Journal of Econometrics 195 (2016) 154–168

Contents lists available at ScienceDirect

Journal of Econometrics

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

Testing a single regression coefficient in high dimensional linear models

Wei Lan [a,*], Ping-Shou Zhong [b], Runze Li [c], Hansheng Wang [d], Chih-Ling Tsai [e]

[a] Statistics School and Center of Statistical Research, Southwestern University of Finance and Economics, Chengdu, PR China
[b] Department of Statistics and Probability, Michigan State University, East Lansing, MI 48823, United States
[c] Department of Statistics and the Methodology Center, The Pennsylvania State University, University Park, PA 16802-2111, United States
[d] Department of Business Statistics and Econometrics, Guanghua School of Management, Peking University, Beijing, 100871, PR China
[e] Graduate School of Management, University of California, Davis, CA 95616-8609, United States

ARTICLE INFO

Article history:
Received 15 February 2016
Received in revised form
15 February 2016
Accepted 21 May 2016
Available online 15 June 2016

Keywords:
Correlated Predictors Screening
False discovery rate
High dimensional data
Single coefficient test

ABSTRACT

In linear regression models with high dimensional data, the classical z-test (or t-test) for testing the significance of each single regression coefficient is no longer applicable. This is mainly because the number of covariates exceeds the sample size. In this paper, we propose a simple and novel alternative by introducing the Correlated Predictors Screening (CPS) method to control for predictors that are highly correlated with the target covariate. Accordingly, the classical ordinary least squares approach can be employed to estimate the regression coefficient associated with the target covariate. In addition, we demonstrate that the resulting estimator is consistent and asymptotically normal even if the random errors are heteroscedastic. This enables us to apply the z-test to assess the significance of each covariate. Based on the p-value obtained from testing the significance of each covariate, we further conduct multiple hypothesis testing by controlling the false discovery rate at the nominal level. Then, we show that the multiple hypothesis testing achieves consistent model selection. Simulation studies and empirical examples are presented to illustrate the finite sample performance and the usefulness of the proposed method, respectively.

1. Introduction

In linear regression models, it is a common practice to employ the z-test (or t-test) to assess whether an individual predictor (or covariate) is significant when the number of covariates (p) is smaller than the sample size (n). This test has been widely applied across various fields (e.g., economics, finance and marketing) and is available in most statistical software. One usually applies the ordinary least squares (OLS) approach to estimate regression coefficients and standard errors for constructing a z-test (or t-test); see, for example, Draper and Smith (1998) and Wooldridge (2002). However, in a high dimensional linear model with p exceeding n, the classical z-test (or t-test) is not applicable because it is infeasible to compute the OLS estimators of p regression coefficients. This motivates us to modify the classical z-test (or t-test) to accommodate high dimensional data.

In high dimensional regression analysis, hypothesis testing has attracted considerable attention (Goeman et al., 2006, 2011; Zhong and Chen, 2011). Since these papers mainly focus on testing a large set of coefficients against a high dimensional alternative, their approaches are not applicable for testing the significance of a single coefficient. Hence, Bühlmann (2013) recently applied the ridge estimation approach and obtained a test statistic to examine the significance of an individual coefficient. His proposed test involves a bias correction, which is different from the classical z-test (or t-test) via the OLS approach. In the meantime, Zhang and Zhang (2014) proposed a low dimensional projection procedure to construct the confidence intervals for a linear combination of a small subset of regression coefficients. The key assumption behind their procedure is the existence of good initial estimators for the unknown regression coefficients and the unknown standard deviation of random errors. To this end, the penalty function with a tuning parameter is required to implement Zhang and Zhang's (2014) procedure. Later, van de Geer et al. (2014) extended the

* Corresponding author.
E-mail addresses: lanwei@swufe.edu.cn (W. Lan), pszhong@stt.msu.edu (P.-S. Zhong), rzli@psu.edu (R. Li), hansheng@gsm.pku.edu.cn (H. Wang), cltsai@ucdavis.edu (C.-L. Tsai).

http://dx.doi.org/10.1016/j.jeconom.2016.05.016
0304-4076/

results of Zhang and Zhang's (2014) to broad models and general loss functions.

Instead of the ridge estimation and low dimensional projection, Fan and Lv (2008) and Fan et al. (2011) used the correlation approach to screen out those covariates that have weak correlations with the response variable. As a result, the total number of predictors that are highly correlated with the response variable is smaller than the sample size. However, Cho and Fryzlewicz (2012) found out that such a screening process via the marginal correlation procedure may not be reliable when the predictors are highly correlated. To this end, they proposed a tilting correlation screening (TCS) procedure to measure the contribution of the target variable to the response. Motivated by the TCS idea of Cho and Fryzlewicz (2012), we develop a new testing procedure that can lead to accurate inferences. Specifically, we adopt the TCS idea and introduce the Correlated Predictors Screening (CPS) method to control for predictors that are highly correlated with the target covariate before a hypothesis test is conducted. It is worth noting that Cho and Fryzlewicz (2012) mainly focus on variable selection, while we aim at hypothesis testing.

If the total number of highly correlated predictors resulting from the CPS procedure is smaller than the sample size, their effects can be profiled out from both the response and the target predictor via projections. Based on the profiled response and the profiled predictor, we are able to employ a classical simple regression model to obtain the OLS estimate of the target regression coefficient. We then demonstrate that the resulting estimator is $\sqrt{n}$-consistent and asymptotically normal, even if the random errors are heteroskedastic as considered by Belloni et al. (2012, 2014). Accordingly, a z-test statistic can be constructed for testing the target coefficient. Under some mild conditions, we show that the p-values obtained by the asymptotic normal distribution satisfy the weak dependence assumption of Storey et al. (2004). As a result, the multiple hypothesis testing procedure of Storey et al. (2004) can be directly applied to control the false discovery rate (FDR). Finally, we demonstrate that the proposed multiple testing procedure achieves model selection consistency.

The rest of the article is organized as follows. Section 2 introduces model notation and proposes the CPS method. The theoretical properties of hypothesis tests via the CPS as well as the FDR procedures are obtained. Section 3 presents simulation studies, while Section 4 provides real data analyses. Some concluding remarks are given in Section 5. All technical details are relegated to Appendix.

2. The methodology

2.1. The CPS method

Let (Y_i, X_i) be a random vector collected from the ith subject $(1 \le i \le n)$, where $Y_i \in \mathbb{R}^1$ is the response variable and $X_i = (X_{i1}, \ldots, X_{ip})^\top \in \mathbb{R}^p$ is the associated p-dimensional predictor vector with $E(X_i) = 0$ and $\mathrm{cov}(X_i) = \Sigma = (\sigma_{j_1 j_2}) \in \mathbb{R}^{p\times p}$. In addition, the response variable has been centralized such that $E(Y_i) = 0$. Unless explicitly stated otherwise, we hereafter assume that $p \gg n$ and n tends to infinity for asymptotic behavior. Then, consider the linear regression model,

$$Y_i = X_i^\top \beta + \varepsilon_i, \tag{2.1}$$

where $\beta = (\beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^p$ is an unknown regression coefficient vector. Motivated by Belloni et al. (2012, 2014), we assume that the error terms ε_i are independently distributed with $E(\varepsilon_i|X_i) = 0$ and finite variance $\mathrm{var}(\varepsilon_i) = \sigma_i^2$ for $i = 1, \ldots, n$. In addition, define the average of error variances as $\bar\sigma_n^2 = n^{-1}\sum_i \sigma_i^2$, and assume that $\bar\sigma_n^2 \to \bar\sigma^2$ as $n \to \infty$ for some finite positive constant $\bar\sigma^2$. To assess the significance of a single coefficient, we test the null hypothesis $H_0 : \beta_j = 0$ for any given j. Without loss of generality, we focus on testing the first regression coefficient. That is,

$$H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_1 : \beta_1 \neq 0, \tag{2.2}$$

and the same testing procedure is applicable to the rest of the individual regression coefficients.

For the sake of convenience, let $\mathbb{Y} = (Y_1, \ldots, Y_n)^\top \in \mathbb{R}^n$ be the vector of responses, $\mathbb{X} = (X_1, \ldots, X_n)^\top \in \mathbb{R}^{n\times p}$ be the design matrix with the jth column $\mathbb{X}_j \in \mathbb{R}^n$, and $\mathcal{E} = (\varepsilon_1, \ldots, \varepsilon_n)^\top \in \mathbb{R}^n$. In addition, let $\mathcal{A}$ be an arbitrary index set with cardinality $|\mathcal{A}|$. Then, define $X_{i\mathcal{A}} = (X_{ij} : j \in \mathcal{A})^\top \in \mathbb{R}^{|\mathcal{A}|}$, $\mathbb{X}_{\mathcal{A}} = (X_{1\mathcal{A}}, \ldots, X_{n\mathcal{A}})^\top = (\mathbb{X}_j : j \in \mathcal{A}) \in \mathbb{R}^{n\times|\mathcal{A}|}$, $\Sigma_{\mathcal{A}} = (\sigma_{j_1 j_2} : j_1 \in \mathcal{A}, j_2 \in \mathcal{A}) \in \mathbb{R}^{|\mathcal{A}|\times|\mathcal{A}|}$, and $\Sigma_{\mathcal{A}j} = \Sigma_{j\mathcal{A}}^\top = (\sigma_{j_1 j_2} : j_1 \in \mathcal{A}, j_2 = j) \in \mathbb{R}^{|\mathcal{A}|}$. Moreover, define $\Sigma_{\mathcal{A}_a\mathcal{A}_b} = (\sigma_{j_1 j_2} : j_1 \in \mathcal{A}_a, j_2 \in \mathcal{A}_b) \in \mathbb{R}^{|\mathcal{A}_a|\times|\mathcal{A}_b|}$ for any two arbitrary index sets $\mathcal{A}_a$ and $\mathcal{A}_b$, which implies $\Sigma_{\mathcal{A}\mathcal{A}} = \Sigma_{\mathcal{A}}$.

Before constructing the test statistic, we first control those predictors that are highly correlated with X_{i1}. Otherwise, they can generate a confounding effect, due to multicollinearity and yield an incorrect estimator of β_1. Specifically, the marginal regression coefficient $(\mathbb{X}_1^\top\mathbb{X}_1^\top)^{-1}\mathbb{X}_1^\top\mathbb{Y} = \beta_1 + (\mathbb{X}_1^\top\mathbb{X}_1^\top)^{-1}\mathbb{X}_1^\top(\mathbb{Y} - \mathbb{X}_1\beta_1)$ is not a consistent estimator of β_1 when $\mathbb{Y} - \mathbb{X}_1\beta_1$ and $\mathbb{X}_1$ have a strong linear relationship. To remove the confounding effect, define $\rho_{1j} = \mathrm{corr}(X_{i1}, X_{ij})$ as the correlation coefficient of X_{i1} and X_{ij} for $j = 2, \ldots, p$, and $\rho_1^* = (|\rho_{12}|, \ldots, |\rho_{1p}|)^\top \in \mathbb{R}^{p-1}$. We also assume that $|\rho_{1j}|$ are distinct. Then, let $\mathcal{S}_k$ be the set of k indices whose associated predictors have the largest absolute correlations with X_{i1}:

$$\mathcal{S}_k = \big\{2 \le j \le p : \ |\rho_{1j}| \text{ is among the first } k \text{ largest} \\ \text{absolute correlations in } \rho_1^*\big\}. \tag{2.3}$$

The choice of k (i.e., $\mathcal{S}_k$) will be discussed in Remark 2. With a slight abuse of notation, we sometimes denote $\mathcal{S}_k$ by $\mathcal{S}$ in the rest of the paper for the sake of convenience. To remove the confounding effect due to $X_{i\mathcal{S}}$, we construct the profiled response and predictor as $\widetilde{\mathbb{Y}} = \mathcal{Q}_{\mathcal{S}}\mathbb{Y}$ and $\widetilde{\mathbb{X}}_1 = \mathcal{Q}_{\mathcal{S}}\mathbb{X}_1$, respectively, where $\mathcal{Q}_{\mathcal{S}} = I_n - \mathbb{X}_{\mathcal{S}}(\mathbb{X}_{\mathcal{S}}^\top\mathbb{X}_{\mathcal{S}})^{-1}\mathbb{X}_{\mathcal{S}}^\top \in \mathbb{R}^{n\times n}$ and $I_n \in \mathbb{R}^{n\times n}$ is the $n \times n$ identity matrix. We next follow the OLS approach and obtain the estimate of the target coefficient β_1,

$$\hat\beta_1 = (\widetilde{\mathbb{X}}_1^\top\widetilde{\mathbb{X}}_1)^{-1}(\widetilde{\mathbb{X}}_1^\top\widetilde{\mathbb{Y}}) = (\mathbb{X}_1^\top\mathcal{Q}_{\mathcal{S}}\mathbb{X}_1)^{-1}(\mathbb{X}_1^\top\mathcal{Q}_{\mathcal{S}}\mathbb{Y}).$$

We refer to the above procedure as the Correlated Predictors Screening (CPS) method, $\hat\beta_1$ as the CPS estimator of β_1, and $\mathcal{S}$ as the CPS set of X_{i1}.

It is of interest to note that the proposed CPS estimator $\hat\beta_1$ is closely related to the estimator obtained via the "added-variable plot" approach (e.g., see Cook and Weisberg, 1998). To illustrate their relationship, let $\mathbb{X}_{-1}$ be the collection of all covariates in $\mathbb{X}$ except for $\mathbb{X}_1$. Then the method of "added-variable plot" essentially takes the residuals from regressing $\mathbb{Y}$ against $\mathbb{X}_{-1}$ as the response and the residuals from regressing $\mathbb{X}_1$ against $\mathbb{X}_{-1}$ as covariates. Although both approaches can be used to assess the effect of $\mathbb{X}_{-1}$ on the estimation of β_1, they are different. Specifically, the "added-variable plot" approach requires regressing $\mathbb{X}_1$ on all remaining covariates, which is not computable when the dimension p is larger than n. By contrast, CPS only considers those predictors in $\mathcal{S}$ that are highly correlated with $\mathbb{X}_1$, which is applicable in high dimensional settings.

Making inferences about β_1 in high dimensional models is challenging because these inferences can depend on the accuracy of estimating the whole vector β; see Belloni et al. (2014), van de Geer et al. (2014) and Zhang and Zhang (2014). The main contribution of our proposed CPS method is employing a simple

marginal regression approach to estimate β_1 after controlling for the predictors that are highly correlated with $\mathbb{X}_1$. As a result, the profiled predictor, $\widetilde{\mathbb{X}}_1$, is approximately independent of the remaining covariates. This allows us to not only directly estimate β_1, but also make inferences about β_1. The theoretical properties of the CPS estimator and associated test statistic are presented below.

2.2. Asymptotic normality of the CPS estimator and test statistic

To make inferences, we study the asymptotic properties of the CPS estimator $\hat{\beta}_1$. Define $\varrho_{j_1j_2}(\mathcal{S}) = \sigma_{j_1j_2} - \Sigma_{\mathcal{S}j_1}^\top \Sigma_{\mathcal{S}}^{-1}\Sigma_{\mathcal{S}j_2} \in \mathbb{R}^1$, which measures the partial covariance of X_{ij_1} and X_{ij_2}, after controlling for the effect of $X_{i\mathcal{S}} = (X_{ij} : j \in \mathcal{S})^\top \in \mathbb{R}^{|\mathcal{S}|}$. Then, we make the following assumptions to facilitate the technical proofs, while admittedly not the weakest possible assumptions.

(C1) *Gaussian condition*. Assume that the X_is are independent and normally distributed with mean 0 and covariance matrix Σ.

(C2) *Bounded diagonal elements*. There exist two finite constants $c_{\max}$ and $c_{\max}^*$ such that the diagonal components of Σ and Σ^{-1} are bounded above by $c_{\max}$ and $c_{\max}^*$, respectively.

(C3) *Predictor dimension*. There exist two positive constants $\hbar < 1$ and $\nu > 0$ such that $\log p \le \nu n^{\hbar}$ for every $n > 0$.

(C4) *Partial covariance*. There exists a constant $\xi > 3/2$ such that $\max_{j\notin\mathcal{S}} |\varrho_{1j}(\mathcal{S})| = O(|\mathcal{S}|^{-\xi})$ as $|\mathcal{S}| \to \infty$.

(C5) *Dimension of the CPS set*. There exist a CPS set $|\mathcal{S}|$ and two positive constants C_a and C_b such that $C_a n^{\nu_1} \le |\mathcal{S}| \le C_b n^{\nu_2}$, where ν_1 and ν_2 are two positive constants with $1/(2\xi) < \nu_1 \le \nu_2 < 1/3$ and $\hbar + 3\nu_2 < 1$, where $\hbar$ is defined in Condition (C3).

(C6) *Regression coefficients*. Assume that $|\beta| = \sum_{j=1}^p |\beta_j| < C_{\max} n^{\varpi}$ for some constant $C_{\max} > 0$ and $\varpi < \min(1/4, \xi\nu_1 - 1/2)$, where ξ and ν_1 are defined in (C4) and (C5), respectively.

Condition (C1) is a common condition used for high dimensional data to simplify theoretical proofs; see for example, Wang (2009) and Zhang and Zhang (2014). This condition can be relaxed to the sub-Gaussian random variables (Wang, 2012; Li et al., 2012) and our theoretical results still hold. Condition (C2) is a mild condition that has been well discussed in Liu (2013). Condition (C3) allows the dimension of predictors p to diverge exponentially with the sample size n, so that p can be much larger than n. Condition (C4) is a technical condition for simplifying the proofs of our theory, and it requires that the partial covariance between the target covariate and any other predictor that does not belong to $\mathcal{S}$, after controlling for the effect of $X_{i\mathcal{S}}$ (i.e., key confounders), and that it converges towards 0 at a fast speed as $|\mathcal{S}| \to \infty$. This condition is satisfied for many typical covariance structures, e.g., diagonal and autoregressive structures. It is worth noting that, according to (C1), the conditional distribution of X_{i1} given $X_{i,-1} = (X_{i2}, \ldots, X_{ip})^\top \in \mathbb{R}^{p-1}$ remains normal. As a result, there exists a coefficient vector $\theta_{(1)} = (\theta_{(1),1}, \ldots, \theta_{(1),p-1})^\top \in \mathbb{R}^{p-1}$ such that $X_{i1} = X_{i,-1}^\top\theta_{(1)} + e_{i1}$, where e_{i1} is a random error that is independent of $X_{i,-1}$. Furthermore, if $\mathcal{S} \supset \mathcal{S}_\theta = \{j : \theta_{(1),j} \neq 0\}$, then it implies that $\max_{j\notin\mathcal{S}} |\varrho_{1j}(\mathcal{S})| = 0$. Hence, Condition (C4) is satisfied. Furthermore, this condition is closely related to the assumption given in Theorem 5 of Zhang and Zhang (2014). Moreover, Condition (C5), together with Condition (C4), ensures that the size of the CPS set is much smaller than the sample size, but it does not imply that the number of regressors highly correlated with $\mathbb{X}_1$ is bounded. Condition (C5) is used to guarantee that $\max_{j\notin\mathcal{S}} |\varrho_{1j}(\mathcal{S})|$ is of order $o(n^{-1/2})$, so that the bias of $\hat{\beta}_1$ vanishes. Note that the size of the CPS set in Condition (C5) depends on the rate of $\max_{j\notin\mathcal{S}} |\varrho_{1j}(\mathcal{S})| \to 0$. Thus, Condition (C5) can be dropped if $\theta_{(1)}$ has finite non-zero elements or Σ follows an autoregressive structure so that $\max_{j\notin\mathcal{S}} |\varrho_{1j}(\mathcal{S})| = O\{\exp(-\bar{\zeta}|\mathcal{S}|^{\bar{\eta}})\}$ for some positive constants $\bar{\zeta}$ and $\bar{\eta}$. Lastly, Condition (C6) is satisfied when β is sparse with only a finite number of nonzero coefficients. Under the above conditions, we obtain the following result.

Theorem 1. *Assume that Conditions* (C1)–(C6) *hold. We then have* $n^{1/2}(\hat{\beta}_1 - \beta_1) \xrightarrow{d} N(0, \sigma_{\beta_1}^2)$, *where* $\sigma_{\beta_1}^2 = \tau_{\beta_1}^2(\sigma_{11} - \Sigma_{1\mathcal{S}}\Sigma_{\mathcal{S}}^{-1}\Sigma_{\mathcal{S}1})^{-1}$, $\tau_{\beta_1}^2 = \beta_{\mathcal{S}^*}^\top(\Sigma_{\mathcal{S}^*} - \Sigma_{\mathcal{S}^*\mathcal{S}^+}\Sigma_{\mathcal{S}^+}^{-1}\Sigma_{\mathcal{S}^+\mathcal{S}^*})\beta_{\mathcal{S}^*} + \bar{\sigma}^2$, $\mathcal{S}^+ = \{1\}\bigcup\mathcal{S}$, *and* $\mathcal{S}^* = \{j : j \notin \mathcal{S}^+\}$.

Using the results of two lemmas in Appendix A, we are able to prove the above theorem; see the detailed proofs in Appendix B. By Theorem 1, we construct the test statistic,

$$Z_1 = n^{1/2}\hat{\beta}_1/\hat{\sigma}_{\beta_1}, \tag{2.4}$$

where $\hat{\sigma}_{\hat{\beta}_1}^2 = \hat{\tau}_{\hat{\beta}_1}^2\big(n^{-1}\mathbb{X}_1^\top\mathcal{Q}_{\mathcal{S}}\mathbb{X}_1\big)^{-1}$, $\hat{\tau}_{\hat{\beta}_1}^2 = (n - |\mathcal{S}^+|)^{-1}\hat{\mathcal{E}}_{\mathcal{S}^+}^\top\hat{\mathcal{E}}_{\mathcal{S}^+}$, $\hat{\mathcal{E}}_{\mathcal{S}^+}$ is the residual vector obtained by regressing Y_i on $X_{i\mathcal{S}^+}$, and $X_{i\mathcal{S}^+} = (X_{ij} : j \in \mathcal{S}^+)^\top$. Applying similar techniques to those used in the proof of Theorem 1 under Conditions (C1)–(C6), together with Slutsky's theorem and the result that $c_{\max}^{*-1} \le n^{-1}\mathbb{X}_1^\top\mathcal{Q}_{\mathcal{S}}\mathbb{X}_1 \le c_{\max}$ obtained from Lemma 3 and Condition (C2), we can verify that $\hat{\sigma}_{\hat{\beta}_1}^2$ is the consistent estimator of $\sigma_{\hat{\beta}_1}^2$. As a result, Z_1 is asymptotically standard normal under H_0, and one can reject the null hypothesis if $|Z_1| > z_{1-\alpha/2}$, where z_α stands for the αth quantile of the standard normal distribution. Note that if $p < n$ and $\mathcal{S} = \{j : j \neq 1\}$, the test statistic Z_1 is the same as the classical z-test statistic.

To make the testing procedure practically useful, one needs to select the CPS set $\mathcal{S}$ among the sets $\mathcal{S}_k$ for $k \ge 1$. Since $\mathcal{S}_k$ in (2.3) is unknown in practice, we consider its estimator

$$\hat{\mathcal{S}}_k = \Big\{2 \le j \le p : \ |\hat{\rho}_{1j}| \text{ is among the } k \text{ largest elements of } \hat{\rho}_1^*\Big\}, \tag{2.5}$$

where $\hat{\rho}_{1j}$ is the sample correlation coefficient of X_{i1} and X_{ij} and $\hat{\rho}_1^* = (|\hat{\rho}_{12}|, \ldots, |\hat{\rho}_{1p}|)^\top \in \mathbb{R}^{p-1}$. The connection between $\mathcal{S}_k$ and its sample counterpart $\hat{\mathcal{S}}_k$ is established in the following proposition.

Proposition 1. *Let* $|\rho_{1j_i}|$ *be the* i*th largest absolute value of* $\{\rho_{1j} : 2 \le j \le p\}$. *For any* $1 \le k \le C_b n^{\nu_2}$ *with* C_b *and* ν_2 *being defined in Condition* (C5), $\min_{i\ge d_{\max}n^{\nu_2/2}}\big(|\rho_{1j_k}| - |\rho_{1j_{k+i}}|\big) > d_{\min}/n^{\nu_2}$ *for some positive constants* $d_{\min}$ *and* $d_{\max}$. *Then, under Conditions* (C1) *and* (C3), *for any CPS set* $\mathcal{S}_{k_0}$ *satisfying* $k_0 \le C_b n^{\nu_2}$, *there exists* $k^* \le k_0 + d_{\max}n^{\nu_2}$ *such that* $P(\hat{\mathcal{S}}_{k^*} \supset \mathcal{S}_{k_0}) \to 1$.

The proof is given in Appendix C. The condition $\min_{i\ge d_{\max}n^{\nu_2/2}}(|\rho_{1j_k}| - |\rho_{1j_{k+i}}|) > d_{\min}/n^{\nu_2}$ for some finite positive constants $d_{\min}$ and $d_{\max}$ is quite mild, and it ensures that the difference between $|\rho_{1j_l}|$ and $|\rho_{1j_m}|$ cannot be too small when $|j_l - j_m|$ is large enough. By Condition (C5), there exists a positive integer $k_0 \in (C_a n^{\nu_1}, C_b n^{\nu_2})$ such that $\mathcal{S}_{k_0}$ is the CPS set. According to Proposition 1, we can then find $k^* \le k_0 + d_{\max}n^{\nu_2} = O(n^{\nu_2})$ satisfying $P(\hat{\mathcal{S}}_{k^*} \supset \mathcal{S}_{k_0}) \to 1$. This indicates that there exists a set among the paths $\hat{\mathcal{S}}_k$ that contains the CPS set. By Condition (C4), we further have that $\max_{j\notin\hat{\mathcal{S}}_{k^*}^+} |\varrho_{1j}(\hat{\mathcal{S}}_{k^*})| = O(|\mathcal{S}_{k_0}|^{-\xi}) = o(n^{-1/2})$. Using this result, one can verify that Theorem 1 holds by replacing $\mathcal{S}_{k_0}$ with $\hat{\mathcal{S}}_{k^*}$.

Proposition 1 indicates that the sequential selection of the CPS set along the paths $\hat{\mathcal{S}}_k$ ($k = 1, \ldots, p-1$) is attainable. In practice, however, k is unknown and needs to be selected effectively. By the results of Corollary 1 of Kalisch and Bühlmann (2007), we have that $|\hat{\varrho}_{1j}(\bar{\mathcal{S}}) - \varrho_{1j}(\bar{\mathcal{S}})| = O_p(n^{-1/2})$ uniformly for any conditional set with size $|\bar{\mathcal{S}}| = O(n^{\nu_2})$, which leads to

$\max_{j\notin\hat{\mathcal{S}}_k^+}|\hat{\varrho}_{1j}(\hat{\mathcal{S}}_k)-\varrho_{1j}(\hat{\mathcal{S}}_k)| = O_p(n^{-1/2})$ for any $k = O(n^{\nu_2})$. This indicates that the sample partial correlation is close to its true partial correlation as the sample size gets large. Motivated by this finding, we propose choosing the CPS set among the paths $\hat{\mathcal{S}}_k$ by sequentially testing the partial correlations. Specifically, for any $k \geq 1$, let $\hat{\varrho}_{1j}(\hat{\mathcal{S}}_k) = \mathbb{X}_1^\top \mathcal{Q}_{\hat{\mathcal{S}}_k}\mathbb{X}_j/n$ be the sample counterpart of $\varrho_{1j}(\mathcal{S}_k)$ and define $\hat{F}_{1j}(\hat{\mathcal{S}}_k) = 2^{-1}\log[\{1+\hat{\varrho}_{1j}(\hat{\mathcal{S}}_k)\}/\{1-\hat{\varrho}_{1j}(\hat{\mathcal{S}}_k)\}]$, which is in the spirit of Fisher's Z-transformation for the purpose of identifying nodes (variables) that have edges connected to the variable X_{i1} in a Gaussian graph (see Kalisch and Bühlmann, 2007). Then, we select the smallest size of k sequentially such that $(n-|\hat{\mathcal{S}}_k|-3)^{1/2}\max_{j\notin\hat{\mathcal{S}}_k^+}|\hat{F}_{1j}(\hat{\mathcal{S}}_k)| < z_{1-\gamma/2}$ (denoted as $\hat{k}$), for every $j\notin\hat{\mathcal{S}}_k^+$, where γ is a pre-specified significance level and $\hat{\mathcal{S}}_k^+ = \{1\}\cup\hat{\mathcal{S}}_k$. Employing Lemma 3 in Kalisch and Bühlmann (2007) that $|\hat{F}_{1j}(\bar{\mathcal{S}})-F_{1j}(\bar{\mathcal{S}})| = O_p(n^{-1/2})$ uniformly for any conditional set with size $|\bar{\mathcal{S}}| = O(n^{\nu_2})$, we then have $\max_{j\notin\hat{\mathcal{S}}_{\hat{k}}^+}|F_{1j}(\hat{\mathcal{S}}_{\hat{k}})| = O(n^{-1/2})$, which immediately leads to $\max_{j\notin\hat{\mathcal{S}}_{\hat{k}}^+}|\varrho_{1j}(\hat{\mathcal{S}}_{\hat{k}})| = O(n^{-1/2})$. Using the result of Proposition 1 and Condition (C4), we further obtain $|\hat{\mathcal{S}}_{\hat{k}}| \leq (C_b+d_{\max})n^{\nu_2}$. Hence, the $\hat{k}$ selected via the sequential testing procedure is of order $O(n^{\nu_2})$, which is directly related to the assumption imposed on $\mathcal{S}_k$ in Condition (C5).

Remark 1. It is worth noting that the proposed CPS method is based on the same idea as the tilting method of Cho and Fryzlewicz (2012), namely controlling the effect of the predictors that could generate a confounding effect. However, there is a difference between these two methods in one of their scaling factors. Specifically, our proposed test statistic is

$$\frac{\mathbb{X}_1^\top\mathcal{Q}_{\mathcal{S}}\mathbb{Y}(1-|\mathcal{S}^+|/n)^{1/2}}{(\mathbb{X}_1^\top\mathcal{Q}_{\mathcal{S}}\mathbb{X}_1)^{1/2}\{\mathbb{Y}^\top\mathcal{Q}_{\mathcal{S}^+}\mathbb{Y}\}^{1/2}}$$

and the tilted correlation of Cho and Fryzlewicz (2012) is

$$\frac{\mathbb{X}_1^\top\mathcal{Q}_{\mathcal{S}}\mathbb{Y}}{(\mathbb{X}_1^\top\mathcal{Q}_{\mathcal{S}}\mathbb{X}_1)^{1/2}(\mathbb{Y}^\top\mathcal{Q}_{\mathcal{S}}\mathbb{Y})^{1/2}}.$$

Note that $\mathcal{S}^+ = \mathcal{S}\cup\{1\}$. The asymptotic properties of these two quantities above can be quite different when $\beta_1 \neq 0$ because the difference between $\mathbb{Y}^\top\mathcal{Q}_{\mathcal{S}^+}\mathbb{Y}$ and $\mathbb{Y}^\top\mathcal{Q}_{\mathcal{S}}\mathbb{Y}$ can be large. This indicates that the tilted correlation approach designed for variable selection may not be appropriate for hypothesis testing.

Remark 2. Based on partial correlation, we construct the CPS set. An alternative approach is via the correlation approach proposed by Cho and Fryzlewicz (2012, Section 3.4), who focused on testing correlations between covariates by controlling the false discovery rate. Although their method is quite useful for variable selection, it raises the following two concerns for our testing procedure. First, Theorem 1 may not be valid via the correlation approach. The reason is that Theorem 1 requires the partial covariance, $\max_j|\varrho_{1j}(\mathcal{S})|$, to converge to 0 at a fast rate so that the bias of $\hat{\beta}_1$ is asymptotically negligible; see the proof of Theorem 1 in Appendix B for details. However, the correlation approach only ensures the convergence of $\max_j|\rho_{1j}|$, but not $\max_j|\varrho_{1j}(\mathcal{S})|$. Hence, $\hat{\beta}_1$ may yield a nontrivial bias by using the correlation approach. Second, their method requires that only a small proportion of the $\rho_{j_1j_2}$s are nonzero. Accordingly, it may not be applicable for our proposed test when correlations among predictors are either non-sparse or less sparse (see the covariance structure with the polynomial decay setting above Example 4).

Remark 3. We use a single screening approach to obtain the CPS set of the target covariate, which yields the CPS estimator of the target regression coefficient. On the other hand, Zhang and Zhang (2014) employed the scaled lasso procedure of Sun and Zhang (2012) to obtain the initial estimators of all regression coefficients and the scale parameter estimator. Then, they apply the classical lasso procedure to find the low dimensional projection vector. In sum, Zhang and Zhang (2014) applied the lasso approach to find the low dimensional projection estimator (LDPE) for the target regression coefficient. When $\mathbb{X}$ has orthogonal columns and $p < n$, both approaches lead to the same parameter estimator as that obtained from the marginal univariate regression (MUR). However, these two approaches are quite different, and it seems nearly impossible to find the exact relationship between the CPS estimator and LDPE when the columns of $\mathbb{X}$ are not orthogonal.

2.3. Controlling the False Discovery Rate (FDR)

In identifying significant coefficients among the high dimensional regression coefficients β_j ($j = 1,\ldots,p$), a multiple testing procedure can be considered by testing $H_{0j}: \beta_j = 0$ simultaneously. Denote the p-value obtained by testing each individual null hypothesis, H_{0j}, as $p_j = 2\{1-\Phi(|Z_j|)\}$, where Z_j is the test statistic and can be constructed similarly to that in Eq. (2.4). To guard against false discoveries, we next develop a procedure to control the false discovery rate (Benjamini and Hochberg, 1995).

Let $\mathcal{N}_0 = \{j: \beta_j = 0\}$ be the set of variables whose associated coefficients are truly zero and $\mathcal{N}_1 = \{j: \beta_j \neq 0\}$ be the set of variables whose associated coefficients are truly nonzero. For any significance level $t \in [0,1]$, let $V(t) = \#\{j\in\mathcal{N}_0 : p_j \leq t\}$ be the number of falsely rejected hypotheses, $S(t) = \#\{j\in\mathcal{N}_1 : p_j \leq t\}$ be the number of correctly rejected hypotheses, and $R(t) = \#\{j: p_j \leq t\}$ be the total number of rejected hypotheses. We adopt the approach of Storey et al. (2004) to implement the multiple testing procedure, which is less conservative than the method of Benjamini and Hochberg (1995) and is applicable under a weak dependence structure (Storey et al., 2004). To this end, define $\mathrm{FDP}(t) = V(t)/[R(t)\vee 1]$ and $\mathrm{FDR}(t) = E\{V(t)/[R(t)\vee 1]\}$, where $R(t)\vee 1 = \max\{R(t),1\}$. Then, the estimator proposed by Storey (2002) is

$$\widehat{\mathrm{FDR}}_\lambda(t) = \frac{\hat{\pi}_0(\lambda)t}{\{R(t)\vee 1\}/p}, \tag{2.6}$$

where $\hat{\pi}_0(\lambda) = \{(1-\lambda)p\}^{-1}\{p-R(\lambda)\}$ is an estimate of $\pi_0 = p_0/p$, $p_0 = |\mathcal{N}_0|$ is the number of true null hypotheses, and $\lambda\in[0,1)$ is a tuning parameter. Then, for any pre-specified significance level q and a fixed λ, consider the cutoff point chosen by the thresholding rule, $t_q(\widehat{\mathrm{FDR}}_\lambda) = \sup\{0\leq t\leq 1 : \widehat{\mathrm{FDR}}_\lambda(t)\leq q\}$. We reject the null hypotheses for those p-values that are less than or equal to $t_q(\widehat{\mathrm{FDR}}_\lambda)$.

To study the theoretical property of $\widehat{\mathrm{FDR}}_\lambda(t)$, we begin by introducing two notations. Let $T_{1,n}(t) = p^{-1}\sum_{j=1}^p P(p_j\leq t)$ be the average probability of all rejected hypotheses and $\mathcal{S}_j$ be the CPS set of covariates X_{ij}. We next demonstrate that $t_q(\widehat{\mathrm{FDR}}_\lambda)$ asymptotically provides strong control of FDR at the pre-specified nominal level q.

Theorem 2. *Assume that $p_0/p \to 1$ as p goes to infinity, $\lim_{n\to\infty} T_{1,n}(t) = T_1(t)$ and, for any $k\in\mathcal{N}_0$, $\sum_{j\in\mathcal{N}_0}\sum_{l\in\mathcal{S}_j}\sigma_{lk}^2 = o(p/\Lambda_0^2)$, where $T_1(t)$ is a continuous function and $\Lambda_0 = \max\{\max_{j\in\mathcal{N}_0}|\mathcal{S}_j|, |\mathcal{N}_1|\}$. Under Conditions* (C1)–(C6), *we have that* $\limsup_{n\to\infty}\mathrm{FDR}\{t_q(\widehat{\mathrm{FDR}}_\lambda)\}\leq q$.

The proof is given in Appendix D. In general, the dependences among the test statistics Z_j become stronger as the overlap among the CPS sets increases. To control the dependences, they must have weaker dependence among covariates as the size of overlap

increases. Accordingly, the condition $\sum_{j\in\mathcal{N}_0}\sum_{l\in\mathcal{S}_j}\sigma_{lk}^2 = o(p/\Lambda_0^2)$ for any $k \in \mathcal{N}_0$, in Theorem 2, controls the overall dependence between the covariates in the union of the CPS sets $\cup_{l\in\mathcal{N}_0}\mathcal{S}_l$ for any fixed $k \in \mathcal{N}_0$; see Fan et al. (2012) for a similar condition on dependence. In addition, Λ_0 provides an upper bound on the size of the overlap among the CPS sets. Assume that Σ follows an autoregressive structure such that the $|\mathcal{S}_j|$s are small compared with n and $\sum_j \sigma_{jk}^2 < \infty$ for any $1 \le k \le p$. Hence, the above condition is satisfied. In sum, $\widehat{\mathrm{FDR}}_\lambda$ in (2.6) is applicable under weak dependence. For a more general dependence structure, one might apply the FDP estimation procedure proposed by Fan et al. (2012).

2.4. Model selection consistency

According to Theorem 2, for any given significance level $q > 0$, the FDR can be controlled asymptotically by setting the threshold at $t = t_q(\widehat{\mathrm{FDR}}_\lambda)$. This result motivates us to further investigate the model selection consistency by letting $q \to 0$. In fact, the model selection consistency in high dimensional linear models has been intensively studied in the variable selection literature. There is a large body of papers discussing the model selection consistency via the penalized likelihood approach (e.g., Meinshausen and Bühlmann, 2006; Zhao and Yu, 2006; Huang et al., 2007). However, the use of p-values for model selection has not received considerable attention. Some exceptions include Bunea et al. (2006) who considered variable selection consistency using p-values under the condition $p = o(n^{1/2})$, and Meinshausen et al. (2009) who investigated the consistency of a two-step procedure involving screening and then a multiple test procedure. It is worth noting that the p-value obtained in Meinshausen et al. (2009) is not designed for assessing the significance of a single coefficient. The aim of this section in our paper is to study the model selection consistency using p-values obtained from the test proposed in Section 2.2.

For any given nominal levels α_n, let $\widehat{\mathcal{N}}_1^{\alpha_n} = \{j : p_j \le \alpha_n\}$ be an estimate of $\mathcal{N}_1$, the set containing all the variables whose associated coefficients are truly nonzero. Assume that $\alpha_n \to 0$ as $n \to \infty$. By Theorem 2, the probability of obtaining false discoveries is $P\{|\hat{\mathcal{N}}_1^{\alpha_n} \cap \mathcal{N}_0| \ge 1\} \le \alpha_n \to 0$, which implies that $P\{\hat{\mathcal{N}}_1^{\alpha_n} \subset \mathcal{N}_1\} \to 1$. Thus, this procedure requires a sure screening property $P\{\mathcal{N}_1 \subset \hat{\mathcal{N}}_1^{\alpha_n}\} \to 1$ to obtain model selection consistency. Before demonstrating this property, two additional assumptions are given below.

(C7) There exist two positive constants κ and C_κ such that $\min_{j\in\mathcal{N}_1}|\beta_j| > C_\kappa n^{-\kappa}$ for $\kappa + \hbar < 1/2$, where $\hbar$ is defined in (C3).

(C8) There exists some positive constant C_e such that for any $\ell > 0$ and $1 \le j \le p$, $P\left(n^{-1}|X_j^\top \mathcal{E}| > \ell\right) \le \exp(-C_e n\ell^2)$.

Condition (C7) is a minimum signal assumption, and similar conditions are commonly considered in the variable screening literature (Fan and Lv, 2008; Wang, 2009). We further assume that the random errors ε_i are independent and normally distributed. Using the fact that $n^{-1}\|\mathbb{X}_j\|^2 \to 1$ and that $n^{-1/2}\mathbb{X}_j^\top\mathcal{E}$ follows a normal distribution with finite variance for $j = 1, \ldots, p$, Condition (C8) is satisfied. The above conditions, together with Conditions (C1)–(C6), lead to the following result.

Theorem 3. *Under Conditions* (C1)–(C8)*, there exists a sequence of significance levels* $\alpha_n \to 0$ *such that* $P(\hat{\mathcal{N}}_1^{\alpha_n} = \mathcal{N}_1) \to 1$.

The proof of Theorem 3 is given in Appendix E. According to the proof of Theorem 3, one can select α_n at the level of $\alpha_n = 2\{1 - \Phi(n^\jmath)\}$ with $\hbar < \jmath < 1/2 - \kappa$. This selection implies that $p\alpha_n/\log(p) \to 0$ as $n \to \infty$, which is similar to the assumption (C_q) in Bunea et al. (2006). Compared with the penalized likelihood method, the proposed testing procedure is able to control the false discovery rate and the family-wise error rate for the given α_n. This is important especially in the finite sample case; see Meinshausen et al. (2009) for a detailed discussion.

3. Simulation studies

To demonstrate the finite sample performance of the proposed methods, we consider four simulation studies with different covariance patterns and distributions among predictors. Each simulation includes three different sample sizes ($n = 100, 200, 500$) and two different dimensions of predictors ($p = 1000$ and 2000). All simulation results presented in this section were based on 1000 realizations. The nominal level α of the CPS test and the significance level q of FDR are both set to 5%. Moreover, to determine the CPS set for each predictor, three different significance levels were considered ($\alpha = 0.01, 0.05$, and 0.10). Since the results were similar, we only report the case with the nominal level $\alpha = 0.05$.

To study the significance of each individual regression coefficient, consider the proposed test statistic Z_{rj} for testing the jth coefficient in the rth simulation, where $j = 1, \ldots, p$ and $r = 1, \ldots, 1000$. Then, define an indicator measure $I_{rj} = I(|Z_{rj}| > z_{1-\alpha/2})$ and compute the empirical rejection probability (ERP) for the jth coefficient test, $\mathrm{ERP}_j = 1000^{-1}\sum_{r=1}^{1000} I_{rj}$. As a result, ERP_j is the empirical size under the null hypothesis $H_{0j} : \beta_j = 0$, while it is the empirical power under the alternative hypothesis. Subsequently, define the average empirical size (ES) and the average empirical power (EP) as $\mathrm{ES} = |\mathcal{N}_0|^{-1}\sum_{j\in\mathcal{N}_0}\mathrm{ERP}_j$ and $\mathrm{EP} = |\mathcal{N}_1|^{-1}\sum_{j\in\mathcal{N}_1}\mathrm{ERP}_j$, respectively. Accordingly, ES and EP provide overall measures for assessing the performance of the single coefficient test. Based on the p-values of the Z_{rj} tests, we next employ the multiple testing procedure of Storey et al. (2004) to study the performance of multiple tests via the empirical FDR discussed in Section 2.3. It is worth noting that we adopt the commonly used tuning parameter $\lambda = 1/2$ in the first two examples, and its robustness is evaluated in Example 3. To assess the effect of model selection consistency, we examine the average true rate $\mathrm{TR} = |\hat{\mathcal{N}}_1^\alpha \cap \mathcal{N}_1|/|\mathcal{N}_1|$ and the average false rate $\mathrm{FR} = |\hat{\mathcal{N}}_1^\alpha \cap \mathcal{N}_0|/|\mathcal{N}_0|$. When the true model can be identified consistently, TR and FR should approach 1 and 0, respectively, as the sample size gets large. For the sake of comparison, we also examine the marginal univariate regression (MUR) test (i.e., the classical t-test obtained from the marginal univariate regression model) and the low dimensional projection estimator (LDPE) proposed by Zhang and Zhang (2014) and van de Geer et al. (2014) in Monte Carlo studies. The tuning parameter of the LDPE method is set to $\{2\log p/n\}^{1/2}$, as suggested by Zhang and Zhang (2014). It is noteworthy that we do not include the method of Bühlmann (2013) for comparison since it is not optimal, as shown by van de Geer et al. (2014).

Example 1: Autocorrelated predictors

Consider a linear regression model with autocorrelated predictors X_i generated from a multivariate normal distribution with mean 0 and covariance $\Sigma = (\sigma_{j_1j_2}) \in \mathbb{R}^{p\times p}$ with $\sigma_{j_1j_2} = 0.5^{|j_1-j_2|}$. Although different predictors are correlated with each other, the correlation decreases to 0 as the distance $|j_1 - j_2|$ between X_{ij_1} and X_{ij_2} increases. The regression coefficient vector β is such that $\beta_{3j+1} = 1$ for any $0 \le j \le d_0$, and $\beta_j = 0$ otherwise. Note that $d_0 = |\mathcal{N}_1|$ represents the number of non-zero regression coefficients. In this example, we consider three different values of d_0 ($d_0 = 10, 50, 100$) to investigate the performance of the proposed test under sparse (i.e., $d_0 = 10$) and less sparse (i.e., $d_0 = 50$ and 100) scenarios. In addition, the average variance of ε_i (i.e., $\bar{\sigma}^2$)

Table 1
Simulation results for Example 1 with $\alpha = 5\%$, $q = 5\%$ and $d_0 = 10$.

p	n	Methods	ES	EP	FDR	TR	FR
1000	100	MUR	0.055	0.981	0.753	0.812	0.004
		LDPE	0.071	0.882	0.396	0.808	0.028
		CPS	0.056	0.495	0.127	0.451	0.000
	200	MUR	0.054	1.000	0.726	1.000	0.007
		LDPE	0.059	0.980	0.354	0.901	0.014
		CPS	0.053	0.792	0.073	0.762	0.000
	500	MUR	0.054	1.000	0.691	1.000	0.011
		LDPE	0.055	1.000	0.201	1.000	0.006
		CPS	0.053	0.982	0.053	0.958	0.000
2000	100	MUR	0.057	0.884	0.717	0.731	0.014
		LDPE	0.078	0.731	0.429	0.690	0.035
		CPS	0.055	0.442	0.128	0.362	0.000
	200	MUR	0.053	1.000	0.793	0.951	0.020
		LDPE	0.059	0.941	0.390	0.892	0.009
		CPS	0.053	0.712	0.078	0.668	0.000
	500	MUR	0.052	1.000	0.826	1.000	0.023
		LDPE	0.055	1.000	0.229	1.000	0.006
		CPS	0.052	0.979	0.056	0.971	0.000

Table 2
Simulation results for Example 2 with $\alpha = 5\%$, $q = 5\%$ and $d_0 = 10$.

p	n	Methods	ES	EP	FDR	TR	FR
1000	100	MUR	0.054	0.929	0.289	0.787	0.005
		LDPE	0.070	0.793	0.179	0.765	0.021
		CPS	0.057	0.386	0.120	0.398	0.000
	200	MUR	0.053	1.000	0.236	1.000	0.009
		LDPE	0.061	0.998	0.147	1.000	0.015
		CPS	0.053	0.947	0.068	0.952	0.000
	500	MUR	0.052	1.000	0.186	1.000	0.011
		LDPE	0.054	1.000	0.098	1.000	0.002
		CPS	0.051	1.000	0.051	1.000	0.000
2000	100	MUR	0.054	0.882	0.282	0.716	0.009
		LDPE	0.074	0.737	0.169	0.724	0.027
		CPS	0.058	0.329	0.126	0.325	0.000
	200	MUR	0.053	1.000	0.229	1.000	0.012
		LDPE	0.060	0.968	0.128	0.954	0.015
		CPS	0.054	0.874	0.071	0.901	0.000
	500	MUR	0.051	1.000	0.156	1.000	0.015
		LDPE	0.056	1.000	0.093	1.000	0.006
		CPS	0.051	1.000	0.055	1.000	0.000

is chosen to generate a theoretical $R^2 = \text{var}(X_i^\top\beta)/\{\text{var}(X_i^\top\beta) + \bar{\sigma}^2\} = 0.5$. Moreover, the variance of ε_i (σ_i^2) is independently generated from a uniform distribution with the lower and upper endpoints $\bar{\sigma}^2/2$ and $3\bar{\sigma}^2/2$, respectively. Accordingly, we generate the heteroscedastic linear regression model.

The results for $d_0 = 10$ are presented in Table 1. Since the results for $d_0 = 50$ and 100 yield a similar pattern to those in Table 1, we provide them in the supplementary material to save space. Table 1 shows that both CPS and MUR control the size well, while MUR has a larger power than CPS. After closely examining MUR's performance, however, we find that its ES can be misleading. For example, $X_{i2} \in \mathcal{N}_0$ is moderately correlated with a non-zero predictor $X_{i1} \in \mathcal{N}_1$. As a result, the empirical size for testing $H_0 : \beta_2 = 0$ obtained from MUR could be as large as 0.90 in almost all realizations. On the other hand, most predictors in $\mathcal{N}_0$ are nearly independent of the predictors in $\mathcal{N}_1$ and the response variable. Accordingly, MUR can have a reasonable average empirical size and a high average true rate (TR). This misleading result can be detected by the empirical false discovery rate (FDR) being much greater than the nominal level. In addition, the average false rate (FR) becomes larger as the sample size increases. Therefore, the MUR approach should be used with caution when testing a single coefficient, conducting multiple hypothesis tests, or selecting variables.

We next study the performance of LDPE. Table 1 indicates that, although LDPE can control the size well at a reasonable level, it fails to control the FDR at the nominal level, particularly in small samples. For instance, when the sample size $n = 100$, the FDR values are 0.396 and 0.429 for $p = 1000$ and $p = 2000$, respectively. In contrast to MUR and LDPE, the CPS approach not only controls the size well, but also leads to FDR converging to the nominal level as the sample size increases. Furthermore, the average TR increases towards 1 and the average FR decreases to 0, both of which are consistent with theoretical findings.

In addition to $d_0 = 10$, the results for $d_0 = 50$ and 100 in Tables S1 and S2 of the supplementary material indicate that CPS is still superior to MUR and LDPE under the less sparse scenario. It is of interest to note that LDPE does not control the size well under less sparse regression models. This finding is not surprising since LDPE depends heavily on the accuracy of estimating the whole vector β. In sum, CPS performs well for testing a single coefficient, and the resulting p-values are reliable for multiple hypothesis tests and model selection.

Example 2: Moving average predictors

In this example, we generate data from a linear regression model with predictors following the moving average model with order 1: $X_i = u_i + 0.5u_{i-1}$ for $i = 2, \ldots, n$ and $X_1 = u_1$, where u_i are independently generated from a multivariate normal distribution with mean 0 and covariance $0.8I_p$ for $i = 1, \ldots, n$. Accordingly, the covariance matrix of X_i can be written as $\Sigma = (\sigma_{j_1j_2}) \in \mathbb{R}^{p\times p}$ with $\sigma_{j_1j_2} = 1$ if $j_1 = j_2$, $\sigma_{j_1j_2} = 0.4$ if $|j_1 - j_2| = 1$, and $\sigma_{j_1j_2} = 0$ otherwise. The regression coefficients β, the number of non-zero coefficients d_0, and the variance of ε_i (σ_i^2), are the same as those in Example 1.

Table 2 reports the results for $d_0 = 10$, and similar findings for $d_0 = 50$ and 100 can be found in Tables S3 and S4, respectively, of the supplementary material. Table 2 shows that both CPS and MUR control the size well. However, MUR fails to control FDR at the nominal level. In fact, its FDR is much greater than the nominal level. In addition, its average false rate (FR) becomes larger as the sample size increases. We next study the performance of LDPE. Table 2 indicates that LDPE fails to control the FDR at the nominal level, particularly in small samples, although it can control the size well at a reasonable level. In addition, LDPE fails to control the size well for less sparse regression models (see Tables S3 and S4 for $d_0 = 50$ and 100, respectively, in the supplementary material). This finding is not surprising since LDPE depends heavily on the accuracy of estimating the whole vector β. In contrast to MUR and LDPE, the resulting p-values obtained by CPS are reliable for multiple hypothesis tests and model selections. Furthermore, CPS performs well even under less sparse models (see Tables S3 and S4 in the supplementary material), and this nice property is not enjoyed by MUR and LDPE.

Example 3: Equally correlated predictors

Consider a model with equally correlated predictors, X_i, generated from a multivariate normal distribution with mean 0 and a compound symmetric covariance matrix $\Sigma = (\sigma_{j_1j_2}) \in \mathbb{R}^{p\times p}$, where $\sigma_{j_1j_2} = 1$ if $j_1 = j_2$ and $\sigma_{j_1j_2} = 0.5$ for any $j_1 \neq j_2$. In addition, the regression coefficients are set as follows: $\beta_j = 5$ for $1 \leq j \leq d_0$, and $\beta_j = 0$ for $j > d_0$. The number of non-zero regression coefficients d_0 and the variance of ε_i (σ_i^2) are the same as those in Example 1.

To save space, we only present the results for $d_0 = 10$ in Table 3, and the results for $d_0 = 50$ and 100 are in Tables S5 and S6, respectively, of the supplementary material. Table 3 indicates that MUR performs poorly in terms of both ES and FDR measures. This

Table 3
Simulation results for Example 3 with $\alpha = 5\%$, $q = 5\%$ and $d_0 = 10$.

p	n	Test	ES	EP	FDR	TR	FR
1000	100	MUR	1.000	1.000	0.997	1.000	1.000
		LDPE	0.054	0.528	0.349	0.477	0.009
		CPS	0.058	0.449	0.141	0.418	0.000
	200	MUR	1.000	1.000	0.998	1.000	1.000
		LDPE	0.054	0.832	0.207	0.719	0.005
		CPS	0.049	0.747	0.051	0.701	0.000
	500	MUR	1.000	1.000	0.998	1.000	1.000
		LDPE	0.051	1.000	0.162	0.981	0.004
		CPS	0.049	0.987	0.048	0.965	0.000
2000	100	MUR	1.000	1.000	0.996	1.000	1.000
		LDPE	0.050	0.465	0.402	0.428	0.008
		CPS	0.055	0.398	0.138	0.366	0.000
	200	MUR	1.000	1.000	0.997	1.000	1.000
		LDPE	0.049	0.776	0.232	0.713	0.005
		CPS	0.051	0.689	0.055	0.664	0.000
	500	MUR	1.000	1.000	0.995	1.000	1.000
		LDPE	0.050	1.000	0.144	0.962	0.002
		CPS	0.050	0.937	0.050	0.915	0.000

finding is not surprising because every predictor in $\mathcal{N}_0$ is equally correlated with those predictors in $\mathcal{N}_1$. As a result, the marginal correlation between any predictor in $\mathcal{N}_0$ and the response variable is bounded well away from 0. Thus, MUR's empirical rejection probability is close to 100%, which leads to highly inflated ES and FDR. Furthermore, FR equals 1 at all sample sizes, which implies that MUR tends to over reject the null hypothesis. Moreover, the results of LDPE are similar to those in Tables 1–2. On the other hand, the ES and FDR of CPS are close to the nominal level, except for the case of CPS with $n = 100$. Moreover, TR and EP increase towards 1 as the sample size gets large, and FR equals 0 at all sample sizes.

From the above simulation studies, we find that FDR plays an important role for examining the reliability of test statistics. Hence, we next study the accuracy for the estimation of FDR discussed in Section 2.3. Since we are interested in the statistical behavior of the number of false discoveries $V(t)$, we follow Fan et al.'s (2012) suggestion and compare $\widehat{\mathrm{FDR}}_\lambda(t)$ in (2.6), with $\lambda = 1/2$, to $\mathrm{FDP}(t)$ calculated via $V(t)/[R(t) \vee 1]$. For the sake of illustration, we consider the same simulation settings as given in Examples 1 and 3 with $n = 100, p = 1000$ and $d_0 = 10$. Panels A and B in Fig. 1 depict $\widehat{\mathrm{FDR}}_\lambda(t)$ and $\mathrm{FDP}(t)$, obtained via the CPS method for Examples 1 and 3, respectively, across various t values. In contrast, Panels C and D are calculated via the MUR approach and Panels E and F are calculated via the LDPE method. Fig. 1 clearly shows that $\widehat{\mathrm{FDR}}_\lambda(t)$ calculated from the p-values of CPS is reliable and consistent with the theoretical finding in Theorem 2. However, MUR and LDPE do not provide accurate estimates of FDP, and they should be used with caution in high dimensional data analysis.

The above three examples have demonstrated that CPS performs well across three commonly used covariance structures. It is worth noting that Conditions (C4) and (C5) hold in the first two examples, while these conditions are invalid in the third example. However, CPS still performs well in Example 3, which shows its robustness. Motivated by an anonymous referee's comments, we present an additional study with the covariance structure $\Sigma = I_p + uu^\top$, where $u = (u_1, \ldots, u_p)^\top \in \mathbb{R}^p, u_j = \delta j^{-2}$ for $j = 1, \ldots, p$, and δ is a finite constant. Accordingly, $\mathrm{cov}(X_{i1}, X_{ij}) = (\delta/j)^2$ so that covariates exhibit polynomial decay and $\rho_{1j} = (\delta/j)^2/\{(1+\delta^2)(1+\delta^2/j^4)\}^{0.5}$. Hence, there are quite a number of predictors that are highly correlated with X_{i1} when δ is large enough. One can also verify that $\max_{j \notin \mathcal{S}} |\varrho_{1j}(\mathcal{S})| = O(|\mathcal{S}|^{-2})$ as $|\mathcal{S}| \to \infty$, and then both Conditions (C4) and (C5) hold. Our simulation results indicate that CPS performs well; see Table S7 in the supplementary material.

Table 4
Simulation results for Example 4 with $\alpha = 5\%$, $q = 5\%$, $d_0 = 10$, Σ as given in Example 1, $\lambda = 0.1$, and X_is being generated from a standardized exponential distribution.

p	n	Methods	ES	EP	FDR	TR	FR
1000	100	MUR	0.055	0.962	0.728	0.833	0.008
		LDPE	0.068	0.881	0.407	0.816	0.021
		CPS	0.057	0.518	0.129	0.492	0.000
	200	MUR	0.055	1.000	0.721	0.975	0.012
		LDPE	0.058	0.995	0.349	0.912	0.014
		CPS	0.052	0.775	0.067	0.744	0.000
	500	MUR	0.052	1.000	0.722	1.000	0.016
		LDPE	0.055	1.000	0.174	1.000	0.006
		CPS	0.052	0.991	0.053	0.971	0.000
2000	100	MUR	0.056	0.882	0.723	0.711	0.015
		LDPE	0.069	0.731	0.441	0.704	0.022
		CPS	0.058	0.449	0.130	0.401	0.000
	200	MUR	0.056	1.000	0.744	0.922	0.018
		LDPE	0.063	0.943	0.386	0.915	0.016
		CPS	0.054	0.738	0.076	0.703	0.000
	500	MUR	0.052	1.000	0.782	1.000	0.022
		LDPE	0.053	1.000	0.196	1.000	0.005
		CPS	0.053	0.988	0.057	0.981	0.000

Example 4: Robustness of covariate distribution and λ parameter

In the first three examples, the covariate vector X_is were generated from a multivariate normal distribution and the tuning parameter λ was set to be 1/2. To assess the robustness of CPS against the covariate distribution and λ, we conduct simulation studies for various λs and three distributions of $X_i = \Sigma^{1/2} Z_i$, where each element of Z_i is randomly generated from the standard normal distribution, the standardized exponential distribution exp(1), and the normal mixture distribution $0.1N(0, 3^2) + 0.9N(0, 1)$, respectively, for $i = 1, \ldots, n$, and the Σs are correspondingly defined in Examples 1–3. Since all results are qualitatively similar, we only report the case when $\lambda = 0.1$, $d_0 = |\mathcal{N}_1| = 10$ and Z_i follows a standardized exponential distribution. The results in Tables 4–6 show similar findings to those in Tables 1–3, respectively. Hence, Monte Carlo studies indicate that the CPS approach is robust against the covariate distribution and the threshold parameter λ.

4. Real data analysis

To illustrate the usefulness of the proposed method, we consider two empirical examples. The first example analyzes financial data and the second example studies supermarket data.

4.1. Index fund data

The data set consists of a total of $n = 155$ observations, in which the response Y_i is the weekly return of the Shanghai composite index. Explanatory variables X_i are $p = 382$ stock returns that traded on the Shanghai stock exchange during the period from Oct. 9, 2010 to Sep. 28, 2013, with $i = 1, \ldots, 155$. We assume that there is a linear relationship between Y_i and X_i, which is $Y_i = X_i^\top \beta + \varepsilon_i$, as given in Eq. (2.1). In addition, both the response and predictors are standardized so that they have zero mean and unit variance. The task of this study is identifying a small number of relevant stocks that financial managers can use to establish a portfolio that tracks the return of the Shanghai composite index.

To identify important stocks (predictors) that are associated with Y_i, we employ the CPS, MUR, and LDPE methods and test the significance of each individual regression coefficient, namely, testing $H_{0j} : \beta_j = 0$ vs. $H_{1j} : \beta_j \neq 0$ for $j = 1, \ldots, 382$. Here,

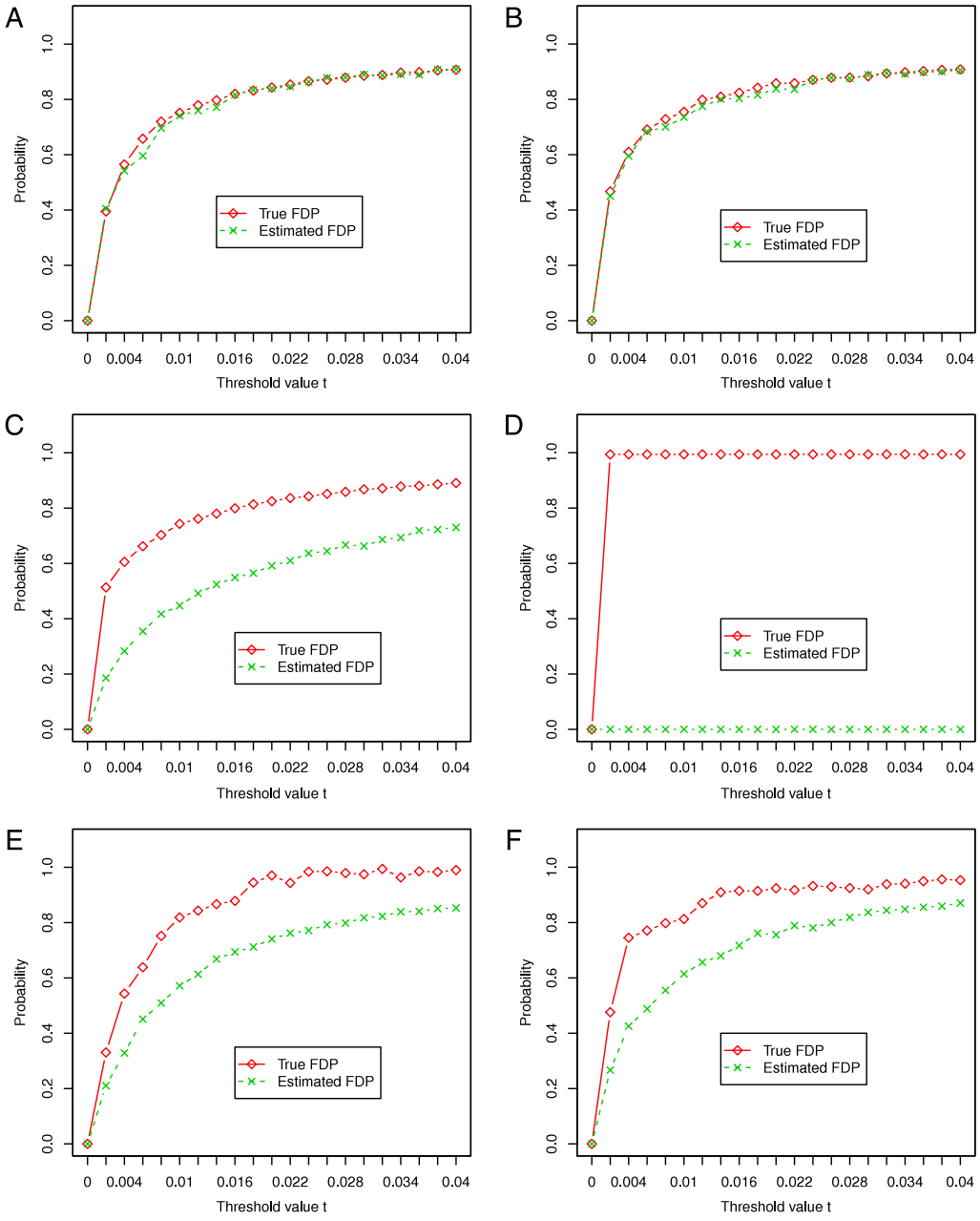

Fig. 1. Panels A and B depict the estimated FDP value (i.e., $\widehat{\mathrm{FDR}}_{0.5}(t)$) compared with the true FDP value obtained via the CPS method for Examples 1 and 3, respectively. Panels C and D are obtained via the MUR approach, and Panels E and F are obtained via the LDPE method.

the tuning parameter of the LDPE method is set to $\{2\log p/n\}^{1/2}$, as suggested by Zhang and Zhang (2014). Since the asymptotic distribution of the p-values obtained from the above test statistics is uniform [0, 1], we use the histogram to effectively illustrate their performances. Fig. 2 depicts the histograms of the p-values for testing H_{0j} ($j = 1, \ldots, 382$) via three tests. Based on the CPS test, we find 32 p-values that are less than the significance level $\alpha = 5\%$. After controlling the false discoveries rate via the method of Storey et al. (2004) at the level of $q = 5\%$, the number of hypotheses H_{0j} being rejected is 12. As a result, we have identified the 12 most important stocks that can be used for index tracking.

In contrast, the histogram of the p-values calculated from the MUR tests is heavily skewed with very thin tails. This suggests that most of its p-values are very small. Consequently, it rejected a total of 161 hypotheses H_{0j} after controlling the FDR at the level of $q = 5\%$. This finding is not surprising since the covariates in the model are highly correlated due to the existence of latent factors, as observed by Fama and French (1993). Analogous results can be found in the histogram of the p-values generated from LDPE. In sum, CPS is able to identify the most relevant stocks from high dimensional data, while MUR and LDPE cannot.

4.2. Supermarket data

This data set contains a total of $n = 464$ daily records. For each record, the response variable (Y_i) is the number of customers and the predictors $(X_{i1}, \ldots, X_{ip})$ are the sales volumes of $p = 6398$ products. Consider a linear relationship between Y_i and $X_i = (X_{i1}, \ldots, X_{ip})^\top \in \mathbb{R}^p$, given by $Y_i = X_i^\top \beta + \varepsilon_i$, where both the response and predictors are standardized so that they have zero mean and unit variance. The purpose of this study is to determine a small number of products that attract the most customers.

We apply the proposed CPS, MUR, and LDPE methods to test the significance of each regression coefficient, namely $H_{0j} : \beta_j = 0$ vs.

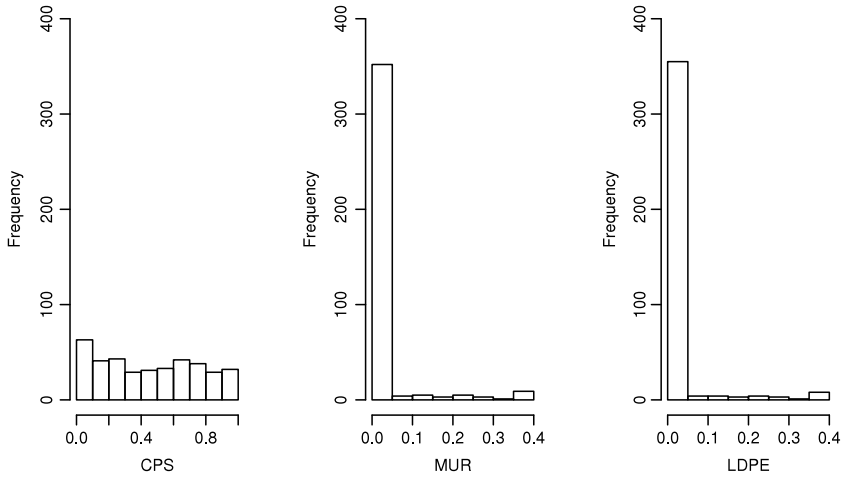

Fig. 2. Index fund data. The histograms of the p-values for the CPS, MUR, and LDPE tests.

Table 5
Simulation results for Example 4 with $\alpha = 5\%$, $q = 5\%$, $d_0 = 10$, Σ as given in Example 2, $\lambda = 0.1$, and X_is being generated from a standardized exponential distribution.

p	n	Test	ES	EP	FDR	TR	FR
1000	100	MUR	0.055	0.938	0.296	0.792	0.009
		LDPE	0.071	0.802	0.185	0.772	0.022
		CPS	0.056	0.404	0.132	0.401	0.000
	200	MUR	0.055	1.000	0.228	1.000	0.012
		LDPE	0.064	1.000	0.152	1.000	0.013
		CPS	0.052	0.943	0.070	0.947	0.000
	500	MUR	0.053	1.000	0.190	1.000	0.017
		LDPE	0.054	1.000	0.103	1.000	0.004
		CPS	0.052	1.000	0.052	1.000	0.000
2000	100	MUR	0.053	0.891	0.290	0.725	0.009
		LDPE	0.072	0.744	0.181	0.731	0.030
		CPS	0.059	0.352	0.133	0.332	0.000
	200	MUR	0.053	1.000	0.242	1.000	0.013
		LDPE	0.059	0.977	0.153	0.973	0.018
		CPS	0.054	0.853	0.069	0.911	0.000
	500	MUR	0.054	1.000	0.134	1.000	0.015
		LDPE	0.053	1.000	0.097	1.000	0.009
		CPS	0.052	1.000	0.058	1.000	0.000

Table 6
Simulation results for Example 4 with $\alpha = 5\%$, $q = 5\%$, $d_0 = 10$, Σ as given in Example 3, $\lambda = 0.1$, and X_is being generated from a standardized exponential distribution.

p	n	Test	ES	EP	FDR	TR	FR
1000	100	MUR	1.000	1.000	0.998	1.000	1.000
		LDPE	0.054	0.521	0.374	0.496	0.005
		CPS	0.059	0.449	0.126	0.433	0.000
	200	MUR	1.000	1.000	0.998	1.000	1.000
		LDPE	0.054	0.832	0.211	0.743	0.002
		CPS	0.049	0.743	0.052	0.683	0.000
	500	MUR	1.000	1.000	0.998	1.000	1.000
		LDPE	0.050	1.000	0.127	0.982	0.001
		CPS	0.049	0.982	0.051	0.970	0.000
2000	100	MUR	1.000	1.000	0.996	1.000	1.000
		LDPE	0.052	0.471	0.412	0.451	0.012
		CPS	0.054	0.429	0.143	0.352	0.000
	200	MUR	1.000	1.000	0.996	1.000	1.000
		LDPE	0.052	0.772	0.217	0.717	0.005
		CPS	0.051	0.712	0.058	0.686	0.000
	500	MUR	1.000	1.000	0.997	1.000	1.000
		LDPE	0.050	1.000	0.143	0.981	0.002
		CPS	0.052	0.952	0.050	0.937	0.000

$H_{1j} : \beta_j \neq 0$. Fig. 3 depicts the three histograms of the p-values computed via the CPS, MUR, and LDPE methods, respectively. As one can observe from the histogram, for the CPS method, the pattern indicates that most of the H_{0j} are true and the p-values are asymptotically valid. There were 1426 p-values that are less than the significance level $\alpha = 5\%$. After controlling the false discoveries rate via the method of Storey et al. (2004) at the level of $q = 5\%$, the number of hypotheses H_{0j} being rejected is 132. In other words, we have identified 132 most important products on which the supermarket decision maker (or manager) might perform further analysis. In contrast, for the MUR method, the histogram of the p-values are extremely skewed with very thin tails. It rejected a total of 5648 hypotheses H_{0j} after controlling the FDR at the level of $q = 5\%$. In addition, the histogram of the p-values generated from the LDPE tests in Fig. 3 shows a flat pattern within the entire interval [0, 1]. As a result, it is not surprising to find that there were a total of 535 p-values that are less than the significance level $\alpha = 5\%$, while none of them were significant after controlling the false discoveries rate at the level of $q = 5\%$.

The above two above examples indicate that the CPS method not only practically provides a simple and efficient approach to compute the p-value for testing a single coefficient in a high dimensional linear model, but also results in reliable p-values for multiple hypothesis testing.

5. Discussion

In linear regression models with high dimensional data, we propose a single screening procedure, Correlated Predictors Screening (CPS), to control for predictors that are highly correlated with the target covariate. This allows us to employ the classical ordinary least squares approach to obtain the parameter estimator. We then demonstrate that the resulting estimator is asymptotically normal. Accordingly, we extend the classical t-test (or z-test) for testing a single coefficient to the high dimensional setting. Based on the p-value obtained from testing the significance of each covariate, the multiple hypothesis testing is established by controlling the false discovery rate at the nominal level. In addition, we show that multiple hypothesis test leads to consistent model selection. Accordingly, the main focus of this paper is on statistical inference rather than variable selection and parameter estimation, which are often the aims of regularization methods such as LASSO (Tibshirani, 1996) and SCAD (Fan and Li, 2001).

The proposed CPS method can be extended for testing a small subset of regression coefficients. Consider the hypothesis:

$$H_0 : \beta_{\mathcal{M}} = 0 \quad \text{vs.} \quad H_1 : \beta_{\mathcal{M}} \neq 0, \tag{5.1}$$

where $\mathcal{M}$ is a pre-specified index set with a fixed size and $\beta_{\mathcal{M}} = (\beta_j : j \in \mathcal{M})^\top \in \mathbb{R}^{|\mathcal{M}|}$ is the subvector of β corresponding

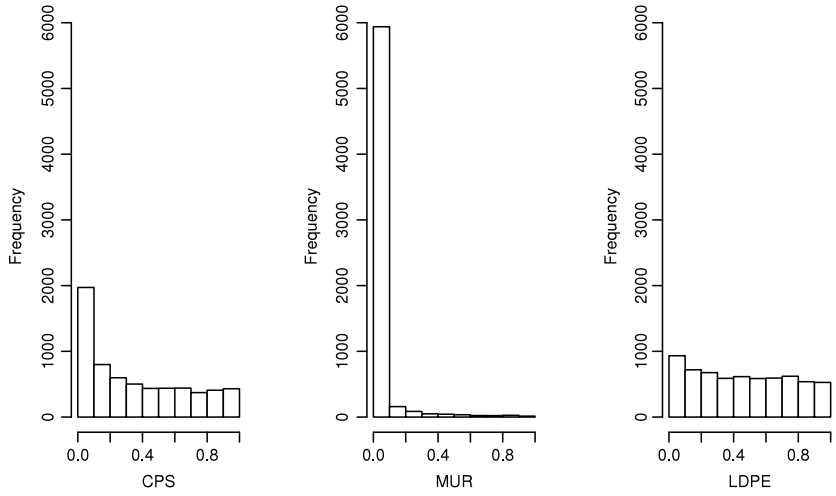

Fig. 3. Supermarket data. The histograms of the p-values for the CPS, MUR, and LDPE tests.

to $\mathcal{M}$. Without loss of generality, we assume that $\mathcal{M} = \{j : 1 \le j \le |\mathcal{M}|\}$ and $1 < |\mathcal{M}| \ll n$. Then, define an overall CPS set of $\mathcal{M}$ as $\mathcal{S}_{\mathcal{M}} = \bigcup_{j\in\mathcal{M}} \mathcal{S}_j$, where $\mathcal{S}_j$ is the CPS set for the jth predictor in $\mathcal{M}$. Accordingly, the target parameter estimator $\beta_{\mathcal{M}} = (\beta_j : j \in \mathcal{M})^\top \in \mathbb{R}^{|\mathcal{M}|}$ can be estimated by $\hat{\beta}_{\mathcal{M}} = (\mathbb{X}_{\mathcal{M}}^\top \mathcal{Q}_{\mathcal{S}_{\mathcal{M}}} \mathbb{X}_{\mathcal{M}})^{-1}(\mathbb{X}_{\mathcal{M}}^\top \mathcal{Q}_{\mathcal{S}_{\mathcal{M}}} \mathbb{Y})$, where $\mathbb{X}_{\mathcal{M}} = (\mathbb{X}_j : j \in \mathcal{M}) \in \mathbb{R}^{n\times|\mathcal{M}|}$. Applying similar techniques to those used in the proof of Theorem 1, we can show that $n^{1/2}(\hat{\beta}_{\mathcal{M}} - \beta_{\mathcal{M}}) \to_d N(0, \Sigma_\beta)$, where $\Sigma_\beta = \sigma_{\mathcal{M}}^2(\Sigma_{\mathcal{M}} - \Sigma_{\mathcal{S}_{\mathcal{M}}\mathcal{M}}^\top \Sigma_{\mathcal{S}}^{-1}\Sigma_{\mathcal{S}_{\mathcal{M}}\mathcal{M}})^{-1}$ with $\sigma_{\mathcal{M}}^2 = \beta_{\mathcal{S}_{\mathcal{M}}^*}^T(\Sigma_{\mathcal{S}_{\mathcal{M}}^*} - \Sigma_{\mathcal{S}_{\mathcal{M}}^*\mathcal{S}_{\mathcal{M}}^+}\Sigma_{\mathcal{S}_{\mathcal{M}}^+}^{-1}\Sigma_{\mathcal{S}_{\mathcal{M}}^+\mathcal{S}_{\mathcal{M}}^*})\beta_{\mathcal{S}_{\mathcal{M}}^*} + \bar{\sigma}^2$, $\mathcal{S}_{\mathcal{M}}^+ = \cup_{j\in\mathcal{M}} \mathcal{S}_j^+$ and $\mathcal{S}_{\mathcal{M}}^* = \{j : j \notin \mathcal{S}_{\mathcal{M}}^+\}$. Consequently, an F-type test statistic can be constructed to test (5.1).

To broaden the usefulness of the proposed method, we conclude the article by discussing three possible research avenues. Firstly, from the model aspect, it would be practically useful to extend the CPS method to generalized linear models, single index models, partial linear models, and survival models. Secondly, from the data aspect, it is important to generalize the proposed CPS method to accommodate category explanatory variables, repeated measurements, and missing observations. Lastly, to control the FDR at the nominal level, we have imposed a weak dependence assumption in Theorem 2. Hence, it would be useful to employ the method of Fan et al. (2012) to adjust for the arbitrary covariance dependence among test statistics Z_j. We believe these extensions would enhance the usefulness of CPS in high dimensional data analysis.

Acknowledgments

Wei Lan's research was supported by National Natural Science Foundation of China (NSFC, 11401482, 71532001). Ping-Shou Zhong's research was supported by a National Science Foundation grant DMS 1309156. Runze Li's research was supported by a National Science Foundation grant DMS 1512422, National Institute on Drug Abuse (NIDA) grants P50 DA039838, P50 DA036107, and R01 DA039854. Hansheng Wang's research was supported in part by National Natural Science Foundation of China (NSFC, 11131002, 11271031, 71532001), the Business Intelligence Research Center at Peking University, and the Center for Statistical Science at Peking University. The authors thank the Editor, the AE and reviewers for their constructive comments, which have led to a dramatic improvement of the earlier version of this paper. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NSF, NIH and NIDA.

Appendix A. Four useful lemmas

Before proving the theoretical results, we present the following four lemmas which are needed in the proofs. The first lemma is directly borrowed from Lemma A.3 of Bickel and Levina (2008), and the second lemma can be found in Bendat and Piersol (1966). As a result, we only verify the third and fourth lemmas.

Lemma 1. *Let* $\hat{\sigma}_{j_1j_2} = n^{-1}\sum_i X_{ij_1}X_{ij_2}$ *and* $\hat{\rho}_{j_1j_2} = \hat{\sigma}_{j_1j_2}/\{\hat{\sigma}_{j_1j_1}\hat{\sigma}_{j_2j_2}\}^{1/2}$, *and assume that Condition* (C1) *holds. Then, there exist three positive constants* $\zeta_0 > 0$, $C_1 > 0$, *and* $C_2 > 0$, *such that* (i) $P(|\hat{\sigma}_{j_1j_2} - \sigma_{j_1j_2}| > \zeta) \le C_1\exp(-C_2n\zeta^2)$ *and* (ii) $P(|\hat{\rho}_{j_1j_2} - \rho_{j_1j_2}| > \zeta) \le C_1\exp(-C_2n\zeta^2)$ *for any* $0 < \zeta < \zeta_0$ *and every* $1 \le j_1, j_2 \le p$.

Lemma 2. *Let* $(U_1, U_2, U_3, U_4)^\top \in \mathbb{R}^4$ *be a* 4*-dimensional normal random vector with* $E(U_j) = 0$ *and* $var(U_j) = 1$ *for* $1 \le j \le 4$. *We then have* $E(U_1U_2U_3U_4) = \delta_{12}\delta_{34} + \delta_{13}\delta_{24} + \delta_{14}\delta_{23}$, *where* $\delta_{ij} = E(U_iU_j)$.

Lemma 3. *Assume that Conditions* (C1)–(C3) *hold, and* $m = O(n^{\nu_2})$ *for some positive constant* ν_2 *which satisfies* $3\nu_2 + \hbar < 1$, *where* $\hbar$ *is given in* (C3). *Then,* $\max_{|\mathcal{S}|\le m} |n^{-1}\mathbb{X}_1^\top \mathcal{Q}_{\mathcal{S}}\mathbb{X}_1 - (\sigma_{11} - \Sigma_{1\mathcal{S}}\Sigma_{\mathcal{S}}^{-1}\Sigma_{\mathcal{S}1})| \to_p 0$.

Proof. Since $n^{-1}\mathbb{X}_1^\top \mathcal{Q}_{\mathcal{S}}\mathbb{X}_1 = n^{-1}\mathbb{X}_1^\top\mathbb{X}_1 - (n^{-1}\mathbb{X}_1^\top\mathbb{X}_{\mathcal{S}})(n^{-1}\mathbb{X}_{\mathcal{S}}^\top\mathbb{X}_{\mathcal{S}})^{-1}(n^{-1}\mathbb{X}_{\mathcal{S}}^\top\mathbb{X}_1)$ and $n^{-1}\mathbb{X}_1^\top\mathbb{X}_1 \to_p \sigma_{11}$, it suffices to show that

$$\max_{|\mathcal{S}|\le m}\Big|(n^{-1}\mathbb{X}_1^\top\mathbb{X}_{\mathcal{S}})(n^{-1}\mathbb{X}_{\mathcal{S}}^\top\mathbb{X}_{\mathcal{S}})^{-1}(n^{-1}\mathbb{X}_{\mathcal{S}}^\top\mathbb{X}_1) - \Sigma_{1\mathcal{S}}\Sigma_{\mathcal{S}}^{-1}\Sigma_{\mathcal{S}1}\Big| \to_p 0. \tag{A.1}$$

Denote $\|A\| = \{tr(AA^\top)\}^{1/2}$ for any arbitrary matrix A. Since $\sigma_{11} - \Sigma_{1\mathcal{S}}\Sigma_{\mathcal{S}}^{-1}\Sigma_{\mathcal{S}1}$ is the conditional variance of X_1 given $X_{\mathcal{S}}$, $\sigma_{11} - \Sigma_{1\mathcal{S}}\Sigma_{\mathcal{S}}^{-1}\Sigma_{\mathcal{S}1} \ge 0$. Then by Condition (C2), we have $\Sigma_{1\mathcal{S}}\Sigma_{\mathcal{S}}^{-1}\Sigma_{\mathcal{S}1} \le \sigma_{11} \le c_{\max} < \infty$. Then, we obtain (A.1) if the following two uniform convergence results hold:

$$\max_{|\mathcal{S}|\le m}\left\|n^{-1}\mathbb{X}_{\mathcal{S}}^\top\mathbb{X}_1 - \Sigma_{\mathcal{S}1}\right\| = o_p(1), \tag{A.2}$$

$$\text{and}\quad \max_{|\mathcal{S}|\le m}\left\|n^{-1}\mathbb{X}_{\mathcal{S}}^\top\mathbb{X}_{\mathcal{S}} - \Sigma_{\mathcal{S}}\right\| = o_p(1). \tag{A.3}$$

Accordingly, it suffices to demonstrate (A.2) and (A.3).

It is noteworthy that, for any $\mathcal{S}$ satisfying $|\mathcal{S}| \le m$, we have

$$\left\|n^{-1}\mathbb{X}_{\mathcal{S}}^{\top}\mathbb{X}_1 - \Sigma_{\mathcal{S}1}\right\| = \left\{\sum_{j\in\mathcal{S}}(\hat{\sigma}_{1j}-\sigma_{1j})^2\right\}^{1/2} \le m^{1/2}\max_{j\in\mathcal{S}}|\hat{\sigma}_{1j}-\sigma_{1j}|.$$

This, together with the Bonferroni inequality, Condition (C1), Lemma 1(i), and the fact that $\#\{\mathcal{S}\subset\{1,\ldots,p\} : |\mathcal{S}| \le m\} \le p^m$, implies

$$\begin{aligned}
&P\Big(\max_{|\mathcal{S}|\le m}\left\|n^{-1}\mathbb{X}_{\mathcal{S}}^{\top}\mathbb{X}_1 - \Sigma_{\mathcal{S}1}\right\| > \epsilon\Big)\\
&\le \sum_{|\mathcal{S}|\le m} P\Big(\left\|n^{-1}\mathbb{X}_{\mathcal{S}}^{\top}\mathbb{X}_1 - \Sigma_{\mathcal{S}1}\right\| > \epsilon\Big)\\
&\le \sum_{|\mathcal{S}|\le m} P\Big(\max_{j\in\mathcal{S}}|\hat{\sigma}_{1j}-\sigma_{1j}| > \epsilon/m^{1/2}\Big)\\
&\le \sum_{|\mathcal{S}|\le m}\sum_{j\in\mathcal{S}} P\Big(|\hat{\sigma}_{1j}-\sigma_{1j}| > \epsilon/m^{1/2}\Big)\\
&\le p^m m C_1\exp(-C_2 n m^{-1}\epsilon^2)\\
&= C_1\exp\Big(-C_2 n m^{-1}\epsilon^2 + \log m + m\log p\Big). \qquad \text{(A.4)}
\end{aligned}$$

Furthermore, by the assumptions in Lemma 3 ($m = O(n^{\nu_2})$) and Condition (C3) ($\log p \le \nu n^{\hbar}$), we have that $m\log p = O(n^{\nu_2+\hbar})$. Moreover, using the assumptions in Lemma 3 again ($3\nu_2+\hbar<1$), the right-hand side of (A.4) converges towards 0 as $n\to\infty$. Hence, we have proved (A.2). Applying similar techniques to those used in the proof of (A.2), we can also demonstrate (A.3). This completes the entire proof.

Lemma 4. *Assume that* (a) $\lim_{p\to\infty} V(t)/p_0 = G_0(t)$ *and* $\lim_{p\to\infty} S(t)/(p-p_0) = G_1(t)$, *where* $G_0(t)$ *and* $G_1(t)$ *are continuous functions;* (b) $0 < G_0(t) \le t$ *for* $t\in(0,1]$; (c) $\lim_{p\to\infty} p_0/p = 1$. *Then, we have* $\limsup_{p\to\infty}\mathrm{FDR}\{t_\alpha(\widehat{\mathrm{FDR}}_\lambda)\} \le \alpha$.

Proof. By slightly modifying the proof of Theorem 4 in Storey et al. (2004), we can demonstrate the result. The detailed proof can be obtained from the authors upon request.

Appendix B. Proof of Theorem 1

Let $T_1 = (\mathbb{X}_1^{\top}\mathcal{Q}_{\mathcal{S}}\mathbb{X}_1)^{-1}\mathbb{X}_1^{\top}\mathcal{Q}_{\mathcal{S}}\mathcal{E}$ and $T_2 = (\mathbb{X}_1^{\top}\mathcal{Q}_{\mathcal{S}}\mathbb{X}_1)^{-1}\mathbb{X}_1^{\top}\mathcal{Q}_{\mathcal{S}}\mathbb{X}_{\mathcal{S}^*}\beta_{\mathcal{S}^*}$. Then, $\hat{\beta}_1-\beta_1 = T_1+T_2$. Using the fact that $E(\varepsilon_i|X_i)=0$, one can show that $\mathrm{cov}(T_1,T_2) = E(T_1T_2)-E(T_1)E(T_2) = 0$. Therefore, T_1 and T_2 are uncorrelated. To prove the theorem, hence, it suffices to show that $(\sqrt{n}T_1, \sqrt{n}T_2)^{\top}$ is asymptotically bivariate normal. By Conditions (C1)–(C3) and Lemma 2, we obtain that $\|n^{-1}\mathbb{X}_1^{\top}\mathbb{X}_{\mathcal{S}} - \Sigma_{1\mathcal{S}}\| \to_p 0$, $\|n^{-1}\mathbb{X}_{\mathcal{S}}^{\top}\mathbb{X}_{\mathcal{S}} - \Sigma_{\mathcal{S}}\| \to_p 0$, and $\max_{\mathcal{S}}\left|n^{-1}\mathbb{X}_1^{\top}\mathcal{Q}_{\mathcal{S}}\mathbb{X}_1 - (\sigma_{11}-\Sigma_{1\mathcal{S}}\Sigma_{\mathcal{S}}^{-1}\Sigma_{\mathcal{S}1})\right| \to_p 0$. Accordingly, we have

$$\begin{aligned}
\sqrt{n}T_1 &= \big\{1+o_p(1)\big\}\big(\sigma_{11}-\Sigma_{1\mathcal{S}}\Sigma_{\mathcal{S}}^{-1}\Sigma_{\mathcal{S}1}\big)^{-1}\\
&\quad\times\big(n^{-1/2}\mathbb{X}_1^{\top}\mathcal{E} - n^{-1/2}\Sigma_{1\mathcal{S}}\Sigma_{\mathcal{S}}^{-1}\mathbb{X}_{\mathcal{S}}^{\top}\mathcal{E}\big)\\
&= \big\{1+o_p(1)\big\}\big(\sigma_{11}-\Sigma_{1\mathcal{S}}\Sigma_{\mathcal{S}}^{-1}\Sigma_{\mathcal{S}1}\big)^{-1}\\
&\quad\times\Big\{n^{-1/2}\sum_{i=1}^{n}(X_{i1}-\Sigma_{1\mathcal{S}}\Sigma_{\mathcal{S}}^{-1}X_{i\mathcal{S}})\varepsilon_i\Big\}.
\end{aligned}$$

Applying the same arguments as those given above, we also obtain that

$$\begin{aligned}
\sqrt{n}T_2 &= \big\{1+o_p(1)\big\}\big(\sigma_{11}-\Sigma_{1\mathcal{S}}\Sigma_{\mathcal{S}}^{-1}\Sigma_{\mathcal{S}1}\big)^{-1}\\
&\quad\times\Big\{n^{-1/2}\sum_{i=1}^{n}(X_{i1}-\Sigma_{1\mathcal{S}}\Sigma_{\mathcal{S}}^{-1}X_{i\mathcal{S}})X_{i\mathcal{S}^*}\beta_{\mathcal{S}^*}\Big\}.
\end{aligned}$$

Let $\xi_{i1} = (X_{i1}-\Sigma_{1\mathcal{S}}\Sigma_{\mathcal{S}}^{-1}X_{i\mathcal{S}})\varepsilon_i$ and $\delta_{i1} = (X_{i1}-\Sigma_{1\mathcal{S}}\Sigma_{\mathcal{S}}^{-1}X_{i\mathcal{S}})X_{i\mathcal{S}^*}\beta_{\mathcal{S}^*}$. Then, it can be shown that $E(\xi_{i1})=0$, $\mathrm{var}(\xi_{i1}) = \sigma_i^2(\sigma_{11}-\Sigma_{1\mathcal{S}}\Sigma_{\mathcal{S}}^{-1}\Sigma_{\mathcal{S}1})$, and $E(\delta_{i1}) = (\Sigma_{1\mathcal{S}^*}-\Sigma_{1\mathcal{S}}\Sigma_{\mathcal{S}}^{-1}\Sigma_{\mathcal{S}\mathcal{S}^*})\beta_{\mathcal{S}^*}$. Using Conditions (C4)–(C6), we further obtain that $\sqrt{n}|(\Sigma_{1\mathcal{S}^*}-\Sigma_{1\mathcal{S}}\Sigma_{\mathcal{S}}^{-1}\Sigma_{\mathcal{S}\mathcal{S}^*})\beta_{\mathcal{S}^*}| < C_{\max}\times n^{1/2+\varpi}\max_{j\notin\mathcal{S}^+}|\varrho_{1j}(\mathcal{S})| \to 0$. Moreover,

$$\begin{aligned}
\mathrm{var}(\delta_{i1}) &\to E(\delta_{i1}^2) = E\Big\{(X_{i1}-\Sigma_{1\mathcal{S}}\Sigma_{\mathcal{S}}^{-1}X_{i\mathcal{S}})^2\beta_{\mathcal{S}^*}^{\top}X_{i\mathcal{S}^*}X_{i\mathcal{S}^*}^{\top}\beta_{\mathcal{S}^*}\Big\}\\
&= E\Big[(X_{i1}-\Sigma_{1\mathcal{S}}\Sigma_{\mathcal{S}}^{-1}X_{i\mathcal{S}})^2\beta_{\mathcal{S}^*}^{\top}E\Big\{X_{i\mathcal{S}^*}X_{i\mathcal{S}^*}^{\top}|X_{i\mathcal{S}^+}\Big\}\beta_{\mathcal{S}^*}\Big]\\
&\to \Big(\sigma_{11}-\Sigma_{1\mathcal{S}}\Sigma_{\mathcal{S}}^{-1}\Sigma_{\mathcal{S}1}\Big)\beta_{\mathcal{S}^*}^{\top}\Big(\Sigma_{\mathcal{S}^*}-\Sigma_{\mathcal{S}^*\mathcal{S}^+}\Sigma_{\mathcal{S}^+}^{-1}\Sigma_{\mathcal{S}^+\mathcal{S}^*}\Big)\beta_{\mathcal{S}^*}.
\end{aligned}$$

The bivariate Central Limit Theorem, together with the above results, implies that

$$\begin{aligned}
\Big(\sqrt{n}T_1,\sqrt{n}T_2\Big)^{\top} &= \big(1+o_p(1)\big)\big(\sigma_{11}-\Sigma_{1\mathcal{S}}\Sigma_{\mathcal{S}}^{-1}\Sigma_{\mathcal{S}1}\big)^{-1}\\
&\quad\times\Big\{n^{-1/2}\sum_{i=1}^{n}(\xi_{i1},\delta_{i1})^{\top}\Big\}
\end{aligned}$$

is asymptotically bivariate normal with mean zero and diagonal covariance matrix $V = \mathrm{Diag}(V_{ii})$. In addition, $V_{11} = \bar{\sigma}^2(\sigma_{11}-\Sigma_{1\mathcal{S}}\Sigma_{\mathcal{S}}^{-1}\Sigma_{\mathcal{S}1})^{-1}$ and $V_{22} = \big(\sigma_{11}-\Sigma_{1\mathcal{S}}\Sigma_{\mathcal{S}}^{-1}\Sigma_{\mathcal{S}1}\big)^{-1}\beta_{\mathcal{S}^*}^{\top}\big(\Sigma_{\mathcal{S}^*}-\Sigma_{\mathcal{S}^*\mathcal{S}^+}\Sigma_{\mathcal{S}^+}^{-1}\Sigma_{\mathcal{S}^+\mathcal{S}^*}\big)\beta_{\mathcal{S}^*}$. Consequently, $\sqrt{n}(T_1+T_2)$ is asymptotically normal with mean zero and variance $V_{11}+V_{22}$, which completes the proof.

Appendix C. Proof of Proposition 1

As defined in (2.3), $\mathcal{S}_{k_0} = \{j_1,\ldots,j_{k_0}\}$ contains the indices whose associated predictors have the k_0 largest absolute correlations with X_{i1}. For a given $\bar{k}$, $\hat{\mathcal{S}}_{\bar{k}}$ is defined as in (2.5). In addition, the event $\{\hat{\mathcal{S}}_{\bar{k}} \not\supset \mathcal{S}_{k_0}\}$ indicates that there exists at least one index, say $j_{i_1}\in\mathcal{S}_{k_0}$ ($i_1\le k_0$), but $j_{i_1}\notin\hat{\mathcal{S}}_{\bar{k}}$. Then, for any $\bar{k}$ satisfying $k_0 + d_{\max}n^{\nu_2}/2 < \bar{k} < k_0 + d_{\max}n^{\nu_2}$ with $1\le k_0\le C_b n^{\nu_2}$, we have $\{\hat{\mathcal{S}}_{\bar{k}}\not\supset\mathcal{S}_{k_0}\} \subset \{$There exist indices $i_1\le k_0$ and $i_2 > \bar{k}$ that satisfy $|\hat{\rho}_{1j_{i_2}}| > |\hat{\rho}_{1j_{i_1}}|\}$. The reasoning is as follows. When $j_{i_1}\notin\hat{\mathcal{S}}_{\bar{k}}$, it implies that there exists some index, say j_{i_2} with $i_2>\bar{k}$, such that $j_{i_2}\in\hat{\mathcal{S}}_{\bar{k}}$. Otherwise, all indices j_k in $\hat{\mathcal{S}}_{\bar{k}}$ satisfy $k\le\bar{k}_0$, which implies that $\hat{\mathcal{S}}_{\bar{k}} = \{j_1,\ldots,j_{\bar{k}}\}$ contains $\mathcal{S}_{k_0}$ as a subset. This yields a contradiction. As a result, we have $P\big(\hat{\mathcal{S}}_{\bar{k}}\not\supset\mathcal{S}_{k_0}\big) \le P\big($There exist indices $i_1\le k_0$ and $i_2>\bar{k}$ that satisfy $|\hat{\rho}_{1j_{i_2}}| > |\hat{\rho}_{1j_{i_1}}|\big)$. Thus,

$$\begin{aligned}
P\big(\hat{\mathcal{S}}_{\bar{k}}\supset\mathcal{S}_{k_0}\big) &= 1-P\big(\hat{\mathcal{S}}_{\bar{k}}\not\supset\mathcal{S}_{k_0}\big)\\
&\ge 1-P\big(\text{There exist indices } i_1\le k_0 \text{ and } i_2>\bar{k}\\
&\qquad\text{that satisfy } |\hat{\rho}_{1j_{i_2}}|>|\hat{\rho}_{1j_{i_1}}|\big).
\end{aligned}$$

After simple calculation, we obtain that

$$\begin{aligned}
&|\hat{\rho}_{1j_{i_2}}|-|\hat{\rho}_{1j_{i_1}}|\\
&\quad= |\rho_{1j_{i_2}}|-|\rho_{1j_{i_1}}| + (|\hat{\rho}_{1j_{i_2}}|-|\rho_{1j_{i_2}}|) - (|\hat{\rho}_{1j_{i_1}}|-|\rho_{1j_{i_1}}|)\\
&\quad\le |\rho_{1j_{i_2}}|-|\rho_{1j_{i_1}}| + |\hat{\rho}_{1j_{i_1}}-\rho_{1j_{i_1}}| + |\hat{\rho}_{1j_{i_2}}-\rho_{1j_{i_2}}|\\
&\quad\le |\rho_{1j_{i_2}}|-|\rho_{1j_{i_1}}| + 2\max_j|\hat{\rho}_{1j}-\rho_{1j}|.
\end{aligned}$$

This, together with Lemma 1(ii) and the assumption in Proposition 1 that $|\rho_{1j_{i_1}}|-|\rho_{1j_{i_2}}| > d_{\min}n^{-\nu_2}$ for any $i_2-i_1 > d_{\max}n^{\nu_2}/2$, leads to

$$P\big(\text{There exist indices } i_1<k_0 \text{ and } i_2>\bar{k} \text{ that satisfy } |\hat{\rho}_{1j_{i_2}}|>|\hat{\rho}_{1j_{i_1}}|\big)$$

$$\le P\big(\text{There exist indices } i_1 < k_0 \text{ and } i_2 > \bar{k} \text{ that satisfy } |\rho_{1j_{i_2}}| - |\rho_{1j_{i_1}}| + 2\max_j |\hat{\rho}_{1j} - \rho_{1j}| > 0\big)$$
$$\le P\big(\max_j |\hat{\rho}_{1j} - \rho_{1j}| > d_{\min} n^{-\nu_2}/2\big)$$
$$\le pP\big(|\hat{\rho}_{1j} - \rho_{1j}| > d_{\min} n^{-\nu_2}/2\big)$$
$$\le C_1 \exp(-C_2 n^{1-2\nu_2} d_{\min}^2/4 + \log p).$$

By Condition (C3), the negative sign of the first term on the right-hand side of the above equation, $C_2 n^{1-2\nu_2} d_{\min}^2/4$, dominates the second term $\log p$. As a result,

$$P\big(\text{There exist indices } i_1 < k_0 \text{ and } i_2 > \bar{k} \text{ that satisfy } |\hat{\rho}_{1j_{i_2}}| > |\hat{\rho}_{1j_{i_1}}|\big) \to 0,$$

which completes the proof.

Appendix D. Proof of Theorem 2

We mainly apply Lemma 4 to prove Theorem 2. To this end, we need to show the following two results,

$$\frac{1}{p}\sum_{j=1}^{p} I(p_j \le t) - T_{1,n}(t) \to 0 \quad \text{a.s.} \quad \text{and} \tag{D.1}$$

$$\frac{1}{p_0}\sum_{j\in\mathcal{N}_0} I(p_j \le t) - G_{0,n}(t) \to 0 \quad \text{a.s.}, \tag{D.2}$$

as $p \to \infty$, where $G_{0,n}(t) = p_0^{-1}\sum_{j\in\mathcal{N}_0} P(p_j \le t)$. Since the proofs for (D.1) and (D.2) are quite similar, we only verify (D.1). By the law of large numbers, it is enough to show that

$$\operatorname{var}\Big\{\frac{1}{p}\sum_{j=1}^{p} I(p_j \le t)\Big\} = O(p^{-\delta}) \quad \text{for any } \delta > 0. \tag{D.3}$$

It is worth noting that the left-hand side of (D.3) is equivalent to

$$\begin{aligned}
&\operatorname{var}\Big\{\frac{1}{p}\sum_{j=1}^{p} I(|Z_j| \ge z_{1-t/2})\Big\} \\
&= \operatorname{var}\Big\{\frac{1}{p}\sum_{j\in\mathcal{N}_0} I(|Z_j| \ge z_{1-t/2})\Big\} + \operatorname{var}\Big\{\frac{1}{p}\sum_{j\in\mathcal{N}_1} I(|Z_j| \ge z_{1-t/2})\Big\} \\
&\quad + \frac{2}{p^2}\sum_{j_1\in\mathcal{N}_0}\sum_{j_2\in\mathcal{N}_1} \operatorname{cov}\Big\{I(|Z_{j_1}| \ge z_{1-t/2}), I(|Z_{j_2}| \ge z_{1-t/2})\Big\} \\
&:= J_1 + J_2 + 2J_3.
\end{aligned} \tag{D.4}$$

Using the fact that $\operatorname{var}\{I(|Z_j| \ge z_{1-t/2})\} \le E\{I(|Z_j| \ge z_{1-t/2})\} \le 1$ and the assumption that $p_0/p \to 1$, together with the Cauchy–Schwarz inequality, we have that $J_2 \le p^{-2}|\mathcal{N}_1| \sum_{j\in\mathcal{N}_1} \operatorname{var}\{I(|Z_j| \ge z_{1-t/2})\} \le (p - p_0)^2/p^2 \to 0$. In addition, applying the Cauchy–Schwarz inequality, we obtain

$$\begin{aligned}
J_3^2 &\le \operatorname{var}\Big\{\frac{1}{p}\sum_{j\in\mathcal{N}_0} I(|Z_j| \ge z_{1-t/2})\Big\}\operatorname{var}\Big\{\frac{1}{p}\sum_{j\in\mathcal{N}_1} I(|Z_j| \ge z_{1-t/2})\Big\} \\
&\le (p - p_0)^2 p_0^2/p^4 \to 0.
\end{aligned}$$

Accordingly, to prove (D.3), we only need to show that $J_1 = O(p^{-\delta})$ for some $\delta > 0$. It can be seen that

$$\begin{aligned}
J_1 &= \frac{1}{p^2}\sum_{j\in\mathcal{N}_0} \operatorname{var}\Big\{I(|Z_j| \ge z_{1-t/2})\Big\} \\
&\quad + \frac{1}{p^2}\sum_{j_1\ne j_2, j_1 j_2\in\mathcal{N}_0} \operatorname{cov}\Big\{I(|Z_{j_1}| \ge z_{1-t/2}), I(|Z_{j_2}| \ge z_{1-t/2})\Big\} \\
&:= J_{11} + J_{12}.
\end{aligned}$$

Since $J_{11} \le p_0/p^2 \to 0$ as $p \to \infty$, it suffices to show that $J_{12} = O(p^{-\delta})$ for some $\delta > 0$. Note that

$$\operatorname{cov}\Big\{I(|Z_{j_1}| \ge z_{1-t/2}), I(|Z_{j_2}| \ge z_{1-t/2})\Big\} := I_1 + I_2 + I_3 + I_4,$$

where $I_1 = E\{I(Z_{j_1} \ge z_{1-t/2})I(Z_{j_2} \ge z_{1-t/2})\} - E\{I(Z_{j_1} \ge z_{1-t/2})\}E\{I(Z_{j_2} \ge z_{1-t/2})\}$, $I_2 = E\{I(Z_{j_1} \ge z_{1-t/2})I(Z_{j_2} \le -z_{1-t/2})\} - E\{I(Z_{j_1} \ge z_{1-t/2})\}E\{I(Z_{j_2} \le -z_{1-t/2})\}$, $I_3 = E\{I(Z_{j_1} \le -z_{1-t/2})I(Z_{j_2} \ge z_{1-t/2})\} - E\{I(Z_{j_1} \le -z_{1-t/2})\}E\{I(Z_{j_2} \ge z_{1-t/2})\}$ and $I_4 = E\{I(Z_{j_1} \le -z_{1-t/2})I(Z_{j_2} \le -z_{1-t/2})\} - E\{I(Z_{j_1} \le -z_{1-t/2})\}E\{I(Z_{j_2} \le -z_{1-t/2})\}$. Since the proofs for I_1 to I_4 are essentially the same, we only focus on I_1.

Applying the asymptotic expansion of Z_j given in the proof of Theorem 1, we have

$$Z_j = \frac{\sqrt{n}\hat{\beta}_j}{\hat{\sigma}_{\beta_j}} = \frac{\sqrt{n}\beta_j}{\sigma_{\beta_j}} + n^{-1/2}\sum_{i=1}^{n} u_{ij} + o_p(1), \tag{D.5}$$

where $u_{ij} = \sigma_{\beta_j}^{-1}(\delta_{ij} + \xi_{ij})$, $\delta_{ij} = (X_{ij} - \Sigma_{j\mathcal{S}_j}\Sigma_{\mathcal{S}_j}^{-1}X_{i\mathcal{S}_j})X_{i\mathcal{S}_j^*}^{\top}\beta_{\mathcal{S}_j^*}$, $\xi_{ij} = (X_{ij} - \Sigma_{j\mathcal{S}_j}\Sigma_{\mathcal{S}_j}^{-1}X_{i\mathcal{S}_j})\varepsilon_i$, and $\sigma_{\beta_j}^2 = (\sigma_{jj} - \Sigma_{j\mathcal{S}_j}\Sigma_{\mathcal{S}_j}^{-1}\Sigma_{\mathcal{S}_j j})\{\beta_{\mathcal{S}_j^*}^{\top}(\Sigma_{\mathcal{S}_j^*} - \Sigma_{\mathcal{S}_j^*\mathcal{S}_j^+}\Sigma_{\mathcal{S}_j^+}^{-1}\Sigma_{\mathcal{S}_j^+\mathcal{S}_j^*})\beta_{\mathcal{S}_j^*} + \sigma^2\}$. As a result, for any $j \in \mathcal{N}_0$, $Z_j = n^{-1/2}\sum_{i=1}^{n} u_{ij} + o_p(1)$ and Z_j can be expressed as a summation of independent and identically distributed (i.i.d.) random variables u_{ij}. In addition, Condition (C1) implies that u_{ij} has an exponential tail. This, together with the bivariate large deviation result (Zhong et al., 2013), leads to

$$\begin{aligned}
&E\Big\{I(Z_{j_1} \ge z_{1-t/2})I(Z_{j_2} \ge z_{1-t/2})\Big\} \\
&\quad = U\big(z_{1-t/2}, z_{1-t/2}; \rho_{j_1 j_2}\big)\big\{1 + o(1)\big\},
\end{aligned}$$

where $\rho_{j_1 j_2} = \operatorname{corr}(u_{ij_1}, u_{ij_2})$ and

$$\begin{aligned}
U(a, b; \rho) &= \Big\{2\pi(1 - \rho^2)^{1/2}\Big\}^{-1} \\
&\quad \times \int_a^{\infty}\int_b^{\infty} \exp\Big\{-\frac{1}{2(1 - \rho^2)}(y_1^2 + y_2^2 - 2\rho y_1 y_2)\Big\} dy_1 dy_2.
\end{aligned}$$

Without loss of generality, we assume that $\rho_{j_1 j_2} = \operatorname{corr}(u_{ij_1}, u_{ij_2}) > 0$ and $z_{1-t/2} > 0$. Then, by using the inequality in Willink (2004), we have

$$\begin{aligned}
\Phi(z_{1-t/2})\Phi(\zeta z_{1-t/2}) &\le U\big(z_{1-t/2}, z_{1-t/2}; \rho_{j_1 j_2}\big) \\
&\le \Phi(z_{1-t/2})\Phi\big(\zeta z_{1-t/2}\big)\big(1 + \rho_{j_1 j_2}\big),
\end{aligned} \tag{D.6}$$

where $\zeta = \{(1 - \rho_{j_1 j_2})/(1 + \rho_{j_1 j_2})\}^{1/2}$. Accordingly, we obtain that

$$\begin{aligned}
&\Phi(z_{1-t/2})\Big\{\Phi(\zeta z_{1-t/2}) - \Phi(z_{1-t/2})\Big\} \le I_1 \\
&\le \Phi(z_{1-t/2})\Big\{\Phi(\zeta z_{1-t/2})(1 + \rho_{j_1 j_2}) - \Phi(z_{1-t/2})\Big\}.
\end{aligned}$$

After algebraic simplification, $(1 - \zeta)/\rho_{j_1 j_2} \to 1$ as $\rho_{j_1 j_2} \to 0$. Hence, $I_1/(C_I \rho_{j_1 j_2}) \to 1$ for a positive constant C_I, which implies that $\operatorname{cov}\{I(|Z_{j_1}| \ge z_{1-t/2}), I(|Z_{j_2}| \ge z_{1-t/2})\} \approx C_I|\rho_{j_1 j_2}|$. Consequently, if $\sum_{j\in\mathcal{N}_0}|\rho_{jk}| = o(p)$ for any $k \in \mathcal{N}_0$, then $J_{12} = O(p^{-\delta})$ for some $\delta > 0$.

To complete the proof, we next verify the above condition $\sum_{j\in\mathcal{N}_0}|\rho_{jk}| = o(p)$ for any $k \in \mathcal{N}_0$. By the Cauchy–Schwarz inequality, we need to show that $\sum_{j\in\mathcal{N}_0}\rho_{jk}^2 = o(p)$. Since $\Sigma_{j\mathcal{S}_j}\Sigma_{\mathcal{S}_j}^{-1}\Sigma_{\mathcal{S}_j j}$ is the conditional variance of X_j given $X_{\mathcal{S}_j}$, $\Sigma_{j\mathcal{S}_j}\Sigma_{\mathcal{S}_j}^{-1}\Sigma_{\mathcal{S}_j j} \ge 0$. Therefore, $\sigma_{jj} - \Sigma_{j\mathcal{S}_j}\Sigma_{\mathcal{S}_j}^{-1}\Sigma_{\mathcal{S}_j j} \le \sigma_{jj} < c_{\max} < \infty$ uniformly by Condition (C2) and $\beta_{\mathcal{S}_j^*}^{\top}(\Sigma_{\mathcal{S}_j^*} - \Sigma_{\mathcal{S}_j^*\mathcal{S}_j^+}\Sigma_{\mathcal{S}_j^+}^{-1}\Sigma_{\mathcal{S}_j^+\mathcal{S}_j^*})\beta_{\mathcal{S}_j^*}$ is bounded,

respectively, for any $j \in \mathcal{N}_0$. Hence, $\max_j \sigma^2_{\hat{\beta}_j} < \infty$. In addition, (D.5) implies $\mathrm{var}(u_{ij}) = 1$. As a result, we only need to demonstrate that $\sum_{j\in\mathcal{N}_0} v_{jk}^2 = o(p)$, where $v_{jk} = \mathrm{cov}(u_{ij}, u_{ik})$.

It can be shown that $v_{jk} = v_{jk,1} + v_{jk,2}$, where $v_{jk,1} = \mathrm{cov}(\xi_{ij}, \xi_{ik})$, $v_{jk,2} = \mathrm{cov}(\delta_{ij}, \delta_{ik})$, and ξ_{ij} and δ_{ij} are defined after Eq. (D.5). Hence, to complete the proof, it suffices to show the following results:

$$p^{-1}\sum_{j\in\mathcal{N}_0} v_{jk,1}^2 = o(1) \quad \text{and} \quad p^{-1}\sum_{j\in\mathcal{N}_0} v_{jk,2}^2 = o(1). \tag{D.7}$$

We begin with proving the first equation of (D.7). Applying the Cauchy–Schwarz inequality, it can be shown that

$$\begin{aligned} v_{jk,1}^2 &= \sigma^4\Big(\sigma_{jk} - \Sigma_{j\mathcal{S}_j}\Sigma_{\mathcal{S}_j}^{-1}\Sigma_{\mathcal{S}_j k} - \Sigma_{j\mathcal{S}_k}\Sigma_{\mathcal{S}_k}^{-1}\Sigma_{\mathcal{S}_k j} \\ &\quad - \Sigma_{j\mathcal{S}_j}\Sigma_{\mathcal{S}_j}^{-1}\Sigma_{\mathcal{S}_j\mathcal{S}_k}\Sigma_{\mathcal{S}_k}^{-1}\Sigma_{\mathcal{S}_k k}\Big)^2 \\ &\le 3\sigma^4\Big\{(\sigma_{jk} - \Sigma_{j\mathcal{S}_j}\Sigma_{\mathcal{S}_j}^{-1}\Sigma_{\mathcal{S}_j k})^2 + (\Sigma_{j\mathcal{S}_k}\Sigma_{\mathcal{S}_k}^{-1}\Sigma_{\mathcal{S}_k j})^2 \\ &\quad + (\Sigma_{j\mathcal{S}_j}\Sigma_{\mathcal{S}_j}^{-1}\Sigma_{\mathcal{S}_j\mathcal{S}_k}\Sigma_{\mathcal{S}_k}^{-1}\Sigma_{\mathcal{S}_k k})^2\Big\}. \end{aligned}$$

We then study the above three components separately.

By definition, we have $\sigma_{jk} - \Sigma_{j\mathcal{S}_j}\Sigma_{\mathcal{S}_j}^{-1}\Sigma_{\mathcal{S}_j k} = \varrho_{jk}(\mathcal{S}_j) \le \max_{k\notin\mathcal{S}_j}\varrho_{jk}(\mathcal{S}_j) = O(|\mathcal{S}_j|^{-\xi})$ uniformly for any $j = 1, \ldots, p$. This, together with Condition (C5), implies that $\sigma_{jk} - \Sigma_{j\mathcal{S}_j}\Sigma_{\mathcal{S}_j}^{-1}\Sigma_{\mathcal{S}_j k} = O(n^{-1/2})$ uniformly for any j. As a result, we have

$$\begin{aligned} \sum_{j\in\mathcal{N}_0}(\sigma_{jk} - \Sigma_{j\mathcal{S}_j}\Sigma_{\mathcal{S}_j}^{-1}\Sigma_{\mathcal{S}_j k})^2 &= \Big(\sum_{\substack{j\in\mathcal{N}_0,\\ k\notin\mathcal{S}_j}} + \sum_{\substack{j\in\mathcal{N}_0,\\ k\in\mathcal{S}_j}}\Big)(\sigma_{jk} - \Sigma_{j\mathcal{S}_j}\Sigma_{\mathcal{S}_j}^{-1}\Sigma_{\mathcal{S}_j k})^2 \\ &= O(n^{-1/2}p) = o(p), \end{aligned}$$

where the second summation on the right-hand side of the above equation is 0 since, for $k \in \mathcal{S}_j$, $\sigma_{jk} - \Sigma_{j\mathcal{S}_j}\Sigma_{\mathcal{S}_j}^{-1}\Sigma_{\mathcal{S}_j k}$ is one of the component of the vector $\Sigma_{j\mathcal{S}_j} - \Sigma_{j\mathcal{S}_j}\Sigma_{\mathcal{S}_j}^{-1}\Sigma_{\mathcal{S}_j} = 0$. We next consider the second term $(\Sigma_{j\mathcal{S}_k}\Sigma_{\mathcal{S}_k}^{-1}\Sigma_{\mathcal{S}_k j})^2$. Using the fact that the conditional variance of X_{ij} is non-negative and then applying Condition (C2), we have $\Sigma_{j\mathcal{S}_k}\Sigma_{\mathcal{S}_k}^{-1}\Sigma_{\mathcal{S}_k j} \le \sigma_{jj} < c_{\max}$. This, together with the assumption in Theorem 2, Condition (C2), and the fact that $|\mathcal{S}_k| \le \Lambda_0$, leads to

$$\begin{aligned} p^{-1}\sum_{j\in\mathcal{N}_0}(\Sigma_{j\mathcal{S}_k}\Sigma_{\mathcal{S}_k}^{-1}\Sigma_{\mathcal{S}_k j})^2 &\le c_{\max}p^{-1}\sum_{j\in\mathcal{N}_0} tr(\Sigma_{\mathcal{S}_k}^{-1})\|\Sigma_{j\mathcal{S}_k}\|^2 \\ &\le (c_{\max}c_{\max}^*)|\mathcal{S}_k|p^{-1}\sum_{j\in\mathcal{N}_0}\|\Sigma_{j\mathcal{S}_k}\|^2 \\ &= (c_{\max}c_{\max}^*)|\mathcal{S}_k|p^{-1}\sum_{j\in\mathcal{N}_0}\sum_{l\in\mathcal{S}_k}\sigma_{jl}^2 = o(1). \end{aligned}$$

Employing similar techniques, we can show that $p^{-1}\sum_j(\Sigma_{j\mathcal{S}_j}\Sigma_{\mathcal{S}_j}^{-1}\Sigma_{\mathcal{S}_j\mathcal{S}_k}\Sigma_{\mathcal{S}_k}^{-1}\Sigma_{\mathcal{S}_k k})^2 = o(1)$. The above results complete the proof of the first equation in (D.7).

Subsequently, we will verify the second equation of (D.7). According to the result in the proof of Theorem 1, $E(\delta_{ij}) \le C_\delta \max_{l\in\mathcal{S}_j^+}|\rho_{jl}(\mathcal{S}_j)| = o(1)$ for some positive constant C_δ. It follows that $p^{-1}\sum_{j\in\mathcal{N}_0}E^2(\delta_{ij})E^2(\delta_{ik}) = o(1)$; hence, we only need to show that $p^{-1}\sum_{j\in\mathcal{N}_0}E^2(\delta_{ij}\delta_{ik}) = o(1)$. After algebraic simplification, we obtain that

$$\begin{aligned} E(\delta_{ij}\delta_{ik}) &= E\big\{(X_{ij} - \Sigma_{j\mathcal{S}_j}\Sigma_{\mathcal{S}_j}^{-1}X_{i\mathcal{S}_j})X_{i\mathcal{S}_j^*}^\top\beta_{\mathcal{S}_j^*}X_{ik}\big\} \\ &\quad - E\big\{(X_{ij} - \Sigma_{j\mathcal{S}_j}\Sigma_{\mathcal{S}_j}^{-1}X_{i\mathcal{S}_j})X_{i\mathcal{S}_j^*}^\top\beta_{\mathcal{S}_j^*}\Sigma_{k\mathcal{S}_k}\Sigma_{\mathcal{S}_k}^{-1}X_{i\mathcal{S}_k}X_{i\mathcal{S}_k^*}^\top\beta_{\mathcal{S}_k^*}\big\} \\ &:= Q_{1j} - Q_{2j}. \end{aligned}$$

For the sake of simplicity, we suppress the subscript i in the rest of the proof.

We first demonstrate $Q_{1j} = o(1)$ for each $j \in \mathcal{N}_0$. By Lemma 2 with some tedious calculations, we obtain that

$$\begin{aligned} Q_{1j} &= \sum_{j_2\in\mathcal{S}_j^*\cap\mathcal{N}_1}\sum_{j_3\in\mathcal{S}_k^*\cap\mathcal{N}_1}\beta_{j_2}\beta_{j_3}\sigma_{j_2j_3}\varrho_{jk}(\mathcal{S}_j) \\ &\quad + \sum_{j_2\in\mathcal{S}_j^*\cap\mathcal{N}_1}\sum_{j_3\in\mathcal{S}_k^*\cap\mathcal{N}_1}\beta_{j_2}\beta_{j_3}\sigma_{kj_3}\varrho_{jj_2}(\mathcal{S}_j) \\ &\quad + \sum_{j_2\in\mathcal{S}_j^*\cap\mathcal{N}_1}\sum_{j_3\in\mathcal{S}_k^*\cap\mathcal{N}_1}\beta_{j_2}\beta_{j_3}\sigma_{j_2k}\varrho_{jj_3}(\mathcal{S}_j) := Q_{1j}^{(1)} + Q_{1j}^{(2)} + Q_{1j}^{(3)}. \end{aligned}$$

By Condition (C2), we have $|\sigma_{j_2j_3}| \le c_{\max}$. As a result.

$$\begin{aligned} Q_{1j}^{(1)} &= \sum_{j_2\in\mathcal{S}_j^*\cap\mathcal{N}_1}\sum_{j_3\in\mathcal{S}_k^*\cap\mathcal{N}_1}\beta_{j_2}\beta_{j_3}\sigma_{j_2j_3}\varrho_{jk}(\mathcal{S}_j) \\ &\le c_{\max}\max_{k\notin\mathcal{S}_j}|\varrho_{jk}(\mathcal{S}_j)|\sum_{j_2\in\mathcal{S}_j^*\cap\mathcal{N}_1}\sum_{j_3\in\mathcal{S}_k^*\cap\mathcal{N}_1}|\beta_{j_2}||\beta_{j_3}| \\ &\le c_{\max}\max_{k\notin\mathcal{S}_j}|\varrho_{jk}(\mathcal{S}_j)|\Big(\sum_j|\beta_j|\Big)^2. \end{aligned}$$

Then employing Condition (C6), we obtain $\sum_j|\beta_j| = O(n^{\varpi})$. In addition, Conditions (C4) and (C5) imply that $\varrho_{jk}(\mathcal{S}_j) = O(n^{-1/2})$. The above results lead to

$$Q_{1j}^{(1)} = O(n^{2\varpi}) \times O(n^{-1/2}) = O(n^{1/2-2\varpi}) = o(1)$$

uniformly for any j. Applying similar techniques, we can also show that $Q_{2j}^{(1)} = o(1)$ and $Q_{3j}^{(2)} = o(1)$, which complete the proof of $Q_{1j} = o(1)$.

We next verify $Q_{2j} = o(1)$ for each $j \in \mathcal{N}_0$. After algebraic calculation, we obtain that

$$\begin{aligned} Q_{2j} &= \sum_{j_2\in\mathcal{S}_j^*\cap\mathcal{N}_1}\sum_{j_4\in\mathcal{S}_k^*\cap\mathcal{N}_1}\beta_{j_2}\beta_{j_4}\Sigma_{k\mathcal{S}_k}\Sigma_{\mathcal{S}_k}^{-1}\Sigma_{\mathcal{S}_k j_4}\varrho_{jj_2}(\mathcal{S}_j) \\ &\quad + \sum_{j_2\in\mathcal{S}_j^*\cap\mathcal{N}_1}\sum_{j_4\in\mathcal{S}_k^*\cap\mathcal{N}_1}\beta_{j_2}\beta_{j_4}\Sigma_{k\mathcal{S}_k}\Sigma_{\mathcal{S}_k}^{-1}\Sigma_{\mathcal{S}_k j_2}\varrho_{jj_4}(\mathcal{S}_j) \\ &\quad + \sum_{j_2\in\mathcal{S}_j^*\cap\mathcal{N}_1}\sum_{j_3\in\mathcal{S}_k\cap\mathcal{N}_1}\sum_{j_4\in\mathcal{S}_k^*\cap\mathcal{N}_1}\beta_{j_2}\beta_{j_4}\sigma_{j_2j_4}(\Sigma_{k\mathcal{S}_k}\Sigma_{\mathcal{S}_k}^{-1})_{j_3}\varrho_{jj_3}(\mathcal{S}_j) \\ &:= Q_{2j}^{(1)} + Q_{2j}^{(2)} + Q_{2j}^{(3)}, \end{aligned}$$

where $(\Sigma_{k\mathcal{S}_k}\Sigma_{\mathcal{S}_k}^{-1})_{j_3}$ represents for the j_3th elements of $\Sigma_{k\mathcal{S}_k}\Sigma_{\mathcal{S}_k}^{-1}$. By Conditions (C2), (C4) and (C5), we have that $|\Sigma_{k\mathcal{S}_k}\Sigma_{\mathcal{S}_k}^{-1}\Sigma_{\mathcal{S}_k j_4}| = |\sigma_{kj_4} - \varrho_{kj_4}(\mathcal{S}_k)| \le |\sigma_{kj_4}| + \max_{j_4\notin\mathcal{S}_k}\varrho_{kj_4}(\mathcal{S}_k) \le c_{\max} + O(n^{-1/2})$ and $\varrho_{jj_2}(\mathcal{S}_j) \le \max_{j_2\notin\mathcal{S}_j}\varrho_{jj_2}(\mathcal{S}_j) = O(n^{-1/2})$. These results, in conjunction with Condition (C6), yield

$$Q_{2j}^{(1)} \le O(n^{-1/2})\left(\sum_j|\beta_j|\right)^2 = o(1).$$

Employing similar techniques, we can also demonstrate that $Q_{2j}^{(2)} = o(1)$ and $Q_{3j}^{(3)} = o(1)$, which lead to $Q_{2j} = o(1)$. This, together with $Q_{1j} = o(1)$, implies that

$$p^{-1}\sum_{j\in\mathcal{N}_0}E^2(\delta_{ij}\delta_{ik}) = p^{-1}\sum_{j\in\mathcal{N}_0}(Q_{1j} - Q_{2j})^2 = o(1),$$

which completes the proof of (D.1).

It is worth noting that $G_{0,n}(t) \to t$. This, in conjunction with (D.1), (D.2), and the assumptions $T_{1,n} \to T_1(t)$ with $T_1(t)$ continuous and $p_0/p \to 1$ as $p \to \infty$, indicates that Conditions (a), (b), and (c) in Lemma 4 hold. Accordingly, the proof of Theorem 2 is complete.

Appendix E. Proof of Theorem 3

Let $Z_j = n^{1/2}\hat{\beta}_j/\hat{\sigma}_{\beta_j}$ be the test statistic and p_j be the corresponding p-value for $j = 1, \ldots, p$. Define $\alpha_n = 2\{1 - \Phi(n^{\jmath})\}$ for some $\hbar < \jmath < 1/2 - \kappa$, where $\hbar$ and κ are given in Condition (C7); hence, $\alpha_n \to 0$ as $n \to \infty$. To prove the theorem, it suffices to show that

$$\lim_{n\to\infty} P\Big\{V(\alpha_n) > 0\Big\} \to 0 \quad \text{and}$$
$$\lim_{n\to\infty} P\Big\{S(\alpha_n)/(p - p_0) = 1\Big\} \to 1.$$

It is worth noting that $n \to \infty$ implicitly implies $p \to \infty$. We demonstrate the above equations in the following two steps accordingly.

STEP I. We show that $P\{V(\alpha_n) > 0\} \to 0$. Using the fact that $\tau_{\beta_j}^2 \geq \bar{\sigma}^2$ for $1 \leq j \leq p$, we have

$$\begin{aligned}
|Z_j| &= \Big\{\Big|(X_j^\top \mathcal{Q}_{\mathcal{S}_j} X_j)^{-1} X_j^\top \mathcal{Q}_{\mathcal{S}_j} \mathcal{E} \\
&\quad + (X_j^\top \mathcal{Q}_{\mathcal{S}_j} X_j)^{-1} X_j^\top \mathcal{Q}_{\mathcal{S}_j} X_{\mathcal{S}_j^*} \beta_{\mathcal{S}_j^*}\Big|\Big\} \Big/ (n^{1/2}\hat{\sigma}_{\beta_j}) \\
&\leq \bar{\sigma}^{-1}\Big|(X_j^\top \mathcal{Q}_{\mathcal{S}_j} X_j)^{-1/2} X_j^\top \mathcal{Q}_{\mathcal{S}_j} \mathcal{E}\Big| \\
&\quad + \bar{\sigma}^{-1}\Big|(X_j^\top \mathcal{Q}_{\mathcal{S}_j} X_j)^{-1} X_j^\top \mathcal{Q}_{\mathcal{S}_j} X_{\mathcal{S}_j^*} \beta_{\mathcal{S}_j^*}\Big|.
\end{aligned}$$

This, together with Bonferroni's inequality, leads to

$$\begin{aligned}
P\Big\{V(\alpha_n) > 0\Big\} &= P\Big(\max_{j\in\mathcal{N}_0} |Z_j| > z_{1-\alpha_n/2}\Big) \\
&\leq P\Big(\max_{j\in\mathcal{N}_0} \big|(X_j^\top \mathcal{Q}_{\mathcal{S}_j} X_j)^{-1/2} X_j^\top \mathcal{Q}_{\mathcal{S}_j} \mathcal{E}/\bar{\sigma}\big| > n^{\jmath}/2\Big) \\
&\quad + P\Big(\max_{j\in\mathcal{N}_0} \big|(X_j^\top \mathcal{Q}_{\mathcal{S}_j} X_j)^{-1} X_j^\top \mathcal{Q}_{\mathcal{S}_j} X_{\mathcal{S}_j^*} \beta_{\mathcal{S}_j^*}\big| > \bar{\sigma} n^{\jmath}/2\Big). \qquad \text{(E.1)}
\end{aligned}$$

Consider the quantity $\big|(X_j^\top \mathcal{Q}_{\mathcal{S}_j} X_j)^{-1/2} X_j^\top \mathcal{Q}_{\mathcal{S}_j} \mathcal{E}/\bar{\sigma}\big|$, which is in the first term of the right-hand side of the above equation. Employing the same technique as used in the proof of Lemma 3, we obtain that $\max_j |n^{-1} X_j^\top \mathcal{Q}_{\mathcal{S}_j} X_j - (\sigma_{jj} - \Sigma_{j\mathcal{S}_j} \Sigma_{\mathcal{S}_j}^{-1} \Sigma_{\mathcal{S}_j j})| = o_p(1)$. By Condition (C2), one can easily verify that $\sigma_{jj} - \Sigma_{j\mathcal{S}_j} \Sigma_{\mathcal{S}_j}^{-1} \Sigma_{\mathcal{S}_j j} \geq c_{\max}^{*-1}$. The above two results lead to $n^{-1} X_j^\top \mathcal{Q}_{\mathcal{S}_j} X_j \geq c_{\max}^{*-1}$ uniformly for any j. Accordingly, there exists some constant C_3 such that

$$\max_{j\in\mathcal{N}_0} \big|(X_j^\top \mathcal{Q}_{\mathcal{S}_j} X_j)^{-1/2} X_j^\top \mathcal{Q}_{\mathcal{S}_j} \mathcal{E}/\bar{\sigma}\big| \leq C_3 \max_{j\in\mathcal{N}_0} n^{-1/2} |X_j^\top \mathcal{E}|.$$

This, in conjunction with Bonferroni's inequality and Condition (C8), yields

$$\begin{aligned}
&P\Big(\max_{j\in\mathcal{N}_0} \big|(X_j^\top \mathcal{Q}_{\mathcal{S}_j} X_j)^{-1/2} X_j^\top \mathcal{Q}_{\mathcal{S}_j} \mathcal{E}/\bar{\sigma}\big| > n^{\jmath}/2\Big) \\
&\leq \sum_{j\in\mathcal{N}_0} P\Big(n^{-1}\big|X_j^\top \mathcal{E}\big| > C_3^{-1} n^{\jmath - 1/2}/2\Big) \\
&\leq 2p \exp\big\{-C_4 n^{2\jmath}\big\} \leq 2\exp\big\{-C_4 n^{2\jmath} + \nu n^{\hbar}\big\}
\end{aligned}$$

for some positive constant C_4. By definition, $\hbar < 2\jmath$. Thus, the first term on the right-hand side of the above equation, $-C_\varepsilon n^{2\jmath}/4$, dominates the second term $\nu n^{\hbar}$, which immediately leads to

$$\lim_{n\to\infty} P\Big(\max_{j\in\mathcal{N}_0} \big|(X_j^\top \mathcal{Q}_{\mathcal{S}_j} X_j)^{-1/2} X_j^\top \mathcal{Q}_{\mathcal{S}_j} \mathcal{E}/\bar{\sigma}\big| > n^{\jmath}/2\Big) \to 0.$$

We next consider the quantity $(X_j^\top \mathcal{Q}_{\mathcal{S}_j} X_j)^{-1/2} X_j^\top \mathcal{Q}_{\mathcal{S}_j} X_{\mathcal{S}_j^*} \beta_{\mathcal{S}_j^*}$, which is in the second term of the right-hand side of Eq. (E.1). It is worth noting that

$$\begin{aligned}
&(X_j^\top \mathcal{Q}_{\mathcal{S}_j} X_j)^{-1/2} X_j^\top \mathcal{Q}_{\mathcal{S}_j} X_{\mathcal{S}_j^*} \beta_{\mathcal{S}_j^*} \\
&= (X_j^\top \mathcal{Q}_{\mathcal{S}_j} X_j/n)^{-1/2} n^{1/2} \sum_{j^*\in\mathcal{S}_j^*} \hat{\varrho}_{jj^*}(\mathcal{S}_j)\beta_{j^*} \\
&\leq C_5 \Big\{\min_j n^{-1} X_j^\top \mathcal{Q}_{\mathcal{S}_j} X_j\Big\}^{-1/2} \max_{j^*\in\mathcal{S}_j^*} |n^{1/2}\hat{\varrho}_{jj^*}(\mathcal{S}_j)|
\end{aligned}$$

for some finite positive constant C_5.

Using the results of $\max_j |n^{-1} X_j^\top \mathcal{Q}_{\mathcal{S}_j} X_j - (\sigma_{jj} - \Sigma_{j\mathcal{S}_j} \Sigma_{\mathcal{S}_j}^{-1} \Sigma_{\mathcal{S}_j j})| = o_p(1)$ and $\sigma_{jj} - \Sigma_{j\mathcal{S}_j} \Sigma_{\mathcal{S}_j}^{-1} \Sigma_{\mathcal{S}_j j} \geq c_{\max}^{*-1}$ discussed after (E.1), we have that $\{\min_j n^{-1} X_j^\top \mathcal{Q}_{\mathcal{S}_j} X_j\}^{-1/2} = O_p(1)$. In addition, Condition (C4), together with the fact that $\mathcal{S}_j$ satisfies Condition (C5), leads to $\max_{j^*\in\mathcal{S}_j^*} |n^{1/2}\varrho_{jj^*}(\mathcal{S}_j)| = o(1)$. By Corollary 1 of Kalisch and Bühlmann (2007), we immediately obtain

$$\max_{j,\mathcal{S}_j^*} P\Big\{\max_{j^*\in\mathcal{S}_j^*} |n^{1/2}\hat{\varrho}_{jj^*}(\mathcal{S}_j)| > O(n^{b/2})\Big\} \to 0$$

for every $\hbar < b < 1$. Taking $b = (\hbar + \jmath)/2$, we then have $\max_{j^*\in\mathcal{S}_j^*} |n^{1/2}\hat{\varrho}_{jj^*}(\mathcal{S}_j)| = o(n^{b/2}) = o(n^{\jmath})$. This, in conjunction with the above result $\{\min_j n^{-1} X_j^\top \mathcal{Q}_{\mathcal{S}_j} X_j\}^{-1/2} = O_p(1)$, results in

$$P\Big(\max_{j\in\mathcal{N}_0} \big|(X_j^\top \mathcal{Q}_{\mathcal{S}_j} X_j)^{-1} X_j^\top \mathcal{Q}_{\mathcal{S}_j} X_{\mathcal{S}_j^*} \beta_{\mathcal{S}_j^*}\big| > \sigma n^{\jmath}/2\Big) \to 0.$$

In sum, we have shown the asymptotic behavior of the first component on the right-hand side of (E.1).

STEP II. We prove that $\lim_{n\to\infty} P\Big\{S(\alpha_n)/(p - p_0) = 1\Big\} \to 1$. By definition, we have

$$(p - p_0)^{-1} S(\alpha_n) = (p - p_0)^{-1} \sum_{j\in\mathcal{N}_1} I\Big(|n^{1/2}\hat{\beta}_j/\hat{\sigma}_{\beta_j}| > n^{\jmath}\Big).$$

Applying the asymptotic result of the first component on the right-hand side of (E.1), we have $\max_j |n^{1/2}(\hat{\beta}_j - \beta_j)/\hat{\sigma}_{\beta_j}| = o(n^{\jmath})$. Then by Condition (C7) that $\min_{j\in\mathcal{N}_1} |\beta_j| \geq C_\kappa n^{-\kappa}$ for some constants $C_\kappa > 0$ and $\kappa > 0$, we can further obtain that $\min_{j\in\mathcal{N}_1} |n^{1/2}\beta_j/\hat{\sigma}_{\beta_j}| = O(n^{1/2-\kappa})$. Moreover, by Bonferroni's inequality and the fact that $\jmath + \kappa < 1/2$, we have

$$\begin{aligned}
P\Big((p - p_0)^{-1} S(\alpha_n) = 1\Big) &= P\Big(\min_{j\in\mathcal{N}_1} |n^{1/2}\hat{\beta}_j/\hat{\sigma}_{\beta_j}| > n^{\jmath}\Big) \\
&\geq P\Big(\min_{j\in\mathcal{N}_1} |n^{1/2}\beta_j/\hat{\sigma}_{\beta_j}| > n^{\jmath}\Big) \\
&\quad - P\Big(\max_{j\in\mathcal{N}_1} |n^{1/2}(\hat{\beta}_j - \beta_j)/\hat{\sigma}_{\beta_j}| > 2n^{\jmath}\Big) \to 1,
\end{aligned}$$

which completes the proof of Step II. Consequently, the entire proof is complete.

Appendix F. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.jeconom.2016.05.016.

References

Belloni, A., Chen, D., Chernozhukov, V., Hansen, C., 2012. Sparse models and methods for optimal instruments with an application to eminent domain. Econometrica 80, 2369–2429.

Belloni, A., Chernozhukov, V., Hansen, C., 2014. Inference on treatment effects after selection amongst high-dimensional controls. Rev. Econom. Stud. 81, 608–650.

Bendat, J.S., Piersol, A.G., 1966. Measurement and Analysis of Random Data. Wiley, New York.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B Stat. Methodol. 57, 289–300.

Bickel, P., Levina, E., 2008. Regularized estimation of large covariances matrix. Ann. Statist. 36, 199–227.
Bühlmann, P., 2013. Statistical significance in high-dimensional linear models. Bernoulli 19, 1212–1242.
Bunea, F., Wegkamp, M., Auguste, A., 2006. Consistent variable selection in high dimensional regression via multiple testing. J. Statist. Plann. Inference 136, 4349–4364.
Cho, H., Fryzlewicz, P., 2012. High dimensional variable selection via tilting. J. R. Stat. Soc. Ser. B Stat. Methodol. 74, 593–622.
Cook, R.D., Weisberg, S., 1998. Residuals and Influence in Regression. Chapman and Hall, New York.
Draper, N.R., Smith, H., 1998. Applied Regression Analysis, third ed. Wiley, New York.
Fama, E.F., French, K.R., 1993. Common risk factors in the return on stocks and bonds. J. Financ. Econ. 33, 3–56.
Fan, J., Han, X., Gu, W., 2012. Estimating false discovery proportion under arbitrary covariance dependence. J. Amer. Statist. Assoc. 107, 1019–1035.
Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Amer. Statist. Assoc. 96, 1348–1360.
Fan, J., Lv, J., 2008. Sure independence screening for ultra-high dimensional feature space (with discussion). J. R. Stat. Soc. Ser. B Stat. Methodol. 70, 849–911.
Fan, J., Lv, J., Qi, L., 2011. Sparse high-dimensional models in economics. The Annual Review of Economics 3, 291–317.
Goeman, J., Geer, V.D., Houwelingen, V., 2006. Testing against a high-dimensional alternative. J. R. Stat. Soc. Ser. B Stat. Methodol. 68, 477–493.
Goeman, J., Houwelingen, V., Finos, L., 2011. Testing against a high dimensional alternative in the generalized linear model: asymptotic type I error control. Biometrika 98, 381–390.
Huang, J., Ma, S., Zhang, C.H., 2007. Adaptive Lasso for sparse high-dimensional regression models. Statist. Sinica 18, 1603–1618.
Kalisch, M., Bühlmann, P., 2007. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. J. Mach. Learn. Res. 8, 613–636.
Li, R., Zhong, W., Zhu, L.P., 2012. Feature screening via distance correlation learning. J. Amer. Statist. Assoc. 107, 1129–1139.
Liu, W.D., 2013. Gaussian graphical model estimation with false discovery rate control. Ann. Statist. 41, 2948–2978.
Meinshausen, N., Bühlmann, P., 2006. High-dimensional graphs and variable selection with the Lasso. Ann. Statist. 34, 1436–1462.
Meinshausen, N., Meier, L., Bühlmann, P., 2009. P-values for high-dimensional regression. J. Amer. Statist. Assoc. 104, 1671–1681.
Storey, J.D., 2002. A direct approach to false discovery rates. J. R. Stat. Soc. Ser. B Stat. Methodol. 64, 479–498.
Storey, J.D., Taylor, J.E., Siegmund, D., 2004. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. J. R. Stat. Soc. Ser. B Stat. Methodol. 66, 187–205.
Sun, T., Zhang, C.H., 2012. Scaled sparse linear regression. Biometrika 99, 879–898.
Tibshirani, R.J., 1996. Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. Ser. B Stat. Methodol. 58, 267–288.
van de Geer, S., Bühlmann, P., Ritov, Y., Dezeure, R., 2014. On asymptotically optimal confidence regions and tests for high-dimensional models. Ann. Statist. 42, 1166–1202.
Wang, H., 2009. Forward regression for ultra-high dimensional variable screening. J. Amer. Statist. Assoc. 104, 1512–1524.
Wang, H., 2012. Factor profiled independence screening. Biometrika 99, 15–28.
Willink, R., 2004. Bounds on the bivariate normal distribution function. Commun. Stat. - Theory Methods 33, 2281–2297.
Wooldridge, J., 2002. Econometric Analysis of Cross Section and Panel Data. MIT Press, USA.
Zhang, C.H., Zhang, S.S., 2014. Confidence intervals for low dimensional parameters in high dimensional linear models. J. R. Stat. Soc. Ser. B Stat. Methodol. 76, 217–242.
Zhao, P., Yu, B., 2006. On model selection consistency of Lasso. J. Mach. Learn. Res. 7, 2541–2563.
Zhong, P.S., Chen, S.X., 2011. Tests for high dimensional regression coefficients with factorial designs. J. Amer. Statist. Assoc. 106, 260–274.
Zhong, P.S., Chen, S.X., Xu, M., 2013. Tests alternative to higher criticism for high dimensional means under sparsity and column-wise dependence. Ann. Statist. 41, 2820–2851.