Statistica Sinica **25** (2015), 1297-1312
doi:http://dx.doi.org/10.5705/ss.2014.048

TWO-SAMPLE BEHRENS-FISHER PROBLEM FOR HIGH-DIMENSIONAL DATA

Long Feng[1], Changliang Zou[1], Zhaojun Wang[1] and Lixing Zhu[2]

[1]*Nankai University and* [2]*Hong Kong Baptist University*

Abstract: This article is concerned with the two-sample Behrens-Fisher problem in high-dimensional settings. A test is proposed that is scale-invariant, asymptotically normal under certain mild conditions, and the dimensionality is allowed to grow in the rate, respectively, from square to cube of the sample size in different scenarios. We explain the necessity of bias correction for existing scale-invariant tests. We also give some examples to show the advantage of the scale-invariant test over scale-variant tests when variances of the two samples are different.

Key words and phrases: Asymptotic normality, Behrens-Fisher problem, high-dimensional data, large-p-small-n, two-sample test.

1. Introduction

We are concerned with the two-sample Behrens-Fisher problem in high-dimensional settings. Assume that $\{\mathbf{X}_{i1}, \ldots, \mathbf{X}_{in_i}\}$ for $i = 1, 2$ are two independent random samples with the sizes n_1 and n_2, from p-variate distributions $F(\mathbf{x}-\boldsymbol{\mu}_1)$ and $G(\mathbf{x}-\boldsymbol{\mu}_2)$ located at p-variate centers $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$. Let $n = n_1+n_2$. We wish to test

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \quad \text{versus} \quad H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2, \tag{1.1}$$

when their respective covariances $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are unknown. For testing (1.1), Bai and Saranadasa (1996) proposed a test statistic based on $M_n = ||\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2||^2$, is developed when $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$. The key feature of the Bai and Saranadasa's proposal is to use the Euclidian norm to replace the Mahalanobis norm since having the inverse of the sample covariance matrix is no longer beneficial when $p/n \to c > 0$. Zhang and Xu (2009) extended this method to the k-sample high-dimensional Behrens-Fisher problem and derived the asymptotic distribution of the test statistic when $p/n \to c < 1$. To allow simultaneous testing for ultra high-dimensional data, Chen and Qin (2010) considered the test statistic

$$W_n = \frac{\sum_{i\neq j}^{n_1} \mathbf{X}_{1i}^T \mathbf{X}_{1j}}{n_1(n_1-1)} + \frac{\sum_{i\neq j}^{n_2} \mathbf{X}_{2i}^T \mathbf{X}_{2j}}{n_2(n_2-1)} - 2\frac{\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \mathbf{X}_{1i}^T \mathbf{X}_{2j}}{n_1 n_2} \tag{1.2}$$

by removing $\sum_{j=1}^{n_i} \mathbf{X}_{ij}^T \mathbf{X}_{ij}$ for $i = 1, 2$ from $||\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2||^2$; these terms impose demands on the dimensionality. The asymptotic normality of W_n is established without imposing any explicit connection between p and n. The restriction on the dimension is that the number of divergent eigenvalues of $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ is not too large, and the divergence rate is not too fast.

As a test statistic, an important requirement is scale-invariance, lest the test suffer from scalar transformations: the same dataset generating different conclusions due to different scalar transformations. This is an obvious limitation of W_n and M_n. Intuitively speaking, both tests take the sum of all p squared mean differences without using the information from the diagonal elements of the sample covariance, and thus their test power depend heavily on the underlying variance magnitudes. In practice, different components can have completely different physical or biological readings and scales that are not similar. It is desirable to develop scalar-transformation-invariant tests for the Behrens-Fisher problem that can integrate the individual information in a relatively "fair" way.

Under the normality, Srivastava and Du (2008) proposed a scalar-transformation-invariant test assuming equality of the covariance matrices. Srivastava, Katayama, and Kano (2013) extended their results to unequal covariance matrices. However, to derive the well-defined asymptotic null distribution, the dimension p must have a smaller order than n^2. Park and Ayyala (2013) also proposed a scale-invariant test from the idea of leave-out cross validation, but their test is not shift-invariant. The test is not powerful for significance level maintainance and power enhancement.

We propose test via standardizing each component of the difference of two-sample means by the corresponding variance and suggest a simple but effective test statistic. The proposed test is invariant under scalar transformations and its asymptotic normality can be derived under some very mild conditions similar to those in Chen and Qin (2010).

For large dimension p gets larger, the assumption that every component of $\mathbf{x}$ is standardized with variance 1 brings too many plug-in sample variances and seriously affects the asymptotic behaviors of the test. For p not of order n^2, existing scale-invariant tests suffer asymptotically from bias-terms, as the sample variance is only root-n consistent, bias cannot be eliminated asymptotically. Thus, when p is of the order n^2 or higher, a calibration or bias correction to make scale-invariant tests useful is needed.

The remainder of the paper is organized as follows. In the next section, the test statistic is constructed and its asymptotic normality is established. A bias correction to the expectation of the proposed test is developed and the associated plug-in estimators are suggested. Simulation comparisons are reported in Section 3. Section 4 contains a sensor detection example to illustrate the proposed test.

Several remarks in Section 5 conclude the paper. Technical details and some additional simulation results are provided in the Appendix in a supplementary file.

2. High-dimensional Two-sample Location Tests

For the univariate Behrens-Fisher problem, Fisher (1935, 1939) considered $\tau = (\bar{x}_1 - \bar{x}_2)/\sqrt{s_1^2/n_1 + s_2^2/n_2}$, where $\bar{x}_i$ and s_i are the i-th sample mean and standard deviation, respectively. For the multivariate Behrens-Fisher problem, James (1954), Yao (1965) and Johansen (1980) proposed test procedures based on:

$$Q_1 = (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)^T \left(\frac{\mathbf{S}_1}{n_1} + \frac{\mathbf{S}_2}{n_2} \right)^{-1} (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2),$$

where $\bar{\mathbf{X}}_i$ and $\mathbf{S}_i$ are the i-th sample mean and covariance matrices, respectively. When the dimension $p > n$, the sample covariance matrices $\mathbf{S}_1$ and $\mathbf{S}_2$ are not positive definite and, accordingly, these test procedures are no longer feasible. Zhang and Xu (2009) simply removed the term $(\mathbf{S}_1/n_1 + \mathbf{S}_2/n_2)^{-1}$ in Q_1. Their test statistic is essentially an L^2-norm and thus not scale-invariant. Although having the inverse of the sample covariance matrix is no longer beneficial when $p/n \to c > 0$ (Bai and Saranadasa (1996)), the information provided by the sample variances should still be useful. Consider

$$Q_2 = (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)^T \left(\frac{\mathbf{D}_1}{n_1} + \frac{\mathbf{D}_2}{n_2} \right)^{-1} (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2),$$

where $\mathbf{D}_i$ is the diagonal matrix of $\mathbf{S}_i$. Q_2 can be written as

$$Q_2 = \sum_{k=1}^{p} \frac{(\hat{\mu}_{1k} - \hat{\mu}_{2k})^2}{\hat{\sigma}_{1k}^2/n_1 + \hat{\sigma}_{2k}^2/n_2},$$

where $\hat{\mu}_{ik}$ and $\hat{\sigma}_{ik}$ are the sample mean and standard deviation of the k-th variable for the i-th sample, respectively. Thus, Q_2 is simply a sum of the p squared univariate Fisher's test statistics. Srivastava, Katayama, and Kano (2013) used the standardized Q_2 as the test statistic, $\hat{q}_n = p^{-1/2}(Q_2 - p)$. If $n = O(p^\delta)$, $\delta > 0.5$ and under normality, this statistic has mean zero asymptotically through the asymptotic equivalence: $\hat{q}_n \to \tilde{q}_n$ in probability, where

$$\tilde{q}_n = p^{-1/2} \left(\sum_{k=1}^{p} \frac{(\hat{\mu}_{1k} - \hat{\mu}_{2k})^2}{\sigma_{1k}^2/n_1 + \sigma_{2k}^2/n_2} - p \right).$$

Srivastava, Katayama, and Kano (2013) claimed in their setting that Q_2 is almost identical to the statistic proposed by Chen and Qin (2010). When p is of

the order of n^2 or higher, this is not true as Q_2 has well-defined limit under the null, the equivalence between $\hat{q}_n$ and $\tilde{q}_n$ is no longer true; there is a non-negligible bias-term between $\hat{q}_n$ and $\tilde{q}_n$, as shown below in a simple scenario. For $n_1 = n_2$ and $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{I}_p$, we have result about $\hat{q}_n$.

Proposition 1. *If Conditions* (C1)−(C5) *below hold, and* $p = n^{2+\alpha}$, $0 < \alpha < 1$, *then under* H_0, $P\left(\hat{q}_n/\sqrt{\mathrm{var}(\hat{q}_n)} > z_\alpha\right) \to 1$, *where* z_α *is the upper* α *quantile of* $N(0,1)$.

The justification follows, almost exactly, the same arguments used to prove Theorem 1 in the Appendix. We can show that $\{\hat{q}_n - E(\hat{q}_n)\}/\sqrt{\mathrm{var}(\hat{q}_n)} \xrightarrow{\mathcal{L}} N(0,1)$, with $E(\hat{q}_n) = 4p^{1/2}n^{-1} + p^{-1/2}n^{-1}\sum_{k=1}^{p}(\kappa_{1k}-\kappa_{2k})^2 + o(1)$ and $\mathrm{var}(\hat{q}_n) = 2 + o(1)$, where $\kappa_{ik} = E(X_{ijk} - \mu_{ik})^3$. Thus, $E(\hat{q}_n)/\sqrt{\mathrm{var}(\hat{q}_n)} = O(n^{\alpha/2})$ even in the normal case with $\sum_{k=1}^{p}(\kappa_{1k} - \kappa_{2k})^2 = 0$.

Thus, even in the simple case, the size of the Srivastava, Katayama, and Kano (2013) test is distorted when the dimension gets higher. They also proposed a ratio consistent estimator of $\widehat{\mathrm{var}}(\hat{q}_n)$ but when $p = O(n^{2+\alpha})$, $\alpha > 0$, their estimator is not ratio-consistent. They proposed a correction term $c_{p,n}$ to adjust the empirical size, but in finite sample cases we conducted, this size-adjusted value $c_{p,n}$ makes the variance estimator always larger than the asymptotic variance of $\hat{q}_n$; see Figure 2.

We can define a test that can be asymptotically unbiased. The details are as follows. As in Bai and Saranadasa (1996), the term $\sum_{j=1}^{n_i} X_{ijk}^2$ in $(\hat{\mu}_{1k} - \hat{\mu}_{2k})^2$ imposes demands on the dimensionality, where $\mathbf{X}_{ij} = (X_{ij1}, \ldots, X_{ijp})^T$. Motivated by Chen and Qin (2010), after removing the terms like $\sum_{j=1}^{n_i} X_{ijk}^2$ in $(\hat{\mu}_{1k} - \hat{\mu}_{2k})^2$, we take

$$Q_3 = \sum_{k=1}^{p} \frac{1}{\hat{\sigma}_{1k}^2/n_1 + \hat{\sigma}_{2k}^2/n_2} \left\{ \frac{1}{n_1(n_1-1)} \sum_{i \neq j} X_{1ik}X_{1jk} + \frac{1}{n_2(n_2-1)} \sum_{i \neq j} X_{2ik}X_{2jk} - \frac{2}{n_1 n_2} \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} X_{1ik}X_{2jk} \right\} \doteq \sum_{k=1}^{p} \frac{A_k}{\hat{\sigma}_{1k}^2/n_1 + \hat{\sigma}_{2k}^2/n_2}.$$

Here $\sum_{k=1}^{p} A_k = W_n$ is the test statistic proposed by Chen and Qin (2010). It is not scale-invariant and Q_3 scales each A_k in W_n by its corresponding variance estimator. A modification of Q_3 can be taken as an initial test statistic:

$$T_n = \frac{1}{n_1} Q_3 = \sum_{k=1}^{p} \frac{A_k}{\hat{\sigma}_{1k}^2 + \gamma\hat{\sigma}_{2k}^2},$$

where $\gamma = n_1/n_2$. It is called an initial test statistic because T_n cannot be directly used as the test for the hypotheses we want to check, but a standardized

version can. While $E(A_k)$ is zero, the expectation of T_n is not since A_k is not independent of $\hat{\sigma}_{1k}^2 + \gamma\hat{\sigma}_{2k}^2$. We need an analysis about T_n as to bias.

The tests statistics proposed by Bai and Saranadasa (1996) and Chen and Qin (2010) are invariant under orthogonal transformations, $\mathbf{X}_{ij} \to \mathbf{P}\mathbf{X}_{ij}$ where $\mathbf{P}$ is an orthogonal matrix. In contrast, T_n is not invariant under orthogonal transformations, but it is invariant under location shifts and scalar transformations, $\mathbf{X}_{ij} \to \mathbf{D}\mathbf{X}_{ij} + \mathbf{c}$ for $i = 1, 2, j = 1, \ldots, n_i$, where $\mathbf{c}$ is a constant vector, $\mathbf{D} = \text{diag}(d_1, \ldots, d_p)$, and $d_1, \ldots, d_p$ are non-zero constants.

As do Bai and Saranadasa (1996) and Chen and Qin (2010), we suppose

$$\mathbf{X}_{ij} = \mathbf{\Gamma}_i \mathbf{z}_{ij} + \boldsymbol{\mu}_i \quad \text{for} \quad j = 1, \ldots, n_i,\ i = 1, 2, \tag{2.1}$$

where each $\mathbf{\Gamma}_i$ is a $p \times m$ matrix for some $m \geq p$ such that $\mathbf{\Gamma}_i \mathbf{\Gamma}_i^T = \mathbf{\Sigma}_i$, and the $\{\mathbf{z}_{ij}\}_{j=1}^{n_i}$ are m-variate i.i.d. random vectors such that

$$\begin{gathered} E(\mathbf{z}_i) = 0,\ \text{var}(\mathbf{z}_i) = \mathbf{I}_m,\ E(z_{il}^4) = 3 + \Delta_i,\ E(z_{il}^8) = m_{i8} \in (0, \infty), \\ E(z_{ik_1}^{\alpha_1} z_{ik_2}^{\alpha_2} \cdots z_{ikq}^{\alpha_q}) = E(z_{ik_1}^{\alpha_1}) E(z_{ik_2}^{\alpha_2}) \cdots E(z_{ikq}^{\alpha_q}), \end{gathered} \tag{2.2}$$

for a positive integer q such that $\sum_{k=1}^{q} \alpha_k \leq 8$ and $k_1 \neq k_2 \cdots \neq k_q$. We need conditions as $n, p \to \infty$.

(C1) $n_1/(n_1 + n_2) \to \lambda \in (0, 1)$.

(C2) $\text{tr}\left(\mathbf{\Lambda}\mathbf{\Sigma}_i\mathbf{\Lambda}^2\mathbf{\Sigma}_j\mathbf{\Lambda}^2\mathbf{\Sigma}_l\mathbf{\Lambda}^2\mathbf{\Sigma}_h\mathbf{\Lambda}\right) = o(\text{tr}^2\{(\mathbf{\Lambda}\mathbf{\Sigma}_1\mathbf{\Lambda} + \mathbf{\Lambda}\mathbf{\Sigma}_2\mathbf{\Lambda})^2\})$ for $i, j, l, h = 1$ or 2, where $\mathbf{\Lambda} = \text{diag}\left\{(\sigma_{11}^2 + \gamma\sigma_{21}^2)^{-1/2}, \ldots, (\sigma_{1p}^2 + \gamma\sigma_{2p}^2)^{-1/2}\right\}$.

(C3) $p^2/n^5 \text{var}(T_n) \to 0$.

(C4) With $\mathbf{\Pi}_{1i} = E(\mathbf{\Lambda}(\mathbf{X}_{ij} - \boldsymbol{\mu}_i)\ (\mathbf{\Lambda}(\mathbf{X}_{ij} - \boldsymbol{\mu}_i))^{3^T})$ and $\mathbf{\Pi}_{2i} = E(\mathbf{\Lambda}(\mathbf{X}_{ij} - \boldsymbol{\mu}_i)(\mathbf{X}_{ij} - \boldsymbol{\mu}_i)^T\mathbf{\Lambda})^3$, $i = 1, 2$, $n_i^{-4}\text{tr}\left(\mathbf{\Pi}_{1i}^2\right) = o(\text{var}(T_n))$ and $n_i^{-4}\text{tr}\left(\mathbf{\Lambda}\mathbf{\Sigma}_i\mathbf{\Lambda}\mathbf{\Pi}_{2i}\right) = o(\text{var}(T_n))$ for $i = 1, 2$.

(C5) $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\mathbf{\Lambda}^2\mathbf{\Sigma}_i\mathbf{\Lambda}^2(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = o(n^{-1}\text{tr}((\mathbf{\Lambda}\mathbf{\Sigma}_1\mathbf{\Lambda} + \mathbf{\Lambda}\mathbf{\Sigma}_2\mathbf{\Lambda})^2))$, for $i = 1, 2$, $((\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\mathbf{\Lambda}^2(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2))^2 = o(n^{-1}\text{tr}((\mathbf{\Lambda}\mathbf{\Sigma}_1\mathbf{\Lambda} + \mathbf{\Lambda}\mathbf{\Sigma}_2\mathbf{\Lambda})^2))$.

Remark 1. Conditions (C1) and (C5) are similar to conditions (3.3) and (3.4) in Chen and Qin (2010). To appreciate Condition (C2), consider the simple case $\sigma_{ik} = \sigma_{jl}$, $i, j = 1, 2$, $k, l = 1, \ldots, p$. Condition (C2) then becomes $\text{tr}\left(\mathbf{\Sigma}_i\mathbf{\Sigma}_j\mathbf{\Sigma}_l\mathbf{\Sigma}_h\right) = o[\text{tr}^2\{(\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2)^2\}]$ for $i, j, l, h = 1$ or 2, which is the same as condition their (3.6). Unlike them, we restrict the relationship between the dimension p and sample size n in Condition (C3) due to the use of the plug-in variances. Such a "step backward" is the price we need to pay for scalar-transformation-invariance. Consider the simple case $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$ with bounded eigenvalues so $\text{var}(T_n) = O(pn^{-2})$. We can allow $p = o(n^3)$, which is much more relaxed than required by Srivastava and Du (2008). Under assumptions

(i)−(iii) in Srivastava and Du (2008), the test statistic T_n is not biased, $\mu_n = o(\sqrt{\mathrm{var}(T_n)})$ under H_0. When the dimension p is large, when $p = O(n^2)$ and $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$ has bounded eigenvalues, we have $\mu_n \sim \sqrt{\mathrm{var}(T_n)}$. Condition (C4) is a technical condition, see specific cases in the Appendix. Intuitively, if all the variables are positively correlated, $E((X_{ijk} - \mu_k)^3(X_{ijl} - \mu_l))$ is dominated by $\sigma_{ik}^2 E((X_{ijk} - \mu_k)(X_{ijl} - \mu_l))$ and then $\mathrm{tr}(\mathbf{\Pi}_{1i}^2)$ is dominated by $\mathrm{tr}((\mathbf{\Lambda\Sigma}_i\mathbf{\Lambda})^2)$. Similarly, $\mathrm{tr}(\mathbf{\Lambda\Sigma}_i\mathbf{\Lambda\Pi}_{2i})$ is dominated by $\mathrm{tr}((\mathbf{\Lambda\Sigma}_i\mathbf{\Lambda})^2)$, and Condition (C4) will hold in this special case.

Theorem 1. *Under* (C1)−(C5), $\{T_n - E(T_n)\}/\sqrt{\mathrm{var}(T_n)} \xrightarrow{\mathcal{L}} N(0,1)$, *as* p, $n \to \infty$.

Then to formulate a testing procedure, $E(T_n)$ and $\mathrm{var}(T_n)$ under H_0 need to be estimated. In the Appendix, under (C1)−(C5), we can show that

$$\begin{aligned}
E(T_n) =& ||\mathbf{\Lambda}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)||^2 + \sum_{k=1}^{p}\Bigg\{\frac{2\sigma_{1k}^4}{n_1(n_1-1)(\sigma_{1k}^2+\gamma\sigma_{2k}^2)^2} \\
&+ \frac{2\gamma\sigma_{2k}^4}{n_2(n_2-1)(\sigma_{1k}^2+\gamma\sigma_{2k}^2)^2} + \frac{2}{n_1^2}\frac{\kappa_{1k}^2}{(\sigma_{1k}^2+\gamma\sigma_{2k}^2)^3} + \frac{2\gamma^2}{n_2^2}\frac{\kappa_{2k}^2}{(\sigma_{1k}^2+\gamma\sigma_{2k}^2)^3} \\
&- \frac{4\gamma}{n_1 n_2}\frac{\kappa_{1k}\kappa_{2k}}{(\sigma_{1k}^2+\gamma\sigma_{2k}^2)^3}\Bigg\} + \sum_{k=1}^{p}\frac{\frac{2}{n_1}\kappa_{1k}(\mu_{2k}-\mu_{1k}) + \frac{2\gamma}{n_2}\kappa_{2k}(\mu_{1k}-\mu_{2k})}{(\sigma_{1k}^2+\gamma\sigma_{2k}^2)^2} \\
&+ \sum_{k=1}^{p}\frac{(\mu_{1k}-\mu_{2k})^2}{(\sigma_{1k}^2+\gamma\sigma_{2k}^2)^3}\left(\frac{1}{n_1}\nu_{1k} + \frac{4}{n_1(n_1-1)}\sigma_{1k}^4 + \frac{\gamma}{n_2}\nu_{2k} + \frac{4\gamma^2}{n_2(n_2-1)}\sigma_{2k}^4\right) \\
&+ o(\sqrt{\mathrm{var}(T_n)}) \\
\doteq& \mu_n + o(\sqrt{\mathrm{var}(T_n)}), \text{ where} \\
\mathrm{var}(T_n) =& \Bigg\{\frac{2}{n_1(n_1-1)}\mathrm{tr}((\mathbf{\Lambda\Sigma}_1\mathbf{\Lambda})^2) + \frac{2}{n_2(n_2-1)}\mathrm{tr}((\mathbf{\Lambda\Sigma}_2\mathbf{\Lambda})^2) \\
&+ \frac{4}{n_1 n_2}\mathrm{tr}(\mathbf{\Lambda\Sigma}_1\mathbf{\Lambda}^2\mathbf{\Sigma}_2\mathbf{\Lambda})\Bigg\}(1+o(1)),
\end{aligned}$$

with $\kappa_{ik} = E(X_{ijk} - \mu_{ik})^3$, $\nu_{ik} = E(X_{ijk} - \mu_{ik})^4$, $i = 1, 2$. Although complicated, the bias-term is estimable and then correctable.

The sample variance $\hat{\sigma}_{ik}^2$ and skewness $\hat{\kappa}_{ik} = n_i^{-1}\sum_{j=1}^{n_i}(X_{ijk} - \hat{\mu}_{ik})^3$ are required. An estimator of $E(T_n)$ under H_0 is obtained by plugging-in relevant estimators,

$$\hat{\mu}_n \doteq \widehat{E(T_n)} = \sum_{k=1}^{p} \left\{ \frac{2\hat{\sigma}_{1k}^4}{n_1(n_1-1)(\hat{\sigma}_{1k}^2+\gamma\hat{\sigma}_{2k}^2)^2} + \frac{2\gamma\hat{\sigma}_{2k}^4}{n_2(n_2-1)(\hat{\sigma}_{1k}^2+\gamma\hat{\sigma}_{2k}^2)^2} + \frac{2}{n_1^2}\frac{\hat{\kappa}_{1k}^2}{(\hat{\sigma}_{1k}^2+\gamma\hat{\sigma}_{2k}^2)^3} + \frac{2\gamma^2}{n_2^2}\frac{\hat{\kappa}_{2k}^2}{(\hat{\sigma}_{1k}^2+\gamma\hat{\sigma}_{2k}^2)^3} - \frac{4\gamma}{n_1 n_2}\frac{\hat{\kappa}_{1k}\hat{\kappa}_{2k}}{(\hat{\sigma}_{1k}^2+\gamma\hat{\sigma}_{2k}^2)^3} \right\}.$$

Next, we estimate the trace terms $\mathrm{tr}((\boldsymbol{\Lambda}\boldsymbol{\Sigma}_i\boldsymbol{\Lambda})^2)$, $i=1,2$ and $\mathrm{tr}(\boldsymbol{\Lambda}\boldsymbol{\Sigma}_1\boldsymbol{\Lambda}^2\boldsymbol{\Sigma}_2\boldsymbol{\Lambda})$ in $\mathrm{var}(T_n)$. Chen and Qin (2010) proposed effective estimators for these terms by applying the "leave-one-out" idea. In a similar spirit, we take

$$\mathrm{tr}(\widetilde{(\boldsymbol{\Lambda}\boldsymbol{\Sigma}_1\boldsymbol{\Lambda})^2}) = \frac{1}{n_1(n_1-1)}\sum_{i\neq j}^{n_1}\left(\sum_{l=1}^{p}\frac{(X_{1il}-\bar{\mu}_{1l(i,j)})(X_{1jl}-\bar{\mu}_{1l(i,j)})}{\hat{\sigma}_{1l}^2+\gamma\hat{\sigma}_{2l}^2}\right)^2,$$

where $\bar{\mu}_{il(j,k)}$ is the i-the sample mean after excluding X_{ijl} and X_{ikl}. Since X_{ijk} is not independent of $\hat{\sigma}_{ik}^2$, using such an estimator yields bias-terms which are not negligible when $n=O(p^{1/2})$. It seems more complex to calculate those terms numerically than the expectation of T_n in a high-dimensional setting. We suggest a remedy, is motivated by Chen and Qin (2010), as

$$\begin{aligned}\mathrm{tr}(\widehat{(\boldsymbol{\Lambda}\boldsymbol{\Sigma}_s\boldsymbol{\Lambda})^2}) &= \frac{1}{2P_{n_s}^4}\sum^{*}(\mathbf{X}_{si_1}-\mathbf{X}_{si_2})^T\mathbf{D}_{s(i_1,i_2,i_3,i_4)}^{-1}(\mathbf{X}_{si_3}-\mathbf{X}_{si_4})\\ &\quad\times(\mathbf{X}_{si_3}-\mathbf{X}_{si_2})^T\mathbf{D}_{s(i_1,i_2,i_3,i_4)}^{-1}(\mathbf{X}_{si_1}-\mathbf{X}_{si_4}),\ s=1,2,\ \text{and}\\ \mathrm{tr}(\boldsymbol{\Lambda}\widehat{\boldsymbol{\Sigma}_1\boldsymbol{\Lambda}^2}\boldsymbol{\Sigma}_2\boldsymbol{\Lambda}) &= \frac{1}{4P_{n_1}^2P_{n_2}^2}\sum_{i_1\neq i_2}^{n_1}\sum_{i_3\neq i_4}^{n_2}\Big((\mathbf{X}_{1i_1}-\mathbf{X}_{1i_2})^T\mathbf{D}_{(i_1,i_2,i_3,i_4)}^{-1}(\mathbf{X}_{2i_3}-\mathbf{X}_{2i_4})\Big)^2,\end{aligned}$$

where

$$\begin{aligned}\mathbf{D}_{1(i_1,i_2,i_3,i_4)} &= \mathrm{diag}(\hat{\sigma}_{11(i_1,i_2,i_3,i_4)}^2+\gamma\hat{\sigma}_{21}^2,\ldots,\hat{\sigma}_{1p(i_1,i_2,i_3,i_4)}^2+\gamma\hat{\sigma}_{2p}^2),\\ \mathbf{D}_{2(i_1,i_2,i_3,i_4)} &= \mathrm{diag}(\hat{\sigma}_{11}^2+\gamma\hat{\sigma}_{21(i_1,i_2,i_3,i_4)}^2,\ldots,\hat{\sigma}_{1p}^2+\gamma\hat{\sigma}_{2p(i_1,i_2,i_3,i_4)}^2),\\ \mathbf{D}_{(i_1,i_2,i_3,i_4)} &= \mathrm{diag}(\hat{\sigma}_{11(i_1,i_2)}^2+\gamma\hat{\sigma}_{21(i_3,i_4)}^2,\ldots,\hat{\sigma}_{1p(i_1,i_2)}^2+\gamma\hat{\sigma}_{2p(i_3,i_4)}^2),\end{aligned}$$

and $\hat{\sigma}_{sk(i_1,\ldots,i_l)}^2$ is the s-th sample variance after excluding X_{si_j}, $j=1,\ldots,l$, $s=1,2$, $l=2,4$, $k=1,\ldots,p$. We use $\sum^{*}$ to denote summations over distinct indexes. Thus, in $\mathrm{tr}(\widehat{(\boldsymbol{\Lambda}\boldsymbol{\Sigma}_1\boldsymbol{\Lambda})^2})$, the summation is over the set $\{i_1\neq i_2\neq i_3\neq i_4\}$, for all $i_1,i_2,i_3,i_4\in\{1,\ldots,n_1\}$ and $P_n^m=n!/(n-m)!$.

Remark 2. The estimators of $\mathrm{tr}(\boldsymbol{\Sigma}_i^2)$ and $\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)$ proposed by Chen and Qin (2010) are not translation-invariant. According to the proof of their Theorem 2, $E(\widehat{\mathrm{tr}(\boldsymbol{\Sigma}_i{}^2)}_{CQ}) = \mathrm{tr}(\boldsymbol{\Sigma}_i{}^2)+\boldsymbol{\mu}'\boldsymbol{\Sigma}_i\boldsymbol{\mu}/(n-2)$. Thus, it is easy to verify that $\widehat{\mathrm{tr}(\boldsymbol{\Sigma}_i{}^2)}_{CQ}$ is different when $\mathbf{X}_{ij}$ is transformed to $\mathbf{X}_{ij}+\boldsymbol{\theta}$. From an asymptotic viewpoint,

the ratio consistency of $\widehat{\mathrm{tr}(\mathbf{\Sigma}_i{}^2)}_{CQ}$ relies on the condition that $\boldsymbol{\mu}'\mathbf{\Sigma}_i\boldsymbol{\mu}/(n-2) = o(\mathrm{tr}(\mathbf{\Sigma}_i{}^2))$. This is fairly restrictive because the expectation of $\mathbf{X}_{ij}$ could be in any scale in two-sample location testing applications. As such, the estimator of $\mathrm{tr}(\mathbf{R}_i^2)$ proposed by Park and Ayyala (2013) is not translation-invariant.

Remark 3. The estimators of Srivastava and Du (2008) and Srivastava, Katayama, and Kano (2013) are for $\mathrm{tr}(\mathbf{R}^2)$ with plug-in sample variances, and suffer from bias-terms in their variance estimators when p is large. See about this in Section 3.

Proposition 2. *If* (C1)−(C4) *hold,* $\hat{\mu}_n = E(T_n) + o_p(\sqrt{\mathrm{var}(T_n)})$, *and*

$$\frac{\mathrm{tr}(\widehat{(\mathbf{\Lambda}\mathbf{\Sigma}_i\mathbf{\Lambda})^2})}{\mathrm{tr}((\mathbf{\Lambda}\mathbf{\Sigma}_i\mathbf{\Lambda})^2)} \xrightarrow{p} 1, i = 1, 2 \quad \text{and} \quad \frac{\mathrm{tr}(\mathbf{\Lambda}\widehat{\mathbf{\Sigma}_1\mathbf{\Lambda}^2\mathbf{\Sigma}_2}\mathbf{\Lambda})}{\mathrm{tr}(\mathbf{\Lambda}\mathbf{\Sigma}_1\mathbf{\Lambda}^2\mathbf{\Sigma}_2\mathbf{\Lambda})} \xrightarrow{p} 1.$$

As a consequence, a ratio-consistent estimator of $\mathrm{var}(T_n)$ under H_0 is

$$\hat{\sigma}_n^2 \doteq \widehat{\mathrm{var}(T_n)}$$
$$= \left\{ \frac{2}{n_1(n_1-1)}\mathrm{tr}(\widehat{(\mathbf{\Lambda}\mathbf{\Sigma}_1\mathbf{\Lambda})^2}) + \frac{2}{n_2(n_2-1)}\mathrm{tr}(\widehat{(\mathbf{\Lambda}\mathbf{\Sigma}_2\mathbf{\Lambda})^2}) + \frac{4}{n_1 n_2}\mathrm{tr}(\mathbf{\Lambda}\widehat{\mathbf{\Sigma}_1\mathbf{\Lambda}^2\mathbf{\Sigma}_2}\mathbf{\Lambda}) \right\}.$$

We are now in a position to set the test statistic $(T_n - \hat{\mu}_n)/\hat{\sigma}_n$, and to shwo its asymptotic normality under the null hypothesis.

Corollary 1. *Under* (C1)−(C4) *and* H_0, $(T_n - \hat{\mu}_n)/\hat{\sigma}_n \xrightarrow{\mathcal{L}} N(0,1)$.

The result suggests rejecting H_0 with α level of significance if $(T_n - \hat{\mu}_n)/\hat{\sigma}_n > z_\alpha$. The ratio-consistent estimator of $\mathrm{var}(T_n)$ appears complex but computes quickly. For example, it takes 5s per iteration in FORTRAN using Inter Core 2.2 MHz CPU for a $n_1 = n_2 = 30, p = 1{,}000$ case for each $\mathrm{tr}(\widehat{(\mathbf{\Lambda}\mathbf{\Sigma}_i\mathbf{\Lambda})^2}), \mathrm{tr}(\mathbf{\Lambda}\widehat{\mathbf{\Sigma}_1\mathbf{\Lambda}^2\mathbf{\Sigma}_2}\mathbf{\Lambda})$ and the entire procedure is generally completed in less than 20s.

According to Theorem 1, the power of the test under the local alternative (C5) is

$$\beta_{BF}(||\mathbf{\Lambda}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)||) = \Phi\left(-z_\alpha + \frac{\tilde{\mu}_n}{\tilde{\sigma}_n}\right),$$

where

$$\tilde{\mu}_n = ||\mathbf{\Lambda}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)||^2 + \sum_{k=1}^{p} \frac{\frac{2}{n_1}\kappa_{1k}(\mu_{2k} - \mu_{1k}) + (2\gamma/n_2)\kappa_{2k}(\mu_{1k} - \mu_{2k})}{(\sigma_{1k}^2 + \gamma\sigma_{2k}^2)^2}$$
$$+ \sum_{k=1}^{p} \frac{(\mu_{1k} - \mu_{2k})^2}{(\sigma_{1k}^2 + \gamma\sigma_{2k}^2)^3}\left(\frac{1}{n_1}\nu_{1k} + \frac{4}{n_1(n_1-1)}\sigma_{1k}^4 + \frac{\gamma}{n_2}\nu_{2k} + \frac{4\gamma^2}{n_2(n_2-1)}\sigma_{2k}^4\right),$$
$$\tilde{\sigma}_n^2 = \frac{2}{n_1(n_1-1)}\mathrm{tr}((\mathbf{\Lambda}\mathbf{\Sigma}_1\mathbf{\Lambda})^2) + \frac{2}{n_2(n_2-1)}\mathrm{tr}((\mathbf{\Lambda}\mathbf{\Sigma}_2\mathbf{\Lambda})^2) + \frac{4}{n_1 n_2}\mathrm{tr}(\mathbf{\Lambda}\mathbf{\Sigma}_1\mathbf{\Lambda}^2\mathbf{\Sigma}_2\mathbf{\Lambda}),$$

and $\Phi(\cdot)$ is the standard normal distribution function. In contrast, Chen and Qin (2010) showed that the power of their proposed test is

$$\beta_{CQ}(||\boldsymbol{\mu}_1-\boldsymbol{\mu}_2||)=\Phi\Big(-z_\alpha+\frac{n\lambda(1-\lambda)||\boldsymbol{\mu}_1-\boldsymbol{\mu}_2||^2}{\sqrt{2\mathrm{tr}(\tilde{\boldsymbol{\Sigma}}^2)}}\Big),$$

where $\tilde{\boldsymbol{\Sigma}}=(1-\lambda)\boldsymbol{\Sigma}_1+\lambda\boldsymbol{\Sigma}_2$.

Theoretically comparing the proposed test with Chen and Qin's under general settings turns out to be difficult. In order to get a rough picture of the asymptotic power comparison between them, we simply assume that $\kappa_{ik}=0$, $k=1,\ldots,p$, $i=1,2$. It is then easy to show that

$$\sum_{k=1}^{p}\frac{(\mu_{1k}-\mu_{2k})^2}{(\sigma_{1k}^2+\gamma\sigma_{2k}^2)^3}\left(\frac{1}{n_1}\nu_{1k}+\frac{4}{n_1(n_1-1)}\sigma_{1k}^4+\frac{\gamma}{n_2}\nu_{2k}+\frac{4\gamma^2}{n_2(n_2-1)}\sigma_{2k}^4\right)$$
$$=o(||\boldsymbol{\Lambda}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)||^2).$$

Now, the power of the proposed test becomes

$$\beta_{BF}(||\boldsymbol{\Lambda}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)||)=\Phi\Big(-z_\alpha+\frac{n\lambda(1-\lambda)||\boldsymbol{\Lambda}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)||^2}{\sqrt{2\mathrm{tr}(\boldsymbol{\Lambda}\tilde{\boldsymbol{\Sigma}}\boldsymbol{\Lambda})^2}}\Big)+o(1),$$

and consider the some representative cases.

(i) $\mu_{1k}-\mu_{2k}=\delta$, $k=1,\ldots,p$. Here

$$\beta_{BF}(||\boldsymbol{\Lambda}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)||)=\Phi\Big(-z_\alpha+\frac{n\lambda(1-\lambda)\delta^2\mathrm{tr}(\boldsymbol{\Lambda}^2)}{\sqrt{2\mathrm{tr}(\boldsymbol{\Lambda}\tilde{\boldsymbol{\Sigma}}\boldsymbol{\Lambda})^2}}\Big)+o(1),$$
$$\beta_{CQ}(||\boldsymbol{\mu}_1-\boldsymbol{\mu}_2||)=\Phi\Big(-z_\alpha+\frac{np\lambda(1-\lambda)\delta^2}{\sqrt{2\mathrm{tr}(\tilde{\boldsymbol{\Sigma}}^2)}}\Big).$$

By the Cauchy inequality, $p^2\mathrm{tr}(\boldsymbol{\Lambda}\tilde{\boldsymbol{\Sigma}}\boldsymbol{\Lambda})^2\le\mathrm{tr}(\tilde{\boldsymbol{\Sigma}}^2)\mathrm{tr}^2(\boldsymbol{\Lambda}^2)$. As a consequence,

$$\beta_{CQ}(||\boldsymbol{\mu}_1-\boldsymbol{\mu}_2||)\le\Phi\Big(-z_\alpha+\frac{n\lambda(1-\lambda)\delta^2\mathrm{tr}(\boldsymbol{\Lambda}^2)}{\sqrt{2\mathrm{tr}(\boldsymbol{\Lambda}\tilde{\boldsymbol{\Sigma}}\boldsymbol{\Lambda})^2}}\Big)\le\beta_{BF}(||\boldsymbol{\Lambda}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)||).$$

When the variances of the components are equal, the tests are equivalently powerful from the asymptotic viewpoint. Otherwise, the proposed test is preferable in terms of asymptotic power under local alternatives.

(ii) $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$, diagonal. The variances of two half of components are ζ_1^2 and ζ_2^2. Assume $\mu_{1k} - \mu_{2k} = \delta$, $k = 1, \ldots, \lfloor p/2 \rfloor$. Then

$$\beta_{BF}(||\mathbf{\Lambda}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)||) = \Phi\Big(- z_\alpha + \frac{n\sqrt{p}\lambda(1-\lambda)\delta^2}{2\sqrt{2}\zeta_1^2}\Big) + o(1),$$

$$\beta_{CQ}(||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||) = \Phi\Big(- z_\alpha + \frac{n\sqrt{p}\lambda(1-\lambda)\delta^2}{2\sqrt{\zeta_1^4 + \zeta_2^4}}\Big).$$

The asymptotic relative efficiency (ARE) of the porposed test with respect to the CQ test is then $\sqrt{\zeta_1^4 + \zeta_2^4}/(\sqrt{2}\zeta_1^2)$, so the proposed test is more powerful than CQ if $\zeta_1^2 < \zeta_2^2$, and vice versa. The ARE has a positive lower bound of $1/\sqrt{2}$ when $\zeta_1^2 >> \zeta_2^2$. It can be arbitrarily large if ζ_1^2/ζ_2^2 is close to zero, showing the need for scale-invariance test.

3. Numerical Studies

Throughout this section, we run 1,000 replications for each experiment so the standard error of size or power entries is bounded by 0.016.

3.1. The bias-term

Here we report on a simulation study designed to evaluate the bias-term of $\hat{q}_n$, as proposed by Srivastava, Katayama, and Kano (2013), of $T_n - \hat{\mu}_n$ as proposed by us and, the quality of the corresponding variance estimator under the null hypothesis. We consider only the case $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{\Sigma} = (a_{ij}), a_{ij} = 0.5^{|i-j|}$, and $\mathbf{X}_{ij}$ independent p-dimensional multivariate normal random vectors. We summarize simulation results for the mean-standard deviation-ratio $E(T)/\sqrt{\text{var}(T)}$ and the variance estimator ratio $\widehat{\text{var}(T)}/\text{var}(T)$, where T denotes either $\hat{q}_n$ or $T_n - \hat{\mu}_n$. Since the explicit forms of $E(T)$ and $\text{var}(T)$ are difficult to calculate, we estimate them by simulation. We consider sample sizes $n_1 = n_2 = 15, 30$ and dimensions $p = 30, 60, 100, 200, 300, 400, 800,$ 1,000.

Figure 1 reports the mean-standard deviation-ratio of the test statistics proposed by Srivastava, Katayama, and Kano (2013) and us. In Figure 1, the bias-term in $\hat{q}_n$ apparently exists, especially when the dimension is high. There is also some bias for our BF test when $n_1 = n_2 = 15$, p =1,000. This last is not strange because the dimension is comparable to the cube of the sample size and (C3) does not hold. In the other cases, the mean-standard deviation-ratio of $T_n - \hat{\mu}_n$ is approximately zero, showing the effectiveness of our bias correction procedure. Figure 2 reports the simulation results of the variance estimator ratio. Here the variance estimator of Srivastava, Katayama, and Kano (2013) is apparently larger than the variance. First, there is a bias-term in the estimator $\widehat{\text{var}}(\hat{q}_n)$ when the dimension is high. Second, the correction term $c_{p,n}$ is always larger then one.

The variance estimator ratio of our test statistic is approximately one, so our variance estimator is effective even when the dimension is very high. Because both ratios are higher than the acceptable level, the empirical sizes of the test statistics proposed by Srivastava, Katayama, and Kano (2013) also deviate from the significance level. See the next subsections for more information.

3.2. Empirical sizes and power comparison

Here we report a simulation study designed to evaluate the performance of our proposed test (abbreviated as BF). To allow a meaningful comparison with the methods proposed by Bai and Saranadasa (1996) (abbreviated as BS), Srivastava and Du (2008) (abbreviated as SD), Chen and Qin (2010) (abbreviated as CQ), and Srivastava, Katayama, and Kano (2013) (abbreviated as SKK), we first considered the unequal covariance matrices assumption, where the assumption of common covariances in Bai and Saranadasa (1996) and Srivastava and Du (2008) does not hold. We considered the moving average model in Chen and Qin (2010):

$$X_{ijk} = \rho_{i1} Z_{ij} + \rho_{i2} Z_{i(j+1)} + \cdots + \rho_{iL_i} Z_{i(j+L_i-1)} + \mu_{ij}$$

for $i = 1, 2$, $j = 1, \ldots, n_i$ and $k = 1, \ldots, p$ where $\{Z_{ijk}\}$ are i.i.d. random variables. We considered Scenario I: the $\{Z_{ijk}\}$ were $N(0, 1)$; Scenario II: the first half components of $\{Z_{ijk}\}_{k=1}^{p}$ were centralized Gamma(4,1) and the second half components were $N(0, 1)$. The coefficients $\{\rho_{il}\}_{l=1}^{L_i}$ were independently $U(2, 3)$ and were kept fixed once generated, through our simulations. The correlations among X_{ijk} and X_{ijl} were determined by $|k - l|$ and L_i. We chose $L_1 = 1$, and $L_2 = 3$ to generate different covariances of $\mathbf{X}_i$. For the alternative hypothesis, we fixed $\boldsymbol{\mu}_1 = \mathbf{0}$ and again chose $\boldsymbol{\mu}_2$ in according to Case A: one allocates all of the components of equal magnitude to be nonzero, or Case B: one randomly allocates half of the components of equal magnitude to be nonzero. To make the power comparable among the configurations of H_1, we set $\eta := ||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||^2 / \sqrt{\mathrm{tr}(\boldsymbol{\Sigma}_1^2) + \mathrm{tr}(\boldsymbol{\Sigma}_2^2)} = 0.1$ throughout the simulation.

For simplicity, sample sizes $n_1 = n_2$ were chosen to be 15, 20 and 30. In the supplemental file, we also present some simulation results with $n_1 \neq n_2$, for which the comparison conclusion revealed below still holds (c.f., Figure S1.1). We chose three dimensions for each sample size $p = 225, 400, 900$. Figure 3 below reports the empirical sizes of five tests. Clearly, when $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$, the sizes of BS and SD are much smaller than the significance level, 0.05, especially when p is ultra-high; CQ and BF have reasonable sizes in most cases. The performance of SKK is not very encouraging as, in many cases, sizes are larger than the significance level, whereas in some cases where $n_1 = n_2 = 15$, $p = 900$, the sizes of SKK were conservative.There were considerable biases in the estimation of the

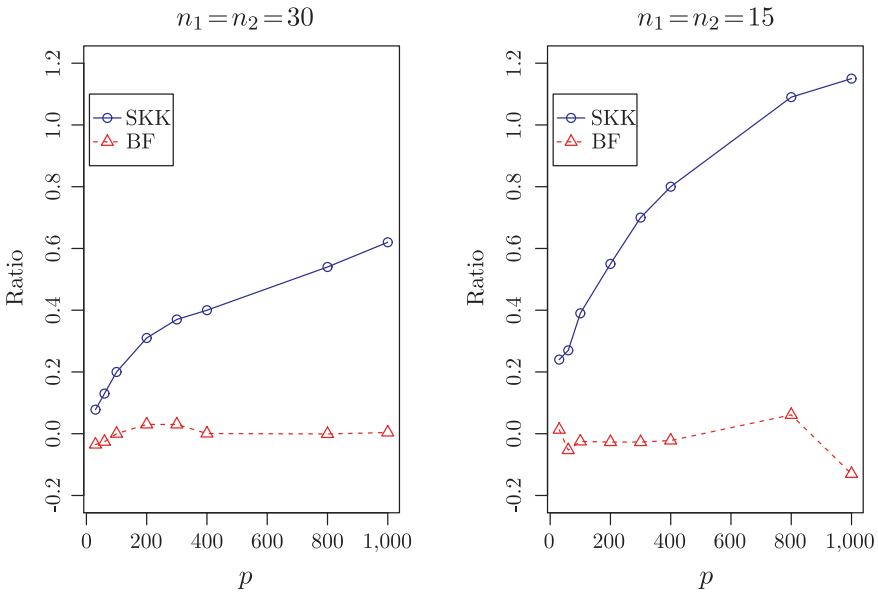

Figure 1. The mean-standard deviation-ratio $E(T)/\sqrt{\text{var}(T)}$ of the test statistics proposed by Srivastava, Katayama, and Kano (2013) (SKK) and us (BF).

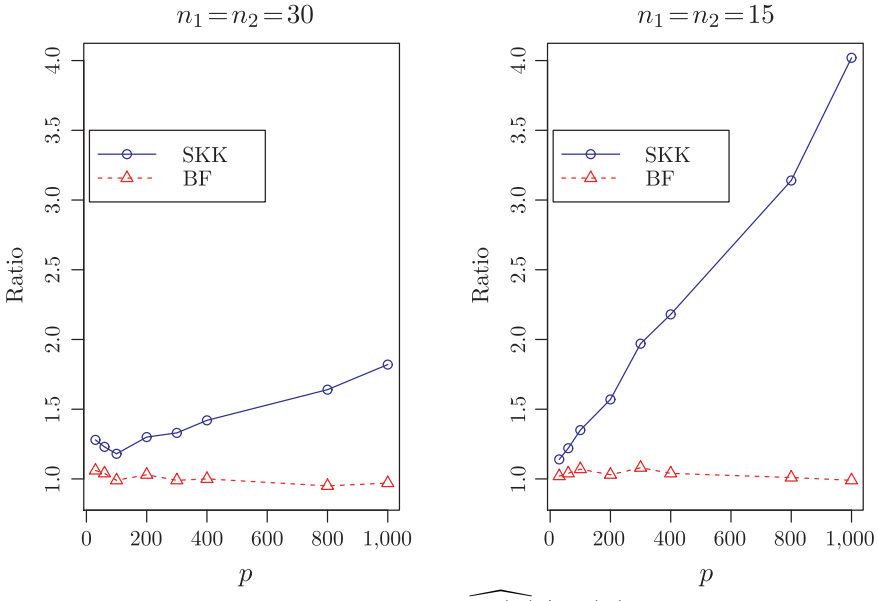

Figure 2. The variance estimator ratio $\widehat{\text{var}(T)}/\text{var}(T)$ of the test statistics proposed by Srivastava, Katayama, and Kano (2013) (SKK) and us (BF).

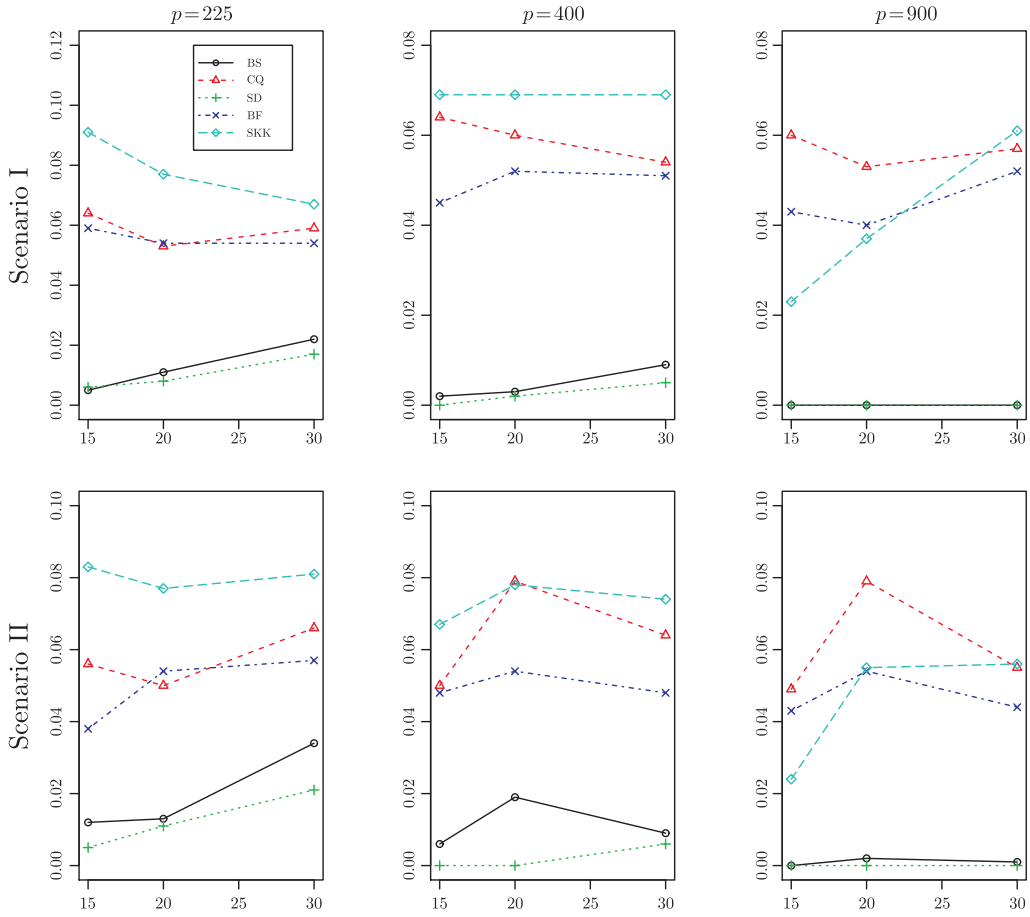

Figure 3. Empirical size comparisons at 5% significance when $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$.

sample correlation matrix when $n = O(p^{1/2})$, so it is very difficult to maintain significance level when a bias-correction is not made.

The powers of the tests were in Figure 4 for a further comparison. The of BS and SD tests are not efficient in most of cases, as expected. Under Scenario I, the variances of components are all equal and the powers of CQ test are slightly larger than the BF test. Under Scenario II, BF outperforms CQ uniformly in all the cases by a large margin. These results suggest that the BF test is scale-invariant, quite efficient, and robust in testing the equality of locations, and is particularly useful when $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$.

4. A Data Example

We applied the proposed methodology to a real date set from a semi-conductor manufacturing process that collects variables from sensors at many measurement

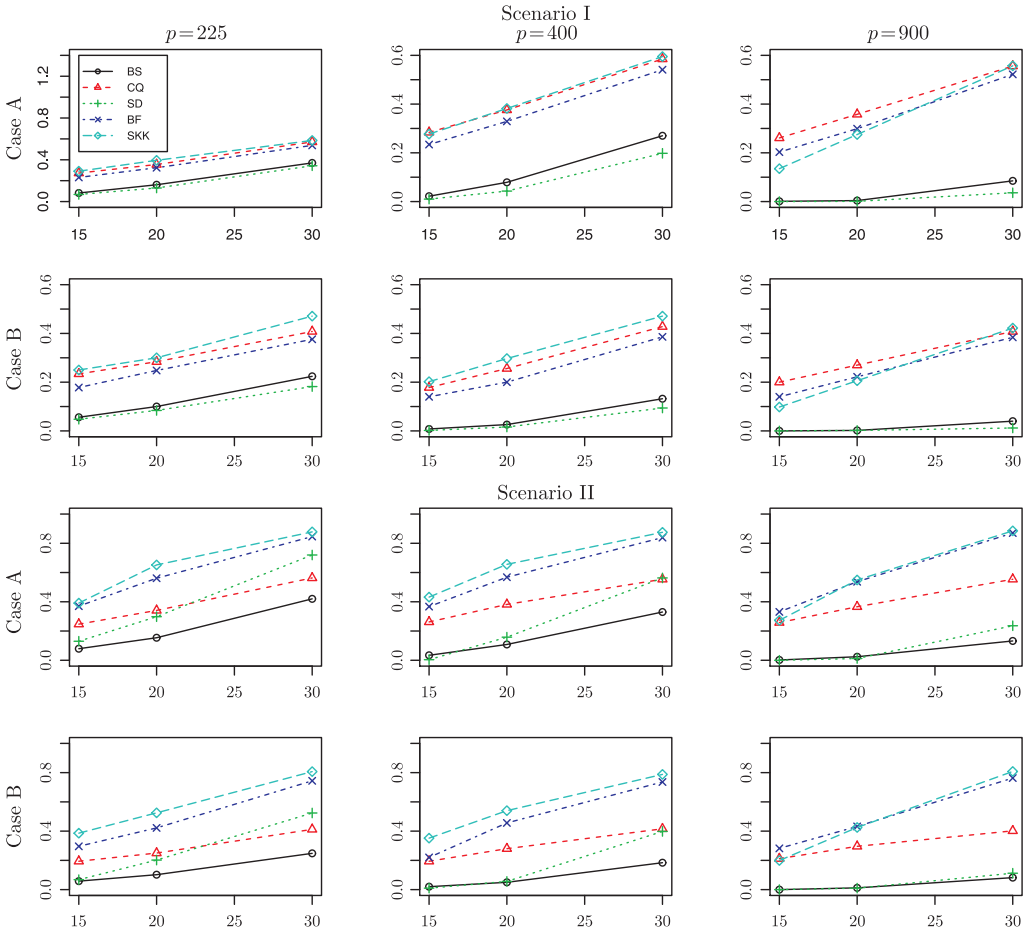

Figure 4. Empirical power comparisons at 5% significance when $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$.

points. The data set contains 1,567 vector observations. For each observation, there are 591 continuous measurements. A categorical label, indicates whether a conforming yield through house-line testing is also provided. The goal is to model and monitor production quality.

In quality control, a critical step is to compare a new sample with reference sample to check if the process is in control. If so, the reference sample is updated. This setp clearly requires powerful testing approach. While one could consider tests that incorporate the information from all the sensors, in many applications, this is not feasible if users want to test the process at start-up stage when only a few reference samples are available. Our method is applicable with rather small sample sizes, and thus appears to be particulary useful in this setting.

The dataset contains null values varying in intensity depending on the individuals features. Since the fraction of missing values is trivial in this dataset, we

Table 1. Empirical power comparisons at 5% significance for the sensor dataset.

n	BF	BS	CQ	SD	SKK
15	0.355	0.067	0.076	0.288	0.297
20	0.611	0.048	0.051	0.616	0.621
30	0.941	0.058	0.059	0.971	0.982

used mean imputation. There are 117 constant features in the data, so just 474 variables are involved. For illustration, we artificially assumed n observations were categorized as nonconforming and conforming and appled the BF, BS, CQ and SD tests. To get a broad picture of performance comparison, we considered a bootstrap-type testing procedure. Two random samples of size n were drawn from 104 nonconforming observations and 1463 conforming observations without replacement. Then all four considered tests are applied to these two samples and the corresponding test results with a significance level 0.05 were recorded. This procedure are repeated 1,000 times and the resulting powers are given in Table 1. The BF and SD tests gave satisfactory results and their powers increased quickly with sample size. The proposed BF test performed slightly better than SD and SKK when the sample size were small ($n = 15$). The BS and CQ tests were ineffective here because they are not scale-invariant, and in this data set, components have varying.

5. Discussion

A natural concern is whether our test can handle ultra-high dimensional scenario with larger p, say at an exponential rate in n. It is difficult, if not impossible, when there is no sparse structure for existing scale-invariant tests to correct bias-terms. It is an open problem as to one can define a test statistic that is (at least) asymptotically unbiased, lacking a sparsity assumption on the data structure. The standardized version, with shrinkage estimation under sparse structure and other conditions, may help; see Cai, Liu, and Xia (2014).

Acknowledgement

The authors thank the Editor, an associate editor, and two referees for their many helpful comments that have resulted in significant improvements in the article. Feng and Zhu were partly supported by a grant from the Research Grants Council of Hong Kong, Hong Kong, China. Zou and Wang was supported by the NNSF of China Grants 11431006, 11131002, 11371202, 11471069, the Foundation for the Author of National Excellent Doctoral Dissertation of PR China 201232.

References

Bai, Z. and Saranadasa, H. (1996). Effect of high dimension: by an example of a two sample problem. *Statist. Sinica* **6**, 311-29.

Cai, T., Liu, W. D. and Xia, Y. (2014). Two-sample test of high dimensional means under dependency. *J. Roy. Statist. Soc. Ser. B* **76**, 349-372.

Chen, S. X. and Qin, Y.-L. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *Ann. Statist.* **38**, 808-835.

Fisher, R. A. (1935). The fiducial argument in statistical inference. *Ann. Eugenics* **11**, 141-172.

Fisher, R. A. (1939). The comparison of samples with possibly unequal variances. *Ann. Eugenics* **9**, 174-180.

James, G. S. (1954). Tests of linear hypotheses in univariate and multivariate analysis when the ratios of the population variances are unknown. *Biometrika* **41**, 19-43.

Johansen, S. (1980). The Welch-James approximation to the distribution of the residual sum of squares in a weighted linear regression. *Biometrika* **67**, 85-95.

Park, J. and Ayyala, D. N. (2013). A test for the mean vector in large dimension and small samples. *J. Statist. Plann. Inference* **143**, 929-943.

Srivastava, M. S. and Du, M. (2008). A test for the mean vector with fewer observations than the dimension. *J. Multivariate Anal.* **99**, 386-402.

Srivastava, M. S., Katayama, S. and Kano, Y. (2013). A two sample test in high dimensional data. *J. Multivar. Anal.* **114**, 349-358.

Yao, Y. (1965). An approximate degrees of freedom solution to the multivariate Behrens-Fisher problem. *Biometrika* **52**, 139-147.

Zhang, J. T. and Xu, J. F. (2009). On the k-sample Behrens-Fisher problem for high-dimensional data. *Sci. China Ser. A-Math.* **52**, 1285-1304.

Institute of Statistics, Nankai University, Tianjin, 300071, China.

E-mail: flnankai@126.com

Institute of Statistics, Nankai University, Tianjin, 300071, China.

E-mail: nk.chlzou@gmail.com

Institute of Statistics, Nankai University, Tianjin, 300071, China.

E-mail: zjwang@nankai.edu.cn

Department of Mathematics, Hong Kong Baptist University, Hong Kong, China.

E-mail: lzhu@hkbu.edu.hk

(Received February 2014; accepted October 2014)