Published in final edited form as:
J Am Stat Assoc. 2010 June 1; 105(490): 578–587. doi:10.1198/jasa.2010.ap09114.

A Rank-Based Test for Comparison of Multidimensional Outcomes

Aiyi Liu[Investigator],
Eunice Kennedy Shriver National Institute of Child Health and Human Development

Qizhai Li[Postdoctoral Fellow],
National Cancer Institute

Chunling Liu[Postdoctoral Fellow],
Eunice Kennedy Shriver National Institute of Child Health and Human Development

Kai Yu[Investigator], and
National Cancer Institute

Kai F. Yu[Senior Investigator]
Eunice Kennedy Shriver National Institute of Child Health and Human Development

Aiyi Liu: liua@mail.nih.gov

Abstract

For comparison of multiple outcomes commonly encountered in biomedical research, Huang et al. (2005) improved O'Brien's (1984) rank-sum tests through the replacement of the ad hoc variance by the asymptotic variance of the test statistics. The improved tests control the Type I error rate at the desired level and gain power when the differences between the two comparison groups in each outcome variable fall into the same direction. However, they may lose power when the differences are in different directions (e.g., some are positive and some are negative). These tests and the popular Bonferroni correction failed to show important significant difference when applied to compare heart rates from a clinical trial to evaluate the effect of a procedure to remove the cardioprotective solution HTK. We propose an alternative test statistic, taking the maximum of the individual rank-sum statistics, which controls the type I error and maintains satisfactory power regardless of the directions of the differences. Simulation studies show the proposed test to be of higher power than other tests in certain alternative parameter space of interest. Furthermore, when used to analyze the heart rates data the proposed test yields more satisfactory results.

Keywords

Autism spectrum disorder; Behrens-Fisher problem; Cardioprotective solution; Case-control studies; Growth hormones; Multiple outcomes; Non-parametrics; Rank-sum statistics

1. INTRODUCTION

Multiple outcomes are frequently encountered in biomedical research, for example, in clinical trials, genome-wide association studies, and disease-exposure association studies. In a randomized clinical trial of diabetes to determine which treatment yields better nerve function, a total of 34 electromyographic variables are measured that jointly define the nerve function (O'Brien 1984). A clinical trial of Coenzyme Q_{10} in treating early Parkinson's disease to slow the functional decline caused by the disease was described in Huang et al. (2005). The functional decline is measured by a number of outcome variables, including mentation, motor, and average daily living scales. Other clinical trial examples can be found

in Pocock, Geller and Tsiatis (1987), Shames et al. (1998), Tilley et al. (1999), and Li, Zhao and Paty (2001).

Under the assumption of multivariate normality, multiple outcomes between groups can be compared using a parametric procedure, such as multivariate ANOVA (Brunner, Munzel and Puri 2002), ordinary least squares, Hotelling's T^2, or the Bonferroni procedure (Pocock, Geller and Tsiatis 1987). Nonparametric procedures that are robust against distributional assumptions such as the rank-sum test (O'Brien, 1984) and the adjusted rank-sum test (Huang et al. 2005) are also available. Most of the methods are obtained based on the null hypothesis that the two or more groups have a common distribution.

O'Brien's (1984) rank-sum-type tests are distribution-free and used to evaluate whether the outcome measures from one treatment group are uniformly better than those from the other group. The null hypothesis is that outcomes from the two treatment groups have a common distribution and the tests are robust for small sample size. The tests have been widely used in biomedical research (Kaufman et al. 1998; Shames et al. 1998; Li, Zhao and Paty 2001). For the generalized nonparametric Behrens-Fisher hypothesis, Huang et al. (2005) showed that O'Brien's (1984) rank-sum tests have an inflated type I error rate when the groups being compared have different variances. They moved on to propose variance-adjusted rank-sum tests that control the type I error of the tests.

The rank-sum-type tests of O'Brien (1984) and Huang et al. (2005) gain power by accumulating evidence across comparisons on each individual outcome, but they may lose power in situations when the differences in the outcomes between the two samples exist but fall in different directions (some differences are positive and some are negative), or in the same direction with relatively much varied magnitudes. As a result, they may fail to reveal important differences between the two groups under investigation. For example, these tests and the popular Bonferroni correction failed to show important significant difference when applied to compare heart rates from a clinical trial to evaluate the effect of a procedure to remove the cardioprotective solution HTK. To overcome this, we propose a simple yet robust test that maintains satisfactory power regardless of the directions of the differences. Simulation results show that our approach controls the type I error rate and is more powerful than O'Brien's (1984) tests and the tests of Huang et al. (2005) in certain parameter space of intertest. Furthermore, when used to analyze the heart rates data the proposed test succeeds in finding important treatment effects. Growth-related hormones data from the Autism/ASD Study are also used to illustrate the application of the proposed test.

2. TWO MOTIVATING DATA EXAMPLES

2.1 Autism/ASD Study

The Growth and Maturation in Children with Autism or Autistic Spectrum Disorder (ASD) Study (the Autism/ASD Study) is a case-control study conducted by the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development in 2002–2005. The study was designed to determine whether children with autism/ASD between the ages of 4 and 8 are larger than matched control children, and if so, what role hormones play. Potential cases were identified from the patient records of the Kelly O'Leary Center for Autism Spectrum Disorders at Cincinnati Children's Hospital Medical Center (CCHMC). Patients evaluated for a possible diagnosis of autism/ASD receive a multidisciplinary evaluation that typically includes a detailed medical history, a physical examination by a developmental pediatrician, and application of the Autism Diagnostic Observation Schedule (ADOS), published by Western Psychological Services. Eighty-one subjects, 75 boys and 6 girls, diagnosed as having autism/ASD were enrolled. Age-matched controls were recruited from one of CCHMC's ENT outpatient surgery facilities. Eighty control children, 59 boys and 21

girls, were enrolled. Three major components of the comparison were: (a) bone age, (b) weight, height, head circumference and body mass index, and (c) growth-related hormones. Blood samples were assayed for insulin-like growth factors (IGF-1, IGF-2), insulin-like growth factor binding protein (IGFBP-3), and growth hormone binding protein (GHBP), as well as for dehydroepiandrosterone (DHEA) and DHEA-sulphate (DHEAS). Details of subject enrollment and data collection were described in Mills et al. (2007).

2.2 Removal of Cardioplegic Solution (HTK Study)

Section 1.3.14 of Brunner, Domhof and Langer (2002) described a clinical trial to evaluate the effect of a procedure to remove the cardioprotective solution HTK. A total of 30 patients undergoing coronary bypass surgery were evenly randomized to the treatment and control groups. The heart rates of these 30 patients were recorded at five particular times, namely, after induction of anesthesia, pre-cardiopulmonary bypass (CPB), 15 minutes after separation from CPB, 3 hours after the end of surgery, and 6 hours after the end of surgery. The original data can be found in Appendix A.15 in Brunner, Domhof and Langer (2002). The question of interest is whether the heart rates are different at any of the time points.

Data from the HTK and the Autism/ASD study will be used to illustrate the performance of the method proposed in the present paper.

3. METHODS

3.1 Rank-MAX Statistics for the Generalized Behrens-Fisher Problem

We consider comparison of p-dimensional outcomes, each measured on a continuous scale, from two groups, $X = (X_1, \cdots, X_p)'$ and $Y = (Y_1, \cdots, Y_p)'$, following distributions F and G, respectively. The marginal distributions corresponding to X_a and Y_a are F_a and G_a, respectively, with $a = 1, \cdots, p$.

Following Huang et al. (2005), we define

$$\theta_a = \Pr(X_a < Y_a) - \Pr(X_a > Y_a), a = 1, \cdots, p.$$

The null hypothesis being tested is

$$H_0: \theta_1 = \cdots = \theta_p = 0.$$

In practice $\theta_a/2 = \Pr(X_a < Y_a) - 1/2$ is called the relative effect of the Y-group with respect to the X-group for the ath outcome variable. The null hypothesis is often referred to as a generalized or nonparametric Behrens-Fisher problem; see Brunner, Domhof and Langer (2002, Chapter 3). Alternatively a more restrictive null hypothesis $H_0^{'}: F = G$, a special case of H_0, is also often used.

Let $x_i = (x_{i1}, \cdots, x_{ip})'$, $i = 1, \cdots, m$, be the outcomes of the ith subject from the X-sample and $y_j = (y_{j1}, \cdots, y_{jp})'$, $j = 1, \cdots, n$, be the outcomes of the jth subject from the Y-sample, and let $N = m + n$, the total number of subjects. For the ath outcome variable, $a = 1, \cdots, p$, we combine the two samples and rank the N observations $x_{1a}, \cdots, x_{ma}, y_{1a}, \cdots, y_{na}$, and denote the midranks of x_{ia} and y_{ja} by R_{xia} and R_{yja}, respectively. Define

$$R_{xi\cdot}=\sum_{a=1}^{p}R_{xia},\ \bar{R}_{x\cdot\cdot}=\frac{1}{m}\sum_{i=1}^{m}R_{xi\cdot},\ \hat{\sigma}^2_{x\cdot\cdot}=\frac{1}{m-1}\sum_{i=1}^{m}(R_{xi\cdot}-\bar{R}_{x\cdot\cdot})^2;$$
$$R_{yj\cdot}=\sum_{a=1}^{p}R_{yja},\ \bar{R}_{y\cdot\cdot}=\frac{1}{n}\sum_{j=1}^{n}R_{yj\cdot},\ \hat{\sigma}^2_{y\cdot\cdot}=\frac{1}{n-1}\sum_{j=1}^{n}(R_{yj\cdot}-\bar{R}_{y\cdot\cdot})^2.$$

Applying the two-sample t-test to the summary rank statistics $\{R_{xi\cdot}\}$ and $\{R_{yj\cdot}\}$, O'Brien (1984) considered two test statistics for testing H_0':

$$T_1=(\bar{R}_{y\cdot\cdot}-\bar{R}_{x\cdot\cdot})/\sqrt{(1/m+1/n)[(m-1)\hat{\sigma}^2_{x\cdot\cdot}+(n-1)\hat{\sigma}^2_{y\cdot\cdot}]/(m+n-2)},$$

and

$$T_2=(\bar{R}_{y\cdot\cdot}-\bar{R}_{x\cdot\cdot})/\sqrt{\hat{\sigma}^2_{x\cdot\cdot}/m+\hat{\sigma}^2_{y\cdot\cdot}/n}.$$

Under H_0' : $F = G$, T_1 and T_2 both follow a t-distribution with degrees of freedom equal to $m + n - 2$ and $[\zeta^2/(m-1)+(1-\zeta)^2/(n-1)]^{-1}$, respectively, where $\zeta=(\hat{\sigma}^2_{x\cdot\cdot}/m)/(\hat{\sigma}^2_{x\cdot\cdot}/m+\hat{\sigma}^2_{y\cdot\cdot}/n)$.

Huang et al. (2005) noticed that under the more restrictive null hypothesis, H_0' : $F = G$, both T_1 and T_2 asymptotically follow the standard normal distribution. However, when $F \neq G$, these two statistics remain asymptotically normally distributed but have nonunit variances. When being used to test the generalized Behrens-Fisher hypothesis H_0, these test statistics can substantially inflate the Type I error rate, as being demonstrated in Huang et al. (2005). To make O'Brien's (1984) test suitable for testing the null hypothesis H_0, Huang et al. (2005) derived the asymptotic variances of the two statistics and suggested using the following two modified test statistics for H_0:

$$T_{h1}=(\bar{R}_{y\cdot\cdot}-\bar{R}_{x\cdot\cdot})/\sqrt{\hat{h}_1(1/m+1/n)[(m-1)\hat{\sigma}^2_{x\cdot\cdot}+(n-1)\hat{\sigma}^2_{y\cdot\cdot}]/(m+n-2)},$$
$$T_{h2}=(\bar{R}_{y\cdot\cdot}-\bar{R}_{x\cdot\cdot})/\sqrt{\hat{h}_2(\hat{\sigma}^2_{x\cdot\cdot}/m+\hat{\sigma}^2_{y\cdot\cdot}/n)},$$

where $\hat{h}_1$ and $\hat{h}_2$ are, respectively, the consistent estimates of

$$h_1=\frac{\sum_{a=1}^{p}\sum_{b=1}^{p}(1+\lambda)^2(c_{ab}+d_{ab}\lambda)}{\sum_{a=1}^{p}\sum_{b=1}^{p}[e_{ab}\lambda^3+(d_{ab}+2f_{ab})\lambda^2+(c_{ab}+2\eta_{ab})\lambda+\xi_{ab}]},$$

$$\text{and } h_2=\frac{\sum_{a=1}^{p}\sum_{b=1}^{p}(1+\lambda)^2(c_{ab}+d_{ab}\lambda)}{\sum_{a=1}^{p}\sum_{b=1}^{p}[d_{ab}\lambda^3+(e_{ab}+2\eta_{ab})\lambda^2+(\xi_{ab}+2f_{ab})\lambda+c_{ab}]},$$

where $c_{ab} = \mathrm{Cov}(G_a(X_a), G_b(X_b))$, $d_{ab} = \mathrm{Cov}(F_a(Y_a), F_b(Y_b))$, $e_{ab} = \mathrm{Cov}(F_a(X_a), F_b(X_b))$, $f_{ab} = \mathrm{Cov}(F_a(X_a), G_b(X_b))$, $\xi_{ab} = \mathrm{Cov}(G_a(Y_a), G_b(Y_b))$, $\eta_{ab} = \mathrm{Cov}(G_a(Y_a), F_b(Y_b)$ and $\lambda = m/n$.

The two-sided test of Huang et al. (2005) rejects the generalized Behrens-Fisher null hypothesis H_0 at level α if $|T_{h1}|$ (or $|T_{h2}|$) exceeds $Z_{1-\alpha/2}$, the $\alpha/2$ upper-tail quartile of the standard normal distribution. The test maintains good power in the alternative parameter space when the θ_as are in the same direction. However in the parameter space when the θ_as fall in different directions, or in the same direction with relatively much varied magnitudes, the test may suffer from substantial loss of power. To see this, we define $\bar{R}_{x\cdot a}=\sum_{i=1}^{m} R_{xia}/m$ and $\bar{R}_{y\cdot a}=\sum_{j=1}^{m} R_{yja}/n$. Then $\bar{R}_{y\cdot\cdot} - \bar{R}_{x\cdot\cdot}=\sum_{a=1}^{p}[\bar{R}_{y\cdot a} - \bar{R}_{x\cdot a}]$, and thus $E(\bar{R}_{y\cdot\cdot} - \bar{R}_{x\cdot\cdot})=\sum_{a=1}^{p}\theta_a$. Clearly the power of the test is asymptotically α for parameters in the alternative subspace defined as $\{\theta_a : \sum_{a=1}^{p}\theta_a=0, \theta_a \neq 0 \text{ for some } a\}$. This is true even when the magnitude of some of the θ_as are large. For example, with $p = 2$ and $\theta_1 = -\theta_2 \gg 0$, the power of the test is still approximately α.

To overcome this, we consider a more robust test statistic,

$$T_{\max}=\max_{a\in\{1,\cdots,p\}}(|\bar{R}_{y\cdot a} - \bar{R}_{x\cdot a}|),$$

and reject H_0 if $T_{\max} > c_{\max}$. The statistic $T_{\max}$ can be viewed as a nonlinear combination of the test statistics $\bar{R}_{y\cdot a} - \bar{R}_{x\cdot a}$, $a = 1, \cdots, p$, while the test of Huang et al. (2005) can be viewed as a linear combination (with equal weights). Because H_0 is rejected if the observed relative effect of the Y-sample with respect to the X-sample is large, the test is expected to maintain satisfactory power regardless of the directions of the relative effects.

3.2 Statistical Significance of $T_{\max}$

In this subsection, we provide a procedure to calculate the critical value $c_{\max}$ and p-values of $T_{\max}$. Theorem 1 below gives the asymptotic joint distribution of $(\bar{R}_{y\cdot 1} - \bar{R}_{x\cdot 1}, \cdots, \bar{R}_{y\cdot p} - \bar{R}_{x\cdot p})'$ under the null hypothesis H_0.

Theorem 1. Under the null hypothesis H_0, $(\bar{R}_{y\cdot 1} - \bar{R}_{x\cdot 1}, \cdots, \bar{R}_{y\cdot p} - \bar{R}_{x\cdot p})'$ follows asymptotically a multivariate normal distribution with mean $(0, \cdots, 0)'$ and correlation coefficient matrix $\Lambda = (\rho_{ab})_{p\times p}$ as $\min\{m, n\} \to \infty$ and $0 < m/n \to \lambda_0 < \infty$, where

$\rho_{ab}=\frac{c_{ab}+\lambda_0 d_{ab}}{\sqrt{[c_{aa}+\lambda_0 d_{aa}][c_{bb}+\lambda_0 d_{bb}]}}$, $c_{ab} = \mathrm{Cov}(G_a(X_a), G_b(X_b))$, and $d_{ab} = \mathrm{Cov}(F_a(Y_a), F_b(Y_b))$.

Huang et al. (2005) provided a computational procedure for estimating c_{ab} and d_{ab}. The details are given in the Appendix. Based on

$$\Pr_{H_0}(T_{\max}>t)=1-\Pr_{H_0}(|\bar{R}_{y\cdot 1} - \bar{R}_{x\cdot 1}|<t, \cdots, |\bar{R}_{y\cdot p} - \bar{R}_{x\cdot p}|<t),$$

and the multivariate integration, we can evaluate the statistical significance of $T_{\max}$.

4. SIMULATION STUDIES

In this section, we conduct simulation studies to explore the performance of the proposed test statistic $T_{\max}$ by comparing its type I error rate and power with that of O'Brien's (1984) and Huang et al.'s (2005) tests. We consider generating data from various scenarios, 4-

variate normal distributions, bivariate exponential distributions, and ordinal variables with 5 levels. Power is compared among tests for two selected configurations of the relative effect parameters θ_a, one when they are in different directions and one when they are in the same direction but with relatively much varied magnitudes.

4.1 Multivariate Normal Distributions

We first conduct a simulation study to evaluate the type I error rate and power of the tests with data from 4-dimensional distributions. To this end, we generate $X = (X_{i1}, X_{i2}, X_{i3}, X_{i4})'$, $i = 1, \ldots, m$, *iid*, from a 4-variate normal distribution with mean $(0, 0, 0, 0)'$ and variance-covariance matrix $(u_{rs})_{4\times4}$, where $u_{rr} = 1$ for $r \in \{1, 2, 3, 4\}$, and $u_{rs} = 0.8$ for $r \neq s \in \{1, 2, 3, 4\}$, and $Y = (Y_{j1}, Y_{j2}, Y_{j3}, Y_{j4})'$, $j = 1, \ldots, n$, *iid*, from a 4-variate normal distribution with mean $(0, 0, 0, 0)'$ and variance-covariance matrix $(\upsilon_{rs})_{4\times4}$, where $\upsilon_{rr} = 16$ for $r \in \{1, 2, 3, 4\}$, and $\upsilon_{rs} = 14.4$ for $r \neq s \in \{1, 2, 3, 4\}$. Clearly the null hypothesis holds with these two distributions, i.e., for any i and j, $Pr(X_{ia} < Y_{ja}) - Pr(X_{ia} > Y_{ja}) = 0$, $a = 1, 2, 3, 4$.

The simulated power is obtained similarly under the same settings with the mean vector of X setting to be $(-0.5, -0.5, 0.5, 0.5)'$. The mean vector of Y is set to be $(0.5, 0.5, -0.5, -0.5)'$ and $(-0.4, -0.4, 0.6, 2.0)'$, respectively, with the former resulting in different directions in θs and the latter yielding same direction for θs with varied magnitudes.

We generate 10,000 replicates for each pair of m and n selected from {50, 100, 200}. The simulated Type I error is the proportion of the null hypothesis H_0 being rejected at a nominal significance level of 0.05 (two-sided).

Table 1 summarizes the empirical type I error and power. Comparing tests with significance level 0.05, we can see, from the table, that the tests of Huang et al. (2005) and the proposed $T_{\max}$ both control the type I error rate at about the nominal level, 0.05, while O'Brien's (1984) tests have a substantially inflated type I error rate. For example, when $m = 100$, $n = 50$, the empirical type I error rates of O'Brien's two tests are 0.117 and 0.066, respectively, and the two tests of Huang et al. give 0.056 and 0.054, respectively, whereas the proposed test gives 0.058. As expected, when the parameters θ_a fall into differene directions, the proposed test is considerably more powerful than the tests of O'Brien and Huang et al. For example when $m = 100$, $n = 100$, O'Brien's two tests have power levels of 0.067 and 0.066 respectively, and the power = 0.052 for both tests of Huang et al.. The proposed test, however, has a power level as high as 0.907. For the second configuration when θ_a are in the same direction but with varied magnitudes, the proposed test still has the highest power, though the other tests gained considerably in power as compared to the first configuration. With $m = n = 100$, O'Brien's two tests yield power level of 0.221 and 0.220 respectively, and the power = 0.179 and 0.178 for both tests of Huang et al.. The proposed test, however, has a power level of 0.830.

4.2 Bivariate Exponential Distribution

The second simulation study generates data from $X = (X_1, X_2)'$ and $Y = (Y_1, Y_2)$, each following a bivariate exponential distribution as defined by Marshall and Olkin (1967a, b): two random variables W_1 and W_2 follow a bivariate exponential distribution, denoted by $\text{BiExp}(\lambda_1, \lambda_2, \lambda_{12})$, if their joint survival function is $\exp(-\lambda_1 w_1 - \lambda_2 w_2 - \lambda_{12} \max(w_1, w_2))$. The correlation coefficient between W_1 and W_2 is then $\lambda_{12}/(\lambda_1 + \lambda_2 + \lambda_{12})$. To draw samples for (W_1, W_2), one can first generate samples from U_1, U_2 and U_{12}, mutually independent and following univariate exponential distributions with parameters λ_1, λ_2 and λ_{12}, respectively, and then set $W_1 = \min(U_1, U_{12})$ and $W_2 = \min(U_2, U_{12})$. We assess the tests under investigation with significance level 0.05 and 10,000 replicates. The type I error rates for the tests are evaluated with $X, Y \sim \text{BiExp}(1, 2, 5)$. The power of the tests is evaluated

respectively with $X \sim \text{BiExp}(1, 2, 1)$ and $Y \sim \text{BiExp}(2, 1, 1)$, and with $X \sim \text{BiExp}(1, 1, 1)$ and $Y \sim \text{BiExp}(2, 1.2, 1)$, yielding θ_1 and θ_2 in opposite and same directions, respectively. The number of subjects in each group is chosen from {50, 100, 200}. Here the rejection region is for two-sided tests.

Table 2 summarizes the empirical type I error and power. From the table, all tests control the type I error rate at about the nominal level, 0.05. For example, when $m = 200$ and $n = 100$, the empirical type I error rates are 0.053 and 0.054 for O'Brien's tests, 0.055 for Huang et al's tests, and 0.052 for the proposed test. However, as expected, the proposed test has the highest power among the tests being assessed when θ_1 and θ_2 are in opposite directions. For example when $m = 100$ and $n = 200$, the power is 0.046 and 0.048 for O'Brien's two tests, and 0.051 and 0.050 for Huang et al.'s tests. The power of the proposed method has the highest value, 0.965. For the second configuration with $\theta_1 = -0.20$ and $\theta_2 = -0.05$ the proposed test still has the highest power though the gain is relatively smaller as compared to the first configuration. With $m = 100$ and $n = 200$, O'Brien's two tests yield power level of 0.557 and 0.536 respectively, and the power = 0.540 and 0.537 for both tests of Huang et al.. The proposed test, however, has a power level of 0.721.

4.3 Ordinal Variables

Ordinal outcomes are common in biomedical research, often representing, for example, the various stages of severity of a disease or levels of improvement of a patient after being treated. For the simulation, we consider $p = 3$ ordinal variables with five different levels, −2, −1, 0, 1, and 2. The outcomes (x_{i1}, x_{i2}, x_{i3}) are generated according to the following formula:

$$x_{ia} = -2I_{\{x'_{ia} < c_1\}} - I_{\{c_1 \le x'_{ia} < c_2\}} + I_{\{c_3 \le x'_{ia} < c_4\}} + 2I_{\{x'_{ia} < c_4\}},$$

where x'_{ia} *iid* are drawn from the uniform distribution $U(-1, 1)$, $a = 1, \cdots, 3$, and c_1, c_2, c_3, c_4 are parameters determining the distributions of the ordinal variables. For assessing the type I error rate, we set (c_1, c_2, c_3, c_4) to (−0.2, −0.1, 0.1, 0.2) for the X-sample, and (−0.9, −0.8, 0.8, 0.9) for the Y-sample.

To evaluate the power we set (c_1, c_2, c_3, c_4) to (0.1, 0.3, 0.5, 0.7) for x_{i1}, (−0.2, 0.0, 0.2, 0.4) for both x_{i2} and x_{i3}, and (−0.2, 0.0, 0.2, 0.4) for y_{i1}, and (0.1, 0.3, 0.5, 0.7) for both y_{i2} and y_{i3} so that θ_a are in different directions and (c_1, c_2, c_3, c_4) to (0, 0.2, 0.4, 0.8) for x_{i1}, (−0.5, 0.2, 0.4, 0.8) for both x_{i2} and x_{i3}, and (−0.9, 0.2, 0.4, 0.9) for y_{i1}, and (−0.6, 0.2, 0.4, 0.8) for both y_{i2} and y_{i3} to yield θ_a in the same direction. Again the type I error rates and power are simulated with 10,000 replicates at significance level 0.05 and the number of subjects in each group is chosen from {50, 100, 200}.

Table 3 shows the simulation results. The table shows that both the proposed method and the methods of Huang et al. maintain the correct type I error rates, which are very close to the nominal significance level, 0.05. However, O'Brien's tests have substantially inflated type I error rates, as expected. For example, when $m = 100$, and $n = 200$, the empirical type I error of O'Brien's two tests are 0.167 and 0.076, respectively, while Huang et al.'s two methods give 0.046 and 0.045, respectively. The proposed method gives 0.051. Furthermore, the proposed test is considerably more powerful than those of O'Brien and of Huang et al. when the θs are in different directions. For example, when $m = 100$, $n = 200$, the power levels for O'Brien's two tests are 0.400 and 0.392, respectively, and are 0.220 and 0.247, respectively, for the two tests of Huang et al. On the other hand, the proposed method achieves power as

high as 0.935. Similar pattern is observed again for θ_a to be in the same direction with varied magnitudes.

5. APPLICATIONS TO DATA EXAMPLES

In this section, we exemplify the methods using the growth-related hormone data from the Austim/ASD Study and removal of cardioplegic solution data from the HTK Study.

5.1 Growth-Related Hormone in Autism

We demonstrate the aforementioned methods with an examination of growth-related hormones in young children from the Autism/ASD Study, as described in the Introduction section. In Mills et al. (2007), it is noted that only the data on the male subjects were used in this analysis because of the small number of girls among the cases. Four boys in the case group did not provide blood samples. Consequently there were only 71 cases and 59 controls in this analysis. We confine our attention to five hormones: insulin-like growth factor-1 (IGF-1), insulin-like growth factor 2 (IGF-2), IGF binding protein (IGFBP-3), growth hormone binding protein (GHBP), and dehydroepiandrosterone (DHEA); their means and standard deviations are given in Table 4. DHEA-sulphate (DHEAS) was not included in the analysis since its levels were undetectable in more than half of the subjects (Mills et al., 2007). We are interested in whether the levels of a growth-related hormone, if any, differ between cases and controls. The null hypothesis is set to be the generalized Behrens-Fisher problem, i.e., there is no (relative) effect for any of the hormones under consideration.

We applied the proposed test and the tests of O'Brien (1984) and Huang et al. (2005) to the five growth-related hormone levels in cases and controls. The P-values were 2.86×10^{-6} and 1.75×10^{-6} for O'Brien's tests and 6.28×10^{-7} and 6.30×10^{-7} for the two tests of Huang et al. In contrast the proposed test yielded a P-value of 1.52×10^{-9}, indicating that the proposed method is more powerful than O'Brien's and Huang et al.'s methods.

5.2 Removal of Cardioplegic Solution (HTK Study)

Due to technical reasons, there were a few missing observations in the study, which are replaced by the mean of each endpoint in each group for our analysis. Means and standard deviations of the heart rates at each time point are presented in the table below.

A univariate comparison of the heart rates between the two groups at each of the five time points by the Wilcoxon rank-sum test yields P-values of 0.663, 0.724, 0.648, 0.110, 0.014, respectively. Using the conservative Bonferroni correction procedure at significance level of 0.05, we would conclude that there are no significant differences at the five time points between the groups.

The P-values are respectively 0.295 and 0.295 for O'Brien's (1984) tests and 0.289 and 0.289 for the tests of Huang et al. (2005). All tests failed to detect any difference at certain time point. On the other hand, for our proposed MAX statistic, the P-value is 0.032, thus effectively detecting a difference.

6. DISCUSSION

For comparing the distributions of two samples with multiple endpoints, we proposed using the MAX statistic and via simulation and real data examples demonstrated its effectiveness, as compared with the methods of O'Brien (1984) and Huang et al. (2005), in maintaining high power in certain parameter space and detecting differences between the two samples at individual endpoints. In the growth hormone example, all methods under investigation produced significant results. In the HTK Study example, our proposed method successfully

detected a difference while the others (including the conservative Bonferroni procedure) failed to do so. Our simulation results demonstrate that the MAX statistic is more efficient than the other methods when the relative effects fall into different directions or in the same direction but with relatively different magnitudes.

The MAX statistic, like the test statistics of O'Brien and Huang et al., can be used for comparison of two multivariate distributions, as a nonparametric alternative to Hotelling's T^2, a popular test statistic in multivariate data analysis (Anderson, 2003; Muirhead, 1982). These procedures are often termed as "global" tests in contrast to tests for each individual marginal distribution. In clinical trial settings where the trial has several endpoints that are equally ordered by their clinical importance, the MAX statistic, or any other proper overall significance tests in that matter, can be used for monitoring of the trial, followed by a step-down method to identify individual endpoints with positive or negative treatment effects; see Jennison and Turnbull (2000, chapter 15). One important implication is that when an overall significance test is used for possible early stopping of the trial, the MAX statistic may lead to early termination (while the others may fail to do so) if indeed the differences between treatment groups are statistically significant but fall into different directions, thus saving sample sizes and reducing study costs.

Often in practice multiple outcomes are ordered hierarchically according to their clinical importance. In this case the MAX statistic could be extended to

$$T_{\max} = \max_{a \in \{1, \cdots, p\}} (w_a |\overline{R}_{y \cdot a} - \overline{R}_{x \cdot a}|),$$

where the weights $w_a \geq 0$, $\sum_{a=1}^{p} w_a = 1$ are properly chosen so that more important outcomes carry larger weights.

The test statistics under investigation can all be viewed as forms of combinations of univariate tests on the individual outcomes. The test statistics of O'Brien (1984) and Huang et al. (2005) are linear combinations of the individual test statistics, while the MAX statistic is a nonlinear combination of the individual test statistics. In general, nonlinear combinations are expected to be more efficient when the parameters lie beyond the multidimensional plane determined by the linear combinations. With $p = 2$ we show analytically in the Appendix that the MAX statistic is more powerful than the test statistic in Huang et al. (2005) in some parameter space. Because the asymptotic distribution of the MAX statistic is quite complicated, more theoretical research is deemed needed to show the superiority of power of the MAX statistic over the other statistics in certain parameter space.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to thank Dr. B.J. Stone for her help. Research of the authors are supported by the Intramural Research Program of the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development (AL, CL, KFY), and the National Cancer Institute (QL, KY), National Institutes of Health. The opinions expressed in the article are not necessarily those of the National Institutes of Health. The authors thank two referees, an associate editor and the editor for their thoughtful comments and suggestions that improve the paper.

REFERENCES

Anderson, TW. An Introduction to Multivariate Statistical Analysis. 3rd ed.. Hoboken: Wiley; 2003.

Brunner E, Munzel U, Puri ML. The multivariate nonparametric Behrens-Fisher problem. Journal of Statistical Planning and Inference. 2002; 108:37–53.

Brunner, E.; Domhof, S.; Langer, F. Nonparametric analysis of longitudinal data in factorial experiments. New York: Wiley; 2002.

Huang P, Tilley BC, Woolson RF, Lipsitz S. Adjusting O'Brien's test to control type I error for the generalized nonparametric Behrens-Fisher problem. Biometrics. 2005; 61:532–539. [PubMed: 16011701]

Jennison, C.; Turnbull, BW. Group Sequential Methods with Applications to Clinical Trials. New York: Chapman Hall/CRC; 2000.

Kaufman KD, Oslen EA, Whiting D, Savin R, Devillez R, Bergfeld W. Finasteride in the treatment of men with androgenetic alopecia. Journal of the American Academy of Dermatology. 1998; 39:578–589. [PubMed: 9777765]

Li DK, Zhao GJ, Paty DW. Randomized controlled trial of interferon-beta-1a in secondary progressive MS: MRI results. Neurology. 2001; 56:1505–1513. [PubMed: 11402107]

Marshall AW, Olkin I. A multivariate exponential distribution. Journal of the American Statistical Association. 1967a; 62:30–44.

Marshall AW, Olkin I. A generalized bivariate exponential distribution. Journal of Applied probability. 1967b; 4:291–302.

Mills JL, Hediger ML, Molloy CA, Chrousos GP, Manning-Courtney P, Yu KF, Brasington M, England LJ. Elevated levels of growth-related hormones in autism and autism spectrum disorder. Clinical Endocrinology. 2007; 67:230–237. [PubMed: 17547689]

Muirhead, RJ. Aspects of Multivariate Statistical Theory. New York: Wiley; 1982.

O'Brien PC. Procedures for comparing samples with multiple endpoints. Biometrics. 1984; 40:1079–1087. [PubMed: 6534410]

Pocock SJ, Geller NL, Tsiatis AA. The analysis of multiple endpoints in clinical trials. Biometrics. 1987; 43:487–498. [PubMed: 3663814]

Shames RS, Heilbron DC, Janson SL, Kishiyama JL, Au DS, Adelman DC. Clinical differences among women with and without self-reported perimenstrual asthma. Annals of Allergy Asthma Immunology. 1998; 81:65–72.

Tilley BC, Pillemer SR, Heyse SP, Li S, Clegg DO, Alarón GS. Global statistical tests for comparing multiple outcomes in rheumatoid arthritis trials. Arthritis and Rheumatism. 1999; 42:1879–1888. [PubMed: 10513802]

APPENDIX: TECHNICAL DETAILS

Proof of Theorem 1. Since

$$\begin{pmatrix} \overline{R}_{y\cdot 1} - \overline{R}_{x\cdot 1} \\ \vdots \\ \overline{R}_{y\cdot p} - \overline{R}_{x\cdot p} \end{pmatrix} = \frac{N}{2mn} \begin{pmatrix} \sum_{i=1}^{m}\sum_{j=1}^{n}[I_{\{x_{i1}<y_{j1}\}} - I_{\{x_{i1}>y_{j1}\}}] \\ \vdots \\ \sum_{i=1}^{m}\sum_{j=1}^{n}[I_{\{x_{ip}<y_{jp}\}} - I_{\{x_{ip}>y_{jp}\}}] \end{pmatrix}$$

is a *U*-statistic, it converges in distribution to a normal distribution as $\min\{m, n\} \to \infty$ and $0 < m/n \to \lambda_0 < \infty$. Denote

$$\xi_{ab}^{(i_1 i_2 j_1 j_2)} = \mathrm{Cov}(I_{\{x_{i_1 a}<y_{j_1 a}\}} - I_{\{x_{i_1 a}>y_{j_1 a}\}}), (I_{\{x_{i_2 b}<y_{j_2 b}\}} - I_{\{x_{i_2 b}>y_{j_2 b}\}}), i_1, i_2 \in \{1, \cdots, m\}$$

and $j_1, j_2, \in \{1, \cdots, n\}$. Then we have

1. when $i_1 \neq i_2, j_1 \neq j_2$, $\xi_{ab}^{(i_1 i_2 j_1 j_2)}=0$.
2. when $i_1 = i_2 = i, j_1 \neq j_2$,

$$\begin{aligned}\xi_{ab}^{(i_1 i_2 j_1 j_2)} &= E\left[\left(I_{\{x_{ia}<y_{j_1 a}\}} - I_{\{x_{ia}>y_{j_1 a}\}}\right)\left(I_{\{x_{ib}<y_{j_2 b}\}} - I_{\{x_{ib}>y_{j_2 b}\}}\right)\right] \\ &= E\left\{E\left[\left(I_{\{x_{ia}<y_{j_1 a}\}} - I_{\{x_{ia}>y_{j_1 a}\}}\right)\left(I_{\{x_{ib}<y_{j_2 b}\}} - I_{\{x_{ib}>y_{j_2 b}\}}\right) | (x_{ia}, x_{ib})\right]\right\} \\ &= E\{[1 - 2G_a(X_a)][1 - 2G_b(X_b)]\} \\ &= 4\mathrm{Cov}(G_a(X_a), G_b(X_b)),\end{aligned}$$

   where the last equality follows from $E[G_a(X_a)] = E[G_b(X_b)] = 1/2$ under the null hypothesis.
3. when $i_1 \neq i_2, j_1 = j_2$, $\xi_{ab}^{(i_1 i_2 j_1 j_2)}=4\mathrm{Cov}(F_a(Y_a), F_b(Y_b))$.
4. when $i_1 = i_2, j_1 = j_2$, $\xi_{ab}^{(i_1 i_2 j_1 j_2)}=\eta * I_{\{a\neq b\}}+I_{\{a=b\}}$, $|\eta| \leq 1$.

Therefore, Cov $(I_{\{x_{i_1 a}<y_{j_1 a}\}} - I_{\{x_{i_1 a}>y_{j_1 a}\}}, I_{\{x_{i_2 b}<y_{i_2 b}\}} - I_{\{x_{i_2 b}>y_{j_2 b}\}})$

$$= \begin{cases} 0, & i_1 \neq i_2, j_1 \neq j_2; \\ 4\mathrm{Cov}(F_a(Y_a), F_b(Y_b)), & i_1 \neq i_2, j_1 = j_2; \\ 4\mathrm{Cov}(G_a(X_a), G_b(X_b)), & i_1 = i_2, j_1 \neq j_2; \\ \eta * I_{\{a\neq b\}}+I_{\{a=b\}}, & i_1 = i_2, j_1 = j_2. \end{cases}$$

It follows that when $a \neq b$,

$$\begin{aligned}\mathrm{Cov}(\overline{R}_{ya} - \overline{R}_{xa}, \overline{R}_{yb} - \overline{R}_{xb}) &= \frac{N^2}{4m^2n^2}\sum_{i_1=1}^{m}\sum_{i_2=1}^{m}\sum_{j_1=1}^{n}\sum_{j_2=1}^{n}\mathrm{Cov}(I_{\{x_{i_1 a}<y_{j_1 a}\}} - I_{\{x_{i_1 a}>y_{j_1 a}\}}, I_{\{x_{i_2 b}<y_{i_2 b}\}} - I_{\{x_{i_2 b}>y_{j_2 b}\}}) \\ &= \frac{N^2[4mn(n-1)\mathrm{Cov}(G_a(X_a), G_b(X_b))+4nm(m-1)\mathrm{Cov}(F_a(Y_a), F_b(Y_b))+mn\eta]}{4m^2n^2} \\ &= \frac{N^2[(n-1)\mathrm{Cov}(G_a(X_a), G_b(X_b))+(m-1)\mathrm{Cov}(F_a(Y_a), F_b(Y_b))+\eta]}{mn},\end{aligned}$$

and when $a = b$,

$$\mathrm{Cov}(\overline{R}_{ya} - \overline{R}_{xa}, \overline{R}_{yb} - \overline{R}_{xb}) = \frac{N^2[(n-1)\mathrm{Cov}(G_a(X_a), G_a(X_a))+(m-1)\mathrm{Cov}(F_a(Y_a), F_a(Y_a))+0.25]}{mn}.$$

Therefore,

$$\begin{aligned}\rho_{ab} &= \mathrm{Cor}_{H_0}(\overline{R}_{ya} - \overline{R}_{xa}, \overline{R}_{yb} - \overline{R}_{xb}) \\ &= \frac{\mathrm{Cov}(\overline{R}_{ya} - \overline{R}_{xa}, \overline{R}_{yb} - \overline{R}_{xb})}{\sqrt{\mathrm{Var}(\overline{R}_{ya} - \overline{R}_{xa}\mathrm{Var}(\overline{R}_{yb} - \overline{R}_{xb})}} \\ &\approx \frac{c_{ab}+\lambda_0 d_{ab}}{\sqrt{[c_{aa}+\lambda_0 d_{aa}][c_{bb}+\lambda_0 d_{bb}]}},\end{aligned}$$

where $c_{ab} = \mathrm{Cov}(G_a(X_a), G_b(X_b))$ and $d_{ab} = \mathrm{Cov}(F_a(Y_a), F_b(Y_b))$.

Estimation of c_{ab} = Cov$(G_a(X_a), G_b(X_b))$ and d_{ab} = Cov$(F_a(Y_a), F_b(Y_b))$. For any $a \in \{1, \cdots, p\}$, define $R_y(x_{ia})$ to be the midrank of x_{ia} among $\{x_{ia}, y_{1a}, \cdots, y_{na}\}$, $R_x(x_{ia})$ the midrank

of x_{ia} among $\{x_{1a}, \cdots, x_{ma}\}$, $R_x(y_{ja})$ the midrank of y_{ja} among $\{x_{1a}, \cdots, x_{ma}, y_{ja}\}$, and $R_y(y_{ja})$ the midrank of y_{ja} among $\{y_{1a}, \cdots, y_{na}\}$. Let

$$\widehat{\theta}_a = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\{I_{\{x_{ia}<y_{ja}\}} - I_{\{x_{ia}>y_{ja}\}}\},$$

$P = (p_{ia})_{m\times p}$ with $p_{ia} = 2R_y(x_{ia}) - 2 - n + n\hat{\theta}_a$, and $Q = (q_{ja})_{n\times p}$ with $q_{ja} = 2R_x(y_{ja}) - 2 - m - m\hat{\theta}_a$, $i = 1, \cdots, m$, $j = 1, \cdots, n$, $a = 1, \cdots, p$. Then the consistent estimates of $\mathrm{Cov}(G_a(X_a), G_b(X_b))$ and $\mathrm{Cov}(F_a(Y_a), F_b(T_b))$, are, in matrix form,

$$\widehat{\mathrm{Cov}}(G_a(X_a), G_b(X_b)) = P'P/(4mn^2), \quad \text{and } \widehat{\mathrm{Cov}}(F_a(Y_a), F_b(Y_b)) = Q'Q/(4m^2n).$$

Efficiency of the MAX statistic ($p = 2$). Let Z_1 and Z_2 be two test statistics (corresponding to the rank test statistic of the first and second endpoint, respectively) with $(Z_1, Z_2)'$ asymptotically following a bivariate normal distribution with mean $(\mu_1, \mu_2)'$ and covariance matrix $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. Under the null hypothesis H_0 we have $\mu_1 = \mu_2 = 0$. We consider alternative parameter space H_1 with $(\mu_1, \mu_2) = (\mu, -\mu - \delta)$ where $\mu > 0$ and $\delta \geq 0$. Thus μ_1 and μ_2 are in opposite directions but their magnitudes differ according to δ. With the above notations the test statistic of Huang et al (2005) can be expressed as $(Z_1+Z_2)/(2\sqrt{1+\rho})$, and the MAX statistic is $\max\{|Z_1|, |Z_2|\}$.

Theorem. Under the above assumptions, for any given significance level α ($\alpha > 0$), there exists $\delta_0 > 0$ such that, when $\delta \in [0, \delta_0)$,

$$\mathrm{Pr}_{H_1}\left((Z_1+Z_2)/\left(2\sqrt{1+\rho}\right) > c\right) < \mathrm{Pr}_{H_1}(\max\{|Z_1|, |Z_2|\} > t),$$

where c and t satisfy

$$\mathrm{Pr}_{H_0}\left((Z_1+Z_2)/\left(2\sqrt{1+\rho}\right) > c\right) = \mathrm{Pr}_{H_0}(\max\{|Z_1|, |Z_2|\} > t) = \alpha.$$

Proof. First consider $\delta = 0$, yielding $\mu_1 + \mu_2 = 0$, and thus

$$\mathrm{Pr}_{H_1}\left((Z_1+Z_2)/\left(2\sqrt{1+\rho}\right) > c\right) = \alpha.$$

We need to show that $\mathrm{Pr}_{H_1}(\max\{|Z_1|, |Z_2|\} > t) > \alpha$.

Because

$$\mathrm{Pr}_{H_1}(\max\{|Z_1|, |Z_2|\} > t) = 1 - \mathrm{Pr}_{H_1}(|Z_1| \leq t, |Z_2| \leq t) = 1 - \int_{-t}^{t}\int_{-t}^{t} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{ - \frac{[(z_1-\mu)^2 + (z_2+\mu)^2 - 2\rho(z_1-\mu)(z_2+\mu)]}{2(1-\rho^2)}\right\} dz_1 dz_2 = 1 - g(\mu),$$

it suffices to show that $1 - g(\mu) > \alpha$ as $\mu > 0$, that is, $g(\mu) < 1 - \alpha$, since

$$g(\mu)=\begin{cases} 1-\alpha, & \mu=0; \\ 0, & \mu\to\infty. \end{cases}$$

Note that

$$\begin{aligned}\frac{dg(\mu)}{d\mu} &= \int_{-t}^{t}\int_{-t}^{t}\frac{1}{2\pi\sqrt{1-\rho^2}}\exp\left\{-\frac{(z_1-\mu)^2+(z_2+\mu)^2-2\rho(z_1-\mu)(z_2+\mu)}{2(1-\rho^2)}\right\}\times\left\{-\frac{-2(z_1-\mu)+2(z_2+\mu)+2\rho(z_2+\mu)-2\rho(z_1-\mu)}{2(1-\rho^2)}\right\}dz_1dz_2\\ &= \int_{-t}^{t}\int_{-t}^{t}\frac{1}{2\pi\sqrt{1-\rho^2}}\exp\left\{-\frac{(z_1-\mu)^2+(z_2+\mu)^2-2\rho(z_1-\mu)(z_2+\mu)}{2(1-\rho^2)}\right\}\times\left\{-\frac{-2(1+\rho)(z_1-\mu)+2(1+\rho)(z_2+\mu)}{2(1-\rho^2)}\right\}dz_1dz_2\\ &= \int_{-t}^{t}\int_{-t}^{t}\frac{1}{2\pi\sqrt{1-\rho^2}}\exp\left\{-\frac{(z_1-\mu)^2+(z_2+\mu)^2-2\rho(z_1-\mu)(z_2+\mu)}{2(1-\rho^2)}\right\}\times\left\{\frac{(z_1-\mu)-(z_2+\mu)}{1-\rho}\right\}dz_1dz_2.\end{aligned}$$

Write $\xi = (z_1 - \mu)$ and $\eta = z_2 + \mu$, then

$$\begin{aligned}\frac{dg(\mu)}{d\mu} &= \int_{-t+\mu}^{t+\mu}\int_{-t-\mu}^{t-\mu}\frac{1}{2\pi\sqrt{1-\rho^2}}\exp\left\{-\frac{\xi^2+\eta^2-2\rho\xi\eta}{2(1-\rho^2)}\right\}\left\{\frac{\xi-\eta}{1-\rho}\right\}d\xi d\eta\\ &= \int_{-t+\mu}^{t+\mu}\int_{-t-\mu}^{t-\mu}\frac{1}{2\pi\sqrt{1-\rho^2}}\exp\left\{-\frac{\xi^2+\eta^2-2\rho\xi\eta}{2(1-\rho^2)}\right\}\left\{\frac{2\xi}{1-\rho}\right\}d\xi d\eta\\ &= E\xi I_{\{-t-\mu\le\xi\le t-\mu\}}I_{\{-t+\mu\le\eta\le t+\mu\}} = -EI_{\{-t-\mu\le\xi\le t-\mu\}}\eta I_{\{-t+\mu\le\eta\le t+\mu\}}.\end{aligned}$$

When $\rho \le 0$, we have

$$\frac{dg(\mu)}{d\mu}=E\left(\xi I_{\{-t-\mu\le\xi\le t-\mu\}}I_{\{-t+\mu\le\eta\le t+\mu\}}\right)\le E\left(\xi I_{\{-t-\mu\le\xi\le t-\mu\}}\right)E\left(I_{\{-t+\mu\le\eta\le t+\mu\}}\right)<0,$$

and when $\rho > 0$, we have

$$\frac{dg(\mu)}{d\mu}=-E\left(I_{\{-t-\mu\le\xi\le t-\mu\}}\eta I_{\{-t+\mu\le\eta\le t+\mu\}}\right)\le -E\left(I_{\{-t-\mu\le\xi\le t-\mu\}}\right)E\left(\eta I_{\{-t+\mu\le\eta\le t+\mu\}}\right)<0.$$

In the above proof, we utilized the fact that the marginal distributions of ξ and η are both standard normal distribution.

Thus $dg(\mu)/d\mu < 0$ as $\mu \ge 0$, implying that the function $g(\mu)$ is strictly decreasing in μ and hence

$$\Pr_{H_1}(\max\{|Z_1|,|Z_2|\}>t)>\alpha=\Pr_{H_1}\left((Z_1+Z_2)/\left(2\sqrt{1+\rho}\right)>c\right).$$

Therefore we have

$$\Pr_{H_1}\left((Z_1+Z_2)/\left(2\sqrt{1+\rho}\right)>c\right)<\Pr_{H_1}(\max\{|Z_1|,|Z_2|\}>t),$$

when $\delta = 0$.

On the other hand, when $\delta \to \infty$ we have

$$\Pr_{H_1}\left((Z_1+Z_2)/\left(2\sqrt{1+\rho}\right)>c\right)=\Pr_{H_1}(\max\{|Z_1|,|Z_2|\}>t)=0.$$

Note that the power functions are continuous in δ and therefore there exists $\delta_0 > 0$, such that when $\delta \in (0, \delta_0)$,

$$\Pr_{H_1}\left((Z_1+Z_2)/\left(2\sqrt{1+\rho}\right)>c\right)<\Pr_{H_1}(\max\{|Z_1|,|Z_2|\}>t).$$

Figure 1 presents the ratios in power of the other four tests to the MAX statistic with significance level 0.05. Power of the test is computed based on 10,000 simulations of $X = (X_{i1}, X_{i2})'$, $i = 1, \ldots, 100$, *iid* from a bivariate normal distribution with mean $(-0.1, 0.1)'$ and covariance matrix $\begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$ and $Y = (Y_{j1}, Y_{j2})'$, $j = 1, \ldots, 100$, *iid*, from a bivariate normal distribution with mean $(0.1 + \delta, -0.1)'$ and covariance matrix $\begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$. A ratio below 1 indicates superiority of the MAX statistic to its counterpart. We observe from the figure that for a relatively large range of δ values the MAX statistic has higher power than the other statistics.

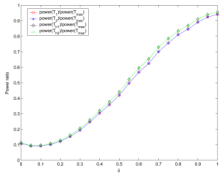

Figure 1.
Ratio of power levels, where power(T) denotes the power of the test statistic T. T_1 and T_2 are O'Brien's test statistics, T_{h1} and T_{h2} are Huang et al.'s test statistics, and $T_{\max}$ is the proposed test statistic.

Table 1

Type I error and power results under significance level, 0.05. (10,000 replicates) (4-dimensional normal distribution)

m	n	T_1	T_2	T_{h1}	T_{h2}	$T_{\max}$
Type I error rate ($\theta_1 = \theta_2 = \theta_3 = \theta_4 = 0$)						
50	50	0.071	0.070	0.050	0.049	0.052
100	100	0.077	0.077	0.056	0.056	0.054
200	200	0.077	0.076	0.052	0.052	0.045
50	100	0.030	0.079	0.048	0.048	0.046
100	200	0.030	0.084	0.048	0.049	0.046
100	50	0.117	0.066	0.056	0.054	0.058
200	100	0.123	0.066	0.052	0.050	0.051
Power($\theta_1 = \theta_2 = 0.19, \theta_3 = \theta_4 = -0.19$)						
50	50	0.060	0.059	0.047	0.046	0.557
100	100	0.067	0.066	0.052	0.052	0.907
200	200	0.068	0.068	0.051	0.052	1.000
50	100	0.022	0.067	0.045	0.045	0.882
100	200	0.024	0.068	0.047	0.047	0.999
100	50	0.120	0.063	0.058	0.055	0.579
200	100	0.116	0.060	0.051	0.048	0.916
Power($\theta_1 = \theta_2 = \theta_3 = 0.02, \theta_4 = 0.28$)						
50	50	0.149	0.147	0.117	0.116	0.519
100	100	0.221	0.220	0.179	0.178	0.830
200	200	0.359	0.359	0.303	0.302	0.993
50	100	0.124	0.231	0.176	0.176	0.808
100	200	0.225	0.372	0.296	0.296	0.989
100	50	0.216	0.137	0.122	0.116	0.531
200	100	0.309	0.214	0.192	0.188	0.839

Table 2

Type I error and power results under significance level, 0.05. (10,000 replicates) (Bivariate exponential distribution)

m	n	T_1	T_2	T_{h1}	T_{h2}	$T_{\max}$
Type I error rate($\theta_1 = \theta_2 = 0$)						
50	50	0.049	0.049	0.049	0.049	0.055
100	100	0.050	0.050	0.051	0.051	0.054
200	200	0.048	0.048	0.048	0.048	0.048
50	100	0.052	0.053	0.055	0.054	0.059
100	200	0.048	0.048	0.049	0.049	0.051
100	50	0.048	0.049	0.051	0.049	0.056
200	100	0.053	0.054	0.055	0.055	0.052
Power ($\theta_1 = -0.20, \theta_2 = 0.20$)						
50	50	0.047	0.047	0.050	0.050	0.596
100	100	0.048	0.048	0.051	0.051	0.891
200	200	0.044	0.044	0.046	0.046	0.997
50	100	0.051	0.052	0.056	0.055	0.732
100	200	0.046	0.048	0.051	0.050	0.965
100	50	0.047	0.049	0.051	0.050	0.727
200	100	0.045	0.046	0.049	0.048	0.966
Power ($\theta_1 = -0.20, \theta_2 = -0.05$)						
50	50	0.247	0.247	0.250	0.250	0.349
100	100	0.451	0.451	0.453	0.452	0.615
200	200	0.740	0.740	0.740	0.740	0.905
50	100	0.330	0.310	0.317	0.311	0.434
100	200	0.557	0.536	0.540	0.537	0.721
100	50	0.313	0.331	0.337	0.334	0.465
200	100	0.561	0.586	0.589	0.587	0.762

Table 3

Type I error and power results under significance level, 0.05. (10,000 replicates) (3-dimensional ordinal variables)

m	n	T_1	T_2	T_{h1}	T_{h2}	$T_{\max}$
Type I error rate ($\theta_1 = \theta_2 = \theta_3 = 0$)						
50	50	0.094	0.091	0.041	0.040	0.059
100	100	0.098	0.097	0.045	0.044	0.051
200	200	0.097	0.096	0.043	0.042	0.047
50	100	0.168	0.081	0.047	0.044	0.059
100	200	0.167	0.076	0.046	0.045	0.051
100	50	0.047	0.117	0.039	0.045	0.050
200	100	0.046	0.116	0.038	0.043	0.046
Power ($\theta_1 = 0.19, \theta_2 = -0.19, \theta = -0.19$)						
50	50	0.182	0.182	0.082	0.082	0.497
100	100	0.316	0.316	0.167	0.167	0.837
200	200	0.560	0.560	0.364	0.364	0.994
50	100	0.225	0.219	0.111	0.126	0.633
100	200	0.400	0.392	0.220	0.247	0.935
100	50	0.233	0.240	0.111	0.130	0.636
200	100	0.398	0.406	0.221	0.257	0.943
Power ($\theta_1 = 0.25, \theta_2 = 0.03, \theta = 0.03$)						
50	50	0.371	0.371	0.266	0.265	0.426
100	100	0.644	0.643	0.520	0.519	0.739
200	200	0.903	0.903	0.833	0.833	0.968
50	100	0.472	0.426	0.323	0.329	0.455
100	200	0.750	0.710	0.614	0.623	0.779
100	50	0.473	0.514	0.374	0.397	0.675
200	100	0.768	0.802	0.682	0.704	0.942

J Am Stat Assoc. Author manuscript; available in PMC 2011 May 26.

Table 4

Mean and Standard Deviation (SD) of the variables for Autism/ASD Study

	DHEA		IGF-1		IGF-2		IGFBP-3		GHBP	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Case	67.66	82.16	150.40	57.53	398.77	99.39	2.52	0.57	1010.77	380.61
Control	64.96	54.86	115.96	45.66	308.47	76.11	2.07	0.42	768.75	287.20

Table 5

Mean and Standard Deviation (SD) of the variables for HTK Study

	AIA		CPB		15MCPB		3HAES		6HAES	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Treatment	65.27	10.02	88.53	12.32	93.53	16.71	98.00	12.74	100.07	10.55
Control	68.40	13.37	90.73	13.74	89.87	11.73	91.33	12.19	89.13	12.84

AIA: after induction of anesthesia
CPB: pre-cardiopulmonary bypass (CPB)
15MCPB: 15 minutes after separation from CPB
3HAES: 3 hours after the end of surgery
6HAES: hours after the end of surgery

J Am Stat Assoc. Author manuscript; available in PMC 2011 May 26.