Discussiones Mathematicae
Probability and Statistics 37 (2017) 147–156
doi:10.7151/dmps.1195

A TWO-SAMPLE TEST BASED ON CLUSTER SUBSPACES FOR EQUALITY OF MEAN VECTORS IN HIGH DIMENSION

Łukasz Smaga

Adam Mickiewicz University
Faculty of Mathematics and Computer Science

e-mail: ls@amu.edu.pl

Abstract

In this paper, a two-sample problem in a high-dimensional setting, where the data dimension is larger than the sample size, is considered. In such setting, the Hotelling's test is not applicable due to singularity of the pooled sample covariance matrix. Recently, Zhang and Pan (2016) proposed a permutation test based on several cluster subspaces of lower dimension, where the Hotelling's statistic can be applied. This paper considers a modification of this test using other dissimilarity measure. To calculate clusters, a cutoff measure is established. The new testing procedure is shown to be invariant under linear transformations of the marginal distributions. Simulation studies indicate that the new test performs comparable to or even better in certain situations than the test of Zhang and Pan (2016) in terms of power.

Keywords: cluster analysis, coefficient of determination, high-dimensional data, Pearson correlation coefficient, two-sample problem.

2010 Mathematics Subject Classification: 62H15, 62H30.

1. Introduction

Let us consider a two-sample problem for multivariate data. We assume that $\mathbf{X}_1, \ldots, \mathbf{X}_{n_1}$ and $\mathbf{Y}_1, \ldots, \mathbf{Y}_{n_2}$ are two random samples generated in an independent and identically distributed manner from independent p-dimensional normal random vectors $\mathbf{X}$ and $\mathbf{Y}$. Moreover, the vectors $\mathbf{X}$ and $\mathbf{Y}$ have mean vectors $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, respectively, and the same covariance matrix $\boldsymbol{\Sigma} > 0$, which are all fixed

and unknown. The problem of interest is to test the null hypothesis $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ against the alternative hypothesis $H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$.

In the case of $p \leq n := n_1 + n_2 - 2$, the Hotelling's test is usually used to solve the above problem (Anderson [1]). However, for $p > n$, this test can not be applied, since the pooled sample covariance matrix is not invertible. Many ideas on how avoid this singularity problem are studied in the literature. Bai and Saranadasa [2], Srivastava and Du [8] and Chen and Qin [3] assumed independence of variables and considered test statistics replacing the sample covariance matrix in the Hotelling's statistic by the identity matrix or diagonal matrix containing sample variances of the variables. Unfortunately, this assumption may be unrealistic, for instance, for gene expressions (Thulin [9]). Moreover, not taking into account the information about the covariance structure may cause significant loss of power when the variables are dependent. More use of the dependence structure was made by Thulin [9] and Zhang and Pan [10]. They considered the tests based on random and cluster subspaces of variables, respectively. In these lower-dimensional subspaces, the Hotelling's statistic is well defined. So, this test statistic is computed in each subspace, and the overall test statistic is the sum of statistics of subspaces. The finale p-value is obtained by permutation method. The testing procedures of Thulin [9] and Zhang and Pan [10] offer higher power than competing ones when the variables are dependent. The results of Zhang and Pan [10] indicate that their test seems to be more powerful and less computationally intensive than that of Thulin [9].

In this paper, we consider a modification of the test of Zhang and Pan [10], which results in better power in certain scenarios. Zhang and Pan [10] used 1-Pearson correlation coefficient as a dissimilarity measure, which first clusters together the highly positively correlated variables. However, highly negatively correlated variables are clustered at the end, which is a little strange. We propose to use 1-coefficient of determination as the dissimilarity measure, which first clusters together highly positively or negatively correlated variables. (By the coefficient of determination, we mean the square of the Pearson correlation coefficient.) At the end, the least correlated variables are clustered. In this way, we treat negatively and positively correlated variables more "fair". For the new dissimilarity measure, we establish a cutoff for calculating the first clusters to restrain the effect of statistical fluctuations for coefficient of determination. Similarly as the solutions of Thulin [9] and Zhang and Pan [10], the new testing procedure is invariant under linear transformations of the marginal distributions. This is important as it is common, for example, for genetic data to be rescaled in the sense of dividing the marginal distributions by their standard deviations. Conducted simulation studies show that the new test is comparable to the testing procedure of Zhang and Pan [10] in terms of size control and power, when all non-zero correlations are positive. On the other hand, the new method may be

more powerful in the presence of negatively correlated variables.

The remainder of the paper is organized as follows: In Section 2, the new testing procedure is proposed. The choice of its parameters as well as its properties are also discussed there. Section 3 presents the description of simulation experiments and discussion of their results. Section 4 concludes the paper.

2. Testing procedure

We consider the following permutation procedure, which is a modified version of that of Zhang and Pan [10]:

1. Input: observed data $\mathbf{X}_1, \ldots, \mathbf{X}_{n_1}$, $\mathbf{Y}_1, \ldots, \mathbf{Y}_{n_2}$, cutoff dissimilarity measure CDM, maximum number of variables in a cluster MNVC $\leq n_1 + n_2 - 2$, number of random permutations B. Choice of CDM and MNVC is described below.
2. Perform hierarchical clustering of variables by using $1-$coefficient of determination as a dissimilarity measure, average linkage and observations of both groups.
3. Calculate clusters based on cutoff dissimilarity measure CDM.
4. Cluster each cluster or sub-cluster consisting of more than MNVC variables, into two sub-clusters as long as each cluster or sub-cluster has more than MNVC variables. Let N_c denote a final number of clusters.
5. Compute the value of the test statistic given by

$$T_{obs} = \sum_{k=1}^{N_c} T_k^2, \tag{1}$$

for the original data, where T_k^2 is the Hotelling's statistic for the k-th cluster, $k = 1, \ldots, N_c$, i.e.,

$$T_k^2 = \frac{n_1 n_2}{n_1 + n_2} (\bar{\mathbf{X}}_k - \bar{\mathbf{Y}}_k)^\top \hat{\mathbf{\Sigma}}_k^{-1} (\bar{\mathbf{X}}_k - \bar{\mathbf{Y}}_k),$$

$\bar{\mathbf{X}}_k = n_1^{-1} \sum_{i=1}^{n_1} \mathbf{X}_{k,i}$, $\bar{\mathbf{Y}}_k = n_2^{-1} \sum_{i=1}^{n_2} \mathbf{Y}_{k,i}$ and

$$n\hat{\mathbf{\Sigma}}_k = \sum_{i=1}^{n_1} (\mathbf{X}_{k,i} - \bar{\mathbf{X}}_k)(\mathbf{X}_{k,i} - \bar{\mathbf{X}}_k)^\top + \sum_{i=1}^{n_2} (\mathbf{Y}_{k,i} - \bar{\mathbf{Y}}_k)(\mathbf{Y}_{k,i} - \bar{\mathbf{Y}}_k)^\top$$

are the sample means and pooled sample covariance matrix of the observations for variables from k-th cluster ($n = n_1 + n_2 - 2$).

6. From all observations $\mathbf{X}_1, \ldots, \mathbf{X}_{n_1}, \mathbf{Y}_1, \ldots, \mathbf{Y}_{n_2}$, select randomly without replacement n_1 observations for the first new sample. The remainder of the observations forms the second new sample.
7. Compute the value of the test statistic (1) for new data created in step 6 using the same clusters obtained in steps 2-4.
8. Repeat steps 6-7 B times. Let $T_1, \ldots, T_B$ denote the obtained values of the test statistic.
9. Output: P-value computed as $B^{-1} \sum_{j=1}^{B} I(T_j \geq T_{obs})$, where $I(S)$ stands for the usual indicator function on the set S (takes value 1 if S is true and 0 otherwise).

Remark 1.

1. Since each cluster contains no more than $n_1 + n_2 - 2$ variables, the pooled sample covariance matrices $\hat{\mathbf{\Sigma}}_k$ computed in Step 5 of the above procedure are in fact invertible. Hence, the test statistic (1) is well defined.
2. As we can notice, in Step 7 of the above procedure, the value of the test statistic is computed by using the same clusters as for the original data. This follows from the fact that permutation of the data does not affect the correlations of variables. Thus, clustering of the data can be performed only for the original observations.

Now, we have to specify how to choose the cutoff dissimilarity measure CDM and the maximum number of variables in a cluster MNVC.

The cutoff dissimilarity measure CDM will be selected to restrain the effect of statistical fluctuations for the sample correlation coefficient and hence also for coefficient of determination. It is well known that, especially for small sample size, the sample correlation coefficient may be quite large when variables are in fact uncorrelated. We have $\binom{p}{2} = p(p-1)/2$ coefficients of determination for p variables. By the normality assumption, the Fisher z-transformation of the sample correlation coefficient r is approximately normally distributed. More precisely, we have

$$z = \frac{1}{2} \log\left(\frac{1+r}{1-r}\right) \sim N\left(\frac{1}{2} \log\left(\frac{1+\rho}{1-\rho}\right), \frac{1}{\sqrt{n-1}}\right),$$

approximately, where log denotes the natural logarithm. Let q_N be the $1 - 1/[p(p-1)]$ quantile of $N(0,1)$ distribution. Therefore, when $\rho = 0$, we obtain

$$P\left(-q'_N \leq \frac{1}{2} \log\left(\frac{1+r}{1-r}\right) \leq q'_N\right) = 1 - \frac{2}{p(p-1)}, \tag{2}$$

approximately, where $q'_N = q_N/(n-1)^{1/2}$. Since the inverse of the Fisher z-transformation given by $r = (e^{2z}-1)/(e^{2z}+1)$ is increasing and odd function, we have

$$
\begin{aligned}
P\left(-q'_N \le \frac{1}{2}\log\left(\frac{1+r}{1-r}\right) \le q'_N\right) &= P\left(\frac{e^{-2q'_N}-1}{e^{-2q'_N}+1} \le r \le \frac{e^{2q'_N}-1}{e^{2q'_N}+1}\right) \\
&= P\left(-\frac{e^{2q'_N}-1}{e^{2q'_N}+1} \le r \le \frac{e^{2q'_N}-1}{e^{2q'_N}+1}\right) \\
&= P\left(|r| \le \frac{e^{2q'_N}-1}{e^{2q'_N}+1}\right).
\end{aligned} \tag{3}
$$

By (2) and (3), we conclude that

$$
P\left(r^2 \le \left(\frac{e^{2q'_N}-1}{e^{2q'_N}+1}\right)^2\right) = 1 - \frac{2}{p(p-1)},
$$

approximately. Thus, when all p variables were uncorrelated, only about one of the all coefficients of determination was greater than $((e^{2q'_N}-1)/(e^{2q'_N}+1))^2$. Moreover, the dissimilarity measure was less than $1-((e^{2q'_N}-1)/(e^{2q'_N}+1))^2$, and the corresponding two variables were clustered together due to statistical fluctuations. That justifies the following choice:

$$
\mathrm{CDM} = 1 - \left(\frac{e^{2q'_N}-1}{e^{2q'_N}+1}\right)^2.
$$

Some of the clusters calculated based on the above cutoff dissimilarity measure may have more variables than n. So, they must be further clustered. More precisely, in Step 4 of the testing procedure, each cluster having more than MNVC variables have to be clustered into two sub-clusters. To be consistent with Zhang and Pan [10] and to see only the effect of our dissimilarity measure, we select the same MNVC as they, i.e., $\mathrm{MNVC} = \lfloor 2n/3 \rfloor$, where $\lfloor x \rfloor$ is the greatest integer that is less than or equal to x. Such choice might result in cluster dimension distributed around $\lfloor n/2 \rfloor$, which is shown by Thulin [9] to be the optimal dimension of subspace giving highest power.

Finally, we establish the invariance property of the new testing procedure. As we noticed in Section 1, the invariance of a procedure under linear transformations of the marginal distributions is important in the high-dimensional setting. Hence, this property should be taken into account when evaluating the testing procedure along with size control and power. The following result states that the new test is in fact invariant under such transformations (both samples are

transformed analogously) similarly to the tests of Srivastava and Du [8], Thulin [9] and Zhang and Pan [10]. However, the other testing procedures mentioned in the introduction are not invariant.

Proposition 2. *Assume that* $\mathbf{C}$ *is an invertible diagonal real matrix of size* $p \times p$ *and* $\mathbf{c} \in \mathbb{R}^p$. *Conditioned on the random permutations chosen, the new testing procedure is invariant under linear transformations* $(\mathbf{X}, \mathbf{Y}) \to (\mathbf{CX} + \mathbf{c}, \mathbf{CY} + \mathbf{c})$, *where* $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_{n_1})$ *and* $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_{n_2})$.

Proof. Since the sample correlation coefficients, and hence also coefficients of determination, are invariant under scaling and shifting transformations as defined in the proposition, the clustering of variables (Steps 2-4 of the testing procedure) is also invariant under such transformations. Moreover, the Hotelling's statistic is also invariant under these transformations. This finishes the proof. ■

3. Simulation studies

In this section, we compare the new testing procedure presented in Section 2 and the test of Zhang and Pan [10], and investigate the randomness of p-values of the new (permutation) test. The simulations of that paper suggest that their test is comparable to or even better than the other testing procedures mentioned in Section 1, so we did not consider them in our simulation studies. Simulation experiments were performed using the R language (R Core Team [6]), and the codes are available from the author.

Our simulation experiments were similar to those of Zhang and Pan [10]. The observations $\mathbf{X}_1, \ldots, \mathbf{X}_{n_1}$ and $\mathbf{Y}_1, \ldots, \mathbf{Y}_{n_2}$ were generated from the multivariate normal and t_4 distributions of dimension p with mean vectors $\boldsymbol{\mu}_1 = \mathbf{0}_p$ and $\boldsymbol{\mu}_2$, respectively, and covariance matrix $\boldsymbol{\Sigma} = (\sigma_{ij})$. We considered $p = 200, 1000$, $n_1 = 30$, $n_2 = 40$, $B = 500$ permutation samples, and the following different covariance matrices:

- $\boldsymbol{\Sigma}_1$: $\sigma_{ij} = 0.6^{|i-j|}$,
- $\boldsymbol{\Sigma}_2$: $\sigma_{ij} = \left[(-1)^{\lceil i/25 \rceil} 0.6\right]^{|i-j| \pmod{25}}$ for $\lceil i/25 \rceil = \lceil j/25 \rceil$, and $\sigma_{ij} = 0$ for $\lceil i/25 \rceil \neq \lceil j/25 \rceil$,
- $\boldsymbol{\Sigma}_3$: $\sigma_{ii} = 1$ for $i = 1, \ldots, p$; $\sigma_{ij} = 0.6$ when $i \neq j$, $\lceil i/25 \rceil = \lceil j/25 \rceil$, and $\lceil i/25 \rceil$ is odd; $\sigma_{ij} = (0.6)^{|i-j| \pmod{25}}$ when $i \neq j$, $\lceil i/25 \rceil = \lceil j/25 \rceil$ and $\lceil i/25 \rceil$ is even; $\sigma_{ij} = 0$ for $\lceil i/25 \rceil \neq \lceil j/25 \rceil$,
- $\boldsymbol{\Sigma}_4$: $\sigma_{ii} = 1$ for $i = 1, \ldots, p$; $\sigma_{ij} = 0.6$ when $i \neq j$, $\lceil i/25 \rceil = \lceil j/25 \rceil$, and $\lceil i/25 \rceil$ is odd; $\sigma_{ij} = (-0.6)^{|i-j| \pmod{25}}$ when $i \neq j$, $\lceil i/25 \rceil = \lceil j/25 \rceil$ and $\lceil i/25 \rceil$ is even; $\sigma_{ij} = 0$ for $\lceil i/25 \rceil \neq \lceil j/25 \rceil$,

$\mathbf{\Sigma}_5$: $\sigma_{ii} = 1$ for $i = 1, \ldots, p$ and $\sigma_{2k-1,2k} = \sigma_{2k,2k-1} = 0.6$ for $k = 1, \ldots, p/2$,

$\mathbf{\Sigma}_6$: $\sigma_{ii} = 1$ for $i = 1, \ldots, p$, $\sigma_{2k-1,2k} = \sigma_{2k,2k-1} = -0.6$ for $k = 1, \ldots, p/4$, and $\sigma_{2k-1,2k} = \sigma_{2k,2k-1} = 0.6$ for $k = p/4 + 1, \ldots, p/2$.

The matrices $\mathbf{\Sigma}_5$ and $\mathbf{\Sigma}_6$ were inspired by those considered in Feng *et al.* [5]. Observe that matrices $\mathbf{\Sigma}_1$, $\mathbf{\Sigma}_3$ and $\mathbf{\Sigma}_5$ have only positive non-zero elements, while the rest of matrices also have some negative ones.

For checking size control of the tests, we consider $\boldsymbol{\mu}_2 = \mathbf{0}_p$, while for power, the following two alternatives are chosen: (1) randomly half of components of $\boldsymbol{\mu}_2$ are distributed from $N(0, 1)$ and the others are zeros. Denote this vector by $\boldsymbol{\mu}_2^{(1)}$. (2) the elements of $\boldsymbol{\mu}_2$ on positions $2k-1$ (resp. $2k$) are equal to one (resp. zero), $k = 1, \ldots, p/2$. Let $\boldsymbol{\mu}_2^{(2)}$ denote this vector. These two alternatives were considered by Feng *et al.* [5]. Similarly to Feng *et al.* [4], to make the power comparable among the configurations of $\boldsymbol{\mu}_2$, we set $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2 \left(\operatorname{tr}(\mathbf{\Sigma}^2)\right)^{1/2} = 0.1$ throughout the simulation. Other alternatives, covariance matrices, sample sizes, etc. were also considered, but the results were similar, and therefore, they are omitted for space saving.

The empirical sizes and powers were obtained based on 1000 simulation replicates. The results are depicted in Table 1. In all cases, both tests maintain the preassigned type I error. Except one case, their empirical sizes belong to the usual 95% significance interval $[3.6, 6.4]$ (see, for example, Smaga [7]). It is also worth noting that although both testing procedures were constructed under normality assumption, simulation results indicate the robustness of them to non-normal distribution under the null hypothesis. This means that the permutation method approximates the null distribution of the test statistic (1) very satisfactorily.
The empirical powers of the testing procedures are generally quite satisfactory. When all non-zero correlations are positive ($\mathbf{\Sigma} = \mathbf{\Sigma}_i$, $i = 1, 3, 5$), the empirical powers of both tests are almost identical. On the other hand, the new test outperforms significantly the testing procedure of Zhang and Pan [10] in case of appearing of negative correlation of variables ($\mathbf{\Sigma} = \mathbf{\Sigma}_i$, $i = 2, 4, 6$). These all can be explained by the fact that the numbers of clusters obtained during performing both tests are similar when $\mathbf{\Sigma} = \mathbf{\Sigma}_i$, $i = 1, 3, 5$, while these numbers for the new test are much smaller than for the test of Zhang and Pan [10] when $\mathbf{\Sigma} = \mathbf{\Sigma}_i$, $i = 2, 4, 6$ (see Table 2 for some examples). So the new testing procedure has the opportunity to use more information about correlation of variables, and hence it may be more powerful. Observe also that in most cases the power of the tests under t_4-distribution is comparable to or even better than under normal one, which can be explained similarly as above.

Finally, we study the randomness of the p-values of the new test estimated based on different numbers of random permutations B. For this purpose, we applied 100 times the new testing procedure to a single data set with $p = 100$,

Table 1. Empirical sizes (rows with $\boldsymbol{\mu}_2 = \mathbf{0}_p$) and powers (rows with $\boldsymbol{\mu}_2 = \boldsymbol{\mu}_2^{(1)}$ or $\boldsymbol{\mu}_2 = \boldsymbol{\mu}_2^{(2)}$), as percentages, of the new testing procedure ("New") and the test of Zhang and Pan [10] ("ZP").

		$p = 200$				$p = 1000$			
		Normal		t_4		Normal		t_4	
$\boldsymbol{\Sigma}$	$\boldsymbol{\mu}_2$	ZP	New	ZP	New	ZP	New	ZP	New
$\boldsymbol{\Sigma}_1$	$\mathbf{0}_p$	5.1	4.9	5.5	5.4	4.5	4.8	5.8	5.0
	$\boldsymbol{\mu}_2^{(1)}$	73.6	73.3	78.2	77.7	64.5	63.7	70.4	70.0
	$\boldsymbol{\mu}_2^{(2)}$	74.4	73.8	81.1	80.8	68.9	67.9	67.5	69.5
$\boldsymbol{\Sigma}_2$	$\mathbf{0}_p$	5.2	5.4	4.9	5.5	5.5	5.2	4.9	4.2
	$\boldsymbol{\mu}_2^{(1)}$	43.3	69.0	72.3	84.5	48.5	66.9	53.8	67.8
	$\boldsymbol{\mu}_2^{(2)}$	47.8	70.1	55.7	80.5	48.2	71.0	47.4	69.2
$\boldsymbol{\Sigma}_3$	$\mathbf{0}_p$	5.2	5.0	5.1	4.9	5.5	5.5	5.2	5.3
	$\boldsymbol{\mu}_2^{(1)}$	98.1	98.2	99.8	99.8	99.3	99.3	99.8	99.6
	$\boldsymbol{\mu}_2^{(2)}$	87.7	87.5	96.7	96.8	89.9	90.2	96.1	95.8
$\boldsymbol{\Sigma}_4$	$\mathbf{0}_p$	5.5	5.5	5.1	4.7	5.5	5.3	6.3	6.7
	$\boldsymbol{\mu}_2^{(1)}$	95.4	98.8	99.5	100.0	97.8	99.4	99.2	99.8
	$\boldsymbol{\mu}_2^{(2)}$	68.2	87.2	86.3	96.6	74.1	88.8	85.1	95.7
$\boldsymbol{\Sigma}_5$	$\mathbf{0}_p$	5.2	5.5	5.5	5.8	4.5	5.1	4.9	4.2
	$\boldsymbol{\mu}_2^{(1)}$	60.9	60.1	68.7	67.2	62.6	62.3	54.6	53.9
	$\boldsymbol{\mu}_2^{(2)}$	65.3	64.6	63.6	64.3	61.1	60.0	53.2	55.2
$\boldsymbol{\Sigma}_6$	$\mathbf{0}_p$	4.3	4.2	5.0	4.2	4.0	4.0	5.1	5.5
	$\boldsymbol{\mu}_2^{(1)}$	43.7	63.3	51.5	63.8	43.7	58.9	42.3	52.9
	$\boldsymbol{\mu}_2^{(2)}$	47.2	63.1	46.1	64.3	42.1	59.0	37.8	53.1

Table 2. Numbers of clusters obtained during performing the new testing procedure ("New") and the test of Zhang and Pan [10] ("ZP") for $p = 200$, $\boldsymbol{\mu}_2 = \mathbf{0}_p$ (For the alternatives $\boldsymbol{\mu}_2^{(1)}$ and $\boldsymbol{\mu}_2^{(2)}$, the results were very similar.).

Distribution	Test	$\boldsymbol{\Sigma}_1$	$\boldsymbol{\Sigma}_2$	$\boldsymbol{\Sigma}_3$	$\boldsymbol{\Sigma}_4$	$\boldsymbol{\Sigma}_5$	$\boldsymbol{\Sigma}_6$
Normal	ZP	87	125	44	88	101	149
	New	91	90	44	47	104	104
t_4	ZP	85	118	47	90	104	144
	New	85	85	46	53	107	106

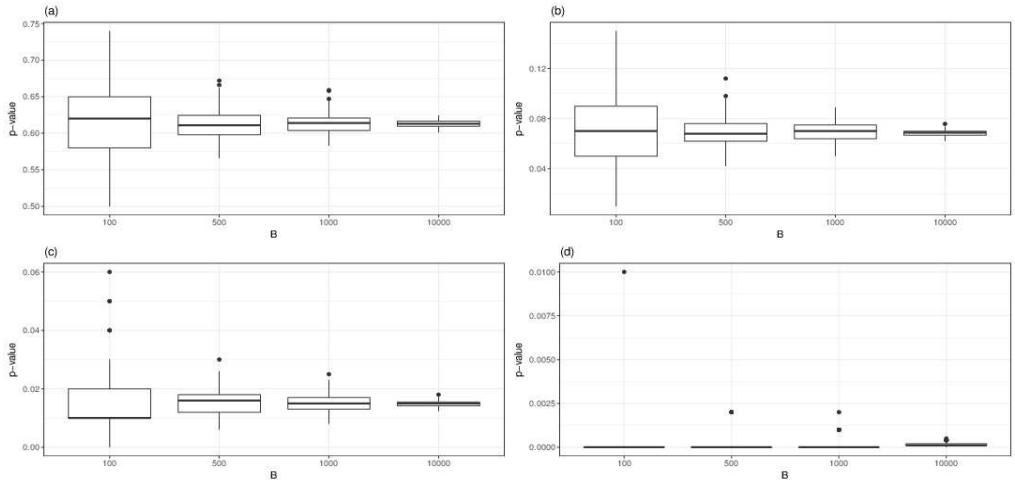

Figure 1. Box-and-whisker plots for the randomness of the p-values of the new test for a single data set with $p = 100$, $n_i = 25$, $i = 1, 2$. For (a) (resp. (b), (c), (d)) the null (resp. alternative) hypothesis was true and the p-value using $B = 100{,}000$ is 0.61316 (resp. 0.06939, 0.01532, 0.00017, respectively). For each B, the testing procedure was applied 100 times.

$n_i = 25$, $i = 1, 2$ under the null or alternative hypothesis for different values of B. The results are depicted on Figure 1. We observe that the variance of the estimator of p-value may be non-negligible for too low B. However, it also depends on the "proper" p-value. The variance decreases when p-value also decreases, which is of practical interest, since the accuracy of larger p-values is less important than the accuracy of small p-values. This observation may also result in reducing the computational cost of permutation test, which may be time-consuming. Namely, we can first perform the test using a small number of random permutations (e.g., $B = 100$) to see which decision is expected, and then use a higher B, especially when the results are inconvincing (e.g., the p-value is close to the nominal level as in Figure 1 (b)). Similar results were obtained by Thulin [9] for its test based on random subspaces.

4. Concluding remarks

For a two-sample problem in high dimension, we have proposed the modification of the testing procedure of Zhang and Pan [10] using different dissimilarity measure in clustering variables. Both tests are invariant under linear transformations of the marginal distributions, but they differ in their finite sample behavior. The power of the two tests closely mimic each other when the variables are non-

negatively correlated, but the new test has higher power when some variables are negatively correlated. We therefore recommend the new test as the default high-dimensional two-sample test.

References

[1] T.W. Anderson, An Introduction to Multivariate Statistical Analysis (3rd ed., Wiley, London, 2003).

[2] Z. Bai and H. Saranadasa, *Effect of high dimension: by an example of a two sample problem*, Statist. Sinica **6** (1996) 311–329.

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

[4] L. Feng, C. Zou, Z. Wang and L. Zhu, *Two sample Behrens-Fisher problem for high-dimensional data*, Statist. Sinica **25** (2015) 1297–1312.

[5] L. Feng, C. Zou, Z. Wang and L. Zhu, *Composite T^2 test for high-dimensional data*, Statist. Sinica (2016).
doi:10.5705/ss.202015.0199

[6] R Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria (2017).
https://www.R-project.org/

[7] Ł. Smaga, *Diagonal and unscaled Wald-type tests in general factorial designs*, Electr. J. Stat. **11** (2017) 2613–2646.

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

[9] M. Thulin, *A high-dimensional two-sample test for the mean using random subspaces*, Comput. Stat. & Data Anal. **74** (2014) 26–38.

[10] J. Zhang and M. Pan, *A high-dimension two-sample test for the mean using cluster subspaces*, Comput. Stat. & Data Anal. **97** (2016) 87–97.

Received 5 October 2017
Accepted 16 November 2017