Computational Statistics and Data Analysis 144 (2020) 106889

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

A high-dimensional spatial rank test for two-sample location problems

Long Feng [a], Xiaoxu Zhang [b], Binghui Liu [b,*]

[a] School of Statistics and Data Science, LPMC and KLMDASR, Nankai University, China
[b] Key Laboratory for Applied Statistics of MOE and School of Mathematics and Statistics, Northeast Normal University, China

ARTICLE INFO

Article history:
Received 10 December 2018
Received in revised form 2 November 2019
Accepted 12 November 2019
Available online 21 November 2019

Keywords:
High-dimensional
Scalar-invariant
Spatial rank
Two-sample location problems

ABSTRACT

In high-dimensional situations, the traditional multivariate sign- or rank-based procedures for the two-sample location testing problems are ineffective, since in the construction of the test statistics, the scatter matrix to be inverted is singular. To solve this problem, many high-dimensional spatial sign or rank tests have been proposed, some of which are very efficient. However, most of these existing tests no longer work in very high dimensional situations, which only allows the dimension of variables to be the square of the sample sizes at most, hence are restrictive for practical applications. On this ground, a new high-dimensional spatial rank test is proposed in this paper, which is invariant under scalar transformations, maintains the efficiency advantage of spatial-rank-based testing methods, and could even allow the dimension to grow almost exponentially with the sample sizes. The theoretical results of the proposed test are established, followed by some convincing numerical results and two real data analyses.

1. Introduction

Nowadays, high-dimensional or even ultra-high dimensional data are becoming increasingly available in many fields of application, as data collection technology evolves very quickly. Here the dimensionality is high when the number of variables is larger than the sample size, while it is ultra-high when the number of variables is one or several orders of magnitude larger than the sample size, as pointed out in Fan and Lv (2008). Researchers from these fields of application urgently need powerful, effective and robust analytic methods to take full advantage of the core value in these data. Although traditional statistical methods still rule the roost, their serious flaws in high-dimensional situations can generally not be ignored, which make them no longer suitable for high-dimensional data. As for hypothesis testing problems, most traditional testing methods are not available in high-dimensional settings, which forces statisticians to make a sustained effort to improve existing high-dimensional testing methods.

In this paper, we consider the high-dimensional two-sample location testing problems. Let $\boldsymbol{X}_{k1}, \ldots, \boldsymbol{X}_{kn_k}$, $k \in \{1, 2\}$, be i.i.d. copies of two independent random vectors respectively, which obey p_n-variate elliptical distributions with the following density function respectively:

$$\det(\boldsymbol{\Sigma}_{kn})^{-1/2} g_n(\|\boldsymbol{\Sigma}_{kn}^{-1/2}(\mathbf{x} - \boldsymbol{\mu}_{kn})\|), \quad k \in \{1, 2\}, \tag{1}$$

* Correspondence to: 5268 Renmin Street, Changchun, Jilin Province, China.
E-mail address: liubh100@nenu.edu.cn (B. Liu).

https://doi.org/10.1016/j.csda.2019.106889
0167-9473/

where $n = n_1 + n_2$, $\boldsymbol{\mu}_{kn}$ are the symmetry centers and $\boldsymbol{\Sigma}_{kn}$ are the positive definite symmetric $p_n \times p_n$ scatter matrices. The two-sample location testing problem can be depicted as follows

$$H_0 : \boldsymbol{\mu}_{1n} = \boldsymbol{\mu}_{2n} \ \text{ versus } \ H_1 : \boldsymbol{\mu}_{1n} \neq \boldsymbol{\mu}_{2n}. \tag{2}$$

There is a lot of existing literature on this problem. As the dimension of the variables is assumed to be fixed and the sample size can grow to infinity, the most familiar method for the two-sample location problem is the Hotelling's T^2 test, with the test statistic:

$$H_n = \frac{n_1 n_2}{n}(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2)^T \mathbf{S}_n^{-1}(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2), \tag{3}$$

where $\bar{\boldsymbol{X}}_1$ and $\bar{\boldsymbol{X}}_2$ are the corresponding sample means and $\mathbf{S}_n$ is the pooled sample covariance matrix. It fails when the dimension becomes larger than the sample size, because in this situation the sample covariance matrix is singular hence cannot be inverted in the construction of test statistic.

On this ground, Bai and Saranadasa (1996) proposed a test statistic by replacing the Mahalanobis norm in the Hotelling's T^2 test statistic with the Euclidian norm, which is based on $M_n = \|\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2\|^2$ by replacing $\mathbf{S}_n$ in H_n with $\mathbf{I}_{p_n}$, the $p_n \times p_n$ identity matrix. Then, to make it available for ultra-high-dimensional situation, Chen et al. (2010) proposed a modified test statistic (abbreviated as CQ hereafter) as follows

$$W_n = \frac{\sum_{i\neq j}^{n_1} \boldsymbol{X}_{1i}^T \boldsymbol{X}_{1j}}{n_1(n_1-1)} + \frac{\sum_{i\neq j}^{n_2} \boldsymbol{X}_{2i}^T \boldsymbol{X}_{2j}}{n_2(n_2-1)} - 2\frac{\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \boldsymbol{X}_{1i}^T \boldsymbol{X}_{2j}}{n_1 n_2}, \tag{4}$$

by removing $\sum_{i=1}^{n_k} \boldsymbol{X}_{ki}^T \boldsymbol{X}_{ki}$ from $M_n = \|\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2\|^2$, because the terms to be removed could have imposed certain demands on the dimensionality. However, neither of these two statistics is invariant under scalar transformations, which may make them suffer from scalar transformations: the same dataset might generate different conclusions due to different scalar transformations. Because of this, under the normality assumption, Srivastava and Du (2008) proposed a scalar-transformation-invariant test under the assumption of the equality of the two covariance matrices. Srivastava et al. (2013) extended the results of Srivastava and Du (2008) to unequal covariance matrices.

Then, to develop a scale-invariant test applicable to high dimensional data, Park and Ayyala (2013) proposed a test statistic (abbreviated as PA hereafter) by leave-out cross validation:

$$\begin{aligned} P_n = \frac{n_1+n_2-6}{n_1+n_2-4}\Bigg(& \frac{1}{n_1(n_1-1)} \sum_{i\neq j}^{n_1} \boldsymbol{X}_{1i}^T \mathbf{D}_{\mathbf{S}^*_{1(i,j)}}^{-1} \boldsymbol{X}_{1j} + \frac{1}{n_2(n_2-1)} \\ & \sum_{i\neq j}^{n_2} \boldsymbol{X}_{2i}^T \mathbf{D}_{\mathbf{S}^*_{2(i,j)}}^{-1} \boldsymbol{X}_{2j} - \frac{2}{n_1 n_2} \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \boldsymbol{X}_{1i}^T \mathbf{D}_{\mathbf{S}^*_{12(i,j)}}^{-1} \boldsymbol{X}_{2j} \Bigg), \end{aligned} \tag{5}$$

where $\mathbf{D}_{\mathbf{S}^*_{1(i,j)}}$, $\mathbf{D}_{\mathbf{S}^*_{2(i,j)}}$ and $\mathbf{D}_{\mathbf{S}^*_{12(i,j)}}$ are the diagonal matrices of $\mathbf{S}^*_{1(i,j)}$, $\mathbf{S}^*_{2(i,j)}$ and $\mathbf{S}^*_{12(i,j)}$ respectively. Here

$$\begin{aligned} \mathbf{S}^*_{1(i,j)} &= \frac{(n_1-3)\mathbf{S}_{1(i,j)} + (n_2-1)\mathbf{S}_{2,n_2}}{n_1+n_2-4}, \\ \mathbf{S}^*_{2(i,j)} &= \frac{(n_1-1)\mathbf{S}_{1,n_1} + (n_2-3)\mathbf{S}_{2(i,j)}}{n_1+n_2-4} \ \text{ and} \\ \mathbf{S}^*_{12(i,j)} &= \frac{(n_1-2)\mathbf{S}_{1(i)} + (n_2-2)\mathbf{S}_{2(j)}}{n_1+n_2-4}, \end{aligned}$$

where $\mathbf{S}_{k,n_k}$ is the covariance matrix of the kth sample, $\mathbf{S}_{k(i,j)}$ is the covariance matrix of the kth sample excluding $\{\boldsymbol{X}_{ki}, \boldsymbol{X}_{kj}\}$ and $\mathbf{S}_{k(i)}$ is the covariance matrix of the kth sample excluding $\{\boldsymbol{X}_{ki}\}$ for each $k \in \{1,2\}$ and each $i, j \in \{1, \ldots, n_k\}$. Unfortunately, this test is not shift-invariant.

The above modified Hotelling's T^2 tests generally have very good performance for data from normal distributions, but deteriorate quickly when the data deviate from normality especially in high dimension situation, hence would perform extremely poorly for heavy-tailed distributions. For this reason, as mentioned in Oja (2010) and Wang et al. (2015), recently much effort has also been devoted to extending the nonparametric tests to the high-dimensional case. When using the nonparametric frameworks, an important issue to be addressed is to deal with the scatter matrix, which is generally not available due to its irreversibility. On this ground, for one-sample location problems, Wang et al. (2015) proposed a high-dimensional nonparametric multivariate test based on spatial signs, and Paindaveine et al. (2016) proposed a high-dimensional spatial sign test under a rotationally symmetric distribution. Although such high-dimensional sign- or rank-based methods for one-sample problems can be instructively enlightening for two-sample problems, they may not be simply extended to the two-sample problems unless the spatial medians would be estimated appropriately.

For example, Feng et al. (2016) proposed a scalar-invariant test statistic based on spatial signs (abbreviated as SS hereafter)

$$R_n = -\frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} U^T(\hat{\mathbb{D}}_{1,i}^{-1/2}(\boldsymbol{X}_{1i} - \hat{\boldsymbol{\mu}}_{2,j}))U(\hat{\mathbb{D}}_{2,j}^{-1/2}(\boldsymbol{X}_{2j} - \hat{\boldsymbol{\mu}}_{1,i})), \tag{6}$$

with the spatial sign function $U(\mathbf{x}) = \mathbf{x}/\|\mathbf{x}\| I(\mathbf{x} \neq \mathbf{0})$ for each $\mathbf{x} \in \mathbb{R}^{p_n}$, where $\hat{\boldsymbol{\mu}}_{k,i}$ and $\hat{\mathbb{D}}_{k,i}$ are the corresponding estimations of the location vectors and the scatter matrices respectively. Here $\hat{\boldsymbol{\mu}}_{k,i}$ and $\hat{\mathbb{D}}_{k,i}$ can be obtained by using the "leave-one-out" samples, $\{\boldsymbol{X}_{kj}\}_{j\neq i}$, with the following recursive algorithm:

(i) $\boldsymbol{\varepsilon}_{kj} \leftarrow \mathbb{D}_k^{-1/2}(\boldsymbol{X}_{kj} - \boldsymbol{\mu}_k)$, $j = 1, \ldots, i-1, i+1, \ldots, n_k$;

(ii) $\boldsymbol{\mu}_k \leftarrow \boldsymbol{\mu}_k + \frac{\mathbb{D}_k^{1/2} \sum_{j=1, j\neq i}^{n_k} U(\boldsymbol{\varepsilon}_{kj})}{\sum_{j=1, j\neq i}^{n_k} \|\boldsymbol{\varepsilon}_{kj}\|^{-1}}$ and

(iii) $\mathbb{D}_k \leftarrow p_n \mathbb{D}_k^{1/2} \mathrm{diag}\{n_k^{-1} \sum_{j=1, j\neq i}^{n_k} U(\boldsymbol{\varepsilon}_{kj}) U(\boldsymbol{\varepsilon}_{kj})^T\} \mathbb{D}_k^{1/2}$.

Unfortunately, as a result of the additional bias caused by estimating the location parameters, this method can only allow the dimension of variables to be at most the square of the sample sizes. In addition, Chakraborty et al. (2017) proposed a two-sample spatial rank test (abbreviated as CC hereafter):

$$C_n = n_1^{-2} n_2^{-2} \sum_{i=1}^{n_1} \sum_{j\neq i}^{n_1} \sum_{s=1}^{n_2} \sum_{l\neq s}^{n_2} U(\boldsymbol{X}_{1i} - \boldsymbol{X}_{2s})^T U(\boldsymbol{X}_{1j} - \boldsymbol{X}_{2l}), \tag{7}$$

which can deal with ultra-high-dimensional data, but is not invariant under scalar transformations. These motivate us to establish a new spatial rank testing method for the high-dimensional data, which could avoid estimating the location parameters in the construction of test statistic, hence be available for ultra-high-dimensional data. Essentially, the proposed high-dimensional spatial rank test is a reworking of C_n (CC) (Chakraborty et al., 2017) by embedding the process of re-scaling and leveraging the leave-out strategy in Feng et al. (2016) to remove bias, which can be invariant under scalar transformations and just has the power to deal with ultra-high-dimensional data, allowing the dimension to grow almost exponentially with the sample sizes.

Specifically, we first estimate the scale of each variable by spatial-rank-based procedures; then on this basis we construct the high-dimensional spatial rank test via the leave-out method as in Feng and Sun (2015) and Feng et al. (2016). By embedding the estimated scales in the test statistic, we aim to treat all the variables in a "fair" way. Unlike the spatial-sign-based methods, there is no need for the proposed test to estimate the location parameters for spatial ranks. As a result, the bias could be avoided, which makes it available even as the dimension grows almost exponential with the sample sizes.

The rest of the paper is organized as follows. We introduce the proposed high-dimensional spatial rank test for high-dimensional data in Section 2, and then establish its theoretical framework, including the limiting null distribution, the power performance under the local alternative and the asymptotic relative efficiency in Section 3. The numerical performance of the proposed test is demonstrated in Section 4, and two real data analyses are demonstrated in Section 5. Finally, we conclude this paper in Section 6 and relegate the technical proofs to Supplementary Material.

2. The proposed spatial rank test

The proposed spatial rank test in this paper is based on the assumption of equal scatter matrix, i.e. $\boldsymbol{\Sigma}_n = \boldsymbol{\Sigma}_{1n} = \boldsymbol{\Sigma}_{2n}$. The testing problem in situation of unequal scatter matrices is not the focus of this paper, which was studied by some other literature, such as Lix et al. (2005). In this paper, we will propose the test statistic and establish its asymptotic properties based on the equal scatter matrix assumption, while just investigating the performance of the proposed test in situation of unequal scatter matrices via some numerical results in later section.

The proposed test statistic is a reworking of C_n (CC) in (7) proposed by Chakraborty et al. (2017). The underlying defect in C_n (CC) is that it is invariant only to the "homogeneous positive scale transformations", saying $\boldsymbol{X}_{ki} \to d\boldsymbol{X}_{ki} + c$, where $d > 0$ is a scalar and c is a vector of constants. The power of CC overly depends on the underlying variance magnitudes. However, in practical applications, different variables may have completely different practical meanings and scales. This motivates us to utilize individual information for each variable in a relatively "fair" way, rather than scaling all the variables in the same magnitude. To accomplish this, one practical solution is to estimate the diagonal matrix of $\boldsymbol{\Sigma}_n$, $\mathbf{D}_n$, which is a diagonal matrix with the same diagonal elements as in $\boldsymbol{\Sigma}_n$.

Specifically, for the kth sample $\boldsymbol{X}_k = (\boldsymbol{X}_{k1}, \ldots, \boldsymbol{X}_{kn_k})^T$, $k \in \{1, 2\}$, we first obtain the diagonal matrix $\mathcal{D}_k$, which satisfies

$$\mathrm{diag}\left\{ \frac{1}{n_k} \sum_{i=1}^{n_k} R_{\mathcal{D}_k^{-1/2}\boldsymbol{X}_k}(\mathcal{D}_k^{-1/2}\boldsymbol{X}_{ki}) R_{\mathcal{D}_k^{-1/2}\boldsymbol{X}_k}(\mathcal{D}_k^{-1/2}\boldsymbol{X}_{ki}) \right\} \propto \mathbf{I}_{p_n},$$

where $\mathcal{D}_k^{-1/2}\boldsymbol{X}_k \triangleq (\mathcal{D}_k^{-1/2}\boldsymbol{X}_{k1}, \ldots, \mathcal{D}_k^{-1/2}\boldsymbol{X}_{kn_k})$. Here $R_{\boldsymbol{Y}}(\mathbf{y}) = n^{-1} \sum_{i=1}^{n} U(\mathbf{y} - \boldsymbol{Y}_i)$ denotes the spatial rank function of any dataset $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)$. The spatial ranks $R_Y(\boldsymbol{Y}_i)$ are automatically centered, i.e. $\sum_{i=1}^{n} R_Y(\boldsymbol{Y}_i) = 0$. Let $RCOV(\boldsymbol{Y}) = n^{-1} \sum_{i=1}^{n} R_Y(\boldsymbol{Y}_i) R_Y(\boldsymbol{Y}_i)^T$ denote the spatial rank covariance matrix of $\boldsymbol{Y}$. To accomplish this, for each $k \in \{1, 2\}$, by taking

the sample variance matrix of $\boldsymbol{X}_k$ as an initial value, we can iteratively update $\mathcal{D}_k$ in the following way:

$$\mathcal{D}_k \leftarrow \mathcal{D}_k^{1/2}\mathrm{diag}\left\{\frac{1}{n_k}\sum_{i=1}^{n_k} R_{\mathcal{D}_k^{-1/2}\boldsymbol{X}_k}(\mathcal{D}_k^{-1/2}\boldsymbol{X}_{ki})R_{\mathcal{D}_k^{-1/2}\boldsymbol{X}_k}(\mathcal{D}_k^{-1/2}\boldsymbol{X}_{ki})\right\}\mathcal{D}_k^{1/2},$$

$$\mathcal{D}_k \leftarrow \frac{p_n}{\mathrm{tr}(\mathcal{D}_k)}\mathcal{D}_k.$$

The above iterative algorithm stops at the $m+1$th iteration, if the F-norm of the difference between the $\mathcal{D}_k$ obtained in the $m+1$th iteration and that obtained in the mth iteration is less than 0.0001. After the algorithm is stopped, we denote the resulting $\mathcal{D}_k$ as $\hat{\mathbf{D}}_{k,n_k}$, for each $k \in \{1, 2\}$. Up to now, although there is still no theoretical proof on the convergence of the above iterative algorithm, even for the low-dimensional case, it is often very effective for practical use. It should be noted that this algorithm is always convergent in the later numerical studies.

Then, we consider the following test statistic:

$$G_n = \frac{1}{n_1(n_1-1)}\frac{1}{n_2(n_2-1)}\sum_{i\neq j}^{n_1}\sum_{i\neq j}^{n_1}\sum_{s\neq l}^{n_2}\sum_{s\neq l}^{n_2} U\big(\hat{\mathbf{D}}_n^{-1/2}(\boldsymbol{X}_{1i}-\boldsymbol{X}_{2s})\big)^T U\big(\hat{\mathbf{D}}_n^{-1/2}(\boldsymbol{X}_{1j}-\boldsymbol{X}_{2l})\big), \tag{8}$$

where $\hat{\mathbf{D}}_n \triangleq \frac{n_1}{n}\hat{\mathbf{D}}_{1,n_1} + \frac{n_2}{n}\hat{\mathbf{D}}_{2,n_2}$. It is not difficult to see that G_n still has a drawback. Due to the dependence between $(\boldsymbol{X}_{1i}, \boldsymbol{X}_{1j}, \boldsymbol{X}_{2s}, \boldsymbol{X}_{2l})$ and $\hat{\mathbf{D}}_n$, a non-negligible bias occurs, which is infeasible to correct, because the bias term depends on $\boldsymbol{\Sigma}_n$ as in Feng et al. (2015). This motivates us to use the leave-out method as in Feng and Sun (2015) to remove the bias.

Therefore, we propose using the following high-dimensional spatial rank test statistic (abbreviated as SR hereafter):

$$T_n = \frac{1}{n_1(n_1-1)}\frac{1}{n_2(n_2-1)}\sum_{i\neq j}^{n_1}\sum_{i\neq j}^{n_1}\sum_{s\neq l}^{n_2}\sum_{s\neq l}^{n_2} U(\hat{\mathbf{D}}_{n(i,j,s,l)}^{-1/2}(\boldsymbol{X}_{1i}-\boldsymbol{X}_{2s}))^T U(\hat{\mathbf{D}}_{n(i,j,s,l)}^{-1/2}(\boldsymbol{X}_{1j}-\boldsymbol{X}_{2l})), \tag{9}$$

where $\hat{\mathbf{D}}_{n(i,j,s,l)} = \frac{n_1}{n}\hat{\mathbf{D}}_{1,n_1(i,j)} + \frac{n_2}{n}\hat{\mathbf{D}}_{2,n_2(s,l)}$. Here $\hat{\mathbf{D}}_{1,n_1(i,j)}$ and $\hat{\mathbf{D}}_{2,n_2(s,l)}$ denote the corresponding estimations of $\mathbf{D}_n$, by applying the leave-two-out method to $\{\boldsymbol{X}_{1a}\}_{a\neq i,j}$ and $\{\boldsymbol{X}_{2b}\}_{b\neq s,l}$, respectively. Obviously, if $\boldsymbol{\mu}_{1n} \neq \boldsymbol{\mu}_{2n}$, $U(\hat{\mathbf{D}}_{n(i,j,k,l)}^{-1/2}(\boldsymbol{X}_{1i}-\boldsymbol{X}_{2k}))$ will deviate from zero hence the value of T_n will not be too small. Accordingly, we may reject the null hypothesis if T_n has a large enough value.

3. Theoretical results

We impose the following conditions for the asymptotic analysis of the proposed test: as $n, p_n \to \infty$,

(C1) $n_1/n \to \kappa \in (0, 1)$;
(C2) $\mathrm{tr}(\mathbf{R}_n^4) = o(\mathrm{tr}^2(\mathbf{R}_n^2))$ where $\mathbf{R}_n \triangleq \mathbf{D}_n^{-1/2}\boldsymbol{\Sigma}_n\mathbf{D}_n^{-1/2}$ and
(C3) $\mathrm{tr}(\mathbf{R}_n^2) - p_n = o(n^{-1}p_n^2)$ and $\log(p_n) = o(n)$.

Condition (C2) is the same as Condition (4) used in Park and Ayyala (2013). Condition (C3) is imposed to ensure the consistency of the diagonal matrix estimators. To control the difference between $\mathbf{D}_n^{-1/2}(\boldsymbol{X}_{ki} - \boldsymbol{\mu}_{kn})$ and $\boldsymbol{\varepsilon}_{ki} \triangleq \boldsymbol{\Sigma}_n{}^{-1/2}(\boldsymbol{X}_{ki} - \boldsymbol{\mu}_{kn})$ for each $k \in \{1, 2\}$ and each $i \in \{1, \ldots, n_k\}$, we need to restrict the correlation between the corresponding variables. To better understand Conditions (C2) and (C3), let $\lambda_{n,1}, \ldots, \lambda_{n,p_n}$ define all the eigenvalues of $\mathbf{R}_n$ and let $\nu_{n,t} = \sum_{i=1}^{p_n}\lambda_{n,i}^t$ for any positive integer t, then it can be concluded that Conditions (C2) and (C3) are equivalent to the following Conditions (C2′) and (C3′) respectively:

(C2′) $\nu_{n,4} = o(\nu_{n,2}^2)$ and
(C3′) $\nu_{n,2} - p_n = o(n^{-1}p_n^2)$ and $\log p_n = o(n)$.

In the special case where $\lambda_{n,1}, \ldots, \lambda_{n,p_n}$ are all bounded, Condition (C2) holds, because in this case $\nu_{n,4} = O(p_n)$ and $\nu_{n,2} = O(p_n)$. Further, in this case, by using Conditions (C2) and (C3) together, we can conclude that $p_n/n \to \infty$.

We then present the asymptotic null distribution of T_n under Conditions (C1)–(C3).

Theorem 1. *Under Conditions (C1)–(C3) and H_0, as $(p_n, n) \to \infty$,*

$$T_n/\sigma_n \xrightarrow{\mathcal{L}} N(0, 1), \tag{10}$$

where $\sigma_n^2 \triangleq \left(\frac{1}{2n_1(n_1-1)p_n{}^2} + \frac{1}{2n_2(n_2-1)p_n{}^2} + \frac{1}{n_1n_2p_n{}^2}\right)\mathrm{tr}(\mathbf{R}_n^2)$.

Note that we relegate the proof of Theorem 1 to the Supplementary Material, and the proofs of the other theorems or propositions are the same.

To estimate σ_n, we need to estimate $\mathrm{tr}(\mathbf{R}_n^2)$ first, whose estimation can be obtained by using one of the following three ratio-consistent estimators:

$$\widehat{\mathrm{tr}(\mathbf{R}_n^2)}_k = \frac{2{p_n}^2}{P_{n_k}^4}\sum^{k,*} U\left(\hat{\mathbf{D}}^{-1/2}_{k,n_k(i_1,i_2,i_3,i_4)}(\boldsymbol{X}_{ki_1}-\boldsymbol{X}_{ki_2})\right)^T U\left(\hat{\mathbf{D}}^{-1/2}_{k,n_k(i_1,i_2,i_3,i_4)}(\boldsymbol{X}_{ki_3}-\boldsymbol{X}_{ki_4})\right)$$
$$U\left(\hat{\mathbf{D}}^{-1/2}_{k,n_k(i_1,i_2,i_3,i_4)}(\boldsymbol{X}_{ki_3}-\boldsymbol{X}_{ki_2})\right)^T U\left(\hat{\mathbf{D}}^{-1/2}_{k,n_k(i_1,i_2,i_3,i_4)}(\boldsymbol{X}_{ki_1}-\boldsymbol{X}_{ki_4})\right),$$

for each $k \in \{1, 2\}$, and

$$\widehat{\mathrm{tr}(\mathbf{R}_n^2)}_3 = \frac{{p_n}^2}{n_1^2 n_2^2}\sum_{i_1\neq i_2}^{n_1}\sum_{i_3\neq i_4}^{n_2}\left(U\left(\hat{\mathbf{D}}^{-1/2}_{1,n_1(i_1,i_2)}(\boldsymbol{X}_{1i_1}-\boldsymbol{X}_{1i_2})\right)^T \boldsymbol{U}\left(\hat{\mathbf{D}}^{-1/2}_{2,n_2(i_3,i_4)}(\boldsymbol{X}_{2i_3}-\boldsymbol{X}_{2i_4})\right)\right)^2,$$

where $P^4_{n_k} \triangleq n_k(n_k-1)(n_k-2)(n_k-3)$, $\hat{\mathbf{D}}_{k,n_k(i_1,i_2,i_3,i_4)}$ are the estimators of $\mathbf{D}_n$ with the leave-four-out samples $\{\boldsymbol{X}_{ks}\}_{s\neq i_1,i_2,i_3,i_4}$ respectively and $\sum^{k,*}$ denotes the summation over $\{(i_1, i_2, i_3, i_4) \subseteq \{1, \ldots, n_k\} : i_1, i_2, i_3$ and i_4 are not equal to each other$\}$.

Then the ratio-consistency of $\widehat{\mathrm{tr}(\mathbf{R}_n^2)}_k$ for each $k \in \{1, 2, 3\}$ can be established as follows.

Proposition 1. *Under Condition (C1)–(C3), as* $(p_n, n) \to \infty$,

$$\frac{\widehat{\mathrm{tr}(\mathbf{R}_n^2)}_k}{\mathrm{tr}(\mathbf{R}_n^2)} \xrightarrow{p} 1, \quad k = 1, 2, 3. \tag{11}$$

Accordingly, we obtain the following ratio-consistent estimator of σ_n^2:

$$\hat{\sigma}_n^2 = \frac{1}{2n_1(n_1-1){p_n}^2}\widehat{\mathrm{tr}(\mathbf{R}_n^2)}_1 + \frac{1}{2n_2(n_2-1){p_n}^2}\widehat{\mathrm{tr}(\mathbf{R}_n^2)}_2 + \frac{1}{n_1 n_2 {p_n}^2}\widehat{\mathrm{tr}(\mathbf{R}_n^2)}_3.$$

On this ground, by taking $T_n/\hat{\sigma}_n$ as the test statistic, we will reject H_0 with the significance level α, once $T_n/\hat{\sigma}_n > z_\alpha$, where z_α is the upper α-quantile of $N(0, 1)$.

Next, we consider the asymptotic distribution of T_n under the alternative hypothesis. To achieve this, we need to use one more condition:

(C4) $c_{0n} = E(\|\mathbf{D}_n^{-1/2}(\boldsymbol{X}_{ki}-\boldsymbol{X}_{kj})\|^{-1})$ exists, $(\boldsymbol{\mu}_{1n}-\boldsymbol{\mu}_{2n})^T\mathbf{D}_n^{-1}(\boldsymbol{\mu}_{1n}-\boldsymbol{\mu}_{2n}) = O(c_{0n}^{-2}\sigma_n)$ and $(\boldsymbol{\mu}_{1n}-\boldsymbol{\mu}_{2n})^T\mathbf{D}_n^{-1/2}\mathbf{R}_n\mathbf{D}_n^{-1/2}(\boldsymbol{\mu}_{1n}-\boldsymbol{\mu}_{2n}) = o(npc_0^{-2}\sigma_n)$.

Condition (C4) constrains that the difference between $\boldsymbol{\mu}_{1n}$ and $\boldsymbol{\mu}_{2n}$ is small enough, which enables the variance of T_n to be asymptotically bounded by σ_n^2. In the special case where $\lambda_{n,1}, \ldots, \lambda_{n,p_n}$ are all bounded, Condition (C4) becomes $(\boldsymbol{\mu}_{1n}-\boldsymbol{\mu}_{2n})^T\mathbf{D}_n^{-1}(\boldsymbol{\mu}_{1n}-\boldsymbol{\mu}_{2n}) = O(n^{-1}p_n^{1/2})$ and $(\boldsymbol{\mu}_{1n}-\boldsymbol{\mu}_{2n})^T\mathbf{D}_n^{-1/2}\mathbf{R}_n\mathbf{D}_n^{-1/2}(\boldsymbol{\mu}_{1n}-\boldsymbol{\mu}_{2n}) = O(n^{-1}p_n^3)$, which implies that $\|\boldsymbol{\mu}_{1n}-\boldsymbol{\mu}_{2n}\|^2 = O(n^{-1}p_n^{1/2})$. Furthermore, if we let $\boldsymbol{\mu}_{1n}-\boldsymbol{\mu}_{2n} \triangleq (\delta_n, \ldots, \delta_n)$, then $\|\boldsymbol{\mu}_{1n}-\boldsymbol{\mu}_{2n}\|^2 = O(n^{-1}p_n^{1/2})$ will reduce to $\delta_n = O(n^{-1/2}p_n^{-1/4})$, which can be viewed as a high-dimensional version of the local alternative hypotheses.

Now, we are ready to present the explicit power expression of the proposed test.

Theorem 2. *Under Conditions (C1)–(C4), as* $(p_n, n) \to \infty$,

$$\frac{T_n - c_{0n}^2(\boldsymbol{\mu}_{1n}-\boldsymbol{\mu}_{2n})^T\mathbf{D}_n^{-1}(\boldsymbol{\mu}_{1n}-\boldsymbol{\mu}_{2n})}{\sigma_n} \xrightarrow{\mathcal{L}} N(0, 1). \tag{12}$$

Based on this, the asymptotic power of the proposed test (abbreviated as SR) is

$$\beta_{\mathrm{SR}} = \Phi\left(-z_\alpha + \frac{2c_{0n}^2 pn\kappa(1-\kappa)(\boldsymbol{\mu}_{1n}-\boldsymbol{\mu}_{2n})^T\mathbf{D}_n^{-1}(\boldsymbol{\mu}_{1n}-\boldsymbol{\mu}_{2n})}{\sqrt{2\mathrm{tr}(\mathbf{R}_n^2)}}\right). \tag{13}$$

Recall that under some specific conditions, Park and Ayyala (2013) have proved that the asymptotic power of P_n (PA) is

$$\beta_{\mathrm{PA}} = \Phi\left(-z_\alpha + \frac{np\kappa(1-\kappa)(\boldsymbol{\mu}_{1n}-\boldsymbol{\mu}_{2n})^T\mathbf{D}_n^{-1}(\boldsymbol{\mu}_{1n}-\boldsymbol{\mu}_{2n})}{E(\|\boldsymbol{\varepsilon}_{ki}\|^2)\sqrt{2\mathrm{tr}(\mathbf{R}_n^2)}}\right). \tag{14}$$

Table 1
Empirical size and power comparison at 5% significance when $p_n < n$, where S_n (TR) and T_n (SR) are the traditional spatial-rank-based test statistic proposed in Oja (2010) and the proposed test statistic in this paper, respectively.

	Size				Dense case				Sparse case			
(n_i, p_n)	(30,24)		(40,32)		(30,24)		(40,32)		(30,24)		(40,32)	
Set-up	S_n	T_n	S_n	T_n	S_n	T_n	S_n	T_n	S_n	T_n	S_n	T_n
(I)	0.016	0.063	0.013	0.054	0.27	0.63	0.41	0.96	0.42	0.58	0.63	0.96
(II)	0.007	0.065	0.018	0.052	0.27	0.64	0.43	0.98	0.42	0.58	0.69	0.96
(III)	0.014	0.064	0.008	0.057	0.19	0.50	0.31	0.82	0.29	0.45	0.46	0.81
(IV)	0.012	0.049	0.011	0.053	0.51	0.60	0.72	0.91	0.29	0.54	0.54	0.59
(V)	0.012	0.053	0.009	0.045	0.18	0.45	0.25	0.78	0.24	0.44	0.43	0.76
(VI)	0.014	0.043	0.013	0.057	0.10	0.79	0.14	0.86	0.26	0.68	0.44	0.77

The asymptotic relative efficiency (ARE) of T_n (SR) with respect to P_n (PA) is

$$\begin{aligned}\mathrm{ARE(SR,\,PA)} =& 2c_{0n}^2 E(\|\boldsymbol{\varepsilon}_{ki}\|^2)\\ \approx& 2\{E(\|\boldsymbol{\varepsilon}_{ki}-\boldsymbol{\varepsilon}_{kj}\|^{-1})\}^2 E(\|\boldsymbol{\varepsilon}_{ki}\|^2)\\ =& \{E(\|\boldsymbol{\varepsilon}_{ki}-\boldsymbol{\varepsilon}_{kj}\|^{-1})\}^2 E(\|\boldsymbol{\varepsilon}_{ki}-\boldsymbol{\varepsilon}_{kj}\|^2) \geq 1, \end{aligned} \tag{15}$$

by using the Cauchy inequality and the fact that $c_{0n} = E(\|\boldsymbol{\varepsilon}_{ki}-\boldsymbol{\varepsilon}_{kj}\|^{-1})(1+o(1))$ under Condition (C4) (see the proof of Theorem 1). Recall that from the definition of $\boldsymbol{\varepsilon}_{ki}$, it can be seen that $\boldsymbol{\varepsilon}_{ki}$ are i.i.d. zero-mean random vectors. In situation where $\|\boldsymbol{\varepsilon}_{ki}-\boldsymbol{\varepsilon}_{kj}\|^2/E(\|\boldsymbol{\varepsilon}_{ki}-\boldsymbol{\varepsilon}_{kj}\|^2) \to^p 1$, T_n (SR) has the same asymptotic efficiency as P_n (PA), otherwise has higher asymptotic efficiency than P_n (PA). Taking the standard multivariate t-distributions for example, for $\nu = 6, 5, 4, 3$, the ARE values are 1.22, 1.31, 1.48 and 1.98 respectively, which suggests that as ν becomes smaller, the distribution becomes more heavy-tailed, and then the ARE of T_n (SR) with respect to P_n (PA) becomes higher. This will be verified via some numerical results in the following section.

4. Numerical results

In this section, we report some numerical results to demonstrate the performance of the proposed test T_n (SR), with comparison to some commonly used two-sample tests, such as the traditional spatial-rank-based test S_n (TR) and those proposed in Chen et al. (2010), Feng et al. (2016), Park and Ayyala (2013) and Chakraborty et al. (2017), abbreviated as W_n (CQ), R_n (SS), P_n (PA), C_n (CC), respectively.

We consider the following commonly studied simulation set-ups:

(I) Multivariate normal distribution. $\boldsymbol{X}_{ki} \sim N(\boldsymbol{\mu}_{kn}, \mathbf{R}_n)$.
(II) Multivariate normal distribution with different component variances. $\boldsymbol{X}_{ki} \sim N(\boldsymbol{\mu}_{kn}, \boldsymbol{\Sigma}_n)$, where $\boldsymbol{\Sigma}_n = \mathbf{D}_n^{1/2}\mathbf{R}_n\mathbf{D}_n^{1/2}$, $\mathbf{D}_n = \mathrm{diag}\{d_1^2, \ldots, d_{p_n}^2\}$, $d_j^2 = 3$ for each $j \leq p_n/2$ and $d_j^2 = 1$, for each $j > p_n/2$.
(III) Multivariate t-distribution $t_{p_n,3}$. $\boldsymbol{X}_{ki}$ are generated from $t_{p_n,3}$ with $\boldsymbol{\Sigma}_n = \mathbf{R}_n$.
(IV) Multivariate t-distribution with different component variances. $\boldsymbol{X}_{ki}$ are generated from $t_{p_n,3}$ and d_j^2 are generated from χ_2^2.
(V) Multivariate mixture normal distribution $\mathrm{MN}_{p_n,\gamma,9}$. $\boldsymbol{X}_{ki}$ are generated from $\gamma f_{p_n}(\boldsymbol{\mu}_{kn}, \mathbf{R}_n) + (1-\gamma) f_{p_n}(\boldsymbol{\mu}_{kn}, 9\mathbf{R}_n)$, denoted by $\mathrm{MN}_{p_n,\gamma,9}$, where $f_{p_n}(\cdot;\cdot)$ is the density function of p_n-variate multivariate normal distribution and γ equals 0.8.
(VI) Multivariate skew t-distribution. $\boldsymbol{X}_{ki}$ are from $St_{p_n}(\boldsymbol{\mu}_{kn}, \mathbf{R}_{kn}, \boldsymbol{\alpha}, 3)$ (Azzalini and Capitanio, 2003) with $\boldsymbol{\alpha} = (1, \ldots, 1)$.

First, we consider the low-dimensional case with $p_n < n$ and compare T_n (SR) with S_n (TR). Here the common correlation matrix is set to be $\mathbf{R}_n = (0.5^{|i-j|})$. For power comparison, we consider the same configurations of H_1: $\eta = \|\mathbf{D}_n^{-1/2}(\boldsymbol{\mu}_{1n}-\boldsymbol{\mu}_{2n})\|^2/\sqrt{\mathrm{tr}(\mathbf{R}_n^2)} = 0.5$. Without loss of generality, under H_1, we fix $\boldsymbol{\mu}_{1n} = 0$ and then set up $\boldsymbol{\mu}_{2n}$ in the following way. Let $\boldsymbol{\mu}_{kn} = (\mu_{kn,1}, \ldots, \mu_{kn,p_n})^T$. The percentages of $\mu_{1n,l} = \mu_{2n,l}$ among $l \in \{1, \ldots, p_n\}$ for the sparse case and the dense case are chosen to be 95% and 50%, respectively. Further, for each percentage level, all the nonzero $\mu_{2n,l}$ are set to be equal. Two combinations of (n_k, p_n) are considered: (30, 24) and (40, 32). Table 1 reports the empirical size and the power results of these two tests, where all the numerical results are obtained based on 2500 replications as well as in the following numerical results reported in the remaining tables. The empirical sizes of S_n (TR) are significantly smaller than the nominal level, by contrast SR can better control the empirical sizes in general. Also, T_n (SR) is much more powerful than S_n (TR) in all the cases. These results are consistent with the previous conclusions in Bai and Saranadasa (1996), which suggests that in large p_n situation classical Mahalanobis distance may lose efficiency due to the contamination bias that is generated when estimating the covariance matrix. And, when $p_n/n \to c \in (0, 1)$, the estimation of the scatter matrix is singular hence cannot be inverted in the construction of the test statistic.

Then, we consider the high-dimensional case with $p_n > n$, and compare T_n (SR) with W_n (CQ), R_n (SS), P_n (PA) and C_n (CC). The sample sizes are chosen as $n_1 = n_2 = 20$, and the dimension has four choices, $p_n = 100, 200, 400, 800$.

Table 2
Empirical size and power comparison at 5% significance with equal scatter matrix, where W_n (CQ), R_n (SS), P_n (PA), C_n (CC), T_n (SR) are the test statistics proposed by Chen et al. (2010), Feng et al. (2016), Park and Ayyala (2013), Chakraborty et al. (2017) and this paper, respectively.

(n_i, p_n)	Size					Dense case					Sparse case				
	W_n	R_n	P_n	C_n	T_n	W_n	R_n	P_n	C_n	T_n	W_n	R_n	P_n	C_n	T_n
Set-up (I)															
(20,100)	0.069	0.048	0.050	0.061	0.063	0.83	0.77	0.77	0.84	0.82	0.79	0.73	0.74	0.75	0.79
(20,200)	0.053	0.034	0.033	0.052	0.050	0.87	0.81	0.83	0.87	0.86	0.86	0.81	0.82	0.86	0.87
(20,400)	0.055	0.025	0.041	0.045	0.047	0.90	0.80	0.85	0.89	0.90	0.89	0.80	0.85	0.91	0.90
(20,800)	0.054	0.012	0.040	0.032	0.045	0.92	0.81	0.89	0.91	0.92	0.91	0.79	0.88	0.88	0.91
Set-up (II)															
(20,100)	0.075	0.050	0.043	0.063	0.060	0.44	0.78	0.78	0.42	0.82	0.40	0.73	0.73	0.36	0.78
(20,200)	0.045	0.031	0.025	0.054	0.044	0.47	0.81	0.82	0.46	0.87	0.44	0.81	0.82	0.43	0.87
(20,400)	0.060	0.025	0.041	0.047	0.057	0.46	0.81	0.86	0.42	0.91	0.42	0.80	0.85	0.38	0.90
(20,800)	0.066	0.012	0.038	0.045	0.056	0.42	0.80	0.85	0.44	0.90	0.45	0.79	0.88	0.44	0.91
Set-up (III)															
(20,100)	0.058	0.043	0.033	0.065	0.062	0.42	0.64	0.29	0.65	0.62	0.37	0.59	0.29	0.57	0.58
(20,200)	0.042	0.030	0.015	0.056	0.047	0.44	0.67	0.34	0.70	0.69	0.41	0.63	0.27	0.64	0.63
(20,400)	0.061	0.020	0.039	0.048	0.057	0.44	0.63	0.32	0.69	0.67	0.41	0.63	0.31	0.70	0.66
(20,800)	0.044	0.007	0.037	0.027	0.053	0.44	0.61	0.32	0.70	0.72	0.44	0.60	0.32	0.68	0.71
Set-up (IV)															
(20,100)	0.058	0.043	0.033	0.098	0.063	0.11	0.68	0.33	0.12	0.67	0.07	0.70	0.33	0.16	0.69
(20,200)	0.075	0.030	0.017	0.087	0.044	0.09	0.73	0.33	0.10	0.73	0.13	0.65	0.28	0.16	0.65
(20,400)	0.064	0.020	0.039	0.078	0.057	0.10	0.64	0.30	0.14	0.69	0.09	0.62	0.29	0.14	0.66
(20,800)	0.057	0.007	0.037	0.065	0.052	0.10	0.60	0.32	0.08	0.73	0.09	0.62	0.32	0.10	0.72
Set-up (V)															
(20,100)	0.075	0.040	0.055	0.051	0.052	0.42	0.59	0.35	0.61	0.60	0.35	0.55	0.30	0.54	0.57
(20,200)	0.065	0.030	0.045	0.055	0.052	0.43	0.60	0.32	0.65	0.63	0.41	0.56	0.32	0.64	0.61
(20,400)	0.061	0.022	0.037	0.053	0.053	0.41	0.57	0.30	0.63	0.64	0.39	0.57	0.28	0.63	0.64
(20,800)	0.055	0.008	0.040	0.026	0.056	0.41	0.54	0.30	0.65	0.68	0.39	0.52	0.28	0.62	0.66
Set-up (VI)															
(20,100)	0.072	0.043	0.055	0.130	0.062	0.37	0.67	0.31	0.75	0.68	0.41	0.58	0.31	0.65	0.57
(20,200)	0.053	0.031	0.038	0.130	0.059	0.30	0.60	0.24	0.65	0.57	0.31	0.48	0.18	0.65	0.47
(20,400)	0.051	0.022	0.047	0.170	0.053	0.42	0.62	0.32	0.89	0.72	0.41	0.69	0.34	0.86	0.74
(20,800)	0.045	0.002	0.050	0.140	0.056	0.46	0.65	0.33	0.93	0.77	0.45	0.34	0.28	0.88	0.69

The other settings are all the same as the low dimensional case. Table 2 reports the empirical size and power results of these five tests under normal and non-normal set-ups. It can be seen that the empirical sizes obtained by W_n (CQ), P_n (PA) and T_n (SR) are generally close to the nominal level for all these set-ups. The empirical sizes of C_n (CC) test are much larger than the nominal level under set-ups (IV) and (VI), because the two tests were constructed under the equal variance assumption. On the other hand, the empirical sizes of SS are too conservative when p_n/n^2 is large. This result is predictable, because R_n (SS) can only allow the dimension p_n to be the square of the sample size n at most. As shown in Feng et al. (2016), when p_n/n^2 is large enough, there would be a non-negligible bias term in the R_n (SS) test statistic, because it includes the estimations of the location parameters. In contrast, demonstrated by these size results in Table 2, T_n (SR) has advantages in dealing with ultra-high-dimensional data, which can allow the dimension to grow almost exponentially with the sample sizes.

Next, according to Table 2, we compare the power performance of the listed methods under the above set-ups as follows.

(1) Under the normal case with equal component variances (Set-up (I)), the power performance of all these methods is generally similar.
(2) Under the normal case with different component variances (Set-up (II)), the power performance of P_n (PA), R_n (SS) and T_n (SR) is generally similar and better than W_n (CQ) and C_n (CC), while W_n (CQ) and C_n (CC) fail because they are not invariant under scalar transformations.
(3) Under the non-normal cases (Set-ups (III)–(VI)), most of the time R_n (SS), C_n (CC) and T_n (SR) have better power performance than P_n (PA) and W_n (CQ), since R_n (SS), C_n (CC) and T_n (SR) are nonparametric methods, which can handle heavy-tailed data. There are also some exceptions. For example, C_n (CC) fails for Set-up (IV) due to the different component variances. In addition, for Set-up (VI), C_n (CC) has larger empirical power values than the other four methods, which, however, does not mean that C_n (CC) performs well in such situation, as its empirical size is out of control.
(4) T_n (SR) generally outperforms the other four methods in power comparison, especially in situation of heavy-tail and different component variances. And, the higher the dimension is, the more obvious such advantage is. However, in

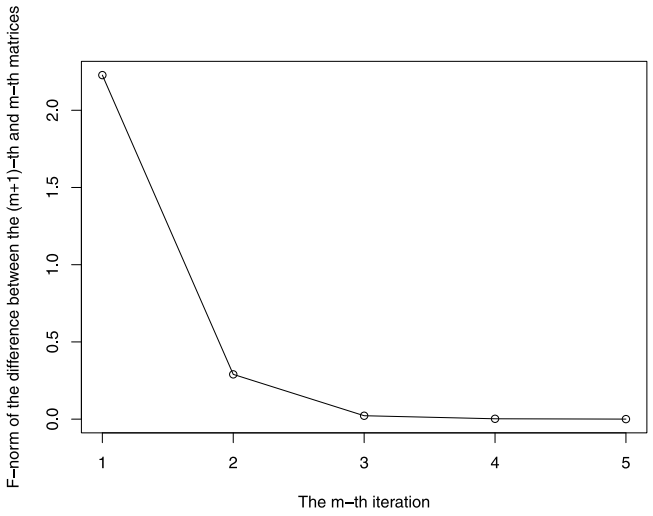

Fig. 1. Convergence path of the iterative algorithm that is employed to obtain $\hat{\mathbf{D}}_{k,n_k}$, under Set-up (III) with $n = 50$ and $p = 100$, where the horizontal axis corresponds to the iteration number m, while the vertical axis corresponds to the F-norm of the difference between the corresponding $\mathcal{D}_k$ in the $m+1$th and mth iterations of the algorithm.

Table 3
Empirical size and power comparison at 5% significance with unequal scatter matrices, where W_n (CQ), R_n (SS), P_n (PA), C_n (CC), T_n (SR) are the test statistics proposed by Chen et al. (2010), Feng et al. (2016), Park and Ayyala (2013), Chakraborty et al. (2017) and this paper, respectively.

Set-up	Size					Dense case					Sparse case				
	W_n	R_n	P_n	C_n	T_n	W_n	R_n	P_n	C_n	T_n	W_n	R_n	P_n	C_n	T_n
$(n_i, p_n) = (20, 200)$															
(I)	0.058	0.028	0.042	0.062	0.065	0.95	0.89	0.93	0.95	0.95	0.92	0.87	0.88	0.92	0.93
(II)	0.065	0.025	0.034	0.071	0.055	0.58	0.91	0.93	0.57	0.95	0.52	0.87	0.88	0.52	0.93
(III)	0.044	0.022	0.025	0.045	0.039	0.55	0.79	0.43	0.78	0.77	0.49	0.74	0.36	0.77	0.72
(IV)	0.066	0.021	0.026	0.082	0.046	0.65	0.80	0.45	0.10	0.79	0.11	0.75	0.40	0.18	0.74
(V)	0.061	0.023	0.052	0.059	0.051	0.53	0.73	0.39	0.77	0.74	0.49	0.72	0.38	0.75	0.75
(VI)	0.047	0.053	0.027	0.140	0.051	0.36	0.64	0.26	0.77	0.62	0.31	0.56	0.22	0.67	0.54
$(n_i, p_n) = (20, 800)$															
(I)	0.072	0.011	0.056	0.048	0.059	0.98	0.91	0.97	0.99	0.98	0.98	0.90	0.98	0.95	0.99
(II)	0.069	0.015	0.052	0.047	0.058	0.64	0.91	0.97	0.60	0.98	0.59	0.90	0.98	0.55	0.99
(III)	0.073	0.013	0.046	0.039	0.046	0.56	0.73	0.36	0.80	0.85	0.52	0.68	0.41	0.79	0.80
(IV)	0.062	0.012	0.041	0.068	0.041	0.70	0.70	0.40	0.09	0.87	0.13	0.58	0.36	0.12	0.74
(V)	0.044	0.015	0.046	0.041	0.048	0.52	0.60	0.35	0.76	0.81	0.51	0.60	0.39	0.74	0.77
(VI)	0.041	0.003	0.032	0.150	0.043	0.45	0.59	0.32	0.88	0.75	0.41	0.60	0.28	0.84	0.65

the case of lower dimension and equal component variances, the power advantage of T_n (SR) is not obvious enough, and sometimes the power of T_n (SR) is a little worse than that of some of the remaining methods.

Then, we consider the case of unequal scatter matrices, where $\mathbf{\Sigma}_{1n} \neq \mathbf{\Sigma}_{2n}$. Let $\mathbf{R}_{1n} = (0.5^{|i-j|})$ and $\mathbf{R}_{2n} = \mathbf{I}_{p_n}$, and the other settings are all the same as in the above equal scatter matrix case except that $\eta = \|\mathbf{D}_n^{-1/2}(\boldsymbol{\mu}_{1n} - \boldsymbol{\mu}_{2n})\|^2/\sqrt{\mathrm{tr}(\mathbf{R}_1^2)} = 0.5$. We consider two choices of (n_k, p_n): (20, 200), (20, 800). The corresponding results are summarized in Table 3, which suggests that T_n (SR) can well control the empirical sizes in this case and the rest results are very similar to Table 2. Interestingly, T_n (SR) has very good performance in the unequal scatter matrices case, though it is constructed based on the equal scatter matrix assumption.

To further investigate the application range of T_n (SR), we consider four additional simulation set-ups with different correlation structures and distributions. The moving average model is considered:

$$X_{kij} = \|\boldsymbol{\rho}_k\|^{-1}(\rho_{k1}Z_{kij} + \rho_{k2}Z_{ki(j+1)} + \cdots + \rho_{kT_k}Z_{ki(j+T_k-1)}) + \mu_{kn,j},$$

for $k = 1, 2$, $i = 1, \ldots, n_k$ and $j = 1, \ldots, p_n$ where $\boldsymbol{\rho}_k = (\rho_{k1}, \ldots, \rho_{kT_k})^T$, $\boldsymbol{\mu}_{kn} = (\mu_{kn,1}, \ldots, \mu_{kn,p_n})^T$ are generated in the same way as in the above elliptical cases, and $\{Z_{kij}\}$ are i.i.d. random variables. Consider four set-ups for the innovation $\{Z_{kij}\}$:

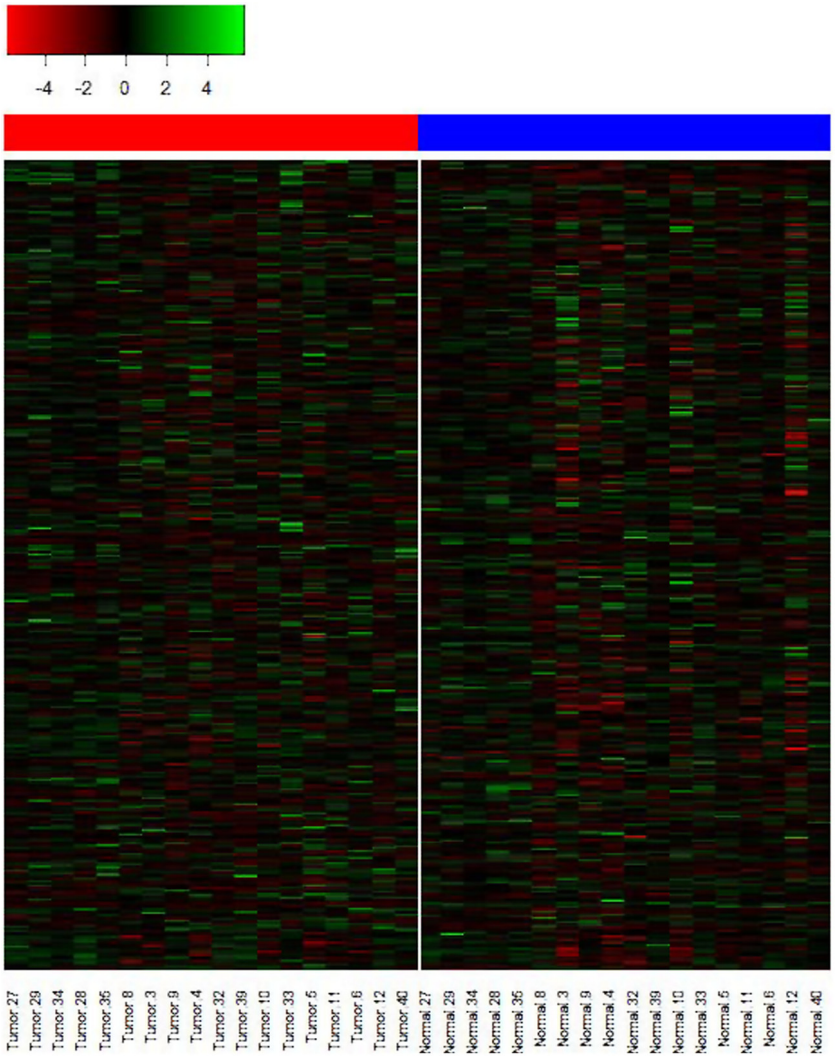

Fig. 2. The heat map of the two samples with all the 7457 genes via row scaling, where samples are arranged horizontally, the red bar corresponds to the tumor group and the blue bar corresponds to the normal group.

(VII) All the Z_{kij} are from $N(0, 1)$;
(VIII) The Z_{kij} for all the $j \in \{1, \ldots, p_n/2\}$ are from centralized Gamma(8,1), and the others are from $N(0, 1)$;
(IX) All the Z_{kij} are from t_3;
(X) All the Z_{kij} are from $0.8N(0, 1) + 0.2N(0, 9)$.

Note that the coefficients $\{\rho_{kl}\}_{l=1}^{T_k}$ are generated independently from $U(2, 3)$ only once, and then are used for all the above four set-ups. The correlations among X_{kij} and X_{kil} are determined by $|j - l|$ and T_k. We consider the "full dependence" for the first sample and the "2-dependence" for the second sample, i.e. $T_1 = p_n$ and $T_2 = 3$, to ensure that $\boldsymbol{X}_{ki}$ have different covariances for different $k \in \{1, 2\}$. For simplicity, set $\eta = \|\boldsymbol{\mu}_{1n} - \boldsymbol{\mu}_{2n}\|^2/\sqrt{\mathrm{tr}(\boldsymbol{\Lambda}_1^2) + \mathrm{tr}(\boldsymbol{\Lambda}_2^2)} = 0.1$ where $\boldsymbol{\Lambda}_k$ is the covariance matrix of $\boldsymbol{X}_{ki}$ and $(n_k, p_n) = (20, 200), (20, 800)$ for each $k \in \{1, 2\}$. Table 4 reports the simulation results for these four moving average models, which are non-elliptical and skewed distributions. It suggests that T_n (SR) generally outperforms all the other methods listed above for these four set-ups, which partly highlights the robustness of the proposed method.

Finally, recall that we employ an iterative algorithm to obtain $\hat{\mathbf{D}}_{k,n_k}$ in Section 2, while establishing the proposed testing statistic. Although the convergence of the algorithm has not been proved theoretically, the algorithm is always convergent in our numerical studies. Fig. 1 presents its convergence path under certain set-up ($n = 50$, $p = 100$, Set-up (III)), where the horizontal axis corresponds to the iteration number m, while the vertical axis corresponds to the F-norm of

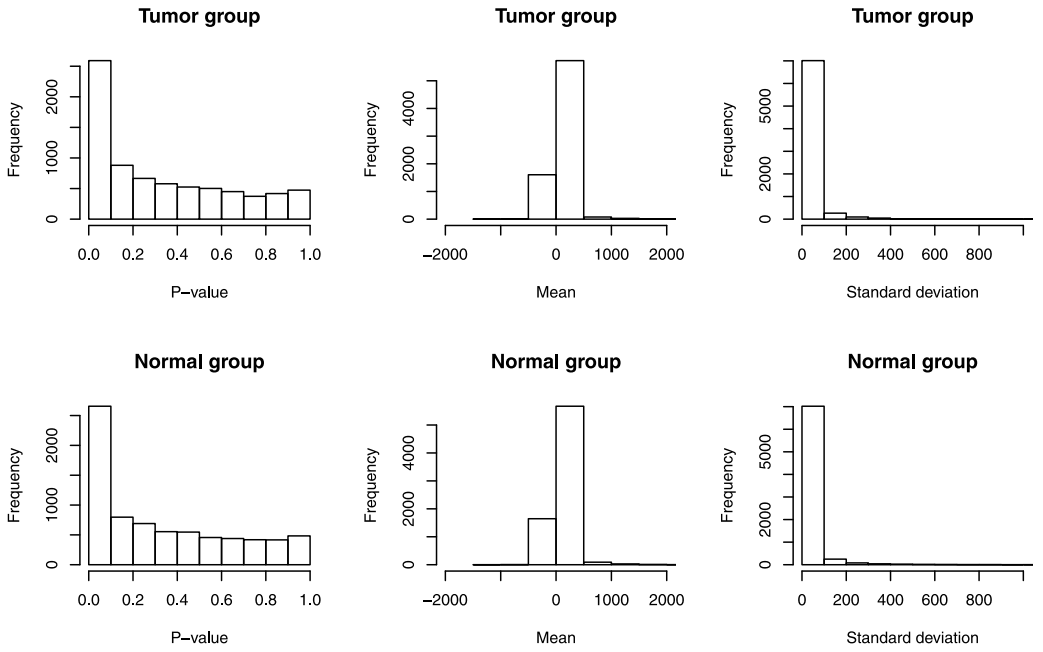

Fig. 3. The histograms of the p-values of the normality tests, the sample means and the sample standard deviations, for the tumor group and the normal group respectively.

Table 4
Empirical size and power comparison at 5% significance with moving average model, where W_n (CQ), R_n (SS), P_n (PA), C_n (CC), T_n (SR) are the test statistics proposed by Chen et al. (2010), Feng et al. (2016), Park and Ayyala (2013), Chakraborty et al. (2017) and this paper, respectively.

Set-up	Size					Dense case					Sparse case				
	W_n	R_n	P_n	C_n	T_n	W_n	R_n	P_n	C_n	T_n	W_n	R_n	P_n	C_n	T_n
$(n_i, p_n) = (20, 200)$															
(VII)	0.067	0.058	0.064	0.071	0.059	0.32	0.28	0.26	0.22	0.31	0.43	0.42	0.36	0.47	0.50
(VIII)	0.039	0.041	0.061	0.026	0.048	0.23	0.32	0.31	0.20	0.33	0.38	0.72	0.69	0.39	0.76
(IX)	0.063	0.037	0.071	0.045	0.051	0.41	0.34	0.35	0.36	0.41	0.42	0.44	0.34	0.51	0.53
(X)	0.065	0.054	0.063	0.061	0.057	0.29	0.27	0.26	0.24	0.33	0.34	0.36	0.28	0.41	0.47
$(n_i, p_n) = (20, 800)$															
(VII)	0.066	0.048	0.067	0.043	0.058	0.32	0.27	0.26	0.24	0.28	0.33	0.35	0.22	0.36	0.40
(VIII)	0.041	0.040	0.063	0.062	0.047	0.37	0.43	0.38	0.31	0.46	0.38	0.72	0.68	0.41	0.74
(IX)	0.059	0.039	0.078	0.085	0.054	0.39	0.32	0.31	0.36	0.42	0.41	0.29	0.31	0.44	0.49
(X)	0.064	0.034	0.062	0.023	0.051	0.34	0.32	0.28	0.29	0.34	0.31	0.31	0.23	0.33	0.35

the difference between the corresponding $\mathcal{D}_k$ in the $m+1$th and mth iterations of the algorithm. It can be seen that the algorithm converges very fast. We note that such fast convergence also occurs under the remaining set-ups.

In summary, all the above numerical results demonstrate that the proposed test method is efficient under a wide range of distributions. These results suggest that the proposed method generally has more advantages than the existing methods in comparison, especially in heavy-tailed and ultra-high-dimensional situations.

5. Real data analyses

5.1. Carcinoma dataset

In this subsection, we first apply the proposed testing method to a carcinoma dataset, which consists of 7457 genes measurements for 18 patients on both tumor and normal tissues. The dataset was previously studied by Notterman et al. (2001) and William et al. (2016), and can be freely downloaded at the following web site: "http://genomics-pubs.princeton.edu/oncology". Below we will apply the proposed method to test the hypothesis that the tissues in the tumor group have the same expression levels, in terms of these 7457 genes, as those in the normal group, where the dimension 7457 is eventually larger than the square of the sample sizes 324. Fig. 2 plots the heat map of the tumor and normal groups with all the 7457 genes via row scaling. From Fig. 2, it is difficult to see the difference between the two

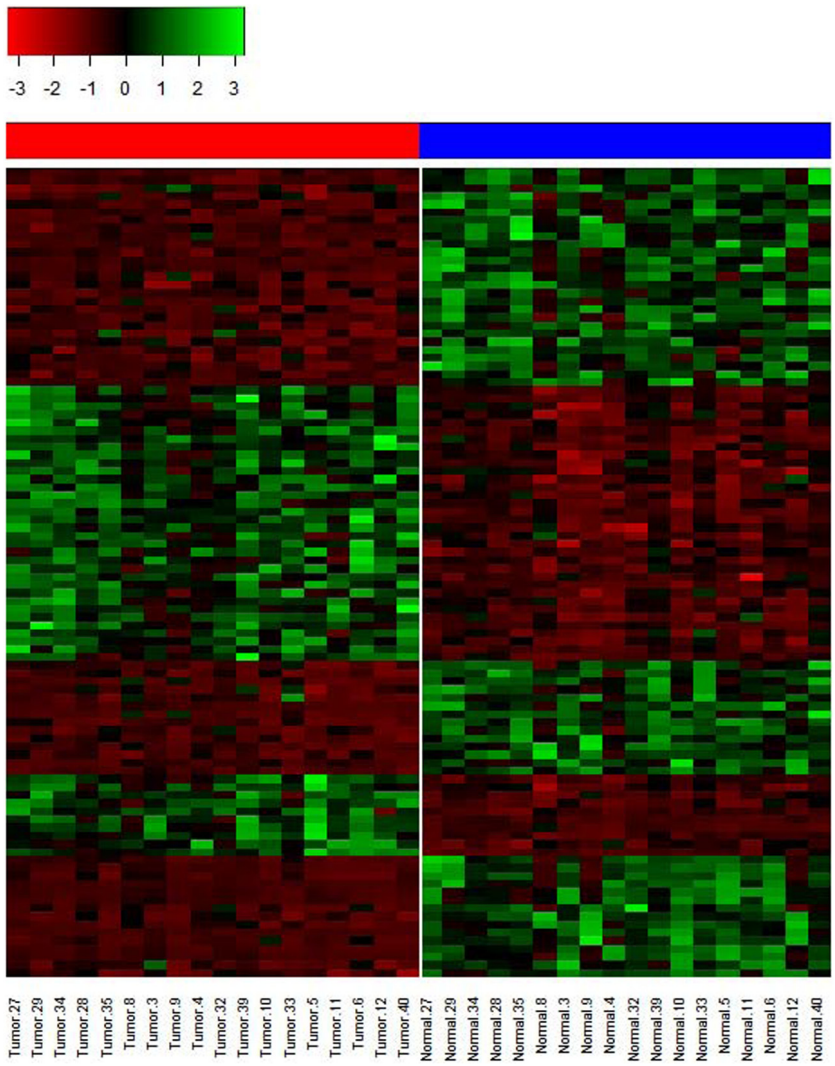

Fig. 4. The heat map of the two samples with a selected group of genes via row scaling.

groups with the naked eye. It should be noted that the 18 normal tissues and 18 tumor tissues are paired data collected from 18 patients, where the dependence between normal and tumor tissues may exist. To demonstrate the proposed method, this dependency is ignored in this application.

Note that the original focus of the research on this dataset is to identify the genes with significant differences between the tumor and normal groups, as in Notterman et al. (2001) and William et al. (2016). In fact, some genetic differences between the two groups have been well recognized by many researchers, while we just use this dataset to test whether the proposed method is available to discover such well-known differences.

First, the normal distribution was tested for each gene, using the Shapiro–Wilk test. The left-most two panels of Fig. 3 present the histograms of the p-values of the normality tests for the tumor group and the normal group respectively, which indicate that for a large number of genes the expression data are non-normal. In fact, under the significance level of 0.05, the overall rejection rates of all the normality tests are 26.64% and 27.56% for the tumor group and the normal group respectively. This motivates us to use a non-parametric approach for testing the above hypothesis.

The rest four panels of Fig. 3 indicate that there exist some genes with very high values of sample mean and sample variance in terms of expression. We see that the sample means vary largely for each of the two groups and recall that the dimension is eventually larger than the square of the sample sizes, which raises a concern that using a spatial sign-based approach may lead to an uncontrollable bias. Hence, in theory, a spatial rank-based approach seems more appropriate for this dataset. Furthermore, the sample standard deviations also vary largely, which suggests that a scalar-invariant approach may be necessary.

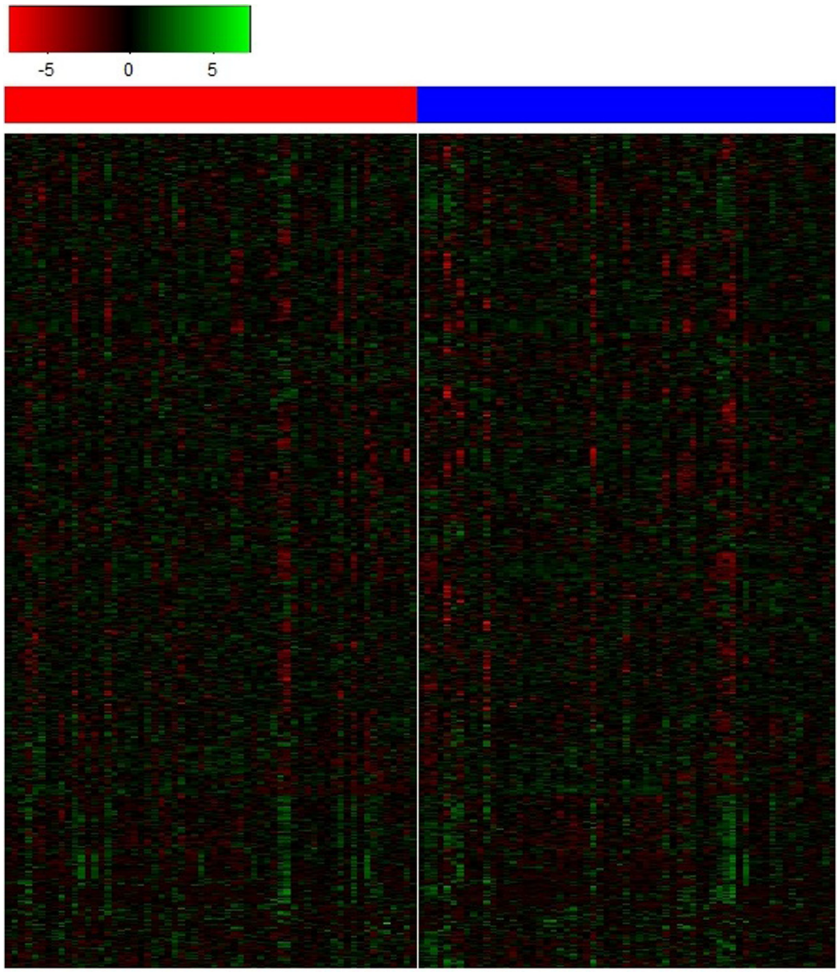

Fig. 5. The heat map of the two samples with all the 672 genes via row scaling, where samples are arranged horizontally, the red bar corresponds to the testing sample and the blue bar corresponds to the training sample.

Based on the above reasons, we apply the proposed T_n (SR) test to this dataset. The test statistic and p-value of the T_n (SR) test are 14.8 and 0.000 respectively, hence the null hypothesis is rejected, which suggests that the gene expression levels of the tumor group are significantly different from the normal group. In particular, we see that the main difference lies in a small number of the genes. For example, we first get a sequence of p-values by applying the univariate t-test to the two samples for each gene, then basing on the 100 genes with the lowest p-values we plot the heat map of the two samples in Fig. 4, where the difference between the two samples is very clear. In addition, the (test statistic, p-value) results of the remaining testing methods are listed as follows: R_n (SS) (7.062, 0.000); W_n (CQ) (12.130, 0.000); P_n (PA) (5.814, 0.000); C_n (CC) (13.781, 0.000), where the null hypothesis is also rejected by these remaining methods.

5.2. Non-small-cell lung cancer dataset

In many fields of medical research and biology, random sample split has been frequently employed. Samples need to be split into a training set and an independent testing set, where the former is used to carry on the statistical inference and the latter to evaluate its performance. The samples corresponding to the two sets should have similar distributions. Otherwise, using the testing sample to measure the performance of the inference based on the training sample may lead to wrong conclusions. On this ground, testing the equality of the distributions or some distribution parameters between the training sample and the testing sample is a necessary step of random sample split.

In this subsection, we will use the proposed method to test the equality of the location/mean parameters between the training and testing samples from a non-small-cell lung cancer dataset. This dataset consists the expression of 672 genes that are associated with invasive activity for the frozen specimens of lung-cancer tissue from 125 randomly selected patients, which was originally studied by Chen et al. (2007) and recently studied by Emura et al. (2019). The dataset can

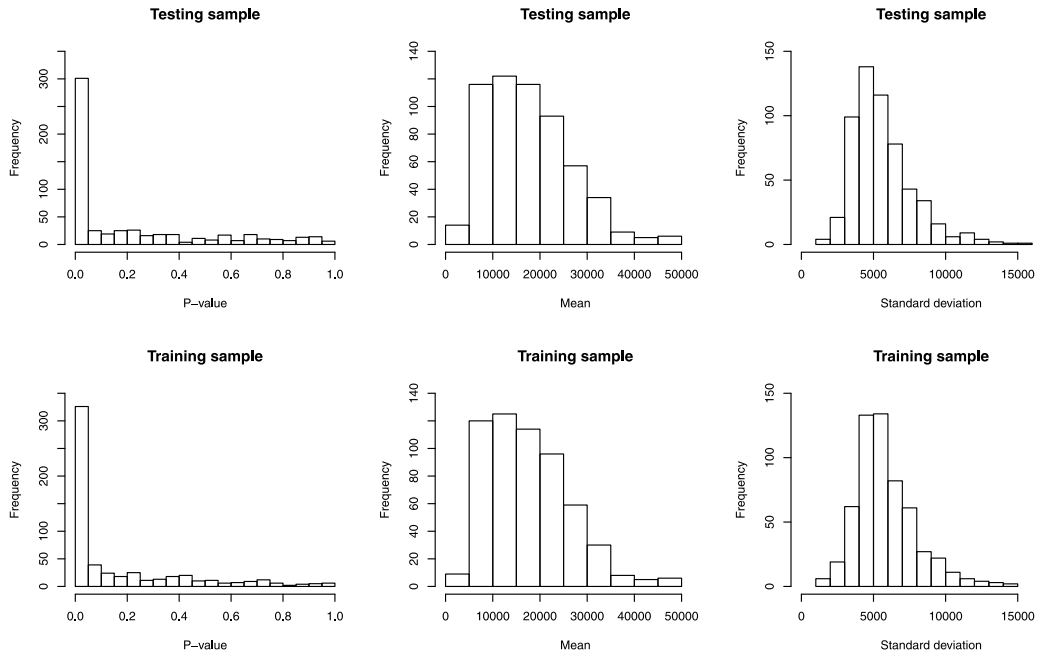

Fig. 6. The histograms of the p-values of the normality tests, the sample means and the sample standard deviations, for the testing sample and the training sample, respectively.

be freely downloaded from ''https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE4882''. Here the training sample with 63 observations and the testing sample with 62 observations are divided in Chen et al. (2007). Fig. 5 plots the heat map of the two samples with all the 672 genes via row scaling, where it is difficult to see the difference between the two samples.

The normal distribution was tested for each gene, using the Shapiro–Wilk test. The left-most two panels of Fig. 6 present the histograms of the p-values of the normality tests for the testing sample and the training sample respectively, which indicate that for a large number of genes the expression data are non-normal. Under the significance level of 0.05, the overall rejection rates of all the normality tests are 52.62% and 56.99% for the testing sample and the training sample respectively. This motivates us to use a non-parametric approach for testing the above hypothesis. The rest four panels of Fig. 6 indicate that the distributions of these sample means and sample standard deviations are not centralized and there exist some genes with very high values of sample mean and sample variance in terms of expression. In particular, the sample standard deviations vary largely, which suggests that a scalar-invariant approach may be necessary.

Based on the above reasons, we apply the proposed T_n (SR) test to this dataset. The test statistic and p-value of the T_n (SR) test are 1.42 and 0.078 respectively, hence the null hypothesis, i.e. the equality of the location/mean parameters between the training and testing samples, is not rejected, which suggests that it is reasonable for Chen et al. (2007) to use such training and testing sets in terms of parameters of the 672 genes between them. It is also reasonable for Emura et al. (2019) to employ the same training and testing partition on some subset of the 672 genes. In addition, the (test statistic, p-value) results of the remaining testing methods are listed as follows: R_n (SS) (0.001, 0.499); W_n (CQ) (-0.033, 0.513); P_n (PA) (0.061, 0.475); C_n (CC) (1.244, 0.107), where the null hypothesis is not rejected for all these remaining methods.

6. Conclusion

In this paper, we propose a novel high-dimensional spatial rank test for two-sample location problem. In comparison with many existing high-dimensional two-sample location testing procedures, the proposed test is highly competitive in efficiency, even in heavy-tailed, non-elliptical and ultra-high dimensional situations. Both theoretical and numerical investigations demonstrate the superiority of the proposed method. As for our future work, we will consider a high-dimensional weighted spatial rank test for two-sample location problem, where each summand in the test statistic will be assigned with a weight related to the corresponding observations. Hopefully, such a weighted test may have better power performance under certain local alternative hypothesis.

Acknowledgments

This work was supported by NSFC, China grants 11501092, 11571068, 11631003, the Special Fund for Key Laboratories of Jilin Province, China grant 20190201285JC, the Fundamental Research Funds for the Central Universities grant 2412017BJ002, the Key Laboratory of Applied Statistics of MOE (KLAS), China grants 130026507 and 130028612.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csda.2019.106889.

References

Azzalini, A., Capitanio, A., 2003. Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. J. R. Stat. Soc. Ser. B Stat. Methodol. 65 (2), 367–389.
Bai, Z., Saranadasa, H., 1996. Effect of high dimension: By an example of a two sample problem. Statist. Sinica 6 (2), 311–329.
Chakraborty, A., Chaudhuri, P., et al., 2017. Tests for high-dimensional data based on means, spatial signs and spatial ranks. Ann. Statist. 45 (2), 771–799.
Chen, S.X., Qin, Y.-L., et al., 2010. A two-sample test for high-dimensional data with applications to gene-set testing. Ann. Statist. 38 (2), 808–835.
Chen, H.Y., Yu, S., Chen, C., Chang, G.C., Chen, C.Y., Yuan, A., Cheng, C.L., Wang, C.H., Terng, H.J., Kao, S.F., et al., 2007. A five-gene signature and clinical outcome in non-small-cell lung Cancer. New Engl. J. Med. 356 (1), 11–20.
Emura, T., Matsui, S., Chen, H., 2019. Compound.cox: Univariate feature selection and compound covariate for predicting survival. Comput. Methods Programs Biomed. 168, 21–37.
Fan, J., Lv, J., 2008. Sure independence screening for ultrahigh dimensional feature space. J. R. Stat. Soc. Ser. B 70 (5), 849–911.
Feng, L., Sun, F., 2015. A note on high-dimensional two-sample test. Statist. Probab. Lett. 105, 29–36.
Feng, L., Zou, C., Wang, Z., 2016. Multivariate-sign-based high-dimensional tests for the two-sample location problem. J. Amer. Statist. Assoc. 111 (514), 721–735.
Feng, L., Zou, C., Wang, Z., Zhu, L., 2015. Two-sample Behrens-Fisher problem for high-dimensional data. Statist. Sinica 25 (4), 1297–1312.
Lix, L.M., Keselman, H., Hinds, A.M., 2005. Robust tests for the multivariate Behrens-Fisher problem.. Comput. Methods Programs Biomed. 77 (2), 129–139.
Notterman, D.A., Alon, U., Sierk, A.J., Levine, A.J., 2001. Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays. Cancer Res. 61 (7), 3124–3130.
Oja, H., 2010. Multivariate Nonparametric Methods with R: An approach based on spatial signs and ranks. Springer Science & Business Media.
Paindaveine, D., Verdebout, T., et al., 2016. On high-dimensional sign tests. Bernoulli 22 (3), 1745–1769.
Park, J., Ayyala, D.N., 2013. A test for the mean vector in large dimension and small samples. J. Statist. Plann. Inference 143 (5), 929–943.
Srivastava, M.S., Du, M., 2008. A test for the mean vector with fewer observations than the dimension. J. Multivariate Anal. 99 (3), 386–402.
Srivastava, M.S., Katayama, S., Kano, Y., 2013. A two sample test in high dimensional data. J. Multivariate Anal. 114, 349–358.
Wang, L., Peng, B., Li, R., 2015. A high-dimensional nonparametric multivariate test for mean vector. J. Amer. Statist. Assoc. 110 (512), 1658–1669.
William, C.I., Timothy, H., Mclain, A.C., 2016. Bayesian nonparametric multiple testing. Comput. Statist. Data Anal. 101, 64–79.