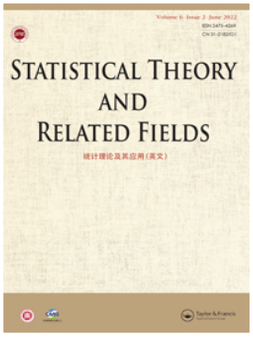

Statistical Theory and Related Fields

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tstf20

High-dimensional proportionality test of two covariance matrices and its application to gene expression data

Long Feng, Xiaoxu Zhang & Binghui Liu

To cite this article: Long Feng, Xiaoxu Zhang & Binghui Liu (2022) High-dimensional proportionality test of two covariance matrices and its application to gene expression data, Statistical Theory and Related Fields, 6:2, 161-174, DOI: 10.1080/24754269.2021.1984373

To link to this article: https://doi.org/10.1080/24754269.2021.1984373

Published online: 06 Oct 2021.

STATISTICAL THEORY AND RELATED FIELDS
2022, VOL. 6, NO. 2, 161–174
https://doi.org/10.1080/24754269.2021.1984373

High-dimensional proportionality test of two covariance matrices and its application to gene expression data

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

[a]School of Statistics and Data Science, LPMC and KLMDASR, Nankai University, Tianjin, People's Republic of China; [b]School of Mathematics and Statistics and KLAS, Northeast Normal University, Changchun, People's Republic of China

ABSTRACT

With the development of modern science and technology, more and more high-dimensional data appear in the application fields. Since the high dimension can potentially increase the complexity of the covariance structure, comparing the covariance matrices among populations is strongly motivated in high-dimensional data analysis. In this article, we consider the proportionality test of two high-dimensional covariance matrices, where the data dimension is potentially much larger than the sample sizes, or even larger than the squares of the sample sizes. We devise a novel high-dimensional spatial rank test that has much-improved power than many existing popular tests, especially for the data generated from some heavy-tailed distributions. The asymptotic normality of the proposed test statistics is established under the family of elliptically symmetric distributions, which is a more general distribution family than the normal distribution family, including numerous commonly used heavy-tailed distributions. Extensive numerical experiments demonstrate the superiority of the proposed test in terms of both empirical size and power. Then, a real data analysis demonstrates the practicability of the proposed test for high-dimensional gene expression data.

ARTICLE HISTORY
Received 4 November 2020
Revised 26 August 2021
Accepted 9 September 2021

KEYWORDS
Covariance matrices; elliptically symmetric distributions; high dimension test; proportionality; spatial rank

1. Introduction

High-dimensional data are nowadays more and more common in bioinformatics, material science, astronomy and other application fields, as data collection technology rapidly evolves (Bühlmann & van de Geer, 2011). However, due to limited resources available to replicate observations, the sample sizes are usually much smaller than the dimension, which makes most traditional statistical approaches no longer appropriate. Under such an embarrassing background, scientists in many application fields urgently need powerful approaches to gather the greatest scientific insight from data. Testing equality of the distributions of two populations is a crucial problem in high-dimensional statistics, which is extremely complex and far more challenging than that for fixed-dimensional data. Due to this extreme complexity, it is usually replaced by a simpler problem, i.e. testing equality of some numerical characteristics, such as means and covariances, of the two populations, which is very useful but much easier to implement.

There is already a large number of literature on detecting the difference between the means of two high-dimensional populations, such as Bai and Saranadasa (1996), Chen and Qinm (2010), and Feng et al. (2016), to name just a few. In contrast, there are much fewer studies on high-dimensional covariance matrix test of two high-dimensional populations. Hence, in this article, we focus on comparing the covariance matrices among two populations, which is strongly motivated for high-dimensional data, as high data dimensions can potentially increase the complexity of the covariance structure (Li & Chen, 2012). In particular, we consider the testing problem of the proportionality of two high-dimensional covariance matrices, which investigates the simplest heteroscedasticity of the population covariance matrices (Xu et al., 2014). It is often a preparation procedure before the case–control analysis of genomic data. Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be two p-dimensional populations with the mean vectors $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and the covariance matrices $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2$, respectively. The proportionality test of two population covariance matrices is formulated as follows:

$$H_0 : \boldsymbol{\Sigma}_1 = c\boldsymbol{\Sigma}_2 \quad \text{versus} \quad H_1 : \boldsymbol{\Sigma}_1 \neq c\boldsymbol{\Sigma}_2, \qquad (1)$$

where c is an unknown scalar.

The proportionality testing problem in (1) has been widely studied in various areas, such as in discriminant analysis and principal component analysis (Flury & Riedwyl, 1988; Schott, 1991), and there is a lot of early literature on its methodological researches, such as Eriksen (1987), Federer (1951), Flury (1986),

CONTACT Binghui Liu liubh100@nenu.edu.cn

Kim (1971), Rao (1983), and Schott (1999). For example, the most traditional test statistic is

$$(n_1+n_2)\sum_{j=1}^{p}\log(\hat{\lambda}_j)-n_1\log(|\hat{\mathbf{\Sigma}}_1|)$$
$$+n_2(p\log(\hat{c}-\log(|\hat{\mathbf{\Sigma}}_2|)))\xrightarrow{d}\chi^2_{(p^2+p-2)/2},$$

where $(\hat{\lambda}_j,\hat{c})$ are obtained by an iterative algorithm proposed in Flury (1986) and $\hat{\mathbf{\Sigma}}_1,\hat{\mathbf{\Sigma}}_2$ are the corresponding sample covariance matrices, respectively. These researches are constructed based on the classical limit theorems, assuming that the sample sizes tend to infinity and the dimension is fixed, hence have difficulties to analyse the high-dimensional data, where the dimension is much larger than the sample sizes. To alleviate such difficulties, Xu et al. (2014) proposed to use a pseudo-likelihood ratio test by extending the traditional likelihood ratio test with the statistic $p\log(p^{-1}\operatorname{tr}(\hat{\mathbf{\Sigma}}_1\hat{\mathbf{\Sigma}}_2^{-1}))-\log|\hat{\mathbf{\Sigma}}_1\hat{\mathbf{\Sigma}}_2^{-1}|$, which allows the dimension to increase proportionally with each sample size; furthermore, Liu et al. (2014) proposed an improved method, which allows the dimension to be larger than one of the sample sizes. In addition, for the special case of $c\equiv 1$ in (1), Li and Chen (2012) proposed a test statistic

$$T_{LC}=A_{n_1}+A_{n_2}-2C_{n_1,n_2},$$

where

$$A_{n_i}=\frac{1}{n_i(n_i-1)}\sum_{k\neq l}(\boldsymbol{X}_{ik}^\top\boldsymbol{X}_{il})^2$$
$$-\frac{2}{n_i(n_i-1)(n_i-2)}\sum_{j,k,l\text{ not equal}}\boldsymbol{X}_{ik}^\top\boldsymbol{X}_{ij}\boldsymbol{X}_{ij}^\top\boldsymbol{X}_{il}$$
$$+\frac{1}{n_i(n_i-1)(n_i-2)(n_i-3)}$$
$$\sum_{j,k,l,t\text{ not equal}}\boldsymbol{X}_{ik}^\top\boldsymbol{X}_{ij}\boldsymbol{X}_{it}^\top\boldsymbol{X}_{il},$$

$$C_{n_1,n_2}=\frac{1}{n_1n_2}\sum_{k=1}^{n_1}\sum_{l=1}^{n_2}(\boldsymbol{X}_{1k}^\top\boldsymbol{X}_{2l})^2$$
$$-\frac{1}{n_1n_2(n_1-1)}\sum_{k\neq l}^{n_1}\sum_{j=1}^{n_2}\boldsymbol{X}_{1k}^\top\boldsymbol{X}_{2j}\boldsymbol{X}_{2j}^\top\boldsymbol{X}_{1l}$$
$$-\frac{1}{n_1n_2(n_2-1)}\sum_{k\neq l}^{n_2}\sum_{j=1}^{n_1}\boldsymbol{X}_{2k}^\top\boldsymbol{X}_{1j}\boldsymbol{X}_{1j}^\top\boldsymbol{X}_{2l}$$
$$+\frac{1}{n_1n_2(n_1-1)(n_2-1)}$$
$$\times\sum_{k\neq t}^{n_1}\sum_{j\neq l}^{n_2}\boldsymbol{X}_{1k}^\top\boldsymbol{X}_{2j}\boldsymbol{X}_{1t}^\top\boldsymbol{X}_{2l}.$$

As mentioned in Li and Chen (2012), T_{LC} is an unbiased estimation of $\operatorname{tr}\{(\mathbf{\Sigma}_1-\mathbf{\Sigma}_2)^2\}$. Despite some progress, there are also drawbacks: first, these methods may have extremely poor performance for heavy-tailed distributions; second, the sample covariance matrices, which need to be inverted in the construction of the test statistic, are singular when the dimension is larger than both of the sample sizes.

To overcome these two drawbacks, more attention has been paid to nonparametric testing methods based on the multivariate sign or rank. Just recently, for testing the proportionality of two high-dimensional covariance matrices, Cheng et al. (2018) proposed to use a test procedure based on the multivariate sign and demonstrated its good performance in high-dimensional data analysis, especially for the heavy-tailed distributions. Recall that for fixed-dimensional data, the multivariate sign and rank are widely used to construct robust tests (Oja, 2010). However, most of these tests cannot be effective for high-dimensional data. Therefore, many researches extend the traditional multivariate sign- or rank-based testing methods to the high-dimension data, such as Feng and Sun (2016) and Wang et al. (2015) for one-sample problems; Feng et al. (2016) for two-sample problems; Feng and Liu (2017) and Zou et al. (2014) for sphericity testing problems. These researches clearly demonstrate the advantages of the high-dimensional multivariate sign- or rank-based methods in high-dimensional and heavy-tailed cases.

Unfortunately, due to the bias caused by estimating the location parameters, the test procedure based on the multivariate sign can only allow the dimension to be the squares of the sample sizes at most (Cheng et al., 2018), which makes the test procedure too restrictive for various practical applications, hence greatly affects the validity of the test procedure. For example, in genomic data analysis, genomic data typically carry thousands of dimensions for measurements on the genome, where the dimension can be much larger than the squares of the sample sizes. Therefore, it is very urgent to develop a new method to deal with the proportionality testing problem in (1) for the high-dimensional data, where the dimension is much higher than the squares of the sample sizes. This is the motivation and intention of this article.

The rest of the article is organized as follows. In Section 2, we introduce the proposed high-dimensional spatial rank test and establish its asymptotic normality under the elliptically symmetric populations. Then, we demonstrate the numerical performance of the proposed test in Sections 3, followed by a real data analysis in Section 4. Finally, we conclude this article in Section 5 and relegate the technical proofs to Appendix.

2. Method

2.1. The proposed test

A p-dimensional random vector $\boldsymbol{Z}$ is said to follow an elliptically symmetric distribution, denoted by

$\mathcal{E}_p(\boldsymbol{\mu}, \boldsymbol{\Lambda}, F_\xi)$, if it has the following stochastic representation:

$$\boldsymbol{Z} = \boldsymbol{\mu} + \xi \mathbf{A}\mathbf{U},$$

where $\boldsymbol{\mu}$ is the p-dimensional mean vector, ξ is a nonnegative random variable, F_ξ is the cumulative distribution function of ξ, $\mathbf{U}$ is independent of ξ and is uniformly distributed on the unit sphere $\mathcal{R}^p$ and $\mathbf{A}$ is a deterministic $p \times p$-dimensional matrix satisfying $\mathbf{A}\mathbf{A}^{\mathrm{T}} = \boldsymbol{\Lambda}$ with $\mathrm{tr}(\boldsymbol{\Lambda}) = 1$. It is known that the covariance matrix $\boldsymbol{\Sigma}$ and shape matrix $\boldsymbol{\Lambda}$ of the elliptical symmetric population $\boldsymbol{Z}$ will satisfy the equation $\boldsymbol{\Sigma} = p^{-1}E(\xi^2)\boldsymbol{\Lambda}$.

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{n_1}$ and $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{n_2}$ denote the samples of two p-dimensional random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$, which are generated from the two independent elliptically symmetric populations $\mathcal{E}_p(\boldsymbol{\mu}_1, \boldsymbol{\Lambda}_1, F_{\xi_1})$ and $\mathcal{E}_p(\boldsymbol{\mu}_2, \boldsymbol{\Lambda}_2, F_{\xi_2})$, respectively. From Section 3.1 in Magyar and Tyler (2014), it is known that $\boldsymbol{\Sigma}_i$ and $\S_i$ have the same eigenvectors for each $i \in \{1, 2\}$ under the assumption of elliptically symmetric distribution. Also, from Equation 3.9 in Magyar and Tyler (2014), it is known that when the eigenvalues of the covariance matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are proportional, the spatial sign covariance matrices $\S_1$ and $\S_2$ have the same eigenvalues. Theorem 1 in Cheng et al. (2018) showed that when $\S_1$ and $\S_2$ have the same eigenvalues, the eigenvalues of $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are proportional. Hence, the hypotheses in (1) are equivalent to the following hypotheses:

$$H_0 : \S_1 = \S_2 \quad \text{versus} \quad H_1 : \S_1 \neq \S_2, \tag{2}$$

where $\S_1 = \mathbb{E}\{U(\boldsymbol{X} - \boldsymbol{\mu}_1)U(\boldsymbol{X} - \boldsymbol{\mu}_1)^{\mathrm{T}}\}$, $\S_2 = \mathbb{E}\{U(\boldsymbol{Y} - \boldsymbol{\mu}_2)U(\boldsymbol{Y} - \boldsymbol{\mu}_2)^{\mathrm{T}}\}$ are the spatial sign covariance matrices of $\boldsymbol{X}$, $\boldsymbol{Y}$, respectively, and $U(\mathbf{z}) = \frac{\mathbf{z}}{\|\mathbf{z}\|} I(\mathbf{z} \neq \mathbf{0})$ for each $\mathbf{z} \in \mathcal{R}^p$ is the spatial sign function with $\|\cdot\|$ denoting the L_2-norm and $I(\cdot)$ denoting the indicator function. On this ground, Cheng et al. (2018) suggested to use a test statistics based on the square Frobenius norm of $\S_1 - \S_2$, i.e. $\mathrm{tr}\{(\S_1 - \S_2)^2\}$.

The proposed spatial rank test in this article is also based on the square Frobenius norm of $\S_1 - \S_2$, which is a high-dimensional extension of Kendall's tau test for the hypotheses in (2) (Oja, 2010). Specifically, the test statistic is

$$\begin{aligned} T_{\mathrm{HT}} = {} & \frac{p}{n_1(n_1-1)(n_1-2)(n_1-3)} \\ & \times \sum\nolimits^{*} \{U(\boldsymbol{X}_i - \boldsymbol{X}_j)^{\mathrm{T}} U(\boldsymbol{X}_k - \boldsymbol{X}_l)\}^2 \\ & + \frac{p}{n_2(n_2-1)(n_2-2)(n_2-3)} \\ & \times \sum\nolimits^{*} \{U(\boldsymbol{Y}_i - \boldsymbol{Y}_j)^{\mathrm{T}} U(\boldsymbol{Y}_k - \boldsymbol{Y}_l)\}^2 \\ & - \frac{2p}{n_1(n_1-1)n_2(n_2-1)} \\ & \times \sum_{i=1}^{n_1}\sum_{j\neq i}^{n_1}\sum_{k=1}^{n_2}\sum_{l\neq k}^{n_2} \{U(\boldsymbol{X}_i - \boldsymbol{X}_j)^{\mathrm{T}} U(\boldsymbol{Y}_k - \boldsymbol{Y}_l)\}^2, \end{aligned} \tag{3}$$

where $\sum^{*}$ denotes summation over distinct indexes $\{i, j, k, l\} \subseteq \{1, \ldots, n_1\}$ or $\{1, \ldots, n_2\}$. Note that recently many developed versions of Kendall's tau test are frequently used on many related issues (Barber & Kolar, 2018; Cai & Zhang, 2016; Han et al., 2017; Leung & Drton, 2018).

In deriving the asymptotic properties of T_{HT}, we impose the following two conditions used in Cheng et al. (2018):

(C1) $n_1/(n_1 + n_2) \to \kappa \in (0, 1)$ as $\min\{n_1, n_2\} \to \infty$;

(C2) $\mathrm{tr}(\boldsymbol{\Lambda}_i\boldsymbol{\Lambda}_j\boldsymbol{\Lambda}_k\boldsymbol{\Lambda}_l) = o\{\mathrm{tr}(\boldsymbol{\Lambda}_i\boldsymbol{\Lambda}_j)\,\mathrm{tr}(\boldsymbol{\Lambda}_k\boldsymbol{\Lambda}_l)\}$ for $i, j, k, l \in \{1, 2\}$.

Note that: (1) Condition (C1) is a commonly used condition in high-dimensional two sample testing problems; (2) Condition (C2) is similar to Condition (A2) in Li and Chen (2012); (3) If all the eigenvalues of $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are bounded, Condition (C2) holds.

Remark 2.1: Note that the above Conditions (C1) and (C2) do not contain any restriction on p and n_1, n_2, since such restriction is not needed to control the following terms:

$$\begin{aligned} & \sum_{i,j,k,l} \{U(\boldsymbol{X}_i - \boldsymbol{X}_j)^{\mathrm{T}} U(\boldsymbol{X}_k - \boldsymbol{X}_l)\}^2 \\ & \quad - \sum\nolimits^{*} \{U(\boldsymbol{X}_i - \boldsymbol{X}_j)^{\mathrm{T}} U(\boldsymbol{X}_k - \boldsymbol{X}_l)\}^2, \\ & \sum_{i,j,k,l} \{U(\boldsymbol{Y}_i - \boldsymbol{Y}_j)^{\mathrm{T}} U(\boldsymbol{Y}_k - \boldsymbol{Y}_l)\}^2 \\ & \quad - \sum\nolimits^{*} \{U(\boldsymbol{Y}_i - \boldsymbol{Y}_j)^{\mathrm{T}} U(\boldsymbol{Y}_k - \boldsymbol{Y}_l)\}^2, \\ & \sum_{i=1}^{n_1}\sum_{j=1}^{n_1}\sum_{k=1}^{n_2}\sum_{l=1}^{n_2} \{U(\boldsymbol{X}_i - \boldsymbol{X}_j)^{\mathrm{T}} U(\boldsymbol{Y}_k - \boldsymbol{Y}_l)\}^2 \\ & \quad - \sum_{i=1}^{n_1}\sum_{j\neq i}^{n_1}\sum_{k=1}^{n_2}\sum_{l\neq k}^{n_2} \{U(\boldsymbol{X}_i - \boldsymbol{X}_j)^{\mathrm{T}} U(\boldsymbol{Y}_k - \boldsymbol{Y}_l)\}^2, \end{aligned}$$

which have been removed from T_{HT}. That is to say, we remove all the items that include at least one pair of identical vectors, such as $\{U(\boldsymbol{X}_i - \boldsymbol{X}_j)^{\mathrm{T}} U(\boldsymbol{X}_i - \boldsymbol{X}_j)\}^2$, $\{U(\boldsymbol{X}_i - \boldsymbol{X}_j)^{\mathrm{T}} U(\boldsymbol{X}_i - \boldsymbol{X}_l)\}^2$ and so on. Such type of strategy was previously used in Chen and Qinm (2010). By removing the terms $\sum_i \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_i$ and $\sum_k \boldsymbol{Y}_k^{\mathrm{T}}\boldsymbol{Y}_k$ from the test statistic proposed by Chen and Qinm (2010), no restriction on p, n_1 and n_2 is needed.

Under the above two conditions, the limiting null distribution of T_{HT} is given in the following theorem.

Theorem 2.1: *Under Conditions (C1), (C2) and H_0, as $n_1, n_2, p \to \infty$,*

$$\sigma_{0,n}^{-1} T_{\mathrm{HT}} \xrightarrow{d} N(0,1),$$

where $\sigma_{0,n}^2 = 4(n_1^{-1} + n_2^{-1})^2 (p+2)^{-2} \operatorname{tr}^2(\boldsymbol{\Lambda}^2)$ with $\boldsymbol{\Lambda} = \boldsymbol{\Lambda}_1 = \boldsymbol{\Lambda}_2$.

Moreover, we obtain the limiting distribution of T_{HT} under H_1.

Theorem 2.2: *Under Conditions (C1), (C2) and H_1, as $n_1, n_2, p \to \infty$,*

$$\sigma_n^{-1}[T_{\mathrm{HT}} - p \operatorname{tr}\{(\S_1 - \S_2)^2\}] \xrightarrow{d} N(0,1),$$

where

$$\begin{aligned}
\sigma_n^2 = {} & \frac{4}{n_1(n_1-1)} \frac{\operatorname{tr}^2(\boldsymbol{\Lambda}_1^2)}{(p+2)^2} + \frac{8}{n_1} \frac{p \operatorname{tr}(\boldsymbol{\Lambda}_1^4) - \operatorname{tr}^2(\boldsymbol{\Lambda}_1^2)}{p^2(p+2)} \\
& \times \frac{4}{n_2(n_2-1)} \frac{\operatorname{tr}^2(\boldsymbol{\Lambda}_2^2)}{(p+2)^2} + \frac{8}{n_2} \frac{p \operatorname{tr}(\boldsymbol{\Lambda}_2^4) - \operatorname{tr}^2(\boldsymbol{\Lambda}_2^2)}{p^2(p+2)} \\
& + \frac{8}{n_1 n_2} \frac{\operatorname{tr}^2(\boldsymbol{\Lambda}_1\boldsymbol{\Lambda}_2)}{(p+2)^2} + \left(\frac{8}{n_1} + \frac{8}{n_2}\right) \\
& \times \frac{p \operatorname{tr}(\boldsymbol{\Lambda}_1\boldsymbol{\Lambda}_2)^2 - \operatorname{tr}^2(\boldsymbol{\Lambda}_1\boldsymbol{\Lambda}_2)}{p^2(p+2)} \\
& - \frac{16}{n_1} \frac{p \operatorname{tr}(\boldsymbol{\Lambda}_1^3\boldsymbol{\Lambda}_2) - \operatorname{tr}(\boldsymbol{\Lambda}_1\boldsymbol{\Lambda}_2) \operatorname{tr}(\boldsymbol{\Lambda}_1^2)}{p^2(p+2)} \\
& - \frac{16}{n_2} \frac{p \operatorname{tr}(\boldsymbol{\Lambda}_2^3\boldsymbol{\Lambda}_1) - \operatorname{tr}(\boldsymbol{\Lambda}_1\boldsymbol{\Lambda}_2) \operatorname{tr}(\boldsymbol{\Lambda}_2^2)}{p^2(p+2)}.
\end{aligned}$$

Due to the fact that $p \operatorname{tr}(\S_l^2) = p^{-1} \operatorname{tr}(\boldsymbol{\Lambda}_l^2)\{1 + o(1)\}$ for $l = 1, 2$ obtained by Cheng et al. (2018), we propose to use the following estimator of $\sigma_{0,n}^2$:

$$\begin{aligned}
\hat{\sigma}_{0,n}^2 = {} & 4(n_1^{-1} + n_2^{-1})^2 (n_1 + n_2)^{-1} (p+2)^{-2} \\
& \times p^2 (n_1 \mathbf{A}_1 + n_2 \mathbf{A}_2),
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{A}_1 = {} & \frac{1}{n_1(n_1-1)(n_1-2)(n_1-3)} \\
& \times \sum\nolimits^{*} \{U(\boldsymbol{X}_i - \boldsymbol{X}_j)^{\mathrm{T}} U(\boldsymbol{X}_k - \boldsymbol{X}_l)\}^2, \\
\mathbf{A}_2 = {} & \frac{1}{n_2(n_2-1)(n_2-2)(n_2-3)} \\
& \times \sum\nolimits^{*} \{U(\boldsymbol{Y}_i - \boldsymbol{Y}_j)^{\mathrm{T}} U(\boldsymbol{Y}_k - \boldsymbol{Y}_l)\}^2.
\end{aligned}$$

As presented by the following proposition, $\hat{\sigma}_{0,n}^2$ is a consistent estimator of $\sigma_{0,n}^2$ under H_0.

Proposition 2.1: *Under Conditions (C1), (C2) and H_0, $\hat{\sigma}_{0,n}^2 / \sigma_{0,n}^2 \to 1$.*

Therefore, the proposed test with a nominal α level of significance rejects H_0 if $T_{\mathrm{HT}} \geq z_\alpha \hat{\sigma}_{0,n}$, where z_α is the upper α-quantile of $N(0,1)$. The asymptotic power function of T_{HT} is

$$\begin{aligned}
& \beta_{n_1,n_2}(\S_1, \S_2, \alpha) \\
& = \Phi\{-\sigma_n^{-1} \sigma_{0,n} z_\alpha + p \sigma_n^{-1} \operatorname{tr}(\S_1 - \S_2)^2\},
\end{aligned}$$

where $\Phi(\cdot)$ denotes the cumulative probability function of $N(0,1)$.

2.2. Relationship with the test proposed in Cheng et al. (2018)

The proposed spatial rank test seems to be more complex than the existing ones, such as the spatial sign test proposed by Cheng et al. (2018). This is a price that we have to pay for making the proposed method powerful in testing the high-dimensional data, where the data dimension is potentially much larger than the squares of the sample sizes, especially for the data generated from heavy-tailed distributions. Below we will explain the motivation of the proposed method in detail.

First, we recall Lemma B.1 in Han and Liu (2018).

Lemma 2.3: *Let $\boldsymbol{X}, \tilde{\boldsymbol{X}} \sim \mathcal{E}_p(\boldsymbol{\mu}, \boldsymbol{\Lambda}, F_\xi)$, where $\boldsymbol{X}$ and $\tilde{\boldsymbol{X}}$ are independent, then*

$$\mathbb{E}\{U(\boldsymbol{X} - \tilde{\boldsymbol{X}})U(\boldsymbol{X} - \tilde{\boldsymbol{X}})^{\mathrm{T}}\} = \mathbb{E}\{U(\boldsymbol{X} - \boldsymbol{\mu})U(\boldsymbol{X} - \boldsymbol{\mu})\}.$$

By Lemma 2.3, we have that

$$\begin{aligned}
& \mathbb{E}\{U(\boldsymbol{X}_i - \boldsymbol{X}_j)U(\boldsymbol{X}_i - \boldsymbol{X}_j)^{\mathrm{T}}\} \\
& = \mathbb{E}\{U(\boldsymbol{X}_i - \boldsymbol{\mu}_1)U(\boldsymbol{X}_i - \boldsymbol{\mu}_1)^{\mathrm{T}}\} = \S_1,
\end{aligned}$$

for each $i, j \in \{1, \ldots, n_1\}$ with $i \neq j$, where $\mathbb{E}\{U(\boldsymbol{X}_i - \boldsymbol{X}_j)U(\boldsymbol{X}_i - \boldsymbol{X}_j)^{\mathrm{T}}\}$ is the so-called population multivariate Kendall's tau matrix of $\boldsymbol{X}$ (Oja, 2010). Similarly,

$$\begin{aligned}
& \mathbb{E}\{U(\boldsymbol{Y}_i - \boldsymbol{Y}_j)U(\boldsymbol{Y}_i - \boldsymbol{Y}_j)^{\mathrm{T}}\} \\
& = \mathbb{E}\{U(\boldsymbol{Y}_i - \boldsymbol{\mu}_2)U(\boldsymbol{Y}_i - \boldsymbol{\mu}_2)^{\mathrm{T}}\} = \S_2,
\end{aligned}$$

for each $i, j \in \{1, \ldots, n_2\}$ with $i \neq j$, where $\mathbb{E}\{U(\boldsymbol{Y}_i - \boldsymbol{Y}_j)U(\boldsymbol{Y}_i - \boldsymbol{Y}_j)^{\mathrm{T}}\}$ is the population multivariate Kendall's tau matrix of $\boldsymbol{Y}$. Lemma 2.3 suggests that for each of the two populations, the population multivariate Kendall's tau matrix is the same as the spatial sign covariance matrix. As a result, testing equality of the two spatial sign covariance matrices is identical to testing equality of the two population multivariate Kendall's tau matrices.

Moreover, it can be seen that the three components of the Frobenius norm of the difference between $\S_1$ and $\S_2$, $\operatorname{tr}\{(\S_1 - \S_2)^2\} = \operatorname{tr}(\S_1^2) + \operatorname{tr}(\S_2^2) - 2\operatorname{tr}(\S_1\S_2)$, have

the following equivalent representations:

$$\mathrm{tr}(\S_1^2) = \mathbb{E}[\{U(\boldsymbol{X}_i - \boldsymbol{X}_j)^{\mathrm{T}} U(\boldsymbol{X}_k - \boldsymbol{X}_l)\}^2],$$

for each $i, j, k, l \in \{1, \ldots, n_1\}$, where i, j, k, l are not equal to each other;

$$\mathrm{tr}(\S_2^2) = \mathbb{E}[\{U(\boldsymbol{Y}_i - \boldsymbol{Y}_j)^{\mathrm{T}} U(\boldsymbol{Y}_k - \boldsymbol{Y}_l)\}^2],$$

for each $i, j, k, l \in \{1, \ldots, n_2\}$, where i, j, k, l are not equal to each other;

$$\mathrm{tr}(\S_1 \S_2) = \mathbb{E}[\{U(\boldsymbol{X}_i - \boldsymbol{X}_j)^{\mathrm{T}} U(\boldsymbol{Y}_k - \boldsymbol{Y}_l)\}^2],$$

for each $i, j \in \{1, \ldots, n_1\}$ with $i \neq j$ and each $k, l \in \{1, \ldots, n_2\}$ with $k \neq l$. These representations finally enlighten us to construct T_{HT} as that in the above subsection, which is actually a consistent estimator of $p\,\mathrm{tr}\{(\S_1 - \S_2)^2\}$.

Unlike the spatial sign covariance matrix, to estimate the multivariate Kendall's tau matrix, it is not necessary to estimate the spatial medians, whose estimators may bring a bias hence strengthens the condition imposed on the dimension p. That is the reason why we propose to use a new test procedure based on the multivariate Kendall's tau matrix rather than the spatial sign covariance matrix. Therefore, the condition imposed on the dimension p can be released to some extent, which makes the proposed test procedure powerful in high-dimensional data, even with the dimension much larger than the sample sizes.

In fact, in the spatial sign test proposed by Cheng et al. (2018), to test the equality of the two spatial sign covariance matrices $\S_1$ and $\S_2$, the test statistic is

$$T_{\mathrm{SS}} = \frac{p}{n_1(n_1-1)} \sum_{i \neq j}^{n_1} (\hat{\boldsymbol{u}}_i^{\mathrm{T}} \hat{\boldsymbol{u}}_j)^2 + \frac{p}{n_2(n_2-1)} \times \sum_{i \neq j}^{n_2} (\hat{\boldsymbol{v}}_i^{\mathrm{T}} \hat{\boldsymbol{v}}_j)^2 - \frac{2p}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n2} (\hat{\boldsymbol{u}}_i^{\mathrm{T}} \hat{\boldsymbol{v}}_j)^2,$$

where $\hat{\boldsymbol{u}}_i = U(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}_1)$ and $\hat{\boldsymbol{v}}_j = U(\boldsymbol{Y}_j - \hat{\boldsymbol{\mu}}_2)$ for $i = 1, \ldots, n_1, j = 1, \ldots, n_2$. Here, $\hat{\boldsymbol{\mu}}_1$ and $\hat{\boldsymbol{\mu}}_2$ are the spatial median estimators of $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively, obtained by using the estimation method proposed in Mottonen and Oja (1995). T_{SS} is an estimator of $p\,\mathrm{tr}\{(\S_1 - \S_2)^2\}$, but unfortunately $\mathbb{E}(T_{\mathrm{SS}}/p) - \mathrm{tr}\{(\S_1 - \S_2)^2\} = \delta_{n_1,n_2} \neq 0$, due to the spatial median estimators $\hat{\boldsymbol{\mu}}_1$ and $\hat{\boldsymbol{\mu}}_2$ (see Lemma 2 in Cheng et al., 2018). To obtain a consistent estimator of the bias δ_{n_1,n_2}, the condition $p = O\{(n_1 + n_2)^2\}$ was imposed in Cheng et al. (2018), which limits the application of T_{SS} for the high-dimensional data where the dimension is much larger than the squares of sample sizes.

3. Simulation study

In this section, we will present some numerical results to demonstrate the performance of the proposed test (abbreviated as HT) in high-dimensional cases, in comparison with two existing popular tests, the test proposed by Li and Chen (2012) (abbreviated as LZ) and the spatial sign test proposed by Cheng et al. (2018) (abbreviated as SS). The following three scenarios are considered.

(I) Multivariate normal distribution: $\boldsymbol{X} \sim N_p(\boldsymbol{0}, \boldsymbol{\Sigma}_1)$ and $\boldsymbol{Y} \sim N_p(\boldsymbol{0}, \boldsymbol{\Sigma}_2)$.
(II) Multivariate t-distribution: $\boldsymbol{X} \sim t_p(\boldsymbol{0}, \boldsymbol{\Sigma}_1, 3)$ and $\boldsymbol{Y} \sim t_p(\boldsymbol{0}, \boldsymbol{\Sigma}_2, 3)$.
(III) Multivariate mixture normal distribution: $\boldsymbol{X} \sim \mathrm{MN}_{p,\gamma,9}(\boldsymbol{0}, \boldsymbol{\Sigma}_1) \triangleq \gamma N_p(\boldsymbol{0}, \boldsymbol{\Sigma}_1) + (1-\gamma) N_p(\boldsymbol{0}, 9\boldsymbol{\Sigma}_1)$, $\boldsymbol{Y} \sim \mathrm{MN}_{q,\gamma,9}(\boldsymbol{0}, \boldsymbol{\Sigma}_2)$, $\gamma = 0.8$.

For all the above scenarios, let $\boldsymbol{\Sigma}_1 = (0.3^{|i-j|})$ and $\boldsymbol{\Sigma}_2 = (\rho^{|i-j|})$ with $\rho = 0.3, 0.6, 0.7$. Then, $\rho = 0.3$ corresponds to the situation where the null hypothesis is true, while $\rho = 0.6$ or 0.7 corresponds to the situation where the alternative hypothesis is true. Note that all the following simulation results are obtained based on 1000 replications.

First, to observe the influence of the dimension p to the potential bias of the methods involved, we summarize the results of the mean-standard deviation-ratio $\mathrm{E}(T)/\sqrt{\mathrm{var}(T)}$ and the variance estimator ratio $\widehat{\mathrm{var}(T)}/\mathrm{var}(T)$ under the null hypothesis in Table 1 for each $T \in \{T_{\mathrm{HT}}, T_{\mathrm{SS}}, T_{\mathrm{LZ}}\}$ with $n_1 = n_2 = 15$ and $p = 100, 200, 400, 800, 1200$, where T_{LZ} is the test statistic proposed in Li and Chen (2012). Since the exact value of $\mathrm{E}(T)$ and $\mathrm{var}(T)$ are difficult to calculate, we replace them with their Monte-Carlo estimators respectively, using 1000 repeated samplings.

Table 1 indicates that SS has worse mean-standard deviation-ratio results than the other two methods in high-dimensional situations, particularly when $p > (n_1 + n_2)^2$. This is most likely due to the fact that in T_{SS} the bias correction process is limited by the condition that $p = O\{(n_1 + n_2)^2\}$. On the other hand, suggested by the variance estimator ratio results of Table 1, the estimated variances of LZ are eventually larger than the real ones, particularly in non-normal situations. In contrast, HT has better performance in these two aspects.

Then, we will compare the performance of the three methods in empirical size and empirical power. Let $n_1 = n_2 = 15, 20, 30$ and $p = 100, 200, 400, 800, 1200$. Tables 2–4 summarize the empirical size and power results of the three methods. First, the empirical size results in Tables 2–4, corresponding to the setting of $\rho = 0.3$, suggest that LZ fails to control the empirical size in the non-normal cases. Moreover, when comparing HT with SS, we find that their performance is very similar, except in the cases where the dimension is comparable to or larger than the squares of the sample sizes, i.e. $1200 > (15 + 15)^2$. In such cases, SS may lose control of the empirical size, which is consistent

Table 1. Comparison of the mean-standard deviation-ratio and the variance estimator ratio at the 5% level with $n_1 = n_2 = 15$ and $p = 100, 200, 400, 800, 1200$.

	$E(T)/\sqrt{\text{var}(T)}$			$\widehat{\text{var}(T)}/\text{var}(T)$		
p	HT	SS	LZ	HT	SS	LZ
Scenario I: multivariate normal distribution						
100	−0.04	−0.01	−0.02	1.06	1.27	1.01
200	−0.03	−0.01	−0.05	1.13	1.17	1.03
400	−0.04	0.03	0.03	1.18	1.09	1.03
800	0.03	0.23	0.03	1.15	1.20	1.04
1200	0.03	0.50	0.06	1.12	1.14	1.15
Scenario II: multivariate t-distribution						
100	−0.07	−0.22	−0.28	1.13	3.75	495
200	−0.03	−0.18	−0.72	1.14	1.63	202
400	0.00	−0.25	−0.18	1.24	4.21	144
800	0.01	−0.35	0.65	1.22	7.50	195
1200	0.05	−0.48	−0.06	1.21	7.60	100
Scenario III: multivariate mixture normal distribution						
100	−0.04	−0.02	−0.03	1.12	1.27	7.91
200	−0.06	−0.03	−0.04	1.13	1.32	7.15
400	−0.01	0.16	−0.01	1.24	1.13	8.63
800	0.03	0.42	0.11	1.19	1.12	8.14
1200	0.00	0.67	−0.06	1.22	1.03	8.16

Table 2. Empirical size and power comparison at the 5% level with $n_1 = n_2 = 15$ and $p = 100, 200, 400, 800, 1200$.

	$\rho = 0.3$			$\rho = 0.6$			$\rho = 0.7$		
p	HT	SS	LZ	HT	SS	LZ	HT	SS	LZ
Multivariate normal distribution									
100	4.1	4.6	3.6	21	20	22	56	53	58
200	3.7	5.0	3.5	22	21	23	59	56	60
400	4.3	4.5	3.7	25	23	23	59	59	61
800	4.5	6.8	3.5	24	28	23	62	65	63
1200	5.2	13	4.6	24	37	24	62	69	62
Multivariate t-distribution									
100	4.7	4.8	19	22	21	33	53	52	47
200	4.8	4.6	20	23	20	32	57	55	46
400	5.7	4.2	23	26	21	34	60	55	48
800	6.0	6.2	22	26	22	35	61	54	47
1200	6.4	8.4	22	25	25	37	63	57	51
Multivariate mixture normal distribution									
100	4.2	4.6	22	25	22	35	54	55	50
200	4.8	5.9	20	21	26	32	56	58	48
400	5.8	6.3	23	26	25	36	59	60	50
800	5.6	9.1	24	26	32	37	59	69	51
1200	6.1	13	24	26	39	34	60	72	48

Table 3. Empirical size and power comparison at the 5% level with $n_1 = n_2 = 20$ and $p = 100, 200, 400, 800, 1200$.

	$\rho = 0.3$			$\rho = 0.6$			$\rho = 0.7$		
p	HT	SS	LZ	HT	SS	LZ	HT	SS	LZ
Multivariate normal distribution									
100	5.1	3.9	5.3	35	29	37	79	75	78
200	4.1	4.1	3.9	34	33	35	81	81	82
400	5.6	6.4	5.1	35	36	35	84	83	84
800	4.3	3.9	4.1	36	36	35	84	86	86
1200	4.2	2.6	2.8	35	37	36	86	84	82
Multivariate t-distribution									
100	5.8	4.1	25	36	31	42	75	77	58
200	5.0	4.2	22	36	35	38	78	79	58
400	6.7	4.6	27	37	34	40	80	82	57
800	5.8	4.8	26	39	33	45	84	80	63
1200	4.8	4.4	26	37	34	46	83	81	63
Multivariate mixture normal distribution									
100	6.5	4.2	25	36	31	45	78	76	61
200	5.1	4.3	26	36	33	42	79	78	61
400	5.9	5.0	24	38	36	43	82	85	60
800	5.2	7.1	25	37	40	43	83	85	62
1200	4.6	6.2	25	39	42	43	84	89	63

Table 4. Empirical size and power comparison at the 5% level with $n_1 = n_2 = 30$ and $p = 100, 200, 400, 800, 1200$.

	$\rho = 0.3$			$\rho = 0.6$			$\rho = 0.7$		
p	HT	SS	LZ	HT	SS	LZ	HT	SS	LZ
Multivariate normal distribution									
100	3.9	4.7	3.1	60	58	60	98	98	99
200	4.2	3.6	4.1	62	61	62	98	98	99
400	5.4	5.2	4.7	64	63	64	99	98	99
800	4.7	5.0	3.8	64	60	64	99	99	99
1200	6.2	5.8	5.1	68	65	66	99	99	99
Multivariate t-distribution									
100	4.4	3.9	27	58	57	53	97	98	72
200	5.7	4.8	27	62	59	49	97	98	72
400	5.2	5.6	29	62	62	51	98	99	73
800	5.9	4.5	28	61	62	51	99	99	74
1200	6.4	5.6	29	64	61	50	99	99	73
Multivariate mixture normal distribution									
100	4.6	5.8	27	59	58	54	97	97	77
200	5.8	6.2	24	60	62	50	98	99	75
400	5.7	6.2	25	62	62	54	98	99	77
800	5.6	7.3	26	63	63	53	98	99	77
1200	5.8	5.4	26	63	66	54	99	99	78

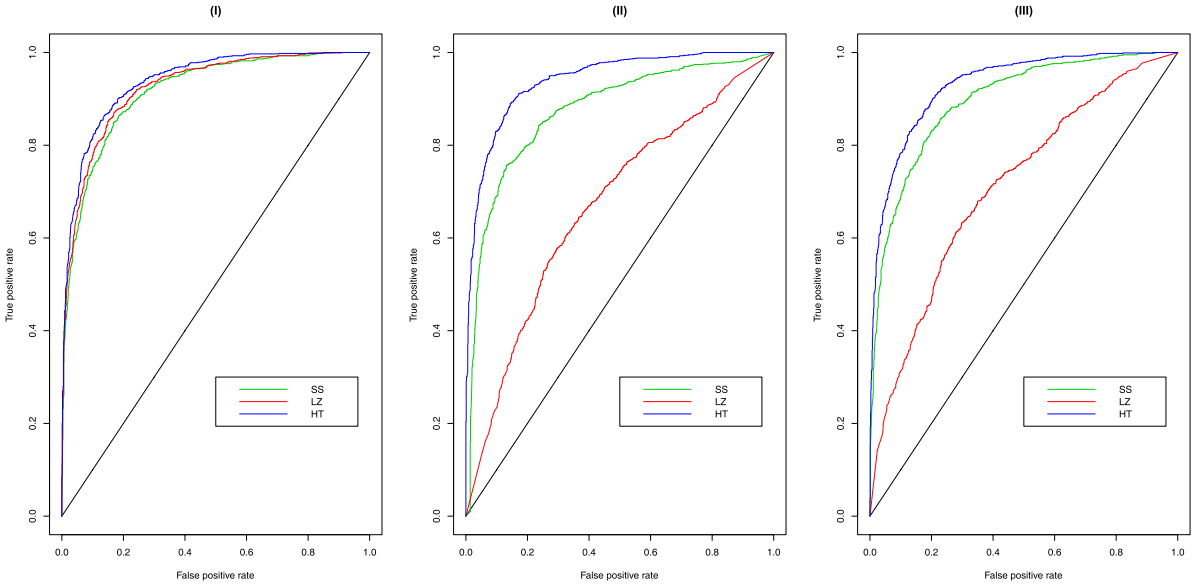

Figure 1. ROC curves of the involved tests under the three scenarios with $(n_1, n_2, p) = (15, 15, 800)$.

with the conclusion made by analysing Table 1. In the above results about the empirical size, in a few cases, the empirical size is slightly larger than 5%, but still within a reasonable range. To comprehensively compare the empirical size and power of the three tests, in Figure 1, we present the receiver operating characteristic curves (ROCs) for the three tests with $(n_1, n_2, p) = (15, 15, 800)$. Suggested by Figure 1, these tests have similar performance under the multivariate normal distributions, while under the remaining heavy-tailed distributions, the area under ROC (AUC) of the proposed HT test is larger than the AUCs of its competitors. This further demonstrates the advantages of the proposed test.

Next, we consider an alternative structure of the covariance matrices, i.e. $\mathbf{\Sigma}_i = (a_{ikl})$ for each $i \in \{1, 2\}$, where

$$a_{ikk} = 1, \quad a_{ik,k+1} = \frac{\rho_i + \rho_i^2}{1 + 2\rho_i^2}, \quad a_{k,k+2} = \frac{\rho_i}{1 + 2\rho_i^2}$$

for each $k \in \{1, \ldots, p\}$ and the remaining entries of $\mathbf{\Sigma}_i$ are all zeros. Note that $\mathbf{\Sigma}_i$ is the corresponding covariance matrix of x_i following the MA(2) model:

$$x_{it} = z_{it} + \rho_i z_{i,t-1} + \rho_i z_{i,t-2},$$

where z_{it}'s are i.i.d. random variables with mean zero and variance $\frac{1}{1+2\rho_i^2}$. Under the null hypothesis, we set $\rho_1 = \rho_2 = 0.7$, while under the alternative hypothesis, we set $\rho_1 = 0.7$ and $\rho_2 = 0.1$ for instance. The other settings are all the same as the above. Tables 5 and 6 report the empirical sizes and power of these three

Table 5. Empirical size comparison at the 5% level with the MA(2) covariance matrices with $n_1 = n_2 = 15, 20, 30$ and $p = 100, 200, 400, 800, 1200$.

	$n_1 = n_2 = 15$			$n_1 = n_2 = 20$			$n_1 = n_2 = 30$		
p	HT	SS	LZ	HT	SS	LZ	HT	SS	LZ
Multivariate normal distribution									
100	2.7	4.1	2.6	4.2	3.7	5.6	3.6	4.3	4.8
200	3.6	2.9	3.4	4.0	3.4	3.8	4.1	3.7	4.3
400	5.1	4.4	4.0	5.7	4.6	5.6	4.7	4.5	4.8
800	4.2	5.3	3.6	3.9	3.8	2.2	4.4	4.1	3.8
1200	5.8	7.8	5.8	3.1	5.6	3.0	3.9	4.3	6.6
Multivariate t-distribution									
100	4.2	3.3	19	6.5	4.0	24	4.8	5.6	29
200	6.0	4.4	19	6.0	3.9	23	5.9	4.0	28
400	6.0	4.0	22	6.5	3.2	25	6.1	4.9	28
800	5.7	7.0	22	6.1	4.2	26	6.7	3.4	27
1200	6.7	7.8	23	6.1	4.9	26	6.9	5.1	30
Multivariate mixture normal distribution									
100	5.0	3.5	22	5.6	3.7	27	4.8	3.5	26
200	5.7	4.0	25	6.3	4.1	26	6.3	4.4	27
400	5.8	5.8	23	6.3	4.3	28	6.2	3.2	26
800	5.4	7.0	24	5.1	7.0	26	5.9	4.7	26
1200	5.9	8.8	24	4.9	7.9	25	5.2	5.2	26

Table 6. Empirical power comparison at the 5% level with the MA(2) covariance matrices with $n_1 = n_2 = 15, 20, 30$ and $p = 100, 200, 400, 800, 1200$.

	$n_1 = n_2 = 15$			$n_1 = n_2 = 20$			$n_1 = n_2 = 30$		
p	HT	SS	LZ	HT	SS	LZ	HT	SS	LZ
Multivariate normal distribution									
100	42	40	44	65	63	66	92	92	93
200	42	41	43	67	64	67	93	93	94
400	45	44	48	67	66	68	94	94	94
800	44	47	41	66	68	67	93	93	94
1200	45	55	44	65	70	64	94	95	94
Multivariate t-distribution									
100	42	40	40	65	62	56	91	92	67
200	42	40	42	64	63	50	91	92	67
400	43	40	46	64	66	51	92	92	65
800	44	45	46	64	64	51	91	94	65
1200	45	45	46	62	61	54	92	94	68
Multivariate mixture normal distribution									
100	43	45	42	64	61	57	90	92	68
200	44	44	44	66	66	56	91	93	73
400	45	48	43	66	68	55	93	93	69
800	43	55	43	65	71	53	92	95	70
1200	43	61	46	62	74	53	92	94	66

methods, respectively. Although Table 6 suggests that the performance of empirical power of the three methods is similar, Table 5 suggests that the abilities of LZ and SS to control the empirical size are weakening much more quickly than HT with the increase of p for fixed n_1 and n_2, especially when the dimension is comparable to or larger than the squares of the sample sizes.

Overall, the comprehensive numerical results suggest that the proposed HT test has obvious advantages in terms of controlling empirical size over the existing two methods. Such gain is especially clear when the original distribution deviates from normality, and when the dimension is larger than the squares of sample sizes.

4. Application

In this section, we apply the proposed testing method to a gene dataset, which contains the expression of the 2000 genes with the highest minimal intensity across the 62 tissues. Each entry in the dataset is a gene intensity derived using the filtering process proposed in Alon et al. (1999). The dataset was previously studied by Alon et al. (1999), and now can be freely downloaded at the following website: http://genomics-pubs.princeton.edu/oncology/affydata/index.html.

Among the 62 tissues, there are 22 normal tissues and 40 tumour colon tissues. We aim to test the hypothesis that the tissues in the tumour group and those in the normal group have the proportional covariance matrices in terms of the expression levels of the 2000 genes, where the dimension 2000 is larger than the squares of the sample sizes, 484 and 1600.

First, the normal distribution was tested for the expression data of each gene, using the Shapiro–Wilk test. The top two panels of Figure 1 present the histograms of the p-values of the normality tests for the tumour 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

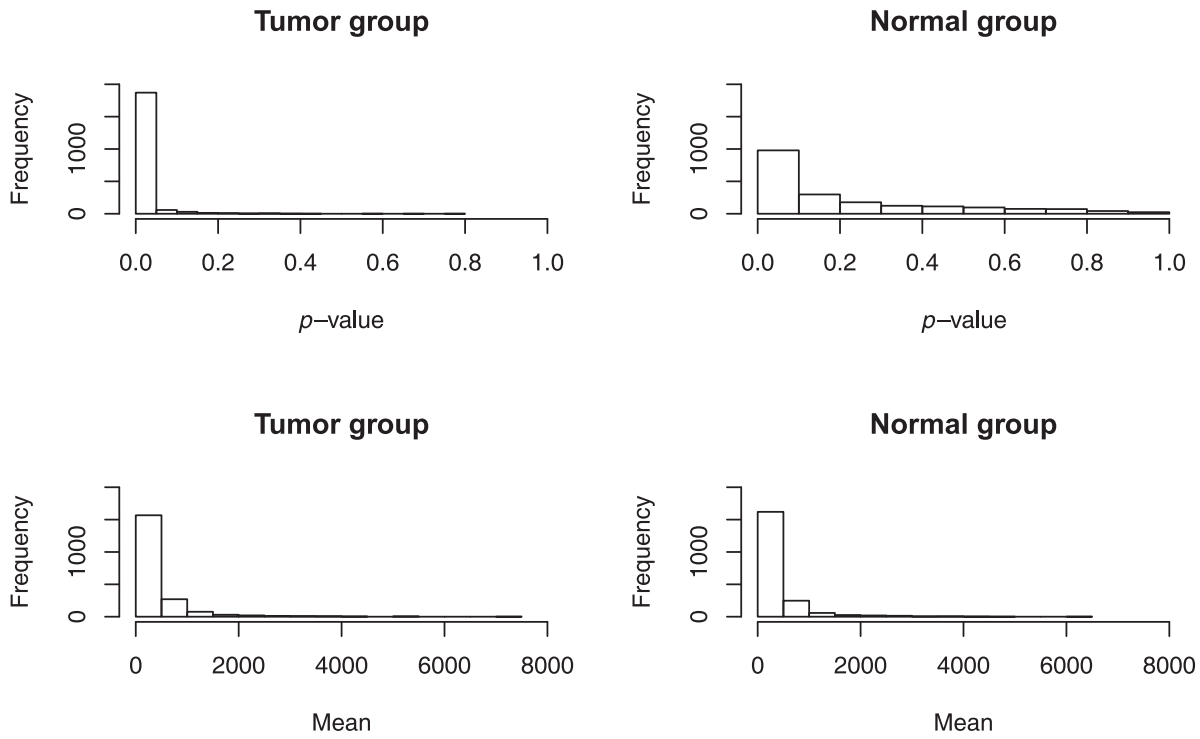

Figure 2. Histograms of the *p*-values of the normality tests and the gene expression means, for the tumour group and the normal group, respectively.

of all the normality tests are 93.55% and 37.75% for the tumour group and the normal group, respectively. This motivates us to use a nonparametric approach for testing the above hypothesis, which can deal with the high-dimensional data from non-normal distributions.

The bottom two panels of Figure 2 indicate that there exist some genes with very high values of sample mean in terms of expression. We see that the sample means vary largely for each of the two groups and recall that the dimension is larger than the squares 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 is more appropriate for this dataset.

Based on the above reasons, we apply the proposed HT test to this dataset. The test statistic and *p*-value of the HT test are 4.823 and 0.000, respectively, hence the null hypothesis is rejected, which suggests that the covariance matrix of the gene expression levels of the tumour group is significantly not proportional to that of the normal group. This result can also be intuitively verified by comparing the sample correlation matrices of the two groups. As a convenience and for demonstration purposes, in Figure 3, we only plot the heatmaps of the sample correlation matrices of the two groups as well as the difference of the two matrices using the first 100 genes in the original data. The heatmaps demonstrate that there are some intuitive differences between the two sample correlation matrices,

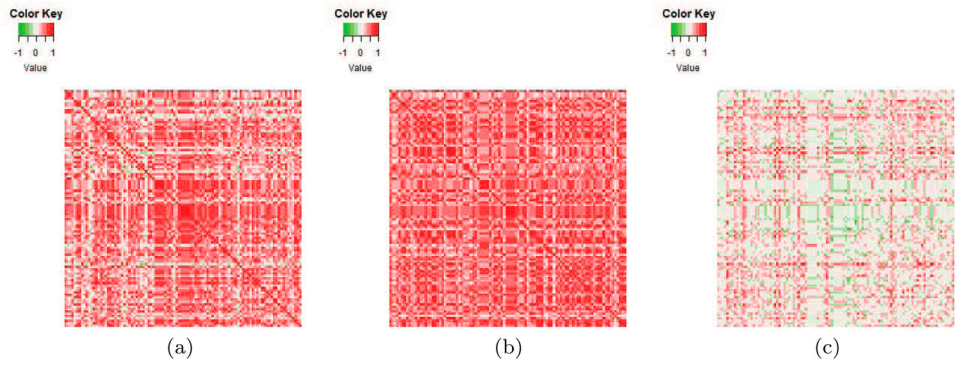

Figure 3. Heatmaps of the sample correlation matrices of the two groups as well as the difference of the two matrices, which are constructed via the first 100 genes in the original data. (a) Normal group, (b) tumour group and (c) difference of two groups.

which tends to support our result of rejecting the null hypothesis.

5. Conclusion

We have proposed the HT test, a new high-dimensional spatial rank test, for the proportionality testing problem of two high-dimensional covariance matrices, which is a high-dimensional extension of Kendall's tau test. It inherits the robustness advantage of the traditional spatial rank-based methods, and also has strong potential in dealing with the high-dimensional data, where the dimension can be potentially much larger than the squares of the sample sizes. We establish the asymptotic distributions of the proposed method rigorously. In comparison with some existing test procedures, the gain in empirical power and empirical size of HT is especially clear in high-dimensional and heavy-tailed data, shown by many numerical evidence. The real data analysis shows the applicability and pertinence of the proposed method to high-dimensional gene expression data.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the National Natural Science Foundation of China [Grant Numbers 11501092, 11571068] and the Special Fund for Key Laboratories of Jilin Province, China [Grant Number 20190201285JC].

References

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., & Levine, D. M. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences of the United States of America*, *96*(12), 6745–6750. https://doi.org/10.1073/pnas.96.12.6745

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

Barber, R. F., & Kolar, M. (2018). Rocket: Robust confidence intervals via Kendall's tau for transelliptical graphical models. *The Annals of Statistics*, *46*(6B), 3422–3450. https://doi.org/10.1214/17-AOS1663

Bühlmann, P, & van de Geer, S. (2011). *Statistics for High-dimensional Data: Methods, Theory and Applications* (1st ed.). Springer Publishing Company, Incorporated.

Cai, T. T., & Zhang, A. (2016). Inference for high-dimensional differential correlation matrices. *Journal of Multivariate Analysis*, *143*(6009), 107–126. https://doi.org/10.1016/j.jmva.2015.08.019

Chen, S. X., & Qinm, Y. L. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *Annals of Statistics*, *38*(2), 808–835.https://doi.org/10.1214/09-AOS716

Cheng, G., Liu, B., Peng, L., Zhang, B., & Zheng, S. (2018). Testing the equality of two high-dimensional spatial sign covariance matrices. *Scandinavian Journal of Statistics*, *46*(1), 257–271. https://doi.org/10.1111/sjos.v46.1

Eriksen, P. S. (1987). Proportionality of covariance matrices. *Annals of Statistics*, *15*(2), 732–748. https://doi.org/10.1214/aos/1176350372

Fang, K. T., Kotz, S., & Ng, K. W. (1990). *Symmetric Multivariate and Related Distributions*. Chapman and Hall.

Federer, W. T. (1951). Testing proportionality of covariance matrices. *Annals of Mathematical Statistics*, *22*(1), 102–106. https://doi.org/10.1214/aoms/1177729697

Feng, L., & Liu, B. (2017). High-dimensional rank tests for sphericity. *Journal of Multivariate Analysis*, *155*, 217–233. https://doi.org/10.1016/j.jmva.2017.01.003

Feng, L., & Sun, F. (2016). Spatial-sign based high-dimensio nal location test. *Electronic Journal of Statistics*, *10*(2), 2420–2434. https://doi.org/10.1214/16-EJS1176

Feng, L., Zou, C., & Wang, Z. (2016). Multivariate-sign-based high-dimensional tests for the two-sample location problem. *Journal of the American Statistical Association*, *111*(514), 721–735. https://doi.org/10.1080/01621459.2015.1035380

Flury, B. K. (1986). Proportionality of *k* covariance matrices. *Statistics and Probability Letters*, *4*(1), 29–33. https://doi.org/10.1016/0167-7152(86)90035-0

Flury, B. K., & Riedwyl, H. (1988). *Multivariate Statistics: A Practical Approach*. Chapman and Hall.

Hall, P. G., & Hyde, C. C. (1980). *Martingale Central Limit Theory and Its Applications*. Academic Press.

Han, F., Chen, S., & Liu, H. (2017). Distribution-free tests of independence in high dimensions. *Biometrika*, *104*(4), 813–828. https://doi.org/10.1093/biomet/asx050

Han, F., & Liu, H. (2018). ECA: High-dimensional elliptical component analysis in non-Gaussian distributions. *Journal of the American Statistical Association*, *113*(521), 252–268. https://doi.org/10.1080/01621459.2016.1246366

Kim, D. Y. (1971). Statistical inference for constants of proportionality between covariance matrices. Technical Report 59, Stanford University.

Leung, D., & Drton, M. (2018). Testing independence in high dimensions with sums of rank correlations. *The Annals of Statistics*, *46*(1), 280–307. https://doi.org/10.1214/17-AOS1550

Li, J., & Chen, S. X. (2012). Two sample tests for high dimensional covariance matrices. *Annals of Statistics*, *40*(2), 908–940.https://doi.org/10.1214/12-AOS993

Liu, B., Xu, L., Zheng, S., & Tian, G. (2014). A new test for the proportionality of two large-dimensional covariance matrices. *Journal of Multivariate Analysis*, *131*(1), 293–308. https://doi.org/10.1016/j.jmva.2014.06.008

Magyar, A., & Tyler, D. (2014). The asymptotic inadmissibility of the spatial sign covariance matrix for elliptically symmetric distributions. *Biometrika*, *101*(3), 673–688. https://doi.org/10.1093/biomet/asu020

Mottonen, J., & Oja, H. (1995). Multivariate spatial sign and rank methods. *Journal of Nonparametric Statistics*, *5*(2), 201–213. https://doi.org/10.1080/10485259508832643

Oja, H. (2010). *Multivariate nonparametric methods with R*. Springer.

Rao, C. R. (1983). Likelihood ratio tests for relationships between two covariance matrices. In S. Karlin, T. Amemiya, & L. A. Goodman (Eds.), *Studies in Econometrics, Time Series and Multivariate Statistics* (pp. 529–543). Academic Press.

Schott, J. R. (1991). Some tests for common principal component subspaces in several groups. *Biometrika*, *78*(4), 771–777. https://doi.org/10.1093/biomet/78.4.771

Schott, J. R. (1999). A test for proportional covariance matrices. *Computational Statistics and Data Analysis*, *32*(2), 135–146. https://doi.org/10.1016/S0167-9473(99)00032-8

Wang, L., Peng, B., & Li, R. (2015). A high-dimensional nonparametric multivariate test for mean vector. *Journal of the American Statistical Association*, *110*(512), 1658–1669. https://doi.org/10.1080/01621459.2014.988215

Xu, L., Liu, B., Zheng, S., & Bao, S. (2014). Testing proportionality of two large-dimensional covariance matrices. *Computational Statistics and Data Analysis*, *78*, 43–55. https://doi.org/10.1016/j.csda.2014.03.014

Zou, C. L., Peng, L. H., Feng, L., & Wang, Z. J. (2014). Multivariate sign-based high-dimensional tests for sphericity. *Biometrika*, *101*(1), 229–236. https://doi.org/10.1093/biomet/ast040

Appendix

Define

$$\mathbf{U}_{X,i} = U(\boldsymbol{X}_i - \boldsymbol{\theta}_X), \quad \mathbf{U}_{Y,i} = U(\boldsymbol{Y}_i - \boldsymbol{\theta}_Y).$$

Before proving the main theorem, below we recall some necessary lemmas.

Lemma A.1: *Under Conditions (C1) and (C2), for any $p \times p$ symmetric matrix $\boldsymbol{W}$,*

$$\mathbb{E}\{(\mathbf{U}_{X,i}^{\mathrm{T}}\mathbf{U}_{X,j})^4\} = O(1)E^2\{\mathbf{U}_{X,i}^{\mathrm{T}}\mathbf{U}_{X,j})^2\},$$
$$\mathbb{E}\{(\mathbf{U}_{X,i}^{\mathrm{T}}\boldsymbol{W}\mathbf{U}_{X,i})^2\} = O(1)E^2(\mathbf{U}_{X,i}^{\mathrm{T}}\boldsymbol{W}\mathbf{U}_{X,i}),$$
$$\mathbb{E}\{(\mathbf{U}_{X,i}^{\mathrm{T}}\boldsymbol{W}\mathbf{U}_{X,j})^2\} = O(1)E^2(\mathbf{U}_{X,i}^{\mathrm{T}}\boldsymbol{W}\mathbf{U}_{X,j}).$$

Note that Lemma A.1 is the same as Lemma 1 of Wang et al. (2015).

Lemma A.2: *Let $\mathbf{U}^* = (U_1^*, \ldots, U_p^*)^{\mathrm{T}}$ be a random vector uniformly distributed on the unit sphere of $\mathcal{R}^p$, then we have that*

(1) $\mathbb{E}(\mathbf{U}^*) = \mathbf{0}$, $\mathrm{Cov}(\mathbf{U}^*) = p^{-1}\mathbf{I}_p$, $\mathbb{E}(U_k^{*4}) = 3p^{-1}(p+2)^{-1}$ *and* $\mathbb{E}(U_k^{*2}U_l^{*2}) = p^{-1}(p+2)^{-1}$ *for each* $k, l \in \{1, \ldots, p\}$ *with* $k \neq l$;

(2) *for any $p \times p$ symmetric matrix $\boldsymbol{W}$, $E\{(\mathbf{U}^{*\mathrm{T}}\boldsymbol{W}\mathbf{U}^*)^2\} = p^{-1}(p+2)^{-1}\{\mathrm{tr}^2(\boldsymbol{W}) + 2\mathrm{tr}(\boldsymbol{W}^2)\}$ and $E\{(\mathbf{U}^{*\mathrm{T}}\boldsymbol{W}\mathbf{U}^*)^4\} = p^{-2}(p+2)^{-2}\{3\mathrm{tr}^2(\boldsymbol{W}^2) + 6\mathrm{tr}(\boldsymbol{W}^2)\}$.*

In Lemma A.2, the first statement has been proved in Section 3.1 of Fang et al. (1990) and the second statement has been proved in Zou et al. (2014).

Now, we are ready to present the proof of Theorem 2.2. Then, the proof of Theorem 2.1 can be directly obtained.

Proof of Theorem 2.2:

Define

$$\boldsymbol{V}_{X,i} \doteq \mathbb{E}\{U(\boldsymbol{X}_i - \boldsymbol{X}_j)|\boldsymbol{X}_i\}, \quad \boldsymbol{V}_{Y,i} \doteq \mathbb{E}\{U(\boldsymbol{Y}_i - \boldsymbol{Y}_j)|\boldsymbol{Y}_i\},$$
$$\boldsymbol{W}_{Y,ij} \doteq U(\boldsymbol{Y}_i - \boldsymbol{Y}_j) - \boldsymbol{V}_{Y,i} + \boldsymbol{V}_{Y,j},$$
$$\boldsymbol{W}_{X,ij} \doteq U(\boldsymbol{X}_i - \boldsymbol{X}_j) - \boldsymbol{V}_{X,i} + \boldsymbol{V}_{X,j},$$
$$\mathbf{B}_1 \doteq \mathbb{E}(\boldsymbol{V}_{X,i}\boldsymbol{V}_{X,i}^{\mathrm{T}}), \quad \mathbf{B}_2 \doteq \mathbb{E}(\boldsymbol{V}_{Y,i}\boldsymbol{V}_{Y,i}^{\mathrm{T}}).$$

Hence we have that $\mathbb{E}(\boldsymbol{V}_{X,i}^{\mathrm{T}}\boldsymbol{V}_{X,j}) = 0$ and $\mathbb{E}(\boldsymbol{V}_{X,i}^{\mathrm{T}}\boldsymbol{W}_{X,ij}) = 0$. According to Lemma 1 in Feng and Liu (2017), we have that $\mathbb{E}(\boldsymbol{W}_{X,ij}^{\mathrm{T}}\boldsymbol{W}_{X,ij}) \to 0$ as p goes to infinity and $\mathbf{B}_1 = 0.5\S_1\{1 + o(1)\}$. The same goes for $\boldsymbol{W}_{Y,ij}$ and $\mathbf{B}_2$. On this ground, by Lemma A.1, we have that

$$\mathbb{E}\{(\boldsymbol{V}_{X,i}^{\mathrm{T}}\boldsymbol{V}_{X,j})^4\} = O(1)E^2\{(\boldsymbol{V}_{X,i}^{\mathrm{T}}\boldsymbol{V}_{X,j})^2\},$$
$$\mathbb{E}\{(\boldsymbol{V}_{X,i}^{\mathrm{T}}\mathbf{A}\boldsymbol{V}_{X,i})^2\} = O(1)E^2(\boldsymbol{V}_{X,i}^{\mathrm{T}}\mathbf{A}\boldsymbol{V}_{X,i}),$$
$$\mathbb{E}\{(\boldsymbol{V}_{X,i}^{\mathrm{T}}\mathbf{A}\boldsymbol{V}_{X,j})^2\} = O(1)E^2(\boldsymbol{V}_{X,i}^{\mathrm{T}}\mathbf{A}\boldsymbol{V}_{X,j}).$$

As a result, the first part of T_{HT} has the following decomposition:

$$\begin{aligned}
&\frac{p}{n_1(n_1-1)(n_1-2)(n_1-3)}\sum\nolimits^*\{U(\boldsymbol{X}_i - \boldsymbol{X}_j)^{\mathrm{T}}U(\boldsymbol{X}_k - \boldsymbol{X}_l)\}^2 \\
&= \frac{4p}{n_1(n_1-1)}\sum\nolimits^*(\boldsymbol{V}_{X,i}^{\mathrm{T}}\boldsymbol{V}_{X,j})^2 \\
&+ \frac{2p}{n_1(n_1-1)(n_1-2)}\sum\nolimits^*(\boldsymbol{V}_{X,i}^{\mathrm{T}}\boldsymbol{W}_{X,kl})^2 \\
&+ \frac{p}{n_1(n_1-1)(n_1-2)(n_1-3)}\sum\nolimits^*(\boldsymbol{W}_{X,ij}^{\mathrm{T}}\boldsymbol{W}_{X,kl})^2 \\
&\doteq J_1 + J_2 + J_3.
\end{aligned}$$

According to Lemma A.2 and the fact that $\mathbb{E}(\boldsymbol{W}_{X,ij}^{\mathrm{T}}\boldsymbol{W}_{X,ij}) \to 0$ as p goes to infinity, we similarly have that $\mathbb{E}(J_2^2) = o\{p^2n^{-3}\,\mathrm{tr}(\S_1^2)\} = o(\sigma_n^2)$ and $\mathbb{E}(J_3^2) = o(p^2n^{-4}) = o(\sigma_n^2)$. Using the similar techniques, we can decompose the rest two parts of T_{HT}, hence conclude that

$$\begin{aligned}
T_{\mathrm{HT}} &= \frac{4p}{n_1(n_1-1)}\sum\nolimits^*(\boldsymbol{V}_{X,i}^{\mathrm{T}}\boldsymbol{V}_{X,j})^2 \\
&+ \frac{4p}{n_2(n_2-1)}\sum\nolimits^*(\boldsymbol{V}_{Y,i}^{\mathrm{T}}\boldsymbol{V}_{Y,j})^2 \\
&- \frac{8p}{n_1n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}(\boldsymbol{V}_{X,i}^{\mathrm{T}}\boldsymbol{V}_{Y,j})^2 + o_p(\sigma_n) \\
&\doteq pA_{n_1} + pB_{n_2} - 2pC_{n_1,n_2} + o_p(\sigma_n).
\end{aligned}$$

Therefore, we have that

$$\begin{aligned}
&\mathrm{var}(T_{\mathrm{HT}})/\sigma_n^2 \\
&= p^2\sigma_n^{-2}\{\mathrm{var}(A_{n_1}) + \mathrm{var}(B_{n_2}) + 4\mathrm{var}(C_{n_1,n_2}) \\
&\quad - 4\mathrm{cov}(A_{n_1}, C_{n_1,n_2}) - 4\mathrm{cov}(B_{n_2}, C_{n_1,n_2})\} + o(1).
\end{aligned}$$

Below we will consider each item in $\mathrm{var}(T_{\mathrm{HT}})/\sigma_n^2$ one by one. Before we can get the further expression of $\mathrm{var}(A_{n_1})$, we need to study $\mathbb{E}(A_{n_1}^2)$ first. We have that

$$\begin{aligned}
\mathbb{E}(A_{n_1}^2) &= \frac{16}{n_1^2(n_1-1)^2}\mathbb{E}\left[\left\{\sum\nolimits^*(\boldsymbol{V}_{X,i}^{\mathrm{T}}\boldsymbol{V}_{X,j})^2\right\}^2\right] \\
&= \frac{16}{n_1^2(n_1-1)^2}[2n_1(n_1-1)\mathbb{E}\{(\boldsymbol{V}_{X,i}^{\mathrm{T}}\boldsymbol{V}_{X,j})^4\} \\
&\quad + 4n_1(n_1-1)(n_1-2)\mathbb{E}\{(\boldsymbol{V}_{X,i}^{\mathrm{T}}\boldsymbol{V}_{X,j})^2(\boldsymbol{V}_{X,i}^{\mathrm{T}}\boldsymbol{V}_{X,k})^2\} \\
&\quad + n_1(n_1-1)(n_1-2)(n_1-3) \\
&\quad \times \mathbb{E}\{(\boldsymbol{V}_{X,i}^{\mathrm{T}}\boldsymbol{V}_{X,j})^2(\boldsymbol{V}_{X,k}^{\mathrm{T}}\boldsymbol{V}_{X,l})^2\}].
\end{aligned}$$

Using the same proof techniques as in Cheng et al. (2018), we can get the following equations:

$$\begin{aligned}
&\mathbb{E}\{(\boldsymbol{V}_{X,i}^{\mathrm{T}}\boldsymbol{V}_{X,j})^4\} \\
&= \frac{1}{4}p^2(p+2)^{-2}\{3\mathrm{tr}^2(\boldsymbol{\Lambda}_1^2) + 6\mathrm{tr}(\boldsymbol{\Lambda}_1^4)\}\{1 + o(1)\},
\end{aligned}$$

$$\mathbb{E}\{(\boldsymbol{V}_{\boldsymbol{X},i}^{\mathrm{T}}\boldsymbol{V}_{\boldsymbol{X},j})^2\}$$
$$= \frac{1}{2}\operatorname{tr}(\S_1^2)\{1+o(1)\} = p^{-2}\operatorname{tr}(\boldsymbol{\Lambda}_1^2)\{1+o(1)\},$$
$$\mathbb{E}\{(\boldsymbol{V}_{\boldsymbol{X},i}^{\mathrm{T}}\boldsymbol{V}_{\boldsymbol{X},j})^2(\boldsymbol{V}_{\boldsymbol{X},i}^{\mathrm{T}}\boldsymbol{V}_{\boldsymbol{X},k})^2\}$$
$$= \frac{1}{4}p^{-3}(p+2)^{-1}\{\operatorname{tr}^2(\boldsymbol{\Lambda}_1^2)+2\operatorname{tr}(\boldsymbol{\Lambda}_1^4)\}\{1+o(1)\}.$$

On this ground, we have that

$$\operatorname{var}(A_{n_1}) = \left\{\frac{4}{n_1(n_1-1)}\frac{\operatorname{tr}^2(\boldsymbol{\Lambda}_1^2)}{p^2(p+2)^2} + \frac{8}{n_1}\frac{p\operatorname{tr}(\boldsymbol{\Lambda}_1^4)-\operatorname{tr}^2(\boldsymbol{\Lambda}_1^2)}{p^4(p+2)}\right\} \times\{1+o(1)\}.$$

Similarly, we have that

$$\operatorname{var}(B_{n_2}) = \left(\frac{4}{n_2(n_2-1)}\frac{\operatorname{tr}^2(\boldsymbol{\Lambda}_2^2)}{p^2(p+2)^2} + \frac{8}{n_2}\frac{p\operatorname{tr}(\boldsymbol{\Lambda}_2^4)-\operatorname{tr}^2(\boldsymbol{\Lambda}_2^2)}{p^4(p+2)}\right) \times\{1+o(1)\},$$

$$\operatorname{var}(C_{n_1,n_2}) = \left[\frac{2}{n_1n_2}\frac{\operatorname{tr}^2(\boldsymbol{\Lambda}_1\boldsymbol{\Lambda}_2)}{p^2(p+2)^2} + \left(\frac{2}{n_1}+\frac{2}{n_2}\right) \times \frac{p\operatorname{tr}\{(\boldsymbol{\Lambda}_1\boldsymbol{\Lambda}_2)^2\}-\operatorname{tr}^2(\boldsymbol{\Lambda}_1\boldsymbol{\Lambda}_2)}{p^4(p+2)}\right]\{1+o(1)\},$$

$$\operatorname{cov}(A_{n_1},C_{n_1,n_2}) = \left\{\frac{4}{n_1}\frac{p\operatorname{tr}(\boldsymbol{\Lambda}_1^3\boldsymbol{\Lambda}_2)-\operatorname{tr}(\boldsymbol{\Lambda}_1\boldsymbol{\Lambda}_2)\operatorname{tr}(\boldsymbol{\Lambda}_1^2)}{p^4(p+2)}\right\}\{1+o(1)\},$$

$$\operatorname{cov}(B_{n_2},C_{n_1,n_2}) = \left\{\frac{4}{n_2}\frac{p\operatorname{tr}(\boldsymbol{\Lambda}_2^3\boldsymbol{\Lambda}_1)-\operatorname{tr}(\boldsymbol{\Lambda}_1\boldsymbol{\Lambda}_2)\operatorname{tr}(\boldsymbol{\Lambda}_2^2)}{p^4(p+2)}\right\}\{1+o(1)\}.$$

To sum up, we conclude that $\operatorname{var}(T_{\mathrm{HT}}) = \sigma_n^2\{1+o(1)\}$.

Define a sequence of random variables $\{\mathbf{z}_1,\ldots,\mathbf{z}_{n_1+n_2}\}$ as follows:

$$\mathbf{z}_i = \boldsymbol{V}_{\boldsymbol{X},i} \text{ for each } i\in\{1,\ldots,n_1\} \quad \text{and}$$
$$\mathbf{z}_{n_1+j} = \boldsymbol{V}_{\boldsymbol{Y},j} \text{ for each } j\in\{1,\ldots,n_2\}.$$

Let $\mathbb{E}_k(\cdot)$ denote the conditional expectation conditional on $\{\mathbf{z}_1,\ldots,\mathbf{z}_k\}$. Define $D_{n,k} = p^{-1}\{\mathbb{E}_k(T_{\mathrm{HT}})-\mathbb{E}_{k-1}(T_{\mathrm{HT}})\}$, then $p^{-1}\{T_{\mathrm{HT}}-\mathbb{E}(T_{\mathrm{HT}})\} = \sum_{k=1}^{n_1+n_2} D_{n,k}$. As a result, the sequence $\{D_{n,1},\ldots,D_{n,n_1+n_2}\}$ constitutes a martingale difference with respect to the σ-fields $\sigma(\mathbf{z}_1,\mathbf{z}_2,\ldots,\mathbf{z}_k)$. To use the martingale central limit theorem, we need to get the following results first:

$$p^2\sum_{k=1}^{n_1+n_2}\sigma_{n,k}^2/\operatorname{var}(T_{\mathrm{HT}}) \xrightarrow{p} 1 \quad \text{and}$$
$$\sum_{k=1}^{n_1+n_2}\mathbb{E}(D_{nk}^4) = p^{-4}o\{\operatorname{var}^2(T_{\mathrm{HT}})\}, \tag{A1}$$

where $\sigma_{n,k}^2 \doteq \mathbb{E}_{k-1}(D_{n,k}^2)$.

Proof of the first part of (A1): As $\mathbb{E}(\sum_{k=1}^{n_1+n_2}\sigma_{n,k}^2) = p^{-2}\times \operatorname{var}(T_{\mathrm{HT}})$, we only need to show that as $\min\{n_1,n_2\}\to\infty$, $\operatorname{var}(\sum_{k=1}^{n_1+n_2}\sigma_{n,k}^2) = o\{p^{-4}\operatorname{var}^2(T_{\mathrm{HT}})\}$. Define $\boldsymbol{\Gamma}_{1,k-1} = \sum_{i=1}^{k-1}(\mathbf{z}_i\mathbf{z}_i'-\mathbf{B}_1)$ for each $k\in\{1,\ldots,n_1-1\}$, and define $\boldsymbol{\Gamma}_{2,n_1+l-1} = \sum_{i=1}^{l-1}(\mathbf{z}_{n_1+i}\mathbf{z}_{n_1+i}'-\mathbf{B}_2)$ for each $l\in\{1,\ldots,n_2-1\}$. For each $k\in\{1,\ldots,n_1\}$, we have that

$$(\mathbb{E}_k-\mathbb{E}_{k-1})(A_{n_1}) = \frac{2}{n_1(n_1-1)}\{\boldsymbol{V}_{\boldsymbol{X},k}^{\mathrm{T}}\boldsymbol{\Gamma}_{1,k-1}\boldsymbol{V}_{\boldsymbol{X},k}-\operatorname{tr}(\boldsymbol{\Gamma}_{1,k-1}\mathbf{B}_1)\} + \frac{2}{n_1}\{\boldsymbol{V}_{\boldsymbol{X},k}^{\mathrm{T}}\mathbf{B}_1\boldsymbol{V}_{\boldsymbol{X},k}-\operatorname{tr}(\mathbf{B}_1^2)\},$$
$$(\mathbb{E}_k-\mathbb{E}_{k-1})(B_{n_2}) = 0,$$

and

$$(\mathbb{E}_k-\mathbb{E}_{k-1})(C_{n_1,n_2}) = \frac{1}{n_1}\{\boldsymbol{V}_{\boldsymbol{X},k}^{\mathrm{T}}\mathbf{B}_2\boldsymbol{V}_{\boldsymbol{X},k}-\operatorname{tr}(\mathbf{B}_1\mathbf{B}_2)\}.$$

For each $k\in\{n_1+1,\ldots,n_1+n_2\}$, we have that

$$(\mathbb{E}_k-\mathbb{E}_{k-1})(A_{n_1}) = 0,$$

$$(\mathbb{E}_k-\mathbb{E}_{k-1})(B_{n_2}) = \frac{2}{n_2(n_2-1)}\{\boldsymbol{V}_{\boldsymbol{Y},k-n_1}^{\mathrm{T}}\boldsymbol{\Gamma}_{2,k-1}\boldsymbol{V}_{\boldsymbol{Y},k-n_1}-\operatorname{tr}(\boldsymbol{\Gamma}_{2,k-1}\mathbf{B}_2)\} + \frac{2}{n_2}\{\boldsymbol{V}_{\boldsymbol{Y},k-n_1}^{\mathrm{T}}\mathbf{B}_2\boldsymbol{V}_{\boldsymbol{Y},k-n_1}-\operatorname{tr}(\mathbf{B}_2^2)\},$$

and

$$(\mathbb{E}_k-\mathbb{E}_{k-1})(C_{n_1,n_2}) = \frac{1}{n_1n_2}\left\{\boldsymbol{V}_{\boldsymbol{Y},k-n_1}^{\mathrm{T}}\left(\sum_{i=1}^{n_1}\boldsymbol{V}_{\boldsymbol{X},i}\boldsymbol{V}_{\boldsymbol{X},i}^{\mathrm{T}}\right)\boldsymbol{V}_{\boldsymbol{Y},k-n_1} - \operatorname{tr}\left(\sum_{i=1}^{n_1}\boldsymbol{V}_{\boldsymbol{X},i}\boldsymbol{V}_{\boldsymbol{X},i}^{\mathrm{T}}\mathbf{B}_2\right)\right\}.$$

Thus, for each $k\in\{1,\ldots,n_1\}$,

$$\begin{aligned}\sigma_{n,k}^2 &= \mathbb{E}_{k-1}\left(\left[\frac{2}{n_1(n_1-1)}\{\boldsymbol{V}_{\boldsymbol{X},k}^{\mathrm{T}}\boldsymbol{\Gamma}_{1,k-1}\boldsymbol{V}_{\boldsymbol{X},k}-\operatorname{tr}(\boldsymbol{\Gamma}_{1,k-1}\mathbf{B}_1)\}\right.\right.\\ &\quad +\frac{2}{n_1}\{\boldsymbol{V}_{\boldsymbol{X},k}^{\mathrm{T}}\mathbf{B}_1\boldsymbol{V}_{\boldsymbol{X},k}-\operatorname{tr}(\mathbf{B}_1^2)\}\\ &\quad \left.\left.-\frac{2}{n_1}\{\boldsymbol{V}_{\boldsymbol{X},k}^{\mathrm{T}}\mathbf{B}_2\boldsymbol{V}_{\boldsymbol{X},k}-\operatorname{tr}(\mathbf{B}_1\mathbf{B}_2)\}\right]^2\right)\\ &= \left(\frac{8}{n_1^2(n_1-1)^2}\frac{p\operatorname{tr}(\boldsymbol{\Gamma}_{1,k-1}\boldsymbol{\Lambda}_1)^2-\operatorname{tr}^2(\boldsymbol{\Gamma}_{1,k-1}\boldsymbol{\Lambda}_1)}{p^2(p+2)}\right.\\ &\quad +\frac{16}{n_1^2(n_1-1)}\frac{p\operatorname{tr}(\boldsymbol{\Gamma}_{1,k-1}\boldsymbol{\Lambda}_1^3)-\operatorname{tr}(\boldsymbol{\Gamma}_{1,k-1}\boldsymbol{\Lambda}_1)\operatorname{tr}(\boldsymbol{\Lambda}_1^2)}{p^3(p+2)}\\ &\quad -\frac{16}{n_1^2(n_1-1)}\\ &\quad \times\frac{p\operatorname{tr}(\boldsymbol{\Gamma}_{1,k-1}\boldsymbol{\Lambda}_1\boldsymbol{\Lambda}_2\boldsymbol{\Lambda}_1)-\operatorname{tr}(\boldsymbol{\Gamma}_{1,k-1}\boldsymbol{\Lambda}_1)\operatorname{tr}(\boldsymbol{\Lambda}_1\boldsymbol{\Lambda}_2)}{p^3(p+2)}\\ &\quad \left.+\frac{8}{n_1^2}\frac{p\operatorname{tr}[\{\boldsymbol{\Lambda}_1(\boldsymbol{\Lambda}_1-\boldsymbol{\Lambda}_2)\}^2]-\operatorname{tr}^2\{\boldsymbol{\Lambda}_1(\boldsymbol{\Lambda}_1-\boldsymbol{\Lambda}_2)\}}{p^4(p+2)}\right)\\ &\quad \times\{1+o(1)\},\end{aligned}$$

and for each $k\in\{n_1+1,\ldots,n_1+n_2\}$,

$$\begin{aligned}\sigma_{n,k}^2 &= \mathbb{E}_{k-1}\left(\left[\frac{2}{n_2(n_2-1)}\{\boldsymbol{V}_{\boldsymbol{Y},k-n_1}^{\mathrm{T}}\boldsymbol{\Gamma}_{2,k-1}\boldsymbol{V}_{\boldsymbol{Y},k-n_1}\right.\right.\\ &\quad \left.-\operatorname{tr}(\boldsymbol{\Gamma}_{2,k-1}\mathbf{B}_2)\}\right.\\ &\quad \left.\left.+\frac{2}{n_2}\{\boldsymbol{V}_{\boldsymbol{Y},k-n_1}^{\mathrm{T}}\mathbf{B}_2\boldsymbol{V}_{\boldsymbol{Y},k-n_1}-\operatorname{tr}(\mathbf{B}_2^2)\}\right]^2\right)\end{aligned}$$

$$
\begin{aligned}
&+\frac{2}{n_1 n_2}\left\{\boldsymbol{V}_{\boldsymbol{Y},k-n_1}^{\mathrm{T}}\left(\sum_{i=1}^{n_1}\boldsymbol{V}_{\boldsymbol{X},i}\boldsymbol{V}_{\boldsymbol{X},i}^{\mathrm{T}}\right)\boldsymbol{V}_{\boldsymbol{Y},k-n_1}\right.\\
&\left.\left.-\mathrm{tr}\left(\sum_{i=1}^{n_1}\boldsymbol{V}_{\boldsymbol{X},i}\boldsymbol{V}_{\boldsymbol{X},i}^{\mathrm{T}}\boldsymbol{B}_2\right)\right\}\right]^2\Bigg)\\
&=\Bigg[\frac{8}{n_2^2(n_2-1)^2}\frac{p\,\mathrm{tr}(\boldsymbol{\Gamma}_{2,k-1}\boldsymbol{\Lambda}_2)^2-\mathrm{tr}^2(\boldsymbol{\Gamma}_{2,k-1}\boldsymbol{\Lambda}_2)}{p^2(p+2)}\\
&+\frac{16}{n_2^2(n_2-1)}\frac{p\,\mathrm{tr}(\boldsymbol{\Gamma}_{2,k-1}\boldsymbol{\Lambda}_2^3)-\mathrm{tr}(\boldsymbol{\Gamma}_{2,k-1}\boldsymbol{\Lambda}_2)\,\mathrm{tr}(\boldsymbol{\Lambda}_2^2)}{p^3(p+2)}\\
&-\frac{16}{n_1 n_2^2(n_2-1)}\\
&\times\frac{\mathrm{tr}\{\boldsymbol{\Gamma}_{2,k-1}\boldsymbol{\Lambda}_2(\sum_{i=1}^{n_1}\boldsymbol{V}_{\boldsymbol{X},i}\boldsymbol{V}_{\boldsymbol{X},i}^{\mathrm{T}})\boldsymbol{\Lambda}_2\}}{p^2(p+2)}\\
&-\frac{\mathrm{tr}(\boldsymbol{\Gamma}_{2,k-1}\boldsymbol{\Lambda}_2)\,\mathrm{tr}\{\boldsymbol{\Lambda}_2(\sum_{i=1}^{n_1}\boldsymbol{V}_{\boldsymbol{X},i}\boldsymbol{V}_{\boldsymbol{X},i}^{\mathrm{T}})\}}{p^2(p+2)}\\
&+\frac{8}{n_2^2}\frac{p\,\mathrm{tr}(\boldsymbol{\Lambda}_2^4)-\mathrm{tr}^2(\boldsymbol{\Lambda}_2^2)}{p^4(p+2)}\\
&-\frac{16}{n_1 n_2^2}\frac{p\,\mathrm{tr}(\sum_{i=1}^{n_1}\boldsymbol{V}_{\boldsymbol{X},i}\boldsymbol{V}_{\boldsymbol{X},i}^{\mathrm{T}}\boldsymbol{\Lambda}_2^3)}{p^3(p+2)}\\
&-\frac{\mathrm{tr}(\sum_{i=1}^{n_1}\boldsymbol{V}_{\boldsymbol{X},i}\boldsymbol{V}_{\boldsymbol{X},i}^{\mathrm{T}}\boldsymbol{\Lambda}_2)\,\mathrm{tr}(\boldsymbol{\Lambda}_2^2)}{p^3(p+2)}\\
&+\frac{8}{n_1^2 n_2^2}\frac{p\,\mathrm{tr}\{(\sum_{i=1}^{n_1}\boldsymbol{V}_{\boldsymbol{X},i}\boldsymbol{V}_{\boldsymbol{X},i}^{\mathrm{T}}\boldsymbol{\Lambda}_2)^2\}}{p^2(p+2)}\\
&-\frac{\mathrm{tr}^2(\boldsymbol{\Lambda}_2\sum_{i=1}^{n_1}\boldsymbol{V}_{\boldsymbol{X},i}\boldsymbol{V}_{\boldsymbol{X},i}^{\mathrm{T}})}{p^2(p+2)}\Bigg]\{1+o(1)\}.
\end{aligned}
$$

Hence

$$
\begin{aligned}
\sum_{k=1}^{n_1+n_2}\sigma_{n,k}^2&=(R_1+R_2+R_3+R_4+R_5\\
&\quad+R_6+R_6+R_7+C_0)\{1+o(1)\},
\end{aligned}
$$

where C_0 is a constant, and

$$
\begin{aligned}
R_1&=\sum_{k=1}^{n_1}\frac{8}{n_1^2(n_1-1)^2}\\
&\quad\times\frac{p\,\mathrm{tr}\{(\boldsymbol{\Gamma}_{1,k-1}\boldsymbol{\Lambda}_1)^2\}-\mathrm{tr}^2(\boldsymbol{\Gamma}_{1,k-1}\boldsymbol{\Lambda}_1)}{p^2(p+2)},\\
R_2&=\sum_{l=1}^{n_2}\frac{8}{n_2^2(n_2-1)^2}\\
&\quad\times\frac{p\,\mathrm{tr}\{(\boldsymbol{\Gamma}_{2,n_1+l-1}\boldsymbol{\Lambda}_2)^2\}-\mathrm{tr}^2(\boldsymbol{\Gamma}_{2,n_1+l-1}\boldsymbol{\Lambda}_2)}{p^2(p+2)},\\
R_3&=\sum_{k=1}^{n_1}\frac{16}{n_1^2(n_1-1)}\\
&\quad\times\Bigg[\frac{p\,\mathrm{tr}\{\boldsymbol{\Gamma}_{1,k-1}(\boldsymbol{\Lambda}_1^3-\boldsymbol{\Lambda}_1\boldsymbol{\Lambda}_2\boldsymbol{\Lambda}_1)\}}{p^3(p+2)},\\
&\qquad-\frac{\mathrm{tr}(\boldsymbol{\Gamma}_{1,k-1}\boldsymbol{\Lambda}_1)\,\mathrm{tr}\{\boldsymbol{\Lambda}_1(\boldsymbol{\Lambda}_1-\boldsymbol{\Lambda}_2)\}}{p^3(p+2)}\Bigg],\\
R_4&=\sum_{l=1}^{n_2}\frac{16}{n_2^2(n_2-1)}\\
&\quad\times\frac{p\,\mathrm{tr}(\boldsymbol{\Gamma}_{2,n_1+l-1}\boldsymbol{\Lambda}_2^3)-\mathrm{tr}(\boldsymbol{\Gamma}_{2,n_1+l-1}\boldsymbol{\Lambda}_2)\,\mathrm{tr}(\boldsymbol{\Lambda}_2^2)}{p^3(p+2)},\\
R_5&=-\sum_{l=1}^{n_2}\frac{16}{n_1 n_2^2(n_2-1)}\\
&\quad\times\Bigg[\frac{p\,\mathrm{tr}\{\boldsymbol{\Gamma}_{2,n_1+l-1}\boldsymbol{\Lambda}_2(\sum_{i=1}^{n_1}\boldsymbol{V}_{\boldsymbol{X},i}\boldsymbol{V}_{\boldsymbol{X},i}^{\mathrm{T}})\boldsymbol{\Lambda}_2\}}{p^2(p+2)}\\
&\qquad-\frac{\mathrm{tr}(\boldsymbol{\Gamma}_{2,n_1+l-1}\boldsymbol{\Lambda}_2)\,\mathrm{tr}\{\boldsymbol{\Lambda}_2(\sum_{i=1}^{n_1}\boldsymbol{V}_{\boldsymbol{X},i}\boldsymbol{V}_{\boldsymbol{X},i}^{\mathrm{T}})\}}{p^2(p+2)}\Bigg],\\
R_6&=\sum_{l=1}^{n_2}\frac{8}{n_1^2 n_2^2}\frac{p\,\mathrm{tr}\{(\sum_{i=1}^{n_1}\boldsymbol{V}_{\boldsymbol{X},i}\boldsymbol{V}_{\boldsymbol{X},i}^{\mathrm{T}}\boldsymbol{\Lambda}_2)^2\}}{p^2(p+2)}\\
&\qquad-\frac{\mathrm{tr}^2(\boldsymbol{\Lambda}_2\sum_{i=1}^{n_1}\boldsymbol{V}_{\boldsymbol{X},i}\boldsymbol{V}_{\boldsymbol{X},i}^{\mathrm{T}})}{p^2(p+2)},\\
R_7&=\sum_{l=1}^{n_2}-\frac{16}{n_1 n_2^2}\frac{p\,\mathrm{tr}(\sum_{i=1}^{n_1}\boldsymbol{V}_{\boldsymbol{X},i}\boldsymbol{V}_{\boldsymbol{X},i}^{\mathrm{T}}\boldsymbol{\Lambda}_2^3)}{p^3(p+2)}\\
&\qquad-\frac{\mathrm{tr}(\sum_{i=1}^{n_1}\boldsymbol{V}_{\boldsymbol{X},i}\boldsymbol{V}_{\boldsymbol{X},i}^{\mathrm{T}}\boldsymbol{\Lambda}_2)\,\mathrm{tr}(\boldsymbol{\Lambda}_2^2)}{p^3(p+2)}.
\end{aligned}
$$

Moreover, to calculate the order of $\mathrm{var}(R_1)$, we need to evaluate $\mathrm{var}[\sum_{k=1}^{n_1}p^{-2}\,\mathrm{tr}\{(\boldsymbol{\Gamma}_{1,k-1}\boldsymbol{\Lambda}_1)^2\}]$. Since

$$
\begin{aligned}
&\mathbb{E}\left(\left[p^{-2}\sum_{k=1}^{n_1}\sum_{i=1}^{k-1}\sum_{j=1}^{k-1}\{(\boldsymbol{V}_{\boldsymbol{X},i}^{\mathrm{T}}\boldsymbol{\Lambda}_1\boldsymbol{V}_{\boldsymbol{X},j})^2-\mathrm{tr}(\boldsymbol{\Lambda}_1\boldsymbol{B}_1)^2\}\right]^2\right)\\
&=p^{-4}\sum_{k=1}^{n_1}\sum_{m=1}^{n_1}\mathbb{E}\left[\sum_{i=1}^{k-1}\sum_{j=1}^{k-1}\{(\boldsymbol{V}_{\boldsymbol{X},i}^{\mathrm{T}}\boldsymbol{\Lambda}_1\boldsymbol{V}_{\boldsymbol{X},j})^2-\mathrm{tr}(\boldsymbol{\Lambda}_1\boldsymbol{B}_1)^2\}\right.
\end{aligned}
$$

$$\times \sum_{l=1}^{m-1}\sum_{h=1}^{m-1}\{(\boldsymbol{V}_{\boldsymbol{X},l}^{\mathrm{T}}\boldsymbol{\Lambda}_1\boldsymbol{V}_{\boldsymbol{X},h})^2 - \mathrm{tr}(\boldsymbol{\Lambda}_1\mathbf{B}_1)^2\}\Bigg]$$

$$\le n_1\sum_{k=1}^{n_1}(k-1)^2p^{-4}\mathbb{E}\{(\boldsymbol{V}_{\boldsymbol{X},i}^{\mathrm{T}}\boldsymbol{\Lambda}_1\boldsymbol{V}_{\boldsymbol{X},j})^4 - \mathrm{tr}^2(\boldsymbol{\Lambda}_1\mathbf{B}_1)^2\} + n_1^2\sum_{k=1}^{n_1}(k-1)^2p^{-4}.$$

$$\mathbb{E}\{(\boldsymbol{V}_{\boldsymbol{X},i}^{\mathrm{T}}\boldsymbol{\Lambda}_1\boldsymbol{V}_{\boldsymbol{X},j})^2(\boldsymbol{V}_{\boldsymbol{X},i}^{\mathrm{T}}\boldsymbol{\Lambda}_1\boldsymbol{V}_{\boldsymbol{X},l})^2 - \mathrm{tr}^2(\boldsymbol{\Lambda}_1\mathbf{B}_1)^2\}$$
$$= C_1n_1^4p^{-4}p^{-4}\{\mathrm{tr}^2(\boldsymbol{\Lambda}_1^4) + \mathrm{tr}(\boldsymbol{\Lambda}_1^8)\} + C_2n_1^5p^{-4}p^{-4}\{\mathrm{tr}(\boldsymbol{\Lambda}_1^8) - p^{-1}\,\mathrm{tr}^2(\boldsymbol{\Lambda}_1^4)\},$$

where C_1 and C_2 are constants, we have that

$$\mathrm{var}(R_1) \le C_1n_1^{-4}p^{-8}\,\mathrm{tr}^2(\boldsymbol{\Lambda}_1^2)\,\mathrm{tr}(\boldsymbol{\Lambda}_1^4) + C_2n_1^{-3}p^{-8}\,\mathrm{tr}(\boldsymbol{\Lambda}_1^4)\{\mathrm{tr}(\boldsymbol{\Lambda}_1^4) - p^{-1}\,\mathrm{tr}^2(\boldsymbol{\Lambda}_1^2)\}.$$

Based on the fact that $\mathrm{tr}(\boldsymbol{\Lambda}_1^4)/\mathrm{tr}^2(\boldsymbol{\Lambda}_1^2) \to 0$ and the following inequality

$$\mathrm{var}^2(T_{\mathrm{HT}}) \ge K\left\{\frac{\mathrm{tr}^4(\boldsymbol{\Lambda}_1^2)}{p^4n_1^4} + \frac{\mathrm{tr}^4(\boldsymbol{\Lambda}_2^2)}{p^4n_2^4}\right\}$$

for some constant K, we conclude that

$$p^4\mathrm{var}(R_1)/\mathrm{var}^2(T_{\mathrm{HT}}) \to 0,$$

which indicates that $\mathrm{var}(R_1) = o\{p^{-4}\mathrm{var}^2(T_{\mathrm{HT}})\}$. By using similar techniques, we conclude that $\mathrm{var}(R_l) = o\{p^{-4} \times \mathrm{var}^2(T_{\mathrm{HT}})\}$ for each $l \in \{1,\ldots,7\}$, based on which we finally conclude that $\mathrm{var}(\sum_{k=1}^{n_1+n_2}\sigma_{n,k}^2) = o\{p^{-4}\mathrm{var}^2(T_{\mathrm{HT}})\}$. ■

Proof of the second part of (A1): For $1 \le k \le n_1$,

$$\sum_{k=1}^{n_1}\mathbb{E}(D_{nk}^4)$$
$$= \sum_{k=1}^{n_1}\mathbb{E}\left(\left[\frac{2}{n_1(n_1-1)}\{\boldsymbol{V}_{\boldsymbol{X},k}^{\mathrm{T}}\boldsymbol{\Gamma}_{1,k-1}\boldsymbol{V}_{\boldsymbol{X},k} - \mathrm{tr}(\boldsymbol{\Gamma}_{1,k-1}\mathbf{B}_1)\} + \frac{2}{n_1}\{\boldsymbol{V}_{\boldsymbol{X},k}^{\mathrm{T}}\mathbf{B}_1\boldsymbol{V}_{\boldsymbol{X},k} - \mathrm{tr}(\mathbf{B}_1^2)\} - \frac{2}{n_1}\{\boldsymbol{V}_{\boldsymbol{X},k}^{\mathrm{T}}\mathbf{B}_2\boldsymbol{V}_{\boldsymbol{X},k} - \mathrm{tr}(\mathbf{B}_1\mathbf{B}_2)\}\right]^4\right)$$
$$\le c_1\left\{n_1^{-3}p^{-8}\,\mathrm{tr}[\{\boldsymbol{\Lambda}_1(\boldsymbol{\Lambda}_1-\boldsymbol{\Lambda}_2)\}^2]\left(\mathrm{tr}[\{\boldsymbol{\Lambda}_1(\boldsymbol{\Lambda}_1-\boldsymbol{\Lambda}_2)\}^2] - p^{-1}\,\mathrm{tr}^2\{\boldsymbol{\Lambda}_1(\boldsymbol{\Lambda}_1-\boldsymbol{\Lambda}_2)\}\right) + n_1^{-5}p^{-8}\,\mathrm{tr}^4(\boldsymbol{\Lambda}_1^2)\right\},$$

where c_1 is some constant. Then

$$p^4\sum_{k=1}^{n_1+n_2}\mathbb{E}(D_{nk}^4)/\mathrm{var}^2(T_{\mathrm{HT}}) \to 0.$$

Similarly, for $n_1 \le k \le n_1+n_2$,

$$\sum_{k=n_1}^{n_1+n_2}\mathbb{E}(D_{nk}^4)$$
$$\le c_2\left(n_1^{-1}n_2^{-4}p^{-8}\,\mathrm{tr}^2(\boldsymbol{\Lambda}_1\boldsymbol{\Lambda}_2)\,\mathrm{tr}\{\boldsymbol{\Lambda}_2(\boldsymbol{\Lambda}_1-\boldsymbol{\Lambda}_2)\}^2 \times [\mathrm{tr}\{\boldsymbol{\Lambda}_2(\boldsymbol{\Lambda}_1-\boldsymbol{\Lambda}_2)\}^2 - p^{-1}\,\mathrm{tr}^2\{\boldsymbol{\Lambda}_2(\boldsymbol{\Lambda}_1-\boldsymbol{\Lambda}_2)\}] + n_2^{-5}p^{-8}\,\mathrm{tr}^4(\boldsymbol{\Lambda}_2^2) + n_1^{-2}n_2^{-4}p^{-8}\,\mathrm{tr}^4(\boldsymbol{\Lambda}_1\boldsymbol{\Lambda}_2)\right),$$

where c_2 is some constant. Then we have that as $n_1, n_2 \to \infty$,

$$\frac{p^4\sum_{k=1}^{n_1+n_2}\mathbb{E}(D_{nk}^4)}{\mathrm{var}^2(T_{\mathrm{HT}})} \to 0.$$

By using the martingale central limit theorem (Hall & Hyde, 1980), we finally conclude that

$$\frac{T_{\mathrm{HT}} - \mathbb{E}(T_{\mathrm{HT}})}{\mathrm{var}(T_{\mathrm{HT}})} \xrightarrow{d} N(0,1).$$

■

Proof of Proposition 2.1:

Using the same techniques as in the proof of Theorem 2.1, we have that $\mathbb{E}(\mathbf{A}_1) = \mathrm{tr}^2(\S_1)$ and

$$\mathrm{Var}\left\{\frac{p^2\mathbf{A}_1}{\mathrm{tr}(\boldsymbol{\Lambda}_1^2)}\right\} = O\left(\frac{p^4}{\mathrm{tr}^2\boldsymbol{\Lambda}_1^2}\left[\frac{2}{n_1(n_1-1)} \times \left\{\frac{3\mathrm{tr}^2(\boldsymbol{\Lambda}_1^2) + 6\mathrm{tr}(\boldsymbol{\Lambda}_1^4)}{p^2(p+2)^2} - \frac{\mathrm{tr}^2(\boldsymbol{\Lambda}_1^2)}{p^4}\right\}\right]\right).$$

Therefore, $\frac{p^2\mathbf{A}_1}{\mathrm{tr}(\boldsymbol{\Lambda}_1^2)} \to 1$, and similarly, $\frac{p^2\mathbf{A}_2}{\mathrm{tr}(\boldsymbol{\Lambda}_2^2)} \to 1$, hence $\frac{\hat{\sigma}_{0,n}^2}{\sigma_{0,n}^2} \to 1$.