Statistica Sinica

Likelihood Ratio Test in Multivariate Linear Regression: from Low to High Dimension

Yinqiu He[1], Tiefeng Jiang[2], Jiyang Wen[3], Gongjun Xu[1]

[1]*Department of Statistics, University of Michigan*
[2]*School of Statistics, University of Minnesota*
[3]*Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health*

Abstract: Multivariate linear regressions are widely to model the associations between multiple related responses and a set of predictors. To infer such associations, researchers often test the structure of the regression coefficients matrix, usually using a likelihood ratio test (LRT). Despite their popularity, classical χ^2 approximations for LRTs are known to fail in high-dimensional settings, where the dimensions of the responses and the predictors (m, p) are allowed to grow with the sample size n. Although various corrected LRTs and other test statistics have been proposed, few studies have examined the important question of when the classic LRT starts to fail. An answer to this would provide insights for practitioners, especially when analyzing data in which m/n and p/n are small, but not negligible. Moreover, the power of the LRT in high-dimensional data analyses remains under-researched. To address these issues, the first part of this work determines the asymptotic boundary at which the classical LRT fails, and develops a corrected limiting distribution for the LRT with a general asymptotic regime. The second part of this work examines the power of the LRT in high-dimensional settings. In addition to advancing the current understanding of the asymptotic behavior of the LRT under an alternative hypothesis, these results motivate the development

of a more powerful LRT. The third part of this work considers the setting in which $p > n$, where the LRT is not well defined. We propose a two-step testing procedure. First, we perform a dimension reduction, and then we apply the proposed LRT. Theoretical properties are developed to ensure the validity of the proposed method, and simulations demonstrate that the method performs well.

Key words and phrases: High dimension, Likelihood ratio test, Multivariate linear regression

1. Introduction

Multivariate linear regressions are widely used in econometrics, financial engineering, psychometrics, and many other areas to model the relationships between multiple related responses and a set of predictors. Suppose we have n observations of m-dimensional responses $\mathbf{y}_i = (y_{i,1}, \ldots, y_{i,m})^\intercal$ and p-dimensional predictors $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,p})^\intercal$, for $i = 1, \ldots, n$. Let $Y = (\mathbf{y}_1, \ldots, \mathbf{y}_n)^\intercal$ be the $n \times m$ response matrix, and $X = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\intercal$ be the $n \times p$ design matrix. The multivariate linear regression model assumes $Y = XB + E$, where B is a $p \times m$ matrix of unknown regression parameters, and $E = (\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_n)^\intercal$ is an $n \times m$ matrix of regression errors, where $\boldsymbol{\epsilon}_i$ is independently sampled from an m-dimensional Gaussian distribution $\mathcal{N}(\mathbf{0}, \Sigma)$.

Under the multivariate linear regression model, we are interested in testing the null hypothesis $H_0 : CB = \mathbf{0}_{r\times m}$, where C is an $r \times p$ matrix of rank $r \leq p$, and $\mathbf{0}_{r\times m}$ is an all-zero matrix of size $r \times m$. This is often called a general linear hypothesis in multivariate analyses, and has been widely used in multivariate analysis

of variance (see, e.g., Muirhead, 2009). The choice of the testing matrix C depends on the application. For instance, if B is partitioned as $B^\intercal = [B_1^\intercal, B_2^\intercal]$, where B_1 is an $r \times m$ matrix, then the null hypothesis of $B_1 = \mathbf{0}_{r\times m}$ is equivalent to taking $C = [I_r, \mathbf{0}_{r\times(p-r)}]$, which can be used to test the significance of the first r predictors of X. Another example is to test the equivalence of the effects of a set of $r+1$ predictors (e.g., different levels of categorical variables), where $C = [I_r, \mathbf{0}_{r\times(p-r-1)}, -\mathbf{1}_r]$, and $\mathbf{1}_r$ represents an r-dimensional vector of ones.

To test $H_0 : CB = \mathbf{0}_{r\times m}$, a popular approach in the literature is the likelihood ratio test (LRT) (Anderson, 2003; Muirhead, 2009). Specifically, when $n > m + p$, Σ is positive definite, and X has rank p, then the LRT statistic is $L_n = \det(S_E)^{n/2}/\{\det(S_E + S_X)^{n/2}\}$. Here $S_E = Y^\intercal[I - X(X^\intercal X)^{-1}X^\intercal]Y$ and $S_X = (C\hat{B})^\intercal[C(X^\intercal X)^{-1}C^\intercal]^{-1}C\hat{B}$ are the residual sum of squares and the regression sum of squares matrices, respectively, and $\hat{B} = (X^\intercal X)^{-1}X^\intercal Y$ is the least squares estimator. Assuming m and p are fixed, it is well known that $-2\log L_n$ converges weakly to a χ^2 distribution as $n \to \infty$ under the null hypothesis (Anderson, 2003).

However, in high-dimensional settings, where the dimension parameters (p, m, r) are allowed to increase with n, the LRT suffers from several issues. First, under the null hypothesis, the limiting distribution of $-2\log L_n$ may longer be a χ^2 distribution. The failure of the χ^2 approximations of LRT distributions under high dimensions has been studied in various model settings. For instance, Bai et al. (2009) examined

two LRTs for covariance matrices. They showed that the χ^2 approximations perform poorly, and thus proposed corrected normal limiting distributions. Jiang and Yang (2013) and Jiang and Qi (2015) studied classical LRTs for sample means and covariance matrices, showing that the χ^2 approximations fail as the dimensions increase. Moreover, Bai et al. (2013) considered the LRT for linear hypotheses in high-dimensional multivariate linear regressions. They demonstrated the failure of the χ^2 approximation and derived a corrected LRT. Note that Bai et al. (2013) only considered high-dimensional settings where m, r, and $n - p$ are proportional to each other, with $m \leq r$. Despite these works, it is still unclear under which asymptotic regimes the χ^2 approximation of a LRT starts to fail. An answer to this question would provide insights for practitioners, especially when analyzing data in which m/n and p/n are small, but not negligible.

The second problem with the LRT is its power performance under high-dimensional alternative hypotheses. When $n > p + m$, $-2 \log L_n = n \sum_{i=1}^{\min\{m,r\}} \log(1 + \lambda_i)$, where λ is an eigenvalue of $S_X^{1/2} S_E^{-1} S_X^{1/2}$. Therefore, we expect the asymptotic power of the LRT to depend on an averaged effect of all eigenvalues. However, few studies have examined the eigenvalues of the random matrix $S_X^{1/2} S_E^{-1} S_X^{1/2}$ under alternative hypotheses.

The third issue with the LRT arises when the dimension parameters p and m are large, such that $n < p+m$. In this situation, the LRT is not well defined, owing to the

singularity of the matrix S_E. This excludes the LRT from many high-dimensional applications with $p > n$ or $m > n$ (e.g., Donoho, 2000; Fan et al., 2014). When $m > n$, the linear hypothesis testing problem has been studied in depth for specific submodels, such as the one-way MANOVA (Srivastava and Fujikoshi, 2006; Hu et al., 2017; Zhou et al., 2017; Cai and Xia, 2014, etc.). Li et al. (2018) recently proposed a modified LRT for general linear hypothesis tests using spectral shrinkage. However, these works assume that p is fixed.

This study aims to address the above problems. First, under the null hypothesis, we derive the asymptotic boundary at which the χ^2 approximation fails as the dimension parameters (p, m, r) increase with the sample size n. Moreover, we develop a corrected limiting distribution for $\log L_n$ in a general asymptotic regime of (p, m, r, n). Second, under alternative hypotheses, we characterize the statistical power of $\log L_n$ in the high-dimensional setting. By analyzing the partial differential equations induced by the test statistic, we show that the LRT is powerful when the trace of the signal matrix $(CB)\Sigma^{-1}(CB)^\intercal$ is large, but that it loses power under a low-rank signal matrix. Given that alternatives tend to be unknown in practice, we propose an enhanced likelihood ratio test that is also powerful against low-rank alternative signal matrices. The power-enhanced test statistic combines the LRT statistic and the largest eigenvalue (Johnstone, 2008, 2009) to further improve the test power against low-rank alternatives. Third, when $n < p$ and the LRT is not well defined, we pro-

pose a two-step testing procedure: first, we reduce the dimensions of the covariates and responses, and then we use the proposed (enhanced) LRT. To control the estimation error induced by the dimension reduction in the first step, we employ a *repeated* data-splitting approach, and show that the asymptotic type-I error is well controlled under the null hypothesis. Simulation results confirm that the proposed approach performs well.

The rest of the paper is organized as follows. In Section 2, we examine when the classic LRT fails under the null hypothesis, and propose a corrected limiting distribution for $\log L_n$. In Section 3, we analyze the power of $\log L_n$ and propose a more powerful test statistic. In Section 4, we discuss the multi-split LRT procedure when $n < p$. Simulation studies and a real dataset analysis on breast cancer are reported in Sections 5 and 6, respectively.

2. When the LRT begins to fail?

In traditional multivariate regression analyses, where the dimension parameters (p, m, r) are considered fixed, the χ^2 approximation of the LRT,

$$-2\log L_n \xrightarrow{D} \chi^2_{mr}, \quad \text{as } n \to \infty, \tag{2.1}$$

is used for $H_0 : CB = \mathbf{0}_{r\times m}$ (Muirhead, 2009; Anderson, 2003), where $\xrightarrow{D}$ denotes the convergence in distribution. However, this χ^2 approximation is known to perform poorly in high-dimensional applications (see, e.g., Bai and Saranadasa, 1996; Jiang et al., 2012; Bai et al., 2009, 2013; Jiang and Yang, 2013).

When the three dimension parameters (m, p, r) are allowed to grow with n, it is of interest to examine the phase transition boundary where the χ^2 approximation fails. This is described in the following theorem.

Theorem 1. *Consider $n > p+m$ and $p \geq r$. Let $\chi^2_{mr}(\alpha)$ denote the upper α-quantile of a χ^2_{mr} distribution.*

(i) When $mr \to \infty$ and $\max\{p, m, r\}/n \to 0$ as $n \to \infty$, $P\{-2\log L_n > \chi^2_{mr}(\alpha)\} \to \alpha$, for any significance level α, if and only if

$$\lim_{n\to\infty} \sqrt{mr}(p + m/2 - r/2)n^{-1} = 0. \tag{2.2}$$

(ii) When mr is finite, $P\{-2\log L_n > \chi^2_{mr}(\alpha)\} \to \alpha$, if and only if $\lim_{n\to\infty} p/n = 0$.

Theorem 1 gives the sufficient and necessary condition on (m, p, r, n) such that the χ^2 approximation (2.1) fails. Note that although (2.2) is obtained when $mr \to \infty$, (2.2) becomes $\lim_{n\to\infty} p/n = 0$ for finite m and r, supporting the conclusion when mr is finite. To further examine the implications of (2.2), we consider two special cases. Specifically, let $m = \lfloor n^\eta \rfloor$ and $p = \lfloor n^\epsilon \rfloor$, with η and $\epsilon \in (0, 1)$, where $\lfloor\cdot\rfloor$ denotes

the floor of a number. When r is fixed, (2.2) implies $\sqrt{m}(p+m/2) = o(n)$; that is, $\max\{\epsilon, \eta\} + \eta/2 < 1$. When $r = p = \lfloor n^{\epsilon} \rfloor$, (2.2) implies $\sqrt{mp}(p+m) = o(n)$; that is, $\max\{\epsilon, \eta\} + (\eta+\epsilon)/2 < 1$. For these two cases, we give two corresponding (η, ϵ)-regions in Figure 1 satisfying constraint (2.2). In these two regions, when ϵ approaches zero, the largest η approaches $2/3$. Therefore, when p is small, the largest m such that (2.2) holds is of order $n^{2/3}$. The same is true for the cases of fixed r and $r = p$, because p is small and $r \leq p$. In addition, when η goes to zero, the largest ϵ-values under fixed r and $r = p$ converge to one and $2/3$, respectively. Thus, when m is small, the largest p-values satisfying (2.2) are of order n and $n^{2/3}$, respectively. Moreover, when $m = p$, the largest orders of m and p for the two cases are $n^{2/3}$ and $n^{1/2}$, respectively.

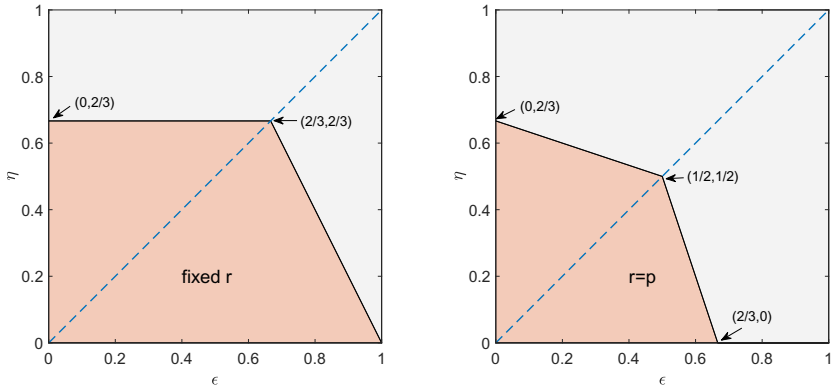

Figure 1: η versus ϵ when r is fixed (left) and $r = p$ (right)

To illustrate this phase transition phenomenon, we present a simple simulation experiment. We set $\Sigma = I_m$, and estimate the type-I errors of the χ^2 approximation (2.1) using 10^4 repetitions under the following four cases: (a) fixed $m = r = 2$ and $p =$

$\lfloor n^\eta \rfloor$; (b) fixed $p = r = 2$ and $m = \lfloor n^\eta \rfloor$; (c) fixed $m = 2$ and $p = r = \lfloor n^\eta \rfloor$; and (d) $p = m = r = \lfloor n^\eta \rfloor$. In all cases, $\eta \in \{1/24, \ldots, 23/24\}$. In Figure 2, we plot the estimated type-I errors against the η-values for $n = 100$ and 300. The plots show consistent patterns with the theoretical results. In particular, when $p = m = r = \lfloor n^\eta \rfloor$, the χ^2 approximation begins to fail for η around 1/2. When p and r are fixed and $m = \lfloor n^\eta \rfloor$ and when m is fixed and $p = r = \lfloor n^\eta \rfloor$, the χ^2 approximation begins to fail for η around 2/3. When m and r are fixed and $p = \lfloor n^\eta \rfloor$, the χ^2 approximation begins to fail for η larger than the other three cases, which is consistent with the theoretical results.

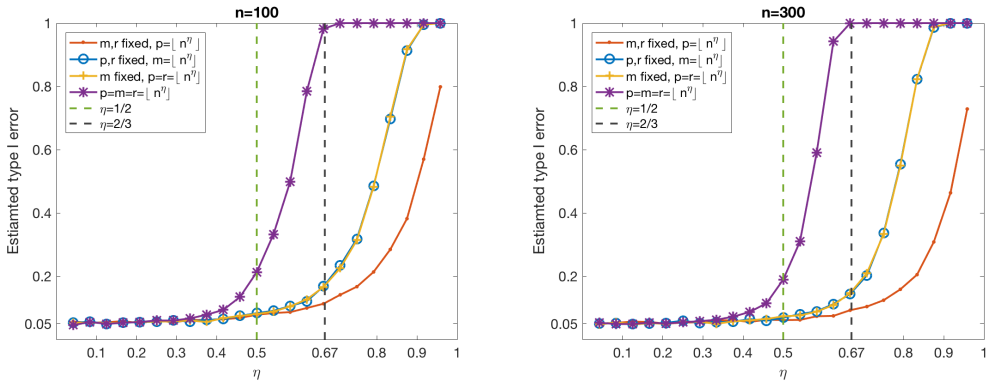

Figure 2: Estimated type I errors using χ^2 approximation (2.1)

Note that the sufficient and necessary constraint (2.2) also characterizes the bias of the χ^2 approximation. Specifically, under the conditions of Theorem 1, $\mathrm{E}(-2\log L_n - \chi^2_{mr})/\sqrt{\mathrm{var}(\chi^2_{mr})} = \sqrt{mr}(p + m/2 - r/2 + 1/2)n^{-1}\{1 + o(1)\}$. Thus, when (p, m, r) are large, such that (2.2) is violated and the χ^2 approximation fails,

the bias of the χ^2 approximation increases with $\sqrt{mr}(p + m/2 - r/2 + 1/2)n^{-1}$. This can be seen in Figure 2, and is supported by the simulations reported in Section 5.

In the classic regime with fixed m and p, researchers have also proposed the Bartlett correction of the LRT, $-2\rho \log L_n \xrightarrow{D} \chi^2_{mr}$, where $\rho = 1 - (p - r/2 + m/2 + 1/2)/n$. In particular, for any $z \in \mathbb{R}$, this corrected approximation gets rid of the first-order approximation error $O(n^{-1})$; that is, for any z, $P(-2\rho \log L_n < z) - P(\chi^2_{mr} < z) = O(n^{-2})$ when m and p are fixed. Similarly to Theorem 1, the χ^2 approximation with the Bartlett correction also fails as m and p increase with n. The phase transition boundary is characterized in the following result.

Theorem 2. *Consider $n > p + m$ and $p \geq r$.*

(i) When $mr \to \infty$ and $\max\{p, m, r\}/n \to 0$ as $n \to \infty$, $P\{-2\rho \log L_n > \chi^2_{mr}(\alpha)\} \to \alpha$, for any significance level α, if and only if $\lim_{n\to\infty} \sqrt{mr}(r^2 + m^2)n^{-2} = 0$.

(ii) When mr is finite, $P\{-2\rho \log L_n > \chi^2_{mr}(\alpha)\} \to \alpha$, if and only if $n - p \to \infty$.

Theorem 2 suggests that when m and r are fixed, the corrected LRT approximation holds when $n - p \to \infty$. When $mr \to \infty$, the phase transition threshold in Theorem 2 only involves m and r. In particular, when r is fixed and $m = \lfloor n^\eta \rfloor$, and when m is fixed and $r = \lfloor n^\eta \rfloor$, the χ^2 approximation with the Bartlett correction fails when $\eta \geq 4/5$; when $m = r = \lfloor n^\eta \rfloor$, the corrected approximation fails when $\eta \geq 2/3$.

To illustrate this phenomenon, we present a numerical experiment on the χ^2 approximation with the Bartlett correction in Figure 3. The setup is the same as

that shown in Figure 2. The results show that when m and r are fixed and $p = \lfloor n^\eta \rfloor$, the type-I errors are well controlled for large η approaching one. Moreover, when p and r are fixed and $m = \lfloor n^\eta \rfloor$ and when m is fixed and $p = r = \lfloor n^\eta \rfloor$, the corrected χ^2 approximation begins to fail around $\eta = 4/5$. When $p = m = r = \lfloor n^\eta \rfloor$, the corrected χ^2 approximation begins to fail around $\eta = 2/3$. These numerical results are consistent with the theory.

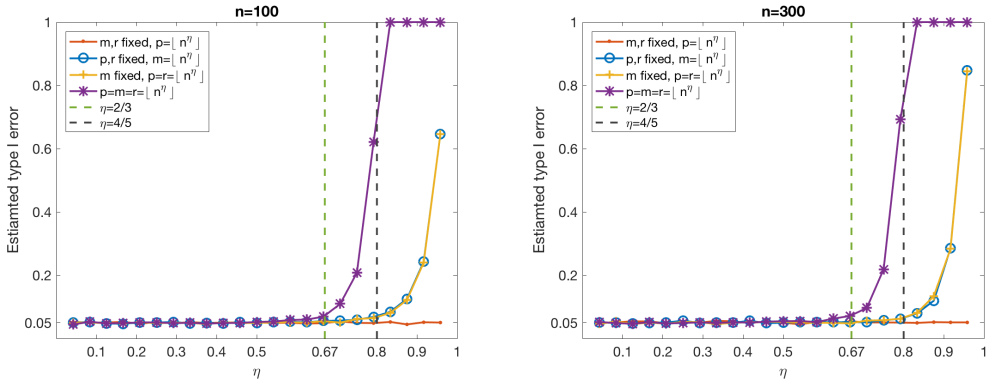

Figure 3: Estimated type-I error using the χ^2 approximation with the Bartlett correction

More generally, to have a unified limiting distribution for analyzing high-dimensional data under a general asymptotic region of (m, p, r, n), we derive a corrected normal limiting distribution for the LRT statistic.

Theorem 3. *When $n > p + m$, $p \geq r$, $mr \to \infty$, and $n - p - \max\{m - r, 0\} \to \infty$ as $n \to \infty$, the LRT statistic L_n has the corrected form T_1 satisfying*

$$T_1 := \frac{-2 \log L_n + \mu_n}{n\sigma_n} \xrightarrow{D} \mathcal{N}(0, 1), \tag{2.3}$$

where $\sigma_n^2 = 2\log(n+r-p-m)(n-p) - 2\log(n-p-m)(n+r-p)$, *and*

$$\mu_n = n(n-m-p-1/2)\log\frac{(n+r-p-m)(n-p)}{(n-p-m)(n+r-p)} + nr\log\frac{(n+r-p-m)}{(n+r-p)} + nm\log\frac{(n-p)}{(n+r-p)}.$$

Theorem 3 covers the asymptotic regime where $mr \to \infty$, $\max\{p, m, r\}/n \to 0$, and the constraint (2.2) holds. Under this region, we can show that $\mu_n \to -mr$ and $(n\sigma_n)^2 \to 2mr$, which are consistent with the mean and variance, respectively, of the χ^2_{mr} approximation. In addition, although Theorem 3 requires $mr \to \infty$, the normal approximation (2.3) could still perform well when m or r is small, as long as mr is sufficiently large. The simulations in Section 5 show that the χ^2 and normal approximations can perform similarly in low dimensions.

Alternatively, under some high-dimensional settings, we can check that no χ^2 or even noncentral χ^2 distribution matches the asymptotic mean and variance of $-2\log L_n$ in Theorem 3. Specifically, if the distribution of $-2\log L_n$ can be approximated by some χ^2 distribution, we should have $-(n\sigma_n)^2/\mu_n \to 2$, which is, however, not satisfied as $p/n, m/n$ and r/n increase. If the distribution of $-2\log L_n$ can be approximated by some noncentral χ^2 distribution with degrees of freedom k_n, then we should have $k_n = -2\mu_n - n^2\sigma_n^2/2$, which can become negative as $p/n, m/n$ and r/n increase. Thus, the χ^2-type approximation for $-2\log L_n$ can fail fundamentally

under high dimensions.

Remark 1. *A similar result on the asymptotic normality of* $\log L_n$ *in Theorem 3 is proved in Zheng (2012) and Bai et al. (2013). However, there are several differences between our result and theirs. First, our asymptotic regime is more general. Specifically, Zheng (2012) and Bai et al. (2013) require that* $m < r$, $\min\{m, r\} \to \infty$, *and* $m/(n-p)$ *converges to a constant in* $(0, 1)$, *whereas we only need* $mr \to \infty$ *and* $n - p - \max\{m - r, 0\} \to \infty$. *Our analysis covers the case when* $m/(n-p) \to 0$, *and even when the limit does not exist. Second, the proofs of Zheng (2012) and Bai et al. (2013) are based on random matrix theory, whereas we prove Theorem 3 using a moment-generating function technique motivated by the work of Jiang and Yang (2013).*

3. Power analysis and an enhanced LRT

Although the limits of LRTs for high-dimensional data have been explored for various problems, the power of these tests is less well studied and remains a challenging problem, as discussed in Jiang and Yang (2013). In this section, we focus on the high-dimensional multivariate linear regression and analyze the power of the LRT statistic. Moreover, based on the theoretical results, we propose a more powerful LRT.

To examine the power of the LRT statistic, we introduce the classic canonical

form of the LRT problem, which expresses $H_0 : CB = \mathbf{0}$ in an equivalent form as follows (Muirhead, 2009). Specifically, consider the matrix decomposition $X = O[I_p, \mathbf{0}_{p\times(n-p)}]^\intercal D$, where O is an $n\times n$ orthogonal matrix, and D is a $p\times p$ nonsingular real matrix. Given D, we have a similar decomposition $CD^{-1} = \mathbb{E}[I_r, \mathbf{0}_{r\times(p-r)}]V$, where $\mathbb{E}$ is an $r\times r$ nonsingular matrix, and V is a $p\times p$ orthogonal matrix. Therefore, $CB = CD^{-1}DB = \mathbb{E}[I_r, \mathbf{0}_{r\times(p-r)}]VDB$, and, thus, $H_0 : CB = \mathbf{0}_{r\times m}$ is equivalent to $M_1 = \mathbf{0}_{r\times m}$, where we define $M_1 = [I_r, \mathbf{0}_{r\times(p-r)}]VDB = \mathbb{E}^{-1}CB$.

We next describe the relationship between M_1 and the LRT statistic through a linear transformation of Y. Let V_1 denote the first r rows of V. Define $Y_1^* = [V_1, \mathbf{0}_{r\times(n-p)}]O^\intercal Y$ and $Y_2^* = [\mathbf{0}_{(n-p)\times p}, I_{n-p}]O^\intercal Y$. We then know that $Y_1^{*\intercal}Y_1^* = S_X$ and $Y_2^{*\intercal}Y_2^* = S_E$. We further define $\tilde{S}_X = \Sigma^{-1/2}S_X\Sigma^{-1/2}$, $\tilde{S}_E = \Sigma^{-1/2}S_E\Sigma^{-1/2}$, and $\Omega = \Sigma^{-1/2}M_1^\intercal M_1\Sigma^{-1/2}$. Then, we can write the LRT statistic $-2\log L_n = n\sum_{i=1}^{\min\{m,r\}}\log(1+\lambda_i)$, where λ is an eigenvalue of $\tilde{S}_E^{-1}\tilde{S}_X$. Given that $\mathrm{E}(\tilde{S}_E^{-1}\tilde{S}_X) = (rI_m + \Omega)/(n-p)$ (Muirhead, 2009), we expect the power of the LRT to depend on an averaged effect of all eigenvalues of Ω.

We focus on the alternatives where the signal matrix Ω is of low rank and (p, m, r) increase proportionally with n. In particular, we assume Ω has a fixed rank m_0, and write $\Omega = n\Delta$, where Δ has fixed nonzero eigenvalues $\delta_1, \ldots, \delta_{m_0}$. Note that this is reasonable when the entries in $M_1\Sigma^{-1/2}$ are $O(1)$, because the entries in Ω could be $O(n)$, with r proportional to n. The following theorem specifies how the power of

the LRT statistic T_1 depends on the eigenvalues of Ω.

Theorem 4. *Consider the setting where (p, m, r) increase proportionally with n, and $p/n = \rho_p$, $m/n = \rho_m$, and $r/n = \rho_r$, where $\rho_p, \rho_m, \rho_r \in (0, 1)$ are fixed constants and $\rho_p + \rho_m < 1$. Given $\Delta = \Omega/n$ with fixed nonzero eigenvalues $\delta_1, \ldots, \delta_{m_0}$, define $W_\Delta = \sum_{j=1}^{m_0} \log[1 + \delta_j(1 + \rho_r - \rho_p)^{-1}]$. There exists a constant $A_1 > 0$ such that $P(T_1 > z_\alpha) \to 1-\Phi(z_\alpha - A_1 W_\Delta)$, where $\Phi(\cdot)$ and z_α denote the cumulative distribution function and the upper α-quantile, respectively, of $\mathcal{N}(0, 1)$.*

Theorem 4 establishes the relationship between the eigenvalues of Ω and the power of T_1 under high-dimensional and low-rank signals. It implies that when W_Δ is large, T_1 has high power. Alternatively, the LRT could be highly underpowered when W_Δ is small. Because in real applications the truth is usually unknown, we require a testing procedure with high statistical power against various alternatives.

To enhance the power of the LRT, we propose combining it with Roy's test statistic based on the largest eigenvalue of $S_E^{-1} S_X$ (Roy, 1953). In particular, Johnstone (2008, 2009) extended Roy's test to high-dimensional settings, and proposed the largest eigenvalue test statistic $T_2 = [\log\{\theta_{n,1}/(1 - \theta_{n,1})\} - \tilde{\mu}_n]/\tilde{\sigma}_n$. Here, $\theta_{n,1} = \lambda_{\max}\{(S_E + S_X)^{-1} S_E\}$, with $\lambda_{\max}(\cdot)$ denoting the largest eigenvalue, and $\tilde{\mu}_n = 2 \times \log\tan\{(\phi + \gamma/2)\}$ and $\tilde{\sigma}_n^3 = 16(n - p + r - 1)^{-2}\{\sin^2(\phi + \gamma)\sin\phi\sin\gamma\}^{-1}$, with $\sin^2(\gamma/2) = \{\min(m, r) - 1/2\}/(n - p + r - 1)$ and $\sin^2(\phi/2) = \{\max(m, r) - 1/2\}/(n - p + r - 1)$. Moreover, Johnstone (2008) proved that under the high-dimensional null

hypothesis, $T_2 \xrightarrow{D} \mathcal{TW}$, where $\mathcal{TW}$ denotes a Tracy–Widom distribution. Under the alternative hypothesis, Dharmawansa et al. (2018) studied the spiked alternative with $\Omega = rUHU^\intercal$, where U is an $m \times m_0$ matrix with orthonormal columns and fixed m_0, and $H = \text{diag}(h_1, \ldots, h_{m_0})$ with $h_1 > \ldots > h_{m_0}$. They showed that the phase transition threshold for h is a constant that depends on the limit of $(p/n, m/n, r/n)$. Note that with fixed r/n, there exists a constant $c_2 > 0$ such that $\delta_1 = c_2 h_1$. This implies that when δ_1 is a sufficiently large constant, the power of T_2 can converge to one, whereas the LRT statistic T_1 may only have power less than one, by Theorem 4. On the other hand, when δ_1 is below the phase transition threshold, T_1 may be more powerful than T_2.

We therefore propose a combined test statistic $T_3 = T_1 + T_2 * I(T_2 \geq F_n)$, where F_n is a positive constant. With properly chosen F_n, the proposed test statistic T_3 may enhance the power of T_1 under alternative hypotheses, whereas $T_3 \xrightarrow{D} \mathcal{N}(0,1)$ under H_0. Specifically, under the null hypothesis, the type-I error rate of T_3 is controlled if $P\{T_2 \geq F_n\} \to 0$. On the other hand, under alternative hypotheses, we have $P(T_3 > z_\alpha) \geq P(T_1 > z_\alpha)$ because $T_2 * I\{T_2 > F_n\} \geq 0$ for $F_n > 0$. This guarantees that the power of T_3 is at least as large as that of the LRT statistic T_1. Moreover, consider the case when W_Δ is relatively small, but δ_1 is significantly above the phase transition threshold, where T_2 is more powerful than T_1. Then if F_n does not grow too quickly, T_3 would also be powerful. Thus, we can choose F_n to be a

slow-varying function, in which case the combined test statistic T_3 may improve the power of T_1 with little size distortion. Through extensive simulation studies, we find $F(n) = \max\{\log\log n, 2\}$ exhibits good performance; please see Section 5.

4. Likelihood ratio test when $p > n$

When the number of predictors is large, such that $p > n$, S_E becomes singular, and the test statistics T_1, T_2, and T_3 cannot be applied directly. To deal with this issue, we propose a multiple data-splitting procedure that repeatedly splits the data into two random subsets. We use the first subset to perform the dimension reduction and obtain a manageable size of predictors. Then we apply the proposed LRT to the second subset. The test statistics from different data splittings are aggregated to provide the final test statistic. The random splits of data ensure correct size control of the test's type-I error. Similar ideas are used in other high-dimensional problems (Meinshausen et al., 2009; Berk et al., 2013, etc.). We next describe the proposed procedure.

Consider the setting when $p > n$ and $m < n$. Denote $B = [\mathbf{b}_1, \ldots, \mathbf{b}_p]^\intercal$ and $\mathcal{M}_* = \{k : \mathbf{b}_k \neq \mathbf{0}, 1 \leq k \leq p\}$. We assume a "sparsity" structure in which the responses depend only on a subset of the predictors (or transformed predictors), such that $n > m + |\mathcal{M}_*|$. Let $X_{\mathcal{M}_*}$ be the $n \times |\mathcal{M}_*|$ submatrix of X with columns indexed by $\mathcal{M}_*$, and let $B_{\mathcal{M}_*}$ be the $|\mathcal{M}_*| \times m$ submatrix of B with rows indexed by $\mathcal{M}_*$.

The underlying model then satisfies $Y = X_{\mathcal{M}_*} B_{\mathcal{M}_*} + E$. Under this model, for any subset $\mathcal{M} \subseteq \{1, \ldots, p\}$ such that $\mathcal{M} \supseteq \mathcal{M}_*$ and $n > m + |\mathcal{M}|$, testing $CB = \mathbf{0}$ is equivalent to $C_{\mathcal{M}} B_{\mathcal{M}} = \mathbf{0}$, and the LRT is then applicable. Here, $C_{\mathcal{M}}$ denotes the $r \times |\mathcal{M}|$ submatrix of C with columns indexed by $\mathcal{M}$, and $B_{\mathcal{M}}$ denotes the $|\mathcal{M}| \times m$ submatrix of B with rows indexed by $\mathcal{M}$.

To obtain such a set $\mathcal{M} \supseteq \mathcal{M}_*$, we propose a screening method for a multivariate linear regression. The seminal work of Fan and Lv (2008) first introduced a sure independence screening procedure that significantly reduces the number of predictors, while preserving the true linear model with an overwhelming probability. This procedure has been extended in various settings (e.g., Fan and Song, 2010; Wang and Leng, 2016; Barut et al., 2016). However, many of these works focus on the settings with a univariate response variable.

To use the joint information from multivariate response variables, we propose a screening method that selects the columns of X based on their canonical correlations with Y. The canonical correlation is a widely used dimension-reduction criterion inferring information from cross-covariance matrices in a multivariate analysis (Muirhead, 2009). Specifically, for each column vector $\mathbf{x}^j = (x_{1,j}, \ldots, x_{n,j})^\intercal$, for $j = 1, \ldots, p$, we first compute its canonical correlation with Y, denoted by

$$\omega_j = \max_{\mathbf{a} \in \mathbb{R}^m} \frac{\mathbf{a}^\intercal (Y - \mathbf{1}_n \bar{Y})^\intercal (\mathbf{x}^j - \bar{x}^j \mathbf{1}_n)}{\sqrt{\{\mathbf{a}^\intercal (Y - \mathbf{1}_n \bar{Y})^\intercal (Y - \mathbf{1}_n \bar{Y}) \mathbf{a}\} \times \{(\mathbf{x}^j - \bar{x}^j \mathbf{1}_n)^\intercal (\mathbf{x}^j - \bar{x}^j \mathbf{1}_n)\}}},$$

where $\bar{x}^j = \sum_{i=1}^n x_{i,j}/n$, $\bar{Y}$ is the row mean vector of Y, and $\mathbf{1}_n$ is an all-one column vector of length n. Then, for $0 < \delta < 1$, we select $\lfloor \delta p \rfloor$ columns of X with the highest canonical correlations with Y, and define the selected column set as $\mathcal{M}_\delta = \{j : |\omega_j|$ is among the largest $\lfloor \delta p \rfloor$ of all, $1 \le j \le p\}$. In practice, we choose an integer $\lfloor \delta p \rfloor$, such that $n_T > \lfloor \delta p \rfloor + m$, to apply the LRT. On the other hand, we keep $\lfloor \delta p \rfloor$ large to increase the probability of $\mathcal{M}_\delta \supseteq \mathcal{M}_*$. The following theoretical result provides the desired screening property that $P(\mathcal{M}_* \subseteq \mathcal{M}_\delta) \to 1$ for properly chosen δ.

Theorem 5. *Under Conditions 1–3 given in Supplementary Material Section S5.1, for some constant $c_0 > 0$, $P(\mathcal{M}_* \subseteq \mathcal{M}_\delta) = 1 - O[\exp\{-c_0 n^{1-\iota}/\log n\}]$, where the constant $\iota < 1$ is defined in Condition 3.*

Remark 2. *When testing the coefficients of the first r predictors of X, such as $C = [I_r, \mathbf{0}_{r\times(p-r)}]$, we can keep the first r predictors, denoted by X_1, in the model, while screen the remaining predictors, denoted by X_2. In particular, we can apply the screening procedure to the residuals $\tilde{R}$, from the regression of Y on X_1, and X_2. More generally. when C is a matrix of rank r, we can use this conditional screening procedure by employing a linear transformation of the data. In particular, given the singular value decomposition $C = UVD^\intercal$, we can transform X and B into $\tilde{X} = XD$ and $\tilde{B} = D^\intercal B$, respectively. Then, testing $H_0 : CB = \mathbf{0}_{r\times m}$ is equivalent to testing $H_0 : [I_r, \mathbf{0}_{r\times(p-r)}]\tilde{B} = \mathbf{0}_{r\times m}$ under the model of the transformed data $Y = \tilde{X}\tilde{B} + E$. A theoretical result similar to that in Theorem 5 can be obtained under properly adjusted*

assumptions.

Remark 3. *The proposed procedure uses the canonical correlation, which is an extension of the marginal correlation in Fan and Lv (2008). The computation of a canonical correlation is fast, and is pre-implemented in many software packages. Moreover, the proposed method aggregates the joint information of the response variables, and thus may be better than simply applying marginal screening to each response variable. On the other hand, the correlation-based method has potential issues when the predictors are highly correlated (Dutta and Roy, 2017). To study the effect of highly correlated predictors, we performed a preliminary simulation, documented in the Supplementary Material, Section S7.4. Here, we compared our method with that of using a lasso with cross-validation to select predictors, which is expected to account for the dependence in the predictors, but not in the responses. Under the considered settings with correlated predictors, our method outperforms the lasso. The comparison results also show that over- and under-selecting predictors can both cause substantial loss of test power. To further enhance this power, we may extend existing high-dimensional screening methods, such as Wang and Leng (2016), to a multivariate regression setting. In this way, we account for the dependence in both the predictors and responses. This topic is left to future research.*

Given a proper screening approach, we propose a data-splitting procedure to apply the LRT. We randomly split n observations into two independent sets: the

screening data $\{X_S, Y_S\}$ of size n_S, and the test data $\{X_T, Y_T\}$ of size n_T. We use $\{X_S, Y_S\}$ to select $\mathcal{M}$, and apply the proposed LRT to $\{X_T, Y_T\}$ using the selected predictors in $\mathcal{M}$. Data splitting avoids the influence of the screening step on the inference step and provides a valid inference, as is widely recognized in the literature (Berk et al., 2013; Taylor and Tibshirani, 2015). We also demonstrate that the type-I error rate cannot be controlled without splitting the data in the simulation studies in Section 5.

The result of a test based on a single random split is known to be sensitive to the arbitrary split choice, making it difficult to reproduce the result (Meinshausen et al., 2009; Meinshausen and Bühlmann, 2010). Therefore we propose using multiple splits and aggregating the results. Note that computing test statistics by splitting the data can be viewed as a resampling method. Such methods usually do not perform well when approximating statistics that depend on the eigenvalues of high-dimensional random matrices (Karoui and Purdom, 2016). Furthermore, the test statistics computed after splitting the data are correlated. As a result, it is challenging to combine the statistics into a valid and efficient method.

In this study, we adopt the general p-value combination method proposed by Meinshausen et al. (2009). Specifically, we randomly split the data J times, and compute the J p-values for different splits. For each $j = 1, \ldots, J$, we compute the p-value $p^{(j)}$ with data splitting. Then, for $\gamma \in (0, 1)$, define $Q(\gamma) = \min\{1, q_\gamma(\{p^{(j)}/\gamma; j =$

$1, \ldots, J\})\}$, where q_γ denotes the empirical γ-quantile function. Because a proper selection of γ may be difficult, we use the adaptive version below. Let $\gamma_{\min} \in (0, 1)$ be a lower bound for γ, and define the adjusted p-value p_t as $p_t = \min\{1, (1 - \log \gamma_{\min}) \inf_{\gamma \in (\gamma_{\min}, 1)} Q(\gamma)\}$. The extra correction factor $1 - \log \gamma_{\min}$ ensures the type-I error is controlled, despite the adaptive search for the best quantile. For the adaptive multi-split adjusted p-value p_t, the null hypothesis is rejected when $p_t < \alpha$, where α is the prespecified threshold. Following the proof of Theorem 3.2 in Meinshausen et al. (2009), we have the proposition below.

Proposition 6. *Under H_0, for any J random sample splits, if Theorem 5 holds for each split, then* $\limsup_{n\to\infty} P(p_t \leq \alpha) \leq \alpha$.

Proposition 6 shows that the multi-split and aggregation procedure can control the type-I error. To apply the multi-split procedure, we need to choose two parameters, J and $\gamma_{\min}$. In practice, we choose J slightly large and of the same order of n. We next discuss the choice of $\gamma_{\min}$. To improve the test power, we want to choose $\gamma_{\min}$ such that $\limsup_{n\to\infty} P(p_t \leq \alpha)$ in Proposition 6 is maximized to be close to α under H_0. By the proof of Proposition 6, it suffices to make $\text{argmax}_{\gamma \in (0,1)} P\{Q(\gamma) \leq \alpha\} \in (\gamma_{\min}, 1)$, because the adaptive search of γ in p_t is adjusted by the correction factor $1 - \log \gamma_{\min}$. Note that $\{Q(\gamma) \leq \alpha\}$ is equivalent to $\{\psi(\alpha\gamma) \geq \gamma\}$, with $\psi(u) = J^{-1} \sum_{j=1}^{J} 1\{p^{(j)} \leq u\}$. It is then equivalent to finding the γ-value such that $P\{\psi(\alpha\gamma) \geq \gamma\}$ is the closest to the upper bound α.

To evaluate this, we consider two extreme cases for a given J. When the $p^{(j)}$ are highly dependent, $P\{\psi(\alpha\gamma) \geq \gamma\} \simeq P\{p^{(1)} \leq \alpha\gamma\} = \alpha\gamma$, which approaches α when γ is close to one. When the $p^{(j)}$ are nearly independent, $J\gamma \leq 1$, and $\alpha\gamma$ is small, $P\{\psi(\alpha\gamma) \geq \gamma\} \simeq P\{\min_j p^{(j)} \leq \alpha\gamma\} \simeq 1 - (1 - \alpha\gamma)^J \simeq J\alpha\gamma$; then, $J\alpha\gamma \to \alpha$ if $\gamma \to J^{-1}$. When the dependence between the $p^{(j)}$ is between these two extreme cases, we expect the maximum $P\{\psi(\alpha\gamma) \geq \gamma\}$ to be achieved at some $\gamma \in [J^{-1}, 1)$. Because the true correlation is unknown in practice, in the simulations, we recommend takking $\gamma_{\min}$ slightly smaller than J^{-1} so that the candidate γ range contains $[J^{-1}, 1)$. We performed a simulation study to illustrate how the value of $P\{\psi(\alpha\gamma) \geq \gamma\}$ depends on the correlations of the p-values. The results are provided in the Supplementary Material, Section S7.3, and are consistent with the theoretical analysis presented here.

The following is a summary of the testing procedure for large p.

Procedure For $j = 1, \ldots, J$,

1. Randomly split the data into a screening data set $\{X_S, Y_S\}$ and a test data set $\{X_T, Y_T\}$.

2. On $\{X_S, Y_S\}$: compute the canonical correlations between Y_S and each column of X_S; then, select the columns with the largest $\lfloor \delta p \rfloor$ corresponding correlations. The selected column indices form a set $\mathcal{S}_\mathcal{C} \subseteq \{1, \ldots, p\}$.

3. On $\{X_T, Y_T\}$: choose the columns of X_T indexed by $\mathcal{S}_\mathcal{C}$ to obtain $X_{\mathcal{S}_\mathcal{C}}$. Use $\{X_{\mathcal{S}_\mathcal{C}}, Y_T\}$ to compute the test statistic T_3 and obtain the p-value $p^{(j)}$.

After obtaining the set of p-values, $\{p^{(j)} : j = 1, \ldots, J\}$, we compute the adjusted p-value p_t. Reject the null hypothesis if $p_t \leq \alpha$.

Remark 4. *When the dimension of the response Y is large ($m > n$), we also need to reduce the dimension of the response vectors in order to apply the LRT. We can use a principal component analysis (PCA) or factor analysis method to perform the dimension reduction. In the simulation studies, we select the first m_0 principal components of Y_S as the columns of a matrix $\hat{W}$, where m_0 satisfies $m_0 + p < n_T$ and can be chosen using a parallel analysis (Buja and Eyuboglu, 1992; Dobriban and Owen, 2017). Then, we transform the responses Y_T in the test data to obtain $\hat{\mu}_T = Y_T\hat{W}$, which only has m_0 columns. We then use the transformed data $\{X_T, \hat{\mu}_T\}$ to examine $CB\hat{W} = \mathbf{0}$. The independence between the screening and test data sets ensures that the test is valid. Under the sparse model setting, the signal matrix $X_{\mathcal{M}_*}B_{\mathcal{M}_*}$ has a low rank decomposition; thus, we expect the dimension-reduction procedure to maintain high power. This is verified by the simulation studies in Section 5, which show that reducing the dimensions of the responses may even boost the power of certain sparse models. Alternatively, other dimension-reduction techniques can be applied (e.g., Yuan et al., 2007; Ma, 2013). When both m and p are large, we can apply the dimension reduction to Y and X simultaneously to reduce both m and p.*

5. Simulations

In this section, we report the results of several simulation studies used to evaluate the theoretical results and proposed methods for $n > p + m$ and $n < p + m$.

5.1 $n > p + m$

For $n > p + m$, we conduct simulations under null and alternative hypotheses to examine the type-I error and power of our proposed test statistics.

In the first setting, we sample the test statistics by simulating data following the canonical form introduced in Section 3. Specifically, we generate random matrices Y_1^* of size $r \times m$ and Y_2^* of size $(n-p) \times m$, where the rows of Y_1^* and Y_2^* are independent m-variate Gaussian with covariance I_m, and $E(Y_1^*) = M_1$ and $E(Y_2^*) = \mathbf{0}$. Under the canonical form, we know H_0 is equivalent to $M_1 = \mathbf{0}$, as discussed. In the following, each simulation is based on 10,000 replications with significance level 0.05.

Under the null hypothesis, we compare the traditional χ^2 approximation (2.1) with the normal approximations for T_1 in (2.3) and T_3. In particular, we study how the dimension parameters (p, m, r) influence these approximations by varying one parameter each time. Figure 4 gives the estimated type-I errors as p increases. The figure shows that as p becomes larger, the χ^2 approximation (2.1) performs poorly, whereas the normal approximations for T_1 and T_3 still control the type-I error well. Other simulation results with varying m or r are given in the Supplement Material

Section S7.1; similar patterns are observed.

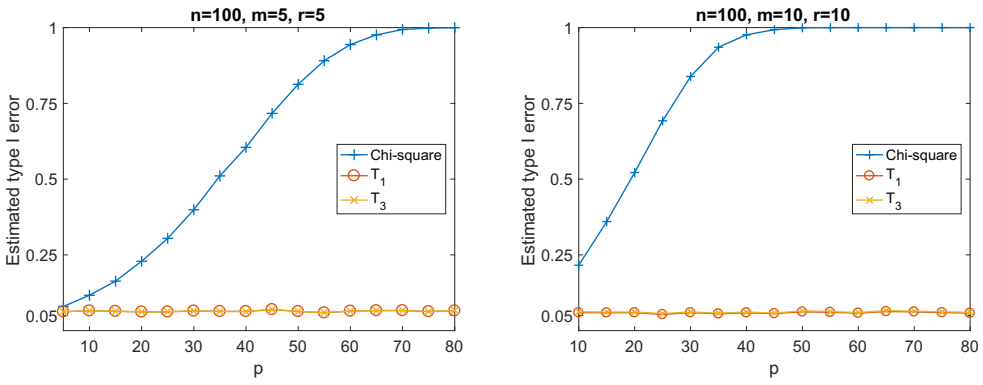

Figure 4: Estimated type I error versus p

Under the alternative hypotheses, we compare the power of the test statistics T_1, T_2, and T_3, and show the power improvement of T_3 over T_1 and T_2. Specifically, under the canonical form, we simulate data with $M_1 = \text{diag}(\delta_1, \ldots, \delta_{r_k}, 0, \ldots, 0)$, that is, a diagonal matrix with r_k nonzero elements. It follows that $\Omega = \text{diag}(\delta_1^2, \ldots, \delta_{r_k}^2, 0, \ldots, 0)$ has rank r_k. Under this setup, we test four cases: (a) $r_k = 1$; (b) $r_k = 2$ and $\delta_1 = \delta_2$; (c) $r_k = 2$ and $\delta_1 = 10\delta_2$; and (d) $r_k = 3$ and $\delta_1 = \delta_2 = \delta_3$. In all cases, $n = 100, m = 20, p = 50$, and $r = 30$. For each case, we plot the estimated power versus $\text{tr}(\Omega)/m$ in Figure 5. The results show that when the rank of Ω, r_k, is small or the significant entries in Ω have low rank, T_2 is more powerful than T_1; however, when r_k or the rank of significant entries in Ω increases, T_1 becomes more powerful. Moreover, in both sparse and nonsparse cases, the combined statistic T_3 has power close to the better of T_1 and T_2, with the type-I error well controlled. These patterns

are consistent with the results of our theoretical power analysis in Section 3.

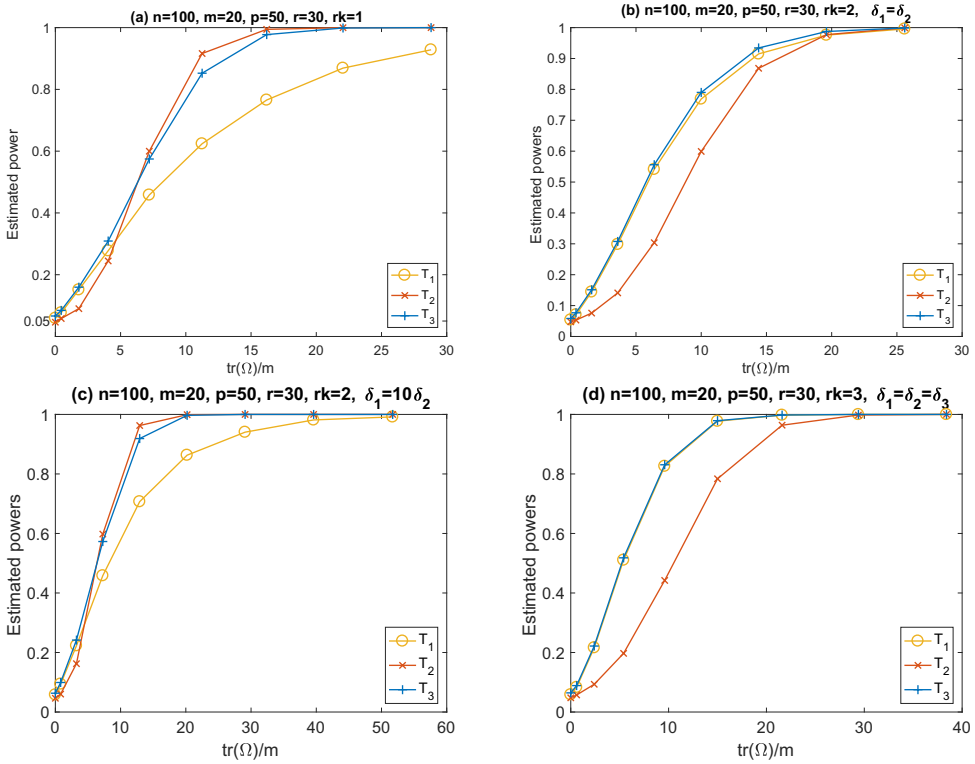

Figure 5: Powers of T_1, T_2 and T_3 versus $\mathrm{tr}(\Omega)/m$

In addition, we conduct simulations when X and Y are generated following $Y = XB + E$, where the rows of E follow multivariate Gaussian distributions. The results are given in the Supplementary Material, Section S7.1, and show that T_3 is powerful under both dense and sparse B cases. Moreover, we conduct similar studies when X and Y take discrete values and when the statistical error follows a heavy-tail t distribution. The results are provided in the Supplementary Material, Section S7.1. We observe similar patterns to the normal cases in Section S7.1, which suggests that

the proposed test statistic is robust to the normal assumption of the statistical error.

5.2 $n < p + m$

This section presents the results of the simulations for $n < p + m$ and evaluates the performance of our proposed procedure in Section 4. Specifically, we take $C = [I_r, \mathbf{0}_{r\times(p-r)}]$, and let B be a $p \times m$ diagonal matrix with σ_s in the first r_k diagonal entries, where σ_s represents the signal size that varies in the simulations. The rows of X and E are independent multivariate Gaussian with covariance matrices $\Sigma_x = (\rho^{|i-j|})_{p\times p}$ and $\Sigma = (\rho^{|i-j|})_{m\times m}$, respectively. We set $n = 100$, $p = 120$, and $r = 120$, and test the cases when $m \in \{20, 120\}, r_k \in \{5, 10\}$, and $\rho \in \{0.3, 0.7\}$. We conduct each simulation with 200 replications, and split the data into screening and test data sets with ratio 3:7 (the ratios 2:8 and 4:6 performed similarly in our simulations). Figure 6 reports the simulation results when $r_k = 10$; all other results are presented in the Supplementary Material, Section S7.2. In Figure 6, "screening" represents the proposed screening procedure on X (with 20% features selected); "PCA" represents the PCA on Y, as in Remark 4; and J represents the number of splits, where $J = 0$ represents a test on the same data without splitting.

Figure 6 shows that when we do not split the data ($J = 0$), the type-I errors cannot be controlled under all cases. If we split the data once ($J = 1$), the type-I errors become closer to the significance level, but can still be unstable. If we use the

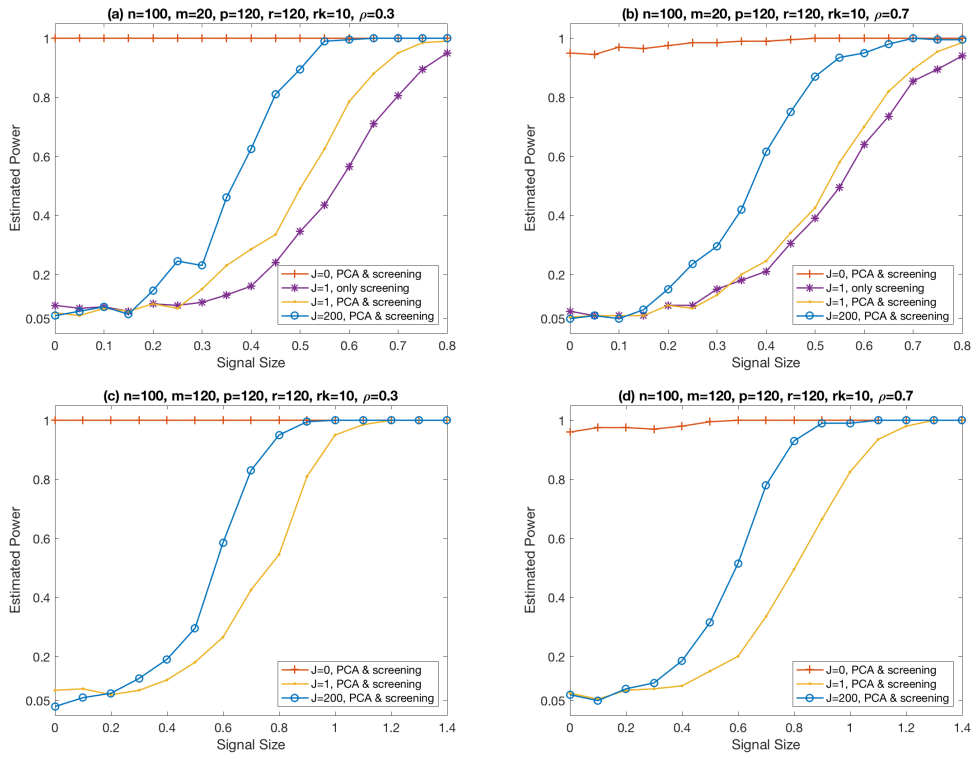

Figure 6: Estimated powers versus signal sizes when $n < m + p$

multi-split method with 200 splits ($J = 200$), the type-I errors become well controlled. The results imply that data splitting is necessary for the proposed two-stage testing procedure, and show that multiple splits help us to obtain stable results. In addition, in the four cases, the multi-split method ($J = 200$) achieves higher power than that of the single split ($J = 1$) as the signal size increases. Moreover, for cases (a) and (b) in Figure 6, with the single split of data ($J = 1$), we also compare the test power when screening only on X with that when performing a dimension reduction on both X and Y. The results are given by the curves "$J = 1$, only screening" and "$J = 1$,

PCA & screening", respectively. We observe that the test power is slightly enhanced by performing a dimension reduction on both X and Y.

In addition, we conduct similar studies when X and Y take discrete values and when the statistical error follows a heavy-tail t distribution; see the Supplementary Material, Section S7.2. We observe similar patterns to those in Figure 6, which suggests that the proposed method is robust to the normal assumption of the statistical error.

6. Real-Data Analysis

We demonstrate our proposed method by analyzing the breast cancer dataset from Chin et al. (2006), which was also studied by Chen et al. (2013) and Molstad and Rothman (2016). The data set is available in the R package `PMA`, and consists of measured gene expression profiles (GEPs) and DNA copy-number variations (CVNs) for $n = 89$ subjects. Prior studies have demonstrated a link between DNA copy-number variations and cancer risk (see, e.g., Peng et al., 2010). Here, we examine the relationship between CNVs and GEPs using a multivariate regression method.

We examine the three chromosomes 8, 17, and 22, and test whether they are related (i.e., $C = I_p$). We report the regression results for the CNVs on the GEPs in this section; we provide the regression results for the GEPs on the CNVs in the Supplementary Material, Section S8, where similar patterns are observed. Here, the

m-variate response is the CNV data and the p-variate predictor is the GEP data, where the dimension parameters are $(p, m) = (673, 138), (1161, 87), (516, 18)$ for the respective chromosomes. Because the parameters p and m are either comparable to or larger than the sample size $n = 89$, we apply the proposed testing procedure in Section 4. In particular, we choose the screening data size $n_S = 26$ and the test data size $n_T = 63$, where $n_S : n_T$ is approximately $3 : 7$. We reduce the dimension of the response CNV data matrix using a parallel analysis, and select the columns of the GEP data matrix using the screening method in Section 4. To include as much information on the predictors as possible, we select between 40 and 50 predictors when screening. For each chromosome, we split the data $J = 2000$ times. Then, we obtain the corresponding p-values, $p^{(j)}$, for $j = 1, \ldots, J$, from the limiting distribution of the test statistic T_3. Lastly, we compute the final p-value, p_t, and reject the null hypothesis if $p_t < \alpha$.

We summarize the test results in Table 1. The column "p_0" indicates the number of selected predictors, and the columns "$k_1 \to k_2$" under "Chromosome pair" indicate that we use GEPs from the k_1th chromosome to predict the CNVs from the k_2th chromosome. For each setting, the symbols "✗" and "✓" indicate that we reject and accept the null hypothesis, respectively. The test results show that the null hypothesis is rejected when the CNVs and GEPs are from the same chromosome, which makes biological sense. On the other hand, if we use GEPs from the eighth chromosome

	Chromosome pair					
p_0	$8 \to 8$	$17 \to 17$	$22 \to 22$	$8 \to 17$	$17 \to 22$	$8 \to 22$
40	×	×	×	×	×	✓
45	×	×	×	×	×	✓
50	×	×	×	×	×	✓

Table 1: Decision results

to predict the CNVs from the 17th chromosome, or use the GEPs from the 17th chromosome to predict the CNVs from the 22nd chromosome, the null hypotheses are rejected; if we use GEPs from the eighth chromosome to predict the CNVs from the 22nd chromosome, the null hypothesis is accepted. These conclusions indicate different relationships between the CNVs and GEPs of different chromosomes, which might deserve closer investigation.

To further illustrate the test results, Figure 7 provides box plots of $\{p^{(j)} : j = 1, \ldots, J\}$ with respect to different chromosome pairs when $p_0 = 45$. We find that the medians of the p-values obtained from the regressions of $8 \to 17$, $17 \to 22$, and the same chromosome pairs are smaller than 0.05, which are consistent with the rejection decisions shown in Table 1. Moreover, for $8 \to 22$, the majority of the p-values are larger than 0.05. This is thus consistent with the decision to accept the null hypothesis when using the GEPs from the eighth chromosome to predict the CNVs from the 22nd chromosome.

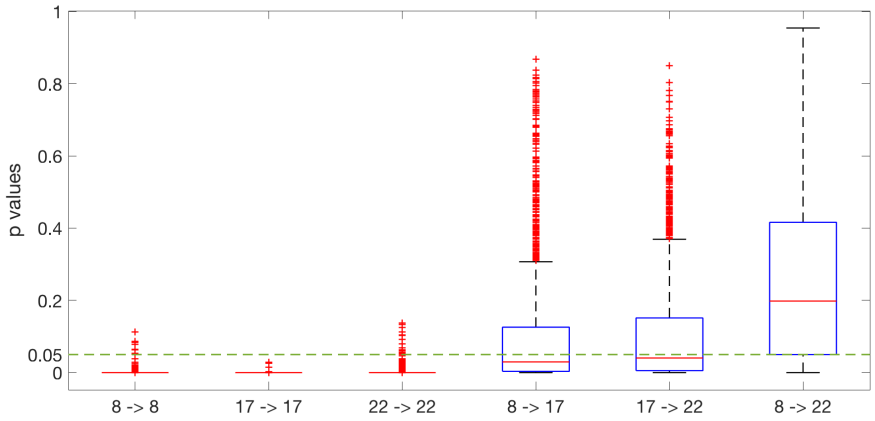

Figure 7: Box plot of p-values for regressions on different chromosome pairs

7. Conclusion

We have examined the LRT for $H_0 : CB = \mathbf{0}_{r\times m}$ in a high-dimensional multivariate linear regression, where p and m are allowed to increase with n. Under the null hypothesis, we derive the asymptotic boundary where the classical χ^2 approximation fails, and propose a corrected limiting distribution for $\log L_n$ in a general asymptotic regime of (p, m, r, n). Under alternative hypotheses, we characterize the statistical power of $\log L_n$ in the high-dimensional setting, and propose a power-enhanced test statistic. In addition, when $n < p + m$ and the LRT is not well defined, we propose using a two-step testing procedure with repeated data-splitting.

This study on the LRT of a multivariate linear regression can be extended to vector nonparametric regression models. Specifically, for $k = 1, \ldots, m$, suppose the kth response variable depends on the p-dimensional predictor vector $\mathbf{x}$ through the

regression equation $y_k = \mathbb{M}_k(\mathbf{x}) + e_k$, where $\mathbb{M}_k$ is an unknown smooth function, and e_k is an error term. We begin with the case when the predictor is univariate. Then, we can model $\mathbb{M}_k(x)$ using regression splines: $\mathbb{M}_k(x) = \sum_{j=1}^{M} b_{k,j}\phi_j(x)$, where $\Phi = (\phi_j : k = 1, \ldots, M)^\intercal$ are some basis functions. Write $\mathbf{y} = (y_1, \ldots, y_m)^\intercal$, $\mathbf{e} = (e_1, \ldots, e_m)^\intercal$, and $B = (b_{k,j})_{M\times m}$; then, $\mathbf{y} = B^\intercal\Phi + \mathbf{e}$, which is in the form of the multivariate linear regression. To test the coefficients B, we can apply the proposed method. More generally, when the predictors are multivariate, additive models (Hastie and Tibshirani, 1986) are commonly used to finesse the "curse of dimensionality". The multivariate functions $\mathbb{M}_k$ are written as $\mathbb{M}_k(\mathbf{x}) = \mathbb{M}_{k,1}(x_1) + \ldots + \mathbb{M}_{k,p}(x_p)$, for $k = 1, \ldots, m$, where $\mathbb{M}_{k,1}(\cdot), \ldots, \mathbb{M}_{k,p}(\cdot)$ are univariate functions. Suppose $\Phi_1, \ldots, \Phi_p$ are the basis functions for $\mathbb{M}_{k,1}(\cdot), \ldots, \mathbb{M}_{k,p}(\cdot)$, respectively. Then, $\mathbf{y} = \tilde{B}^\intercal\tilde{\Phi}+\mathbf{e}$, where $\tilde{B} = (B_1^\intercal, \ldots, B_p^\intercal)^\intercal$ and $\tilde{\Phi} = (\Phi_1^\intercal, \ldots, \Phi_p^\intercal)^\intercal$. Therefore, we can apply the proposed LRT method to test the structure of the coefficient matrix $\tilde{B}$.

This work establishes its theoretical results under the assumption that the error terms E follow Gaussian distributions; nevertheless, we expect our conclusions to hold over a larger range of distributions. Numerically, we conduct simulations when the error terms follow discrete distributions or heavy-tail t distributions, which are provided in the Supplementary Material. The simulation results show similar patterns to the Gaussian cases, implying that the theoretical results may be valid. Theoretically, Bai et al. (2013) showed that the linear spectral of the F-matrix $S_1S_2^{-1}$

also has an asymptotic normal distribution, without specifying that the distributions of the entries of S_1 and S_2 must be normal. However, they assumed that entries of S_1 and S_2 are independent and identically distributed, which is usually not satisfied in a general multivariate regression analysis. Recently, Li et al. (2018) proposed a modified LRT using a nonlinear spectral shrinkage, and established its asymptotic normality without the normal assumption on E when m is proportional to n. However, they assumed that p, the number of predictors, is fixed. Thus, the asymptotic distribution of $\log L_n$ for general high-dimensional nonGaussian cases remains an open question.

Supplementary Material

The online Supplementary Material includes proofs and additional simulations.

Acknowledgments The authors thank the co-editor Dr. Hans-Georg Müller, an associate editor and two anonymous referees for their constructive comments. The authors also thank Prof. Xuming He for the helpful discussions. This research was partially supported by the National Science Foundation grants DMS-1406279, DMS-1712717, SES-1659328, and SES-1846747.

References

Anderson, T. W. (2003). *An introduction to multivariate statistical analysis.* Wiley New York.

Bai, Z., Jiang, D., Yao, J.-F., and Zheng, S. (2009). Corrections to LRT on large-dimensional covariance matrix by RMT. *Ann. Statist.*, 37(6B):3822–3840.

Bai, Z., Jiang, D., Yao, J.-F., and Zheng, S. (2013). Testing linear hypotheses in high-dimensional regressions. *Statistics*, 47(6):1207–1223.

Bai, Z. and Saranadasa, H. (1996). Effect of high dimension: By an example of a two sample problem. *Statistica Sinica*, 6(2):311–329.

Barut, E., Fan, J., and Verhasselt, A. (2016). Conditional sure independence screening. *Journal of the American Statistical Association*, 111(515):1266–1277. PMID: 28360436.

Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013). Valid post-selection inference. *Ann. Statist.*, 41(2):802–837.

Buja, A. and Eyuboglu, N. (1992). Remarks on parallel analysis. *Multivariate behavioral research*, 27(4):509–540.

Cai, T. T. and Xia, Y. (2014). High-dimensional sparse MANOVA. *Journal of Multivariate Analysis*, 131:174–196.

Chen, K., Dong, H., and Chan, K.-S. (2013). Reduced rank regression via adaptive nuclear norm penalization. *Biometrika*, 100(4):901–920.

Chin, K., DeVries, S., Fridlyand, J., Spellman, P. T., Roydasgupta, R., Kuo, W.-L., Lapuk, A., Neve, R. M., Qian, Z., Ryder, T., et al. (2006). Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer cell*, 10(6):529–541.

Dharmawansa, P., Johnstone, I. M., and Onatski, A. (2018). Local asymptotic normality of the spectrum of high-dimensional spiked F-ratios. *arXiv preprint arXiv:1411.3875*.

Dobriban, E. and Owen, A. B. (2017). Deterministic parallel analysis. *arXiv preprint arXiv:1711.04155*.

Donoho, D. L. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. *AMS math challenges lecture*, 1(2000):32.

Dutta, S. and Roy, V. (2017). A note on marginal correlation based screening. *arXiv preprint arXiv:1707.08143*.

Fan, J., Han, F., and Liu, H. (2014). Challenges of big data analysis. *National science review*, 1(2):293–314.

Fan, J. and Lv, J. (2008). Sure independence screening for ultra-high dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911.

Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *Ann. Statist.*, 38(6):3567–3604.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statist. Sci.*, 1(3):297–310.

Hu, J., Bai, Z., Wang, C., and Wang, W. (2017). On testing the equality of high dimensional mean vectors with unequal covariance matrices. *Annals of the Institute of Statistical Mathematics*, 69(2):365–387.

Jiang, D., Jiang, T., and Yang, F. (2012). Likelihood ratio tests for covariance matrices of high-dimensional normal distributions. *Journal of Statistical Planning and Inference*, 142(8):2241–2256.

Jiang, T. and Qi, Y. (2015). Likelihood ratio tests for high-dimensional normal distributions. *Scandinavian Journal of Statistics*, 42(4):988–1009.

Jiang, T. and Yang, F. (2013). Central limit theorems for classical likelihood ratio tests for high-dimensional normal distributions. *Ann. Statist.*, 41(4):2029–2074.

Johnstone, I. M. (2008). Multivariate analysis and Jacobi ensembles: Largest eigenvalue, Tracy–Widom limits and rates of convergence. *Ann. Statist.*, 36(6):2638–2716.

Johnstone, I. M. (2009). Approximate null distribution of the largest root in multivariate analysis. *Ann. Appl. Stat.*, 3(4):1616–1633.

Karoui, N. E. and Purdom, E. (2016). The bootstrap, covariance matrices and PCA in moderate and high-dimensions. *arXiv preprint arXiv:1608.00948*.

Li, H., Aue, A., and Paul, D. (2018). High-dimensional general linear hypothesis tests via non-linear spectral shrinkage. *arXiv preprint arXiv:1810.02043*.

Ma, Z. (2013). Sparse principal component analysis and iterative thresholding. *Ann. Statist.*, 41(2):772–801.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473.

Meinshausen, N., Meier, L., and Bühlmann, P. (2009). P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488):1671–1681.

Molstad, A. J. and Rothman, A. J. (2016). Indirect multivariate response linear regression. *Biometrika*, 103(3):595–607.

Muirhead, R. J. (2009). *Aspects of multivariate statistical theory*, volume 197. John Wiley & Sons.

Peng, J., Zhu, J., Bergamaschi, A., Han, W., Noh, D.-Y., Pollack, J. R., and Wang, P. (2010). Regularized

multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *The annals of applied statistics*, 4(1):53.

Roy, S. N. (1953). On a heuristic method of test construction and its use in multivariate analysis. *Ann. Math. Statist.*, 24(2):220–238.

Srivastava, M. S. and Fujikoshi, Y. (2006). Multivariate analysis of variance with fewer observations than the dimension. *Journal of Multivariate Analysis*, 97(9):1927–1940.

Taylor, J. and Tibshirani, R. J. (2015). Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, 112(25):7629–7634.

Wang, X. and Leng, C. (2016). High dimensional ordinary least squares projection for screening variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(3):589–611.

Yuan, M., Ekici, A., Lu, Z., and Monteiro, R. (2007). Dimension reduction and coefficient estimation in multivariate linear regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(3):329–346.

Zheng, S. (2012). Central limit theorems for linear spectral statistics of large dimensional F-matrices. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 48, pages 444–476. Institut Henri Poincaré.

Zhou, B., Guo, J., and Zhang, J.-T. (2017). High-dimensional general linear hypothesis testing under heteroscedasticity. *Journal of Statistical Planning and Inference*, 188:36–54.