Statistica Sinica **22** (2012), 531-554
doi:http://dx.doi.org/10.5705/ss.2010.134

PROFILED FORWARD REGRESSION FOR ULTRAHIGH DIMENSIONAL VARIABLE SCREENING IN SEMIPARAMETRIC PARTIALLY LINEAR MODELS

Hua Liang, Hansheng Wang and Chih-Ling Tsai

University of Rochester, Peking University and University of California, at Davis

Abstract: In partially linear model selection, we develop a profiled forward regression (PFR) algorithm for ultrahigh dimensional variable screening. The PFR algorithm effectively combines the ideas of nonparametric profiling and forward regression. This allows us to obtain a uniform bound for the absolute difference between the profiled predictors and their estimators. Based on this finding, we are able to show that the PFR algorithm uncovers all relevant variables within a few fairly short steps. Numerical studies are presented to illustrate the performance of the proposed method.

Key words and phrases: Forward regression, partially linear model, profiled forward regression, screening consistency, ultrahigh dimensional predictor.

1. Introduction

In regression analysis, the linear model has been commonly used to link a response variable to explanatory variables for data analysis. One major reason for this is that its resulting ordinary least squares (OLS) estimates have a closed form that is easy to compute. However, there are two scenarios in which OLS does not apply: (i) there is an additional explanatory variable that is not easily parameterized, and (ii) the linear model is appropriate but the number of linear predictors d is greater than the sample size n. As a result, the closed form breaks down, and the computation of parameter estimates under these scenarios is not straightforward.

In the first scenario, Heckman (1986); Engle et al. (1986); Speckman (1988) and Wahba (1984) considered the partially linear model,

$$Y = X^\top \beta + g(U) + \varepsilon, \tag{1.1}$$

where $Y \in \mathbb{R}^1$ is the response variable, $X \in \mathbb{R}^d$ is a predictor vector, $\beta \in \mathbb{R}^d$ is an unknown parameter vector, U is an univariate explanatory variable in $[0, 1]$ (for simplicity), and $g(U)$ is an unknown smooth function of U. We also assume that $(X^\top, U)^\top$ and ε are independent. This model is important in the context of semiparametric regression. Detailed information on parameter estimators and their properties can be found in Härdle, Liang, and Gao (2000), while useful discussions on variable selection and parameter shrinkage are given in Bunea (2004) and Xie and Huang (2009), respectively. It is noteworthy that although Xie and Huang studied high-dimensional partially linear models (i.e., $d \to \infty$ as $n \to \infty$), they required $d^2/n \to 0$ as $n \to \infty$.

Under the second scenario, Wold (1966) proposed a partial least squares algorithm that has been widely used in the field of chemometrics. However, the theoretical properties of partial least squares estimates are not well established. Furthermore, due to technology developments as well as theoretical and practical demands across various fields (e.g., engineering, medicine, and business), high-dimensional data analysis has come to play an increasingly critical role (see Fan and Li (2006)). Hence, it is not surprising that researchers have recently proposed novel approaches to analyze ultrahigh (or high) dimensional data with $d \gg n$. Useful references can be found in Fan and Lv (2008); Candes and Tao (2007); Paul et al. (2008) and Wang (2009). All of these studies focus primarily on linear regression models. Lately, Fan, Feng, and Song (2010) extended Fan and Lv's (2008) sure independence screening (SIS) approach to a general parametric model. Moreover, Witten and Tibshirani (2009) developed the covariance-regularized regression for generalized linear models. As a result, we are able to analyze high-dimensional data for broad parametric models.

In practice, we may encounter the challenging situation where both scenarios appear. This motives us to study ultrahigh dimensional partial linear models. To this end, we employ Fan and Huang's (2005) profile least squares approach to convert the partial linear model (1.1) to the classical linear regression model. In addition, we obtain a uniform bound of the absolute difference between the profiled predictors and their estimators when $d \gg n$. This finding allows us to apply Wang's (2009) forward regression (FR) algorithm to develop a profiled forward regression (PFR) procedure for ultrahigh dimensional screening. We show that the FPR algorithm is able to detect relevant predictors within a limited number of steps. Moreover, we obtain a better detecting rate than that of Wang (2009) without imposing the normality assumption on predictors.

The rest of the article is organized as follows. Section 2 introduces the PFR algorithm, whose asymptotic properties are presented in Section 3. Extensive numerical studies are reported in Section 4, while a short discussion is given in Section 5. All technical proofs are left to the Appendix.

2. Profiled Forward Regression

2.1. Model and notations

Let (X_i, U_i, Y_i) be independent and identically distributed as (X, U, Y) for

$1 \le i \le n$ and $X_i = (X_{i1}, \dots, X_{id})^\top \in \mathbb{R}^d$. Let $\mathbb{Y} = (Y_1, \dots, Y_n)^\top \in \mathbb{R}^n$ be the response vector and $\mathbb{X} = (X_1, \dots, X_n)^\top \in \mathbb{R}^{n \times d}$ the matrix of explanatory variables. We then refer to X_{ij} as a relevant (irrelevant) predictor if $\beta_j \neq 0$ ($\beta_j = 0$), and use a generic notation $\mathcal{M} = \{j_1, \dots, j_{d^*}\}$ to denote an arbitrary model with $X_{ij_1}, \dots, X_{ij_{d^*}}$ as relevant predictors. Accordingly, $\mathcal{M}_F = \{1, \dots, d\}$ and $\mathcal{M}_T = \{j : \beta_j \neq 0\}$ represent the full model and the true model, respectively. We denote the size of model $\mathcal{M}$ (i.e., the number of predictors in model $\mathcal{M}$) by $|\mathcal{M}|$. Hence, $|\mathcal{M}_F| = d$ and $|\mathcal{M}_T| = d_0$, where d_0 is the number of relevant predictors in the true model. Moreover, for any candidate model $\mathcal{M}$, the notation $X_{i(\mathcal{M})} = \{X_{ij} : j \in \mathcal{M}\}$ stands for the subvector of X_i that yields the submatrix $\mathbb{X}_{(\mathcal{M})}$ of $\mathbb{X}$.

2.2. Profiled responses and predictors

In regression analysis, the profile least squares approach is useful to convert the semiparametric model to the least squares setting Fan and Huang (2005). Following this approach, we can easily verify that

$$Y_i - E(Y_i|U_i) = \sum_{j=1}^{d} \beta_j \Big\{ X_{ij} - E(X_{ij}|U_i) \Big\} + \varepsilon_i. \tag{2.1}$$

Then, we define the profiled response and the profiled predictor as $Y_i^* = Y_i - E(Y_i|U_i)$ and $X_i^* = X_i - E(X_i|U_i) = (X_{i1}^*, \dots, X_{id}^*)^\top \in \mathbb{R}^d$, respectively, where $X_{ij}^* = X_{ij} - E(X_{ij}|U_i)$ for $i = 1, \dots, n$ and $j = 1, \dots, d$. As a result, the partial linear model (1.1) reduces to the classical linear regression model

$$Y_i^* = X_i^{*\top} \beta + \varepsilon_i, \tag{2.2}$$

in which the variance of X_i^* is denoted by Σ. To implement the linear model (2.2) in practice, however, the unknown functions $E(Y_i|U_i)$ and $E(X_i|U_i)$ need to be estimated nonparametrically. For this purpose, we employ the local linear regression technique (Fan and Gijbels (1996)). To illustrate this process, we estimate $E(Y_i|U_i)$ as follows. Let

$$Q(\alpha_{1i}, \alpha_{2i}) = \sum_{j=1}^{n} \Big\{ Y_j - \alpha_{1i} - \alpha_{2i}(U_j - U_i) \Big\}^2 K_h(U_i - U_j),$$

where $K_h(u) = K(u/h)/h$, $K(\cdot)$ is a density function symmetric about 0, and $h > 0$ is a bandwidth. Let $(\hat{\alpha}_{1i}, \hat{\alpha}_{2i}) = \operatorname{argmin}_{\alpha_{1i}, \alpha_{2i}} Q(\alpha_{1i}, \alpha_{2i})$ and take $U_{ji} = U_j - U_i$. Then $E(Y_i|U_i)$ can be estimated by $\hat{\alpha}_{1i} = s_i^\top \mathbb{Y}$, where $s_i = (s_{i1}, \dots, s_{in})^\top \in \mathbb{R}^n$ and

$s_{ij} =$

$$\left[\Big\{\sum_{j=1}^n U_{ji}^2 K_h(U_{ji})\Big\}\Big\{\sum_{j=1}^n Y_j K_h(U_{ji})\Big\} - \Big\{\sum_{j=1}^n U_{ji} K_h(U_{ji})\Big\}\Big\{\sum_{j=1}^n U_{ji} Y_j K_h(U_{ji})\Big\}\right]$$
$$\times\left[\Big\{\sum_{j=1}^n U_{ji}^2 K_h(U_{ji})\Big\}\Big\{\sum_{j=1}^n K_h(U_{ji})\Big\} - \Big\{\sum_{j=1}^n U_{ji} K_h(U_{ji})\Big\}\Big\{\sum_{j=1}^n U_{ji} K_h(U_{ji})\Big\}\right]^{-1}.$$

Similarly, one can estimate $E(X_{ij}|U_i)$ by $s_i^\top \mathbb{X}_j$, where $\mathbb{X}_j$ is the jth column of $\mathbb{X}$. Accordingly, the profiled response and predictors can be estimated by $\widehat{\mathbb{Y}} = (\widehat{Y}_1, \ldots, \widehat{Y}_n)^{\mathrm{T}} = (I_n - S)\mathbb{Y}$ and $\widehat{\mathbb{X}} = (\widehat{X}_1, \ldots, \widehat{X}_n)^{\mathrm{T}} = (I_n - S)\mathbb{X}$, respectively, where $I_n \in \mathbb{R}^{n\times n}$ is an identity matrix and $S = (s_1, \ldots, s_n)^\top \in \mathbb{R}^{n\times n}$ is the smoothing matrix. Note that we use S to represent the smoothing matrix, which is a slight abuse of notation.

2.3. The PFR algorithm

In linear regression modeling, Wang (2009) proposed a forward regression (FR) method for ultrahigh dimensional variable screening. Based on the profiled estimators of $\widehat{\mathbb{Y}}$ and $\widehat{\mathbb{X}}$, we adopt FR to develop the PFR algorithm for model (1.1). For the sake of completeness, the algorithm is briefly described below.

Step (1) (*Initialization*). Initiate a null model as $\mathcal{M}^{(0)} = \emptyset$.

Step (2) (*Forward Regression*).

(2.A) (*Evaluation*). In the kth step ($k \geq 1$), the model $\mathcal{M}^{(k-1)}$ is given *a priori*. Then, for every $j \in \mathcal{M}_F \backslash \mathcal{M}^{(k-1)}$, a candidate model is constructed as $\mathcal{M}_j^{(k-1)} = \mathcal{M}^{(k-1)} \bigcup \{j\}$, whose lack of fit can be quantified as $\mathrm{RSS}_j^{(k-1)} = \widehat{\mathbb{Y}}^\top \{I_n - H_{\mathcal{M}_j^{(k-1)}}\}\widehat{\mathbb{Y}}$, where

$$H_{\mathcal{M}_j^{(k-1)}} = \widehat{\mathbb{X}}_{(\mathcal{M}_j^{(k-1)})}\Big\{\widehat{\mathbb{X}}^\top_{(\mathcal{M}_j^{(k-1)})}\widehat{\mathbb{X}}_{(\mathcal{M}_j^{(k-1)})}\Big\}^{-1}\widehat{\mathbb{X}}^\top_{(\mathcal{M}_j^{(k-1)})}.$$

(2.B) (*Screening*). Subsequently, the next most promising predictor is discovered as $a_k = \mathrm{argmin}_{j\in\mathcal{M}_F\backslash\mathcal{M}^{(k-1)}} \mathrm{RSS}_j^{(k-1)}$, and the candidate model is updated accordingly, that is, $\mathcal{M}^{(k)} = \mathcal{M}^{(k-1)} \bigcup \{a_k\}$.

Step (3) (*Solution Path*). By iterating Step (2) n times, a total of n nested candidate models is obtained with the solution path $\mathbb{S} = \{\mathcal{M}^{(k)} : 1 \leq k \leq n\}$, where $\mathcal{M}^{(k)} = \{a_1, \ldots, a_k\}$.

After extending FR to model (1.1), we study the theoretical properties of PFR to assure that it is applicable in practice.

3. Theoretical Properties

3.1. The technical conditions

To gain theoretical insights into PFR, we consider standard technical conditions. We relabel the profiled response variable Y^* and predictors X_j^* as V_0 and V_j, respectively, for $j = 1, \ldots, d$; this allows us to avoid imposing conditions on response variable and predictors separately. In addition, we write $G_0(t) = E(Y|t)$ and $G_j(t) = E(X_j|t)$. Let $\widehat{G}_j(t)$ be the estimator of $G_j(t)$ and $M_j(u)$ be the generating functions of V_j for $j = 0, \ldots, d$.

(C1) (*Normality*) The error, ε, is normal.

(C2) (*Covariance Matrix*) If $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$, respectively, are the smallest and largest eigenvalues of a positive definite matrix A, then, there exist two positive constants $0 < \tau_{\min} < \tau_{\max} < \infty$ such that $2\tau_{\min} < \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) < 2^{-1}\tau_{\max}$.

(C3) (*Regression Coefficients*) $\|\beta\| \leq C_\beta$ for some constant $C_\beta > 0$ and $\beta_{\min} \geq \nu_\beta n^{-\xi_{\min}}$ for some $\xi_{\min} > 0$ and $\nu_\beta > 0$, where $\| \cdot \|$ denotes the standard L_2 norm and $\beta_{\min} = \min_{j \in \mathcal{M}_T} |\beta_j|$.

(C4) (*Divergence Speed of d and d_0*) There exist positive constants ξ, ξ_0, and ν, such that $\log d \leq \min(\nu n^{\xi}, n^{3/10})$, $d_0 \leq \nu n^{\xi_0}$, and $\xi + 6\xi_0 + 12\xi_{\min} < 1$.

(C5) (*Smoothness Constraint*) $G_j(\cdot)$, $j = 0, 1, \ldots, d$, are uniformly Lipschitz continuous of order one.

(C6) (*Local Weights*) The weight functions $\omega_{nk}(\cdot)$ satisfy, with probability one,

(i) $\max_{1 \leq k \leq n} \sum_{i=1}^{n} \omega_{nk}(U_i) = O(1)$,

(ii) $\max_{1 \leq i,k \leq n} \omega_{nk}(U_i) = O(b_n)$,

(iii) $\max_{1 \leq i \leq n} \sum_{k=1}^{n} \omega_{nk}(U_i) I(|U_i - U_k| > c_n) = O(c_n)$,

where $b_n = n^{-4/5}$ and $c_n = n^{-2/5} \log n$.

(C7) (*Moment Constraint*) $\max_{0 \leq j \leq d} E\{\exp(u|V_j|)\} < \infty$ for all $0 \leq u \leq t_0/\sigma_v$, where $t_0 > 0$ and $\sigma_v^2 > 0$ are positive constants; the generating functions $M_j(u)$ $(j = 0, \ldots, d)$ satisfy

$$\max_{0 \leq j \leq d} \sup_{0 \leq u \leq t_0} \left| \frac{d^3 \log\{M_j(u)\}}{du^3} \right| < \infty;$$

$\max_{0 \leq j \leq d} E|V_j|^{2k} \leq \sigma_v^2$ for some $k > 2$.

Conditions (C1)–(C4) have customarily been assumed in the model selection literature, see Wang (2009). Conditions (C5) and (C6) are the standard assumptions for nonparametric regression, and can be easily ensured or verified; see Härdle, Liang, and Gao (2000). Condition (C7) is used for obtaining an exponential inequality of a sum of random variables (Chernoff (1952)). Finally, it is noteworthy that most of the literature has imposed the normality assumption (Zhang and Huang (2008)) or the spherically symmetric distribution (Fan and Lv (2008)) on X, while we do not need it.

3.2. The profiled consistency

In nonparametric regression analysis, one needs an accurate estimator for nonparametric functions. When $d = 1$, Mack and Silverman (1982) established a uniform convergence result. Following their approach, we are able to obtain a uniform convergence rate when d is finite. If $d \gg n$, this becomes a challenging task. This motivates us to develop a number of novel theoretical results in the Appendix. These findings provide theoretical justifications for PFR, but they also facilitate the development of ultrahigh dimensional methods in the nonparametric (or semiparametric) context.

Based on (C6), $G_j(t)$ can be expressed as $\sum_{k=1}^n \omega_{nk}(t)Y_k$ if $j = 0$, and $\sum_{k=1}^n \omega_{nk}(t)X_{jk}$ if $j = 1, \ldots, d$, for some weight function $\omega_{nk}(\cdot)$. To assure the good performance of PFR, we obtain the profiled consistency of the estimators $\widehat{G}_j(t)$ (or the profiled estimators, $\hat{Y}$ and $\hat{X}_j$) for $j = 0, \ldots, d$.

Theorem 1. *Under* (C4)$-$(C7), *we have*

$$\max_{0\le j\le d}\max_{1\le i\le n}\left|\widehat{G}_j(U_i) - \sum_{k=1}^n \omega_{nk}(U_i)G_j(U_k)\right| = o_p\Big(n^{-1/4}\log^{-1} n\Big). \tag{3.1}$$

This result indicates that the profiled estimators have the uniform convergence rate $n^{-1/4}\log^{-1} n$, independent of d as long as $\log d < n^{3/10}$. This finding suggests that, under a fairly mild restriction, the performance of PFR can be asymptotically as good as FR. Although the rate in Theorem 1 is slower than the optimal uniform convergence rate of $n^{-2/5}\log n$ established by Mack and Silverman (1982) when $d = 1$, $n^{-1/4}\log^{-1} n$ is sufficient for establishing theoretical analyses on the variable screening method.

3.3. Screening consistency

In the variable screening process, it is important to identify all relevant variables within a rather limited number of steps (Fan and Lv (2008)). To this end, we take the solution path $\mathbb{S}$ to be screening consistent if

$$P\Big(\mathcal{M}_T \subset \mathcal{M}^{(k)} \in \mathbb{S} \ \text{ for some } 1 \le k \le n\Big) \to 1. \tag{3.2}$$

Then, we establish PFR's screening consistency as follows.

Theorem 2. *Under* (C1)−(C7)*, we have that, as $n \to \infty$,*

$$P\Big(\mathcal{M}_T \subset \mathcal{M}^{([Kn^{\xi_0+4\xi_{\min}}])}\Big) \to 1,$$

where the constant $K = 6\tau_{\max}\tau_{\min}^{-2}C_\beta^2\nu_\beta^{-4}\nu$ is independent of n, with componenet constants from (C2), (C3)*, and* (C4)*, and where $[t]$ denotes the smallest integer not less than t.*

This theorem indicates that, with probability tending to one, the PFR algorithm is able to detect all relevant predictors within $O(n^{\xi_0+4\xi_{\min}})$ steps. This number is much smaller than the sample size n under Condition (C4). Furthermore, if the size of the true model is fixed, $\xi_0 = 0$, and the size of the smallest nonzero regression coefficient is bounded away from 0, $\xi_{\min} = 0$, then only a finite number of steps are needed to discover the entire relevant variable set. Moreover, the rate demonstrated in Theorem 2, $O(n^{\xi_0+4\xi_{\min}})$, is sharper than the rate of $O(n^{2\xi_0+4\xi_{\min}})$ given in Wang (2009), which is due to our approach; see Appendix B.

Theorem 2 provides a theoretical basis for PFR that enables us to empirically select the best model from $\mathbb{S}$. We therefore consider two BIC criteria

$$\text{BIC}_1(\mathcal{M}) = \log \hat{\sigma}^2_{(\mathcal{M})} + n^{-1}|\mathcal{M}|\Big(\log n + 2\log d\Big), \tag{3.3}$$

$$\text{and} \ \ \text{BIC}_2(\mathcal{M}) = \log \hat{\sigma}^2_{(\mathcal{M})} + n^{-1}|\mathcal{M}|\Big(2\log d\Big), \tag{3.4}$$

where $\mathcal{M}$ is an arbitrary candidate model with $|\mathcal{M}| \le n$, $\hat{\sigma}^2_{(\mathcal{M})} = n^{-1}\text{RSS}(\mathcal{M})$, $\text{RSS}(\mathcal{M}) = \widehat{\mathbb{Y}}^\top\{I_n - H_{\mathcal{M}}\}\widehat{\mathbb{Y}}$, and $H_{\mathcal{M}} = \widehat{\mathbb{X}}_{(\mathcal{M})}\{\widehat{\mathbb{X}}^\top_{(\mathcal{M})}\widehat{\mathbb{X}}_{(\mathcal{M})}\}^{-1}\widehat{\mathbb{X}}^\top_{(\mathcal{M})}$. The first of these has been considered by Chen and Chen (2008) and Wang (2009), while the second one has been investigated by An et al (2008). For either criterion, we select the best model $\widehat{\mathcal{M}} = \mathcal{M}^{(\hat{m})}$, where $\hat{m} = \text{argmin}_{1\le m\le n}\text{BIC}_k(\mathcal{M}^{(m)})$. Applying the same techniques as those in Wang (2009), we are able to show that $P(\mathcal{M}_T \subset \widehat{\mathcal{M}}) \to 1$. Consequently, the selected model $\widehat{\mathcal{M}}$ is screening consistent and its size is usually considerably smaller than n. Nevertheless, we typically do not expect that $P(\mathcal{M}_T = \widehat{\mathcal{M}}) \to 1$ since even in the classical large sample size and fixed dimension setup, forward regression is not consistent in selection (Wang (2009)). We present numerical evidence in the next section.

4. Numerical Studies

4.1. Simulation settings and performance measures

We give examples to demonstrate the finite sample performance of PFR. In each simulated model, we generated the index variable U from a uniform distribution on $[0,1]$; it is independent of the predictors X. The variance of ε was selected so that the resulting theoretical $R^2 = \text{var}\{X_i^\top\beta + g(U_i)\}/\text{var}(Y_i)$ was approximately 75%. This ensured that the signal-to-noise ratio was not weak. In addition, four sample sizes (n = 50, 100, 150, and 200) and three predictor dimensions (d = 500, 1,000, 2,000) were considered, and a total of $N = 200$ realizations were conducted. For the sake of comparison, the FR method proposed by Wang (2009) was also evaluated.

Example 1. (Independent Predictors) The linear component, $X^\top\beta$, of this model is given by Fan and Lv (2008). Accordingly, the linear predictors were independent and standard normal random variables. The size of the true model was $d_0 = 8$ with $\beta_j = (-1)^{I(\tilde{a}_j>0.6)}(4\log n/\sqrt{n} + |Z_j|)$ for every $1 \le j \le d_0$, with $\tilde{a}_j$ uniformly distributed on $[0,1]$ and Z_j standard normal. In addition to the linear component, the nonlinear element was $g(U) = 20UI(U < 0.5) + 10(U \ge 0.5)$, a piecewise linear function with structural change.

Example 2. (Autoregressive Correlation) The predictors associated with the linear component are correlated with each other in an autoregressive manner; see Tibshirani (1996). Specifically, the X_i $(i = 1,\ldots,n)$ were generated from a multivariate normal distribution with mean 0 and $\text{cov}(X_{ij_1}, X_{ij_2}) = 0.5^{|j_1-j_2|}$ for $1 \le j_1, j_2 \le d$. The 1st, 4th, and 7th elements of β were 3, 1.5, and 2, respectively. The remaining elements of β were 0. Consequently, we had $d_0 = 3$. Furthermore, the nonlinear component was $g(U) = \exp(3U)$, a nonlinear and increasing function.

Example 3. (Compound Symmetry) The covariance structure of predictors in the linear component is compound symmetry (Fan and Lv (2008)). Specifically, the X_i $(i = 1,\ldots,n)$ were simulated from a multivariate normal distribution with mean 0, $\text{var}(X_{ij}) = 1$, and $\text{cov}(X_{ij_1}, X_{ij_2}) = 0.5$ for $j_1 \ne j_2$, where $j = 1,\ldots,d$ and $1 \le j_1, j_2 \le d$. Furthermore, $\beta_j = 5$ for $j \le d_0 = 3$, while $\beta_j = 0$ for $j > d_0$. Moreover, the nonlinear component was $g(U) = 10\sin(2\pi U)$, which represents a nonlinear and non-monotonic function.

In the above examples, let $\{\boldsymbol{X}_{i(k)} : 1 \le i \le n\}$ be the simulated predictors from the kth realization, where $k = 1,\ldots,N$. Analogously, let $\hat{\beta}_{(k)} = (\hat{\beta}_{1(k)},\ldots,\hat{\beta}_{d(k)})^\top \in \mathbb{R}^d$ and $\hat{\Sigma}_{(k)} = n^{-1}\sum_{i=1}^n \boldsymbol{X}_{i(k)}\boldsymbol{X}_{i(k)}^\top$ be the estimators of β and Σ, respectively, and let the selected model be $\widehat{\mathcal{M}}_{(k)} = \{j : |\hat{\beta}_{j(k)}| > 0\}$.

To study the finite sample performance of FR and PFR, we computed six performance measures: (i.) the coverage probability ($100\% \times N^{-1}\sum_{k=1}^{N} I(\widehat{\mathcal{M}}_{(k)} \supset \mathcal{M}_T)$); (ii.) the percentage of correct fit ($100\% \times N^{-1}\sum_{k=1}^{N} I(\widehat{\mathcal{M}}_{(k)} = \mathcal{M}_T)$); (iii.) the percentage of correct zeros ($100\% \times \{(d-d_0)N\}^{-1}\{\sum_{k=1}^{N}\sum_{j=1}^{d} I(\hat{\beta}_{j(k)} = 0) \times I(\beta_j = 0)\}$); (iv.) the percentage of incorrect zeros ($100\% \times (d_0 N)^{-1}\{\sum_{k=1}^{N}\sum_{j=1}^{d} I(\hat{\beta}_{j(k)} = 0) \times I(\beta_j \neq 0)\}$); (v.) the average size of the selected model ($N^{-1}\sum_{k=1}^{N} |\widehat{\mathcal{M}}_{(k)}|$); and (vi.) the relative estimation error ($100\% \times N^{-1}\sum_{k=1}^{N}\{(\hat{\beta}_{(k)} - \beta)^\top \hat{\Sigma}_{(k)}(\hat{\beta}_{(k)} - \beta)\}/\{\beta^\top \hat{\Sigma}_{(k)}\beta\}$)).

4.2. Simulation results

The detailed simulation results of Examples 1 to 3 are presented in Tables 1−3, respectively. Because U and X are independent, we can show that FR is screening consistent under the appropriate conditions from Wang (2009). As a result, the finite sample performance of the FR method is qualitatively similar to that of the PFR method. However, Tables 1−3 show that PFR outperforms FR. This finding is not surprising since FR does not control the nonlinear component $g(U)$. To save space, we focus only on PFR in the following discussions.

For the fixed dimension d, the PFR's finite sample performance improves substantially as the sample size increases. For example, with a reasonably large sample size (e.g., $n = 200$), the resulting coverage probability approaches 100%. These results numerically confirm that PFR is screening consistent; see Theorem 2. As a byproduct of the screening consistency property, the percentage of incorrect zeros moves quickly toward 0 as the sample size increases.

It is noteworthy that focusing solely on the screening consistency is insufficient for assessing performance. This is because an excellent coverage probability can be attained by some naïve or useless method. For example, if one always selects the full model $\mathcal{M}_F$ as the "best" model, the resulting coverage probability is 100%. Although this is an excellent coverage probability, its cost is a huge model size. This motivates us to evaluate the PFR's capability in identifying the correct sparse solutions. Tables 1−3 indicate that PFR together with the BIC criteria (3.3) and (3.4) identifies correct zeros almost 100% of cases. As a result, the average model size is not only small, but also closes to the true model size d_0 when the sample size is large. Consequently, the relative estimation error steadily decreases as the sample size increases.

We also note that the finite sample performances of the two BIC criteria are qualitatively similar. Because BIC_1 uses a slightly larger penalty than does BIC_2, its capability in identifying sparse solutions is stronger while its underfitting effect is more serious. As a result, BIC_1 typically yields the larger percentage of

Table 1. Simulation Results for Example 1.

		Model	Coverage	Correct	% of	% of	Average	Relative
		Selection	Probability	Selection	Incorrect	Correct	Model	Estimation
d	n	Criterion	(%)	(%)	Zeros	Zeros	Size	Error (%)
				PFR				
2,000	100	BIC_1	20.2	0.0	79.8	100.0	1.7	66.3
		BIC_2	48.9	2.0	51.1	99.9	5.2	39.6
	150	BIC_1	74.8	24.5	25.2	100.0	6.0	16.7
		BIC_2	90.2	33.5	9.8	100.0	7.8	8.9
	200	BIC_1	90.1	54.0	9.9	100.0	7.2	6.3
		BIC_2	97.1	57.5	2.9	100.0	8.2	4.1
1,000	100	BIC_1	30.3	1.0	69.7	100.0	2.5	55.1
		BIC_2	67.3	6.0	32.7	99.9	6.6	26.7
	150	BIC_1	79.5	29.0	20.5	100.0	6.4	13.0
		BIC_2	92.2	38.5	7.8	99.9	7.9	7.2
	200	BIC_1	92.7	55.5	7.3	100.0	7.5	5.0
		BIC_2	97.2	59.5	2.8	100.0	8.1	3.8
500	100	BIC_1	44.1	6.5	55.9	100.0	3.6	42.0
		BIC_2	76.2	15.5	23.8	99.8	7.1	19.5
	150	BIC_1	83.9	37.5	16.1	100.0	6.7	10.7
		BIC_2	95.2	47.5	4.8	99.9	8.1	6.0
	200	BIC_1	95.2	69.5	4.8	100.0	7.6	3.7
		BIC_2	98.6	62.5	1.4	99.9	8.4	3.4
				FR				
2,000	100	BIC_1	12.6	0.0	87.4	100.0	1.1	78.0
		BIC_2	31.3	0.0	68.7	100.0	3.5	59.9
	150	BIC_1	39.2	0.5	60.8	100.0	3.2	48.9
		BIC_2	69.4	8.0	30.6	100.0	6.2	29.1
	200	BIC_1	63.0	8.5	37.0	100.0	5.1	27.7
		BIC_2	84.1	21.0	15.9	100.0	7.1	16.5
1,000	100	BIC_1	16.2	0.0	83.8	100.0	1.4	72.3
		BIC_2	41.6	1.5	58.4	99.9	3.9	49.8
	150	BIC_1	48.4	3.0	51.6	100.0	3.9	40.9
		BIC_2	74.6	7.5	25.4	99.9	6.5	24.0
	200	BIC_1	72.6	15.0	27.4	100.0	5.8	22.0
		BIC_2	89.1	32.0	10.9	100.0	7.5	14.2
500	100	BIC_1	20.6	0.0	79.4	100.0	1.7	68.1
		BIC_2	51.2	2.0	48.8	99.9	4.8	43.2
	150	BIC_1	56.4	7.0	43.6	100.0	4.5	34.9
		BIC_2	81.3	14.5	18.7	99.9	6.9	20.4
	200	BIC_1	77.1	19.5	22.9	100.0	6.2	18.9
		BIC_2	91.4	40.5	8.6	99.9	7.7	12.8

Table 2. Simulation Results for Example 2.

d	n	Model Selection Criterion	Coverage Probability (%)	Correct Selection (%)	% of Incorrect Zeros	% of Correct Zeros	Average Model Size	Relative Estimation Error (%)
				PFR				
2,000	100	BIC_1	64.5	15.5	35.5	100.0	2.0	19.2
		BIC_2	74.5	21.5	25.5	100.0	2.8	18.3
	150	BIC_1	81.8	48.0	18.2	100.0	2.5	9.2
		BIC_2	90.3	52.5	9.7	100.0	3.2	8.8
	200	BIC_1	92.3	75.5	7.7	100.0	2.8	4.0
		BIC_2	96.7	64.5	3.3	100.0	3.3	4.8
1,000	100	BIC_1	66.2	15.5	33.8	100.0	2.0	17.9
		BIC_2	78.5	31.5	21.5	99.9	2.9	15.7
	150	BIC_1	83.8	52.5	16.2	100.0	2.6	8.0
		BIC_2	90.7	54.0	9.3	100.0	3.1	7.6
	200	BIC_1	93.8	79.5	6.2	100.0	2.8	3.7
		BIC_2	98.7	71.0	1.3	100.0	3.3	3.6
500	100	BIC_1	67.7	18.0	32.3	100.0	2.1	17.4
		BIC_2	79.7	27.5	20.3	99.9	2.9	15.5
	150	BIC_1	87.2	61.0	12.8	100.0	2.6	6.4
		BIC_2	93.8	59.0	6.2	99.9	3.1	6.0
	200	BIC_1	94.8	83.0	5.2	100.0	2.9	3.3
		BIC_2	98.3	71.0	1.7	99.9	3.3	3.8
				FR				
2,000	100	BIC_1	18.7	0.0	81.3	100.0	0.6	72.9
		BIC_2	30.8	0.0	69.2	100.0	1.4	70.4
	150	BIC_1	39.3	0.5	60.7	100.0	1.2	44.8
		BIC_2	52.3	3.0	47.7	100.0	1.9	38.3
	200	BIC_1	51.3	2.0	48.7	100.0	1.6	30.9
		BIC_2	63.7	10.5	36.3	100.0	2.2	26.0
1,000	100	BIC_1	24.8	0.0	75.2	100.0	0.8	63.8
		BIC_2	36.0	0.5	64.0	100.0	1.5	59.5
	150	BIC_1	39.5	0.5	60.5	100.0	1.2	44.9
		BIC_2	55.3	3.5	44.7	100.0	2.1	38.7
	200	BIC_1	54.5	3.5	45.5	100.0	1.7	29.2
		BIC_2	68.5	14.5	31.5	100.0	2.4	25.3
500	100	BIC_1	27.0	0.0	73.0	100.0	0.9	62.8
		BIC_2	39.8	2.5	60.2	99.9	1.7	58.8
	150	BIC_1	46.3	1.5	53.7	100.0	1.4	37.7
		BIC_2	59.5	6.5	40.5	99.9	2.1	30.4
	200	BIC_1	58.0	5.5	42.0	100.0	1.8	26.2
		BIC_2	73.0	22.5	27.0	99.9	2.6	23.4

Table 3. Simulation Results for Example 3.

		Model	Coverage	Correct	% of	% of	Average	Relative
		Selection	Probability	Selection	Incorrect	Correct	Model	Estimation
d	n	Criterion	(%)	(%)	Zeros	Zeros	Size	Error (%)
				PFR				
2,000	100	BIC$_1$	48.0	5.5	52.0	100.0	2.1	19.7
		BIC$_2$	55.0	12.0	45.0	100.0	2.6	17.1
	150	BIC$_1$	84.7	59.5	15.3	100.0	2.7	5.6
		BIC$_2$	91.0	66.0	9.0	100.0	3.1	4.3
	200	BIC$_1$	97.0	90.5	3.0	100.0	3.0	1.7
		BIC$_2$	98.3	81.5	1.7	100.0	3.2	1.9
1,000	100	BIC$_1$	55.0	11.0	45.0	99.9	2.2	16.4
		BIC$_2$	63.5	20.5	36.5	99.9	2.7	13.8
	150	BIC$_1$	87.0	63.5	13.0	100.0	2.7	4.9
		BIC$_2$	94.2	71.0	5.8	100.0	3.2	3.6
	200	BIC$_1$	97.0	90.5	3.0	100.0	3.0	1.6
		BIC$_2$	98.7	81.5	1.3	100.0	3.2	1.7
500	100	BIC$_1$	63.0	20.5	37.0	99.9	2.2	13.9
		BIC$_2$	73.7	31.0	26.3	99.9	2.9	10.7
	150	BIC$_1$	90.7	73.0	9.3	100.0	2.8	3.8
		BIC$_2$	95.5	69.5	4.5	99.9	3.2	3.3
	200	BIC$_1$	98.0	92.0	2.0	100.0	3.0	1.4
		BIC$_2$	99.3	81.5	0.7	100.0	3.2	1.5
				FR				
2,000	100	BIC$_1$	27.5	0.0	72.5	100.0	1.7	32.3
		BIC$_2$	31.8	0.5	68.2	99.9	2.2	29.8
	150	BIC$_1$	53.3	8.5	46.7	100.0	2.1	18.3
		BIC$_2$	62.3	18.0	37.7	100.0	2.7	15.6
	200	BIC$_1$	72.5	29.5	27.5	100.0	2.4	10.4
		BIC$_2$	82.2	48.5	17.8	100.0	3.0	8.2
1,000	100	BIC$_1$	34.2	0.0	65.8	99.9	1.8	28.3
		BIC$_2$	38.8	1.0	61.2	99.9	2.3	25.9
	150	BIC$_1$	58.0	7.5	42.0	100.0	2.1	16.2
		BIC$_2$	68.3	24.0	31.7	99.9	2.8	13.5
	200	BIC$_1$	76.0	34.0	24.0	100.0	2.5	9.1
		BIC$_2$	87.0	51.5	13.0	100.0	3.1	6.5
500	100	BIC$_1$	39.5	1.0	60.5	99.9	1.8	26.0
		BIC$_2$	46.7	5.0	53.3	99.8	2.5	22.9
	150	BIC$_1$	64.2	15.5	35.8	99.9	2.2	13.6
		BIC$_2$	75.8	30.5	24.2	99.9	2.9	10.9
	200	BIC$_1$	83.7	52.5	16.3	100.0	2.6	6.4
		BIC$_2$	91.0	59.0	9.0	99.9	3.1	5.1

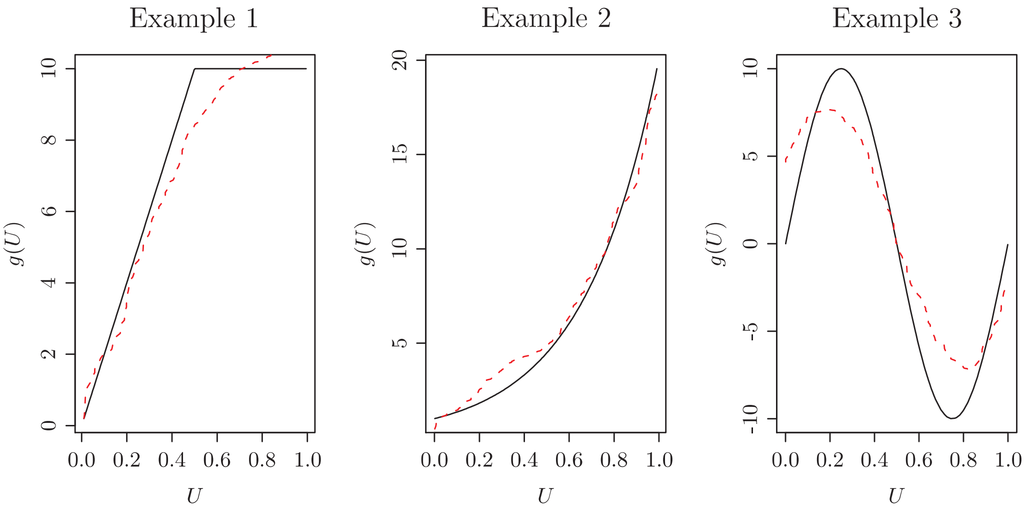

Figure 1. The true nonparametric component (the solid line) and its estimate (the dashed line).

incorrect zeros, while BIC_2 leads to the bigger average model size. In addition, Examples 2 and 3 show that BIC_1 slightly outperforms BIC_2 for identifying the correct fit. However, since PFR is not a consistent selection method, we do not expect that the percentage of correct fit will converge to 100% as $n \to \infty$.

For the fixed sample size, we compare the PFR's performance across various dimensions of predictors. As expected, Tables 1−3 demonstrate that the larger dimension of predictors leads to worse finite sample performance. It is of interest to note that the performance of PFR does not deteriorate rapidly when d increases. For example, Table 3 shows that if the dimension of predictors increases from 500 to $4 \times 500 = 2,000$ with $n = 100$, the coverage probability drops from 63.0% to 48.0%. In contrast, if we fix the predictor dimension to be $2,000$, but increase the sample size from $n = 50$ to $n = 4 \times 50 = 200$, the coverage probability increases from 8.3% to 97.0%. These findings suggest that the sample size plays a more important role than the dimension of predictors in ultrahigh dimensional variable screening. In sum, the numerical results corroborate our theoretical findings, and PFR performs well.

Finally, to gain some intuitive understanding of the estimated nonlinear component, we fixed the sample size to be $n = 200$ and the predictor dimension to be $p = 500$. Then, used BIC_1 as the stopping rule to select linear component. Figure 1 shows that the resulting estimates approximate the true nonlinear component fairly well.

4.3. Empirical example

To illustrate the usefulness of the PFR method, we consider the data set provided by a major domestic supermarket chain located in northern China. It contains a total of 159 daily observations collected between 07/26/2008 and 12/31/2008. The response is the supermarket's daily profit in log-scale, while the predictor vector X_i is of 1,795 dimensions. The first 6 component of X_i (i.e., $X_{i1}, \ldots, X_{i6}$) are dummy variables associated with Tuesday, Wednesday, Thursday, Friday, Saturday, and Sunday, respectively; Monday is used as a benchmark. Subsequently, X_{ij} ($j = 7$ to $j = 1,795$) correspond to the supermarket's products, in log-scale, which have been advertised for sales promotion during the above time period. For the sake of convenience, both response and predictors are standardized to have zero mean and unit variance. Because the time U of the sales promotion commonly affects the profit in a nonlinear form, we also include it as the nonparametric component $g(U)$ in our study.

We first conducted an out-of-sample performance to illustrate the usefulness of PFR in forecasting. To this end, we considered the training sample size $n_0 = 120$, and created a total of $n - n_0 = 39$ moving windows $\mathcal{W}_l = \{(X_i, U_i) : l \leq i \leq l + 119\}$, $l = 1, \ldots, 39$. Based on $\mathcal{W}_l$, we fit the data with the partially linear model via PFR. For the sake of comparison, we also fit the corresponding data (ignoring U_i) with the linear model via FR. The resulting models were then used to predict the value of Y_{l+120}, denoted as $\hat{Y}_{l+120}$. To compare the accuracy of predictions between FR and PFR, we computed the absolute prediction error (APE), $|\hat{Y}_{l+120} - Y_{l+120}|$. The results suggest that PFR is considerably better than FR in terms of both the mean (0.5646 vs. 0.6981) and the median (0.4587 vs. 0.6813) calculated from 39 APEs. This indicates that the nonlinear component $g(U)$ plays a useful role in PFR for forecasting.

We next applied the PFR method to the whole dataset, and identified four relevant variables. Among them, two were the dummies associated with Saturday and Sunday. Their corresponding regression coefficients of 0.30 (Saturday) and 0.38 (Sunday), respectively, perfectly match with the common conception that weekend sales and profits are substantially higher than those of Monday. In addition, PFR identified two food products for which sales promotion were very effective. Moreover, Figure 2 presents the estimated nonparametric function $g(U)$ together with 90% pointwise confidence bands. It depicts a nonlinear curve that rapidly increases as time approaches the end of the promotion period. Note that the end of sales promotion was 12/31/2008; one day before the New Year holidays. In sum, the partially linear model in conjunction with the PFR method provides insightful findings on supermarket promotions.

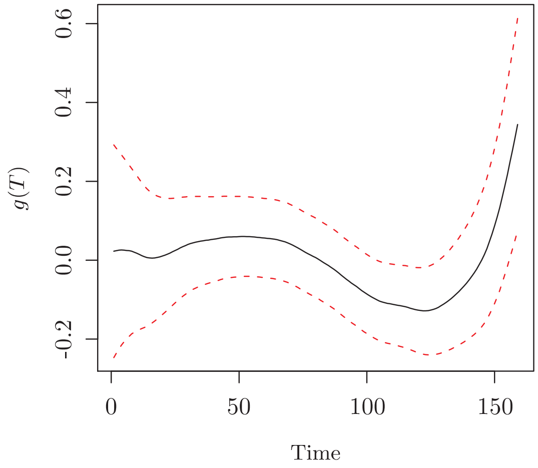

Figure 2. The estimated nonparametric component (the solid line) and its 90% pointwise confidence bands (the dashed lines).

5. Conclusion

We have proposed a PFR algorithm for ultrahigh dimensional variable screening in partially linear models. It combines the classical forward regression with the nonparametric profiling estimator (Fan and Huang (2005)). We have shown that PFR is screening consistent, and numerical results suggest that PFR performs well.

In the recent literature on variable screening or selection, there are two closely related forward regression methods, LARS by Efron et al. (2004), and stagewise orthogonal matching pursuit (StoMP), by Donoho et al. (2006). LARS search for the next variable in an equiangular direction with all current variables, while StoMP selects predictors having significant correlation with the current residual at each stage, recursively, and can deal with the case where the data may be noiseless. We cannot rule out the possibility that LARS and StoMP might also have the screening consistency, as enjoyed by SIS, FR, and PFR. However, various simulation studies conducted by Wang (2009) suggest that the LARS finite sample performance in variable screening is much worse than that of FR. Such a result is not surprising because the LARS estimate is closely related to the LASSO estimate, which does not have the oracle property (Zou (2006); Fan and Li (2001)). Our unreported numerical results suggest that, for partial linear regression models, PFR outperforms the profiled LARS algorithm. Furthermore, Stodden (2006) found that StoMP is inferior to LARS/LASSO in recovering the underlying model when sparsity levels are high. We therefore expect PFR to be superior to profiled StoMP.

We identify three research areas that could broaden the usefulness of PFR. The first is to generalize PFR to the varying coefficient model (Cai, Fan, and Li (2000); Hastie and Tibshirani (1993)) and the single-index model (Xia (2006); Hristache, Juditsky and Spokoiny (2001)) with their various extensions (Carroll et al. (1997)). The second is to develope a data driven bandwidth selection method for PFR in an ultrahigh dimensional setting. The third is to employ the property of profiled consistency in Theorem 1 to extend the existing screening methods (e.g., LARS, Efron et al. (2004) and SIS, Fan and Lv (2008)) to semiparametric models. We believe that these efforts would enhance the usefulness of ultrahigh dimensional variable screening in nonparametric data analysis.

Acknowledgements

Liang's research was supported by NSF grants DMS-0806097 and DMS-1007167. Wang's research was supported in part by a grant from NSFC (10771006), a grant from the Fundamental Research Funds for the Central Universities, and the Research Funds of Renmin University of China (Grant No. 10XNL007). The authors are grateful to an associate editor and a referee for valuable comments and suggestions that lead to a much improved paper.

Appendix

Appendix A. Technical Lemmas

Before proving the two theorems, we present four lemmas. Lemmas A.1 and A.2 are used for the proof of Theorem 1, while Lemma A.3 is used for Lemma A.4, which is needed in the proof of Theorem 2.

Lemma A.1. *Suppose* (C5), (C6.i), *and* (C6.iii) *hold. Then*

$$\max_{0\le j\le d}\max_{1\le i\le n}\Big|G_j(U_i)-\sum_{k=1}^n \omega_{nk}(U_i)G_j(U_k)\Big| = O(c_n).$$

Proof. For a given $G_j(U_i)$ ($j=1,\ldots,d$ and $i=1,\ldots,n$), we have

$$\begin{aligned}
&G_j(U_i)-\sum_{k=1}^n \omega_{nk}(U_i)G_j(U_k)\\
&=\sum_{k=1}^n \omega_{nk}(U_i)\{G_j(U_i)-G_j(U_k)\}\\
&=\sum_{k=1}^n \omega_{nk}(U_i)\{G_j(U_i)-G_j(U_k)\}I(|U_i-U_k|>c_n)
\end{aligned}$$

$$+\sum_{k=1}^{n}\omega_{nk}(U_i)\{G_j(U_i)-G_j(U_k)\}I(|U_i-U_k|\le c_n).$$

This, together with Condition (iii) of (C6), implies that

$$\max_{0\le j\le d}\max_{1\le i\le n}\left|\sum_{k=1}^{n}\omega_{nk}(U_i)\{G_j(U_k)-G_j(U_i)\}I(|U_i-U_k|>c_n)\right|$$
$$\le C\max_{0\le j\le d}\max_{1\le i\le n}\left|\sum_{k=1}^{n}\omega_{nk}(U_i)I(|U_i-U_k|>c_n)\right|=O(c_n).$$

Applying the Lipschitz continuity assumption (C5) and Condition (i) of (C6), we obtain

$$\max_{0\le j\le d}\max_{1\le i\le n}\left|\sum_{k=1}^{n}\omega_{nk}(U_i)\{G_j(U_k)-G_j(U_i)\}I(|U_i-U_k|\le c_n)\right|$$
$$\le\max_{0\le j\le d}\max_{1\le i\le n}\left|\sum_{k=1}^{n}\omega_{nk}(U_i)c_n\right|=O(c_n).$$

This completes the proof.

Lemma A.2. *Let $W_1,\ldots,W_n$ be i.i.d. with constant variance σ^2. Take $Z_k=(W_k-EW_k)/\sigma$, let $M(u)=E\{\exp(uZ_k)\}$ be the generating function of Z_k for $k=1,\ldots,n$, and assume that there is a positive constant t_0 such that $E\{\exp(t|W_k|)\}<\infty$ for $0\le t\le t_0/\sigma$. Let $a_{nk}, 1\le k\le n$, be a sequence of constants and $A, A_1, A_2,\ldots,$ be a sequence of constants satisfying $A_n\ge\sum_{k=1}^{n}a_{nk}^2\sigma^2$ and $A\ge\max_k|a_{nk}\sigma|/A_n$. If*

$$M^*\doteq\sup_{0\le u\le t_0}\left|\frac{d^3\log M(u)}{dt^3}\right|<\infty, \tag{A.1}$$

then, for $0<\zeta<t_0/A$, we have

$$P\left\{\left|\sum_{k=1}^{n}a_{nk}(W_k-EW_k)\right|>\zeta\right\}\le\exp\left\{-\frac{\zeta^2}{2A_n}\left(1-\frac{1}{3}AM^*\zeta\right)\right\}. \tag{A.2}$$

Proof. Take $t_\zeta=\zeta/A_n$ for $\zeta\le t_0/A$. Then $|a_{nk}\sigma t_\zeta|=|a_{nk}\sigma\zeta/A_n|\le A\zeta\le t_0$. Applying a Taylor expansion to $\log\{M(u)\}$ at $u=0$, we have, for $0\le u\le t_0$,

$$\log M(u)=\log M(0)$$
$$+t\frac{d\log M(u)}{dt}\Big|_{u=0}+\frac{u^2}{2}\frac{d^2\log M(u)}{du^2}\Big|_{u=0}+\frac{u^3}{6}\frac{d^3\log M(u)}{du^3}\Big|_{u=u_*},$$

where u_* lies between u and 0. It is noteworthy that $\log M(0) = 0$,

$$\frac{d\log M(u)}{du}\Big|_{u=0} = E(Z_1) = 0, \ \ \frac{d^2\log M(u)}{du^2}\Big|_{u=0} = 1, \text{ and } \left|\frac{d^3\log M(u)}{du^3}\Big|_{u=u_*}\right| \leq M^*.$$

It follows that

$$\begin{aligned}\log\Big\{M\Big(a_{nk}\sigma t_\zeta\Big)\Big\} &\leq 2^{-1}\left(\frac{a_{nk}\sigma\zeta}{A_n}\right)^2 + \frac{1}{6}\left|\frac{a_{nk}\sigma\zeta}{A_n}\right|^3 \cdot M^* \\ &\leq \frac{a_{nk}^2\sigma^2\zeta^2}{2A_n^2}\left(1 + \frac{1}{3}AM^*\zeta\right).\end{aligned}$$

After algebraic simplification, we have

$$\begin{aligned}&\log P\left\{\sum_{k=1}^n a_{nk}(W_k - EW_k) > \zeta\right\} = \log P\left\{\sum_{k=1}^n a_{nk}\sigma Z_k > \zeta\right\} \\ &\leq \log E\left[\exp\left\{t_\zeta\left(\sum_{k=1}^n a_{nk}\sigma Z_k - \zeta\right)\right\}\right] = -\zeta t_\zeta + \sum_{k=1}^n \log M(a_{nk}\sigma t_\zeta) \\ &\leq -\frac{\zeta^2}{A_n} + \sum_{k=1}^n \frac{a_{nk}^2\sigma^2\zeta^2}{2A_n}\left(1 + \frac{1}{3}AM^*\zeta\right) \\ &\leq -\frac{\zeta^2}{A_n} + \frac{\zeta^2}{2A_n^2}\left(1 + \frac{1}{3}AM^*\zeta\right) = -\frac{\zeta^2}{2A_n}\left(1 - \frac{1}{3}AM^*\zeta\right).\end{aligned}$$

As a result,

$$P\left\{\sum_{k=1}^n a_{nk}(W_k - EW_k) > \zeta\right\} \leq \exp\left\{-\frac{\zeta^2}{A_n}\left(1 - \frac{1}{3}AM^*\zeta\right)\right\}. \tag{A.3}$$

Analogously, we can show that

$$P\left\{\sum_{k=1}^n a_{nk}(W_k - EW_k) < -\zeta\right\} \leq \exp\left\{-\frac{\zeta^2}{A_n}\left(1 - \frac{1}{3}AM^*\zeta\right)\right\}. \tag{A.4}$$

The results (A.3) and (A.4) complete the proof.

Lemma A.3 (Bernstein's Inequality). *Let $\{R_k, 1 \leq k \leq n\}$ be independent random variables with $E(R_k) = 0$ and $var(R_k) = \sigma_k^2$. If $E|R_k|^l \leq (l!/2)\sigma_k^2 c^{l-2}$, for $1 \leq k \leq n$, $0 < c < \infty$, and some $l > 2$, then*

$$P\left\{\sum_{k=1}^n R_k > \delta\right\} \leq \exp\left\{-\frac{\delta^2}{2(\sum_{k=1}^n \sigma_k^2 + c\delta)}\right\} \quad \textit{for } \delta > 0.$$

It can be easily shown that the sufficient condition in Lemma A.3 holds when R_k are independent and identically distributed normal random variable or $P\{|R_k| \le c\} = 1$ for $1 \le k \le n$. A detailed proof of this lemma can be found in Pollard (1984).

Lemma A.4. *Take* $\hat{\Sigma} = n^{-1}\widehat{\mathbb{X}}^{\top}\widehat{\mathbb{X}}$ *and* $\Sigma^{*} = n^{-1}\mathbb{X}^{*\top}\mathbb{X}^{*}$ *and, for a subset model* $\mathcal{M}$*, let* $\hat{\Sigma}_{(\mathcal{M})}$ *and* $\Sigma^{*}_{(\mathcal{M})}$ *be the submatrices of* $\hat{\Sigma}$ *and* Σ^{*}*, respectively. If* (C2) *and* (C4)$-$(C7) *hold and* $\tilde{m} = O(n^{2\xi_0 + 4\xi_{\min}})$*, then, with probability tending to one, we have*

$$\tau_{\min} \le \min_{|\mathcal{M}| \le \tilde{m}} \lambda_{\min}\Big\{\hat{\Sigma}_{(\mathcal{M})}\Big\} \le \max_{|\mathcal{M}| \le \tilde{m}} \lambda_{\max}\Big\{\hat{\Sigma}_{(\mathcal{M})}\Big\} \le \tau_{\max}. \tag{A.5}$$

Proof. Let $r = (r_1, \ldots, r_d)^{\top} \in \mathbb{R}^d$ be an arbitrary d-dimensional vector and $r_{(\mathcal{M})}$ be the subvector corresponding to $\mathcal{M}$. By (C2), we immediately have

$$\begin{aligned} 2\tau_{\min} &\le \min_{\mathcal{M} \subset \mathcal{M}_F} \inf_{\|r_{(\mathcal{M})}\|=1} r_{(\mathcal{M})}^{\top} \Sigma_{(\mathcal{M})} r_{(\mathcal{M})} \\ &\le \max_{\mathcal{M} \subset \mathcal{M}_F} \sup_{\|r_{(\mathcal{M})}\|=1} r_{(\mathcal{M})}^{\top} \Sigma_{(\mathcal{M})} r_{(\mathcal{M})} \le 2^{-1}\tau_{\max}. \end{aligned}$$

Therefore, (A.5) follows if we are able to show that

$$P\Bigg(\max_{|\mathcal{M}| \le \tilde{m}} \sup_{\|r_{(\mathcal{M})}\|=1} \Big| r_{(\mathcal{M})}^{\top} \Big\{ \hat{\Sigma}_{(\mathcal{M})} - \Sigma_{(\mathcal{M})} \Big\} r_{(\mathcal{M})} \Big| > \tilde{\epsilon} \Bigg) \to 0, \tag{A.6}$$

where $\tilde{\epsilon} > 0$ is an arbitrary positive number. Note that

$$\begin{aligned} \widehat{\Sigma}_{(\mathcal{M})} - \Sigma^{*}{}_{(\mathcal{M})} &= \frac{1}{n}\Big\{ \widehat{\mathbb{X}}_{(\mathcal{M})}^{\top} \widehat{\mathbb{X}}_{(\mathcal{M})} - \mathbb{X}_{(\mathcal{M})}^{*\top} \mathbb{X}_{(\mathcal{M})}^{*} \Big\} \\ &= \frac{1}{n}\Big\{ \widehat{\mathbb{X}}_{(\mathcal{M})} - \mathbb{X}_{(\mathcal{M})}^{*} \Big\}^{\top} \Big\{ \widehat{\mathbb{X}}_{(\mathcal{M})} - \mathbb{X}_{(\mathcal{M})}^{*} \Big\} \\ &\quad + \frac{1}{n}\Big\{ \widehat{\mathbb{X}}_{(\mathcal{M})} - \mathbb{X}_{(\mathcal{M})}^{*} \Big\}^{\top} \mathbb{X}_{(\mathcal{M})}^{*} + \frac{1}{n}\mathbb{X}_{(\mathcal{M})}^{*\top} \Big\{ \widehat{\mathbb{X}}_{(\mathcal{M})} - \mathbb{X}_{(\mathcal{M})}^{*} \Big\}. \end{aligned}$$

Under (C4)$-$(C7), it follows from Theorem 1 and the Cauchy inequality that

$$\frac{1}{n} \max_{|\mathcal{M}| \le \tilde{m}} \sup_{\|r_{(\mathcal{M})}\|=1} \Big| r_{(\mathcal{M})}^{\top} \Big[\mathbb{X}_{(\mathcal{M})}^{*\top} \{ \widehat{\mathbb{X}}_{(\mathcal{M})} - \mathbb{X}_{(\mathcal{M})}^{*} \} \Big] r_{(\mathcal{M})} \Big| = o_p\Big(n^{-1/4} \log^{-1} n \Big).$$

In the same way, we can prove that

$$\frac{1}{n} \max_{|\mathcal{M}| \le \tilde{m}} \sup_{\|r_{(\mathcal{M})}\|=1} \Big| r_{(\mathcal{M})}^{\top} \Big[\{ \widehat{\mathbb{X}}_{(\mathcal{M})} - \mathbb{X}_{(\mathcal{M})}^{*} \}^{\top} \mathbb{X}_{(\mathcal{M})}^{*} \Big] r_{(\mathcal{M})} \Big| = o_p\Big(n^{-1/4} \log^{-1} n \Big),$$

$$\frac{1}{n} \max_{|\mathcal{M}| \le \tilde{m}} \sup_{\|r_{(\mathcal{M})}\|=1} \Big| r_{(\mathcal{M})}^{\top} \Big[\{ \widehat{\mathbb{X}}_{(\mathcal{M})} - \mathbb{X}_{(\mathcal{M})}^{*} \}^{\top} \{ \widehat{\mathbb{X}}_{(\mathcal{M})} - \mathbb{X}_{(\mathcal{M})}^{*} \} \Big] r_{(\mathcal{M})} \Big| = o_p\Big(n^{-1/4} \log^{-1} n \Big).$$

The above results indicate that

$$P\left(\max_{|\mathcal{M}|\leq\tilde{m}}\sup_{\|r_{(\mathcal{M})}\|=1}\left|r_{(\mathcal{M})}^{\top}\left\{\widehat{\Sigma}_{(\mathcal{M})}-\Sigma^{*}{}_{(\mathcal{M})}\right\}r_{(\mathcal{M})}\right|>\tilde{\epsilon}\right)\to 0.$$

To show (A.6), then, it suffices to prove that

$$P\left(\max_{|\mathcal{M}|\leq\tilde{m}}\sup_{\|r_{(\mathcal{M})}\|=1}\left|r_{(\mathcal{M})}^{\top}\left\{\Sigma^{*}{}_{(\mathcal{M})}-\Sigma_{(\mathcal{M})}\right\}r_{(\mathcal{M})}\right|>\tilde{\epsilon}\right)\to 0. \qquad (A.7)$$

Under (C2), (C4), (C7), and calling on Lemma A.3, (A.7) can be proved in a similar manner as Lemma 1 in Wang (2009). This completes the proof.

Appendix B. Proofs of the Theorems

Proof of Theorem 1. We focus on the case $j\geq 1$ since the proof for $j=0$ is simpler. It is noteworthy that $\widehat{G}_j(U_i)=\sum_{k=1}^{n}\omega_{nk}(U_i)G_j(U_k)+\sum_{k=1}^{n}\omega_{nk}(U_i)V_{jk}$, where V_{jk} is the kth sample of V_j (i.e., $V_{jk}=X_{jk}^{*}$). Then

$$\begin{aligned}&\max_{1\leq j\leq d}\max_{1\leq i\leq n}|\widehat{G}_j(U_i)-G_j(U_i)|\\&\leq\max_{1\leq j\leq d}\max_{1\leq i\leq n}\left|\sum_{k=1}^{n}\omega_{nk}(U_i)G_j(U_k)-G_j(U_i)\right|+\max_{1\leq j\leq d}\max_{1\leq i\leq n}\left|\sum_{k=1}^{n}\omega_{nk}(U_i)V_{jk}\right|. \quad (B.1)\end{aligned}$$

By (C5) and (C6) along with Lemma A.1, the first term on the right side of the above equation is bounded by $O(c_n)$, where c_n is defined in (C6).

We next show the asymptotic convergence of the second term on the right side of (B.1). To this end, we introduce a positive constant independent of n, called $\mathbb{C}$, and take $a_{nk}=\omega_{nk}(U_i)$ and $\zeta=n^{-1/4}\log^{-1}n$. Let A be a constant such that $A_n=\mathbb{C}\sigma^2 b_n$, where b_n is defined in (C6) and $A\geq\max_{1\leq k\leq n}\omega_{nk}(U_i)/\mathbb{C}b_n$. By (C6), it is easily verified that $A_n\geq\sum_{k=1}^{n}a_{nk}^2\sigma^2$ and $A\geq\max_k|a_{nk}\sigma|/A_n$. Furthermore, it can be seen that the assumption (A.1) in Lemma A.2 satisfies (C7) for each given j. Taken together with (C4), this leads to

$$\begin{aligned}&P\left\{\max_{1\leq j\leq d}\max_{1\leq i\leq n}\left|\sum_{k=1}^{n}\omega_{nk}(U_i)V_{jk}\right|>\zeta\right\}\\&\quad\leq dn\max_{1\leq j\leq d}\max_{1\leq i\leq n}P\left\{\left|\sum_{k=1}^{n}\omega_{nk}(U_i)V_{jk}\right|>\zeta\right\}\\&\quad\leq 2dn\exp\left\{-\frac{\zeta^2}{2A_n}(1+AM_v\zeta)\right\}\leq 2dn\exp\left\{-\frac{\zeta^2}{4b_n\mathbb{C}}\right\}\end{aligned}$$

$$= 2\exp\left\{-n^{3/10}\log^{-2}\frac{n}{\mathbb{C}} + \log(dn)\right\} \to 0.$$

Consequently, $\max_{1\le j\le d}\max_{1\le i\le n} |\widehat{G}_j(U_i) - G_j(U_i)| = o_p(n^{-1/4}\log^{-1} n)$. This completes the proof of Theorem 1.

Proof of Theorem 2. It is important to note that although we adapt the approach of Wang (2009), Theorem 1 and Lemma A.4 are critical to our proof of Theorem 2. For every $k \le [Kn^{\xi_0+4\xi_{\min}}]$, we have

$$\Omega(k) \doteq \mathrm{RSS}(\mathcal{M}^{(k)}) - \mathrm{RSS}(\mathcal{M}^{(k+1)}) = \left\| H^{(k)}_{a_{k+1}} Q_{(\mathcal{M}^{(k)})}\widehat{\mathbb{Y}}\right\|^2, \tag{B.2}$$

where $Q_{(\mathcal{M}^{(k)})} = I_n - H_{\mathcal{M}^{(k)}}$, $H_{\mathcal{M}^{(k)}} = \widehat{\mathbb{X}}_{(\mathcal{M}^{(k)})}\{\widehat{\mathbb{X}}^\top_{(\mathcal{M}^{(k)})}\widehat{\mathbb{X}}_{(\mathcal{M}^{(k)})}\}^{-1}\widehat{\mathbb{X}}^\top_{(\mathcal{M}^{(k)})}$, $H^{(k)}_{a_{k+1}} = \widehat{\mathbb{X}}^{(k)}_{a_{k+1}}\widehat{\mathbb{X}}^{(k)\top}_{a_{k+1}}\|\widehat{\mathbb{X}}^{(k)}_{a_{k+1}}\|^{-2}$, and $\widehat{\mathbb{X}}^{(k)}_{a_{k+1}} = \{I_n - H_{\mathcal{M}^{(k)}}\}\widehat{\mathbb{X}}_{a_{k+1}}$. Suppose that $\mathcal{M}_T \not\subset \mathcal{M}^{[Kn^{\xi_0+4\xi_{\min}}]}$. This leads to

$$\begin{aligned}\Omega(k) &\ge \max_{j\in\mathcal{M}^*_k}\left\| H^{(k)}_j Q_{(\mathcal{M}^{(k)})}\widehat{\mathbb{Y}}\right\|^2\\ &\ge 3^{-1}\max_{j\in\mathcal{M}^*_k}\left\| H^{(k)}_j Q_{(\mathcal{M}^{(k)})}\Big\{\mathbb{X}^*_{(\mathcal{M}_T)}\beta_{(\mathcal{M}_T)}\Big\}\right\|^2 - \max_{j\in\mathcal{M}_T}\left\| H^{(k)}_j Q_{(\mathcal{M}^{(k)})}\mathcal{E}\right\|^2\\ &\quad - \max_{j\in\mathcal{M}_T}\left\| H^{(k)}_j Q_{(\mathcal{M}^{(k)})}(\widehat{\mathbb{Y}} - \mathbb{Y}^*)\right\|^2,\end{aligned} \tag{B.3}$$

where $\mathcal{M}^*_k \doteq \mathcal{M}_T\backslash\mathcal{M}^{(k)} \ne \emptyset$, $\mathcal{E} = (\varepsilon_1,\ldots,\varepsilon_n)^\top \in \mathbb{R}^n$, and $H^{(k)}_j$ is the $H^{(k)}_{a_{k+1}}$ above. Under (C4)–(C7) and applying Theorem 1, $\max_i |\widehat{Y}_i - Y^*_i| = o_p(n^{-1/4}\log^{-1} n)$. In addition, both $H^{(k)}_j$ and $Q_{(\mathcal{M}^{(k)})}$ are projection matrices. Thus, the third term on the right side of (B.3) is bounded by $\|\widehat{\mathbb{Y}} - \mathbb{Y}^*\|^2 = n\cdot o_p(n^{-1/2})$. As a result, we need only focus on the first two terms on the right side of (B.3). After Theorem 1, we have

$$\max_{i,j}\left\|\widehat{X}_{ij} - X^*_{i,j}\right\| = o_p\Big(n^{-1/4}\log^{-1} n\Big). \tag{B.4}$$

We then can show that the first term (ignoring the constant) satisfies

$$\begin{aligned}&\max_{j\in\mathcal{M}^*_k}\left\| H^{(k)}_j Q_{(\mathcal{M}^{(k)})}\Big\{\mathbb{X}^*_{(\mathcal{M}_T)}\beta_{(\mathcal{M}_T)}\Big\}\right\|^2\\ &\quad = \max_{j\in\mathcal{M}^*_k}\left\| H^{(k)}_j Q_{(\mathcal{M}^{(k)})}\Big\{\mathbb{X}^*_{(\mathcal{M}^*_k)}\beta_{(\mathcal{M}^*_k)}\Big\}\right\|^2\\ &\quad \ge \Big\{\max_{j\in\mathcal{M}_T}\|\widehat{\mathbb{X}}_j\|^2\Big\}^{-1}\Big[\max_{j\in\mathcal{M}^*_k}\left|\widehat{\mathbb{X}}^\top_j Q_{(\mathcal{M}^{(k)})}\Big\{\mathbb{X}^*_{(\mathcal{M}^*_k)}\beta_{(\mathcal{M}^*_k)}\Big\}\right|^2\Big].\end{aligned} \tag{B.5}$$

Moreover, the Cauchy inequality and (C3) lead to

$$\begin{aligned}
&\left\|Q_{(\mathcal{M}^{(k)})}\left\{\mathbb{X}^*_{(\mathcal{M}^*_k)}\beta_{(\mathcal{M}^*_k)}\right\}\right\|^2 \\
&\quad= \sum_{j\in\mathcal{M}^*_k}\beta_j\left(\widehat{\mathbb{X}}_j^\top Q_{(\mathcal{M}^{(k)})}\left\{\mathbb{X}^*_{(\mathcal{M}^*_k)}\beta_{(\mathcal{M}^*_k)}\right\}\right) \\
&\quad\le \Big(\sum_{j\in\mathcal{M}^*_k}\beta_j^2\Big)^{1/2}\left(\sum_{j\in\mathcal{M}^*_k}\left[\widehat{\mathbb{X}}_j^\top Q_{(\mathcal{M}^{(k)})}\left\{\mathbb{X}^*_{(\mathcal{M}^*_k)}\beta_{(\mathcal{M}^*_k)}\right\}\right]^2\right)^{1/2} \\
&\quad\le C_\beta\cdot|\mathcal{M}_T|^{1/2}\max_{j\in\mathcal{M}^*_k}\left|\widehat{\mathbb{X}}_j^\top Q_{(\mathcal{M}^{(k)})}\left\{\mathbb{X}^*_{(\mathcal{M}^*_k)}\beta_{(\mathcal{M}^*_k)}\right\}\right|. \qquad \text{(B.6)}
\end{aligned}$$

Then, (B.6) together with (C1)–(C3) and Lemma A.4, demonstrates that the right side of (B.5) can be further bounded by

$$\begin{aligned}
&\Big\{\max_{j\in\mathcal{M}_T}\|\widehat{\mathbb{X}}_j\|^2\Big\}^{-1}\cdot\left[\left\|Q_{(\mathcal{M}^{(k)})}\left\{\mathbb{X}^*_{(\mathcal{M}^*_k)}\beta_{(\mathcal{M}^*_k)}\right\}\right\|^2\cdot|\mathcal{M}_T|^{-1/2}C_\beta^{-1}\right]^2 \\
&\quad\ge n\tau_{\max}^{-1}\tau_{\min}^2\beta_{\min}^4|\mathcal{M}_T|^{-1}C_\beta^{-2}\ge\tau_{\max}^{-1}\tau_{\min}^2C_\beta^{-2}\nu_\beta^4\nu^{-1}\cdot n^{1-\xi_0-4\xi_{\min}}. \qquad \text{(B.7)}
\end{aligned}$$

Consider the second term of (B.3). By (C1) and (C2) along with Lemma A.4, we have $\|\widehat{\mathbb{X}}_j^{(k)}\|^2\ge n\tau_{\min}$. After algebraic simplification with (B.4), we obtain

$$\left\|H_j^{(k)}Q_{(\mathcal{M}^{(k)})}\mathcal{E}\right\|^2\le\tau_{\min}^{-1}n^{-1}\max_{j\in\mathcal{M}_T}\max_{|\mathcal{M}|\le m^*}\left(\mathbb{X}_j^{*\top}Q_{(\mathcal{M})}\mathcal{E}\right)^2, \qquad \text{(B.8)}$$

where $m^*=Kn^{\xi_0+4\xi_{\min}}$. It is noteworthy that $\mathbb{X}_j^{*\top}Q_{(\mathcal{M})}\mathcal{E}$ is a normal random variable with mean 0 and variance $\|Q_{(\mathcal{M})}\mathbb{X}_j^*\|^2\le\|\mathbb{X}_j^*\|^2$. Accordingly, the right side of (B.8) is bounded by $\tau_{\min}^{-1}n^{-1}\max_{j\in\mathcal{M}_T}\|\mathbb{X}_j^*\|^2\cdot\max_{j\in\mathcal{M}_T}\max_{|\mathcal{M}|\le m^*}\chi_1^2$, which can be shown to be less than $3K\nu n^{\xi+\xi_0+4\xi_{\min}}$ with probability tending to one. This, in conjunction with (B.3) and (B.7), yields

$$\begin{aligned}
n^{-1}\Omega(k)\ge{}&3^{-1}\tau_{\max}^{-1}\tau_{\min}^2C_\beta^{-2}\nu_\beta^4\nu^{-1}n^{-\xi_0-4\xi_{\min}} \\
&\times\Big\{1-9K\nu^2\tau_{\max}\tau_{\min}^{-2}C_\beta^2\nu_\beta^{-4}n^{\xi+2\xi_0+8\xi_{\min}-1}\Big\}\{1+o_p(1)\} \qquad \text{(B.9)}
\end{aligned}$$

uniformly for every $k\le Kn^{\xi_0+4\xi_{\min}}$. Under (C4), we have

$$n^{-1}\|\widehat{\mathbb{Y}}\|^2\ge n^{-1}\sum_{k=1}^{[Kn^{\xi_0+4\xi_{\min}}]}\Omega(k)\to 2. \qquad \text{(B.10)}$$

Without loss of generality, we further assume that $\mathrm{var}(Y_i^*)=1$. Then, according to Theorem 1, we have $n^{-1}\|\widehat{\mathbb{Y}}\|^2\to_p 1$. This contradicts the result of (B.10), which implies that it is impossible to have $\mathcal{M}^{(k)} \bigcap \mathcal{M}_T = \emptyset$ for every $1 \leq k \leq Kn^{\xi_0 + 4\xi_{\min}}$. Consequently, with probability tending to one, all relevant variables are identified within a total of $Kn^{\xi_0 + 4\xi_{\min}}$ steps. This completes the proof.

References

An, H., Huang, D., Yao, Q. and Zhang, C.-H. (2008). Stepwise searching for feature variables in high-dimensional linear regression. Technical report. Department of Statistics, London School of Economics, London.

Bunea, F. (2004). Consistent covariate selection and post model selection inference in semiparametric regression. *Ann. Statist.* **32**, 898-927.

Cai, Z., Fan, J. and Li, R. (2000). Efficient estimation and inferences for varying-coefficient models. *J. Amer. Statist. Assoc.* **95**, 888-902.

Candes, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.* **35**, 2313-2351.

Carroll, R., Fan, J., Gijbels, I. and Wand, M. P. (1997). Generalized partially linear single-index models. *J. Amer. Statist. Assoc.* **92**, 477-489.

Chen, J. and Chen, Z. (2008). Extended bayesian information criterion for model selection with large model spaces. *Biometrika* **95**, 959-771.

Chernoff, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Statist.* **23**, 493-507.

Donoho, D. L., Drori, I., Tsaig, Y. and Starck, J. L. (2006). Sparse solution of underdetermined linear equations by stagewise orthogonal matching pursuit. Technical report. Department of Statistics, Stanford University.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Ann. Statist.* **32**, 407-499.

Engle, R., Granger, C., Rice, J. and Weiss, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *J. Amer. Statist. Assoc.* **81**, 310-320.

Fan, J., Feng, Y. and Song, R. (2010). Nonparametric independence screening in sparse ultra-high dimensional additive models. *Ann. Statist.* **38**, to appear.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*, Vol. 66 of *Monographs on Statistics and Applied Probability*, Chapman & Hall, London.

Fan, J. and Huang, T. (2005). Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli* **11**, 1031-1057.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348-1360.

Fan, J. and Li, R. (2006). Statistical challenges with high dimensionality: feature selection in knowledge discovery. *International Congress of Mathematicians.* Vol. III, Eur. Math. Soc., Zürich, 595-622.

Fan, J. and Lv, J. (2008). Sure independence screening for ultra-high-dimensional feature space (with discussion). *J. Roy. Statist. Soc. Ser. B* **70**, 849-911.

Härdle, W., Liang, H. and Gao, J. T. (2000). *Partially Linear Models.* Springer Physica, Heidelberg.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models (with discussion). *J. Roy. Statist. Soc. Ser. B* **55**, 757-796.

Heckman, N. E. (1986). Spline smoothing in partly linear models. *J. Roy. Statist. Soc. Ser. B* **48**, 244-248.

Hristache, M., Juditsky, A. and Spokoiny, V. (2001). Direct estimation of the index coefficient in a single-index model, *Ann. Statist.* **29**, 595-623.

Mack, Y. P. and Silverman, B. W. (1982). Weak and strong uniform consistency of kernel regression estimates, *Z. Wahrsch. Verw. Gebiete* **61**, 405-415.

Paul, D., Bair, E., Hastie, T. and Tibshirani, R. (2008). "Preconditioning" for feature selection and regression in high-dimensional problems. *Ann. Statist.* **36**, 1595-1618.

Pollard, D. (1984). *Convergence of Stochastic Processes*, Springer-Verlag, New York.

Speckman, P. E. (1988). Regression analysis for partially linear models. *J. Roy. Statist. Soc. Ser. B* **50**, 413-436.

Stodden, V. (2006). Model selection when the number of variables exceeds the number of observations. PhD thesis, Standford University.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58**, 267-288.

Wahba, G. (1984). Partial spline models for the semiparametric estimation of functions of several variables. *Statistical Analyses for Time Series*, Institute of Statistical Mathematics, Tokyo, 319-329. Japan-US Joint Seminar.

Wang, H. (2009). Forward regression for ultra-high dimensional variable screening, *J. Amer. Statist. Assoc.* **104**, 1512-1524.

Witten, D. and Tibshirani, R. (2009). Covariance-regularized regression and classification for high-dimensional problems. *J. Roy. Statist. Soc. Ser. B* **71**, 615-636.

Wold, H. (1966). Nonlinear estimation by iterative least square procedures. *Research Papers in Statistics (Festschrift J. Neyman)*, 411-444. John Wiley, London.

Xia, Y. (2006). Asymptotic distributions for two estimators of the single-index model. *Econometric Theory* **22**, 1112-1137.

Xie, H. and Huang, J. (2009). SCAD-penalized regression in high-dimensional partially linear models. *Ann. Statist.* **37**, 673-696.

Zhang, C.-H. and Huang, J. (2008). The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Ann. Statist.* **36**, 1567-1594.

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101**, 1418-1429.

Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY 14642, USA.

E-mail: hliang@bst.rochester.edu

Guanghua School of Management, Peking University, Beijing, 100871, P. R. China.

E-mail: hansheng@gsm.pku.edu.cn

Graduate School of Management, University of California, Davis, CA 95616, USA.

E-mail: cltsai@ucdavis.edu

(Received June 2010; accepted January 2011)