HHS Public Access

Author manuscript

J Am Stat Assoc. Author manuscript; available in PMC 2017 January 15.

Published in final edited form as:

J Am Stat Assoc. 2015 ; 110(512): 1658–1669. doi:10.1080/01621459.2014.988215.

A High-Dimensional Nonparametric Multivariate Test for Mean Vector

Lan Wang,

Associate Professor, School of Statistics, University of Minnesota, Minneapolis, MN 55455

Bo Peng, and

Graduate student, School of Statistics, University of Minnesota, Minneapolis, MN 55455

Runze Li

Distinguished Professor, Department of Statistics and the Methodology Center, the Pennsylvania State University, University Park, PA 16802-2111

Lan Wang: wangx346@umn.edu; Runze Li: rzli@psu.edu

Abstract

This work is concerned with testing the population mean vector of nonnormal high-dimensional multivariate data. Several tests for high-dimensional mean vector, based on modifying the classical Hotelling T^2 test, have been proposed in the literature. Despite their usefulness, they tend to have unsatisfactory power performance for heavy-tailed multivariate data, which frequently arise in genomics and quantitative finance. This paper proposes a novel high-dimensional nonparametric test for the population mean vector for a general class of multivariate distributions. With the aid of new tools in modern probability theory, we proved that the limiting null distribution of the proposed test is normal under mild conditions when p is substantially larger than n. We further study the local power of the proposed test and compare its relative efficiency with a modified Hotelling T^2 test for high-dimensional data. An interesting finding is that the newly proposed test can have even more substantial power gain with large p than the traditional nonparametric multivariate test does with finite fixed p. We study the finite sample performance of the proposed test via Monte Carlo simulations. We further illustrate its application by an empirical analysis of a genomics data set.

Keywords

Asymptotic relative efficiency; High dimensional multivariate data; Hotelling T^2 test; Nonparametric multivariate test

1 Introduction

Let $X_1, \ldots, X_n$ be independent and identically distributed (iid) p-dimensional random vectors from the model $X_i = \mu + \varepsilon_i$ where ε_i is the random error to be specified later. In this paper, we consider a novel nonparametric procedure for testing the hypothesis

Runze Li is the corresponding author.

$$H_0:\mu=0 \text{ versus } H_1:\mu \neq 0, \quad (1)$$

when p is potentially much larger than n. Here and throughout this paper, p stands for the number of variables (or features) of the data, and n for the sample size.

The above testing problem is motivated by recent advances in genomics. There is growing evidence that most biological processes involve the regulation of multiple genes; and that analysis focusing on individual genes often suffer from low power to detect important genetic variation and poor reproduceability (Vo et al., 2007). As a result, increasing attention has been focused on the analysis of gene sets/pathways, which are groups of genes sharing common biological functions, chromosomal locations or regulations. In some important applications, the problem of evaluating whether a group of genes are differentially expressed can be formulated as the hypothesis in (1), where X_i represents a vector of summary statistics computed on each of the p genes, such as the log-intensity ratios of the red over green channels; or the log ratios of the gene expression levels between control and treatment chips (or before and after drug treatment). For example, the data set we analyzed in Section 3.2 contains microarray measurements from diabetic patients before and after insulin treatment (Wu et al., 2007, 2011).

Testing the hypothesis in (1) becomes very challenging for high-dimensional data. The traditional Hotelling's T^2 test is not well defined as the inverse of sample covariance matrix may not exist when p is larger than n. It has been observed in Bai and Saranadasa (1996) that the power of the Hotelling's T^2 test can be adversely affected even when $p < n$, if the sample covariance matrix is nearly singular, see also Pan and Zhou (2011). Recently, there has been great interest in extending Hotelling's test to the $p > n$ setting, see Bai and Saranadasa (1996, $p/n \to c\ \varepsilon\ (0,1)$), Srivastava and Du (2008, $n = O(p^{\delta})$ for some $1/2 < \delta \leq 1$), Srivastava (2009, $n = O(p^{\delta})$ for some $0 < \delta \leq 1$), Lee et al. (2012, $p/n \to c > 0$), Srivastava et al. (2013, $n = O(p^{\delta})$, $\delta > 1/2$), Chen and Qin (2010, $\mathrm{Tr}(\Sigma^4) = o(\mathrm{Tr}^2(\Sigma^2))$). Thulin (2014) proposed a more computing-intensive extension by combining Hotelling's tests from a large number of lower-dimensional random subspaces. A shared drawback of the aforementioned tests is that they tend to have unsatisfactory power performance when the multivariate distribution is heavy-tailed and is very sensitive to outlying observations.

In many microarray experiments, most genes are expressed at very low levels, few genes are expressed at high levels. The distribution of intensities tends to be nonnormal even after log transformation, regardless of the normalization methods (e.g., Purdom and Holmes, 2005). For the data example in Section 3.2, it is observed that the marginal distributions of the microarray expressions are nonnormal and have heavy tails based on values of their marginal kurtosises. Furthermore, in microarray experiments, outliers frequently arise due to the array chip artifacts such as uneven spray of reagents within arrays and other reasons. This motivates us to develop a nonparametric test for high-dimensional population mean vector or the location parameter without the multivariate normality assumption.

We propose a new test for hypothesis (1) based on spatial signs of the observations, and further study its asymptotic theory. Comparing with the extensions of Hotelling's T^2 test (Chen and Qin, 2010), the theory for the nonparametric test with $p > n$ is considerably more

challenging. To derive the asymptotic theory, we employ new probability tools on the concentration properties of certain quadratic forms, which may be of independent interest and have potential applications in developing the theory for other related high-dimensional nonparametric procedures. The proposed nonparametric test has several appealing properties. First it is directly applicable for the setting with $p > n$, and it is computationally simple. Second, the new test is shown to lose little efficiency when the underlying data are multivariate normal and to have potentially significant efficiency gain for heavy-tailed multivariate distributions. This is verified by deriving its asymptotic relative efficiency. From our Monte Carlo simulation, significant efficiency gain can be achieved at small or moderate sample size.

Nonparametric statistical procedures have been explored little in the high dimensional setting. An open question is whether their power advantage continues to hold (and if hold, to what extent) in high dimension. This work takes a substantial step towards understanding the merits of nonparametric procedures when $p > n$ by providing both theoretical justification and numerical evidence. Our theoretical analysis reveals a striking phenomenon: the efficiency gain of the new nonparametric test in the high-dimensional setting can be more substantial comparing with the well known traditional nonparametric tests efficiency gain in the "classical" framework where p is fixed and n goes to infinity. For example, consider the p-dimensional multivariate t-distribution with 3 degrees of freedom, which is heavy-tailed. For this distribution, it is well known that the asymptotic relative efficiency of the spatial sign test versus Hotelling's T^2 test is 1.9 for $p = 1$, 2.02 for $p = 3$, and 2.09 for $p = 10$. This implies an increasing trend as the dimension p increases. The theory established in this paper suggests that when $p > n$, the asymptotic relative efficiency of the proposed new nonparametric test versus Chen and Qin's extension of Hotelling's T^2 test is about 2.54. This result provides strong support for the usefulness of nonparametric tests in high-dimensional problems.

It is worth noting that we do not impose structural constrains, such as sparsity, on the alternative hypothesis. Hence, it allows for a dense alternative, where many components of the vector contribute to the signal. In fact, this is one of the main motivations for gene set analysis. For many complex diseases, such as depression and diabetes, evidence from medical literature suggests that many of the genes from a biological pathway contribute small signals which are hard to detect individually. Cook et al. (2012) discussed other applications of similar nature, where sparsity may not be the reality. In the simulations, we demonstrated that the test based on marginal p values with Bonferroni or FDR correction may have low power to detect the global signal. On the other hand, in some other applications involving high-dimensional testing, there may be reasons to believe the alternative is sparse, for which case the existing tests can be further tuned to increase the power performance, see the recent work by Hall and Jin (2010), Zhong, Chen and Xu (2013) and Cai, Liu and Xia (2014), among others. It is noted that these tests use sample means as basic building blocks and hence are expected to suffer from power loss for heavy-tailed multivariate data. The new test we propose has the potential to be extended to the sparse alternative setting with the promise of improved power performance.

We introduce the high-dimensional nonparametric test in Section 2.1. We derive its limiting null distribution under a set of weak conditions in Section 2.2, and investigate its power performance under local alternatives and study the asymptotic relative efficiency in Section 2.3, some important extensions are discussed in Section 2.4. We conduct Monte Carlo simulations and analyze the gene sets from a genomics study in Section 3. Section 4 concludes the paper and discusses relevant issues. Technical proofs are given in the Appendix. The Supplemental Material include additional technical and numerical results.

2 A high-dimensional nonparametric test

We first focus on the case that the random vector X_i follows an elliptical distribution. Extensions to beyond the elliptical distribution family are discussed in Section 2.4.

The class of elliptical distributions encompasses many useful non-Gaussian multivariate distributions such as multivariate *t* distribution, multivariate logistic distribution, Kotz-type multivariate distribution, Pearson II type multivariate distribution and many others. The family of elliptical distributions is well studied in the statistical literature (e.g., Fang, Kotz and Ng, 1990). Recently, this family becomes important for modeling finance data (McNeil, Frey and Embrechts, 2005) due to its potential to accommodate tail dependence (the phenomenon of simultaneous extremes), which is important in quantitative finance but is not allowed by the multivariate normal distribution (Schmidt, 2002).

An elliptically distributed random vector X_i has the following convenient stochastic representation:

$$X_i=\mu+\varepsilon_i, \text{ and } \varepsilon_i=\Gamma R_i U_i, \quad (2)$$

where Γ is a $p \times p$ matrix, U_i is a random vector uniformly distributed on the unit sphere in $\mathbb{R}^p$, and R_i is a nonnegative random variable independent of U_i. The distribution of X_i depends on Γ only through $\Gamma\Gamma^T$ (Fang, Kotz and Ng, 1989). Thus, we denote $\Omega = \Gamma\Gamma^T$ for easy future reference. An important special case of (2) is the multivariate normal distribution with mean μ and covariance matrix Σ, for which R_i^2 has a chi-square distribution with p degrees of freedom and $\Omega = \Sigma$. In general, X_i's covariance matrix Σ is related to Ω by $\sum = p^{-1}E(R_i^2)\Omega$.

2.1 The test statistic

Our test statistic T_n is based on the spatial sign function of the observed data. The spatial sign function of X_i is defined as $Z_i=\frac{X_i}{\|X_i\|}$ if $X_i \neq 0$; and $Z_i = 0$ if $X_i = 0$, where $\|X_i\|$ denotes the L_2 norm of X_i. The spatial sign vector is simply the unit vector in the direction of X_i. In the univariate case, it reduces to the familiar sign function.

We propose the following new nonparametric test statistics:

$$T_n = \sum_{i=1}^{n} \sum_{j=1}^{i-1} Z_i^T Z_j, \quad (3)$$

which indeed is a *U*-statistic. Under H_0, $E(Z_i) = 0$ which implies $E(T_n) = 0$. The above test statistic has an intuitive connection with the work of Bai and Saranadasa (1996) and Chen and Qin (2010), particularly the latter one. To see this, we note that the test statistic of Bai and Saranadasa (1996) for testing (1) is based on $\|\overline{X}\|^2$, while the one of Chen and Qin (2010) is based on $\sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} X_i^T X_j$. By removing the diagonal elements in the statistic of Bai and Saranadasa (1996), Chen and Qin (2010) was able to considerably relax the restrictive condition on p and n. In this spirit, we also dismiss the diagonal elements in defining T_n. Our test statistic hence can be deemed as a nonparametric extension of Chen and Qin (2010).

From another perspective, the new test generalizes the multivariate spatial sign test (e.g., Brown, 1983; Chaudhuri, 1992; Möttönen and Oja, 1995) to the high-dimensional setting. In the classical setting of $p < n$, Möttönen, Oja and Tienari (1997) derived the asymptotic relative efficiency (ARE) of the spatial sign test versus Hotelling's T^2 test and established its theoretical advantage for heavy-tailed distributions. For example, when the underlying distribution is a 10-dimensional *t* distribution with ν degrees of freedom, the ARE of the spatial sign test versus the Hotelling's T^2 test is 2.42 when $\nu = 3$, and is 0.95 when $\nu = \infty$ (multivariate normality). However, similarly as Hotelling's T^2 test, the multivariate spatial sign test is not defined when $p > n$. It is an open question whether we can modify it in a way such that its efficiency advantage can be preserved in the high-dimensional setting. This paper provides an affirmative answer.

Remark 1—It is interesting to compare with Bai and Saranadasa (1996) and Chen and Qin (2010), both of which adopt a factor model structure and a type of pseudo-independence assumption. It is noted that their model assumption excludes some commonly-used multivariate distributions such as the multivariate *t* distribution. However, we can show that Chen and Qin's test remain valid for the multivariate *t* distribution (see the Supplementary Material); but could suffer from substantial power loss. In Section 2.4, we also extend the new test to some important models in the Chen and Qin's class.

2.2 The limiting null distribution

Despite the simple form of T_n, deriving its asymptotic distribution when $p > n$ is by no means straightforward. As for any other high-dimensional inference, the most challenging issue lies in characterizing the underlying conditions for the asymptotic theory. In Bai and Saranadasa (1996) and Chen and Qin (2010), the key condition is stated through the behavior of the population covariance matrix $\Sigma = \mathrm{Cov}(X_i)$. In Bai and Saranadasa (1996), it is assumed that $\lambda_{\max}(\sum) = o\{\sqrt{\mathrm{Tr}^2(\sum^2)}\}$, where $\lambda_{\max}(\cdot)$ denotes the largest eigenvalue of a matrix and $\mathrm{Tr}(\cdot)$ denotes the trace. In Chen and Qin (2010), it is assumed that $\mathrm{Tr}(\Sigma^4) = o\{\mathrm{Tr}^2(\Sigma^2)\}$, which is satisfied under quite relaxed conditions on the eigenvalues of Σ. For

the nonparametric test T_n, it is desirable to characterize the underlying conditions in a similar fashion. However, this is challenging as the building blocks of T_n are the transformations Z_i's, which are not directly related to Σ.

In deriving the asymptotic properties of T_n, moment conditions directly related to Z_i's naturally arise. Lemma 2.1 below plays an important role in this paper. It establishes some of the key properties of the moments of Z_i's under a set of relaxed conditions on Σ. More specifically, we impose the following two conditions:

$$\mathrm{Tr}(\textstyle\sum^4)=o\{\mathrm{Tr}^2(\sum^2)\}. \quad \text{(C1)}$$

$$\frac{\mathrm{Tr}^4(\sum)}{\mathrm{Tr}^2(\sum^2)}\exp\left\{-\frac{\mathrm{Tr}^2(\sum)}{128p\lambda_{\max}^2(\sum)}\right\}=o(1). \quad \text{(C2)}$$

Lemma 2.1—Suppose that conditions (C1) and (C2) hold. Let $B=E\left(\frac{\varepsilon_i\varepsilon_i^T}{\|\varepsilon_i\|^2}\right)$. Then under H_0,

$$E\{(Z_1^TZ_2)^4\}=O(1)\left(E^2\{(Z_1^TZ_2)^2\}\right), \quad (4)$$

$$E\{(Z_1^TBZ_1)^2\}=O(1)\left(E^2(Z_1^TBZ_1)\right), \quad (5)$$

$$E\{(Z_1^TBZ_2)^2\}=o(1)\left(E^2(Z_1^TBZ_1)\right). \quad (6)$$

The above result is established by using a recent probability tool developed by El Karoui (2009) on the concentration inequality for the quadratic form of a random vector that has a uniform distribution on the unit sphere of $\mathbb{R}^P$.

Some intuition on T_n's asymptotic behavior under H_0 can be gained by observing its first two moments. First, it is evident that $E(T_n) = 0$. To calculate its variance, we write $T_n=\sum_{i=2}^n Y_i$, where $Y_i=\sum_{j=1}^{i-1}Z_i^TZ_j$. It follows from direct calculation that

$$\begin{aligned}E(Y_i^2)&=\sum_{j=1}^{i-1}\sum_{k=1}^{i-1}E(Z_i^TZ_jZ_i^TZ_k)=\sum_{j=1}^{i-1}E((Z_i^TZ_j)^2)\\&=(i-1)\mathrm{Tr}(E(Z_1Z_1^T)E(Z_2Z_2^T))=(i-1)\mathrm{Tr}(B^2),\end{aligned}$$

where B is defined in Lemma 2.1. Hence, $\mathrm{Var}(T_n)=\frac{n(n-1)}{2}\mathrm{Tr}(B^2)$. Although T_n has a U-statistics structure, the classical central limit theorem for U-statistics does not apply because the dimension p may depend on the sample size n. By applying Lemma 2.1 and exploring

the martingale structure of T_n, we can establish the asymptotic normality of $\frac{T_n}{\sqrt{\mathrm{Var}(T_n)}}$. The limiting null distribution of T_n is given in the following theorem.

Theorem 2.2—Assume conditions (C1) and (C2) hold. Then under H_0, as n, p $\to \infty$,

$$\frac{T_n}{\sqrt{\frac{n(n-1)}{2}Tr(B^2)}} \to N(0,1)$$ in distribution.

Remark 2: Condition (C1) holds trivially if all eigenvalues of Σ are bounded away from 0 and ∞. It is noted that the bounded eigenvalues assumption is commonly adopted in the literature of estimating high-dimensional covariance matrices (e.g., Bickel and Levina, 2008). It has also been shown that (C1) holds under some general conditions if some of the eigenvalues are unbounded (Chen and Qin, 2010).

Remark 3: Condition (C2) is new but quite relaxed. In particular, it is generally weaker than those conditions in the literature which explicitly imposed a relationship between *n* and *p* such as $p = o(n^2)$. Condition (C2) holds if all eigenvalues of Σ are bounded away from 0 and ∞. It also permits the eigenvalues to be unbounded as the exponential term is expected to converge to zero quickly if $\frac{\mathrm{Tr}(\sum)}{\sqrt{p}\lambda_{\max}(\sum)}$ diverges to ∞. To see this, let $\lambda_1 < \lambda_2 \le \cdots \le \lambda_p$ be ordered eigenvalues of Σ. Assume that as $p \to \infty$, k_1 eigenvalues converge to 0; k_2 eigenvalues diverge to ∞; and $p - k_1 - k_2$ eigenvalues remain bounded with lower bound $c_1 > 0$ and upper bound $c_2 < \infty$. Then

$$\frac{\mathrm{Tr}(\sum)}{\sqrt{p}\lambda_{\max}(\sum)} \ge \frac{k_1\lambda_1 + c_1(p-k_1-k_2) + k_2\lambda_{p-k_2+1}}{\sqrt{p}\lambda_p},$$
$$\frac{\mathrm{Tr}^2(\sum)}{\mathrm{Tr}(\sum^2)} \le \frac{k_2^2\lambda_p^2 + (p-k_2)^2c_2^2 + 2k_2(p-k_2)c_2\lambda_p}{k_1\lambda_1^2 + (p-k_1)c_1^2}.$$

Assume $\lambda_1 = p^{-b_1}$ and $\lambda_p = p^{b_2}$ for $b_1 > 0$, $b_2 > 0$. If both k_1 and k_2 are bounded, then it is easy to see that condition (C2) is satisfied if $b_2 < \frac{1}{2}$. It is noted that (C2) can still hold under some extra conditions on the rate of λ_1 and λ_p even if both k_1 and K_2 diverge to infinity at appropriate rate.

Remark 4: To apply T_n in practice, we need an estimator of $\mathrm{Tr}(B^2)$. Following Chen and Qin (2010), we may estimate $\mathrm{Tr}(B^2)$ using the cross-validation approach as follows:

$$\widehat{Tr(B^2)} = \{n(n-1)\}^{-1}\mathrm{Tr}\left\{\sum_{1\le j\ne k\le n}(Z_j - \overline{Z}_{(j,k)})Z_j^T(Z_k - \overline{Z}_{(j,k)})Z_k^T\right\}, \quad (7)$$

where $\overline{Z}_{(j,k)}$ is the sample mean after excluding Z_j and Z_k. It is noteworthy that the estimator in Chen and Qin can be computationally intensive for large *p* as each term inside the *U*-statistic involves multiplying high-dimensional matrices. In contrast, the computational

burden of the estimator in (7) can be substantially reduced by observing that $\|Z_j\|^2 = 1$. Let $\overline{Z}^* = (n-2)^{-1}\sum_{m=1}^{n} Z_m$. In the Appendix, it is derived that

$$\widehat{\mathrm{Tr}(B^2)} = -\frac{n}{(n-2)^2} + \frac{(n-1)}{n(n-2)^2}\mathrm{Tr}\{(\sum_{j=1}^{n} Z_j Z_j^T)^2\} + \frac{1-2n}{n(n-1)}\overline{Z}^{*T}(\sum_{j=1}^{n} Z_j Z_j^T)\overline{Z}^* + \frac{2}{n}\|\overline{Z}^*\|^2 + \frac{(n-2)^2}{n(n-1)}\|\overline{Z}^*\|^4. \quad (8)$$

In Figure 1(a), we plot the empirical distribution of $\frac{T_n}{\sqrt{\frac{n(n-1)}{2}\widehat{\mathrm{Tr}(B^2)}}}$ and compare it with the $N(0, 1)$ density curve for $n = 50$, $p = 1000$, where the data are generated from the $N_p(0, \Sigma)$ distribution with the (i, j) th entry of Σ equal to $0.8^{|i-j|}$. The two curves are very close to each other, which suggests that the standard normal distribution provides a satisfactory approximation of the null distribution.

2.3 Local power analysis

We now turn our attention to the power analysis of T_n under contiguous sequences of alternative hypotheses. This analysis enables us to further investigate the asymptotic relative efficiency of T_n with respect to Chen and Qin's test (referred to as CQ test in the sequel). Some interesting findings are revealed, which suggests promising efficiency gain of the new test for heavy-tailed multivariate distributions in the high-dimensional setting.

For the local power analysis, we impose the following additional conditions.

$$\exp\left(-\frac{\mathrm{Tr}^2(\sum)}{256p\lambda_{\max}^2(\sum)}\right) = o\left(\min\left(\frac{\lambda_{\max}(\sum)}{\mathrm{Tr}(\sum)}, \frac{\lambda_{\min}(\sum)}{\lambda_{\max}(\sum)}\right)\right). \quad \text{(C3)}$$

$$\lambda_{\max}(\sum) = o(\mathrm{Tr}(\sum)). \quad \text{(C4)}$$

$$\|\mu\|^2 E(\|\varepsilon\|^{-2}) = o\left(\min\left(n^{-1}\frac{\mathrm{Tr}(\sum^2)}{\lambda_{\max}(\sum)\mathrm{Tr}(\sum)}, n^{-1/2}\frac{\mathrm{Tr}^{1/2}(\sum^2)}{\mathrm{Tr}(\sum)}\right)\right). \quad \text{(C5)}$$

$$\text{For some } 0<\delta<1, \|\mu\|^{2\delta} E(\|\varepsilon\|^{-2-2\delta}) = o(E^2(\|\varepsilon\|^{-1})). \quad \text{(C6)}$$

Remark 5—Conditions (C3) and (C4) are concerned with the properties of the population covariance matrix Σ. These two conditions are relatively weak. In particular, they are satisfied when the eigenvalues of Σ are bounded away from 0 and ∞. Conditions (C5) and (C6) can be viewed as high-dimensional local-alternative statements for $p > n$. To gain some insight into the local alternative, we consider the case the eigenvalues of Σ are bounded away from 0 and ∞, then the right-hand side of (C5) is $o(n^{-1/2}p^{-1/2})$. For p-dimensional

spherical *t*-distribution with ν degrees of freedom, $\frac{p}{\|\varepsilon\|^2} \sim F(v,p)$. It is easy to show that $E(\|\varepsilon\|^{-2})=\frac{1}{p-2}$. A slightly more involved calculation based on the properties of *F*-distribution reveals that $E(\|\varepsilon\|)^{-1} = O(p^{-1/2}$ and $E(\|\varepsilon\|)^{-2-2\delta} = O(p^{-1-\delta})$. Then the conditions in (C4) and (C5) amount to $\|\mu\|^2 = o(n^{-1/2}\, p^{1/2})$ and $\|\mu\|^{2\delta} = o(p^{\delta})$ for some $0 < \delta < 1$. If $\delta = 1/2$, then the condition further reduces to $\|\mu\| = o(n^{-1/4}p^{1/4})$. If we consider the local alternatives such that all components of μ are equal to κ, then we have $\kappa = o(n^{-1/4}p^{-1/4})$, which when $p > n$ is of smaller order of $n^{-1/2}$, the usual local alternative rate for Hotelling's test with fixed dimension. The faster rate of local alternative can be viewed as a blessing of high dimensionality, where more information can be gained to distinguish subtle deviation from the null hypothesis.

Theorem 2.3—Assume conditions (C1)–(C6) hold. Letting $A=E\left\{\frac{1}{\|\varepsilon_i\|}(I_p - \frac{\varepsilon_i\varepsilon_i^T}{\|\varepsilon_i\|^2})\right\}$. Then as n, p $\to \infty$, $\frac{T_n - \frac{n(n-1)}{2}\mu^T A^2\mu(1+o(1))}{\sqrt{\frac{n(n-1)}{2}Tr(B^2)}} \to N(0,1)$ in distribution.

Theorem 2.3 implies that under the local alternatives, the proposed level α test has the local power $\beta_n=\Phi\left(-z_\alpha+\sqrt{\frac{n(n-1)}{2}}\frac{\mu^T A^2\mu(1+o(1))}{\sqrt{\mathrm{Tr}(B^2)}}\right)$, where $\Phi(\cdot)$ and z_α denote the cumulative distribution function and the upper α quantile of the *N* (0, 1) distribution, respectively. Let $\eta=\frac{\mu^T A^2\mu}{\sqrt{\mathrm{Tr}(B^2)}}$. Figure 1(b) plots the empirical power of the proposed test as a function of η (for $n = 50$, $p = 1000$) and compare it with the theoretical power given by the above formula. The data are generated from the multivariate *t* distribution with mean vector μ, covariance matrix Σ and 3 degrees of freedom, where the (*i*, *j*) th entry of Σ is $0.8^{|i-j|}$ and μ has all elements equal to κ with κ being chosen according to the value of η. The plot suggests that the theoretical formula of the local power provides a reasonable approximation to the empirical power.

On the other hand, the test of CQ test has the local power $\beta_n^{\mathrm{CQ}}=\Phi\left(-z_\alpha+\frac{n\|\mu\|^2}{\sqrt{2\mathrm{Tr}(\sum^2)}}\right)$. The asymptotic relative efficiency (ARE) of T_n versus the CQ test is

$$\mathrm{ARE}_{T_n,\mathrm{CQ}}=\frac{\mu^T A^2\mu}{\|\mu\|^2}\sqrt{\frac{\mathrm{Tr}(\sum^2)}{\mathrm{Tr}(B^2)}}(1+o(1)). \quad (9)$$

To appreciate the implication of the above result, we consider the asymptotic relative efficiency when the data arise from a spherical *p*-dimensional *t* distribution with ν degrees of freedom ($\nu > 2$). In this case, $A=E[\|\varepsilon\|^{-1}]\frac{p-1}{p}I_p$ where I_p denotes the $p \times p$ identity

matrix, $\mathrm{Tr}(B^2) = p^{-1}$ and $\mathrm{Tr}(\Sigma^2) = p^{-1}E^2[\|\varepsilon\|^2]$. Hence, $\mathrm{ARE}_{Tn,\mathrm{CQ}} = p^{-2}(p-1)2E^2[\|\varepsilon\|^{-1}]E[\|\varepsilon\|^2]$. For the t distribution, we have $E[\|\varepsilon\|^2]=\frac{pv}{v-2}$ and $E[\|\varepsilon\|^{-1}]=\sqrt{\frac{2}{pv}}\frac{\Gamma((v+1)/2)}{\Gamma(v/2)}$, where $\Gamma(t)=\int_0^\infty u^{t-1}e^{-u}du$ denotes the gamma function. For large p, the asymptotic relative efficiency thus is approximately $\mathrm{ARE}_{Tn,\mathrm{CQ}} \approx \frac{2}{v-2}\left(\frac{\Gamma((v+1)/2)}{\Gamma(v/2)}\right)^2$. For $\nu = 3$, this value is about 2.54; for $\nu = 4$, it is about 1.76; for $\nu = 5$, it is about 1.51; for $\nu = 6$, it is about 1.38; for $\nu = \infty$ (corresponding to multivariate normal distribution), by noting that $\Gamma((v+1)/2) \approx \Gamma(v/2)\sqrt{\frac{v}{2}}$ as $\nu \to \infty$, we have that the ARE has limit one. Theoretically, the efficiency loss of the new test under multivariate normality is little, but the efficiency gain can be substantial for heavy-tailed distribution. Recall that for $v = 3$, the ARE of the classical spatial sign test versus Hotelling's T^2 is 2.02 for $p = 3$ and 2.09 for $p = 10$ in the fixed dimensional case. This suggests that nonparametric test may have more substantial power gain in the high-dimensional case.

2.4 Extensions to beyond the elliptical distribution family

In this paper, we focus on the family of elliptical distributions because its popularity and flexibility for modeling non-normal multivariate data. Our results have the potential to extend to some useful multivariate distributions beyond the family of elliptical distributions.

One such class of distributions are those generated from the *symmetric independent component models* (e.g., Ilmonen and Paindaveine, 2011). That is,

$$X_i=\mu+\Gamma Z_i, \quad (10)$$

where Γ is a full rank $p \times p$ positive definite matrix; $Z_i = (Z_{i1}, \ldots, Z_{ip})'$ has independent components Z_{ij} and Z_{ij} is symmetric about zero. The independent components model assumes that the observed random vector can be written as linear combinations of independent random variables. This model has received broad attentions in signal processing and machine learning (Hyvärinen, Karhunen and Oja, 2001). For example, independent component analysis with exponential power marginal density ($p(x) \propto \exp(-|x|^q)$ for some $q > 0$) is popular for analyzing image and sound signals. It is noted that this class of models encompass many of the practically useful distributions from Bai and Saranadasa (1996) and Chen and Qin (2010).

We assume that Z_{ij} are standardized such that $Var(Z_{ij}) = 1$. Thus $\mathrm{Var}(X_i) = \Gamma\Gamma^T$. We also assume that Z_{ij} has a sub-exponential distribution with exponent α, that is, there exist constants $a > 0$, $b > 0$ such that for all $t > 0$, $P(|Z_{ij}| \geq t^\alpha) \leq a\exp(-bt)$. If $\alpha = 1/2$, then Z_{ij} is sub-gaussian. The class of sub-exponential distributions include many practically used heavy-tailed distributions. The fact that our proposed test is still valid for this class is summarized in the following theorem, whose proof is given in the Supplementary Material.

Theorem 2.4—Assume (C1) and (C2) hold for model (10), $E(Z_{ij}^4)$ for some positive constant c for all i, j, $E(Z_{ij}^4) \leq c$ and $\mathrm{Tr}(\Sigma^2) = o(\mathrm{Tr}^2(\Sigma))$. Then $\frac{T_n}{\sqrt{\frac{n(n-1)}{2}Tr(B^2)}} \to N(0,1)$ in distribution under H_0, as n, p $\to \infty$, where B has the same expression as in Lemma 2.1.

Another interesting extension of the elliptical distributions involves generating random variables from (2) but allowing R_i to be negative and depend on U_i. This yields the so-called family of *generalized elliptical distributions.* The asymptotic results of this paper also hold for this class by observing that $\frac{X_i}{\|X_i\|} = \frac{U_i}{\|U_i\|}$ under H_0. This class of models recently caught the attentions of researchers in finance, see Branco and Dey (2001), Frahm (2004), among others. A representative example of this class is the collection of *multi-tail elliptical distributions*, where R_i is a positive random variable whose tail parameter depends on ΓU_i (e.g., Kring et al, 2009; Rachev et al, 2011). The multi-tail elliptical distributions are particularly useful for modeling asset returns in finance.

Not surprisingly, generalizations to other multivariate distributions are possible although a case-by-case consideration may be needed. Particularly, the requirement that the U_i in (2) is uniformly distributed on the L_2 sphere can be relaxed. For example, one possible extension is to allow U_i to be from the class of distributions discussed in Gupta and Song (1997) and Szablowski (1998). Concentration inequalities similar to that given in Lemma A.2, which plays an important role in the proof, can be obtained for random vectors that satisfy certain concentration of measure properties (El Karoui, 2009).

3 Numerical studies

3.1 Monte Carlo simulations

We compare the performance of the new test with four alternatives: the test of Chen and Qin (CQ test, 2010), the test of Srivastava, Katayama and Kano (SKK test, 2013), the test based on multiple comparison with Bonferroni correction (BF test), and the test based on multiple comparison with FDR control (FDR test, Benjamini and Hochberg, 1995). The SKK test is constructed using the inverse of the diagonalized version of the sample covariance matrix and is computationally attractive as it involves a simple estimator for the asymptotic covariance. The BF test controls the family error rate at 0.05 and the FDR test controls the false discovery rate at 0.05. Both the BF test and FDR test are computed using the p-values from the *t* tests for the marginal hypotheses and reject H_0 if at least one marginal test is significant. The performance of the five tests are evaluated on 1000 simulation runs. We consider $n = 20, 50$ and $p = 200, 1000$ and 2000. To save space, we report the results for $p = 1000$ and 2000 here. The results for $p = 50, 100$ and 200 are reported in the Supplemental Material.

Example 1—In this example, random data were generated from $N_p(\mu, \Sigma)$. We consider three different choices for μ and three different choices for $\Sigma = (\sigma_{ij})$.

The three choices for μ are: (1) the null hypothesis $\mu_0 = (0, \ldots, 0)^T$; (2) the alternative $\mu_1 = (0.25, 0.25, \ldots, 0.25)^T$; and (3) the alternative $\mu_2 = (\mu_{21}, \ldots, \mu_{2p})^T$ with $\mu_{21} = \ldots = \mu_{2\frac{p}{3}} = 0$, $\mu_{2(\frac{p}{3}+1)} = \ldots = \mu_{2(\frac{2p}{3})} = 0.25$ and $\mu_{2(\frac{2p}{3}+1)} = \ldots = \mu_{2p} = -0.25$ The three choices for Σ are: (1) $\sigma_{ii} = 1$ and $\sigma_{ij} = 0.2$ ($i \neq j$); (2) $\sigma_{ij} = 0.8^{|i-j|}$; and (3) $\Sigma = DRD$, where $D = \text{diag}(d_1, \ldots, d_p)$ with $d_i = 2 + (p - i + 1)/p$, $R = (r_{ij})$ with $r_{ii} = 1$ and $r_{ij} = (-1)^{i+j}(0.2)^{|i-j|}$ for $i \neq j$. In the tables, we denote these three choices for Σ by Σ_1, Σ_2 and Σ_3, respectively. It is noted that Σ_3 was considered in Srivastava, Katayama and Kano (2013).

Table 1 summarizes the simulations results for different choices of Σ, μ, n and p. We observe that the five tests have nominal levels reasonably close to 0.05, especially when $n = 50$. For the alternative μ_1, the performance of the new test is very close to that of the CQ test and the SKK test, which are significantly better than the BF test and the FDR test. The latter two tests have especially low power when $n = 20$. For the alternative μ_2, we first note that the BF test and the FRD test perform fine for Σ_1 when $n = 50$ but has significantly lower power in all other settings. We also observe that the new test, the CQ test and the SKK test perform similarly for Σ_3; the new test has somewhat better performance for Σ_1; and the SKK test has somewhat better performance for Σ_2 for $p = 1000$.

Example 2—We simulate X_i from a p-variate t distribution with mean vector μ, covariance matrix Σ and 3 degrees of freedom. The choices of μ and Σ are set to be the same as those in Example 1. The distribution is heavy-tailed in this example.

We summarize the simulation results in Table 2. Both the new test and the CQ test have empirical levels close to 0.05 under the null hypothesis μ_0 while the other three tests tend to be conservative. In this example, the BF test and FDR test perform unsatisfactorily under the alternatives μ_1 and μ_2. It is observed that the new test has the best power performance in all settings; which is often substantially higher than (sometimes more than twofold) the second best performed test (the CQ test in this example). For example, the new test (and the CQ test) has power 0.83 (and 0.36) for the setting with $\mu = \mu_2$, $\Sigma = \Sigma_1$, $n = 20$ and $p = 2000$; 0.98 (and 0.47) for the setting with $\mu = \mu_1$, $\Sigma = \Sigma_2$, $n = 50$ and $p = 1000$; 0.88 (and 0.51) for the setting with $\mu = \mu_1$, $\Sigma = \Sigma_3$, $n = 20$ and $p = 2000$.

Example 3—We simulate X_i from a scale mixture of two multivariate normal distributions $0.9 * N_p(\mu, \Sigma) + 0.1 * N_p(\mu, 9\Sigma)$, where we consider the same choices of μ and Σ as in Example 1. The distribution in this example also has heavy tails.

We summarize the simulation results in Table 3. Similarly as in Example 2, the new test significantly outperforms the four contending approaches. For example, the new test (and the CQ test) has power 0.88 (and 0.49) for the setting with $\mu = \mu_2$, $\Sigma = \Sigma_1$, $n = 20$ and $p = 2000$; 0.80 (and 0.42) for the setting with $\mu = \mu_2$, $\Sigma = \Sigma_2$, $n = 50$ and $p = 1000$; 0.91 (and 0.70) for the setting with $\mu = \mu_1$, $\Sigma = \Sigma_3$, $n = 20$ and $p = 2000$.

3.2 An application

Type 2 diabetes is one of the most common chronic diseases. Insulin resistance in skeletal muscle, which is the major site of glucose disposal, is a prominent feature of Type 2

diabetes. To study insulins ability to regulate gene expression, an experiment performed microarray analysis using the Affymetrix Hu95A chip of human skeletal muscle biopsies from 15 diabetic patients both before and after insulin treatment (Wu et al., 2007). The gene expression alterations are promising to provide insights on new therapeutic targets for the treatment of this common disease. Hence, we are interested in testing the hypothesis in (1), where μ represents the average change of the gene expression level due to the treatment.

The underlying genetics of Type 2 diabetes were recognized to be very complex. It is believed that Type 2 diabetes is resulted from interactions between many genetic factors and the environment. The data were normalized by quantile normalization. When multiple probes are associated with the same gene, their expression values are consolidated by taking the average. In our analysis, we considered 2519 curated gene sets. The gene sets we used are from the C2 collection of the GSEA online pathway databases (http://www.broadinstitute.org/gsea/msigdb/collection details.jsp#C2). The largest gene set contains 1607 genes, which makes the hypothesis testing problem a high-dimensional one.

We applied both the new test and the CQ test at 5% significance level with the Bonferroni correction to control the family-wise error rate at 0.05 level. For the CQ method, 520 gene sets (20.64% of all candidates) are identified as significant; and for the new method, 954 gene sets (37.87% of all candidates) are selected as significant. We observe that the significant gene sets selected by the new test include those identified by the CQ test with only one exception (HASLINGER_B_CLL_WITH_CHROMOSOME_12_TRISOMY).

Table 4 displays the top 10 significant gene sets identified by the two tests and their corresponding test statistics values. The “NA” values in the table correspond to gene sets in the top 10 list of one test but not the other. We observe these two lists share 7 common gene sets. Among these seven gene sets, ZWANG_CLASS_2_TRANSIENTLY_INDUCED_BY_EGF, NAGASHIMA_EGF_SIGNALING_UP, AMIT_EGF_RESPONSE_60_HELA, AMIT_SERUM_RESPONSE_40_MCF10A and AMIT_SERUM_RESPONSE_60_MCF10A are known to be biologically related to insulin effect on human cells. We also observe that for 9 out of the top 10 gene sets, the new test has a smaller p-value than the CQ test does. The gene set SEMENZA_HIF1_TARGETS is only on the top ten list of the new test and was also found to be biologically related to insulin effect on human cells. Most of those significant gene sets are induced by Epidermal growth factor (EGF) or insulin-like growth factor (IGF).

It is interesting to point out that exploratory analysis of the gene expression data suggests the multivariate normality assumption is questionable. In fact, we investigated each of the top 10 gene sets identified by the new test and found that the multivariate normal distribution is plausible for none of them. For example, Figure 2 displays the histogram of the marginal kurtosises of the difference of each gene expression levels (before/after the treatment) of all genes in MCCLUNG_CREB1_TARGETS_DN gene set, which was selected among the top 10 gene sets by the new method but not by the CQ method. Figure 2 clearly shows that some gene expression levels have heavy tails as their kurtosises are much larger than 3, the kurtosis of a normal distribution.

4 Conclusion and discussions

The paper proposes a new spatial sign based nonparametric test for testing a hypothesis about the location parameter of a high-dimensional random vector. The goal is to improve the power performance when the underlying distribution of the data deviates from multivariate normality. We investigate the asymptotic properties of the new test and compare it with alternative tests based on extending Hotelling's T^2 test. A remarkable finding is the power improvement in the large p setting can be more substantial than that in the classical fixed p setting. The proposed test can be used as a basic building block to develop nonparametric tests in other important settings such as testing for sparse alternative or testing a hypothesis on coefficients in high-dimensional factorial designs (Zhong and Chen, 2011). A spatial sign based test was proposed for sphericity when $p = O(n^2)$ in Zou et al. (2014), and spatial sign tests were proposed for testing uniformity on the unit sphere and other related null hypotheses when $p/n \to c$ for some positive constant c in Paindaveinez and Verdebout (2013). The techniques related to sign tests have the potential to be used to develop the high-dimensional theory for other classical nonparametric multivariate testing procedures, such as those based on spatial sign ranks (e.g., Möttönen and Oja, 1995) and ranks (e.g., Hallin and Davy Paindaveine, 2006).

For reasons discussed in the introduction section, detecting the significance of a gene set is often of independent interest. In particular, finding significant gene sets/pathways can improve our understanding of the biological processes associated with a specific disease. The proposed method can also be incorporated into a multi-step procedure, in combination with various gene-level testing procedures and multiple tests correction methods, to further identify a short list of top genes for the biologists. This kind of multi-step procedure is expected to have better power to identify important individual genes as the gene set acts as a dimension reduction from potentially thousands of genes.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Wang and Peng's research is supported by a NSF grant DMS1308960. Li's research is supported by NIDA, NIH grants P50 DA10075 and P50 DA036107. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIDA or the NIH. We thank the Editor, the AE and three referees for their constructive comments which help us significantly improve the paper. We also thank Professor Tiefeng Jiang for helpful discussions.

Appendix: Technical proofs

Appendix 1: Some useful lemmas

We present below several useful technical lemmas, the proof for which can be found in the online supplementary material.

Lemma A.1

Let $U = (U_1, \ldots, U_p)^T$ be a random vector uniformly distributed on the unit sphere in $\mathbb{R}^p$. Then

1. $E(U) = 0$, $\mathrm{Var}(U) = p^{-1} I_p$, $E(U_j^4) = \frac{3}{p(p+2)}$, and $E(U_j^2 U_k^2) = \frac{1}{p(p+2)}$ for $j \neq k$.
2. Let M be a deterministic real-valued matrix. Assume that $\|M\|_2 \leq k$, where $\|M\|_2$ denotes the spectral norm of M. Then,

$$\forall t>0, P\left(|U^T M U - p^{-1} Tr(M)|>t\right) \leq 2\exp\left(-\frac{(p-1)(t-c_p)^2}{8k^2}\right), \text{ where } c_p = \sqrt{\frac{8\pi k^2}{p-1}}.$$

Lemma A.2

(A concentration inequality) Assume $W = \Gamma U$, where U is uniformly distributed on the unit sphere in $\mathbb{R}^p$. Let $\Omega = \Gamma\Gamma^T$ and consider the event $A = \{\frac{Tr(\Omega)}{2p} \leq \|W\|^2 \leq \frac{3Tr(\Omega)}{2p}\}$. Then $P(A) \geq 1 - c_1 \exp\left(-\frac{Tr^2(\sum)}{128 p \lambda_{\max}^2(\sum)}\right)$, for all $p > 1$, where $c_1 = 2\exp(\pi/2)$ is a finite constant.

Lemma A.3

For any p-dimensional vectors X and μ, we have (1) $\left\|\frac{X-\mu}{\|X-\mu\|} - \frac{X}{\|X\|}\right\| \leq 2\frac{\|\mu\|}{\|X\|}$; and (2) $\left\|\frac{X-\mu}{\|X-\mu\|} - \frac{X}{\|X\|} - \frac{1}{\|X\|}\left(I_p - \frac{XX^T}{\|X\|}\right)\mu\right\| \leq c_2 \frac{\|\mu\|^{1+\delta}}{\|X\|^{1+\delta}}$, for all $0 < \delta < 1$, where c_2 is a constant that does not depend on X or μ.

Lemma A.4

Let B be the matrix defined in Lemma 2.1. Assume condition (C3) holds, then

$$\lambda_{\max}(B) \leq \frac{2\lambda_{\max}(\sum)}{Tr(\sum)}(1+o(1)).$$

Lemma A.5

Let A be the matrix defined in Theorem 2.3 and $D = E\left\{\frac{1}{\|\varepsilon_1\|^2}\left(I_p - \frac{\varepsilon_1\varepsilon_1^T}{\|\varepsilon_1\|^2}\right)\right\}$ then $\lambda_{\max}(A) \leq E(\|\varepsilon_1\|^{-1})$ and $\lambda_{\max}(D) \leq E(\|\varepsilon_1\|^{-2})$. Furthermore, if conditions (C3) and (C4) hold, then $\lambda_{\max}(A) \geq \frac{1}{\sqrt{3}} E(\|\varepsilon_1\|^{-1})(1 - o(1))$.

Appendix 2: Proof of main theorems

We use *c* or *C* to denote generic positive constants, which may vary from line to line.

Proof of Theorem 2.2

Let $S_n^2 = Var(T_n) = \frac{n(n-1)}{2}\mathrm{Tr}(B^2) = \frac{n(n-1)}{2}E\{(Z_1^T Z_2)^2\}$. Let $V_n^2 = \sum_{i=2}^n E(Y_i^2 | Z_1, \ldots, Z_{i-1})$ and $Y_i = \sum_{j=1}^{i-1} Z_i^T Z_j$. To apply the martingale central limit theorem (Hall and Heyde, 1980), it is sufficient to check two conditions:

$$S_n^{-4}\sum_{i=2}^n E(Y_i^4) \to 0 \ \text{ as } n, p \to \infty, \quad \text{(A.1)}$$

$$S_n^{-2}V_n^2 \to 1 \ \text{ in probability as } \ n, p \to \infty. \quad \text{(A.2)}$$

To check (A.1), note that under H_0,

$$E(Y_i^4) = E\left\{\left(\sum_{j=1}^{i-1} Z_i^T Z_j\right)^4\right\} = \sum_{j=1}^{i-1} E\left\{\left(Z_i^T Z_j\right)^4\right\} + 3\sum_{\substack{1\le j,k\le i-1 \\ j\ne k}} E\left\{\left(Z_i^T Z_j\right)^2\left(Z_i^T Z_k\right)^2\right\}$$
$$= (i-1)E\{(Z_1^T Z_2)^4\} + 3(i-1)(i-2)E\left\{\left(Z_1^T Z_2\right)^2\left(Z_i^T Z_3\right)^2\right\}.$$

Hence, $\sum_{i=1}^n E(Y_i^4) \le c\left[n^2 E\{(Z_1^T Z_2)^4\} + n^3 E\{(Z_1^T Z_2)^2 (Z_1^T Z_3)^2\}\right] \le cn^3 E\{(Z_1^T Z_2)^4\}$ by Hölder's inequality. By Lemma 2.1, we have $E\{(Z_1^T Z_2)^4\} = o\left(nE^2\{(Z_1^T Z_2)^2\}\right)$. Therefore, (A.1) holds.

To prove (A.2), it is sufficient to verify that $\frac{E(V_n^2 - S_n^2)^2}{S_n^4} \to 0$ as $n, p \to \infty$. We write $V_n^2 = \sum_{i=2}^n V_{ni}$, where $V_{ni} = E(Y_i^2 | Z_1, \ldots, Z_{i-1})$. We have

$$V_{ni} = \sum_{j=1}^{i-1}\sum_{k=1}^{i-1} E(Z_i^T Z_j Z_i^T Z_k | Z_1, \ldots, Z_{i-1}) = \sum_{j=1}^{i-1}\sum_{k=1}^{i-1} \mathrm{Tr}(Z_j Z_k^T B) = \sum_{j=1}^{i-1}\sum_{k=1}^{i-1} Z_j^T B Z_k$$
$$= 2\sum_{1\le j<k\le i-1} Z_j^T B Z_k + \sum_{j=1}^{i-1} Z_j^T B Z_j.$$

If $j_1 \le k_1$ and $j_2 \le k_2$, then

$$E(Z_{j1}^T B Z_{k1} Z_{j2}^T B Z_{k2}) = E\left\{\left(Z_1^T B Z_1\right)^2\right\} I\{j_1 = k_1 = j_2 = k_2\} + E^2(Z_1^T B Z_1) I\{j_1 = k_1 \ne j_2 = k_2\}$$
$$+ E\left\{\left(Z_1^T B Z_2\right)^2\right\} I\{j_1 = j_2, k_1 = k_2, j_1 < k_1\}.$$

Therefore, for $i_1 < i_2$,

$$E(V_{ni_1}V_{ni_2})$$
$$=4\sum_{1\le j<k\le i_1-1} E\left\{(Z_1^TBZ_2)^2\right\}+\sum_{j=1}^{i_1-1}\sum_{k=1}^{i_2-1}E^2(Z_1^TBZ_1)+\sum_{j=1}^{i_1-1}[E\left\{(Z_1^TBZ_1)^2\right\}-E^2(Z_1^TBZ_1)]$$
$$=2(i_1-1)(i_1-2)E\left\{(Z_1^TBZ_2)^2\right\}+(i_1-1)(i_2-1)E^2(Z_1^TBZ_1)+(i_1-1)\,Var(Z_1^TBZ_1).$$

Consequently,

$$E(V_n^4)=E\{(\sum_{i=2}^n V_{ni})^2\}=2\sum_{2\le i<j\le n}E(V_{ni}V_{nj})+\sum_{j=2}^n E(V_{ni}^2)$$
$$=2\sum_{i=2}^n(i-1)(i-2)(2n-2i+1)E\{(Z_1^TBZ_2)^2\}+\sum_{i=2}^n(i-1)(2n-2i+1)\,Var(Z_1^TBZ_1)$$
$$+\{n(n-1)E(Z_1^TBZ_1)/2\}^2.$$

Note that $E(Z_1^TBZ_1)=\mathrm{Tr}(B^2)$ and $S_n^2=\frac{n(n-1)}{2}\mathrm{Tr}(B^2)$ Hence,

$$E\left\{(V_n^2-S_n^2)^2\right\}=E(V_n^4)-S_n^4$$
$$=2\sum_{i=2}^n(i-1)(i-2)(2n-2i+1)E\{(Z_1^TBZ_2)^2\}+\sum_{i=2}^n(i-1)(2n-2i+1)\,Var(Z_1^TBZ_1)$$
$$\le c\left[n^4E\{(Z_1^TBZ_2)^2\}+n^3E\{(Z_1^TBZ_1)^2\}\right].$$

Hence, a sufficient condition for $S_n^{-4}E(V_n^2-S_n^2)^2\to 0$. is

$\frac{n^4E\{(Z_1^TBZ_2)^2\}+n^3E\{(Z_1^TBZ_1)^2\}}{n^4E^2\{Z_1^TBZ_1\}}\to 0$. This condition holds by Lemma 2.1. This finishes the proof of Theorem 2.2. □

Proof of Theorem 2.3

Under the local alternatives,

$$T_n=\sum_{i=1}^n\sum_{\substack{j=1\\ j<i}}^n\left\{\frac{\varepsilon_i}{\|\varepsilon_i\|}+\left(\frac{\varepsilon_i+\mu}{\|\varepsilon_i+\mu\|}-\frac{\varepsilon_i}{\|\varepsilon_i\|}\right)\right\}^T\left\{\frac{\varepsilon_j}{\|\varepsilon_j\|}+\left(\frac{\varepsilon_j+\mu}{\|\varepsilon_j+\mu\|}-\frac{\varepsilon_j}{\|\varepsilon_j\|}\right)\right\}$$
$$=T_{n1}+T_{n2}+T_{n3},$$

where $T_{n1}=\sum_{i=1}^{n}\sum_{\substack{j=1\\j<i}}^{n}\frac{\varepsilon_i^T\varepsilon_j}{\|\varepsilon_i\|\|\varepsilon_j\|}$, $T_{n2}=\sum_{i=1}^{n}\sum_{\substack{j=1\\j\neq i}}^{n}\left(\frac{\varepsilon_i+\mu}{\|\varepsilon_i+\mu\|}-\frac{\varepsilon_i}{\|\varepsilon_i\|}\right)^T\frac{\varepsilon_j}{\|\varepsilon_i\|}$, and $T_{n3}=\sum_{i=1}^{n}\sum_{\substack{j=1\\j<i}}^{n}\left(\frac{\varepsilon_i+\mu}{\|\varepsilon_i+\mu\|}-\frac{\varepsilon_i}{\|\varepsilon_i\|}\right)^T\left(\frac{\varepsilon_j+\mu}{\|\varepsilon_j+\mu\|}-\frac{\varepsilon_j}{\|\varepsilon_j\|}\right)$. By Theorem 2.2, $T_{n1}/\sqrt{\frac{n(n-1)}{2}\mathrm{Tr}(B^2)}\to N(0,1)$.

To analyze T_{n2}, we write T_{n2} = T_{n21} + T_{n22}, where $T_{n21}=\sum_{i<j}\left(\frac{\varepsilon_i+\mu}{\|\varepsilon_i+\mu\|}-\frac{\varepsilon_i}{\|\varepsilon_i\|}\right)^T\frac{\varepsilon_j}{\|\varepsilon_j\|}$ and $T_{n22}=\sum_{j<i}\left(\frac{\varepsilon_i+\mu}{\|\varepsilon_i+\mu\|}-\frac{\varepsilon_i}{\|\varepsilon_i\|}\right)^T\frac{\varepsilon_j}{\|\varepsilon_j\|}$. Note that $E(T_{n21})=0$, and

$$\begin{aligned}E(T_{n21}^2)&=\sum_{i_1<j_1}\sum_{i_2<j_2}E\left\{\left(\frac{\varepsilon_{i_1}+\mu}{\|\varepsilon_{i_1}+\mu\|}-\frac{\varepsilon_{i_1}}{\|\varepsilon_{i_1}\|}\right)^T\frac{\varepsilon_{j_1}}{\|\varepsilon_{j_1}\|}\frac{\varepsilon_{j_2}^T}{\|\varepsilon_{j_2}\|}\left(\frac{\varepsilon_{i_2}+\mu}{\|\varepsilon_{i_2}+\mu\|}-\frac{\varepsilon_{i_2}}{\|\varepsilon_{i_2}\|}\right)\right\}\\&=\sum_{i_1<ji_2<j}\sum E\left\{\left(\frac{\varepsilon_{i_1}+\mu}{\|\varepsilon_{i_1}+\mu\|}-\frac{\varepsilon_{i_1}}{\|\varepsilon_{i_1}\|}\right)^T B\left(\frac{\varepsilon_{i_2}+\mu}{\|\varepsilon_{i_2}+\mu\|}-\frac{\varepsilon_{i_2}}{\|\varepsilon_{i_2}\|}\right)\right\}\\&\le\lambda_{\max}(B)\sum_{i_1<ji_2<j}\sum E\left\{\frac{4\|\mu\|^2}{\|\varepsilon_{i_1}\|\|\varepsilon_{i_2}\|}\right\}\\&\le 8\frac{\lambda_{\max}(\sum)}{\mathrm{Tr}(\sum)}(1+o(1))\|\mu\|^2\left\{\sum_{i<j}E(\|\varepsilon_{i_1}\|^{-2})+\sum_{i_1<ji_2<j}\sum E^2(\|\varepsilon_{i_1}\|^{-1})\right\}\\&\le O(n^3\|\mu\|^2)\frac{\lambda_{\max}(\sum)}{\mathrm{Tr}(\sum)}E(\|\varepsilon\|^{-2}),\end{aligned}$$

where the first inequality uses Lemma A.3, and the second inequality uses Lemma A.4. In the derivation in Lemma 2.1, we derived that $\mathrm{Tr}(B^2)\ge\frac{4}{9}\frac{\mathrm{Tr}(\sum^2)}{\mathrm{Tr}^2(\sum)}(1-o(1))$. Hence, it follows by condition (C5) that

$$\frac{E(T_{n21}^2)}{\frac{n(n-1)}{2}\mathrm{Tr}(B^2)}\le O(n\|\mu\|^2)\frac{\lambda_{\max}(\sum)\mathrm{Tr}(\sum)}{\mathrm{Tr}(\sum^2)}E(\|\varepsilon\|^{-2}=o(1)).$$

This implies $T_{n21}/\sqrt{\frac{n(n-1)}{2}\mathrm{Tr}(B^2)}=o_p(1)$. Similarly, $T_{n22}/\sqrt{\frac{n(n-1)}{2}\mathrm{Tr}(B^2)}=op(1)$.

Finally, we analyze T_{n3}. Denote

$$\begin{aligned}T_{n31}&=\tfrac{n(n-1)}{2}E\left(\tfrac{\varepsilon_1+\mu}{\|\varepsilon_1+\mu\|}\right)^T E\left(\tfrac{\varepsilon_2+\mu}{\|\varepsilon_2+\mu\|}\right),\\T_{n32}&=\sum_{j\neq i}\mathrm{E}\left(\frac{\varepsilon_i+\mu}{\|\varepsilon_i+\mu\|}\right)^T\left\{\frac{\varepsilon_j+\mu}{\|\varepsilon_j+\mu\|}-\frac{\varepsilon_j}{\|\varepsilon_j\|}-E\left(\frac{\varepsilon_j+\mu}{\|\varepsilon_j+\mu\|}\right)\right\},\\T_{n33}&=\sum_{j<i}\left\{\frac{\varepsilon_i+\mu}{\|\varepsilon_i+\mu\|}-\frac{\varepsilon_i}{\|\varepsilon_i\|}-E\left(\frac{\varepsilon_i+\mu}{\|\varepsilon_i+\mu\|}\right)\right\}^T\left\{\frac{\varepsilon_j+\mu}{\|\varepsilon_j+\mu\|}-\frac{\varepsilon_j}{\|\varepsilon_j\|}-E\left(\frac{\varepsilon_j+\mu}{\|\varepsilon_j+\mu\|}\right)\right\}.\end{aligned}$$

Then it follows that

$$T_{n3}=\sum_{i=1}^{n}\sum_{\substack{j=1\\ j<i}}^{n}\left[E\left(\frac{\varepsilon_i+\mu}{\|\varepsilon_i+\mu\|}\right)+\left\{\frac{\varepsilon_i+\mu}{\|\varepsilon_i+\mu\|}-\frac{\varepsilon_i}{\|\varepsilon_i\|}-E\left(\frac{\varepsilon_i+\mu}{\|\varepsilon_i+\mu\|}\right)\right\}\right]^T \times\left[E\left(\frac{\varepsilon_j+\mu}{\|\varepsilon_j+\mu\|}\right)+\left\{\frac{\varepsilon_j+\mu}{\|\varepsilon_j+\mu\|}-\frac{\varepsilon_j}{\|\varepsilon_j\|}-E\left(\frac{\varepsilon_j+\mu}{\|\varepsilon_j+\mu\|}\right)\right\}\right] =T_{n31}+T_{n32}+T_{n33}.$$

To analyze T_{n31}, by Lemma *A*.3 (2), we can write $E\left(\frac{\varepsilon_1+\mu}{\|\varepsilon_1+\mu\|}\right)=-A\mu+E(Q_1)$, where $Q_1=\frac{\varepsilon_1+\mu}{\|\varepsilon_1+\mu\|}-\frac{\varepsilon_1}{\|\varepsilon_1\|}+\frac{1}{\|\varepsilon_1\|}(I_p-\frac{\varepsilon_1\varepsilon_1^T}{\|\varepsilon_1\|^2})\mu$ satisfies $E(\|Q_1\|^2)\le c_3\|\mu\|^{2+2\delta}E(\|\varepsilon_1\|^{-2-2\delta})$ for all $0<\delta<1$, where c_3 is a constant that does not depend on ε_1 or μ. Hence,

$$E\left(\frac{\varepsilon_1+\mu}{\|\varepsilon_1+\mu\|}\right)^T E\left(\frac{\varepsilon_2+\mu}{\|\varepsilon_2+\mu\|}\right)=\mu^TA^2\mu-\mu^TAE(Q_1)-\mu^TAE(Q_2)+E(Q_1)^TE(Q_2).$$

Note that by Lemma A.5 and condition (C6), the last three terms on the right-hand side of the above expression are bounded by

$$2c_3^{1/2}\|\mu\|^{1+\delta}\|\mu^TA\|\cdot E^{1/2}\left(\|\varepsilon_1\|^{-2-2\delta}\right)+c_3\|\mu\|^{2+2\delta}E\left(\|\varepsilon_1\|^{-2-2\delta}\right) \le c\|\mu\|^2o(E^2(\|\varepsilon_1\|^{-1}))=o(\mu^TA^2\mu).$$

Therefore, $T_{n31}=\frac{n(n-1)}{2}\mu^TA^2\mu(1+o(1))$.

To evaluate Tn_{32}, we observe that $E(T_{n32})=0$ and that

$$E(T_{n32}^2)=O(n^3)E\left(\frac{\varepsilon_2+\mu}{\|\varepsilon_2+\mu\|}\right)^TE\{\left(\frac{\varepsilon_1+\mu}{\|\varepsilon_1+\mu\|}-\frac{\varepsilon_1}{\|\varepsilon_1\|}-E\left(\frac{\varepsilon_1+\mu}{\|\varepsilon_1+\mu\|}\right)\right) \left(\frac{\varepsilon_1+\mu}{\|\varepsilon_1+\mu\|}-\frac{\varepsilon_1}{\|\varepsilon_1\|}-E\left(\frac{\varepsilon_1+\mu}{\|\varepsilon_1+\mu\|}\right)\right)^T\}E\left(\frac{\varepsilon_3+\mu}{\|\varepsilon_3+\mu\|}\right).$$

Note that by Lemma A.3 (2),

$$\frac{\varepsilon_1+\mu}{\|\varepsilon_1+\mu\|}-\frac{\varepsilon_1}{\|\varepsilon_1\|}-E\left(\frac{\varepsilon_1+\mu}{\|\varepsilon_1+\mu\|}\right)=-\{\frac{1}{\|\varepsilon_1\|}\left(I_p-\frac{\varepsilon_1\varepsilon_1^T}{\|\varepsilon_1\|^2}\right)-A\}\mu+(Q_1-E(Q_1)).$$

Applying the above decomposition, we obtain

$$\lambda_{\max}\left[E\{\left(\frac{\varepsilon_1+\mu}{\|\varepsilon_1+\mu\|}-\frac{\varepsilon_1}{\|\varepsilon_1\|}-E\left(\frac{\varepsilon_1+\mu}{\|\varepsilon_1+\mu\|}\right)\right)\left(\frac{\varepsilon_1+\mu}{\|\varepsilon_1+\mu\|}-\frac{\varepsilon_1}{\|\varepsilon_1\|}-E\left(\frac{\varepsilon_1+\mu}{\|\in_1+\mu\|}\right)\right)^T\}\right]$$
$$\leq 2\|\mu\|^2\lambda_{\max}(D)+C\|\mu\|^{2+2\delta}E\left(\|\varepsilon_1\|^{-2-2\delta}\right),$$
$$\leq 2\|\mu\|^2E\left(\|\varepsilon_1\|^{-2}\right)+C\|\mu\|^{2+2\delta}E\left(\|\varepsilon_1\|^{-2-2\delta}\right),$$

by Lemma A.5, where $D=E\{\frac{1}{\|\varepsilon_1\|^2}\left(I_p-\frac{\varepsilon_1\varepsilon_1^T}{\|\varepsilon_1\|^2}\right)\}$. Therefore, by Lemma A.5, conditions (C5), (C6), and observing that $\mathrm{Tr}(B)^2\geq\frac{4\mathrm{Tr}(\sum^2)}{9\mathrm{Tr}^2(\sum)}(1-o(1))$, we have

$$E(T_{n32}^2)\leq O(n^3)\{2\|\mu\|^2E(\|\varepsilon_1\|^{-2})+C\|\mu\|^{2+2\delta}E(\|\varepsilon_1\|^{-2-2\delta})\}\|\mu\|^2E^2(\|\varepsilon_1\|^{-1})$$
$$\leq O(n^3)\|\mu\|^4E^2(\|\varepsilon_1\|^{-2})=o(n^2\mathrm{Tr}(B^2)).$$

Therefore, $T_{n32}/\sqrt{\frac{n(n-1)}{2}\mathrm{Tr}(B^2)}=o_p(1)$.

To evaluate T_{n33}, we observe that $E(T_{n33}) = 0$ and that

$$E(T_{n33}^2)$$
$$=O(n^2)E\left[\left\{\frac{\varepsilon_1+\mu}{\|\varepsilon_1+\mu\|}-\frac{\varepsilon_1}{\|\varepsilon_1\|}-E\left(\frac{\varepsilon_1+\mu}{\|\varepsilon_1+\mu\|}\right)\right\}^T\left\{\frac{\varepsilon_2+\mu}{\|\varepsilon_2+\mu\|}-\frac{\varepsilon_2}{\|\varepsilon_2\|}-E\left(\frac{\varepsilon_2+\mu}{\|\varepsilon_2+\mu\|}\right)\right\}\right.$$
$$\left.\left\{\frac{\varepsilon_2+\mu}{\|\varepsilon_2+\mu\|}-\frac{\varepsilon_2}{\|\varepsilon_2\|}-E\left(\frac{\varepsilon_2+\mu}{\|\varepsilon_2+\mu\|}\right)\right\}^T\left\{\frac{\varepsilon_1+\mu}{\|\varepsilon_1+\mu\|}-\frac{\varepsilon_1}{\|\varepsilon_1\|}-E\left(\frac{\varepsilon_1+\mu}{\|\varepsilon_1+\mu\|}\right)\right\}\right]$$
$$\leq O(n^2)\left\{2\|\mu\|^2\lambda_{\max}(D)+C\|\mu\|^{2+2\delta}E\left(\|\varepsilon_1\|^{-2-2\delta}\right)\right\}$$
$$\times E\left[\|\frac{\varepsilon_1+\mu}{\|\varepsilon_1+\mu\|}-\frac{\varepsilon_1}{\|\varepsilon_1\|}-E\left(\frac{\varepsilon_1+\mu}{\|\varepsilon_1+\mu\|}\right)\|^2\right]$$
$$\leq O(n^2)\left\{2\|\mu\|^2\lambda_{\max}(D)+(C)\|\mu\|^{2+2\delta}E\left(\|\varepsilon_1\|^{-2-2\delta}\right)\right\}^2$$
$$\leq O(n^2)\|\mu\|^4E^2\left(\|\varepsilon_1\|^{-2}\right)=o(n^2\mathrm{Tr}(B^2))$$

by Lemma A.5, conditions (C5) and (C6). Therefore, $T_{n33}/\sqrt{\frac{n(n-1)}{2}\mathrm{Tr}(B^2)}=o_p(1)$.

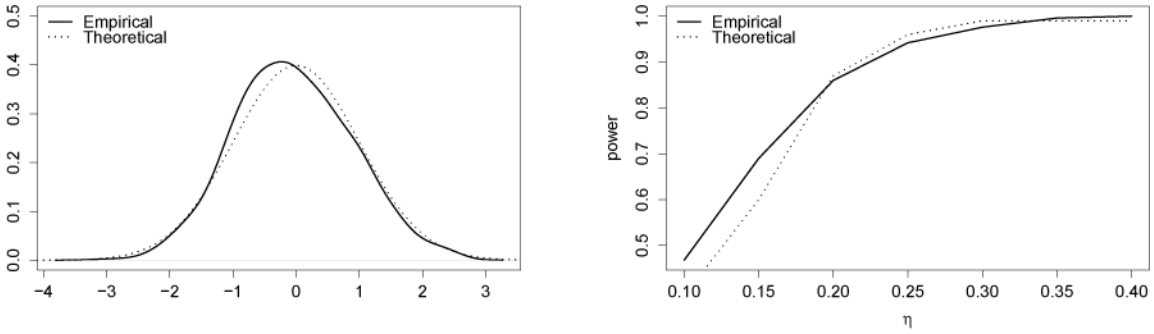
Summarizing the above, $T_{n3}/\sqrt{\frac{n(n-1)}{2}\mathrm{Tr}(B^2)}=\frac{\frac{n(n-1)}{2}\mu^TA^2\mu(1+o(1))}{\sqrt{\frac{n(n-1)}{2}\mathrm{Tr}(B^2)}}+o_p(1)$. This finishes the proof. □

References

1. Bai Z, Sarandasa H. Effect of High Dimension: By an Example of a Two Sample Problem. Statistica Sinica. 1996; 6:311–329.
2. Benjamini, Yoav; Hochberg, Yosef. Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society, Series B. 1995; 57:289?00.
3. Bickel PJ, Levina E. Covariance Regularization by Thresholding. Annals of Statistics. 2008; 36:2577–2604.

4. Branco MD, Dey DK. A General Class of Multivariate Skew-elliptical Distributions. Journal of Multivariate Analysis. 2001; 79:99–113.
5. Brown BM. Statistical Uses of the Spatial Median. Journal of the Royal Statistical Society, Series B. 1983; 45:25–30.
6. Cai T, Liu W, Xia Y. Two-sample Test of High Dimensional Means under Dependence. Journal of the Royal Statistical Society, Series B. 2014; 76:349–372.
7. Chaudhuri P. Multivariate Location Estimation Using Extension of R-estimates through U-statistics Type Approach. Annals of Statistics. 1992; 20:897–916.
8. Chen SX, Qin YL. A Two-sample Test for High-dimensional Data with Application to Gene-Set Testing. Annals of Statistics. 2010; 38:808–835.
9. Cook RD, Forzani L, Rothman AJ. Estimating Sufficient Reductions of the Predictors in Abundant High-dimensional Regressions. Annals of Statistics. 2012; 40(353):84.
10. El Karoui N. Concentration of Measure and Spectra of Random Matrices: with Applications to Correlation Matrices, Elliptical Distributions and Beyond. The Annals of Applied Probability. 2009; 19:2362–2405.
11. Fang, KT.; Kotz, S.; Ng, KW. Symmetric Multivariate and Related Distributions. Chapman and Hall; London: 1990.
12. Frahm, G. Ph.D. thesis. University of Cologne; Germany: 2004. Generalized Elliptical Distributions: Theory and Applications.
13. Gupta AK, Song D. L_p-norm Spherical Distributions. Journal of Statistical Planning and Inference. 1997; 100:241–260.
14. Hallin M, Paindaveine D. Semiparametrically Efficient Rank-based Inference for Shape I. Optimal Rank-based Tests for Sphericity. The Annals of Statistics. 2006; 34:2707–2756.
15. Hall, P.; Heyde, C. Martingale Limit Theory and Applications. Academic Press; New York: 1980.
16. Hall P, Jin J. Innovated Higher Criticism for Detecting Sparse Signals in Correlated Noise. The Annals of Statistics. 2010; 38:1686–1732.
17. Hyvärinen, A.; Karhunen, J.; Oja, E. Independent Component Analysis. John Wiley & Sons; New York: 2001.
18. Ilmonen P, Paindaveine D. Semiparametrically Efficient Inference based on Signed Ranks in Symmetric Independent Component Models. Annals of Statistics. 2011; 39:2448–2476.
19. Kring S, Rachev ST, Hchsttter M, Fabozzi FJ, Bianchi ML. Multitail Generalized Elliptical Distributions for Asset Returns. The Econometrics Journal. 2009; 12(272):91.
20. Lee SH, Limb J, Li E, Vannuccid M, Petkova E. Order test for high-dimensional two-sample means. Journal of Statistical Planning and Inference using random subspaces. 2012; 142:2719–2725.
21. Ledoux, M. The Concentration of Measure Phenomenon. American Mathematical Society; Providence, Rhode Island: 2001.
22. Mcneil, AJ.; Frey, R.; Embrechts, P. Quantitative Risk Management: Concepts, Techniques and Tools. Princeton University Press; Princeton, NJ: 2005.
23. Möttönen J, Oja H. Multivariate Spatial Sign and Rank Methods. Journal of Nonparametric Statistics. 1995; 5:201–213.
24. Möttönen J, Oja H, Tienari J. On the Efficiency of Multivariate Spatial Sign and Rank Tests. Annals of Statistics. 1997; 25:542–552.
25. Oja, H. Multivariate Nonparametric Methods with R. Springer; 2010.
26. Paindaveine, D.; Verdebout, T. Universal Asymptotics for High-dimensional Sign Tests. Université libre de Bruxellesz; 2013. technical report
27. Pan GM, Zhou W. Central Limit Theorem for Hotelling's T^2 Statistic under Large Dimension. Annals of Applied Probability. 2011; 21:1860–1910.
28. Purdom E, Holmes SP. Error Distribution for Gene Expression Data. Statistical Applications in Genetics and Molecular Biology. 2005; 4 Article 16.
29. Rachev, ST.; Kim, YS.; Bianchi, ML.; Fabozzi, FJ. Financial Models with Lévy Processes and Volatility Clustering. John Wiley & Sons; Hoboken, NJ, USA: 2011. Multi-Tail t-Distribution.

30. Schmidt R. Tail Dependence for Elliptically Contoured Distributions. Mathematical Methods of Operations Research. 2002; 55:301–327.
31. Srivastava M. A Test for the Mean Vector with Fewer Observations than the Dimension under Non-normality. Journal of Multivariate Analysis. 2009; 100:386–402.
32. Srivastava MS, Du M. A Test for the Mean Vector with Fewer Observations than the Dimension. Journal of Multivariate Analysis. 2008; 99:386–402.
33. Srivastava MS, Katayama S, Kano Y. A two sample test in high dimensional data. Journal of Multivariate Analysis. 2013; 114:349–358.
34. Szabowski PJ. Uniform Distributions on Spheres in Finite-dimensional L_a and Their Generalization. Journal of Multivariate Analysis. 1998; 64:103–117.
35. Thulin M. 11A high-dimensional two-sample test for the mean using random subspaces. Computational Statistics & Data Analysis. 2014; 74:26–38.
36. Vo, T.; Phan, J.; Huynh, K.; Wang, M. Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE. 2007. Reproducibility of Differential Gene Detection across Multiple Microarray Studies; p. 4231?234
37. Wu X, Wang J, Cui X, Maianu L, et al. The Effect of Insulin on Expression of Genes and Biochemical Pathways in Human Skeletal Muscle. Endocrine. 2007; 31:5–17. [PubMed: 17709892]
38. Zhong PS, Chen SX. Tests for High Dimensional Regression Coefficients with Factorial Designs. Journal of the American Statistical Association. 2011; 106:260–274.
39. Zhong PS, Chen SX, Xu MY. Tests Alter- native to Higher Criticism for High Dimensional Means under Sparsity and Column-wise Dependence. The Annals of Statistics. 2013; 41:2703–3110.
40. Zou CL, Peng LH, Wang ZJ. Multivariate Sign-based High-dimensional Tests for Sphericity. Biometrika. 2014; 101:229–236.



(a) Density curves under the null hypothesis (b) Power curves

Figure 1.

Comparing the empirical distribution of the new test with the theoretical distribution (n = 50, p = 1000)

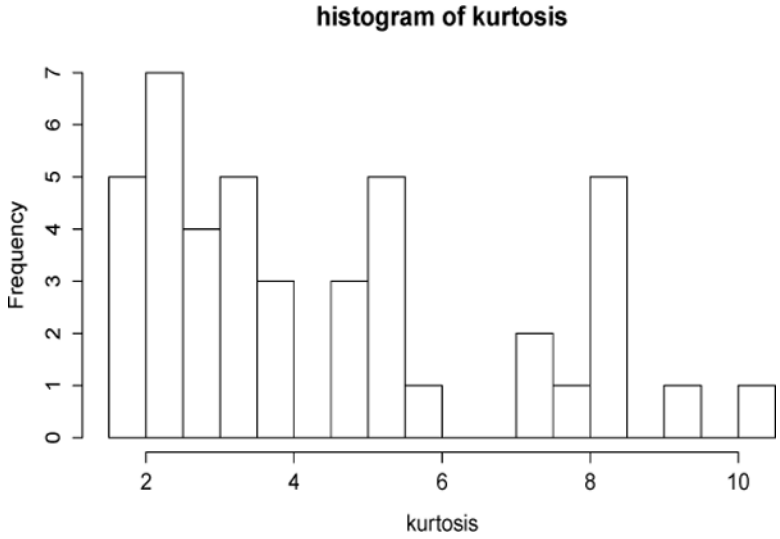


Figure 2.
The histogram of marginal kurtosises for all genes in MCCLUNG_CREB1_TARGETS_DN gene set.

J Am Stat Assoc. Author manuscript; available in PMC 2017 January 15.

Table 1

Example 1: multivariate normal distribution

Σ	***μ***	***n***	***p***	**New**	**CQ**	**SKK**	**BF**	**FDR**
Σ_1	μ_0	20	1000	0.066	0.069	0.061	0.046	0.046
		20	2000	0.073	0.070	0.046	0.052	0.053
		50	1000	0.059	0.060	0.043	0.035	0.038
		50	2000	0.058	0.061	0.029	0.043	0.047
	μ_1	20	1000	0.723	0.723	0.692	0.405	0.471
		20	2000	0.720	0.729	0.638	0.385	0.447
		50	1000	0.975	0.976	0.962	0.842	0.890
		50	2000	0.970	0.976	0.945	0.850	0.901
	μ_2	20	1000	0.951	0.826	0.650	0.382	0.443
		20	2000	0.954	0.821	0.567	0.404	0.464
		50	1000	1.000	1.000	1.000	0.964	0.997
		50	2000	1.000	1.000	1.000	0.973	0.998
Σ_2	μ_0	20	1000	0.052	0.051	0.047	0.038	0.041
		20	2000	0.058	0.059	0.023	0.047	0.047
		50	1000	0.048	0.050	0.051	0.043	0.048
		50	2000	0.060	0.060	0.052	0.054	0.061
		50	1000	0.795	0.797	0.815	0.122	0.138
	μ_1	20	2000	0.969	0.968	0.930	0.134	0.145
		50	1000	0.999	0.999	0.999	0.357	0.430
		50	2000	1.000	1.000	1.000	0.416	0.479
	μ_2	20	1000	0.540	0.549	0.579	0.092	0.102
		20	2000	0.790	0.788	0.695	0.102	0.112
		50	1000	0.992	0.991	0.994	0.265	0.289
		50	2000	1.000	1.000	1.000	0.343	0.381
Σ_3	μ_0	20	1000	0.055	0.055	0.067	0.052	0.052

Σ	μ	n	p	New	CQ	SKK	BF	FDR
		20	2000	0.059	0.061	0.048	0.042	0.044
		50	1000	0.052	0.052	0.061	0.042	0.044
		50	2000	0.045	0.045	0.052	0.048	0.048
	μ_1	20	1000	0.490	0.438	0.514	0.127	0.132
		20	2000	0.646	0.594	0.556	0.113	0.119
		50	1000	1.000	0.998	1.000	0.296	0.331
		50	2000	1.000	1.000	1.000	0.331	0.369
	μ_2	20	1000	0.242	0.225	0.335	0.100	0.104
		20	2000	0.342	0.310	0.318	0.084	0.092
		50	1000	0.932	0.862	0.982	0.239	0.260
		50	2000	0.991	0.987	1.000	0.269	0.296

Table 2

Example 2: multivariate t-distribution

Σ	***μ***	***n***	***p***	**New**	**CQ**	**SKK**	**BF**	**FDR**
Σ_1	μ_0	20	1000	0.083	0.088	0.012	0.011	0.013
		20	2000	0.064	0.072	0.007	0.010	0.010
		50	1000	0.053	0.063	0.015	0.011	0.012
		50	2000	0.069	0.076	0.008	0.010	0.010
	μ_1	20	1000	0.633	0.472	0.222	0.153	0.183
		20	2000	0.631	0.468	0.171	0.117	0.138
		50	1000	0.941	0.721	0.493	0.424	0.491
		50	2000	0.921	0.736	0.448	0.438	0.492
	μ_2	20	1000	0.815	0.371	0.076	0.129	0.150
		20	2000	0.830	0.363	0.040	0.107	0.122
		50	1000	1.000	0.803	0.333	0.427	0.485
		50	2000	1.000	0.825	0.288	0.493	0.571
Σ_2	μ_0	20	1000	0.052	0.053	0.000	0.011	0.012
		20	2000	0.058	0.060	0.000	0.013	0.013
		50	1000	0.052	0.053	0.000	0.015	0.015
		50	2000	0.059	0.060	0.000	0.020	0.021
	μ_1	20	1000	0.682	0.349	0.013	0.029	0.033
		20	2000	0.883	0.512	0.002	0.038	0.040
		50	1000	0.996	0.780	0.120	0.094	0.102
		50	2000	1.000	0.933	0.091	0.118	0.128
	μ_2	20	1000	0.441	0.228	0.001	0.027	0.029
		20	2000	0.654	0.339	0.000	0.027	0.028
		50	1000	0.942	0.570	0.037	0.067	0.071
		50	2000	0.998	0.785	0.018	0.077	0.086
Σ_3	μ_0	20	1000	0.054	0.058	0.001	0.023	0.023

Σ	μ	n	p	New	CQ	SKK	BF	FDR
		20	2000	0.061	0.057	0.000	0.011	0.011
		50	1000	0.056	0.057	0.002	0.015	0.015
		50	2000	0.048	0.042	0.001	0.015	0.015
	μ_1	20	1000	0.355	0.174	0.001	0.033	0.034
		20	2000	0.488	0.224	0.003	0.022	0.022
		50	1000	0.979	0.465	0.021	0.080	0.084
		50	2000	0.998	0.624	0.015	0.095	0.103
	μ_2	20	1000	0.198	0.113	0.001	0.034	0.034
		20	2000	0.251	0.141	0.000	0.018	0.019
		50	1000	0.766	0.249	0.010	0.059	0.064
		50	2000	0.922	0.349	0.005	0.070	0.073

Table 3

Example 3: mixture of multivariate normal distributions

Σ	μ	n	p	New	CQ	SKK	BF	FDR
Σ_1	μ_0	20	1000	0.063	0.070	0.007	0.014	0.015
		20	2000	0.045	0.049	0.007	0.015	0.018
		50	1000	0.063	0.066	0.014	0.020	0.021
		50	2000	0.042	0.040	0.007	0.015	0.016
	μ_1	20	1000	0.649	0.548	0.277	0.209	0.259
		20	2000	0.627	0.542	0.229	0.193	0.231
		50	1000	0.941	0.859	0.730	0.600	0.682
		50	2000	0.964	0.867	0.701	0.619	0.687
	μ_2	20	1000	0.870	0.449	0.109	0.175	0.201
		20	2000	0.882	0.492	0.089	0.224	0.243
		50	1000	1.000	0.966	0.700	0.663	0.759
		50	2000	1.000	0.968	0.577	0.698	0.800
Σ_2	μ_0	20	1000	0.046	0.063	0.006	0.019	0.019
		20	2000	0.050	0.047	0.004	0.017	0.018
		50	1000	0.039	0.049	0.000	0.021	0.023
		50	2000	0.041	0.039	0.001	0.020	0.021
	μ_1	20	1000	0.678	0.485	0.093	0.059	0.067
		20	2000	0.914	0.700	0.126	0.069	0.070
		50	1000	0.998	0.943	0.230	0.157	0.177
		50	2000	1.000	0.995	0.167	0.190	0.211
	μ_2	20	1000	0.437	0.285	0.059	0.050	0.053
		20	2000	0.687	0.478	0.088	0.047	0.050
		50	1000	0.953	0.759	0.045	0.111	0.118
		50	2000	0.998	0.940	0.040	0.150	0.169
Σ_3	μ_0	20	1000	0.054	0.053	0.008	0.020	0.020

Σ	μ	n	p	New	CQ	SKK	BF	FDR
		20	2000	0.039	0.045	0.002	0.022	0.022
		50	1000	0.050	0.050	0.000	0.026	0.026
		50	2000	0.035	0.034	0.000	0.021	0.022
	μ_1	20	1000	0.342	0.207	0.045	0.058	0.061
		20	2000	0.493	0.307	0.072	0.046	0.049
		50	1000	0.995	0.730	0.050	0.124	0.129
		50	2000	1.000	0.879	0.035	0.111	0.119
	μ_2	20	1000	0.178	0.130	0.030	0.046	0.050
		20	2000	0.249	0.173	0.043	0.047	0.048
		50	1000	0.797	0.421	0.012	0.100	0.110
		50	2000	0.947	0.565	0.011	0.092	0.093

Table 4

The top 10 significant gene sets selected by the new test and the CQ test

Gene set	New test	CQ test
ZWANG_CLASS_2_TRANSIENTLY_INDUCED_BY_EGF	24.34	20.11
NAGASHIMA_EGF_SIGNALING_UP	22.44	17.01
SHIPP_DLBCL_CURED_VS_FATAL_DN	22.34	18.24
WILLERT_WNT_SIGNALING	19.66	NA
UZONYI_RESPONSE_TO_LEUKOTRIENE_AND_THROMBIN	19.63	18.46
PID_HIF2PATHWAY	19.46	15.65
PHONG_TNF_TARGETS_UP	19.21	18.90
AMIT_EGF_RESPONSE_60_HELA	18.64	16.38
MCCLUNG_CREB1_TARGETS_DN	18.43	NA
SEMENZA_HIF1_TARGETS	18.34	NA
AMIT_SERUM_RESPONSE_40_MCF10A	NA	15.98
AMIT_SERUM_RESPONSE_60_MCF10A	NA	15.43
PLASARI_TGFB1_TARGETS_1HR_UP	NA	15.00