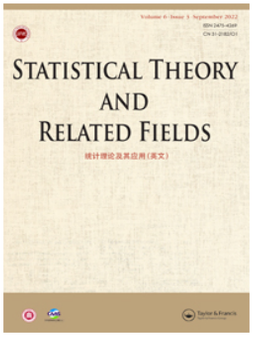

Statistical Theory and Related Fields

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tstf20

Variable screening in multivariate linear regression with high-dimensional covariates

Shiferaw B. Bizuayehu, Lu Li & Jin Xu

To cite this article: Shiferaw B. Bizuayehu, Lu Li & Jin Xu (2022) Variable screening in multivariate linear regression with high-dimensional covariates, Statistical Theory and Related Fields, 6:3, 241-253, DOI: 10.1080/24754269.2021.1982607

To link to this article: https://doi.org/10.1080/24754269.2021.1982607

Published online: 06 Oct 2021.

STATISTICAL THEORY AND RELATED FIELDS
2022, VOL. 6, NO. 3, 241–253
https://doi.org/10.1080/24754269.2021.1982607

Variable screening in multivariate linear regression with high-dimensional covariates

Shiferaw B. Bizuayehu[a], Lu Li[b] and Jin Xu[a,c]

[a]School of Statistics, East China Normal University, Shanghai, People's Republic of China; [b]School of Mathematical Sciences, Shanghai Jiao Tong University, Shanghai, People's Republic of China; [c]Key Laboratory of Advanced Theory and Application in Statistics and Data Science – MOE, East China Normal University, Shanghai, People's Republic of China

ABSTRACT

We propose two variable selection methods in multivariate linear regression with high-dimensional covariates. The first method uses a multiple correlation coefficient to fast reduce the dimension of the relevant predictors to a moderate or low level. The second method extends the univariate forward regression of Wang [(2009). Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association*, *104*(488), 1512–1524. https://doi.org/10.1198/jasa.2008.tm08516] in a unified way such that the variable selection and model estimation can be obtained simultaneously. We establish the sure screening property for both methods. Simulation and real data applications are presented to show the finite sample performance of the proposed methods in comparison with some naive method.

ARTICLE HISTORY
Received 1 January 2021
Revised 2 September 2021
Accepted 9 September 2021

KEYWORDS
Dimension reduction; forward regression; multiple correlation coefficient; multivariate regression; variable selection

1. Introduction

High-dimensional multivariate regression has been widely applied in bioinformatics, chemometrics, and medical image analysis where many of the response variables are highly correlated (Cai et al., 2013; Ferte et al., 2013; Jia et al., 2017; Peng et al., 2010; Smith & Fahrmeir, 2007). For instance, in genetics study, we are interested in the association between correlated phenotypes (involved in biological pathways) and genotypes, as genetic effects and their possible interaction have been recognized as an important component for the genetic architecture of each complex phenotype (Yi, 2010). For this kind of problem, the number of covariates or explanatory variables is much larger than the number of observations or samples. Traditional methods of subset selection and stepwise procedure become infeasible when confronted with high dimensionality (Breiman, 1995).

Statistical methods and theories have been developed to solve this problem through various approaches such as network-based regularization method (C. Li & Li, 2008; Ren et al., 2019, 2017), graphical model (B. Li, Chuns et al., 2012; Yin & Li, 2011), correlation-based screening (B. Li, Chuns et al., 2012; Song et al., 2016) and group lasso (Y. Li et al., 2015; J. Wang et al., 2019; Yang & Zou, 2015).

Variable selection methods for regression models with a univariate response have been proposed in the past. Some popular methods include the bridge regression (Frank & Friedman, 1993; Fu, 1998), LASSO (Tibshirani, 1996), SCAD (Fan & Li, 2001), LARS (Efron et al., 2004), elastic net (Zou & Hastie, 2005), adaptive LASSO (H. H. Zhang & Lu, 2007; Zou, 2006), and Dantzig selector (Candes & Tao, 2007; Y. Kong et al., 2016), among others. On the other hand, variable screening procedures have been developed to reduce the dimensionality from an ultrahigh dimension to a lower dimension which is smaller than the sample size (Fan & Lv, 2008; X. Kong et al., 2017; G. Li, Peng et al., 2012; H. Wang, 2009; Zhu et al., 2011).

For variable selection under multivariate regression models, one simple approach is to apply some variable selection method to univariate regression of each response separately. Such an approach may produce sub-optimal results since it does not utilize the joint information among the responses (Breiman & Friedman, 1997; Kim et al., 2009). To improve the estimation, various attempts have been made. One approach is to use dimension reduction techniques such as the reduced rank regression (Chen & Huang, 2012; He et al., 2018; Zhao et al., 2017) and the sliced inverse regression (Setdji & Cook, 2004; N. Zhang et al., 2019). Another approach is to use a block-structured regularization method to select a subset which can be used as predictors for all outcome variables (Obozinski et al., 2011; Peng et al., 2010; Turlach et al., 2005). The latter approach assumes that a covariate affects either all or none of the responses. However, this assumption may be too strong when each response variable is affected by different sets of predictors. Rothman

CONTACT Jin Xu jxu@stat.ecnu.edu.cn
Contributed equally

et al. (2010) proposed a penalized framework to estimate multivariate regression coefficient and covariance matrix simultaneously under ℓ_1 penalty. Lee and Liu (2012) further improved Rothman et al.'s (2010) work by using a weighted ℓ_1 regularization. Cai et al. (2013) proposed a method to first estimate the regression coefficients in a column-wise fashion with Dantzig selector and then to estimate the precision matrix by solving a constrained ℓ_1 minimization problem.

In high-dimensional setting, most of the aforementioned multivariate regression methods use the technique of regularization to estimate the regression coefficient matrix (Obozinski et al., 2011; Peng et al., 2010; Turlach et al., 2005). However, a well-chosen penalty requires an efficient exploration of the correlation structure of the responses. It is reported that simultaneously estimating covariance and selecting variables via joint optimization can be numerically unstable in high-dimensional cases (Deshpande et al., 2019; Pecanka et al., 2019; Ren et al., 2019).

In this study, we propose two methods in parallel for variable screening and variable selection, namely the multiple correlation coefficient (MCC) screening (Section 3) and the unified forward regression (UFR) (Section 4). The first method is for dimension reduction which filters out covariates that have weak correlation with the response variables. It significantly reduces the feature space to a moderate or low dimension that covers the set of relevant predictors almost certainly. The second method is for variable selection which uses an extended forward regression (FR) (H. Wang, 2009) to identify all relevant predictors consistently under mild conditions. By MCC all relevant predictors are identified or screened, whereas by UFR both variable selection and model estimation are obtained. We illustrate the finite sample performance of the proposed methods in comparison with a naive method by simulation (Section 5) and a real data application (Section 6). We conclude the paper in Section 7 and defer the technical proofs in Appendix.

2. Notation and assumptions

Let $\mathbf{y} = (y_1, y_2, \ldots, y_q)^\top$ denote the q-dimensional response vector of interest. Let $\mathbf{x} = (x_1, x_2, \ldots, x_p)^\top$ denote the p-dimensional covariates or predictors. Denote the covariance matrices of $\mathbf{y}$ and $\mathbf{x}$ by Σ_y and $\Sigma_x = (\sigma_{ij})$, respectively. Without loss of generality, assume that $\mathrm{E}(x_k) = 0$ and $\mathrm{var}(x_k) = 1$ for $k = 1, \ldots, p$ and that $\mathrm{E}(y_j) = 0$ for $j = 1, \ldots, q$. In practice, these can be achieved by standardization and centralization.

Consider the multivariate linear regression model

$$\mathbf{y} = B^\top \mathbf{x} + \vec{\varepsilon}, \tag{1}$$

where B is a $p \times q$ matrix of coefficients and $\vec{\varepsilon}$ is the random error vector which is independent with $\mathbf{x}$. For $j = 1, \ldots, q$ and $k = 1, \ldots, p$, denote $\vec{\beta}_j$ as the jth column vector of B and $\vec{\beta}_{(k)}$ as the kth row vector of B. If $\vec{\beta}_{(k)} \neq \mathbf{0}$, x_k is referred to as a relevant predictor.

Let $F = \{1, \ldots, p\}$ denote the full model of predictors. Let $S = \{k : \vec{\beta}_{(k)} \neq \mathbf{0}\}$ denote the true model. Denote the compliment of S by S^c. Denote the cardinalities of F and S as $|F| = p$ and $|S| = p_0$, respectively. Throughout, let $\|\cdot\|$ denote the Euclidean norm of a vector.

Let $\{(\mathbf{y}_i, \mathbf{x}_i) : i = 1, \ldots, n\}$ denote independent and identically distributed samples of $(\mathbf{y}, \mathbf{x})$. Denote $X_{n\times p} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top$ and $Y_{n\times q} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)^\top$. For $j = 1, \ldots, q$, let $\mathbf{y}_{(j)}$ denote the jth column of Y.

Assume that $\mathbf{x}$ is high dimensional with p being much larger than the sample size n (in the sense of Cai & Lv, 2007). Assume that the response vector is associated with only a small portion of predictors, i.e., p_0/p is small and p_0 is $\mathrm{O}(n)$ (Fan & Lv, 2008). This sparsity principle is frequently adopted and deemed useful in analysis.

3. Multiple correlation coefficient

We first propose to use a multiple correlation coefficient (MCC) to identify S. It is known that the multiple correlation coefficient between $\mathbf{y}$ and x_k is defined as $\rho_k = \max_{\vec{\alpha}\in\mathbf{R}^q} \mathrm{corr}(\vec{\alpha}^\top \mathbf{y}, x_k)$ and its square can be further expressed as

$$\rho_k^2 = \mathrm{E}(\vec{\gamma}_k^\top \mathbf{y} x_k), \tag{2}$$

where $\vec{\gamma}_k = \Sigma_y^{-1}\mathrm{E}(\mathbf{y}x_k)$ (Anderson, 2003, Section 12.2).

Given the standardized samples, we estimate ρ_k^2 by

$$\widehat{\rho}_k^2 = \frac{1}{n}\sum_{i=1}^{n} \widehat{\vec{\gamma}}_k^\top \mathbf{y}_i x_{ik}, \tag{3}$$

where $\widehat{\vec{\gamma}}_k = (\frac{1}{n}\sum_{i=1}^n \mathbf{y}_i\mathbf{y}_i^\top)^{-1}\sum_{i=1}^n \mathbf{y}_i x_{ik}$. Note that the computation of $\widehat{\rho}_k^2$ is simple and fast through matrix algebra and does not involve any iteration. Then, we estimate S by $\widehat{S}_{\mathrm{MCC}} = \{k : \widehat{\rho}_k^2 \geq \tau\}$, where τ is the threshold which determines the size of the estimated predictors. Here we adopt the threshold of Fan and Lv (2008) by choosing $\tau = \widehat{\rho}^2_{(p-d_n+1)}$, where $\widehat{\rho}^2_{(1)} \leq \cdots \leq \widehat{\rho}^2_{(p)}$ are the order statistics and $d_n = \lceil n/\log(n)\rceil$ ($\lceil\cdot\rceil$ is the ceiling function), so that d_n predictors with the largest values of $\widehat{\rho}_k^2$ are retained. The naive correlation coefficient (NCC) method of Fan and Lv (2008) estimates S by $\widehat{S}_{\mathrm{NCC}} = \cup_{j=1}^p \{k : \widehat{\rho}_{k,j}^2 \geq \tau_j\}$, where $\widehat{\rho}_{k,j}$ is the sample correlation coefficient between y_j and x_k and τ_j is determined in the same way as in MCC with respect to the jth response.

We now show that the MCC-based screening procedure has the sure screening property (i.e., the probability of selecting all true relevant predictors tends to one)

and reduces the dimensionality of predictors below the sample size.

We state some assumptions first.

Assumption 3.1: Let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the smallest and largest eigenvalue of a positive definite matrix A, respectively. Assume that there exist two positive constants $\tau_{\min} < \tau_{\max}$ such that

$$2\tau_{\min} < \lambda_{\min}(\Sigma_y^{-1}) \leq \lambda_{\max}(\Sigma_y^{-1}) < 2^{-1}\tau_{\max}$$

and

$$2\tau_{\min} < \lambda_{\min}(\Sigma_x) \leq \lambda_{\max}(\Sigma_x) < 2^{-1}\tau_{\max}.$$

Assumption 3.2: Assume that (i) for $j = 1, \ldots, q$, $\|\vec{\beta}_j\| \leq C_B$ for some positive constant C_B and that (ii) for $k = 1, \ldots, p$, $\beta_{\min} = \min_{k\in S} \min_j |\beta_{kj}| \geq \nu_B n^{-\xi_{\min}}$ for some positive constants $\xi_{\min}$ and ν_B.

Assumption 3.3: Assume that there exist positive constants $0 < \eta < 4^{-1}$, K such that (i) $n^{-1}\log(pq) \leq \eta$, and (ii) $\mathrm{E}(e^{tx_k^2}) \leq K$ for $|t| \leq \eta$ and all $k = 1, \ldots, p$.

Assumption 3.1 requires the matrix X to be well behaved. Assumption 3.2 requires the smallest nonzero regression coefficient does not converge too fast. Otherwise, it cannot be consistently identified. (See Fan & Peng, 2004 for more discussions.) Assumption 3.3 ensures the exponential convergence rate of arbitrary order moments of $\mathbf{x}$ and $\vec{\varepsilon}$ (Cai et al., 2011) which is superior to the polynomial type counterpart (Ravikumar et al., 2010).

Theorem 3.1: *Under Assumptions* 3.1–3.3, *if* $\rho_k^2 \geq \tau$ *for all* $k \in S$, *then* $\mathrm{P}(S \subset \widehat{S}_{\mathrm{MCC}}) \to 1$ *as* $n \to \infty$.

Theorem 3.1 reveals that for a properly chosen threshold τ, the probability that MCC detects all relevant predictors tends to one.

4. Unified forward regression

In this section, we propose a unified forward regression (UFR) for variable selection. It extends Wang's (2009) forward regression method for the multivariate response case.

Let $M = \{k_1, \ldots, k_t\}$ denote a generic subset of F with $|M| = t$. Denote $\mathbf{x}_{(M)} = (x_{k_1}, \ldots, x_{k_t})^\top$ and denote $X_{(M)} = (\mathbf{x}_{1(M)}, \ldots, \mathbf{x}_{n(M)})^\top$ as the subset of X corresponding to M. We first describe a naive forward regression (NFR) method that combines the selected variables obtained by repeatedly applying Wang's (2009) forward regression method to univariate regressions with respect to every response. The procedure is summarized as follows. Initially, set $S_{(j)}^{(0)} = \emptyset$ for $j = 1, \ldots, q$. Perform forward regression with respect to the jth response by iterating the following two steps for $\ell = 1, \ldots, n$.

(i) For every $k \in F \backslash S_{(j)}^{(\ell-1)}$, let $M_{kj}^{(\ell-1)} = S_{(j)}^{(\ell-1)} \bigcup \{k\}$. Compute the sum square of residuals $\mathrm{RSS}_{kj}^{(\ell-1)} = \mathbf{y}_{(j)}^\top (I_n - \widetilde{H}_k^{(\ell-1)})\mathbf{y}_{(j)}$, where $\widetilde{H}_{kj}^{(\ell-1)} = X_{(M_{kj}^{(\ell-1)})} (X_{(M_{kj}^{(\ell-1)})}^\top X_{(M_{kj}^{(\ell-1)})})^{-1} X_{(M_{kj}^{(\ell-1)})}^\top$. Let $a_{\ell j} = \mathrm{argmin}_{k\in F\backslash S_{(j)}^{(\ell-1)}} \mathrm{RSS}_{kj}^{(\ell-1)}$.

(ii) Update $S_{(j)}^{(\ell)} = S_{(j)}^{(\ell-1)} \bigcup \{a_{\ell j}\}$.

The solution path of NFR is obtained by $\{S_{\mathrm{NFR}}^{(\ell)} = \cup_{j=1}^q S_{(j)}^{(\ell)} : \ell = 1, \ldots, n\}$.

Next, we propose the unified forward regression to select predictors by applying a modified forward regression algorithm that makes use of all response variables simultaneously. The procedure is modified from the previous one as follows. Initially, set $S^{(0)} = \emptyset$. Perform a modified forward regression by iterating the following two steps for $\ell = 1, \ldots, n$.

(i) For every $k \in F \backslash S^{(\ell-1)}$, let $M_k^{(\ell-1)} = S^{(\ell-1)} \bigcup \{k\}$. Compute the sum square of residuals $\mathrm{RSS}_k^{(\ell-1)} = \mathrm{tr}\{Y^\top (I_n - \widetilde{H}_k^{(\ell-1)})Y\}$, where $\widetilde{H}_k^{(\ell-1)} = X_{M_k^{(\ell-1)}} (X_{M_k^{(\ell-1)}}^\top X_{M_k^{(\ell-1)}})^{-1} X_{M_k^{(\ell-1)}}^\top$. Let $a_\ell = \mathrm{argmin}_{k\in F\backslash S^{(\ell-1)}} \mathrm{RSS}_k^{(\ell-1)}$.

(ii) Update $S^{(\ell)} = S^{(\ell-1)} \bigcup \{a_\ell\}$.

The solution path of UFR is obtained by $\{S_{\mathrm{UFR}}^{(\ell)} = S^{(\ell)} : \ell = 1, \ldots, n\}$. Notice that both NFR and UFR terminate automatically after n iterations. It is seen that the UFR algorithm makes use of all response variables simultaneously by the trace operator. It has nearly one qth computation cost of NFR.

We show that the proposed UFR method also possesses the sure screening property. Also, we add a few more assumptions to facilitate the development of the theory.

Assumption 4.1: Assume that (i) $\mathbf{x}$ follows elliptically contoured distribution, whose density admits the form $|\Sigma_x|^{-1/2} g\{(\mathbf{x} - \vec{\mu})^\top \Sigma_x^{-1} (\mathbf{x} - \vec{\mu})\}$ with $\vec{\mu} = \mathrm{E}\mathbf{x}$ and $g(\cdot) > 0$, denoted by $\mathrm{EC}(\vec{\mu}, \Sigma_x, g)$, and that (ii) the distribution of $\vec{\epsilon}$ is normal.

Assumption 4.2: There exist positive constants ξ, ξ_0 and ν such that (i) $\log(p) \leq \nu n^{\xi}$, (ii) $p_0 \leq \nu n^{\xi_0}$, and (iii) $\xi + 6\xi_0 + 12\xi_{\min} < 1$.

Assumption 4.3: The row vectors of B, i.e, $\vec{\beta}_{(k)}$, $k = 1, \ldots, p$, have the same 'all-or-nothing' structure, i.e., the entries of $\vec{\beta}_{(k)}$ are either all zero or none-zero.

Usually, the normality assumption of $\mathbf{x}$ is imposed to facilitate theory development (Fan & Lv, 2008; H. Wang, 2009). Here in Assumption 4.1, we relax it to elliptically contoured distribution and show its sufficiency to obtain Lemma 1 of H. Wang (2009) in Appendix. Assumption 4.1, together with Assumption 3.1, ensures the sparse Riesz assumption (C. Zhang & Huang, 2008) to derive some key inequalities in proving Theorem 4.1. Assumption 4.2 has been popularly assumed in the literature of ultra-high dimensional inference (Fan & Lv, 2008; H. Wang, 2009). It implies that the dimension of the covariates diverges to infinity at an exponential rate (Fan & Lv, 2008). Assumption 4.3 implies that all responses are associated with the same covariates (Turlach et al., 2005). It warrants the row-wise selection of B by UFR in contrast to the element-wise selection by NFR, which enables UFR to reach the sure screening property in fewer steps than NFR.

Define $\mathcal{K} = 2\tau_{\max}\nu C_B^2 \tau_{\min}^{-2} \nu_B^{-4}$, where the factors are defined in Assumptions 3.1, 3.2 and 4.2. Applying Theorem 1 of H. Wang (2009), we can readily get $\mathrm{P}(S \subset S_{\mathrm{NFR}}^{(q\mathcal{K}\nu n^{2\xi_0+4\xi_{\min}})}) \to 1$, i.e., the NFR selects all relevant predictors with high probability after $q\mathcal{K}\nu n^{2\xi_0+4\xi_{\min}}$ steps for the multivariate regression setting. While the following theorem shows that the UFR can do the job in much fewer (one qth) steps.

Theorem 4.1: *Under Assumptions* 3.1–4.3, $\mathrm{P}(S \subset S_{\mathrm{UFR}}^{(K\nu n^{2\xi_0+4\xi_{\min}})}) \to 1$ *as* $n \to \infty$.

We adopt the BIC criteria to select the best subset of variables from a solution path (Liang et al., 2012; H. Wang, 2009). Let

$$\mathrm{BIC}(M) = \log\left[n^{-1}\mathrm{tr}\left\{Y^\top (I_n - H_{(M)})Y\right\}\right] + n^{-1}|M|(\log n + 2\log p), \quad (4)$$

where $H_{(M)} = X_{(M)}\{X_{(M)}^\top X_{(M)}\}^{-1}X_{(M)}^\top$. We then choose the subset of the variables from the solution path which minimizes $\mathrm{BIC}(M)$. The selection consistency was showed by H. Wang (2009) and Sofer et al. (2014).

Note that the UFR method is consistent if $p = O(n^\alpha)$ for some $\alpha > 0$, while the MCC method works with $\log(p) = O(n^\alpha)$. In this sense, the MCC method can handle higher dimensional variable screening than the UFR method. Secondly, the UFR method is computationally more expensive than the MCC method as the former involves $n-1$ more times of matrix inversion operation than the latter. On the other hand, when p and n are of the same order, both MCC and UFR perform well in terms of coverage probability and UFR performs better in yielding a parsimonious model with high specificity in terms of model size and correct fit (defined later) as seen in simulation.

5. Simulation

We conduct numerical studies to investigate the finite sample performance of the proposed methods, i.e., MCC and UFR, in comparison with the naive correlation coefficient (NCC) method and the naive forward regression (NFR).

5.1. Models

Consider five models for generating the p-dimensional covariates $\mathbf{x}$ in Table 1, which are adopted from Examples 1 and 2 of Fan and Lv (2008), Example 1 of Tibshirani (1996), and Examples 4 and 5 of H. Wang (2009), respectively.

For models 1 to 3, $\mathbf{x}$ follows a multivariate normal distribution with zero mean vector and covariance matrix Σ_x of the structure of identity, autoregressive and compound symmetry, respectively. In model 4, $\mathbf{x}$ is generated by $x_r = (z_r + w_r)/\sqrt{2}$ for $r = 1, \ldots, p_0$ and $x_r = (z_r + \sum_{r'=1}^{p_0} z_{r'})/2$ for $r = p_0 + 1, \ldots, p$, where z_r and w_r are independent standard normal variables (H. Wang, 2009). Note that model 4 is a challenging case as the correlation coefficient of the relevant predictors and the response variables are much smaller than the correlation coefficient of irrelevant predictors and the response variables. The details of Σ_x are provided in Table 1. In model 5, we generate independent components of both $\mathbf{x}$ and $\vec{\epsilon}$ to be $e-1$ where e is an exponential random variable with parameter one. This model is used to examine the robustness of the proposed methods against the departure from the normality assumption. Consider the number of predictors p to be 1000, 5000, and 10,000, respectively, which are all much larger than the sample sizes considered in the five models. Recall p_0 is the number of relevant predictors. Denote the first p_0 rows of B by B_0. We generate independent entries of B_0 from distributions given in the last column of Table 1, where $N(4\log(n)/\sqrt{n}, 1)$ is a normal random variable with mean $4\log(n)/\sqrt{n}$ and variance 1, $\Gamma(2,1)$ denotes a random variable of gamma distribution with shape parameter 2 and scale parameter 1, and exp(9) is an exponential random variable with parameter 9. They are all independent with $\mathbf{x}$. Set the remaining entries (the last $p - p_0$ rows) of B to be zero.

Table 1. Five models.

Model	n	q	p_0	Σ_x	Entries of B_0
1	200	4	8	I_p	$N(4\log(n)/\sqrt{n}, 1)$
2	75	5	3	$0.5I_p + 0.5\mathbf{1}_p\mathbf{1}_p^\top$	$N(4\log(n)/\sqrt{n}, 1)$
3	200	3	3	$\sigma_{kr} = 0.5^{\lvert k-r\rvert}, 1 \le k, r \le p$	$\Gamma(2,1)$
4	300	2	5	$\mathrm{diag}\{I_{p_0}, 4^{-1}(I_{p-p_0} + p_0\mathbf{1}_{(p-p_0)}\mathbf{1}_{(p-p_0)}^\top)\}$	exp(9)
5	200	6	10	I_p	exp(9)

Table 2. Measures for the finite sample performance of variable selection.

Model size	$\mathrm{MS} = \lvert\widehat{S}\rvert$
Coverage probability	$\mathrm{CP} = \mathrm{P}(S \subset \widehat{S})$
% of correctly fitted model	$\mathrm{CF} = \mathrm{P}(S = \widehat{S})$
% of correct zero	$\mathrm{CZ} = (p - p_0)^{-1}\lvert S^c \cap \widehat{S}^c\rvert$
% of incorrect zero	$\mathrm{IZ} = p_0^{-1}\lvert S \cap \widehat{S}^c\rvert$

For the multivariate response case, the signal-to-noise ratio is given by

$$R^2 = \frac{\mathrm{tr}\{\mathrm{var}(B^\top \mathbf{x})\}}{\mathrm{tr}\{\mathrm{var}(\mathbf{y})\}} = \frac{\mathrm{tr}(B^\top \Sigma_x B)}{p\sigma^2 + \mathrm{tr}(B^\top \Sigma_x B)}.$$

We chose the values of σ^2 such that the signal-to-noise ratios are 30%, 60%, and 90%, respectively.

Throughout, set the number of replications N to be 1000.

5.2. Evaluation criteria

For MCC screening, we use Fan and Lv's (2008) hard threshold method to retain the relevant predictors. For both NFR and UFR, we use the BIC criterion (4) to determine the relevant predictors.

We adopt five measures as described in Table 2 to evaluate the finite sample performance of the proposed methods, where the model size (MS) is the number of the selected relevant predictors, the coverage probability (CP) measures how likely all the relevant predictors are identified, the percentage of correctly fitted (CF) measures the capability in identifying the true model correctly, the correct zero (CZ) characterizes the capability in producing sparse solution, and the incorrect zero (IZ) characterizes the method's under-fitting effects. Ideally, we wish a method to have MS close to p_0, CP, CF, CZ all close to 100% and IZ close to zero.

For $b = 1, \ldots, N$, let $\widehat{B}^{(b)}$ denote the estimate of B under the bth replication. The corresponding selected model is denoted by $\widehat{S}^{(b)} = \{k : \widehat{\beta}^{(b)}_{(k)} \neq \mathbf{0}, k = 1, \ldots p\}$. The empirical MS is computed as $\mathrm{MS} = N^{-1}\sum_{b=1}^{N} \lvert\widehat{S}^{(b)}\rvert$ and the empirical values of the other measures are similarly computed.

5.3. Results

Tables 3–7 report the finite sample performance of the four competing methods in terms of the measures

Table 3. Five measures of the performance of variable selection defined in Table 2 obtained by the four competing methods under various numbers of covariates (p) and signal-to-noise ratio (R^2) for Model 1 in Table 1 with $(n, q, p_0) = (200, 4, 8)$.

Method	p	R^2 (%)	MS	CP (%)	CF (%)	CZ (%)	IZ (%)
MCC	1000	30	38	75.9	0	98.4	24.8
		60	38	96.7	0	98.6	3.3
		90	38	99.5	0	98.6	0.5
	5000	30	38	58.2	0	99.6	41.8
		60	38	93.2	0	99.7	6.8
		90	38	98.5	0	99.7	1.5
	10,000	30	38	50.8	0	99.8	49.2
		60	38	90.6	0	99.9	9.4
		90	38	98.1	0	99.9	1.9
NCC	1000	30	38	26.2	0	97.9	73.8
		60	38	40.6	0	98.9	59.2
		90	38	50.5	0	98.2	49.5
	5000	30	38	13.9	0	99.6	86.1
		60	38	30.7	0	99.6	69.3
		90	38	41.8	0	99.6	58.2
	10,000	30	38	10.9	0	99.8	89.1
		60	38	25.6	0	99.8	74.4
		90	38	37.9	0	99.8	62.0
NFR	1000	30	1.1	1.5	0	99.9	98.5
		60	8.1	12.9	0	99.9	87.1
		90	8.3	99.7	90.9	99.9	0.3
	5000	30	2.1	0.5	0	99.9	99.4
		60	6.5	6.2	0	99.9	93.8
		90	7.8	95.3	76.9	100	4.7
	10,000	30	3.1	0.4	0	100	99.6
		60	7.4	4.6	0	100	95.4
		90	10.3	87.9	65.8	100	12.1
UFR	1000	30	0.3	4.1	0	100	95.9
		60	7.6	94.9	71.9	100	5.1
		90	8.0	100	100	100	0
	5000	30	0.1	1.1	0	100	98.9
		60	6.9	86.4	49.9	100	13.6
		90	8.0	100	99.9	100	0
	10,000	30	0.1	0.7	0	100	99.3
		60	5.9	73.8	34.2	100	26.1
		90	8.0	100	99.9	100	0

Table 4. Five measures of the performance of variable selection defined in Table 2 obtained by the four competing methods under various numbers of covariates (p) and signal to noise ratio (R^2) for Model 2 in Table 1 with $(n, q, p_0) = (75, 5, 3)$.

Method	p	R^2 (%)	MS	CP (%)	CF (%)	CZ (%)	IZ (%)
MCC	1000	30	18	80.7	0	98.4	19.3
		60	18	98.1	0	98.5	1.9
		90	18	99.7	0	98.5	0.3
	5000	30	18	68.1	0	99.7	31.9
		60	18	95.5	0	99.7	4.5
		90	18	99.6	0	99.7	0.4
	10,000	30	18	61.3	0	99.8	38.7
		60	18	95.1	0	99.8	4.9
		90	18	99.7	0	99.9	0.3
NCC	1000	30	18	79.7	0	96.5	23.3
		60	18	95.8	0	96.5	4.7
		90	18	98.9	0	96.5	1.4
	5000	30	18	68.9	0	99.3	39.1
		60	18	94.3	0	99.3	7.3
		90	18	98.7	0	99.3	3.4
	10,000	30	18	68.1	0	99.6	51.9
		60	18	89.9	0	99.6	10.1
		90	18	95.3	0	99.7	2.1
NFR	1000	30	1.6	7.9	0.1	99.9	92.1
		60	3.2	51.6	23.7	99.9	48.4
		90	3.4	99.2	93.2	99.9	0.8
	5000	30	1.2	3.5	0	99.9	96.5
		60	3.9	35.6	11.3	99.9	64.4
		90	4.6	97.8	88.1	99.9	2.2
	10,000	30	1.6	2.3	0	99.9	97.4
		60	3.8	28.6	6.7	99.9	71.3
		90	4.7	96.4	84.7	99.9	3.6
UFR	1000	30	1.5	45.6	11.7	99.9	54.4
		60	3.0	99.8	99.1	100	0.2
		90	3.0	100	100	100	0
	5000	30	1.0	28.8	2.4	99.9	71.2
		60	2.9	98.9	97.5	100	1.0
		90	3.0	100	100	100	0
	10,000	30	0.9	24.4	2.0	99.9	75.6
		60	3.0	98.6	95.6	100	1.4
		90	3.0	100	100	100	0

Table 5. Five measures of the performance of variable selection defined in Table 2 obtained by the four competing methods under various numbers of covariates (p) and signal-to-noise ratio (R^2) for Model 3 in Table 1 with $(n, q, p_0) = (200, 3, 3)$.

Method	p	R^2 (%)	MS	CP (%)	CF (%)	CZ (%)	IZ (%)
MCC	1000	30	38	100	0	96.5	0
		60	38	100	0	96.5	0
		90	38	100	0	96.5	0
	5000	30	38	100	0	99.3	0
		60	38	100	0	99.3	0
		90	38	100	0	96.3	0
	10,000	30	38	99.9	0	99.6	0.1
		60	38	100	0	99.6	0
		90	38	100	0	99.6	0
NCC	1000	30	38	29.5	0	98.3	70.5
		60	38	47.8	0	98.3	52.2
		90	38	54.8	0	98.4	45.2
	5000	30	38	18.2	0	99.7	81.8
		60	38	38.1	0	99.6	61.9
		90	38	52.0	0	99.7	48.0
	10,000	30	38	13.3	0	99.8	86.7
		60	38	36.0	0	99.8	63.9
		90	38	50.6	0	99.8	49.4
NFR	1000	30	2.2	99.9	0.2	100	60.1
		60	3.0	67.8	17.8	99.9	32.2
		90	4.7	90.6	71.6	99.9	9.4
	5000	30	2.2	37.0	0	100	63.0
		60	3.0	65.0	13.7	100	35.0
		90	3.7	89.8	69.2	100	10.2
	10,000	30	2.1	36.2	0.1	100	63.8
		60	3.9	63.1	11.7	100	36.9
		90	5.7	89.2	67.4	100	10.8
UFR	1000	30	2.1	71.5	22.6	100	28.5
		60	2.9	96.4	89.1	100	3.6
		90	3.0	99.9	99.6	100	0
	5000	30	2.0	67.2	14.4	100	32.8
		60	2.8	94.1	82.3	100	5.9
		90	3.0	99.9	99.7	100	0.1
	10,000	30	1.9	63.7	10.9	100	36.3
		60	2.8	93.6	80.6	100	6.4
		90	3.0	99.9	99.5	100	0

Table 6. Five measures of the performance of variable selection defined in Table 2 obtained by the four competing methods under various numbers of covariates (p) and signal to noise ratio (R^2) for Model 4 in Table 1 with $(n, q, p_0) = (300, 2, 5)$.

Method	p	R^2 (%)	MS	CP (%)	CF (%)	CZ (%)	IZ (%)
MCC	1000	30	53	21.7	0	97.1	78.3
		60	53	27.8	0	97.1	72.2
		90	53	37.7	0	97.2	79.7
	5000	30	53	19.8	0	99.4	80.2
		60	53	26.7	0	99.4	73.1
		90	53	36.1	0	99.4	63.9
	10,000	30	53	17.7	0	99.7	82.3
		60	53	25.2	0	99.7	74.8
		90	53	35.1	0	99.7	64.9
NCC	1000	30	53	19.8	0	97.1	80.2
		60	53	20.7	0	97.1	79.2
		90	53	21.8	0	97.1	78.1
	5000	30	53	16.8	0	99.4	83.1
		60	53	18.6	0	99.4	81.3
		90	53	19.7	0	99.4	80.3
	10,000	30	53	16.2	0	99.7	83.8
		60	53	17.3	0	99.7	82.7
		90	53	20.0	0	99.7	80.0
NFR	1000	30	2.7	15.6	0	99.9	84.4
		60	4.9	34.3	0	99.9	65.7
		90	5.0	65.1	5.0	99.9	34.9
	5000	30	2.6	13.2	0	99.9	86.8
		60	4.8	27.5	0	99.9	72.5
		90	5.5	57.8	3.1	99.9	42.3
	10,000	30	2.0	12.1	0	99.9	87.9
		60	4.8	26.6	0	99.9	73.4
		90	6.5	54.7	1.9	99.9	45.3
UFR	1000	30	2.4	35.9	0	99.9	64.1
		60	4.4	75.8	10.3	99.9	24.2
		90	5.0	96.8	70.8	99.9	3.2
	5000	30	2.1	28.9	0	99.9	71.1
		60	4.2	70.9	7.2	99.9	29.1
		90	5.0	95.5	63.2	99.9	4.5
	10,000	30	2.0	24.7	0	99.9	75.3
		60	4.1	69.3	4.9	99.9	30.7
		90	5.0	95.6	61.9	99.9	4.4

given in Table 2 under various numbers of covariates p and signal strength R^2. We summarize the findings as follows. (i) The MCC method is uniformly superior to the NCC method with larger coverage probability (CP), better estimation of sparsity (with larger CZ and smaller IZ), as expected. (ii) As we adopted the fixed threshold procedure for MCC and NCC, these two methods produce conservatively large coverage of predictors at the cost of large model size. For the same reason, the percentage of incorrect zero is larger than the other two regression-based methods (UFR and NFR). So the resulting percentages of correctly fitted models for MCC and NCC are zero. (iii) When comparing UFR with NFR, the UFR demonstrates its superiority over NFR uniformly in all five measures across all five models (including Model 5 with the non-normal distribution). This corroborates the advantage of UFR in utilizing the correlation within responses over NFR. When comparing UFR with PWL, both methods perform comparably when the signal strength is as small as 30%. When the signal strength is as large as 60% or 90%, UFR outperforms PWL in all five measures in general. (iv) The UFR method performs inferior to the MCC method in cases of ultra-high dimensional covariates especially under lower signal strength, as pointed out earlier. For instance, in Model 1 of Table 3, the coverage probability of UFR reduces by 83%, while the counterpart of MCC reduces by 33% when the dimension of predictors p increases from 1000 to 10,000 at the signal strength of 30%. (v) As for the impact of the signal strength, the percentage of incorrect zeros rises under the weak signal strength cases from those under the strong signal strength cases. It is consistent with the findings for the univariate case in H. Wang (2009) and Y. Li et al. (2017). However, as the signal strength increases (e.g., from 30% to 90%), the percentages of coverage probability (CP) and probability of correct fit (CF) increase significantly (e.g., 61.9% to 98.3% and 28.8% to 58.8%, respectively, with $p = 5000$) and the percentage of incorrect zeros (IZ) drops quickly (e.g., from 53.7% to 2.35% with $p = 5000$) by both NFR and UFR as seen in Table 3. (vi) To examine the impact of the sample size, Table 8 reports the performance of the proposed methods under Model 1 with a number of covariates p fixed at 5000 and varying sample size n to be 100, 200, and 400, respectively. It is seen that

Table 7. Five measures of the performance of variable selection defined in Table 2 obtained by the four competing methods under various numbers of covariates (p), and signal-to-noise ratio (R^2) for Model 5 in Table 1 with $(n, q, p_0) = (200, 6, 10)$.

Method	p	R^2 (%)	MS	CP (%)	CF (%)	CZ (%)	IZ (%)
MCC	1000	30	38	85.2	0	97.0	14.8
		60	38	97.2	0	97.1	2.8
		90	38	99.8	0	97.2	0.2
	5000	30	38	70.1	0	99.4	29.9
		60	38	93.3	0	99.4	6.7
		90	38	99.5	0	99.4	0.6
	10,000	30	38	63.0	0	99.7	37.0
		60	38	90.6	0	99.7	9.4
		90	38	99.3	0	99.7	0.8
NCC	1000	30	38	61.2	0	96.8	38.8
		60	38	78.8	0	96.9	21.2
		90	38	85.8	0	97	14.2
	5000	30	38	40.8	0	99.3	59.2
		60	38	64.9	0	99.4	24.3
		90	38	75.7	0	99.4	24.3
	10,000	30	38	33.0	0	99.6	66.9
		60	38	60.0	0	99.7	39.9
		90	38	72.5	0	99.7	27.5
NFR	1000	30	3.8	7.6	0	99.9	92.4
		60	10.2	50.9	0.2	99.9	49.0
		90	10.7	96.5	63.9	99.9	3.5
	5000	30	2.5	4.3	0	99.9	95.7
		60	6.4	40.2	0.1	99.9	59.8
		90	10.6	94.6	49.9	99.9	5.4
	10,000	30	2.4	3.3	0	99.9	96.7
		60	10.6	35.3	0.1	99.9	64.7
		90	11.6	93.9	44.3	99.9	6.1
UFR	1000	30	2.7	27.3	0	100	72.7
		60	9.9	99.3	92.8	100	0.7
		90	10.0	100	99.9	100	0
	5000	30	1.3	13.3	0	100	86.7
		60	4.1	98.2	84.5	100	1.8
		90	10.0	100	99.9	100	0
	10,000	30	1.0	10.1	0	100	89.9
		60	9.8	97.8	81.3	100	2.2
		90	10.0	100	99.9	100	0

Table 8. Five measures of the performance of variable selection defined in Table 2 obtained by the four competing methods under various sample sizes (n) and signal-to-noise ratio (R^2) for Model 1 in Table 1 with $(p, q, p_0) = (5000, 4, 8)$.

Method	n	R^2 (%)	MS	CP (%)	CF (%)	CZ (%)	IZ (%)
MCC	100	30	22	39.3	0	99.6	60.3
		60	22	72.6	0	99.7	27.4
		90	22	92.3	0	99.7	7.7
	200	30	38	81.2	0	99.4	18.8
		60	38	95.5	0	99.4	4.5
		90	38	98.9	0	99.4	1.1
	400	30	67	96.7	0	98.8	3.3
		60	67	99.3	0	98.8	0.7
		90	67	99.8	0	98.8	0.1
NCC	100	30	22	24.7	0	99.6	75.3
		60	22	45.9	0	99.6	54.0
		90	22	57.4	0	99.6	42.6
	200	30	38	48.3	0	99.3	51.8
		60	38	65.2	0	99.3	34.8
		90	38	70.9	0	99.4	29.1
	400	30	67	62.1	0	98.8	37.9
		60	67	72.2	0	98.8	27.8
		90	67	76.7	0	98.8	23.3
NFR	100	30	2.1	1.4	0	99.9	98.6
		60	6.2	14.8	0	99.9	85.2
		90	8.3	78.8	14.2	99.9	21.2
	200	30	3.0	10.6	0	99.9	89.4
		60	8.0	50.0	0.2	99.9	49.9
		90	8.9	84.8	24.0	100	15.2
	400	30	7.7	33.3	30.0	100	66.6
		60	8.0	65.9	2.2	100	34.1
		90	9.0	86.8	30.7	100	13.1
UFR	100	30	1.0	3.2	0	100	96.8
		60	5.6	69.2	12.0	100	30.8
		90	8.0	99.9	99.1	100	0.1
	200	30	2.8	35.5	0.1	100	64.5
		60	7.6	95.6	69.8	100	4.4
		90	8.0	99.9	99.7	100	0
	400	30	6.4	79.5	12.3	100	20.5
		60	7.9	98.4	88.1	100	1.5
		90	8.0	100	100	100	0

the measures of model size (MS), coverage probability (CP), probability of correct fit (CF) and probability of incorrect zero (IZ) are sensitive to sample size. The improvement of performance is significant. For instance, when the sample size increases from $n = 100$ to $n = 200$ with signal strength $R^2 = 60\%$, the CP increases from 52.2% to 80.4% on average and the percentage of incorrect zero drops from 47.8% to 19.7% on average.

In conclusion, the MCC method performs better when the dimension of covariates is ultra-high ($\log(p) = O(n^\alpha)$) with respect to the sample size and the UFR method outperforms the MCC method when the dimension of covariates is of polynomial order ($p = O(n^\alpha)$).

6. Real data application

We apply the proposed methods to a real data set regarding bone mineral density (BMD) (Reppe et al., 2010). The data were collected from 84 postmenopausal Caucasian women aged from 50 to 86. For each subject, there are two responses, namely the body mass index and total hip z-score (a measure of how strong the bone in the hip), and 8649 gene expression levels in trans-iliacal bone biopsies served as covariates. It is known that low bone mineral density is usually related to fragile bone and osteoporosis and progressive reduction of bone strength which leads to increasing susceptibility of bone fractures (Cooper, 1997; Reppe et al., 2010). The goal of the study is to identify the genes that are related to BMD.

Table 9 reports the genes identified by the five competing methods. The MCC method identified 19 genes which include all 13 genes identified by NFR except gene TNK2. The PWL method identified 12 genes which all identified by NFR except PAIP1. And the UFR found 10 significant genes which are all contained in the set identified by NFR.

To examine the quality of variable selection of these methods, we compare the prediction mean square error ($\mathrm{E}\|\mathbf{y} - B_S^\top \mathbf{x}_S\|^2$) obtained by the three methods. To this end, we randomly split the data into a training set of 60 samples and a testing set of the remaining 24 samples. The average prediction mean square errors over the 100 replications for MCC, NCC, NFR and UFR are 273.7, 293.5, 271.0 and 241.6, respectively. Clearly, the UFR method is the winner. All the

Table 9. Selected genes for the BMD data.

method	genes
MCC	ACSL3, NIPSNAP3B, DLEU2, C1ORF61, DKK1, SOST, ABCA8, AFFX-M27830-M-at, RRNF216, PLIN5, PACS2, MUM1, CRYGS, RABEP2, PEX14, USP2, FKBP14, FAM55C, MAPK8
NCC	ACSL3, NIPSNAP3B, DLEU2, C1ORF61, DKK1, SOST, ABCA8, AFFX-M27830-M-at, RRNF216, PLIN5, MEPE, COPS4, CRYGS RABEP2, CTSE, PEX14, FKBP14, NARG1, PAIP1
NFR	ACSL3, NIPSNAP3B, DLEU2, C1ORF61, DKK1, SOST, ABCA8, AFFX-M27830-M-at, RNF216, PLIN5, FKBP14, FAM55C, TNK2
UFR	ACSL3, NIPSNAP3B, DLEU2, C1ORF61, DKK1, SOST, ABCA8, AFFX-M27830-M-at, RNF216, PLIN5

eight genes (ACSL3, NIPSNAP3B, DLEU2, C1ORF61, DKK1, SOST, ABCA8, and AFFX-M27830-M-at) identified by Reppe et al. (2010) were selected by the four competing methods. The UFR method discovered two more genes, RNF216 and PLIN5, with the smallest prediction mean square errors. Similar to the findings in simulation, both MCC and NCC selected more genes than NFR and UFR with larger prediction error.

7. Conclusion

We propose two methods for variable screening in high-dimensional multivariate linear regression. The MCC method has the advantage of computational ease and can provide fast variable screening to obtain an accurate subset with a dimension below the ample size. The proposed UFR method has the feature of discovering all relevant predictors consistently at nearly the same computational cost as the univariate forward regression. The performance of UFR is sensitive to the dimensionality and signal strength. Our theory assumes Gaussian distribution for the response variables. The numerical study also shows the robustness of the proposed methods against non-normality. It is of interest to investigate the problem under more general non-homogeneously sparse assumption and nonlinear models.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

Anderson, T. (2003). *An introduction to statistical multivariate analysis* (3rd ed.). Wiley.

Bickel, P. J., & Levina, E. (2008). Regularized estimation of large covariance matrices. *The Annals of Statistics*, *36*(1), 199–227. https://doi.org/10.1214/009053607000000758

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics*, *37*(4), 373–384. https://doi.org/10.1080/00401706.1995.10484371

Breiman, L., & Friedman, J. H. (1997). Predicting multivariate responses in multiple linear regression. *Journal of the Royal Statistical Society: Series B*, *59*(1), 3–54. https://doi.org/10.1111/rssb.1997.59.issue-1

Cai, T., Li, H., Liu, W., & Xie, J. (2013). Covariate–adjusted precision matrix estimation with an application in genetical genomics. *Biometrika*, *100*(1), 139–156. https://doi.org/10.1093/biomet/ass058

Cai, T., Liu, W., & Luo, X. (2011). A constrained l_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, *106*(494), 594–607. https://doi.org/10.1198/jasa.2011.tm10155

Cai, T., & Lv, J. (2007). Discussion: The Dantzig selector: Statistical estimation when *p* is much larger than *n*. *Annals of Statistics*, *35*(6), 2365–2369. https://doi.org/10.1214/009053607000000442

Candes, E., & Tao, T. (2007). The Dantzig selector: Statistical estimation when *p* is much larger than *n*. *Annals of Statistics*, *35*(6), 2313–2351. https;//doi.org/10.1214/009053606000001523

Chen, L., & Huang, J. Z. (2012). Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *Journal of the American Statistical Association*, *107*(500), 1533–1545. https://doi.org/10.1080/01621459.2012.734178

Cooper, C. (1997). The crippling consequences of fractures and their impact on quality of life. *American Journal of Medicine*, *103*(2), 12–19. https://doi.org/10.1016/S0002-9343(97)90022-X

Deshpande, S., Rockova, V., & George, E. (2019). Simultaneous variable and covariance selection with the multivariate Spike- and Slab lasso. *Journal of Computational and Graphical Statistics*, *28*(4), 921–931. https://doi.org/10.1080/10618600.2019.1593179

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, *32*(2), 407–499. https://doi.org/10.1214/009053604000000067

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*(456), 1348–1360. https://doi.org/10.1198/016214501753382273

Fan, J., & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion). *Journal of the Royal Statistical Society: Series B*, *70*(5), 849–911. https://doi.org/10.1111/rssb.2008.70.issue-5

Fan, J., & Peng, H. (2004). Non-concave penalized likelihood with a diverging number of parameters. *Annals of Statistics*, *32*(3), 928–961. https://doi.org/10.1214/009053604000000256

Fang, K.-T., Kotz, S., & Ng, K. W. (2018). *Symmetric multivariate and related distributions*. Chapman and Hall/CRC.

Ferte, C., Trister, A. D., Erich, H., & Bot, B. (2013). Impact of bioinformatic procedures in the development and translation of high-throughput molecular classifiers in oncology. *Clinical Cancer Research*, *19*(16), 4315–4325. https://doi.org/10.1158/1078-0432.CCR-12-3937

Frank, L., & Friedman, J. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, *35*(2), 10–135. https://doi.org/10.1080/00401706.1993.10485033

Fu, W. J. (1998). Penalized regressions: The bridge versus the lasso. *Journal of Computational and Graphical Statistics*, *7*(3), 397–416. https://doi.org/10.1080/10618600.1998.10474784

He, K., Lian, H., Ma, S., & Huang, J. Z. (2018). Dimensionality reduction and variable selection in multivariate varying-coefficient models with a large number of covariates. *Journal of Statistical Planning and Inference*, *113*(522), 746–754. https://doi.org/10.1080/01621459.2017.1285774

Jia, B., Xu, S., Xiao, G., & Lambda, V. (2017). Learning gene regulatory networks from next generation sequencing

data. *Biometrics*, *73*(4), 1221–1230. https://doi.org/10.1111/biom.v73.4

Kim, S., Sohn, K., & Xing, E. (2009). A multivariate regression approach to association analysis of a quantitative trait network. *Bioinformatics (Oxford, England)*, *25*(12), 204–212. https://doi.org/10.1093/bioinformatics/btp218

Kong, X., Liu, Z., Yao, Y., & Zhou, W. (2017). Sure screening by ranking the canonical correlations. *Test*, *26*(1), 46–70. https://doi.org/10.1007/s11749-016-0497-z

Kong, Y., Zheng, Z., & Lv, J. (2016). The constrained Dantzing selector with enhanced consistency. *Journal of Machine Learning Research*, *17*(123), 1–22.

Lee, W., & Liu, Y. (2012). Simultaneous multiple response regression and inverse covariate matrix estimation via penalized Gaussian maximum likelihood. *Journal of Multivariate Analysis*, *111*, 241–255. https://doi.org/10.1016/j.jmva.2012.03.013

Li, B., Chuns, H., & Zhao, H. (2012). Sparse estimation of conditional graphical models with application to gene networks. *Journal of American Statistical Association*, *107*(497), 152–167. https://doi.org/10.1080/01621459.2011.644498

Li, C., & Li, H. (2008). Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics (Oxford, England)*, *24*(9), 1175–1182. https://doi.org/10.1093/bioinformatics/btn081

Li, G., Peng, H., Zhang, J., & Zhu, L. (2012). Robust rank correlation based screening. *Annals of Statistics*, *40*(3), 1846–1877. https://doi.org/10.1214/12-AOS1024

Li, Y., Li, G., Lian, H., & Tong, T. (2017). Profile forward regression screening for ultra-high dimensional semiparametric varying coefficient partially linear models. *Journal of Multivariate Analysis*, *155*, 133–150. https://doi.org/10.1016/j.jmva.2016.12.006

Li, Y., Nan, B., & Zhu, J. (2015). Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure. *Biometrics*, *71*(2), 354–363. https://doi.org/10.1111/biom.v71.2

Liang, H., Wang, H., & Tsai, C.-L. (2012). Profiled forward regression for ultrahigh dimensional variable screening in semiparametric partially linear model. *Statistica Sinica*, *22*(2), 531–554. https://doi.org/10.5705/ss.2010.134

Obozinski, G., Wainwright, M. J., & Jordan, M. I. (2011). Support union recovery in high-dimensional multivariate regression. *Annals of Statistics*, *39*(1), 1–47. https://doi.org/10.1214/09-AOS776

Pecanka, J., van der Vaart, A. W., & Marianne, J. (2019). Modeling association between multivariate correlated and high-dimensional sparse covariates: The adaptive SVS method. *Journal of Applied Statistics*, *46*(5), 893–913. https://doi.org/10.1080/02664763.2018.1523377

Peng, J., Zhu, J., Bergamaschi, A., Han, W., Noh, D.-Y., Pollack, J. R., & Wang, P. (2010). Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *Annals of Applied Statistics*, *4*(1), 53–77. https://doi.org/10.1214/09-AOAS271

Ravikumar, P., Wainwright, M., & Lafferty, J. (2010). High-dimensional Ising model selection using l_1 regularized logistic regression. *Annals of Statistics*, *38*(3), 1287–1319. https://doi.org/10.1214/09-AOS691

Ren, J., Du, Y., Li, S., Ma, S., Jiang, Y., & Wu, C. (2019). Robust network-based regularization and variable selection for high-dimensional genomic data in cancer prognosis. *Genetic Epidemiology*, *43*(3), 276–291. https://doi.org/10.1002/gepi.2018.43.issue-3

Ren, J., He, T., Li, Y., Liu, S., Du, Y., Jiang, Y., & Wu, C. (2017). Network-based regularization for high dimensional SNP data in the case-control study of type 2 diabetes. *BMC Genetics*, *18*(1), 44. https://doi.org/10.1186/s12863-017-0495-5

Reppe, S., Refvem, H., Gautvik, V. T., Olstad, O. K., H∅vring, P. I., Reinholt, F. P., Holden, M., Frigessi, A., Jemtland, R., & Gautvik, K. M. (2010). Eight genes are highly associated with BMD variation in postmenopausal Caucasian women. *Bone*, *46*(3), 604–612. https://doi.org/10.1016/j.bone.2009.11.007

Rothman, A. J., Levina, E., & Zhu, J. (2010). Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics*, *19*(4), 947–962. https://doi.org/10.1198/jcgs.2010.09188

Saulis, L., & Statulevicius, V. (1991). *Limit theorems for large deviations* (Vol. 73). Springer Science & Business Media.

Setdji, C. M., & R. D. Cook (2004). *K*-means inverse regression. *Technometrics*, 46(4), 421–429. https://doi.org/10.1198/004017004000000437

Smith, M., & Fahrmeir, L. (2007). Spatial Bayesian variable selection with application to functional magnetic resonance imaging. *Journal of American Statistical Association*, *102*(478), 417–431. https://doi.org/10.1198/016214506000001031

Sofer, T., Dicker, L., & Lin, X. (2014). Variable selection for high dimensional multivariate outcomes. *Statistica Sinica*, *24*(4), 1633–1654. https://doi.org/10.5705/ss.2013.019

Song, Y., Schreier, P. J., Ramirez, D., & Hasija, T. (2016). Canonical correlation analysis of high-dimensional data with very small sample support. *Signal Processing*, *128*, 449–458. https://doi.org/10.1016/j.sigpro.2016.05.020

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Turlach, B., Venables, W., & Wright, S. (2005). Simultaneous variable selection. *Technometrics*, *47*(3), 349–363. https://doi.org/10.1198/004017005000000139

Wang, H. (2009). Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association*, *104*(488), 1512–1524. https://doi.org/10.1198/jasa.2008.tm08516

Wang, J., Zhang, Z., & Ye, J. (2019). Two-layer feature reduction for sparse-group lasso via decomposition of convex sets. *Journal of Machine Learning*, *20*(163), 1–42.

Yang, Y., & Zou, H. (2015). A fast unified algorithm for solving group-lasso penalize learning problems. *Statistics and Computing*, *25*(6), 1129–1141. https://doi.org/10.1007/s11222-014-9498-5

Yi, N. (2010). Statistical analysis of genetic interactions. *Genetics Research*, *92*(5–6), 443–459. https://doi.org/10.1017/S0016672310000595

Yin, J., & Li, H. (2011). A sparse conditional Gaussian graphical model for analysis of genetical genomics data. *Annals of Applied Statistics*, *5*(4), 2630–2650. https://doi.org/10.1214/11-AOAS494

Zhang, C., & Huang, J. (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. *Annals of Statistics*, *36*(4), 1567–1594. https://doi.org/10.1214/07-AOS520

Zhang, H. H., & Lu, W. (2007). Adaptive lasso for Cox's proportional hazards model. *Biometrika*, *94*(3), 691–703. https://doi.org/10.1093/biomet/asm037

Zhang, N., Yu, Z., & Wu, Q. (2019). Overlapping sliced inverse regression for dimension reduction. *Analysis and Applications*, *17*(5), 715–736. https://doi.org/10.1142/S0219530519400013

Zhao, W., Lian, H., & Ma, S. (2017). Robust reduced–rank modeling via rank regression. *Journal of Statistical Planning and Inference*, *180*, 1–12. https://doi.org/10.1016/j.jspi.2016.08.009

Zhu, L., Li, L., Li, R., & Zhu, L. (2011). Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, *106*(496), 1464–1475. https://doi.org/10.1198/jasa.2011.tm10563

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418–1429. https://doi.org/10.1198/016214506000000735

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, *67*(2), 301–320. https://doi.org/10.1111/rssb.2005.67.issue-2

Appendix

A.1 Proof of Theorem 3.1

Let $C_K = \tau + \eta^{-1}K$ and $\vartheta = 3\eta^{-1}C_K\sqrt{n\log(pq)}$. When $\rho_k^2 \geq \tau$ for all $k \in S$, it suffices to show that under Assumptions 3.1–3.3

$$\begin{aligned}
&\mathrm{P}\left(\max_{1\leq k\leq p}|\widehat{\rho}_k^2 - \rho_k^2| \geq \vartheta\right) \\
&\quad = \mathrm{P}\left(\max_{1\leq k\leq p}\left|\frac{1}{n}\sum_{i=1}^{n}\widehat{\vec{\gamma}}_k^{\top}\mathbf{y}_i x_{ik} - \mathrm{E}\vec{\gamma}_k^{\top}\mathbf{y}x_k\right| \geq \vartheta\right) \\
&\quad < 3p^{-\tau}q^{-\tau}. \qquad \text{(A1)}
\end{aligned}$$

Observe that

$$\begin{aligned}
&|\widehat{\rho}_k^2 - \rho_k^2| \\
&\quad = \left|(\widehat{\vec{\gamma}}_k - \vec{\gamma}_k)^{\top}\frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i x_{ik} + \vec{\gamma}_k^{\top}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i x_{ik} - \mathrm{E}\mathbf{y}x_k\right)\right| \\
&\quad \leq \left|(\widehat{\vec{\gamma}}_k - \vec{\gamma}_k)^{\top}\frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i x_{ik}\right| + \left|\vec{\gamma}_k^{\top}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i x_{ik} - \mathrm{E}\mathbf{y}x_k\right)\right| \\
&\quad \triangleq A_1 + A_2. \qquad \text{(A2)}
\end{aligned}$$

First, we compute the rate of convergence for A_2. For $k = 1, \ldots, p$, let $\omega_k = n^{-1}\sum_{i=1}^{n}\vec{\gamma}_k^{\top}(\mathbf{y}_i x_{ik} - \mathrm{E}\mathbf{y}x_k)$. (So $A_2 = |\omega_k|$.) Let $t_1 = \eta\sqrt{n^{-1}\log(pq)}$. Applying the inequalities $\mathrm{P}(|U| \geq V) \leq e^{-t_1 V}\mathrm{E}e^{t_1|U|}$ for any $V > 0$ and $|e^s - 1 - s| \leq s^2 e^{\max(s,0)}$ for any $s \in \mathbb{R}$, we obtain

$$\begin{aligned}
&\mathrm{P}\left(A_2 \geq \eta^{-1}C_K\sqrt{n\log(pq)}\right) \\
&\quad \leq e^{-C_K\log(pq)}\mathrm{E}\exp(t_1|\omega_k|) \\
&\quad \leq \exp\{-C_K\log(pq) + t_1^2\mathrm{E}\omega_k^2\exp(t_1|\omega_k|)\} \\
&\quad \leq \exp\{-C_K\log(pq) + \eta^{-1}K\log(pq)\} \\
&\quad = p^{-\tau}q^{-\tau}. \qquad \text{(A3)}
\end{aligned}$$

Second, we compute the rate of convergence for A_1. Notice that

$$\left|(\widehat{\gamma}_k - \vec{\gamma}_k)^{\top}\frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i x_{ik}\right| \leq \left|\widehat{\gamma}_k^{\top}\frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i x_{ik}\right| + \left|\gamma_k^{\top}\frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i x_{ik}\right|. \qquad \text{(A4)}$$

Parallel to (A3), we have

$$\begin{aligned}
&\mathrm{P}\left(\left|\widehat{\gamma}_k^{\top}\frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i x_{ik}\right| \geq \eta^{-1}C_K\sqrt{n\log(pq)}\right) \\
&\quad \leq e^{-C_K\log(pq)}\mathrm{E}\exp\left\{t_1\left(\left|\widehat{\gamma}_k^{\top}\frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i x_{ik}\right|\right)\right\} \\
&\quad \leq \exp\left\{-C_K\log(pq) + t_1^2\mathrm{E}\left(\widehat{\gamma}_k^{\top}\frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i x_{ik}\right)^2\right. \\
&\qquad \left.\times\exp\left(t_1\left|\widehat{\gamma}_k^{\top}\frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i x_{ik}\right|\right)\right\} \\
&\quad \leq \exp\{-C_K\log(pq) + \eta^{-1}K\log(pq)\} \\
&\quad = p^{-\tau}q^{-\tau}. \qquad \text{(A5)}
\end{aligned}$$

Similar to (A5),

$$\mathrm{P}\left(\left|\gamma_k^{\top}\frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i x_{ik}\right| \geq \eta^{-1}C_K\sqrt{n\log(pq)}\right) \leq p^{-\tau}q^{-\tau}. \qquad \text{(A6)}$$

Combining (A4), (A5) and (A6),

$$\begin{aligned}
&\mathrm{P}\left(A_1 = \left|(\widehat{\gamma}_k - \vec{\gamma}_k)^{\top}\frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i x_{ik}\right| \geq 2\eta^{-1}C_K\sqrt{n\log(pq)}\right) \\
&\quad \leq 2p^{-\tau}q^{-\tau}. \qquad \text{(A7)}
\end{aligned}$$

At last, combining (A3) and (A7), we get

$$\mathrm{P}\left(A_1 + A_2 \geq 3\eta^{-1}C_K\sqrt{n\log(pq)}\right) \leq 3p^{-\tau}q^{-\tau},$$

which, coupled with (A2), implies (A1). This completes the proof.

A.2 A lemma used to prove theorem 4.1

Lemma A.1: *Suppose Assumptions* 3.1, 4.1 *and* 4.2 *hold. Let* $\Sigma_{x(M)}$ *denote the submatrix of* Σ_x *after* M. *Let* $\widehat{\Sigma}_{x(M)}$ *and* $\widehat{\Sigma}_x$ *denote the corresponding estimators, respectively. Suppose* $m = O(n^{2\xi_0} + 4\xi_{\min})$, *where* ξ_0 *and* $\xi_{\min}$ *are defined in Assumption* 4.2. *Then, with probability tending to one, we have*

$$\tau_{\min} \leq \min_{|M|\leq m}\lambda_{\min}(\widehat{\Sigma}_{x(M)}) \leq \max_{|M|\leq m}\lambda_{\max}(\widehat{\Sigma}_{x(M)}) \leq \tau_{\max}. \qquad \text{(A8)}$$

Proof: The proof is similar to that for Lemma 1 of H. Wang (2009). Here, we relax the normality assumption of $\mathbf{x}$ to the elliptically contoured distribution.

First, for $i = 1, \ldots, n$, $j, k = 1, \ldots, p$, let $U_i = (z_{ij} + z_{ik})/\sqrt{2(1+\rho_{jk})}$ and $V_i = (z_{ij} - z_{ik})/\sqrt{2(1-\rho_{jk})}$. By Assumption 4.1 and the additive property of elliptical contoured distribution (Fang et al., 2018), $(z_{ij}, z_{ik}) \sim \mathrm{EC}_2(0, 0, 1, 1, \rho_{jk}, g)$, $U_i \overset{\text{i.i.d.}}{\sim} \mathrm{EC}(0, 1, g)$ and $V_i \overset{\text{i.i.d.}}{\sim} \mathrm{EC}(0, 1, g)$.

Second, observe that

$$\begin{aligned}
\sum_{i=1}^{n}(z_{ij}z_{ik} - \rho_{jk}) = \frac{1}{4}\left[\sum_{i=1}^{n}\left\{(z_{ij} + z_{ik})^2 - 2(1+\rho_{jk})\right\}\right. \\
\left. - \sum_{i=1}^{n}\left\{(z_{ij} - z_{ik})^2 - 2(1-\rho_{jk})\right\}\right].
\end{aligned}$$

Then, following Lemma A.3 of Bickel and Levina (2008), we have

$$\begin{aligned}
&\mathrm{P}(|\widehat{\sigma}_{ij}-\sigma_{ij}|\geq nv)\\
&\quad=\mathrm{P}\left(\left|\sum_{i=1}^{n}(x_{ij}x_{ik}-\sigma_{jk})\right|\geq nv\right)\\
&\quad=\mathrm{P}\left(\left|\sum_{i=1}^{n}(z_{ij}z_{ik}-\rho_{jk})\right|\geq \frac{nv}{(\sigma_{jj}\sigma_{kk})^{1/2}}\right)\\
&\quad=\mathrm{P}\left[\left|\sum_{i=1}^{n}\left\{(z_{ij}+z_{ik})^2-2(1+\rho_{jk})\right\}\right.\right.\\
&\qquad\left.\left.-\sum_{i=1}^{n}\left\{(z_{ij}-z_{ik})^2-2(1-\rho_{jk})\right\}\right|\geq\frac{4nv}{(\sigma_{jj}\sigma_{kk})^{1/2}}\right]\\
&\quad\leq\mathrm{P}\left[\left|\sum_{i=1}^{n}\left\{(z_{ij}+z_{ik})^2-2(1+\rho_{jk})\right\}\right|\right.\\
&\qquad\left.+\left|\sum_{i=1}^{n}\left\{(z_{ij}-z_{ik})^2-2(1-\rho_{jk})\right\}\right|\geq\frac{4nv}{(\sigma_{jj}\sigma_{kk})^{1/2}}\right]\\
&\quad\leq\mathrm{P}\left[\left|\sum_{i=1}^{n}\left\{(z_{ij}+z_{ik})^2-2(1+\rho_{jk})\right\}\right|\geq\frac{4nv}{(\sigma_{jj}\sigma_{kk})^{1/2}}\right]\\
&\qquad+\mathrm{P}\left[\left|\sum_{i=1}^{n}\left\{(z_{ij}-z_{ik})^2-2(1-\rho_{jk})\right\}\right|\geq\frac{4nv}{(\sigma_{jj}\sigma_{kk})^{1/2}}\right]\\
&\quad=\mathrm{P}\left\{\left|\sum_{i=1}^{n}(U_i^2-1)\right|\geq\frac{2nv}{(1+\rho_{jk})(\sigma_{jj}\sigma_{kk})^{1/2}}\right\}\\
&\qquad+\mathrm{P}\left\{\left|\sum_{i=1}^{n}(V_i^2-1)\right|\geq\frac{2nv}{(1-\rho_{jk})(\sigma_{jj}\sigma_{kk})^{1/2}}\right\}. \quad \text{(A9)}
\end{aligned}$$

Further, for $i=1,\ldots,n$, let $W_i=U_i^2-1$ and $B_n^2=\sum_{i=1}^{n}\mathrm{var}(W_i)$. Observe that by Jensen inequality there exist positive constants $c_1,\ldots,c_n$, C_3 and C_4 such that

$$\left|\frac{\ln\{\mathrm{E}\exp(\zeta W_i)\}}{\zeta^2}\right|\leq c_i^2 \text{ for } |\zeta|<C_3 \quad \text{and}$$

$$\overline{\lim}_{n\to\infty}\frac{1}{B_n^2}\sum_{i=1}^{n}c_n^2\leq C_4,$$

satisfying condition (P) of Saulis and Statulevicius (1991). The same result holds when $W_i=V_i^2-1$. By Theorem 3.2 of Saulis and Statulevicius (1991), the first and second terms of (A9) are bounded, respectively, by

$$2\exp\left\{-\frac{2nv^2}{(1+\rho_{jk})^2(\sigma_{jj}\sigma_{kk})+2v(1+\rho_{jk})(\sigma_{jj}\sigma_{kk})^{1/2}}\right\}$$

and

$$2\exp\left\{-\frac{2nv^2}{(1-\rho_{jk})^2(\sigma_{jj}\sigma_{kk})+2v(1-\rho_{jk})(\sigma_{jj}\sigma_{kk})^{1/2}}\right\}.$$

Therefore,

$$\begin{aligned}
&\mathrm{P}(|\widehat{\sigma}_{ij}-\sigma_{ij}|\geq nv)\\
&\quad\leq 4\max\left[\exp\left\{\frac{-2nv^2}{(1+\rho_{jk})^2(\sigma_{jj}\sigma_{kk})+2v(1+\rho_{jk})(\sigma_{jj}\sigma_{kk})^{1/2}}\right\},\right.\\
&\qquad\left.\times\ \exp\left\{\frac{-2nv^2}{(1-\rho_{jk})^2(\sigma_{jj}\sigma_{kk})+2v(1-\rho_{jk})(\sigma_{jj}\sigma_{kk})^{1/2}}\right\}\right].
\end{aligned}$$

This, together with $\lambda_{\max}(\Sigma_x)<2^{-1}\tau_{\max}$, implies that

$$\mathrm{P}(|\widehat{\sigma}_{ij}-\sigma_{ij}|\geq nv)\leq C_1\exp(-C_2nv^2)\quad \text{for}|v|\leq\delta, \quad \text{(A10)}$$

where the positive constants C_1, C_2 and δ all depend on $\tau_{\max}$ alone (Bickel & Levina, 2008, Lemma A.3).

The rest of the proof follows exactly the same as that for Lemma 1 of H. Wang (2009). ■

A.3 Proof of theorem 4.1

Our proof follows similar arguments as in the proof of Theorem 1 of H. Wang (2009).

Assume that no relevant predictor has been discovered in the first ℓ iterations, i.e., $S\not\subset S^{(\ell)}$. We evaluate the probability that at least one relevant will be identified in the $(\ell+1)$'s iteration or equivalently its complementary probability that the predictor selected by the $(\ell+1)$'s iteration is still an irrelevant one.

Let $X_{(S)}=(\mathbf{x}_{1(S)},\ldots,\mathbf{x}_{n(S)})$ denote the subset of X corresponding to S. Let $B_{(S)}$ denote the coefficient matrix under the true model.

Denote

$$\begin{aligned}
H_{(S^{(\ell)})}&=X_{(S^{(\ell)})}\{X_{(S^{(\ell)})}^{\top}X_{(S^{(\ell)})}\}^{-1}X_{(S^{(\ell)})}^{\top},\\
\widetilde{H}_k^{(\ell)}&=X_{M_k^{(\ell)}}(X_{M_k^{(\ell)}}^{\top}X_{M_k^{(\ell)}})^{-1}X_{M_k^{(\ell)}}^{\top},\\
\mathbf{x}_k^{(\ell)}&=(I_n-H_{S^{(\ell)}})\mathbf{x}_k,\\
H_k^{(\ell)}&=\mathbf{x}_k^{(\ell)}\mathbf{x}_k^{(\ell)^{\top}}\|\mathbf{x}_k^{(\ell)}\|^{-2}.
\end{aligned}$$

Observe that

$$\begin{aligned}
\Omega(\ell)&=\mathrm{RSS}(S^{(\ell)})-\mathrm{RSS}(S^{(\ell+1)})\\
&=\mathrm{tr}\left\{Y^{\top}(I_n-\widetilde{H}_k^{(\ell)})Y\right\}-\mathrm{tr}\left\{Y^{\top}(I_n-\widetilde{H}_k^{(\ell+1)})Y\right\}\\
&=\mathrm{tr}\left\{Y^{\top}(\widetilde{H}_k^{(\ell+1)}-\widetilde{H}_k^{(\ell)})Y\right\}\\
&=\mathrm{tr}\left\{Y^{\top}(I_n-H_{(S^{(\ell)})})^{\top}H_{a_{\ell+1}}^{(\ell)^{\top}}H_{a_{\ell+1}}^{(\ell)}(I_n-H_{(S^{(\ell)})})Y\right\}.
\end{aligned} \quad \text{(A11)}$$

Assume $a_{\ell+1}\notin S$. We have

$$\begin{aligned}
\Omega(\ell)&\geq\max_{j\in S}\mathrm{tr}\left\{Y^{\top}(I_n-H_{(S^{(\ell)})})^{\top}H_k^{(\ell)^{\top}}H_k^{(\ell)}(I_n-H_{(S^{(\ell)})})Y\right\}\\
&\geq\mathrm{tr}\left\{Y^{\top}(I_n-H_{(S^{(\ell)})})^{\top}H_{\widehat{k}}^{(\ell)^{\top}}H_{\widehat{k}}^{(\ell)}(I_n-H_{(S^{(\ell)})})Y\right\},
\end{aligned} \quad \text{(A12)}$$

where

$$\begin{aligned}
\widehat{k}=\mathrm{argmax}_{k\in S}\mathrm{tr}&\left\{B_{(S)}^{\top}X_{(S)}(I_n-H_{(S^{(\ell)})})^{\top}\right.\\
&\left.H_k^{(\ell)^{\top}}H_k^{(\ell)}(I_n-H_{(S^{(\ell)})})X_{(S)}^{\top}B_{(S)}\right\}.
\end{aligned}$$

Further, observe that the last inequality of (A12) is no less than

$$\begin{aligned}
&\mathrm{tr}\left\{B_{(S)}^{\top}X_{(S)}(I_n-H_{(S^{(\ell)})})^{\top}H_{\widehat{k}}^{(\ell)^{\top}}H_{\widehat{k}}^{(\ell)}(I_n-H_{(S^{(\ell)})})X_{(S)}^{\top}B_{(S)}\right\}\\
&\quad-\mathrm{tr}\left\{\boldsymbol{\varepsilon}^{\top}(I_n-H_{(S^{(\ell)})})^{\top}H_{\widehat{k}}^{(\ell)^{\top}}H_{\widehat{k}}^{(\ell)}(I_n-H_{(S^{(\ell)})})\boldsymbol{\varepsilon}\right\}\\
&\geq\max_{k\in S}\mathrm{tr}\left\{B_{(S)}^{\top}X_{(S)}(I_n-H_{(S^{(\ell)})})^{\top}H_k^{(\ell)^{\top}}\right.\\
&\quad\left.H_k^{(\ell)}(I_n-H_{(S^{(\ell)})})X_{(S)}^{\top}B_{(S)}\right\}\\
&\quad-\max_{k\in S}\mathrm{tr}\left\{\boldsymbol{\varepsilon}^{\top}(I_n-H_{(S^{(\ell)})})^{\top}H_k^{(\ell)^{\top}}H_k^{(\ell)}(I_n-H_{(S^{(\ell)})})\boldsymbol{\varepsilon}\right\},
\end{aligned} \quad \text{(A13)}$$

where $\boldsymbol{\varepsilon} = (\vec{\epsilon}_1, \ldots, \vec{\epsilon}_n)^\top \in \mathbb{R}^{n\times q}$.

In what follows, we study the two terms in (A13) separately.

Step 1: The first term of (A13). Define $Q_{(S^{(k)})} = I_n - H_{(S^{(k)})}$. And denote $\mathbf{x}_k^{(\ell)} = \mathbf{x}_k^\top Q_{(S^{(\ell)})}$. Then, the first term in (A13) can be expressed as

$$\begin{aligned}
&\max_{k\in S} \operatorname{tr}\left\{B_{(S)}^\top X_{(S)} Q_{(S^{(\ell)})}^\top H_k^{(\ell)^\top} H_k^{(\ell)} Q_{(S^{(\ell)})} X_{(S)}^\top B_{(S)}\right\} \\
&\quad = \max_{j\in S} \operatorname{tr}\left\{B_{(S)}^\top X_{(S)} Q_{(S^{(\ell)})}^\top \frac{\mathbf{x}_k^{(\ell)}\mathbf{x}_k^{(\ell)^\top}\mathbf{x}_k^{(\ell)}\mathbf{x}_k^{(\ell)^\top}}{\|\mathbf{x}_k^{(\ell)}\|^4}\right. \\
&\qquad \left. \times\, Q_{(S^{(\ell)})} X_{(S)}^\top B_{(S)}\right\} \\
&\quad = \max_{k\in S} \operatorname{tr}\left\{B_{(S)}^\top X_{(S)} Q_{(S^{(\ell)})}^\top \frac{\mathbf{x}_k^{(\ell)}\mathbf{x}_k^{(\ell)^\top}}{\|\mathbf{x}_k^{(\ell)}\|^2} Q_{(S^{(\ell)})} X_{(S)}^\top B_{(S)}\right\}.
\end{aligned} \tag{A14}$$

Define

$$k^* = \operatorname{argmax}_{k\in S} \operatorname{tr}\left\{(X_{(S)}B_{(S)})^\top Q_{(S^{(\ell)})}^\top \mathbf{x}_k^{(\ell)}\mathbf{x}_k^{(\ell)^\top} Q_{(S^{(\ell)})} X_{(S)} B_{(S)}\right\}.$$

Thus, the RHS of (A14) is no less than

$$\begin{aligned}
&\|\mathbf{x}_{k^*}^{(\ell)}\|^{-2} \operatorname{tr}\left\{B_{(S)}^\top X_{(S)} Q_{(S^{(\ell)})}^\top \mathbf{x}_{k^*}^{(\ell)}\mathbf{x}_{k^*}^{(\ell)^\top} Q_{(S^{(\ell)})} X_{(S)}^\top B_{(S)}\right\} \\
&\quad \geq \min_{j\in S} \|\mathbf{x}_k^{(\ell)}\|^{-2} \\
&\qquad \times \operatorname{tr}\left\{B_{(S)}^\top X_{(S)} Q_{(S^{(\ell)})}^\top \mathbf{x}_{k^*}^{(\ell)}\mathbf{x}_{k^*}^{(\ell)^\top} Q_{(S^{(\ell)})} X_{(S)}^\top B_{(S)}\right\} \\
&\quad = (\max_{k\in S} \|\mathbf{x}_k^{(\ell)}\|^2)^{-1} \\
&\qquad \times \max_{k\in S} \operatorname{tr}\left\{B_{(S)}^\top X_{(S)} Q_{(S^{(\ell)})}^\top \mathbf{x}_k^{(\ell)}\mathbf{x}_k^{(\ell)^\top} Q_{(S^{(\ell)})} X_{(S)}^\top B_{(S)}\right\} \\
&\quad \geq (\max_{k\in S} \|\mathbf{x}_k\|^2)^{-1} \\
&\qquad \times \max_{k\in S} \operatorname{tr}\left\{B_{(S)}^\top X_{(S)} Q_{(S^{(\ell)})}^\top \mathbf{x}_k^{(\ell)}\mathbf{x}_k^{(\ell)^\top} Q_{(S^{(\ell)})} X_{(S)}^\top B_{(S)}\right\},
\end{aligned} \tag{A15}$$

where the last inequality follows after the fact $\|\mathbf{x}_k\| \geq \|\mathbf{x}_k^{(\ell)}\|$.

On the other hand, observe that

$$\begin{aligned}
&\operatorname{tr}\left\{B_{(S)}^\top X_{(S)} Q_{(S^{(\ell)})}^\top Q_{(S^{(\ell)})} X_{(S)}^\top B_{(S)}\right\} \\
&\quad = \operatorname{tr}\left\{B_{(S)}^\top X_{(S)} Q_{(S^{(\ell)})} X_{(S)}^\top B_{(S)}\right\} \\
&\quad = \sum_{j=1}^q \left(\sum_{k\in S} \beta_{kj} \mathbf{x}_k^\top Q_{(S^{(\ell)})} X_{(S)}^\top \vec{\beta}_j\right) \\
&\quad \leq \sum_{j=1}^q \left(\sum_{k\in S} \beta_{kj}^2\right)^{1/2} \left\{\sum_{k\in S} \left(\mathbf{x}_k^\top Q_{(S^{(\ell)})} X_{(S)}^\top \vec{\beta}_j\right)^2\right\}^{1/2} \\
&\quad \leq q \times \max_j \|\vec{\beta}_j\| \operatorname{tr}\left\{B_{(S)}^\top X_{(S)} Q_{(S^{(\ell)})}^\top \mathbf{x}_k^{(\ell)}\mathbf{x}_k^{(\ell)^\top}\right. \\
&\qquad \left. Q_{(S^{(\ell)})} X_{(S)}^\top B_{(S)}\right\}^{1/2} \times \sqrt{p_0}.
\end{aligned} \tag{A16}$$

Applying (A16) to (A15) and (A14), using the fact that $\max_{k\in S} \|\mathbf{x}_k\|^2/n \leq \tau_{\max}$ with probability tending to one, and by Assumptions 3.1, 4.1 and Lemma 1 of H. Wang (2009), we obtain

$$\begin{aligned}
&\max_{k\in S} \operatorname{tr}\left\{B_{(S)}^\top X_{(S)} Q_{(S^{(\ell)})}^\top H_k^{(\ell)^\top} H_k^{(\ell)} Q_{(S^{(\ell)})} X_{(S)}^\top B_{(S)}\right\} \\
&\quad \geq \frac{\left[\operatorname{tr}\left\{B_{(S)}^\top X_{(S)} Q_{(S^{(\ell)})} X_{(S)}^\top B_{(S)}\right\}\right]^2}{q^2 n \tau_{\max} p_0 \times \max_j \|\vec{\beta}_j\|^2}.
\end{aligned} \tag{A17}$$

Define $\zeta_{(S^{(\ell)})} = (X_{(S^{(\ell)})}^\top X_{(S^{(\ell)})})^{-1} X_{(S)}^\top B_{(S)}$. We have

$$\begin{aligned}
&\operatorname{tr}\left\{B_{(S)}^\top X_{(S)} Q_{(S^{(\ell)})} X_{(S)}^\top B_{(S)}\right\} \\
&\quad = \operatorname{tr}\left(B_{(S)}^\top X_{(S)} X_{(S)}^\top B_{(S)} - \zeta_{(S^{(\ell)})}^\top X_{(S^{(\ell)})}^\top X_{(S^{(\ell)})} \zeta_{(S^{(\ell)})}\right).
\end{aligned}$$

Recall the assumption at the beginning of the proof that $S \not\subset S^{(k)}$. Then, by Assumptions 3.1, 3.2 and 4.1, and Lemma 1 of H. Wang (2009), we get

$$\operatorname{tr}\left\{B_{(S)}^\top X_{(S)} Q_{(S^{(\ell)})} X_{(S)}^\top B_{(S)}\right\} \geq q n \tau_{\min} \beta_{\min}^2 \tag{A18}$$

with probability tending to one. Applying (A18) to (A17) and using Assumptions 3.2, 4.1 and 4.2, we obtain

$$\begin{aligned}
&\max_{k\in S} \operatorname{tr}\left\{B_{(S)}^\top X_{(S)} Q_{(S^{(\ell)})}^\top H_k^{(\ell)^\top} H_k^{(\ell)} Q_{(S^{(\ell)})} X_{(S)}^\top B_{(S)}\right\} \\
&\quad \geq n \tau_{\max}^{-1} p_0^{-1} (\max_j \|\vec{\beta}_j\|)^{-2} \tau_{\min}^2 \beta_{\min}^4 \\
&\quad \geq \tau_{\max}^{-1} \nu^{-1} C_B^{-2} \tau_{\min}^2 \nu_B^4 n^{1-\xi_0-4\xi_{\min}}.
\end{aligned} \tag{A19}$$

Step 2: The second term of (A13). Recall

$$\mathbf{x}_k^{(\ell)} = \mathbf{x}_k - H_{(S^{(\ell)})}\mathbf{x}_k = \mathbf{x}_k - X_{(S^{(\ell)})}\theta_{k(S^{(\ell)})}$$

where

$$\theta_{k(S^{(\ell)})} = \left(X_{(S^{(\ell)})}^\top X_{(S^{(\ell)})}\right)^{-1} X_{(S^{(\ell)})}^\top \mathbf{x}_k.$$

Then, by Assumptions 3.1 and 3.2 and Lemma 1 of H. Wang (2009), we have $\|\mathbf{x}_k^{(\ell)}\|^2 \geq n\tau_{\min}$.

Moreover, since $\mathbf{x}_k^{(\ell)} = (I_n - H_{(S^{(\ell)})})\mathbf{x}_k$, we have

$$\begin{aligned}
&\operatorname{tr}\left\{\vec{\varepsilon}^\top (I_n - H_{(S^{(\ell)})})^\top H_k^{(\ell)^\top} H_k^{(\ell)} (I_n - H_{(S^{(\ell)})})\vec{\varepsilon}\right\} \\
&\quad = \|\mathbf{x}_k^{(\ell)}\|^{-2} \operatorname{tr}\left(\vec{\varepsilon}^\top \mathbf{x}_k \mathbf{x}_k^\top \vec{\varepsilon} - \vec{\varepsilon}^\top H_{(S^{(\ell)})}^\top \mathbf{x}_k \mathbf{x}_k^\top H_{(S^{(\ell)})} \vec{\varepsilon}\right) \\
&\quad \leq \tau_{\min}^{-1} n^{-1} \operatorname{tr}\left(\vec{\varepsilon}^\top \mathbf{x}_k \mathbf{x}_k^\top \vec{\varepsilon} - \vec{\varepsilon}^\top H_{(S^{(\ell)})}^\top \mathbf{x}_k \mathbf{x}_k^\top H_{(S^{(\ell)})} \vec{\varepsilon}\right) \\
&\quad = \tau_{\min}^{-1} n^{-1} \operatorname{tr}\left(\varepsilon^\top Q_{(S^{(\ell)})} \mathbf{x}_k \mathbf{x}_k^\top Q_{(S^{(\ell)})} \vec{\varepsilon}\right) \\
&\quad \leq \tau_{\min}^{-1} n^{-1} \max_{k\in S} \max_{|M|\leq m^*} \operatorname{tr}\left(\vec{\varepsilon}^\top Q_{(M)} \mathbf{x}_k \mathbf{x}_k^\top Q_{(M)} \vec{\varepsilon}\right) \\
&\quad = \tau_{\min}^{-1} n^{-1} \max_{k\in S} \max_{|M|\leq m^*} \sum_{j=1}^q \left(\vec{\varepsilon}_j^\top Q_{(M)} \mathbf{x}_k \mathbf{x}_k^\top Q_{(M)} \vec{\varepsilon}_j\right),
\end{aligned} \tag{A20}$$

where $m^* = K\nu n^{2\xi_0+4\xi_{\min}}$.

Notice that $\mathbf{x}_k^\top Q_{(M)} \vec{\varepsilon}_j$ is a normal random variable with mean zero and variance given by $\|Q_{(M)}\mathbf{x}_k\|^2 \leq \|\mathbf{x}_k\|^2$. Thus, the RHS of (A20) is no greater than

$$q \tau_{\min}^{-1} n^{-1} \max_{k\in S} \|\mathbf{x}_k\|^2 \max_{k\in S} \max_{|M|\leq m^*} \chi_1^2, \tag{A21}$$

where χ_1^2 stands for a chi-square random variable with one degree of freedom. By Assumptions 3.1 and 4.1, and Lemma 1 of H. Wang (2009), we get that $n^{-1} \max_{k\in S} \|\mathbf{x}_k\|^2 \leq \tau_{\max}$ with probability tending to one.

On the other hand, the total number of combinations for $k \in S$ and $|M| \le m^*$ is no more than d^{m^*+2}. Then, by Assumption 4.2, we get

$$\begin{aligned}&\max_{k\in S}\max_{|M|=\ell}\chi_1^2\\&\quad\le 2(m^*+2)\log(p)\\&\quad\le 3K\nu n^{2\xi_0+4\xi_{\min}}\times \nu n^{\xi}=3K\nu^2 n^{\xi+2\xi_0+4\xi_{\min}}\end{aligned}$$

with probability tending to one. Therefore, (A21) is bounded by

$$q\tau_{\min}^{-1}\tau_{\max}3K\nu^2 n^{\xi+2\xi_0+4\xi_{\min}-1}. \tag{A22}$$

Combining (A13), (A19) and (A22), we have

$$\begin{aligned}n^{-1}\Omega(\ell) &\ge \tau_{\max}^{-1}\nu^{-1}C_B^{-2}\tau_{\min}^2\nu_B^4 n^{1-\xi_0-4\xi_{\min}}\\&\quad - q\tau_{\min}^{-1}\tau_{\max}3K\nu^2 n^{\xi+2\xi_0+4\xi_{\min}-1}\\&= \tau_{\max}^{-1}\nu^{-1}C_B^{-2}\tau_{\min}^2\nu_B^4 n^{1-\xi_0-4\xi_{\min}}\\&\quad\times\left(1-q\tau_{\max}^2\nu^3C_B^2\tau_{\max}^{-3}\nu_B^{-4}3Kn^{\xi+3\xi_0+8\xi_{\min}-1}\right)\end{aligned} \tag{A23}$$

uniformly for every $\ell \le Kn^{\xi_0+4\xi_{\min}}$. Recall $K = 2\tau_{\max}\nu C_B^2\tau_{\min}^{-2}\nu_B^{-4}$ defined in Section 4. Then, by Assumption 4.2 and 4.3, we have

$$\begin{aligned}n^{-1}\|Y\|_F^2 &\ge n^{-1}\sum_{\ell=1}^{Kn^{\xi_0+4\xi_{\min}}}\Omega(\ell)\\&\ge 2\left(1-q\tau_{\max}^2\nu^3C_B^2\tau_{\max}^{-3}\nu_B^{-4}3Kn^{\xi+3\xi_0+8\xi_{\min}-1}\right)\\&\xrightarrow{p} 2,\end{aligned} \tag{A24}$$

where $\|\cdot\|_F$ is the Frobenius norm.

Without loss of generality, we can assume $\mathrm{var}(y_{i1})+\cdots+\mathrm{var}(y_{iq})=1$, and we have $n^{-1}\|Y\|_F^2 \xrightarrow{p} 1$. (Otherwise, we can standardize it by letting $y_{ik}^* = y_{ik}/\sqrt{\mathrm{var}(y_{i1})+\cdots+\mathrm{var}(y_{iq})}$ for $i=1,\ldots,n$ and $k=1,\ldots,q$.) Thus, it is impossible to have $S^{(k)}\bigcup M_t=\emptyset$ for every $1\le k\le Kn^{\xi_0+4\xi_{\min}}$, which implies that at least one relevant variable will be discovered within $Kn^{\xi_0+4\xi_{\min}}$ steps. This completes the proof.