STATISTICAL THEORY AND RELATED FIELDS, 2017
VOL. 1, NO. 2, 234–245
https://doi.org/10.1080/24754269.2017.1396432

Taylor & Francis
Taylor & Francis Group

Sequential profile Lasso for ultra-high-dimensional partially linear models

Yujie Li[a,b], Gaorong Li[b] and Tiejun Tong[c]

[a]College of Applied Sciences, Beijing University of Technology, Beijing, China; [b]Beijing Institute for Scientific and Engineering Computing, Beijing University of Technology, Beijing, China; [c]Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong

ABSTRACT

In this paper, we study ultra-high-dimensional partially linear models when the dimension of the linear predictors grows exponentially with the sample size. For the variable screening, we propose a sequential profile Lasso method (SPLasso) and show that it possesses the screening property. SPLasso can also detect all relevant predictors with probability tending to one, no matter whether the ultra-high models involve both parametric and nonparametric parts. To select the best subset among the models generated by SPLasso, we propose an extended Bayesian information criterion (EBIC) for choosing the final model. We also conduct simulation studies and apply a real data example to assess the performance of the proposed method and compare with the existing method.

ARTICLE HISTORY
Received 12 October 2017
Revised 24 October 2017
Accepted 21 October 2017

KEYWORDS
Sequential profile Lasso; partially linear model; extended Bayesian information criterion; screening property; ultra-high-dimensional data

1. Introduction

High-dimensional data are becoming increasingly popular in the past two decades, and they have wide applications in various fields such as genomics, economics, finance and epidemiology. As one example, genome-wide association studies usually encompass hundreds of thousands, or millions, of single nucleotide polymorphism (SNP) at the same time, and that pose new computational and statistical challenges. To analyse ultra-high-dimensional data, Fan and Lv (2008) proposed the sure independence screening (SIS) method to ultra-high-dimensional linear models. Owing to its great success, the SIS method was further extended to more general models in the recent literature. To name a few, Fan, Feng, and Song (2011) proposed the nonparametric independence screening for ultra-high-dimensional additive models. Fan, Ma, and Dai (2014) extended the nonparametric independence screening to ultra-high-dimensional varying coefficient models. Li, Peng, Zhang, and Zhu (2012) developed a robust rank correlation screening based on Kendall's rank correlation. Li, Zhong, and Zhu (2012) proposed a sure independence screening method based on the distance correlation for general parametric models. Zhu, Li, Li, and Zhu (2011) proposed the variable screening method under a unified model framework. Cui, Li, and Zhong (2015) proposed a model free variable screening method for categorical response variable. Note also that Wang (2009) applied the forward regression (FR) in Weisberg (1980) to ultra-high-dimensional linear regression models. Cheng, Honda, and Zhang (2016) further extended the FR method to ultra-high-dimensional varying coefficient models.

Partially linear models are important semiparametric models and are widely used in practice, which possess both the flexibility of nonparametric models and the ease of interpretation of linear regression models. Partially linear models have been extensively studied in the literature (see for example Härdle, Liang, and Gao (2000) and Li, Zhang, and Feng (2016)). In this paper, we propose to analyse partially linear models when the dimension of the linear predictors grows exponentially with the sample size. Specifically, letting Y be the response variable, we consider the partially linear model as follows:

$$Y = g(U) + \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon, \tag{1}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is a p-dimensional vector of unknown regression coefficients, U is a univariate variable, $g(\cdot)$ is an unknown smooth function, and ε follows a distribution with mean 0 and variance σ^2. We assume that ε is independent of the associated covariates $(U, \boldsymbol{X}^{\mathrm{T}})$, and that the predictor variable $\boldsymbol{X}$ has a ultra-high-dimensionality or a nonpolynomial dimensionality such that $\ln p = O(n^{\kappa})$ for some $\kappa > 0$, where p is the dimension and n is the sample size.

Variable selection for model (1) is very challenging because it involves both parametric and nonparametric parts. When p is fixed, variable selection and parameter shrinkage are conventional and there are many related methods in the literature such as Bunea (2004), Liang and Li (2009), Liu, Wang, and Liang (2011), Mammen and van de Geer (1997) and Wang, Liu, Liang, and Carroll (2011). When p grows with n, Xie and Huang (2009) applied the SCAD penalty to partially linear models and studied the asymptotic properties of the

CONTACT Gaorong Li ligaorong@bjut.edu.cn, ligaorong@gmail.com

proposed estimators. Sherwood and Wang (2016) considered partially linear additive quantile regression models and studied the oracle property for a general class of nonconvex penalty functions. We note however that, due to the challenges in computational expediency, statistical accuracy and algorithm stability, the aforementioned penalised variable selection methods may not work well for model (1). To overcome the challenges, Liang, Wang, and Tsai (2012) proposed the profile forward regression (PFR) algorithm to perform the variable screening for model (1). Li, Li, Lian, and Tong (2017) extended the PFR algorithm to ultra-high-dimensional varying coefficient partially linear models.

The L_1 penalty or Lasso proposed by Tibshirani (1996) is a popular variable selection method. When p diverges to infinity faster than n but not too fast, under the irrepresentability condition, Zhao and Yu (2006) established the selection consistency for the fixed design, and Meinshausen and Bühlmann (2006) established the selection consistency for the random design. To alleviate the irrepresentability condition, Zou (2006) proposed the adaptive Lasso and showed that the adaptive Lasso has the oracle property when p is fixed. Luo (2012) and Luo and Chen (2014) proposed the sequential Lasso which chooses the largest tuning parameter in the sequentially partially penalised least squares objective function to assure at least one (mostly just one) of the regression coefficients being estimated as nonzero. Furthermore, under the partial positive cone condition, they proved the set of the predictors which maximise the correlation with the current residual, i.e., the response vector projected onto the orthogonal complement of the space spanned by the currently selected predictors, is the set of nonzero elements of the solution. This means that the next predictors being selected by the sequential Lasso can be chosen from the set of the predictors which maximise the correlation with the current residual. Such a method enjoys the expected theoretical properties including the screening property, meanwhile, it has some advantages from the numerical aspects.

Inspired by the above advantages, in this paper, we also apply the profile technique to convert model (1) to a linear model, and apply the sequential Lasso to develop a sequential profile Lasso (SPLasso) procedure. To select the best subset among the models generated by SPLasso, we propose an extended Bayesian information criterion (EBIC) for choosing the final model. We further show that our proposed SPLasso method can identify all relevant predictors with probability tending to one, and that the resulting model determined by EBIC possesses the screening property.

The rest of this paper is organised as follows. In Section 2, the SPLasso procedure is introduced for model (1). In Section 3, the asymptotic properties are derived under some regularity conditions. In Section 4, simulation studies are carried out to evaluate the finite sample performance of our proposed method and to compare it with existing method. Section 5 presents the application of our proposed method to a real data set. The technical proofs of the two theorems, together with some lemmas, are given in the Appendix.

2. Sequential profile Lasso

To avoid confusion, we specify in the beginning that the boldface roman $\mathbf{B}$ represents a matrix, and the boldface italics $\boldsymbol{B}$ represents a vector. Throughout this paper, we denote $\gamma_{\min}(\mathbf{B})$ and $\gamma_{\max}(\mathbf{B})$ as the smallest and largest eigenvalues of an arbitrary matrix $\mathbf{B}$, respectively. Suppose that $\{(Y_i, \boldsymbol{X}_i^{\mathrm{T}}, U_i), 1 \le i \le n\}$ are independent and identically distributed copies of $(Y, \boldsymbol{X}^{\mathrm{T}}, U)$ that are generated from model (1). For ease of notation, we denote $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}} \in \mathbb{R}^n$ as the response vector, $\mathbf{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^{\mathrm{T}} \in \mathbb{R}^{n\times p}$ as the matrix of explanatory variables, where $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^{\mathrm{T}} \in \mathbb{R}^p$ is the predictor vector, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^{\mathrm{T}}$ as the vector of random errors. We write $\mathcal{M}_{\mathcal{F}} = \{1, \ldots, p\}$ and $\mathcal{M}_{\mathcal{T}} = \{j : \beta_j \neq 0\}$ as the index sets of the full and true predictors, respectively. Let also $|\mathcal{M}|$ denote the number of the elements of a candidate model $\mathcal{M}$, where $\mathcal{M}$ is the index set of the predictors in the corresponding candidate model. Thus, $|\mathcal{M}_{\mathcal{F}}| = p$ and $|\mathcal{M}_{\mathcal{T}}| = p_0$, where p_0 is the size of the true model or the number of relevant predictors in the true model. For any candidate model $\mathcal{M}$, we use $\boldsymbol{X}_{i(\mathcal{M})} = \{X_{ij} : j \in \mathcal{M}\}$ to represent the subvector of $\boldsymbol{X}_i$ corresponding to $\mathcal{M}$, and $\mathbf{X}_{\mathcal{M}} = \{X_{ij}, i = 1, \ldots, n, j \in \mathcal{M}\}$ to denote the matrix consisting of the column of $\mathbf{X}$ with indices in $\mathcal{M}$. Similarly, let $\boldsymbol{\beta}_{\mathcal{M}}$ denote the vector consisting of the corresponding components of $\boldsymbol{\beta}$. For any candidate model $\mathcal{M}$, let $\mathcal{M}^c$ be the complement of $\mathcal{M}$ in the full model $\mathcal{M}_{\mathcal{F}}$.

By model (1) and the fact that $g(U_i) = \mathrm{E}(Y_i|U_i) - \mathrm{E}(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}|U_i)$, we have

$$Y_i - \mathrm{E}(Y_i|U_i) = \{\boldsymbol{X}_i - \mathrm{E}(\boldsymbol{X}_i|U_i)\}^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon_i. \tag{2}$$

For simplicity, we define the profile response as $Y_i^* = Y_i - \mathrm{E}(Y_i|U_i)$ and the profile predictor vector as $\boldsymbol{X}_i^* = \boldsymbol{X}_i - \mathrm{E}(\boldsymbol{X}_i|U_i) = (X_{i1}^*, \ldots, X_{ip}^*)^{\mathrm{T}}$, where $X_{ij}^* = X_{ij} - \mathrm{E}(X_{ij}|U_i)$ for $i = 1, \ldots, n$ and $j = 1, \ldots, p$. By (2), model (1) reduces to the following linear regression model:

$$Y_i^* = \boldsymbol{X}_i^{*\mathrm{T}}\boldsymbol{\beta} + \varepsilon_i. \tag{3}$$

Note that model (3) contains the unknown functions $\mathrm{E}(Y_i|U_i)$ and $\mathrm{E}(\boldsymbol{X}_i|U_i)$ and they need to be estimated in practice. In this paper, we approximate $\mathrm{E}(Y_i|U_i)$ locally by a linear function, and consider the following objective function (Fan & Gijbels, 1996):

$$\sum_{i=1}^{n} (Y_i - \alpha_1 - \alpha_2(U_i - u))^2 K_h(U_i - u),$$

where $K_h(\cdot) = K(\cdot/h)/h$ with $K(\cdot)$ a kernel function and h a bandwidth. By minimising the above weighted

least squares objective function, we can obtain the local linear regression estimator of $\mathrm{E}(Y_i|U_i = u)$ as follows:

$$\sum_{k=1}^{n} w_k(u) Y_k,$$

where

$$w_k(u) = \frac{\{S_{n2}(u) - (U_k - u)S_{n1}(u)\}K_h(U_k - u)}{S_{n2}(u)S_{n0}(u) - S_{n1}^2(u)},$$

and $S_{n\ell} = \sum_{i=1}^{n} K_h(U_i - u)(U_i - u)^{\ell}$ for $\ell = 0, 1, 2$. We note that the above method can also be applied to estimate $\mathrm{E}(X_{ij}|U_i = u)$ for $1 \le j \le p$. To facilitate the notation, we write the estimator of the profile response as $\widetilde{\boldsymbol{Y}} = (\widetilde{Y}_1, \ldots, \widetilde{Y}_n)^{\mathrm{T}}$ with $\widetilde{Y}_i = Y_i - \sum_{k=1}^{n} w_k(U_i)Y_k$, and the estimators of the profile predictors as $\widetilde{\mathbf{X}} = (\widetilde{\boldsymbol{X}}_1, \ldots, \widetilde{\boldsymbol{X}}_n)^{\mathrm{T}} = (\widetilde{\boldsymbol{x}}_1, \ldots, \widetilde{\boldsymbol{x}}_p)$ with $\widetilde{\boldsymbol{X}}_i = \boldsymbol{X}_i - \sum_{k=1}^{n} w_k(U_i)\boldsymbol{X}_k$. This gives rise to the following linear model:

$$\widetilde{Y}_i \approx \widetilde{\boldsymbol{X}}_i^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon_i. \tag{4}$$

In what follows, we introduce our SPLasso procedure. At the initial step, the SPLasso method minimises the following penalised least squares objective function:

$$\mathcal{L}_1 = (\widetilde{\boldsymbol{Y}} - \widetilde{\mathbf{X}}\boldsymbol{\beta})^{\mathrm{T}}(\widetilde{\boldsymbol{Y}} - \widetilde{\mathbf{X}}\boldsymbol{\beta}) + \lambda_1 \sum_{j=1}^{p} |\beta_j|, \tag{5}$$

where λ_1 is the largest value such that at least one (mostly just one) β_j will be estimated as nonzero. The index set of the selected predictors with nonzero estimated coefficient is labelled as $\mathcal{M}_1$. Let $\mathcal{M}_k$ be the index set of the predictors being selected until step k. At the $(k+1)$th step, we consider the following partially penalised objective function:

$$\mathcal{L}_{k+1} = (\widetilde{\boldsymbol{Y}} - \widetilde{\mathbf{X}}\boldsymbol{\beta})^{\mathrm{T}}(\widetilde{\boldsymbol{Y}} - \widetilde{\mathbf{X}}\boldsymbol{\beta}) + \lambda_{k+1} \sum_{j \notin \mathcal{M}_k} |\beta_j|, \tag{6}$$

where λ_{k+1} is the largest value such that at least one (mostly just one) β_j, $j \notin \mathcal{M}_k$, will be estimated as nonzero. According to the Karush–Kuhn–Tucker (KKT) condition (see Proposition 3.3.1 in Bertsekas, 1999), the solution of $\mathcal{L}_{k+1}$ is equivalent to the minimisation of

$$(\widehat{\boldsymbol{Y}} - \widehat{\mathbf{X}}_{\mathcal{M}_k^c}\boldsymbol{\beta}_{\mathcal{M}_k^c})^{\mathrm{T}}(\widehat{\boldsymbol{Y}} - \widehat{\mathbf{X}}_{\mathcal{M}_k^c}\boldsymbol{\beta}_{\mathcal{M}_k^c}) + \lambda_{k+1} \sum_{j \in \mathcal{M}_k^c} |\beta_j|,$$

where $\widehat{\boldsymbol{Y}} = \left(\mathbf{I}_n - \mathbf{H}_{\mathcal{M}_k}\right)\widetilde{\boldsymbol{Y}}$, $\widehat{\mathbf{X}} = \left(\mathbf{I}_n - \mathbf{H}_{\mathcal{M}_k}\right)\widetilde{\mathbf{X}} = (\widehat{\boldsymbol{x}}_1, \ldots, \widehat{\boldsymbol{x}}_p)$, $\mathbf{I}_n$ is an $n \times n$ identity matrix, and $\mathbf{H}_{\mathcal{M}_k} = \widetilde{\mathbf{X}}_{\mathcal{M}_k}\{\widetilde{\mathbf{X}}_{\mathcal{M}_k}^{\mathrm{T}}\widetilde{\mathbf{X}}_{\mathcal{M}_k}\}^{-1}\widetilde{\mathbf{X}}_{\mathcal{M}_k}^{\mathrm{T}}$. Let also $\mathcal{S}_{k+1} = \{j : |\widehat{\boldsymbol{x}}_j^{\mathrm{T}}\widehat{\boldsymbol{Y}}| = \max_{\ell \notin \mathcal{M}_k} |\widehat{\boldsymbol{x}}_\ell^{\mathrm{T}}\widehat{\boldsymbol{Y}}|\}$. Following Luo and Chen (2014), we consider the new predictors selected by the $(k+1)$th step are from the set $\mathcal{S}_{k+1}$. For this, we consider the following objective function:

$$(\widehat{\boldsymbol{Y}} - \widehat{\mathbf{X}}_{\mathcal{S}_{k+1}}\boldsymbol{\beta}_{\mathcal{S}_{k+1}})^{\mathrm{T}}(\widehat{\boldsymbol{Y}} - \widehat{\mathbf{X}}_{\mathcal{S}_{k+1}}\boldsymbol{\beta}_{\mathcal{S}_{k+1}}) + \lambda_{k+1} \sum_{j \in \mathcal{S}_{k+1}} |\beta_j|. \tag{7}$$

If $\mathcal{S}_{k+1}$ has only one element, then the jth predictor with $j \in \mathcal{S}_{k+1}$ is the predictor with nonzero estimated coefficient in the minimisation of $\mathcal{L}_{k+1}$. If $\mathcal{S}_{k+1}$ has more than one element, we need to minimise the objective function (7) by applying the `R` function '`glmpath`' developed by Park and Hastie (2007). Our proposed SPLasso is as follows:

[S1] Let $\mathcal{S}_1 = \{j : |\widetilde{\boldsymbol{x}}_j^{\mathrm{T}}\widetilde{\boldsymbol{Y}}| = \max_{1 \le \ell \le p} |\widetilde{\boldsymbol{x}}_\ell^{\mathrm{T}}\widetilde{\boldsymbol{Y}}|\}$. If $\mathcal{S}_1$ has only one element, we update $\mathcal{M}_1 = \mathcal{S}_1$. Otherwise, use '`glmpath`' to $\widetilde{\boldsymbol{Y}}$ and $\widetilde{\mathbf{X}}_{\mathcal{S}_1}$ and obtain the solution path. Let $\mathcal{M}_1$ be the index of the first predictor with nonzero estimated coefficient in the solution path.

[S2] At the $(k+1)$th step, let $\mathcal{S}_{k+1} = \{j : |\widehat{\boldsymbol{x}}_j^{\mathrm{T}}\widehat{\boldsymbol{Y}}| = \max_{\ell \notin \mathcal{M}_k} |\widehat{\boldsymbol{x}}_\ell^{\mathrm{T}}\widehat{\boldsymbol{Y}}|\}$, where $\widehat{\boldsymbol{Y}} = \left(\mathbf{I}_n - \mathbf{H}_{\mathcal{M}_k}\right)\widetilde{\boldsymbol{Y}}$ and $\widehat{\boldsymbol{x}}_j = \left(\mathbf{I}_n - \mathbf{H}_{\mathcal{M}_k}\right)\widetilde{\boldsymbol{x}}_j$. If $\mathcal{S}_{k+1}$ has only one element, we update $\mathcal{M}_{k+1} = \mathcal{M}_k \bigcup \mathcal{S}_{k+1}$. Otherwise, use '`glmpath`' to $\widehat{\boldsymbol{Y}}$ and $\widehat{\mathbf{X}}_{\mathcal{S}_{k+1}}$, where $\widehat{\mathbf{X}}_{\mathcal{S}_{k+1}} = \left(\mathbf{I}_n - \mathbf{H}_{\mathcal{M}_k}\right)\widetilde{\mathbf{X}}_{\mathcal{S}_{k+1}}$, and obtain the solution path. We add the index of the first predictor with nonzero estimated coefficient in the solution path in the current model $\mathcal{M}_k$, and write the new model as $\mathcal{M}_{k+1}$.

[S3] Iterate the S2 step for n times to obtain a total of n nested candidate models by the solution path $\mathbb{S} = \{\mathcal{M}_k : 1 \le k \le n\}$.

Note that we can update $\mathbf{I}_n - \mathbf{H}_{\mathcal{M}_{k+1}}$ from $\mathbf{I}_n - \mathbf{H}_{\mathcal{M}_k}$. Suppose the predictors with indices $\{j_k: k = 1, \ldots, M\}$ are added to the current model at the $(k+1)$th step, and denote the index set as $\mathcal{B}_m = \{j_1, \ldots, j_m\}$ for $m \ge 1$ and let $\mathcal{B}_0 = \emptyset$. The recursive formula is given by

$$\mathbf{I}_n - \mathbf{H}_{\mathcal{M}_k \cup \mathcal{B}_m} = \left(\mathbf{I}_n - \mathbf{H}_{\mathcal{M}_k \cup \mathcal{B}_{m-1}}\right) \times \left\{\mathbf{I}_n - \frac{\widetilde{\boldsymbol{x}}_{j_m}\widetilde{\boldsymbol{x}}_{j_m}^{\mathrm{T}}\left(\mathbf{I}_n - \mathbf{H}_{\mathcal{M}_k \cup \mathcal{B}_{m-1}}\right)}{\widetilde{\boldsymbol{x}}_{j_m}^{\mathrm{T}}\left(\mathbf{I}_n - \mathbf{H}_{\mathcal{M}_k \cup \mathcal{B}_{m-1}}\right)\widetilde{\boldsymbol{x}}_{j_m}}\right\}.$$

By the above discussion, it is evident that our proposed SPLasso procedure has the advantage of reducing the computational burden by avoiding the computation of the inverse matrices.

3. Asymptotic properties

In this section, we establish the screening property of SPLasso. We use an EBIC to obtain the best model in the solution path $\mathbb{S}$ and show that this model contains the true model $\mathcal{M}_{\mathcal{T}}$ with probability tending to one. To derive the theoretical results, we need some regularity conditions.

(C1) There exist two positive constants $\tau_{\min}$ and $\tau_{\max}$, such that $2\tau_{\min} < \gamma_{\min}(\Sigma) \le \gamma_{\max}(\Sigma) < 2^{-1}\tau_{\max}$, where Σ is the covariance matrix of the profile predictor $\boldsymbol{X}_i^*$.

(C2) Assume that $\|\boldsymbol{\beta}\|_\infty \le C_{\boldsymbol{\beta}}$ for some positive constant $C_{\boldsymbol{\beta}}$ and $\beta_{\min} \ge \nu_{\boldsymbol{\beta}} n^{-\xi_{\min}}$ for some positive constants $\xi_{\min}$ and $\nu_{\boldsymbol{\beta}}$, where $\beta_{\min} = \min_{j\in\mathcal{M}_\mathcal{T}} |\beta_j|$.

(C3) There exist positive constants ξ, ξ_0 and ν, such that $\ln p \le \min(\nu n^{\xi}, n^{3/10})$, $p_0 \le \nu n^{\xi_0}$, and $\xi + 3\xi_0 + 6\xi_{\min} < 1$.

(C4) $\mathrm{E}(\boldsymbol{X}|U=u)$ and $\mathrm{E}(Y|U=u)$ are uniformly Lipschitz continuous of order one.

(C5) The weight functions $w_k(\cdot)$ satisfy, with probability tending to one, $\max_{1\le k\le n}\sum_{i=1}^n w_k(U_i) = O(1)$, $\max_{1\le i,k\le n} w_k(U_i) = O(b_n)$ with $b_n = n^{-4/5}$, and $\max_{1\le i\le n}\sum_{k=1}^n w_k(U_i) I(|U_i - U_k| > c_n) = O(c_n)$ with $c_n = n^{-2/5}\ln n$.

(C6) Assume that $\max\{\mathrm{E}\exp(u|Y_i^*|), \max_{1\le j\le p}\mathrm{E}\exp(u|X_{ij}^*|)\} < \infty$ for all $0 \le u \le t_0/\sigma_\nu$, where t_0 and σ_ν are positive constants, and that the moment generating functions $M_j(u)$ of X_{ij}^* for $j = 1, \ldots, p$ and $M_0(u)$ of Y_i^* satisfy

$$\max_{0\le j\le p}\sup_{0\le u\le t_0}\left|\frac{d^3}{du^3}\ln\{M_j(u)\}\right| < \infty.$$

Furthermore, assume that $\max\{\mathrm{E}|Y_i^*|^{2k}, \max_{1\le j\le p}\mathrm{E}|X_{ij}^*|^{2k}\} \le \sigma_v^2$ for some $k > 2$, and that ε follows a normal distribution.

Conditions (C1)–(C3) are technical requirements for model selection (see Li et al., 2017; Liang et al., 2012; Wang, 2009). Conditions (C4) and (C5) are commonly used in the semiparametric regression and can be easily verified (see Härdle et al., 2000). Condition (C6) follows from Liang et al. (2012) to obtain an exponential inequality of a sum of random variables. It is worth noting that we do not pose any restriction on the distribution on $\boldsymbol{X}$, whereas the SIS method in Fan and Lv (2008) requires it to be the spherically symmetric distribution, and the FR method in Wang (2009) and Lasso in Zhang and Huang (2008) require it to be the normal distribution. In addition, we replace the L_2 norm with the maximum norm of $\boldsymbol{\beta}$ that is slightly different from Li et al. (2017), Liang et al. (2012) and Wang (2009).

Theorem 3.1: *Suppose that regularity conditions (C1)–(C6) hold, and let*

$$K_n = \left[\frac{(1+\lambda_0)^2\tau_{\max}^2 C_{\boldsymbol{\beta}}^2 \nu}{\tau_{\min}^2 \nu_{\boldsymbol{\beta}}^2} n^{\xi_0+2\xi_{\min}}\right] + 1,$$

where [t] denotes the largest integer less than t and λ_0 is a constant larger than 1. Then,

$$\Pr\left(\mathcal{M}_\mathcal{T} \subset \mathcal{M}_{K_n}\right) \to 1, \tag{8}$$

where $\mathcal{M}_{K_n}$ denotes the selected K_nth model in the solution path $\mathbb{S}$.

Theorem 3.1 shows that the proposed SPLasso procedure can identify all relevant predictors within $O(n^{\xi_0+2\xi_{\min}})$ steps with probability tending to one, which is better than the order of $O(n^{\xi_0+4\xi_{\min}})$ derived in Theorem 2 of Liang et al. (2012). Since the models generated by SPLasso are nested, we need to determine which model should be used for further statistical inference. To this end, we consider the EBIC as follows:

$$\mathrm{EBIC}(\mathcal{M}) = \ln(\widehat{\sigma}_{(\mathcal{M})}^2) + n^{-1}|\mathcal{M}|\left(\ln n + 2\eta\ln p\right), \tag{9}$$

where η is a fixed positive constant, $\mathcal{M}$ is any candidate model with $|\mathcal{M}| \le n$, and

$$\widehat{\sigma}_{(\mathcal{M})}^2 = n^{-1}\mathrm{RSS}(\mathcal{M}) = \widetilde{\boldsymbol{Y}}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{H}_\mathcal{M})\widetilde{\boldsymbol{Y}}/n,$$

where $\mathbf{H}_\mathcal{M} = \widetilde{\mathbf{X}}_\mathcal{M}\{\widetilde{\mathbf{X}}_\mathcal{M}^{\mathrm{T}}\widetilde{\mathbf{X}}_\mathcal{M}\}^{-1}\widetilde{\mathbf{X}}_\mathcal{M}^{\mathrm{T}}$. Note that for $\zeta = 1$, EBIC has been used in Chen and Chen (2008), Liang et al. (2012) and Wang (2009). Let $\widehat{k} = \arg\min_{1\le k\le n}\mathrm{EBIC}(\mathcal{M}_k)$, then the resulting model is $\mathcal{M}_{\widehat{k}}$. In the following, we show that $\mathcal{M}_{\widehat{k}}$ contains the true model with probability tending to one.

Theorem 3.2: *Under regularity conditions (C1)–(C6), as $n \to \infty$, we have*

$$\Pr(\mathcal{M}_\mathcal{T} \subset \mathcal{M}_{\widehat{k}}) \longrightarrow 1. \tag{10}$$

4. Simulation study

In this section, we present the results of Monte Carlo simulations to evaluate the finite sample performance of the proposed SPLasso procedure. We employ the Epanechnikov kernel $K(u) = 0.75(1-u^2)_+$ and the bandwidth $h = 1.5\widehat{\sigma}_U n^{-1/5}$, where $\widehat{\sigma}_U$ is the sample standard deviation of U. In all examples, the variable U is generated from the uniform distribution on $[0, 1]$. We consider n =100, 150 and 200, p =500, 1000 and 2000, and compare SPLasso with the PFR method in Liang et al. (2012).

Let $\widehat{\boldsymbol{\beta}}^{(k)} = (\widehat{\beta}_{1k}, \ldots, \widehat{\beta}_{pk})^{\mathrm{T}} \in \mathbb{R}^p$ be the estimator obtained from the kth simulation, and the resulting model be $\widehat{\mathcal{M}}^{(k)} = \{j : |\widehat{\beta}_{jk}| > 0\}$. We consider the following eight performance measures to evaluate the performance of SPLasso: (1) AMS: the average model size of the resulting model based on 200 simulations; (2) CP: the average coverage probability that all relevant predictors are detected among 200 simulations; (3) CZ: the proportion of correct identified zeros among 200 simulations; (4) IZ: the proportion of incorrect identified zeros among 200 simulations; (5) CF: the average of correctly fitted that all relevant predictors are detected and no irrelevant predictors are contained in the resulting model among 200 simulations; (6) AEE: the average estimation error is computed as $\sum_{k=1}^{200}\|\widehat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}\|_2/200$; (7) PDR: the average positive discovery rate is computed as $\sum_{k=1}^{200}|\widehat{\mathcal{M}}^{(k)}\bigcap\mathcal{M}_\mathcal{T}|/(200p_0)$; and (8) FDR: the average false discovery rate is computed as $\sum_{k=1}^{200}|\widehat{\mathcal{M}}^{(k)}\bigcap\mathcal{M}_\mathcal{T}^c|/(200|\widehat{\mathcal{M}}^{(k)}|)$.

Example 4.1: In this example, we consider that the relevant predictors are independent and are generated from

Table 1. Simulation results for Example 4.1.

p	n	AMS	PDR	FDR	AEE	CP	CZ	IZ	CF
				Method:	PFR				
500	100	13.810	0.848	0.378	1.926	0.025	0.989	0.152	0.000
	150	14.225	0.916	0.350	1.170	0.325	0.989	0.084	0.000
	200	13.930	0.974	0.293	0.498	0.770	0.991	0.026	0.000
1000	100	12.315	0.812	0.337	2.804	0.015	0.996	0.188	0.000
	150	13.200	0.880	0.330	1.817	0.150	0.996	0.120	0.000
	200	13.530	0.953	0.290	0.881	0.590	0.996	0.047	0.000
2000	100	12.295	0.803	0.343	3.159	0.000	0.998	0.198	0.000
	150	13.240	0.860	0.347	2.348	0.065	0.998	0.141	0.000
	200	13.450	0.927	0.305	1.295	0.385	0.998	0.073	0.000
				Method:	SPLasso				
500	100	11.490	1.000	0.124	0.274	1.000	0.997	0.000	0.000
	150	11.285	1.000	0.111	0.137	1.000	0.997	0.000	0.000
	200	11.155	1.000	0.103	0.064	1.000	0.998	0.000	0.000
1000	100	11.060	0.990	0.096	0.121	0.995	0.999	0.001	0.000
	150	11.035	1.000	0.094	0.060	1.000	0.999	0.000	0.000
	200	11.015	1.000	0.092	0.037	1.000	0.999	0.000	0.000
2000	100	11.085	0.970	0.097	0.123	0.990	0.999	0.008	0.000
	150	11.015	1.000	0.092	0.067	1.000	0.999	0.000	0.000
	200	11.030	1.000	0.093	0.045	1.000	0.999	0.000	0.000

the standard normal distribution. The regression coefficient vector of relevant predictors is (3, 3.75, 4.5, 5.25, 6, 6.75, 7.5, 8.25, 9, 9.75). The irrelevant predictors are generated as

$$X_j = 0.25Z_j + \sqrt{0.75}\sum_{k\in\mathcal{M}_\mathcal{T}} X_k, \text{ for } j \in \mathcal{M}_\mathcal{T}^c,$$

where Z_j are independent standard normal random variables, and are independent of the relevant predictors. The nonlinear function is $g(U) = 4\sin(2\pi U)$, a non-monotonic function. The noises ε are generated from the standard normal distribution. The simulation results are reported in Table 1.

Example 4.2: In this simulation, we consider a compound symmetry structure for the covariance of the relevant predictors. Specially, the relevant predictors follow the p_0-dimensional multivariate normal distribution $\mathcal{N}(0, \Sigma_0)$. The size of the true model is set to $p_0 = 8$. The covariance matrix Σ_0 has $\sigma_{ii} = 1$ and $\sigma_{ij} = 0.5$ for $1 \le i \ne j \le p_0$. The irrelevant predictors X_j are generated as

$$X_j = \epsilon_j + \frac{1}{p_0}\sum_{1\le k\le p_0} X_k,$$

where ϵ_j are independent and identically distributed with $\mathcal{N}(0, 0.08)$. The nonzero coefficients are generated as $(-1)^V(4n^{-0.15} + |T|)$, where V is a binary random variable with $\Pr(V = 1) = 0.4$ and T is a normal random variable with mean 0 and satisfies $\Pr(|T| \ge 0.1) = 0.25$. Let the nonlinear component $g(U) = \exp(3U)$, a monotone function. The noises ε are independent and identically distributed with $\mathcal{N}(0, \sigma^2)$. The variance σ^2 is chosen such that $\text{SNR} = \text{var}(\boldsymbol{X}^\text{T}\boldsymbol{\beta} + g(U))/\text{var}(Y)$ is approximately 80%. The simulation results are reported in Table 2.

From Tables 1 and 2, we have the comparison results in what follows.

(1) The irrelevant predictors are equally and almost highly correlated with the relevant predictors in both examples, and SPLasso is always better than PFR. For example, when $n = 200$, PFR has the CP values at 0.770, 0.590 and 0.385, while SPLasso has the CP values all at 1, for $p = 500$, 1000 and 2000, respectively, in Table 1.
(2) Note that the larger PDR values and the smaller FDR values are, the better the associated procedure performs. From this point of view, SPLasso also behaves better than PFR.
(3) For the fixed p, SPLasso performs better as the sample size increases. It is clear that the coverage probability changes substantially as the sample size increases. In addition, the coverage probability approaches 1 as long as the sample size is enough large. For the fixed n, the finite sample performance of SPLasso becomes worse as the dimension of predictors increases. However, from the variation rate, we note that the performance does not deteriorate rapidly as the dimension p increases. This means that the sample size is more important than the dimension of predictors in ultra-high-dimensional variable screening.
(4) Note that the proportion of correctly zeros is almost 1. As a result, the average model size is small, and is close to the true model size when the sample size increases. Consequently, the average estimation error decreases as the sample size increases.

In conclusion, the numerical results are in line with the theoretical results that SPLasso contains the true model with probability tending to one. These results

Table 2. Simulation results for Example 4.2.

p	n	AMS	PDR	FDR	AEE	CP	CZ	IZ	CF
				Method:	PFR				
500	100	6.875	0.467	0.490	111.060	0.050	0.994	0.533	0.025
	150	7.255	0.638	0.354	80.577	0.385	0.996	0.363	0.290
	200	7.490	0.658	0.335	55.135	0.540	0.995	0.340	0.500
1000	100	7.375	0.464	0.516	144.029	0.055	0.996	0.536	0.050
	150	6.980	0.523	0.455	114.385	0.230	0.997	0.478	0.180
	200	7.230	0.603	0.380	90.892	0.435	0.998	0.397	0.345
2000	100	9.150	0.426	0.608	205.817	0.010	0.997	0.574	0.005
	150	7.500	0.557	0.452	126.259	0.265	0.998	0.443	0.155
	200	7.475	0.591	0.406	95.332	0.370	0.999	0.409	0.280
				Method:	SPLasso				
500	100	5.865	0.570	0.281	98.091	0.345	0.997	0.430	0.195
	150	7.205	0.757	0.218	62.354	0.615	0.998	0.243	0.410
	200	7.760	0.840	0.165	34.328	0.730	0.998	0.160	0.500
1000	100	6.145	0.561	0.315	116.950	0.355	0.998	0.439	0.155
	150	6.580	0.623	0.306	100.934	0.425	0.998	0.378	0.270
	200	7.410	0.741	0.239	70.525	0.590	0.999	0.259	0.380
2000	100	6.250	0.486	0.375	144.353	0.220	0.999	0.514	0.085
	150	7.08	0.666	0.305	106.756	0.500	0.999	0.334	0.250
	200	7.295	0.726	0.250	72.007	0.570	0.999	0.274	0.355

demonstrate that SPLasso is one of the best variable screening methods and it can be useful in real data analysis.

5. Real data analysis

We demonstrate the effectiveness of SPLasso by an application to a breast cancer data. As reported in Stewart and Wild (2014), breast cancer is one of the leading causes of cancer death among women, and there were about 1.7 million new cases (25% of all cancers in women) and 0.5 million cancer deaths (15% of all cancer deaths in women) in 2012. Breast cancer is the most common cancer diagnosis in women in 140 countries and is the most frequent cause of cancer mortality in 101 countries. van't Veer et al. (2002) collected the samples from a total of 97 lymph node-negative breast cancer patients under 55 years old. The collected dataset consists of expression levels for 24,481 gene probes and seven clinical risk factors including age, tumour size, histological grade, angioinvasion, lymphocytic infiltration, estrogen receptor (ER) and progesterone receptor status for the 97 participators in this study. Among the 97 participators, 46 developed distant metastases within 5 years and 51 remained metastases free for more than 5 years. Yu, Li, and Ma (2012) proposed a receiver operating characteristic approach to rank the genes by adjusting the clinical risk factors. In addition, they removed the severe missing genes, and obtained an effective number of 24,188 genes. Each data vector is normalised to have sample mean 0 and standard deviation 1.

Knight, Livingston, Gregory, and McGuire (1977) found the absence of estrogen receptor in primary breast tumours is associated with the early recurrence. We are interested in finding genes that are related to the estrogen receptor. We consider the following partially linear model to fit the data:

$$\mathrm{ER} = g(U) + \sum_{j=1}^{24,188} \beta_j \mathrm{GE}_j + \varepsilon, \tag{11}$$

where U is the age of the patients, GE_j is the jth gene.

As in Section 4, the Epanechnikov kernel $K(u) = 0.75(1-u^2)_+$ and the bandwidth $h = 1.5\widehat{\sigma}_U n^{-1/5}$ are adopted to fit the nonlinear function, where $\widehat{\sigma}_U$ denotes the sample standard deviation of U. We first compare SPLasso with PFR in terms of the prediction mean squared errors (PMSE) based on 100 random partitions. For each partition, we randomly select 90 observations as the training set and the remaining seven observations as the test set. Based on the training set, we fit the data with the partially linear model (11) via SPLasso and PFR. The resulting models are used to predict the value of the seven observations in the test set. The five-number summary of the prediction mean squared errors is listed in Table 3. From Table 3, we know that SPLasso is better than PFR in terms of the PMSE.

We observe that different models are often selected for different random partitions. Table 4 shows the top five genes selected by SPLasso and PFR in the 100 random partitions. Gene 15835 is detected as important at each time among the 100 random partitions by both methods. We note that gene 15835 was also identified in Cheng et al. (2016). In addition, gene 1279 is also identified by both methods. From all these findings, we conclude that gene 15835 is associated with the estrogen receptor.

To investigate the estimated nonlinear effects of the patient's age based on one random split in which SPLasso identified two genes 15835 and 14117, we present the estimated nonparametric function in Figure 1. It shows that the patient's age almost has a positive impact on the estrogen receptor. We note that the value of effect first decreases (up to about 5.2), then

Table 3. Five-number summary of PMSEs for PFR and SPLasso.

Method	Minimum	First quartile	Median	Third quartile	Maximum
PFR	0.139	0.403	0.602	1.839	3.494
SPLasso	0.130	0.305	0.368	0.478	1.906

Table 4. Top five genes selected among 100 random partitions.

PFR	1279	1690	5985	15835	23134
SPLasso	1279	6257	9918	14117	15835

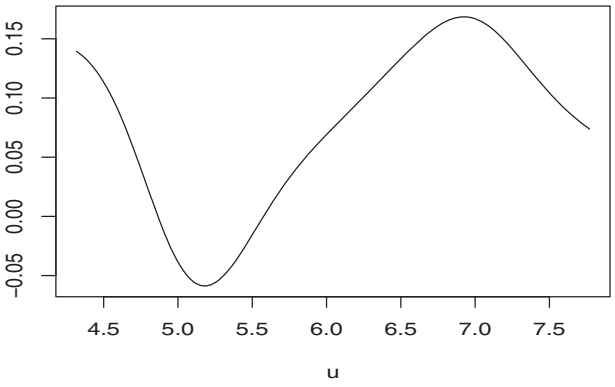

Figure 1. The fitted nonlinear function $\hat{g}(u)$.

increases (up to about 7), and then decreases again. Hence, from a practical point of view, we have demonstrated that our proposed SPLasso method can be an efficient method for analysing partially linear models.

6. Conclusion and discussion

In this paper, we propose a SPLasso procedure to screen predictors for ultra-high-dimensional partially linear models, and we further show that SPLasso can identify all relevant predictors with probability tending to one, and that it provides a satisfactory performance in finite samples.

SPLasso selects the next predictor which has the highest correlation with the current residual. It is interesting to point out that LARS proposed by Efron, Hastie, Johnstone, and Tibshirani (2004) also selects the next predictor like SPLasso. However, the current residual of LARS is based on a shrunken estimation of the regression coefficients. The effect of the selected predictors on response variable is not fully used in this estimation. Consequently, this gives a chance for the predictors that have high spurious correlation with predictors in the current model. Simulation studies in Wang (2009) also show that the finite sample performance of LARS is worse than FR. Fitting the response by adding one predictor to the current model, FR selects the next predictor which minimises the residual sum of squares. This amounts to select the next predictor with the largest partial correlation $\{\widetilde{\boldsymbol{x}}_j^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\widetilde{\boldsymbol{x}}_j\}^{-1/2}|\widetilde{\boldsymbol{x}}_j^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\widetilde{\boldsymbol{Y}}|$ with $\mathbf{Q}_{\mathcal{M}_k} = \mathbf{I}_n - \mathbf{H}_{\mathcal{M}_k}$. It is clear that the difference between SPLasso and PFR is the factor $\{\widetilde{\boldsymbol{x}}_j^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\widetilde{\boldsymbol{x}}_j\}^{-1/2}$. As a result, if $\widetilde{\boldsymbol{x}}_j$ has a higher correlation with the predictors in $\mathcal{M}_k$, it will have priority to be selected by PFR. Therefore, we expect that SPLasso will perform better than the profile LARS (or its variants) and the profile FR.

Finally, we note that the errors are assumed to be homogeneous in the current paper. However, as the heterogeneity is often presented in ultra-high-dimensional data, we will investigate the heterogeneity in our future study by combining the quantile regression with the sequential Lasso. Another interesting direction is to extend SPLasso to other semiparametric models including generalised semiparametric models, varying coefficient models and semi-varying coefficient models.

Acknowledgments

The authors thank the editor, the associate editor and two reviewers for their constructive comments that have led to a substantial improvement of the paper.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

Gaorong Li's research was supported in part by the National Natural Science Foundation of China [number 11471029]. Tiejun Tong's research was supported in part by the National Natural Science Foundation of China [number 11671338] and the Hong Kong Baptist University grants [grant number FRG2/15-16/019], [grant number FRG1/16-17/018].

Notes on contributors

Yujie Li is a Ph.D. candidate in College of Applied Sciences at Beijing University of Technology. Her research interests are high-dimensional statistics and nonparametric statistics.

Gaorong Li is Professor of Statistics, Beijing Institute for Scientific and Engineering Computing at Beijing University of Technology. His major research interests are high-dimensional statistics, nonparametric statistics, statistics learning, empirical likelihood, longitudinal data analysis and measurement error models.

Tiejun Tong is Associate Professor of Statistics in the Department of Mathematics at Hong Kong Baptist University. His major research interests are medical statistics and meta-analysis, high-dimensional data analysis, and nonparametric and semiparametric regression.

References

Bertsekas, D. P. (1999). *Nonlinear programming*. Belmont, MA: Athena Scientific.

Bunea, F. (2004). Consistent covariate selection and post model selection inference in semiparametric regression. *The Annals of Statistics*, *32*, 898–927.

Chen, J. H., & Chen, Z. H. (2008). Extended Bayesian information criterion for model selection with large model spaces. *Biometrika*, *95*, 759–771.

Cheng, M. Y., Honda, T., & Zhang, J. T. (2016). Forward variable selection for sparse ultra-high dimensional varying coefficient models. *Journal of the American Statistical Association*, *111*, 1209–1221.

Cui, H. J., Li, R. Z., & Zhong, W. (2015). Model-free feature screening for ultra-high dimensional discriminant analysis. *Journal of the American Statistical Association*, *110*, 630–641.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, *32*, 407–499.

Fan, J. Q., Feng, Y., & Song, R. (2011). Nonparametric independence screening in sparse ultra-high dimensional additive models. *Journal of the American Statistical Association*, *106*, 544–557.

Fan, J. Q., & Gijbels, I. (1996). *Local polynomial modeling and its applications*. London: Chapman and Hall.

Fan, J. Q., & Lv, J. C. (2008). Sure independence screening for ultra-high dimensional feature space (with discussion). *Journal of the Royal Statistical Society, Series B*, *70*, 849–911.

Fan, J. Q., Ma, Y. B., & Dai, W. (2014). Nonparametric independence screening in sparse ultra-high-dimensional varying coefficient models. *Journal of the American Statistical Association*, *109*, 1270–1284.

Härdle, W., Liang, H., & Gao, J. T. (2000). *Partially linear models*. Heidelberg: Springer Physica.

Knight, W. A., Livingston, R. B., Gregory, E. J., & McGuire, W. L. (1977). Estrogen receptor as an independent prognostic factor for early recurrence in breast cancer. *Cancer Research*, *37*, 4669–4671.

Li, Y. J., Li, G. R., Lian, H., & Tong, T. J. (2017). Profile forward regression screening for ultra-high dimensional semiparametric varying coefficient partially linear models. *Journal of Multivariate Analysis*, *155*, 133–150.

Li, G. R., Peng, H., Zhang, J., & Zhu, L. X. (2012). Robust rank correlation based screening. *The Annals of Statistics*, *40*, 1846–1877.

Li, G. R., Zhang, J., & Feng, S. Y. (2016). *Modern measurement error models*. Beijing: Science Press.

Li, R. Z., Zhong, W., & Zhu, L. P. (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association*, *107*, 1129–1139.

Liang, H., & Li, R. Z. (2009). Variable selection for partially linear models with measurement errors. *Journal of the American Statistical Association*, *104*, 234–248.

Liang, H., Wang, H. S., & Tsai, C. L. (2012). Profile forward regression for ultrahigh dimensional variable screening in semiparametric partially linear models. *Statistica Sinica*, *22*, 531–554.

Liu, X., Wang, L., & Liang, H. (2011). Estimation and variable selection for semiparametric additive partial linear models. *Statistica Sinica*, *21*, 1225–1248.

Luo, S. (2012). *Feature selection in high-dimensional studies* (Doctoral dissertation). National University of Singapore, Singapore.

Luo, S., & Chen, Z. H. (2014). Sequential Lasso cum EBIC for feature selection with ultra-high dimensional feature space. *Journal of the American Statistical Association*, *109*, 1229–1240.

Mammen, E., & van de Geer, S. (1997). Penalized quasi-likelihood estimation in partial linear models. *The Annals of Statistics*, *25*, 1014–1035.

Meinshausen, N., & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*, *34*, 1436–1462.

Park, M. Y., & Hastie, T. (2007). L_1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society, Series B*, *69*, 659–677.

Sherwood, W., & Wang, L. (2016). Partially linear additive quantile regression in ultra-high dimension. *The Annals of Statistics*, *44*, 288–317.

Stewart, B. W., & Wild, C. P. (2014). *World cancer report 2014*. Lyon: International Agency for Research on Cancer.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, *58*, 267–288.

van't Veer, L. J., Dai, H. Y., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., … Friens, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, *415*, 530–536.

Wang, H. S. (2009). Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association*, *104*, 1512–1524.

Wang, L., Liu, X., Liang, H., & Carroll, R. J. (2011). Estimation and variable selection for generalized additive partial linear models. *The Annals of Statistics*, *39*, 1827–1851.

Weisberg, S. (1980). *Applied linear regression*. New York: Wiley.

Xie, H. L., & Huang, J. (2009). SCAD-penalized regression in high-dimensional partially linear models. *The Annals of Statistics*, *37*, 673–696.

Yu, T., Li, J. L., & Ma, S. G. (2012). Adjusting confounders in ranking biomarkers: A model-based ROC approach. *Briefings in Bioinformatics*, *13*, 513–523.

Zhang, C. H., & Huang, J. (2008). The sparsity and bias of the Lasso selection in high dimensional linear regression. *The Annals of Statistics*, *37*, 1567–1594.

Zhao, P., & Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research*, *7*, 2541–2563.

Zhu, L. P., Li, L. X., Li, R. Z., & Zhu, L. X. (2011). Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, *106*, 1464–1475.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, *101*, 1418–1429.

Appendices

We introduce the following notation to simplify our presentation. Let $\partial|x|$ be r if $x = 0$, where r is an arbitrary number with $|r| \leq 1$, otherwise be $\mathrm{sgn}(x)$. For any set $\mathcal{M} \subset \mathcal{M}_{\mathcal{F}}$, let $\mathcal{M}^- = \mathcal{M}^c \bigcap \mathcal{M}_{\mathcal{T}}$, $\mathbf{Q}_{\mathcal{M}} = \mathbf{I}_n - \mathbf{H}_{\mathcal{M}}$ and $\psi_n(j, \mathcal{M}, \boldsymbol{\beta}) = \widetilde{\boldsymbol{x}}_j^{\mathrm{T}} \mathbf{Q}_{\mathcal{M}} \widetilde{\mathbf{X}} \boldsymbol{\beta}/n$. For any n-dimensional vectors $\boldsymbol{\nu}_1$ and $\boldsymbol{\nu}_2$, define $\triangle^{\boldsymbol{\nu}_1}(\mathcal{M}) = \boldsymbol{\nu}_1^{\mathrm{T}} \mathbf{Q}_{\mathcal{M}} \boldsymbol{\nu}_1$ and $\triangle^{\boldsymbol{\nu}_1, \boldsymbol{\nu}_2}(\mathcal{M}) = \boldsymbol{\nu}_1^{\mathrm{T}} \mathbf{Q}_{\mathcal{M}} \boldsymbol{\nu}_2$. Furthermore, define $\boldsymbol{\mu} = \widetilde{\mathbf{X}} \boldsymbol{\beta}$, $\widehat{\Sigma} = \widetilde{\mathbf{X}}^{\mathrm{T}} \widetilde{\mathbf{X}}/n$ and $\widehat{\Sigma}_{(\mathcal{M})} = \{\widehat{\Sigma}_{ij} : i, j \in \mathcal{M}\}$ for any candidate model $\mathcal{M}$.

Appendix 1. Some lemmas

Lemma A.1: *Suppose that regularity conditions (C3)–(C6) hold. We have*

$$\max_{1\le i\le n}\Big\{\max\Big(\max_{1\le j\le p}\big|\widehat{\mathrm{E}(X_j|U_i)}-\mathrm{E}(X_j|U_i)\big|,\ \big|\widehat{\mathrm{E}(Y|U_i)}-\mathrm{E}(Y|U_i)\big|\Big)\Big\}=o_P(c_n),\quad(\mathrm{A}.1)$$

where $\widehat{\mathrm{E}(X_j|U_i)}$ *is the estimator of* $E(X_j|U_i)$, $\widehat{\mathrm{E}(Y|U_i)}$ *is the estimator of* $E(Y|U_i)$ *for* $1\le i\le n$ *and* $1\le j\le p$, *and* $c_n=n^{-1/4}\ln^{-1}n$.

Lemma A.2: *Suppose that regularity conditions (C1) and (C3)–(C6) hold, and let* $\tilde m=O(n^{2\xi_0+4\xi_{\min}})$ *with probability tending to one. Then,*

$$\tau_{\min}<\min_{|\mathcal{M}|\le\tilde m}\gamma_{\min}\big(\widehat\Sigma_{(\mathcal{M})}\big)\le\max_{|\mathcal{M}|\le\tilde m}\gamma_{\max}\big(\widehat\Sigma_{(\mathcal{M})}\big)<\tau_{\max}.\quad(\mathrm{A}.2)$$

The proofs of Lemmas A.1 and A.2 can be found in Liang et al. (2012), and hence we omit the details.

Lemma A.3: *Suppose that regularity conditions (C1) and (C3)–(C6) hold. Then, we have, for* $1\le k\le K_n$,

(1) $\tilde{\boldsymbol{x}}_j^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\boldsymbol{\varepsilon}/n=O_P(n^{-1/2}\ln p)$, *uniformly for all* $j\in\mathcal{M}_k^c$.
(2) $\max_{j\in\mathcal{M}_k^-}|\psi_n(j,\mathcal{M}_k,\boldsymbol{\beta})|\ge D_nn^{-1/2}\ln p$ *with* $D_n\to\infty$.

Proof: By Lemma A.2, it is easy to obtain that $\|\tilde{\boldsymbol{x}}_j^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\|\le\|\tilde{\boldsymbol{x}}_j\|\le\sqrt{n\tau_{\max}}$. Hence, we have

$$\Pr\left(\frac1n\left|\tilde{\boldsymbol{x}}_j^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\boldsymbol{\varepsilon}\right|\ge\sigma\tau_{\max}^{1/2}n^{-1/2}\ln p\right)\le2\exp(-(\ln p)^2/2).$$

By the Bonferroni inequality,

$$\Pr\left(\max_{j\in\mathcal{M}_k^c}\frac1n\left|\tilde{\boldsymbol{x}}_j^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\boldsymbol{\varepsilon}\right|\ge\sigma\tau_{\max}^{1/2}n^{-1/2}\ln p\right)\le2\exp(-(\ln p)^2/2+\ln p)\to0.$$

This leads to the result of (1). Next we prove (2). By some simple calculations, we have

$$\boldsymbol{\beta}^{\mathrm{T}}\tilde{\mathbf{X}}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\tilde{\mathbf{X}}\boldsymbol{\beta}=\boldsymbol{\beta}_{\mathcal{M}_k^-}^{\mathrm{T}}\tilde{\mathbf{X}}_{\mathcal{M}_k^-}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\tilde{\mathbf{X}}\boldsymbol{\beta}\le n\|\boldsymbol{\beta}_{\mathcal{M}_k^-}\|_1\times\max_{j\in\mathcal{M}_k^-}|\psi_n(j,\mathcal{M}_k,\boldsymbol{\beta})|\quad(\mathrm{A}.3)$$

and

$$\begin{aligned}\boldsymbol{\beta}^{\mathrm{T}}\tilde{\mathbf{X}}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\tilde{\mathbf{X}}\boldsymbol{\beta}&=\boldsymbol{\beta}_{\mathcal{M}_k^-}^{\mathrm{T}}\tilde{\mathbf{X}}_{\mathcal{M}_k^-}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\tilde{\mathbf{X}}_{\mathcal{M}_k^-}\boldsymbol{\beta}_{\mathcal{M}_k^-}\\&\ge\gamma_{\min}\left(\tilde{\mathbf{X}}_{\mathcal{M}_k^-}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\tilde{\mathbf{X}}_{\mathcal{M}_k^-}\right)\|\boldsymbol{\beta}_{\mathcal{M}_k^-}\|^2\\&\ge\gamma_{\min}\left(\tilde{\mathbf{X}}_{\mathcal{M}_{0k}}^{\mathrm{T}}\tilde{\mathbf{X}}_{\mathcal{M}_{0k}}\right)\|\boldsymbol{\beta}_{\mathcal{M}_k^-}\|_2^2\end{aligned}\quad(\mathrm{A}.4)$$

with $\mathcal{M}_{0k}=\mathcal{M}_k\bigcup\mathcal{M}_{\mathcal{T}}$. Inequality (A.4) follows the fact that $(\tilde{\mathbf{X}}_{\mathcal{M}_k^-}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\tilde{\mathbf{X}}_{\mathcal{M}_k^-})^{-1}$ is a sub-matrix of $(\tilde{\mathbf{X}}_{\mathcal{M}_0}^{\mathrm{T}}\tilde{\mathbf{X}}_{\mathcal{M}_0})^{-1}$ by the inverse of the block matrix. Combining (A.3) and (A.4), we have

$$\max_{j\in\mathcal{M}_k^-}|\psi_n(j,\mathcal{M}_k,\boldsymbol{\beta})|\ge\frac1n\gamma_{\min}\left(\tilde{\mathbf{X}}_{\mathcal{M}_0}^{\mathrm{T}}\tilde{\mathbf{X}}_{\mathcal{M}_0}\right)\frac{\|\boldsymbol{\beta}_{\mathcal{M}_k^-}\|_2^2}{\|\boldsymbol{\beta}_{\mathcal{M}_k^-}\|_1}\doteq D_nn^{-1/2}\ln p,$$

where $D_n=\frac{n^{1/2}}{\ln p}\tau_{\min}\beta_{\min}$. By conditions (C2)–(C3) along with Lemma A.2, we have $D_n\to\infty$. Hence, the proof of (2) is completed. □

Lemma A.4: *Let* $\mathcal{A}_k$ *be the index set of the variables being added at the* $(k+1)$*th step of the SPLasso method. There exists a vector* $\widehat{\boldsymbol{\beta}}_{\mathcal{A}_k}^{(k+1)}$ *with componentwise nonzero elements such that* $|\partial(\widehat{\boldsymbol{\beta}}_j^{(k+1)})|\le1$, *for* $j\in\mathcal{M}_{k+1}^c$, *where*

$$\begin{aligned}\partial(\widehat{\boldsymbol{\beta}}_j^{(k+1)})&=2(\lambda_{k+1}^*)^{-1}\tilde{\boldsymbol{x}}_j^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_{k+1}}\tilde{\mathbf{Y}}\\&\quad+\tilde{\boldsymbol{x}}_j^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\tilde{\mathbf{X}}_{\mathcal{A}_k}\{\tilde{\mathbf{X}}_{\mathcal{A}_k}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\tilde{\mathbf{X}}_{\mathcal{A}_k}\}^{-1}\partial(\widehat{\boldsymbol{\beta}}_{\mathcal{A}_k}^{(k+1)}),\end{aligned}$$

and $\partial(\widehat{\boldsymbol{\beta}}_{\mathcal{A}_k}^{(k+1)})=\partial(\tilde{\mathbf{X}}_{\mathcal{A}_k}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\tilde{\mathbf{Y}})$.

Proof: Differentiating the objective function $\mathcal{L}_{k+1}$ with respect to $\boldsymbol{\beta}_{\mathcal{M}_k}$, we have

$$\frac{\partial\mathcal{L}_{k+1}}{\partial\boldsymbol{\beta}_{\mathcal{M}_k}}=-2\tilde{\mathbf{X}}_{\mathcal{M}_k}^{\mathrm{T}}\tilde{\mathbf{Y}}+2\tilde{\mathbf{X}}_{\mathcal{M}_k}^{\mathrm{T}}\tilde{\mathbf{X}}_{\mathcal{M}_k}\boldsymbol{\beta}_{\mathcal{M}_k}+2\tilde{\mathbf{X}}_{\mathcal{M}_k}^{\mathrm{T}}\tilde{\mathbf{X}}_{\mathcal{M}_k^c}\boldsymbol{\beta}_{\mathcal{M}_k^c}.$$

Let the above derivative equal to zero, we can get the following solution:

$$\widehat{\boldsymbol{\beta}}_{\mathcal{M}_k}=\left\{\tilde{\mathbf{X}}_{\mathcal{M}_k}^{\mathrm{T}}\tilde{\mathbf{X}}_{\mathcal{M}_k}\right\}^{-1}\tilde{\mathbf{X}}_{\mathcal{M}_k}^{\mathrm{T}}\left(\tilde{\mathbf{Y}}-\tilde{\mathbf{X}}_{\mathcal{M}_k^c}^{\mathrm{T}}\tilde{\mathbf{X}}_{\mathcal{M}_k^c}\boldsymbol{\beta}_{\mathcal{M}_k^c}\right).\quad(\mathrm{A}.5)$$

Substituting (A.5) into $\mathcal{L}_{k+1}$, we obtain $\mathcal{L}_{k+1}$ is equivalent to solve the following objective function:

$$\left(\tilde{\mathbf{Y}}-\tilde{\mathbf{X}}_{\mathcal{M}_k^c}\boldsymbol{\beta}_{\mathcal{M}_k^c}\right)^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\left(\tilde{\mathbf{Y}}-\tilde{\mathbf{X}}_{\mathcal{M}_k^c}\boldsymbol{\beta}_{\mathcal{M}_k^c}\right)+\lambda\sum_{j\in\mathcal{M}_k^c}|\beta_j|.\quad(\mathrm{A}.6)$$

By the KKT condition, (A.6) can reach its minimum at $\widehat{\boldsymbol{\beta}}_{\mathcal{M}_k^c}$ if and only if

$$2\tilde{\mathbf{X}}_{\mathcal{M}_k^c}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\left(\tilde{\mathbf{Y}}-\tilde{\mathbf{X}}_{\mathcal{M}_k^c}\widehat{\boldsymbol{\beta}}_{\mathcal{M}_k^c}\right)=\lambda\partial(\widehat{\boldsymbol{\beta}}_{\mathcal{M}_k^c}).\quad(\mathrm{A}.7)$$

Noting that $\partial(\widehat{\boldsymbol{\beta}}_{\mathcal{M}_k^c})^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{\mathcal{M}_k^c}=\|\widehat{\boldsymbol{\beta}}_{\mathcal{M}_k^c}\|_1$, we have

$$\lambda=\frac{2\left(\tilde{\mathbf{Y}}-\tilde{\mathbf{X}}_{\mathcal{M}_k^c}\widehat{\boldsymbol{\beta}}_{\mathcal{M}_k^c}\right)^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\tilde{\mathbf{X}}_{\mathcal{M}_k^c}\widehat{\boldsymbol{\beta}}_{\mathcal{M}_k^c}}{\|\widehat{\boldsymbol{\beta}}_{\mathcal{M}_k^c}\|_1}.\quad(\mathrm{A}.8)$$

Plugging (A.8) into (A.7), we have

$$\begin{aligned}&\tilde{\mathbf{X}}_{\mathcal{M}_k^c}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\left(\tilde{\mathbf{Y}}-\tilde{\mathbf{X}}_{\mathcal{M}_k^c}\widehat{\boldsymbol{\beta}}_{\mathcal{M}_k^c}\right)\\&=\frac{\left(\tilde{\mathbf{Y}}-\tilde{\mathbf{X}}_{\mathcal{M}_k^c}\widehat{\boldsymbol{\beta}}_{\mathcal{M}_k^c}\right)^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\tilde{\mathbf{X}}_{\mathcal{M}_k^c}\widehat{\boldsymbol{\beta}}_{\mathcal{M}_k^c}}{\|\widehat{\boldsymbol{\beta}}_{\mathcal{M}_k^c}\|_1}\partial(\widehat{\boldsymbol{\beta}}_{\mathcal{M}_k^c}).\end{aligned}\quad(\mathrm{A}.9)$$

As $\mathcal{A}_k$ is the set of variables being added at the $(k+1)$th step, (A.7) holds for some $\widehat{\boldsymbol{\beta}}_{\mathcal{M}_k^c}$ which satisfies $\widehat{\boldsymbol{\beta}}_{\mathcal{M}_k^c}=(0,\ldots,0,\widehat{\boldsymbol{\beta}}_{\mathcal{A}_k}^{(k+1)},0,\ldots,0)^{\mathrm{T}}$ and $|\partial(\widehat{\boldsymbol{\beta}}_j^{(k+1)})|\le1$ for any $j\in\mathcal{M}_{k+1}^c$. Let λ_{k+1}^* be the corresponding λ in

(A.8) for this estimator $\widehat{\boldsymbol{\beta}}_{\mathcal{M}_k^c}$. Noting that this particular $\widehat{\boldsymbol{\beta}}_{\mathcal{M}_k^c}$ also satisfies (A.7) and (A.9), we have

$$\widehat{\boldsymbol{\beta}}_{\mathcal{A}_k}^{(k+1)} = \{\widetilde{\mathbf{X}}_{\mathcal{A}_k}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\widetilde{\mathbf{X}}_{\mathcal{A}_k}\}^{-1} \times \left\{\widetilde{\mathbf{X}}_{\mathcal{A}_k}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\widetilde{\boldsymbol{Y}} - \frac{\lambda_{k+1}^*}{2}\partial(\widehat{\boldsymbol{\beta}}_{\mathcal{A}_k}^{(k+1)})\right\}, \quad \text{(A.10)}$$

$$\partial(\widehat{\boldsymbol{\beta}}_j^{(k+1)}) = \frac{2}{\lambda_{k+1}^*}\widetilde{\boldsymbol{x}}_j^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\left(\widetilde{\boldsymbol{Y}} - \widetilde{\mathbf{X}}_{\mathcal{A}_k}\widehat{\boldsymbol{\beta}}_{\mathcal{A}_k}^{(k+1)}\right), \quad \forall j \in \mathcal{M}_{k+1}^c. \quad \text{(A.11)}$$

By (A.10) and the definition of $\partial(\cdot)$, it is easy to see that $\partial(\widehat{\boldsymbol{\beta}}_{\mathcal{A}_k}^{(k+1)}) = \partial(\widetilde{\mathbf{X}}_{\mathcal{A}_k}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\widetilde{\boldsymbol{Y}})$. Furthermore, plugging (A.10) into (A.11), it can be shown that $\partial(\widehat{\boldsymbol{\beta}}_j^{(k+1)})$ is equal to

$$-2\frac{\widetilde{\boldsymbol{x}}_j^{\mathrm{T}}\mathbf{P}^{k+1}\widetilde{\boldsymbol{Y}}}{\lambda_{k+1}^*} + \widetilde{\boldsymbol{x}}_j^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\widetilde{\mathbf{X}}_{\mathcal{A}_k}\{\widetilde{\mathbf{X}}_{\mathcal{A}_k}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\widetilde{\mathbf{X}}_{\mathcal{A}_k}\}^{-1}\partial(\widehat{\boldsymbol{\beta}}_{\mathcal{A}_k}^{(k+1)}), \quad \text{(A.12)}$$

where $\mathbf{P}^{k+1} = \mathbf{Q}_{\mathcal{M}_k}\widetilde{\mathbf{X}}_{\mathcal{A}_k}\{\widetilde{\mathbf{X}}_{\mathcal{A}_k}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\widetilde{\mathbf{X}}_{\mathcal{A}_k}\}^{-1}\widetilde{\mathbf{X}}_{\mathcal{A}_k}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k} - \mathbf{I}_n$. By some algebraic operations, we have $\mathbf{P}^{k+1} = -\mathbf{Q}_{\mathcal{M}_{k+1}}$. Plugging this result into (A.12), the proof of Lemma A.4 is completed. □

Lemma A.5: *Suppose that regularity conditions (C1)–(C6) hold. If $\mathcal{M}_{k+1}^- \neq \emptyset$, for $1 \le k \le K_n - 1$, we have*

$$\frac{\triangle^{\boldsymbol{\mu}}(\mathcal{M}_k) - \triangle^{\boldsymbol{\mu}}(\mathcal{M}_{k+1})}{\ln n} \to \infty.$$

Proof: We prove this conclusion by contradiction. Assume that there exists k such that $\triangle^{\boldsymbol{\mu}}(\mathcal{M}_k) - \triangle^{\boldsymbol{\mu}}(\mathcal{M}_{k+1}) = O(\ln n)$. Let $a_{k+1} = \arg\max_{m\in\mathcal{M}_{k+1}^-}|\psi_n(m, \mathcal{M}_{k+1}, \boldsymbol{\beta})|$. If a_{k+1} is not unique, we choose one of them. Note that a_{k+1} may be the index of the predictor being selected at the $(k+2)$th step. Let the two terms of $\partial(\hat{\beta}_{a_{k+1}}^{(k+1)})$ in Lemma A.4 as I_1 and I_2, respectively. By Lemma A.3, we have

$$\frac{\sqrt{n}|\psi_n(a_{k+1}, \mathcal{M}_{k+1}, \boldsymbol{\beta})|}{\ln p} \ge D_n \to \infty. \quad \text{(A.13)}$$

By Lemma A.1, it is easy to see that, with probability tending to one,

$$\begin{aligned}\mathrm{RSS}(\mathcal{M}_k) - \mathrm{RSS}(\mathcal{M}_{k+1}) &= \{\triangle^{\boldsymbol{\mu}}(\mathcal{M}_k) - \triangle^{\boldsymbol{\mu}}(\mathcal{M}_{k+1})\} \\ &\quad +2\{\triangle^{\boldsymbol{\mu},\boldsymbol{\varepsilon}}(\mathcal{M}_k) - \triangle^{\boldsymbol{\mu},\boldsymbol{\varepsilon}}(\mathcal{M}_{k+1})\} \\ &\quad +\{\triangle^{\boldsymbol{\varepsilon}}(\mathcal{M}_k) - \triangle^{\boldsymbol{\varepsilon}}(\mathcal{M}_{k+1})\}. \quad \text{(A.14)}\end{aligned}$$

For the last two terms of (A.14), if $\mathcal{M}_{k+1}^- \neq \emptyset$, we have, for any $t > 0$,

$$\begin{aligned}&\Pr\left(\max_{1\le k\le K_n-1}(\triangle^{\boldsymbol{\varepsilon}}(\mathcal{M}_k) - \triangle^{\boldsymbol{\varepsilon}}(\mathcal{M}_{k+1})) \ge t\right) \\ &\le K_n \Pr(\chi^2(1) \ge t) \\ &\le \frac{c_1}{\sqrt{t}}\exp(-t/2 + c_2\ln n)\end{aligned}$$

with some positive constants c_1 and c_2. Letting $t = 4\ln n$, the right-hand side of the above inequality converges to 0 as $n \to \infty$. Since $\frac{\triangle^{\boldsymbol{\mu},\boldsymbol{\varepsilon}}(\mathcal{M}_k) - \triangle^{\boldsymbol{\mu},\boldsymbol{\varepsilon}}(\mathcal{M}_{k+1})}{\sqrt{\triangle^{\boldsymbol{\mu}}(\mathcal{M}_k) - \triangle^{\boldsymbol{\mu}}(\mathcal{M}_{k+1})}}$ follows the standard normal distribution, we have

$$\begin{aligned}&\max_{1\le k\le K_n-1}\frac{\triangle^{\boldsymbol{\mu},\boldsymbol{\varepsilon}}(\mathcal{M}_k) - \triangle^{\boldsymbol{\mu},\boldsymbol{\varepsilon}}(\mathcal{M}_{k+1})}{\triangle^{\boldsymbol{\mu}}(\mathcal{M}_k) - \triangle^{\boldsymbol{\mu}}(\mathcal{M}_{k+1})} \\ &= \sqrt{\frac{O_P(\ln n)}{\triangle^{\boldsymbol{\mu}}(\mathcal{M}_k) - \triangle^{\boldsymbol{\mu}}(\mathcal{M}_{k+1})}} = O_P(1).\end{aligned}$$

This leads to $\mathrm{RSS}(\mathcal{M}_k) - \mathrm{RSS}(\mathcal{M}_{k+1}) = O_P(\ln n)$. By the fact that

$$\begin{aligned}&\mathrm{RSS}(\mathcal{M}_k) - \mathrm{RSS}(\mathcal{M}_{k+1}) \\ &\ge \|\widetilde{\boldsymbol{Y}}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\widetilde{\mathbf{X}}_{\mathcal{A}_k}\|_2^2\gamma_{\max}^{-1}\left(\widetilde{\mathbf{X}}_{\mathcal{A}_k}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\widetilde{\mathbf{X}}_{\mathcal{A}_k}\right),\end{aligned}$$

we have

$$\begin{aligned}\|\widetilde{\boldsymbol{Y}}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\widetilde{\mathbf{X}}_{\mathcal{A}_k}\|_2 &= \gamma_{\max}\left(\widetilde{\mathbf{X}}_{\mathcal{A}_k}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\widetilde{\mathbf{X}}_{\mathcal{A}_k}\right)O_P(\ln n) \\ &= O_P(n\ln n).\end{aligned}$$

On the other hand, multiplying $\partial(\widehat{\boldsymbol{\beta}}_{\mathcal{A}_k}^{(k+1)})$ on both sides of (A.10), by the last equation of Lemma A.4 along with the positivity of the left-hand side, we have

$$\begin{aligned}\lambda_{k+1}^* &< \frac{2\widetilde{\boldsymbol{Y}}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\widetilde{\mathbf{X}}_{\mathcal{A}_k}\left\{\widetilde{\mathbf{X}}_{\mathcal{A}_k}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\widetilde{\mathbf{X}}_{\mathcal{A}_k}\right\}^{-1}\widetilde{\mathbf{X}}_{\mathcal{A}_k}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\widetilde{\boldsymbol{Y}}}{\partial(\widehat{\boldsymbol{\beta}}_{\mathcal{A}_k}^{(k+1)})^{\mathrm{T}}\left\{\widetilde{\mathbf{X}}_{\mathcal{A}_k}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\widetilde{\mathbf{X}}_{\mathcal{A}_k}\right\}^{-1}\partial(\widehat{\boldsymbol{\beta}}_{\mathcal{A}_k}^{(k+1)})\|\widetilde{\mathbf{X}}_{\mathcal{A}_k}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\widetilde{\boldsymbol{Y}}\|_1} \\ &= 2\|\widetilde{\mathbf{X}}_{\mathcal{A}_k}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\widetilde{\boldsymbol{Y}}\|_1. \quad \text{(A.15)}\end{aligned}$$

Noting that $\|\widetilde{\mathbf{X}}_{\mathcal{A}_k}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\widetilde{\boldsymbol{Y}}\|_1$ and $\|\widetilde{\mathbf{X}}_{\mathcal{A}_k}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\widetilde{\boldsymbol{Y}}\|_2$ have the same order, by (A.13)–(A.15), we have

$$|I_1| \ge \frac{n|\psi_n(a_{k+1}, \mathcal{M}_{k+1}, \boldsymbol{\beta})|}{\|\widetilde{\mathbf{X}}_{\mathcal{A}_k}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\widetilde{\boldsymbol{Y}}\|_1} \to \infty. \quad \text{(A.16)}$$

Note that

$$\begin{aligned}&\widetilde{\boldsymbol{x}}_{a_{k+1}}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\widetilde{\mathbf{X}}_{\mathcal{A}_k}\left\{\widetilde{\mathbf{X}}_{\mathcal{A}_k}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\widetilde{\mathbf{X}}_{\mathcal{A}_k}\right\}^{-1}\widetilde{\mathbf{X}}_{\mathcal{A}_k}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\widetilde{\boldsymbol{x}}_{a_{k+1}} \\ &\ge \|\widetilde{\boldsymbol{x}}_{a_{k+1}}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\widetilde{\mathbf{X}}_{\mathcal{A}_k}\|_2^2\gamma_{\max}^{-1}\left(\widetilde{\mathbf{X}}_{\mathcal{A}_k}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\widetilde{\mathbf{X}}_{\mathcal{A}_k}\right) \\ &\ge (n\tau_{\max})^{-1}\|\widetilde{\boldsymbol{x}}_{a_{k+1}}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\widetilde{\mathbf{X}}_{\mathcal{A}_k}\|_2^2. \quad \text{(A.17)}\end{aligned}$$

Hence, $\|\widetilde{\boldsymbol{x}}_{a_{k+1}}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\widetilde{\mathbf{X}}_{\mathcal{A}_k}\|_2 = O(n)$. By the above results, we have

$$\begin{aligned}|I_2| &\le \|\widetilde{\boldsymbol{x}}_{a_{k+1}}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\widetilde{\mathbf{X}}_{\mathcal{A}_k}\left\{\widetilde{\mathbf{X}}_{\mathcal{A}_k}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\widetilde{\mathbf{X}}_{\mathcal{A}_k}\right\}^{-1}\|_1 \\ &\le \|\widetilde{\boldsymbol{x}}_{a_{k+1}}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\widetilde{\mathbf{X}}_{\mathcal{A}_k}\|_1\|\left\{\widetilde{\mathbf{X}}_{\mathcal{A}_k}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\widetilde{\mathbf{X}}_{\mathcal{A}_k}\right\}^{-1}\|_1 \\ &\le |\mathcal{A}_k|\|\widetilde{\boldsymbol{x}}_{a_{k+1}}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\widetilde{\mathbf{X}}_{\mathcal{A}_k}\|_2\|\left\{\widetilde{\mathbf{X}}_{\mathcal{A}_k}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\widetilde{\mathbf{X}}_{\mathcal{A}_k}\right\}^{-1}\|_2 \\ &\le Cn^{-1}\|\widetilde{\boldsymbol{x}}_{a_{k+1}}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\widetilde{\mathbf{X}}_{\mathcal{A}_k}\|_2 = O(1).\end{aligned}$$

Therefore, we have $|\widehat{\beta}_{a_{k+1}}^{(k+1)}| = |I_1 + I_2| = \infty$, which contradicts with Lemma A.4, that is, the assumption is false. This completes the proof of Lemma A.5. □

Appendix 2. Proof of Theorem 3.1

To prove Theorem 3.1, we consider what happens if $\mathcal{M}_{K_n}^- \neq \emptyset$, that is, there still exist relevant variables, which are not identified after K_n steps. For simplicity, we assume that the variables enter the model one by one. We first focus on the two terms in $\partial(\widehat{\beta}_{a_{k+1}}^{(k+1)})$ as in

Lemma A.5. Note that

$$\begin{aligned}&\mathrm{RSS}(\mathcal{M}_k)-\mathrm{RSS}(\mathcal{M}_{k+1})\\&=\widetilde{\boldsymbol{Y}}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\widetilde{\mathbf{X}}_{\mathcal{A}_k}\left\{\widetilde{\mathbf{X}}_{\mathcal{A}_k}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\widetilde{\mathbf{X}}_{\mathcal{A}_k}\right\}^{-1}\widetilde{\mathbf{X}}_{\mathcal{A}_k}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\widetilde{\boldsymbol{Y}}\\&\geq\|\widetilde{\boldsymbol{Y}}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\widetilde{\mathbf{X}}_{\mathcal{A}_k}\|_2^2\gamma_{\max}^{-1}\left(\widetilde{\mathbf{X}}_{\mathcal{A}_k}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\widetilde{\mathbf{X}}_{\mathcal{A}_k}\right)\\&\geq(n\tau_{\max})^{-1}\|\widetilde{\boldsymbol{Y}}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\widetilde{\mathbf{X}}_{\mathcal{A}_k}\|_2^2.\end{aligned}\tag{A.18}$$

On the other hand, by Lemma A.5, we have

$$\begin{aligned}\mathrm{RSS}(\mathcal{M}_k)-\mathrm{RSS}(\mathcal{M}_{k+1})=&[\triangle^{\boldsymbol{\mu}}(\mathcal{M}_k)\\&-\triangle^{\boldsymbol{\mu}}(\mathcal{M}_{k+1})](1+o_P(1)).\end{aligned}\tag{A.19}$$

Therefore, by (A.3), (A.18), (A.19), with probability tending to one,

$$\begin{aligned}|I_1|&\geq\frac{n|\psi_n(a_{k+1},\mathcal{M}_{k+1},\boldsymbol{\beta})|}{\|\widetilde{\boldsymbol{Y}}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\widetilde{\mathbf{X}}_{\mathcal{A}_k}\|_1}\\&\geq\frac{n|\psi_n(a_{k+1},\mathcal{M}_{k+1},\boldsymbol{\beta})|}{\sqrt{n\tau_{\max}(\triangle^{\boldsymbol{\mu}}(\mathcal{M}_k)-\triangle^{\boldsymbol{\mu}}(\mathcal{M}_{k+1}))}}\\&>\frac{\triangle^{\boldsymbol{\mu}}(\mathcal{M}_{k+1})}{\|\boldsymbol{\beta}_{\mathcal{M}_{k+1}^-}\|_1\sqrt{n\tau_{\max}(\triangle^{\boldsymbol{\mu}}(\mathcal{M}_k)-\triangle^{\boldsymbol{\mu}}(\mathcal{M}_{k+1}))}}\triangleq t_{k,n}.\end{aligned}\tag{A.20}$$

By some simple calculations, we have

$$\begin{aligned}&\|\widetilde{\boldsymbol{Y}}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\widetilde{\mathbf{X}}_{\mathcal{A}_k}\|_1\|\widetilde{\boldsymbol{x}}_{a_{k+1}}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\widetilde{\mathbf{X}}_{\mathcal{A}_k}\left\{\widetilde{\mathbf{X}}_{\mathcal{A}_k}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\widetilde{\mathbf{X}}_{\mathcal{A}_k}\right\}^{-1}\|_1\\&\leq\|\widetilde{\boldsymbol{Y}}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\widetilde{\mathbf{X}}_{\mathcal{A}_k}\|_2\|\widetilde{\boldsymbol{x}}_{a_{k+1}}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\widetilde{\mathbf{X}}_{\mathcal{A}_k}\left\{\widetilde{\mathbf{X}}_{\mathcal{A}_k}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\widetilde{\mathbf{X}}_{\mathcal{A}_k}\right\}^{-1}\|_2\\&\leq\|\widetilde{\boldsymbol{Y}}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\widetilde{\mathbf{X}}_{\mathcal{A}_k}\|_2\|\widetilde{\boldsymbol{x}}_{a_{k+1}}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\widetilde{\mathbf{X}}_{\mathcal{A}_k}\|_2\gamma_{\min}^{-1}\left(\widetilde{\mathbf{X}}_{\mathcal{A}_k}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\widetilde{\mathbf{X}}_{\mathcal{A}_k}\right)\\&\leq\sqrt{n\tau_{\max}(\triangle^{\boldsymbol{\mu}}(\mathcal{M}_k)-\triangle^{\boldsymbol{\mu}}(\mathcal{M}_{k+1}))}\frac{\gamma_{\max}\left(\widetilde{\mathbf{X}}_{\mathcal{A}_k}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\widetilde{\mathbf{X}}_{\mathcal{A}_k}\right)}{\gamma_{\min}\left(\widetilde{\mathbf{X}}_{\mathcal{A}_k}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\widetilde{\mathbf{X}}_{\mathcal{A}_k}\right)},\end{aligned}\tag{A.21}$$

where the last inequality is obtained by (A.17) and (A.18). Let

$$\lambda_0\geq\lambda_{0,k}=\gamma_{\max}\left(\widetilde{\mathbf{X}}_{\mathcal{A}_k}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\widetilde{\mathbf{X}}_{\mathcal{A}_k}\right)\gamma_{\min}^{-1}\left(\widetilde{\mathbf{X}}_{\mathcal{A}_k}^{\mathrm{T}}\mathbf{Q}_{\mathcal{M}_k}\widetilde{\mathbf{X}}_{\mathcal{A}_k}\right)\geq 1.$$

By (A.20) and (A.21), we have

$$\frac{|I_1|}{|I_2|}\geq\frac{n|\psi_n(a_{k+1},\mathcal{M}_{k+1},\boldsymbol{\beta})|}{\lambda_0\sqrt{n\tau_{\max}(\triangle^{\boldsymbol{\mu}}(\mathcal{M}_k)-\triangle^{\boldsymbol{\mu}}(\mathcal{M}_{k+1}))}}>\frac{t_{k,n}}{\lambda_0}.$$

If $t_{k,n}\to\infty$, we have $|\widehat{\beta}_{a_{k+1}}^{(k+1)}|\to\infty$, which contradicts with Lemma A.4. This means that $t_{k,n}<\infty$. Then, we have $|I_2|\leq\frac{\lambda_0}{t_{k,n}}|I_1|$. Together with Lemma A.4, we have

$$\begin{aligned}1\geq|\partial(\widehat{\beta}_{a_{k+1}}^{(k+1)})|\geq|I_1|-|I_2|&\geq\left(1-\frac{\lambda_0}{t_{k,n}}\right)|I_1|\\&\geq t_{k,n}-\lambda_0.\end{aligned}$$

By this result, we have $0<t_{k,n}<\lambda_0+1$. Using the definition of $t_{k,n}$ in (A.20), we have

$$\triangle^{\boldsymbol{\mu}}(\mathcal{M}_k)-\triangle^{\boldsymbol{\mu}}(\mathcal{M}_{k+1})\geq\frac{\triangle^{\boldsymbol{\mu}}(\mathcal{M}_{k+1})^2}{n\tau_{\max}(1+\lambda_0)^2\|\boldsymbol{\beta}_{\mathcal{M}_{k+1}^-}\|_1^2}.\tag{A.22}$$

If $\mathcal{M}_k^-\neq\emptyset$ for $0\leq k\leq K_n-1$, by the Cauchy–Schwarz inequality and (A.4), we can further get

$$\begin{aligned}&\frac{\triangle^{\boldsymbol{\mu}}(\mathcal{M}_{k+1})^2}{n\tau_{\max}(1+\lambda_0)^2\|\boldsymbol{\beta}(\mathcal{M}_{k+1}^-)\|_1^2}\\&\geq\frac{n^2\tau_{\min}^2\|\boldsymbol{\beta}_{\mathcal{M}_{k+1}^-}\|_2^4}{n\tau_{\max}(1+\lambda_0)^2\|\boldsymbol{\beta}_{\mathcal{M}_{k+1}^-}\|_2^2|\mathcal{M}_{k+1}^-|}\\&\geq\frac{n^2\tau_{\min}^2\|\boldsymbol{\beta}_{\mathcal{M}_{k+1}^-}\|_1^2}{n\tau_{\max}(1+\lambda_0)^2|\mathcal{M}_{k+1}^-|^2}\geq\frac{n\tau_{\min}^2\beta_{\min}^2}{\tau_{\max}(1+\lambda_0)^2}.\end{aligned}$$

Then,

$$\begin{aligned}\triangle^{\boldsymbol{\mu}}(\mathcal{M}_0)-\triangle^{\boldsymbol{\mu}}(\mathcal{M}_{K_n})&=\sum_{k=0}^{K_n-1}\left(\triangle^{\boldsymbol{\mu}}(\mathcal{M}_k)-\triangle^{\boldsymbol{\mu}}(\mathcal{M}_{k+1})\right)\\&\geq\frac{nK_n\beta_{\min}^2\tau_{\min}^2}{\tau_{\max}(1+\lambda_0)^2},\end{aligned}\tag{A.23}$$

where $\mathcal{M}_0=\emptyset$. On the other hand, we have

$$\begin{aligned}\triangle^{\boldsymbol{\mu}}(\mathcal{M}_0)-\triangle^{\boldsymbol{\mu}}(\mathcal{M}_{K_n})&\leq\triangle^{\boldsymbol{\mu}}(\mathcal{M}_0)\leq n\tau_{\max}\|\boldsymbol{\beta}\|_2^2\\&\leq n\tau_{\max}p_0C_{\boldsymbol{\beta}}^2.\end{aligned}\tag{A.24}$$

Under regular conditions (C1)–(C3), and by the definition of K_n, we have

$$\frac{nK_n\tau_{\min}^2\beta_{\min}^2}{\tau_{\max}(1+\lambda_0)^2}\geq\tau_{\max}np_0C_{\boldsymbol{\beta}}^2.$$

It contradicts with the results (A.23) and (A.24), and this means that $\mathcal{M}_{\mathcal{T}}\subset\mathcal{M}_{K_n}$ with probability tending to one. Hence, the proof of Theorem 3.1 is completed.

Appendix 3. Proof of Theorem 3.2

Define $k_{\min}=\min_{1\leq k\leq n}\{k:\mathcal{M}_{\mathcal{T}}\subset\mathcal{M}_k\}$. By Theorem 3.1, $k_{\min}$ is well defined and satisfies $k_{\min}\leq\mathcal{M}_{K_n}$. For any $1\leq k\leq k_{\min}$, $\mathcal{M}_k$ are underfitted models such that $\mathcal{M}_{\mathcal{T}}\not\subset\mathcal{M}_k$ and $\mathcal{M}_k$ are nested. Therefore, if we can prove $\Pr(\widehat{k}<k_{\min})\to 0$, the conclusion (10) will follow. According to Theorem 3.1, with probability tending to one,

$$\begin{aligned}&\mathrm{EBIC}(\mathcal{M}_k)-\mathrm{EBIC}(\mathcal{M}_{k+1})\\&=\ln\left(\frac{\widehat{\sigma}_{(\mathcal{M}_k)}^2}{\widehat{\sigma}_{(\mathcal{M}_{k+1})}^2}\right)-n^{-1}(\ln n+2\zeta\ln p)\\&\geq\ln\left(1+\frac{\widehat{\sigma}_{(\mathcal{M}_k)}^2-\widehat{\sigma}_{(\mathcal{M}_{k+1})}^2}{\widehat{\sigma}_{(\mathcal{M}_{k+1})}^2}\right)-n^{-1}(1+2\zeta)\ln p\\&\geq\ln\left(1+\frac{\triangle^{\boldsymbol{\mu}}(\mathcal{M}_k)-\triangle^{\boldsymbol{\mu}}(\mathcal{M}_{k+1})}{4\max\{\triangle^{\boldsymbol{\mu}}(\mathcal{M}_{k+1}),\triangle^{\boldsymbol{\varepsilon}}(\mathcal{M}_{k+1})\}}\right)\\&\quad-n^{-1}(1+2\zeta)\ln p.\end{aligned}\tag{A.25}$$

Next, we study (A.25) under the following two cases. First, if $\max\{\triangle^{\boldsymbol{\mu}}(\mathcal{M}_{k+1}),\triangle^{\boldsymbol{\varepsilon}}(\mathcal{M}_{k+1})\}=\triangle^{\boldsymbol{\mu}}(\mathcal{M}_{k+1})$, then by (A.4) and (A.22), with probability tending to one, the right-hand side of (A.25) is bounded below by

$$\ln\left(1+\frac{n\tau_{\min}\beta_{\min}}{4(1+\lambda_0)^2n\tau_{\max}}\right)-n^{-1}(1+2\zeta)\ln p.\tag{A.26}$$

According to the inequality $\ln(1+x) \geq \min(\ln 2, x/2)$ and Lemma A.2, the right-hand side of (A.26) is bounded below by, with probability tending to one,

$$\min\left\{\ln 2, \frac{\tau_{\min}\nu_{\beta}n^{-\xi_{\min}}}{8(1+\lambda_0)^2\tau_{\max}}\right\} - n^{-1}(1+2\zeta)\nu n^{\xi}. \tag{A.27}$$

Under condition (C3), the right-hand side of (A.27) is positive with probability tending to one uniformly for $k \leq k_{\min}$.

Second, if $\max\{\triangle^{\mu}(\mathcal{M}_{k+1}), \triangle^{\epsilon}(\mathcal{M}_{k+1})\} = \triangle^{\epsilon}(\mathcal{M}_{k+1})$, by the fact $\triangle^{\epsilon}(\mathcal{M}_{k+1}) = n(1+o_P(1))$ along with (A.4) and (A.22), with probability tending to one, the right-hand side of (A.25) is bounded below by

$$\ln\left(1+\frac{n^2\tau_{\min}^2\beta_{\min}^3}{4(1+\lambda_0)^2n^2\tau_{\max}}\right) - n^{-1}(1+2\zeta)\ln p. \tag{A.28}$$

By Lemma A.2, with probability tending to one, the right-hand side of (A.26) is further bounded below by

$$\min\left\{\ln 2, \frac{\tau_{\min}^2\nu_{\beta}^3n^{-3\xi_{\min}}}{8(1+\lambda_0)^2)}\right\} - n^{-1}(1+2\zeta)\nu n^{\xi}. \tag{A.29}$$

Under condition (C3), the right-hand side of (A.25) is positive with probability tending to one uniformly for $k \leq k_{\min}$. This finishes the proof of Theorem 3.2.