Jackknife Empirical Likelihood Test for Equality of Two High Dimensional Means

Ruodu Wang[1], Liang Peng[1] and Yongcheng Qi[2]

Abstract

It has been a long history to test the equality of two multivariate means. One popular test is the so-called Hotelling T^2 test. However, as the dimension diverges, the Hotelling T^2 test performs poorly due to the possible inconsistency of the sample covariance estimation. To overcome this issue and allow the dimension to diverge as fast as possible, Bai and Saranadasa (1996) and Chen and Qin (2010) proposed tests without the sample covariance involved, and derived the asymptotic limits which depend on whether the dimension is fixed or diverges under a specific multivariate model. In this paper, we propose a jackknife empirical likelihood test which has a chi-square limit independent of the dimension, and the conditions are much weaker than those in the existing methods. A simulation study shows that the proposed new test has a very robust size with respect to the dimension, and is powerful too.

Keywords: Jackknife empirical likelihood, high dimensional mean, hypothesis test.

1 Introduction

Suppose $X_1 = (X_{1,1}, \cdots, X_{1,d})^T, \cdots, X_{n_1} = (X_{n_1,1}, \cdots, X_{n_1,d})^T$ and $Y_1 = (Y_{1,1}, \cdots, Y_{1,d})^T, \cdots, Y_{n_2} = (Y_{n_2,1}, \cdots, Y_{n_2,d})^T$ are two independent random samples with means μ_1 and μ_2, respectively. It has been a long history to test $H_0 : \mu_1 = \mu_2$ against $H_a : \mu_1 \neq \mu_2$ for a fixed dimension d. When both X_1 and Y_1 have a multivariate normal distribution with equal covariance, the well-known test is the

[1] School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332-0160, USA. Email: ruodu.wang@math.gatech.edu, peng@math.gatech.edu

[2] Department of Mathematics and Statistics, University of Minnesota–Duluth, 1117 University Drive, Duluth, MN 55812, USA. Email: yqi@d.umn.edu

so-called Hotelling T^2 test defined as

$$T^2 = \eta(\bar{\mathbf{X}} - \bar{\mathbf{Y}})^T A_n^{-1}(\bar{\mathbf{X}} - \bar{\mathbf{Y}}), \tag{1}$$

where $\eta = \frac{(n_1+n_2-2)n_1n_2}{n_1+n_2}$, $\bar{\mathbf{X}} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i$, $\bar{\mathbf{Y}} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i$ and $A_n = \sum_{i=1}^{n_1}(X_i - \bar{\mathbf{X}})(X_i - \bar{\mathbf{X}})^T + \sum_{i=1}^{n_2}(Y_i - \bar{\mathbf{Y}})(Y_i - \bar{\mathbf{Y}})^T$. However, when $d = d(n_1, n_2) \to \infty$, the Hotelling T^2 test performs poorly due to the possible inconsistency of the sample covariance estimation. When $d/(n_1+n_2) \to c \in (0,1)$, Bai and Saranadasa (1996) derived the asymptotic power of T^2. To overcome the restriction $c < 1$, Bai and Saranadasa (1996) proposed to employ

$$M_n = (\bar{\mathbf{X}} - \bar{\mathbf{Y}})^T(\bar{\mathbf{X}} - \bar{\mathbf{Y}}) - \eta^{-1}\mathrm{tr}(A_n)$$

instead of T^2 under a special multivariate model without assuming multivariate normality while keeping the condition of equal covariance, and derived the asymptotic limit when $d/n \to c > 0$. Recently Chen and Qin (2010) proposed to use the following test statistic

$$CQ = \frac{\sum_{i\neq j}^{n_1} X_i^T X_j}{n_1(n_1-1)} + \frac{\sum_{i\neq j}^{n_2} Y_i^T Y_j}{n_2(n_2-1)} - 2\frac{\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} X_i^T Y_j}{n_1n_2} \tag{2}$$

in order to allow d to be a possible larger order than that in Bai and Saranadasa (1996). Again, the asymptotic limit of the proposed test statistic CQ depends on whether the dimension is fixed or diverges, which results in either a normal limit or a chi-square limit, and special models for $\{X_i\}$ and $\{Y_i\}$ are employed. Another modification of Hotelling T^2 test is proposed by Srivastava and Du (2008) and Srivastava (2009) with the covariance matrix replaced by a diagonal matrix. Rates of convergence for high dimensional means are studied by Kuelbs and Vidyashankar (2010). For nonasymptotic studies of high dimensional means, we refer to Arlot, Blanchard and Roquain (2010a,b). Here, we are interested in seeking a test which does not need to distinguish whether the dimension is fixed or diverges.

By noting that $\mu_1 = \mu_2$ is equivalent to $(\mu_1 - \mu_2)^T(\mu_1 - \mu_2) = 0$, one may think of applying an empirical likelihood test to the estimating equation $\mathbb{E}\{(X_{i_1} - Y_{j_1})^T(X_{i_2} - Y_{j_2})\} = 0$ for $i_1 \neq i_2$

and $j_1 \neq j_2$. If one directly applies the empirical likelihood method based on estimating equations proposed in Qin and Lawless (1994) by using the samples $X_1, \cdots, X_{n_1}$ and $Y_1, \cdots, Y_{n_2}$, then one may define the empirical likelihood function as

$$\sup\{\{\prod_{i=1}^{n_1}(n_1p_i)\}\{\prod_{j=1}^{n_2}(n_2q_j)\} : p_1 \geq 0, \cdots, p_{n_1} \geq 0, q_1 \geq 0, \cdots, q_{n_2} \geq 0, \sum_{i=1}^{n_1} p_i = 1,$$
$$\sum_{j=1}^{n_2} q_j = 1, \sum_{i_1=1}^{n_1}\sum_{i_2\neq i_1}\sum_{j_1=1}^{n_2}\sum_{j_2\neq j_1}(p_{i_1}X_{i_1} - q_{j_1}Y_{j_1})^T(p_{i_2}X_{i_2} - q_{j_2}Y_{j_2}) = 0\},$$

which makes the minimization unsolvable. The reason is that the estimating equation defines a nonlinear functional, and in general one has to linearize the nonlinear functional before applying the empirical likelihood method. For more details on empirical likelihood methods, we refer to Owen (2001) and the review paper of Chen and Van Keilegom (2009). Recently, Jing, Yuan and Zhou (2009) proposed a so-called jackknife empirical likelihood method to construct confidence regions for nonlinear functionals with a particular focus on U-statistics. Using this idea, one needs to construct a jackknife sample based on the following estimator $n_1^{-1}(n_1-1)^{-1}n_2^{-1}(n_2-1)^{-1}\sum_{i_1\neq i_2}^{n_1}\sum_{j_1\neq j_2}^{n_2}(X_{i_1} - Y_{j_1})^T(X_{i_2} - Y_{j_2})$, which equals the statistic CQ given in (2). However, in order to have the jackknife empirical likelihood method work, one has to show that $\sqrt{n_1n_2}n_1^{-1}(n_1-1)^{-1}n_2^{-1}(n_2-1)^{-1}\sum_{i_1\neq i_2}^{n_1}\sum_{j_1\neq j_2}^{n_2}(X_{i_1} - Y_{j_1})^T(X_{i_2} - Y_{j_2})$ has a normal limit when $\mu_1 = \mu_2$. Consider $n_1 = n_2 = n, d = 1$ and $\mu_1 = \mu_2$. Then it is easy to see that

$$\begin{aligned} & n^{-1}(n-1)^{-2}\sum_{i_1\neq i_2}^{n}\sum_{j_1\neq j_2}^{n}(X_{i_1} - Y_{j_1})^T(X_{i_2} - Y_{j_2}) \\ = & \frac{1}{n-1}\{\sum_{i=1}^{n}(X_i - Y_i)\}^2 - \frac{1}{n-1}\sum_{i=1}^{n}(X_i - Y_i)^2 + \frac{2}{n(n-1)}\sum_{i=1}^{n}X_i\sum_{j=1}^{n}Y_j - \frac{2}{(n-1)}\sum_{i=1}^{n}X_iY_i \\ \overset{d}{\to} & (\chi_1^2 - 1)\mathbb{E}(X_1 - Y_1)^2 \end{aligned}$$

as $n \to \infty$, where χ_1^2 denotes a random variable having a chi-square distribution with 1 degree of freedom. Obviously, the above limit is not normally distributed. Hence a direct application of the jackknife empirical likelihood method to the statistic CQ will not lead to a chi-square limit.

In this paper, we propose a novel way to formulate a jackknife empirical likelihood test for testing $H_0 : \mu_1 = \mu_2$ against $H_a : \mu_1 \neq \mu_2$ by dividing the samples into two parts. The proposed new test has no need to distinguish whether the dimension is fixed or goes to infinity. It turns out that the asymptotic limit of the new test under H_0 is a chi-square limit independent of the dimension, the conditions on p and random vectors $\{X_i\}$ and $\{Y_j\}$ are weaker too. A simulation study shows that the size of the new test is quite stable with respect to the dimension and the proposed test is powerful as well.

We organize this paper as follows. In Section 2, the new methodology and main results are given. Section 3 presents a simulation study and a real data analysis. All proofs are put in Section 4.

2 Methodology

Throughout assume $X_i = (X_{i,1}, \cdots, X_{i,d})^T$ for $i = 1, \cdots, n_1$ and $Y_j = (Y_{j,1}, \cdots, Y_{j,d})^T$ for $j = 1, \cdots, n_2$ are two independent random samples with means μ_1 and μ_2, respectively. Assume $\min\{n_1, n_2\}$ goes to infinity. The question is to test $H_0 : \mu_1 = \mu_2$ against $H_a : \mu_1 \neq \mu_2$. Since $\mu_1 = \mu_2$ is equivalent to $(\mu_1 - \mu_2)^T(\mu_1 - \mu_2) = 0$ and $\mathbb{E}(X_{i_1} - Y_{j_1})^T(X_{i_2} - Y_{j_2}) = (\mu_1 - \mu_2)^T(\mu_1 - \mu_2)$ for $i_1 \neq i_2$ and $j_1 \neq j_2$, we propose to apply the jackknife empirical likelihood method to the above estimating equation. As explained in the introduction, a direct application fails to have a chi-square limit. Here we propose to split the samples into two groups as follows.

Put $m_1 = [n_1/2],\ m_2 = [n_2/2],\ m = m_1 + m_2$, $\bar{X}_i = X_{i+m_1}$ for $i = 1, \cdots, m_1$, and $\bar{Y}_j = Y_{j+m_2}$ for $j = 1, \cdots, m_2$. First we propose to estimate $(\mu_1 - \mu_2)^T(\mu_1 - \mu_2)$ by

$$\frac{1}{m_1 m_2} \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} (X_i - Y_j)^T (\bar{X}_i - \bar{Y}_j), \tag{3}$$

which is less efficient than the statistic CQ. However, it allows us to add more estimating equations and to employ the empirical likelihood method without estimating the asymptotic covariance. By noting that $\mathbb{E}\{(X_i - Y_j)^T(\bar{X}_i - \bar{Y}_j)\} = (\mu_1 - \mu_2)^T(\mu_1 - \mu_2) = ||\mu_1 - \mu_2||^2$ instead of $O(||\mu_1 - \mu_2||)$, one may expect that a test based on (3) will not be powerful for a small value of $||\mu_1 - \mu_2||$, confirmed by a brief simulation study. In order to improve the power, we propose to apply the jackknife empirical likelihood method in Jing, Yuan and Zhou (2009) to both (3) and a linear functional such as

$$\frac{1}{m_1m_2}\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\{\alpha^T(X_i-Y_j)+\alpha^T(\bar{X}_i-\bar{Y}_j)\} \tag{4}$$

rather than only (3), where $\alpha\in\mathbb{R}^d$ is a vector chosen based on prior information. Theoretically, when no additional information is available, any linear functional is a possible choice and $\alpha=(1,\cdots,1)^T\in\mathbb{R}^d$ is a convenient one. Note that more linear functionals in equation (4) with different choices of α can be added to further improve the power. In the simulation study we apply the jackknife empirical likelihood to (3) and (4) with $\alpha=(1,\cdots,1)\in\mathbb{R}^d$, which results in a test with good power and quite robust size with respect to the dimension.

As in Jing, Yuan and Zhou (2009), based on (3) and (4), we formulate the jackknife sample as $Z_k=(Z_{k,1},Z_{k,2})^T$ for $k=1,\cdots,m$, where

$$\left\{\begin{array}{rl} Z_{k,1}= & \frac{m_1+m_2}{m_1m_2}\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}(X_i-Y_j)^T(\bar{X}_i-\bar{Y}_j)\\ & -\frac{m_1+m_2-1}{(m_1-1)m_2}\sum_{i\neq k,i=1}^{m_1}\sum_{j=1}^{m_2}(X_i-Y_j)^T(\bar{X}_i-\bar{Y}_j)\\ Z_{k,2}= & \frac{m_1+m_2}{m_1m_2}\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\{\alpha^T(X_i-Y_j)+\alpha^T(\bar{X}_i-\bar{Y}_j)\}\\ & -\frac{m_1+m_2-1}{(m_1-1)m_2}\sum_{i\neq k,i=1}^{m_1}\sum_{j=1}^{m_2}\{\alpha^T(X_i-Y_j)+\alpha^T(\bar{X}_i-\bar{Y}_j)\}\end{array}\right.$$

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

$$\left\{\begin{array}{rl} Z_{k,1}= & \frac{m_1+m_2}{m_1m_2}\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}(X_i-Y_j)^T(\bar{X}_i-\bar{Y}_j)\\ & -\frac{m_1+m_2-1}{m_1(m_2-1)}\sum_{i=1}^{m_1}\sum_{j\neq k-m_1,j=1}^{m_2}(X_i-Y_j)^T(\bar{X}_i-\bar{Y}_j)\\ Z_{k,2}= & \frac{m_1+m_2}{m_1m_2}\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\{\alpha^T(X_i-Y_j)+\alpha^T(\bar{X}_i-\bar{Y}_j)\}\\ & -\frac{m_1+m_2-1}{m_1(m_2-1)}\sum_{i=1}^{m_1}\sum_{j\neq k-m_1,j=1}^{m_2}\{\alpha^T(X_i-Y_j)+\alpha^T(\bar{X}_i-\bar{Y}_j)\}\end{array}\right.$$

for $k=m_1+1,\cdots,m$. Based on this jackknife sample, the jackknife empirical likelihood ratio function for testing $H_0:\mu_1=\mu_2$ is defined as

$$L_m=\sup\{\prod_{i=1}^m(mp_i):p_1\geq 0,\cdots,p_m\geq 0,\sum_{i=1}^m p_i=1,\sum_{i=1}^m p_iZ_i=(0,0)^T\}.$$

By the Lagrange multiplier technique, we have $p_i=m^{-1}\{1+\beta^TZ_i\}^{-1}$ for $i=1,\cdots,m$ and

$$l_m:=-2\log L_m=2\sum_{i=1}^m\log\{1+\beta^TZ_i\},$$

where β satisfies

$$\frac{1}{m}\sum_{i=1}^{m}\frac{Z_i}{1+\beta^T Z_i}=(0,0)^T. \tag{5}$$

Write $\Sigma=(\sigma_{i,j})_{1\le i\le d,1\le j\le d}=\mathbb{E}\{(X_1-\mu_1)(X_1-\mu_1)^T\}$, the covariance matrix of X_1, and use $\lambda_1\le\cdots\le\lambda_d$ to denote the d eigenvalues of the matrix Σ. Similarly, write $\bar{\Sigma}=(\bar{\sigma}_{i,j})_{1\le i\le d,1\le j\le d}=\mathbb{E}\{(Y_1-\mu_2)(Y_1-\mu_2)^T\}$ and use $\bar{\lambda}_1\le\cdots\le\bar{\lambda}_d$ to denote the d eigenvalues of the matrix $\bar{\Sigma}$. Also write

$$\rho_1=\sum_{i,j=1}^{d}\sigma_{i,j}^2=\mathbf{tr}(\Sigma^2),\ \rho_2=\sum_{i,j=1}^{d}\bar{\sigma}_{i,j}^2=\mathbf{tr}(\bar{\Sigma}^2),\ \tau_1=2\alpha^T\Sigma\alpha,\ \tau_2=2\alpha^T\bar{\Sigma}\alpha. \tag{6}$$

Here $\mathbf{tr}$ means trace for a matrix. Note that $\rho_1=\mathbb{E}[(X_1-\mu_1)^T(X_2-\mu_1)]^2$, $\rho_2=\mathbb{E}[(Y_1-\mu_2)^T(Y_2-\mu_2)]^2$, $\tau_1=2\mathbb{E}[\alpha^T(X_1-\mu_1)]^2$ and $\tau_2=2\mathbb{E}[\alpha^T(Y_1-\mu_2)]^2$, and these quantities may depend on n_1,n_2 since d may depend on n_1,n_2.

Theorem 1. *Assume* $\min\{n_1,n_2\}\to\infty$*,* τ_1 *and* τ_2 *in (6) are positive, and for some* $\delta>0$*,*

$$\frac{\mathbb{E}|(X_1-\mu_1)^T(\bar{X}_1-\mu_1)|^{2+\delta}}{\rho_1^{(2+\delta)/2}}=o(m_1^{\frac{\delta+\min(\delta,2)}{4}}), \tag{7}$$

$$\frac{\mathbb{E}|(Y_1-\mu_2)^T(\bar{Y}_1-\mu_2)|^{2+\delta}}{\rho_2^{(2+\delta)/2}}=o(m_2^{\frac{\delta+\min(\delta,2)}{4}}), \tag{8}$$

$$\frac{\mathbb{E}|\alpha^T(X_1+\bar{X}_1-2\mu_1)|^{2+\delta}}{\tau_1^{(2+\delta)/2}}=o(m_1^{\frac{\delta+\min(\delta,2)}{4}}), \tag{9}$$

and

$$\frac{\mathbb{E}|\alpha^T(Y_1+\bar{Y}_1-2\mu_2)|^{2+\delta}}{\tau_2^{(2+\delta)/2}}=o(m_2^{\frac{\delta+\min(\delta,2)}{4}}). \tag{10}$$

Then, under $H_0:\mu_1=\mu_2$*,* l_m *converges in distribution to a chi-square distribution with two degrees of freedom as* $\min\{n_1,n_2\}\to\infty$*.*

Based on the above theorem, one can test $H_0:\mu_1=\mu_2$ against $H_a:\mu_1\ne\mu_2$ by rejecting H_0 when $l_m\ge\chi^2_{2,\gamma}$, where $\chi^2_{2,\gamma}$ denotes the $(1-\gamma)-$quantile of a chi-square distribution with two degrees of freedom and γ is the significant level.

Remark 1. In equations (7)–(10), the restrictions are put on $\mathbb{E}|W|^{2+\delta}/(\mathbb{E}W^2)^{(2+\delta)/2}$ for some random variables W, which are necessary for the CLT to hold for random arrays. Later we will see those conditions are satisfied by imposing some conditions on the higher-order moments or special dependence structure.

Remark 2. The proposed test has the following merits:

1. The limiting distribution is always chi-square without estimating the asymptotic covariance.

2. It does not require any specific structure such as the one used in Bai and Saranadasa (1996) and Chen and Qin (2010), which will be discussed later.

3. With higher-order moment condition or special dependence structure of $\{X_i\}$ and $\{Y_i\}$, d can be very large.

4. There is no restriction imposed on the relation between n_1 and n_2 except that $\min\{n_1, n_2\} \to \infty$. That is, no need to assume a limit or bound on the ratio n_1/n_2. Moreover, no assumptions are needed on ρ_1/ρ_2 or τ_1/τ_2. Hence the covariance matrices Σ_1 and Σ_2 can be arbitrary as long as τ_1,$\tau_2 > 0$, which are simply equivalent to saying that both $\alpha^T X_1$ and $\alpha^T Y_1$ are non-degenerate random variables.

Next we verify Theorem 1 by imposing conditions on the moments and dimension of the random vectors.

A1: $0 < \liminf_{n_1\to\infty} \lambda_1 \le \limsup_{n_1\to\infty} \lambda_d < \infty$ and $0 < \liminf_{n_2\to\infty} \bar{\lambda}_1 \le \limsup_{n_2\to\infty} \bar{\lambda}_d < \infty$;

A2: For some $\delta > 0$, $\frac{1}{d}\sum_{i=1}^d \mathbb{E}\{|X_{1,i} - \mu_{1,i}|^{2+\delta} + |Y_{1,i} - \mu_{2,i}|^{2+\delta}\} = O(1)$;

A3: $d = o(m^{\frac{\delta+\min(\delta,2)}{2(2+\delta)}})$.

Corollary 1. *Assume* $\min\{n_1, n_2\} \to \infty$ *and conditions* **A1**–**A3** *hold. Then, under* $H_0 : \mu_1 = \mu_2$, *conditions* (7) – (10) *are satisfied, i.e., Theorem 1 holds.*

Condition **A3** is a somewhat restrictive condition for the dimension d. Note that conditions **A1** and **A2** are related only to the covariance matrices and some higher moments on the components of the random vectors. The higher moments we have, the less restriction is imposed on d. Condition **A3** can be removed for models with some special dependence structures. For comparisons, we prove the Wilks theorem for the proposed jackknife empirical likelihood test under the following model **B** considered by Bai and Saranadasa (1996), Chen, Peng and Qin (2009) and Chen and Qin (2010):

B (Factor model). $X_i = \Gamma_1 B_i + \mu_1$ for $i = 1, \cdots, n_1$, $Y_j = \Gamma_2 \bar{B}_j + \mu_2$ for $j = 1, \cdots, n_2$, where Γ_1, Γ_2 are $d \times k$ matrices with $\Gamma_1 \Gamma_1^T = \Sigma$, $\Gamma_2 \Gamma_2^T = \bar{\Sigma}$, $\{B_i = (B_{i,1}, \cdots, B_{i,k})^T\}_{i=1}^{n_1}$ and $\{\bar{B}_j = (\bar{B}_{j,1}, \cdots, \bar{B}_{j,k})^T\}_{i=1}^{n_2}$ are two independent random samples satisfying that $\mathbb{E}B_i = \mathbb{E}\bar{B}_i = 0$, $\mathrm{Var}(B_i) = \mathrm{Var}(\bar{B}_i) = I_{k\times k}$, $\mathbb{E}B_{i,j}^4 = 3 + \xi_1 < \infty$, $\mathbb{E}\bar{B}_{i,j}^4 = 3 + \xi_2 < \infty$, $\mathbb{E}\prod_{l=1}^k B_{i,l}^{\nu_l} = \prod_{l=1}^k \mathbb{E}B_{i,l}^{\nu_l}$ and $\mathbb{E}\prod_{l=1}^k \bar{B}_{i,l}^{\nu_l} = \prod_{l=1}^k \mathbb{E}\bar{B}_{i,l}^{\nu_l}$ whenever $\nu_1 + \cdots + \nu_k = 4$ for distinct nonnegative integers ν_l's.

Theorem 2. *Assume* $\min\{n_1, n_2\} \to \infty$*, and* τ_1 *and* τ_2 *in (6) are positive. Under model* ***B*** *and* $H_0 : \mu_1 = \mu_2$*,* l_m *converges in distribution to a chi-square distribution with two degrees of freedom as* $\min\{n_1, n_2\} \to \infty$*.*

Remark 3. It can be seen from the proof of Theorem 2 that assumptions $\mathbb{E}B_{i,j}^4 = 3 + \xi_1 < \infty$ and $\mathbb{E}\bar{B}_{i,j}^4 = 3 + \xi_2 < \infty$ in model **B** can be replaced by the much weaker conditions $\max_{1\le j\le k} \mathbb{E}B_{1,j}^4 = o(m)$ and $\max_{1\le j\le k} \mathbb{E}\bar{B}_{1,j}^4 = o(m)$. Unlike Bai and Saranadasa (1996) and Chen and Qin (2010), there is no restriction on d and k for our proposed method. The only constraint imposed on matrices Γ_1 and Γ_2 is that both $\alpha^T \Sigma \alpha$ and $\alpha^T \bar{\Sigma} \alpha$ are positive, which is very weak.

3 Simulation study and data analysis

3.1 Simulation study

We investigate the finite sample behavior of the proposed jackknife empirical likelihood test (JEL) and compare it with the test statistic in (2) proposed by Chen and Qin (2010) in terms of both size and power.

Let $W_1, \cdots, W_d$ be iid random variables with the standard normal distribution function, and let $\bar{W}_1, \cdots, \bar{W}_d$, independent of $W_i's$ be iid random variables with distribution function $t(8)$. Put $X_{1,1} = W_1, X_{1,2} = W_1 + W_2, \cdots, X_{1,d} = W_{d-1} + W_d$, $Y_{1,1} = \bar{W}_1 + \mu_{2,1}, Y_{1,2} = \bar{W}_1 + \bar{W}_2 + \mu_{2,2}, \cdots, Y_{1,d} = \bar{W}_{d-1} + \bar{W}_d + \mu_{2,d}$, where $\mu_{2,i} = c_1$ if $i \leq [c_2 d]$, and $\mu_{2,i} = 0$ if $i > [c_2 d]$. That is, $100c_2\%$ of the components of Y_1 have a shifted mean compared to that of X_1.

Since we test $H_0 : \mathbb{E}X_1 = \mathbb{E}Y_1$ against $H_a : \mathbb{E}X_1 \neq \mathbb{E}Y_1$, the case of $c_1 = 0$ denotes the size of tests. By drawing $1,000$ random samples of sizes $n_1 = 30, 100, 150$ from $X = (X_{1,1}, \cdots, X_{1,d})^T$ and independently drawing $1,000$ random samples of sizes $n_2 = 30, 100, 200$ from $Y = (Y_{1,1}, \cdots, Y_{1,d})^T$ with $d = 10, 20, \cdots, 100, 300, 500$, $c_1 = 0, 0.1$ and $c_2 = 0.25, 0.75$, we calculate the powers of the two tests mentioned above.

In Tables 1–3, we report the empirical sizes and powers for the proposed jackknife empirical likelihood test with $\alpha = (1, \cdots, 1)^T \in \mathbb{R}^d$ and the test in Chen and Qin (2010) at level 5%. Results for level 10% are similar. From these three tables, we observe that (i) the sizes of both tests, i.e., results for $c_1 = 0$ are comparable and quite stable with respect to the dimension d; (ii) the proposed jackknife empirical likelihood test is more powerful than the test in Chen and Qin (2010) for the case of $c_2 = 0.75$ and the case when the data is sparse, but d is large (i.e., the case of $c_1 = 0.1, c_2 = 0.25$). Since equation (4) has nothing to do with sparsity, it is expected that the proposed jackknife empirical likelihood method is not powerful when the data is sparse. Hence, it would be of interest to connect sparsity with some estimating equations so as to improve the power of the proposed jackknife empirical likelihood test.

In conclusion, the proposed jackknife empirical likelihood test has a very stable size with respect to the dimension and is powerful as well. Moreover, the new test is easy to compute, flexible to take other information into account, and works for both fixed dimension and divergent dimension. In comparison with the test in Chen and Qin (2010), the new test has a comparable size, but is more powerful when the data is not very sparse. Some further research on formulating sparsity into estimating equations will be pursued in future.

3.2 Data analysis

In this subsection we employ the proposed method to the Colon data with 2000 gene expression levels on 22 (n_1) normal colon tissues and 40 (n_2) tumor colon tissues. This data set is available from the R package 'plsgenomics', and has been anlayzed by Alon et al. (1999) and Srivastava and Du (2009). The p-values of three tests proposed by Srivastava and Du (2009) equal to 1.38e-06, 4.48e-11 and 0.00000, which clearly reject the null hypothesis that the tumor group has the same gene expression levels as the normal group. A direct application of the proposed jackknife empirical likelihood method and the CQ test for testing the equality of means gives p-values 1.36e-01 and 5.06e-09, respectively, which clearly result in a contradiction. Although the test in Chen and Qin (2010) and the three tests in Srivastava and Du (2009) clearly reject the null hypothesis, the p-values are significantly different. A closer look at the difference of sample means (see Figure 1) shows that some genes have a significant difference of sample means and a high variability, which may play an important role in the CQ test and the equation (4) with $\alpha = (1, \cdots, 1)^T \in \mathbb{R}^d$ of the proposed jackknife empirical likelihood method. To examine this effect, we apply the methods to those genes without the significant difference in sample means and the logarithms of the 2000 gene expression levels.

First we apply the proposed jackknife empirical likelihood method and the CQ test to those genes satisfying $|n_1^{-1}\sum_{i=1}^{n_1} X_{i,j} - n_2^{-1}\sum_{i=1}^{n_2} Y_{i,j}| \le c_3$ for some given threshold $c_3 > 0$. In Table 4, we report the p-values for these two methods for different c_3, which confirm that some genes play an important role in rejecting the equality of means in the CQ test.

Figure 1 clearly shows that some genes have a large positive mean and some genes have a large negative mean, and the equation (4) with the simple $a = (1, 1, \cdots, 1)$ can not catch this characteristic.

Here, we propose to replace (4) by

$$\frac{1}{m_1 m_2}\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}(X_i - Y_j)\{I(\bar{X}_i - \bar{Y}_j > 0) - I(\bar{X}_i - \bar{Y}_j < 0)\}, \tag{11}$$

which results in the P-value 2.63e-03, and so we reject the null hypothesis that the tumor group has the same gene expression levels as the normal group. Although the derived theorems based on (3) and (4) can not be applied to (3) and (11) directly, results hold under some similar conditions by noting the boundedness of $I(\bar{X}_i - \bar{Y}_j > 0) - I(\bar{X}_i - \bar{Y}_j < 0)$.

It is well known that gene expression data are quite noisy and some transformation and normalization are needed before doing statistical analysis; see Chapter 6 of Lee (2004). Here we apply the CQ test and the proposed empirical likelihood methods based on (3) and (4), and (3) and (11) to testing the equality of means of the logarithms of the 2000 gene expression levels on normal colon tissues and tumor colon tissues, which give p-values 0.184, 0.206 and 0.148, respectively. That is, one can not reject the equality of means of the logarithms of the 2000 gene expression levels. We plot the differences of sample means of the logarithms in Figure 2, which are much less volatile than the differences of sample means in Figure 1.

In summary, carefully choosing α in the empirical likelihood method is needed when it is applied to the colon data, which has a small sample size and a large variation. However simply choosing $\alpha = (1, \cdots, 1)^T$ in the empirical likelihood method gives a similar result as the test in Chen and Qin (2010) for testing the equal means of the logarithms of Colon data.

4 Proofs

In the proofs we use $||\cdot||$ to denote the L_2 norm of a vector or matrix. Since $\mu_1 - \mu_2$ is our target and under the null hypothesis $\mu_1 - \mu_2 = 0$, without loss of generality we assume $\mu_1 = \mu_2 = 0$. Write $u_{ij} = (X_i - Y_j)^T(\bar{X}_i - \bar{Y}_j)$ and $v_{ij} = \alpha^T(X_i - Y_j) + \alpha^T(\bar{X}_i - \bar{Y}_j)$ for $1 \le i \le m_1, 1 \le j \le m_2$. Then it is easily verified that for $1 \leq i, k \leq m_1, 1 \leq j, l \leq m_2$,

$$\mathbb{E}(u_{ij}) = \mathbb{E}(v_{kl}) = \mathbb{E}(u_{ij}v_{kl}) = 0,$$

$$\mathrm{Var}(u_{kl}) = \sum_{i,j=1}^{d} (\sigma_{i,j}^2 + \bar{\sigma}_{i,j}^2) = \rho_1 + \rho_2,$$

and

$$\mathrm{Var}(v_{kl}) = 2\alpha^T(\Sigma + \bar{\Sigma})\alpha = \tau_1 + \tau_2.$$

Lemma 1. *Under conditions of Theorem 1, we have as* $\min\{n_1, n_2\} \to \infty$

$$\frac{1}{\sqrt{m_1}} \sum_{i=1}^{m_1} \frac{X_i^T \bar{X}_i}{\sqrt{\rho_1}} \xrightarrow{d} N(0,1), \tag{12}$$

$$\frac{1}{\sqrt{m_2}} \sum_{j=1}^{m_2} \frac{Y_j^T \bar{Y}_j}{\sqrt{\rho_2}} \xrightarrow{d} N(0,1), \tag{13}$$

$$\frac{1}{\sqrt{m_1}} \sum_{i=1}^{m_1} \frac{\alpha^T(X_i + \bar{X}_i)}{\sqrt{\tau_1}} \xrightarrow{d} N(0,1), \tag{14}$$

and

$$\frac{1}{\sqrt{m_2}} \sum_{j=1}^{m_2} \frac{\alpha^T(Y_j + \bar{Y}_j)}{\sqrt{\tau_2}} \xrightarrow{d} N(0,1). \tag{15}$$

Proof. Since $\mathrm{Var}(X_i^T \bar{X}_i) = \rho_1$ and $X_1^T \bar{X}_1, \cdots, X_{m_1}^T \bar{X}_{m_1}$ are i.i.d. for fixed m_1, equation (12) follows from (7) and the Lyapunov central limit theorem. The rest can be shown in the same way. □

From now on we denote

$$\rho = \frac{m}{m_1}\rho_1 + \frac{m}{m_2}\rho_2 \quad \text{and} \quad \tau = \frac{m}{m_1}\tau_1 + \frac{m}{m_2}\tau_2.$$

Lemma 2. *Under conditions of Theorem 1, we have as* $\min\{n_1, n_2\} \to \infty$

$$\frac{\sqrt{m}}{m_1 m_2 \sqrt{\rho}} \sum_{i=1}^{m_1} X_i^T \sum_{j=1}^{m_2} \bar{Y}_j \xrightarrow{p} 0, \tag{16}$$

$$\frac{1}{m_1\sqrt{\tau}} \sum_{i=1}^{m_1} \alpha^T X_i \xrightarrow{p} 0, \tag{17}$$

$$\frac{1}{m_2\sqrt{\tau}}\sum_{j=1}^{m_2}\alpha^T Y_j \xrightarrow{p} 0, \tag{18}$$

$$\frac{1}{m_1}\sum_{i=1}^{m_1}\frac{(X_i^T\bar{X}_i)^2}{\rho_1} \xrightarrow{p} 1, \tag{19}$$

$$\frac{1}{m_2}\sum_{j=1}^{m_2}\frac{(Y_j^T\bar{Y}_j)^2}{\rho_2} \xrightarrow{p} 1, \tag{20}$$

$$\frac{1}{m_1}\sum_{i=1}^{m_1}\frac{[\alpha^T(X_i+\bar{X}_i)]^2}{\tau_1} \xrightarrow{p} 1, \tag{21}$$

$$\frac{1}{m_2}\sum_{j=1}^{m_2}\frac{[\alpha^T(Y_j+\bar{Y}_j)]^2}{\tau_2} \xrightarrow{p} 1, \tag{22}$$

$$\frac{1}{m_1}\sum_{i=1}^{m_1}\frac{X_i^T\bar{X}_i[\alpha^T(X_i+\bar{X}_i)]}{\sqrt{\rho_1\tau_1}} \xrightarrow{p} 0, \tag{23}$$

and

$$\frac{1}{m_2}\sum_{i=1}^{m_2}\frac{Y_j^T\bar{Y}_j[\alpha^T(Y_j+\bar{Y}_j)]}{\sqrt{\rho_2\tau_2}} \xrightarrow{p} 0. \tag{24}$$

Proof. Note that $\mu_1=\mu_2=0$ is assumed in Section 4. Then (16) follows from the fact that

$$\begin{aligned}
\mathrm{Var}(\frac{\sqrt{m}}{m_1m_2\sqrt{\rho}}\sum_{i=1}^{m_1}X_i^T\sum_{j=1}^{m_2}Y_j) &= \mathbb{E}\left[\frac{m}{m_1^2m_2^2\rho}\left(\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}X_i^TY_j\right)^2\right] \\
&= \mathbb{E}\left[\frac{m}{m_1^2m_2^2\rho}\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}(X_i^TY_j)^2\right] \\
&= \mathbb{E}\left[\frac{m}{m_1m_2\rho}(X_1^TY_1)^2\right] \\
&= \frac{m}{m_1m_2\rho}\sum_{i,j=1}^{d}\sigma_{ij}\bar{\sigma}_{ij} \\
&\le \frac{m}{m_1+m_2}\frac{\rho_1+\rho_2}{2\rho} \\
&\le \frac{1}{2}(\frac{1}{m_1}+\frac{1}{m_2}) \\
&= o(1).
\end{aligned}$$

In the same way, we can show (17) and (18).

To show (19), write $u_i = X_i^T \bar{X}_i$. We need to estimate $\mathbb{E}|\sum_{i=1}^{m_1} u_i^2 - m_1\rho_1|^{(2+\delta)/2}$. Note that $\rho_1 = \mathbb{E}u_1^2$. When $0 < \delta \leq 2$, it follows from von Bahr and Esseen (1965) that

$$\mathbb{E}|\sum_{i=1}^{m_1} u_i^2 - m_1\rho_1|^{(2+\delta)/2} \leq 2m_1\mathbb{E}|u_1^2 - \mathbb{E}(u_1^2)|^{(2+\delta)/2} = O(m_1\mathbb{E}|u_1|^{2+\delta}). \tag{25}$$

When $\delta > 2$, it follows from Dharmadhikari and Jogdeo (1969) that

$$\mathbb{E}|\sum_{i=1}^{m_1} u_i^2 - m_1\rho_1|^{(2+\delta)/2} = O(m_1^{(2+\delta)/4}\mathbb{E}|u_1^2 - \mathbb{E}(u_1^2)|^{(2+\delta)/2}) = O(m_1^{(2+\delta)/4}\mathbb{E}|u_1|^{2+\delta}). \tag{26}$$

Therefore, by (25), (26) and (7) we have for any $\varepsilon > 0$

$$\begin{aligned} & \mathbb{P}(|\frac{\sum_{i=1}^{m_1} u_i^2}{m_1\rho_1} - 1| > \varepsilon) \\ \leq \quad & \varepsilon^{-(2+\delta)/2}\frac{\mathbb{E}|\sum_{i=1}^{m_1} u_i^2 - m\rho_1|^{(2+\delta)/2}}{(m_1\rho_1)^{(2+\delta)/2}} \\ = \quad & O(m_1^{-(\delta+\min(\delta,2))/4}\mathbb{E}|\frac{u_1}{\sqrt{\rho_1}}|^{2+\delta}) \\ = \quad & o(1), \end{aligned}$$

which implies (19). The rest can be shown in the same way. □

Lemma 3. *Under conditions of Theorem 1, we have as* $\min\{n_1, n_2\} \to \infty$

$$\frac{\sqrt{m}}{m_1m_2}\sum_{j=1}^{m_2}\sum_{i=1}^{m_1}\begin{pmatrix} \frac{u_{ij}}{\sqrt{\rho}} \\ \frac{v_{ij}}{\sqrt{\tau}} \end{pmatrix} \xrightarrow{d} N(0, I_2), \tag{27}$$

$$\frac{m}{m_1^2m_2^2\rho}\sum_{k=1}^{m_1}\left(\sum_{j=1}^{m_2} u_{kj}\right)^2 - \frac{m\rho_1}{m_1\rho} \xrightarrow{p} 0, \tag{28}$$

$$\frac{m}{m_1^2m_2^2\rho}\sum_{k=1}^{m_2}\left(\sum_{i=1}^{m_1} u_{ik}\right)^2 - \frac{m\rho_2}{m_2\rho} \xrightarrow{p} 0, \tag{29}$$

$$\frac{m}{m_1^2m_2^2\tau}\sum_{k=1}^{m_1}\left(\sum_{j=1}^{m_2} v_{kj}\right)^2 - \frac{m\tau_1}{m_1\tau} \xrightarrow{p} 0, \tag{30}$$

$$\frac{m}{m_1^2m_2^2\tau}\sum_{k=1}^{m_2}\left(\sum_{i=1}^{m_1} v_{ik}\right)^2 - \frac{m\tau_2}{m_1\tau} \xrightarrow{p} 0, \tag{31}$$

$$\frac{m}{m_1^2 m_2^2 \sqrt{\rho\tau}} \sum_{k=1}^{m_1} \left(\sum_{i=1}^{m_2} u_{ki} \sum_{j=1}^{m_2} v_{kj} \right) \xrightarrow{p} 0, \tag{32}$$

$$\frac{m}{m_1^2 m_2^2 \sqrt{\rho\tau}} \sum_{k=1}^{m_2} \left(\sum_{i=1}^{m_1} u_{ik} \sum_{j=1}^{m_1} v_{jk} \right) \xrightarrow{p} 0, \tag{33}$$

where I_2 is the 2×2 identity matrix.

Proof. It follows from Lemma 2 that

$$\begin{aligned}
\frac{\sqrt{m}}{m_1 m_2} \sum_{j=1}^{m_2} \sum_{i=1}^{m_1} \frac{u_{ij}}{\sqrt{\rho}} &= \frac{\sqrt{m}}{m_1 m_2 \sqrt{\rho}} \sum_{j=1}^{m_2} \sum_{i=1}^{m_1} (X_i^T \bar{X}_i + Y_j^T \bar{Y}_j - X_i^T \bar{Y}_j - Y_j^T \bar{X}_i) \\
&= \frac{\sqrt{m}}{m_1 \sqrt{\rho}} \sum_{i=1}^{m_1} X_i^T \bar{X}_i + \frac{\sqrt{m}}{m_2 \sqrt{\rho}} \sum_{j=1}^{m_2} Y_j^T \bar{Y}_j - \frac{\sqrt{m}}{m_1 m_2 \sqrt{\rho}} \sum_{j=1}^{m_2} \sum_{i=1}^{m_1} (X_i^T \bar{Y}_j + Y_j^T \bar{X}_i) \\
&= \frac{\sqrt{m\rho_1}}{\sqrt{m_1 \rho}} \frac{1}{\sqrt{m_1}} \sum_{i=1}^{m_1} \frac{X_i^T \bar{X}_i}{\sqrt{\rho_1}} + \frac{\sqrt{m\rho_2}}{\sqrt{m_2 \rho}} \frac{1}{\sqrt{m_2}} \sum_{j=1}^{m_2} \frac{Y_j^T \bar{Y}_j}{\sqrt{\rho_2}} + o_p(1) \\
&= a_m A_m + b_m B_m + o_p(1),
\end{aligned}$$

where $a_m = \frac{\sqrt{m\rho_1}}{\sqrt{m_1 \rho}}$, $b_m = \frac{\sqrt{m\rho_2}}{\sqrt{m_2 \rho}}$, $A_m = \frac{1}{\sqrt{m_1}} \sum_{i=1}^{m_1} \frac{X_i^T \bar{X}_i}{\sqrt{\rho_1}} \xrightarrow{d} N(0,1)$ and $B_m = \frac{1}{\sqrt{m_2}} \sum_{j=1}^{m_2} \frac{Y_j^T \bar{Y}_j}{\sqrt{\rho_2}} \xrightarrow{d} N(0,1)$. Obviously $a_m^2 + b_m^2 = 1$ and A_m, B_m are independent. Denote the characteristic functions of A_m and B_m by Φ_m and Ψ_m, respectively. Then,

$$\begin{aligned}
\mathbb{E} \exp(it(a_m A_m + b_m B_m)) &= \mathbb{E} \exp(it a_m A_m) \mathbb{E} \exp(it b_m B_m) \\
&= \Phi_m(t a_m) \Psi_m(t b_m) \\
&= [\exp(-\frac{(t a_m)^2}{2}) + o(1)][\exp(-\frac{(t b_m)^2}{2}) + o(1)] \\
&= \exp(-\frac{t^2}{2}) + o(1),
\end{aligned}$$

i.e.,

$$\frac{\sqrt{m}}{m_1 m_2} \sum_{j=1}^{m_2} \sum_{i=1}^{m_1} \frac{u_{ij}}{\sqrt{\rho}} \xrightarrow{d} N(0,1). \tag{34}$$

Similarly, we have

$$\frac{\sqrt{m}}{m_1 m_2} \sum_{j=1}^{m_2} \sum_{i=1}^{m_1} \frac{v_{ij}}{\sqrt{\tau}} = \frac{\sqrt{m\tau_1}}{\sqrt{m_1 \tau}} \frac{1}{\sqrt{m_1}} \sum_{i=1}^{m_1} \frac{\alpha^T (X_i + \bar{X}_i)}{\sqrt{\tau_1}} - \frac{\sqrt{m\tau_2}}{\sqrt{m_2 \tau}} \frac{1}{\sqrt{m_2}} \sum_{j=1}^{m_2} \frac{\alpha^T (Y_j + \bar{Y}_j)}{\sqrt{\tau_2}} \xrightarrow{d} N(0,1).$$

Let a and b be two real numbers with $a^2+b^2 \neq 0$. Note that

$$\begin{aligned}
& \frac{\sqrt{m}}{m_1 m_2}\sum_{j=1}^{m_2}\sum_{i=1}^{m_1}(a\frac{u_{ij}}{\sqrt{\rho}} + b\frac{v_{ij}}{\sqrt{\tau}}) \\
= \;& a\left(\frac{\sqrt{m\rho_1}}{\sqrt{m_1\rho}}\frac{1}{\sqrt{m_1}}\sum_{i=1}^{m_1}\frac{X_i^T\bar{X}_i}{\sqrt{\rho_1}} - \frac{\sqrt{m\rho_2}}{\sqrt{m_2\rho}}\frac{1}{\sqrt{m_2}}\sum_{j=1}^{m_2}\frac{Y_j^T\bar{Y}_j}{\sqrt{\rho_2}}\right) \\
& +b\left(\frac{\sqrt{m\tau_1}}{\sqrt{m_1\tau}}\frac{1}{\sqrt{m_1}}\sum_{i=1}^{m_1}\frac{\alpha^T(X_i+\bar{X}_i)}{\sqrt{\tau_1}} + \frac{\sqrt{m\tau_2}}{\sqrt{m_2\tau}}\frac{1}{\sqrt{m_2}}\sum_{j=1}^{m_2}\frac{\alpha^T(Y_j+\bar{Y}_j)}{\sqrt{\tau_2}}\right) + o_p(1) \\
= \;& \left(\frac{a\sqrt{m\rho_1}}{\sqrt{m_1\rho}}\frac{1}{\sqrt{m_1}}\sum_{i=1}^{m_1}\frac{X_i^T\bar{X}_i}{\sqrt{\rho_1}} + \frac{b\sqrt{m\tau_1}}{\sqrt{m_1\tau}}\frac{1}{\sqrt{m_1}}\sum_{i=1}^{m_1}\frac{\alpha^T(X_i+\bar{X}_i)}{\sqrt{\tau_1}}\right) \\
& +\left(\frac{a\sqrt{m\rho_2}}{\sqrt{m_2\rho}}\frac{1}{\sqrt{m_2}}\sum_{j=1}^{m_2}\frac{Y_j^T\bar{Y}_j}{\sqrt{\rho_2}} - \frac{b\sqrt{m\tau_2}}{\sqrt{m_2\tau}}\frac{1}{\sqrt{m_2}}\sum_{j=1}^{m_2}\frac{\alpha^T(Y_j+\bar{Y}_j)}{\sqrt{\tau_2}}\right) + o_p(1) \\
= \;& I_1 + I_2 + o_p(1).
\end{aligned}$$

Since $\frac{\sqrt{m\rho_1}}{\sqrt{m_1\rho}}, \frac{\sqrt{m\rho_2}}{\sqrt{m_2\rho}}, \frac{\sqrt{m\tau_1}}{\sqrt{m_1\tau}}, \frac{\sqrt{m\tau_2}}{\sqrt{m_2\tau}}$ are all bounded by one, it is easy to check that I_1 and I_2 satisfy the Lyapunov condition by (7) - (10). Therefore

$$\{a^2\frac{m\rho_1}{m_1\rho} + b^2\frac{m\tau_1}{m_1\tau}\}^{-1/2} I_1 \xrightarrow{d} N(0,1)$$

and

$$\{a^2\frac{m\rho_2}{m_2\rho} + b^2\frac{m\tau_2}{m_2\tau}\}^{-1/2} I_2 \xrightarrow{d} N(0,1).$$

Since $X_i's$ are independent of $Y_i's$, it follows from the same arguments in proving (34) that

$$I_1 + I_2 \xrightarrow{d} N(0, a^2+b^2),$$

i.e., (27) holds.

To prove (28), we write

$$\begin{aligned}
& \tfrac{m}{m_1^2 m_2^2\rho}\textstyle\sum_{k=1}^{m_1}\left(\sum_{j=1}^{m_2}u_{kj}\right)^2 \\
= \;& \tfrac{m}{m_1^2 m_2^2\rho}\textstyle\sum_{k=1}^{m_1}\left(\sum_{j=1}^{m_2}(X_k^T\bar{X}_k + Y_j^T\bar{Y}_j - Y_j^T\bar{X}_k - X_k^T\bar{Y}_j)\right)^2 \qquad (35) \\
= \;& \tfrac{m}{m_1^2\rho}\textstyle\sum_{k=1}^{m_1}\left(X_k^T\bar{X}_k + \frac{1}{m_2}\sum_{j=1}^{m_2}Y_j^T\bar{Y}_j - \frac{1}{m_2}\sum_{j=1}^{m_2}Y_j^T\bar{X}_k - X_k^T\frac{1}{m_2}\sum_{j=1}^{m_2}\bar{Y}_j\right)^2.
\end{aligned}$$

Since $m\rho_1/m_1\rho \le 1$, it follows from Lemma 2 that

$$\frac{m}{m_1^2\rho}\sum_{k=1}^{m_1}(X_k^T\bar{X}_k)^2 - \frac{m\rho_1}{m_1\rho} \xrightarrow{p} 0. \tag{36}$$

By Lemma 1, we have

$$\frac{m}{m_1^2\rho}\sum_{k=1}^{m_1}\left(\frac{1}{m_2}\sum_{j=1}^{m_2}Y_j^T\bar{Y}_j\right)^2 = O_p(\frac{m\rho_2}{m_1m_2\rho}) = o_p(1). \tag{37}$$

A direct calculation shows that

$$\begin{aligned}
& \mathbb{E}\{\tfrac{1}{m_2}\textstyle\sum_{j=1}^{m_2}Y_j^T\bar{X}_k\}^2 \\
= \;& \mathbb{E}\{(\tfrac{1}{m_2}\textstyle\sum_{j=1}^{m_2}Y_j^T)\bar{X}_k\bar{X}_k^T(\tfrac{1}{m_2}\textstyle\sum_{j=1}^{m_2}Y_j)\} \\
= \;& \mathbb{E}\mathbf{tr}\{\tfrac{1}{m_2}\textstyle\sum_{j=1}^{m_2}Y_j^T\bar{X}_k\bar{X}_k^T(\tfrac{1}{m_2}\textstyle\sum_{j=1}^{m_2}Y_j)\} \\
= \;& \mathbb{E}\mathbf{tr}\{\bar{X}_k\bar{X}_k^T(\tfrac{1}{m_2}\textstyle\sum_{j=1}^{m_2}Y_j)(\tfrac{1}{m_2}\textstyle\sum_{i=1}^{m_2}Y_i^T)\} \\
= \;& \mathbf{tr}\mathbb{E}\{\bar{X}_k\bar{X}_k^T(\tfrac{1}{m_2}\textstyle\sum_{j=1}^{m_2}Y_j)(\tfrac{1}{m_2}\textstyle\sum_{i=1}^{m_2}Y_i^T)\} \\
= \;& \mathbf{tr}\{\Sigma\tfrac{1}{m_2}\bar{\Sigma}\} \\
= \;& O(\tfrac{\rho_1+\rho_2}{m_2}) \\
= \;& O(\tfrac{m_1\rho}{m_2m}) + O(\tfrac{\rho_2}{m_2}) \\
= \;& o(\tfrac{m_1\rho}{m}),
\end{aligned}$$

which implies that

$$\frac{m}{m_1^2\rho}\sum_{k=1}^{m_1}\{\frac{1}{m_2}\sum_{j=1}^{m_2}Y_j^T\bar{X}_k\}^2 = o_p(1). \tag{38}$$

Similarly we have

$$\frac{m}{m_1^2\rho}\sum_{k=1}^{m_1}\{X_k^T\frac{1}{m_2}\sum_{j=1}^{m_2}\bar{Y}_j\}^2 = o_p(1). \tag{39}$$

It follows from (36) and (38) that

$$\begin{aligned}
& |\tfrac{m}{m_1^2\rho}\textstyle\sum_{k=1}^{m_1}(X_k^T\bar{X}_k)(\tfrac{1}{m_2}\textstyle\sum_{j=1}^{m_2}Y_j^T\bar{X}_k)| \\
\le \;& \{\tfrac{m}{m_1^2\rho}\textstyle\sum_{k=1}^{m_1}(X_k^T\bar{X}_k)^2\}^{1/2}\{\tfrac{m}{m_1^2\rho}\textstyle\sum_{k=1}^{m_1}(\tfrac{1}{m_2}\textstyle\sum_{j=1}^{m_2}Y_j^T\bar{X}_k)^2\}^{1/2} \qquad (40) \\
= \;& O_p(1)o_p(1) = o_p(1).
\end{aligned}$$

Similarly we can show that

$$\begin{cases} \frac{m}{m_1^2\rho}\sum_{k=1}^{m_1}(X_k^T\bar{X}_k)(\frac{1}{m_2}\sum_{j=1}^{m_2}Y_j^T\bar{Y}_j)=o_p(1) \\ \frac{m}{m_1^2\rho}\sum_{k=1}^{m_1}(X_k^T\bar{X}_k)(X_k^T\frac{1}{m_2}\sum_{j=1}^{m_2}\bar{Y}_j^T)=o_p(1) \\ \frac{m}{m_1^2\rho}\sum_{k=1}^{m_1}(\frac{1}{m_2}\sum_{j=1}^{m_2}Y_j^T\bar{Y}_j)(\frac{1}{m_2}\sum_{i=1}^{m_2}Y_i^T\bar{X}_k)=o_p(1) \\ \frac{m}{m_1^2\rho}\sum_{k=1}^{m_1}(\frac{1}{m_2}\sum_{j=1}^{m_2}Y_j^T\bar{Y}_j)(X_k^T\frac{1}{m_2}\sum_{i=1}^{m_2}\bar{Y}_i)=o_p(1) \\ \frac{m}{m_1^2\rho}\sum_{k=1}^{m_1}(\frac{1}{m_2}\sum_{j=1}^{m_2}Y_j^T\bar{X}_k)(X_k^T\frac{1}{m_2}\sum_{i=1}^{m_2}Y_i)=o_p(1). \end{cases} \tag{41}$$

Hence (28) follows from (35)–(41). The rest can be shown in the same way as proving (28). □

Lemma 4. *Under conditions of Theorem 1, we have as* $\min\{n_1,n_2\}\to\infty$

$$\frac{1}{\sqrt{m}}\sum_{k=1}^{m}\begin{pmatrix}\frac{Z_{k,1}}{\sqrt{\rho}}\\ \frac{Z_{k,2}}{\sqrt{\tau}}\end{pmatrix}\xrightarrow{d}N(0,I_2), \tag{42}$$

$$\frac{1}{m\rho}\sum_{k=1}^{m}Z_{k,1}^2-1\xrightarrow{p}0, \tag{43}$$

$$\frac{1}{m\tau}\sum_{k=1}^{m}Z_{k,2}^2-1\xrightarrow{p}0, \tag{44}$$

$$\frac{1}{m\sqrt{\rho\tau}}\sum_{k=1}^{m}Z_{k,1}Z_{k,2}\xrightarrow{p}0. \tag{45}$$

Moreover, we have

$$\max_{1\le k\le m}|\frac{Z_{k,1}}{\sqrt{\rho}}|=o_p(m^{1/2})\quad \textit{and}\quad \max_{1\le k\le m}|\frac{Z_{k,2}}{\sqrt{\tau}}|=o_p(m^{1/2}). \tag{46}$$

Proof. Note that for $1\le k\le m_1$,

$$Z_{k,1}=\frac{-1}{(m_1-1)m_1}\sum_{j=1}^{m_2}\sum_{i=1}^{m_1}u_{ij}+\frac{m_1+m_2-1}{(m_1-1)m_2}\sum_{j=1}^{m_2}u_{kj},$$

$$Z_{k,2}=\frac{-1}{(m_1-1)m_1}\sum_{j=1}^{m_2}\sum_{i=1}^{m_1}v_{ij}+\frac{m_1+m_2-1}{(m_1-1)m_2}\sum_{j=1}^{m_2}v_{kj},$$

and for $m_1+1\le k\le m$,

$$Z_{k,1}=\frac{-1}{(m_2-1)m_2}\sum_{j=1}^{m_2}\sum_{i=1}^{m_1}u_{ij}+\frac{m_1+m_2-1}{(m_2-1)m_1}\sum_{i=1}^{m_1}u_{i,k-m_1},$$

$$Z_{k,2} = \frac{-1}{(m_2-1)m_2}\sum_{j=1}^{m_2}\sum_{i=1}^{m_1} u_{ij} + \frac{m_1+m_2-1}{(m_2-1)m_1}\sum_{i=1}^{m_1} v_{i,k-m_1}.$$

Thus

$$\begin{aligned}
\frac{1}{\sqrt{m}}\sum_{k=1}^{m}\frac{Z_{k,1}}{\sqrt{\rho}} &= \frac{1}{\sqrt{m}}(\frac{-1}{m_2-1} + \frac{-1}{m_1-1} + \frac{m_1+m_2-1}{(m_1-1)m_2} + \frac{m_1+m_2-1}{(m_2-1)m_1})\sum_{j=1}^{m_2}\sum_{i=1}^{m_1}\frac{u_{ij}}{\sqrt{\rho}} \\
&= \frac{\sqrt{m}}{m_1m_2}\sum_{j=1}^{m_2}\sum_{i=1}^{m_1}\frac{u_{ij}}{\sqrt{\rho}}
\end{aligned}$$

and

$$\frac{1}{\sqrt{m}}\sum_{k=1}^{m}\frac{Z_{k,2}}{\sqrt{\tau}} = \frac{\sqrt{m}}{m_1m_2}\sum_{j=1}^{m_2}\sum_{i=1}^{m_1}\frac{v_{ij}}{\sqrt{\tau}},$$

which imply (42) by using Lemma 3.

It follows from Lemma 3 that

$$\begin{aligned}
& \frac{1}{m\rho}\sum_{k=1}^{m} Z_{k,1}^2 \\
= \;& \frac{1}{m\rho}\sum_{k=1}^{m_1}\left(\frac{-1}{(m_1-1)m_1}\sum_{j=1}^{m_2}\sum_{i=1}^{m_1}u_{ij} + \frac{m_1+m_2-1}{(m_1-1)m_2}\sum_{j=1}^{m_2}u_{kj}\right)^2 \\
& +\frac{1}{m\rho}\sum_{k=1}^{m_2}\left(\frac{-1}{(m_2-1)m_2}\sum_{j=1}^{m_2}\sum_{i=1}^{m_1}u_{ij} + \frac{m_1+m_2-1}{(m_2-1)m_1}\sum_{i=1}^{m_1}u_{ik}\right)^2 \\
= \;& \{\frac{1}{(m_1-1)\sqrt{m_1m\rho}}\sum_{j=1}^{m_2}\sum_{i=1}^{m_1}u_{ij}\}^2 + \frac{(m-1)^2}{(m_1-1)^2m_2^2m\rho}\sum_{k=1}^{m_1}(\sum_{j=1}^{m_2}u_{kj})^2 \\
& -2\{(\frac{m-1}{m\rho(m_1-1)^2m_1m_2})^{1/2}\sum_{j=1}^{m_2}\sum_{i=1}^{m_1}u_{ij}\}^2 + \{\frac{1}{(m_2-1)\sqrt{m_2m\rho}}\sum_{j=1}^{m_2}\sum_{i=1}^{m_1}u_{ij}\}^2 \\
& +\frac{(m-1)^2}{m\rho(m_2-1)^2m_1^2}\sum_{k=1}^{m_2}(\sum_{i=1}^{m_1}u_{ik})^2 - 2\{(\frac{m-1}{m\rho(m_2-1)^2m_2m_1})^{1/2}\sum_{j=1}^{m_2}\sum_{i=1}^{m_1}u_{ij}\}^2 \\
= \;& \{O_p(\frac{1}{m_1\sqrt{m_1m}}\frac{m_1m_2}{\sqrt{m}})\}^2 + \frac{(m-1)^2m_1^2}{(m_1-1)^2m^2}\{\frac{m\rho_1}{m_1\rho}+o_p(1)\} + \{O_p(\frac{1}{m_1\sqrt{m_1m_2}}\frac{m_1m_2}{\sqrt{m}})\}^2 \\
& +\{O_p(\frac{1}{m_2\sqrt{m_2m}}\frac{m_1m_2}{\sqrt{m}})\}^2 + \frac{(m-1)^2m_2^2}{m^2(m_2-1)^2}\{\frac{m\rho_2}{m_2\rho}+o_p(1)\} + \{O_p(\frac{1}{m_2\sqrt{m_2m_1}}\frac{m_1m_2}{\sqrt{m}})\}^2 \\
= \;& \frac{m\rho_1}{m_1\rho} + \frac{m\rho_2}{m_2\rho} + o_p(1) \\
= \;& 1 + o_p(1),
\end{aligned}$$

i.e., (43) holds. Similarly we can show (44) and (45).

Since $\mathbb{E}((\sum_{i=1}^{m_1} u_{ij})^2) = \mathrm{Var}(\sum_{i=1}^{m_1} u_{ij}) = m_1(\rho_1+\rho_2)$, we have

$$\mathbb{E}(\max_{1\le j\le m_2}(\sum_{i=1}^{m_1} u_{ij})^2) \le \sum_{j=1}^{m_2}\mathbb{E}((\sum_{i=1}^{m_1} u_{ij})^2) = m_1m_2(\rho_1+\rho_2)$$

which implies that

$$\max_{1\le j\le m_2}|\sum_{i=1}^{m_1} u_{ij}| = O_p(\sqrt{m_2m_1(\rho_1+\rho_2)}).$$

Similarly we have

$$\max_{1\le 1\le m_1}|\sum_{j=1}^{m_2} u_{ij}| = O_p(\sqrt{m_2m_1(\rho_1+\rho_2)}).$$

Hence by Lemma 3 and the expression for $Z_{k,1}$, we have

$$\begin{aligned}\max_{1\le k\le m}|\tfrac{Z_{k,1}}{\sqrt{\rho}}| &\le \tfrac{1}{(m_1-1)m_1}|\textstyle\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\tfrac{u_{ij}}{\sqrt{\rho}}| + \max_{1\le k\le m_1}|\tfrac{m-1}{(m_1-1)m_2}\sum_{j=1}^{m_2}\tfrac{u_{kj}}{\sqrt{\rho}}|\\ &\quad+\tfrac{1}{(m_2-1)m_2}|\textstyle\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\tfrac{u_{ij}}{\sqrt{\rho}}| + \max_{1\le k\le m_2}|\tfrac{m-1}{(m_2-1)m_1}\sum_{j=1}^{m_1}\tfrac{u_{jk}}{\sqrt{\rho}}|\\ &= o_p(1) + O_p(\tfrac{m-1}{(m_1-1)m_2\sqrt{\rho}}\{m_1m_2(\rho_1+\rho_2)\}^{1/2})\\ &\quad+o_p(1) + O_p(\tfrac{m-1}{(m_2-1)m_1\sqrt{\rho}}\{m_1m_2(\rho_1+\rho_2)\}^{1/2})\\ &= o_p(1) + O_p(\tfrac{m^{1/2}}{(\min(m_1,m_2))^{1/2}})\\ &= o_p(m^{1/2}).\end{aligned}$$

Similarly we can show that

$$\max_{1\le k\le m}|\frac{Z_{k,2}}{\sqrt{\tau}}| = o_p(m^{1/2}).$$

□

Proof of Theorem 1. It follows from Lemma 4 and the standard arguments in empirical likelihood method (see Owen (1990)). □

To show Corollary 1 and Theorem 2, we first prove the following lemmas.

Lemma 5. *$\mathbf{tr}(\Sigma^4) = O\big((\mathbf{tr}(\Sigma^2))^2\big)$, $\rho_1 = \sum_{j=1}^d \lambda_j^2$, and $2||\alpha||^2\lambda_1 \le \tau_1 \le 2||\alpha||^2\lambda_d$.*

Proof. Since $\mathbf{tr}(\Sigma^j) = \sum_{i=1}^d \lambda_i^j$ for any positive integer j, the first equality follows immediately. The second equality follows since $\rho_1 = \mathbf{tr}(\Sigma^2)$. The third pair of inequalities on τ_1 come from the definition of τ_1. □

Lemma 6. *For any $\delta > 0$*

$$\mathbb{E}|X_1^T \bar{X}_1|^{2+\delta} \leq d^\delta \left(\sum_{i=1}^d \mathbb{E}|X_{1,i}|^{2+\delta} \right)^2$$

and

$$\mathbb{E}|\alpha^T (X_1 + \bar{X}_1)|^{2+\delta} \leq 2^{4+\delta} ||\alpha||^{2+\delta} d^{\delta/2} \sum_{i=1}^d \mathbb{E}|X_{1,i}|^{2+\delta}.$$

Proof. It follows from the Cauchy-Schwarz inequality that

$$|X_1^T \bar{X}_1|^2 \leq ||X_1||^2 ||\bar{X}_1||^2.$$

Then by using the C_r inequality we conclude that

$$\begin{aligned}
\mathbb{E}|X_1^T \bar{X}_1|^{2+\delta} &\leq \mathbb{E}\left(\sum_{i=1}^d X_{1,i}^2 \right)^{(2+\delta)/2} \mathbb{E}\left(\sum_{i=1}^d \bar{X}_{1,i}^2 \right)^{(2+\delta)/2} \\
&= \left(\mathbb{E}\left(\sum_{i=1}^d X_{1,i}^2 \right)^{(2+\delta)/2} \right)^2 \\
&\leq \left(d^{\delta/2} \sum_{i=1}^d \mathbb{E}|X_{1,i}|^{2+\delta} \right)^2 \\
&= d^\delta \left(\sum_{i=1}^d \mathbb{E}|X_{1,i}|^{2+\delta} \right)^2 .
\end{aligned}$$

Similarly, from the C_r inequality we have

$$\begin{aligned}
\mathbb{E}|\alpha^T (X_1 + \bar{X}_1)|^{2+\delta} &\leq 2^{4+\delta} \mathbb{E}|\alpha^T X_1|^{2+\delta} \\
&\leq 2^{4+\delta} ||\alpha||^{2+\delta} \mathbb{E}(||X_1||^{2+\delta}) \\
&= 2^{4+\delta} ||\alpha||^{2+\delta} \mathbb{E}\left| \sum_{i=1}^d |X_{1,i}|^2 \right|^{1+\delta/2} \\
&\leq 2^{4+\delta} ||\alpha||^{2+\delta} d^{\delta/2} \sum_{i=1}^d \mathbb{E}|X_{1,i}|^{2+\delta}.
\end{aligned}$$

This completes the proof. □

Proof of Corollary 1. Equations (7) and (9) follow from conditions **A1**–**A3** by using Lemmas 5 and 6. So do equations (8) and (10), since we have the same assumptions on $\{X_i\}$ and $\{Y_j\}$. □

Proof of Theorem 2. If suffices to verify conditions (7) and (9) with $\delta = 2$ in Theorem 1. Recall we assume that $\mu_1 = \mu_2 = 0$. Note that $\mathrm{Var}(X_1) = \Sigma = \Gamma_1 \Gamma_1^T$. Denote $\alpha^T \Gamma_1 = (a_1, \cdots, a_k)$ and $\Sigma' = \Gamma_1^T \Gamma_1 = (\sigma'_{j,l})_{1 \le j,l \le k}$. Then

$$X_1^T \bar{X}_1 = \sum_{j=1}^{k} \sum_{l=1}^{k} \sigma'_{j,l} B_{1,j} B_{1+m_1,l},$$

and

$$\alpha^T (X_1 + \bar{X}_1) = \sum_{j=1}^{k} a_j (B_{1,j} + B_{1+m_1,j}).$$

Set $\delta_{j_1,j_2,j_3,j_4} = \mathbb{E}(B_{1,j_1} B_{1,j_2} B_{1,j_3} B_{1,j_4})$. Then δ_{j_1,j_2,j_3,j_4} equals $3 + \xi_1$ if $j_1 = j_2 = j_3 = j_4$, equals 1 if j_1, j_2, j_3 and j_4 form two different pairs of integers, and is zero otherwise. By Lemma 5, we have

$$\begin{aligned}
\mathbb{E}(X_1^T \bar{X}_1)^4 &= \sum_{j_1,j_2,j_3,j_4=1}^{k} \sum_{l_1,l_2,l_3,l_4=1}^{k} \sigma'_{j_1,l_1} \sigma'_{j_2,l_2} \sigma'_{j_3,l_3} \sigma'_{j_4,l_4} \delta_{j_1,j_2,j_3,j_4} \delta_{l_1,l_2,l_3,l_4} \\
&= O\left(|\textstyle\sum_{j_1 \neq j_2} \sum_{l_1 \neq l_2} \sigma'_{j_1,l_1} \sigma'_{j_1,l_2} \sigma'_{j_2,l_1} \sigma'_{j_2,l_2}| \right) + O\left(\textstyle\sum_{j_1 \neq j_2} \sum_{l=1}^{k} \sigma'^2_{j_1,l} \sigma'^2_{j_2,l} \right) \\
&\quad + O\left(\textstyle\sum_{j=1}^{k} \sum_{l_1 \neq l_2} \sigma'^2_{j,l_1} \sigma'^2_{j,l_2} \right) + O\left(\textstyle\sum_{j=1}^{k} \sum_{l=1}^{k} \sigma'^4_{j,l} \right) \\
&= O\left(|\textstyle\sum_{j_1=1}^{k} \sum_{j_2=1}^{k} \sum_{l_1=1}^{k} \sum_{l_2=1}^{k} \sigma'_{j_1,l_1} \sigma'_{j_1,l_2} \sigma'_{j_2,l_1} \sigma'_{j_2,l_2}| \right) \\
&\quad + O\left(\textstyle\sum_{j_1=1}^{k} \sum_{j_2=1}^{k} \sum_{l=1}^{k} \sigma'^2_{j_1,l} \sigma'^2_{j_2,l} \right) + O\left(\textstyle\sum_{j=1}^{k} \sum_{l=1}^{k} \sigma'^4_{j,l} \right) \\
&= O\left(\mathbf{tr}(\Sigma'^4) \right) + O\left(\textstyle\sum_{j=1}^{k} \sum_{l=1}^{k} \sigma'^2_{j,l})^2 \right) \\
&= O\left(\mathbf{tr}(\Sigma'^4) \right) + O\left((\mathbf{tr}(\Sigma'^2))^2 \right) \\
&= O\left(\mathbf{tr}(\Sigma^4) \right) + O\left((\mathbf{tr}(\Sigma^2))^2 \right) \\
&= O(\rho_1^2),
\end{aligned}$$

i.e., (7) holds with $\delta = 2$.

Similarly we have

$$\begin{aligned}
\mathbb{E}(\alpha^T(X_1+\bar{X}_1))^4 &\le 2^4\mathbb{E}\Big(\sum_{j=1}^k a_j B_{1,j}\Big)^4 \\
&= O\left(\sum_{j_1,j_2=1}^k a_{j_1}^2 a_{j_2}^2\right) + O\left(\sum_{j=1}^k a_j^4\right) \\
&= O\left(\Big(\sum_{j=1}^k a_j^2\Big)^2\right) \\
&= O\left(\left(\alpha^T\Gamma_1\Gamma_1^T\alpha\right)^2\right) \\
&= O\left(\tau_1^2\right),
\end{aligned}$$

which yields (9) with $\delta = 2$. Equations (8) and (10) can be shown in the same way. Hence Theorem 2 follows from Theorem 1. □

Acknowledgment. We thank an associate editor and two reviewers for helpful comments. Wang's research was supported by the Bob Price fellowship in the School of Mathematics at Georgia Tech. Peng's research was supported by NSF Grant DMS-1005336 and Qi's research was supported by NSF grant DMS-1005345.

References

[1] Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S. Mack, D. and Levin A.J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. U.S.A. 96, 6745–6750.*

[2] Arlot, S., Blanchard, G. and Roquain, E. (2010a). Some nonasymptotic results on resampling in high dimension I: confidence regions. *Ann. Statist. 38, 51–82.*

[3] Arlot, S., Blanchard, G. and Roquain, E. (2010b). Some nonasymptotic results on resampling in high dimension II: multiple tests. *Ann. Statist. 38, 83–99.*

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

[5] Chen, S.X., Peng, L. and Qin, Y. (2009). Empirical likelihood methods for high dimension. *Biometrika 96, 711–722.*

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

[7] Chen, S. and Van Keilegom, I. (2009). A review on empirical likelihood methods for regression. *Test 18, 415–447.*

[8] Dharmadhikari, S.W. and Jogdeo, K. (1969). Bounds on moments of certain random variables. *Ann. Math. Statist. 40, 1506-1508.*

[9] Jing, B.Y., Yuan, J.Q. and Zhou, W. (2009). Jackknife empirical likelihood. *J. Amer. Statist. Assoc. 104, 1224–1232.*

[10] Lee, M.L. (2004). Analysis of Microarray Gene Expression Data. *Kluwer.*

[11] Kuelbs, J. and Vidyashankar, A.N. (2010). Asymptotic inference for high dimensional data. *Ann. Statist. 38, 836–869.*

[12] Owen, A. (1990). Empirical likelihood ratio confidence regions. *Ann. Statist. 18, 90–120.*

[13] Owen, A. (2001). *Empirical Likelihood.* Chapman & Hall/CRC.

[14] Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *Ann. Statist. 22, 300–325.*

[15] Srivastava, M.S. (2009). A test for the mean vector with fewer observations than the dimension under non-normality. *J. Multi. Anan. 100, 518–532.*

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

[17] von Bahr, B. and Esseen, C.G. (1965). Inequality for the rth absolute moment of a sum of random variables, $1 \leq r \leq 2$. *Ann. Math. Statist. 36, 299–393.*

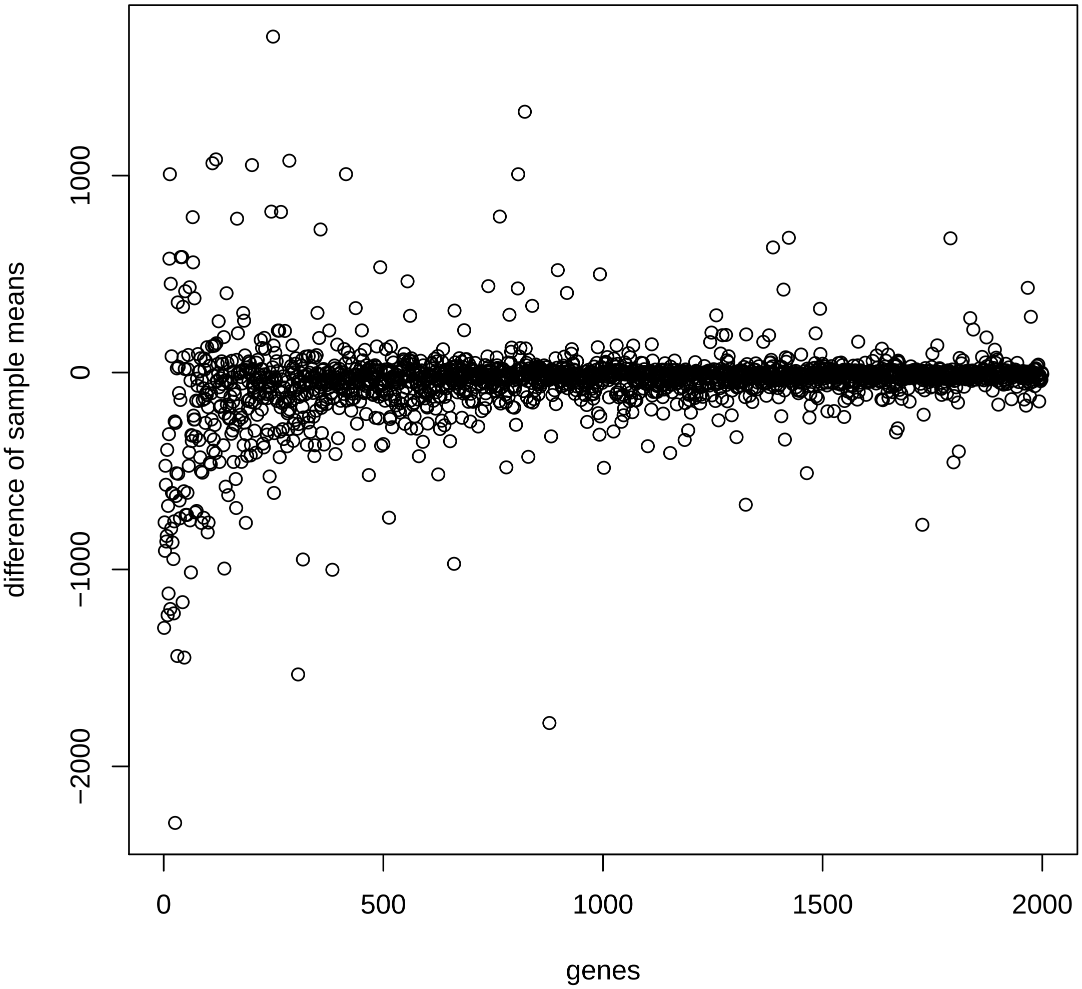

Figure 1: Colon data: differences of the sample means are plotted against each gene.

Table 1: Sizes and powers of the proposed jackknife empirical likelihood test (JEL) and the test in Chen and Qin (2010) (CQ) are reported for the case of $(n_1, n_2) = (30, 30)$ at level 5%.

d	JEL	CQ	JEL	CQ	JEL	CQ
	$c_1 = 0$	$c_1 = 0$	$c_1 = 0.1$	$c_1 = 0.1$	$c_1 = 0.1$	$c_1 = 0.1$
	$c_2 = 0.25$	$c_2 = 0.25$	$c_2 = 0.25$	$c_2 = 0.25$	$c_2 = 0.75$	$c_2 = 0.75$
10	0.070	0.049	0.071	0.049	0.072	0.062
20	0.056	0.037	0.057	0.049	0.096	0.060
30	0.064	0.047	0.066	0.049	0.113	0.066
40	0.070	0.052	0.069	0.058	0.116	0.072
50	0.067	0.049	0.083	0.054	0.138	0.067
60	0.063	0.039	0.069	0.043	0.174	0.055
70	0.053	0.053	0.076	0.065	0.190	0.081
80	0.056	0.059	0.063	0.067	0.191	0.082
90	0.056	0.044	0.080	0.054	0.204	0.071
100	0.066	0.060	0.082	0.064	0.229	0.091
300	0.056	0.045	0.114	0.054	0.537	0.092
500	0.049	0.051	0.160	0.063	0.731	0.110

Table 2: Sizes and powers of the proposed jackknife empirical likelihood test (JEL) and the test in Chen and Qin (2010) (CQ) are reported for the case of $(n_1, n_2) = (100, 100)$ at level 5%.

d	JEL	CQ	JEL	CQ	JEL	CQ
	$c_1 = 0$	$c_1 = 0$	$c_1 = 0.1$	$c_1 = 0.1$	$c_1 = 0.1$	$c_1 = 0.1$
	$c_2 = 0.25$	$c_2 = 0.25$	$c_2 = 0.25$	$c_2 = 0.25$	$c_2 = 0.75$	$c_2 = 0.75$
10	0.074	0.054	0.072	0.063	0.099	0.090
20	0.043	0.047	0.053	0.055	0.145	0.098
30	0.047	0.047	0.056	0.063	0.191	0.115
40	0.051	0.050	0.063	0.062	0.264	0.125
50	0.055	0.040	0.077	0.061	0.326	0.131
60	0.055	0.044	0.077	0.067	0.374	0.151
70	0.043	0.051	0.063	0.086	0.395	0.150
80	0.042	0.059	0.082	0.079	0.474	0.171
90	0.043	0.040	0.098	0.065	0.527	0.163
100	0.049	0.054	0.091	0.088	0.575	0.194
300	0.048	0.054	0.217	0.102	0.974	0.389
500	0.049	0.041	0.353	0.115	0.999	0.544

Table 3: Sizes and powers of the proposed jackknife empirical likelihood test (JEL) and the test in Chen and Qin (2010) (CQ) are reported for the case of $(n_1, n_2) = (150, 200)$ at level 5%.

d	JEL	CQ	JEL	CQ	JEL	CQ
	$c_1 = 0$	$c_1 = 0$	$c_1 = 0.1$	$c_1 = 0.1$	$c_1 = 0.1$	$c_1 = 0.1$
	$c_2 = 0.25$	$c_2 = 0.25$	$c_2 = 0.25$	$c_2 = 0.25$	$c_2 = 0.75$	$c_2 = 0.75$
10	0.048	0.054	0.054	0.062	0.129	0.116
20	0.055	0.042	0.078	0.075	0.237	0.166
30	0.052	0.054	0.079	0.081	0.330	0.207
40	0.039	0.035	0.070	0.068	0.430	0.212
50	0.039	0.048	0.071	0.094	0.480	0.231
60	0.047	0.051	0.092	0.095	0.598	0.273
70	0.046	0.051	0.086	0.107	0.658	0.309
80	0.042	0.047	0.113	0.109	0.753	0.327
90	0.046	0.043	0.148	0.098	0.781	0.346
100	0.048	0.059	0.141	0.117	0.821	0.365
300	0.044	0.040	0.370	0.163	1	0.703
500	0.047	0.045	0.555	0.235	1	0.899

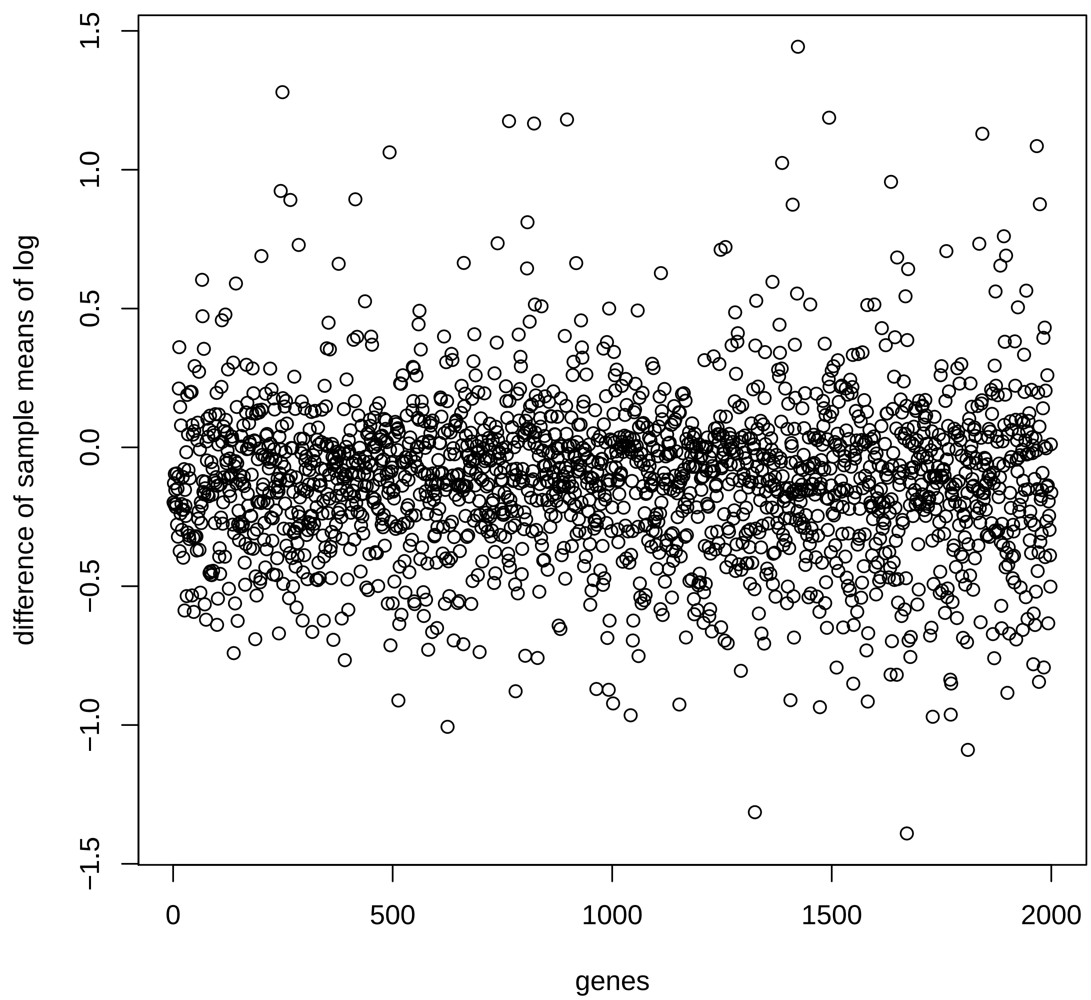

Figure 2: Colon data: differences of the sample means of logarithms of gene expression levels are plotted against each gene.

Table 4: 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	CQ	JEL
50	1158	2.94e-01	2.13e-01
100	1501	5.63e-01	2.82e-01
200	1742	7.21e-01	3.87e-01
500	1913	2.71e-02	3.75e-01
1000	1978	6.79e-05	3.40e-01
3000	2000	5.06e-09	1.36e-01