Least Conservative Critical Boundaries of Multiple Hypothesis Testing in a Range of Correlation Values

Jiangtao Gou
Department of Mathematics and Statistics, Villanova University, Villanova, PA

ABSTRACT
Under suitable assumptions we prove that there does not exist a perfect exact multiple test procedure that would apply simultaneously to any positive correlation coefficient even with a known distribution of test statistics. This nonexistence theorem holds for all simple tests under normal distribution, and holds for all tests under Ferguson’s distribution. Given the nonexistence of a perfect exact test, we provide least conservative tests using three parametric models. The average conservativeness of these tests can be reduced to as low as 1/8 of that of the widely used Simes test assuming normal distribution. Power analysis indicates that the newly proposed tests are useful in practice.

1. Introduction

Multiple test procedures almost always keep the probability of rejecting a null hypothesis below the desired significance level. For instance, the bivariate normal Simes (1986) test is a valid test with two uncorrelated or positively correlated outcomes, but is only exact when the correlation coefficient $\rho = 0$ or $\rho = 1$. The Simes test is conservative when $0 < \rho < 1$ because the bound in the corresponding Simes inequality is not tight (Finner, Roters, and Strassburger 2017). As a matter of course, the Simes test can be modified to have exact size $\alpha$, if we know the $\rho$ value. Examples of such modifications include Dunnett and Tamhane (1992), Cai and Sarkar (2006), and Gou and Tamhane (2014). Unfortunately, the correlation coefficient $\rho$ is unknown in most cases. In practice, we only have limited information on $\rho$ such as the sign of the value. Thus, we have to apply a conservative test to control the Type I error rate below $\alpha$ for all possible $\rho$ values, even though the probability of rejecting the null hypothesis never attains its desired level for almost all situations.

A natural question to ask is whether it is possible to obtain a perfect test that controls the Type I error exactly at level $\alpha$ for a wide range of correlation coefficients. Here in this article we prove that, such a perfect hypothesis test is mostly unlikely. However, we can aim for subperfect tests which are less conservative than the commonly used multiple hypothesis tests in a given range of correlation coefficients.

The present article focuses on bivariate tests where the number of hypotheses $n = 2$, and includes discussion on multivariate tests in the last section. In the next section, we formulate the problem and prove that there is no simple function solution of the perfect bivariate normal test. In addition, we prove that there is no solution of the perfect test that includes all positive dependence cases when the test statistics follow the Ferguson distribution. In Section 3, we discuss three parametric models for searching the least conservative critical boundary curves. Assuming normal distribution, we calculate the least conservative tests under these three models. Some statistical power analyses are presented in Section 4, followed by discussions about potential generalizations in Section 5. Proofs of all the results are included in the online Appendix.

2. Theoretical Results

The intersection hypothesis testing is the basis of multiple test procedures (Henning and Westfall 2015; Tamhane and Gou 2018; Tamhane, Gou, and Dmitrienko 2020; Zhang and Gou 2021). For example, the Hochberg (1988) procedure that controls the familywise error rate (FWER) and Benjamini and Hochberg (1995) procedure that controls the false discovery rate (FDR) are both based on the Simes (1986) test that reject the intersection hypothesis $\bigcap_{i=1}^n H_i$ if and only if at least one $p(i) \leq \alpha/n$ ($1 \leq i \leq n$), where $\{H_i\}_{i=1}^n$ is a set of null hypotheses with corresponding $p$-values $\{p_i\}_{i=1}^n$, and the ordered $p$-values are denoted by $p(1) \leq \cdots \leq p(n)$. The Simes test is a valid $\alpha$-level test when the Simes inequality holds: $\Pr \left( \bigcap_{i=1}^n \{p(i) \leq \alpha/n\} \right) \leq \alpha$. Samuel-Cahn (1996) showed that the Simes inequality holds for two outcomes which follow a bivariate normal distribution with zero or positive correlation coefficient. The upper bound $\alpha$ is not attainable when $0 < \rho < 1$. This result is later generalized to multivariate cases by Sarkar and Chang (1997) and Sarkar (1998). It is worth noting that the condition of multivariate totally positive of order two (MTP2) or positively dependent through stochastic ordering (PDS) in Sarkar and Chang (1997) and Sarkar (1998) is much stricter than the condition of positive correlation (Gou and Tamhane 2018). A multivariate normal distributed random variable whose correlation matrix has all positive elements may not satisfy this condition. For example, the trivariate normal distributed random variable with
correlation matrix \( \{\rho_{ij}\}_{i,j=1,2,3} \), where \( \rho_{11} = \rho_{22} = \rho_{33} = 1, \rho_{12} = \rho_{21} = 0.1, \) and \( \rho_{23} = \rho_{32} = \rho_{31} = \rho_{13} = 0.6 \), does not satisfy the MTP2 condition. In this article, we mainly focus on bivariate distributions, especially the bivariate normal distribution, since the central limit theorem guarantees the normal distribution in the large sample size limit.

The Simes test has been shown to be conservative under certain positive dependence structures. For example, using the bivariate normal Simes test with nominal significance level \( \alpha = 5\% \), the true Type I error rate can be as low as 4.1\%. If the dependence structure and correlation coefficient could be clearly pre-specified, there are various ways to modify the Simes test and make it exact. However, in practice, it is rare to have a complete understanding of the dependence structure between two endpoints in the design stage. It is more likely that people only know two outcomes of interest are positively or negatively correlated. So it is practical to obtain new tests which are less conservative than the Simes test for all possible situations. Furthermore, a perfect test that is always exact is our optimal and ambitious aim. If such a perfect test does not exist and it is impossible to construct a test for all \( \rho \) values, then we realize that some compromises have to be made.

In what follows, we restrict attention to testing the intersection hypothesis \( H_1 \cap H_2 \) of two null elementary hypotheses. We define admissible critical boundary curves to avoid the possibility that less extreme \( p \)-value is rejected but more extreme one is accepted.

**Definition 1 (Admissible critical boundary curve).** Function \( f(u) \), \( u \in [0,1] \) gives an admissible critical boundary curve on \( (p_1,p_2) \in [0,1]^2 \) if \( f(u) \) is a nonincreasing function, and bounded between 0 and 1.

An admissible critical boundary curve divides \( [0,1]^2 \) into a rejection region at the bottom left and an acceptance region at the top right. The property of admissibility guarantees that if \( (p_1,p_2) \) is in the rejection region, then \( (p_1',p_2') \) is also in the rejection region for all pairs of \( p \)-values that satisfy \( p_1' \leq p_1 \) and \( p_2' \leq p_2 \). A symmetric admissible curve is both admissible and symmetric with respect to \( f(u) = u \). Common \( p \)-value combination tests include the Simes (1986) test, Fisher’s (1925) probability test, and Stouffer et al.’s (1949) Z-score test. The corresponding symmetric admissible critical boundary curves of these tests are

- **Simes:** \( f(u) = 1_{[0,a/2]}(u) + \alpha 1_{[a/2,1]}(u) + (a/2) 1_{[1,a]}(u) \),
- **Fisher:** \( f(u) = \exp(-F_{\chi^2}^{-1}(1 - \alpha, 4)/2)/u \),
- **Stouffer:** \( f(u) = \Phi(\sqrt{2} \Phi^{-1}(\alpha) - \Phi^{-1}(u)) \),

where \( 1_A(u) \) is an indicator function having the value 1 for all \( u \)'s in \( A \), \( F_{\chi^2}^{-1}(p,k) \) is the quantile function of the chi-square distribution with \( k \) degrees of freedom, and \( \Phi^{-1} \) is the inverse cumulative distribution function of the standard normal distribution. These curves are shown in Figure 1, where a large \( \alpha = 0.20 \) is applied to demonstrate the differences among these three tests.

The top left part of the critical boundary curve that satisfies \( f(u) \geq u \) is referred to as left curve, and the bottom right part that satisfies \( f(u) \leq u \) is referred to as bottom curve.

The rejection region is divided by \( u = f(u) \) into the left and bottom regions. The probability that falls in the left rejection region is denoted by \( I_f^l \), and that falls in the bottom region is \( I_f^b \). Consequently, the probability of rejection is \( I_f^l + I_f^b \). For a symmetric admissible curve, \( I_f^l = I_f^b \), and we denote both of them by \( I_f \) for simplicity.

We first consider the symmetric admissible curve where \( f \) is a simple function, which is a finite linear combination of indicator functions. In addition, a simple test is defined to be a hypothesis test with a simple function as its symmetric admissible critical boundary curve. The majority of the most commonly used multiple test methods, including the Bonferroni (Dunn 1958), Sidak (1967), and Simes (1986) tests, are examples of simple tests.

**Theorem 1.** For any \( \alpha \)-level simple bivariate normal test that is exact at \( \rho = 1 \), there exists a correlation coefficient \( \rho_0 < 1 \), and the test is either conservative or anti-conservative when \( \rho \in (\rho_0, 1) \).

Theorem 1 states that there is no simple bivariate normal test that is always exact for all \( \rho \in [1-\epsilon, 1] \), where \( \epsilon > 0 \). To put it another way, the Type I error rate of this test is either strictly less than or strictly greater than \( \alpha \), but does not equal \( \alpha \) for all \( \rho \in (1-\epsilon, 1) \). Furthermore, we call a bivariate normal test perfectly exact if the test is exact for all situations we considered.

**Corollary 1.** Is directly followed from Theorem 1, and concludes that the functional equation of \( f \)

\[
\int_{-\infty}^{+\infty} \int_{-\infty}^{\Phi^{-1}(\Phi(x))} \phi_{\rho}(x,y)dydx = \alpha, \tag{1}
\]

\( \rho \in [0,1] \) or \( \rho \in [-1,1] \)

![Figure 1. Critical boundary curves of the Stouffer, Fisher, and Simes tests (\( \alpha = 20\%) \]
Corollary 1. There is no perfect exact simple bivariate normal test for all positive correlations \( \rho \in [0, 1] \) or all correlations \( \rho \in [-1, 1] \).

To measure the distance between a perfect exact test and a non-perfect one, we calculate the average conservativeness using the difference between the significance level \( \alpha \) and the Type I error rate of the test, and compare it with \( \alpha \). We call this measure relative conservativeness, which depends on the distribution of \( \rho \).

**Definition 2 (Relative conservativeness).** We define the relative conservativeness by \( R_f(g) = [\alpha - \int_{-1}^{1} (U_f(\rho) + I_f(\rho)) g(\rho) d\rho] / \alpha \) for a valid \( \alpha \)-level test on \( \rho \in \{\rho : g(\rho) > 0\} \), where \( g(\rho) \) is the density function of \( \rho \), and \( f \) is the admissible critical boundary curve of the test. When the type of correlation coefficient need to be specified, the relative conservativeness is denoted by \( R_f(\rho \sim g) \).

The relative conservativeness defined in Definition 2 is served as the target function for searching the least conservative boundary in the corresponding optimization problems. If \( \rho \) is given, the density function \( g(\rho) \) is reduced to a Dirac delta function. For a test with symmetric curve \( f \), the expression of the relative conservativeness can be simplified as \( R_f(g) = [\alpha - \int_{-1}^{1} 2I_f(\rho) g(\rho) d\rho] / \alpha \). The beta distribution and its functions can be applied to describe the distribution of \( \rho \). For example, when using a flat distribution on all positive correlation coefficients, the relative conservativeness \( R_f(U(0, 1)) = [\alpha - \int_{0}^{1} 2I_f(\rho) d\rho] / \alpha \). Similarly, the relative conservativeness \( R_f(U(-1, 1)) = [\alpha - \int_{-1}^{1} I_f(\rho) d\rho] / \alpha \) using a uniform distribution of \( \rho \) on \([-1, 1]\).

Consider the relative conservativeness of the Simes and Bonferroni tests using uniform distributions of \( \rho \). The Simes bivariate normal test is conservative with positive \( \rho \)’s. Numerical calculation shows that \( R_{\text{Simes}}(U(0, 1)) \) is a decreasing function of \( \alpha \). For example, \( R_{\text{Simes}}(U(0, 1); \alpha = 5\%) = 6.67\% \), and \( R_{\text{Simes}}(U(0, 1); \alpha = 1\%) = 5.30\% \). A rough upper bound of \( R_{\text{Simes}}(U(0, 1)) \) is \( 1/(3 \log(1/\alpha)) \). The Bonferroni test is conservative under any dependence. For the Bonferroni bivariate normal test, \( R_{\text{Bonf}}(U(-1, 1)) \) is also a decreasing function of \( \alpha \), with a rough upper bound \( 1/(4 \log(1/\alpha)) \). For instance, \( R_{\text{Bonf}}(U(-1, 1); \alpha = 5\%) = 6.49\% \), and \( R_{\text{Bonf}}(U(-1, 1); \alpha = 1\%) = 4.64\% \).

Corollary 1 shows that there is no simple function admissible critical boundary curve \( f \) that satisfies the functional equation in (1). However, when test statistics are not normally distributed, or correlation coefficients are limited to some subsets on \([-1, 1]\), some solutions may exist.

Ferguson (1995) proposed a class of bivariate uniform distribution models, and a special type of Ferguson’s model is proposed by Gou and Tamhane (2018). We denote this distribution by the FGT distribution.

**Definition 3 (FGT distribution).** Ferguson distribution (Goul-Tamhane type) is a bivariate distribution of \( (U, V) \) on the unit square \([0, 1]^2\), with density function \( p(u, v) = h(u-v)+h(1-(1-(u+v)))/2 \) for \( u, v \in [0, 1] \), where \( h(w) = \frac{1}{2}[\theta(w)/\theta] \) or \( h(w) = \frac{1}{2}[1-\theta,1](w)/(1-\theta) \) for \( \theta \in [0, 1] \).

Ferguson’s model and bivariate normal copula are both bivariate distributions on the unit square with uniform marginal distributions. Their density functions are compared in Figure 2 for the positive dependence case \( \rho = 0.5 \) and the negative dependence case \( \rho = -0.2 \). Generally speaking, Ferguson’s model and other step-function models provide good approximations to normal models for power analysis (Zhang and Gou 2016).

Remark 1. The marginal distribution of the FGT distribution is \( U[0, 1] \), so it is a bivariate copula.

Remark 2. Suppose \( (U, V) \) follows the FGT distribution and let \( \tau = \text{corr}(U, V) \). When \( h(w) = 1_{[0,\theta]}(w)/\theta \equiv h^+(w) \), \( U \) and \( V \) are positively correlated with \( \tau = 1 - 2\theta^2 + \theta^3 \). When \( h(w) = 1_{[1-\theta,1]}(w)/(1-\theta) \equiv h^-(w) \), \( U \) and \( V \) are negatively correlated with \( \tau = -1 + 2\theta^2 - \theta^3 \). Random variables \( U \) and \( V \) are independent if and only if \( \theta = 1 \).

In this article, we denote the correlation coefficient of copula function by \( \tau \), and that of normal distribution by \( \rho \). When calculating the relative conservativeness \( R_f(g) \), we use \( \rho \) for normal distributed random variables, and \( \tau \) for other random variables, if not otherwise specified.

Remark 3. Assume \( (X, Y) \) are standard normal distributed random variables with correlation coefficient \( \rho \). It follows that \( (\Phi(X), \Phi(Y)) \) is a copula. Let \( \tau = \text{corr}(\Phi(X), \Phi(Y)) \). Coefficient \( \tau \) is an increasing function of \( \rho \) that is \( \tau = 12 \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \Phi(x)\Phi(y)\phi(x,y)dydx - 3 \). As some numerical illustrations, \( \tau = 0 \) when \( \rho = 0 \), \( \tau = \pm 1 \) when \( \rho = \pm 1 \), and \( \tau = \pm 0.483 \) when \( \rho = \pm 0.5 \).

The Simes normal test is exact only when \( \rho = 0 \) and \( \rho = \pm 1 \). For the FGT distributed test statistics, the Simes test is exact on a wide range of \( \tau \), and the relative conservativeness for all positive \( \tau \)’s is considerably small.

**Theorem 2.** The \( \alpha \)-level Simes test is exact on \( \tau \in [0, 1 - 8\alpha^2(1 - \alpha)] \cup [1] \) and conservative on \( \tau \in (1 - 8\alpha^2(1 - \alpha), 1) \) when the test statistics follow the FGT distribution. The relative conservativeness on \( \tau \in [0, 1] \) is \( R_{\text{Simes}}(\tau \sim U(0, 1)) = \alpha^2/2 + o(\alpha^2) \).

With FGT distributed test statistics, the Simes test is exact for almost all positive \( \tau \) values. The Simes test is conservative only when \( \tau \) belongs to a narrow interval close to 1, as shown in the left panel of Figure 3. When \( \alpha = 5\% \), the Simes test is conservative only if \( \tau \in (0.9810, 1) \), and the relative conservativeness \( R_{\text{Simes}}(\tau \sim U(0, 1)) = 0.12\% \). If \( \alpha = 1\% \), the Simes test is conservative only when \( \tau \in (0.9992, 1) \), and \( R_{\text{Simes}}(\tau \sim U(0, 1)) = 0.01\% \).

Bonferroni test is independence assumption free, but is only exact when \( \rho = -1 \) for bivariate normally distributed test.
Figure 2. Density function comparison; $\rho = 0.5$ (top panel) and $\rho = -0.2$ (bottom panel), normal model (left panel), and Ferguson’s model (right panel).

Figure 3. Type I error rate versus $\tau$ for FGT distribution (solid curve) and normal distribution (dotted curve). Left panel: Simes test. Right panel: Bonferroni test ($\alpha = 5\%$).
statistics. The multiple test procedures using the Bonferroni technique, including the Holm (1979) procedure and graphical approach (Bretz et al. 2009), share similar properties with the Bonferroni test. However, using the FGT distributed test statistics, the error rate control of the Bonferroni test becomes notably different and summarized in Theorem 3.

**Theorem 3.** The $\alpha$-level Bonferroni test is exact on $\tau \in [-1, -\alpha(1 + \alpha - \alpha^2)]$ and conservative on $\tau \in (-\alpha(1 + \alpha - \alpha^2), 1]$ using the FGT distributed test statistics. The relative conservativeness on $\tau \in [-1, 1]$ is $R_{\text{Bonf}}(\tau \sim U(-1, 1)) = 5\alpha/16 + o(\alpha^2)$.

Under FGT distribution, the Bonferroni test is exact for almost all negative $\tau$’s except a small range near $\tau = 0$, as shown in the right panel of Figure 3. When $\alpha = 5\%$, the Bonferroni test is conservative if $\tau \in (-0.0524, 1]$, and the relative conservativeness $R_{\text{Bonf}}(\tau \sim U(-1, 1)) = 1.56\%$. If $\alpha = 1\%$, the Bonferroni test is conservative when $\tau \in (-0.00101, 1]$, and $R_{\text{Bonf}}(\tau \sim U(-1, 1)) = 0.31\%$.

From Figure 3, we observe that the Simes test is almost perfectly exact for $\tau \in [0, 1]$ under FGT distribution. A natural question arises about whether a perfect exact test exists for all positive $\tau$ values. The following theorem gives a negative answer.

**Theorem 4.** For $\alpha < 1/2$, there is no perfect exact test for positive correlations $\tau \in [0, 1]$ when the test statistics follow FGT distribution.

**Corollary 1** states the nonexistence of simple normal test that is exact for all positive correlation coefficients. Theorem 4 further concludes no perfect exact test for all positive correlation coefficients under FGT distribution. Alternatively, we may consider minimizing the conservativeness in a given range of correlation coefficients despite the nonexistence of perfect exact test. We have observed that the conservativeness of the Simes test, measured by $R_I(U(0, 1))$, is extremely small under FGT distribution, but is not negligible when using normal test statistics. In the following section, we propose three families of intersection hypothesis tests which are less conservative than the Simes test averagely speaking. The relative conservativeness is calculated by following the normal distribution that applies asymptotically to vast types of data.

### 3. Conservativeness Minimization

The symmetric admissible critical boundary curves are modeled with three types of parametric functions on $(p_1, p_2) \in [0, 1]^2$. The bivariate probability density function under normal distribution is

$$
\frac{1}{\sqrt{1 - \rho^2}} \exp \left( -\frac{\rho}{2(1 - \rho^2)} (\rho \Phi^{-1}(p_1)^2 + \rho \Phi^{-1}(p_2)^2 - 2\Phi^{-1}(p_1)\Phi^{-1}(p_2)) \right), \quad (p_1, p_2) \in [0, 1]^2.
$$

When $\rho = \pm 1$, the probability density concentrates on $p_1 = \pm p_2$, as the complete positive (negative) dependence situation (Song 2000; Zhang and Gou 2016).

#### 3.1. Power-Function Model

Power-function model has a power function admissible critical boundary curve, where the left curve is $f(u) = b(c - u)^k + c$ for $u \leq c$, and the bottom curve is $f(u) = c - (1/b)(u - c)^{1/k}$ for $u \geq c$. As a symmetric curve, we only consider the left part.

We first consider two types of power-function models with constraints that (1) tests are exact at $\rho = \pm 1$ and (2) at $\rho = 0$ and 1. Valid $\alpha$-level tests may not exist for all $\rho$ values when applying these two types of models. Details are included in the online Appendix A.2.1. For these anticonservative tests, we need to relax the assumption of having exact tests under independence, and allow these tests to be conservative when $\rho = 0$. For simplicity, we omit the subscript $f$ in $I_1(\rho)$ unless we need to specify the expression of the boundary curve. Assume $I(\rho = 0) = \alpha_0/2 \leq \alpha/2$ and $\Delta \alpha_0 = \alpha - \alpha_0 \geq 0$. The corresponding test with Type I error rate $\alpha_0$ when $\rho = 0$ and with Type I error rate exactly $\alpha$ when $\rho = 1$ is

$$
f(u) = (1 - \alpha) \left( \frac{2k}{k + 1} \right)^k \left( 1 + \frac{\Delta \alpha_0}{\alpha(1 - \alpha)} \right)^{-k} \left( \frac{1 - u}{\alpha} \right)^k + \alpha,
$$

for $k \geq 1 + \Delta \alpha_0/(\alpha(1 - \alpha))$.

Numerical calculation shows that the tests in (2) are valid $\alpha$-level tests for all positive and negative $\rho$ values, as shown in Figure 4. Table 1 lists the Type I error rate under independence $2I(\rho = 0)$, average Type I error rate for positive correlation coefficients $\int_0^1 2I(\rho)d\rho$, and relative conservativeness $R_{\text{pw}}(\rho \sim U(0, 1))$.

**Table 1.** Examples of power-function model based tests ($\alpha = 5\%$).

| $k$     | $2I(\rho = 0)$ | $\int_0^1 2I(\rho)d\rho$ | $\alpha - \int_0^1 2I(\rho)d\rho/\alpha$ |
|---------|----------------|---------------------------|---------------------------------------------|
| $k = 5$ | 0.04402        | 0.04816                   | 3.67%                                       |
| $k = 7.8$ | 0.04780       | 0.04879                   | 2.42%                                       |
| $k = 10$ | 0.04876       | 0.04869                   | 2.62%                                       |
| $k = 20$ | 0.04980       | 0.04802                   | 3.95%                                       |
| $k = 49.5$ | 0.05000       | 0.04729                   | 5.43%                                       |
| $k = +\infty$ (Simes) | 0.05000 | 0.04667 | 6.67% |

**Figure 4.** Type I error rate versus correlation coefficient, power function models ($\alpha = 5\%$).
$U(0, 1) = (\alpha - f^1_0 2I(\rho)d\rho)/\alpha$ of various power-function model based tests, comparing with the Simes test. When $\alpha = 5\%$, the minimum of $R_{pwe}(\rho \sim U(0, 1))$ is 2.42%, achieved at $k = 7.8$.

### 3.2. Linear-Function Model

Linear function provides another three-parameter model for the critical boundary curves with left curve $f(u) = \frac{\pi}{2}[0,a](u) + (1 - (1 - c)(u - a)/(b - a))\frac{\pi}{2}[a,b](u) + c\frac{\pi}{2}[b,c](u)$ and bottom curve $f(u) = b - (a - b)(u - c)/(1 - c), u \in [c, 1]$. To have a valid test for all $\rho$ values, the true significance level at $\rho = 0$ may be set less than $\alpha$. Assuming $a^2/2 \leq I(\rho = 0) = a_0/2 \leq a/2$ and $\Delta a_0 = \alpha - a_0 \geq 0$, the critical boundary curve is

$$f(u) = 1 - (u - [\alpha - d - \Delta a_0/(1 - \alpha)]/2) \cdot (1 - \alpha)/d,$$

where $[\alpha - d - \Delta a_0/(1 - \alpha)]/2 \leq u \leq [\alpha + d - \Delta a_0/(1 - \alpha)]/2$.

We include the Type I error rate under independence, average Type I error rate for positive correlation coefficients and the corresponding relative conservativeness of some linear-function model based tests in Table 2. The relations between the type I error rate and the correlation coefficient under linear function models are presented in Figure 5. The average relative conservativeness of 5%-level test for all positive $\rho$ values can be as small as 3.44% when $d = 0.0117$.

### 3.3. Tangent-Function Model

Trigonometric functions can also be applied to model the critical boundary curves. A three-parameter tangent-function model is left curve $f(u) = b \tan((c - x)/h) + c$ for $u \leq c$, and bottom curve $f(u) = c \tan((x - c)/b)$ for $u \geq c$.

Let $I(\rho = 0) = a_0/2 \leq a/2$ and $\Delta a_0 = \alpha - a_0 \geq 0$. The critical boundary curve is

$$f(u) = b \tan\left(\frac{\alpha - u}{h(b)}\right) + \alpha,$$

where

$$h(b) = \frac{\Delta a_0 + \alpha(1 - \alpha)}{2(1 - \alpha) \arctan\left(\frac{1 - \alpha}{b}\right) - b \log\left(1 + \left(\frac{1 - \alpha}{b}\right)^2\right)},$$

where $u \in [\alpha - h(b) \arctan((1 - \alpha)/b), \alpha]$ and parameter $b$ satisfies $\alpha + b \tan(a/h(b)) \leq 1$ or $\pi h(b)/(2a) \leq 1$. Figure 6 shows how the type I error rate varies with the correlation coefficient using tangent function models. The Type I error rate under independence, average Type I error rate for positive correlation coefficients and the corresponding relative conservativeness of several 5%-level tangent-function model based tests are reported in Table 3. The smallest average relative conservativeness on $\rho \in [0, 1]$ is 0.89%, which is smaller than the minimum values using the power-function and linear function models.

### 4. Power Comparisons

To investigate the statistical powers of proposed intersection hypothesis tests, numerical integration is used. Assume that the $(x, y)$ is bivariate normal with mean $(\mu_x, \mu_y)$ and correlation $\rho$: \((x, y) \sim N\left((\mu_x, \mu_y), \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)\).
Then the joint density function can be written as $f_{\mu}(x, y; 
abla x, \nabla y) = f(x - \mu_x) f(y - \mu_y - \rho(x - \mu_x))/\sqrt{1 - \rho^2}/\sqrt{1 - \rho^2}$.

The probability that falls in the left rejection region of $(p_1 = \Phi(x), p_2 = \Phi(y))$ with critical boundary curve $f$ is:

$$I(p_1, \mu_x, \mu_y) = \int_{-\infty}^{\Phi^{-1}(a)} \int_{-\infty}^{\Phi^{-1}(1)} f(x, y; \mu_x, \mu_y) dy dx + \int_{-\infty}^{\Phi^{-1}(a)} \int_{-\infty}^{\Phi^{-1}(1)} f(x, y; \mu_x, \mu_y) dy dx + \int_{-\infty}^{\Phi^{-1}(a)} \int_{-\infty}^{\Phi^{-1}(1)} f(x, y; \mu_x, \mu_y) dy dx$$

$$= \int_{-\infty}^{\Phi^{-1}(a)} \Phi\left(\frac{\Phi^{-1}(f(x)) - \mu_y - \rho(x - \mu_x)}{\sqrt{1 - \rho^2}}\right) \phi(x - \mu_x) dx$$

where the left critical boundary curve passes $(p_1 = \mu_0, p_2 = 1)$. For a test with symmetric admissible curve, the statistical power is $I(p_1; \mu_x, \mu_y) + I(p_1; \mu_y, \mu_x)$. We further define the average power of a test with symmetric admissible curve $f$ by

$$P_f^\text{avg}(\rho \sim g) = \int_{-1}^{1} \left(I(p_1; \mu_x, \mu_y) + I(p_1; \mu_y, \mu_x)\right) g(\rho) d\rho,$$  

where $g(\rho)$ is a density function of $\rho$ values.

Table 4 compares the average power values $P_f^\text{avg}(\rho \sim U(0, 1))$ of the Simes tests and six newly proposed tests from three model families. Power-function model with $k = 49.5$, linear-function model with $d = 0.00149$ and tangent-function model with $b = 0.00295$ lead to three tests which are less conservative than the Simes test for any $\rho$. The tests based on the power-function model with $k = 7.8$, linear-function model with $d = 0.0117$ and tangent-function model with $b = 0.0206$ achieve smaller relative conservativeness $R_f(U(0, 1))$ than that using the Simes test. The power gains over the Simes test have been observed if $\mu_x$ and $\mu_y$ are not remarkably different. Bold font indicates the highest average power among the listed tests for a given combination of $\mu_x$ and $\mu_y$. Tests based on tangent-function model are more powerful than those based on power-function and linear-function models. Since we use the left reject region for the one-sided hypothesis testing in calculation, the mean vector $(\mu_x, \mu_y)$ under the alternative hypothesis are negative values when computing the power.

Note that the Simes test is actually anticonservative under negative $\rho$ values, and our newly proposed tests control the Type I error rate at level $\alpha$ for any correlation coefficients. For a fair comparison, we may consider the c-Simes test that is a valid $\alpha$-level test for both positive and negative $\rho$ values, proposed by Gou and Tamhane (2018), as a conservative version of the Simes test. For a bivariate test, the c-Simes test compares $p_{12}$ with $\alpha$, and compares $p_{11}$ with $\alpha/2 - \alpha^2/2$. We can show that the 5% level tests based on the power-function model with $k = 49.5$, linear-function model with $d = 0.00149$ and tangent-function model with $b = 0.00295$ are uniformly more powerful than the 5%-level c-Simes test, since they all have larger rejection regions than that of the c-Simes test. Generally speaking, the power-function and tangent function based models show more power gain than the linear-function based models, and therefore are recommended in this simulation setting.

To gauge the significance of a certain increase in power, we compare the sample size needed to achieve a given power for normally distributed observations. When only one hypothesis is involved, the relation between the sample size $m$ and power $1 - \beta$ satisfies $m \propto (\Phi^{-1}(\alpha) + \Phi^{-1}(\beta))^2$, and therefore the differentials of $m$ and $\beta$ satisfy the relation

$$\frac{dm}{m} = \frac{2d\beta}{\Phi(\Phi^{-1}(\beta) - (\Phi^{-1}(\alpha) + \Phi^{-1}(\beta))^2)}.$$  

| $(\mu_x, \mu_y)$ | Simes | c-Simes | Tests with $dI/d\rho = 0$ at $\rho = 0$ | Tangent | Tests with minimum $R_f(U(0, 1))$ |
|----------------|-------|--------|---------------------------------|--------|--------------------------------|
| $(0, 1)$       | 18.13 | 17.58  | 18.25                          | 18.24  | 18.34                          | 18.38 |
| $(0, 2)$       | **52.11** | 51.25  | 52.09                          | 52.12  | 52.06                          | 51.18 |
| $(0, 3)$       | **85.20** | 84.69  | 85.12                          | 85.11  | 85.13                          | 84.12 |
| $(0, 4)$       | **97.95** | 97.83  | 97.92                          | 97.91  | 97.93                          | 97.64 |
| $(1, 1)$       | 27.80 | 27.22  | 28.11                          | 28.01  | 28.35                          | 29.12 |
| $(1, 2)$       | 55.89 | 55.10  | 56.22                          | 56.15  | 56.29                          | 57.06 |
| $(1, 3)$       | 86.02 | 85.55  | 86.09                          | 86.10  | 86.09                          | 85.99 |
| $(1, 4)$       | **98.03** | 97.93  | 98.03                          | 98.03  | 98.03                          | 97.90 |
| $(2, 2)$       | 69.72 | 69.17  | 70.12                          | 69.96  | 70.48                          | 71.48 |
| $(2, 3)$       | 88.99 | 88.63  | 89.19                          | 89.13  | 89.27                          | 89.79 |
| $(2, 4)$       | 98.33 | 98.25  | 98.36                          | 98.36  | 98.38                          | 98.39 |
| $(3, 3)$       | 94.62 | 94.47  | 94.76                          | 94.70  | 94.90                          | 95.21 |
| $(3, 4)$       | 98.91 | 98.86  | 98.94                          | 98.93  | 98.96                          | 99.05 |
| $(4, 4)$       | 99.61 | 99.60  | 99.63                          | 99.62  | 99.69                          | 99.70 |

| Power gain | Min | Max | 0.08 | -0.09 | -0.07 | -1.07 | -1.39 | -0.77 | 1.76 | 1.31 | 2.24 |
| Power gain | Min | Max | 0.03 | 0.02  | 0.04  | -0.57 | -0.88 | -0.17 | 2.31 | 1.93 | 2.80 |
For a fixed differential $d\beta$, the relative sample size reduction \[ |dm|/m \] is an increasing function of power $1 - \beta$ when \[ \beta \leq \Phi\left(-\Phi^{-1}(\alpha)/2 - \sqrt{1 + (\Phi^{-1}(\alpha))^2}/4\right), \] where the upper bound of $\beta$ is 0.318 when $\alpha = 5\%$. For example, the increasing power from 0.50 to 0.51 represents an efficiency gain of $m_{1-\beta=0.51}/m_{1-\beta=0.50} = 1.031$, the increasing power from 0.70 to 0.71 shows an efficiency gain of $m_{1-\beta=0.71}/m_{1-\beta=0.70} = 1.027$, and the increasing power from 0.98 to 0.99 results an efficiency gain of $m_{1-\beta=0.99}/m_{1-\beta=0.98} = 1.153$. The two-hypothesis case is more complex than the one-hypothesis case. Consider testing two one-sided hypotheses \( H_k : \eta_k = 0 \) and \( H_y : \eta_y = 0 \). The test statistic $X$ follows a normal distribution with mean $\eta_k$ and variance $\sigma_k^2$ and test statistic $Y \sim N(\eta_y, \sigma_y^2)$. The correlation between $X$ and $Y$ is denoted by $\rho$. Given significance level $\alpha$, the relation between the sample size $m$ and power $1 - \beta$ is

$$I(\rho; \sqrt{m} \cdot \eta_k/\sigma_k, \sqrt{m} \cdot \eta_y/\sigma_y) + I(\rho; \sqrt{m} \cdot \eta_y/\sigma_y, \sqrt{m} \cdot \eta_k/\sigma_k) = 1 - \beta,$$

where $I(\rho; \mu_k, \mu_y)$ is defined in (5). There is no explicit expression of sample size $m$ as a function of Type II error $\beta$, except a few special cases. For example, using the critical boundary curve $f$ of the Bonferroni test and assuming independence $\rho = 0$ and equal effect size $\eta/\sigma = \eta_k/\sigma_k = \eta_y/\sigma_y$, the sample size $m$ can be written as a function of power $1 - \beta$, which is $m = (\eta/\sigma)^2 \cdot (\Phi^{-1}(\alpha/2) + \Phi^{-1}(\sqrt{\beta}))^2$, and the relative sample size reduction is

$$\frac{dm}{m} = \frac{d\beta}{\beta \cdot \Phi(\Phi^{-1}(\sqrt{\beta})) \cdot (\Phi^{-1}(\alpha/2) + \Phi^{-1}(\sqrt{\beta}))^2}.$$

When $\alpha = 5\%$, the increasing powers from 0.50 to 0.51, from 0.70 to 0.71, and from 0.98 to 0.99 represent efficiency gains of 1.029, 1.025, and 1.142, respectively. In a general case, $|dm|/m$ depends on the critical boundary curve $f$, correlation $\rho$ and effect size $\eta_k/\sigma_k$ and $\eta_y/\sigma_y$. For a given power $1 - \beta$, we calculate the required sample sizes for tests with different critical boundary curves $f$ under various choices of correlations and effect sizes. The tests are compared using the relative sample size defined by a ratio of the sample size of the Simes test to that of a particular test, where the significance level $\alpha = 5\%$, power $1 - \beta = 80\%$, ratios of the effect size in testing $H_k$ to that in testing $H_y$ are 1:1 and 1:2, and correlation coefficients between $X$ and $Y$ are $-0.9$, $-0.5$ (negative dependence), $0$ (independence), $0.5$ and $0.9$ (positive dependence). The results are listed in Table 5. For example, when $(\eta_k/\sigma_k) : (\eta_y/\sigma_y) = 1 : 1$, $\rho = 0.9$, the relative sample size of using the test based on the tangent-function model with parameter $b = 0.0206$ is 1.072, which indicates that a trial design of 536 participants using the Simes test only requires 500 participants if the tangent-function model based least conservative test is applied. A remark here is that the Simes test is an $\alpha$-level test under independence and positive $\rho$ values, and least conservative test proposed in Section 3 are valid for all $\rho$ values. The least conservative tests are more powerful than the Simes test in most cases, even under negative dependence.

### 5. Discussion

We have proved that there is no simple function solution under normal distribution and no solution under FGT distribution of the integral equation about the exact test for all positive correlation coefficients. For bivariate normal distributed test statistics, the conclusion may not be directly generalized from simple function to arbitrary function, because $dI_f(\rho)/d\rho$ does not converge uniformly on $[0,1]$, where \( f_n(u) \) is a pointwise increasing sequence of nonnegative simple functions and $f = \lim_{n \to +\infty} f_n$, and $I_f(\rho)$ is the rejection probability defined in Section 2. However, an extensive numerical search suggests that there is no solution of any form to the equation as listed in (1). Since the test using the FGT distributed statistics also have no solution, we conjecture that there is no perfect exact test for all positive dependence cases regardless of the distribution of test statistics.

If a bivariate perfect exact test does not exist, there is little practical meaning to multivariate perfect exact tests. The reasons are as follows. The intersection hypothesis tests are developed to construct multiple test procedures for each individual hypothesis. For example, the closure principle can be used to design the FWER-control multiple test procedure for $\{H_i\}_{i=1}^n$ based on the hypothesis tests of the intersection hypotheses $\cap_{i \in B} H_i$, where $B$ is a subset of $\{1, \ldots, n\}$ (Marcus, Peritz, and Gabriel 1976; Hochberg and Tamhane 1987; Gou et al. 2014). Without a perfect exact bivariate test of intersection hypothesis, there is no chance to establish a perfect exact multiple test procedure. Nevertheless, searching a least conservative surface of intersection hypothesis tests with more than two elementary hypotheses can be achieved under some parametric models, using similar methods introduced in this article. For example, we may search the least conservative critical boundary surfaces on $(\rho_1, \rho_2, \rho_3) \in [0,1]^3$ for all possible correlation matrices. We can define the symmetric admissible critical boundary surface.

### Table 5. Relative sample sizes of tests based on power-function, linear-function, and tangent-function models, significance level $\alpha = 5\%$, power $1 - \beta = 80\%$.

| $(\eta_k/\sigma_k) : (\eta_y/\sigma_y)$ | $\rho$ | Power ($k = 7.8$) | Linear ($d = 0.0117$) | Tangent ($b = 0.0206$) |
|--------------------------------------|-------|------------------|---------------------|---------------------|
| 1:1                                 | $\rho = -0.9$ | 1.066           | 1.051               | 1.076               |
|                                     | $\rho = -0.5$ | 1.047           | 1.039               | 1.053               |
|                                     | $\rho = 0$    | 1.041           | 1.033               | 1.047               |
|                                     | $\rho = 0.5$  | 1.043           | 1.033               | 1.052               |
|                                     | $\rho = 0.9$  | 1.046           | 1.027               | 1.072               |
| 1:2                                 | $\rho = -0.9$ | 1.062           | 1.050               | 1.068               |
|                                     | $\rho = -0.5$ | 1.039           | 1.034               | 1.042               |
|                                     | $\rho = 0$    | 1.028           | 1.024               | 1.031               |
|                                     | $\rho = 0.5$  | 1.020           | 1.019               | 1.019               |
|                                     | $\rho = 0.9$  | 1.010           | 1.019               | 0.997               |
in a similar way, where the surface is symmetric to planes $p_1 = p_2$, $p_2 = p_3$, and $p_3 = p_1$, as shown in Figure 7. The Simes test for three hypotheses is also plotted in Figure 7 as a reference critical boundary surface. Power gain over the Simes test is also expected for tests with rejection regions defined by least conservative surfaces.

Acknowledgments

We thank the associate editor and two referees for their comments which helped to improve the article.

References

Benjamini, Y., and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," Journal of the Royal Statistical Society, Series B, 57, 289–300. [559]
Bretz, F., Maurer, W., Brannath, W., and Posch, M. (2009), "A Graphical Approach to Sequentially Rejective Multiple Test Procedures," Statistics in Medicine, 28, 586–604. [563]
Cai, G., and Sarkar, S. (2006), "Modified Simes’ Critical Values Under Positive Dependence," Journal of Statistical Planning and Inference, 136, 4129–4146. [559]
Dunn, O. J. (1958), "Estimation of the Means of Dependent Variables," The Annals of Mathematical Statistics, 29, 1095–1111. [560]
Dunnett, C. W., and Tamhane, A. C. (1992), "A Step-Up Multiple Test Procedure," Journal of the American Statistical Association, 87, 162–170. [559]

Ferguson, T. S. (1995), "A Class of Symmetric Bivariate Uniform Distributions," Statistical Papers, 36, 31–40. [561]
Finner, H., Roters, M., and Strassburger, K. (2017), "On the Simes Test Under Dependence," Statistical Papers, 58, 775–789. [559]
Fisher, R. A. (1925), Statistical Methods for Research Workers (1st ed.), Edinburgh: Oliver & Boyd. [560]
Gou, J., and Tamhane, A. C. (2014), "On Generalized Simes Critical Constants," Biometrical Journal, 56, 1035–1054. [559]
Hochberg, Y., and Tamhane, A. C. (2018), "Hochberg Procedure Under Negative Dependence," Statistica Sinica, 28, 339–362. [559,561,565]
Gou, J., Tamhane, A. C., Xi, D., and Rom, D. M. (2014), "A Class of Improved Hybrid Hochberg-Hommel Type Step-Up Multiple Test Procedures," Biometrika, 101, 899–911. [566]
Henning, K. S. S., and Westfall, P. H. (2015), "Closed Testing in Pharmaceutical Research: Historical and Recent Developments," Statistics in Biopharmaceutical Research, 7, 126–147. [559]
Hochberg, Y. (1988), "A Sharper Bonferroni Procedure for Multiple Tests of Significance," Biometrika, 75, 800–802. [559]
Hochberg, Y., and Tamhane, A. C. (1987), Multiple Comparison Procedures, New York: Wiley. [566]
Holm, S. (1979), "A Simple Sequentially Rejective Multiple Test Procedure," Scandinavian Journal of Statistics, 6, 65–70. [563]
Marcus, R., Peritz, E., and Gabriel, K. R. (1976), "On Closed Testing Procedures With Special Reference to Ordered Analysis of Variance," Biometrika, 63, 655–660. [566]
Samuel-Cahn, E. (1996), "Is the Simes Improved Bonferroni Procedure Conservative?" Biometrika, 83, 928–933. [559]
Sarkar, S. K. (1998), "Some Probability Inequalities for Ordered MTP2 Random Variables: A Proof of the Simes Conjecture," The Annals of Statistics, 26, 494–504. [559]
Sarkar, S. K., and Chang, C. K. (1997), "The Simes Method for Multiple Hypothesis Testing With Positively Dependent Test Statistics," Journal of the American Statistical Association, 92, 1601–1608. [559]
Sidak, Z. (1967), "Rectangular Confidence Regions for the Means of Multivariate Normal Distributions," Journal of the American Statistical Association, 62, 626–633. [560]
Simes, R. J. (1986), "An Improved Bonferroni Procedure for Multiple Tests of Significance," Biometrika, 73, 751–754. [559,560]
Song, P. X.-K. (2000), "Multivariate Dispersion Models Generated From Gaussian Copula," Scandinavian Journal of Statistics, 27, 305–320. [563]
Stouffer, S. A., Suchman, E. A., Devinney, L. C., Star, S. A., and Williams, R. M., J. (1949), The American Soldier: Adjustment During Army Life, Studies in Social Psychology in World War II (Vol. 1), Princeton, NJ: Princeton University Press. [560]
Tamhane, A. C., and Gou, J. (2018), "Advances in p-Value Based Multiple Test Procedures," Journal of Biopharmaceutical Statistics, 28, 10–27. [559]
Tamhane, A. C., Gou, J., and Dmitrienko, A. (2020), "Some Drawbacks of the Simes Test in the Group Sequential Setting," Statistics in Biopharmaceutical Research, 12, 390–393. [559]
Zheng, F., and Gou, J. (2016), "A p-Value Model for Theoretical Power Analysis and Its Applications in Multiple Testing Procedures," BMC Medical Research Methodology, 16, 135. [561,563]
——— (2021), "Refined Critical Boundary With Enhanced Statistical Power for Non-Directional Two-Sided Tests in Group Sequential Designs With Multiple Endpoints," Statistical Papers (to appear), DOI: 10.1007/s00362-019-01134-7. [559]