Comput Stat (2016) 31:451–464
DOI 10.1007/s00180-015-0605-7

ORIGINAL PAPER

On high-dimensional two sample mean testing statistics: a comparative study with a data adaptive choice of coefficient vector

Soeun Kim[1] · Jae Youn Ahn[2] · Woojoo Lee[3]

Received: 1 September 2014 / Accepted: 1 July 2015 / Published online: 18 July 2015

Abstract The key issues involved in two sample tests in high dimensional problems arise due to large dimension of the mean vector for a relatively small sample size. Recently, Wang et al. (Stat Sin 23:667–690, 2013) proposed a jackknife empirical likelihood test that works under weak assumptions on the dimension of variables (p), and showed that the test statistic has a chi-square limit regardless of whether p is finite or diverges. The sufficient condition required for this statistic is still restrictive. In this paper we significantly relax the sufficient condition for the asymptotic chi-square limit with models allowing flexible dependence structures and derive simpler alternative statistics for testing the equality of two high dimensional means. The proposed statistics have a chi-squared distribution or the maximum of two independent chi-square statistics as their limiting distributions, and the asymptotic results hold for either finite or divergent p. We also propose a data-adaptive method to select the coefficient vector, and compare the various methods in simulation studies. The proposed choice of coefficient vector substantially increases power in the simulation.

Electronic supplementary material The online version of this article (doi:10.1007/s00180-015-0605-7) contains supplementary material, which is available to authorized users.

✉ Woojoo Lee
lwj221@gmail.com

Soeun Kim
Soeun.Kim@uth.tmc.edu

Jae Youn Ahn
jaeyahn@ewha.ac.kr

1 Department of Biostatistics, University of Texas Health Science Center, Houston, TX, USA

2 Department of Statistics, Ewha Womans University, Seoul, Korea

3 Department of Statistics, Inha University, Incheon, Korea

Keywords High dimension · Two sample mean test · Coefficient vector · Data adaptive

1 Introduction

High dimensional problems have become increasingly important with the advancement of technology that increased the capacity to collect high dimensional data. For genomic or microarray data, thousands of gene expressions can be collected for each subject, in which case the number of variables is very large compared to sample size. When there is an interest in identifying mean differences for gene sets in two samples, this leads to a simultaneous testing problem of differences in the means of two different gene sets with large number of genes in each set (Nettleton et al. 2008; Newton et al. 2007). In such fields there is a growing demand for methods for handling high dimensional data when the number of variables is very large compared to sample size.

Key issues involved in two sample tests in high dimensional problems arise due to a large dimension of the mean vector for a relatively small sample size, and because Hotelling's T^2 statistic has poor performance with singular covariance matrix. In order to overcome this issue, a test statistic using Moore–Penrose inverse is proposed by Srivastava and Khatri (1979), and covariance shrinkage techniques are introduced to be able to work with positive definite sample covariance matrix (Ledoit and Wolf 2004). The question of the possibility of getting around the issues related to inverse covariance matrix remains for two sample tests in high dimensional problems. Among the various contributions made to the literature, Bai and Saranadasa (2004) proposed to modify Hotelling's T^2 statistic by excluding sample covariance matrix under the assumption $p/n \to c$, where c is a constant with $c \leq \infty$. However, in high dimensional problems, it is in general too restrictive to assume that p/n will converge to c. Chen and Qin (2010) noted this and constructed a test statistic that allows p to be arbitrarily large without restriction of p being of the same order of n, under given assumptions. Wang et al. (2013) proposed a jackknife empirical likelihood (JEL) test, which works under weaker conditions than those proposed by Chen and Qin, and results in good statistical power. Wang et al. (2013) showed that their proposed statistic has a chi-square limit regardless of whether p is finite or diverges. The key idea in Wang et al. (2013)'s methodology is (1) to split the samples into two independent groups, (2) the use of empirical likelihood, and (3) the use of the jackknife samples. Point (1) is essential for the derivation of the necessary asymptotic results. However, the extent to which (2) and (3) contribute to power has not been explored in detail. This understanding is important because this leads to an insight on what should be considered primarily in more complicated problems. There are sufficient conditions that need to be met for Wang et al. (2013)'s methodology as given in the "Appendix", and the conditions are restrictive, requiring the rate of increase of p to be controlled. Relaxing these conditions would lead to important improvement in the methodology.

In this paper, the restrictive conditions in Wang et al. (2013)'s approach are significantly relaxed in our proposed model allowing flexible dependence structures. The explicit form of the model is given in (2.1). In addition, we derive simpler alternative statistics for testing the equality of two high dimensional means and study the

contribution of the use of empirical likelihood and jackknife samples. The proposed statistics result in a chi-square or the maximum of two independent chi-square statistics as limit distributions, and the asymptotic results hold regardless of whether p is finite or diverges.

To study the contribution of the jackknife samples, we investigate one statistic based on the jackknife sample, and another that is not. The proposed statistics are not based on the empirical likelihood, and does not require any optimization procedure. A simulation study is performed to compare the performance of the new statistics not based on empirical likelihood with Wang et al. (2013)'s statistics. In the simulation study, we consider various factors that can affect the performance of the two sample test, including the skewness of distribution, correlation between variables, sample size, the number of variables, and the sign of the mean shifts. It turns out that the sign of mean shifts is critical in obtaining good power. In order to take into account that mean shifts can be in different directions, we investigate different choices of coefficient vectors. In Wang et al. (2013)'s approach, the coefficient vector is chosen a priori to boost the statistical power, taking $(1, \ldots, 1)$ as a convenient choice. Although this method is useful in some settings, the simulation results show that the use of $(1, \ldots, 1)$ does not always yield good power in some practical settings. This vector can be chosen based on prior information (Wang et al. 2013), but we often do not have such information in practice. In this paper, we propose a data-adaptive method to select the coefficient vector, and show by simulation that the proposed choice substantially improves the power. The simple statistics proposed in this paper together with the data adaptive choice of coefficient vector yields good power, and can be used for high dimensional problems in various areas of research.

The organization of this paper is as follows. In Sect. 2, we review Wang et al. (2013)'s approach in detail and explain how we derive the new alternative statistic for two sample tests. A numerical study is given in Sect. 3. The data-adaptive choice of the coefficient vector is explained in Sect. 4 followed by simulation results. We apply the methods in the analysis of gene expression data in Sect. 5. Concluding remarks are given in Sect. 6. All the details of the simulation results are given in the supplementary document.

2 Review of Wang et al. (2013) and proposed statistics

Since our proposed statistics are closely related to the setting in Wang et al. (2013) in the sense that it requires common asymptotic results, we start this section with a detailed review of the JEL approach in Wang et al. (2013). To avoid introducing additional difficulty for the readers, we intend to keep most of notation used in Wang et al. (2013). Assume that $X_i = (X_{i1}, \ldots, X_{ip})^T$ $(i = 1, \ldots, n_1)$ and $Y_j = (Y_{j1}, \ldots, Y_{jp})^T$ $(j = 1, \ldots, n_2)$, where p denotes the dimension of the variables and n_1 and n_2 are the sample sizes of each group, respectively. X_i and Y_j are assumed to be two independent random samples with mean vectors μ_1 and μ_2, respectively. In this paper, we are concerned with testing $H_0 : \mu_1 = \mu_2$ while allowing different covariances for the two groups. Note that this null hypothesis is equivalent to testing $H_0 : (\mu_1 - \mu_2)^T(\mu_1 - \mu_2) = 0$.

Let $m_1 = [n_1/2], m_2 = [n_2/2], m = m_1 + m_2$, and let $\tilde{X}_i = X_{i+m_1}$ for $i = 1 \ldots, m_1$, and $\tilde{Y}_j = Y_{j+m_2}$ for $j = 1 \ldots, m_2$.

To test H_0, Wang et al. (2013) proposed a JEL method. The jackknife sample is formulated as

$$Z_{k,1} = \frac{m_1+m_2}{m_1 m_2} \sum_{i=1}^{m_1}\sum_{j=1}^{m_2}(X_i - Y_j)^T(\tilde{X}_i - \tilde{Y}_j) - \frac{m_1+m_2-1}{(m_1-1)m_2} \sum_{i=1, i\neq k}^{m_1}\sum_{j=1}^{m_2}(X_i - Y_j)^T(\tilde{X}_i - \tilde{Y}_j)$$

$$Z_{k,2} = \frac{m_1+m_2}{m_1 m_2} \sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\left\{\alpha^T(X_i - Y_j) + \alpha^T(\tilde{X}_i - \tilde{Y}_j)\right\}$$
$$- \frac{m_1+m_2-1}{(m_1-1)m_2} \sum_{i=1, i\neq k}^{m_1}\sum_{j=1}^{m_2}\left\{\alpha^T(X_i - Y_j) + \alpha^T(\tilde{X}_i - \tilde{Y}_j)\right\}$$

for $k = 1, \ldots, m_1$, and

$$Z_{k,1} = \frac{m_1+m_2}{m_1 m_2} \sum_{i=1}^{m_1}\sum_{j=1}^{m_2}(X_i - Y_j)^T(\tilde{X}_i - \tilde{Y}_j) - \frac{m_1+m_2-1}{m_1(m_2-1)} \sum_{i=1}^{m_1}\sum_{j=1, j\neq k-m_1}^{m_2}(X_i - Y_j)^T(\tilde{X}_i - \tilde{Y}_j)$$

$$Z_{k,2} = \frac{m_1+m_2}{m_1 m_2} \sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\left\{\alpha^T(X_i - Y_j) + \alpha^T(\tilde{X}_i - \tilde{Y}_j)\right\}$$
$$- \frac{m_1+m_2-1}{m_1(m_2-1)} \sum_{i=1}^{m_1}\sum_{j=1, j\neq k-m_1}^{m_2}\left\{\alpha^T(X_i - Y_j) + \alpha^T(\tilde{X}_i - \tilde{Y}_j)\right\}$$

for $k = m_1 + 1, \ldots, m$. Here, α denotes the coefficient vector previously referred to in the introduction. The JEL ratio function for testing $H_0 : \mu_1 = \mu_2$ is given by

$$L_m = \sup\left\{\prod_{i=1}^{m}(mp_i) : p_i \geq 0, \sum_{i=1}^{m} p_i = 1, \quad \sum_{i=1}^{m} p_i Z_i = (0,0)^T\right\}.$$

where $Z_i = (Z_{i,1}, Z_{i,2})^T$. Under either condition A1 or A2 in "Appendix", Wang et al. (2013) showed that $-2\log L_m \to \chi_2^2$ in distribution. A remarkable property of this statistic is that the asymptotic result holds regardless of whether p is finite or diverges. However, this works under restrictive situations since the sufficient conditions A1 and A2 are required for this property to hold. For instance, to satisfy condition A2, we need $p = o\left(m^{\frac{\delta+\min(\delta,2)}{2(2+\delta)}}\right)$ for some $\delta > 0$. Since $\frac{\delta+\min(\delta,2)}{2(2+\delta)} \leq 1/2$ for any $\delta > 0$, p should increase at slower rate than $m^{1/2}$. In order to relax these restrictive conditions, we consider models that allow flexible dependence structures. Let

$$\Sigma = E\left[(X_1 - \mu_1)(X_1 - \mu_1)^T\right]$$
$$\tilde{\Sigma} = E\left[(Y_1 - \mu_2)(Y_1 - \mu_2)^T\right]$$
$$\rho_1 = tr(\Sigma^2) = E\left[\left((X_1 - \mu_1)^T(X_2 - \mu_1)\right)^2\right]$$
$$\rho_2 = tr(\tilde{\Sigma}^2) = E\left[\left((Y_1 - \mu_2)^T(Y_2 - \mu_2)\right)^2\right].$$

Our model assumes

$$X_i - \mu_1 = \Sigma^{1/2}\varepsilon_i \quad \text{and} \quad Y_i - \mu_2 = \tilde{\Sigma}^{1/2}\tilde{\varepsilon}_i \tag{2.1}$$

where the elements in ε_i and $\tilde{\varepsilon}_i$ are i.i.d random variables with mean 0 and finite fourth moment. If

$$\frac{\lambda_p^4 p^2}{m_1\rho_1^2} = o(1), \quad \frac{\tilde{\lambda}_p^4 p^2}{m_2\rho_2^2} = o(1).$$

hold where λ_p and $\tilde{\lambda}_p$ are the largest eigenvalues of Σ and $\tilde{\Sigma}$, respectively, then the asymptotic chi-square limiting distribution is obtained as described before.

In fact, our model provides a significantly relaxed condition on p. Specifically, if we impose the boundedness on all the eigenvalues of Σ and $\tilde{\Sigma}$ as in Wang et al. (2013),

$$\frac{\lambda_p^4 p^2}{m_1\rho_1^2} = O\left(\frac{1}{m_1}\right) \to 0, \quad \frac{\tilde{\lambda}_p^4 p^2}{m_2\rho_2^2} = O\left(\frac{1}{m_2}\right) \to 0.$$

Therefore, the asymptotic chi-square limiting distribution is obtained for any order of p. More details and proofs are given in "Appendix 2 and 3".

Now we propose a new statistic for testing the equality of two high dimensional means that can be used instead of the JEL. This is a simpler statistic that follows from an intermediate step instead of deriving the JEL.

Denote $U_1 = \frac{1}{m_1 m_2}\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}(X_i - Y_j)^T(\tilde{X}_i - \tilde{Y}_j)$ and $U_2 = \frac{1}{m_1 m_2}\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}$ $\alpha^T(X_i - Y_j) + \alpha^T(\tilde{X}_i - \tilde{Y}_j)$. Under either condition A1 or A2 or B1 in the "Appendix" and $H_0 : \mu_1 = \mu_2$, we have

$$\sqrt{m}\begin{pmatrix} \frac{U_1}{\sqrt{\rho}} \\ \frac{U_2}{\sqrt{\tau}} \end{pmatrix} \to_d N(0, I_2) \tag{2.2}$$

where $\rho = \frac{m}{m_1}\rho_1 + \frac{m}{m_2}\rho_2$ and $\tau = \frac{m}{m_1}\tau_1 + \frac{m}{m_2}\tau_2$,

$$\tau_1 = 2tr(\alpha^T\Sigma\alpha) = 2E\left((\alpha^T(X_1 - \mu_1))^2\right)$$
$$\tau_2 = 2tr(\alpha^T\tilde{\Sigma}\alpha) = 2E\left((\alpha^T(Y_1 - \mu_2))^2\right)$$

We immediately have

$$m\left(\frac{U_1^2}{\rho} + \frac{U_2^2}{\tau}\right) \to_d \chi_2^2.$$

Replacing ρ and τ with their consistent estimators provides us with simple statistics for two sample high dimensional testing without introducing the empirical likelihood. In

fact, Wang et al. (2013) developed two consistent estimators for ρ and τ. First, denote $\hat{\rho}_{jack} = \frac{1}{m}\sum_{k=1}^{m} Z_{k,1}^2$ and $\hat{\tau}_{jack} = \frac{1}{m}\sum_{k=1}^{m} Z_{k,2}^2$. The subscript "jack" highlights the use of the jackknife samples. Following Wang et al. (2013), we have

$$\frac{\hat{\rho}_{\text{jack}}}{\rho} \to_p 1 \quad \text{and} \quad \frac{\hat{\tau}_{\text{jack}}}{\tau} \to_p 1. \tag{2.3}$$

Secondly, let $\hat{\rho}_{ss} = \frac{m}{m_1^2 m_2^2}\sum_{k=1}^{m_1}(\sum_{j=1}^{m_2} u_{kj})^2 + \frac{m}{m_1^2 m_2^2}\sum_{k=1}^{m_2}(\sum_{i=1}^{m_1} u_{ik})^2$ and $\hat{\tau}_{ss} = \frac{m}{m_1^2 m_2^2}\sum_{k=1}^{m_1}(\sum_{j=1}^{m_2} v_{kj})^2 + \frac{m}{m_1^2 m_2^2}\sum_{k=1}^{m_2}(\sum_{i=1}^{m_1} v_{ik})^2$ where $u_{ij} = (X_i - Y_j)^T(\tilde{X}_i - \tilde{Y}_j)$ and $v_{ij} = \alpha^T(X_i - Y_j) + \alpha^T(\tilde{X}_i - \tilde{Y}_j)$. Then, we have

$$\frac{\hat{\rho}_{ss}}{\rho} \to_p 1 \quad \text{and} \quad \frac{\hat{\tau}_{ss}}{\tau} \to_p 1 \tag{2.4}$$

By combing the results above, we have the following:

Theorem 2.1 *Under either condition A1 or A2 or B1 in "Appendix" and* $H_0 : \mu_1 = \mu_2$, *for both* $i = jack$ *and* $i = ss$, *we have*

$$m\left(\frac{U_1^2}{\widehat{\rho}_i} + \frac{U_2^2}{\widehat{\tau}_i}\right) \to_d \chi_2^2. \tag{2.5}$$

Proof Assume that

$$\frac{\hat{\rho}}{\rho} \to_p 1 \quad \text{and} \quad \frac{\hat{\tau}}{\tau} \to_p 1.$$

By (2.2) and Slutsky's theorem, we have

$$\sqrt{m}\begin{pmatrix} \frac{U_1}{\sqrt{\widehat{\rho}}} \\ \frac{U_2}{\sqrt{\widehat{\tau}}} \end{pmatrix} \to_d N(0, I_2).$$

which in turn concludes (2.5) along with (2.3) and (2.4). □

We will call these simple χ_2^2 statistics $S1$ and $S2$ where (ρ, τ) is replaced by $(\widehat{\rho}_{\text{jack}}, \widehat{\tau}_{\text{jack}})$ and $(\widehat{\rho}_{ss}, \widehat{\tau}_{ss})$, respectively. Furthermore, by exploiting the asymptotic independence of U_1 and U_2,

$$\max\left(\frac{mU_1^2}{\widehat{\rho}}, \frac{mU_2^2}{\widehat{\tau}}\right)$$

can be used for testing $H_0 : \mu_1 = \mu_2$ as well. Here, the null distribution is the maximum of two independent chi-square statistics with one degree of freedom. We call these maximum statistics $M1$ and $M2$ where (ρ, τ) is replaced by $(\widehat{\rho}_{\text{jack}}, \widehat{\tau}_{\text{jack}})$ and $(\widehat{\rho}_{ss}, \widehat{\tau}_{ss})$, respectively.

3 Simulation study

In this section we compare several approaches in a simulation study by investigating the sizes and powers of our proposed methods (S1,S2,M1,M2) as well as the JEL test (JEL). For comparison, we consider simulation settings that are similar to those in Wang et al. (2013), but diversify the factors that can affect the statistical power. Assume that $W_1, \ldots, W_p$ are iid random variables, and $\bar{W}_1, \ldots, \bar{W}_p$ are iid random variables independent of W_i's. We consider eight different simulation settings: Four settings (Setting I) are under independence assumption between the variables, and the other four (Setting II) are correlated settings. Within the four independent or correlated settings, we investigate the differences in skewness, and also allow the mean shifts to have opposite signs. In each of these settings, $100c_2\,\%$ of the components of Y_1 have a shifted mean compared to the mean of X_1. Detailed descriptions are given below for each simulation setting:

- Setting I (Independent cases)
- Setting I-1: Let $W_i \sim N(0,1)$ and $\bar{W}_i \sim t(8)$. Define $X_{1,1} = W_1, X_{1,2} = W_2, \ldots, X_{1,p} = W_p, Y_{1,1} = \bar{W}_1 + \mu_{2,1}, Y_{1,2} = \bar{W}_2 + \mu_{2,2}, \ldots, Y_{1,p} = \bar{W}_p + \mu_{2,p}$, where $\mu_{2,i} = c_1$ if $i \leq [c_2 p]$, and $\mu_{2,i} = 0$ if $i > [c_2 p]$.
- Setting I-2: The same setting as I-1, except that $\mu_{2,i} = c_1$ for odd i, $\mu_{2,i} = -c_1$ for even i.
- Setting I-3: $\bar{W}_i \sim \chi^2(1) - 1$ where $\mu_{2,i} = c_1$ if $i \leq [c_2 p]$, and $\mu_{2,i} = 0$ if $i > [c_2 p]$.
- Setting I-4: The same setting as I-3, except that $\mu_{2,i} = c_1$ for odd i, $\mu_{2,i} = -c_1$ for even i.
- Setting II (Correlated cases used in Wang et al. (2013))
- Setting II-1 : Let $W_i \sim N(0,1)$ and $\bar{W}_i \sim t(8)$. Define $X_{1,1} = W_1, X_{1,2} = W_1 + W_2, \ldots, X_{1,p} = W_{p-1} + W_p, Y_{1,1} = \bar{W}_1 + \mu_{2,1}, Y_{1,2} = \bar{W}_1 + \bar{W}_2 + \mu_{2,2}, \ldots, Y_{1,p} = \bar{W}_{p-1} + \bar{W}_p + \mu_{2,p}$, where $\mu_{2,i} = c_1$ if $i \leq [c_2 p]$, and $\mu_{2,i} = 0$ if $i > [c_2 p]$.
- Setting II-2: The same setting as II-1, except that $\mu_{2,i} = c_1$ for odd i, $\mu_{2,i} = -c_1$ for even i.
- Setting II-3: $\bar{W}_i \sim \chi^2(1) - 1$, where $\mu_{2,i} = c_1$ if $i \leq [c_2 p]$, and $\mu_{2,i} = 0$ if $i > [c_2 p]$.
- Setting II-4: The same setting as II-3, except that $\mu_{2,i} = c_1$ for odd i, $\mu_{2,i} = -c_1$ for even i.

In this simulation, the null hypothesis to be tested is $H_0 : E(X_1) = E(Y_1)$. Note that if c_1 is zero there is no shift in the mean vector, so the size of tests can be investigated in this case. After generating 1000 random samples of sizes $n_1 = 30, 100, 150$ from $X = (X_{1,1}, \ldots, X_{1,p})^T$ and independently generating 1000 random samples of sizes $n_2 = 30, 100, 200$ from $Y = (Y_{1,1}, \ldots, Y_{1,p})^T$ with $p = 10, 20, \ldots, 100, 300, 500$, $c_1 = 0, 0.1$ and $c_2 = 0.25, 0.75$, we compute the powers of the five tests.

For comparisons of the five methods, we report the empirical sizes and powers for each simulation setting. The results are given in Tables S1 to S8 in the supplementary document, showing the proportion of rejecting the null $H_0 : \mu_1 = \mu_2$ out of 1000 replications. Each table is divided into three sections where the top part shows the

sizes and powers of our proposed tests (S1,M1,S2,M2) and the JEL for (n_1, n_2) = (30, 30), the middle part shows the results for (n_1, n_2) = (100, 100), and the bottom part for $(n_1, n_2) = (150, 200)$. For M1 and M2, a rejection is declared when the statistic is larger than 5, which corresponds to the 95% quantile of the maximum of two independent χ^2_1. Assuming that no prior information is available, $\alpha = (1, \ldots, 1)$ was used. Key findings from the simulation study can be summarized as below:

- S2 and M2 yield very low statistical power in (n_1, n_2) = (30, 30) and (100,100), illustrating that the use of jackknife samples is critical in boosting the statistical power when the sample sizes are not so high.
- Type I error of JEL is a little higher than its nominal counterpart (0.05) when (n_1, n_2) = (30, 30), whereas that of S1 and M1 is a little lower than 0.05. This explains the reason why JEL has power that is a little higher than S1 and M1 when (n_1, n_2) = (30, 30). The performances of JEL and S1 are comparable for (n_1, n_2) = (100, 100) and (150,200). Thus, the use of empirical likelihood does not seem to be critical.
- When the mean shifts have opposite signs, all the statistics have extremely low power. See Tables S2 and S4.
- The skewness of the distribution of data does not seem to affect the power much. This can be seen by comparing Tables S1 and S3.

4 The choice of α

The simulation study in Sect. 3 showed that both JEL as well as our proposed statistics perform badly when the shifted means have opposite signs. The reason for low statistical power in this case is mainly due to inappropriate choice of α. To understand this, suppose that $\alpha = (1, \ldots, 1)$ and the signs of $\mu_1 - \mu_2$ alternate. Then $\alpha^T(X_i - Y_i) \approx \alpha^T(\mu_1 - \mu_2) \approx 0$ because positive and negative mean shifts cancel each other out. We expect that the choice $\alpha = (1, \ldots, 1)$ is effective only when either positive or negative shifts dominate in $\mu_1 - \mu_2$. Otherwise we need a clever choice for α so that the mean shifts don't cancel each other out.

In particular, we consider the situation where there is no strong prior knowledge on the variables. Our strategy is to estimate the signs of the shifted means from the data. We first split the samples into three independent parts instead of splitting into two. The first two parts will be used to construct the two sample statistics as described in Sect. 2, and the remaining part will be used to estimate the signs. Let

$$\alpha^* = I\left(\tilde{\tilde{X}} - \tilde{\tilde{Y}} > 0\right) - I\left(\tilde{\tilde{X}} - \tilde{\tilde{Y}} < 0\right)$$

where $\tilde{\tilde{X}}$ and $\tilde{\tilde{Y}}$ correspond to the part of the dataset that is used to estimate the signs. Since α^* is independent of the construction of the two sample statistics, the choice of α^* does not change the asymptotic property of JEL and our proposed statistics under some regularity conditions. This can be rephrased as the following:

Corollary 1 *Suppose that for any* $s \in S$, $Var(s^T X_i) > 0$ *and* $Var(s^T Y_i) > 0$ *where* $S = \{(s_1, \ldots, s_p) | s_i = \pm 1\}$. *Under either condition A1 or A2 or B1 and* $H_0 : \mu_1 = \mu_2$, (2.5) *holds by conditioning on* $\alpha = \alpha^*$.

$\mathrm{Var}(s^T X_i) > 0$ and $\mathrm{Var}(s^T Y_i) > 0$ simply requires that X_i and Y_i should not be degenerate for any sign combination s. A different estimation method for α is possible as discussed in Wang et al. (2013). However, as they pointed out, the derived theorems cannot be applied to their choice directly, while they can be applied to our proposed choice directly.

A simulation study was implemented to evaluate the performance of the proposed approach. We use the same setting as in Settings II-2 and II-4 except that we now have $c_1 = 0.5$. The proportion of rejecting $H_0 : \mu_1 = \mu_2$ are shown in Tables S9–S12 in the supplementary file. Statistical power can be improved with larger c_1 even when $\alpha = (1, \ldots, 1)$, so c_1 should be fixed when comparing the results from $\alpha = (1, \ldots, 1)$ and the data-adaptive α. Tables S9 and S10 provide the results for $\alpha = (1, \ldots, 1)$, and Tables S11 and S12 are the results when α is estimated from 10 % of the dataset that was randomly selected. A substantial increase in the statistical power is observed in the results by using our data-adaptive method.

5 Analysis of Gene expression data

There are two major categories of gene set tests: competitive gene set tests and self-contained gene set tests (Goeman and Bühlmann 2007). Competitive gene set tests are concerned with the comparison of the set of genes of interest, say G, with the complementary set of genes which are not in G. On the other hand, self-contained gene set tests focus on the gene set of interest itself without reference to the complementary set of genes. An example of the former is Wu and Smyth (2012), which considered inter-gene correlation. The proposed two sample statistics in this paper belong to the category of self-contained gene set tests.

We analyze the Colon data available in the R-package "plsgenomics". This data set is from the microarray experiments of Colon tissue samples of Alon et al. (1999), and has 2000 gene expression levels where 22 of them are (n_1) normal tissues and 40 are (n_2) tumor tissues. To see the effect of genes with significant difference of sample means, Wang et al. (2013) applied those genes satisfying

$$\left| \frac{1}{n_1} \sum_{i=1}^{n_1} X_{ij} - \frac{1}{n_2} \sum_{i=1}^{n_2} Y_{ij} \right| \leq c_3$$

for some given threshold $c_3 > 0$. We report the p values for testing the equality of means of the genes with absolute difference in the sample means less than the threshold c_3 given in Table 1.

For $c_3 = 3000$, the p values of JEL and S1 are 0.136 and 0.182, respectively. These results are based on $\alpha = (1, \ldots, 1)$ as the coefficient vector. However, since the direction of differentially expressed genes can be inconsistent, it would be reasonable to apply the data-adaptive choice for α for the analysis. Two tissues are randomly selected from the normal group, and four tissues from tumor group, and α is computed based on these selected tissues. Then the two sample mean test is performed on the remaining samples with 20 from the normal group, and 36 in tumor group. The results are given in Table 1. When c_3 is greater than 1000 and α is estimated, p values from

Table 1 Colon data: p values for testing equal means of those genes with the absolute difference of sample means less than the threshold c_3

c_3	Number of genes	S1	M1	JEL	S1 (sign)	M1 (sign)	JEL (sign)
50	1158	0.25348	0.20599	0.21311	0.85591	0.84307	0.81001
100	1501	0.26248	0.24469	0.28235	0.18407	0.14564	0.13870
200	1742	0.29280	0.38249	0.38669	0.47060	0.43985	0.44518
500	1913	0.31903	0.37390	0.37484	0.23655	0.21193	0.17051
1000	1978	0.36252	0.30409	0.34012	0.00119	0.00059	0.00000
3000	2000	0.18160	0.30019	0.13591	0.00119	0.00124	0.00001

Table 2 Colon data (logarithm scale): p values for testing equal means of those genes with the absolute difference of sample means less than the threshold c_3

c_3	Number of genes	S1	M1	JEL	S1 (sign)	M1 (sign)	JEL (sign)
50	1158	0.07402	0.05116	0.02228	0.82515	0.86620	0.83694
100	1501	0.08650	0.07259	0.04217	0.17510	0.16022	0.14834
200	1742	0.09718	0.10602	0.07158	0.39625	0.31875	0.37152
500	1913	0.13240	0.18140	0.13368	0.07160	0.08697	0.03647
1000	1978	0.16000	0.23252	0.17692	0.00689	0.00340	0.00034
3000	2000	0.17991	0.26182	0.20634	0.00022	0.00006	0.00000

all the methods show highly significant results. Although this is an encouraging result, this should be interpreted carefully because few observations with large differences can have a large influence on the test results. In order to see whether the results are still significant when the effects by the large observations are removed, we apply log transformations to the 2000 gene expression levels. The results are given in Table 2. For testing the equality of means of the logarithms of the 2000 gene expression levels on normal colon tissues and tumor colon tissues, JEL and S1 with $\alpha = (1, \ldots, 1)$ give p values of 0.206 and 0.180, respectively. However, when we consider the data-adaptive α, all the results are highly significant. Normal and tumor tissues seem to have different mean vectors, but instead of making a quick judgement based on the results, it is recommended to investigate how such a large difference can be obtained from experiments, and to check the possibility of the biological justification for the mean difference.

6 Conclusion

In this paper, we propose alternative statistics for testing the equality of two high dimensional means, and study their finite sample properties. In our simulation study, we observe that the use of jackknife samples is substantial to gaining good statistical power, but the contribution of the empirical likelihood does not seem substantial. We

propose a new statistic that does not involve the empirical likelihood, eliminating the need for optimization procedures. We also provide significantly relaxed the sufficient conditions compared to what was required by Wang et al. (2013). Simulation results show that the choice of the coefficient vector is critical in all of the proposed methods. In many practical settings, $\alpha = (1, \ldots, 1)$ is a naive choice, so we propose a simple data-adaptive estimation for α. A numerical study shows substantial increase in statistical power for the practical settings that was considered, and this is also observed in the analysis results of the gene expression data.

There are some issues that remain as possible future research topics. First, we may consider different functional forms for U_2 to complement U_1, but to keep the necessary asymptotic theory simple, they are needed to have mean zero and correlation zero with U_1 under H_0. Otherwise, new theoretical developments will be required. It would be interesting to see whether power will increase substantially by using different functional forms of U_2. Second, there are enormous amount of accumulated biological information in modern research environment, and it would be interesting to incorporate the biological information to estimate α.

Acknowledgments This work was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT and Future Planning (NRF-2013R1A1A1061332) and INHA UNIVERSITY Research Grant.

Appendix 1: Condition A

In the following we give sufficient conditions for Theorem 2.1. For complete mathematical details, we refer to Wang et al. (2013):

$\min(n_1, n_2) \to \infty$
$\tau_1 = 2\alpha^T \Sigma \alpha > 0$
$\tau_2 = 2\alpha^T \tilde{\Sigma} \alpha > 0$
For some $\delta > 0$,

$$\frac{E[(X_1 - \mu_1)^T(\tilde{X}_1 - \mu_1)]^{2+\delta}}{\rho_1^{(2+\delta)/2}} = o\left(m_1^{\frac{\delta+\min(\delta,2)}{4}}\right),$$

$$\frac{E[(Y_1 - \mu_2)^T(\tilde{Y}_1 - \mu_2)]^{2+\delta}}{\rho_2^{(2+\delta)/2}} = o\left(m_2^{\frac{\delta+\min(\delta,2)}{4}}\right),$$

$$\frac{E[\alpha^T(X_1 + \tilde{X}_1 - 2\mu_1)]^{2+\delta}}{\tau_1^{(2+\delta)/2}} = o\left(m_1^{\frac{\delta+\min(\delta,2)}{4}}\right),$$

$$\frac{E[\alpha^T(Y_1 + \tilde{Y}_1 - 2\mu_2)]^{2+\delta}}{\rho_2^{(2+\delta)/2}} = o\left(m_2^{\frac{\delta+\min(\delta,2)}{4}}\right).$$

We call this condition A1.

Suppose that $\lambda_1 \le \cdots \le \lambda_p$ are the p eigenvalues of $\Sigma (= E((X_1 - \mu_1)(X_1 - \mu_1)^T))$, and $\tilde{\lambda}_1 \le \cdots \le \tilde{\lambda}_p$ are the p eigenvalues of $\tilde{\Sigma} (= E((Y_1 - \mu_2)(Y_1 - \mu_2)^T))$.

From Wang et al. (2013), it can be shown that if

$$0 < \liminf_{n_1 \to \infty} \lambda_1 \leq \liminf_{n_1 \to \infty} \lambda_p < \infty$$
$$0 < \liminf_{n_2 \to \infty} \tilde{\lambda}_1 \leq \liminf_{n_2 \to \infty} \tilde{\lambda}_p < \infty$$
$$\text{For some } \delta > 0, \frac{1}{p} \sum_{i=1}^{p} E \left\{ |X_{1,i} - \mu_{1,i}|^{2+\delta} + |Y_{1,i} - \mu_{2,i}|^{2+\delta} \right\} = O(1)$$
$$p = o\left(m^{\frac{\delta + \min(\delta, 2)}{2(2+\delta)}} \right).$$

holds, Theorem 2.1 holds. Thus, these four conditions (condition A2) can replace condition A1.

Consider the following Factor model as described in Wang et al. (2013): $X_i = \Gamma_1 B_i + \mu_1$ for $i = 1, \ldots, n_1$, $Y_j = \Gamma_2 \tilde{B}_j + \mu_2$ for $j = 1, \ldots, n_2$ where Γ_1 and Γ_2 are $p \times k$ with $\Gamma_1 \Gamma_1^T = \Sigma$ and $\Gamma_2 \Gamma_2^T = \tilde{\Sigma}$, $B_i = (B_{i1}, \ldots, B_{ip})$ $(i = 1, \ldots, n_1)$ and $\tilde{B}_j = (\tilde{B}_{j1}, \ldots, \tilde{B}_{jp})$ $(j = 1, \ldots, n_2)$. These are two independent random samples with $EB_i = E\tilde{B}_i = 0$, $Var(B_i) = Var(\tilde{B}_i) = I_{k \times k}$. $E(B_{i,j}^4) = 3 + \xi_1 < \infty$, $E(\tilde{B}_{i,j}^4) = 3 + \xi_2 < \infty$, $E \prod_{l=1}^{k} B_{il}^{\nu_l} = \prod_{l=1}^{k} E(B_{il}^{\nu_l})$ and $E \prod_{l=1}^{k} \tilde{B}_{il}^{\nu_l} = \prod_{l=1}^{k} E(\tilde{B}_{il}^{\nu_l})$ whenever $\nu_1 + \cdots + \nu_k = 4$ for distinct nonnegative integer ν_l's.

Wang et al. (2013) showed that under this condition, their asymptotic arguments hold without any restriction on p. Likewise, under this factor model assumption, Theorem 2.1 also holds for arbitrary p.

Appendix 2: Condition B

$\min(n_1, n_2) \to \infty$

$\tau_1 = 2\alpha^T \Sigma \alpha > 0$

$\tau_2 = 2\alpha^T \tilde{\Sigma} \alpha > 0$

$X_i - \mu_1 = \Sigma^{1/2} \varepsilon_i$ and $Y_i - \mu_2 = \tilde{\Sigma}^{1/2} \tilde{\varepsilon}_i$ where the elements in ε_i and $\tilde{\varepsilon}_i$ are i.i.d random variables with mean 0 and finite fourth moment.

$$\frac{\lambda_p^4 p^2}{m_1 \rho_1^2} = o(1), \quad \frac{\tilde{\lambda}_p^4 p^2}{m_2 \rho_2^2} = o(1).$$

We call this condition B1.

The following two boundedness conditions on the eigenvalues are called condition B2.

$$0 < \liminf_{n_1 \to \infty} \lambda_1 \leq \liminf_{n_1 \to \infty} \lambda_p < \infty$$
$$0 < \liminf_{n_2 \to \infty} \tilde{\lambda}_1 \leq \liminf_{n_2 \to \infty} \tilde{\lambda}_p < \infty$$

Appendix 3: Proof of the asymptotic chi-square limiting distribution for arbitrary order of p

Without loss of generality, we assume that $E(X_i) = E(\tilde{X}_i) = 0$. The key step in Wang et al. (2013) showing that the condition A1 is sufficient for Theorem 2.2 is obtained from the following two results.

(1) For $0 < \delta \leq 2$,

$$E\left[\sum_{i=1}^{m_1}(X_i^T\tilde{X}_i)^2 - m_1\rho_1\right]^{(2+\delta)/2} \leq O(m_1 E|X_1^T\tilde{X}_1|^{2+\delta})$$

(2) For $\delta > 2$,

$$E\left[\sum_{i=1}^{m_1}\left(X_i^T\tilde{X}_i\right)^2 - m_1\rho_1\right]^{(2+\delta)/2} \leq O\left(m_1^{(2+\delta)/4}E|X_1^T\tilde{X}_1|^{2+\delta}\right).$$

Instead of these, we directly evaluate

$$\begin{aligned} E\left[\sum_{i=1}^{m_1}\left(X_i^T\tilde{X}_i\right)^2 - m_1\rho_1\right]^2 &= m_1 E\left[\left(X_1^T\tilde{X}_1\right)^2 - \rho_1\right]^2 \\ &= m_1\left(E\left[\left(X_1^T\tilde{X}_1\right)\right]^4 - \rho_1^2\right). \end{aligned}$$

Let $X_1 = \Sigma^{1/2}\varepsilon_1$ and $\tilde{X}_1 = \Sigma^{1/2}\tilde{\varepsilon}_1$ where the elements in ε_1 and $\tilde{\varepsilon}_1$ are i.i.d random variables with mean 0 and finite fourth moment. Then,

$$\begin{aligned} E[(X_1^T\tilde{X}_1)]^4 &= E[(\varepsilon_1^T\Sigma\bar{\varepsilon}_1)]^4 \leq \lambda_p^4 E[(\varepsilon_1^T\bar{\varepsilon}_1)]^4 \\ &= \lambda_p^4\left(\sum_{j=1}^{p}E(\varepsilon_{1j}\bar{\varepsilon}_{1j})^4 + 6\sum_{j\neq k}E[(\varepsilon_{1j}\bar{\varepsilon}_{1j})^2(\varepsilon_{1k}\bar{\varepsilon}_{1k})^2]\right) \\ &= \lambda_p^4\left(\sum_{j=1}^{p}(E(\varepsilon_{1j})^4)^2 + 6\sum_{j\neq k}(E(\varepsilon_{1j}\varepsilon_{1k})^2)^2\right) \\ &= O(\lambda_p^4 p^2). \end{aligned}$$

Thus,

$$P\left(\left|\frac{\sum_{i=1}^{m_1}\left(X_i^T\tilde{X}_i\right)^2}{m_1\rho_1} - 1\right| > \varepsilon\right) \leq O\left(\frac{m_1\lambda_p^4 p^2}{m_1^2\rho_1^2}\right).$$

If the condition B1 holds,

$$\frac{\sum_{i=1}^{m_1}(X_i^T \tilde{X}_i)^2}{m_1 \rho_1} \to_p 1.$$

The rest can be shown in the same way as proved in Wang et al. (2013).

Note that if the condition B2 holds, then $\lambda_p^4/\rho_1^2 = O(1/p^2)$. Thus, for any order of p, since

$$\frac{\lambda_p^4 p^2}{m_1 \rho_1^2} = O\left(\frac{p^2}{m_1 p^2}\right) = O\left(\frac{1}{m_1}\right) \to 0, \quad \frac{\tilde{\lambda}_p^4 p^2}{m_2 \rho_2^2} = O\left(\frac{1}{m_2}\right) \to 0,$$

Theorem 2.2 holds.

References

Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci USA 96(12):6745750

Bai Z, Saranadasa H (2004) Effect of high dimension: by an example of a two sample problem. Stat Sin 6:311–329

Chen S, Qin Y (2010) A two-sample test for high-dimensional data with appliations to gene-set testing. Ann Stat 38(2):808–835

Goeman JJ, Bühlmann P (2007) Analyzing gene expression data in terms of gene sets: methodological issues. Bioinformatics 23:980–987

Ledoit O, Wolf M (2004) Honey, I Shrunk the Sample Covariance Matrix. J Portf Manag 30(4):110–119

Nettleton D, Recknor J, Reecy J (2008) Identification of differentially expressed gene categories in microarray studies using nonparametric multivariate analysis. Bioinformatics 24:192–201

Newton M, Quintana F Den, Boon J, Sengupta S, Ahlquist P (2007) Random set methods identify distinct aspects of the enrichment signal in gene-set analysis. Ann Appl Stat 1:85–106

Srivastava MS, Khatri CG (1979) An introduction to multivariate statistics. North Holland, New York

Wang R, Peng L, Qi Y (2013) Jackknife empirical likelihood test for equality of two high dimensional means. Stat Sin 23:667–690

Wu D, Smyth GK (2012) Camera: a competitive gene set test accounting for inter-gene correlation. Nucleic Acids Res 40(17):e133