Electronic Journal of Statistics
Vol. 7 (2013) 2131–2149
ISSN: 1935-7524
DOI: 10.1214/13-EJS839

Rank-based score tests for high-dimensional regression coefficients[†]

Long Feng, Changliang Zou, Zhaojun Wang*

LPMC and School of Mathematical Sciences, Nankai University, Tianjin, China
e-mail: flnankai@126.com*;* chlzou@yahoo.com.cn*;* zjwangnk@126.com

and

Bin Chen

School of Mathematics and Statistics, Jiangsu Normal University, Xuzhou, China
e-mail: bchen@jsnu.edu.cn

Abstract: This article is concerned with simultaneous tests on linear regression coefficients in high-dimensional settings. When the dimensionality is larger than the sample size, the classic F-test is not applicable since the sample covariance matrix is not invertible. Recently, [5] and [17] proposed testing procedures by excluding the inverse term in F-statistics. However, the efficiency of such F-statistic-based methods is adversely affected by outlying observations and heavy tailed distributions. To overcome this issue, we propose a robust score test based on rank regression. The asymptotic distributions of the proposed test statistic under the high-dimensional null and alternative hypotheses are established. Its asymptotic relative efficiency with respect to [17]'s test is closely related to that of the Wilcoxon test in comparison with the t-test. Simulation studies are conducted to compare the proposed procedure with other existing testing procedures and show that our procedure is generally more robust in both sizes and powers.

AMS 2000 subject classifications: Primary 62H15; secondary 62G20, 62J05.
Keywords and phrases: Asymptotic normality, high-dimensional data, large p, small n, rank regression, wicoxon test.

Received December 2012.

1. Introduction

With the development of technology, high dimensional data was generated in many areas, such as hyperspectral imagery, internet portals, microarray analysis and finance. A frequently encountered challenge in high-dimensional regression is the detection of relevant variables. Identifying significant sets of genes which

*Corresponding author.
[†]Supported by the NNSF of China Grants No. 11001138, 11071128, 11131002, 11101306, 70931004, 11271169, the RFDP of China Grant No. 20110031110002, Foundation for the Author of National Excellent Doctoral Dissertation of PR China H0512101, New Century Excellent Talents in University NCET-12-0276 and PAPD of Jiangsu Higher Education Institutions.

are associated with certain clinical outcome is very important in genomic studies, see [14, 2] and [11]. The main challenge of high-dimensional data is that the dimension p is much larger than the sample sizes n. When this happens, many traditional statistical methods and theories may not necessarily work since they assume that p keeps unchanged as n increases. Recently, many efforts have been devoted to solve this problem. See for instance [3] and [8] for variable selection and screening. In hypothesis testing, it can be advantageous to look for influence not at the level of individual variables but rather at the level of clusters of variables. Thus, a simultaneous test on linear regression coefficients in high-dimensional settings is needed.

In this article, we consider the following linear regression model

$$Y_i = \mathbf{X}_i^T \boldsymbol{\beta} + \varepsilon_i, \; i = 1, \ldots, n, \tag{1.1}$$

where $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are independent and identically distributed (i.i.d.) p-dimensional covariates, $Y_1, \ldots, Y_n$ are independent responses, $\boldsymbol{\beta}$ is the vector of regression coefficients, and ε_i is independent of $\mathbf{X}_i$ and has the finite Fisher information quantity, i.e., $\int f^{-1}(x)[f^{'}(x)]^2 dx < \infty$ with the density function $f(\cdot)$. As a convention in the literature, we assume that $\mathbf{X}_i$ are normalized first, that is, $E(\mathbf{X}_i) = \mathbf{0}_p$, and $\mathrm{var}(X_{ik}) = 1$, $k = 1, \ldots, p$. To make $\boldsymbol{\beta}$ identifiable, we assume that $\boldsymbol{\Sigma} = \mathrm{var}(\mathbf{X}_i)$ is positive definite. Our interest is in testing a high-dimensional hypothesis

$$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0 \quad \text{vs} \quad H_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta}_0 \tag{1.2}$$

for a specific $\boldsymbol{\beta}_0 \in \mathbb{R}^p$. Without loss of generality, we set $\boldsymbol{\beta}_0 = \mathbf{0}_p$ which aims to investigate the association of the covariates $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^T$ with the response $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$.

When $p < n$, a classical method to deal with this problem is the famous F-test statistic. However, [17] showed that the power of F-test is adversely impacted by an increased dimension even $p < n-1$, reflecting a reduced degree of freedom in estimating σ^2 when the dimensionality is close to the sample size. Moreover, the F-test statistics is undefined when the dimension of data is greater than the within sample degrees of freedom since the pooled sample covariance matrices are not positive definite. In order to overcome this issue, [5, 4] proposed an Empirical Bayes test with the test statistic

$$G_n = \frac{\mathbf{Y}^T \mathbf{X} \mathbf{X}^T \mathbf{Y}}{\mathbf{Y}^T \mathbf{Y}} \tag{1.3}$$

which is formulated via a score test on the hyper parameter of a prior distribution assumed on the regression coefficients. The key feature of their method is to use Euclidian norm to replace the Mahalanobis norm since having $(\mathbf{X}^T\mathbf{X})^{-1}$ is no longer beneficial when p is larger than n. That is, exclude the inverse term $(\mathbf{X}^T\mathbf{X})^{-1}$ in the F-statistic. [17] further consider a U-statistic

$$Z_n = \frac{1}{4P_n^4} \sum^{*} (\mathbf{X}_{i_1} - \mathbf{X}_{i_2})^T (\mathbf{X}_{i_3} - \mathbf{X}_{i_4})(Y_{i_1} - Y_{i_2})(Y_{i_3} - Y_{i_4}) \tag{1.4}$$

where we use $\sum^*$ to denote summations over distinct indexes and $P_n^4 = n!/(n-4)!$. They established the asymptotic normality of Z_n under the diverging factor model [1].

Both [5] and [17] methods can be regarded as variants of the F-statistic in which the original observations Y_i are used. Therefore, their statistical properties, designed to perform "best" under the normality assumption, could potentially be (highly) affected when the errors are far away from normal or the data contain some outliers. To this end, in Section 2.1 we propose a robust score test based on rank regression. The asymptotic distributions of the proposed test statistic under the high-dimensional null and alternative hypotheses are established in Section 2.2. Theoretical studies show that its asymptotic relative efficiency with respect to [17]'s test is the same as that of the Wilcoxon test in comparison with the t-test. Numerical studies in Section 3 show that the proposed procedure is generally more robust in both sizes and powers for non-normal errors. All technical details are provided in the Appendix. The R code for implementing the proposed procedure is given in a supplemental file.

2. A robust test

2.1. Test statistics

Denote $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$. Consider, first, the following general rank-based pseudo-norm (Chapter 1, [7])

$$||\boldsymbol{\varepsilon}||_\psi = \sum_{i=1}^n a(R(\varepsilon_i))\varepsilon_i,$$

where $R(\varepsilon_i)$ is the rank of ε_i and $a(1) \le a(2) \le \cdots \le a(n)$ is a set of scores generated as $a(i) = \psi(i/(n+1))$ for some nondecreasing score function $\psi(u)$ defined on the interval $(0,1)$ and standardized such that $\int \psi(u)du = 0$ and $\int \psi^2(u)du = 1$. With respect to model (1.1), when $p < n$, the classic R-estimator of $\boldsymbol{\beta}$ is defined by minimizing the following dispersion function,

$$D_\psi(\boldsymbol{\beta}) = ||Y - \mathbf{X}\boldsymbol{\beta}||_\psi.$$

Based on the data, the most acceptable value of $\boldsymbol{\beta}$ is the value at which the gradient $\frac{\partial D_\psi}{\partial \boldsymbol{\beta}}$ is zero. Hence, large values of $\frac{\partial D_\psi}{\partial \boldsymbol{\beta}}\big|_{\boldsymbol{\beta}=\mathbf{0}}$ will reject the null hypothesis. Note that the variance of $\frac{\partial D_\psi}{\partial \boldsymbol{\beta}}\big|_{\boldsymbol{\beta}=\mathbf{0}}$ is $-\frac{\partial^2 D_\psi}{\partial \boldsymbol{\beta}\boldsymbol{\beta}^T}\big|_{\boldsymbol{\beta}=\mathbf{0}}$ under the null hypothesis. Accordingly, the rank-based score test statistic (Chapter 3, [7]) of the hypothesis (1.2) is defined by

$$\frac{\partial D_\psi}{\partial \boldsymbol{\beta}}\left(-\frac{\partial^2 D_\psi}{\partial \boldsymbol{\beta}\boldsymbol{\beta}^T}\right)^{-1}\frac{\partial D_\psi}{\partial \boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\mathbf{0}} = a^T(R(Y))\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T a(R(Y)), \tag{2.1}$$

where $a(R(Y)) = (a(R(Y_1)), \ldots, a(R(Y_n)))^T$. However, when $p > n$, $\mathbf{X}^T\mathbf{X}$ is not invertible and thus the score test above is not well defined. Thus, motivated by [5] and [17], we use $\mathbf{I}_p$ to replace $(\mathbf{X}^T\mathbf{X})^{-1}$ in (2.1) and propose the following test statistic

$$R_n = \frac{2}{n(n-1)}\sum_{i<j}\mathbf{X}_i^T\mathbf{X}_j a(R(Y_i))a(R(Y_j))$$

after removing the term $\mathbf{X}_i^T\mathbf{X}_i a(R(Y_i))^2$ in $a(R(Y))^T\mathbf{X}\mathbf{X}^T a(R(Y))$. One of the most commonly used score function in the literature is the so-called Wilcoxon score function, i.e., $\psi(u) = \sqrt{12}(u-1/2)$, which is used in this article. As a consequence, the Wilcoxon form of R_n can be written as

$$W_n = \frac{2}{n(n-1)}\sum_{i<j}\mathbf{X}_i^T\mathbf{X}_j e_i e_j,$$

where $e_i = \sqrt{12}(\frac{R(Y_i)}{n+1} - \frac{1}{2})$. Note that in (1.1), it is not required to assume $E(\varepsilon) = 0$ because R_n is constructed by the order statistics of Y. An intercept term can be reflected by the location parameter of the distribution of ε.

Remark 1. The suggested test statistic can be viewed as a specific Empirical Bayes test statistic [5] with a pseudo Wilcoxon likelihood function [9, 16]. Consider the following pseudo-likelihood

$$L_W(\boldsymbol{\beta}) \propto \exp\left(-D_{\boldsymbol{\psi}}(\boldsymbol{\beta})/\tau_\epsilon\right), \tag{2.2}$$

where $\psi(\cdot)$ is chosen as the Wilcoxon score function and $\tau_\epsilon = (\sqrt{12}\int f^2(x)dx)^{-1}$ is the Wilcoxon constant of the error distribution. The main motivation for using (2.2) as a pseudo-likelihood is two-folded: on one hand, it can be shown that under the null hypothesis $\boldsymbol{\beta} = \mathbf{0}$ with fixed p, the test statistic $2\tau_\epsilon(L_W(\mathbf{0}) - L_W(\boldsymbol{\beta}))$ is asymptotically χ^2 distributed, as for the likelihood ratio test; on the other hand, the loss function $D_{\boldsymbol{\psi}}(\boldsymbol{\beta})$ is analogous to the least-squares procedure except that the Euclidean norm is substituted by the rank-based norm. Using similar procedure and arguments in Section 6 of [5], the $a^T(R(Y))\mathbf{X}\mathbf{X}^T a(R(Y))$ is essentially equivalent to the Empirical Bayes test statistic with the pseudo-likelihood (2.2).

2.2. Asymptotic property

In order to establish the asymptotic normality of W_n, we assume, like [1], the following diverging factor model:

$$\mathbf{X}_i = \boldsymbol{\Gamma}\mathbf{z}_i,$$

where $\boldsymbol{\Gamma}$ is a $p\times m$ matrix for some $m \geq p$ such that $\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T = \boldsymbol{\Sigma}$ and $\{\mathbf{z}_i\}_{i=1}^n$ are m-variate independent and identically distributed random vectors such that

$$\begin{gathered} E(\mathbf{z}_i) = 0,\ \mathrm{var}(\mathbf{z}_i) = \mathbf{I}_m,\ E(z_{il}^4) = 3+\Delta,\ E(z_{il}^8) < +\infty, \\ E(z_{ik_1}^{\alpha_1} z_{ik_2}^{\alpha_2}\dots z_{ikq}^{\alpha_q}) = E(z_{ik_1}^{\alpha_1})E(z_{ik_2}^{\alpha_2})\dots E(z_{ikq}^{\alpha_q}), \end{gathered} \tag{2.3}$$

whenever Δ is a constant, α_k are some positive integer with $\sum_{k=1}^{q} \alpha_k \le 8$ and $k_1 \ne k_2 \ne \cdots \ne k_q$. Additionally, we need the following condition (C1) which is used in [17] to regulate for the "large p, small n",

(C1) $\mathrm{tr}(\mathbf{\Sigma}^4) = o(\mathrm{tr}^2(\mathbf{\Sigma}^2))$.
(C2) The density function $f(\cdot)$ has uniformly bounded derivatives.

Remark 2. The (C1) is the same as the last part of condition (2.8) in [17]. If all the eigenvalues of $\mathbf{\Sigma}$ are bounded, (C1) is trivially true for any p. Intuitively speaking, in the correlation matrix if there are a large amount of entries whose values are very large, (C1) will not hold and so does the asymptotic normality of W_n. Thus, the validity of asymptotic normality relies on how strong dependencies among the variables. [5, 4] provide us an alternative way to handle the large p problems. In fact, according to Theorem 1 in [4], we could similarly develop some approximate representation for our quadratic form test statistic $a(R(Y))^T \mathbf{X}\mathbf{X}^T a(R(Y))$ without (C1). As a consequence, the computation-based method proposed by [4] could be used instead of asymptotic normality when the validity of (C1) is doubtful. A systematic comparison between these two methods (in terms of approximation of nominal level) is of interest. This is out of the scope of this paper but which deserves further study. (C2) means that $\sup_x |f'(x)| < M$ with a positive constant M. It is used to control the terms A_{n32} and A_{n33} in the proof of Theorem 1.

In order to study the asymptotic power of our test, we define the following local alternatives in which $\boldsymbol{\beta}$ satisfies

$$\boldsymbol{\beta}^T \mathbf{\Sigma} \boldsymbol{\beta} = o(n^{-1/2}), \boldsymbol{\beta}^T \mathbf{\Sigma}^2 \boldsymbol{\beta} = o(\sqrt{n^{-1}\mathrm{tr}(\mathbf{\Sigma}^2)}), \boldsymbol{\beta}^T \mathbf{\Sigma}^3 \boldsymbol{\beta} = o(n^{-1}\mathrm{tr}(\mathbf{\Sigma}^2)) \quad (2.4)$$

We refer to [17] for detailed discussion on this condition. The following theorem establishes the asymptotic normality of W_n under the null or local alternative (2.4) hypothesis.

Theorem 1. *Assume the condition (C1)–(C2) hold, then under either H_0 or the local alternative (2.4), as $p \to \infty$ and $n \to \infty$,*

$$\frac{n}{\sqrt{2\mathrm{tr}(\mathbf{\Sigma}^2)}} \left(W_n - \boldsymbol{\beta}^T \mathbf{\Sigma}^2 \boldsymbol{\beta} / \tau_\epsilon^2 \right) \xrightarrow{d} N(0,1). \quad (2.5)$$

Remark 3. Define the scale parameter $\tau_\psi = (\int \psi(u)\psi_f(u)du)^{-1}$, where $\psi_f(u) = \frac{f'(F^{-1}(u))}{f(F^{-1}(u))}$ and $F(\cdot)$ is the distribution function of ε. Under the assumption of finite Fisher information, τ_ψ is well defined. Then, similar to the proof of Theorem 1, we can also show that

$$\frac{n}{\sqrt{2\mathrm{tr}(\mathbf{\Sigma}^2)}} \left(R_n - \boldsymbol{\beta}^T \mathbf{\Sigma}^2 \boldsymbol{\beta} / \tau_\psi^2 \right) \xrightarrow{d} N(0,1).$$

Note that, if $\psi(u) = \sqrt{12}(u - 1/2)$, then $\tau_\psi = \tau$ and thus the conclusion in Theorem 1 is a special case of the result in this Remark.

To formulate a test procedure based on Theorem 1, we need to estimate $\mathrm{tr}(\mathbf{\Sigma}^2)$ appeared in the asymptotic variance. Similar to [17], we use the following ratio consistent estimator of $\mathrm{tr}(\mathbf{\Sigma}^2)$,

$$\widehat{\mathrm{tr}(\mathbf{\Sigma}^2)} = \frac{1}{2P_n^4} \sum^{*} (\mathbf{X}_{i_1} - \mathbf{X}_{i_2})^T (\mathbf{X}_{i_3} - \mathbf{X}_{i_4})(\mathbf{X}_{i_3} - \mathbf{X}_{i_2})^T (\mathbf{X}_{i_1} - \mathbf{X}_{i_4}).$$

By Theorem 1 and the Slutsky Theorem, the proposed test rejects H_0 at a significant level α if

$$nW_n \geq \sqrt{2\widehat{\mathrm{tr}(\mathbf{\Sigma}^2)}}\, z_\alpha, \tag{2.6}$$

where z_α is the upper-α quantile of $N(0,1)$.

Next, we discuss the power properties of the proposed test. According to Theorem 1, the power of our proposed test under the local alternative (2.4) is

$$\beta_{W_n}(||\boldsymbol{\beta}||) = \Phi(-z_\alpha + n\boldsymbol{\beta}^T \mathbf{\Sigma}^2 \boldsymbol{\beta} / \sqrt{2\mathrm{tr}(\mathbf{\Sigma}^2)}\tau_\epsilon^2),$$

where Φ is the standard normal distribution function. In comparison, [17] show that the power of their proposed test is

$$\beta_{Z_n}(||\boldsymbol{\beta}||) = \Phi(-z_\alpha + n\boldsymbol{\beta}^T \mathbf{\Sigma}^2 \boldsymbol{\beta} / \sqrt{2\mathrm{tr}(\mathbf{\Sigma}^2)}\sigma^2).$$

Thus, the Pitman asymptotic relative efficiency of W_n with respect to the Z_n is

$$\mathrm{ARE}(Z_n, W_n) = \sigma^2/\tau_\epsilon^2.$$

Based on this result, the robust feature of our proposed test is clear. In our procedure, no moment assumption on ε is made, such as $\sigma^2 < \infty$. For instance, if ε comes from Cauchy or the t-distribution with two degrees of freedom ($t(2)$), our test procedure still works but the moment-based methods such as [17] would generally fail (the ARE goes to infinity). Note that the above ARE is essentially the same as the ARE of the signed-rank Wilcoxon test in comparison with the t-test and it has a lower bound 0.864 [7]. It is well known in the literature of rank analysis that the ARE is as high as 0.955 for normal error distribution, and can be significantly higher than one for many heavier-tailed distributions [7]. For instance, this quantity is 1.5 for the double exponential distribution, and 1.9 for the $t(3)$. Furthermore, we can also show the Pitman asymptotic relative efficiency of R_n with respect to the Z_n is $\mathrm{ARE}(Z_n, R_n) = \sigma^2/\tau_\psi^2$.

3. Simulation

In this section, we present two simulation examples that evaluate the finite sample performance of the rank-based score test (abbreviated as RS). The Empirical Bayes test proposed by [5] (abbreviated as EB) and the modified F-test proposed by [17] (abbreviated as ZC) are included for comparison purposes. The following two examples are considered.

TABLE 1

Power comparison of the tests with AR Model when $p < n$; RS*: The proposed rank-based score test test;* ZC*: [17] test; EB: [5]test; F: the classical F test*

		Errors															
		$N(0,1)$				$t(3)$				Lognormal				$T(0.1,10)$			
(n,p)	SNR	RS	ZC	EB	F	RS	ZC	EB	F	RS	ZC	EB	F	RS	ZC	EB	F
non-sparse case																	
(50, 47)	0	0.05	0.05	0.04	0.05	0.05	0.04	0.05	0.04	0.05	0.05	0.06	0.04	0.06	0.03	0.05	0.05
	1	0.21	0.24	0.20	0.05	0.33	0.27	0.24	0.06	0.58	0.33	0.29	0.06	0.70	0.31	0.31	0.06
	2	0.37	0.41	0.37	0.06	0.51	0.46	0.44	0.07	0.73	0.50	0.46	0.07	0.82	0.49	0.47	0.08
	3	0.49	0.55	0.51	0.08	0.64	0.60	0.58	0.09	0.81	0.63	0.61	0.08	0.87	0.59	0.58	0.09
(100, 95)	0	0.05	0.05	0.04	0.05	0.06	0.06	0.04	0.05	0.05	0.04	0.04	0.05	0.05	0.03	0.05	0.04
	1	0.23	0.25	0.23	0.05	0.37	0.27	0.26	0.07	0.74	0.34	0.30	0.08	0.88	0.29	0.28	0.07
	2	0.43	0.47	0.41	0.09	0.64	0.49	0.51	0.08	0.90	0.56	0.51	0.08	0.96	0.54	0.49	0.07
	3	0.60	0.64	0.59	0.07	0.82	0.68	0.64	0.09	0.95	0.70	0.68	0.09	0.98	0.68	0.62	0.10
(200, 190)	0	0.05	0.05	0.04	0.04	0.06	0.04	0.06	0.06	0.05	0.05	0.06	0.05	0.05	0.04	0.05	0.06
	1	0.21	0.23	0.24	0.06	0.44	0.28	0.28	0.06	0.82	0.30	0.30	0.07	0.94	0.27	0.25	0.06
	2	0.47	0.48	0.46	0.07	0.74	0.56	0.52	0.08	0.98	0.58	0.52	0.08	1.00	0.52	0.51	0.09
	3	0.65	0.70	0.65	0.09	0.89	0.74	0.75	0.10	0.99	0.75	0.70	0.12	1.00	0.71	0.73	0.11
sparse case																	
(50, 47)	1	0.18	0.19	0.18	0.06	0.33	0.26	0.25	0.08	0.52	0.31	0.29	0.07	0.63	0.28	0.28	0.07
	2	0.33	0.37	0.34	0.06	0.51	0.45	0.38	0.07	0.67	0.49	0.45	0.09	0.75	0.46	0.43	0.08
	3	0.46	0.50	0.48	0.08	0.63	0.56	0.58	0.08	0.76	0.60	0.56	0.09	0.79	0.56	0.55	0.08
(100, 95)	1	0.24	0.26	0.20	0.07	0.38	0.26	0.25	0.08	0.66	0.33	0.29	0.06	0.78	0.28	0.25	0.06
	2	0.40	0.47	0.38	0.08	0.65	0.48	0.47	0.09	0.82	0.56	0.51	0.11	0.89	0.47	0.46	0.10
	3	0.58	0.63	0.56	0.09	0.77	0.65	0.67	0.10	0.89	0.69	0.62	0.12	0.92	0.63	0.62	0.10
(200, 190)	1	0.20	0.22	0.22	0.06	0.42	0.26	0.27	0.09	0.79	0.28	0.29	0.07	0.89	0.26	0.27	0.08
	2	0.44	0.46	0.46	0.09	0.72	0.53	0.52	0.10	0.93	0.54	0.53	0.10	0.98	0.50	0.48	0.09
	3	0.64	0.66	0.64	0.12	0.87	0.71	0.69	0.16	0.98	0.72	0.73	0.15	0.99	0.69	0.67	0.13

3.1. Example 3.1 (AR model)

Define $\mathbf{\Sigma} = (\sigma_{ij}) \in \mathbb{R}^{p\times p}$ with $\sigma_{ij} = 0.5^{|i-j|}$. The predictors $\mathbf{X}_i \sim N(\mathbf{0}, \mathbf{\Sigma})$. We consider $\boldsymbol{\beta} = \kappa(\beta_1, \ldots, \beta_p)$. Similar to [17], we consider the following two scenarios. Sparse case: $\beta_i (1 \le i \le 5)$ are generated from standard normal distribution and $\beta_i = 0$, $i > 5$; Non-sparse case: $\beta_i (1 \le i \le p/2)$ are generated from standard normal distribution and $\beta_i = 0$, $i > p/2$. The coefficient κ is selected so that the signal-to-noise ratio (SNR) of ZC equals r, i.e., $n||\mathbf{\Sigma}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)||^2/(\sqrt{2\text{tr}(\mathbf{\Sigma}^2)}\sigma^2) = r$. $r = 0$(size), 1, 2, 3 are considered. Four error distributions are chosen: (i) $N(0,1)$; (ii) Student's t-distribution with three degrees of freedom, $t(3)$; (iii) Lognormal distribution; (iv) Tukey contaminated normal $T(0.1,10)(0.9N(0,1)+0.1N(0,100))$. All these four errors are standardized with mean zero and variance one.

Table 2
Power comparison of the tests with AR Model when $p > n$

		Errors											
		$N(0,1)$			$t(3)$			Lognormal			$T(0.1,10)$		
(n,p)	SNR	RS	ZC	EB	RS	ZC	EB	RS	ZC	EB	RS	ZC	EB
non-sparse case													
(50, 100)	0	0.06	0.06	0.04	0.05	0.05	0.05	0.06	0.05	0.06	0.05	0.04	0.05
	1	0.19	0.21	0.21	0.29	0.23	0.24	0.44	0.27	0.28	0.59	0.29	0.28
	2	0.32	0.35	0.35	0.45	0.39	0.42	0.60	0.44	0.44	0.70	0.44	0.41
	3	0.42	0.48	0.47	0.57	0.52	0.51	0.69	0.56	0.56	0.76	0.56	0.51
(100, 200)	0	0.05	0.05	0.04	0.05	0.05	0.05	0.05	0.05	0.04	0.06	0.03	0.06
	1	0.20	0.22	0.21	0.37	0.27	0.25	0.62	0.29	0.30	0.76	0.26	0.27
	2	0.38	0.41	0.41	0.61	0.48	0.47	0.80	0.51	0.46	0.89	0.45	0.46
	3	0.54	0.57	0.58	0.75	0.64	0.65	0.88	0.67	0.64	0.93	0.59	0.63
(200, 400)	0	0.05	0.05	0.04	0.05	0.05	0.04	0.05	0.05	0.05	0.06	0.04	0.05
	1	0.22	0.24	0.23	0.40	0.28	0.27	0.76	0.27	0.28	0.90	0.26	0.24
	2	0.45	0.47	0.44	0.70	0.50	0.52	0.94	0.51	0.53	0.98	0.51	0.49
	3	0.63	0.66	0.65	0.87	0.70	0.72	0.98	0.72	0.71	1.00	0.69	0.64
sparse case													
(50, 100)	1	0.16	0.18	0.19	0.26	0.22	0.21	0.41	0.27	0.27	0.50	0.24	0.26
	2	0.27	0.30	0.30	0.41	0.36	0.36	0.54	0.42	0.42	0.60	0.38	0.36
	3	0.37	0.42	0.42	0.49	0.47	0.48	0.60	0.50	0.52	0.65	0.47	0.48
(100, 200)	1	0.20	0.21	0.18	0.32	0.22	0.23	0.57	0.27	0.25	0.68	0.27	0.25
	2	0.38	0.41	0.38	0.53	0.42	0.44	0.75	0.48	0.46	0.81	0.46	0.45
	3	0.54	0.56	0.51	0.68	0.58	0.56	0.82	0.61	0.58	0.84	0.58	0.55
(200, 400)	1	0.23	0.25	0.23	0.40	0.26	0.23	0.69	0.30	0.26	0.84	0.25	0.25
	2	0.43	0.45	0.44	0.68	0.50	0.47	0.86	0.52	0.52	0.94	0.50	0.44
	3	0.60	0.63	0.63	0.82	0.68	0.65	0.93	0.68	0.64	0.96	0.65	0.61

3.2. Example 3.2 (Moving average model)

We also consider the moving average model used in [17]. The predictor vector is given by $\mathbf{X}_i$ for $i = 1, \ldots, n$ and $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})$ are generated according to the following moving average model

$$X_{ij} = \rho_1 V_{ij} + \rho_2 V_{i(j+1)} + \cdots + \rho_T V_{i(j+T-1)}, j = 1, \ldots, p$$

for some $T < p$. Here $\boldsymbol{V}_i = (V_{i1}, \ldots, V_{i(p+T-1)})$ is a $(p+T-1)$-dimensional $N(0, \mathbf{I}_{p+T-1})$ random vector and we only consider the case $T = 10$. The coefficients $\{\rho_l\}_{l=1}^{T}$ are generated independently from the Uniform (0, 1) distribution and are kept fixed once generated. The regression coefficients, errors, and sample sizes are the same as those in Example 3.1.

For each experiment we run 1, 000 replications. Tables 1 and 2 summarize the empirical sizes and powers at a 5% significance level for Example 3.1 under the settings $p < n$ and $p > n$, respectively. The simulation results for Example 3.2

TABLE 3
Power comparison of the tests with MA Model when $p < n$

		Errors															
		$N(0,1)$				$t(3)$				Lognormal				$T(0.1, 10)$			
(n,p)	SNR	RS	ZC	EB	F	RS	ZC	EB	F	RS	ZC	EB	F	RS	ZC	EB	F
non-sparse case																	
(50, 47)	0	0.05	0.06	0.05	0.05	0.06	0.06	0.05	0.04	0.05	0.04	0.06	0.06	0.06	0.03	0.04	0.04
	1	0.23	0.25	0.19	0.06	0.32	0.29	0.26	0.06	0.63	0.33	0.30	0.07	0.73	0.31	0.29	0.06
	2	0.38	0.41	0.36	0.06	0.55	0.46	0.44	0.07	0.77	0.52	0.49	0.07	0.86	0.50	0.48	0.07
	3	0.51	0.56	0.48	0.07	0.67	0.61	0.58	0.09	0.83	0.66	0.61	0.07	0.89	0.62	0.57	0.09
(100, 95)	0	0.06	0.06	0.05	0.05	0.06	0.06	0.04	0.05	0.06	0.06	0.05	0.05	0.05	0.05	0.06	0.06
	1	0.25	0.27	0.21	0.05	0.41	0.30	0.28	0.07	0.78	0.34	0.31	0.05	0.88	0.31	0.28	0.08
	2	0.46	0.50	0.43	0.09	0.70	0.55	0.49	0.08	0.93	0.57	0.52	0.07	0.96	0.51	0.46	0.08
	3	0.63	0.67	0.58	0.07	0.86	0.73	0.65	0.09	0.97	0.74	0.69	0.09	0.99	0.67	0.63	0.08
(200, 190)	0	0.06	0.06	0.04	0.04	0.05	0.06	0.06	0.06	0.05	0.05	0.06	0.05	0.05	0.06	0.05	0.05
	1	0.23	0.27	0.24	0.06	0.45	0.29	0.26	0.06	0.88	0.31	0.28	0.08	0.97	0.27	0.25	0.06
	2	0.48	0.51	0.43	0.07	0.79	0.56	0.50	0.08	0.98	0.59	0.54	0.09	1.00	0.53	0.51	0.07
	3	0.68	0.71	0.69	0.09	0.93	0.75	0.73	0.09	0.99	0.75	0.71	0.10	1.00	0.72	0.72	0.09
sparse case																	
(50, 47)	1	0.20	0.21	0.17	0.06	0.32	0.28	0.23	0.07	0.50	0.30	0.28	0.07	0.63	0.31	0.30	0.06
	2	0.32	0.35	0.30	0.08	0.49	0.45	0.41	0.08	0.65	0.48	0.41	0.09	0.74	0.47	0.42	0.90
	3	0.44	0.47	0.43	0.08	0.60	0.57	0.52	0.09	0.71	0.59	0.54	0.09	0.79	0.57	0.49	0.10
(100, 95)	1	0.24	0.24	0.20	0.06	0.35	0.25	0.24	0.06	0.66	0.29	0.27	0.08	0.76	0.26	0.22	0.08
	2	0.40	0.44	0.35	0.08	0.60	0.48	0.44	0.08	0.81	0.50	0.48	0.10	0.87	0.47	0.45	0.09
	3	0.55	0.58	0.51	0.09	0.75	0.63	0.60	0.11	0.87	0.64	0.58	0.11	0.88	0.61	0.54	0.10
(200, 190)	1	0.24	0.25	0.22	0.06	0.44	0.28	0.24	0.07	0.79	0.28	0.23	0.08	0.87	0.28	0.23	0.08
	2	0.45	0.46	0.43	0.11	0.72	0.54	0.48	0.10	0.93	0.53	0.48	0.12	0.94	0.51	0.47	0.10
	3	0.63	0.67	0.61	0.12	0.85	0.71	0.65	0.14	0.95	0.70	0.65	0.16	0.95	0.67	0.62	0.14

are reported in Tables 3–4. The F test is also conducted under the setting of $p < n$. The comparison results for the two examples are similar: Firstly, we observe that the empirical sizes for the three tests RS, ZC and EB, are reasonable in most of the cases. Secondly, the proposed method is highly efficient for all the distributions under consideration. Its powers compared to [17] and [5] tests are high and are uniformly larger than those of the other tests except for the normal distribution. For the three non-normal errors, the proposed test outperforms its three competitors by a quite large margin. Even for normal, its powers are merely slightly smaller than [17] test. Thirdly, both [17] and [5] tests are more powerful than the F test in such high-dimensional settings as we would expect, while the former performs little better than the latter which is consistent with the findings in [17].

4. A real-data application

To further demonstrate the usefulness of our RS test in real practice, we consider a Cardiomyopathy microarray data [12] with $(n, p) = (30, 6319)$, which are

Table 4
Power comparison of the tests with MA Model when $p > n$

		Errors											
		$N(0,1)$			$t(3)$			Lognormal			$T(0.1, 10)$		
(n, p)	SNR	RS	ZC	EB	RS	ZC	EB	RS	ZC	EB	RS	ZC	EB
non-sparse case													
(50, 100)	0	0.05	0.06	0.04	0.06	0.06	0.06	0.05	0.04	0.05	0.05	0.03	0.05
	1	0.22	0.25	0.20	0.35	0.30	0.26	0.60	0.33	0.28	0.69	0.32	0.27
	2	0.38	0.41	0.34	0.56	0.50	0.44	0.74	0.53	0.49	0.83	0.50	0.45
	3	0.51	0.55	0.51	0.67	0.61	0.55	0.82	0.64	0.58	0.87	0.60	0.57
(100, 200)	0	0.06	0.06	0.05	0.05	0.05	0.06	0.06	0.05	0.05	0.05	0.04	0.05
	1	0.24	0.26	0.20	0.41	0.28	0.28	0.72	0.32	0.30	0.84	0.28	0.29
	2	0.42	0.46	0.46	0.69	0.53	0.50	0.89	0.55	0.51	0.96	0.51	0.52
	3	0.58	0.61	0.58	0.82	0.72	0.65	0.95	0.70	0.66	0.98	0.67	0.61
(200, 400)	0	0.05	0.06	0.05	0.05	0.06	0.05	0.05	0.05	0.04	0.06	0.05	0.06
	1	0.26	0.27	0.24	0.46	0.30	0.25	0.87	0.33	0.28	0.96	0.26	0.27
	2	0.47	0.50	0.45	0.78	0.57	0.54	0.98	0.59	0.55	1.00	0.53	0.52
	3	0.67	0.70	0.67	0.93	0.75	0.73	1.00	0.78	0.74	1.00	0.72	0.70
sparse case													
(50, 100)	1	0.18	0.21	0.16	0.28	0.25	0.20	0.46	0.28	0.25	0.53	0.26	0.26
	2	0.30	0.37	0.34	0.43	0.42	0.35	0.59	0.45	0.39	0.63	0.40	0.37
	3	0.42	0.47	0.41	0.54	0.51	0.44	0.64	0.55	0.52	0.67	0.50	0.49
(100, 200)	1	0.21	0.23	0.19	0.32	0.23	0.24	0.57	0.29	0.22	0.69	0.25	0.23
	2	0.36	0.39	0.35	0.54	0.43	0.42	0.74	0.48	0.43	0.80	0.46	0.40
	3	0.50	0.53	0.47	0.66	0.57	0.52	0.80	0.60	0.55	0.83	0.59	0.53
(200, 400)	1	0.20	0.22	0.20	0.39	0.24	0.25	0.69	0.27	0.26	0.80	0.26	0.27
	2	0.40	0.43	0.41	0.66	0.51	0.46	0.85	0.49	0.48	0.87	0.46	0.44
	3	0.58	0.61	0.58	0.79	0.66	0.64	0.89	0.66	0.64	0.90	0.61	0.59

from a transgenic mouse model of dilated cardiomyopathy. This dataset consists of a $n \times p$ matrix of gene expression values $X = (X_{ij})$, where X_{ij} is the expression level of the j-th gene for the i-th mouse. Each mouse also provides an outcome (Ro1) measure Y_i. The goal is to identify the most influential genes for overexpression of a G protein-coupled receptor (Ro1) in mice. Thus, the regression model (1.1) is useful. Due to confidentiality reasons, both the response and predictors have been standardized to be zero mean and unit variance.

Firstly, we select the important predictors by the sure independence screening (SIS) and iterative SIS (ISIS) methods [3] from which we get 8 and 5 potentially relevant variables respectively. Then we consider two hypotheses to check the selection accuracy of these two methods: (i) whether the selected predictors are correlated with the response? (ii) whether the eliminated predictors by the screening methods are indeed irrelevant? Clearly, we can adopt the tests, RS, EB, ZC, and F for the first hypothesis as $n > p$ in this case. For the second hypothesis, we test whether the eliminated predictors are correlated with the

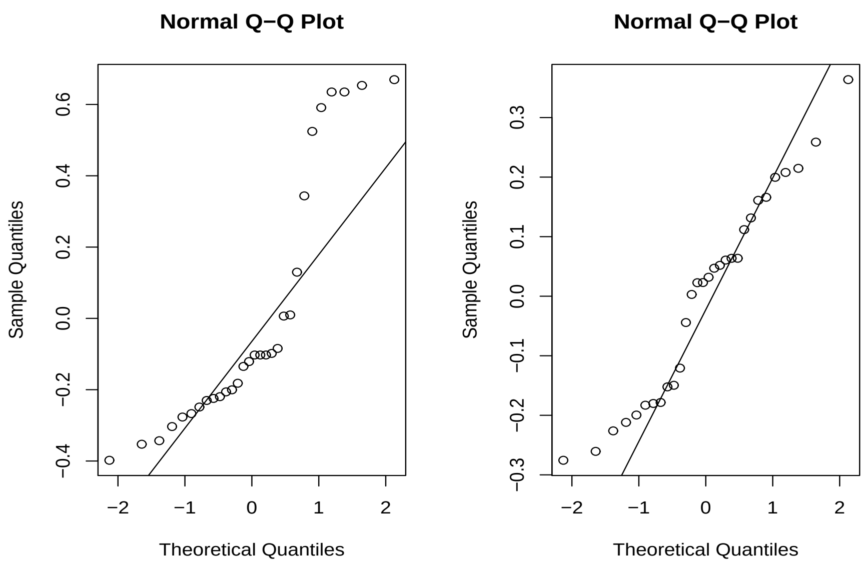

FIG 1. *The QQ plot of residuals of SIS and ISIS for the Cardiomyopathy microarray data, respectively.*

TABLE 5
p-values of each tests

	RS	EB	ZC	F
SIS-selected	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
SIS-eliminated	2e-5 (0.00)	3e-4 (0.00)	0.48 (0.35)	–
ISIS-selected	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
ISIS-eliminated	0.37 (0.27)	0.28 (0.28)	0.50 (0.54)	–

Note: the bootstrapped p-values are given in parentheses.

corresponding R-residuals (Chapter 3, [7]) and then the three tests, ZC, EB and RS, are all applicable. Figure 1 shows the QQ-plot of residuals of SIS and ISIS for the Cardiomyopathy microarray data. We observe that the residuals are somewhat heavy-tailed, which motives us to use our proposed RS test.

Table 5 shows the p-values of the four tests for the considered hypotheses. Furthermore, because the sample size is only 30, we also consider a bootstrapped-based method to obtain p-values (10, 000 resampling). We observe that both selected predictors by SIS and ISIS would be relevant since all the p-values are extremely small. Since we cannot reject the hypothesis (ii) for the ISIS-eliminated predictors, we may conclude that the ISIS method is able to correctly select the relevant predictors. However, the RS and EB tests reject the null hypothesis that the SIS-eliminated predictors are relevant and hence some important predictors may be missed by the SIS method in this case. In contrast, the ZC test is not significant under the hypothesis (ii) for the SIS-eliminated predictors, which may demonstrate the robustness of our proposed method to certain degree.

5. Appendix

Firstly, we restate some results on the quadratic forms in the following lemma [13].

Lemma 1. *Under the model (2.3), for any $m \times m$ symmetric matrix $A = (a_{ij})$ and $B = (b_{ij})$ of constants, we have*

$$E[(\mathbf{z}_i^T A\mathbf{z}_i)^2] = \Delta \sum_{i=1}^{m} a_{ii}^2 + 2\mathrm{tr}(A^2) + (\mathrm{tr}(A))^2,$$
$$\mathrm{var}[(\mathbf{z}_i^T A\mathbf{z}_i)^2] = \Delta \sum_{i=1}^{p} a_{ii}^2 + 2\mathrm{tr}(A^2),$$
$$E[(\mathbf{z}_i^T A\mathbf{z}_i)(\mathbf{z}_i^T B\mathbf{z}_i)] = \Delta \sum_{i=1}^{m} a_{ii}b_{ii} + 2\mathrm{tr}(AB) + \mathrm{tr}(A)\mathrm{tr}(B).$$

5.1. Proof of Theorem 1

Recall the definition $e_i = a(R(Y_i))$ and denote $\xi_i = \sqrt{12}(\frac{R(\varepsilon_i)}{n+1} - \frac{1}{2})$. We can rewrite W_n as follows

$$\begin{aligned} W_n &= \frac{1}{n(n-1)} \sum_{i \neq j} \mathbf{X}_i^T \mathbf{X}_j e_i e_j, \\ &= \frac{1}{n(n-1)} \sum_{i \neq j} \mathbf{X}_i^T \mathbf{X}_j \xi_i \xi_j + \frac{2}{n(n-1)} \sum_{i \neq j} \mathbf{X}_i^T \mathbf{X}_j (e_i - \xi_i)\xi_j, \\ &\quad + \frac{1}{n(n-1)} \sum_{i \neq j} \mathbf{X}_i^T \mathbf{X}_j (e_i - \xi_i)(e_j - \xi_j) \\ &\doteq A_{n1} + A_{n2} + A_{n3}. \end{aligned}$$

We consider the last part A_{n3} firstly. Note that

$$\begin{aligned} &E(A_{n3}|\mathbf{X}) \\ &= \frac{12}{n(n-1)(n+1)^2} E_X\Big\{ \sum_{i \neq j} \mathbf{X}_i^T \mathbf{X}_j \sum_{l=1}^{n}\sum_{k=1}^{n} (I(\varepsilon_l \leq \varepsilon_i + (\mathbf{X}_i - \mathbf{X}_l)^T \boldsymbol{\beta}) \\ &- I(\varepsilon_l \leq \varepsilon_i)) \times I(\varepsilon_k \leq \varepsilon_j + (\mathbf{X}_j - \mathbf{X}_k)^T \boldsymbol{\beta}) - I(\varepsilon_k \leq \varepsilon_j)) \Big\} \\ &= \frac{12}{n(n-1)(n+1)^2} E_X\Big\{ \sum_{i \neq j} \mathbf{X}_i^T \mathbf{X}_j \sum_{l=1}^{n}\sum_{k=1}^{n} E\Big[(I(\varepsilon_l \leq \varepsilon_i + (\mathbf{X}_i - \mathbf{X}_l)^T \boldsymbol{\beta}) \\ &- I(\varepsilon_l \leq \varepsilon_i)) \times I(\varepsilon_k \leq \varepsilon_j + (\mathbf{X}_j - \mathbf{X}_k)^T \boldsymbol{\beta}) - I(\varepsilon_k \leq \varepsilon_j)) \big| \varepsilon_i, \varepsilon_j \Big] \Big\} \end{aligned}$$

$$
\begin{aligned}
&= \frac{12}{n(n-1)(n+1)^2} E_X \Big\{ \sum_{i \neq j} \mathbf{X}_i^T \mathbf{X}_j \sum_{l=1}^n \sum_{k=1}^n (F(\varepsilon_i + (\mathbf{X}_i - \mathbf{X}_l)^T \boldsymbol{\beta}) - F(\varepsilon_i)) \\
&\times (F(\varepsilon_j + (\mathbf{X}_j - \mathbf{X}_k)^T \boldsymbol{\beta}) - F(\varepsilon_j)) \Big\} \\
&= \frac{12}{n(n-1)(n+1)^2} E_X \Big\{ \sum_{i \neq j} \mathbf{X}_i^T \mathbf{X}_j \sum_{l=1}^n \sum_{k=1}^n \Big(\int f^2(x) dx ((\mathbf{X}_i - \mathbf{X}_l)^T \boldsymbol{\beta}) \\
&+ \int f^{'}(\xi_{il}) f(x) dx ((\mathbf{X}_i - \mathbf{X}_l)^T \boldsymbol{\beta})^2 \Big) \\
&\times \Big(\int f^2(x) dx ((\mathbf{X}_j - \mathbf{X}_k)^T \boldsymbol{\beta}) + \int f^{'}(\xi_{jk}) f(x) dx ((\mathbf{X}_j - \mathbf{X}_k)^T \boldsymbol{\beta})^2 \Big) \Big\} \\
&= \frac{12}{n(n-1)(n+1)^2} E_X \Big\{ \sum_{i \neq j} \mathbf{X}_i^T \mathbf{X}_j \sum_{l=1}^n \sum_{k=1}^n \Big(\int f^2(x) dx \Big)^2 \\
&\times (\mathbf{X}_i - \mathbf{X}_l)^T \boldsymbol{\beta} (\mathbf{X}_j - \mathbf{X}_k)^T \boldsymbol{\beta} \Big\} \\
&+ \frac{24}{n(n-1)(n+1)^2} E_X \Big\{ \sum_{i \neq j} \mathbf{X}_i^T \mathbf{X}_j \sum_{l=1}^n \sum_{k=1}^n \int f^2(x) dx \int f^{'}(\xi_{jk}) f(x) dx \\
&\times (\mathbf{X}_i - \mathbf{X}_l)^T \boldsymbol{\beta} ((\mathbf{X}_j - \mathbf{X}_k)^T \boldsymbol{\beta})^2 \Big\} \\
&+ \frac{12}{n(n-1)(n+1)^2} E_X \Big\{ \sum_{i \neq j} \mathbf{X}_i^T \mathbf{X}_j \sum_{l=1}^n \sum_{k=1}^n (((\mathbf{X}_i - \mathbf{X}_l)^T \boldsymbol{\beta})^2 ((\mathbf{X}_j - \mathbf{X}_k)^T \boldsymbol{\beta})^2) \\
&\times \Big(\int f^{'}(\xi_{il}) f(x) dx \int f^{'}(\xi_{jk}) f(x) dx \Big) \Big\} \\
&\doteq A_{n31} + A_{n32} + A_{n33},
\end{aligned}
$$

where E_X denotes the conditional expectation given $\mathbf{X}$ $E(\cdot|\mathbf{X})$ and ξ_{jk} is a variable between ε_j and $\varepsilon_j + (\mathbf{X}_j - \mathbf{X}_k)^T \boldsymbol{\beta}$. Obviously, $E(A_{n31}) = \boldsymbol{\beta}^T \boldsymbol{\Sigma}^2 \boldsymbol{\beta} / \tau_\epsilon^2 + o(\sqrt{n^{-2} \mathrm{tr}(\boldsymbol{\Sigma}^2)})$. Note that

$$
\begin{aligned}
E(A_{n32}) &\leq C_1 E(\mathbf{X}_i^T \mathbf{X}_j (\mathbf{X}_i - \mathbf{X}_l)^T \boldsymbol{\beta} ((\mathbf{X}_j - \mathbf{X}_k)^T \boldsymbol{\beta})^2) \\
&= C_1 E(\boldsymbol{\beta}^T \boldsymbol{\Sigma} \mathbf{X}_j (\mathbf{X}_j^T \boldsymbol{\beta})^2) \\
&\leq C_2 \sqrt{E(\boldsymbol{\beta}^T \boldsymbol{\Sigma} \mathbf{X}_j \mathbf{X}_j^T \boldsymbol{\Sigma} \boldsymbol{\beta}) E((\mathbf{X}_j^T \boldsymbol{\beta})^4)} \\
&\leq C_3 \sqrt{\boldsymbol{\beta}^T \boldsymbol{\Sigma}^3 \boldsymbol{\beta} (\boldsymbol{\beta}^T \boldsymbol{\Sigma} \boldsymbol{\beta})^2}
\end{aligned}
$$

and

$$
\begin{aligned}
E(A_{n33}) &\leq C_4 E(\mathbf{X}_i^T \mathbf{X}_j ((\mathbf{X}_i - \mathbf{X}_l)^T \boldsymbol{\beta})^2 ((\mathbf{X}_j - \mathbf{X}_k)^T \boldsymbol{\beta})^2) \\
&= C_4 E((\mathbf{X}_i^T \boldsymbol{\beta})^2 \mathbf{X}_i^T \mathbf{X}_j (\mathbf{X}_j^T \boldsymbol{\beta})^2)
\end{aligned}
$$

$$\leq C_5\sqrt{E((\mathbf{X}_i^T\mathbf{X}_j)^2)E((\mathbf{X}_i^T\boldsymbol{\beta})^4(\mathbf{X}_j^T\boldsymbol{\beta})^4)}$$
$$\leq C_6\sqrt{\mathrm{tr}(\boldsymbol{\Sigma}^2)(\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta})^4}.$$

where C_i, $i=1,\ldots,6$ are all some positive constants which are independent of the samples. Thus, $E(A_{n3})=\boldsymbol{\beta}^T\boldsymbol{\Sigma}^2\boldsymbol{\beta}/\tau_\epsilon^2+o(\sqrt{n^{-2}\mathrm{tr}(\boldsymbol{\Sigma}^2)})$ under the alternative. Next, we consider the variance of A_{n3} which can be written as

$$\begin{aligned}
&\mathrm{var}(A_{n3})=\mathrm{var}\left(\frac{1}{n(n-1)}\sum_{i\neq j}\mathbf{X}_i^T\mathbf{X}_j(e_i-\xi_i)(e_j-\xi_j)\right)\\
&=O(n^{-2})E((\mathbf{X}_i^T\mathbf{X}_j)^2(e_i-\xi_i)^2(e_j-\xi_j)^2)\\
&\quad+O(n^{-1})E(\mathbf{X}_i^T\mathbf{X}_j\mathbf{X}_i^T\mathbf{X}_s(e_i-\xi_i)^2(e_j-\xi_j)(e_s-\xi_s))+O(n^{-1})E^2(A_{n3})\\
&\doteq B_{n1}+B_{n2}+o(n^{-2}\mathrm{tr}(\boldsymbol{\Sigma}^2)).
\end{aligned}$$

Similar to the arguments in the calculation of $E(A_{n3})$, we can obtain that

$$\begin{aligned}
B_{n1}=O(n^{-2})\Big\{&E\left((\mathbf{X}_i^T\mathbf{X}_j)^2((\mathbf{X}_i-\mathbf{X}_l)^T\boldsymbol{\beta})^2((\mathbf{X}_j-\mathbf{X}_k)^T\boldsymbol{\beta})^2\right)\\
&+E\left((\mathbf{X}_i^T\mathbf{X}_j)^2((\mathbf{X}_i-\mathbf{X}_l)^T\boldsymbol{\beta})^2((\mathbf{X}_j-\mathbf{X}_k)^T\boldsymbol{\beta})^4\right)\\
&+E\left((\mathbf{X}_i^T\mathbf{X}_j)^2((\mathbf{X}_i-\mathbf{X}_l)^T\boldsymbol{\beta})^4((\mathbf{X}_j-\mathbf{X}_k)^T\boldsymbol{\beta})^4\right)\Big\}\\
\doteq O(n^{-2})&(B_{n11}+B_{n12}+B_{n13}).
\end{aligned}$$

Firstly,

$$\begin{aligned}
B_{n11}&=E\left((\mathbf{X}_i^T\mathbf{X}_j)^2((\mathbf{X}_i-\mathbf{X}_l)^T\boldsymbol{\beta})^2((\mathbf{X}_j-\mathbf{X}_k)^T\boldsymbol{\beta})^2\right)\\
&=E\left((\mathbf{X}_i^T\mathbf{X}_j)^2(\mathbf{X}_i^T\boldsymbol{\beta})^2(\mathbf{X}_j^T\boldsymbol{\beta})^2\right)+2E\left((\mathbf{X}_i^T\mathbf{X}_j)^2(\mathbf{X}_l^T\boldsymbol{\beta})^2(\mathbf{X}_j^T\boldsymbol{\beta})^2\right)\\
&\quad+E\left((\mathbf{X}_i^T\mathbf{X}_j)^2(\mathbf{X}_l^T\boldsymbol{\beta})^2(\mathbf{X}_k^T\boldsymbol{\beta})^2\right)
\end{aligned}$$

By Lemma 1,

$$\begin{aligned}
&E\left((\mathbf{X}_i^T\mathbf{X}_j)^2(\mathbf{X}_i^T\boldsymbol{\beta})^2(\mathbf{X}_j^T\boldsymbol{\beta})^2\right)\\
&\quad=E((\mathbf{X}_i^T\boldsymbol{\beta})^2E((\mathbf{X}_j^T\boldsymbol{\beta})^2(\mathbf{X}_i^T\mathbf{X}_j)^2|\mathbf{X}_i))\\
&\quad\leq CE((\mathbf{X}_i^T\boldsymbol{\beta})^2(\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta}\mathbf{X}_i^T\boldsymbol{\Sigma}\mathbf{X}_i+\boldsymbol{\beta}^T\boldsymbol{\Sigma}\mathbf{X}_i\mathbf{X}_i^T\boldsymbol{\Sigma}\boldsymbol{\beta}))\\
&\quad=O(\boldsymbol{\beta}^T\boldsymbol{\Sigma}^3\boldsymbol{\beta}(\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta})+\mathrm{tr}(\boldsymbol{\Sigma}^2)(\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta})^2+(\boldsymbol{\beta}^T\boldsymbol{\Sigma}^2\boldsymbol{\beta})^2)\\
&E\left((\mathbf{X}_i^T\mathbf{X}_j)^2(\mathbf{X}_l^T\boldsymbol{\beta})^2(\mathbf{X}_j^T\boldsymbol{\beta})^2\right)\\
&\quad=\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta}E(\mathbf{X}_j^T\boldsymbol{\Sigma}\mathbf{X}_j\mathbf{X}_j^T\boldsymbol{\beta}\boldsymbol{\beta}^T\mathbf{X}_j)\\
&\quad=\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta}E(\mathbf{z}_j^T\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}\boldsymbol{\Gamma}\mathbf{z}_j\mathbf{z}_j^T\boldsymbol{\Gamma}^T\boldsymbol{\beta}\boldsymbol{\beta}^T\boldsymbol{\Gamma}\mathbf{z}_j)\\
&\quad=O(\boldsymbol{\beta}^T\boldsymbol{\Sigma}^3\boldsymbol{\beta}(\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta})+\mathrm{tr}(\boldsymbol{\Sigma}^2)(\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta})^2)\\
&E\left((\mathbf{X}_i^T\mathbf{X}_j)^2(\mathbf{X}_l^T\boldsymbol{\beta})^2(\mathbf{X}_k^T\boldsymbol{\beta})^2\right)\\
&\quad=\mathrm{tr}(\boldsymbol{\Sigma}^2)(\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta})^2.
\end{aligned}$$

So $B_{n11} = o(\mathrm{tr}(\mathbf{\Sigma}^2))$ under the alternative (2.4). Taking the same procedure as B_{n11}, we can show that

$$
\begin{aligned}
B_{n12} &= O(\boldsymbol{\beta}^T\mathbf{\Sigma}^3\boldsymbol{\beta}(\boldsymbol{\beta}^T\mathbf{\Sigma}\boldsymbol{\beta})^2 + (\boldsymbol{\beta}^T\mathbf{\Sigma}^2\boldsymbol{\beta})^2\boldsymbol{\beta}^T\mathbf{\Sigma}\boldsymbol{\beta} + \mathrm{tr}(\mathbf{\Sigma}^2)(\boldsymbol{\beta}^T\mathbf{\Sigma}\boldsymbol{\beta})^3)\\
B_{n13} &= O(\boldsymbol{\beta}^T\mathbf{\Sigma}^3\boldsymbol{\beta}(\boldsymbol{\beta}^T\mathbf{\Sigma}\boldsymbol{\beta})^3 + (\boldsymbol{\beta}^T\mathbf{\Sigma}^2\boldsymbol{\beta})^2(\boldsymbol{\beta}^T\mathbf{\Sigma}\boldsymbol{\beta})^2 + \mathrm{tr}(\mathbf{\Sigma}^2)(\boldsymbol{\beta}^T\mathbf{\Sigma}\boldsymbol{\beta})^4)\\
B_{n2} &= O(n^{-1}(\boldsymbol{\beta}^T\mathbf{\Sigma}^3\boldsymbol{\beta}\boldsymbol{\beta}^T\mathbf{\Sigma}\boldsymbol{\beta} + (\boldsymbol{\beta}^T\mathbf{\Sigma}^2\boldsymbol{\beta})^2))
\end{aligned}
$$

Thus, under the alternative (2.4), $\mathrm{var}(A_{n3}) = o(n^{-2}\mathrm{tr}(\mathbf{\Sigma}^2))$ and then $A_{n3} = ||\mathbf{\Sigma}\boldsymbol{\beta}||/\tau_\epsilon^2 + o_p(n^{-1}\sqrt{\mathrm{tr}(\mathbf{\Sigma}^2)})$.

Similarly, we can obtain that $E(A_{n2}) = 0$ and

$$
\mathrm{var}(A_{n2}) = O(n^{-1}\boldsymbol{\beta}^T\mathbf{\Sigma}^3\boldsymbol{\beta}) + O(\boldsymbol{\beta}^T\mathbf{\Sigma}\boldsymbol{\beta})O(n^{-1}\boldsymbol{\beta}^T\mathbf{\Sigma}^3\boldsymbol{\beta}) = o(n^{-2}\mathrm{tr}(\mathbf{\Sigma}^2))
$$

under the alternative (2.4). Thus, $A_{n2} = o_p(n^{-1}\sqrt{\mathrm{tr}(\mathbf{\Sigma}^2)})$.

Define $\eta_i = \sqrt{12}(F(\varepsilon_i) - \frac{1}{2})$. Next, we will prove that

$$
\frac{1}{n(n-1)}\sum_{i\neq j}\mathbf{X}_i^T\mathbf{X}_j\xi_i\xi_j = \frac{1}{n(n-1)}\sum_{i\neq j}\mathbf{X}_i^T\mathbf{X}_j\eta_i\eta_j + o_p(n^{-1}\sqrt{\mathrm{tr}(\mathbf{\Sigma}^2)}). \quad (5.1)
$$

Obviously,

$$
\begin{aligned}
&\frac{1}{n(n-1)}\sum_{i\neq j}\mathbf{X}_i^T\mathbf{X}_j(\xi_i\xi_j - \eta_i\eta_j)\\
&\qquad = \frac{1}{n(n-1)}\sum_{i\neq j}\mathbf{X}_i^T\mathbf{X}_j(\xi_i - \eta_i)\xi_j + \frac{1}{n(n-1)}\sum_{i\neq j}\mathbf{X}_i^T\mathbf{X}_j\eta_i(\xi_j - \eta_j)\\
&\qquad \doteq D_{n1} + D_{n2}.
\end{aligned}
$$

After some simple calculation, we can obtain that $\mathrm{var}(D_{n1}) = O(n^{-3}\mathrm{tr}(\mathbf{\Sigma}^2))$ and $\mathrm{var}(D_{n2}) = O(n^{-3}\mathrm{tr}(\mathbf{\Sigma}^2))$ from which (5.1) follows immediately.

Next, we will show that

$$
\frac{n}{\sqrt{2\mathrm{tr}(\mathbf{\Sigma}^2)}}\frac{2}{n(n-1)}\sum_{i<j}\mathbf{X}_i^T\mathbf{X}_j\eta_i\eta_j \xrightarrow{d} N(0,1).
$$

Define $\tilde{W}_{nk} = \sum_{i=2}^k Z_{ni}$ where $Z_{ni} = \sum_{j=1}^{i-1}\eta_i\eta_j\mathbf{X}_i^T\mathbf{X}_j/\sqrt{\frac{n(n-1)}{2}}$. Let

$$
\mathcal{F}_i = \sigma\{(\mathbf{X}_1^T, \varepsilon_1)^T, \ldots, (\mathbf{X}_i^T, \varepsilon_i)^T\}
$$

be the σ-field generated by $\{(\mathbf{X}_j^T, \varepsilon_j)^T, j \leq i\}$.

It is easy to show that $E(Z_{ni}|\mathcal{F}_{i-1}) = 0$ and it follows that $\{\tilde{W}_{nk}, \mathcal{F}_k; 2 \leq k \leq n\}$ is a zero mean martingale. Let $v_{ni} = E(Z_{ni}^2|\mathcal{F}_{i-1}), 2 \leq i \leq n$ and $V_n = \sum_{i=2}^n v_{ni}$. The central limit theorem will hold [6] if we can show

$$
\frac{V_n}{\mathrm{var}(\tilde{W}_{nn})} \xrightarrow{p} 1 \quad (5.2)
$$

and for any $\epsilon > 0$,

$$\sum_{i=2}^{n} \mathrm{tr}^{-1}(\mathbf{\Sigma}^2) E\left[Z_{ni}^2 I\left(|Z_{ni}| > \epsilon\sqrt{\mathrm{tr}(\mathbf{\Sigma}^2)}\right)|\mathcal{F}_{i-1}\right] \xrightarrow{p} 0. \tag{5.3}$$

It can be shown that

$$v_{ni} = \frac{2}{n(n-1)}\left(\sum_{j=1}^{i-1} \eta_j \mathbf{X}_j^T \mathbf{\Sigma} \mathbf{X}_j + 2\sum_{1\le j<k<i} \eta_j\eta_k \mathbf{X}_j^T \mathbf{\Sigma}\mathbf{X}_k\right).$$

Then,

$$\begin{aligned}\frac{V_n}{\mathrm{var}(\tilde{W}_{nn})} &= \frac{4}{n^2(n-1)^2\mathrm{tr}(\mathbf{\Sigma}^2\sigma^2)}\left\{\sum_{j=1}^{n-1} j\eta_j^2 \mathbf{X}_j^T\mathbf{\Sigma}\mathbf{X}_j + 2\sum_{1\le j<k\le n}\eta_j\eta_k\mathbf{X}_j^T\mathbf{\Sigma}\mathbf{X}_k\right\}\\ &\doteq C_{n1} + C_{n2}.\end{aligned}$$

Simple algebras lead to

$$\begin{aligned} E(C_{n1}) &= 1,\\ \mathrm{var}(C_{n1}) &= \frac{16}{n^4(n-1)^4\mathrm{tr}^2(\mathbf{\Sigma}^2)} E\left(\sum_{j=1}^{n-1} j^2(\eta_j^4(\mathbf{X}_j^T\mathbf{\Sigma}\mathbf{X}_j)^2 - \mathrm{tr}^2(\mathbf{\Sigma}^2))\right).\end{aligned}$$

Define $\mathbf{\Gamma}^T\mathbf{\Sigma}\mathbf{\Gamma} = (\omega_{kl})_{1\le k,l\le m}$. Under the diverging factor model,

$$\begin{aligned} E((\mathbf{X}_j^T\mathbf{\Sigma}\mathbf{X}_j)^2) = E((\mathbf{z}_j^T\mathbf{\Gamma}^T\mathbf{\Sigma}\mathbf{\Gamma}\mathbf{z}_j)^2) &= E\left(\left(\sum_{k=1}^m\sum_{l=1}^m \omega_{kl} z_{jk} z_{jl}\right)^2\right)\\ &= \sum_{k=1}^m\sum_{l=1}^m\sum_{s=1}^m\sum_{t=1}^m \omega_{kl}\omega_{st} E(z_{jk}z_{jl}z_{js}z_{jt})\\ &= (3+\Delta)\sum_{k=1}^m \omega_{kk}^2 + \sum_{k\ne l}^m \omega_{kl}^2\\ &= (2+\Delta)\sum_{k=1}^m \omega_{kk}^2 + \mathrm{tr}(\mathbf{\Sigma}^4) \le (3+\Delta)\mathrm{tr}(\mathbf{\Sigma}^4). \end{aligned} \tag{5.4}$$

Under the condition (C1), $E((\mathbf{X}_j^T\mathbf{\Sigma}\mathbf{X}_j)^2) = o(\mathrm{tr}^2(\mathbf{\Sigma}^2))$. Hence, $\mathrm{var}(C_{n1}) \to 0$ and then $C_{n1} \xrightarrow{p} 1$. Similarly, $E(C_{n2}) = 0$ and

$$\mathrm{var}(C_{n2}) = \frac{32}{n^4(n-1)^4}\left(\sum_{i=3}^{n}\frac{i(i-1)}{2} + \sum_{i=3}^{n-1}\frac{i(n-i)(i-1)}{2}\right)\frac{\mathrm{tr}(\mathbf{\Sigma}^4)}{\mathrm{tr}^2(\mathbf{\Sigma}^2)} \to 0$$

implies $C_{n2} \xrightarrow{p} 0$. Thus, (5.2) holds. It remains to show (5.3). Since

$$E\left[Z_{ni}^2 I\left(|Z_{ni}| > \epsilon\sqrt{\mathrm{tr}(\mathbf{\Sigma}^2)}\right)|\mathcal{F}_{i-1}\right] \le E(Z_{ni}^4|\mathcal{F}_{i-1})/(\epsilon^2\mathrm{tr}(\mathbf{\Sigma}^2))$$

we only need to show that

$$\sum_{i=2}^{n} E(Z_{ni}^4) = o(\mathrm{tr}^2(\mathbf{\Sigma}^2)).$$

Note that

$$\sum_{i=2}^{n} E(Z_{ni}^4) = O(n^{-4}) \sum_{i=2}^{n} E\left(\left(\sum_{j=1}^{i-1} \eta_i\eta_j \mathbf{X}_i^T\mathbf{X}_j\right)^4\right)$$

which can be decomposed as $3Q + P$ where

$$Q = O(n^{-4}) \sum_{i=2}^{n} \sum_{s\neq t}^{i-1} E\left(\mathbf{X}_i^T\mathbf{X}_s\mathbf{X}_s^T\mathbf{X}_i\mathbf{X}_i^T\mathbf{X}_t\mathbf{X}_t^T\mathbf{X}_i\right),$$
$$P = O(n^{-4}) \sum_{i=2}^{n} \sum_{s=1}^{i-1} E\left((\mathbf{X}_i^T\mathbf{X}_s)^4\right).$$

Note that $Q = O(n^{-1})E((\mathbf{X}_i^T\mathbf{\Sigma}\mathbf{X}_i)^2) = o(\mathrm{tr}^2(\mathbf{\Sigma}^2))$ by similar arguments in (5.4). Next, we consider the part P. Define $\mathbf{\Gamma}^T\mathbf{\Gamma} = (\nu_{kl})_{1\le k,l\le m}$.

$$\begin{aligned}
P &= O(n^{-4}) \sum_{i=2}^{n} \sum_{s=1}^{i-1} E\left((\mathbf{z}_i^T\mathbf{\Gamma}^T\mathbf{\Gamma}\mathbf{z}_s)^4\right) = O(n^{-4}) \sum_{i\neq j} E\left(\left(\sum_{k,l=1}^{m} \nu_{kl} z_{ik} z_{jl}\right)^4\right) \\
&= O(n^{-2})\Bigg(\sum_{k,l=1}^{m} \nu_{kl}^4 E(z_{ik}^4)E(z_{jl}^4) + \sum_{k\neq l}^{m}\sum_{s\neq t}^{m} v_{kl}^2 v_{st}^2 E(z_{ik}^2)E(z_{is}^2)E(z_{jl}^2)E(z_{jt}^2) \\
&\quad + 2\sum_{k=1}^{m}\sum_{s\neq t}^{m} v_{ks}^2 v_{kt}^2 E(z_{ik}^4)E(z_{js}^2 z_{jt}^2) \\
&\quad + \sum_{k\neq l}^{m}\sum_{s\neq t}^{m} v_{kl}v_{kt}v_{st}v_{sl} E(z_{ik}^2)E(z_{jl}^2)E(z_{is}^2)E(z_{jt}^2)\Bigg)
\end{aligned}$$

Note that $\mathrm{tr}^2(\mathbf{\Sigma}^2) = (\sum_{s,t} \nu_{st}^2)^2 = \sum_{k,l,s,t} \nu_{st}^2\nu_{kl}^2$ and

$$\sum_{k,l=1}^{m} \nu_{kl}^4 \le \left(\sum_{k,l} \nu_{kl}^2\right)^2, \quad \sum_{k=1}^{m}\sum_{s\neq t}^{m} v_{ks}^2 v_{kt}^2 \le \left(\sum_{k,l} \nu_{kl}^2\right)^2,$$
$$\sum_{k\neq l}^{m}\sum_{s\neq t}^{m} v_{kl}^2 v_{st}^2 \le \sum_{k,l,s,t} \nu_{st}^2\nu_{kl}^2, \quad \sum_{k\neq l}^{m}\sum_{s\neq t}^{m} v_{kl}v_{kt}v_{st}v_{sl} \le \sum_{k\neq l} \omega_{kl}^2 \le \sum_{k,l} \omega_{kl}^2 = \mathrm{tr}(\mathbf{\Sigma}^4)$$

Thus, under the condition (C1), $P = o(\mathrm{tr}^2(\mathbf{\Sigma}^2))$ and then (5.3) follows immediately. This complete the proof.

Acknowledgement

The authors thank the editor, associate editor, and two anonymous referees for their many helpful comments that have resulted in significant improvements in the article.

References

[1] Bai, Z. and Saranadasa, H. (1996). Effect of high dimension: by an example of a two sample problem. *Statistica Sinica*, **6**, 311–329. MR1399305

[2] Efron, B. and Tibshirani, R. (2007). On testing the significance of sets of genes. *The Annals of Applied Statistics*, **1**, 107–129. MR2393843

[3] Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society, Series B*, **70**, 849–911. MR2530322

[4] Goeman, J., van Houwelingen, J. C., and Finos, L. (2011). Testing against a high dimensional alternative in the generalized linear model: asymptotic type I error control. *Biometrika*, **98**, 381–390. MR2806435

[5] Goeman, J., Van De Geer, S. A., and Houwelingen, V. (2006). Testing against a highdimensional alternative. *Journal of the Royal Statistical Society, Series B*, **68**, 477–493. MR2278336

[6] Hall, P. and Heyde, C. C. (1980). *Martingale Limit Theory and Its Application*. New York: Academic Press. MR0624435

[7] Hettmansperger, T. P. and McKean, J. W. (1998). *Robust Nonparametric Statistical Methods*. London: Arnold. MR1604954

[8] Li, G. R., Peng, H., Zhang, J., and Zhu, L. X. (2012). Robust rank correlation based screening. *Annals of Statistics*, **40**, 1846–1877. MR3015046

[9] McKean, J. W. and Hettmansperger, T. P. (1976). Tests of hypotheses of the general linear models based on ranks. *Communications in Statistics-Theory and Methods*, **5**, 693–709. MR0436452

[10] Meinshausen, N. (2008). Hierarchical testing of variable importance. *Biometrika*, **95**, 265–278. MR2521583

[11] Newton, M., Quintana, F., Den Boon, J., Sengupta, S., and Ahlquist, P. (2007). Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. *The Annals of Applied Statistics*, **1**, 85–106. MR2393842

[12] Redfern, C. H., Coward, P., Degtyarev, M. Y., Lee, E. K., Kwa, A. T., Hennighausen, L., Bujard, H., Fishman, G. I., and Conklin, B. R. (1999). Conditional expression and signaling of a specifically designed Gi-coupled receptor in transgenic mice. *Nature Biotechnology*, **17**, 165–169.

[13] Srivastava, M. S. (2009). A test for the mean vector with fewer observations than the dimension under non-normality. *Journal of Multivariate Analysis*, **100**, 518–532. MR2483435

[14] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, **102**, 15545–15550.

[15] Wang, H. (2009). Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association*, **104**, 1512–1524. MR2750576

[16] Wang, L. (2009). Wilcoxon-type generalized Bayesian information criterion. *Biometrika*, **96**, 163–173. MR2482142

[17] 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. MR2816719