Acta Mathematica Sinica, English Series
Apr., 2022, Vol. 38, No. 4, pp. 683–708
Published online: April 15, 2022
https://doi.org/10.1007/s10114-022-0435-9
http://www.ActaMath.com

Acta Mathematica Sinica, English Series

Projection-based High-dimensional Sign Test

Hui CHEN **Chang Liang ZOU**[1)]
School of Statistics and Data Science, Nankai University, Tianjin 300071, P. R. China
E-mail: *chenhuiah@163.com* *nk.chlzou@gmail.com*

Run Ze LI
Department of Statistics and Methodology Center, Pennsylvania State University, 201 Old Main, University Park, PA 16802, USA
E-mail: *rzli@psu.edu*

Abstract This article is concerned with the high-dimensional location testing problem. For high-dimensional settings, traditional multivariate-sign-based tests perform poorly or become infeasible since their Type I error rates are far away from nominal levels. Several modifications have been proposed to address this challenging issue and shown to perform well. However, most of modified sign-based tests abandon all the correlation information, and this results in power loss in certain cases. We propose a projection weighted sign test to utilize the correlation information. Under mild conditions, we derive the optimal direction and weights with which the proposed projection test possesses asymptotically and locally best power under alternatives. Benefiting from using the sample-splitting idea for estimating the optimal direction, the proposed test is able to retain type-I error rates pretty well with asymptotic distributions, while it can be also highly competitive in terms of robustness. Its advantage relative to existing methods is demonstrated in numerical simulations and a real data example.

Keywords High dimensional location test problem, locally optimal test, nonparametric test, sample-splitting, spatial sign test

MR(2010) Subject Classification 62H15

1 Introduction

With rapid development in data collection and storage techniques, high-dimensional data are frequently collected in many scientific fields such as genomic studies and finance. In this paper, we study testing hypotheses on high-dimensional locations. Let us start with the one-sample mean problem. Suppose that $\mathbf{X}_1, \ldots, \mathbf{X}_n$ is a random sample from a p-dimensional population $\mathbf{X}$ with mean $E(\mathbf{X}) = \boldsymbol{\mu}$ and finite positive definite covariance matrix $\mathrm{cov}(\mathbf{X}) = \boldsymbol{\Omega}$. Of interest is to test the hypothesis:

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0 \ \leftrightarrow \ H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0 \tag{1.1}$$

Received August 15, 2020, revised February 21, 2021, accepted June 24, 2021
Supported by NNSF of China Grants (Grant Nos. 11925106, 11690015, 11931001 and 11971247), NSF of Tianjin Grant (Grant Nos. 18JCJQJC46000 and 18ZXZNGX00140), 111 Project B20016, and National Science Foundation (Grant Nos. DMS 1820702, DMS 1953196 and DMS 2015539)
1) Corresponding author

for some known $\boldsymbol{\mu}_0$. The well-known Hotelling T^2 test was developed for the one sample problem in the multivariate analysis. However, the Hotelling test is undefined when $p > n$ because the sample covariance matrix is singular. [1] showed that the performance of the Hotelling test is adversely impacted by an increased dimension even when $p < n$, reflecting a reduced degree of freedom in estimation when the dimensionality is close to the sample size. They further illustrated that the contamination bias, which grows rapidly with p, would make the test unreliable for a large p. They also provided certain asymptotic justifications and contained some numerical evidence. This challenge calls for new statistical tests to deal with high-dimensional data and many authors are devoted to developing new tests for location problems. See, e.g., [5, 6, 12, 15, 26–28, 31, 33], and references therein.

Although essentially nonparametric in spirit, the statistical performance of the moment-based tests mentioned above would be degraded when the non-normality is severe, especially for heavy-tailed distributions. This motivates some authors to consider using multivariate sign-and/or-rank-based approaches to construct robust tests for high-dimensional location problems. Existing methods were mainly developed under the framework of the spatial signs and ranks (see [22]). [23] considered several testing problems in multivariate analysis, directional statistics, and multivariate time series analysis, and they showed that, under appropriate symmetry assumptions, the fixed-p multivariate sign tests would remain valid in the high-dimensional case. [30] and [9] proposed high-dimensional spatial-sign-based tests for one-sample and two-sample problems, respectively. The test statistic proposed in [30] is essentially in a similar fashion to [5] test statistic. [9] proposed a scalar-invariant two-sample test and thus is particularly useful when different components have different scales in high-dimensional data. To circumvent the difficulty of estimating additional biases yielded by using the estimation of location parameter to replace the true one, they suggested a "leave-one-out" test statistic which may be computationally extensive when n is not too small.

However, all those sign-based methods do not include correlation information in the construction of statistics to make them robust to high dimensionality. A natural variant is the so-called regularized T^2-type test ([4]), by adding a regularization term to the sample covariance matrix so that the estimation is well-posed in high-dimensional settings, but it does not work well in large p cases ([9]). To overcome these issues, we aim to develop a new sign-based test based on the projection method. There exist some projection tests which project the original sample to a lower-dimensional space and carry out tests with the projected sample, such as [17, 20]. In particular, [19] proposed a moment-based projection test using data-driven estimation of the optimal projection direction. [16] generalized the Cramér–von Mises statistic via projection-averaging to obtain a robust test for the multivariate two-sample problem.

In this paper, we proposed a weighted-sign-based projection test statistic under elliptical symmetric distribution and independent component models respectively, and derive the optimal direction and weights with which the test possesses the asymptotically and locally best power under alternatives. Interestingly, both the optimal direction and weights have simple closed forms, and they do not depend on the underlying distribution. We further propose an estimation procedure for the optimal direction by using the sample-splitting strategy. Under H_0 in (1.1), we show that the proposed tests with the estimated optimal direction asymptotically follows a χ^2-distribution, and its finite sample P-value can be approximated by an F-distribution. We demonstrate that the proposed projection tests with the estimated optimal direction retain type-I error rates pretty well with asymptotic distributions, while it also can achieve high power in a wide range of distributions and alternatives. Under H_1, we derive the power function of the proposed test under local alternative, and asymptotic relative efficiency of the proposed test to several ones proposed in the recent literature. We further apply the projection weighted sign test for two sample mean problems and derive its optimal direction. We further examine the finite sample performance of the proposed tests via Monte Carlo simulations. Our simulation results clearly demonstrate the superiority of the proposed test over existing ones.

The remainder of this article is organized as follows. In Section 2, we present the new test statistic and its theoretical properties, as well as some discussions on the practical implementation of our proposed test. Extensions to the two-sample problem are given in Section 3. Section 4 consists of simulation studies and a real data example. Some concluding remarks are given in Section 5, and theoretical proofs are given in the Appendix. Some additional simulation results are provided in the Supplementary Material.

2 Projection Tests Based on Weighted Signs

We firstly develop tests for (1.1) under the elliptically symmetric assumption which is commonly considered in the literature of multivariate-sign-based methods (see [22]). We further develop projection test under independent component models in Section 2.3. The proposed approach will be further extended to two-sample problem in Section 3.

2.1 Optimal Projection Direction and Weight Function

Suppose that $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ is an independent and identically distributed (i.i.d.) random sample from a p-dimensional elliptical distribution $E_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with the density function $\det(\boldsymbol{\Sigma})^{-1/2}$ $g(\|\boldsymbol{\Sigma}^{-1/2}(\mathbf{x}-\boldsymbol{\mu})\|)$, where $\boldsymbol{\mu}$ is the symmetry center and $\boldsymbol{\Sigma}$ is a positive definite symmetric $p\times p$ scatter matrix. The spatial sign function of a vector $\mathbf{X}$ is defined as $U(\mathbf{X}) = \|\mathbf{X}\|^{-1}\mathbf{X}I(\mathbf{X} \neq \mathbf{0})$. Let $\boldsymbol{\Gamma}$ be a $p\times p$ nonsingular matrix so that $\boldsymbol{\Gamma}^\top\boldsymbol{\Gamma} = \boldsymbol{\Sigma}^{-1}$, then $\boldsymbol{\varepsilon}_i = \boldsymbol{\Gamma}(\mathbf{X}_i-\boldsymbol{\mu}) \sim E_p(\mathbf{0}, \mathbf{I}_p)$. From the theory of elliptical distribution ([8]), the modulus $\|\boldsymbol{\varepsilon}_i\|$ and the direction vector $\mathbf{U}_i = U(\boldsymbol{\varepsilon}_i)$ are independent, and the direction $\mathbf{U}_i$ follows the uniform distribution over the p-dimensional unit sphere. It is well known that $E(\mathbf{U}_i) = \mathbf{0}$ and $\mathrm{cov}(\mathbf{U}_i) = p^{-1}\mathbf{I}_p$. Without loss of generality, we set $\boldsymbol{\mu}_0 = \mathbf{0}$ in (1.1).

Let $\mathbf{S}$ be an appropriate estimate of the scatter matrix ([29]), and define $\widehat{\mathbf{U}}_i = U(\mathbf{S}^{-1/2}\mathbf{X}_i)$, and $\bar{\mathbf{U}} = n^{-1}\sum_{i=1}^n \widehat{\mathbf{U}}_i$. The following so-called "inner standardization" sign-based statistic, $W_n = np\bar{\mathbf{U}}^\top\bar{\mathbf{U}}$, is quite popular under the setting in which p is finite and fixed (e.g., [24]). W_n is affine-invariant and can be regarded as a nonparametric counterpart of Hotelling's T^2 statistic by using the spatial-signs instead of the original observations $\mathbf{X}_i$'s. [9] showed that W_n would not work well when p and n are close.

Consider the weighted sign $\mathbf{V}_i = K(\|\boldsymbol{\Gamma}\mathbf{X}_i\|)U(\boldsymbol{\Gamma}\mathbf{X}_i)$, where $K(\cdot)$ is a positive weight function. $\mathbf{V}_i$ combines the information on the direction of $\boldsymbol{\Gamma}\mathbf{X}_i$ given by its spatial sign $U(\boldsymbol{\Gamma}\mathbf{X}_i)$ with the information on the magnitude of $r_i \stackrel{\text{def}}{=} \|\boldsymbol{\Gamma}\mathbf{X}_i\|$. This weighted version of sign statistic is essentially in a similar spirit of multivariate signed-rank statistic, $h(R_i/(n+1))U(\boldsymbol{\Gamma}\mathbf{X}_i)$, where R_i is the rank of $\|\boldsymbol{\Gamma}\mathbf{X}_i\|$ among $\|\boldsymbol{\Gamma}\mathbf{X}_1\|, \ldots, \|\boldsymbol{\Gamma}\mathbf{X}_n\|$ and $h(\cdot)$ is a non-negative, nondecreasing, uniformly bounded continuous function over $[0,1]$ (see [14, 21]). Clearly, our $\mathbf{V}_i$ includes the signed-rank statistic as a special case. Now, the original hypothesis is equivalent to

$$H_0 : E(\mathbf{V}_i) = \mathbf{0} \longleftrightarrow H_1 : E(\mathbf{V}_i) \neq \mathbf{0}.$$

Consider constructing a test statistic with a projection matrix, $\mathbf{D}$, which is a $p \times k$ full rank matrix with $k \ll p$ and $k < n$. Accordingly, $\mathbf{D}^\top \mathbf{V}_1, \dots, \mathbf{D}^\top \mathbf{V}_n$ are i.i.d. k-variate random vectors. Under H_0, $E(\mathbf{D}^\top \mathbf{V}_i) = \mathbf{0}$ and $\text{cov}(\mathbf{D}^\top \mathbf{V}_i) = p^{-1} E\{K^2(r_i)\} \mathbf{D}^\top \mathbf{D}$. The projection-based test is defined by

$$Q_{\mathbf{D}} = np[E\{K^2(r_i)\}]^{-1} \bar{\mathbf{V}}^\top \mathbf{D}(\mathbf{D}^\top \mathbf{D})^{-1} \mathbf{D}^\top \bar{\mathbf{V}},$$

where $\bar{\mathbf{V}} = n^{-1} \sum_{i=1}^n \mathbf{V}_i$. Apparently, large $Q_{\mathbf{D}}$ leads to the rejection of the null hypothesis.

The following theorem sheds lights on how to determine the weight function $K(\cdot)$ and projection direction $\mathbf{D}$ so that $Q_{\mathbf{D}}$ maximizes the power under a local alternative hypothesis. In the asymptotic analysis, we let $p \to \infty$ as $n \to \infty$. We need the follow assumptions to facilitate the derivation.

Assumption 1 The moment $E(\|\boldsymbol{\varepsilon}_i\|^{-8})$ exists for large enough p.

Assumption 2 For any $t > 0$, the function $K(\cdot)$ satisfies that $|K(x+t) - K(x)|/K(x) \leq Mt$ when $x > 0$ is sufficiently large, where $M > 0$ is some constant.

Assumptions 1–2 ensure the validity of second-order expansions we use. The moments $E(\|\boldsymbol{\varepsilon}_i\|^{-k})$ may not exist for a fixed p. For example, for standard multivariate normal and t distributions, $E(\|\boldsymbol{\varepsilon}_i\|^{-2})$ is equal to $1/(p-2)$ and thus the second moment exists only when $p > 3$. [34] verified this assumption for three commonly used elliptical distributions, the multivariate normal, the multivariate t distributions, and scale mixtures of multivariate normal distributions. They also formulated this assumption using the g_p that fixes the distribution of the modulus $\|\boldsymbol{\varepsilon}_i\|$. The existence of $E(\|\boldsymbol{\varepsilon}_i\|^{-k})$ is guaranteed if $r^{p-1-k} g_p(r)$ is bounded for $r \in (0, \epsilon)$. Most commonly used functions satisfy Assumption 2, such as x^a, $\exp(x)$, $\log(x)$ and $x/(1+x)$.

Theorem 2.1 *Consider the local alternative hypothesis,*

$$H_{1n} : E(\|\boldsymbol{\varepsilon}_i\|^{-2}) \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} = O\{(np)^{-1}\}.$$

Suppose that the function $K(\cdot)$ satisfies Assumption 2 and $E\{K^4(r_i)\}/[E\{K^2(r_i)\}]^2 \to \zeta \in [1, \infty)$ as $p \to \infty$, where ζ is a constant. The projection test based on $Q^2_{\mathbf{D}}$ reaches its asymptotic best power at $k = 1$ and $\mathbf{D} = E(\mathbf{V})$ with the weight function $K(t) = t^{-1}$.

Under this local alternative, the asymptotic distribution of $Q_{\mathbf{D}}$ is a noncentral chi-squared distribution. The difference between $\boldsymbol{\mu}$ and $\mathbf{0}$, quantified by $\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}$, goes to zero as $n, p \to \infty$. Consequently, the asymptotic variance of $\sqrt{n} \mathbf{D}^\top \bar{\mathbf{V}}$ is still $p^{-1} E\{K^2(r_i)\} \mathbf{D}^\top \mathbf{D}$ so that an explicit power expression can be obtained. Theorem 2.1 enlightens us to construct a locally most powerful weighted sign test by using

$$Q_{\mathbf{d}} = np\{E(r_i^{-2})\}^{-1} \bar{\mathbf{V}}^\top \mathbf{d}(\mathbf{d}^\top \mathbf{d})^{-1} \mathbf{d}^\top \bar{\mathbf{V}}, \tag{2.1}$$

where $\mathbf{d} = E(\mathbf{V}_i)$ is the "oracle" optimal direction and $\mathbf{V}_i = U(\boldsymbol{\Gamma} \mathbf{X}_i)/r_i$. Note that when $\mathbf{D} = \mathbf{d}$, $E(\mathbf{D}^\top \mathbf{V}_i) = \mathbf{0}$ is equivalent to $E(\mathbf{V}_i) = \mathbf{0}$, and consequently $Q_{\mathbf{d}}$ provides a method for testing (1.1). It is also worth pointing out that [13] proposed a class of optimal tests based on interdirections and pseudo-Mahalanobis ranks. Depending on the score function considered, locally optimal tests were obtained at selected densities. It is interesting to see that the optimal weight function is free of the density $g(\cdot)$, which is a unique feature of high-dimensional problems.

2.2 Test with Sample-splitting-based Direction Estimation

In practice, $Q_{\mathbf{d}}$ can be carried out with an estimated $\mathbf{d}$ and $\mathbf{\Gamma}$. Following [19] single sample-splitting strategy (c.f., [32]), we partition the random sample into two separate sets: $\mathcal{U}_1$ and $\mathcal{U}_2$. We use $\mathcal{U}_1$ for estimation of the optimal direction and $\mathcal{U}_2$ to construct $Q_{\mathbf{d}}$. To be more specific, let $\mathbf{S}_1$ be the sample covariance matrix computed from $\mathcal{U}_1$. Note that $\mathbf{S}_1$ may not be invertible when p is greater than n. A simple remedy is to use the ridge-like estimator $\mathbf{S}_1 + \lambda \mathbf{D}_{\mathbf{S}_1}$, where $\mathbf{D}_{\mathbf{S}_1} = \mathrm{diag}(\mathbf{S}_1)$, the diagonal matrix of $\mathbf{S}_1$, and λ is a ridge parameter. As alternative to the ridge-like estimator, we propose to consistently estimate $\mathbf{\Sigma}^{-1}$ by the thresholding approach based on $\mathcal{U}_1$ under some sparsity assumptions ([2, 3]). Let $\widehat{\mathbf{\Sigma}}^{-1}$ be a proper estimator of $\mathbf{\Sigma}^{-1}$. Define $\widehat{\mathbf{\Gamma}}^\top \widehat{\mathbf{\Gamma}} = \widehat{\mathbf{\Sigma}}^{-1}$ and $\widehat{\mathbf{V}}_i = U(\widehat{\mathbf{\Gamma}}\mathbf{X}_i)/\|\widehat{\mathbf{\Gamma}}\mathbf{X}_i\|$. Then $\mathbf{d}$ can be estimated by $\widehat{\mathbf{d}} = n_1^{-1}\sum_{i=1}^{n_1} \widehat{\mathbf{V}}_i$. Similar to $\widehat{\mathbf{\Sigma}}^{-1}$, another option of forming $\widehat{\mathbf{d}}$ is to estimate $\boldsymbol{\mu}$ via thresholding. As shown in Section 2.4, the effect of replacing by thresholded estimators is asymptotically negligible and the efficiency of the projection-based test can still be achieved under certain conditions.

The projection-based test statistic is given by

$$\tilde{Q}_{\widehat{\mathbf{d}}} = n_2 p \{\widehat{E}(r_i^{-2})\}^{-1} \tilde{\mathbf{V}}^\top \widehat{\mathbf{d}} (\widehat{\mathbf{d}}^\top \widehat{\mathbf{d}})^{-1} \widehat{\mathbf{d}}^\top \tilde{\mathbf{V}},$$

where $\tilde{\mathbf{V}} = n_2^{-1}\sum_{i=n_1+1}^{n} \widehat{\mathbf{V}}_i$ and $\widehat{E}(r_i^{-2})$ is an estimator of $E(r_i^{-2})$. Since $\widehat{\mathbf{d}}$ is independent of $\mathcal{U}_2$, $\tilde{Q}_{\widehat{\mathbf{d}}}$ based on $\widehat{\mathbf{d}}^\top \tilde{\mathbf{V}}$ is asymptotically χ_1^2 distributed if $\widehat{\mathbf{\Gamma}}$ is instead of $\mathbf{\Gamma}$. However, due to the use of $\widehat{\mathbf{\Gamma}}$ in $\tilde{\mathbf{V}}$, the variance of $\tilde{Q}_{\widehat{\mathbf{d}}}$ would be generally larger than 2, yielding a little liberal test.

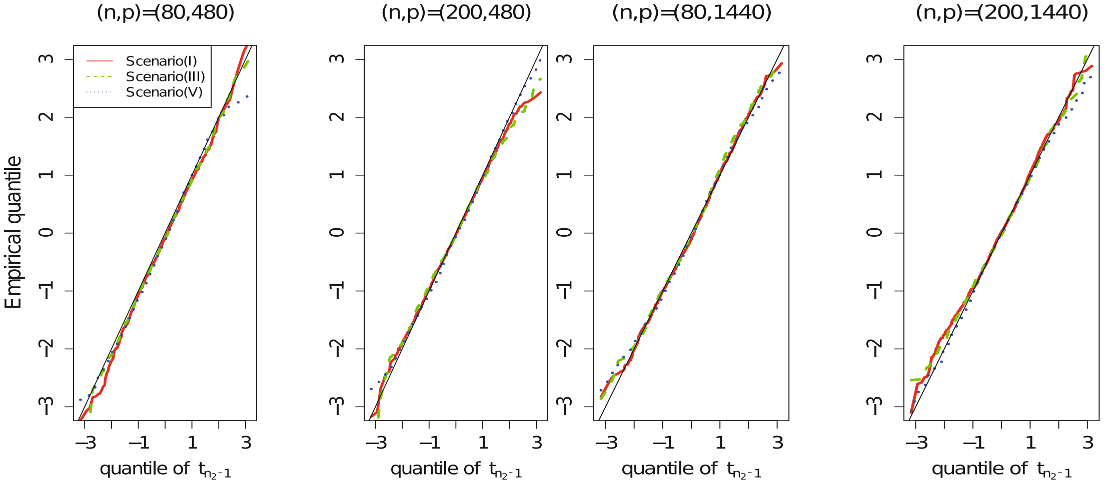


Figure 1 Q-Q plots of our test statistics $T_{\widehat{\mathbf{d}}}$ under the autoregressive correlation structure

Noting that $\tilde{Q}_{\widehat{\mathbf{d}}}$ is essentially of the form $c(\widehat{\mathbf{d}}^\top \tilde{\mathbf{V}})^2$. Conditional on $\mathcal{U}_1$, $\widehat{\mathbf{d}}^\top \widehat{\mathbf{V}}_{n_1+1}, \ldots, \widehat{\mathbf{d}}^\top \widehat{\mathbf{V}}_n$ are i.i.d. random variables with mean zero. Accordingly, we suggest the following simple modification

$$T_{\widehat{\mathbf{d}}}^2 = n_2(\widehat{\mathbf{d}}^\top \tilde{\mathbf{V}})^2/\widehat{\sigma}^2, \tag{2.2}$$

where $\widehat{\sigma}^2 = (n_2 - 1)^{-1} \sum_{i=n_1+1}^n (\widehat{\mathbf{d}}^\top \widehat{\mathbf{V}}_i - \widehat{\mathbf{d}}^\top \tilde{\mathbf{V}})^2$. $T_{\widehat{\mathbf{d}}}^2$ is similar to a classical F-test based on $\widehat{\mathbf{d}}^\top \widehat{\mathbf{V}}_{n_1+1}, \dots, \widehat{\mathbf{d}}^\top \widehat{\mathbf{V}}_n$ and is asymptotically χ_1^2 distributed. Its finite-sample p-value can be approximated by using the F-distribution with 1 and $n_2 - 1$ degrees of freedom. Figure 1 displays the Q-Q plots of $T_{\widehat{\mathbf{d}}}$ under various settings. Generally, the t distribution, t_{n_2-1}, is a good fit of the distribution of $T_{\widehat{\mathbf{d}}}$ as for moderate sample size.

2.3 Extension to Independent Component Model

In the above discussion, we focus on the family of elliptical distributions, while our results can be readily extended to some other distribution families. One such class of distributions is the so-called *independent component model* (see Section 2.3 in [22]); that is $\mathbf{X}_i$ is generated by $\mathbf{X}_i = \boldsymbol{\mu} + \boldsymbol{\Upsilon}\mathbf{Z}_i$, where $\boldsymbol{\Upsilon}$ is a $p \times p$ positive-definite matrix and the standardized and centered p-vector $\mathbf{Z}_i$ has independent and symmetric components. It is noted that this class of models is closely related to the linear transformation models which has received broad attentions in high-dimensional tests ([1, 5]). The following theorem summarizes the fact that the result in Theorem 2.1 is still valid under independent component model. Let $\lambda_{\max}(\mathbf{A})$ and $\lambda_{\min}(\mathbf{A})$ denote the largest and smallest eigenvalues of a matrix $\mathbf{A}$, respectively.

Theorem 2.2 *Suppose all the conditions in Theorem* 2.1 *hold, and* $\lambda_{\min}$ $(\boldsymbol{\Upsilon}\boldsymbol{\Upsilon}^\top)$ *and* $\lambda_{\max}$ $(\boldsymbol{\Upsilon}\boldsymbol{\Upsilon}^\top)$ *are bounded away from both zero and infinity. Under the independent component model, the projection test based on* $Q_{\mathbf{D}}^2$ *reaches its best local power at* $k = 1$ *and* $\mathbf{D} = E(\mathbf{V})$ *with the weight function* $K(t) = t^{-1}$.

Generalizations to other multivariate distributions, such as the family of generalized elliptical distributions are possible, which warrant future research.

2.4 Asymptotic Power Comparison

Theorem 2.1 enables us to compare the proposed test, the weighted-sign-based projection test (abbreviated as WSP) with some existing works, such as [19] and [30], in terms of the limiting efficiency. Consider the local alternative $\boldsymbol{\mu} = n^{-1/2}\boldsymbol{\delta}$ and assume that $\sqrt{n_2/n} \to b > 0$ as $n \to \infty$. Under the alternative, the asymptotic power function of the test based on $Q_{\mathbf{d}}$ in (2.1) at a given significant level α is

$$\beta_{\mathrm{WSP}} = \Phi\bigg(-z_{\alpha/2} + \sqrt{\frac{n_2 p}{E(\|\boldsymbol{\varepsilon}_i\|^{-2})}} \|E(\mathbf{V})\| \bigg),$$

where $\Phi(\cdot)$ is the standard normal distribution function and z_α is the upper α quantile of $N(0,1)$. By the first-order Taylor expansion of $U(\boldsymbol{\Gamma}\mathbf{X}_i)$ (see Appendix), β_{WSP} reduces to

$$\beta_{\mathrm{WSP}} = \Phi(-z_{\alpha/2} + b\sqrt{p\eta E(\|\boldsymbol{\varepsilon}_i\|^{-2})}), \tag{2.3}$$

where we denote $\eta = \boldsymbol{\delta}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\delta}$.

A natural but technical question to address is that when $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$ admit some structures, are the thresholded estimators accurate enough so that the feasible $T_{\widehat{\mathbf{d}}}^2$ in (2.2) is able to inherit optimal properties of $Q_{\mathbf{d}}$ in certain degree? The answer is affirmative, at least when certain sparsity degree is fulfilled. Let $\widehat{\boldsymbol{\mu}} = T_{h_n}(\bar{\mathbf{X}}_1)$, where $h_n = M'\sqrt{\log p/n}$ for some large M' and

$T_h(\cdot)$ is the thresholding operator, defined by $T_t(\boldsymbol{\mu}) = \{\mu_1 I(|\mu_1| \geq t), \ldots, \mu_p I(|\mu_p| \geq t)\}^\top$ and $\bar{\mathbf{X}}_1$ is the sample mean of the set $\mathcal{U}_1$. Let $\widehat{\mathbf{d}} = \widehat{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\mu}}$, where $\widehat{\boldsymbol{\Gamma}}$ is obtained based on the thresholded estimator of $\boldsymbol{\Sigma}$ ([2]). To establish the results shown below, we need some additional conditions.

Assumption 3 Suppose $\lambda_{\min}(\boldsymbol{\Sigma})$ and $\lambda_{\max}(\boldsymbol{\Sigma})$ are bounded away from both zero and infinity. Moreover, the components of $\boldsymbol{\Sigma}$, $(\boldsymbol{\Sigma})_{ij}$, satisfy $(\boldsymbol{\Sigma})_{ii} \leq M$ and $\sum_{j=1}^{p} |(\boldsymbol{\Sigma})_{ij}|^q \leq s_{p,\boldsymbol{\Sigma}}$ for some $M > 0$ and $1 > q \geq 0$, where $s_{p,\boldsymbol{\Sigma}}(\log p/n)^{(1-q)/2} = o(1)$.

Assumption 4 Suppose that the components of the mean vector $\boldsymbol{\mu}$ satisfy $|\mu_j| \leq M, \sum_{j=1}^{p} |\mu_j|^q \leq s_{p,\boldsymbol{\mu}}$ for some $M > 0$ and $1 > q \geq 0$. Furthermore, $s_{p,\boldsymbol{\mu}}$ satisfies $s_{p,\boldsymbol{\mu}}(\log p/n)^{(1-q)/2} = o(1)$.

Assumption 5 The components of $\mathbf{X}_i$ are sub-Gaussian variables.

Assumptions 3 and 4 characterize the sparsity of $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$ by $s_{p,\boldsymbol{\Sigma}}$ and $s_{p,\boldsymbol{\mu}}$, respectively. These two assumptions along with Assumption 5 are imposed for the purpose of establishing consistency of the thresholded estimators $\widehat{\boldsymbol{\Sigma}}^{-1}$ ([2]) and $T_{h_n}(\bar{\mathbf{X}}_1)$. The above assumptions, perhaps somewhat restrictive, substantially facilitates our technical analysis. In fact, they are commonly used in the literature of high-dimensional estimation and test (e.g., [7] and the references therein).

Theorem 2.3 *Suppose that Assumptions* 1–5 *hold. Under either elliptical distributions or independent component model,*

$$\sqrt{n_2}\widehat{\mathbf{d}}^\top \tilde{\mathbf{V}}/\widehat{\sigma} - \sqrt{n_2 p E\{\|\boldsymbol{\varepsilon}_i\|^{-2}\}}(\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})^{1/2} \xrightarrow{d} N(0,1),$$

provided that $\max\{s_{p,\boldsymbol{\Sigma}}, s_{p,\boldsymbol{\mu}}\}(\log p/n)^{(1-q)/2} \to 0$.

By this theorem, it is easy to verify that the power function in (2.3) is still valid for $T^2_{\widehat{\mathbf{d}}}$. If $s_{p,\boldsymbol{\mu}}$ and $s_{p,\boldsymbol{\Sigma}}$ are bounded, p can increase at an exponential rate of n. This should not be surprising to us due to the benefit of using sample-splitting strategy; that is, the magnitudes of the terms like $\mathbf{u}_i^\top(\widehat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1})\mathbf{u}_i$ could be much more easily and accurately controlled when $\mathbf{u}_i$ is independent of $\widehat{\boldsymbol{\Sigma}}^{-1}$ than in the case without sample-splitting. Of course, the ratio "b" in (2.3) represents the price to pay for estimating a large error covariance matrix with a subset of samples.

Remark 2.4 [18] has studied eigenvalues of large dimension matrix via Stieltjes transform of the empirical spectral measure for sample covariance matrix which equals to $\mathrm{tr}(\mathbf{S} + \lambda\mathbf{I})^{-1}$, where $\mathbf{S}$ is the sample covariance, $\mathbf{I}$ the identity matrix, and λ represents tuning parameter of ridge estimator. A natural variant is the so-called regularized T^2-type test ([4]) for high dimensional mean test, by adding a regularization term to the sample covariance matrix so that the estimation is well-posed in high-dimensional settings. Though there is no sparse constraint, the relationship between n and p is required to be $p/n \to c$ with $0 < c < \infty$.

2.4.1 Comparison with the Projection Test without Weights

Similarly, the asymptotic power of the sign-based projection test without weights (abbreviated as SP), say $K(t) = 1$, is given by

$$\beta_{\mathrm{SP}} = \Phi(-z_{\alpha/2} + b\sqrt{\eta p}E(\|\boldsymbol{\varepsilon}_i\|^{-1})).$$

By Cauchy inequality, $\beta_{\rm WSP}(\eta)\ge\beta_{\rm SP}(\eta)$ and the asymptotic relative efficiency (ARE) of the WSP with respect to SP is

$$\mathrm{ARE}(\mathrm{WSP},\mathrm{SP})=\frac{E(\|\boldsymbol{\varepsilon}_i\|^{-2})}{\{E(\|\boldsymbol{\varepsilon}_i\|^{-1})\}^2}.$$

When $\boldsymbol{\varepsilon}_i\sim N_p(\mathbf{0},\mathbf{I}_p)$, this ARE value is equal to one. When $\boldsymbol{\varepsilon}_i\sim t_v(\mathbf{0},\mathbf{I}_p)$, multivariate t distribution with degrees of freedom v, we can see that

$$\mathrm{ARE}(\mathrm{WSP},\mathrm{SP})=\frac{(p-1)v\Gamma^2(v/2)}{2(p-2)\Gamma^2((v+1)/2)}.$$

For $v=3$, its value is about 1.18 and $\mathrm{ARE}(\mathrm{WSP},\mathrm{SP})\to1$ as $v\to\infty$. When $\boldsymbol{\varepsilon}_i\sim MN(\epsilon,\sigma,\mathbf{I}_p)$, the multivariate mixture normal distribution with density function $(1-\epsilon)f_p(\mathbf{0},\mathbf{I}_p)+\epsilon f_p(\mathbf{0},\sigma^2\mathbf{I}_p)$, where $f_p(\mathbf{0},\mathbf{I}_p)$ denotes the probability density function of multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{I}_p$, we have

$$\mathrm{ARE}(\mathrm{WSP},\mathrm{SP})=\frac{(p-1)(1-\epsilon+\epsilon/\sigma^2)}{(p-2)(1-\epsilon+\epsilon/\sigma)^2}.$$

The ARE(WSP,SP) will converge to $1/(1-\epsilon)$ as $\sigma^2\to\infty$ and $p\to\infty$. When $\sigma=10$ and $\epsilon=0.8$, this value is about 2.65. Hence, the efficiency gain of WSP compared with SP may be substantial for heavy-tailed distribution.

2.4.2 Comparison with [19] Projection Test

Note that under the elliptical assumption, $\mathrm{cov}(\mathbf{X})=E(\|\boldsymbol{\varepsilon}_i\|^2)\boldsymbol{\Sigma}/p$. Thus, the asymptotic power function of [19] moment-based projection test (abbreviated as MP) is given by

$$\beta_{\rm MP}=\Phi\bigg(-z_{\alpha/2}+b\sqrt{\frac{p\eta}{E(\|\boldsymbol{\varepsilon}_i\|^2)}}\bigg).$$

Consequently, $\mathrm{ARE}(\mathrm{WSP},\mathrm{MP})=E(\|\boldsymbol{\varepsilon}_i\|^2)E(\|\boldsymbol{\varepsilon}_i\|^{-2})\ge1$.

When $\boldsymbol{\varepsilon}_i\sim N_p(\mathbf{0},\mathbf{I}_p)$, we have $\mathrm{ARE}(\mathrm{WSP},\mathrm{MP})=p/(p-2)\to1$ as $p\to\infty$; two tests are equivalently powerful. When the dimension p is fixed, it can be expected that the proposed test, using the direction of an observation from the origin, should be outperformed by the test constructed with original observations as the MP test. However, as $p\to\infty$ as $n\to\infty$, the disadvantage diminishes, at least from viewpoints of local power. For $\boldsymbol{\varepsilon}_i\sim t_v(\mathbf{0},\mathbf{I}_p)$ and $\boldsymbol{\varepsilon}_i\sim MN(\epsilon,\sigma,\mathbf{I}_p)$, $\mathrm{ARE}(\mathrm{WSP},\mathrm{MP})$ becomes $pv/\{(p-2)(v-2)\}$ and $\{(1-\epsilon)+\epsilon/\sigma^2\}\{(1-\epsilon)p+\epsilon p\sigma^2\}/(p-2)$, yielding the values of 3 and 16.7 for $t_3(\mathbf{0},\mathbf{I}_p)$ and $MN(0.8,10,\mathbf{I}_p)$, respectively. Clearly, the WSP test is more powerful than MP when the distributions are heavy-tailed.

2.4.3 Comparison with [30] Sign Test

[30] spatial-sign-based test (abbreviated as SS) is constructed based on $U(\mathbf{X}_i)$ rather than $U(\boldsymbol{\Gamma}\mathbf{X}_i)$. Hence, the WSP and SS tests are not directly comparable because their power functions depend on $\boldsymbol{\delta}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}$ and $\boldsymbol{\delta}^\top\boldsymbol{\delta}$, respectively. Different choices of $\boldsymbol{\Sigma}$ and $\boldsymbol{\delta}$ would usually yield quite different comparison results. Please refer to [9] for related discussions. To avoid this difficulty, we replace $U(\mathbf{X}_i)$ with $U(\boldsymbol{\Gamma}\mathbf{X}_i)$ in the SS test for a clear comparison. By Theorem 2.3 in [30], the asymptotic power of SS (with $U(\boldsymbol{\Gamma}\mathbf{X}_i)$) is

$$\beta_{\rm SS}\approx\Phi(-z_\alpha+\sqrt{p/2}\eta E^2(\|\boldsymbol{\varepsilon}_i\|^{-1})).$$

When $\boldsymbol{\varepsilon}_i \sim N_p(\mathbf{0}, \mathbf{I}_p)$ or $\boldsymbol{\varepsilon}_i \sim t_v(\mathbf{0}, \mathbf{I}_p)$, we have $E^2(\|\boldsymbol{\varepsilon}_i\|^{-1}) \le 1/(p-2)$. Accordingly, $\beta_{\mathrm{SS}} \le \Phi(-z_\alpha + \eta/\sqrt{2(p-2)})$. When $\sqrt{\eta}b - \eta/\sqrt{2(p-2)} \ge z_{\alpha/2} - z_\alpha$, $\beta_{\mathrm{WSP}} \ge \beta_{\mathrm{SS}}$. This trivially holds if

$$\sqrt{2(p-2)}b/\sqrt{\eta} \to \infty \tag{2.4}$$

as $p \to \infty$. For many commonly used $\boldsymbol{\Sigma}$, $\eta \asymp \|\boldsymbol{\delta}\|^2$ and thus (2.4) is valid as long as $\|\boldsymbol{\delta}\|^2 = o_p(p)$. This implies that when the signal in $\boldsymbol{\mu}$ is weak, our proposed test could be more powerful than the SS test. For example, assume that $(\boldsymbol{\Sigma})_{ij} = \varpi^{|i-j|}$ for $\varpi \in (0,1)$. Suppose that $\boldsymbol{\delta} = c(\mathbf{1}_s^\top, \mathbf{0}_{p-s}^\top)^\top$ for $c \ne 0$. Then $\eta = \|\boldsymbol{\delta}\|^2\{(1+\varpi^2)s - 2\varpi(s-1) - \varpi^2\}/(1-\varpi^2)$. It is straightforward to see that $\eta \asymp \|\boldsymbol{\delta}\|^2$ is true for any $\varpi > 0$ and $s > 1$.

It should be emphasized that this asymptotic comparison between the WSP and SS tests may not be fair, because the conditions on which the validity of asymptotic results of the two tests rely, are rather different. Our proposed test requires the good behavior of the sample-splitting estimation, which is clearly more stringent than the conditions used in [30]. Our numerical results in Section 4 demonstrate that the WSP test is capable of reducing multiplicity and correlation, and thus obtain significant power improvement over SS when the $\boldsymbol{\Sigma}$ is not sparse or heterogeneity of variance exists.

3 Extension to Two-sample Location Problem

3.1 Optimal Projection Test

Assume that $\mathbf{X}_1, \ldots, \mathbf{X}_n$ is an independent sample with sample size n, and $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$ is an independent sample with sample size m, from p-variate distributions $E_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $E_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, respectively. Consider two sample problem:

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \longleftrightarrow H_1 : \boldsymbol{\mu}_1 \ne \boldsymbol{\mu}_2. \tag{3.1}$$

Assume that $n/N \to w \in (0,1)$ as $N \to \infty$, where $N = n + m$. Denote $\boldsymbol{\varepsilon}_{1i} = \boldsymbol{\Gamma}(\mathbf{X}_i - \boldsymbol{\mu}_1)$, $\boldsymbol{\varepsilon}_{2i} = \boldsymbol{\Gamma}(\mathbf{Y}_i - \boldsymbol{\mu}_2)$ and consider the weighted signs $\mathbf{V}_{1i} = K(r_{1i})U(\boldsymbol{\Gamma}(\mathbf{X}_i - \boldsymbol{\mu}))$ and $\mathbf{V}_{2i} = K(r_{2i})U(\boldsymbol{\Gamma}(\mathbf{Y}_i - \boldsymbol{\mu}))$, where $r_{1i} = \|\boldsymbol{\Gamma}(\mathbf{X}_i - \boldsymbol{\mu})\|$, $r_{2i} = \|\boldsymbol{\Gamma}(\mathbf{Y}_i - \boldsymbol{\mu})\|$ and $\boldsymbol{\mu}$ is a "working" parameter. It is tantamount to considering the test problem,

$$H_0 : E(\mathbf{V}_{1i}) = E(\mathbf{V}_{2i}) \longleftrightarrow H_1 : E(\mathbf{V}_{1i}) \ne E(\mathbf{V}_{2i}).$$

Consider a $p \times k$ full rank projection matrix $\mathbf{A}$ with $k \ll p$ and $k < n$. Under H_0, if $\boldsymbol{\mu} = \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$,

$$\sqrt{\frac{nmp}{n+m}} E^{-1/2}\{K^2(r_{1i})\}(\mathbf{A}^\top\mathbf{A})^{-1/2}\mathbf{A}^\top(\bar{\mathbf{V}}_1 - \bar{\mathbf{V}}_2) \xrightarrow{d} N(\mathbf{0}, \mathbf{I}_k),$$

where $\bar{\mathbf{V}}_1 = n^{-1}\sum_{i=1}^n \mathbf{V}_{1i}$ and $\bar{\mathbf{V}}_2 = m^{-1}\sum_{i=1}^m \mathbf{V}_{2i}$. Hence, we could use

$$Q_{\mathbf{A}} = \frac{nmp}{n+m} E^{-1}\{K^2(r_{1i})\}(\bar{\mathbf{V}}_1 - \bar{\mathbf{V}}_2)^\top \mathbf{A}(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top(\bar{\mathbf{V}}_1 - \bar{\mathbf{V}}_2)$$

to assess the equality of $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$.

Consider the local alternative hypothesis

$$H_{1N} : E(\|\boldsymbol{\varepsilon}_{1i}\|^{-2})(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = O\{(Np)^{-1}\}.$$

$Q_{\mathbf{A}}$ is asymptotically $\chi_k^2(\delta_{\mathbf{A}})$ distributed under H_{1N}, where

$$\delta_{\mathbf{A}} = \frac{nmp}{n+m} E^{-1}\{K^2(r_{1i})\}\{E(\mathbf{V}_{1i}) - E(\mathbf{V}_{2i})\}^\top \mathbf{A}(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top\{E(\mathbf{V}_{1i}) - E(\mathbf{V}_{2i})\}.$$

Similar to Theorem 2.1, we have the following result which helps us to determine $\mathbf{A}$.

Theorem 3.1 *Suppose that Assumption* 1 *holds for* $\boldsymbol{\varepsilon}_{1i}$ *and the conditions on the function* $K(\cdot)$ *in Theorem* 2.1 *hold. Set* $\boldsymbol{\mu} = (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2$ *in* $Q_{\mathbf{A}}$. *The projection test* $Q_{\mathbf{A}}$ *reaches its asymptotic best power under* H_{1N} *at* $k = 1$ *and* $\mathbf{A} = E(\mathbf{V}_{1i}) - E(\mathbf{V}_{2i})$ *with the weight function* $K(t) = t^{-1}$.

Let $\mathbf{a} = E(\mathbf{V}_{1i}) - E(\mathbf{V}_{2i})$ be the "oracle" optimal direction. Similarly, $Q_{\mathbf{a}}$ can be carried out with estimated $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\mathbf{a}$ and $\boldsymbol{\Gamma}$. We divide the random sample into two separate sets: $\mathcal{U}_j$ with the sample sizes (n_j, m_j) for $j = 1, 2$. We use $\mathcal{U}_1$ to estimate the parameters and $\mathcal{U}_2$ to construct $Q_{\mathbf{a}}$. Let $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\Sigma}}$ be the pooled sample mean and covariance matrix (with the ridge-modification) computed from $\mathcal{U}_1$, respectively. Define $\widehat{\boldsymbol{\Gamma}}^\top\widehat{\boldsymbol{\Gamma}} = \widehat{\boldsymbol{\Sigma}}^{-1}$, $\widehat{\mathbf{V}}_{1i} = U(\widehat{\boldsymbol{\Gamma}}(\mathbf{X}_i - \widehat{\boldsymbol{\mu}}))/\|\widehat{\boldsymbol{\Gamma}}(\mathbf{X}_i - \widehat{\boldsymbol{\mu}})\|$, and $\widehat{\mathbf{V}}_{2i} = U(\widehat{\boldsymbol{\Gamma}}(\mathbf{Y}_i - \widehat{\boldsymbol{\mu}}))/\|\widehat{\boldsymbol{\Gamma}}(\mathbf{Y}_i - \widehat{\boldsymbol{\mu}})\|$. Correspondingly, $\mathbf{a}$ is estimated by

$$\widehat{\mathbf{a}} = n_1^{-1}\sum_{i=1}^{n_1}\widehat{\mathbf{V}}_{1i} - m_1^{-1}\sum_{i=1}^{m_1}\widehat{\mathbf{V}}_{2i}.$$

Let $\tilde{\mathbf{V}}_1 = n_2^{-1}\sum_{i=n_1+1}^{n}\widehat{\mathbf{V}}_{1i}$ and $\tilde{\mathbf{V}}_2 = m_2^{-1}\sum_{i=m_1+1}^{m}\widehat{\mathbf{V}}_{2i}$. Similar to (2.2), we use

$$T_{\widehat{\mathbf{a}}}^2 = \frac{n_2 m_2}{n_2 + m_2}\{\widehat{\mathbf{a}}^\top(\tilde{\mathbf{V}}_1 - \tilde{\mathbf{V}}_2)\}^2/\widehat{\sigma}_2^2$$

where

$$\widehat{\sigma}_2^2 = (n_2 + m_2 - 2)^{-1}\bigg\{\sum_{i=n_1+1}^{n}(\widehat{\mathbf{a}}^\top\widehat{\mathbf{V}}_{1i} - \widehat{\mathbf{a}}^\top\tilde{\mathbf{V}}_1)^2 + \sum_{i=m_1+1}^{m}(\widehat{\mathbf{a}}^\top\widehat{\mathbf{V}}_{2i} - \widehat{\mathbf{a}}^\top\tilde{\mathbf{V}}_2)^2\bigg\}.$$

$T_{\widehat{\mathbf{a}}}$ is similar to a classical two-sample t-test. Its finite-sample p-value can be approximated by using the t-distribution with $n_2 + m_2 - 2$ degrees of freedom.

3.2 Asymptotic Power Comparison

Consider a local alternative hypothesis $H_1 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \boldsymbol{\delta}/\sqrt{N}$. We assume that $\sqrt{N_2/N} \to b > 0$ as $N \to \infty$, where $N_2 = n_2 + m_2$. Under certain conditions, the ARE of the proposed WSP test with respect to [19]'s moment-based projection test in the two-sample setting is $\mathrm{ARE}(\mathrm{WSP}, \mathrm{MP}) = E(\|\boldsymbol{\varepsilon}_{1i}\|^2)E(\|\boldsymbol{\varepsilon}_{1i}\|^{-2})$, which is the same as that for the one-sample problem as discussed in Section 2.4.2, implying that the proposed test could be more powerful for heavy-tailed distributions. Similarly, the comparison between the proposed test and [9] test is basically similar to that between WSP and [30]'s SS test in Section 2.4.3. Therefore, details are not elaborated here.

4 Numerical Studies

In this section, we demonstrate our method with numerical studies. All the simulation results are based on 1000 independent replications and all the comparisons are made at 0.05 significance level. For $\rho \in (0, 1)$, we set two structures for the correlation matrix $\mathbf{R}$: (i) the compound symmetry correlation $\mathbf{R}_1$, say all off-diagonal elements of the correlation matrix $\mathbf{R}$ are ρ and all diagonal elements are 1; (ii) the autoregressive structure, $\mathbf{R}_2 = (\rho^{|i-j|})$.

4.1 Influence of the Tuning Parameters

Here we study how to determine the sample sizes n_1 and n_2 and the λ in the sample-splitting based direction estimation proposed in Section 2.2.

Let the sample sizes of $\mathcal{U}_1$ and $\mathcal{U}_2$ be $n_1 = \lfloor n\kappa \rfloor$ and $n_2 = n - n_1$, where $\kappa \in (0, 1)$ is the sample splitting percentage and $\lfloor \cdot \rfloor$ denotes the rounding operator. We examine the influence of κ on the performance of our new test by setting a grid of κ which are over $(0, 1)$. We generate a random sample of size n with mean $\boldsymbol{\mu}$, the percentage of $\mu_j = 0$, for $j = 1, \ldots, p$ being chosen to be 50% or 95% which are the dense case or sparse case. For simplicity, all the non-zero components are equal to a constant c. Consider the testing power functions with correlations 0.95, 0.75, 0.5, 0.25 and 0. Figure 2 depicts the power as a function of κ with two different covariance structures for sparse case when $c = 0.24$, $n = 80$ and $p = 480$. The conclusions for other cases are similar. We can observe that different choices of κ indeed affect the power and most peaks locate at the range from 0.4 to 0.6. 0.4 appears to be an appropriate choice and we will use it in other simulations throughout this paper.

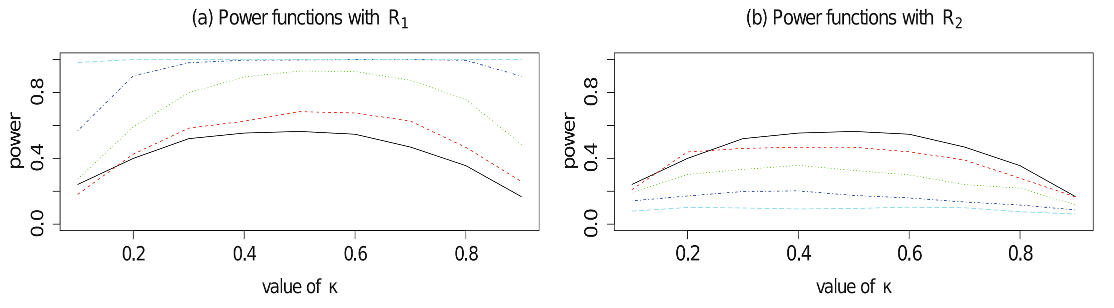
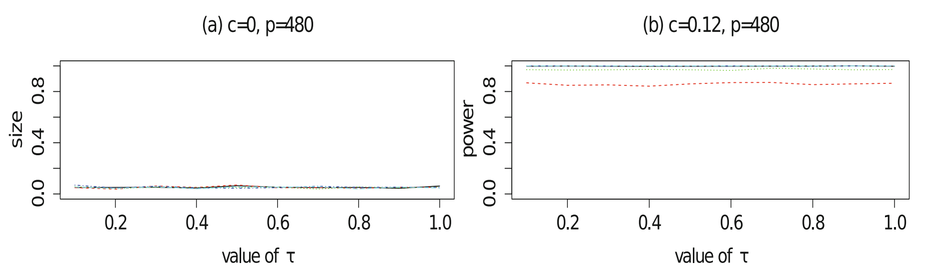


Figure 2 Power of WSP test as a function with respect to κ with $(n, p, c) = (80, 480, 0.24)$. Random sample are from normal distribution. The testing power functions with correlations 0.95, 0.75, 0.5, 0.25 and 0 are represented by bold dashed, bold dotdash, bold dotted, dashed and solid lines.

Next, we consider the choice of the tuning parameter, λ, in $(\mathbf{S}_1 + \lambda \mathbf{D}_{\mathbf{S}_1})^{-1}$. As we require $\lambda \to 0$ as $n_1 \to \infty$, we set $\lambda = n_1^{-\tau}$. Figure 3 depicts the power as a function with respect to τ with $\mathbf{R} = \mathbf{R}_1$ when $n = 80$, $p = 480$ and $n_1 = 0.4n$ with dense case for $c = 0.12$, and sparse cases for two other non-zero values of c. Clearly, the value of τ has little effect on the performance of the WSP test. Therefore, we set $\lambda = n_1^{-0.5}$ in the simulations.



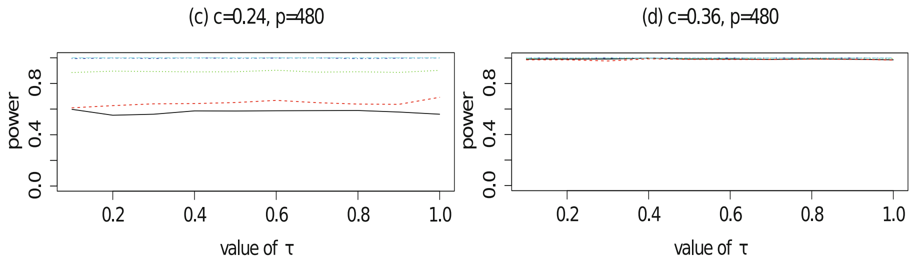


Figure 3 Power of our WSP test as a function with respect to τ with $n = 80$, $p = 480$ and covariance matrix $\mathbf{R}_1$

Though we have theoretically verified that our WSP test works well with thresholded estimation of $\mathbf{d}$ and $\boldsymbol{\Sigma}$ under certain sparsity assumptions in Section 2.4, our numerical results indicate that using ridge-type estimate of $\boldsymbol{\Sigma}$ and $\widehat{\mathbf{d}} = n_1^{-1}\sum_{i=1}^{n_1}\widehat{\boldsymbol{V}}_i$ performs reasonably well for both the sparse and dense models considered here. Hence, throughout the main text, our method is constructed with the non-thresholded estimators. Some additional results with thresholded estimators are available in the Supplementary Material.

4.2 Size and Power Comparison in One-sample Problem

Here we study size and power performance. We choose the following representative examples for illustration.

(I) Multivariate normal distribution $\mathbf{X}_i \sim N(\boldsymbol{\mu}, \mathbf{R})$.

(II) Multivariate normal distribution with different component variances. $\mathbf{X}_i \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = \mathbf{D}_{\boldsymbol{\Sigma}}^{1/2}\mathbf{R}\mathbf{D}_{\boldsymbol{\Sigma}}^{1/2}$ and $\mathbf{D}_{\boldsymbol{\Sigma}} = \mathrm{diag}\{d_1^2, \dots, d_p^2\}$, d_j^2's are generated from χ_4^2.

(III) Multivariate t-distribution $t_3(\boldsymbol{\mu}, \mathbf{R})$.

(IV) Multivariate t-distribution with different component variances. $\mathbf{X}_i \sim t_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = \mathbf{D}_{\boldsymbol{\Sigma}}^{1/2}\mathbf{R}\mathbf{D}_{\boldsymbol{\Sigma}}^{1/2}$, d_j^2's are generated from χ_4^2.

(V) Multivariate mixture normal distribution $MN(0.8, 10, \mathbf{R})$.

In our simulation, we also consider the following moving average model:

$$X_{ij} = \|\varrho\|^{-1}(\varrho_1 Z_{ij} + \varrho_2 Z_{i(j+1)} + \cdots \varrho_T Z_{i(j+T-1)}) + \mu_j$$

for $i = 1, \dots, n$ and $j = 1, \dots, p$, where $\varrho = (\varrho_1, \dots, \varrho_T)^\top$ and $\{Z_{ij}\}$ are i.i.d. random variables. The following three scenarios for the innovation $\{Z_{ij}\}$ are considered:

(VI) All the $\{Z_{ij}\}$'s are from $N(0, 1)$;

(VII) All the $\{Z_{ij}\}$'s are from t_3;

(VIII) All the $\{Z_{ij}\}$'s are from $0.8N(0, 1) + 0.2N(0, 16)$.

$\mathbf{X}_i$'s are then obtained by multiplying $\mathbf{D}_0^{1/2}$, where $\mathbf{D}_0 = \mathrm{diag}\{d_1^2, \dots, d_p^2\}$ and d_j^2's are generated from χ_3^2. The coefficients $\{\varrho_i\}_{l=1}^T$ are generated from $U(0, 1)$ independently and are kept fixed in our simulations once generated with $T = 2$. Note that the last two scenarios are clearly non-elliptical setting and can be framed into the independent component model in Section 2.3.

In this section we compare our proposed WSP test with some existing methods, including [19]'s moment-based projection test (MP), [30]'s spatial sign test (SS) and the projection test without weights (SP) as described in Section 2.4. [19] have made comprehensive investigation about the performances of the MP test relative to some other existing high-dimensional tests, such as the tests proposed by [1], [5] and [27], and thus those tests are not included in our study. Both dense and sparse cases are considered. Because the power of our WSP test depends on $n^{-1}(p-2)\eta E(\|\boldsymbol{\varepsilon}_i\|^{-2})$ as shown in Section 2.4.1, we adjust $\boldsymbol{\mu}$ for each scenario so that this value is fixed as 2 and 1 for the compound symmetry correlation and autoregressive structures, respectively. For Scenarios (VI)–(VIII), we set $\boldsymbol{\mu}^\top\{\mathrm{diag}(\boldsymbol{\Omega})\}^{-1}\boldsymbol{\mu} = 0.007$.

	Size				Dense				Sparse			
	WSP	SP	MP	SS	WSP	SP	MP	SS	WSP	SP	MP	SS
$(n,p)=(80,480)$												
(I)	5.4	5.6	5.5	6.7	76.9	76.6	76.0	16.9	71.7	71.2	70.6	10.8
(II)	4.3	4.4	4.2	7.5	74.0	73.9	73.3	9.4	73.3	73.0	72.1	9.8
(III)	4.7	6.4	5.8	8.6	69.1	64.2	39.1	14.4	68.2	63.8	39.9	8.5
(IV)	5.6	5.2	4.7	7.7	71.3	64.3	38.2	10.2	66.6	61.6	38.9	9.5
(V)	3.5	4.6	5.3	7.9	57.5	29.4	9.1	9.3	58.3	32.4	8.5	9.1
$(n,p)=(80,1440)$												
(I)	5.7	5.4	5.3	6.1	30.6	28.9	29.8	9.5	31.9	31.5	31.4	8.5
(II)	5.1	5.2	5.3	6.2	30.8	30.9	30.8	9.4	32.8	32.3	31.7	6.8
(III)	4.3	5.1	4.2	5.9	29.3	26.5	15.6	9.0	28.8	24.8	15.1	8.4
(IV)	4.8	5.1	5.4	7.1	30.0	26.2	15.9	7.8	30.5	26.1	15.8	8.7
(V)	5.4	4.7	4.8	6.9	21.6	12.6	6.3	8.8	22.6	11.7	5.9	8.9
$(n,p)=(200,1440)$												
(I)	4.5	4.6	4.6	6.8	97.8	97.8	97.6	14.0	98.1	98.1	98.0	8.9
(II)	4.6	4.5	4.2	5.2	97.9	97.8	97.7	9.8	97.8	97.8	97.8	8.7
(III)	4.5	4.4	4.9	6.2	97.8	95.7	72.6	11.5	97.0	94.6	68.4	7.4
(IV)	6.2	5.0	3.3	6.1	97.8	95.9	71.4	9.0	95.4	93.2	68.6	7.2
(V)	5.1	5.2	4.6	6.5	94.0	66.5	16.5	11.0	94.7	67.4	17.5	8.8

Table 1 Empirical size and power (%) comparison under Scenarios (I)–(V) with the compound symmetry correlation structure ($\mathbf{R}_1$) when $\rho = 0.5$

Three combinations of (n,p) are used: $(80,480)$, $(80,1440)$ and $(200,1440)$. The results of empirical size and power comparison under the compound symmetry, autoregressive and moving average structures are given in Tables 1–3, respectively. We observe that the sizes of the WSP test are close to the nominal level under all the scenarios. Generally, under Scenarios (I) and (II), the WSP, MP and SP tests have almost identical performances, since the underlying distribution is multivariate normal. This can be understood from the discussions in Section 2.4:

the three tests are equivalently efficient as $p \to \infty$. Under the other elliptical scenarios (III)–(V), the WSP test is clearly more efficient than the MP test, and the difference is quite remarkable. Certainly, this is not surprising as the asymptotic relative efficiency of WSP with respect to MP is quite high under these distributions. Similarly, the WSP outperforms the projection test without weights, which again concurs with the asymptotic comparison in Section 2.4.1. The SS test is also able to maintain sizes in a reasonable range, though in some cases it appears to be a little liberal. The SS test is outperformed by the other three tests when the correlation matrix is dense, such as in Table 1. The advantages of the three projection tests are partially due to the utilization of the correlation information without sacrificing the null performances. When the correlation matrix possesses certain sparseness, such as the autoregressive structure, the SS test could be much more powerful than the other three tests; see Scenarios (I), (III) and (V) in Table 2. However, this superiority would be largely compromised when the variances of components are not identical, such as the cases of Scenarios (II) and (IV).

	Size				Dense				Sparse			
	WSP	SP	MP	SS	WSP	SP	MP	SS	WSP	SP	MP	SS
					$(n,p)=(80,480)$							
(I)	5.2	5.0	4.8	5.6	84.1	84.2	84.7	99.3	79.4	80.0	79.7	99.3
(II)	4.8	4.8	4.2	6.2	56.6	57.2	57.1	63.6	62.6	62.8	62.5	78.4
(III)	5.2	4.4	4.0	6.1	81.0	76.0	48.0	98.5	74.8	69.0	45.2	97.6
(IV)	4.7	5.2	4.7	7.6	52.3	49.1	30.2	55.1	54.5	50.7	28.7	60.7
(V)	4.6	4.2	4.0	8.9	71.8	42.4	12.2	65.2	62.2	37.4	10.7	56.4
					$(n,p)=(80,1440)$							
(I)	5.6	5.3	5.4	4.5	52.9	53.2	53.2	92.4	55.5	55.2	55.2	92.3
(II)	5.9	5.8	5.8	5.2	24.5	24.2	24.1	22.7	23.3	23.7	23.7	25.0
(III)	5.0	5.3	3.9	5.1	49.1	44.2	26.1	83.1	50.2	45.6	26.6	82.3
(IV)	4.5	4.8	4.8	6.3	20.7	18.2	12.2	18.4	21.8	20.7	10.5	18.1
(V)	5.3	4.8	5.0	12.5	40.9	23.2	7.3	35.2	38.5	21.6	8.1	35.0
					$(n,p)=(200,1440)$							
(I)	4.4	4.3	4.3	5.6	99.4	99.4	99.3	100.0	99.2	99.2	99.3	100.0
(II)	4.4	4.6	4.5	4.6	79.4	79.5	79.3	61.4	80.1	79.8	80.1	67.0
(III)	3.9	4.2	5.0	4.5	98.9	97.2	80.4	100.0	97.9	97.2	77.5	100.0
(IV)	5.4	4.7	4.0	5.9	79.2	72.9	42.6	53.1	76.1	70.6	42.1	55.6
(V)	4.6	4.1	4.1	8.1	98.2	78.0	22.7	88.4	98.0	73.3	21.8	83.2

Table 2 Empirical size and power comparison under Scenarios (I)–(V) with the autoregressive structure ($\mathbf{R}_2$) when $\rho = 0.5$

Table 3 reports empirical sizes and power comparison under Scenarios (VI)–(VIII). We observe that the WSP, SP and MP tests perform equivalently well in all the cases, though the WSP appears to be slightly better. In fact, this should not be surprising to us because the ARE values $\mathrm{ARE}(\mathrm{WSP}, \mathrm{SP}) = E(\|\boldsymbol{\varepsilon}_i\|^{-2})/\{E(\|\boldsymbol{\varepsilon}_i\|^{-1})\}^2$ and $\mathrm{ARE}(\mathrm{WSP}, \mathrm{MP}) = E(\|\boldsymbol{\varepsilon}_i\|^2)E(\|\boldsymbol{\varepsilon}_i\|^{-2})$ are very close to one under these three models.

	Size				Dense				Sparse			
	WSP	SP	MP	SS	WSP	SP	MP	SS	WSP	SP	MP	SS
					$(n,p)=(80,480)$							
(VI)	3.5	3.6	3.6	5.4	95.6	95.7	95.5	76.1	95.1	95.2	95.1	86.5
(VII)	4.5	4.5	4.7	5.5	41.1	41.1	39.1	18.4	39.7	39.3	37.6	25.0
(VIII)	4.9	4.6	4.6	5.3	21.0	21.7	21.2	17.3	23.1	22.9	22.3	22.2
					$(n,p)=(80,1440)$							
(VI)	4.0	4.1	3.7	5.1	100.0	100.0	100.0	66.4	100.0	100.0	100.0	98.6
(VII)	5.4	5.4	5.1	5.8	81.1	81.6	80.8	39.7	81.2	81.0	79.2	37.0
(VIII)	4.2	4.6	4.6	5.4	56.0	55.8	55.7	14.4	52.6	52.3	51.8	26.5

Table 3 Empirical size and power comparison under Scenarios (VI)–(VIII)

R		Size			Dense			Sparse		
		WSP	MP	SS	WSP	MP	SS	WSP	MP	SS
					$(n,p)=(80,480)$					
	(I)	4.9	5.7	5.3	83.5	84.2	14.2	82.6	84.7	9.7
	(III)	4.5	4.7	5.9	77.6	42.5	12.3	75.6	42.4	9.5
	(V)	5.0	4.9	5.8	30.3	8.1	6.9	29.5	10.1	6.3
$\mathbf{R}_1$					$(n,p)=(80,1440)$					
	(I)	5.0	4.7	4.7	40.0	38.1	9.1	39.2	40.1	7.4
	(III)	4.6	4.8	5.4	36.3	20.7	8.3	37.7	20.5	7.7
	(V)	5.1	4.6	5.1	14.2	7.0	6.5	14.3	6.5	5.6
					$(n,p)=(80,480)$					
	(I)	5.7	5.3	5.4	79.9	79.6	99.8	75.3	73.9	99.4
	(III)	5.4	4.0	5.3	75.5	42.3	98.7	71.3	39.0	97.1
	(V)	5.0	5.3	4.9	39.7	12.7	26.1	35.8	10.2	24.4
$\mathbf{R}_2$					$(n,p)=(80,1440)$					
	(I)	5.4	5.3	5.7	54.3	54.1	92.7	52.3	51.9	91.2
	(III)	5.5	4.5	5.8	47.7	27.9	78.7	43.8	25.5	76.2
	(V)	5.2	5.9	5.3	26.2	7.3	10.5	25.5	8.0	10.1

Table 4 Empirical size and power comparison in two-sample problem under Scenarios (I), (III) and (V) when $\rho = 0.5$

4.3 Size and Power Comparison in Two-sample Problem

Here we compare our proposed WSP test with [19]'s moment-based projection test (MP), and [9]'s spatial sign test (SS). We consider $\mathbf{X}_i$ and $\mathbf{Y}_i$ follow multivariate normal, multivariate t and multivariate mixture normal distributions like Scenarios (I), (III) and (V) in one-sample problem with location parameters $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, respectively. Again, the two covariance structures, the compound symmetry $\mathbf{R}_1$ and the autoregressive correlation $\mathbf{R}_2$, are used. Without loss of generality, we choose $n = m$. Under the alternative hypothesis, set $\boldsymbol{\mu}_1 = 0$ and choose $\boldsymbol{\mu}_2$ such that the percentage of $\mu_{1j} = \mu_{2j}$ for $j = 1, \ldots, p$ are 50% (dense case) or 95% (sparse case). We fix $(p-2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)E(\|\boldsymbol{\varepsilon}_{1i}\|^{-2})$ as 4 and 2 for $\mathbf{R}_1$ and $\mathbf{R}_2$, respectively.

The results were reported in Table 4. Similar conclusions to the one-sample problem can be drawn: the improvement of WSP over the other two tests under heavy-tailed errors could be quite substantial.

4.4 A Real-data Application

In this section, we apply the proposed methodology to a real data set. The diffuse large B-cell lymphoma (DLBCL) data set (http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi) includes the expression of the 661 genes across the 93 symbols with 42 ones in group one and 51 in group two. The DLBCL is a highly heterogeneous disease and the two groups have different clinical manifestation. We want to test whether group one and group two have the same gene expression levels or not. This is a high-dimensional test problem with dimensions 661. From the top two panels of Figure 4, we observe that many variables of the two groups are not normal distributed. Thus, we could expect that WSP test would be more robust. Furthermore, from the bottom panels of Figure 4, many between-variables correlations of these two samples are quite large, which illustrates that a test which is capable of utilizing correlations would be more desirable. The p-values of the WSP and MP tests are $4.58e-13$ and $2.01e-11$, respectively. These results suggest the rejection of the null hypothesis. However, the p-value of WSP test is smaller for this data set, proving more strong evidence that the gene expression levels of the two groups are significantly different.

Furthermore, we artificially re-arrange data to see how robust our proposed method is. Firstly, we perform 661 marginal t-tests to study the significance of each variable and then divide all the variables into two separate parts by p-values: smaller than 0.01 and larger than 0.01, which contain 240 and 421 variables, respectively. Then we compute the empirical size by randomly drawing $p = 350$ variables from the second part and performing the two-sample WSP and MP tests on those 350 variables. Similarly, the empirical power is approximated by drawing 10 variables from the first part and 340 variables from the second part. Figure 5 shows the ROC curves (power against significance level) of both the two tests with 1000 replications. The advantage of our WSP test is clear.

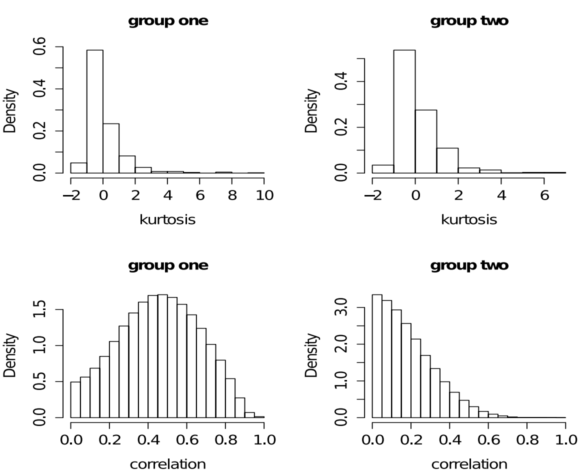


Figure 4 Histograms of marginal kurtosis values and between-variables correlations

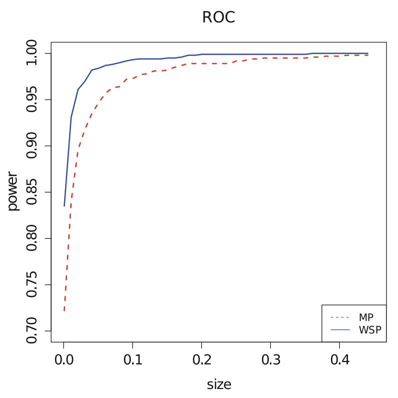


Figure 5 ROC curves of the proposed test and MP test on the diffuse large B-cell lymphoma data

5 Concluding Remarks

A natural concern is whether the newly proposed test in this paper can handle the ultra-high dimensional scenarios with larger p, say, at an exponential rate of n. Unfortunately, it is very difficult, if not impossible, when there is no sparse structure for all existing scale-invariant tests to correct bias-terms. Thus in general, it is still an open problem for whether we can define a test statistic that is (at least) asymptotically effective without sparsity assumption on the data structure. This paper offers the insight on how to find the optimal weights and directions on high-dimensional sign paradigms, and the results in this paper also serve as a preliminary step for us to carefully take this challenge into consideration. One may be wary of the problem of the single sample-splitting method, say its sensitivity with respect to the choice of splitting the entire sample: sample splits may lead to different decisions. To avoid the so-called "p-value lottery", we may consider employing some bagging technique to aggregate results from multiple sample-splitting. In the Supplementary Material, we provide some preliminary analysis but some future study is definitely needed.

Appendix

This appendix contains several lemmas and the proofs of Theorems 2.1–2.3 and 3.1.

Proof of Theorem 2.1 $\mathbf{D}^\top\mathbf{V}_1,\dots,\mathbf{D}^\top\mathbf{V}_n$ are i.i.d. k-variate random vectors. Denote $\mathbf{D}_1 = (\mathbf{D}^\top\mathbf{D})^{-1/2}\mathbf{D}^\top$ and $\sigma_n^2 = (np)^{-1}E\{K^2(r_i)\}$. By the Central Limit Theorem, we have under the null hypothesis,

$$\sqrt{npE^{-1}\{K^2(r_i)\}}\mathbf{D}_1\bar{\mathbf{V}} \xrightarrow{d} N_k(\mathbf{0},\mathbf{I}_k).$$

Under H_{1n}, according to the Taylor expansion, we have

$$U(\mathbf{\Gamma}\mathbf{X}_i) = \mathbf{u}_i + \|\boldsymbol{\varepsilon}_i\|^{-1}(\mathbf{I}_p - \mathbf{u}_i\mathbf{u}_i^\top)\mathbf{\Gamma}\boldsymbol{\mu} + \boldsymbol{Q}_i, \tag{A.1}$$

$$\|\mathbf{\Gamma}\mathbf{X}_i\| = \|\boldsymbol{\varepsilon}_i\| + J_i, \tag{A.2}$$

where $\mathbf{u}_i = U(\boldsymbol{\varepsilon}_i)$. By triangle inequality, the remainder term $\boldsymbol{Q}_i$ satisfies

$$\|\boldsymbol{Q}_i\| \le C\|\boldsymbol{\varepsilon}_i\|^{-2}\|\mathbf{\Gamma}\boldsymbol{\mu}\|^2, \tag{A.3}$$

and J_i satisfies $|J_i| \le \|\mathbf{\Gamma}\boldsymbol{\mu}\|$. Consider

$$\mathbf{D}_1\{\bar{\mathbf{V}} - E(\mathbf{V})\} = n^{-1}\mathbf{D}_1 \sum_{i=1}^n K(r_i)\mathbf{u}_i + n^{-1}\mathbf{D}_1 \sum_{i=1}^n \{\mathbf{V}_i - K(r_i)\mathbf{u}_i - E(\mathbf{V}_i)\}.$$

Denote $\mathbf{I}_1 = n^{-1}\mathbf{D}_1 \sum_{i=1}^n K(r_i)\mathbf{u}_i$ and $\mathbf{I}_2 = n^{-1}\mathbf{D}_1 \sum_{i=1}^n\{\mathbf{V}_i - K(r_i)\mathbf{u}_i - E(\mathbf{V}_i)\}$. First, $\mathbf{I}_1/\sigma_n \xrightarrow{d} N_k(\mathbf{0}, \mathbf{I}_k)$. Denote

$$\begin{aligned}\mathbf{R}_0 = \mathbf{D}_1 n^{-1} \sum_{i=1}^n \{&\|\boldsymbol{\varepsilon}_i\|^{-1} K(r_i)(\mathbf{I}_p - \mathbf{u}_i\mathbf{u}_i^\top)\mathbf{\Gamma}\boldsymbol{\mu} \\ &- E\{\|\boldsymbol{\varepsilon}_i\|^{-1}K(r_i)\}\mathbf{\Gamma}\boldsymbol{\mu} + K(r_i)\boldsymbol{Q}_i - E\{K(r_i)\boldsymbol{Q}_i\}\}.\end{aligned}$$

Using (A.1)–(A.2), $\mathbf{I}_2 = \mathbf{D}_1\mathbf{R}_0$. By (A.3),

$$\begin{aligned}E(\|\mathbf{R}_0\|^2) \le A[&n^{-1}E\{\|\boldsymbol{\varepsilon}_i\|^{-2}K^2(r_i)\}\boldsymbol{\mu}^\top\mathbf{\Sigma}^{-1}\boldsymbol{\mu} \\ &+ n^{-1}C^2E\{\|\boldsymbol{\varepsilon}_i\|^{-4}K^2(r_i)\}(\boldsymbol{\mu}^\top\mathbf{\Sigma}^{-1}\boldsymbol{\mu})^2]\{1 + o(1)\},\end{aligned}$$

where A is some constant.

By Cauchy inequality and H_{1n},

$$\begin{aligned}\frac{pE\{\|\boldsymbol{\varepsilon}_i\|^{-2}K^2(r_i)\}\boldsymbol{\mu}^\top\mathbf{\Sigma}^{-1}\boldsymbol{\mu}}{E\{K^2(r_i)\}} &\le \frac{p[E(\|\boldsymbol{\varepsilon}_i\|^{-4})E\{K^4(r_i)\}]^{1/2}\boldsymbol{\mu}^\top\mathbf{\Sigma}^{-1}\boldsymbol{\mu}}{E\{K^2(r_i)\}} \\ &= O\{pE(\|\boldsymbol{\varepsilon}_i\|^{-2})(\boldsymbol{\mu}^\top\mathbf{\Sigma}^{-1}\boldsymbol{\mu})\} = O(n^{-1}).\end{aligned}$$

Similarly,

$$\begin{aligned}\frac{pE\{\|\boldsymbol{\varepsilon}_i\|^{-4}K^2(r_i)\}(\boldsymbol{\mu}^\top\mathbf{\Sigma}^{-1}\boldsymbol{\mu})^2}{E\{K^2(r_i)\}} &\le \frac{p[E(\|\boldsymbol{\varepsilon}_i\|^{-8})E\{K^4(r_i)\}]^{1/2}(\boldsymbol{\mu}^\top\mathbf{\Sigma}^{-1}\boldsymbol{\mu})^2}{E\{K^2(r_i)\}} \\ &= O\{pE^2(\|\boldsymbol{\varepsilon}_i\|^{-2})(\boldsymbol{\mu}^\top\mathbf{\Sigma}^{-1}\boldsymbol{\mu})^2\} = O(n^{-1}).\end{aligned}$$

By noting that $\lambda_{\max} = 1$, it follows that $E(\|\mathbf{I}_2\|^2)/\sigma_n^2 \le \lambda_{\max}E(\|\mathbf{R}_0\|^2)/\sigma_n^2 = o(1)$, where $\lambda_{\max}$ is the maximum eigenvalue of $\mathbf{D}(\mathbf{D}^\top\mathbf{D})^{-1}\mathbf{D}^\top$. Hence

$$\mathbf{D}_1\{\bar{\mathbf{V}} - E(\mathbf{V})\}/\sigma_n \xrightarrow{d} N_k(\mathbf{0}, \mathbf{I}_k).$$

Denote $\delta_{\mathbf{D}} = \sigma_n^{-2}[E(\mathbf{V})]^T\mathbf{D}(\mathbf{D}^\top\mathbf{D})^{-1}\mathbf{D}^\top E(\mathbf{V})$. When $p \to \infty$, $\|\boldsymbol{\varepsilon}_i\| \xrightarrow{p} \infty$. Assumption 2 implies that $K(r_i) = K(\|\boldsymbol{\varepsilon}_i\|)\{1 + O_p(J_i)\} = K(\|\boldsymbol{\varepsilon}_i\|)\{1 + O(\|\boldsymbol{\mu}^\top\mathbf{\Sigma}^{-1}\boldsymbol{\mu}\|)\}$. Using Lemmas A.1–A.2 again, we have

$$\begin{aligned}\delta_{\mathbf{D}} &= np\frac{E^2\{K(r_i)\|\boldsymbol{\varepsilon}_i\|^{-1}\}}{E\{K^2(r_i)\}}\boldsymbol{\mu}^\top\mathbf{\Gamma}^\top\mathbf{D}(\mathbf{D}^\top\mathbf{D})^{-1}\mathbf{D}^\top\mathbf{\Gamma}\boldsymbol{\mu}\{1 + o(1)\}, \\ &= np\frac{E^2\{K(\|\boldsymbol{\varepsilon}_i\|)\|\boldsymbol{\varepsilon}_i\|^{-1}\}}{E\{K^2(\|\boldsymbol{\varepsilon}_i\|)\}}\boldsymbol{\mu}^\top\mathbf{\Gamma}^\top\mathbf{D}(\mathbf{D}^\top\mathbf{D})^{-1}\mathbf{D}^\top\mathbf{\Gamma}\boldsymbol{\mu}\{1 + o(1)\}.\end{aligned}$$

By the continuous mapping theorem,

$$Q_{\mathbf{D}} \xrightarrow{d} \chi_k^2\left(np\frac{E^2\{K(\|\boldsymbol{\varepsilon}_i\|)\|\boldsymbol{\varepsilon}_i\|^{-1}\}}{E\{K^2(\|\boldsymbol{\varepsilon}_i\|)\}}\boldsymbol{\mu}^\top\mathbf{\Gamma}^\top\mathbf{D}(\mathbf{D}^\top\mathbf{D})^{-1}\mathbf{D}^\top\mathbf{\Gamma}\boldsymbol{\mu}\right).$$

Hence, for any given k and $\mathbf{D}$, the power of the projection test depends only on

$$\frac{E^2\{K(\|\boldsymbol{\varepsilon}_i\|)\|\boldsymbol{\varepsilon}_i\|^{-1}\}}{E\{K^2(\|\boldsymbol{\varepsilon}_i\|)\}} \le \frac{E\{K^2(\|\boldsymbol{\varepsilon}_i\|)\}E(\|\boldsymbol{\varepsilon}_i\|^{-2})}{E\{K^2(\|\boldsymbol{\varepsilon}_i\|)\}} = E(\|\boldsymbol{\varepsilon}_i\|^{-2}),$$

where the inequality is due to the Cauchy inequality. The maximum value is $E(\|\boldsymbol{\varepsilon}_i\|^{-2})$ with maximizer $K(t) = t^{-1}$. For fixed $k > 0$, the probability $\Pr(\chi^2_k(\gamma) > c)$, for any fixed c, is increasing function with respect to the non-centrality parameter γ ([10]). Thus, $K(t) = t^{-1}$ is the optimal weight function so that the test is locally and asymptotically most powerful.

For given $K(t) = t^{-1}$, as $E(\|\boldsymbol{\varepsilon}_i\|^{-2})$ does not depend on $\mathbf{D}$, we just need to study $E(\mathbf{V})^\top \mathbf{D}(\mathbf{D}^\top\mathbf{D})^{-1}\mathbf{D}^\top \cdot E(\mathbf{V})$ to get optimal projection direction. It follows that $\mathbf{D} = E(\mathbf{V})$ is the best choice to reach asymptotic best power by using Theorem 2 in [11]. The remaining proof is exactly the same as that of Theorem 1 in [19] and is omitted here. □

Proof of Theorem 2.2 The proof of this theorem is very similar to that of Theorem 2.1, and thus we only highlight the main difference. We need an addition lemma.

Lemma 5.1 ([30]) *Let $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{ip})^\top$ be such that Z_{ij} are independent subGaussian random variables and have mean* 0 *and variance* 1. *Then*

$$\Pr\left\{\frac{\mathrm{tr}(\boldsymbol{\Sigma})}{2} \le \|\boldsymbol{\Gamma}\mathbf{Z}_i\|^2 \le \frac{3\mathrm{tr}(\boldsymbol{\Sigma})}{2}\right\} \ge 1 - c\exp\left[-c'\left\{\frac{\mathrm{tr}^2(\boldsymbol{\Sigma})}{\mathrm{tr}(\boldsymbol{\Sigma}^2)}\right\}^{1/6}\right],$$

where c' and c are positive finite constants.

Define $\mathbf{B} = E\{K^2(\|\boldsymbol{\varepsilon}_i\|)\mathbf{u}_i\mathbf{u}_i^\top\}$. Write

$$\begin{aligned}\sqrt{n}(\mathbf{D}^\top\mathbf{B}\mathbf{D})^{-1/2}\mathbf{D}^\top\{\bar{\mathbf{V}} - E(\mathbf{V})\} &= \sqrt{n}(\mathbf{D}^\top\mathbf{B}\mathbf{D})^{-1/2}\mathbf{D}^\top n^{-1}\sum_{i=1}^n K(r_i)\mathbf{u}_i \\ &\quad + \sqrt{n}(\mathbf{D}^\top\mathbf{B}\mathbf{D})^{-1/2}\mathbf{D}^\top n^{-1}\sum_{i=1}^n\{\mathbf{V}_i - K(r_i)\mathbf{u}_i - E(\mathbf{V}_i)\}.\end{aligned}$$

Let

$$\mathbf{I}_1 = \sqrt{n}(\mathbf{D}^\top\mathbf{B}\mathbf{D})^{-1/2}\mathbf{D}^\top n^{-1}\sum_{i=1}^n K(r_i)\mathbf{u}_i$$

and

$$\mathbf{I}_2 = \sqrt{n}(\mathbf{D}^\top\mathbf{B}\mathbf{D})^{-1/2}\mathbf{D}^\top n^{-1}\sum_{i=1}^n\{\mathbf{V}_i - K(r_i)\mathbf{u}_i - E(\mathbf{V}_i)\}.$$

Denote $\mathbf{I}_2 = \sqrt{n}(\mathbf{D}^\top\mathbf{B}\mathbf{D})^{-1/2}\mathbf{D}^\top\mathbf{R}_0$, where $\mathbf{R}_0$ is defined in the proof of Theorem 2.1. Note that

$$\begin{aligned}E(\|\mathbf{I}_2\|^2) &\le n\|\mathbf{B}^{-1}\|_2 E(\|\mathbf{R}_0\|^2) \\ &= n\|\mathbf{B}^{-1}\|_2 o(\sigma_n^2) \\ &= p^{-1}\|\mathbf{B}^{-1}\|_2 E\{K^2(\|\boldsymbol{\varepsilon}_i\|)\}o(1).\end{aligned}$$

Thus, $\mathbf{I}_2 = o_p(\sigma_n)$.

By Lemma 5.1, we have with probability tending to one, $p/2 \le \|\boldsymbol{\varepsilon}_i\|^2 \le 3p/2$. Further, by the continuity assumption of $K(\cdot)$, we have $K^2(\|\boldsymbol{\varepsilon}_i\|) \ge K^2(\eta)$, where $\eta \in [p/2, 3p/2]$. Accordingly, we have $\lambda_{\min}(\mathbf{B})$ is lower bounded by $2K^2(\eta)/(3p)$ with probability tending to one. Thus, the noncentral parameter of the chi-square distribution is

$$\begin{aligned}\delta_{\mathbf{D}} &= nE^2\{K(\|\boldsymbol{\varepsilon}_i\|)\|\boldsymbol{\varepsilon}_i\|^{-1}\}\boldsymbol{\mu}^\top\boldsymbol{\Gamma}^\top\mathbf{D}(\mathbf{D}^\top\mathbf{B}\mathbf{D})^{-1}\mathbf{D}^\top\boldsymbol{\Gamma}\boldsymbol{\mu}\{1 + o(1)\}, \\ &\le 3np/\{2K^2(\eta)\}E^2\{K(\|\boldsymbol{\varepsilon}_i\|)\|\boldsymbol{\varepsilon}_i\|^{-1}\}\boldsymbol{\mu}^\top\boldsymbol{\Gamma}^\top\mathbf{D}(\mathbf{D}^\top\mathbf{D})^{-1}\mathbf{D}^\top\boldsymbol{\Gamma}\boldsymbol{\mu}\{1 + o(1)\},\end{aligned}$$

from which we can establish the optimal $\mathbf{D}$ is $\boldsymbol{\Gamma}\boldsymbol{\mu}$ as in the proof of Theorem 2.1. Using Lemma 5.1 again, $\mathrm{tr}(\boldsymbol{\Sigma}^{-1})/2 \le \|\boldsymbol{\Gamma}\boldsymbol{\varepsilon}_i\|^2 \le 3\mathrm{tr}(\boldsymbol{\Sigma}^{-1})/2$ and consequently

$$\mathrm{tr}\left[\boldsymbol{\mu}^\top\boldsymbol{\Gamma}^\top E\{K^2(\|\boldsymbol{\varepsilon}_i\|)\mathbf{u}_i\mathbf{u}_i^\top\}\boldsymbol{\Gamma}\boldsymbol{\mu}\right] \le 3E\{K^2(\|\boldsymbol{\varepsilon}_i\|)\}\mathrm{tr}(\boldsymbol{\Sigma}^{-1})\|\boldsymbol{\mu}\|^2/p.$$

The power of the project test depends on $\delta_{\mathbf{D}}$ which is thus lower bounded by

$$\frac{E^2\{K(\|\boldsymbol{\varepsilon}_i\|)\|\boldsymbol{\varepsilon}_i\|^{-1}\}}{3E\{K^2(\|\boldsymbol{\varepsilon}_i\|)\}}np\{\mathrm{tr}(\boldsymbol{\Sigma}^{-1})\}^{-1}\|\boldsymbol{\mu}\|^2.$$

Taking $K(t) = t^{-1}$ would result in the largest power. □

Next, we state several necessary lemmas for proving Theorem 2.3.

Lemma 5.2 ([9]) *For any $p \times p$ symmetric matrix $\mathbf{M}$, we have*

$$E(\mathbf{u}_i^\top\mathbf{M}\mathbf{u}_i)^2 = \{\mathrm{tr}^2(\mathbf{M}) + 2\mathrm{tr}(\mathbf{M}^2)\}/(p^2+2p),$$
$$E(\mathbf{u}_i^\top\mathbf{M}\mathbf{u}_i)^4 = \{3\mathrm{tr}^2(\mathbf{M}^2) + 6\mathrm{tr}(\mathbf{M}^4)\}/\{p(p+2)(p+4)(p+6)\}.$$

Lemma 5.3 ([2]) *Under Assumptions* 3 *and* 5*, we have* $\|\widehat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1}\|_2 = O_p(s_{p,\boldsymbol{\Sigma}}\sqrt{\log p/n})$, *where* $\|\mathbf{A}\|_2$ *denotes the spectral norm of a matrix* $\mathbf{A}$.

The next lemma is a parallel result of Lemma 5.3, establishing the consistency of the thresholded estimator of $\boldsymbol{\mu}$.

Lemma 5.4 *Suppose Assumptions* 4 *and* 5 *hold. If we take* $h_n = M'\sqrt{\log p/n}$ *for large* M', *then*

$$\|T_{h_n}(\bar{\mathbf{X}}_1) - \boldsymbol{\mu}\| = O_p\{s_{p,\boldsymbol{\mu}}(\log p/n)^{(1-q)/2}\}.$$

Proof By Assumption 5 and limit theorems for large deviations in [25], we have

$$\max_j |\bar{X}_j - \mu_j| = O_p(\sqrt{\log p/n}).$$

Write $\|T_{h_n}(\bar{\mathbf{X}}) - \boldsymbol{\mu}\| \le \|T_{h_n}(\boldsymbol{\mu}) - \boldsymbol{\mu}\| + \|T_{h_n}(\bar{\mathbf{X}}) - T_{h_n}(\boldsymbol{\mu})\|$. On one hand,

$$\|T_{h_n}(\boldsymbol{\mu}) - \boldsymbol{\mu}\| \le \sum_{j=1}^p |\mu_j| I(|\mu_j| \le t) = \sum_{j=1}^p |\mu_j|^{1-q}|\mu_j|^q I(|\mu_j| \le t) \le s_p t^{1-q}.$$

On the other hand, we observe that

$$\|T_{h_n}(\bar{\mathbf{X}}) - T_{h_n}(\boldsymbol{\mu})\| \le L_1 + L_2 + L_3,$$

where

$$L_1 = \sum_{j=1}^p |\bar{X}_j| I(|\bar{X}_j| \ge t, |\mu_j| < t),$$
$$L_2 = \sum_{j=1}^p |\mu_j| I(|\bar{X}_j| < t, |\mu_j| \ge t),$$
$$L_3 = \sum_{j=1}^p |\bar{X}_j - \mu_j| I(|\bar{X}_j| \ge t, |\mu_j| \ge t).$$

By Assumption 4, it is easy to see that

$$L_3 \le \max_j |\bar{X}_j - \mu_j| \sum_{j=1}^p |\mu_j|^{-q}|\mu_j|^q |I(|\mu_j| \ge t) = s_p t^{-q} O_p(\sqrt{\log p/n}).$$

Denote

$$L_4 = \sum_{j=1}^{p} |\bar{X}_j - \mu_j| I(|\bar{X}_j| \geq t, |\mu_j| < t),$$

$$L_5 = \sum_{j=1}^{p} |\mu_j| I(|\mu_j| < t) \leq t^{1-q} c_0(p).$$

Thus $L_1 \leq L_4 + L_5$. Take $0 < \gamma < 1$, Then

$$\begin{aligned} L_4 &\leq \sum_{j=1}^{p} |\bar{X}_j - \mu_j| I(|\bar{X}_j| \geq t, |\mu_j| \leq \gamma t) + \sum_{j=1}^{p} |\bar{X}_j - \mu_j| I(|\bar{X}_j| \geq t, \gamma t < |\mu_j| < t), \\ &= \left\{ \sum_{j=1}^{p} I(|\bar{X}_j - \mu_j| > (1-\gamma)t) \right\} O_p\left(\sqrt{\frac{\log p}{n}}\right) + s_p (\gamma t)^{-q} O_p\left(\sqrt{\frac{\log p}{n}}\right). \end{aligned}$$

Since

$$\begin{aligned} &\Pr\left(\left\{ \sum_{j=1}^{p} I(|\bar{X}_j - \mu_j| > (1-\gamma)t) \right\} > 0\right) = \Pr(\max_j |\bar{X}_j - \mu_j| > (1-\gamma)t) \\ &\leq p \exp\{-n\delta(1-\gamma)^2 t^2\}, \end{aligned}$$

for some $\delta > 0$. Thus, $L_4 = O_p\{s_p t^{-q}(\sqrt{\log p/n})\}$. Accordingly,

$$L_1 = O_p\{s_p t^{-q}(\sqrt{\log p/n}) + s_p t^{1-q}\}.$$

Similarly,

$$\begin{aligned} L_2 &\leq \sum_{j=1}^{p} |\bar{X}_j - \mu_j| I(|\bar{X}_j| \leq t, |\mu_j| > t) + \sum_{j=1}^{p} |\bar{X}_j| I(|\bar{X}_j| \leq t, |\mu_j| > t), \\ &\leq \max_j |\bar{X}_j - \mu_j| \sum_{j=1}^{p} I(|\mu_j| > t) + t \sum_{j=1}^{p} I(|\mu_j| > t), \\ &= O_p\left\{ s_p t^{-q} \sqrt{\frac{\log p}{n}} + s_p t^{1-q} \right\}. \end{aligned}$$

If we take $t = M' \sqrt{\log p/n}$, the lemma follows from the above three terms L_1, L_2 and L_3. □

In the following derivation, we will simply take $q = 0$ without loss of generality.

Lemma 5.5 *For any p-dimensional vectors* $\mathbf{X}$ *and* $\boldsymbol{\mu}$*, we have*

$$\left\| \frac{\mathbf{X} + \boldsymbol{\mu}}{\|\mathbf{X} + \boldsymbol{\mu}\|^2} - \frac{\mathbf{X}}{\|\mathbf{X}\|^2} - \|\mathbf{X}\|^{-2} \left(\mathbf{I}_p - 2 \frac{\mathbf{X}\mathbf{X}^\top}{\|\mathbf{X}\|^2} \right) \boldsymbol{\mu} \right\| \leq C \frac{\|\boldsymbol{\mu}\|^2}{\|\mathbf{X}\|^3}.$$

Proof This lemma can be easily shown by using the Taylor theorem and thus omitted here. See [22, Lemma 6.2] for a similar result. □

Lemma 5.6 *Suppose Assumptions* 3–5 *hold. We have*

$$\|\widehat{\mathbf{d}} - \mathbf{d}\| = O_p\{s_{p,\boldsymbol{\Sigma}} s_{p,\boldsymbol{\mu}} \sqrt{\log p/n}\}.$$

Proof By the proof of Theorem 2.1, we have

$$\mathbf{d} \propto E\mathbf{V} = E\{\|\boldsymbol{\varepsilon}_i\|^{-1} \mathbf{u}_i + \|\boldsymbol{\varepsilon}_i\|^{-2} (\mathbf{I}_p - \mathbf{u}_i \mathbf{u}_i^\top) \boldsymbol{\Gamma} \boldsymbol{\mu} + \|\boldsymbol{\varepsilon}_i\|^{-1} \mathbf{Q}_i\},$$

$$= \frac{p-1}{p} E\{\|\boldsymbol{\varepsilon}_i\|^{-2}\} \left\{ \boldsymbol{\Gamma}\boldsymbol{\mu} + \frac{p}{p-1} E(\|\boldsymbol{\varepsilon}_i\|^{-1}\mathbf{Q}_i)/E(\|\boldsymbol{\varepsilon}_i\|^{-2}) \right\},$$

where $\mathbf{Q}_i$ is defined in (A.1).

As $\mathbf{d}$ is a projection direction, we write

$$\mathbf{d} = \boldsymbol{\Gamma}\boldsymbol{\mu} + \frac{p}{p-1} E(\|\boldsymbol{\varepsilon}_i\|^{-1}\mathbf{Q}_i)/E(\|\boldsymbol{\varepsilon}_i\|^{-2}).$$

Then, we have

$$\begin{aligned}
\|\widehat{\mathbf{d}} - \mathbf{d}\| &\le \|\widehat{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\mu}} - \boldsymbol{\Gamma}\boldsymbol{\mu}\| + \frac{p}{p-1} E(\|\boldsymbol{\varepsilon}_i\|^{-1}\|\mathbf{Q}_i\|)/E(\|\boldsymbol{\varepsilon}_i\|^{-2}),\\
&\le \|(\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma})\boldsymbol{\mu} + \boldsymbol{\Gamma}(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}) + (\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma})(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu})\| + CE(\|\boldsymbol{\varepsilon}_i\|^{-3})\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}/E(\|\boldsymbol{\varepsilon}_i\|^{-2}),\\
&\le \{\|(\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma})\boldsymbol{\mu}\| + \|\boldsymbol{\Gamma}(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu})\| + \|(\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma})(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu})\|\} + \{E(\|\boldsymbol{\varepsilon}_i\|^{-1})\}^{-1}O\{(np)^{-1}\},\\
&= O_p\left(s_{p,\boldsymbol{\Sigma}}s_{p,\boldsymbol{\mu}}\sqrt{\frac{\log p}{n}}\right) + O_p\left(\frac{1}{n\sqrt{p}}\right),\\
&= O_p\left(s_{p,\boldsymbol{\Sigma}}s_{p,\boldsymbol{\mu}}\sqrt{\frac{\log p}{n}}\right). \qquad \square
\end{aligned}$$

Proof of Theorem 2.3 Firstly, from the proof of Theorem 2.1, we know that

$$\mathbf{d} = \boldsymbol{\Gamma}\boldsymbol{\mu}(1 + o(1)). \tag{A.4}$$

Denote $\mathbf{d}_s = \mathbf{d}/\|\mathbf{d}\|$ and $\widehat{\mathbf{d}}_s = \widehat{\mathbf{d}}/\|\widehat{\mathbf{d}}\|$. The proof of this theorem can be divided into three key steps. In the first step, we show that

$$\sigma_{n_2}^{-1}\mathbf{d}_s^\top\widehat{\mathbf{V}} - \sqrt{n_2 p E\{\|\boldsymbol{\varepsilon}_i\|^{-2}\}}(\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})^{1/2} \xrightarrow{d} N(0,1). \tag{A.5}$$

Lemma 5.5 leads to

$$\begin{aligned}
\frac{U(\widehat{\boldsymbol{\Gamma}}\mathbf{X}_i)}{\|\widehat{\boldsymbol{\Gamma}}\mathbf{X}_i\|} &= \|\boldsymbol{\varepsilon}_i\|^{-1}\mathbf{u}_i + \|\boldsymbol{\varepsilon}_i\|^{-1}(\mathbf{I}_p - \mathbf{u}_i\mathbf{u}_i^\top)(\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma})\boldsymbol{\Gamma}^{-1}\mathbf{u}_i + \|\boldsymbol{\varepsilon}_i\|^{-2}(\mathbf{I}_p - \mathbf{u}_i\mathbf{u}_i^\top)(\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma})\boldsymbol{\mu}\\
&\quad + \|\boldsymbol{\varepsilon}_i\|^{-2}(\mathbf{I}_p - \mathbf{u}_i\mathbf{u}_i^\top)\boldsymbol{\Gamma}\boldsymbol{\mu} + \boldsymbol{\upsilon}_i,
\end{aligned}$$

where the remainder term $\boldsymbol{\upsilon}_i$ satisfies

$$\|\boldsymbol{\upsilon}_i\| \le C\|\boldsymbol{\varepsilon}_i\|^{-3}\|(\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma})(\mathbf{X}_i - \boldsymbol{\mu}) + (\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma})\boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\mu}\|^2.$$

Therefore, we have

$$\begin{aligned}
\mathbf{d}_s^\top\widehat{\mathbf{V}} &= \frac{1}{n_2}\sum_{i=n_1+1}^{n} \|\boldsymbol{\varepsilon}_i\|^{-1}\mathbf{d}_s^\top\mathbf{u}_i + \frac{1}{n_2}\sum_{i=n_1+1}^{n} \|\boldsymbol{\varepsilon}_i\|^{-2}\mathbf{d}_s^\top(\mathbf{I}_p - \mathbf{u}_i\mathbf{u}_i^\top)\boldsymbol{\Gamma}\boldsymbol{\mu}\\
&\quad + \frac{1}{n_2}\sum_{i=n_1+1}^{n} \|\boldsymbol{\varepsilon}_i\|^{-1}\mathbf{d}_s^\top(\mathbf{I}_p - \mathbf{u}_i\mathbf{u}_i^\top)(\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma})\boldsymbol{\Gamma}^{-1}\mathbf{u}_i\\
&\quad + \frac{1}{n_2}\sum_{i=n_1+1}^{n} \|\boldsymbol{\varepsilon}_i\|^{-2}\mathbf{d}_s^\top(\mathbf{I}_p - \mathbf{u}_i\mathbf{u}_i^\top)\mathbf{d}_s^\top(\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma})\boldsymbol{\mu}\\
&\quad + \frac{1}{n_2}\sum_{i=n_1+1}^{n} \mathbf{d}_s^\top\boldsymbol{\upsilon}_i\\
&=: M_1 + M_2 + R_1 + R_2 + R_3.
\end{aligned}$$

Note that $\sigma_{n_2}^{-1}M_1 \xrightarrow{d} N(0,1)$. Next, we show that $R_k = o_p(\sigma_{n_2})$ for $k = 1,2,3$. Write $R_1 = \vartheta_1 + \vartheta_2$, where $\vartheta_1 = n_2^{-1}\sum_{i=n_1+1}^{n}\|\boldsymbol{\varepsilon}_i\|^{-1}\mathbf{d}_s^\top(\boldsymbol{\Gamma}-\widehat{\boldsymbol{\Gamma}})\boldsymbol{\Gamma}^{-1}\mathbf{u}_i$ and $\vartheta_2 = -n_2^{-1}\sum_{i=n_1+1}^{n}\|\boldsymbol{\varepsilon}_i\|^{-1}\mathbf{d}_s^\top \mathbf{u}_i\mathbf{u}_i^\top(\boldsymbol{\Gamma}-\widehat{\boldsymbol{\Gamma}})\boldsymbol{\Gamma}^{-1}\mathbf{u}_i$. Denote $\mathbf{H}_1 = \boldsymbol{\Gamma}^{-1}(\boldsymbol{\Gamma}-\widehat{\boldsymbol{\Gamma}})\mathbf{d}_s\mathbf{d}_s^\top(\boldsymbol{\Gamma}-\widehat{\boldsymbol{\Gamma}})\boldsymbol{\Gamma}^{-1}$. By Cauchy inequality and (A.4),

$$\begin{aligned}
E(\vartheta_1^2) &= n_2^{-1}E\{\|\boldsymbol{\varepsilon}_i\|^{-1}(\mathbf{u}_i^\top\boldsymbol{\Gamma}^{-1}(\boldsymbol{\Gamma}-\widehat{\boldsymbol{\Gamma}})\mathbf{d}_s\mathbf{d}_s^\top(\boldsymbol{\Gamma}-\widehat{\boldsymbol{\Gamma}})\boldsymbol{\Gamma}^{-1}\mathbf{u}_i)\},\\
&\le n_2^{-1}\{E(\|\boldsymbol{\varepsilon}_i\|^{-4})\}^{1/2}\{E(\mathbf{u}_i^\top\mathbf{H_1}\mathbf{u}_i)^2\}^{1/2},\\
&= n_2^{-1}\{E(\|\boldsymbol{\varepsilon}_i\|^{-4})\}^{1/2}O(\sqrt{\operatorname{tr}\{E(\mathbf{H}_1^2)\}}p^{-1}),\\
&= \{E(\|\boldsymbol{\varepsilon}_i\|^{-4})\}^{1/2}O\left(\frac{1}{n_2p}\frac{s_{p,\boldsymbol{\Sigma}}^2\log p}{n_2}\right) = o(\sigma_{n_2}^2),
\end{aligned}$$

where the second equality is due to Lemma 5.2 and the third equality comes from Lemma 5.3. Also,

$$\begin{aligned}
E(\vartheta_2^2) &\le \{E(\|\boldsymbol{\varepsilon}_i\|^{-8})\}^{1/4}\{E(\mathbf{u}_i^\top\mathbf{d}_s\mathbf{d}_s^\top\mathbf{u}_i)^4\}^{1/4}[E\{\mathbf{u}_i^\top\boldsymbol{\Gamma}^{-1}(\boldsymbol{\Gamma}-\widehat{\boldsymbol{\Gamma}})\mathbf{u}_i\}^4]^{1/2},\\
&= \{E(\|\boldsymbol{\varepsilon}_i\|^{-8})\}^{1/4}O\left(\frac{1}{n_2p}\frac{\operatorname{tr}^{1/2}(\boldsymbol{\Sigma}^2)s_{p,\boldsymbol{\Sigma}}^2\log p}{p^2}\right) = o(\sigma_{n_2}^2).
\end{aligned}$$

By taking a similar procedure, we can verify that $R_2 = o_p(\sigma_{n_2})$. The proof of $R_3 = o_p(\sigma_{n_2})$ is also similar but more tedious, and thus the technical details are provided in the Supplementary Material. Using (A.4) again, we have

$$E(\sigma_{n_2}^{-1}M_2) = \sqrt{n_2pE\{\|\boldsymbol{\varepsilon}_i\|^{-2}\}}(\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})^{1/2}(1+o(1)).$$

Similar to R_1, it can be verified that $\operatorname{var}(M_2) = o(\sigma_{n_2}^2)$, from which we establish (A.5).

The second step is to show that

$$(\widehat{\mathbf{d}}_s - \mathbf{d}_s)^\top\widehat{\mathbf{V}} = o_p(\sigma_{n_2}). \tag{A.6}$$

By (A.1),

$$\begin{aligned}
\|\widehat{\mathbf{d}}_s - \mathbf{d}_s\| &\le 2\|\widehat{\mathbf{d}}-\mathbf{d}\|/\|\mathbf{d}\|\\
&= O_p(r^{-1}s_{p,\boldsymbol{\Sigma}}s_{p,\boldsymbol{\mu}}\sqrt{\log p/n_2})\\
&= O_p(\max\{s_{p,\boldsymbol{\Sigma}}, s_{p,\boldsymbol{\mu}}\}\sqrt{\log p/n_2}).
\end{aligned}$$

Thus, by using similar arguments for showing $R_1 = o_p(\sigma_{n_2})$, we can establish $(\widehat{\mathbf{d}}_s - \mathbf{d}_s)^\top\widehat{\mathbf{V}} = o_p(\sigma_{n_2})$.

The third step is to prove

$$n_2^{-1}\widehat{\sigma}^2/(\sigma_{n_2}^2\widehat{\mathbf{d}}^\top\widehat{\mathbf{d}}) \xrightarrow{p} 1. \tag{A.7}$$

Notice that given $\widehat{\mathbf{d}}$, $\widehat{\mathbf{d}}^\top\widehat{\mathbf{V}}_{n_1+1},\ldots,\widehat{\mathbf{d}}^\top\widehat{\mathbf{V}}_n$ are i.i.d. random variables. Thus, $\widehat{\sigma}^2/\operatorname{var}(\widehat{\mathbf{d}}^\top\widehat{\mathbf{V}}_i) \xrightarrow{p} 1$ by the WLLN. By taking the same procedure as in the first step, we can see that

$$(n_2\widehat{\mathbf{d}}^\top\widehat{\mathbf{d}})^{-1/2}\widehat{\mathbf{d}}^\top\widehat{\mathbf{V}}_i = n_2^{-1/2}\|\boldsymbol{\varepsilon}_i\|^{-1}(\widehat{\mathbf{d}}^\top\widehat{\mathbf{d}})^{-1/2}\widehat{\mathbf{d}}^\top\mathbf{u}_i + o_p(\sigma_{n_2}).$$

Accordingly, $\frac{\operatorname{var}(\widehat{\mathbf{d}}^\top\widehat{\mathbf{V}}_i)}{n_2\sigma_{n_2}^2\widehat{\mathbf{d}}^\top\widehat{\mathbf{d}}} \to 1$ from which we can conclude that (A.7) holds.

Combining (A.5), (A.6) and (A.7), Theorem 2.3 follows immediately. □

Proof of $R_3 = o_p(\sigma_n)$ Recall that we aim to see that

$$R_3 = \frac{1}{n}\sum_{i=1}^{n} \mathbf{d}_s^\top \boldsymbol{v}_i = o_p(\sigma_n),$$

where the term $\boldsymbol{v}_i$ satisfies

$$\|\boldsymbol{v}_i\| \le C\|\boldsymbol{\varepsilon}_i\|^{-3}\|(\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma})(\mathbf{X}_i - \boldsymbol{\mu}) + (\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma})\boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\mu}\|^2.$$

It follows from that

$$E\bigg(\frac{1}{n}\sum_{i=1}^{n} \mathbf{d}_s^\top \boldsymbol{v}_i\bigg)^2 = O\{E(\boldsymbol{v}_i^\top \mathbf{d}_s^\top \mathbf{d}_s \boldsymbol{v}_i)\},$$

where

$$\begin{aligned}
E\{\boldsymbol{v}_i^\top \mathbf{d}_s^\top \mathbf{d}_s \boldsymbol{v}_i\} &\le E(\boldsymbol{v}_i^\top \boldsymbol{v}_i),\\
&\le E\{\|\boldsymbol{\varepsilon}_i\|^{-6}\|(\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma})(\mathbf{X}_i - \boldsymbol{\mu}) + (\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma})\boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\mu}\|^4\},\\
&= O(E\{\|\boldsymbol{\varepsilon}_i\|^{-6}(\|(\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma})(\mathbf{X}_i - \boldsymbol{\mu})\|^4 + \|(\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma})\boldsymbol{\mu}\|^4 + \|\boldsymbol{\Gamma}\boldsymbol{\mu}\|^4)\}),\\
&= O\bigg(\sum_{i=1}^{3}\Delta_i\bigg).
\end{aligned}$$

Note that

$$\begin{aligned}
\Delta_1 &= E\{\|\boldsymbol{\varepsilon}_i\|^{-2}(\mathbf{u}_i^\top \boldsymbol{\Gamma}^{-1}(\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma})^2\boldsymbol{\Gamma}^{-1}\mathbf{u}_i)^2\},\\
&\le [E\{\|\boldsymbol{\varepsilon}_i\|^{-4}\}]^{1/2}[E\{\mathbf{u}_i^\top \boldsymbol{\Gamma}^{-1}(\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma})^2\boldsymbol{\Gamma}^{-1}\mathbf{u}_i\}^4]^{1/2},\\
&= O\bigg(E\{\|\boldsymbol{\varepsilon}_i\|^{-2}\}\bigg\{\frac{1}{p^4}\mathrm{tr}(E(\mathbf{H}^4))\bigg\}^{1/2}\bigg),\\
&= O\bigg(E\{\|\boldsymbol{\varepsilon}_i\|^{-2}\}\frac{1}{np}\frac{\log^2 p}{np}\mathrm{tr}^{1/2}(\boldsymbol{\Sigma}^4)\bigg),\\
&= o(\sigma_n^2),
\end{aligned}$$

where $\mathbf{H} = \boldsymbol{\Gamma}^{-1}(\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma})^2\boldsymbol{\Gamma}^{-1}$. Similarly,

$$\begin{aligned}
\Delta_2 &= E\{\|\boldsymbol{\varepsilon}_i\|^{-6}(\boldsymbol{\mu}^\top \boldsymbol{\Gamma}\mathbf{H}\boldsymbol{\Gamma}\boldsymbol{\mu})^2\},\\
&= E\{\|\boldsymbol{\varepsilon}_i\|^{-6}\}[\boldsymbol{\mu}^\top \boldsymbol{\Gamma}\{E(\mathbf{H})\}\boldsymbol{\Gamma}\boldsymbol{\mu}]^{\mathbf{2}},\\
&= O\{E(\|\boldsymbol{\varepsilon}_i\|^{-6})(\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})^2\},\\
&= O\{E(\|\boldsymbol{\varepsilon}_i\|^{-2})\{E(\|\boldsymbol{\varepsilon}_i\|^{-2})(\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})\}^2\},\\
&= O\{E(\|\boldsymbol{\varepsilon}_i\|^{-2})(np)^{-2}\},\\
&= o(\sigma_n^2).
\end{aligned}$$

The arguments for showing $\Delta_3 = o(\sigma_n^2)$ is similar. Hence, we can establish $R_3 = o_p(\sigma_n)$. □

Proof of Theorem 3.1 By the Taylor approximation of $U(\boldsymbol{\Gamma}(\mathbf{X}_i - \boldsymbol{\mu}))$ and $U(\boldsymbol{\Gamma}(\mathbf{Y}_i - \boldsymbol{\mu}))$ and the assumption on $\boldsymbol{\mu}$, we get

$$U(\boldsymbol{\Gamma}(\mathbf{X}_i - \boldsymbol{\mu})) = U(\boldsymbol{\varepsilon}_{1i}) + \|\boldsymbol{\varepsilon}_{1i}\|^{-1}(\mathbf{I}_p - U(\boldsymbol{\varepsilon}_{1i})U^\top(\boldsymbol{\varepsilon}_{1i}))\boldsymbol{\Gamma}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}) + \boldsymbol{Q}_{1i},$$

$$U(\boldsymbol{\Gamma}(\mathbf{Y}_i - \boldsymbol{\mu})) = U(\boldsymbol{\varepsilon}_{2i}) - \|\boldsymbol{\varepsilon}_{2i}\|^{-1}(\mathbf{I}_p - U(\boldsymbol{\varepsilon}_{2i})U^\top(\boldsymbol{\varepsilon}_{2i}))\boldsymbol{\Gamma}(\boldsymbol{\mu} - \boldsymbol{\mu}_2) + \boldsymbol{Q}_{2i},$$

where $\boldsymbol{Q}_{1i}$ and $\boldsymbol{Q}_{2i}$ are the Taylor remainder terms satisfying

$$\|\boldsymbol{Q}_{1i}\| \le C\|\boldsymbol{\varepsilon}_{1i}\|^{-2}\|\boldsymbol{\Gamma}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\|^2,$$
$$\|\boldsymbol{Q}_{2i}\| \le C\|\boldsymbol{\varepsilon}_{2i}\|^{-2}\|\boldsymbol{\Gamma}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\|^2,$$

and C is a constant that does not depend on $\boldsymbol{\Gamma}\mathbf{X}_i$ and $\boldsymbol{\Gamma}\mathbf{Y}_i$ or $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$.

Using similar arguments in the proof of Theorem 2.1, we can show that $Q_{\mathbf{A}} \xrightarrow{d} \chi^2_k(\delta_{\mathbf{A}})$, where

$$\delta_{\mathbf{A}} = \frac{nmp}{n+m} E^{-1}\{K^2(\|\boldsymbol{\varepsilon}_{1i}\|)\} E^2\{K(\|\boldsymbol{\varepsilon}_{1i}\|)\|\boldsymbol{\varepsilon}_{1i}\|^{-1}\}$$
$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Gamma}^\top \mathbf{A}(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top \boldsymbol{\Gamma}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\{1 + o(1)\}.$$

By using Cauchy inequality again, the assertion follows immediately. □

Acknowledgements The authors thank the Editor, Associate Editor and anonymous referees for their many helpful comments that have resulted in significant improvements in the article.

References

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

[2] Bickel, P. J., Levina, E.: Covariance regularization by thresholding. *Ann. Statist.*, **36**, 2577–2604 (2008)

[3] Cai, T. T., Yuan, M.: Adaptive covariance matrix estimation through block thresholding. *Ann. Statist.*, **40**, 2014–2042 (2012)

[4] Chen, L. S., Paul, D., Prentice, R. L., et al.: A regularized Hotelling's T^2 test for pathway analysis in proteomic studies. *J. Amer. Statist. Assoc.*, **106**, 1345–1360 (2011)

[5] Chen, S., Qin, Y.: A two-sample test for high-dimensional data with applications to gene-set testing. *Ann. Statist.*, **38**, 808–835 (2010)

[6] Cui, H., Guo, W., Zhong, W., et al.: Test for high-dimensional regression coefficients using refitted cross-validation variance estimation. *Ann. Statist.*, **46**, 958–988 (2018)

[7] Fan, J., Liao, Y., Yao, J.: Power enhancement in high-dimensional cross-sectional tests. *Econometrica*, **83**, 1497–1541 (2015)

[8] Fang, K., Kotz, S., Ng., K.: Symmetric Multivariate and Related Distributions, Chapman & Hall, London, 1990

[9] Feng, L., Zou, C., Wang, Z.: Multivariate-sign-based high-dimensional tests for the two-sample location problem. *J. Amer. Statist. Assoc.*, **111**, 721–735 (2016)

[10] Ghosh, B. K.: Sequential tests of statistical hypotheses, Reading: Addison-Wesley, 1970

[11] Ghosh, B. K.: Some monotonicity theorems for χ^2, F and t distributions with applications. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **35**, 480–492 (1973)

[12] Gregory, K. B., Carroll, R. J., Baladandayuthapani, V., et al.: A two-sample test for equality of means in high dimension. *J. Amer. Statist. Assoc.*, **110**, 837–849 (2015)

[13] Hallin, M., Paindaveine, D.: Optimal tests for multivariate location based on interdirections and pseudo-Mahalanobis ranks. *Ann. Statist.*, **30**, 1103–1133 (2002)

[14] Hallin, M., Paindaveine, D.: Semiparametrically efficient rank-based inference for shape. I. Optimal rank-based tests for sphericity. *The Ann. Statist.*, **34**, 2707–2756 (2006)

[15] Hu, Z., Tong, T., Genton, M. G.: Diagonal likelihood ratio test for equality of mean vectors in high-dimensional data. *Biometrics*, **75**, 256–267 (2019)

[16] Kim, I., Balakrishnan, S., Wasserman, L., et al.: Robust multivariate nonparametric tests via projection averaging. *Ann. Statist.*, **48**, 3417–3441 (2020)

[17] Lauter, J.: Exact t and F tests for analyzing studies with multiple endpoints. *Biometrics*, **52**(3), 964–970 (1996)

[18] Ledoit, O., Péché, S.: Eigenvectors of some large sample covariance matrix ensembles. *Probab. Theory Related Fields*, **151**, 233–264 (2011)

[19] Li, R., Huang, Y., Wang, L., et al.: Projection Test for High-dimensional Mean Vectors with Optimal Direction, Manuscript (2015)

[20] Lopes, M., Jacob, L., Wainwright, M. J.: A more powerful two-sample test in high dimensions using random projection. arXiv: 1108.2401 (2011)

[21] Mahfoud, Z. R., Randles, R. H.: On multivariate signed rank tests. *J. Nonparametr. Stat.*, **17**, 201–216 (2005)

[22] Oja, H.: Multivariate Nonparametric Methods with R: An Approach Based on Spatial Signs and Ranks, Springer Science & Business Media, New York, 2010

[23] Paindaveine, D., Verdebout, T.: On high-dimensional sign tests. *Bernoulli*, **22**, 1745–1769 (2016)

[24] Randles, R. H.: A simpler, affine-invariant, multivariate, distribution-free sign test. *J. Amer. Statist. Assoc.*, **95**, 1263–1268 (2000)

[25] Saulis, L., Statulevicius, V. A.: Limit theorems for large deviations, Vol. 73, Springer Science & Business Media, 1991

[26] Srivastava, M. S., Du, M.: A test for the mean vector with fewer observations than the dimension. *J. Multivariate Anal.*, **99**, 386–402 (2008)

[27] Srivastava, M. S., Katayama, S., Kano, Y.: A two sample test in high dimensional data. *J. Multivariate Anal.*, **114**, 349–358 (2013)

[28] Srivastava, R., Li, P., Ruppert, D.: RAPTT: An exact two-sample test in high dimensions using random projections. *J. Comput. Graph. Statist.*, **25**, 954–970 (2016)

[29] Tyler, D. E.: A distribution-free M-estimator of multivariate scatter. *Ann. Statist.*, **15**, 234–251 (1987)

[30] Wang, L., Peng, B., Li, R.: A high-dimensional nonparametric multivariate test for mean vector. *J. Amer. Statist. Assoc.*, **110**, 1658–1669 (2015)

[31] Wang, S., Cui, H.: Generalized F test for high dimensional linear regression coefficients. *J. Multivariate Anal.*, **117**, 134–149 (2013)

[32] Wasserman, L., Roeder, K.: High dimensional variable selection. *Ann. Statist.*, **37**, 2178–2201 (2009)

[33] Xue, K., Yao, F.: Distribution and correlation-free two-sample test of high-dimensional means. *Ann. Statist.*, **48**, 1304–1328 (2020)

[34] Zou, C., Peng, L., Feng, L., et al.: Multivariate sign-based high-dimensional tests for sphericity. *Biometrika*, **101**, 229–236 (2013)