I_1 uniformly in $\|T\| = C$.

Using Taylor expansion,

$$\begin{aligned}I_3 = &\sum_{j=1}^{s} \frac{n}{L} p_{\lambda_j}^{'}\left(\left\|\theta_{j0}\right\|_2\right)\left(\left\|\theta_{j0}\right\|_2\right)^{'}\delta T_{2j}\\ &+ \frac{n}{2L}\left\{p_{\lambda_j}^{''}\left(\left\|\theta_{j0}\right\|_2\right)\left[\left(\left\|\theta_{j0}\right\|_2\right)^{'}\delta T_{2j}\right]^2\right\}\\ &\times\left(1 + o\left(1\right)\right)\\ \equiv\ &I_{31} + I_{32}.\end{aligned} \quad \text{(A.7)}$$

By simple calculations, we have that

$$\begin{aligned}\left|I_{31}\right| &\le \frac{n}{L}\delta a_n l_1 \sum_{j=1}^{s}\left\|T_{2j}\right\| \le \sqrt{s}\frac{n}{L}\delta a_n l_1\|T\|\\ &= O_p\left(n^{r/(2r+1)} a_n + n^{2r/(2r+1)} a_n^2\right)\|T\|,\end{aligned} \quad \text{(A.8)}$$

$$\left|I_{32}\right| \le \frac{n}{2L}\delta^2 b_n l_2 \sum_{j=1}^{s}\left\|T_{2j}\right\|^2 \le \frac{n}{2L}\delta^2 b_n l_2\|T\|^2,$$

where l_1 and l_2 are some positive constants. We can find that I_{31} is also dominated by I_2 uniformly in $\|T\| = C$, and under the condition (C5), we have

$$0 < \left|I_{32}\right| \le o_p\left(1 + 2n^{r/(2r+1)} a_n + n^{2r/(2r+1)} a_n^2\right)\|T\|^2. \quad \text{(A.9)}$$

Hence, by choosing a sufficiently large C, $P\{\inf_{\|T\|=C} F(\beta, \theta, \nu) > F(\beta_0, \theta_0, \nu_0)\} \ge 1 - \epsilon$, which implies that with probability at least $1 - \epsilon$ there exists a local minimum of $F(\beta, \theta, \nu)$ in the ball $\{\beta_0 + \delta T_1 : \|T_1\| \le C\}$. Denote the local minimizer as $\widehat{\beta}$, then

$$\left\|\widehat{\beta} - \beta_0\right\| = O_p\left(\delta\right) = O_p\left(n^{-r/(2r+1)} + a_n\right). \quad \text{(A.10)}$$

With the same argument as above, there exists a local minimum in the ball $\{\theta_0 + \delta T_2 : \|T_2\| \le C\}$, and the local minimizer $\widehat{\theta}$ satisfies that

$$\left\|\widehat{\theta} - \theta_0\right\| = O_p\left(n^{-r/(2r+1)} + a_n\right). \quad \text{(A.11)}$$

For the nonparametric component $g(\cdot)$, noticing that

$$\begin{aligned}\left\|\widehat{g}_j - g_{j0}\right\|^2 &= E\left\{\widehat{g}_j\left(\widetilde{Z}^{(j)}\right) - g_{j0}\left(\widetilde{Z}^{(j)}\right)\right\}^2\\ &= E\left\{\Psi_j\left(\widetilde{Z}^{(j)}\right)\widehat{\theta}_j - \Psi_j\left(\widetilde{Z}^{(j)}\right)\theta_{j0} + R_{j0}\left(\widetilde{Z}^{(j)}\right)\right\}^2\\ &\le 2E\left\{\Psi_j\left(\widetilde{Z}^{(j)}\right)\widehat{\theta}_j - \Psi_j\left(\widetilde{Z}^{(j)}\right)\theta_{j0}\right\}^2\\ &\quad + 2E\left\{R_{j0}\left(\widetilde{Z}^{(j)}\right)\right\}^2\\ &= 2\left(\widehat{\theta}_j - \theta_{j0}\right)^T\left(\widehat{\theta}_j - \theta_{j0}\right) + 2E\left\{R_{j0}\left(\widetilde{Z}^{(j)}\right)\right\}^2,\end{aligned} \quad \text{(A.12)}$$

it is known that $\|R_j(\widetilde{Z}^{(j)})\| = O(L^{-r})$, so

$$E\left\{R_{j0}\left(\widetilde{Z}^{(j)}\right)\right\}^2 = O_p\left(n^{-2r/(2r+1)}\right). \quad \text{(A.13)}$$

Thus, we get

$$\left\|\widehat{g}_j - g_{j0}\right\| = O_p\left(n^{-r/(2r+1)} + a_n\right). \quad \text{(A.14)}$$

Similarly, there exists a local minimizer $\widehat{\nu}$ satisfies that $\|\widehat{\nu} - \nu_0\| = O_p(n^{-r/(2r+1)} + a_n)$. Then we can get $\|\widehat{f} - f_0\| = O_p(n^{-r/(2r+1)} + a_n)$.

A.3. Proof for Theorem 2. When $\lambda_{\max} \to 0$, $a_n = 0$ for large n by the form of $p_{\lambda}^{'}(w)$. Then by Theorem 1, it is sufficient to show that: with probability tending to 1 as $n \to \infty$, for any β, it satisfies $\|\beta - \beta_0\| = O_p(n^{-r/(2r+1)})$, θ_j satisfies $\|\theta_j - \theta_{j0}\| = O_p(n^{-r/(2r+1)})$ with $j = 1, \ldots, s$, and ν satisfies $\|\nu - \nu_0\| = O_p(n^{-r/(2r+1)})$, for some small $\iota_n = Cn^{-r/(2r+1)}$,

$$\begin{aligned}&\frac{\partial F\left(\beta, \theta, \nu\right)}{\partial \theta_j} > 0, \quad \text{for } 0 < \theta_j < \iota_n,\ j = s+1, \ldots, K_0,\\ &\frac{\partial F\left(\beta, \theta, \nu\right)}{\partial \theta_j} < 0, \quad \text{for } -\iota_n < \theta_j < 0,\ j = s+1, \ldots, K_0.\end{aligned} \quad \text{(A.15)}$$

So the minimizer of $F(\beta, \theta, \nu)$ is obtained at $\theta_j = 0,\ j = s + 1, \ldots, K_0$.

In fact,

$$\begin{aligned}
\frac{\partial F(\beta, \theta, \nu)}{\partial \theta_j}
&= -2 \sum_{i=1}^{n} \Psi_j\left(\widetilde{Z}_i^{(j)}\right)\left(Y_i - \beta^T X_i - \theta^T \Psi\left(\widetilde{Z}_i\right) - \nu^T \Psi_0(U_i)\right) \\
&\quad + n p'_{\lambda_j}\left(\|\theta_j\|_2\right)\left(\|\theta_j\|_2\right)' \\
&= -2 \sum_{i=1}^{n} \Psi_j\left(\widetilde{Z}_i^{(j)}\right)\left(\zeta_{K_0 i} + R\left(\widetilde{Z}_i, U_i\right)\right) \\
&\quad - 2 \sum_{i=1}^{n} \Psi_j\left(\widetilde{Z}_i^{(j)}\right) X_i^T (\beta_0 - \beta) \\
&\quad - 2 \sum_{i=1}^{n} \Psi_j\left(\widetilde{Z}_i^{(j)}\right) \Psi^T\left(\widetilde{Z}_i\right)(\theta_0 - \theta) \\
&\quad - 2 \sum_{i=1}^{n} \Psi_j\left(\widetilde{Z}_i^{(j)}\right) \Psi_0^T(U_i)(\nu_0 - \nu) \\
&\quad + n p'_{\lambda_j}\left(\|\theta_j\|_2\right)\left(\|\theta_j\|_2\right)' \\
&= n\lambda_j \left\{ O_p\left(\lambda_j^{-1} n^{-r/(2r+1)}\right) + \lambda_j^{-1} p'_{\lambda}\left(\|\theta_j\|_2\right) \frac{\theta_j}{\|\theta_j\|_2} \right\}.
\end{aligned} \tag{A.16}$$

Under the conditions $\liminf_{n \to \infty} \liminf_{\|\theta_{j0}\|_2 \to 0} \lambda_j^{-1} p'_{\lambda_j} \cdot (\|\theta_{j0}\|_2) = C > 0$ and $\lambda_j n^{r/(2r+1)} > \lambda_{\min} n^{r/(2r+1)} \to \infty$, then $\partial F(\beta, \theta, \nu)/\partial \theta_j = O_p(n\lambda_j(\theta_j/\|\theta_j\|_2))$. So the sign of the derivative is determined by θ_j.

So with probability tending to 1, $\widehat{\theta}_j = 0$, $j = s + 1, \ldots, K_0$. Then under $\sup_Z \|\Psi_j(\widetilde{Z}^{(j)})\| = O(1)$, $\widehat{g}_j(\widetilde{Z}^{(j)}) = \widehat{\theta}_j^T \Psi_j(\widetilde{Z}^{(j)}) \equiv 0$, $j = s + 1, \ldots, K_0$.

A.4. Proof for Theorem 4. By Theorems 1 and 2, we know that, as $n \to \infty$, with probability tending to 1, $F(\beta, \theta, \nu)$ attains the local minimum value at $\widehat{\beta}$ and $(\widehat{\theta}^{*T}, 0)^T$ and $\widehat{\nu}$. Let $F_{1n}(\beta, \theta, \nu) = \partial F(\beta, \theta, \nu)/\partial \beta$, $F_{2n}(\beta, \theta, \nu) = \partial F(\beta, \theta, \nu)/\partial \theta^*$ and $F_{3n}(\beta, \theta, \nu) = \partial F(\beta, \theta, \nu)/\partial \nu$, then

$$\begin{aligned}
&\frac{1}{n} F_{1n}\left(\widehat{\beta}, \left(\widehat{\theta}^{*T}, 0\right)^T, \widehat{\nu}\right) \\
&\quad = \frac{1}{n} \sum_{i=1}^{n} X_i \left(Y_i - \widehat{\beta}^T X_i - \widehat{\theta}^{*T} \Psi_i^* - \widehat{\nu}^T \Psi_{0i}\right) = 0,
\end{aligned} \tag{A.17}$$

$$\begin{aligned}
&\frac{1}{n} F_{2n}\left(\widehat{\beta}, \left(\widehat{\theta}^{*T}, 0\right)^T, \widehat{\nu}\right) \\
&\quad = \frac{1}{n} \sum_{i=1}^{n} \Psi_i^* \left(Y_i - \widehat{\beta}^T X_i - \widehat{\theta}^{*T} \Psi_i^* - \widehat{\nu}^T \Psi_{0i}\right) \\
&\qquad + \sum_{j=1}^{s} p'_{\lambda_j}\left(\|\widehat{\theta}_j\|_2\right) \frac{\widehat{\theta}_j}{\|\widehat{\theta}_j\|_2} = 0,
\end{aligned} \tag{A.18}$$

$$\begin{aligned}
&\frac{1}{n} F_{3n}\left(\widehat{\beta}, \left(\widehat{\theta}^{*T}, 0\right)^T, \widehat{\nu}\right) \\
&\quad = \frac{1}{n} \sum_{i=1}^{n} \Psi_{0i} \left(Y_i - \widehat{\beta}^T X_i - \widehat{\theta}^{*T} \Psi_i^* - \widehat{\nu}^T \Psi_{0i}\right) = 0.
\end{aligned} \tag{A.19}$$

From (A.17), it yields that

$$\begin{aligned}
&\frac{1}{n} \sum_{i=1}^{n} X_i \left(\left(\beta_0 - \widehat{\beta}\right)^T X_i + \left(\theta_0^* - \widehat{\theta}^*\right)^T \Psi_i^* \right. \\
&\qquad \left. + (\nu_0 - \widehat{\nu})^T \Psi_{0i} + R^*\left(\widetilde{Z}_i, U_i\right) + \zeta_{K_0 i} \right) = 0,
\end{aligned} \tag{A.20}$$

where $R^*(\widetilde{Z}_i, U_i) = \sum_{j=1}^{s} R_j^*(\widetilde{Z}_i) + R_0(U_i)$. Applying the Taylor expansion, we get

$$\begin{aligned}
&p'_{\lambda_j}\left(\|\widehat{\theta}_j\|_2\right) \\
&\quad = p'_{\lambda_j}\left(\|\widehat{\theta}_{j0}\|_2\right) + \left\{ p''_{\lambda_j}\left(\|\widehat{\theta}_{j0}\|_2\right) \frac{\widehat{\theta}_j}{\|\widehat{\theta}_j\|_2} + o_p(1) \right\} \left(\widehat{\theta}^* - \theta_0^*\right).
\end{aligned} \tag{A.21}$$

Furthermore, condition (C5) implies that $p''_{\lambda_j}(\|\widehat{\theta}_{j0}\|_2) = o_p(1)$, and noting that $p'_{\lambda_j}(\|\widehat{\theta}_{j0}\|_2) = 0$ as $\lambda_{\max} \to 0$, then $p'_{\lambda_j}(\|\widehat{\theta}_j\|_2) = o_p(\widehat{\theta}^* - \theta_0^*)$. So from (A.18), it yields

$$\begin{aligned}
&\frac{1}{n} \sum_{i=1}^{n} \Psi_i^* \left(\left(\beta_0 - \widehat{\beta}\right)^T X_i + \left(\theta_0^* - \widehat{\theta}^*\right)^T \Psi_i^* \right. \\
&\qquad \left. + (\nu_0 - \widehat{\nu})^T \Psi_{0i} + R^*\left(\widetilde{Z}_i, U_i\right) + \zeta_{K_0 i} \right) \\
&\quad + o_p\left(\theta_0^* - \widehat{\theta}^*\right) = 0.
\end{aligned} \tag{A.22}$$

Let $\Phi_n = n^{-1} \sum_{i=1}^{n} \Psi_i^* \Psi_i^{*T}$, $\Gamma_n = n^{-1} \sum_{i=1}^{n} \Psi_i^* X_i^T$ and $\Pi_n = n^{-1} \sum_{i=1}^{n} \Psi_i^* \Psi_{0i}^T$, then we have

$$\begin{aligned}
\widehat{\theta}^* - \theta_0^* &= \left[\Phi_n + o_p(1)\right]^{-1} \\
&\quad \times \left\{ \Gamma_n\left(\beta_0 - \widehat{\beta}\right) + \Pi_n(\nu_0 - \widehat{\nu}) \right. \\
&\quad \left. \times \frac{1}{n} \sum_{i=1}^{n} \Psi_i^* \left(R^*\left(\widetilde{Z}_i\right) + \zeta_{K_0 i}\right) \right\}.
\end{aligned} \tag{A.23}$$

Substituting (A.23) into (A.20), it yields

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} X_i \left\{ \left(\Psi_{0i} - \Pi_n^T \Phi_n^{-1}\Psi_i^*\right)^T \left(\widehat{\beta}-\beta_0\right)\right.\\
&\qquad\qquad \left. + \left(X_i - \Gamma_n^T\Phi_n^{-1}\Psi_i^*\right)^T \left(\widehat{\nu}-\nu_0\right)\right\}\\
&\quad + o_p\left(\widehat{\beta}-\beta_0\right) + o_p\left(\widehat{\nu}-\nu_0\right) \qquad\qquad (\text{A.24})\\
&= \frac{1}{n}\sum_{i=1}^{n} X_i \left\{\zeta_{K_0 i} + R^*\left(\widetilde{Z}_i, U_i\right)\right.\\
&\qquad\qquad \left. -\Psi_i^{*T}\left(\Phi_n^{-1} + o_p(1)\right)\Lambda_n\right\},
\end{aligned}$$

where $\Lambda_n = n^{-1}\sum_{i=1}^{n} \Psi_i^*(R^*(\widetilde{Z}_i, U_i) + \zeta_{K_0 i})$.

From (A.19), it yields that

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} \Psi_{0i}\left(\left(\beta_0-\widehat{\beta}\right)^T X_i + \left(\theta_0^* - \widehat{\theta}^*\right)^T \Psi_i^*\right.\\
&\qquad \left. + \left(\nu_0 - \widehat{\nu}\right)^T \Psi_{0i} + R^*\left(\widetilde{Z}_i, U_i\right) + \zeta_{K_0 i}\right) = 0.
\end{aligned} \qquad (\text{A.25})$$

Substituting (A.23) into (A.25), it yields

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} \Psi_{0i}\left\{\left(X_i - \Gamma_n^T\Phi_n^{-1}\Psi_i^*\right)^T\left(\widehat{\beta}-\beta_0\right)\right.\\
&\qquad \left. + \left(\Psi_{0i} - \Pi_n^T\Phi_n^{-1}\Psi_i^*\right)^T\left(\widehat{\nu}-\nu_0\right)\right\}\\
&\quad + o_p\left(\widehat{\beta}-\beta_0\right) + o_p\left(\widehat{\nu}-\nu_0\right)\\
&= \frac{1}{n}\sum_{i=1}^{n} \Psi_{0i}\left\{\zeta_{K_0 i} + R^*\left(\widetilde{Z}_i, U_i\right) - \Psi_i^{*T}\left(\Phi_n^{-1} + o_p(1)\right)\Lambda_n\right\}.
\end{aligned} \qquad (\text{A.26})$$

Noting that

$$\begin{aligned}
&n^{-1}\sum_{i=1}^{n} \Pi_n^T\Phi_n^{-1}\Psi_i^*\left\{\Psi_{0i}^T - \Psi_i^{*T}\Phi_n^{-1}\Pi_n\right\} = 0,\\
&n^{-1}\sum_{i=1}^{n} \Pi_n^T\Phi_n^{-1}\Psi_i^*\left\{\zeta_{K_0 i} + R^*\left(\widetilde{Z}_i, U_i\right) - \Psi_i^{*T}\Phi_n^{-1}\Lambda_n\right\} = 0,
\end{aligned} \qquad (\text{A.27})$$

Equation (A.26) can be rewritten as

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} \Psi_{0i}\breve{X}_i^T\left(\widehat{\beta}-\beta_0\right) + o_p\left(\widehat{\beta}-\beta_0\right)\\
&\quad + \frac{1}{n}\sum_{i=1}^{n} \breve{\Psi}_{0i}\breve{\Psi}_{0i}^T\left(\widehat{\nu}-\nu_0\right)\\
&\quad + o_p\left(\widehat{\nu}-\nu_0\right)\\
&= \frac{1}{n}\sum_{i=1}^{n} \breve{\Psi}_{0i}\left\{\zeta_{K_0 i} + R^*\left(\widetilde{Z}_i, U_i\right) - \Psi_i^{*T}\left(\Phi_n^{-1} + o_p(1)\right)\Lambda_n\right\},
\end{aligned} \qquad (\text{A.28})$$

where $\breve{X}_i = X_i - \Gamma_n^T\Phi_n^{-1}\Psi_i^*$, $\breve{\Psi}_{0i} = \Psi_{0i} - \Pi_n^T\Phi_n^{-1}\Psi_i^*$. Let $\Xi_n = n^{-1}\sum_{i=1}^{n}\breve{\Psi}_{0i}\breve{\Psi}_{0i}^T$, then we have

$$\begin{aligned}
\widehat{\nu} - \nu_0 &= \Xi_n^{-1}\frac{1}{n}\sum_{i=1}^{n}\Psi_{0i}\breve{X}_i^T\left(\beta_0 - \widehat{\beta}\right)\\
&\quad + \Xi_n^{-1}\frac{1}{n}\sum_{i=1}^{n}\breve{\Psi}_{0i}\left(\zeta_{K_0 i} + R^*\left(\widetilde{Z}_i, U_i\right) - \Psi_i^*\Phi_n^{-1}\Lambda_n\right)\\
&\quad + o_p\left(\widehat{\beta}-\beta_0\right).
\end{aligned} \qquad (\text{A.29})$$

Substituting (A.29) into (A.24), and noting that

$$\begin{aligned}
&n^{-1}\sum_{i=1}^{n}\Gamma_n^T\Phi_n^{-1}\Psi_i^*\left\{X_i - \Psi_i^{*T}\Phi_n^{-1}\Gamma_n\right\} = 0,\\
&n^{-1}\sum_{i=1}^{n}\Gamma_n^T\Phi_n^{-1}\Psi_i^*\left\{\zeta_{K_0 i} + R^*\left(\widetilde{Z}_i\right) - \Psi_i^{*T}\Phi_n^{-1}\Lambda_n\right\} = 0,\\
&n^{-1}\sum_{i=1}^{n}\breve{\Psi}_{0i}X_i^T = n^{-1}\sum_{i=1}^{n}\Psi_{0i}\breve{X}_i^T,
\end{aligned} \qquad (\text{A.30})$$

it is easy to show that

$$\begin{aligned}
&\left(\breve{\Phi}_n - \Upsilon_n^T\Xi_n^{-1}\Upsilon_n + o_p(1)\right)\sqrt{n}\left(\widehat{\beta}-\beta_0\right)\\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\breve{X}_i - \Upsilon_n^T\Xi_n^{-1}\breve{\Psi}_{0i}\right)\\
&\qquad \times\left(\zeta_{K_0 i} + R^*\left(\widetilde{Z}_i, U_i\right) - \Psi_i^{*T}\left[\Phi_n^{-1} + o_p(1)\right]\Lambda_n\right)\\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\widetilde{X_i}\zeta_{K_0 i}\\
&\quad - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\widetilde{X_i}\Psi_i^{*T}\left[\Phi_n^{-1} + o_p(1)\right]\Lambda_n\\
&\quad + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\widetilde{X_i}R^*\left(\widetilde{Z}_i, U_i\right)\\
&\equiv I_1 + I_2 + I_3,
\end{aligned} \qquad (\text{A.31})$$

where $\breve{\Phi}_n = n^{-1}\sum_{i=1}^{n}\breve{X}_i\breve{X}_i^T$, $\Upsilon_n = n^{-1}\sum_{i=1}^{n}\breve{\Psi}_{0i}X_i^T$, $\widetilde{X_i} = \breve{X}_i - \Upsilon_n^T\Xi_n^{-1}\breve{\Psi}_{0i}$.

Using the Central Limit Theorem, we can obtain

$$I_1 \xrightarrow{\mathscr{D}} N\left(0, \sigma_{K_0}^2\Sigma_0\right), \qquad (\text{A.32})$$

where "$\xrightarrow{\mathscr{D}}$" means the convergence in distribution and

$$\begin{aligned}
\Sigma_0 &= E\left(XX^T\right) - E\left(X\Psi^{*T}\right)E^{-1}\left(\Psi^*\Psi^{*T}\right)E\left(\Psi^*X^T\right)\\
&\quad - E\left(X\breve{\Psi}_0^T\right)E^{-1}\left(\breve{\Psi}_0\breve{\Psi}_0^T\right)E\left(\breve{\Psi}_0X^T\right).
\end{aligned} \qquad (\text{A.33})$$

In addition, noting that $\sum_{i=1}^{n} \breve{X}_i \Psi_i^{*T} = 0$ and $\sum_{i=1}^{n} \breve{\Psi}_{0i} \Psi_i^{*T} = 0$, we have $I_2 = 0$. Furthermore, we have

$$\begin{aligned} I_3 = {} & \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ X_i - E\left(\Gamma_n^T\right) E^{-1}\left(\Phi_n\right) \Psi_i^* \right\} R^*\left(\widetilde{Z}_i, U_i\right) \\ & + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ E\left(\Gamma_n^T\right) E^{-1}\left(\Phi_n\right) - \Gamma_n^T \Phi_n^{-1} \right\} \Psi_i^* R^*\left(\widetilde{Z}_i, U_i\right) \\ & - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Upsilon_n^T \Xi_n^{-1} \breve{\Psi}_{0i} R^*\left(\widetilde{Z}_i, U_i\right) \\ \equiv {} & I_{31} + I_{32} + I_{33}. \end{aligned} \tag{A.34}$$

Invoking $E\{[X_i - E(\Gamma_n^T)E^{-1}(\Phi_n)\Psi_i^*]\Psi_i^{*T}\} = 0$, then by Zhao and Xue [4], we have

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(X_i - E\left(\Gamma_n^T\right) E^{-1}\left(\Phi_n\right) \Psi_i^* \right) \Psi_i^{*T} = O_p\left(1\right). \tag{A.35}$$

This together with $\|\Psi_j(\widetilde{Z}^{(j)})\| = O(1)$ and $\|R(\widetilde{Z}, U)\| = o(1)$, we get $I_{31} = o_p(1)$. Similarly, $I_{32} = o_p(1)$. Noting that $(1/\sqrt{n}) \sum_{i=1}^{n} \Upsilon_n^T \Xi^{-1} \Psi_{0i}^* \Psi_i^{*T} = 0$, so as above, we have $I_{33} = o_p(1)$. Hence, we get that $I_3 = o_p(1)$.

By the law of large numbers, we have $(1/n) \sum_{i=1}^{n} \widetilde{X}_i \widetilde{X}_i^T \xrightarrow{P} \Sigma_0$, where "$\xrightarrow{P}$" means the convergence in probability. Then using the Slutsky theorem, we get $\sqrt{n}(\widehat{\beta} - \beta_0) \xrightarrow{\mathscr{D}} N(0, \sigma_{K_0}^2 \Sigma_0^{-1})$.

Acknowledgment

Lin and Zeng's research are supported by NNSF projects (11171188, 10921101, and 11231005) of China, NSF and SRRF projects (ZR2010AZ001 and BS2011SF006) of Shandong Province of China and K C Wong-HKBU Fellowship Programme for Mainland China Scholars 2010-11. Wang's research is supported by NSF project (ZR2011AQ007) of Shandong Province of China.

References

[1] X. T. Shen, H.-C. Huang, and J. Ye, "Inference after model selection," *Journal of the American Statistical Association*, vol. 99, no. 467, pp. 751–762, 2004.

[2] Y. Gai, L. Lin, and X. Wang, "Consistent inference for biased sub-model of high-dimensional partially linear model," *Journal of Statistical Planning and Inference*, vol. 141, no. 5, pp. 1888–1898, 2011.

[3] Y. Zeng, L. Lin, and X. Wang, "Multi-step-adjustment consistent inference for biased sub-model of multidimensional linear regression," *Acta Mathematica Scientia*, vol. 32, no. 6, pp. 1019–1031, 2012 (Chinese).

[4] P. Zhao and L. Xue, "Variable selection for semiparametric varying coefficient partially linear models," *Statistics & Probability Letters*, vol. 79, no. 20, pp. 2148–2157, 2009.

[5] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.

[6] L. Wang, G. Chen, and H. Li, "Group SCAD regression analysis for microarray time course gene expression data," *Bioinformatics*, vol. 23, no. 12, pp. 1486–1494, 2007.

[7] L. Wang, H. Li, and J. Z. Huang, "Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements," *Journal of the American Statistical Association*, vol. 103, no. 484, pp. 1556–1569, 2008.

[8] E. F. Simas Filho and J. M. Seixas, "Nonlinear independent component analysis: theoretical review and applications," *Learning and Nonlinear Models*, vol. 5, no. 2, pp. 99–120, 2007.

[9] J. Fan, Y. Feng, and R. Song, "Nonparametric independence screening in sparse ultra-high-dimensional additive models," *Journal of the American Statistical Association*, vol. 106, no. 494, pp. 544–557, 2011.

[10] C. J. Stone, "Optimal global rates of convergence for nonparametric regression," *The Annals of Statistics*, vol. 10, no. 4, pp. 1040–1053, 1982.

[11] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, John Wiley & Sons, 3rd edition, 2003.

[12] P. Rütimann and P. Bühlmann, "High dimensional sparse covariance estimation via directed acyclic graphs," *Electronic Journal of Statistics*, vol. 3, pp. 1133–1160, 2009.

[13] T. Cai and W. D. Liu, "Adaptive thresholding for sparse covariance matrix estimation," *Journal of the American Statistical Association*, vol. 106, no. 494, pp. 672–684, 2011.

[14] H. Zou, "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1418–1429, 2006.

[15] A. Hyvärinen and E. Oja, "A fast fixed-point algorithm for independent component analysis," *Neural Computation*, vol. 9, no. 7, pp. 1483–1492, 1997.

[16] W. Härdle, H. Liang, and J. T. Gao, *Partially Linear Models*, Physica, Heidelberg, Germany, 2000.