Statistica Sinica **32** (2022), 1381-1409
doi:https://doi.org/10.5705/ss.202020.0267

CONDITIONAL TEST FOR ULTRAHIGH DIMENSIONAL LINEAR REGRESSION COEFFICIENTS

Wenwen Guo[1], Wei Zhong[2], Sunpeng Duan[3] and Hengjian Cui[1]

[1]*Capital Normal University,* [2]*Xiamen University and* [3]*University of California at Santa Barbara*

Abstract: This paper presents a conditional test for the overall significance of the regression coefficients in ultrahigh-dimensional linear models, conditional on a subset of predictors. We first propose a conditional U-statistic test (CUT) based on an estimated U-statistic for a moderately high-dimensional linear regression model, and derive its asymptotic distributions under some mild assumptions. However, the empirical power of the CUT is inversely affected by the dimensionality of the predictors. To this end, we further propose a two-stage CUT with screening (CUTS) procedure based on a random data-splitting strategy to enhance the empirical power. In the first stage, we divide the data randomly into two parts and apply conditional sure independence screening to the first part to reduce the dimensionality. In the second stage, we apply the CUT to the reduced model using the second part of the data. To eliminate the effect of data-splitting randomness and to further enhance the empirical power, we also develop a powerful ensemble CUTS_M algorithm based on multiple data-splitting. We then prove that the family-wise error rate is asymptotically controlled at a given significance level. We demonstrate the excellent finite-sample performance of the proposed conditional tests usig Monte Carlo simulations and two real-data analysis examples.

Key words and phrases: Hypothesis testing, linear regression coefficients, random data splitting, ultrahigh dimensionality, variable screening.

1. Introduction

Linear regression is commonly used to explore the relationship between a response and many predictors for ultrahigh-dimensional data, where the predictor dimension p is much larger than the sample size n. On the one hand, existing studies or researchers' beliefs may provide prior information that some subset of predictors is important for the response. On the other hand, feature screening approaches and regularization methods may identify significant predictors for the response. A natural question is whether given the subset of identified predictors, the remaining ultrahigh-dimensional variables are still able to contribute to the

Corresponding author: Wei Zhong, Xiamen University, Xiamen 361005, China. Email: wzhong@ xmu. edu.cn.

response. If the answer is no, it is adequate to consider the linear model based only on the subset of identified predictors. For example, Scheetz et al. (2006) analyzed gene expression microarray data on 120 12-week-old male rats to gain a broad perspective of gene regulation in the mammalian eye. As a result, they detected 22 gene probes (refer to Table 2 in Scheetz et al. (2006)) relevant to human eye disease from 18,976 different gene probes. We consider a linear regression model of the response gene TRIM32, which has been proven to cause the retinal disease Bardet–Biedl syndrome, against the other 18,975 gene probes. It is interesting to test the overall significance of the regression coefficients of the remaining ultrahigh-dimensional gene probes, conditioning on the subset of 22 identified gene probes. If the null hypothesis is significantly rejected, we need to search for additional important gene probes from the remaining ultrahigh-dimensional candidates. This motivates us to explore a new conditional test procedure for ultrahigh-dimensional linear regression coefficients.

We consider a linear regression model

$$Y_i = \alpha + \mathbf{X}_{0i}^{\mathrm{T}}\boldsymbol{\beta}_0 + \mathbf{X}_{1i}^{\mathrm{T}}\boldsymbol{\beta}_1 + \varepsilon_i, \tag{1.1}$$

where $Y_i \in \mathbb{R}^1$ is the ith response variable, and $\mathbf{X}_i = (\mathbf{X}_{0i}^{\mathrm{T}}, \mathbf{X}_{1i}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^p$ is the associated p-dimensional predictor vector, for $1 \le i \le n$. Based on some prior information, we assume that a subset of the predictors, denoted by $X_{0i} \in \mathbb{R}^q$, are known in the linear model, where $\mathbf{X}_{1i} \in \mathbb{R}^{p-q}$ represents the vector of all remaining covariates for the ith observation. Here, α is a nuisance intercept parameter, $\boldsymbol{\beta}_0 \in \mathbb{R}^q$ and $\boldsymbol{\beta}_1 \in \mathbb{R}^{p-q}$ denote vectors of the regression coefficients corresponding to $\mathbf{X}_{0i}$ and $\mathbf{X}_{1i}$, respectively, and ε_i is a random error with mean zero and finite variance σ^2. We assume that p is much greater than the sample size n, and that q is smaller than n. Our main goal is to test, for a given parameter vector $\boldsymbol{\beta}_{10} \in \mathbb{R}^{p-q}$,

$$H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_{10} \qquad \text{versus} \qquad H_1 : \boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_{10}. \tag{1.2}$$

In particular, rejecting $H_0 : \boldsymbol{\beta}_1 = \mathbf{0}$ indicates an overall significant effect of all remaining predictors on the response variable, conditional on the subset of known predictors.

In the literature, unconditional tests for the overall significance of linear regression coefficients have been well studied. In a classic multivariate analysis, the conventional F-test is generally used when the predictor dimension p is fixed and less than the sample size n. However, the power of the F-test has been shown by Zhong and Chen (2011) to be adversely affected by an increased dimension, even when $p < n - 1$. Wang and Cui (2013) generalized the F-test for moderately high-dimensional linear regression coefficients, but it still fails when $p > n$, owing to the singular sample covariance matrix. Geoman, Van de Geer and Van Houwelingen (2006) proposed an empirical Bayes test for a high-dimensional linear regression. Zhong and Chen (2011) developed a novel test statistic based on a U-statistic of order four, and derived its null asymptotic distribution under the pseudo-independence assumption to accommodate high dimensionality. Moreover, Cui, Guo and Zhong (2018) suggested an estimated U-statistic of order two and enhance the test power using the refitted cross-validation (RCV) approach. Wang and Cui (2015) proposed a test for part of the regression coefficients in high-dimensional linear models based on the idea of Zhong and Chen (2011). However, when the predictor dimension is much larger than n in ultrahigh-dimensional data, the power of the aforementioned significance tests for ultrahigh-dimensional sparse linear models might deteriorate remarkably. Here, sparsity means that only a small subset of predictors are truly important to the response. This motivates us to study how to enhance the power of the conditional significance test under the sparsity assumption.

In this study, we develop a conditional test procedure based on random data splitting to test the overall significance of the remaining ultrahigh-dimensional predictors, given a subset of predictors in the linear model. This study makes the following three main contributions. First, we propose a conditional U-statistic test (CUT) based on an estimated U-statistic for a high-dimensional linear regression model, and show that its asymptotic null distribution is normal, and can be used directly to compute the critical region and the p-value when n is sufficiently large. Second, in order to handle the ultrahigh dimensionality, we propose an efficient two-stage testing procedure based on random data splitting, called the conditional U-statistic test with screening (CUTS), to enhance the testing power under the sparsity condition. Data-splitting techniques have been used for various applications in the literature. Wasserman and Roeder (2009) used a data-splitting strategy to control the family-wise error rate, leading to a powerful variable selection procedure. Fan, Guo and Hao (2012) proposed a consistent refitted cross-validation estimator for the error variance in an ultrahigh-dimensional linear model based on a data-splitting technique. Simulations show that the two-stage testing procedure performs much better for ultrahigh-dimensional sparse linear models. Third, to eliminate the effect of single random data splitting, and to further enhance both the empirical power and the algorithm stability, we also develop a powerful ensemble algorithm, CUTS_M, based on a multiple splitting strategy. Motivated by the idea of Meinshausen, Meier and Buhlmann (2009), we demonstrate that the family-wise error rate of the CUTS_M testing procedure is asymptotically controlled at a given significance level. Note that random data splitting is crucial to eliminate the effect of spurious correlations due to ultrahigh dimensionality, and to avoid an inflation of the type-I error.

This work is also partially related to works in the post-selection inference literature. Lockhart et al. (2014) proposed a covariance test for testing the significance of a variable that enters the active set in the Lasso solution path (Tibshirani (1996)). Lee et al. (2016) developed an approach for constructing valid confidence intervals for the selected coefficients after model selection by the Lasso. Moreover, Zhang and Zhang (2014) constructed confidence intervals for low-dimensional parameters in high-dimensional linear models with homoscedastic variance using the low-dimensional projection and regularization methods. Wang, Zhong and Cui (2018) further proposed empirical likelihood ratio tests for low-dimensional parameters in high-dimensional heteroscedastic linear models. Compared with these existing methods, our proposed CUTS procedure has several different features. First, we focus on testing the overall significance of the remaining ultrahigh-dimensional predictors, conditional on a given subset of predictors, whereas the aforementioned methods tend to form valid confidence intervals for a single coefficient or for low-dimensional ones. Second, it is not necessary for the conditioning set in our CUTS procedure to be the variable subset selected by model selection using, for example, the Lasso. It can be a subset of predictors based on researchers' experience or a brief that are independent of the current data. Third, we consider ultrahigh dimensionality in which spurious correlations play an important and nonignorable role in the significance test.

The remainder of the paper is organized as follows. In Section 2, we develop the new conditional test and study its asymptotic distributions. We introduce the two-stage CUTS procedure in Section 3. Section 4 examines the finite-sample performance of the proposed procedure using Monte Carlo simulations and real-data examples. A brief discussion is given in Section 5. All technical proofs are relegated to the Appendix.

2. A New Conditional Test

2.1. Test statistic

Let $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^\top$, $\mathbf{X}_0 = (\mathbf{X}_{01}, \ldots, \mathbf{X}_{0n})^\top$, $\mathbf{X}_1 = (\mathbf{X}_{11}, \ldots, \mathbf{X}_{1n})^\top$, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^\top$. The linear model (1.1) can be rewritten as

$$\mathbf{Y} = \alpha + \mathbf{X}_0\boldsymbol{\beta}_0 + \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}. \tag{2.1}$$

To motivate the test statistic, we first assume that $\boldsymbol{\beta}_0$ is known and $\alpha = 0$, and that the ordinary least squares estimator for $\boldsymbol{\beta}_1$ is $\widehat{\boldsymbol{\beta}}_1 = (\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1)^{-1}\mathbf{X}_1^{\mathrm{T}}(\mathbf{Y} - \mathbf{X}_0\boldsymbol{\beta}_0)$. Note that $\widehat{\boldsymbol{\beta}}_1$ is infeasible for high-dimensional data where $p - q > n$, because $\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1$ is not invertible. To test $H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_{10}$, we naturally consider the difference between $\widehat{\boldsymbol{\beta}}_1$ and $\boldsymbol{\beta}_{10}$. Because $\widehat{\boldsymbol{\beta}}_1 = \boldsymbol{\beta}_{10}$ implies that $\mathbf{X}_1^{\mathrm{T}}(\mathbf{Y} - \mathbf{X}_0\boldsymbol{\beta}_0 - \mathbf{X}_1\boldsymbol{\beta}_{10}) = 0$, we can use $E\|\mathbf{X}_{1i}(Y_i - \mathbf{X}_{0i}^{\mathrm{T}}\boldsymbol{\beta}_0 - \mathbf{X}_{1i}^{\mathrm{T}}\boldsymbol{\beta}_{10})\|^2$ as an effective measure of the discrepancy between $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_{10}$. Following Zhong and Chen (2011), we first use a U-statistic with $\mathbf{X}_{1i}^{\mathrm{T}}\mathbf{X}_{1j}(Y_i - \mathbf{X}_{0i}^{\mathrm{T}}\boldsymbol{\beta}_0 - \mathbf{X}_{1i}^{\mathrm{T}}\boldsymbol{\beta}_{10})(Y_j - \mathbf{X}_{0j}^{\mathrm{T}}\boldsymbol{\beta}_0 - \mathbf{X}_{1j}^{\mathrm{T}}\boldsymbol{\beta}_{10})$, for $i \neq j$, as the kernel to estimate $E\|\mathbf{X}_{1i}(Y_i - \mathbf{X}_{0i}^{\mathrm{T}}\boldsymbol{\beta}_0 - \mathbf{X}_{1i}^{\mathrm{T}}\boldsymbol{\beta}_{10})\|^2$ when $\alpha = 0$, and the mean of $\mathbf{X}_{1i}$ is $\boldsymbol{\mu}_1 = \mathbf{0}$. Then, we remove the effect of nonzero $\boldsymbol{\mu}_1$ and α by centralizing both $\mathbf{X}_{1i}$ and $Y_i - \mathbf{X}_{0i}^{\mathrm{T}}\boldsymbol{\beta}_0 - \mathbf{X}_{1i}^{\mathrm{T}}\boldsymbol{\beta}_{10}$. Define

$$\Delta_{i,j}(\mathbf{X}_1) = (\mathbf{X}_{1i} - \overline{\mathbf{X}}_1)^{\mathrm{T}}(\mathbf{X}_{1j} - \overline{\mathbf{X}}_1) + \frac{\|\mathbf{X}_{1i} - \mathbf{X}_{1j}\|^2}{2n} \tag{2.2}$$

$$\Delta_{i,j}(\mathbf{Y}^*) = (Y_i^* - \overline{Y}^*)(Y_j^* - \overline{Y}^*) + \frac{|Y_i^* - Y_j^*|^2}{2n}, \tag{2.3}$$

where $\mathbf{Y}^* = \mathbf{Y} - \mathbf{X}_0\widehat{\boldsymbol{\beta}}_0 - \mathbf{X}_1\boldsymbol{\beta}_{10}$ and $\widehat{\boldsymbol{\beta}}_0$ is the ordinary least squares estimator after regressing $Y - \mathbf{X}_1^{\mathrm{T}}\boldsymbol{\beta}_{10}$ against $\mathbf{X}_0$. Note that the second terms in (2.2) and (2.3) are proposed to correct the bias due to centralization, which implies that $E[\Delta_{i,j}(\mathbf{X}_1)] = 0$ and $E[\Delta_{i,j}(\mathbf{Y}^*)] = 0$. Then, we define a new test statistic as

$$T_n = \left(1 - \frac{2}{n}\right)^{-2} \binom{n}{2}^{-1} \sum_{i=2}^{n}\sum_{j=1}^{i-1} \Delta_{i,j}(\mathbf{X}_1)\Delta_{i,j}(\mathbf{Y}^*). \tag{2.4}$$

Because the conditional test statistic (2.4) is based on the estimated U-statistic of order two, we call it the conditional U-statistic test (CUT). This extends the work of Cui, Guo and Zhong (2018) to the conditional testing problem.

2.2. Asymptotic distributions

We let $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}_{00}$, and $\boldsymbol{\Sigma}_{11}$ be the covariance matrices of the covariate vectors $\mathbf{X}_i$, $\mathbf{X}_{0i}$, and $\mathbf{X}_{1i}$, respectively, and let $\boldsymbol{\Sigma}_{01} = \boldsymbol{\Sigma}_{10}^{\mathrm{T}}$ be the covariance matrix of $\mathbf{X}_{0i}$ and $\mathbf{X}_{1i}$. Next, we study the asymptotic null distribution of the test statistic $T_{n,p}$ under some technical assumptions.

(C1) $(p - q) \to \infty$ as $n \to \infty$; $\boldsymbol{\Sigma}_{11} > 0$, $tr(\boldsymbol{\Sigma}_{11}^4) = o\{tr^2(\boldsymbol{\Sigma}_{11}^2)\}$.

(C2) Suppose $\mathbf{X}_i$ follows a p-dimensional elliptical contoured distribution, $\mathbf{X}_i = \boldsymbol{\mu} + \boldsymbol{\Gamma} R_i \mathbf{U}_i$, where $\boldsymbol{\Gamma}$ is a $p \times p$ matrix, $\mathbf{U}_i$ is a random vector uniformly distributed on the unit sphere in $\mathbb{R}^p$, R_i is a nonnegative random variable independent of $\mathbf{U}_i$, and $E(R_i^2) = p, Var(R_i^2) = O(p)$. We also denote $\mathbf{X}_{1i} = \boldsymbol{\mu}_1 + \boldsymbol{\Gamma}_1 R_i U_i$ and $\mathbf{X}_{0i} = \boldsymbol{\mu}_0 + \boldsymbol{\Gamma}_0 R_i U_i$.

(C3) $q = O(n^{\kappa})$, for $0 \leq \kappa < 1/3$, and the eigenvalues of $\boldsymbol{\Sigma}_{00}$ are bounded.

(C4) $tr(\boldsymbol{\Sigma}_{01}\boldsymbol{\Sigma}_{11}\boldsymbol{\Sigma}_{10}) = o(n^{-2\kappa} tr(\boldsymbol{\Sigma}_{11}^2))$.

Condition (C1) assumes the dimensionality of $\mathbf{X}_{1i}$, $p - q$, goes to infinity as the sample size increases to infinity. Thus, it can accommodate (at least moderately) high-dimensional problems. The second part of (C1) assumes the positive definiteness of $\boldsymbol{\Sigma}_{11}$ to ensure the identification of the regression coefficients of $\mathbf{X}_{1i}$. (C1) is similar to Assumption (2.8) in Zhong and Chen (2011). The elliptical countered distribution in (C2) is widely assumed in multivariate statistical analysis, and includes the multivariate normal distribution and multivariate t-distribution as special cases. Condition (C3) requires that the dimension of the known covariates, q, should be small or cannot increase faster than $n^{1/3}$. Condition (C4) is a technical assumption on the dependency between $\mathbf{X}_{0i}$ and $\mathbf{X}_{1i}$. Theorem 1 presents the asymptotic null distribution of the new CUT statistic T_n in (2.4).

Theorem 1. *Assume conditions* (C1)–(C4) *hold. Then, under* H_0 *in* (1.2),

$$\frac{nT_n}{\sigma^2\sqrt{2tr(\boldsymbol{\Sigma}_{11}^2)}} \xrightarrow{D} N(0,1) \tag{2.5}$$

as $n \to \infty$, *where* $\xrightarrow{D}$ *denotes convergence in distribution.*

The asymptotic null distribution of T_n can be used to compute the critical region or empirical p-value when the sample size is relatively large. The null hypothesis $H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_{10}$ is rejected at the significance level α if

$$nT_n \geq \widehat{\sigma}^2\sqrt{2\widehat{tr(\boldsymbol{\Sigma}_{11}^2)}}z_{\alpha}, \tag{2.6}$$

where z_{α} is the α upper-tailed critical value of the standard normal distribution, and $\widehat{\sigma}^2$ and $\widehat{tr(\boldsymbol{\Sigma}_{11}^2)}$ are estimators of σ^2 and $tr(\boldsymbol{\Sigma}_{11}^2)$, respectively. We can also compute the p-value by $P(Z > nt_n/\widehat{\sigma}^2\sqrt{2\widehat{tr(\boldsymbol{\Sigma}_{11}^2)}})$, where t_n is the observed test statistic, and Z is a standard normal random variable. In practice, $\widehat{\sigma}^2$ can be the sample variance of the response, as in Zhong and Chen (2011), or the refitted cross-validation variance estimator of Fan, Guo and Hao (2012) and Cui, Guo and Zhong (2018). In addition, $\widehat{tr(\mathbf{\Sigma}_{11}^2)}$ can be estimated unbiasedly by $S_{1n} - 2S_{2n} + S_{3n}$, where $S_{1n} = (n-2)!(n!)^{-1}\sum_{i\neq j}(\mathbf{X}_{1i}^{\mathrm{T}}\mathbf{X}_{1j})^2$, $S_{2n} = (n-3)!(n!)^{-1}\sum_{i\neq j\neq k}(\mathbf{X}_{1i}^{\mathrm{T}}\mathbf{X}_{1j}\mathbf{X}_{1j}^{\mathrm{T}}\mathbf{X}_{1k})$, and $S_{3n} = (n-4)!(n!)^{-1}\sum_{i\neq j\neq k\neq l}(\mathbf{X}_{1i}^{\mathrm{T}}\mathbf{X}_{1j}\mathbf{X}_{1k}^{\mathrm{T}}\mathbf{X}_{1l})$.

Next, we study the asymptotic distribution of T_n under a class of local alternatives (2.7) that prescribe a small discrepancy between $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_{10}$. Similar local alternatives are also considered in Zhong and Chen (2011) and in Cui, Guo and Zhong (2018).

$$\begin{aligned}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10})^{\mathrm{T}}\mathbf{\Sigma}_{11}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}) &= o\left(n^{-\kappa}\right), \\ (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10})^{\mathrm{T}}\mathbf{\Sigma}_{11}^3(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}) &= o\left\{n^{-1-\kappa}tr(\mathbf{\Sigma}_{11}^2)\right\}, \\ (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10})^{\mathrm{T}}\mathbf{\Sigma}_{10}\mathbf{\Sigma}_{01}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}) &= o(n^{-1+\kappa}).\end{aligned} \tag{2.7}$$

Theorem 2. *Assume conditions* (C1)–(C4) *hold. Then under the local alternatives* (2.7),

$$\frac{n[T_n - (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10})^{T}\mathbf{\Sigma}_{11}^2(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10})]}{\sigma^2\sqrt{2tr(\mathbf{\Sigma}_{11}^2)}} \xrightarrow{D} N(0,1) \tag{2.8}$$

as $n \to \infty$, where $\xrightarrow{D}$ denotes convergence in distribution.

Theorem 2 implies that the asymptotic power under the local alternatives (2.7) of the CUT is

$$\Psi_n^{\mathrm{CUT}} = \Phi\left(-z_\alpha + \frac{n(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10})^{\mathrm{T}}\mathbf{\Sigma}_{11}^2(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10})}{\sigma^2\sqrt{2tr(\mathbf{\Sigma}_{11}^2)}}\right), \tag{2.9}$$

where $\Phi(\cdot)$ denotes the distribution function of the standard normal distribution. If the signal-to-noise ratio $(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10})^{\mathrm{T}}\mathbf{\Sigma}_{11}^2(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10})/\sigma^2\sqrt{2tr(\mathbf{\Sigma}_{11}^2)}$ has a higher order than n^{-1}, the asymptotic power tends to one as the sample size increases to infinity, and thus the CUT is consistent. Its asymptotic power is the same as that of the conditional test in Wang and Cui (2015). Let $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_{p-q}$ be the eigenvalues of $\mathbf{\Sigma}_{11}$, and suppose all the eigenvalues are bounded from zero and infinity. Similarly to Zhong and Chen (2011), we find that a sufficient condition for ensuring nontrivial power of the CUT is $\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}\| = O(n^{-1/2}\lambda_1^{-1}(\sum_{j=1}^{p-q}\lambda_j^2)^{1/4}) = O(n^{-1/2}(p-q)^{1/4})$. If we further define $\delta_{\boldsymbol{\beta}_1} = \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}\|/\sqrt{p-q}$ as the average "signal strength," then the previous sufficient condition becomes $\delta_{\boldsymbol{\beta}_1} = O(n^{-1/2}(p-q)^{-1/4})$.

3. Conditional Test with Screening

3.1. A two-stage testing procedure

Although the CUT is able to accommodate moderately high-dimensional problems, it performs unsatisfactorily for ultrahigh-dimensional sparse linear models. The sparsity assumption means that only a small subset of predictors are significant to the response. We denote the small set of predictors $\mathbf{X}_1$ by $\mathcal{M}_1 = \{j : \beta_{1j} \neq 0, j = 1, \ldots, p-q\}$, which are truly relevant to the response. Let $s = |\mathcal{M}_1|$ be the cardinality of the significant subset $\mathcal{M}_1$. Under the sparsity assumption, we define $\delta_{\boldsymbol{\beta}_{1\mathcal{M}_1}} = \|\boldsymbol{\beta}_{1\mathcal{M}_1} - \boldsymbol{\beta}_{10\mathcal{M}_1}\|/\sqrt{s} = \sqrt{\sum_{j\in\mathcal{M}_1}(\beta_{1j} - \beta_{10j})^2/s}$ as the average "signal strength," where $\boldsymbol{\beta}_{1\mathcal{M}_1} = \{\beta_j : j \in \mathcal{M}_1\}$. A sufficient condition for the CUT to have nontrivial power is $\delta_{\boldsymbol{\beta}_{1\mathcal{M}_1}} = O(n^{-1/2}s^{-1/2}(p-q)^{1/4})$. If p increases faster than $O(n^2s^2)$, for example, if $p = O(\exp(n^a))$ for some $a > 0$, this sufficient condition is difficult to satisfy.

To reduce the unfavorable effect of the ultrahigh dimensionality and to enhance the testing power of the CUT, we propose the two-stage CUTS algorithm based on a random data-splitting technique under the sparsity assumption. In the first stage, we split the data randomly into two parts $\mathcal{S}_1$ and $\mathcal{S}_2$, and then apply conditional sure independence screening (CSIS; Barut, Fan and Verhasselt (2016)) to the first part $\mathcal{S}_1$ to select a submodel. In the second stage, we apply the proposed CUT to test the significance of the selected submodel, conditional on $\mathbf{X}_0$, based on the second sample $\mathcal{S}_2$. The CUTS algorithm is summarized in Algorithm 1.

Algorithm 1 Conditional U-statistic Test with Screening (CUTS)

Step 1. (Random Data Splitting) Split the sample $\{(Y_i, \mathbf{X}_{0i}, \mathbf{X}_{1i}), i = 1, 2, \ldots, n\}$ randomly into two parts, $\mathcal{S}_1$ with sample size n_1 and $\mathcal{S}_2$ with sample size n_2. In practice, we can let $n_1 = [n/2]$, the integer of $n/2$.

Step 2. (Conditional Sure Independence Screening) Regress $\mathbf{Y}$ against the union of $\mathbf{X}_0$ and each predictor X_{1j} of $\mathbf{X}_1$ using $\mathcal{S}_1$, i.e., $Y = \alpha + \mathbf{X}_0\boldsymbol{\beta}_0 + \beta_{1j}X_{1j} + \xi$, and obtain the estimators $\widehat{\beta}_{1j}$ for each $j = 1, \ldots, p-q$. Then, select the submodel $\widehat{\mathcal{M}}_1 = \{j : |\widehat{\beta}_{1j}|$ is among the top d_n largest ones$\}$, where d_n is a prespecified threshold, e.g., set $d_n = [n_1/\log(n_1)]$.

Step 3. (Conditional U-statistic Test) Apply the CUT to test the significance of $\mathbf{X}_{1\widehat{\mathcal{M}}_1}$ for the response conditional on $\mathbf{X}_0$ based on the rejection rule (2.6) in $\mathcal{S}_2$ at the significance level α, where $\mathbf{X}_{1\widehat{\mathcal{M}}_1} = \{X_{1j} : j \in \widehat{\mathcal{M}}_1\}$.

In Step 2, the CSIS proposed by Barut, Fan and Verhasselt (2016) is used to eliminate noisy variables and reduce the ultrahigh dimensionality. The sure screening property of CSIS that demonstrates $P(\mathcal{M}_1 \subset \widehat{\mathcal{M}}_1) \to 1$ as $n \to \infty$ ensures the power enhancement of the CUTS under the sparsity assumption. When $\mathcal{M}_1 \subset \widehat{\mathcal{M}}_1$ holds, the original hypothesis (1.2), $H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_{10}$ versus $H_1 : \boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_{10}$, is equivalent to $H_0 : \boldsymbol{\beta}_{1\widehat{\mathcal{M}}_1} = \boldsymbol{\beta}_{10\widehat{\mathcal{M}}_1}$ versus $H_1 : \boldsymbol{\beta}_{1\widehat{\mathcal{M}}_1} \neq \boldsymbol{\beta}_{10\widehat{\mathcal{M}}_1}$, where $\boldsymbol{\beta}_{1\widehat{\mathcal{M}}_1} = \{\beta_{1j} : j \in \widehat{\mathcal{M}}_1\}$. Therefore, the first-stage CSIS helps us to transform an ultrahigh-dimensional testing problem into an asymptotically equivalent low-dimensional testing one, which can be tested efficiently using the CUT in the second stage.

Given the submodel $\widehat{\mathcal{M}}_1$ in the first stage, Theorem 2 implies that the asymptotic power in terms of n_2 under the local alternatives (2.7) of the CUTS test procedure is

$$\Psi_n^{\text{CUTS}}(\widehat{\mathcal{M}}_1) = \Phi\left(-z_\alpha + \frac{n_2(\boldsymbol{\beta}_{1\widehat{\mathcal{M}}_1} - \boldsymbol{\beta}_{10\widehat{\mathcal{M}}_1})^{\mathrm{T}}\boldsymbol{\Sigma}^2_{11\widehat{\mathcal{M}}_1}(\boldsymbol{\beta}_{1\widehat{\mathcal{M}}_1} - \boldsymbol{\beta}_{10\widehat{\mathcal{M}}_1})}{\sigma^2\sqrt{2tr(\boldsymbol{\Sigma}^2_{11\widehat{\mathcal{M}}_1})}}\right), \quad (3.1)$$

where $\boldsymbol{\Sigma}^2_{11\widehat{\mathcal{M}}_1}$ denotes the covariance matrix of selected predictors, indexed by $\widehat{\mathcal{M}}_1$. Assume that all eigenvalues of $\boldsymbol{\Sigma}_{11}$ satisfy $c < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_{p-q} \leq C$, where c, C are two constants. By Fatou's lemma, the upper and lower limits of the mean power function are controlled by

$$\begin{aligned}
\liminf E\Psi_n^{\text{CUTS}}(\widehat{\mathcal{M}}_1) &\geq E \liminf \Psi_n^{\text{CUTS}}(\widehat{\mathcal{M}}_1) \\
&\geq \liminf \Phi\left(-z_\alpha + \frac{n_2\|\boldsymbol{\Sigma}_{11\mathcal{M}_1}(\boldsymbol{\beta}_{1\mathcal{M}_1} - \boldsymbol{\beta}_{10\mathcal{M}_1})\|^2}{\sigma^2\sqrt{2Cd_n}}\right), \\
\limsup E\Psi_n^{\text{CUTS}}(\widehat{\mathcal{M}}_1) &\leq E \limsup \Psi_n^{\text{CUTS}}(\widehat{\mathcal{M}}_1) \\
&\leq \limsup \Phi\left(-z_\alpha + \frac{n_2\|\boldsymbol{\Sigma}_{11\mathcal{M}_1}(\boldsymbol{\beta}_{1\mathcal{M}_1} - \boldsymbol{\beta}_{10\mathcal{M}_1})\|^2}{\sigma^2\sqrt{2cd_n}}\right),
\end{aligned} \quad (3.2)$$

where the second and the fourth inequalities hold because $P(\mathcal{M}_1 \subset \widehat{\mathcal{M}}_1) \to 1$ as $n \to \infty$. We define $\delta_{\boldsymbol{\beta}_{1\mathcal{M}_1}} = \|\boldsymbol{\beta}_{1\mathcal{M}_1} - \boldsymbol{\beta}_{10\mathcal{M}_1}\|/\sqrt{|\mathcal{M}_1|}$ as the average "signal strength." Then, the sufficient condition for nontrivial power becomes $\delta_{\boldsymbol{\beta}_{1\mathcal{M}_1}} = O(n^{-1/2}s^{-1/2}d_n^{1/4})$. Furthermore, we can compare the asymptotic power of the WC test (Wang and Cui (2015)) and that of the CUTS with $n_2/n = O(1)$ by comparing their signal-to-noise (SNR) ratios:

$$\frac{\text{SNR}^{\text{CUTS}}}{\text{SNR}^{\text{WC}}} = O(1)\frac{\|\boldsymbol{\Sigma}_{11\mathcal{M}_1}(\boldsymbol{\beta}_{1\mathcal{M}_1} - \boldsymbol{\beta}_{10\mathcal{M}_1})\|^2}{\|\boldsymbol{\Sigma}_{11}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10})\|^2}\sqrt{\frac{p-q}{d_n}} = O((p-q)^{1/2}d_n^{-1/2}).$$

In the mean sense, the asymptotic power of the CUTS is greater than that of the WC test if $d_n = o(p-q)$ and the sure screening property holds.

In addition, we add three remarks on the CUTS.

Remark 1. The goal of the first-stage screening is to reduce noisy signals and to enhance the power of the test in the second-stage under the sparsity assumption. Related noise reduction ideas have been investigated in the literature on hypothesis testing. For example, Lan et al. (2016) introduced the key confounder controlling (KFC) method, similar to the screening idea in Fan and Lv (2008), to first control for predictors that are highly correlated with the target covariate before testing the significance of the single regression coefficient in high-dimensional linear models. A related idea is the thresholding test in which sufficiently small signals are truncated to zero. Fan (1996) proposed a wavelet thresholding test for the mean of random vectors. Zhong, Chen and Xu (2013) and Chen, Li and Zhong (2019) tested for a one-sample mean vector and two-sample mean vectors, respectively, of high-dimensional populations using thresholding to remove the nonsignal-bearing dimensions. Another idea is to only consider the maximum signal component as the test statistic. For example, Cai, Liu and Xia (2014) proposed a maximum-norm test statistic for comparing high-dimensional two-sample means with sparsity. However, thresholding and maximum-norm tests may suffer from size inflation due to spurious correlations in ultrahigh-dimensional data.

Remark 2. The sure screening property is not necessary for the nontrivial power of the CUTS procedure. To ensure nontrivial power, we require a less restrictive necessary condition that at least one truly relevant predictor is selected, that is, $\mathcal{M}_1 \cap \widehat{\mathcal{M}}_1 \neq \emptyset$. We suppose that the eigenvalues of $\mathbf{\Sigma}_{11}$ are bounded from zero and infinity. It can be shown that given $\widehat{\mathcal{M}}_1$, if $\|\boldsymbol{\beta}_{1(\mathcal{M}_1\cap\widehat{\mathcal{M}}_1)} - \boldsymbol{\beta}_{10(\mathcal{M}_1\cap\widehat{\mathcal{M}}_1)}\|^2$ is not less than $O(\sqrt{d_n/(p-q)})\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}\|^2$, the asymptotic power of the CUTS in terms of n_2 is no less than that of the WC test. In other words, when H_1 is true, once the first-stage screening is able to identify some important predictors, the second-stage test could be statistically significant to reject H_0.

Remark 3. Note that random data splitting is useful to eliminate the effect of spurious correlation due to the ultrahigh dimensionality and to control the type-I error rates. Fan, Guo and Hao (2012) pointed out that spurious correlations are inherent in ultrahigh-dimensional data analysis. That is, the maximum sample correlation between the response and irrelevant predictors increases as the predictor dimension increases. Some irrelevant predictors may be detected as significant owing to spurious correlations, even under $H_0 : \boldsymbol{\beta}_1 = \mathbf{0}$. If we do not split the data, the type-I error rates of the second-stage testing procedure will be severely inflated, because the submodel $\widehat{\mathcal{M}}_1$ contains spuriously significant predictors. However, the random data splitting prevents the inflation of the type-I error rates. To appreciate why, we suppose that the sample correlation between an irrelevant predictor and the response is high over the first half of the data and, thus, this predictor is selected by the screening procedure. Because the two halves of the data are independent, it is unlikely that this predictor is also highly correlated with the response over the second half of the data, and thus has a negligible influence on the testing result.

3.2. An ensemble testing procedure

Although the random data splitting is useful to avoid type-I error rates, the testing power may be affected by the randomness and the sample reduction. As Lockhart et al. (2014) mentioned, sample splitting can result in a loss of power in significance testing. To this end, we introduce a more powerful ensemble CUTS algorithm based on multiple random data splitting to further enhance both the empirical power and the algorithm stability. This idea is motivated by Meinshausen, Meier and Buhlmann (2009), who proposed aggregating inference results across multiple random splits to control both the family-wise error rate and the false discovery rate. The ensemble CUTS algorithm based on multiple random data splitting, denoted by CUTS_M, is summarized in Algorithm 2. In Proposition 1, we demonstrate that the family-wise error rate of the CUTS_M is asymptotically controlled at a given significance level $\alpha \in (0, 1)$.

Proposition 1. *For a significance level $\alpha \in (0, 1)$, the family-wise error rate of the $CUTS_M$ is asymptotically controlled at level α. That is,*

$$\limsup_{n\to\infty} P(Q^* \leq \alpha | H_0) \leq \alpha. \tag{3.3}$$

4. Numerical Studies

4.1. Simulations

This section investigates the finite-sample performance of the WC test (Wang and Cui (2015)), CUTS, and CUTS_M for ultrahigh-dimensional linear regression coefficients using Monte Carlo simulations. In the simulations, we set $M = 20$ times for the CUTS_M.

Example 1. We generate the predictors $(X_1, X_2, \ldots, X_p)^T$ from two distributions: (i) a multivariate normal distribution $\mathcal{N}(\mathbf{0}, \Sigma)$, or (ii) a multivariate t-distribution $\sqrt{1 - 2/q}\, t_q(0, \Sigma, q)$ with $q = 5$, where $\Sigma = (\sigma_{ij})_{p\times p}$ with

Algorithm 2 CUTS$_M$ Algorithm based on multiple random data splitting.

Step 1. (Conditional U-statistic Test with Screening) Split the sample $\{(Y_i, \mathbf{X}_{0i}, \mathbf{X}_{1i}), i = 1, 2, \ldots, n\}$ randomly into two equal parts, $\mathcal{S}_1$ and $\mathcal{S}_2$, and apply Algorithm 1 to obtain a p-value, denoted by p_1.

Step 2. (Multiple Data Splitting) Repeat Step 1 m times and obtain m p-values, denoted by $\{p_1, \ldots, p_m\}$.

Step 3. (Compute Adjusted P-value) Compute the adjusted p-value

$$Q^* = \min\left\{1, (1 - \log\gamma_{\min}) \inf_{\gamma \in (\gamma_{\min}, 1)} Q(\gamma)\right\},$$

where $Q(\gamma) = \min[1, q_\gamma(\{p_k/\gamma; k = 1, \ldots, m\})]$ for a constant $\gamma \in (\gamma_{\min}, 1)$, $q_\gamma(\{p_k/\gamma\})$ is the γth quantile of $\{p_k/\gamma; k = 1, \ldots, m\}$, and $\gamma_{\min}$ is a prespecified constant in (0,1).

Step 4. (Rejection) The null hypothesis H_0 (1.2) is rejected at the significance level α if $Q^* \leq \alpha$.

$\sigma_{ij} = 0.5^{|i-j|}$. The regression model is set as

$$Y = 0.7X_1 + 0.8X_2 + 0.6X_3 - X_4 + \beta_{11}X_{11} + \beta_{12}X_{12} + \beta_{13}X_{13} + \beta_{14}X_{14} + \varepsilon,$$

where the error term ε is independently generated from two distributions: (i) a standard normal distribution $\mathcal{N}(0, 1)$, or (ii) a standard log-normal distribution $(lnorm(0, 1) - e^{1/2})/\sqrt{e(e-1)}$. Assume that the known conditional set is $\mathcal{M}_0 = \{1, 2, 3, 4, 5\}$. We want to test the overall significance of the remaining regression coefficients, given the subset $\mathcal{M}_0$, that is, $H_0 : \boldsymbol{\beta_1} = \mathbf{0}$ versus $H_1 : \boldsymbol{\beta_1} \neq \mathbf{0}$, where $\boldsymbol{\beta_1} = (\beta_6, \ldots, \beta_p)^{\mathrm{T}}$. We set $\beta_j = c/2, j = 11, \ldots, 14$, where the signal strength $c^2 \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$, and $c = 0$ corresponds to the null hypothesis H_0. The sample size $n = 100$ and the predictor dimension $p =$ 1,000 or 2,000. We run the simulations 500 times and compare the empirical size or power of the three tests, WC, CUTS, and CUTS$_M$, at the significance level $\alpha = 0.05$. All simulation results are summarized in Table 1. We observe that the two-stage testing procedures enhance the empirical power substantially based on random data splitting under sparsity. In particular, the CUTS$_M$ approach based on the multiple splitting strategy is more powerful and algorithmically stable than is the single-splitting CUTS. The family-wise error rate of the CUTS$_M$ approach is also favorably controlled under the significance level $\alpha = 0.05$.

Example 2. We further consider the power performance for dense signals. We generate the predictors from the same multivariate normal distribution as that

Table 1. Empirical size and power of WC, CUTS, CUTS_M in Example 1.

(n,p)	c^2	$\varepsilon\sim$Normal			$\varepsilon\sim$Log-normal		
		WC	CUTS	CUTS_M	WC	CUTS	CUTS_M
(1) $\mathbf{X}_i\sim\mathcal{N}_p(\mu,\Sigma)$							
$(100, 1{,}000)$	0.0	0.062	0.058	0.034	0.044	0.028	0.038
	0.1	0.172	0.252	0.400	0.242	0.424	0.596
	0.2	0.326	0.604	0.844	0.366	0.664	0.836
	0.3	0.410	0.838	0.968	0.494	0.836	0.944
	0.4	0.550	0.918	0.994	0.580	0.910	0.962
	0.5	0.620	0.970	0.996	0.616	0.938	0.988
$(100, 2{,}000)$	0.0	0.054	0.054	0.046	0.040	0.040	0.038
	0.1	0.118	0.168	0.278	0.146	0.308	0.466
	0.2	0.196	0.506	0.768	0.256	0.626	0.794
	0.3	0.260	0.784	0.946	0.344	0.778	0.884
	0.4	0.348	0.892	0.990	0.370	0.884	0.950
	0.5	0.412	0.960	0.998	0.388	0.908	0.978
(2) $\mathbf{X}_i\sim\sqrt{1-2/q}\,t_q(\mu,\Sigma,q)$							
$(100, 1{,}000)$	0.0	0.042	0.048	0.024	0.038	0.058	0.030
	0.1	0.176	0.168	0.300	0.220	0.296	0.520
	0.2	0.278	0.462	0.682	0.368	0.576	0.790
	0.3	0.336	0.656	0.892	0.458	0.716	0.884
	0.4	0.470	0.832	0.970	0.476	0.818	0.938
	0.5	0.508	0.904	0.992	0.522	0.870	0.966
$(100, 2{,}000)$	0.0	0.038	0.052	0.046	0.054	0.046	0.042
	0.1	0.130	0.128	0.250	0.126	0.214	0.378
	0.2	0.190	0.346	0.612	0.254	0.524	0.734
	0.3	0.260	0.580	0.876	0.296	0.678	0.856
	0.4	0.336	0.760	0.950	0.352	0.776	0.918
	0.5	0.364	0.852	0.976	0.398	0.846	0.954

in Example 1. Consider the linear regression

$$Y = 0.7X_1 + 0.8X_2 + 0.6X_3 - X_4 + \mathbf{X}_1\boldsymbol{\beta}_1 + \varepsilon,$$

where $\boldsymbol{\beta}_\mathbf{1} = (\beta_6,\ldots,\beta_p)^\top$, $\beta_j = c/2$ for $j = 11,\ldots,20$, $\beta_j = c/\sqrt{6}$ for $j = 21,\ldots,30$, $\beta_j = c/2\sqrt{2}$ for $j = 31,\ldots,40$, $\beta_j = 0.01$ for $j = 41,\ldots,p/2$, and $\beta_j = 0$ otherwise, where the signal strength $c^2\in\{0.1,0.2,0.3,0.4,0.5\}$. The other settings are as in Example 1. Table 2 shows the empirical power of each test at $\alpha = 0.05$. The WC test performs generally well in terms of detecting the weak and dense signals. Although the CUTS with the single data split has nontrivial power, it performs worse than the WC test under the dense signal settings. This result is consistent with Remark 2. However, the ensemble CUTS_M procedure

Table 2. Empirical power of WC, CUTS, CUTS_M in Example 2.

(n,p)	c^2	$\varepsilon \sim$Normal			$\varepsilon \sim$Log-normal		
		WC	CUTS	CUTS_M	WC	CUTS	CUTS_M
(100, 1,000)	0.1	0.880	0.558	0.816	0.856	0.628	0.806
	0.2	0.956	0.796	0.954	0.934	0.802	0.926
	0.3	0.968	0.862	0.986	0.966	0.892	0.968
	0.4	0.984	0.926	0.992	0.982	0.928	0.986
	0.5	0.992	0.918	0.996	0.986	0.936	0.990
(100, 2,000)	0.1	0.664	0.420	0.658	0.678	0.486	0.656
	0.2	0.812	0.628	0.852	0.808	0.612	0.832
	0.3	0.854	0.728	0.932	0.860	0.706	0.912
	0.4	0.886	0.768	0.950	0.862	0.804	0.940
	0.5	0.890	0.826	0.968	0.894	0.858	0.976

with multiple data splitting enhances the power when the signals are not small.

Example 3. We consider a linear model similar to that of Fan and Lv (2008):

$$Y = k_0X_1 + k_0X_2 + k_0X_3 - 3k_0\sqrt{\rho}X_4 + \varepsilon,$$

where each X_j is generated from a standard normal distribution, all X_j for $j = 1, 2, 3, 5, \ldots, 10$ are equally correlated with the correlation coefficient ρ, and the correlation between X_4 and each other predictor X_j for $j = 1, 2, 3, 5, \ldots, 10$ is $\sqrt{\rho}$. All other predictors are independent and ε follows an independent standard normal distribution. It can be demonstrated that the marginal correlation between X_4 and Y is zero and that the sure independence screening (SIS) cannot detect X_4. Fan and Lv (2008) proposed the iterative SIS (ISIS) to identify X_4. In our simulations, we aim to test the overall significance of the regression coefficients of the remaining predictors, given a subset of important predictors $\mathcal{M}_0 = \{1, 2, 3\}$ or $\{1, 2, 3, 4\}$. When $\mathcal{M}_0 = \{1, 2, 3, 4\}$, $H_0 : \boldsymbol{\beta}_1 = \mathbf{0}$ is true. We set the sample size $n = 200$, the dimension $p = 2{,}000$ or $5{,}000$, and the signal strength $k_0 = 1, 2, 3$. Table 3 shows that all tests retain the nominal size $\alpha = 0.05$ well when $\mathcal{M}_0 = \{1, 2, 3, 4\}$. If $\mathcal{M}_0 = \{1, 2, 3\}$ and there is only one important variable X_4 left in the remaining high-dimensional variables, both the CUTS and the CUTS_M perform much better in terms of rejecting H_0. Thus, the result shows that the ISIS is necessary to recruit additional important variables. This example illustrates that the conditional test is useful for checking whether the variable screening procedures adequately identify all important variables in the selected submodel under the sparsity assumption.

Table 3. Empirical size and power of WC, CUTS, CUTS_M in Example 3.

$\mathcal{M}_0$	k_0	$p = 2{,}000$			$p = 5{,}000$		
		WC	CUTS	CUTS_M	WC	CUTS	CUTS_M
$\{1,2,3,4\}$	3	0.050	0.058	0.044	0.044	0.048	0.038
$\{1,2,3\}$	1	0.630	0.986	1.000	0.408	0.980	1.000
	2	0.658	0.992	1.000	0.416	0.988	1.000
	3	0.666	0.998	1.000	0.420	0.994	1.000

4.2. Real-data analysis

Example 4. Scheetz et al. (2006) used expression quantitative trait locus mapping to gain a broad perspective of gene regulation in the mammalian eye of 120 12-week-old male rats. They identified 22 of 18,976 gene probes as important in regulating mammalian eye gene expression. Among them, seven genes showed evidence of contiguous regulation alone, four had both contiguous and noncontiguous linkages, and 11 had evidence of only noncontiguous linkages (refer to Table 2 in Scheetz et al. (2006)). We consider a linear regression model of the response gene TRIM32, which relates to the retinal disease called Bardet–Biedl syndrome, against the remaining 18,975 probes. A natural question is whether the remaining ultrahigh-dimensional variables still contribute to the response, given a subset of the identified significant genes.

We apply the WC test, the CUTS test with single data splitting, and the CUTS_M algorithm with $M = 50$ to test the overall significance of the regression coefficients of the remaining ultrahigh-dimensional gene probes, conditional on various subsets of the 22 genes identified in Scheetz et al. (2006). We delete one outlier (the 58th observation) in our analysis and report the p-values in Table 4. If the conditioning set contains all 22 identified genes ($\mathcal{M}_0(1:22)$), the tests are all nonsignificant and conclude that the remaining ultrahigh-dimensional genes may not contribute to the response, given these 22 genes. Conditional on the seven genes with only contiguous regulation ($\mathcal{M}_0(1:7)$) or the four genes with both contiguous and noncontiguous linkages ($\mathcal{M}_0(8:11)$), all three tests are statistically significant at the level $\alpha = 0.01$, implying that there are more important genes for the response in the remaining ones. However, when the conditioning set includes the first 11 genes with contiguous linkages ($\mathcal{M}_0(1:11)$) or the last 11 genes with only noncontiguous linkages ($\mathcal{M}_0(12:22)$), only CUTS_M rejects the null H_0. In addition, we also report the adjusted R^2 of the linear regressions of the response against various subsets of the 22 genes in Table 4. The linear model with all 22 genes produces the largest adjusted R^2. This data

Table 4. P-values of WC, CUTS, CUTS_M in Example 4.

Conditioning Set	$\mathcal{M}_0(1:7)$	$\mathcal{M}_0(8:11)$	$\mathcal{M}_0(1:11)$	$\mathcal{M}_0(12:22)$	$\mathcal{M}_0(1:22)$
P-value (WC)	0.0011	<0.0001	0.3006	0.0781	0.7016
P-value (CUTS)	0.0046	<0.0001	0.1371	0.1314	0.9410
P-value (CUTS_M)	<0.0001	<0.0001	0.0002	<0.0001	1
Adjusted R^2	0.291	0.231	0.354	0.270	0.417

Notes: $\mathcal{M}_0(1:7)$ denotes the subset of seven genes with only contiguous linkages; $\mathcal{M}_0(8:11)$ denotes the subset of four genes with both contiguous and noncontiguous linkages; $\mathcal{M}_0(1:11)$ is the union of $\mathcal{M}_0(1:7)$ and $\mathcal{M}_0(8:11)$; $\mathcal{M}_0(12:22)$ denotes the subset of 11 genes with only noncontiguous linkages. The number of random data splits for the CUTS_M is M=50.

Table 5. P-values and power of WC and CUTS_M in Example 5.

Conditioning Set	Top 1	Top 1:2	Top 1:3	Top 1:4	Top 1:5	Random 4 Genes
			P-value			Power
WC	<0.0001	0.0034	0.0093	0.0084	0.1630	0.795
CUTS_M	<0.0001	0.0045	0.0003	0.0053	0.0752	0.820
Adjusted R^2	0.584	0.776	0.781	0.773	0.778	0.168(0.157)

Notes: Top 1:k denotes the subset of top k genes ranked by the DC-SIS. Random 4 Genes denotes the subset of four genes selected randomly from all genes except the top 40 genes ranked by the DC-SIS. The last column is based on 200 repetitions, and 0.168(0.157) denotes the average adjusted R^2 and its standard deviation. The number of random data splits for the CUTS_M is M=50.

analysis supports the power enhancement of the CUTS_M.

Example 5. Li, Zhong and Zhu (2012) used distance correlation (DC-SIS) to rank the most influential genes for the expression level of a G protein-coupled receptor (Ro1) in a cardiomyopathy microarray data set (Segal, Dahlquist and Conklin (2003)). In this data set, we have only 30 observations, but the dimension of the genes as predictors is 6,319. We set the conditioning set as the subset of the top k genes ranked by the DC-SIS, and test the overall significance of the remaining ultrahigh-dimensional genes using the WC test and the CUTS_M with $M = 50$. We do not include the CUTS with a single data split because the sample size is only 30 and the result of the CUTS is not stable and depends heavily on the data splits. This drawback can be addressed by using the ensemble CUTS_M procedure, as discussed before. For the power comparison, we set the conditioning set as a subset of four genes randomly selected from all genes except the top 40 genes ranked by the DC-SIS. In this case, the null hypothesis is not true because the top 40 genes should contain important genes for the response Ro1. We repeat this 200 times and compute the empirical power of the WC and CUTS_M tests at the significance level $\alpha = 0.05$. In addition, we report the adjusted R^2 of the linear regressions of the response against the conditioning sets of genes. Table 5 summarizes the results. WC and CUTS_M have similar results, implying that conditional on the top four genes selected by the DC-SIS, the remaining 6315 genes are not statistically significant in the linear regression. Moreover, the CUTS_M has better empirical power in terms of rejecting the null hypothesis conditional on the four random genes.

5. Discussion

We have proposed a two-stage conditional U-statistic test with screening (CUTS) procedure for testing the overall significance of the regression coefficients of the remaining ultrahigh-dimensional predictors, given a subset of known predictors. The procedure reduces the dimensionality under the sparsity assumption and enhances the empirical power using a random data-splitting strategy. The ensemble CUTS_M algorithm based on a multiple splitting strategy is demonstrated to be powerful in simulations. This two-stage testing procedure can be applied directly to unconditional tests of ultrahigh-dimensional linear regression coefficients by setting the conditional set as an empty set, and is able to improve the power performance of the tests in Zhong and Chen (2011) and Cui, Guo and Zhong (2018) under the sparsity condition.

It is also interesting to extend the thresholding tests of Zhong, Chen and Xu (2013) and Chen, Li and Zhong (2019) to test a high-dimensional linear regression. We let

$$\Delta_{i,j}(X_1^{(k)}) = (X_{1i}^{(k)} - \overline{X}_1^{(k)})^{\mathrm{T}}(X_{1j}^{(k)} - \overline{X}_1^{(k)}) + \frac{|X_{1i}^{(k)} - X_{1j}^{(k)}|^2}{2n}.$$

Then, the test statistic in (2.4) can be written as $T_n = \sum_{k=1}^{p-q} T_n^{(k)}$, where

$$T_n^{(k)} = \left(1 - \frac{2}{n}\right)^{-2} \binom{n}{2}^{-1} \sum_{i=2}^{n} \sum_{j=1}^{i-1} \Delta_{i,j}(X_1^{(k)}) \Delta_{i,j}(\mathbf{Y}^*).$$

To remove nonsignal bearing $T_n^{(k)}$ and keep those with signals, we define the thresholding test statistic as

$$L_n(\lambda_n) = \sum_{k=1}^{p-q} n T_n^{(k)} I\{n T_n^{(k)} \geq \lambda_n\},$$

where $I(\cdot)$ is the indicator function, and λ_n is the thresholding level. It is worth investigating the power performance and theoretical properties of the thresholding test for high-dimensional sparse linear regressions in future research. We will also study how to determine the thresholding level and check how spurious correlations affect the thresholding test for ultrahigh-dimensional cases.

Acknowledgments

We thank the editor, associate editor and two referees for their constructive comments. Guo's research was supported by a National Natural Science Foundation of China (NNSFC) grant (11901406), and the Science and Technology Project of Beijing Municipal Commission of Education (KM202110028017). Zhong's research was supported by NNSFC grants (11922117, 11671334), the Fujian Provincial Science Fund for Distinguished Young Scholars (2019J06004), and the Basic Scientific Center Project of National Science Foundation of China (71988101) and the 111 Project (B13028). Cui's research was supported by the State Key Program of the National Natural Science Foundation of China (12031016), NNSFC grants (11971324, 11471223), the Beijing Science and Technology Innovation Platform Construction Project funding, the Interdisciplinary Construction of Bioinformatics and Statistics, and the Academy for Multidisciplinary Studies, Capital Normal University.

Appendix

A. Technical Proofs

The test statistic (2.4) is invariant to location shifts in both $\mathbf{X}_i$ and Y_i, so we assume, without loss of generality, that $\alpha = 0$ and $\boldsymbol{\mu} = \mathbf{0}$ in the rest of the article. For convenience, we denote $\boldsymbol{\delta}_{\beta_0} = \boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}_0$, $\boldsymbol{\delta}_{\beta_1} = \boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}$, $B_i = \boldsymbol{\delta}_{\beta_1}^{\mathrm{T}} \boldsymbol{\Sigma}_{11}^i \boldsymbol{\delta}_{\beta_1}$, and $c_i, i = 1, 2, 3, \ldots$ are some positive constants which are independent of the samples. We first present two lemmas which have been shown in Cui, Guo and Zhong (2018).

Lemma 1. *Let* $\mathbf{U} = (U_1, \ldots, U_p)^T$ *be a random vector uniformly distributed on the unit sphere in* $\mathbb{R}^p$*. Then* $E(\mathbf{U}) = 0$, $Var(\mathbf{U}) = p^{-1}\mathbf{I}_p$, $E(U_j^4) = 3/p(p+2), \forall\ j = 1, \ldots, p$, *and* $E(U_j^2 U_k^2) = 1/(p(p+2))$ *for* $j \neq k$.

Lemma 2. *Suppose condition* (C2) *holds, then we have* $E(\mathbf{U_1 U_1^{\mathrm{T}} M U_1 U_1^{\mathrm{T}}}) = 1/(p(p+2))\left(2\mathbf{M} + tr(\mathbf{M})\mathbf{I}_p\right)$, *where* $\mathbf{M}$ *is a* $p \times p$ *symmetric matrix.*

Lemma 3. *Suppose conditions* (C2)–(C3) *hold, then we have* $\|\widehat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_0\|^2 = O_P(n^{-1+\kappa})$ *under the local alternatives* (2.7).

Proof of Lemma 3. The ordinary least squared estimator of $\boldsymbol{\beta}_0$ implies that

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_0 &= (\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0)^{-1}\mathbf{X}_0^{\mathrm{T}}(\mathbf{Y}-\mathbf{X}_1\boldsymbol{\beta}_{10})\\ &= \boldsymbol{\beta}_0 + \left(\frac{1}{n}\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0\right)^{-1}\frac{1}{n}\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_1(\boldsymbol{\beta}_1-\boldsymbol{\beta}_{10}) + \left(\frac{1}{n}\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0\right)^{-1}\frac{1}{n}\mathbf{X}_0^{\mathrm{T}}\boldsymbol{\varepsilon}\\ &=: \boldsymbol{\beta}_0 + W_1 + W_2.\end{aligned}$$

Under condition (C3), it is obtained that $((1/n)\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0)^{-1}$ converges to $\boldsymbol{\Sigma}_{00}^{-1}$ in probability. Write $W_1^* = (1/n)\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_1(\boldsymbol{\beta}_1-\boldsymbol{\beta}_{10})$ and $W_2^* = (1/n)\mathbf{X}_0^{\mathrm{T}}\boldsymbol{\varepsilon}$. Then we have $E\|W_1^*\|^2 = (n+1/n)\boldsymbol{\delta}_{\beta_1}^{\mathrm{T}}\boldsymbol{\Sigma}_{10}\boldsymbol{\Sigma}_{01}\boldsymbol{\delta}_{\beta_1} + (1/n)\mathrm{tr}(\boldsymbol{\Sigma}_{00})\boldsymbol{\delta}_{\beta_1}^{\mathrm{T}}\boldsymbol{\Sigma}_{10}\boldsymbol{\Sigma}_{01}\boldsymbol{\delta}_{\beta_1}$ and $E\|W_2^*\|^2 = (1/n)\sigma^2\mathrm{tr}(\boldsymbol{\Sigma}_{00})$, which imply that $\|W_1\|^2 = O_P(n^{-1+\kappa})$ and $\|W_1\|^2 = O_P(n^{-1+\kappa})$ under condition (C3) and the local alternatives (2.7). Then this lemma follows.

Proof of Theorems 1 and 2.

It is easy to see that the local alternatives (2.7) is satisfied naturally under the null hypothesis. Then Theorem 1 could be considered as a special case of Theorem 2. Therefore it is just needed to prove Theorem 2. In order to simplify the calculation, we re-formulate $\triangle_{i,j}$ as follows:

$$\begin{aligned}&\frac{n}{n-2}\triangle_{i,j}(\mathbf{X}_1)\\ &= \left(1-\frac{1}{n}\right)\mathbf{X}_{1i}^{\mathrm{T}}\mathbf{X}_{1j} - \frac{1}{2n}\left(\mathbf{X}_{1i}^{\mathrm{T}}\mathbf{X}_{1i}+\mathbf{X}_{1j}^{\mathrm{T}}\mathbf{X}_{1j}-2E(\mathbf{X}_{11}^{\mathrm{T}}\mathbf{X}_{11})\right)\\ &\quad -\left(1-\frac{2}{n}\right)\overline{\mathbf{X}}_1^{(i,j)\mathrm{T}}(\mathbf{X}_{1i}+\mathbf{X}_{1j}) + \left(1-\frac{2}{n}\right)\left[\overline{\mathbf{X}}_1^{(i,j)\mathrm{T}}\overline{\mathbf{X}}_1^{(i,j)} - \frac{E(\mathbf{X}_{11}^{\mathrm{T}}\mathbf{X}_{11})}{n-2}\right]\\ &=: M_{ij}^{(1)} + M_{ij}^{(2)} + M_{ij}^{(3)} + M_{ij}^{(4)}, \qquad \text{(A.1)}\end{aligned}$$

where $\overline{\mathbf{X}}_1^{(i,j)} = 1/(n-2)\sum_{-(i,j)} X_{1k}$, that is the average of $X_k's$ with deleting the i-th and j-th samples respectively. Let $\mathbf{H} = \mathbf{Y}-\mathbf{X}_1\boldsymbol{\beta}_{10}-\mathbf{X}_0\boldsymbol{\beta}_0 = \alpha+\mathbf{X}_1(\boldsymbol{\beta}_1-\boldsymbol{\beta}_{10})+\boldsymbol{\varepsilon}$, and thus $\mathbf{Y}^* = \mathbf{Y}-\mathbf{X}_0\widehat{\boldsymbol{\beta}}_0-\mathbf{X}_1\boldsymbol{\beta}_{10} = \mathbf{H}+\mathbf{X}_0(\boldsymbol{\beta}_0-\widehat{\boldsymbol{\beta}}_0)$. Furthermore, we obtain that

$$\begin{aligned}&\Delta_{i,j}(\mathbf{Y}^*) - \Delta_{i,j}(\mathbf{H})\\ &= (1-n^{-1})(\mathbf{X}_{0i}-\overline{\mathbf{X}}_0)^{\mathrm{T}}\boldsymbol{\delta}_{\beta_0}(H_j-\overline{H}) + (1-n^{-1})(H_i-\overline{H})(\mathbf{X}_{0j}-\overline{\mathbf{X}}_0)^{\mathrm{T}}\boldsymbol{\delta}_{\beta_0}\\ &\quad + (1-n^{-1})\boldsymbol{\delta}_{\beta_0}^{\mathrm{T}}(\mathbf{X}_{0i}-\overline{\mathbf{X}}_0)(\mathbf{X}_{0j}-\overline{\mathbf{X}}_0)^{\mathrm{T}}\boldsymbol{\delta}_{\beta_0} + n^{-1}(H_i-\overline{H})(\mathbf{X}_{0i}-\overline{\mathbf{X}}_0)^{\mathrm{T}}\boldsymbol{\delta}_{\beta_0}\\ &\quad + n^{-1}(H_j-\overline{H})(\mathbf{X}_{0j}-\overline{\mathbf{X}}_0)^{\mathrm{T}}\boldsymbol{\delta}_{\beta_0} + (2n)^{-1}\boldsymbol{\delta}_{\beta_0}^{\mathrm{T}}(\mathbf{X}_{0i}-\overline{\mathbf{X}}_0)(\mathbf{X}_{0i}-\overline{\mathbf{X}}_0)^{\mathrm{T}}\boldsymbol{\delta}_{\beta_0}\end{aligned}$$

$$+ (2n)^{-1}\boldsymbol{\delta}_{\beta_0}^{\mathrm{T}}(\mathbf{X}_{0j} - \overline{\mathbf{X}}_0)(\mathbf{X}_{0j} - \overline{\mathbf{X}}_0)^{\mathrm{T}}\boldsymbol{\delta}_{\beta_0} =: \sum_{k=1}^{7} K_k.$$

Write $T_0 = 2/(n(n-1))\sum_{i>j}\triangle_{i,j}(\mathbf{X}_1)\triangle_{i,j}(\mathbf{H})$, and by Theorem 3.2 and 3.4 in Cui, Guo and Zhong (2018), we obtain that

$$\frac{n[T_0 - (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10})^{\mathrm{T}}\boldsymbol{\Sigma}_{11}^2(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10})]}{\sigma^2\sqrt{2tr(\boldsymbol{\Sigma}_{11}^2)}} \xrightarrow{D} N(0,1) \tag{A.2}$$

hold under conditions (C1) and (C2) together with the local alternatives (2.7). Then, the proof is complete if we prove that $T_k = 2/(n(n-1))\sum_{i>j}\Delta_{i,j}(\mathbf{X}_1)K_k = o(n^{-1}\sqrt{tr(\boldsymbol{\Sigma}_{11}^2)})$, for $k = 1, 2, \ldots, 7$ under the conditions given in this theorem. In the following, we often simply write the constant coefficients with the order of n^{-k} as $O(n^{-k})$. Firstly, we may rewrite

$$\begin{aligned} T_1 = {} & O(n^{-2})\sum_{i>j}\Delta_{i,j}(\mathbf{X}_1)(\mathbf{X}_{0i} - \overline{\mathbf{X}}_0)^{\mathrm{T}}\boldsymbol{\delta}_{\beta_0}(\mathbf{X}_{1j} - \overline{\mathbf{X}}_1)^{\mathrm{T}}\boldsymbol{\delta}_{\beta_1} \\ & + O(n^{-2})\sum_{i>j}\Delta_{i,j}(\mathbf{X}_1)(\mathbf{X}_{0i} - \overline{\mathbf{X}}_0)^{\mathrm{T}}\boldsymbol{\delta}_{\beta_0}(\varepsilon_j - \overline{\varepsilon}) =: T_{11} + T_{12}. \end{aligned}$$

Then, by the expression in (A.1), we can write

$$\begin{aligned} & T_{11} \\ & = O(n^{-2})\sum_{i>j}\mathbf{X}_{1i}^{\mathrm{T}}\mathbf{X}_{1j}(\mathbf{X}_{1j} - \overline{\mathbf{X}}_1)^{\mathrm{T}}\boldsymbol{\delta}_{\beta_1}(\mathbf{X}_{0i} - \overline{\mathbf{X}}_0)^{\mathrm{T}}\boldsymbol{\delta}_{\beta_0} \\ & \quad + O(n^{-3})\sum_{i>j}[\mathbf{X}_{1i}^{\mathrm{T}}\mathbf{X}_{1i} + \mathbf{X}_{1j}^{\mathrm{T}}\mathbf{X}_{1j} - 2E(\mathbf{X}_{11}^{\mathrm{T}}\mathbf{X}_{11})](\mathbf{X}_{1j} - \overline{\mathbf{X}}_1)^{\mathrm{T}}\boldsymbol{\delta}_{\beta_1}(\mathbf{X}_{0i} - \overline{\mathbf{X}}_0)^{\mathrm{T}}\boldsymbol{\delta}_{\beta_0} \\ & \quad + O(n^{-2})\sum_{i>j}\overline{\mathbf{X}}_1^{(i,j)^{\mathrm{T}}}(\mathbf{X}_{1i} + \mathbf{X}_{1j})(\mathbf{X}_{1j} - \overline{\mathbf{X}}_1)^{\mathrm{T}}\boldsymbol{\delta}_{\beta_1}(\mathbf{X}_{0i} - \overline{\mathbf{X}}_0)^{\mathrm{T}}\boldsymbol{\delta}_{\beta_0} \\ & \quad + O(n^{-2})\sum_{i>j}\left[\overline{\mathbf{X}}_1^{(i,j)^{\mathrm{T}}}\overline{\mathbf{X}}_1^{(i,j)} - \frac{E(\mathbf{X}_{11}^{\mathrm{T}}\mathbf{X}_{11})}{(n-2)}\right](\mathbf{X}_{1j} - \overline{\mathbf{X}}_1)^{\mathrm{T}}\boldsymbol{\delta}_{\beta_1}(\mathbf{X}_{0i} - \overline{\mathbf{X}}_0)^{\mathrm{T}}\boldsymbol{\delta}_{\beta_0} \\ & =: T_{111} + T_{112} + T_{113} + T_{114}. \end{aligned}$$

Denote $T_{111}^{(1)} := O(n^{-2})\sum_{i>j}\mathbf{X}_{1i}^{\mathrm{T}}\mathbf{X}_{1j}\mathbf{X}_{1j}^{\mathrm{T}}\boldsymbol{\delta}_{\beta_1}\mathbf{X}_{0i}$, $T_{111}^{(2)} := O(n^{-2})\sum_{i>j}\mathbf{X}_{1i}^{\mathrm{T}}\mathbf{X}_{1j}\overline{\mathbf{X}}_1^{\mathrm{T}}\boldsymbol{\delta}_{\beta_1}\mathbf{X}_{0i}$, $T_{111}^{(3)} := O(n^{-2})\sum_{i>j}\mathbf{X}_{1i}^{\mathrm{T}}\mathbf{X}_{1j}\mathbf{X}_{1j}^{\mathrm{T}}\boldsymbol{\delta}_{\beta_1}\overline{\mathbf{X}}_0$ and $T_{111}^{(4)} := O(n^{-2})\sum_{i>j}\mathbf{X}_{1i}^{\mathrm{T}}\mathbf{X}_{1j}\overline{\mathbf{X}}_1^{\mathrm{T}}\boldsymbol{\delta}_{\beta_1}\overline{\mathbf{X}}_0$. Then, using Lemmas 1–2, we can obtain that

$$E[\|T_{111}^{(1)}\|^2] = O(1)E(\boldsymbol{\delta}_{\beta_1}^{\mathrm{T}}\mathbf{X}_{12}\mathbf{X}_{12}^{\mathrm{T}}\mathbf{X}_{11}\mathbf{X}_{01}^{\mathrm{T}}\mathbf{X}_{03}\mathbf{X}_{13}^{\mathrm{T}}\mathbf{X}_{14}\mathbf{X}_{14}^{\mathrm{T}}\boldsymbol{\delta}_{\beta_1})$$

$$
\begin{aligned}
&+O(n^{-1})E(\boldsymbol{\delta}_{\beta_1}^{\mathrm{T}}\mathbf{X}_{12}\mathbf{X}_{12}^{\mathrm{T}}\mathbf{X}_{11}\mathbf{X}_{01}^{\mathrm{T}}\mathbf{X}_{01}\mathbf{X}_{11}^{\mathrm{T}}\mathbf{X}_{13}\mathbf{X}_{13}^{\mathrm{T}}\boldsymbol{\delta}_{\beta_1})\\
&+O(n^{-1})E(\boldsymbol{\delta}_{\beta_1}^{\mathrm{T}}\mathbf{X}_{12}\mathbf{X}_{12}^{\mathrm{T}}\mathbf{X}_{11}\mathbf{X}_{01}^{\mathrm{T}}\mathbf{X}_{03}\mathbf{X}_{13}^{\mathrm{T}}\mathbf{X}_{12}\mathbf{X}_{12}^{\mathrm{T}}\boldsymbol{\delta}_{\beta_1})\\
&+O(n^{-1})E(\boldsymbol{\delta}_{\beta_1}^{\mathrm{T}}\mathbf{X}_{12}\mathbf{X}_{12}^{\mathrm{T}}\mathbf{X}_{11}\mathbf{X}_{01}^{\mathrm{T}}\mathbf{X}_{02}\mathbf{X}_{12}^{\mathrm{T}}\mathbf{X}_{13}\mathbf{X}_{13}^{\mathrm{T}}\boldsymbol{\delta}_{\beta_1})\\
&+O(n^{-2})E(\boldsymbol{\delta}_{\beta_1}^{\mathrm{T}}\mathbf{X}_{12}\mathbf{X}_{12}^{\mathrm{T}}\mathbf{X}_{11}\mathbf{X}_{01}^{\mathrm{T}}\mathbf{X}_{01}\mathbf{X}_{11}^{\mathrm{T}}\mathbf{X}_{12}\mathbf{X}_{12}^{\mathrm{T}}\boldsymbol{\delta}_{\beta_1})\\
\le{}& c_1B_3 + c_2\frac{q}{n}B_3 + c_3\frac{1}{n}tr(\boldsymbol{\Sigma}_{11}^2)B_1 + c_4\frac{\sqrt{q}}{n}B_3\\
&+c_5\frac{1}{n}\sqrt{qtr(\boldsymbol{\Sigma}_{11}^2)B_1B_3} + c_6\frac{q}{n^2}tr(\boldsymbol{\Sigma}_{11}^2)B_1,
\end{aligned}
$$

where the inequality follows by simple calculation and Cauchy-Schtwarz inequality. As for the term $T_{111}^{(2)}$, we have

$$
\begin{aligned}
&E\|T_{111}^{(2)}\|^2\\
&= E\|O(n^{-3})\sum_{i>j}[\mathbf{X}_{1i}^{\mathrm{T}}\mathbf{X}_{1j}\mathbf{X}_{1i}^{\mathrm{T}}\boldsymbol{\delta}_{\beta_1}\mathbf{X}_{0i} + \mathbf{X}_{1i}^{\mathrm{T}}\mathbf{X}_{1j}\mathbf{X}_{1j}^{\mathrm{T}}\boldsymbol{\delta}_{\beta_1}\mathbf{X}_{0i}\\
&\quad + \sum_{k\notin\{i,j\}}\mathbf{X}_{1i}^{\mathrm{T}}\mathbf{X}_{1j}\mathbf{X}_{1k}^{\mathrm{T}}\boldsymbol{\delta}_{\beta_1}\mathbf{X}_{0i}]\|^2\\
&\le O(n^{-2})E(\mathbf{X}_{01}^{\mathrm{T}}\mathbf{X}_{01})E(\boldsymbol{\delta}_{\beta_1}^{\mathrm{T}}\mathbf{X}_{11}\mathbf{X}_{11}^{\mathrm{T}}\mathbf{X}_{12}\mathbf{X}_{12}^{\mathrm{T}}\mathbf{X}_{11}\mathbf{X}_{11}^{\mathrm{T}}\boldsymbol{\delta}_{\beta_1})\\
&\quad + \big[O(n^{-1})E(\boldsymbol{\delta}_{\beta_1}^{\mathrm{T}}\mathbf{X}_{12}\mathbf{X}_{12}^{\mathrm{T}}\mathbf{X}_{11}\mathbf{X}_{01}^{\mathrm{T}}\mathbf{X}_{03}^{\mathrm{T}}\mathbf{X}_{13}^{\mathrm{T}}\mathbf{X}_{14}\mathbf{X}_{14}^{\mathrm{T}}\boldsymbol{\delta}_{\beta_1})\\
&\quad + O(n^{-2})E(\mathbf{X}_{01}^{\mathrm{T}}\mathbf{X}_{01})E(\boldsymbol{\delta}_{\beta_1}^{\mathrm{T}}\mathbf{X}_{11}\mathbf{X}_{11}^{\mathrm{T}}\mathbf{X}_{12}\mathbf{X}_{12}^{\mathrm{T}}\mathbf{X}_{11}\mathbf{X}_{11}^{\mathrm{T}}\boldsymbol{\delta}_{\beta_1})\big]\\
&\quad +O(n^{-2})E(\mathbf{X}_{01}^{\mathrm{T}}\mathbf{X}_{01})E(\boldsymbol{\delta}_{\beta_1}^{\mathrm{T}}\mathbf{X}_{13}\mathbf{X}_{12}^{\mathrm{T}}\mathbf{X}_{11}\mathbf{X}_{11}^{\mathrm{T}}\mathbf{X}_{12}\mathbf{X}_{13}^{\mathrm{T}}\boldsymbol{\delta}_{\beta_1})\\
&\le O(n^{-2})tr(\boldsymbol{\Sigma}_{00})(2B_3 + tr(\boldsymbol{\Sigma}_{11}^2)B_1) + O(n^{-1})B_3\\
&\le c_1\frac{q}{n^2}B_3 + c_2\frac{q}{n^2}B_1tr(\boldsymbol{\Sigma}_{11}^2) + c_3\frac{1}{n}B_3,
\end{aligned}
$$

With the same methods, similar results can be obtained for $T_{111}^{(k)}, k = 3, 4$. Combining with Lemma 3, $T_{111} = o_P(n^{-1}\sqrt{\mathrm{tr}(\boldsymbol{\Sigma}_{11}^2)})$ follows under the local alternatives (2.7). As for T_{112}, write $T_{112}^{(1)} := O(n^{-3})\sum_{i>j}[\mathbf{X}_{1i}^{\mathrm{T}}\mathbf{X}_{1i} + \mathbf{X}_{1j}^{\mathrm{T}}\mathbf{X}_{1j} - 2E(\mathbf{X}_{11}^{\mathrm{T}}\mathbf{X}_{11})]\mathbf{X}_{1j}^{\mathrm{T}}\boldsymbol{\delta}_{\beta_1}\mathbf{X}_{0i}$, $T_{112}^{(2)} := O(n^{-3})\sum_{i>j}[\mathbf{X}_{1i}^{\mathrm{T}}\mathbf{X}_{1i} + \mathbf{X}_{1j}^{\mathrm{T}}\mathbf{X}_{1j} - 2E(\mathbf{X}_{11}^{\mathrm{T}}\mathbf{X}_{11})]$ $\overline{\mathbf{X}}_1^{\mathrm{T}}\boldsymbol{\delta}_{\beta_1}\mathbf{X}_{0i}$, $T_{112}^{(3)} := O(n^{-3})\sum_{i>j}[\mathbf{X}_{1i}^{\mathrm{T}}\mathbf{X}_{1i} + \mathbf{X}_{1j}^{\mathrm{T}}\mathbf{X}_{1j} - 2E(\mathbf{X}_{11}^{\mathrm{T}}\mathbf{X}_{11})]\mathbf{X}_{1j}^{\mathrm{T}}\boldsymbol{\delta}_{\beta_1}\overline{\mathbf{X}}_0$ and $T_{112}^{(4)} := O(n^{-3})\sum_{i>j}[\mathbf{X}_{1i}^{\mathrm{T}}\mathbf{X}_{1i}+\mathbf{X}_{1j}^{\mathrm{T}}\mathbf{X}_{1j}-2E(\mathbf{X}_{11}^{\mathrm{T}}\mathbf{X}_{11})]\overline{\mathbf{X}}_1^{\mathrm{T}}\boldsymbol{\delta}_{\beta_1}\overline{\mathbf{X}}_0$. Then, we obtain that

$$
\begin{aligned}
E\|T_{112}^{(1)}\|^2 &\le E\|O(n^{-3})\sum_{i>j}[\mathbf{X}_{1i}^{\mathrm{T}}\mathbf{X}_{1i} + \mathbf{X}_{1j}^{\mathrm{T}}\mathbf{X}_{1j} - 2E(\mathbf{X}_{11}^{\mathrm{T}}\mathbf{X}_{11})]\mathbf{X}_{1j}^{\mathrm{T}}\boldsymbol{\delta}_{\beta_1}\mathbf{X}_{0i}\|^2\\
&\le O(n^{-2})Var(\mathbf{X}_{11}^{\mathrm{T}}\mathbf{X}_{11})E(\boldsymbol{\delta}_{\beta_1}^{\mathrm{T}}\mathbf{X}_{11}\mathbf{X}_{11}^{\mathrm{T}}\boldsymbol{\delta}_{\beta_1}\mathbf{X}_{02}^{\mathrm{T}}\mathbf{X}_{02})\\
&= O(n^{-2})tr(\boldsymbol{\Sigma}_{00})tr(\boldsymbol{\Sigma}_{11}^2)B_1 = o(n^{-2}tr(\boldsymbol{\Sigma}_{11}^2)),
\end{aligned}
$$

$$
\begin{aligned}
E\|T_{112}^{(2)}\| &\le O(n^{-1})\left[Var(\mathbf{X}_{11}^{\mathrm{T}}\mathbf{X}_{11})E(\boldsymbol{\delta}_{\beta_1}^{\mathrm{T}}\overline{\mathbf{X}}_1\overline{\mathbf{X}}_1^{\mathrm{T}}\boldsymbol{\delta}_{\beta_1}\mathbf{X}_{01}^{\mathrm{T}}\mathbf{X}_{01})\right]^{1/2}\\
&= O(n^{-3/2})\left[tr(\boldsymbol{\Sigma}_{00})tr(\boldsymbol{\Sigma}_{11}^2)B_1 + O(n^{-1})tr(\boldsymbol{\Sigma}_{11}^2)B_1\right]^{1/2}\\
&= o(n^{-1}\sqrt{tr(\boldsymbol{\Sigma}_{11}^2)}).
\end{aligned}
$$

Similar results can be obtained for $T_{112}^{(k)}, k = 3, 4$. Then using Lemma 3, it is obtained that $T_{112} = o_P(n^{-1}\sqrt{tr(\boldsymbol{\Sigma}_{11}^2)})$ under the local alternatives (2.7). Similarly, for T_{113}, denote $T_{113}^{(1)} := O(n^{-2})\sum_{i>j}\overline{\mathbf{X}}_1^{(i,j)^{\mathrm{T}}}(\mathbf{X}_{1i}+\mathbf{X}_{1j})\mathbf{X}_{1j}^{\mathrm{T}}\boldsymbol{\delta}_{\beta_1}\mathbf{X}_{0i}$, $T_{113}^{(2)} := O(n^{-2})\sum_{i>j}\overline{\mathbf{X}}_1^{(i,j)^{\mathrm{T}}}(\mathbf{X}_{1i}+\mathbf{X}_{1j})\overline{\mathbf{X}}_1^{\mathrm{T}}\boldsymbol{\delta}_{\beta_1}\mathbf{X}_{0i}$, $T_{113}^{(3)} := O(n^{-2})\sum_{i>j}\overline{\mathbf{X}}_1^{(i,j)^{\mathrm{T}}}(\mathbf{X}_{1i}+\mathbf{X}_{1j})\mathbf{X}_{1j}^{\mathrm{T}}\boldsymbol{\delta}_{\beta_1}\overline{\mathbf{X}}_0$, and $T_{113}^{(4)} := O(n^{-2})\sum_{i>j}\overline{\mathbf{X}}_1^{(i,j)^{\mathrm{T}}}(\mathbf{X}_{1i}+\mathbf{X}_{1j})\overline{\mathbf{X}}_1^{\mathrm{T}}\boldsymbol{\delta}_{\beta_1}\overline{\mathbf{X}}_0$. Then, calculating the expectations of $\|T_{113}^{(k)}\|$ or $\|T_{113}^{(k)}\|^2$, we have

$$
\begin{aligned}
E\|T_{113}^{(1)}\|^2 &= E\|O(n^{-2})\sum_{i>j}(\overline{\mathbf{X}}_1^{(i,j)^{\mathrm{T}}}\mathbf{X}_{1i}\mathbf{X}_{1j}^{\mathrm{T}}\boldsymbol{\delta}_{\beta_1}\mathbf{X}_{0i} + \overline{\mathbf{X}}_1^{(i,j)^{\mathrm{T}}}\mathbf{X}_{1j}\mathbf{X}_{1j}^{\mathrm{T}}\boldsymbol{\delta}_{\beta_1}\mathbf{X}_{0i})\|^2\\
&\le O(n^{-1})[E\|\overline{\mathbf{X}}_1^{(1,2)^{\mathrm{T}}}\mathbf{X}_{11}\mathbf{X}_{12}^{\mathrm{T}}\boldsymbol{\delta}_{\beta_1}\mathbf{X}_{01}\|^2 + E\|\overline{\mathbf{X}}_1^{(1,2)^{\mathrm{T}}}\mathbf{X}_{12}\mathbf{X}_{12}^{\mathrm{T}}\boldsymbol{\delta}_{\beta_1}\mathbf{X}_{01}\|^2]\\
&\le 2O(n^{-1})E(\overline{\mathbf{X}}_1^{(1,2)^{\mathrm{T}}}\mathbf{X}_{11}\mathbf{X}_{11}^{\mathrm{T}}\overline{\mathbf{X}}_1^{(1,2)})E(\boldsymbol{\delta}_{\beta_1}^{\mathrm{T}}\mathbf{X}_{12}\mathbf{X}_{12}^{\mathrm{T}}\boldsymbol{\delta}_{\beta_1}\mathbf{X}_{01}^{\mathrm{T}}\mathbf{X}_{01})\\
&= O(n^{-2})tr(\boldsymbol{\Sigma}_{00})tr(\boldsymbol{\Sigma}_{11}^2)B_1 = o(n^{-2}tr(\boldsymbol{\Sigma}_{11}^2)),
\end{aligned}
$$

under the local alternatives (2.7). Rewrite

$$
\begin{aligned}
T_{113}^{(2)} &= O(n^{-3})\sum_{i>j}\overline{\mathbf{X}}_1^{(i,j)^{\mathrm{T}}}(\mathbf{X}_{1i}+\mathbf{X}_{1j})(\mathbf{X}_{1i}+\mathbf{X}_{1j})^{\mathrm{T}}\boldsymbol{\delta}_{\beta_1}\mathbf{X}_{0i}\\
&\quad + O(n^{-3})\sum_{i>j}\overline{\mathbf{X}}_1^{(i,j)^{\mathrm{T}}}\mathbf{X}_{1i}\left(\sum_{k\notin\{i,j\}}\mathbf{X}_{1k}\right)^{\mathrm{T}}\boldsymbol{\delta}_{\beta_1}\mathbf{X}_{0i}\\
&\quad + O(n^{-3})\sum_{i>j}\overline{\mathbf{X}}_1^{(i,j)^{\mathrm{T}}}\mathbf{X}_{1j}\left(\sum_{k\notin\{i,j\}}\mathbf{X}_{1k}\right)^{\mathrm{T}}\boldsymbol{\delta}_{\beta_1}\mathbf{X}_{0i},
\end{aligned}
$$

then we obtain that

$$
\begin{aligned}
E\|T_{113}^{(2)}\|^2 &\le O(n^{-2})tr(\boldsymbol{\Sigma}_{11}^2)B_1 + O(n^{-2})tr(\boldsymbol{\Sigma}_{00})tr(\boldsymbol{\Sigma}_{11}^2)B_1\\
&\quad + O(n^{-2})B_3 + O(n^{-2})tr(\boldsymbol{\Sigma}_{11}^2)B_1 + O(n^{-2})tr(\boldsymbol{\Sigma}_{00})tr(\boldsymbol{\Sigma}_{11}^2)B_1\\
&\quad + O(n^{-3})tr(\boldsymbol{\Sigma}_{00})tr(\boldsymbol{\Sigma}_{11}^2)B_1 + O(n^{-4})tr(\boldsymbol{\Sigma}_{00})B_3 = o(n^{-2}tr(\boldsymbol{\Sigma}_{11}^2)).
\end{aligned}
$$

Similar results can be obtained for $T_{113}^{(k)}, k = 3, 4$. Thus, we have $T_{113} = o(n^{-1}$

$\sqrt{tr(\mathbf{\Sigma}_{11}^2)})$. As for T_{114}, write

$$T_{114}^{(1)} = O(n^{-2})\sum_{i>j}\left(\overline{\mathbf{X}}_1^{(i,j)\mathrm{T}}\overline{\mathbf{X}}_1^{(i,j)} - \frac{E(\mathbf{X}_{11}^{\mathrm{T}}\mathbf{X}_{11})}{(n-2)}\right)\mathbf{X}_{1j}^{\mathrm{T}}\boldsymbol{\delta}_{\beta_1}\mathbf{X}_{0i},$$
$$T_{114}^{(2)} = O(n^{-2})\sum_{i>j}\left(\overline{\mathbf{X}}_1^{(i,j)\mathrm{T}}\overline{\mathbf{X}}_1^{(i,j)} - \frac{E(\mathbf{X}_{11}^{\mathrm{T}}\mathbf{X}_{11})}{(n-2)}\right)\overline{\mathbf{X}}_1^{\mathrm{T}}\boldsymbol{\delta}_{\beta_1}\mathbf{X}_{0i},$$
$$T_{114}^{(3)} = O(n^{-2})\sum_{i>j}\left(\overline{\mathbf{X}}_1^{(i,j)\mathrm{T}}\overline{\mathbf{X}}_1^{(i,j)} - \frac{E(\mathbf{X}_{11}^{\mathrm{T}}\mathbf{X}_{11})}{(n-2)}\right)\mathbf{X}_{1j}^{\mathrm{T}}\boldsymbol{\delta}_{\beta_1}\overline{\mathbf{X}}_0,$$
$$T_{114}^{(4)} = O(n^{-2})\sum_{i>j}\left(\overline{\mathbf{X}}_1^{(i,j)\mathrm{T}}\overline{\mathbf{X}}_1^{(i,j)} - \frac{E(\mathbf{X}_{11}^{\mathrm{T}}\mathbf{X}_{11})}{(n-2)}\right)\overline{\mathbf{X}}_1^{\mathrm{T}}\boldsymbol{\delta}_{\beta_1}\overline{\mathbf{X}}_0$$

Then, we have

$$\begin{aligned}E\|T_{114}^{(1)}\| &\le [Var(\overline{\mathbf{X}}_1^{(1,2)\mathrm{T}}\overline{\mathbf{X}}_1^{(1,2)})E(\boldsymbol{\delta}_{\beta_1}^{\mathrm{T}}\mathbf{X}_{12}\mathbf{X}_{12}^{\mathrm{T}}\boldsymbol{\delta}_{\beta_1}\mathbf{X}_{01}^{\mathrm{T}}\mathbf{X}_{01})]^{1/2}\\ &= [O(n^{-2})tr(\mathbf{\Sigma}_{00})tr(\mathbf{\Sigma}_{11}^2)B_1]^{1/2} = o\left(n^{-1}\sqrt{tr(\mathbf{\Sigma}_{11}^2)}\right),\end{aligned}$$

and $E\|T_{114}^{(k)}\| = o(n^{-1}\sqrt{tr(\mathbf{\Sigma}_{11}^2)})$, for $k = 2,3,4$. Thus, we have $T_{114} = o_P(n^{-1}\sqrt{tr(\mathbf{\Sigma}_{11}^2)})$. For the term T_{12}, write

$$T_{121} = O(n^{-2})\sum_{i>j}\mathbf{X}_{1i}^{\mathrm{T}}\mathbf{X}_{1j}(\mathbf{X}_{0i}-\overline{\mathbf{X}}_0)^{\mathrm{T}}\boldsymbol{\delta}_{\beta_0}(\varepsilon_j-\overline{\varepsilon}),$$
$$T_{122} = O(n^{-3})\sum_{i>j}[\mathbf{X}_{1i}^{\mathrm{T}}\mathbf{X}_{1i}+\mathbf{X}_{1j}^{\mathrm{T}}\mathbf{X}_{1j}-2E(\mathbf{X}_{11}^{\mathrm{T}}\mathbf{X}_{11})](\mathbf{X}_{0i}-\overline{\mathbf{X}}_0)^{\mathrm{T}}\boldsymbol{\delta}_{\beta_0}(\varepsilon_j-\overline{\varepsilon}),$$
$$T_{123} = O(n^{-2})\sum_{i>j}\overline{\mathbf{X}}_1^{(i,j)\mathrm{T}}(\mathbf{X}_{1i}+\mathbf{X}_{1j})(\mathbf{X}_{0i}-\overline{\mathbf{X}}_0)^{\mathrm{T}}\boldsymbol{\delta}_{\beta_0}(\varepsilon_j-\overline{\varepsilon}),$$
$$T_{124} = O(n^{-2})\sum_{i>j}\left(\overline{\mathbf{X}}_1^{(i,j)\mathrm{T}}\overline{\mathbf{X}}_1^{(i,j)} - \frac{E(\mathbf{X}_{11}^{\mathrm{T}}\mathbf{X}_{11})}{(n-2)}\right)(\mathbf{X}_{0i}-\overline{\mathbf{X}}_0)^{\mathrm{T}}\boldsymbol{\delta}_{\beta_0}(\varepsilon_j-\overline{\varepsilon})$$

Re-formulate T_{121} as $T_{121}^{(1)} := O(n^{-2})\sum_{i>j}\mathbf{X}_{1i}^{\mathrm{T}}\mathbf{X}_{1j}\varepsilon_j\mathbf{X}_{0i}$, $T_{121}^{(2)} := O(n^{-2})\sum_{i>j}\mathbf{X}_{1i}^{\mathrm{T}}\mathbf{X}_{1j}\overline{\varepsilon}\mathbf{X}_{0i}$, $T_{121}^{(3)} := O(n^{-2})\sum_{i>j}\mathbf{X}_{1i}^{\mathrm{T}}\mathbf{X}_{1j}\varepsilon_j\overline{\mathbf{X}}_0$, and $T_{121}^{(4)} := O(n^{-2})\sum_{i>j}\mathbf{X}_{1i}^{\mathrm{T}}\mathbf{X}_{1j}\overline{\varepsilon}\overline{\mathbf{X}}_0$. Then, we have

$$\begin{aligned}E\|T_{121}^{(1)}\|^2 &= O(n^{-4})\sum_{i>j}\sum_{k>l}E(\varepsilon_j\varepsilon_l\mathbf{X}_{1i}^{\mathrm{T}}\mathbf{X}_{1j}\mathbf{X}_{0i}^{\mathrm{T}}\mathbf{X}_{0k}\mathbf{X}_{1k}^{\mathrm{T}}\mathbf{X}_{1l})\\ &\le O(n^{-2})(q\mathrm{tr}(\mathbf{\Sigma}_{11}^2)+\mathrm{tr}(\mathbf{\Sigma}_{11}^2)) + O(n^{-1})\mathrm{tr}(\mathbf{\Gamma}_1^{\mathrm{T}}\mathbf{\Gamma}_1\mathbf{\Gamma}_0^{\mathrm{T}}\mathbf{\Gamma}_0\mathbf{\Gamma}_1^{\mathrm{T}}\mathbf{\Gamma}_1)\end{aligned}$$

$$E\|T_{121}^{(2)}\|^2 = O(n^{-4})\sum_{i>j}\sum_{k>l} E(\mathbf{X}_{1i}^{\mathrm{T}}\mathbf{X}_{1j}\bar{\varepsilon}^2\mathbf{X}_{0i}^{\mathrm{T}}\mathbf{X}_{0k}\mathbf{X}_{1k}^{\mathrm{T}}\mathbf{X}_{1l})$$
$$\le O(n^{-2})\mathrm{tr}(\mathbf{\Sigma}_{11}^2) + O(n^{-3})q\mathrm{tr}(\mathbf{\Sigma}_{11}^2),$$

and also $E\|T_{121}^{(k)}\|^2 = o(n^{-1-\kappa}tr(\mathbf{\Sigma}_{11}^2))$, for $k = 3, 4$. Thus, using Lemma 3, we have $T_{121} = o_P(n^{-1}\sqrt{tr(\mathbf{\Sigma}_{11}^2)})$ under conditions (C3)–(C4). For T_{122}, write $T_{122}^{(1)} = O(n^{-3})\sum_{i>j}[\mathbf{X}_{1i}^{\mathrm{T}}\mathbf{X}_{1i} + \mathbf{X}_{1j}^{\mathrm{T}}\mathbf{X}_{1j} - 2E(\mathbf{X}_{11}^{\mathrm{T}}\mathbf{X}_{11})](\varepsilon_j - \bar{\varepsilon})\mathbf{X}_{0i}$ and $T_{122}^{(2)} = O(n^{-3})\sum_{i>j}[\mathbf{X}_{1i}^{\mathrm{T}}\mathbf{X}_{1i} + \mathbf{X}_{1j}^{\mathrm{T}}\mathbf{X}_{1j} - 2E(\mathbf{X}_{11}^{\mathrm{T}}\mathbf{X}_{11})](\varepsilon_j - \bar{\varepsilon})\overline{\mathbf{X}}_0$. Then using lemmas 1–2, we have

$$E\|T_{122}^{(1)}\| \le O(n^{-1})[Var(\mathbf{X}_{11}^{\mathrm{T}}\mathbf{X}_{11})E(\mathbf{X}_{01}^{\mathrm{T}}\mathbf{X}_{01})]^{1/2} \le O(n^{-1/2})\sqrt{\frac{qtr(\mathbf{\Sigma}_{11}^2)}{n}},$$
$$E\|T_{122}^{(2)}\| \le O(n^{-1})\sqrt{\frac{qtr(\mathbf{\Sigma}_{11}^2)}{n}} = o\left(n^{-1}\sqrt{tr(\mathbf{\Sigma}_{11}^2)}\right).$$

Therefore, we obtain that $T_{122} = o_P(n^{-1}\sqrt{tr(\mathbf{\Sigma}_{11}^2)})$. Similarly, for the term T_{123}, denote $T_{123}^{(1)} = O(n^{-2})\sum_{i>j}\overline{\mathbf{X}}_1^{(i,j)^{\mathrm{T}}}(\mathbf{X}_{1i}+\mathbf{X}_{1j})\varepsilon_j\mathbf{X}_{0i}$, $T_{123}^{(2)} = O(n^{-2})\sum_{i>j}\overline{\mathbf{X}}_1^{(i,j)^{\mathrm{T}}}$ $(\mathbf{X}_{1i}+\mathbf{X}_{1j})\bar{\varepsilon}\mathbf{X}_{0i}$, and $T_{123}^{(3)} = O(n^{-2})\sum_{i>j}\overline{\mathbf{X}}_1^{(i,j)^{\mathrm{T}}}(\mathbf{X}_{1i}+\mathbf{X}_{1j})\overline{\mathbf{X}}_0^{\mathrm{T}}(\varepsilon_j - \bar{\varepsilon})$. Then, under conditions given in this theorem, we obtain that

$$E\|T_{123}^{(1)}\|^2 = O(n^{-4})\sum_{i>j}\sum_{k>l}E[\varepsilon_j\varepsilon_l\overline{\mathbf{X}}_1^{(i,j)^{\mathrm{T}}}(\mathbf{X}_{1i}+\mathbf{X}_{1j})\mathbf{X}_{0i}^{\mathrm{T}}\mathbf{X}_{0k}\overline{\mathbf{X}}_1^{(k,l)^{\mathrm{T}}}(\mathbf{X}_{1k}+\mathbf{X}_{1l})]$$
$$\le O(n^{-1})E[\overline{\mathbf{X}}_1^{(1,2)^{\mathrm{T}}}(\mathbf{X}_{11}+\mathbf{X}_{12})\mathbf{X}_{01}^{\mathrm{T}}\mathbf{X}_{01}\overline{\mathbf{X}}_1^{(1,2)^{\mathrm{T}}}(\mathbf{X}_{11}+\mathbf{X}_{12})]$$
$$\le O(n^{-2})[tr(\mathbf{\Sigma}_{11}^2) + qtr(\mathbf{\Sigma}_{11}^2)],$$
$$E\|T_{123}^{(2)}\| \le O(n^{-1/2})[E(\overline{\mathbf{X}}_1^{(1,2)^{\mathrm{T}}}(\mathbf{X}_{11}+\mathbf{X}_{12})(\mathbf{X}_{11}+\mathbf{X}_{12})^{\mathrm{T}}\overline{\mathbf{X}}_1^{(1,2)})E(\mathbf{X}_{01}^{\mathrm{T}}\mathbf{X}_{01})]^{1/2}$$
$$\le O(n^{-1/2})\left[\frac{qtr(\mathbf{\Sigma}_{11}^2)}{n}\right]^{1/2},$$
$$E(\|T_{123}^{(3)}\|) \le O(1)[E(\overline{\mathbf{X}}_1^{(1,2)^{\mathrm{T}}}(\mathbf{X}_{11}+\mathbf{X}_{12})(\mathbf{X}_{11}+\mathbf{X}_{12})^{\mathrm{T}}\overline{\mathbf{X}}_1^{(1,2)})E(\overline{\mathbf{X}}_0^{\mathrm{T}}\overline{\mathbf{X}}_0)]^{1/2}$$
$$\le O(n^{-1/2})\left[\frac{qtr(\mathbf{\Sigma}_{11}^2)}{n}\right]^{1/2}.$$

Thus, by the definition of T_{123}, condition (C3) and Lemma 3, $T_{123} = o_P(n^{-1}$ $\sqrt{tr(\mathbf{\Sigma}_{11}^2)})$ follows. Denote

$$T_{124}^{*} = O(n^{-2})\sum_{i>j}\left(\overline{\mathbf{X}}_1^{(i,j)^{\mathrm{T}}}\overline{\mathbf{X}}_1^{(i,j)} - \frac{E(\mathbf{X}_{11}^{\mathrm{T}}\mathbf{X}_{11})}{(n-2)}\right)(\varepsilon_j - \bar{\varepsilon})(\mathbf{X}_{0i} - \overline{\mathbf{X}}_0),$$

Then calculate the expectation of the absolute value of T_{124},

$$
\begin{aligned}
E\|T^*_{124}\| &\le O(1)[Var(\overline{\mathbf{X}}_1^{(1,2)^{\mathrm{T}}}\overline{\mathbf{X}}_1^{(1,2)})E(\mathbf{X}_{0i}-\overline{\mathbf{X}}_0)^{\mathrm{T}}(\mathbf{X}_{0i}-\overline{\mathbf{X}}_0)]^{1/2}\\
&\le O\left(\sqrt{qn^{-2}tr(\mathbf{\Sigma}_{11}^2)}\right).
\end{aligned}
$$

Then $T_{124}=o_P(n^{-1}\sqrt{tr(\mathbf{\Sigma}_{11}^2)})$ follows by Lemma 3.

Write $T^*_3 = O(n^{-2})\sum_{i>j}\Delta_{i,j}(\mathbf{X}_1)(\mathbf{X}_{0i}-\overline{\mathbf{X}}_0)(\mathbf{X}_{0j}-\overline{\mathbf{X}}_0)^{\mathrm{T}}$. Write $T^*_{31} = O(n^{-2})\sum_{i>j}\Delta_{i,j}(\mathbf{X}_1)\mathbf{X}_{0i}\mathbf{X}_{0j}^{\mathrm{T}}$, $T^*_{32} = O(n^{-2})\sum_{i>j}\Delta_{i,j}(\mathbf{X}_1)\overline{\mathbf{X}}_0\mathbf{X}_{0j}^{\mathrm{T}}$ and $T^*_{33} = O(n^{-2})\sum_{i>j}\Delta_{i,j}(\mathbf{X}_1)\overline{\mathbf{X}}_0\overline{\mathbf{X}}_0^{\mathrm{T}}$. For T^*_{31}, reconstruct it as

$$
\begin{aligned}
T^*_{311} &= O(n^{-2})\sum_{i>j}\mathbf{X}_{1i}^{\mathrm{T}}\mathbf{X}_{1j}\mathbf{X}_{0i}\mathbf{X}_{0j}^{\mathrm{T}},\\
T^*_{312} &= O(n^{-3})\sum_{i>j}[\mathbf{X}_{1i}^{\mathrm{T}}\mathbf{X}_{1j}+\mathbf{X}_{1j}^{\mathrm{T}}\mathbf{X}_{1j}-E(\mathbf{X}_{11}^{\mathrm{T}}\mathbf{X}_{11})]\mathbf{X}_{0i}\mathbf{X}_{0j}^{\mathrm{T}},\\
T^*_{313} &= O(n^{-2})\sum_{i>j}\overline{\mathbf{X}}_1^{(i,j)^{\mathrm{T}}}(\mathbf{X}_{1i}+\mathbf{X}_{1j})\mathbf{X}_{0i}\mathbf{X}_{0j}^{\mathrm{T}},\\
T^*_{314} &= O(n^{-2})\sum_{i>j}\left[\overline{\mathbf{X}}_1^{(i,j)^{\mathrm{T}}}\overline{\mathbf{X}}_1^{(i,j)}-\frac{E(\mathbf{X}_{11}^{\mathrm{T}}\mathbf{X}_{11})}{(n-2)}\right]\mathbf{X}_{0i}\mathbf{X}_{0j}^{\mathrm{T}}.
\end{aligned}
$$

Then using Lemmas 1–2, we obtain that

$$
\begin{aligned}
E\|T^*_{311}\|^2 &= O(n^{-4})\sum_{i>j}\sum_{k>l}E(\mathbf{X}_{1i}^{\mathrm{T}}\mathbf{X}_{1j}\mathbf{X}_{0j}^{\mathrm{T}}\mathbf{X}_{0k}\mathbf{X}_{1k}^{\mathrm{T}}\mathbf{X}_{1l}\mathbf{X}_{0l}^{\mathrm{T}}\mathbf{X}_{0i})\\
&\le O(1)\mathrm{tr}[(\Sigma_{10}\mathbf{\Sigma}_{01})^2]+O(n^{-1})\mathrm{tr}(\mathbf{\Sigma}_{00})\mathrm{tr}(\mathbf{\Gamma}_1^{\mathrm{T}}\mathbf{\Sigma}_{10}\mathbf{\Sigma}_{01}\mathbf{\Gamma}_1),\\
E\|T^*_{312}\| &= O(n^{-1})[Var(\mathbf{X}_{11}^{\mathrm{T}}\mathbf{X}_{11})\mathrm{tr}(\mathbf{\Sigma}_{00}^2)]^{1/2}\\
&\le O(n^{-1})[\mathrm{tr}(\mathbf{\Sigma}_{00}^2)]^{1/2}[\mathrm{tr}(\mathbf{\Sigma}_{11}^2)]^{1/2}\\
E\|T^*_{313}\| &\le O(n^{-1/2})[\mathrm{tr}(\mathbf{\Sigma}_{11}^2)\mathrm{tr}(\mathbf{\Sigma}_{00}^2)]^{1/2},\\
E\|T^*_{314}\| &\le O(n^{-1})[\mathrm{tr}(\mathbf{\Sigma}_{11}^2)\mathrm{tr}(\mathbf{\Sigma}_{00}^2)]^{1/2}.
\end{aligned}
$$

By condition (C3) and Lemma 3, we have $T_{31}=O_P(n^{-1}\sqrt{tr(\mathbf{\Sigma}_{11}^2)})$.

$$
E\|T^*_{32}\| \le O(1)[Var(\Delta_{1,2}(\mathbf{X}_1))E(\overline{\mathbf{X}}_0^{\mathrm{T}}\mathbf{X}_{01}\mathbf{X}_{01}^{\mathrm{T}}\overline{\mathbf{X}}_0)]^{1/2} = O\left(\frac{1}{n}\mathrm{tr}(\mathbf{\Sigma}_{11}^2)\mathrm{tr}(\mathbf{\Sigma}_{00}^2)\right)
$$

where $Var(\Delta_{1,2}(\mathbf{X}_1))=O(tr(\mathbf{\Sigma}_{11}^2))$. Furthermore, we have

$$
\begin{aligned}
E\|T^*_{33}\| &\le O(1)[Var(\Delta_{1,2}(\mathbf{X}_1))E(\overline{\mathbf{X}}_0^{\mathrm{T}}\overline{\mathbf{X}}_0\overline{\mathbf{X}}_0^{\mathrm{T}}\overline{\mathbf{X}}_0)]^{1/2}\\
&\le [O(n^{-2})(\mathrm{tr}(\mathbf{\Sigma}_{00}^2)+\mathrm{tr}^2(\mathbf{\Sigma}_{00}))\mathrm{tr}(\mathbf{\Sigma}_{11}^2)]^{1/2}.
\end{aligned}
$$

Using Lemma 3 and condition (C3), $T_3 = o_P(n^{-1}\sqrt{tr(\mathbf{\Sigma}_{11}^2)})$ is true. By similar analysis, write $T_4^* = O(n^{-3})\sum_{i>j}\Delta_{i,j}(\mathbf{X}_1)(H_i - \overline{H})(\mathbf{X}_{0i} - \overline{\mathbf{X}}_0)$ and rewrite is with $T_{41}^* = O(n^{-3})\sum_{i>j}\Delta_{i,j}(\mathbf{X}_1)(\mathbf{X}_{1i} - \overline{\mathbf{X}}_1)^{\mathrm{T}}\boldsymbol{\delta}_{\beta_1}(\mathbf{X}_{0i} - \overline{\mathbf{X}}_0)$ and $T_{42}^* = O(n^{-3})\sum_{i>j}$ $\Delta_{i,j}(\mathbf{X}_1)(\varepsilon_i - \overline{\varepsilon})(\mathbf{X}_{0i} - \overline{\mathbf{X}}_0)$. Using Lemmas 1–2, we obtain that

$$\begin{aligned} E\|T_{41}^*\| &\le O(n^{-1})\{Var(\Delta_{1,2}(\mathbf{X}_1))E[\boldsymbol{\delta}_{\beta_1}^{\mathrm{T}}(\mathbf{X}_{11} - \overline{\mathbf{X}}_1)(\mathbf{X}_{11} - \overline{\mathbf{X}}_1)^{\mathrm{T}} \\ &\quad \times\boldsymbol{\delta}_{\beta_1}(\mathbf{X}_{01} - \overline{\mathbf{X}}_0)^{\mathrm{T}}(\mathbf{X}_{01} - \overline{\mathbf{X}}_0)]\}^{1/2} \\ &\le O(n^{-1})[tr(\mathbf{\Sigma}_{00})tr(\mathbf{\Sigma}_{11}^2)B_1]^{1/2} \end{aligned}$$

and

$$\begin{aligned} E\|T_{42}^*\| &\le O(n^{-1})[Var(\Delta_{1,2}(\mathbf{X}_1))E(\mathbf{X}_{01} - \overline{\mathbf{X}}_0)^{\mathrm{T}}(\mathbf{X}_{01} - \overline{\mathbf{X}}_0)]^{1/2} \\ &\le O(n^{-1})[tr(\mathbf{\Sigma}_{00})tr(\mathbf{\Sigma}_{11}^2)]^{1/2}. \end{aligned}$$

Combine the last two results, we have $T_4 = o_P(n^{-1}\sqrt{tr(\mathbf{\Sigma}_{11}^2)})$ under condition (C3) and the local alternatives (2.7). For T_6, write $T_6^* = O(n^{-3})\sum_{i>j}\Delta_{i,j}(\mathbf{X}_1)$ $(\mathbf{X}_{0i} - \overline{\mathbf{X}}_0)(\mathbf{X}_{0i} - \overline{\mathbf{X}}_0)^{\mathrm{T}}$, and then calculate the expectation of the absolute value

$$\begin{aligned} &E\|T_6^*\| \\ &\le O(n^{-1})[Var(\Delta_{i,j}(\mathbf{X}_1))E(\mathbf{X}_{0i} - \overline{\mathbf{X}}_0)^{\mathrm{T}}(\mathbf{X}_{0i} - \overline{\mathbf{X}}_0)(\mathbf{X}_{0i} - \overline{\mathbf{X}}_0)^{\mathrm{T}}(\mathbf{X}_{0i} - \overline{\mathbf{X}}_0)]^{1/2} \\ &\le O(n^{-1})[(\mathrm{tr}(\mathbf{\Sigma}_{00}^2) + tr(\mathbf{\Sigma}_{00})^2)]^{1/2}[\mathrm{tr}(\mathbf{\Sigma}_{11}^2)]^{1/2} \le O(n^{-1+\kappa}[\mathrm{tr}(\mathbf{\Sigma}_{11}^2)]^{1/2}). \end{aligned}$$

Then by Lemma 3, $T_6 = o_P(n^{-1}\sqrt{tr(\mathbf{\Sigma}_{11}^2)})$ follows. Finally, notice that the analysis of the terms T_5 and T_7 are quite similar to that of T_4 and T_6 respectively. This completes the proof.

Proof of Proposition 1. As discussed in Meinshausen, Meier and Buhlmann (2009), we also omit the function $min\{1,\cdot\}$ from the definition of $Q(\gamma)$ and Q^*. Then it is sufficient to show that

$$P\left\{(1 - \log\gamma_{min})\inf_{\gamma\in(\gamma_{min},1)} Q(\gamma) \le \alpha\right\} \le \alpha.$$

Define $\pi(u)$ as the fraction of samples of p_k satisfying $p_k \le u$, that is $\pi(u) = m^{-1}\sum_{k=1}^m I(p_k \le u)$. Then, the two events $\{Q(\gamma) \le \alpha\}$ and $\pi(\alpha\gamma) \ge \gamma$ are equivalent. Therefore,

$$P(Q(\gamma) \le \alpha) = P(\pi(\alpha\gamma) \ge \gamma) = P\left\{m^{-1}\sum_{k=1}^m I(p_k \le \alpha\gamma) \ge \gamma\right\}$$

$$\leq (\gamma m)^{-1} \sum_{k=1}^{m} P(p_k \leq \alpha\gamma),$$

where the last inequality is applied by Markov's inequality. Using the fact that the obtained p-values p_k's follow a uniform distribution conditional under the null hypothesis H_0, we have $P(p_k \leq \alpha\gamma | H_0) = \alpha\gamma$, which implies that $P(Q(\gamma) \leq \alpha | H_0) \leq \alpha$.

Since p_k's follow a uniform distribution under the null hypothesis H_0,

$$E\left\{ \sup_{\gamma \in (\gamma_{min}, 1)} \gamma^{-1} I(p_k \leq \alpha\gamma) \right\} = \int_0^{\alpha\gamma_{min}} \gamma_{min}^{-1} du + \int_{\alpha\gamma_{min}}^{\alpha} \frac{\alpha}{u} du = \alpha(1 - \log\gamma_{min}).$$

Again using Markov's inequality,

$$E\left(\sup_{\gamma \in (\gamma_{min}, 1)} I(\pi(\alpha\gamma) \geq \gamma) \right) = E\left(\sup_{\gamma \in (\gamma_{min}, 1)} I\left(m^{-1} \sum_{k=1}^{m} I(p_k \leq \alpha\gamma) \geq \gamma \right) \right) \leq \alpha(1 - \log\gamma_{min}).$$

It implies that $P(\inf_{\gamma \in (\gamma_{min}, 1)} Q(\gamma) \leq \alpha) \leq \alpha(1 - \log\gamma_{min})$ holds. by replacing $\alpha(1 - \log\gamma_{min})$ with α, we obtain $\limsup_{n \to \infty} P(Q^* \leq \alpha | H_0) \leq \alpha$. This completes the proof.

References

Barut, E., Fan, J. and Verhasselt, A. (2016). Conditional sure independence screening. *Journal of the American Statistical Association* **111**, 1266–1277.

Cai, T., Liu, W. and Xia, Y. (2014). Two-sample test of high dimensional means under dependence. *Journal of the Royal Statistical Society, Series B (Statistical Methodological)* **76**, 349–372.

Chen, S. X., Li, J. and Zhong, P. S. (2019). Two-sample and ANOVA tests for high dimensional means. *The Annals of Statistics* **47**, 1443–1474.

Cui, H., Guo, W. and Zhong, W. (2018). Test for high dimensional regression coefficients using refitted cross-validation variance estimation. *The Annals of Statistics* **46**, 958–988.

Efron, B. and Tibshirani, R. (2007). On testing the significance of sets of genes. *Annals of Applied Statistics* **1**, 107–129.

Geoman, J. J., Van de Geer, S. and Van Houwelingen, J. C. (2006). Testing against a high dimensional alternative. *Journal of the Royal Statistical Sociey, Series B (Statistical Methodological)* **68**, 477–493.

Fan, J. (1996). Test of significance based on wavelet thresholding and Neyman's truncation. *Journal of the American Statistical Association* **91**, 674–688.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion). *Journal of the Royal Statistical Society, Series B (Statistical Methodological)* **70**, 849–911.

Fan, J., Guo, S. and Hao, N. (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of the Royal Statistical Sociey, Series B (Statistical Methodological)* **74**, 37–65.

Lan, W., Zhong, P. S., Li, R., Wang, H. and Tsai, C. L. (2016). Testing a single regression coefficient in high dimensional linear models. *Journal of Econometrics* **195**, 154–168.

Lee, J. D., Sun, D. L., Sun, Y. and Taylor, J. E. (2016). Exact post-selection inference, with application to the Lasso. *The Annals of Statistics* **44**, 907–927.

Lockhart, R., Taylor, J., Tibshirani, R. and Tibshirani, R. (2014). A significance test for the Lasso (with discussion). *The Annals of Statistics* **42**, 413–468.

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

Meinshausen, N., Meier L. and Bühlmann P. (2009) p-values for high-dimensional regression. *Journal of the American Statistical Association* **104**, 1671–1681.

Scheetz, T. E., Kim, K.-Y. A., Swiderski, R. E., Philp1, A. R., Braun, T. A., Knudtson, K. L. et al. (2006). Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proceeding of the National Academy of Sciences* **103**, 14429–14434.

Segal, M. R., Dahlquist, K. D. and Conklin, B. R. (2003). Regression approach for microarray data analysis. *Journal of Computational Biology* **10**, 961–980.

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

Wang, S. and Cui, H. (2013). Generalized F test for high dimensional linear regression coefficients. *Journal of Multivariate Analysis* **117**, 134–149.

Wang, S., and Cui, H. (2015). A new test for part of high dimensional regression coefficients. *Journal of Multivariate Analysis* **137**, 187–203.

Wang, H., Zhong, P. S. and Cui, Y. (2018). Empirical likelihood ratio tests for coefficients in high-dimensional heteroscedastic linear models. *Statistica Sinica* **28**, 2409–2433.

Wasserman, L. and Roeder, K. (2009). High-dimensional variable selection. *The Annals of Statistics* **37**, 2178–2201.

Zhang, C.-H. and Zhang, S. (2014). Confidence intervals for low-dimensional parameters with high-dimensional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodological)* **76**, 217–242.

Zhong, P. S. and Chen, S. X. (2011). Tests for high-dimensional regression coefficients with factorial designs. *Journal of the American Statistical Association* **106**, 260–274.

Zhong, P. S., Chen, S. X. and Xu, M. (2013) Tests alternative to higher criticism for high-dimensional means under sparsity and column-wise dependence. *The Annals of Statistics* **41**, 2802-2851.

Wenwen Guo

School of Mathematical Science, Capital Normal University, Beijing 100048, China.

E-mail: wwguo@cnu.edu.cn

Wei Zhong

MOE Lab of Econometrics, Wang Yanan Institute for Studies in Economics, Department of Statistics, School of Economics, and Fujian Key Lab of Statistical Science, Xiamen University, Xiamen 361005, China.

E-mail: wzhong@xmu.edu.cn

Sunpeng Duan

Department of Statistics and Applied Probability University of California at Santa Barbara Santa Barbara, CA, 93106, USA.

E-mail: fredduan.dsp@gmail.com

Hengjian Cui

School of Mathematical Science, Capital Normal University, Beijing 100048, China.

E-mail: hjcui@bnu.edu.cn

(Received August 2019; accepted December 2020)