文章编号:2095-6134(2019)02-0155-07

A powerful procedure for multiple outcomes comparison with covariate adjustment and its application to genomic data*

ZHANG Shenghu[1,2], ZHU Jiayan[3], ZHANG Sanguo[1,2†]

(1 *School of Mathematical Sciences*, *University of Chinese Academy of Sciences*, *Beijing* 100049, *China*;
2 *Key Laboratory of Big Data Mining and Knowledge Management of Chinese Academy of Sciences*, *Beijing* 100049, *China*;
3 *School of Information and Communication*, *Wuhan College*, *Wuhan* 430212, *China*)

(Received 12 January 2018; Revised 9 March 2018)

Zhang S H, Zhu J Y, Zhang S G.A powerful procedure for multiple outcomes comparison with covariate adjustment and its application to genomic data[J]. Journal of University of Chinese Academy of Sciences, 2019,36(2):155-161.

Abstract Although there are many procedures developed for handling multiple outcomes comparison in the literature, the nonparametric methodology for group comparison with covariate adjustment is still in its infancy. One can use rank-sum test, adjusted rank-sum test, or max-type test by analyzing the processed data orthogonal to the space spanned by covariates. However, the power is not satisfactory. In this work, we combine the adjusted rank-sum test and pseudo F test and then construct a MIN2 test to handle this issue. The performances of MIN2 are thoroughly explored by extensive computer simulations and a real example.

Keywords multiple outcomes comparison; covariate adjustment; pseudo F test; power
CLC number:O212;Q348 **Document code**: A **doi**:10.7523/j.issn.2095-6134.2019.02.002

一个带协变量调整多响应比较的高效方法及其在基因组数据中的应用

张胜虎[1,2],朱家砚[3],张三国[1,2]

(1 中国科学院大学数学科学学院, 北京 100049; 2 中国科学院大数据挖掘与知识管理重点实验室, 北京 100049;
3 武汉学院信息系, 武汉 430212)

摘 要 目前在文献中有很多关于多响应比较的研究方法,但是对带协变量调整的非参数检验的研究较少。一种直观的想法是将数据先投影到协变量的正交空间中,然后再利用秩和检验、调整的秩和检验或最大值检验方法。但是,功效普遍不高。在调整的秩和检验和伪 F 检验两种方法基础上,构建 MIN2 检验。大量模拟和实际数据表明,MIN2 检验的效果优于现有的非参数检验方法。

关键词 多响应比较;协变量调整;伪 F 检验;功效

* Supported by Special Fund of University of Chinese Academy of Sciences for Scientific Research Cooperation(Y652022Y00)

† Corresponding author, E-mail: sgzhang@ucas.ac.cn

Multiple outcomes comparison is frequently encountered in many research areas. For example, in a plasma-renin activity clinical trail[1-2], investigators aimed to see whether the drug fenoterol increases or reduces plasma-renin activity, five endpoints described by five occasions (after 0, 2, 6, 8, and 12 h) were measured. In genetics, in order to see whether the genetic variants increase the risk of disease occurrence, investigators often collect two groups of individuals with the case group suffering from disease and the control group being healthy, and many outcomes described by genetic variants are genotyped on them. In genomics, to investigate the age-dependent regulation of gene expression in human brain, RNAs harvested from postmortem samples of 30 individuals were analyzed using Affymetrix gene chips and the aim was to see whether the gene expression patterns varied among two groups categorized by age of 73 with adjusting for gender[3].

Many procedures including parametric and nonparametric ones have been developed in the literature. A classic method is the Hotelling's T^2 test (HT)[4], which is the optimal invariant test when data follow multinormal distribution with homoscedasticity. If the normal assumption is violated, the non-parametric methods are rank-sum test (RST)[5], adjusted rank-sum test (ARST)[6], and rank-maximum test (MAX)[7]. The above tests were derived without adjusting for covariates. However, in a real application such as the ageing human brain data[3] analyzed later, the investigators want to see the differences of multiple patterns between cases and controls after adjusting for gender. At this point, covariate adjustment is essential, and it reduces the bias and improves the precision of the comparison (see Refs. [8-12]).

In this work, we combine a version of ARST and the pseudo F test, which was developed to handle the ecologic data in Ref. [13] and can be thought as a non-parametric version of multivariate regression model[14-17], and propose a MIN2 test.

1 The MIN2 test

Consider two groups, group 1 and group 2. Suppose that there are n_1 and n_2 subjects sampled from the two groups, respectively, and k ($k>1$) outcomes are measured on each individual. Let $n=n_1+n_2$, $\boldsymbol{Y}_1$ and $\boldsymbol{X}_1$ be the response and covariate matrices for group 1, with dimension of $n_1\times k$ and $n_1\times d$, respectively, and $\boldsymbol{Y}_2$ and $\boldsymbol{X}_2$ be the response and covariate matrices for group 2, with dimension of $n_2\times k$ and $n_2\times d$, respectively. Define $\boldsymbol{Y}=\begin{pmatrix}\boldsymbol{Y}_1\\ \boldsymbol{Y}_2\end{pmatrix}$ and $\boldsymbol{X}=\begin{pmatrix}\boldsymbol{X}_1\\ \boldsymbol{X}_2\end{pmatrix}$. The problem of interest is the extent to which the differences between the two groups are maintained after covariate adjustment. Therefore, the null hypothesis can be expressed as follows:

H_0: There is no difference between the two groups after covariate adjustment.

When considering the covariates, the HT, RST, ARST, and MAX can not be applied directly. So we project $\boldsymbol{Y}$ orthogonal to the space spanned by covariates $\boldsymbol{X}$ to get the residual matrix $\boldsymbol{E}$, that is,

$$\boldsymbol{E}=(\boldsymbol{I}_n-\boldsymbol{H}_X)\boldsymbol{Y}=\begin{pmatrix}\boldsymbol{E}_1\\ \boldsymbol{E}_2\end{pmatrix},$$

where $\boldsymbol{I}_n$ is n-dimensional identity matrix, $\boldsymbol{H}_X=\boldsymbol{X}(\boldsymbol{X}^\tau\boldsymbol{X})^{-1}\boldsymbol{X}^\tau$, $\boldsymbol{E}_1$ and $\boldsymbol{E}_2$ are matrices with dimension of $n_1\times k$ and $n_2\times k$, respectively. Denote $\boldsymbol{E}=(e_{uv})$, $u=1,2,\cdots,n$, and $v=1,2,\cdots,k$. The marginal distributions corresponding to e_{1v} and e_{nv} are F_v and G_v, respectively. At this point, the tests such as HT, RST, ARST, and MAX mentioned above can be obtained based on $\boldsymbol{E}$.

The ARST was proposed in Ref. [6] to accommodate the null hypothesis

$\widetilde{H}_0:\theta_v=\Pr(e_{1v}<e_{nv})-\Pr(e_{1v}>e_{nv})=0, v=1,\cdots,k$. For the v_{th} outcomes, let R_{1iv} and R_{2jv} be the mid-ranks of e_{iv} and e_{jv}, $i=1,2,\cdots,n_1$, $j=n_1+1,n_1+2,\cdots,n$. Define $R_{1i}=\sum_{v=1}^{k}R_{1iv}$, $R_{2j}=\sum_{v=1}^{k}R_{2jv}$, $\bar{R}_1=\frac{1}{n_1}\sum_{i=1}^{n_1}R_{1i}$, $\bar{R}_2=\frac{1}{n_2}\sum_{j=n_1+1}^{n}R_{2j}$, $\hat{\sigma}_1^2=\frac{1}{n_1}\sum_{i=1}^{n_1}(R_{1i}-\bar{R}_1)^2$, $\hat{\sigma}_2^2=\frac{1}{n_2}\sum_{j=n_1+1}^{n}(R_{2j}-\bar{R}_2)^2$, and $\hat{\sigma}^2=\frac{1}{n-2}((n_1-1)\hat{\sigma}_1^2+(n_2-1)\hat{\sigma}_2^2)$. Then the ARST can be written as

$$T_h = \frac{\bar{R}_2 - \bar{R}_1}{\hat{\sigma}\sqrt{\hat{h}(1/n_1 + 1/n_2)}}, \quad (1)$$

where $\hat{h}$ is a consistent estimate (see Ref.[6]) of

$$h = \frac{\sum_{u=1}^{k}\sum_{v=1}^{k}(1+\lambda)^2(a_{uv} + b_{uv}\lambda)}{\sum_{u=1}^{k}\sum_{v=1}^{k}[e_{uv}\lambda^3 + (b_{uv} + 2f_{uv})\lambda^2 + (a_{uv} + 2q_{uv})\lambda + p_{uv}]},$$

where

$a_{uv} = \mathrm{cov}(G_u(e_{1u}), G_v(e_{1v}))$, $b_{uv} = \mathrm{cov}(F_u(e_{nu}), F_v(e_{nv}))$, $e_{uv} = \mathrm{cov}(F_u(e_{1u}), F_v(e_{1v}))$, $f_{uv} = \mathrm{cov}(F_u(e_{1u}), G_v(e_{nv}))$, $p_{uv} = \mathrm{cov}(G_u(e_{nu}), G_v(e_{nv}))$, $q_{uv} = \mathrm{cov}(G_u(e_{nu}), F_v(e_{1v}))$, and $\lambda = n_1/n_2$.

The ARST maintains good power in the alternative parameter space when θ_vs lie in the same direction. When θ_vs lie in different directions or the magnitudes of some of θ_vs are large, the test may suffer from substantial loss of power. So, a MAX test was proposed in Ref. [7] to address this issue. However its power is not optimistic when most of the endpoints provide evidences and these evidences are not so strong.

When θ_vs lie in different directions or the magnitudes of some of θ_vs are large, the Kendall τ distance is applied in identifying this difference very well. The Kendall τ distance between two groups of observations is defined as the total number of discordant pairs. The larger the distance, the more dissimilar both groups are. Let $\boldsymbol{D} = (d_{lm})_{n\times n}$ be the Kendall τ distance matrix based on $\boldsymbol{Y}$ and $S = (s_{lm})_{n\times n}$ with $s_{lm} = -\frac{1}{2}d_{lm}^2, l, m = 1, 2, \cdots, n$. Denote $\boldsymbol{G} = (G_1, G_2, \cdots, G_n)^{\mathrm{T}}$, which is the group status column vector, with $G_i = 1$ for group 1 and $G_j = 0$ for group 2, $i = 1, 2, \cdots, n_1, j = n_1 + 1, n_2 + 2, \cdots, n$. Let $\boldsymbol{Z} = (\boldsymbol{X}, \boldsymbol{G})$ be the design matrix, $\boldsymbol{H}_X = \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}$, $\boldsymbol{H}_Z = \boldsymbol{Z}(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z})^{-1}\boldsymbol{Z}^{\mathrm{T}}$, and $\boldsymbol{C} = \boldsymbol{I}_n - n^{-1}\boldsymbol{J}\boldsymbol{J}^{\mathrm{T}}$ be the centering matrix, where $\boldsymbol{I}_n$ and $\boldsymbol{J}$ are the $n \times n$ identity matrix and the n-dimensional column vector of 1_s, respectively. The pseudo F statistic [13] based on Kendall τ distance can be expressed as

$$T_F = \frac{\mathrm{tr}[(\boldsymbol{H}_Z - \boldsymbol{H}_X)\boldsymbol{C}\boldsymbol{S}\boldsymbol{C}]}{\mathrm{tr}[\boldsymbol{I}_n - \boldsymbol{H}_Z)\boldsymbol{C}\boldsymbol{S}\boldsymbol{C}]}. \quad (2)$$

Let p_1 be the p-value of T_h and p_2 be the p-value of T_F, where p_1 can be obtained by the normal distribution and p_2 is obtained by permutation procedure [14]. We propose an MIN2 as

$$\mathrm{MIN2} = \min(p_1, p_2). \quad (3)$$

The MIN2 test integrates the superiorities of T_h and T_F and is thus more robust than T_h and T_F. However, the asymptotical distribution of MIN2 is not known. We recommend to use the permutation procedure to get the p-value of MIN2:

1) set a large number B, for example $B = 1\,000$, and calculate the MIN2 statictic using the observations, denote it by $\eta^{(0)}$;

2) for b from 1 to B, randomly permute n observations and arrange the first n_1 samples to group 1 and other n_2 samples to group 2, and calculate the MIN2 statictic, denote it by $\eta^{(b)}$;

3) the p-value of the MIN2 statictic is calculated as $p\text{-value} = \frac{\#\{\eta^{(b)} \leqslant \eta^{(0)}: b = 1, 2, \cdots, B\}}{B}$.

2 Simulation studies

We conduct simulation studies to evaluate the performance of the proposed MIN2 with HT, RST, ARST, and MAX. The empirical type I error rates and powers are simulated using data from two distributions: Log-normal and Laplace distributions. To study the influence of small sample size, we consider $n_1 = n_2 \in \{20, 25, 30, 35, 40, 45, 50\}$. Assume

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\gamma} + \boldsymbol{G}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad (4)$$

where $\boldsymbol{X} \sim N(0, \Sigma)$, $\Sigma = (\sigma_{ij})$ with $\sigma_{ii} = 1$ and $\sigma_{ij} = 0.2 (i \neq j)$ for $i, j \in \{1, 2, 3, 4\}$, $\boldsymbol{G}$ is a column of the group status indicator, $\boldsymbol{\gamma}$ is the matrix with the element being 1, $\boldsymbol{\epsilon}$ follows two distributions: (i) multivariate Log-normal with logs having mean vector 0 and covariance matrix $\boldsymbol{\Delta}$; and (ii) Laplace distribution with mean vector 0 and covariance matrix $\boldsymbol{\Delta}$. $\boldsymbol{\Delta}$ is a 10-dimensional positive definite matrix with $\Delta_{ii} = 1$ for $i \in \{1, 2, \cdots, 10\}$ and $\Delta_{ij} = \rho = 0.3 (0.7)$ for $i \neq j \in \{1, 2, \cdots, 10\}$. Then the null hypothesis testing on $\theta_v = 0$ can be transformed as $\beta_v = 0, v = 1, 2, \cdots, 10$.

To evaluate the type I error rate, we set $\beta_v = 0$,

$v = 1, 2, \cdots, 10.\ 1\ 000$ replicates are conducted and the nominal significance level is set to be 0.05. The results for Log-normal and Laplace distributions are summarized in Table 1 and Table 2, respectively. In Table 1, it is seen that the HT is a little bit conservative with the empirical type I error rates being less than 0.05 and the MAX is optimistic when the sample size is small. The other three tests maintain good type I error rates, which are close to 0.05. Similar phenomena are observed in Table 2. For example, when the data are Log-normal distributed with the sample size of 40, the empirical type I error rates of HT, RST, ARST, MAX, and MIN2 are 0.038, 0.048, 0.047, 0.069, and 0.046, respectively, as $\rho = 0.3$ and 0.034, 0.045, 0.045, 0.065, and 0.048, respectively, as $\rho = 0.7$. When the data are generated from the Laplace distribution with the sample size of 40, the empirical type I error rates of HT, RST, ARST, MAX, and MIN2 are 0.054, 0.046, 0.047, 0.072, and 0.046, respectively, as $\rho = 0.3$ and 0.054, 0.049, 0.050, 0.066, and 0.052, respectively, as $\rho = 0.7$.

Table 1 The empircial type I error rates of HT, RST, ARST, MAX, and MIN2 when the data are generated from ten-dimensional Log-normal distribution

ρ	$n_1 = n_2$	HT	RST	ARST	MAX	MIN2
0.3	20	0.035	0.048	0.048	0.102	0.049
	25	0.037	0.051	0.052	0.087	0.058
	30	0.034	0.054	0.054	0.081	0.054
	35	0.033	0.047	0.046	0.075	0.057
	40	0.038	0.048	0.047	0.069	0.046
	45	0.034	0.046	0.046	0.060	0.047
	50	0.037	0.052	0.054	0.058	0.057
0.7	20	0.031	0.043	0.045	0.087	0.044
	25	0.038	0.050	0.053	0.080	0.055
	30	0.035	0.045	0.047	0.076	0.043
	35	0.034	0.058	0.059	0.068	0.053
	40	0.034	0.045	0.045	0.065	0.048
	45	0.037	0.051	0.055	0.063	0.049
	50	0.037	0.052	0.054	0.058	0.057

Note: The nominal significance level is 0.05 and 1 000 replicates are conducted.

To make power comparison, two types of alternatives are considered:

(a) $\boldsymbol{\beta} = (0.3, 0.3, 0.3, 0.3, 0.3, -0.3, -0.3, -0.3, -0.3, -0.3)^{\mathrm{T}}$;

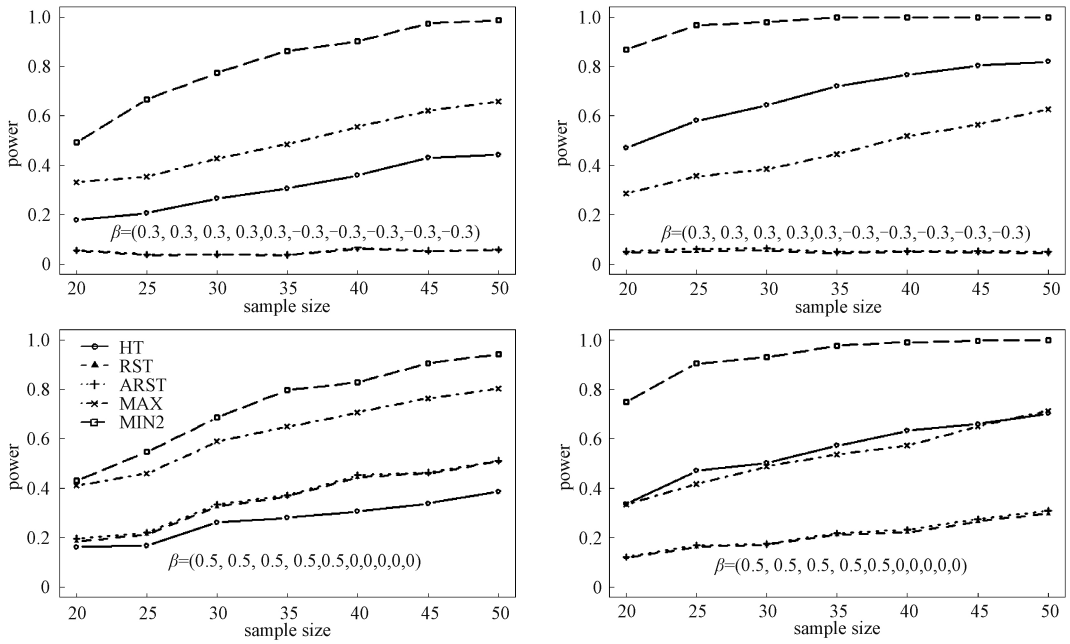
(b) $\boldsymbol{\beta} = (0.5, 0.5, 0.5, 0.5, 0.5, 0, 0, 0, 0, 0)^{\mathrm{T}}$.

The results for Log-normal and Laplace distributions are displayed in Figs. 1 and 2, respectively, with $\rho = 0.3$ on the left panels and $\rho = 0.7$ on the right panels. For scenario (a), we can see that the RST and ARST have smallest powers, which are close to 0.05. It is reasonable since the differences among the outcomes are counteracted. With the increase in the sample size, the powers of HT, MAX, and MIN2 increase. The MIN2 is the most powerful among all the tests. Sometimes, the power increase for MIN2 reaches more than 30% compared with other tests. For example, for Log-normal distribution, when $n_1 = n_2 = 30$ and $\rho = 0.3$, the empirical powers of HT, RST, ARST, MAX, and MIN2 are 0.266, 0.038, 0.040, 0.427, and 0.775, respectively. When $n_1 = n_2 = 30$ and $\rho = 0.7$, the empirical powers of HT, RST, ARST, MAX, and MIN2 are 0.643, 0.055, 0.064, 0.384, and 0.980, respectively. For Laplace distribution, the power increase for MIN2 is sometimes more than 10% over other tests. For example, for scenario (b) when $n_1 = n_2 = 25$ and $\rho = 0.3$, the empirical powers of HT, RST, ARST, MAX, and MIN2 are 0.439, 0.144, 0.151, 0.498, and 0.72, respectively. When $n_1 = n_2 = 25$ and $\rho = 0.7$, the empirical powers of HT, RST, ARST, MAX, and MIN2 are 0.895, 0.182, 0.190, 0.538, and 0.997, respectively.

Table 2 The empircial type I error rates of HT, RST, ARST, MAX, and MIN2 when the data are generated from ten-dimensional Laplace distribution

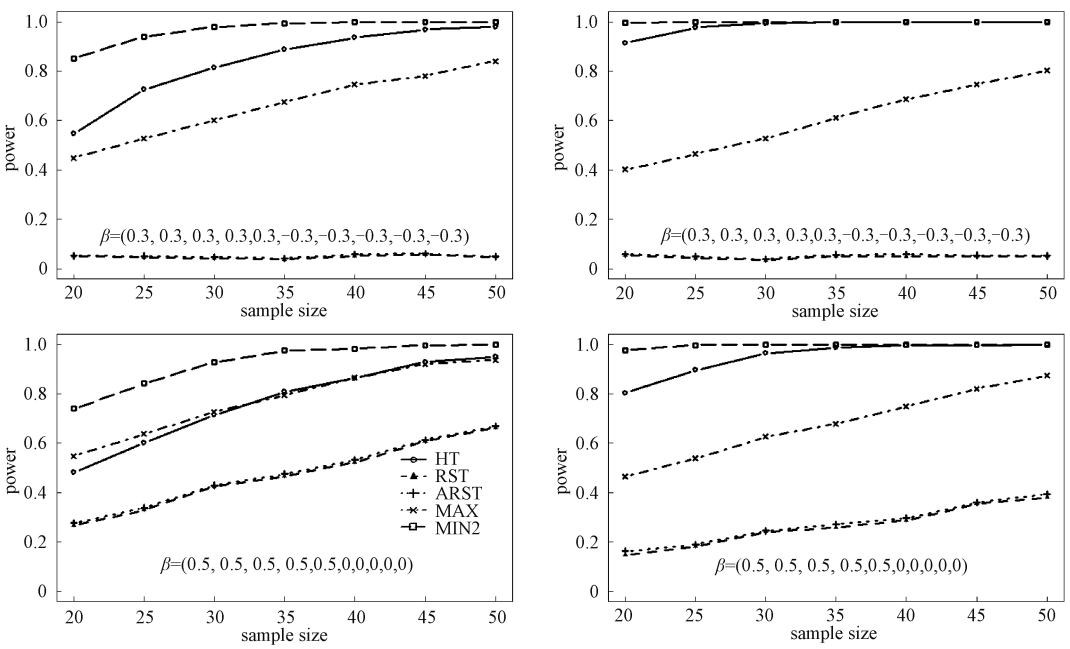
ρ	$n_1 = n_2$	HT	RST	ARST	MAX	MIN2
0.3	20	0.049	0.048	0.050	0.108	0.056
	25	0.046	0.039	0.041	0.086	0.046
	30	0.045	0.049	0.050	0.082	0.053
	35	0.046	0.052	0.055	0.076	0.055
	40	0.054	0.046	0.047	0.072	0.046
	45	0.042	0.045	0.047	0.067	0.050
	50	0.046	0.058	0.058	0.062	0.043
0.7	20	0.038	0.054	0.059	0.083	0.045
	25	0.046	0.040	0.043	0.059	0.046
	30	0.045	0.047	0.051	0.067	0.048
	35	0.046	0.058	0.058	0.065	0.055
	40	0.054	0.049	0.050	0.066	0.052
	45	0.042	0.057	0.061	0.066	0.054
	50	0.046	0.045	0.047	0.064	0.048

Note: The nominal significance level is 0.05 and 1 000 replicates are conducted.



Left: $\rho = 0.3$, right: $\rho = 0.7$; the nominal significance level is 0.05 and 1 000 replicates are conducted.

Fig.1　The power values of HT, RST, ARST, MAX, and MIN2 when the data are generated from the ten-dimensional Log-normal distribution



Left: $\rho = 0.3$, right: $\rho = 0.7$; the nominal significance level is 0.05 and 1 000 replicates are conducted.

Fig.2　The empirical power values of HT, RST, ARST, MAX, and MIN2 when the data are generated from the ten-dimensional Laplace distribution

Next, we consider the influences of large sample size and dimension on the efficacy of our MIN2. The simulation data are generated from (4), where ϵ follows Laplace distribution with mean vector 0 and covariance matrix (Δ_{ij}), where $\Delta_{ii} = 1$ for $i \in \{1,2,\cdots,k\}$ and $\Delta_{ij} = \rho = 0.3(0.7)$ for $i \neq j \in \{1,2,\cdots,k\}$. Here we consider two types of alternative hypotheses:

(c) $(\beta_1 = \cdots = \beta_{k/2} = 0.05, \beta_{k/2+1} = \cdots = \beta_k = -0.05)^{\mathrm{T}}$;

(d) $(\beta_1 = \cdots = \beta_{k/2} = 0.1, \beta_{k/2+1} = \cdots = \beta_k = 0)^{\mathrm{T}}$.

The results are shown in Table 3. It can be seen

that the power of MIN2 increases with the sample size when the dimension is fixed. For example, when $k = 20$ and $\rho = 0.3$ under scenario (c), the empirical powers of HT, RST, ARST, MAX, and MIN2 are 0.137, 0.055, 0.054, 0.120, and 0.262 for $n_1 = n_2 = 100$ and 0.270, 0.052, 0.052, 0.151, and 0.609 for $n_1 = n_2 = 200$, respectively. Similarly, the power of MIN2 rises drastically with the increment of the dimension relative to other tests when the sample size is fixed. For example, when $n_1 = n_2 = 100$ and $\rho = 0.7$ under scenario (d), the powers of all the tests are 0.234, 0.088, 0.091, 0.148, and 0.480 for $k = 10$, 0.344, 0.093, 0.092, 0.153, and 0.657 for $k = 20$, and 0.434, 0.105, 0.106, 0.155, and 0.808 for $k = 30$, respectively. So, the performance of our MIN2 is superior to the other tests in the two cases when the sample size or dimension becomes larger.

Table 3 The empircial power results of HT, RST, ARST, MAX, and MIN2 when the data are generated from the k-dimensional Laplace distribution with large sample size

$n_1 = n_2$	k	$\rho = 0.3$					$\rho = 0.7$				
		HT	RST	ARST	MAX	MIN2	HT	RHT	ARST	MAX	MIN2
Power ($\beta_1 = \cdots = \beta_{k/2} = 0.05, \beta_{k/2+1} = \cdots = \beta_k = -0.05$)											
100	10	0.119	0.048	0.047	0.110	0.186	0.230	0.053	0.055	0.108	0.479
	20	0.137	0.055	0.054	0.120	0.262	0.333	0.042	0.042	0.113	0.675
	30	0.163	0.058	0.057	0.124	0.348	0.441	0.055	0.055	0.114	0.789
200	10	0.199	0.046	0.047	0.134	0.396	0.454	0.052	0.051	0.141	0.816
	20	0.270	0.052	0.052	0.151	0.609	0.678	0.053	0.053	0.145	0.983
	30	0.381	0.044	0.044	0.157	0.793	0.811	0.055	0.055	0.156	0.996
Power ($\beta_1 = \cdots = \beta_{k/2} = 0.1, \beta_{k/2+1} = \cdots = \beta_k = 0$)											
100	10	0.116	0.123	0.124	0.162	0.231	0.234	0.088	0.091	0.148	0.480
	20	0.156	0.131	0.134	0.197	0.305	0.344	0.093	0.092	0.153	0.657
	30	0.166	0.148	0.148	0.212	0.360	0.434	0.105	0.106	0.155	0.808
200	10	0.224	0.184	0.184	0.249	0.426	0.441	0.099	0.101	0.221	0.809
	20	0.309	0.188	0.190	0.279	0.613	0.691	0.104	0.104	0.241	0.979
	30	0.391	0.205	0.205	0.308	0.803	0.810	0.123	0.124	0.267	0.996

Note: The nominal significance level is 0.05 and 1 000 replicates are conducted.

3 Application

We apply the HT, RST, ARST, MAX, and MIN2 to the data on the aging human brain [3]. A total of 30 samples are divided into two groups on account of age. The ages of subjects in group 1 are less than 73 and those in group 2 are larger than 73, where the threshold of 73 is suggested by the authors of Ref.[3]. The aim is to investigate the difference between the two groups with gender as a covariate. Six gene chips (accession number), CYP11B1, CYP11B2, D26561, CEACAM7, ESRRB, and MMP15, are treated as multiple endpoints. Figure 3 shows the box-plots of six gene chips after removing the effect of gender. It is most likely that there exists difference between the two groups. Furthermore, the average values of three gene chips, CY11B1, D26561, and CEACAM7, of group 1 are less than those of group 2, but for the other gene chips, CYP11B2, ESRRB and MMP15, the results are contrary. We carry out the HT, RST, ARST, MAX, and MIN2 to detect the difference between the two groups and the p-values of the above five tests are 0.409, 0.782, 0.794, 0.243, and 0.024, respectively. Evidently, except for MIN2, the other tests fail to detect the difference between the two groups after removing effect of gender at the nominal level of 0.05.

4 Conclusion

Studies involving multiple outcomes are fairly common in many research areas. Many procedures including parametric and nonparametric ones have been developed in the literature without considering covariates. Actually in applications, the auxiliary covariates may often be recorded on each subject. If some covariates are associated with outcomes, the precision may be improved by adjusting for this

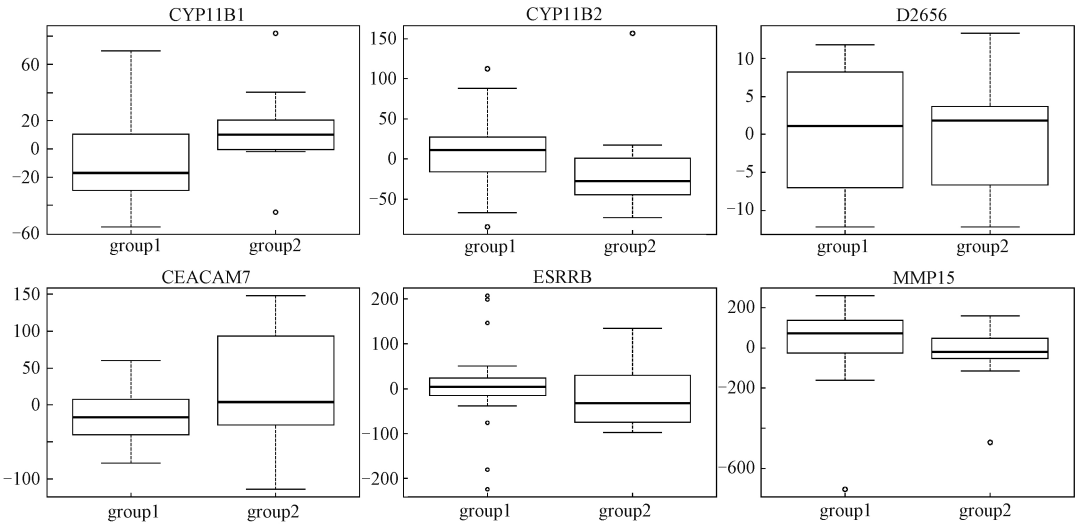


Fig.3　The box-plots of six gene chips after covariate adjustment in the aging human brain

relationship. In this work, we propose an MIN2 test to compare the difference between two groups for multiple outcomes with covariate adjustment. Through the simulation studies, the MIN2 test controls the type I error rate very well and has superior power to other existing nonparametric tests.

References

[1] Brunner E, Domhof S and Langer F. Nonparametric analysis of longitudinal data in factorial Experiments[M]. New York: Wiley, 2002.

[2] Li Z, Cao F, Zhang J, et al. Summation of absolute value test for multiple outcome comparison with moderate effect[J]. Journal of Systems Science and Complexity, 2013, 26(3): 462-469.

[3] Lu T, Pan Y, Kao S Y, et al. Gene regulation and DNA damage in the ageing human brain[J]. Nature, 2004, 429(6 994): 883-891.

[4] Hotelling H. The generalization of Student's ratio[J]. The Annals of Mathematical Statistics, 1931, 2(3): 360-378.

[5] O'Brien P C. Procedures for comparing samples with multiple endpoints[J]. Biometrics, 1984, 40(4): 1 079-1 087.

[6] Huang P, Tilley B C, Woolson R F, et al. Adjusting O'Brien's test to control type I error for the generalized nonparametric Behrens-Fisher problem[J]. Biometrics, 2005, 61(2): 532-539.

[7] Liu A, Li Q, Liu C, et al. A rank-based test for comparison of multidimensional outcomes[J]. Journal of the American Statistical Association, 2010, 105(490): 578-587.

[8] Grouin J M, Day S, Lewis J. Adjustment for baseline covariates: an introductory note[J]. Statistics in medicine, 2004, 23(5): 697-699.

[9] Koch G G, Tangen C M, Jung J W, et al. Issues for covariance analysis of dichotomous and ordered categorical data from randomized clinical trials and non-parametric strategies for addressing them[J]. Statistics in medicine, 1998, 17(15/16): 1 863-1 892.

[10] Lesaffre E, Senn S. A note on non-parametric ANCOVA for covariate adjustment in randomized clinical trials[J]. Statistics in medicine, 2003, 22(23): 3 583-3 596.

[11] Tsiatis A A, Davidian M, Zhang M, et al. Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach[J]. Statistics in medicine, 2008, 27(23): 4 658-4 677.

[12] Zhang M, Tsiatis A A, Davidian M. Improving efficiency of inferences in randomized clinical trials using auxiliary covariates[J]. Biometrics, 2008, 64(3): 707-715.

[13] McArdle B H, Anderson M J. Fitting multivariate models to community data: a comment on distance-based redundancy analysis[J]. Ecology, 2001, 82(1): 290-297.

[14] Li Q, Wacholder S, Hunter D J, et al. Genetic background comparison using distance-based regression, with applications in population stratification evaluation and adjustment[J]. Genetic epidemiology, 2009, 33(5): 432-441.

[15] Pan W. Relationship between genomic distance-based regression and kernel machine regression for multi-marker association testing[J]. Genetic epidemiology, 2011, 35(4): 211-216.

[16] Wessel J, Schork N J. Generalized genomic distance-based regression methodology for multilocus association analysis[J]. The American Journal of Human Genetics, 2006, 79(5): 792-806.

[17] Zapala M A, Schork N J. Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables[J]. Proceedings of the national academy of sciences, 2006, 103(51): 19 430-19 435.