Journal of Multivariate Analysis 183 (2021) 104733

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Variable selection for partially linear models via Bayesian subset modeling with diffusing prior

Jia Wang [a,1], Xizhen Cai [b,1], Runze Li [a,1,*]

[a] Department of Statistics, Pennsylvania State University, University Park, PA 16802, USA
[b] Department of Mathematics and Statistics, Williams College, Williamstown, MA 01267, USA

ARTICLE INFO

Article history:
Received 4 March 2020
Received in revised form 31 January 2021
Accepted 31 January 2021
Available online 13 February 2021

AMS 2010 subject classifications:
primary 62G08
secondary 62J05

Keywords:
Bayesian variable selection
Difference-based method
Selection consistency
Semiparametric modeling

ABSTRACT

Most existing methods of variable selection in partially linear models (PLM) with ultrahigh dimensional covariates are based on partial residuals, which involve a two-step estimation procedure. While the estimation error produced in the first step may have an impact on the second step, multicollinearity among predictors adds additional challenges in the model selection procedure. In this paper, we propose a new Bayesian variable selection approach for PLM. This new proposal addresses those two issues simultaneously as (1) it is a one-step method which selects variables in PLM, even when the dimension of covariates increases at an exponential rate with the sample size, and (2) the method retains model selection consistency, and outperforms existing ones in the setting of highly correlated predictors. Distinguished from existing ones, our proposed procedure employs the difference-based method to reduce the impact from the estimation of the nonparametric component, and incorporates Bayesian subset modeling with diffusing prior (BSM-DP) to shrink the corresponding estimator in the linear component. The estimation is implemented by Gibbs sampling, and we prove that the posterior probability of the true model being selected converges to one asymptotically. Simulation studies support the theory and the efficiency of our methods as compared to other existing ones, followed by an application in a study of supermarket data.

1. Introduction

Semiparametric model attracts considerable attention in the literature since it retains the interpretability of the parametric models and keeps some flexibility of the nonparametric models. In this paper, we study a type of commonly used semiparametric model, the partially linear model (PLM). The PLM assumes that, the response Y depends both linearly on some covariates $\mathbf{X} \in \mathbb{R}^p$ of interest, and nonparametrically on another univariate continuous covariate U defined on $[0, 1]$. Suppose that the observed data $\{(Y_i, \mathbf{X}_i, U_i)\}, i \in \{1, \ldots, n\}$, is a random sample from the following PLM

$$Y = f(U) + \mathbf{X}^\top \boldsymbol{\beta} + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2). \tag{1}$$

This PLM specifies a parsimonious linear function in the parametric part, while allowing a nonparametric component to be unconstrained and subject to empirical estimation. In this paper, a new one-step Bayesian approach is proposed to select variables for PLM with ultrahigh dimensional covariate $\mathbf{X}$, that is $\ln p = o(n)$. Specifically, our proposed method simplifies

* Corresponding author.
E-mail addresses: jzw88@psu.edu (J. Wang), xc2@williams.edu (X. Cai), rzli@psu.edu (R. Li).

[1] All authors have equally contributed to this work, and are listed in the order of seniority.

https://doi.org/10.1016/j.jmva.2021.104733
0047-259X/

the procedure by avoiding estimating the infinite dimensionality brought by the nonparametric component and results in sparsity in the linear component.

The estimation procedure for PLM with fixed dimension p of $\mathbf{X}$ has been extensively studied. Engle et al. used the penalized least squares method to estimate $\boldsymbol{\beta}$ and the nuisance function $f(\cdot)$ simultaneously by adding a penalty on the roughness of $f(\cdot)$, which was referred as the partial smoothing splines [6,7,12,23]. Since $\boldsymbol{\beta}$ is of primary interest, some other methods brilliantly avoid the estimation of $f(\cdot)$. For example, Robinson [24] introduced a profile least squares estimator based on the idea of partial residual, which later became one of the commonly used approaches to eliminate the nonparametric component in PLM. Another type of approach to eliminate the nonparametric component is the difference-based method [27,29]. It estimates the coefficients in linear component by taking differences of the ordered observations. The resulting estimator is proven to be asymptotically efficient under finite dimensions. See Section 2.1 for more details about the difference-based method.

Variable selection for PLM can be accomplished by adding another penalty function on $\boldsymbol{\beta}$ to the loss function of the aforementioned partial smoothing splines method. The least absolute shrinkage and selection operator (LASSO) [25], the nonnegative garrote [2,31], the smoothly clipped absolute deviation (SCAD) [8], the elastic net [34], and the minimum concave penalty (MCP) [32] are all among popular choices of penalty functions. Xie and Huang [28] used the SCAD penalty to achieve the sparsity in the linear part and used the polynomial splines to estimate the nonparametric component simultaneously. The resulting estimator $\widehat{\boldsymbol{\beta}}$ was shown to be consistent with $p = o(\sqrt{n})$. An alternative is to use a two-step procedure, in which Y and $\mathbf{X}$ are first regressed on U separately to get partial residuals, then the variable selection is further applied on the transformed model. For example, the consistency for both the linear and the nonparametric components has been well studied by Zhu et al. [33] under the regime of $p = o(\sqrt{n})$. Liang and Li [18] discussed this approach in the presence of measurement errors. More recently, Liu et al. [20] proposed a selection procedure via recursively test on the partial correlation among the partial residuals and among the covariates when $\ln p = o(n)$. The method was referred as the thresholded partial correlation on partial residuals (TPC-PR). However, to the best of our knowledge, there is nearly no literature on variable selection in the high-dimensional setting based on the extension of difference-based method.

Bayesian approach puts priors on the parameters and the model space, and selects the model with the highest posterior probability. There have been multiple developments for variable selection using Bayesian approach with linear and generalized linear models. George and McCulloch [11] proposed a milestone method of Bayesian variable selection via stochastic search. They introduced a latent binary vector to indicate the inclusion of variables in linear models, and then placed a mixture spike and slab prior on each coefficient conditioning on this latent vector. Following this approach, many other selection procedures with similar structure have been proposed. The distinction between them is mostly in the form of the spike and slab priors, or in the form of the prior on the model space. To alleviate the difficulty in choosing specific prior parameters, several approaches have been proposed, see [10,14,30]. However, these papers focused on small-scaled questions and did not discuss any possible extension to the high-dimensional setting. More recently, Ishwaran and Rao [15] established the oracle property of the posterior mean as n converges to infinity with fixed p under certain conditions on the prior variances for linear models. Johnson and Rossell [16] proved selection consistency under $p = O(n)$ for a non-local prior in linear model settings. Liang et al. [19] proposed a point-mass spike prior with a slab prior depending on the model size, and proved the posterior consistency under $\ln p = o(n)$ in generalized linear models, but the corresponding conditions are relatively strong. Additionally, the step-wise estimation procedure is not efficient. Narisetty and He [21] also used Gaussian prior but argued the prior should be sample-size dependent, referred to as Bayesian shrinking and diffusing priors (BASAD), and obtained strong selection consistency when $\ln p = o(n)$ for linear models under mild assumptions. However, BASAD is not computationally practical for large-p problems, since it requires to update $\boldsymbol{\beta}$ from a p-dimensional multivariate normal distribution in each iteration. Recently Narisetty et al. [22] proposed Skinny Gibbs (SG) algorithm to address this computation issue via sparsifying the precision matrix. They referred to this kind of update as Skinny Gibbs (SG) and argued that it is a scalable method, namely the required computation time grows approximately linearly in p. The selection consistency was proved for the logistic regression. While spike and slab priors have been widely used in applications for its attractive interpretability, the theory for spike and slab models has not caught up with the applications. Again, all the aforementioned papers focused on linear or generalized linear models, and the corresponding work on semiparametric or nonparametric models under high dimensional setting is limited.

In this paper, we propose a Bayesian subset selection procedure for the partially linear model. We incorporate the difference-based method in the prior for the nonparametric component. For the parametric component, we adopt a modified version of Bayesian shrinking and diffusing priors (BASAD) [21] and propose the novel Bayesian subset modeling with diffusing prior (BSM-DP). We use a normal distribution with a diverging variance as the slab prior and a normal distribution with a small variance as the spike prior. Differently from BASAD, the response variable in our model only depends on the active covariates. This conveniently allows us to sample coefficients separately for the active and the inactive sets during the estimation. In fact, the spike prior has no impact on the theoretical result, so any proposal including a point mass will work. As a practical note, we recommend a Gaussian distribution with a small variance, which allows more flexibility for the Markov chain to explore the model space, and hence avoids local trap. As a result, the proposed methods are more computationally efficient than BASAD. We also notice that the Skinny Gibbs (SG) [22] is a special case of BSM-DP when the variance of spike prior is set to be proportional to $1/n$. Their original paper [22] discussed logistic regression only. We establish the selection consistency for the parametric component in partially linear models when $\ln p = o(n)$ under mild conditions.

The rest of the paper is organized as follows. In Section 2 we present the Bayesian subset modeling with diffusing prior (BSM-DP) and discuss variable selection for partially linear model, followed by the estimation procedure, regularity conditions and theoretical results. Performances of several numerical studies are presented in Section 3 to demonstrate reliability of the proposed model. We further apply the proposed method on the supermarket data set. Proofs for lemmas and theorems are given in Section 4, followed by discussions in Section 5.

2. Bayesian subset modeling with diffusing prior

2.1. Model and notation

Suppose that $\{(Y_i, \mathbf{X}_i, U_i)\}$, $i \in \{1, \ldots, n\}$ is a random sample from PLM (1) with high-dimensional covariates $\mathbf{X} \in \mathbb{R}^{p_n}$ and univariate covariate $U \in [0, 1]$, where we use p_n to emphasize that the number of variables is allowed to diverge with sample size n. Assume that the random error ϵ is independent of $(\mathbf{X}^\top, U)$ and each observation $\mathbf{X}_i$ has the same distribution with mean 0 and covariance $\mathbf{\Sigma}$. Denote $f(U_i)$ as α_i, and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^\top$ as a vector with size n. Notation $\mathbf{Y}$ is the corresponding size n vector, and $\mathbf{X}$ is the design matrix with size $n \times p_n$.

We will propose a prior for the nonparametric function (i.e. the $\boldsymbol{\alpha}$) in our proposed Bayesian subset selection based on difference-based method. Assume the observation $\{(Y_i, \mathbf{X}_i, U_i)\}_{1\le i\le n}$ is ordered by the increasing order of $\{U_i\}_{1\le i\le n}$. The difference in observed value for contingent Y can be written as

$$Y_i - Y_{i-1} = \{f(U_i) - f(U_{i-1})\} + (\mathbf{X}_i - \mathbf{X}_{i-1})^\top \boldsymbol{\beta} + \epsilon_i - \epsilon_{i-1}, \quad i \in \{2, \ldots, n\}.$$

If U_{i-1} and U_i are close and $f(\cdot)$ is smooth enough, $f(U_i)$ should also be close to $f(U_{i-1})$. So the nonparametric part tends to be canceled out. In this case, the ordinary least squares estimate can be applied on the differenced data, as long as $\mathbf{X}$ is not perfectly correlated with U. Define the mth higher order difference sequence to be $\{d_i\}_{i\in\{1,\ldots,m+1\}}$, which satisfies $\sum_{i=1}^{m+1} d_i = 0$ and $\sum_{i=1}^{m+1} d_i^2 = 1$. So the mth order difference operation reduces the sample size to $n - m$ by defining

$$Y_i^* = C^{1/2} \sum_{t=1}^{m+1} d_t Y_{i+m+1-t}, \mathbf{X}_i^* = C^{1/2} \sum_{t=1}^{m+1} d_t \mathbf{X}_{i+m+1-t}, \delta_i = C^{1/2} \sum_{t=1}^{m+1} d_t f(U_{i+m+1-t}), \omega_i = C^{1/2} \sum_{t=1}^{m+1} d_t \epsilon_{i+m+1-t},$$

for $i \in \{1, \ldots, n - m\}$, where C is some positive constant. Define the difference matrix $\mathbf{D}$ as

$$\mathbf{D} = \begin{pmatrix} d_{m+1} & \ldots & d_1 & 0 & \ldots & \ldots & 0 \\ 0 & d_{m+1} & \ldots & d_1 & 0 & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & \ldots & 0 & d_{m+1} & \ldots & d_1 & 0 \\ 0 & \ldots & \ldots & 0 & d_{m+1} & \ldots & d_1 \end{pmatrix} \in \mathbb{R}^{(n-m)\times n}. \tag{2}$$

Therefore the PLM (1) can be rewritten in matrix form as

$$\mathbf{Y}^* = \mathbf{X}^* \boldsymbol{\beta} + \boldsymbol{\delta} + \boldsymbol{\omega},$$

where $\mathbf{Y}^* = C^{1/2}\mathbf{D}\mathbf{Y} \in \mathbb{R}^{n-m}$, $\mathbf{X}^* = C^{1/2}\mathbf{D}\mathbf{X} \in \mathbb{R}^{(n-m)\times p_n}$, $\boldsymbol{\delta} = C^{1/2}\mathbf{D}\boldsymbol{\alpha} \in \mathbb{R}^{n-m}$, $\boldsymbol{\omega} = C^{1/2}\mathbf{D}\boldsymbol{\epsilon} \in \mathbb{R}^{n-m}$. Under some smoothness conditions on $f(\cdot)$ with fixed p, Yatchew [29] and Wang et al. [27] showed that the ordinary least square estimator $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^{*\top}\mathbf{X}^*)^{-1}\mathbf{X}^{*\top}\mathbf{Y}^*$ is asymptotically efficient when $m \to \infty$, if $\mathbf{X}$ and U are independent. This indicates that the effect of the nonparametric component is negligible after applying a high order difference operation on the data.

In the literature, U_i's are either from a fixed design e.g., $U_i = i/n$, or observations from a distribution on $[0, 1]$ with density function bounded away from 0. In this paper, we only consider the case when $\mathbf{X}$ and U are independent under a dense design with a constraint on $\max_{2\le i\le n} |U_i - U_{i-1}|$.

We use X_j as the notation for the jth covariate. A size-p_n latent binary random vector is introduced as $\boldsymbol{\gamma}$. The jth entry γ_j indicates whether X_j is included in the model (1 = present, 0 = not present). Therefore, the model space is fully specified by $\boldsymbol{\gamma}$, and we use $\boldsymbol{\gamma}$ and $\mathcal{M}$ as notations for models interchangeably. The true model is denoted as $\mathcal{A}$. The cardinality of model $\mathcal{M}$, denoted by $|\mathcal{M}|$, is the size of the model. Consequently, if $\boldsymbol{\beta}_{\mathcal{M}}$ is the subvector of $\boldsymbol{\beta}$ with size $|\mathcal{M}|$, $\mathbf{X}_{\mathcal{M}}$ is the submatrix of $\mathbf{X}$ with respect to model $\mathcal{M}$, and $\mathbf{\Sigma}_{\mathcal{M}}$ is the $|\mathcal{M}| \times |\mathcal{M}|$ covariance matrix for $\mathbf{X}_{\mathcal{M}}$. Other notations used in the paper are unified as follows.

- Model operation: $\mathcal{M}_1 \wedge \mathcal{M}_2$ and $\mathcal{M}_1 \vee \mathcal{M}_2$ are defined as the intersection and union of model $\mathcal{M}_1$ and $\mathcal{M}_2$, for example $\mathcal{M}_1 \wedge \mathcal{M}_2 = \{i : i \in \mathcal{M}_1 \text{ and } i \in \mathcal{M}_2\}$.
- Rate: $a_n \preceq b_n$ or $b_n \succeq a_n$ means $a_n = O(b_n)$, $a_n \prec b_n$ or $b_n \succ a_n$ means $a_n = o(b_n)$. And $a_n \sim b_n$ refers to $a_n/b_n \longrightarrow c$ for some positive constant.
- Matrix and matrix operation: $n \times n$ identity matrix is denoted as $\mathbf{I}_n$. For a matrix $\mathbf{M}$, $\|\mathbf{M}\|$ is the spectral norm, which is the largest singular value of $\mathbf{M}$. The Moore–Penrose inverse of $\mathbf{M}$ is denoted by $\mathbf{M}^+$, which is the unique generalized inverse. And if $\mathbf{M}$ is a positive definite matrix we use $\lambda_{\min}(\mathbf{M})$ and $\lambda_{\max}(\mathbf{M})$ as the notation for the minimum and the maximum eigenvalues of $\mathbf{M}$.

Here and hereafter, the densities are conditional on $\mathbf{X}$ and U. The working model for variable selection in the partially linear model in (1) via Bayesian subset modeling with diffusing prior (BSM-DP) is proposed as

$$\begin{aligned} \pi(\mathbf{Y}|\boldsymbol{\alpha},\boldsymbol{\gamma},\boldsymbol{\beta},\sigma) &= \mathcal{N}(\boldsymbol{\alpha}+\mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}},\sigma^2\mathbf{I}_n),\\ \pi(\beta_j|\boldsymbol{\gamma},\sigma) &= \begin{cases}\mathcal{N}(0,\sigma_{1n}^2\sigma^2) & \gamma_j=1\\ \mathcal{N}(0,\sigma_0^2\sigma^2) & \gamma_j=0\end{cases}, \quad j\in\{1,\ldots,p_n\},\\ \pi(\boldsymbol{\gamma}) &= \prod_{j=1}^{p_n} q_n^{\gamma_j}(1-q_n)^{1-\gamma_j}, \quad \pi(\boldsymbol{\alpha}|\sigma)=\mathcal{N}(0,\sigma^2\boldsymbol{\Sigma}_{0n}), \quad \pi(\sigma^2)=IG(a_0,b_0), \end{aligned} \tag{3}$$

where, $\boldsymbol{\Sigma}_{0n}=\left\{(\mathbf{I}_n-C\mathbf{D}^\top\mathbf{D})^{-1}-\mathbf{I}_n\right\}^+$, $\mathbf{D}$ is the difference matrix defined in (2).

We choose the classical Inverse Gamma distribution as the prior for σ^2 as it is the most commonly used conjugate prior. Other choices of prior could be used, and it can be shown that Theorem 5 applies to a wider family of priors, including some commonly choices like improper non-informative prior and the class of folded-noncentral-t prior (see Remark 8). Meanwhile, independent Bernoulli distribution with probability q_n is used as the prior for each γ_j. So the preliminary marginal inclusion probability for each variable is q_n. It is natural to assume that when the dimension p_n diverges with the sample size, q_n should converge at some rate to 0. Each β_i has a mixture normal distribution. Conditioning on $\gamma_j=1$, β_j has a normal distribution with a relative large variance $\sigma_{1n}^2\sigma^2$. This corresponds to a very wide and flat distribution, usually referred to as a slab prior. We call it the diffusing prior as named in [21]. Within its variance, σ_{1n} depends on the sample size, and diverges at some certain rate when the sample size goes to infinity. Conditioning on $\gamma_j=0$, β_j has a normal distribution with variance $\sigma_0^2\sigma^2$. As the choice of σ_0^2 would not influence the asymptotic results, it can be chosen depending on sample size or simply as a fixed value.

With a partially linear model, we will need to accommodate the nonparametric part. A conjugate prior of normal distribution with a semi-definite covariance matrix $\sigma^2\boldsymbol{\Sigma}_{0n}$ is proposed for $\boldsymbol{\alpha}=(f(U_1),\ldots,f(U_n))$. The covariance matrix $\sigma^2\boldsymbol{\Sigma}_{0n}$ is further taken as a function of the difference matrix $\mathbf{D}$ thus to eliminate the effect of nonparametric function. More intuitions about the choice of $\boldsymbol{\Sigma}_{0n}$ will be discussed later. The error term is assumed to be normally distributed. Therefore, conditional on the latent indicator $\boldsymbol{\gamma}$, coefficients $\boldsymbol{\beta}$, nonparametric component $\boldsymbol{\alpha}$, and the variance for the error σ, $\mathbf{Y}$ has a normal distribution.

Remark 1 (*Comparison with BASAD [21] and SG [22]*)**.** As mentioned earlier, the inclusion of $\boldsymbol{\gamma}$ in the conditional distribution of Y distinguishes our model from BASAD. This difference allows us to sample separately for the active and the inactive groups.

In our working model (3), the response variable Y is conditioned on $\boldsymbol{\gamma}$, hence only depends on the active covariates $\mathbf{X}_{\boldsymbol{\gamma}}$. But in BASAD, Y depends on both the active and nonactive part of the covariates. As a result, the full conditional distributions for $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ and $\boldsymbol{\beta}_{\boldsymbol{\gamma}^c}$ are not independent in BASAD. Therefore, to update $\boldsymbol{\beta}$ in MCMC, each iteration requires sampling a size-p vector from a multivariate normal distribution. This will increase the computational time quickly under large p. On the other hand, the full conditional distributions for $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ and $\boldsymbol{\beta}_{\boldsymbol{\gamma}^c}$ are independent in our proposal. So in each iteration, we only need to sample a size-$|\boldsymbol{\gamma}|$ vector from a multivariate normal distribution and sample $(p-|\boldsymbol{\gamma}|)$ scalars from independent univariate normal distributions. The current active model size $|\boldsymbol{\gamma}|$ is usually small after several iterations if the true model is sparse. In fact, just as SG, the proposed BSM-DP is scalable in high dimensional problems, which means that the computation time growing approximately linearly with the dimension p. The computational complexity for each iteration in the estimation procedure is $n(p\vee|\boldsymbol{\gamma}|^2\vee n^2)$. We have also validated this claim in the simulation study for PLM and more simulation studies about linear models as compared with BASAD and SG in the supplementary material. It can be shown that SG is equivalent to our Bayesian subset modeling by taking the variance of spike prior to be $\sigma_0^2=(n+\tau_{0n}^{-2})^{-1}$, where τ_{0n}^2 is the variance for spike prior in SG.

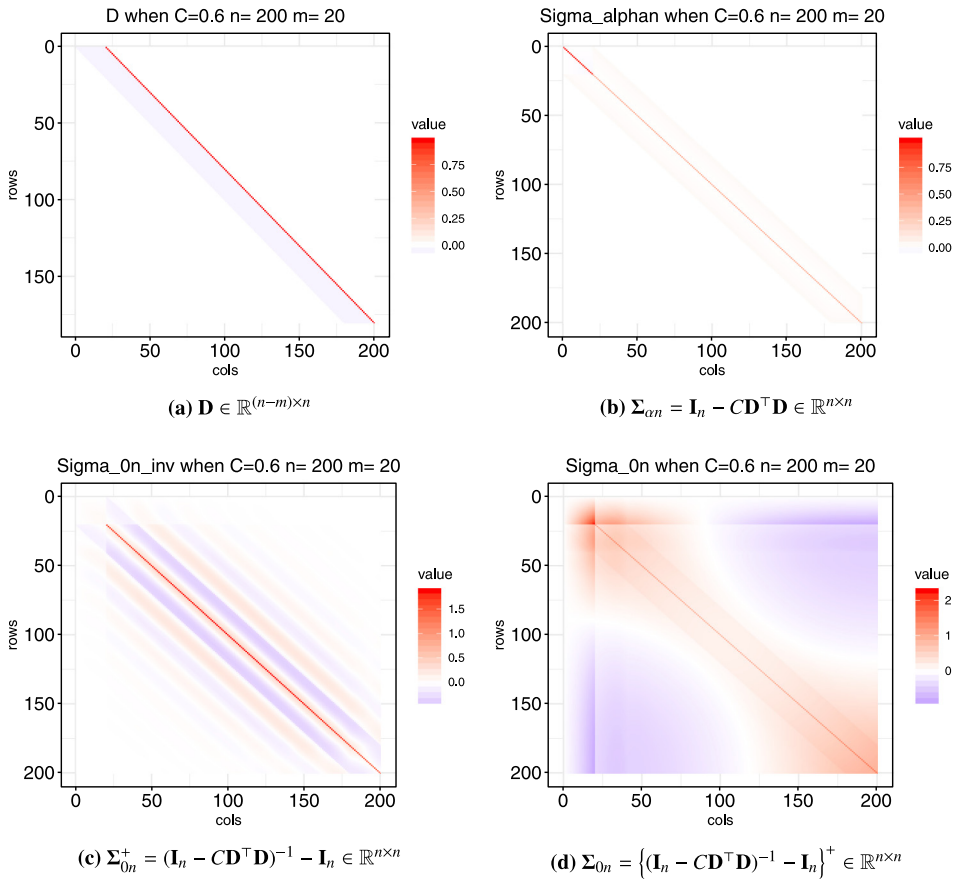
2.2. Estimation procedure

Gibbs sampling is used to update our parameters iteratively. In each iteration, we draw samples from those full conditional distributions.

1. Update γ_k from a Bernoulli distribution:
   Full conditional distribution of γ_k is a Bernoulli distribution with probability $\Pr(\gamma_k=1|\boldsymbol{\beta},\boldsymbol{\alpha},\sigma^2,\mathbf{Y},\gamma_{-k})=p_1/(p_1+p_2)$, and

$$\frac{p_1}{p_2}=\frac{q_n\sigma_0}{(1-q_n)\sigma_{1n}}\exp\left\{-\frac{\beta_k^2}{2\sigma_{1n}^2\sigma^2}+\frac{\beta_k^2}{2\sigma_0^2\sigma^2}+\beta_k X_k^\top\left(\mathbf{Y}-\boldsymbol{\alpha}-\mathbf{X}_{\widehat{\mathcal{A}}_k}\boldsymbol{\beta}_{\widehat{\mathcal{A}}_k}\right)/\sigma^2-\beta_k^2X_k^\top X_k/\left(2\sigma^2\right)\right\},$$

   where X_i is the ith column of $\mathbf{X}$, and the index $\widehat{\mathcal{A}}_k$ is the collection of current active covariates after removing the kth covariate, that is $\widehat{\mathcal{A}}_k=\{i:\gamma_i=1,i\neq k\}$.



(a) $\mathbf{D} \in \mathbb{R}^{(n-m)\times n}$ **(b)** $\boldsymbol{\Sigma}_{\alpha n} = \mathbf{I}_n - C\mathbf{D}^\top\mathbf{D} \in \mathbb{R}^{n\times n}$

(c) $\boldsymbol{\Sigma}_{0n}^{+} = (\mathbf{I}_n - C\mathbf{D}^\top\mathbf{D})^{-1} - \mathbf{I}_n \in \mathbb{R}^{n\times n}$ **(d)** $\boldsymbol{\Sigma}_{0n} = \left\{(\mathbf{I}_n - C\mathbf{D}^\top\mathbf{D})^{-1} - \mathbf{I}_n\right\}^{+} \in \mathbb{R}^{n\times n}$

Fig. 1. Visualizations to display the magnitude of values in the difference matrix $\mathbf{D}$, the covariance matrix $\boldsymbol{\Sigma}_{\alpha n}$ used for the update of $\boldsymbol{\alpha}$ and the covariance matrix for prior $\boldsymbol{\Sigma}_{0n}$. All plots are taking constant $C = 0.6$, sample size $n = 200$ and difference order $m = 20$. On the graph, red color indicates positive values on the corresponding locations of the matrix, and purple color represents negative values. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2. Update $\boldsymbol{\beta}$ from multivariate normal distributions:
   In each iteration, we divide $\boldsymbol{\beta}$ into the active group and the inactive group based on the current $\boldsymbol{\gamma}$. Denote $\widehat{\mathcal{A}}$ as the collection of covariates with $\gamma_j = 1$, and $\widehat{\mathcal{I}}$ as the collection of covariates with $\gamma_j = 0$. Rewrite $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_{\widehat{\mathcal{A}}}, \boldsymbol{\beta}_{\widehat{\mathcal{I}}}\right)$, and we can update those two groups separately.

   I. Update the active group $\widehat{\mathcal{A}}$: $\boldsymbol{\beta}_{\widehat{\mathcal{A}}} \sim \mathcal{N}\left(\mathbf{V}\mathbf{X}_{\widehat{\mathcal{A}}}^\top(\mathbf{Y} - \boldsymbol{\alpha}), \sigma^2\mathbf{V}\right)$, where $\mathbf{V} = \left(\mathbf{X}_{\widehat{\mathcal{A}}}^\top\mathbf{X}_{\widehat{\mathcal{A}}} + \frac{1}{\sigma_{1n}^2}\mathbf{I}_{|\widehat{\mathcal{A}}|}\right)^{-1}$.

   II. Update the inactive group $\widehat{\mathcal{I}}$: $\boldsymbol{\beta}_{\widehat{\mathcal{I}}} \sim \mathcal{N}\left(0, \sigma_0^2\sigma^2\mathbf{I}_{|\widehat{\mathcal{I}}|}\right)$.

3. Update σ^2 from an Inverse Gamma distribution:
   $\sigma^2 \sim IG(a, b)$, where

$$a = a_0 + (n + p_n)/2,$$
$$b = b_0 + \boldsymbol{\beta}_{\widehat{\mathcal{A}}}^\top\boldsymbol{\beta}_{\widehat{\mathcal{A}}}/\left(2\sigma_{1n}^2\right) + \boldsymbol{\beta}_{\widehat{\mathcal{I}}}^\top\boldsymbol{\beta}_{\widehat{\mathcal{I}}}/(2\sigma_0^2) + \left(\mathbf{Y} - \boldsymbol{\alpha} - \mathbf{X}_{\widehat{\mathcal{A}}}\boldsymbol{\beta}_{\widehat{\mathcal{A}}}\right)^\top\left(\mathbf{Y} - \boldsymbol{\alpha} - \mathbf{X}_{\widehat{\mathcal{A}}}\boldsymbol{\beta}_{\widehat{\mathcal{A}}}\right)/2 + \boldsymbol{\alpha}^\top\boldsymbol{\Sigma}_{0n}\boldsymbol{\alpha}/2.$$

4. Update $\boldsymbol{\alpha}$ from a multivariate normal distribution:
   $\boldsymbol{\alpha} \sim \mathcal{N}\left(\boldsymbol{\Sigma}_{\alpha n}\left(\mathbf{Y} - \mathbf{X}_{\widehat{\mathcal{A}}}\boldsymbol{\beta}_{\widehat{\mathcal{A}}}\right), \sigma^2\boldsymbol{\Sigma}_{\alpha n}\right)$, where $\boldsymbol{\Sigma}_{\alpha n} = \left(\boldsymbol{\Sigma}_{0n}^{+} + \mathbf{I}_n\right)^{-1}$, furthermore by Condition A, $\boldsymbol{\Sigma}_{\alpha n} = \mathbf{I}_n - C\mathbf{D}^\top\mathbf{D}$.

In the literature, the nonparametric function $f(\cdot)$ is usually assumed to be smooth, which means $f(x)$ and $f(y)$ should be close if x and y are close enough. This dependency among $f(U_1), f(U_2), \ldots, f(U_n)$ suggests that the covariance matrix $\boldsymbol{\Sigma}_{0n}$ of $\boldsymbol{\alpha}$ has to be a dense matrix. Here we take $\boldsymbol{\Sigma}_{0n}$ to be $\boldsymbol{\Sigma}_{0n} = \left\{(\mathbf{I}_n - C\mathbf{D}^\top\mathbf{D})^{-1} - \mathbf{I}_n\right\}^{+}$ where C is some positive constant and $\mathbf{D}$ is the difference matrix defined in (2). We will show the reason for this specific choice of $\boldsymbol{\Sigma}_{0n}$ in Remark 2. Fig. 1. shows the intuitive structure of difference matrix $\mathbf{D}$, the matrix $\boldsymbol{\Sigma}_{\alpha n}$ used for the update of $\boldsymbol{\alpha}$, and covariance matrix for the prior $\boldsymbol{\Sigma}_{0n}$ when constant $C = 0.6$ with sample size $n = 200$ and difference order $m = 20$. As demonstrated in the figure,

$\mathbf{D} \in \mathbb{R}^{(n-m)\times n}$ is a general upper triangular band matrix with bandwidth m. The update matrix $\mathbf{\Sigma}_{\alpha n}$ is also a band matrix with bandwidth m. The covariance matrix for the prior of the nonparametric component $\mathbf{\Sigma}_{0n} \in \mathbb{R}^{n\times n}$ is a dense matrix with positive larger values near the diagonal, then decays gradually to 0 and negative when moving further. The reason why it has negative off-diagonal values is because the difference sequence $\{d_i\}_{1\le i\le m+1}$ is standardized to be centered at 0. Theoretically, we do require the difference order m, which is also the bandwidth, goes to infinity as $n \longrightarrow \infty$ at some slow rate. In this way, the effect of the nonparametric component can be removed without over-smoothing the nonparametric function $f(\cdot)$, so the selection consistency for the linear component holds.

2.3. Selection procedure

In the typical Bayesian variable selection approach, the model with the highest posterior is selected as the final model, referred to as maximum a posterior model (MAP): $\widehat{\mathcal{M}} = \text{argmax}_{\mathcal{M}} \Pr(\boldsymbol{\gamma} = \mathcal{M}|Y)$. With the spike and slab prior, the posterior of the model space is usually reflected by the posterior probability of the latent variable $\boldsymbol{\gamma}$. Alternatively, another way is to consider the marginal probability of $\Pr(\gamma_j = 1|Y)$. Specifically, one will select the jth covariate if $\Pr(\gamma_j = 1|Y)$ is equal to or greater than a certain threshold. A threshold of 0.5 is a natural choice. This is known as the median probability model (MPM). It has been shown that MPM has good predictive power [1]. Although it is likely that these two approaches may produce different results in practice, it can be shown those two selection methods are asymptotically the same under strong selection consistency, which will be shown in Section 2.4. Moreover, some other data-driven criteria could also be used in determining the threshold, e.g. AIC, BIC and EBIC [4].

2.4. Theoretical results

Variable selection procedures typically aim to achieve selection consistency, and under Bayesian framework, it means conditional on observed data, the probability of the true model $\mathcal{A}$ being selected goes to 1 in probability.

$$\Pr(\widehat{\mathcal{M}} = \mathcal{A} \mid Y) \xrightarrow{P} 1 \quad \text{as } n \to \infty.$$

That is, the true model is selected consistently. Note that the posterior of model space is fully specified by $\boldsymbol{\gamma}$. If the model is selected via MAP: $\widehat{\mathcal{M}} = \text{argmax}_{\mathcal{M}} \Pr(\boldsymbol{\gamma} = \mathcal{M}|Y)$, then the selection consistency only requires that the posterior probability of the true model, i.e. $\Pr(\gamma = \mathcal{A}|Y)$ is no less than that of any other models, i.e. $\Pr(\gamma = \mathcal{M}|Y)$. But the difference in their posterior probabilities could still shrink to 0. In this paper, we will consider the following strong selection consistency

$$\Pr(\boldsymbol{\gamma} = \mathcal{A} \mid Y) \xrightarrow{P} 1 \quad \text{as } n \to \infty.$$

It indicates the difference for the posterior probabilities of the true model and any other model is 1. This non-zero difference indicates a stronger conclusion than selection consistency. We first present the following regularity conditions for the selection consistency of the linear component in the PLM, and we then start with the case when σ^2 is known as it provides intuitive interpretation for the proposed method.

Condition A (On the dimension and priors). The dimension p_n satisfies that $\ln(p_n) = o(n)$. The prior probability that a coefficient is nonzero q_n satisfies that $q_n \sim 1/p_n$. The variance for slab prior $\sigma_{1n}^2 \to \infty$ as $n \to \infty$, and $n\sigma_{1n}^2\lambda_1 \sim p_n^{2+3\delta}$ for some $\delta > 0$, where λ_1 is defined in Conditions C. The covariance for the prior of nonparametric component $\mathbf{\Sigma}_{0n} = \left\{(\mathbf{I}_n - C\mathbf{D}^\top\mathbf{D})^{-1} - \mathbf{I}_n\right\}^+$, where C is a positive constant, with values no greater than $\min\left\{1, 1/\lambda_{\max}(\mathbf{D}^\top\mathbf{D})\right\}$, and $\mathbf{D}$ is the difference matrix defined in (2).

Condition B (Identifiability). There exists $K > 1 + 4/\delta$ such that

$$\Delta_n(K) = \inf_{\mathcal{M}:|\mathcal{M}|\le K|\mathcal{A}|, \mathcal{A}\not\subseteq\mathcal{M}} \|(\mathbf{I} - \mathbf{P}_{\mathcal{M}})\mathbf{X}_{\mathcal{A}}^{*\top}\boldsymbol{\beta}_{\mathcal{A}}\|_2^2 > \sigma^2|\mathcal{A}|(4 + 4\delta + \kappa)\ln p_n,$$

where $\mathbf{X}^* = C^{1/2}\mathbf{D}\mathbf{X}$ and the projection matrix $\mathbf{P}_{\mathcal{M}} = \mathbf{X}_{\mathcal{M}}^*(\mathbf{X}_{\mathcal{M}}^{*\top}\mathbf{X}_{\mathcal{M}}^*)^{-1}\mathbf{X}_{\mathcal{M}}^{*\top}$.

Condition C (Regularity of the design). Define

$$\lambda_1 = \min_{\mathcal{M}:|\mathcal{M}|\le m_n+|\mathcal{A}|} \lambda_{\min}\left(\frac{1}{n}\mathbf{X}_{\mathcal{M}}^\top\mathbf{X}_{\mathcal{M}}\right), \quad \lambda_2 = \max_{\mathcal{M}:\mathcal{M}\subseteq\mathcal{A}} \lambda_{\max}\left(\frac{1}{n}\mathbf{X}_{\mathcal{M}}^\top\mathbf{X}_{\mathcal{M}}\right),$$

then $(\lambda_2/\lambda_1)^{|\mathcal{A}|} \preceq p_n^\kappa$, for some $0 < \kappa < \delta$, where m_n is defined in Condition D.

Condition D (On the true model). Let $\boldsymbol{\beta}_{\mathcal{A}}$ consist of all nonzero elements of $\boldsymbol{\beta}$. That is, $\mathbf{X}_{\mathcal{A}}$ consists of all active predictors. The size of $\mathcal{A}$ satisfies that $|\mathcal{A}| = o(m_n)$, $m_n = cn/\ln p_n$, where $c < \delta/\{(4+\delta)(2+\delta)\}$. Further assume that $U_\infty = \max_{2\le i\le n}|U_i - U_{i-1}| = O\left(n^{-c_1}\right)$ for some $0 < c_1 \le 1$.

Condition E (On the difference matrix). Let $\mathbf{D}$ be the difference matrix, as defined in (2). Denote $h_k = \sum_{i=1}^{m+1-k} d_i d_{i+k}$, then the mth difference sequence $d_1, \ldots, d_{m+1}$ satisfies,

$$\sum_{i=1}^{m+1} d_i = 0, \quad \sum_{i=1}^{m+1} d_i^2 = 1, \quad \sum_{k=1}^{m} h_k^2 = O(m^{-1}), \quad 1 - d_1^2 = O(m^{-1}).$$

Furthermore $m \longrightarrow \infty$, $m = o(n^{c_2})$ for some $0 < c_2 < c_1 \le 1$, where c_1 is defined in Condition D.

Condition F (On the nonparametric component $f(\cdot)$). Suppose $f(\cdot) \in \Lambda^k(M)$ for some $k > \frac{1}{2(c_1-c_2)}$, where c_1 and c_2 are defined in Conditions D, E. The Lipschitz ball $\Lambda^k(M)$ is defined as

$$\Lambda^k(M) = \left\{ f : \text{for all } 0 \le x, y \le 1, \ \ |f^{(i)}(x)| \le M, i \in \{0, \ldots, \lfloor k \rfloor - 1\}, |f^{(\lfloor k \rfloor)}(x) - f^{(\lfloor k \rfloor)}(y)| \le M|x-y|^{k'} \right\},$$

where $\lfloor k \rfloor$ is the largest integer less than k and $k' = k - \lfloor k \rfloor$.

The convergence and divergence rates of the parameters in the priors and the dimension of the variables are stated in Condition A. Identifiability Condition B is needed to distinguish active covariates out of spurious ones. Condition C gives the regularity condition of the design matrix. Instead of requiring bounded eigenvalues, we will need the minimal eigenvalues to decay slower than some rate and the maximal eigenvalues to diverge slower than some rate. We would like to point that, if the size of the true model is not too large, the condition holds even with the extreme case when $\mathbf{X}$ is sampled from normal distribution with compound symmetric covariance matrix when correlation among predictors $\rho \longrightarrow 1$. Theoretically the model still works even under nearly perfectly correlated covariates. Condition D states the normality assumption for the error and we do allow infinitely many active variables.

Conditions E–F control the error in estimating the nonparametric component. Condition E is about the difference matrix. A sequence satisfying the above conditions, for example, is $d_1 = \sqrt{\frac{m}{m+1}}, d_2 = d_3 = \cdots = d_{m+1} = -\sqrt{\frac{1}{m(m+1)}}$. As argued in [23] about partial smoothing spline method for PLM, higher ordered difference operation gives lower approximation error. We do assume $m \longrightarrow \infty$, so the approximation error becomes ignorable. In the estimation of the partial residuals, it requires that the nonparametric estimators of $\mathrm{E}(Y|U)$ and $\mathrm{E}(\mathbf{X}|U)$ to converge sufficiently fast so that their substitutions in the OLS estimator do not affect its asymptotic distribution. Similarly, our upper bound for the growth rate of difference order m reflects this. Finally, the commonly used smoothness assumption for nonparametric nuisance function is stated in Condition F.

The first step is to derive posterior probability of any model $\mathcal{M}$.

Lemma 1. *Under fixed σ^2, for any model $\mathcal{M}$, the posterior probability has the following explicit form:*

$$\Pr(\boldsymbol{\gamma} = \mathcal{M} \mid \mathbf{Y}, \sigma^2) \propto q_n^{|\mathcal{M}|} T_{\mathcal{M}} \exp\left(-\frac{1}{2\sigma^2} R_{\mathcal{M}}\right),$$

where

$$R_{\mathcal{M}} = \mathbf{Y}^\top \left\{ \boldsymbol{\Sigma}_{1n} - \boldsymbol{\Sigma}_{1n}\mathbf{X}_{\mathcal{M}} \left(\mathbf{X}_{\mathcal{M}}^\top \boldsymbol{\Sigma}_{1n} \mathbf{X}_{\mathcal{M}} + \mathbf{I}_{|\mathcal{M}|}/\sigma_{1n}^2 \right)^{-1} \mathbf{X}_{\mathcal{M}}^\top \boldsymbol{\Sigma}_{1n} \right\} \mathbf{Y},$$

$$T_{\mathcal{M}} = \sigma_{1n}^{-|\mathcal{M}|} \left| \mathbf{X}_{\mathcal{M}}^\top \boldsymbol{\Sigma}_{1n} \mathbf{X}_{\mathcal{M}} + \mathbf{I}_{|\mathcal{M}|}/\sigma_{1n}^2 \right|^{-1},$$

and $\boldsymbol{\Sigma}_{1n} = \mathbf{I}_n - (\boldsymbol{\Sigma}_{0n}^+ + \mathbf{I}_n)^{-1}$, $\boldsymbol{\Sigma}_{0n}$ is the covariance matrix for the prior of $f(U)$.

Furthermore, define the likelihood ratio between model $\mathcal{M}$ and the true model $\mathcal{A}$ as $PR(\mathcal{M}, \mathcal{A})$. If Conditions A, C hold, $PR(\mathcal{M}, \mathcal{A})$ is bounded by

$$PR(\mathcal{M}, \mathcal{A}) = \frac{\Pr(\boldsymbol{\gamma} = \mathcal{M}|\mathbf{Y}, \sigma)}{\Pr(\boldsymbol{\gamma} = \mathcal{A}|\mathbf{Y}, \sigma)} \preceq p_n^{-1.5\delta(|\mathcal{M}|-|\mathcal{A}|)+0.5\kappa} \exp\left\{ -\frac{1}{2\sigma^2}(R_{\mathcal{M}} - R_{\mathcal{A}}) \right\}. \tag{4}$$

Remark 2. Lemma 1 gives the explicit form of the posterior probability for any given model $\mathcal{M}$ and puts an upper bound on the likelihood ratio between model $\mathcal{M}$ and the true model $\mathcal{A}$. Intuitively from (4), when σ_{1n}^2 is sufficiently large, $R_{\mathcal{M}}$ is close to

$$R_{\mathcal{M}}^* = \mathbf{Y}^\top \left\{ \boldsymbol{\Sigma}_{1n} - \boldsymbol{\Sigma}_{1n}\mathbf{X}_{\mathcal{M}} \left(\mathbf{X}_{\mathcal{M}}^\top \boldsymbol{\Sigma}_{1n} \mathbf{X}_{\mathcal{M}} \right)^{-1} \mathbf{X}_{\mathcal{M}}^\top \boldsymbol{\Sigma}_{1n} \right\} \mathbf{Y}.$$

So if $\boldsymbol{\Sigma}_{1n}$ is taken to be $C\mathbf{D}^\top\mathbf{D}$, which means $\boldsymbol{\Sigma}_{0n} = \left\{ (\mathbf{I}_n - C\mathbf{D}^\top\mathbf{D})^{-1} - \mathbf{I}_n \right\}^+$, then $R_{\mathcal{M}}^*$ is proportional to the sum of squared residuals under model $\mathcal{M}$, after applying the difference-based method. It could be interpreted as the goodness of fit. Additionally, the first term $p_n^{-1.5\delta(|\mathcal{M}|-|\mathcal{A}|)}$ in (4) could be regarded as the penalty on the model size. So it is mostly analogous to a L_0 penalized method. As σ_{1n} diverges fast, we can directly work with $R_{\mathcal{A}}^*$ and $R_{\mathcal{M}}^*$ instead of $R_{\mathcal{M}} - R_{\mathcal{A}}$. The following two lemmas present some properties of $R_{\mathcal{A}}^*$ and $R_{\mathcal{M}}^*$.

Lemma 2. *For any model $\mathcal{M}$ containing the true model, i.e $\mathcal{A} \subseteq \mathcal{M}$, if conditions A, E, F hold, then*

$$R_{\mathcal{M}}^* \sim C\sigma^2 \chi^2_{n-m-|\mathcal{M}|} \quad a.s.,$$
$$R_{\mathcal{A}}^* - R_{\mathcal{M}}^* \sim C\sigma^2 \chi^2_{|\mathcal{M}|-|\mathcal{A}|} \quad a.s..$$

Lemma 3. *Suppose that Conditions A, C, D are satisfied, then for any $g_n \to \infty$ and $\epsilon > 0$,*

(i) $\Pr(R_{\mathcal{A}} - R_{\mathcal{A}}^* > g_n) \le \exp(-c' n\lambda_1 \sigma_{1n}^2 g_n)$, *for some $c' > 0$, where λ_1 is defined in Condition C.*
(ii) $\Pr(|R_{\mathcal{A}}^*/(Cn\sigma^2) - 1| > \epsilon) \le \exp(-cn)$, *for some $c > 0$.*

Remark 3. Lemma 2 shows for over-fitted models, after applying the difference operation, the sum of squared residuals has an asymptotic χ^2 distribution with the degrees of freedom as $n - m - |\mathcal{M}|$. It also gives the asymptotic distribution of $R^*_{\mathcal{A}} - R^*_{\mathcal{M}}$. The difference of $R_{\mathcal{A}}$ and $R^*_{\mathcal{A}}$ is further bounded in Lemma 3. Lemma 3(ii) is straightforward by using the tail bound with χ^2 distributions.

Theorem 4 (*Strong Selection Consistency Under Fixed σ^2*)**.** *Suppose that Conditions A, B, C, D, E, F hold for the partially linear model in (1) with $\ln p_n = o(n)$, and $|\mathcal{A}| = o(m_n)$ is valid, we have*

$$\Pr\left(\gamma = \mathcal{A} \mid \mathbf{Y}, \sigma^2, |\gamma| \le m_n + |\mathcal{A}|\right) \xrightarrow{P} 1 \quad as \quad n \to \infty,$$

where $m_n = cn/\ln p_n$, as defined in Condition D.

Remark 4. It suffices to show

$$\sum_{\mathcal{M}\neq\mathcal{A}, |\mathcal{M}|\le m_n+|\mathcal{A}|} PR(\mathcal{M}, \mathcal{A}) = \sum_{\mathcal{M}\neq\mathcal{A}, |\mathcal{M}|\le m_n+|\mathcal{A}|} \frac{\Pr(\gamma = \mathcal{M}|\mathbf{Y}, \sigma)}{\Pr(\gamma = \mathcal{A}|\mathbf{Y}, \sigma)} \xrightarrow{P} 0.$$

Recall that, by (4) in Lemma 1, we have $PR(\mathcal{M}, \mathcal{A}) \preceq p_n^{-1.5\delta(|\mathcal{M}|-|\mathcal{A}|)+0.5\kappa} \exp\left\{-\frac{1}{2\sigma^2}(R_{\mathcal{M}} - R_{\mathcal{A}})\right\}$. Inspired by [21], we first divide the model space into 3 disjoint parts, $\mathcal{P}_i$, $i = 1, 2, 3$ defined as below. In each group, we prove the sum of likelihood ratio converges to 0 in probability.

1. Consider the set of overfitted models $\mathcal{P}_1 = \{\mathcal{M} : \mathcal{A} \subseteq \mathcal{M}, |\mathcal{M}| \le m_n + |\mathcal{A}|\}$. Model $\mathcal{M}$ in this group contains all active variables, so $R_{\mathcal{A}} - R_{\mathcal{M}}$ might be large. However since $|\mathcal{M}| - |\mathcal{A}| > 0$, the number of extra spurious variables is penalized at the rate of p_n^{-c}.
2. For large and including some inactive variables models $\mathcal{P}_2 = \{\mathcal{M} : \mathcal{A} \nsubseteq \mathcal{M}, K|\mathcal{A}| < |\mathcal{M}| \le m_n + |\mathcal{A}|\}$, let $\mathcal{M} \vee \mathcal{A}$ be the union of models $\mathcal{M}$ and $\mathcal{A}$, then any model $\mathcal{M}$ in this group has $\mathcal{M} \vee \mathcal{A} \in \mathcal{P}_1$. Although size of $\mathcal{M} \vee \mathcal{A}$ may exceed $m_n + |\mathcal{A}|$, but since m_n dominates $|\mathcal{A}|$ so it is reasonable to assume the difference is negligible. Model $\mathcal{M} \vee \mathcal{A}$ will also have a better fit than $\mathcal{M}$. Thus we can control $\exp\left\{-\frac{1}{2\sigma^2}(R_{\mathcal{M}} - R_{\mathcal{A}})\right\}$ by bounding the value of $\exp\left\{-\frac{1}{2\sigma^2}(R_{\mathcal{M}\vee\mathcal{A}} - R_{\mathcal{A}})\right\}$. Since the size of models in $\mathcal{P}_2$ is large, the growth of $\exp\left\{-\frac{1}{2\sigma^2}(R_{\mathcal{M}} - R_{\mathcal{A}})\right\}$ is under control.
3. Now consider the set of underfitted models missing some active variables, which is formulated as $\mathcal{P}_3 = \{\mathcal{M} : \mathcal{A} \nsubseteq \mathcal{M}, |\mathcal{M}| \le K|\mathcal{A}|\}$, where K is the constant defined in Condition B. For any model $\mathcal{M}$ in this group, at least one active variable is missed. By Condition B on the identifiability of the active variables, $R_{\mathcal{M}} - R_{\mathcal{A}}$ will be large since model $\mathcal{M}$ does not have a good fit, thus the value of $PR(\mathcal{M}, \mathcal{A})$ is controlled.

Remark 5. In this paper, we only consider models that are not unreasonably large, that is, $|\mathcal{M}| \le m_n + |\mathcal{A}|$, where m_n is at the order of $n/\ln p_n$. There is a reason behind the choice of m_n. It can be shown that the marginal probability for the inclusion of a single variable i.e. $q = \Pr(\gamma_i = 1)$ somehow controls the size of the selected model. Let $\widehat{\mathcal{M}} = \underset{\mathcal{M}}{\operatorname{argmax}} \Pr(\gamma = \mathcal{M}|Y, \sigma^2)$. For any fixed choice of q, we can derive the upper bound on the selected model size

$$|\widehat{\mathcal{M}}| \preceq |\mathcal{A}| + \frac{c|\mathcal{A}| \ln p_n}{-\ln q} + \frac{c'n}{-\ln q}.$$

Since we require $q_n \sim p_n^{-1}$ and $|\mathcal{A}| = o\left(\frac{n}{\ln p_n}\right)$ in Conditions A, D, it reduces to

$$|\widehat{\mathcal{M}}| = O\left(\frac{n}{\ln p_n}\right).$$

Theorem 5 (*Strong Selection Consistency Under Unknown σ^2*)**.** *Suppose that Conditions A, B, C, D, E, F hold for the partially linear model in (1) with $\ln p_n = o(n)$, and $|\mathcal{A}| = o(m_n)$ is valid, we have*

$$\Pr\left(\gamma = \mathcal{A} \mid \mathbf{Y}, |\gamma| \le m_n + |\mathcal{A}|\right) \xrightarrow{P} 1 \quad as \ n \to \infty,$$

where $m_n = cn/\ln p_n$, as defined in Condition D.

By further integrating out σ^2 and applying some inequalities, the problem reduces to intermediate steps in Theorem 4. Please refer to Section 4 for the proof.

Remark 6. As linear models are special cases of partially linear models, the proposed BSM-DP variable selection method may be directly applicable to linear models. We have also studied theoretical properties, finite sample performance and computational time of the BSM-DP under the setting of linear models and compared with BASAD [21] and SG [22]. To save space, all material related to BSM-DP for linear models are put in the supplementary material.

3. Numerical study

3.1. Simulation study

3.1.1. Simulation settings and the choice of hyperparameters

In this section, we compare the performance of the proposed method with several other existing methods including penalized methods on partial residuals and methods based on partial correlation of partial residuals. The penalized methods include the famous LASSO [25] and SCAD [8] tuned by BIC. The R packages `msgps`, `ncvreg` are used for LASSO and SCAD. Methods based on partial correlation include PC-simple algorithm on the partial residuals (PC-PR) [3] and threshold partial correlation on partial residuals (TPC-PR) [20]. Both PC-PR and TPC-PR select variables based on the magnitude of the partial correlation between the partial residuals of the response and the corresponding predictors, while the difference is on the threshold used for partial correlations. TPC-PR uses a threshold depending on the kurtosis of $\mathbf{X}$, so the normality assumption for $\mathbf{X}$ is not necessary. For TPC-PR, we also consider a fine tuning on the critical value $cT(\alpha, n, \widehat{\kappa}, m)$, where c is the tuning parameter chosen by EBIC [4]. The method is denoted as TPC-PR.EBIC.

First we need to specify the hyper-parameters $\sigma_0, \sigma_{1n}, q_n, m, \alpha_0, \beta_0$. Partially refer to the choice in [21], we use $\alpha_0 = 2$, $\beta_0 = 5$, and $\sigma_0 = 0.1$ for our proposed method (BSM-DP). The order of the difference operator is set to be $m = \lfloor 5n^{1/5} \rfloor$. Additionally, the variance for the diffusing prior is set as $\sigma_{1n} = \sqrt{\max\left\{\frac{p_n^{2.1}}{100n}, \ln(n)\right\}}$, and we choose $q_n = \Pr(\gamma_i = 1)$ such that $\Pr(\sum_{j=1}^{p_n} \gamma_j > K) = 0.1$, for a prespecified value of K. The value of K can be our preliminary guess for the size of active set, and for example, we can use the size of active set selected by LASSO. In this paper we simply set $K = 10$. In each of the following case, we allow 6000 iterations, and treat the first 3000 as burn-in samples. We report simulation results based on both MAP and MPM for our proposed method.

We fix $n = 200$, $p = 1000$ and the true active set $\mathcal{A} = (1, 2, 5, 8)$ with coefficients of $\boldsymbol{\beta}_{\mathcal{A}} = (1.5, 2.0, 2.5, 3.0)$. The error ϵ_i is drawn from standard normal distribution $\mathcal{N}(0, 1)$. Fixed design of $U_i = i/n,\ i \in \{1, \ldots, n\}$ is used with three different types of $\mathbf{X}$:

Case 1. Type I normal distribution with autoregressive covariance matrix: $\mathbf{X} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}_{ij} = \rho^{|i-j|}$.
Case 2. Type II normal distribution with compound symmetric covariance matrix: $\mathbf{X} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}_{ij} = 1$ for $i = j$ and $\boldsymbol{\Sigma}_{ij} = \rho$ for $i \neq j$.
Case 3. Type III mixture of normals: $\mathbf{X}$ is sampled from $\mathcal{N}(0, \boldsymbol{\Sigma})$ with probability 0.9 and from $\mathcal{N}(0, 9\boldsymbol{\Sigma})$ with probability 0.1, where $\boldsymbol{\Sigma}$ is the compound symmetric correlation matrix with correlation ρ.

In each case, we also consider low and high correlations at $\rho = 0.2$ and $\rho = 0.8$ separately, with two choices of the nonparametric component $f(U) = U^2$ and $f(U) = \sin(2\pi U)$. The following evaluation criteria are used for comparing methods based on 500 replications:

$p_{\mathcal{A}^c}^{\max}$ and $p_{\mathcal{A}}^{\min}$: the average for the maximal of marginal posterior probabilities on true inactive covariates, and the average for the minimal of marginal posterior probabilities on true active covariates.
$p_{\mathcal{A}=\mathcal{M}}$: the proportion of replications with the exact model being selected.
$p_{\mathcal{A}\in\mathcal{M}}$: the proportion of replications with all true active variables being selected.
p_i: the proportion of replications that the ith true active variable is selected successfully , $i \in \{1, 4\}$.
TP (true positive): the average number of true active variables selected.
FP (false positive): the average number of selected variables that are actually inactive.
ME (model error): $(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \mathrm{cov}(\mathbf{X})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$.

Note that for those existing methods, partial residuals are firstly obtained. Getting the partial residuals involves the value of $\mathrm{E}(Y|U)$, which is estimated by the local linear regression, followed by [20]. The bandwidth is chosen via plug-in methods using R package `KernelSmooth`. Afterwards, with LASSO and SCAD, we are able to estimate $\boldsymbol{\beta}$ and achieve variable selection simultaneously in the second step. While for the partial correlation methods including PC and TPC, an estimation of the active set $\widehat{\mathcal{A}}$ needs to be obtained first, then $\widehat{\boldsymbol{\beta}}$ is estimated by regressing the partial residuals on $\widehat{\mathcal{A}}$ through the least squares method. Details can be found in [20]. In this simulation, we use the posterior mean of $\boldsymbol{\beta}$ as $\widehat{\boldsymbol{\beta}}$ for the proposed method.

3.1.2. Simulation results

Tables 1–3 record the mean results from those 500 replications. *Case 1* is with the decayed autoregressive covariance matrix. Under the case with low correlation $\rho = 0.2$, all methods perform well regardless of the type of the nonparametric function. For the situation with high correlation of $\rho = 0.8$, LASSO is prone to overfit the model, with the exact model being selected only around 20% of the time.

It gets more challenging to identify true covariates under dense correlation, which is *Case 2* with the compound symmetric covariance matrix. The exact-fit rates are much lower for most of the methods as compared to that in *Case 1*. It is noteworthy that under high correlation $\rho = 0.8$, except for our proposed BSM-DP, most other methods do not perform well. LASSO consistently selects a larger model with about 16 spurious covariates on average, while SCAD tuned

Table 1
Summarized simulation results for *Case 1*: $p = 1000, n = 200$ and $\mathbf{X}$ is sampled from normal distribution with autoregressive correlation matrix. The reported values are means of different performance measures averaged over 500 replications. The methods compared include LASSO and SCAD on partial residuals tuned by BIC (LASSO.BIC, SCAD.BIC), PC-simple algorithm on the partial residuals (PC-PR), threshold partial correlation on partial residuals (TPC-PR, TPC-PR.EBIC), proposed method with model selected by MAP and MPM (BSM-DP.MAP, BSM-DP.MPM). The details of different methods and measures are provided in Section 3.1.1.

	Method	$p_{\mathcal{A}^c}^{\max}$	$p_{\mathcal{A}}^{\min}$	$p_{\mathcal{A}=\mathcal{M}}$	$p_{\mathcal{A}\in\mathcal{M}}$	p_1	p_4	TP	FP	ME
$f(u) = u^2$ $\rho = 0.2$	LASSO.BIC	0.073	1.000	0.760	1.000	1.000	1.000	4.000	0.256	0.313
	SCAD.BIC			0.932	1.000	1.000	1.000	4.000	0.082	0.026
	PC-PR			0.966	0.994	0.994	1.000	3.994	0.030	0.040
	TPC-PR			0.964	0.994	0.994	1.000	3.994	0.032	0.041
	TPC-PR.EBIC			0.994	1.000	1.000	1.000	4.000	0.010	0.027
	BSM-DP.MAP (new)			0.974	1.000	1.000	1.000	4.000	0.032	0.026
	BSM-DP.MPM (new)			0.974	1.000	1.000	1.000	4.000	0.030	0.026
$f(u) = \sin(2\pi u)$ $\rho = 0.2$	LASSO.BIC	0.074	1.000	0.738	1.000	1.000	1.000	4.000	0.288	0.334
	SCAD.BIC			0.924	1.000	1.000	1.000	4.000	0.120	0.029
	PC-PR			0.996	0.994	0.994	1.000	3.994	0.032	0.045
	TPC-PR			0.958	0.994	0.994	1.000	3.994	0.040	0.045
	TPC-PR.EBIC			0.982	1.000	1.000	1.000	4.000	0.018	0.029
	BSM-DP.MAP (new)			0.974	1.000	1.000	1.000	4.000	0.028	0.026
	BSM-DP.MPM (new)			0.970	1.000	1.000	1.000	4.000	0.032	0.026
$f(u) = u^2$ $\rho = 0.8$	LASSO.BIC	0.053	1.000	0.222	1.000	1.000	1.000	4.000	1.266	0.309
	SCAD.BIC			0.984	0.998	0.998	1.000	3.998	0.020	0.032
	PC-PR			0.644	0.652	0.768	1.000	3.652	0.094	0.344
	TPC-PR			0.652	0.660	0.774	1.000	3.660	0.092	0.338
	TPC-PR.EBIC			0.846	0.858	0.886	1.000	3.858	0.038	0.149
	BSM-DP.MAP (new)			0.990	1.000	1.000	1.000	4.000	0.010	0.023
	BSM-DP.MPM (new)			0.994	1.000	1.000	1.000	4.000	0.006	0.023
$f(u) = \sin(2\pi u)$ $\rho = 0.8$	LASSO.BIC	0.055	1.000	0.172	1.000	1.000	1.000	4.000	1.362	0.321
	SCAD.BIC			0.986	0.998	0.998	1.000	3.998	0.014	0.040
	PC-PR			0.650	0.656	0.796	1.000	3.998	0.014	0.348
	TPC-PR			0.662	0.668	0.804	1.000	3.666	0.132	0.337
	TPC-PR.EBIC			0.840	0.854	0.906	1.000	3.854	0.064	0.159
	BSM-DP.MAP (new)			0.982	1.000	1.000	1.000	4.000	0.018	0.023
	BSM-DP.MPM (new)			0.984	1.000	1.000	1.000	4.000	0.016	0.023

by BIC is prone to select a smaller model. PC and TPC work evidently slow under dense correlation, when $\rho = 0.2$, it takes around 12 h for PC, TPC and TPC-EBIC to finish one replication. More than 48 h are needed when $\rho = 0.8$, thus we mark it as stars (***) since 500 replications cannot be done in timely manner. Our proposed method (BSM-DP) gives the best results, with high exact-fit rates (above 95%) even under the high dense correlation situation.

In *Case 3*, $\mathbf{X}$ is generated from a mixture normal distribution, with a heavier tail than the normal distribution. Since PC-PR relies heavily on the normality of the covariates, it gives poor results. The updated version of TPC without assuming normality shows improvement. Our proposed method (BSM-DP) still stands out in the comparison with around 95% perfect exact-fit rates.

Overall, when correlation increases, LASSO tuned by BIC tends to overfit the model while SCAD tuned by BIC is more likely to select a smaller model. When $\mathbf{X}$ is normally distributed, PC and TPC are similar. But when the normality assumption is violated, TPC performs better than PC. The newly proposed method BSM-DP performs consistently the best, regardless of the correlation strength and distribution of $\mathbf{X}$. The exact-fit rates for all cases are all above 90%.

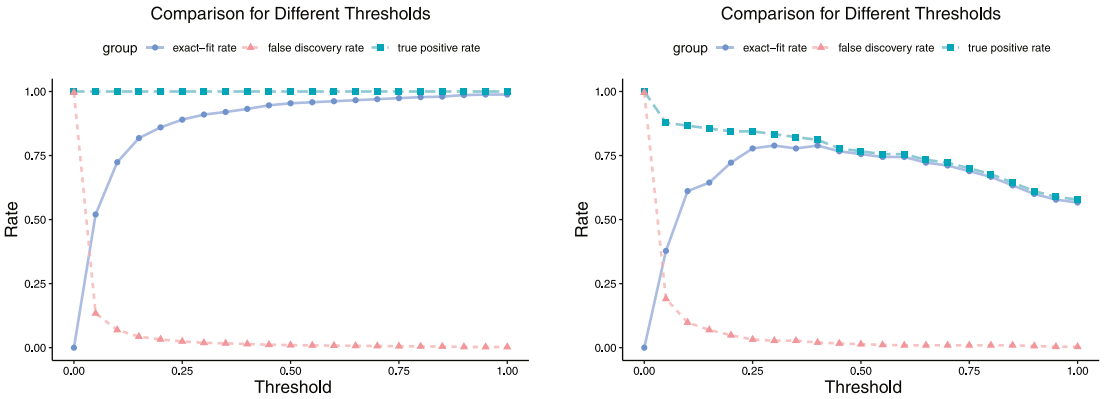
Remark 7 (*On Model Selection Procedure*)**.** In our simulation study, models selected by MAP and MPM are very similar to each other. We do not need to select the threshold with MAP. With the MPM, the jth variable is selected if its posterior probability $\Pr(\gamma_j = 1|Y) \geq 0.5$. In order to investigate the impact of different thresholds other than 0.5, we further explore one simulation setting *Case 2* with $\rho = 0.8$, and consider various threshold values from 0 to 1. The results are presented in Fig. 2. With a smaller threshold, more spurious variables are likely to enter the model, so the false discovery rate (dotted red line) is higher. While with a larger threshold which associates with a more stringent selection criterion, we have a higher chance to miss active variables. It is worth noting that although for our simulation setting with $\boldsymbol{\beta}_{\mathcal{A}} = (1.5, 2.0, 2.5, 3.0)$ shown in Fig. 2(a), the true positive rate (dashed green line) is consistently high as all active variables have marginal inclusion probabilities as 1, but generally we may expect a drop when threshold approaches to 1 for most cases as in Fig. 2(b) with lower signal $\boldsymbol{\beta}_{\mathcal{A}} = (0.8, 1.3, 1.8, 2.3)$. The exact-fit rate (solid blue line), which gives the proportion of replications with exact model being selected, is reasonably good, as long as the threshold is neither too large nor too small. Overall, 0.5 looks like a good choice.

We also compare the computation time required by different methods. Among the five methods mentioned above, LASSO and SCAD are the fastest which take about 2 min in R to finish estimation for one replication. PC and TPC are fast as well when covariate $\mathbf{X}$ has decayed covariance matrix. They become dramatically slow with high and dense correlation.

Table 2
Summarized simulation results for *Case 2*: $p = 1000, n = 200$ and $\mathbf{X}$ is sampled from normal distribution with compound symmetric correlation matrix. The reported values are means of different performance measures averaged over 500 replications. The methods compared include LASSO and SCAD on partial residuals tuned by BIC (LASSO.BIC, SCAD.BIC), PC-simple algorithm on the partial residuals (PC-PR), threshold partial correlation on partial residuals (TPC-PR, TPC-PR.EBIC), proposed method with model selected by MAP and MPM (BSM-DP.MAP, BSM-DP.MPM). The details of different methods and measures are provided in Section 3.1.1. Results under high correlation $\rho = 0.8$ are highlighted.

	Method	$p_{\mathcal{A}^c}^{\max}$	$p_{\mathcal{A}}^{\min}$	$p_{\mathcal{A}=\mathcal{M}}$	$p_{\mathcal{A}\in\mathcal{M}}$	p_1	p_4	TP	FP	ME
$f(u) = u^2$ $\rho = 0.2$	LASSO.BIC	0.077	1.000	0.220	1.000	1.000	1.000	4.000	1.652	0.509
	SCAD.BIC			0.938	1.000	1.000	1.000	4.000	0.070	0.028
	PC-PR			0.420	0.998	0.998	1.000	3.998	0.766	0.063
	TPC-PR			0.406	0.998	0.998	1.000	3.998	0.782	0.063
	TPC-PR.EBIC			0.862	0.998	0.998	1.000	3.998	0.140	0.040
	BSM-DP.MAP (new)			0.970	1.000	1.000	1.000	4.000	0.030	0.024
	BSM-DP.MPM (new)			0.970	1.000	1.000	1.000	4.000	0.024	0.024
$f(u) = \sin(2\pi u)$ $\rho = 0.2$	LASSO.BIC	0.087	1.000	0.202	1.000	1.000	1.000	4.000	1.746	0.515
	SCAD.BIC			0.950	1.000	1.000	1.000	4.000	0.054	0.027
	PC-PR			0.368	0.984	0.984	1.000	3.984	0.870	0.098
	TPC-PR			0.358	0.986	0.986	1.000	3.986	0.884	0.095
	TPC-PR.EBIC			0.826	0.990	0.990	1.000	3.990	0.208	0.060
	BSM-DP.MAP (new)			0.970	1.000	1.000	1.000	4.000	0.036	0.032
	BSM-DP.MPM (new)			0.976	1.000	1.000	1.000	4.000	0.026	0.033
$f(u) = u^2$ $\rho = 0.8$	LASSO.BIC	0.105	1.000	**0.000**	1.000	1.000	1.000	4.000	16.174	0.390
	SCAD.BIC			**0.386**	0.386	0.412	1.000	3.386	0.000	0.585
	PC-PR			***	***	***	***	***	***	***
	TPC-PR			***	***	***	***	***	***	***
	TPC-PR.EBIC			***	***	***	***	***	***	***
	BSM-DP.MAP (new)			**0.942**	1.000	1.000	1.000	4.000	0.068	0.034
	BSM-DP.MPM (new)			**0.956**	1.000	1.000	1.000	4.000	0.048	0.037
$f(u) = \sin(2\pi u)$ $\rho = 0.8$	LASSO.BIC	0.093	1.000	**0.000**	1.000	1.000	1.000	4.000	16.114	0.401
	SCAD.BIC			**0.426**	0.426	0.448	1.000	3.426	0.000	0.574
	PC-PR			***	***	***	***	***	***	***
	TPC-PR			***	***	***	***	***	***	***
	TPC-PR.EBIC			***	***	***	***	***	***	***
	BSM-DP.MAP (new)			**0.970**	1.000	1.000	1.000	4.000	0.036	0.032
	BSM-DP.MPM (new)			**0.976**	1.000	1.000	1.087	4.000	0.026	0.033

*** Stands for the cases when a single replication takes more than 48 h. So 500 replications cannot be done in timely manner.



(a) $\boldsymbol{\beta}_{\mathcal{A}} = (1.5, 2.0, 2.5, 3.0)$ **(b)** $\boldsymbol{\beta}_{\mathcal{A}} = (0.8, 1.3, 1.8, 2.3)$

Fig. 2. Impact of model selection performance using different threshold values for the posterior probability. The criteria displayed to measure performance include exact-fit rate (solid blue line), false discovery rate (dotted red line) and true positive rate (dashed green line). A range of threshold values from 0 and 1 are used to plot the curve for each criterion. Two signal values with different strengths are considered: $\boldsymbol{\beta}_{\mathcal{A}} = (1.5, 2.0, 2.5, 3.0)$ and $\boldsymbol{\beta}_{\mathcal{A}} = (0.8, 1.3, 1.8, 2.3)$.

More than 48 h is needed for PC using R to finish one replication when the covariance of $\mathbf{X}$ is compound symmetric with $\rho = 0.8$. It may get worse with higher dimensional covariates and larger active groups, since the computational time for both PC and TPC grows polynomially with them. Time is recorded based on Macbook Pro early 2015 with 2.7 GHZ, Intel i5 and 8 GB.

On the contrary, the computation burden for the newly proposed Bayesian method (BSM-DP) is moderate. Among all simulation settings, the slowest one takes about 12 min to finish 6000 iterations for one replication using Julia

Table 3
Summarized simulation results for *Case 3*: $p = 1000, n = 200$ and $\mathbf{X}$ is sampled from mixture of normals with compound symmetric matrix. The reported values are means of different performance measures averaged over 500 replications. The methods compared include LASSO and SCAD on partial residuals tuned by BIC (LASSO.BIC, SCAD.BIC), PC-simple algorithm on the partial residuals (PC-PR), threshold partial correlation on partial residuals (TPC-PR, TPC-PR.EBIC), proposed method with model selected by MAP and MPM (BSM-DP.MAP, BSM-DP.MPM). The details of different methods and measures are provided in Section 3.1.1. Results under high correlation $\rho = 0.8$ are highlighted.

	Method	$p^{\max}_{\mathcal{A}^c}$	$p^{\min}_{\mathcal{A}}$	$p_{\mathcal{A}=\mathcal{M}}$	$p_{\mathcal{A}\in\mathcal{M}}$	p_1	p_4	TP	FP	ME
$f(u) = u^2$ $\rho = 0.2$	LASSO.BIC	0.046	1.000	0.156	1.000	1.000	1.000	4.000	2.434	0.475
	SCAD.BIC			0.994	1.000	1.000	1.000	4.000	0.008	0.022
	PC-PR			0.476	0.690	0.724	1.000	3.676	1.304	0.927
	TPC-PR			0.574	0.574	0.636	1.000	3.552	0.622	1.252
	TPC-PR.EBIC			0.680	0.694	0.726	1.000	3.680	0.612	0.891
	BSM-DP.MAP (new)			0.984	1.000	1.000	1.000	4.000	0.016	0.015
	BSM-DP.MPM (new)			0.986	1.000	1.000	1.000	4.000	0.014	0.015
$f(u) = \sin(2\pi u)$ $\rho = 0.2$	LASSO.BIC	0.050	1.000	0.228	1.000	1.000	1.000	4.000	2.194	0.449
	SCAD.BIC			0.990	1.000	1.000	1.000	4.000	0.012	0.018
	PC-PR			0.572	0.768	0.802	1.000	3.776	0.982	0.670
	TPC-PR			0.618	0.622	0.682	1.000	3.602	0.544	1.125
	TPC-PR.EBIC			0.748	0.768	0.804	1.000	3.764	0.438	0.667
	BSM-DP.MAP (new)			0.986	1.000	1.000	1.000	4.000	0.016	0.014
	BSM-DP.MPM (new)			0.988	1.000	1.000	1.000	4.000	0.012	0.014
$f(u) = u^2$ $\rho = 0.8$	LASSO.BIC	0.092	1.000	**0.000**	0.998	0.998	1.000	3.998	14.060	0.395
	SCAD.BIC			**0.380**	0.382	0.432	1.000	3.374	0.120	0.652
	PC-PR			***	***	***	***	***	***	***
	TPC-PR			***	***	***	***	***	***	***
	TPC-PR.EBIC			***	***	***	***	***	***	***
	BSM-DP.MAP (new)			**0.962**	1.000	1.000	1.000	4.000	0.042	0.025
	BSM-DP.MPM (new)			**0.964**	1.000	1.000	1.000	4.000	0.040	0.025
$f(u) = \sin(2\pi u)$ $\rho = 0.8$	LASSO.BIC	0.082	1.000	**0.000**	0.996	0.996	1.000	3.996	14.002	0.418
	SCAD.BIC			**0.422**	0.426	0.492	1.000	3.416	0.022	0.672
	PC-PR			***	***	***	***	***	***	***
	TPC-PR			***	***	***	***	***	***	***
	TPC-PR.EBIC			***	***	***	***	***	***	***
	BSM-DP.MAP (new)			**0.960**	1.000	1.000	1.000	4.000	0.042	0.020
	BSM-DP.MPM (new)			**0.970**	1.000	1.000	1.000	4.000	0.032	0.020

*** Stands for the cases when a single replication takes more than 48 h. So 500 replications cannot be done in timely manner.

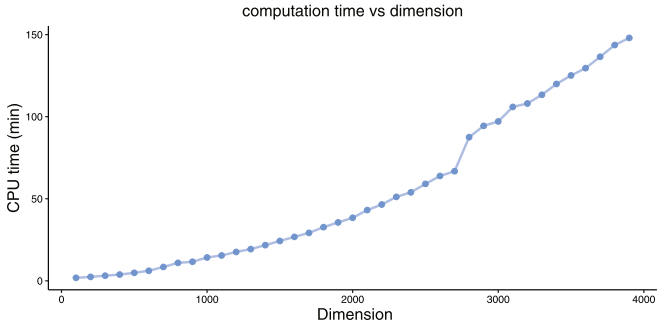


Fig. 3. Change in computational time (in minutes) when the dimension of covariates increases from 100 to 4000. The CPU time is estimated by the median computation time consumed among 10 replications for each dimension setting.

0.6. As discussed in Remark 1 Section 2.1, the Bayesian subset modeling is scalable, and the computational time only grows approximately linearly with the dimension of covariates. Based on the estimation procedure, the computational complexity for each iteration is $n(p \vee |\boldsymbol{\gamma}|^2 \vee n^2)$, where $|\boldsymbol{\gamma}|$ is the current active model size. To explore the change in the computation time for different (and especially higher) dimensions of covariates, we record the CPU time to finish 6000 iterations with various p from 100 to 4000, in the simulation setting *Case 1* $\rho = 0.8$. The result is presented in Fig. 3. The computation time increases nearly linearly with the dimension p. It is not perfectly linear since the number of iterations until convergence seems to grow with p. We notice that for small p (e.g. $p < 500$), it usually only takes a few iterations to converge and ends up with a small $|\boldsymbol{\gamma}|$, but it often requires more iterations when p gets larger. There is one place in the above plot which shows a large jump. This might relate to the caching limit on the hardware, especially when p is large.

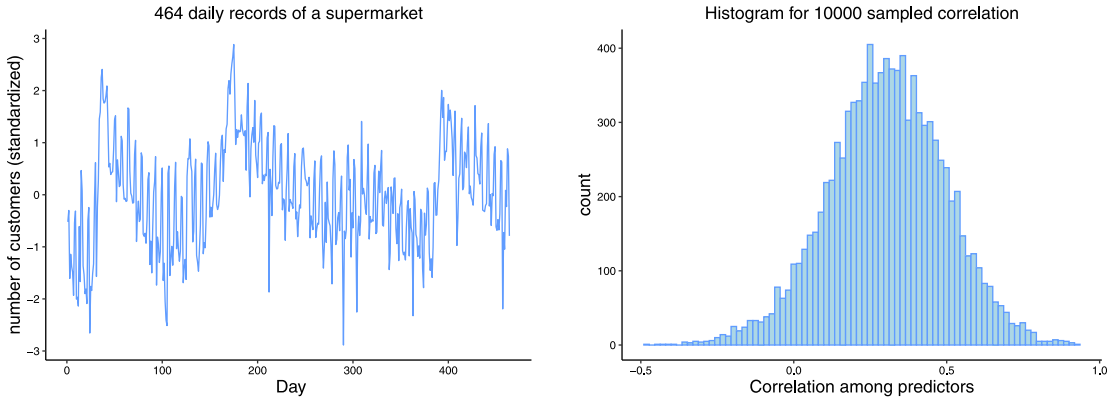


Fig. 4. Visualizations to display features of the (standardized) supermarket data set. The plot on the left gives the trend of daily number of customers entering a supermarket for 464 days. The histogram on the right describes the distribution of the correlation among sampled predictors.

Table 4
Comparisons of the resulting model size and the mean squared errors by different methods. The values in the table are the means and the corresponding standard errors (in the parenthesis) over the 100 replications. The methods compared are LASSO and SCAD on partial residuals tuned by BIC (SIS-LASSO.BIC, SIS-SCAD.BIC), PC-simple algorithm on the partial residuals (SIS-PC-PR), threshold partial correlation on partial residuals (SIS-TPC-PR, SIS-TPC-PR.EBIC), proposed method with model selected by MPM tuned by EBIC (SIS-BSM-DP).

Method	Model size (s.e.)	MSE on training set (s.e.)	MSE on testing set (s.e.)
SIS-LASSO.BIC	28.80 (3.32)	0.0575 (0.0030)	0.0836 (0.0120)
SIS-SCAD.BIC	15.75 (5.44)	0.0647 (0.0062)	0.0907 (0.0143)
SIS-PC-PR	12.94 (1.03)	0.0497 (0.0031)	0.0847 (0.0112)
SIS-TPC-PR	9.81 (0.92)	0.0540 (0.0034)	0.0860 (0.0109)
SIS-TPC-PR.EBIC	8.50 (0.98)	0.0559 (0.0038)	0.0867 (0.0117)
SIS-BSM-DP.EBIC (new)	7.95 (1.91)	0.0610 (0.0055)	0.0864 (0.0122)

3.2. A real data example — supermarket data analysis

In this section, the proposed method is applied to analyze a supermarket data set mentioned in [5,20,26]. The data set contains $n = 464$ daily records of the number of customers, which is the response variable and the sales of $p = 6398$ products which are predictors. Both the response variable and the predictors are standardized to have zero mean and unit variance. The plot on the left of Fig. 4 shows the relationship between the number of customers and the days. The periodicity of Y is obvious, thus it is reasonable to model Y with a PLM, which takes time variation into account. A covariate $U_i = i/n$ is introduced to represent time. To check the correlation among predictors, we plot the histogram of the sampled correlation in Fig. 4., which shows some moderate correlation. We randomly select 75% of the observations $(\mathbf{X}_i, Y_i, U_i), i \in \{1, \ldots, 464\}$ as the training set and keep the remaining 25% as the testing set. The PLM is fitted and variables are selected using the training data set, then the mean squared errors of the selected model on the testing data set are calculated to evaluate the model fit. This procedure is repeated 100 times, and Table 4 summarizes the average size of the selected models and the mean squared errors.

To take the dimension down to a moderate scale, we first apply the SIS [9] on the partial residual to only keep the top 2000 predictors as the set subjected to variable selection. We also implement the LASSO.BIC, SCAD.BIC, PC-PR, TPC-PR and TPC-PR-EBIC on the same data set as comparisons. The choice of the hyperparameters for BSM-DP is the same as the set up in the simulation. We complete 10 000 iterations, with the first 6000 as burn-in samples, and rank covariates by their marginal posterior probabilities $\Pr(\gamma_j = 1|Y)$. The candidate set is further selected by EBIC.

After obtaining partial residuals, with LASSO and SCAD, we are able to obtain $\widehat{\boldsymbol{\beta}}$ and select variables simultaneously. But for PC, TPC, and BSM-DP, an estimation of the active set $\widehat{\mathcal{A}}$ is first obtained, then $\widehat{\boldsymbol{\beta}}$ is estimated by regressing the partial residuals on $\widehat{\mathcal{A}}$ through OLS. Since we derive the theoretical property of $\boldsymbol{\gamma}$, we only use BSM-DP to select the set of active covariates, $\widehat{\boldsymbol{\beta}}$ is also obtained by regressing the partial residuals on the selected covariate set.

As shown in Table 4, LASSO gives the most conservative result with an average size of selected models as 28.80. On the other hand, the MSE is much smaller on the training than that on the testing. This suggests that the LASSO may be overfitting. SCAD selects smaller models with size 15.75 on average, but having large error on testing set. PC, TPC, TPC.EBIC and the newly proposed BSM-DP all select even smaller models with the average size less than 10. Among all, our proposed BSM-DP selects the smallest number of covariates with very similar value of the mean squared error on the testing data set. Fig. 5 illustrates a comparison on the estimated nonparametric function obtained by different methods.

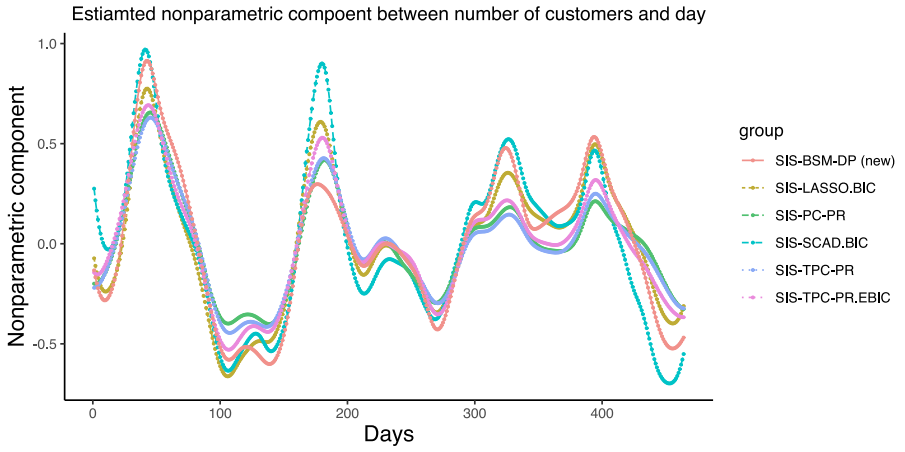


Fig. 5. The estimates of the nonparametric function for the supermarket data set by different methods including: LASSO and SCAD on partial residuals tuned by BIC (SIS-LASSO.BIC, SIS-SCAD.BIC), PC-simple algorithm on the partial residuals (SIS-PC-PR), threshold partial correlation on partial residuals (SIS-TPC-PR, SIS-TPC-PR.EBIC), and proposed method with model selected by MPM tuned by EBIC (SIS-BSM-DP.EBIC).

4. Technical proofs

This section includes technical proofs for Lemmas 1–3 and Theorems 4–5.

Proof of Lemma 1. Note that, as $q_n \longrightarrow 0$, which is stated in Condition A, we first write out the posterior distribution for parameters as

$$\begin{aligned}\Pr(\boldsymbol{\gamma}=\mathcal{M},\boldsymbol{\beta},\sigma,\boldsymbol{\alpha}|\mathbf{Y}) \propto& \pi(\mathbf{Y}|\boldsymbol{\alpha},\boldsymbol{\gamma},\boldsymbol{\beta},\sigma)\pi(\boldsymbol{\alpha}|\sigma)\pi(\boldsymbol{\beta}|\boldsymbol{\gamma},\sigma)\pi(\boldsymbol{\gamma})\pi(\sigma^2)\\ \propto& \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{Y}-\boldsymbol{\alpha}-\mathbf{X}_{\mathcal{M}}\boldsymbol{\beta}_{\mathcal{M}})^{\top}(\mathbf{Y}-\boldsymbol{\alpha}-\mathbf{X}_{\mathcal{M}}\boldsymbol{\beta}_{\mathcal{M}})-\frac{1}{2\sigma^2}\boldsymbol{\alpha}^{\top}\boldsymbol{\Sigma}_{0n}^{+}\boldsymbol{\alpha}\right\}\\ & q_n^{|\mathcal{M}|}\sigma^{-2n-p_n}\sigma_{1n}^{-|\mathcal{M}|}\sigma_0^{-(p_n-|\mathcal{M}|)}\exp\left\{-\frac{1}{2\sigma_{1n}^2\sigma^2}\boldsymbol{\beta}_{\mathcal{M}}^{\top}\boldsymbol{\beta}_{\mathcal{M}}-\frac{1}{2\sigma_0^2\sigma^2}\boldsymbol{\beta}_{\mathcal{M}^c}^{\top}\boldsymbol{\beta}_{\mathcal{M}^c}\right\}\pi(\sigma^2).\end{aligned}$$

We first integrate out $\boldsymbol{\alpha}$, and it follows that

$$\begin{aligned}\Pr(\boldsymbol{\gamma}=\mathcal{M},\boldsymbol{\beta},\sigma|\mathbf{Y}) \propto& q_n^{|\mathcal{M}|}\sigma^{-n-p_n}\sigma_{1n}^{-|\mathcal{M}|}\sigma_0^{-(p_n-|\mathcal{M}|)}\exp\left\{-\frac{1}{2\sigma_{1n}^2\sigma^2}\boldsymbol{\beta}_{\mathcal{M}}^{\top}\boldsymbol{\beta}_{\mathcal{M}}-\frac{1}{2\sigma_0^2\sigma^2}\boldsymbol{\beta}_{\mathcal{M}^c}^{\top}\boldsymbol{\beta}_{\mathcal{M}^c}\right\}\\ & \exp\left[-\frac{1}{2\sigma^2}\left(\mathbf{Y}-\mathbf{X}_{\mathcal{M}}\boldsymbol{\beta}_{\mathcal{M}}\right)^{\top}\left\{\mathbf{I}_n-(\boldsymbol{\Sigma}_{0n}^{+}+\mathbf{I}_n)^{-1}\right\}\left(\mathbf{Y}-\mathbf{X}_{\mathcal{M}}\boldsymbol{\beta}_{\mathcal{M}}\right)\right]|\boldsymbol{\Sigma}_{0n}^{+}+\mathbf{I}_n|^{-1/2}\pi(\sigma^2).\end{aligned}$$

Denote $\boldsymbol{\Sigma}_{1n}=\mathbf{I}_n-(\boldsymbol{\Sigma}_{0n}^{+}+\mathbf{I}_n)^{-1}$. It follows by integrating out $\boldsymbol{\beta}$ that

$$\Pr(\boldsymbol{\gamma}=\mathcal{M},\sigma|\mathbf{Y}) \propto q_n^{|\mathcal{M}|}\sigma^{-n}\sigma_{1n}^{-|\mathcal{M}|}|\boldsymbol{\Sigma}_{0n}^{+}+\mathbf{I}_n|^{-1/2}|\mathbf{X}_{\mathcal{M}}^{\top}\boldsymbol{\Sigma}_{1n}\mathbf{X}_{\mathcal{M}}+\mathbf{I}_{|\mathcal{M}|}/\sigma_{1n}^2|^{-1/2}\exp\left(-\frac{1}{2\sigma^2}R_{\mathcal{M}}\right)\pi(\sigma^2), \tag{5}$$

where

$$R_{\mathcal{M}}=\mathbf{Y}^{\top}\left\{\boldsymbol{\Sigma}_{1n}-\boldsymbol{\Sigma}_{1n}\mathbf{X}_{\mathcal{M}}\left(\mathbf{X}_{\mathcal{M}}^{\top}\boldsymbol{\Sigma}_{1n}\mathbf{X}_{\mathcal{M}}+\mathbf{I}_{|\mathcal{M}|}/\sigma_{1n}^2\right)^{-1}\mathbf{X}_{\mathcal{M}}^{\top}\boldsymbol{\Sigma}_{1n}\right\}\mathbf{Y}.$$

Let $\mathbf{X}^{*}=C^{1/2}\mathbf{D}\mathbf{X}$, $\lambda_1^{*}=\min_{\mathcal{M}:|\mathcal{M}|\le m_n+|\mathcal{A}|}\lambda_{\min}\left(\frac{1}{n-m}X_{\mathcal{M}}^{*\top}X_{\mathcal{M}}^{*}\right)$, $\lambda_2^{*}=\max_{\mathcal{M}:\mathcal{M}\subseteq\mathcal{A}}\lambda_{\max}\left(\frac{1}{n-m}X_{\mathcal{M}}^{*\top}X_{\mathcal{M}}^{*}\right)$. As $\mathbf{D}^{\top}\mathbf{D}\longrightarrow\begin{pmatrix}\mathbf{0}_{\mathbf{m}\times\mathbf{m}} & \mathbf{0}_{\mathbf{m}\times(\mathbf{n}-\mathbf{m})}\\ \mathbf{0}_{(\mathbf{n}-\mathbf{m})\times\mathbf{m}} & \mathbf{I}_{n-m}\end{pmatrix}$, for any model $\mathcal{M}$, $X_{\mathcal{M}}^{*\top}X_{\mathcal{M}}^{*}$ is asymptotically equal to $CX_{\mathcal{M}'}^{\top}X_{\mathcal{M}'}$, where $\mathcal{M}'$ is the subset of $\mathcal{M}$ by taking last $n-m$ elements. By Conditions C, E, it can be shown that $\lambda_1^{*}\sim\lambda_1$ and $\lambda_2^{*}\sim\lambda_2$. Now we put bound on $T_{\mathcal{M}}/T_{\mathcal{A}}$. For any model $\mathcal{M}$:

$$\begin{aligned}\frac{T_{\mathcal{M}}}{T_{\mathcal{M}\wedge\mathcal{A}}}&=\left|\mathbf{I}_{n-m}+\sigma_{1n}^2\mathbf{X}_{\mathcal{M}}^{*}\mathbf{X}_{\mathcal{M}}^{*\top}\right|^{-\frac{1}{2}}\left|\mathbf{I}_{n-m}+\sigma_{1n}^2\mathbf{X}_{\mathcal{M}\wedge\mathcal{A}}^{*}\mathbf{X}_{\mathcal{M}\wedge\mathcal{A}}^{*\top}\right|^{\frac{1}{2}}\preceq(n\sigma_{1n}^2\lambda_1^{*})^{-|\mathcal{M}|/2}(n\sigma_{1n}^2\lambda_2^{*})^{|\mathcal{M}\wedge\mathcal{A}|/2}\\ \frac{T_{\mathcal{M}\wedge\mathcal{A}}}{T_{\mathcal{A}}}&=|\mathbf{I}_{n-m}+\sigma_{1n}^2\mathbf{X}_{\mathcal{M}\wedge\mathcal{A}}^{*}\mathbf{X}_{\mathcal{M}\wedge\mathcal{A}}^{*\top}|^{-\frac{1}{2}}|\mathbf{I}_{n-m}+\sigma_{1n}^2\mathbf{X}_{\mathcal{A}}^{*}\mathbf{X}_{\mathcal{A}}^{*\top}|^{\frac{1}{2}}\preceq|\mathbf{I}_{n-m}+\sigma_{1n}^2\mathbf{X}_{\mathcal{A}\wedge\mathcal{M}^c}^{*}\mathbf{X}_{\mathcal{A}\wedge\mathcal{M}^c}^{*\top}|^{\frac{1}{2}}\preceq(n\sigma_{1n}^2\lambda_2^{*})^{|\mathcal{A}\wedge\mathcal{M}^c|/2}.\end{aligned}$$

Thus, $\frac{T_{\mathcal{M}}}{T_{\mathcal{A}}} \preceq (n\sigma_{1n}^2\lambda_1)^{-(|\mathcal{M}|-|\mathcal{A}|)/2}\left(\frac{\lambda_2}{\lambda_1}\right)^{|\mathcal{A}|/2}$, so

$$\begin{aligned} PR(\mathcal{M},\mathcal{A}) = \frac{\Pr(\gamma=\mathcal{M}|\mathbf{Y},\sigma)}{\Pr(\gamma=\mathcal{A}|\mathbf{Y},\sigma)} &= q_n^{|\mathcal{M}|-|\mathcal{A}|}\frac{T_{\mathcal{M}}}{T_{\mathcal{A}}}\exp\left\{-\frac{1}{2\sigma^2}(R_{\mathcal{M}}-R_{\mathcal{A}})\right\} \\ &\preceq q_n^{|\mathcal{M}|-|\mathcal{A}|}(n\sigma_{1n}^2\lambda_1)^{-(|\mathcal{M}|-|\mathcal{A}|)/2}\left(\frac{\lambda_2}{\lambda_1}\right)^{|\mathcal{A}|/2}\exp\left\{-\frac{1}{2\sigma^2}(R_{\mathcal{M}}-R_{\mathcal{A}})\right\} \\ &\preceq p_n^{-1.5\delta(|\mathcal{M}|-|\mathcal{A}|)+0.5\kappa}\exp\left\{-\frac{1}{2\sigma^2}(R_{\mathcal{M}}-R_{\mathcal{A}})\right\}. \quad \square \end{aligned}$$

Proof of Lemma 2. By Condition A, if $\boldsymbol{\Sigma}_{0n}$ is taken to be $\boldsymbol{\Sigma}_{0n} = \left\{(\mathbf{I}_n - C\mathbf{D}^\top\mathbf{D})^{-1} - \mathbf{I}_n\right\}^+$, since $\boldsymbol{\Sigma}_{1n}$ is defined as $\boldsymbol{\Sigma}_{1n} = \mathbf{I}_n - (\boldsymbol{\Sigma}_{0n}^+ + \mathbf{I}_n)^{-1}$, so we have

$$\boldsymbol{\Sigma}_{1n} = C\mathbf{D}^\top\mathbf{D},$$

where $\mathbf{D} \in \mathbb{R}^{(n-m)\times n}$ is the difference matrix defined in (2) and $0 < C \le \min\left\{1, 1/\lambda_{\max}(\mathbf{D}^\top\mathbf{D})\right\}$ is a constant, thus $\boldsymbol{\Sigma}_{\alpha n} = \mathbf{I}_n - \boldsymbol{\Sigma}_{1n}$ and $\boldsymbol{\Sigma}_{0n} = (\boldsymbol{\Sigma}_{\alpha n}^{-1} - \mathbf{I}_n)^+$ are semi-positive definite. By taking difference operation on each side, we have

$$\mathbf{Y}^* = \mathbf{X}^* + \boldsymbol{\delta} + \boldsymbol{\omega},$$

where $\mathbf{Y}^* = C^{1/2}\mathbf{D}\mathbf{Y} \in \mathbb{R}^{n-m}$, $\mathbf{X}^* = C^{1/2}\mathbf{D}\mathbf{X} \in \mathbb{R}^{(n-m)\times p_n}$, $\boldsymbol{\delta} = C^{1/2}\mathbf{D}\boldsymbol{\alpha} \in \mathbb{R}^{n-m}$, $\boldsymbol{\omega} = C^{1/2}\mathbf{D}\boldsymbol{\epsilon} \in \mathbb{R}^{n-m}$. The projection matrix is defined as $\mathbf{P}_{\mathcal{M}} = \mathbf{X}_{\mathcal{M}}^*(\mathbf{X}_{\mathcal{M}}^{*\top}\mathbf{X}_{\mathcal{M}}^*)^{-1}\mathbf{X}_{\mathcal{M}}^{*\top}$ and furthermore we denote $\mathbf{Q}_{\mathcal{M}} = \mathbf{I}_{n-m} - \mathbf{P}_{\mathcal{M}}$. Under the true model, we have $\mathbf{Y} = \boldsymbol{\alpha} + \mathbf{X}_{\mathcal{A}}\boldsymbol{\beta}_{\mathcal{A}} + \boldsymbol{\epsilon}$, then for any model containing the true model $\mathcal{A} \subseteq \mathcal{M}$,

$$\begin{aligned} R_{\mathcal{M}}^* &= \mathbf{Y}^\top\left\{\boldsymbol{\Sigma}_{1n} - \boldsymbol{\Sigma}_{1n}\mathbf{X}_{\mathcal{M}}\left(\mathbf{X}_{\mathcal{M}}^\top\boldsymbol{\Sigma}_{1n}\mathbf{X}_{\mathcal{M}}\right)^{-1}\mathbf{X}_{\mathcal{M}}^\top\boldsymbol{\Sigma}_{1n}\right\}\mathbf{Y} = \mathbf{Y}^{*\top}(\mathbf{I}_{n-m} - \mathbf{P}_{\mathcal{M}})\mathbf{Y}^* \\ &= (\boldsymbol{\delta}+\boldsymbol{\omega})^\top\mathbf{Q}_{\mathcal{M}}(\boldsymbol{\delta}+\boldsymbol{\omega}) = \boldsymbol{\omega}^\top\mathbf{Q}_{\mathcal{M}}\boldsymbol{\omega} + 2\boldsymbol{\omega}^\top\mathbf{Q}_{\mathcal{M}}\boldsymbol{\delta} + \boldsymbol{\delta}^\top\mathbf{Q}_{\mathcal{M}}\boldsymbol{\delta}. \end{aligned}$$

It suffices to show $\boldsymbol{\delta}^\top\mathbf{Q}_{\mathcal{M}}\boldsymbol{\delta} \longrightarrow 0$ a.s., $\boldsymbol{\omega}^\top\mathbf{Q}_{\mathcal{M}}\boldsymbol{\delta} \longrightarrow 0$ a.s. and $\sigma^{-2}\boldsymbol{\omega}^\top\mathbf{Q}_{\mathcal{M}}\boldsymbol{\omega} \sim C\chi^2_{n-m-|\mathcal{M}|}$ a.s..

Step 1: To show that $\boldsymbol{\delta}^\top\mathbf{Q}_{\mathcal{M}}\boldsymbol{\delta} \longrightarrow 0$ a.s.: note that

$$0 \le \boldsymbol{\delta}^\top\mathbf{Q}_{\mathcal{M}}\boldsymbol{\delta} = \|\boldsymbol{\delta}\|_2^2 - \|\mathbf{P}_{\mathcal{M}}\boldsymbol{\delta}\|_2^2 \le \|\boldsymbol{\delta}\|_2^2,$$

where $\delta_i = C^{1/2}\sum_{t=1}^{m+1} d_t f(U_{i+m+1-t})$. Thus by $\sum d_i = 0$, $\sum d_i^2 = 1$, we have

$$\begin{aligned} \delta_i^2 &= C\left\{\sum_{t=1}^{m+1} d_t f(U_{i+m+1-t})\right\}^2 \\ &= C\left[d_2\left\{f(U_{i+m-1}) - f(U_{i+m})\right\} + d_3\left\{f(U_{i+m-2}) - f(U_{i+m})\right\} + \cdots + d_{m+1}\left\{f(U_i) - f(U_{i+m})\right\}\right]^2. \end{aligned}$$

By Cauchy–Schwarz inequality and Condition F, for any $f(\cdot) \in \Lambda^k(M)$,

$$\delta_i^2 \le C(1-d_1^2)\left(\sum_{t=1}^{m} M^2\|U_{i+m-t} - U_{i+m}\|^{2(k\wedge 1)}\right) \le O\left(m^{-1}\right)m^{1+2(k\wedge 1)}U_\infty^{2(k\wedge 1)} = O\left((mU_\infty)^{2(k\wedge 1)}\right).$$

By Conditions D, E, F, $1 + (c_2 - c_1)2(k \wedge 1) < 0$, it can be shown that $\|\boldsymbol{\delta}\|_2^2 = O\left(n(mU_\infty)^{2(k\wedge 1)}\right) \longrightarrow 0$. Then $\boldsymbol{\delta}^\top\mathbf{Q}_{\mathcal{M}}\boldsymbol{\delta} \longrightarrow 0$ a.s..

Step 2: To show that $\boldsymbol{\omega}^\top\mathbf{Q}_{\mathcal{M}}\boldsymbol{\delta} \longrightarrow 0$ a.s.: under fixed design, we have

$$\boldsymbol{\omega}^\top\mathbf{Q}_{\mathcal{M}}\boldsymbol{\delta} \mid \boldsymbol{\delta} \sim \mathcal{N}(0, C\sigma^2\boldsymbol{\delta}^\top\mathbf{Q}_{\mathcal{M}}\mathbf{D}\mathbf{D}^\top\mathbf{Q}_{\mathcal{M}}\boldsymbol{\delta}).$$

Since $\mathbf{D}\mathbf{D}^\top \longrightarrow \mathbf{I}_{n-m}$ and $\boldsymbol{\delta}^\top\mathbf{Q}_{\mathcal{M}}\boldsymbol{\delta} \longrightarrow 0$ a.s. from Step 1, so $\boldsymbol{\omega}^\top\mathbf{Q}_{\mathcal{M}}\boldsymbol{\delta} \longrightarrow 0$ a.s.

Step 3: To show that $\sigma^{-2}\boldsymbol{\omega}^\top\mathbf{Q}_{\mathcal{M}}\boldsymbol{\omega} \sim C\chi^2_{n-m-|\mathcal{M}|}$ a.s.: by Condition E, $\boldsymbol{\omega}^\top\mathbf{Q}_{\mathcal{M}}\boldsymbol{\omega} = C\boldsymbol{\epsilon}^\top(\mathbf{D}^\top\mathbf{Q}_{\mathcal{M}}\mathbf{D})\boldsymbol{\epsilon}$. Let $J = (\mathbf{0}_{(\mathbf{n-m})\times\mathbf{m}}, \mathbf{I}_{\mathbf{n-m}}) \in \mathbb{R}^{(\mathbf{n-m})\times\mathbf{n}}$, then $\boldsymbol{\omega} = C^{1/2}\mathbf{D}\boldsymbol{\epsilon} = C^{1/2}J\boldsymbol{\epsilon} + C^{1/2}(\mathbf{D}-J)\boldsymbol{\epsilon} = \boldsymbol{\omega}_1 + \boldsymbol{\omega}_2$. Since $\mathbf{D} \longrightarrow J$, so $\boldsymbol{\omega}_2$ is negligible as compared to $\boldsymbol{\omega}_1$ as n goes to infinity. Furthermore $\boldsymbol{\omega}_1^\top\mathbf{Q}_{\mathcal{M}}\boldsymbol{\omega}_1 = C\boldsymbol{\epsilon}^\top J^\top\mathbf{Q}_{\mathcal{M}}J\boldsymbol{\epsilon} \sim C\sigma^2\chi^2_{n-m-|\mathcal{M}|}$, therefore $\sigma^{-2}\boldsymbol{\omega}^\top\mathbf{Q}_{\mathcal{M}}\boldsymbol{\omega} \sim C\chi^2_{n-m-|\mathcal{M}|}$ a.s..

Overall we have $R_{\mathcal{M}}^* \sim C\sigma^2\chi^2_{n-m-|\mathcal{M}|}$ a.s.. Similarly, write $R_{\mathcal{A}}^* - R_{\mathcal{M}}^* = \boldsymbol{\omega}^\top(\mathbf{Q}_{\mathcal{A}} - \mathbf{Q}_{\mathcal{M}})\boldsymbol{\omega} + 2\boldsymbol{\omega}^\top(\mathbf{Q}_{\mathcal{A}} - \mathbf{Q}_{\mathcal{M}})\boldsymbol{\delta} + \boldsymbol{\delta}^\top(\mathbf{Q}_{\mathcal{A}} - \mathbf{Q}_{\mathcal{M}})\boldsymbol{\delta}$. It can be proven that the second and third terms are almost surely 0, so $R_{\mathcal{A}}^* - R_{\mathcal{M}}^* \sim \boldsymbol{\omega}^\top(\mathbf{Q}_{\mathcal{A}} - \mathbf{Q}_{\mathcal{M}})\boldsymbol{\omega} \sim C\sigma^2\chi^2_{|\mathcal{M}|-|\mathcal{A}|}$. $\square$

Proof of Lemma 3. We first prove part (i). Note that $R_{\mathcal{A}} = \mathbf{Y}^{*\top}\left\{\mathbf{I}_{n-m} - \mathbf{X}^*_{\mathcal{A}}\left(\frac{1}{\sigma^2_{1n}}\mathbf{I} + \mathbf{X}^{*\top}_{\mathcal{A}}\mathbf{X}^*_{\mathcal{A}}\right)^{-1}\mathbf{X}^{*\top}_{\mathcal{A}}\right\}\mathbf{Y}^*$, $R^*_{\mathcal{A}} = \mathbf{Y}^{*\top}\left\{\mathbf{I}_{n-m} - \mathbf{X}^*_{\mathcal{A}}\left(\mathbf{X}^{*\top}_{\mathcal{A}}\mathbf{X}^*_{\mathcal{A}}\right)^{-1}\mathbf{X}^{*\top}_{\mathcal{A}}\right\}\mathbf{Y}^*$, thus

$$\begin{aligned}0 \le R_{\mathcal{A}} - R^*_{\mathcal{A}} &= \mathbf{Y}^{*\top}\mathbf{X}^*_{\mathcal{A}}\left\{(\mathbf{X}^{*\top}_{\mathcal{A}}\mathbf{X}^*_{\mathcal{A}})^{-1} - \left(\frac{1}{\sigma^2_{1n}}\mathbf{I} + \mathbf{X}^{*\top}_{\mathcal{A}}\mathbf{X}^*_{\mathcal{A}}\right)^{-1}\right\}\mathbf{X}^{*\top}_{\mathcal{A}}\mathbf{Y}^* \\ &= \mathbf{Y}^{*\top}\mathbf{X}^*_{\mathcal{A}}(\mathbf{X}^{*\top}_{\mathcal{A}}\mathbf{X}^*_{\mathcal{A}})^{-1}\left\{\sigma^2_{1n}\mathbf{I} + (\mathbf{X}^{*\top}_{\mathcal{A}}\mathbf{X}^*_{\mathcal{A}})^{-1}\right\}^{-1}(\mathbf{X}^{*\top}_{\mathcal{A}}\mathbf{X}^*_{\mathcal{A}})^{-1}\mathbf{X}^{*\top}_{\mathcal{A}}\mathbf{Y}^* \le \sigma^{-2}_{1n}\mathbf{Y}^{*\top}\mathbf{X}^*_{\mathcal{A}}(\mathbf{X}^{*\top}_{\mathcal{A}}\mathbf{X}^*_{\mathcal{A}})^{-2}\mathbf{X}^{*\top}_{\mathcal{A}}\mathbf{Y}^*,\end{aligned}$$

where the first equality is due to the Woodbury matrix identity $\mathbf{A}^{-1} - (\mathbf{A}+\mathbf{UCV})^{-1} = \mathbf{A}^{-1}\mathbf{U}\left(\mathbf{C}^{-1} + \mathbf{VA}^{-1}\mathbf{U}\right)^{-1}\mathbf{VA}^{-1}$. Denote $\mathbf{M} = n\mathbf{X}^*_{\mathcal{A}}(\mathbf{X}^{*\top}_{\mathcal{A}}\mathbf{X}^*_{\mathcal{A}})^{-2}\mathbf{X}^{*\top}_{\mathcal{A}}$, which has rank $|\mathcal{A}|$ and $\lambda_{\max}(\mathbf{M}) \le 1/\lambda^*_1 \sim 1/\lambda_1$. By [13], we can derive the tail bound for the quadratic term:

$$\Pr(R_{\mathcal{A}} - R^*_{\mathcal{A}} \ge g_n) \le \Pr(\mathbf{Y}^{*\top}\mathbf{MY}^* \ge n\sigma^2_{1n}g_n) \le \exp(-cn\sigma^2_{1n}\lambda_1 g_n)$$

We next prove part (ii). By Lemma 2, $R^*_{\mathcal{A}}/(C\sigma^2) \sim \chi^2_{n-m-|\mathcal{A}|}$, by the tail bound for χ^2 distribution in [17], for any positive x, we have

$$\Pr\left\{\left|\frac{R^*_{\mathcal{A}}}{C\sigma^2} - (n-m-|\mathcal{A}|)\right| \ge 2(n-m-|\mathcal{A}|)(\sqrt{x}+2x)\right\} \le 2\exp\{-(n-|\mathcal{A}|)x\}.$$

Furthermore, since $m = o(n)$, thus for any fixed $\epsilon > 0$, there exists a constant $c > 0$, such that

$$\Pr\left(\left|\frac{R^*_{\mathcal{A}}}{nC\sigma^2} - 1\right| > \epsilon\right) \le \exp(-cn). \quad \square$$

Proof of Theorem 4. We will use strategy related to that of Theorem 4.1 in [21] to establish Theorem 4.

For overfitted models $\mathcal{P}_1 = \{\mathcal{M} : \mathcal{A} \subseteq \mathcal{M},\ \ |\mathcal{M}| \le m_n + |\mathcal{A}|\}$, we first put bound on $R_{\mathcal{M}} - R_{\mathcal{A}}$. By Lemma 2, for $\mathcal{M} \in \mathcal{P}_1$, we have $R^*_{\mathcal{A}} - R^*_{\mathcal{M}} \sim C\sigma^2\chi^2_{|\mathcal{M}|-|\mathcal{A}|}$. For any $x > 0$, $\sqrt{2/3} < w < 1$, there exists some constant $c > 0$, such that

$$\begin{aligned}\Pr\left\{R^*_{\mathcal{A}} - R^*_{\mathcal{M}} > C\sigma^2(2+3x)(|\mathcal{M}|-|\mathcal{A}|)\ln p_n\right\} &= \Pr\left\{\chi^2_{|\mathcal{M}|-|\mathcal{A}|} > (2+3x)(|\mathcal{M}|-|\mathcal{A}|)\ln p_n\right\} \\ &\le \Pr\left\{\chi^2_{|\mathcal{M}|-|\mathcal{A}|} - (|\mathcal{M}|-|\mathcal{A}|) > (2+3w^2x)(|\mathcal{M}|-|\mathcal{A}|)\ln p_n\right\} \\ &\le c\exp\{-(1+x)(|\mathcal{M}|-|\mathcal{A}|)\ln p_n\} \le cp_n^{-(1+x)(|\mathcal{M}|-|\mathcal{A}|)}.\end{aligned}$$

Define events $A(\mathcal{M}) = \{R_{\mathcal{A}} - R_{\mathcal{M}} > 2C\sigma^2(1+2s)(|\mathcal{M}|-|\mathcal{A}|)\ln p_n\}$, $U(d) = \bigcup_{\mathcal{M}\in\mathcal{P}_1:|\mathcal{M}|=d} A(\mathcal{M})$.

Then for any fixed $s > 0$, there exists some $c, c' > 0$, such that

$$\begin{aligned}\Pr\{U(d)\} &= \Pr\left\{\bigcup_{\mathcal{M}\in\mathcal{P}_1:|\mathcal{M}|=d}\left\{R_{\mathcal{A}} - R_{\mathcal{M}} > 2C\sigma^2(1+2s)(|\mathcal{M}|-|\mathcal{A}|)\ln p_n\right\}\right] \\ &= \Pr\left[\bigcup_{\mathcal{M}\in\mathcal{P}_1:|\mathcal{M}|=d}\left\{R_{\mathcal{A}} - R_{\mathcal{M}} + R^*_{\mathcal{M}} - R^*_{\mathcal{M}} > 2C\sigma^2(1+2s)(|\mathcal{M}|-|\mathcal{A}|)\ln p_n\right\}\right] \\ &\le \Pr\left[\bigcup_{\mathcal{M}\in\mathcal{P}_1:|\mathcal{M}|=d}\left\{R_{\mathcal{A}} - R^*_{\mathcal{M}} > 2C\sigma^2(1+2s)(|\mathcal{M}|-|\mathcal{A}|)\ln p_n\right\}\right] \\ &\le \Pr\left[\bigcup_{\mathcal{M}\in\mathcal{P}_1:|\mathcal{M}|=d}\left\{R^*_{\mathcal{A}} - R^*_{\mathcal{M}} > C\sigma^2(2+3s)(d-|\mathcal{A}|)\ln p_n\right\}\right] \\ &\quad + \Pr\left[R_{\mathcal{A}} - R^*_{\mathcal{A}} > C\sigma^2 s(d-|\mathcal{A}|)\ln p_n\right],\end{aligned}$$

where the first inequality is due to the fact that $R_{\mathcal{M}} - R^*_{\mathcal{M}} \ge 0$. By Lemma 3 and Condition A, it follows that

$$\Pr\{U(d)\} \le c'p_n^{-(1+s)(d-|\mathcal{A}|)}p_n^{d-|\mathcal{A}|} + \exp\left\{-cn\sigma^2_{1n}\lambda_1(d-|\mathcal{A}|)\ln p_n\right\} \le 2c'p_n^{-s(d-|\mathcal{A}|)}.$$

Then,

$$\Pr\left\{\bigcup_{d>|\mathcal{A}|}U(d)\right\} \le \sum_{d>|\mathcal{A}|}\Pr\{U(d)\} \le \frac{2c'}{p_n^s - 1} \longrightarrow 0.$$

So consider the high probability event $\left\{\bigcap_{d>|\mathcal{A}|} U(d)^c\right\}$, we have

$$\begin{aligned}\sum_{\mathcal{M}\in\mathcal{P}_1} PR(\mathcal{M},\mathcal{A}) &\preceq \sum_{\mathcal{M}\in\mathcal{P}_1} q_n^{|\mathcal{M}|-|\mathcal{A}|} p_n^{-(1+\delta)(|\mathcal{M}|-|\mathcal{A}|)} \exp\left\{-\frac{1}{2\sigma^2}(R_{\mathcal{M}}-R_{\mathcal{A}})\right\}\\ &\preceq \sum_{\mathcal{M}\in\mathcal{P}_1} q_n^{|\mathcal{M}|-|\mathcal{A}|} p_n^{-(1+\delta)(|\mathcal{M}|-|\mathcal{A}|)} \exp\left\{C(1+2s)(|\mathcal{M}|-|\mathcal{A}|)\ln p_n\right\}\\ &\preceq \sum_{\mathcal{M}\in\mathcal{P}_1} q_n^{|\mathcal{M}|-|\mathcal{A}|} p_n^{-(1+\delta)(|\mathcal{M}|-|\mathcal{A}|)} p_n^{C(1+2s)(|\mathcal{M}|-|\mathcal{A}|)}.\end{aligned}$$

As $0<C<1$, set $0<s<\delta/2$, then there exists $c>0$ such that

$$\sum_{\mathcal{M}\in\mathcal{P}_1} PR(\mathcal{M},\mathcal{A}) \preceq p_n^{-c} \sum_{|\mathcal{M}|-|\mathcal{A}|=1}^{p_n} \binom{|\mathcal{M}|-|\mathcal{A}|}{p_n} p_n^{-(|\mathcal{M}|-|\mathcal{A}|)} \preceq p_n^{-c} \longrightarrow 0.$$

Consider large and missing some active variables models $\mathcal{P}_2=\{\mathcal{M}:\mathcal{A}\nsubseteq\mathcal{M}, K|\mathcal{A}|<|\mathcal{M}|\le m_n+|\mathcal{A}|\}$. Define events

$$\begin{aligned}B(\mathcal{M}) &= \left\{R_{\mathcal{A}}-R_{\mathcal{M}}>2C\sigma^2(1+2s)(|\mathcal{M}|-|\mathcal{A}|)\ln p_n\right\}\\ &\subseteq \left\{R_{\mathcal{A}}-R_{\mathcal{M}\vee\mathcal{A}}>2C\sigma^2(1+2s)(|\mathcal{M}|-|\mathcal{A}|)\ln p_n\right\},\\ V(d) &= \bigcup_{\mathcal{M}\in\mathcal{P}_2:|\mathcal{M}|=d} B(\mathcal{M}).\end{aligned}$$

Similar to the proof for $\mathcal{P}_1$, there exists $c'>0$, such that

$$\begin{aligned}\Pr\{V(d)\} &\le P\left[\bigcup_{\mathcal{M}\in\mathcal{P}_2:|\mathcal{M}|=d}\left\{R_{\mathcal{A}}-R_{\mathcal{M}\vee\mathcal{A}}>2C\sigma^2(1+2s)(|\mathcal{M}|-|\mathcal{A}|)\ln p_n\right\}\right]\\ &\le 2c' p_n^{-(1+s)(d-|\mathcal{A}|)} p_n^d \le 2c' p_n^{-(1+w')d} p_n^d \le 2c' p_n^{-w'd}.\end{aligned}$$

Take $s=\delta/4$, we can find such $w'>0$ as long as $\frac{K-1}{K}(1+s)>1$. That is, $K>1+4/\delta$, which is stated in Condition B. It follows that

$$\Pr\left\{\bigcup_{d>K|\mathcal{A}|} V(d)\right\} \le \sum_{d>K|\mathcal{A}|} \Pr\{V(d)\} \le 2c' p_n^{-w'K|\mathcal{A}|} \longrightarrow 0.$$

Then consider the high probability event $\left\{\bigcap_{d>K|\mathcal{A}|} V(d)^c\right\}$,

$$\begin{aligned}\sum_{\mathcal{M}\in\mathcal{P}_2} PR(\mathcal{M},\mathcal{A}) &\preceq \sum_{\mathcal{M}\in\mathcal{P}_2} q_n^{|\mathcal{M}|-|\mathcal{A}|} p_n^{-(1+\delta)(|\mathcal{M}|-|\mathcal{A}|)} \exp\left\{C(1+2s)(|\mathcal{M}|-|\mathcal{A}|)\ln p_n\right\}\\ &\preceq \sum_{\mathcal{M}\in\mathcal{P}_2} q_n^{|\mathcal{M}|-|\mathcal{A}|} p_n^{-\delta(\mathcal{M}-1)/2} \preceq p_n^{-\delta(K-1)/2} \sum_{|\mathcal{M}|-|\mathcal{A}|=1}^{p_n} \binom{|\mathcal{M}|-|\mathcal{A}|}{p_n} p_n^{-(|\mathcal{M}|-|\mathcal{A}|)} \preceq p_n^{-\delta(K-1)/2} \longrightarrow 0.\end{aligned}$$

For any model $\mathcal{M}$ belonging to the group of underfitted models $\mathcal{P}_3=\{\mathcal{M}:\mathcal{A}\nsubseteq\mathcal{M}, |\mathcal{M}|\le K|\mathcal{A}|\}$, it follows that

$$\begin{aligned}R^*_{\mathcal{M}}-R^*_{\mathcal{M}\vee\mathcal{A}} &= \|(\mathbf{P}_{\mathcal{M}}-\mathbf{P}_{\mathcal{M}\vee\mathcal{A}})\mathbf{Y}^*\|_2^2 = \|(\mathbf{P}_{\mathcal{M}}-\mathbf{P}_{\mathcal{M}\vee\mathcal{A}})(\mathbf{X}^*_{\mathcal{A}}\boldsymbol{\beta}_{\mathcal{A}}+\boldsymbol{\delta}+\boldsymbol{\omega})\|_2^2\\ &\ge \left\{\|(\mathbf{P}_{\mathcal{M}}-\mathbf{P}_{\mathcal{M}\vee\mathcal{A}})\mathbf{X}^*_{\mathcal{A}}\boldsymbol{\beta}_{\mathcal{A}}\|_2 - \|(\mathbf{P}_{\mathcal{M}}-\mathbf{P}_{\mathcal{M}\vee\mathcal{A}})(\boldsymbol{\delta}+\boldsymbol{\omega})\|_2\right\}^2.\end{aligned}$$

By Condition B,

$$\|(\mathbf{P}_{\mathcal{M}}-\mathbf{P}_{\mathcal{M}\vee\mathcal{A}})\mathbf{X}^*_{\mathcal{A}}\boldsymbol{\beta}_{\mathcal{A}}\|_2 = \|(\mathbf{I}-\mathbf{P}_{\mathcal{M}})\mathbf{X}^*_{\mathcal{A}}\boldsymbol{\beta}_{\mathcal{A}}\|_2 \ge \sqrt{\Delta_n(K)}.$$

And on the other hand, $\|(\mathbf{P}_{\mathcal{M}}-\mathbf{P}_{\mathcal{M}\vee\mathcal{A}})(\boldsymbol{\delta}+\boldsymbol{\omega})\|_2^2 = \|(\mathbf{P}_{\mathcal{M}}-\mathbf{P}_{\mathcal{M}\vee\mathcal{A}})\boldsymbol{\delta}\|_2^2 + \|(\mathbf{P}_{\mathcal{M}}-\mathbf{P}_{\mathcal{M}\vee\mathcal{A}})\boldsymbol{\omega}\|_2^2 - 2\boldsymbol{\omega}^\top(\mathbf{P}_{\mathcal{M}}-\mathbf{P}_{\mathcal{M}\vee\mathcal{A}})\boldsymbol{\delta}$. Among them, $0\le\|(\mathbf{P}_{\mathcal{M}}-\mathbf{P}_{\mathcal{M}\vee\mathcal{A}})\boldsymbol{\delta}\|_2^2 \le 2(\|\mathbf{P}_{\mathcal{M}}\boldsymbol{\delta}\|_2^2+\|\mathbf{P}_{\mathcal{M}\vee\mathcal{A}}\boldsymbol{\delta}\|_2^2) \le 4\|\boldsymbol{\delta}\|_2^2 \longrightarrow o(1)$. By the similar trick in Lemma 2, it can be shown $\boldsymbol{\omega}^\top(\mathbf{P}_{\mathcal{M}}-\mathbf{P}_{\mathcal{M}\vee\mathcal{A}})\boldsymbol{\delta}=0$ a.s.

For any $0<w<1$,

$$\begin{aligned}&\Pr\left[\bigcup_{\mathcal{M}\in\mathcal{P}_3}\left\{R^*_{\mathcal{M}}-R^*_{\mathcal{M}\vee\mathcal{A}}<(1-w)^2\Delta_n(K)\right\}\right] \le \Pr\left[\bigcup_{\mathcal{M}\in\mathcal{P}_3}\left\{\|(\mathbf{P}_{\mathcal{M}}-\mathbf{P}_{\mathcal{M}\vee\mathcal{A}})\boldsymbol{\omega}\|_2>w/2\sqrt{\Delta_n(K)}\right\}\right]\\ &\le \Pr\left\{\|\mathbf{P}_{\mathcal{A}}\boldsymbol{\omega}\|_2>w/2\sqrt{\Delta_n(K)}\right\} = \Pr\left\{C\sigma^2\chi^2_{|\mathcal{A}|}>w^2/4\Delta_n(K)\right\} \le \exp\left\{-c\Delta_n(K)/|\mathcal{A}|\right\}.\end{aligned}$$

The last step follows by the bound for tail with quadratic form. For any $0 < w' < 1$,

$$\begin{aligned}
&\Pr\left[\bigcup_{\mathcal{M}\in\mathcal{P}_3}\left\{R_{\mathcal{M}} - R_{\mathcal{M}\vee\mathcal{A}} < (1-w')\Delta_n(K)\right\}\right]\\
=&\Pr\left[\bigcup_{\mathcal{M}\in\mathcal{P}_3}\{(R^*_{\mathcal{M}} - R^*_{\mathcal{M}\vee\mathcal{A}}) + (R^*_{\mathcal{M}\vee\mathcal{A}} - R_{\mathcal{M}\vee\mathcal{A}}) + (R_{\mathcal{M}} - R^*_{\mathcal{M}}) < (1-w')\Delta_n(K)\}\right]\\
\le&\Pr\left[\bigcup_{\mathcal{M}\in\mathcal{P}_3}\{R^*_{\mathcal{M}} - R^*_{\mathcal{M}\vee\mathcal{A}}\} < (1-w'/2)\Delta_n(K)\right] + \Pr\left[\bigcup_{\mathcal{M}\in\mathcal{P}_3}\{R^*_{\mathcal{M}\vee\mathcal{A}} - R_{\mathcal{M}\vee\mathcal{A}}\} < -w'/2\Delta_n(K)\right]\\
\le&2\exp\left\{-c'\Delta_n(K)/|\mathcal{A}|\right\},
\end{aligned}$$

where the first inequality is due to the fact that $R_{\mathcal{M}} - R^*_{\mathcal{M}} \ge 0$, and the last inequality follows by the exponential tails of $n\sigma_{1n}^2\lambda_1(R_{\mathcal{M}\vee\mathcal{A}} - R^*_{\mathcal{M}\vee\mathcal{A}})$. The proof is similar to Lemma 2(1).

Let $c = 2w$, it follows that

$$\begin{aligned}
\Pr\left[\bigcup_{\mathcal{M}\in\mathcal{P}_3}\{R_{\mathcal{M}} - R_{\mathcal{A}} < (1-c)\Delta_n(K)\}\right] \le& \Pr\left[\bigcup_{\mathcal{M}\in\mathcal{P}_3}\{R_{\mathcal{M}} - R_{\mathcal{M}\vee\mathcal{A}} < (1-w)\Delta_n(K)\}\right]\\
&+ \Pr\left[\bigcup_{\mathcal{M}\in\mathcal{P}_3}\{R_{\mathcal{A}} - R_{\mathcal{M}\vee\mathcal{A}} > w\Delta_n(K)\}\right]\\
\le& 2\exp\left\{-c'\Delta_n(K)/|\mathcal{A}|\right\} + \Pr\left[\bigcup_{\mathcal{M}\in\mathcal{P}_3}\{R^*_{\mathcal{A}} - R^*_{\mathcal{M}\vee\mathcal{A}} > w\Delta_n(K)\}\right]\\
&+ \Pr\left\{R_{\mathcal{A}} - R^*_{\mathcal{A}} > w\Delta_n(K)\right\}\\
\le& 3\exp\left\{-c'\Delta_n(K)/|\mathcal{A}|\right\} + \Pr\left\{C\sigma^2\chi^2_{K|\mathcal{A}|} > w\Delta_n(K)\right\}\\
\le& 4\exp\left\{-c'\Delta_n(K)/|\mathcal{A}|\right\} \longrightarrow 0,
\end{aligned}$$

where the second inequality holds because $R_{\mathcal{A}} - R_{\mathcal{M}\vee\mathcal{A}} = R^*_{\mathcal{A}} - R^*_{\mathcal{M}\vee\mathcal{A}} + R_{\mathcal{A}} - R^*_{\mathcal{A}} + R^*_{\mathcal{M}\vee\mathcal{A}} - R_{\mathcal{M}\vee\mathcal{A}}$, and the last step follows by Condition B. Then consider the high probability event $\left\{\bigcap_{\mathcal{M}\in\mathcal{P}_3}(R_{\mathcal{M}} - R_{\mathcal{A}}) > (1-c)\Delta_n(K)\right\}$,

$$\begin{aligned}
\sum_{\mathcal{M}\in\mathcal{P}_3} PR(\mathcal{M},\mathcal{A}) \preceq& \sum_{\mathcal{M}\in\mathcal{P}_3} q_n^{|\mathcal{M}|-|\mathcal{A}|}(n\sigma_{1n}^2\lambda_1)^{|\mathcal{A}|/2}(\lambda_2/\lambda_1)^{|\mathcal{A}|/2}\exp\left\{-(1-c)\Delta_n(K)/(2\sigma^2)\right\}\\
\preceq& \exp\left[-\frac{1}{2\sigma^2}\left\{(1-c)\Delta_n(K) - \sigma^2|\mathcal{A}|(2+3\delta)\ln p_n - \sigma^2|\mathcal{A}|(2+\kappa)\ln p_n\right\}\right].
\end{aligned}$$

By Condition B, $\Delta_n(K) > \sigma^2|\mathcal{A}|\ln p_n(4+4\delta+\kappa)$, so, we can find $0 < c < 1$, $w' > 0$, such that $(1-c)\Delta_n(K) - \sigma^2|\mathcal{A}|\ln p_n(4+3\delta+\kappa) = w'\sigma^2|\mathcal{A}|\ln p_n$. Thus,

$$\sum_{\mathcal{M}\in\mathcal{P}_3} PR(\mathcal{M},\mathcal{A}) \preceq \exp\left(-\frac{1}{2\sigma^2}w'\sigma^2|\mathcal{A}|\ln p_n\right) \longrightarrow 0. \quad \square$$

Proof of Theorem 5. Similar to Theorem 4.2 in [21], we start from (5) and integrate out σ^2, thus the posterior probability under unknown σ^2 is

$$\begin{aligned}
\Pr(\boldsymbol{\gamma} = \mathcal{M}|\mathbf{Y}) \propto& \int q_n^{|\mathcal{M}|}\sigma^{-n}\sigma_{1n}^{-|\mathcal{M}|}|\mathbf{X}_{\mathcal{M}}^\top\mathbf{X}_{\mathcal{M}} + \mathbf{I}_n/\sigma_{1n}^2|^{-1/2}\exp\left(-\frac{1}{2\sigma^2}R_{\mathcal{M}}\right)\pi(\sigma^2)d\sigma^2\\
\propto& q_n^{|\mathcal{M}|}\sigma_{1n}^{-|\mathcal{M}|}|\mathbf{X}_{\mathcal{M}}^\top\mathbf{X}_{\mathcal{M}} + \mathbf{I}_n/\sigma_{1n}^2|^{-1/2}(2b_0 + R_{\mathcal{M}})^{-a_0-n/2}.
\end{aligned}$$

Remark 8. Inside the integration, $\sigma^{-n}\exp\left(-\frac{1}{2\sigma^2}R_{\mathcal{M}}\right)$ is the dominant term as the sum of squared residuals $R_{\mathcal{M}}$ has an order of $O_p(n)$. The theorem applies to a wider family of prior as long as $\pi(\sigma^2)$ is $O_p(1)$ and the support is not too strange. This includes some commonly used priors like improper non-informative prior $\pi(\sigma^2) \propto \sigma^{-2}$ and the class of folded-noncentral-t prior with fixed hyper-parameters $\pi(\sigma) \propto \left(1 + \frac{\sigma^2}{\nu^2 a^2}\right)^{-(\nu+1)/2}$.

By Lemma 1, we have

$$\frac{\Pr(\gamma=\mathcal{M}|Y)}{\Pr(\gamma=\mathcal{A}|Y)} \propto p_n^{-1.5\delta(|\mathcal{M}|-|\mathcal{A}|)+0.5\kappa}\left(\frac{2b_0+R_{\mathcal{M}}}{2b_0+R_{\mathcal{A}}}\right)^{-a_0-\frac{n}{2}}.$$

Define $\rho_n=\frac{R_{\mathcal{A}}+2b_0}{nC\sigma^2}-1$, then $\rho_n=o_p(1)$, since

$$\frac{R^*_{\mathcal{A}}+2b_0}{nC\sigma^2}-1<\rho_n=\frac{R_{\mathcal{A}}+2b_0}{nC\sigma^2}-1=\frac{R^*_{\mathcal{A}}+2b_0}{nC\sigma^2}-1+\frac{R_{\mathcal{A}}-R^*_{\mathcal{A}}}{nC\sigma^2},$$
$$\Pr(|\rho_n|>2\epsilon)\le\Pr\left(\left|\frac{R^*_{\mathcal{A}}}{nC\sigma^2}-1\right|>\epsilon\right)+\Pr\left(\left|\frac{R_{\mathcal{A}}-R^*_{\mathcal{A}}}{nC\sigma^2}\right|>\epsilon\right)\le 2\exp(-cn).$$

First consider overfitted models $\mathcal{M}\in\mathcal{P}_1$. Define $x_n=(|\mathcal{M}|-|\mathcal{A}|)\ln p_n/n$, $t_n=-\ln\{1-2(1+4s)x_n\}$. Then by Condition D, $x_n\le\frac{\delta}{(4+\delta)(2+\delta)}$. Since $-\ln(1-x)\le\frac{x}{1-x}$, then for any $s<\delta/16$, we have

$$t_n=-\ln\{1-2(1+4s)x_n\}<\frac{2(1+4s)x_n}{1-2(1+4s)x_n}<2(1+\delta/2)x_n.$$

Similar to the proof in Theorem 4(1), consider the high probability event $\left\{\bigcap_{d>|\mathcal{A}|}U(d)^c\right\}\bigcap\{|\rho_n|<\epsilon^*\}$, where $(1+4s)(1-\epsilon^*)>(1+2s)$, then

$$\begin{aligned}\left(\frac{2b_0+R_{\mathcal{M}}}{2b_0+R_{\mathcal{A}}}\right)^{-a_0-\frac{n}{2}}&=\left\{1+\frac{R_{\mathcal{M}}-R_{\mathcal{A}}}{nC\sigma^2(1+\rho_n)}\right\}^{-a_0-\frac{n}{2}}\preceq\exp\left\{(a_0+\frac{n}{2})t_n\right\}\\&\preceq\exp\left\{(a_0+\frac{n}{2})2(1+\delta/2)(|\mathcal{M}|-|\mathcal{A}|)\ln p_n/n\right\}\\&\preceq\exp\{(1+\delta/2)(|\mathcal{M}|-|\mathcal{A}|)\ln p_n\}\preceq p_n^{(1+\delta/2)(|\mathcal{M}|-|\mathcal{A}|)}.\end{aligned}$$

The problem reduces to the same problem in Theorem 4(1). And for $\mathcal{M}\in\mathcal{P}_2$, we can use the same trick, thus we have

$$\sum_{\mathcal{M}\in\mathcal{P}_1\cup\mathcal{P}_2}\frac{\Pr(\gamma=\mathcal{M}|Y)}{\Pr(\gamma=\mathcal{A}|Y)}\xrightarrow{P}0.$$

For underfitted models $\mathcal{M}\in\mathcal{P}_3$, similar to the proof in Theorem 4(3), consider the high probability event $\left\{\bigcap_{\mathcal{M}\in\mathcal{P}_3}[R_{\mathcal{M}}-R_{\mathcal{A}}>(1-c)\Delta_n(K)]\right\}\bigcap\{|\rho_n|<\epsilon^*\}$.

If $\Delta_n(K)=o(n)$, by $\lim_{n\to\infty}(1+1/n)^n=e$,

$$\begin{aligned}\left(\frac{2b_0+R_{\mathcal{M}}}{2b_0+R_{\mathcal{A}}}\right)^{-a_0-\frac{n}{2}}&=\left\{1+\frac{R_{\mathcal{M}}-R_{\mathcal{A}}}{nC\sigma^2(1+\rho_n)}\right\}^{-a_0-\frac{n}{2}}\preceq\left\{1+\frac{(1-c)\Delta_n(K)}{nC\sigma^2(1+\rho_n)}\right\}^{-a_0-\frac{n}{2}}\\&\preceq\exp\left\{-(\frac{n}{2}+a_0)\frac{(1-c)\Delta_n(K)}{nC\sigma^2(1+\epsilon^*)}\right\}\preceq\exp\left\{-\frac{(1-c)\Delta_n(K)}{2\sigma^2(1+\epsilon^*)}\right\}.\end{aligned}$$

It reduces to the same problem in Theorem 4(3).

While if $\Delta_n(K)\succeq n$, it follows that

$$\left(\frac{2b_0+R_{\mathcal{M}}}{2b_0+R_{\mathcal{A}}}\right)^{-a_0-\frac{n}{2}}\preceq\left\{1+\frac{(1-c)\Delta_n(K)}{nC\sigma^2(1+\rho_n)}\right\}^{-a_0-\frac{n}{2}}\preceq\exp\left\{-c'n\ln\frac{\Delta_n(K)}{n}\right\}\preceq\exp\left(-c'n\right),$$

which converges even faster to 0 as $n\to\infty$. □

5. Discussion

Inspired by the difference-based method, we have proposed a new Bayesian approach to select variables in the linear component of the partially linear models. We modify the Bayesian shrinking and diffusing priors (BASAD) [21], and propose the new Bayesian subset modeling with diffusing prior (BSM-DP). The idea is extended from linear models to the partially linear model with the help of the difference-based method. Model selection consistency is proved under the setting with ultra-high dimensional covariates. Compared to BASAD, BSM-DP performances better in identifying the low signal covariates and in shorter computation time, as shown in the supplementary material. Results in the simulation studies show that our method has higher tolerance on the correlation among predictors and requires mild conditions on covariates, compared with other existing methods for variable selection on PLM. The proposed model is less likely to overfit the model, which is also illustrated by the real data example about supermarket. However, like all other Bayesian methods, it has some price to pay. We do need specific assumptions on the error distribution. The computation is relative intense as compared to frequentist penalized methods. Finally, similar to frequentist methods, although we showed the required rates for the hyperparameters of the priors, the practical choices of them in finite sample applications still need fine tuning.

Acknowledgments

The authors are grateful to the Editor-in-Chief, an Associate Editor and the referees for comments and suggestions that led to significant improvements. This research was supported by NSF, USA grants DMS 1820702, DMS 1953196, DMS 2015539 and NIH, USA grants R01CA229542 and R01 ES019672. The content is solely the responsibility of the authors and does not necessarily represent the official views of NSF and NIH.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jmva.2021.104733. The online supplement provides some extra technical details and additional simulation results.

References

[1] M.M. Barbieri, J.O. Berger, Optimal predictive model selection, Ann. Statist. 32 (2004) 870–897.
[2] L. Breiman, Better subset regression using the nonnegative garrote, Technometrics 37 (1995) 373–384.
[3] P. Bühlmann, M. Kalisch, M.H. Maathuis, Variable selection in high-dimensional linear models: partially faithful distributions and the pc-simple algorithm, Biometrika 97 (2010) 261–278.
[4] J. Chen, Z. Chen, Extended bic for small-n-large-p sparse glm, Statist. Sinica 22 (2012) 555–574.
[5] Z. Chen, J. Fan, R. Li, Error variance estimation in ultrahigh-dimensional additive models, J. Amer. Statist. Assoc. 113 (2018) 315–327.
[6] H. Chen, J.-J.H. Shiau, A two-stage spline smoothing method for partially linear models, J. Statist. Plann. Inference 27 (1991) 187–201.
[7] R.F. Engle, C.W.J. Granger, J. Rice, A. Weiss, Semiparametric estimates of the relation between weather and electricity sales, J. Amer. Statist. Assoc. 81 (1986) 310–320.
[8] J. Fan, R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, J. Amer. Statist. Assoc. 96 (2001) 1348–1360.
[9] J. Fan, J. Lv, Sure independence screening for ultrahigh dimensional feature space, J. R. Stat. Soc. Ser. B Stat. Methodol. 70 (2008) 849–911.
[10] E.I. George, D.P. Foster, Calibration and empirical bayes variable selection, Biometrika 87 (2000) 731–747.
[11] E.I. George, R.E. McCulloch, Variable selection via gibbs sampling, J. Amer. Statist. Assoc. 88 (1993) 881–889.
[12] N.E. Heckman, Spline smoothing in a partly linear model, J. R. Stat. Soc. Ser. B Methodol. 48 (1986) 244–248.
[13] D. Hsu, S. Kakade, T. Zhang, A tail inequality for quadratic forms of subgaussian random vectors, Electron. Commun. Probab. 17 (2012) 1–6.
[14] H. Ishwaran, J.S. Rao, Spike and slab variable selection: frequentist and bayesian strategies, Ann. Statist. 33 (2005) 730–773.
[15] H. Ishwaran, J.S. Rao, Consistency of spike and slab regression, Statist. Probab. Lett. 81 (2011) 1920–1928.
[16] V.E. Johnson, D. Rossell, Bayesian model selection in high-dimensional settings, J. Amer. Statist. Assoc. 107 (2012) 649–660.
[17] B. Laurent, P. Massart, Adaptive estimation of a quadratic functional by model selection, Ann. Statist. 28 (2000) 1302–1338.
[18] H. Liang, R. Li, Variable selection for partially linear models with measurement errors, J. Amer. Statist. Assoc. 104 (2009) 234–248.
[19] F. Liang, Q. Song, K. Yu, Bayesian subset modeling for high-dimensional generalized linear models, J. Amer. Statist. Assoc. 108 (2013) 589–606.
[20] J. Liu, L. Lou, R. Li, Variable selection for partially linear models via partial correlation, J. Multivariate Anal. 167 (2018) 418–434.
[21] N.N. Narisetty, X. He, Bayesian variable selection with shrinking and diffusing priors, Ann. Statist. 42 (2014) 789–817.
[22] N.N. Narisetty, J. Shen, X. He, Skinny gibbs: A consistent and scalable gibbs sampler for model selection, J. Amer. Statist. Assoc. 114 (2019) 1205–1217.
[23] J. Rice, Convergence rates for partially splined models, Statist. Probab. Lett. 4 (1986) 203–208.
[24] P.M. Robinson, Root-n-consistent semiparametric regression, Econometrica 56 (1988) 931–954.
[25] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. Ser. B Methodol. 58 (1996) 267–288.
[26] H. Wang, Forward regression for ultra-high dimensional variable screening, J. Amer. Statist. Assoc. 104 (2009) 1512–1524.
[27] L. Wang, L.D. Brown, T.T. Cai, A difference based approach to the semiparametric partial linear model, Electron. J. Stat. 5 (2011) 619–641.
[28] H. Xie, J. Huang, Scad-penalized regression in high-dimensional partially linear models, Ann. Statist. 37 (2009) 673–696.
[29] A. Yatchew, An elementary estimator of the partial linear model, Econ. Lett. 57 (1997) 135–143.
[30] M. Yuan, Y. Lin, Efficient empirical bayes variable selection and estimation in linear models, J. Amer. Statist. Assoc. 100 (2005) 1215–1225.
[31] M. Yuan, Y. Lin, On the non-negative garrotte estimator, J. R. Stat. Soc. Ser. B Stat. Methodol. 69 (2007) 143–161.
[32] C.-H. Zhang, Nearly unbiased variable selection under minimax concave penalty, Ann. Statist. 38 (2010) 894–942.
[33] L. Zhu, R. Li, H. Cui, Robust estimation for partially linear models with large-dimensional covariates, Sci. China Math. 56 (2013) 2069–2088.
[34] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. R. Stat. Soc. Ser. B Stat. Methodol. 67 (2005) 301–320.