TESTING MULTIPLE HYPOTHESES WITH COMMON EFFECT DIRECTION USING THE CLOSURE METHOD

by

Richard M. Bittman
Joseph P. Romano
Carlos Vallarino
Michael Wolf

October 2006

Department of Statistics
STANFORD UNIVERSITY
Stanford, California 94305-4065
TESTING MULTIPLE HYPOTHESES WITH COMMON EFFECT DIRECTION USING THE CLOSURE METHOD

by

Richard M. Bittman
Bittman Biostat, Inc.
Glencoe, IL

Joseph P. Romano
Department of Statistics
Stanford University

Carlos Vallarino
Takeda Pharmaceuticals
Lincolnshire, IL

Michael Wolf
Inst. For Empirical Research in Economics
University of Zurich

Technical Reports No. 2006-18
October 2006

This research was supported in part by the National Science Foundation grant DMS 010392

Department of Statistics
STANFORD UNIVERSITY
Stanford, California 94305-4065

http://www-stat.stanford.edu
Testing Multiple Hypotheses with Common Effect Direction
Using the Closure Method

Richard M. Bittman
Bittman Biostat, Inc.
766 Grove Street, Suite 600
Glencoe, IL 60022
E-mail: rmb@bittmanbiostat.com

Joseph P. Romano
Department of Statistics
Stanford University
Stanford, CA 94305
E-mail: romano@stanford.edu

Carlos Vallarino
Medical Outcomes Research
Takeda Pharmaceuticals
Lincolnshire, IL 60069
E-mail: cvallarino@tpna.com

Michael Wolf *
Institute for Empirical Research in Economics
University of Zurich
CH-8006 Zurich, Switzerland
E-mail: mwolf@iew.unizh.ch

October 10, 2006

Abstract

Consider the problem of testing s hypotheses simultaneously. Typically, it is known how to construct tests of the individual hypotheses, and the problem is how to combine them into a multiple test procedure that controls the familywise error rate. A very useful approach is the closure method. Due to the generality of the closure method, several largely ad hoc approaches exist. The purpose of this paper is to present a theoretical basis for an approach which is designed to perform well when testing related endpoints. We also emphasize the role of consonant procedures, from an interpretive as well as a theoretical viewpoint, and introduce a new procedure, which is consonant and has a maximin property under the normal model. The results are then applied to PROactive, a clinical trial designed to investigate the effectiveness of a glucose-lowering drug on macrovascular outcomes among patients with type 2 diabetes.

KEY WORDS: Closure Method, Consonance, Familywise Error Rate, Multiple Endpoints,
Multiple Testing, O’Brien’s Method, WLW Method, Permutation Test,
Stepdown Procedure.

*Research partially supported by the Spanish Ministry of Science and Technology and FEDER, grant BMF2003-03324.
1 Introduction

Consider the problem of simultaneously testing $s$ different but possibly related null hypotheses, $H_1, \ldots, H_s$. For example, suppose an experiment is designed to compare a treatment to a control with $s > 1$ related measures of effectiveness, or "endpoints" in the parlance of clinical trials. To fix ideas, consider—as in Davison and Hinkley (1997, Example 4.12)—the role of maintenance chemotherapy for leukemia patients in remission. Suppose that $s = 2$ measures of effectiveness are important: time until relapse and survival time. For either endpoint, a log-rank test (discussed in detail later) can be used to compare the treatment to the control. The main problem we consider in this paper is how to combine these individual tests into a multiple testing procedure. In particular, the overall procedure should be designed to be sensitive or powerful when the endpoints are related and correlated.

A fine balancing act is involved in controlling the familywise error rate (FWE)—the probability of rejecting at least one true null (Hochberg and Tamhane, 1987)—and achieving statistically significant results from every false hypothesis. If the experimenter can confidently project highly statistically significant results from every hypothesis test, the Bonferroni correction to the observed individual p-values may suffice. Effectively, the Bonferroni procedure consists of multiplying each individual p-value by $s$. Any adjusted p-value that is still less than or equal to $\alpha$, the original size of the test, is considered significant. Many alternative strategies are available, however. We offer a brief review of several of these options. Holm (1979) proposed a stepwise testing strategy that is uniformly more powerful than Bonferroni. Order the individual p-values so that $p_1 \leq p_2 \leq \ldots \leq p_s$. Compare the smallest p-value $p_1$ to $\alpha/s$; if it’s less than or equal, then reject the corresponding hypothesis and proceed to compare the second smallest p-value $p_2$ to $\alpha/(s - 1)$. Continue in this manner until the first non-rejection, at which point the comparisons cease and all untested hypotheses are retained. The global null hypothesis $H_{1, \ldots, s}$ stipulates that every $H_i$ is true. Simes (1986) devised the following procedure to test the global null: reject $H_{1, \ldots, s}$ if $p_i \leq i\alpha/s$ for at least one $i$ such that $1 \leq i \leq s$. However, Simes’ test strongly controls the probability of Type I error only under the assumption of independence or certain types of positive dependence (Sarkar and Chang, 1997). A procedure featuring strong control of the FWE—that is, control of the Type I error for any true hypothesis, regardless of whether the others are true—was given by Hommel (1988), who applied the closure method of Marcus, Peritz and Gabriel (1976) to propose the following test: compute $j = \max\{i \in \{1, \ldots, s\} : p_{s-i+k} > k\alpha/i \text{ for } k = 1, \ldots, i\}$. If the set over which the maximum is taken is empty, then reject all $H_i$ for $i = 1, \ldots, s$; otherwise reject all $H_i$ for which $p_i \leq \alpha/j$. This test, however, requires strictly independent test statistics.

The closure method of Marcus et al. (1976) provides the experimenter with a quite general strategy for constructing tests that control the FWE. Consider the general situation of testing
individual hypotheses \( H_1, \ldots, H_s \). Data \( X \) is available, whose distribution is given by a model \( P = \{ P_\theta, \ \theta \in \Omega \} \). The parameter space \( \Omega \) can be parametric, semiparametric or nonparametric, since \( \theta \) merely indexes the parameter space. In general, hypothesis \( H_i \) is specified by a subset \( \omega_i \) of \( \Omega \); that is, \( H_i \) is true if and only if \( \theta \in \omega_i \).

In order to devise a procedure which controls the FWE, the closure method reduces the problem to constructing single tests that control the usual probability of Type 1 error. Specifically, for a subset \( K \subseteq \{1, \ldots, s\} \), let \( H_K \) denote the intersection hypothesis defined by

\[
H_K = \omega_K = \bigcap_{i \in K} \omega_i ;
\]

that is, \( H_K \) is true if and only if \( \theta \in \bigcap_{i \in K} \omega_i \). Of course, \( H_i = H_{\{i\}} \). Suppose \( \phi_K \) is an \( \alpha \) level test of \( H_K \), i.e.

\[
\sup_{\theta \in \omega_K} E_\theta[\phi_K(X)] \leq \alpha .
\]

Then, the decision rule that rejects \( H_i \) if and only if \( H_K \) is rejected for all subsets \( K \) for which \( \{i\} \subseteq K \) strongly controls the FWE. So, in order for \( H_i \) to be deemed significant, every intersection hypothesis which includes \( H_i \) must be deemed significant.

The main problem we study here is the choice of tests of \( H_K \). Even in the case \( s = 2 \), little formal theory exists in the design of tests of \( H_K \). However, closed testing has led to some innovative methods for testing the global null hypothesis, and by consequence the intermediate intersection hypotheses. Some aim at reducing the number of null hypotheses to be tested, which number \( 2^s - 1 \) under a full closed testing algorithm. Westfall and Young (1993) discuss a host of resampling-based methods for incorporating the correlation structure of the observed \( p \)-values to increase the power of the test, such as bootstrapping for the asymptotic setting and permutation methods for finite samples. Westfall and Wolfinger (1997) further develop the concept in the case of discrete distributions, while Westfall et al. (1999) aid the applied statistician by describing how to implement some of these ideas in SAS through PROC MULTTEST and other procedures. Since the graphical depiction of a closed test involves stepping down through the ever smaller intersection hypotheses, some closed tests are also known as stepdown tests. Westfall and Young (1993) introduce a stepdown procedure that, for any given intersection hypothesis, bootstraps the distribution of the corresponding largest test statistic, or minimum \( p \)-value, under their condition of subset pivotality. The condition of subset pivotality can be removed, as shown by van der Laan et al. (2004) and Romano and Wolf (2005); the latter paper develops both finite sample and large sample theory for such stepdown tests that applies to general approaches to resampling, such as the bootstrap, subsampling, permutation and randomization tests. These approaches incorporate the dependence structure of the data and improve on Holm’s method. Stepdown tests based on the maximum test statistic yield
multiple test procedures which satisfy a property called *consonance*. A testing method is *consonant* when the rejection of an intersection hypothesis implies the rejection of at least one of its component hypotheses. An associated concept is that of *coherence*, which states that the non-rejection of an intersection hypothesis implies the non-rejection of any subset hypothesis it implies. Coherence is *de facto* true in any closed testing method. These concepts are more formally defined in Section 3. It is important to note that Sonnemann and Finner (1988) showed that any incoherent procedure can be improved by a coherent one, and so there is no restriction to consider only coherent procedures. Furthermore, Sonnemann (1982) showed that all coherent procedures which control the FWE must be obtained by the closure method. Therefore, our restriction to procedures based on the closure method is no restriction at all. A thorough review of the topic of multiple testing, in particular emphasizing closed tests in a clinical trial setting, can be found in Bauer (1991).

In a clinical setting it is common to specify several hypotheses. Classically these have been characterized as primary (usually one hypothesis or "endpoint") and secondary (several endpoints to be tested if the primary endpoint is significant). More frequently now, clinical trials will have multiple, co-primary endpoints, the significance of any of which will be the basis for a claim of efficacy. Therefore, for both scientific and regulatory reasons, the overall Type I error is to be controlled. It may occur that only one or two endpoints will be statistically significant, but it may also be reasonable to expect that each co-primary endpoint will exhibit an effect of treatment, possibly some to a greater degree than others. This is the common effect direction alluded to in the title of this paper.

The desire to focus power in such a specific direction led O'Brien (1984) to combine multiple test statistics into a single hypothesis test. Under a normal model assumption, O'Brien derived an ordinary least squares (OLS) test statistic and a generalized least squares (GLS) test statistic that are more powerful than Hotelling's $T^2$ statistic in the case of positively correlated endpoints. Lehmacher et al. (1991) point out that Bonferroni—and by extension, stepdown tests based on the maximum test statistic—is useful for detecting one highly significant difference, or treatment effect, among a group of otherwise barely or non-significant differences; while O'Brien's tests, based on the sum, succeed in rejecting the global null against alternatives closer to the diagonal, meaning a group of similar treatment effects, none of which may achieve significance.

While O'Brien's tests can be applied directly in the ANOVA setting, that is, when detecting differences among normal means, it was Pocock et al. (1987) who broadened the concept to a general situation of asymptotically normal test statistics. Wei et al. (1989) further extended the concept into the multivariate survival analysis and censored data setting by introducing an approach in which the treatment effect is again measured by $s$ distinct endpoints and
the marginal distributions are fit in a standard way, such as the Cox (1972) proportional hazards regression model. This technique is commonly called the WLW method, after Wei and coauthors Lin and Weissfeld. Then inference across endpoints is accomplished using a "sandwich" estimate of the covariance matrix of the parameter estimates, a concept found in Lin and Wei (1989) and traced to Liang and Zeger (1986). Briefly, this estimator corrects for the fact that the inverted second derivative of the log-likelihood function, call it $V$, does not yield a valid covariance matrix in the case of model misspecification, such as excluding an important covariate or departures from the assumed covariance structure (Lachin, 2000), as in the case of correlated endpoints (Lin, 1994). The formula consists of inserting, or "sandwiching", the term $R'R$, where $R$ is the matrix of score residuals (Cain and Lange, 1984), to obtain a more robust, indeed consistent estimate of the form $V(R'R)V$. This approach specializes to either the Wei and Lachin (1984) test, a survival analog of the Hotelling test, with similarly low power, or more powerful single degree of freedom tests analogous to the O'Brien (1984) approach.

Nowadays, the WLW model can be easily fitted by at least two statistical packages, including SAS (2004, ch. 54) Version 9. However, fitting this model with an O'Brien-like approach to multivariate inference has been cited infrequently in the literature. We hope to illustrate its potential for greater use and provide a theoretical basis for it. Our objective then is to develop tests which will direct more power toward the detection of alternatives that involve relatively similar effects and which take advantage of their inherent correlation structure. The underlying theory is generally applicable in a multiple testing situation, but it was especially motivated by a problem involving censored data, analyzed later in this paper.

Despite the largely ad hoc approaches to construction of multiple tests, we wish to develop methods that are both reliable in control of the FWE and which are designed to have good power. In order to study the problem more formally, Section 2 introduces a maximin sum test under the normal model and certain restrictions on the parameter space. The normal model, while restrictive, is detailed for two main reasons. First, even in such a setting, very little optimality theory is known for the multiple testing problem outside tests that are one-sided; see Lehmann, Romano and Shafer (2005). Second, it is common for the individual test statistics to be jointly asymptotically normal. Section 3 presents a consonant maximin sum test under similar assumptions. Section 4 applies these concepts to data from the PROactive clinical trial. The Appendix gives proofs of the results.
2 Rationale for the Sum Test

2.1 A Stylized Version of the Problem

In this section, we consider a stylized version of the problem. The parametric structure we now assume is an asymptotic approximation to the more general nonparametric framework. Think of $X_i$ as denoting a test statistic for the $i$th hypothesis, and assume $(X_1, \ldots, X_s)$ is multivariate normal with $X_i \sim N(\theta_i, 1)$ and known covariance matrix $\Sigma$. Let $\theta = (\theta_1, \ldots, \theta_s)$. For testing one-sided alternatives in this parametric model, the parameter space is given by

$$\Omega = \{ \theta : \bigcap_{i=1}^{s} \{ \theta_i \geq 0 \} \} .$$

(2)

However, we will also consider two-sided alternatives, but with the restriction that alternatives $(\theta_1, \ldots, \theta_s)$ are such that all $\theta_i$ have the same sign (possibly negative); that is, we will also consider the larger parameter space

$$\Omega' = \{ \theta : \bigcap_{i=1}^{s} \{ \theta_i \geq 0 \} \} \cup \{ \theta : \bigcap_{i=1}^{s} \{ \theta_i \leq 0 \} \} .$$

(3)

We will apply the closure method to this parametric setup, and derive some optimal max-min tests for this problem. The results are largely based on classical methodology and, at the risk of being pedantic, we include several propositions that clearly state the optimal tests under suitable conditions.

2.2 One-sided Testing in the Case $s = 2$

Consider testing the null hypotheses $H_i : \theta_i = 0$ against the one-sided alternatives $H_i' : \theta_i > 0$, $i = 1, 2$. In this subsection and the next, we continue to assume $Var(X_i) = 1$ and we let $\rho$ denote the correlation between $X_1$ and $X_2$. With minor changes, the results in this section apply to composite null hypotheses of the form $H_i : \theta_i \leq 0$, but we will focus on the simple null hypotheses case for concreteness.

For the multiple testing problem, we wish to apply the closure method. The closure method states that we can construct a decision rule for each of the hypotheses while maintaining control of the FWE, in the following manner.

First, for testing $H_i$ individually, the natural choice of test rejects $H_i$ if $X_i > z_{1-\alpha}$, where $z_{\alpha}$ is the $\alpha$ quantile of the standard normal distribution. In fact, this test is uniformly most powerful level $\alpha$, even in the presence of the nuisance parameter $\theta_2$.

**Proposition 2.1** For the above model, the test that rejects $H_i$ if $X_i > z_{1-\alpha}$ is uniformly most powerful (UMP) at level $\alpha$. 
PROOF. All proofs are deferred to the appendix.

Because of Proposition 2.1, we will use the UMP tests for testing each of the individual hypotheses $H_i$. Now, we need to decide upon a test for the intersection hypothesis $H : \theta_1 = \theta_2 = 0$ versus the alternative hypothesis that at least one $\theta_i$ is $> 0$. So, we now focus on the problem of testing the single null (intersection) hypothesis $H : \theta_1 = \theta_2 = 0$ against the alternative that at least one $\theta_i$ is positive. Here are some facts.

**Proposition 2.2** For testing $(\theta_1, \theta_2) = (0, 0)$ against the fixed alternative $(\theta'_1, \theta'_2)$, the most powerful test rejects for large values of

$$(\theta'_1 - \rho \theta'_2)X_1 + (\theta'_2 - \rho \theta'_1)X_2.$$  \hspace{1cm} (4)

In particular, no UMP test exists against all alternatives in $\Omega \setminus (0, 0)$. However, if we restrict attention in the alternative hypothesis parameter space to $(\theta_1, \theta_2) \in \Omega$ such that $\theta_1 = \theta_2$, then a UMP level $\alpha$ test exists and it rejects when $X_1 + X_2 > z_{1-\alpha}(2 + 2\rho)^{1/2}$.

In fact, for the whole alternative parameter space $\Omega$, a uniformly most powerful unbiased test does not exist; see Lehmann and Romano (2005), Problem 5.21 in the case $\rho = 0$.

Notice that the sum test, i.e. the test that rejects for large $X_1 + X_2$ is UMP against alternatives $\theta$ such that $\theta_1 = \theta_2$ regardless of the value of $\rho$; however, the critical value for this test depends on the value of $\rho$. While this justifies the sum test, it strictly holds only for a one-dimensional subset of the alternative parameter space. On the other hand, if there exist positive effects at all, then one might expect such a test would perform fairly well as long as the $\theta_i$ do not differ too much.

**Remark 2.1** It is interesting to note that one of the coefficients of $X_i$ in the optimal test statistic (4) can be negative, even when $\rho$, $\theta'_1$ and $\theta'_2$ are all positive. For example, if $\theta'_1 = 2$, $\theta'_2 = 8$ and $\rho = 1/2$, then the optimal test statistic is $-2X_1 + 7X_2$. It is perhaps surprising that this test is not monotone in each of the $X_i$. That is, decreasing $X_1$ increases the value of the test statistic and the test rejects with probability tending to one as $X_1 \to -\infty$. Similarly, Pocock et al. (1987) observed that a test statistic based on a linear combination of sample means, derived from O'Brien's (1984) generalized least squares estimate, could have negative coefficients and found it "unteenable from a practical viewpoint".

We now obtain an alternative, more formal justification of the sum test. We find the test that maximizes the minimum power over a larger parameter space, i.e., a maximin test. Because the minimum power would be $\alpha$ over the entire alternative parameter space, it is necessary and customary to remove values of $\theta$ near the origin; see Chapter 8 of Lehmann and
Romano (2005). So, for any \( \epsilon > 0 \), consider the part of the alternative parameter space where both parameters are at least \( \epsilon \), that is the subset \( \omega_1(\epsilon) \) of \( \Omega \) defined by

\[
\omega_1(\epsilon) = \{ \theta : \bigcap_i \{ \theta_i \geq \epsilon \} \} .
\]  

(5)

**Proposition 2.3** For testing \( \theta = (0, 0) \) against \( \theta \in \omega_1(\epsilon) \), the test that rejects when \( X_1 + X_2 > z_{1-\alpha}(2 + 2\rho)^{1/2} \) is maximin level \( \alpha \), for any \( \epsilon \).

In fact, the proof shows that this test continues to be maximin over the larger region \( \omega_2(2\epsilon) \), where \( \omega_2(\epsilon) \) is defined by

\[
\omega_2(\epsilon) = \{ \theta : \theta_1 + \theta_2 \geq \epsilon \} .
\]

Furthermore, the test is maximin over any region which is a subset of \( \omega_2(2\epsilon) \) and contains the point \((\epsilon, \epsilon)\). Indeed, the least favorable distribution concentrates on the point \((\epsilon, \epsilon)\) in all these cases. So, for example, the stated test is also maximin over the region defined by \((\theta_1, \theta_2)\) with \( \theta_i > 0 \) and \( \theta_1 \theta_2 \geq \epsilon^2 \).

On the other hand, if one were most concerned with testing for any positive effect (where possibly one of the \( \theta_i \) is positive and the other zero), one might also consider finding the maximin test over the region \( \omega_3(\epsilon) \) defined by

\[
\omega_3(\epsilon) = \{ \theta : \theta_i \geq \epsilon \text{ for at least one } i \} .
\]

While this set of alternatives may not be the most relevant when we expect the effects to be similar, we study this case for completeness. For this problem, we expect the least favorable distribution to assign equal weight to the two points

\[
\omega_4(\epsilon) = \{(\epsilon, 0), (0, \epsilon)\} .
\]

**Proposition 2.4** (i) Assume \( |\rho| \neq 1 \). For testing \( \theta = (0, 0) \) against the two point set \( \omega_4(\epsilon) \), the test that rejects for large values of

\[
T = T(X_1, X_2) = \exp\left\{ \frac{\epsilon}{1-\rho^2} [X_1 - \rho X_2] \right\} + \exp\left\{ \frac{\epsilon}{1-\rho^2} [X_2 - \rho X_1] \right\}
\]

is maximin.

(ii) For the same problem with \( \rho = 1 \), the test that rejects when \( |X_1 - X_2| > 0 \) is maximin. When, \( \rho = -1 \), the test that rejects when \( X_1 + X_2 > 0 \) is maximin.

(iii). For any \(-1 < \rho \leq 0\), the test that rejects for large \( T \) given by (6) is maximin over \( \omega_3(\epsilon) \).

(iv). For \( 0 < \rho < 1 \), the test that rejects for large \( T \) given by (6) is not monotone in \( X_i \), i.e., it will reject for large negative and large positive values of either \( X_i \).
(v). For $|\rho| < 1$, the test that rejects for large values of $T$ given by (6) is approximately, as $\epsilon \to 0$, equivalent to the test that rejects for large values of $X_1 + X_2$, i.e. the test that rejects for large $X_1 + X_2$ is locally maxmin.

(vi). For $|\rho| < 1$, the test that rejects for large values of $T$ given by (6) is approximately, as $\epsilon \to \infty$, equivalent to the test that rejects for large values of $\max(X_1 - \rho X_2, X_2 - \rho X_1)$.

**Remark 2.2** Note that the most powerful test for testing $(0, 0)$ against $(\epsilon, 0)$ is given by Proposition 2.2 and rejects for large values of $X_1 - \rho X_2$. Similarly, the most powerful test against the alternative $(0, \epsilon)$ rejects for large values of $X_2 - \rho X_1$. The maximin test against the two point alternative $\omega_4(\epsilon)$ combines these two test statistics. Indeed, $T^{1/p}$ is actually the usual $\ell_p$ norm of the vector $(\exp(X_1 - \rho X_2), \exp(X_2 - \rho X_1))$, where $p = \epsilon/(1 - \rho^2)$. Statement (vi) of the proposition merely recalls the $\ell_\infty$ approximation to $\ell_p$ for large $p$.

### 2.3 Restricted Two-sided Testing in the Case $s = 2$

Now we consider testing null hypotheses $H_i : \theta_i = 0$ against two-sided alternatives, with $(\theta_1, \theta_2) \in \Omega'$, where $\Omega'$ was defined in (3).

**Proposition 2.5** Assume $\theta \in \Omega'$. For testing the null hypothesis $H_i : \theta_i = 0$, the test that rejects $H_i$ if $|X_i| > z_{1-\frac{\alpha}{2}}$ is uniformly most powerful unbiased (UMPU) at level $\alpha$.

Because of Proposition 2.5, we will use the UMPU tests for testing each of the individual hypotheses $H_i$. As before, in order to apply the closure method, we now need to decide upon a test for the intersection hypothesis $H : \theta_1 = \theta_2 = 0$ versus the alternative hypothesis that $(\theta_1, \theta_2) \in \Omega'/(0,0)$. So, we now focus on the problem of testing the single null hypothesis $H : \theta_1 = \theta_2 = 0$ against alternatives such that $\theta_1 \theta_2$ is positive.

First, we consider the restricted class of alternatives where $\theta_1 = \theta_2$, but now allow that both may be negative or both may be positive. Because of the two-sided nature, no UMP test exists; however, a UMPU test does exist.

**Proposition 2.6** Assume $|\rho| < 1$. For testing $(\theta_1, \theta_2) = (0, 0)$ against the restricted class of alternatives where $\theta_1 = \theta_2$, where $\theta_1$ can be any nonzero real number, then a UMPU level $\alpha$ test exists and it rejects when $|X_1 + X_2| > z_{1-\frac{\alpha}{2}}(2 + 2\rho)^{1/2}$.

Notice that this test is UMPU regardless of the value of $\rho$; however, the critical value for this test depends on the value of $\rho$. While this justifies the (absolute) sum test, it strictly holds only for a one-dimensional subset of the alternative parameter space. On the other hand, if there exist effects at all, then one might expect such a test to perform fairly well as long as the $\theta_i$ do not differ too much.
We now obtain an alternative, more formal justification of the (absolute) sum test. As before, we find the test that maximizes the minimum power over a larger parameter space, i.e., a maximin test. For any $\epsilon > 0$, consider the subset $\omega'_1(\epsilon)$ of $\Omega'$ defined by

$$\omega'_1(\epsilon) = \{\theta : \theta_i \geq \epsilon, i = 1, 2\} \bigcup \{\theta : \theta_i \leq -\epsilon, i = 1, 2\}.$$  \hfill (7)

So, $\omega'_1(\epsilon)$ considers the part of the alternative parameter space where the absolute size of each of the $\theta_i$ is at least $\epsilon$ (and of the same sign).

**Proposition 2.7** For testing $\theta = (0, 0)$ against $\theta \in \omega'_1(\epsilon)$, the test that rejects when $|X_1 + X_2| > z_{1-\frac{\alpha}{2}}(2 + 2\rho)^{1/2}$ is maximin level $\alpha$, for any $\epsilon$.

As in the one-sided case, the argument shows that the same test continues to be maximin against other regions, such as

$$\omega'_2(\epsilon) = \{\theta \in \Omega' : |\theta_1 + \theta_2| \geq \epsilon\}.$$

As expected, the least favorable distribution now concentrates on the two points $\{(\epsilon, \epsilon), (-\epsilon, -\epsilon)\}$. As another example, the stated test is now maximin over the set of $\theta \in \Omega'$ such that $\theta_1\theta_2 \geq \epsilon^2$.

### 2.4 One-sided Testing: General $s$

Assume $(X_1, \ldots, X_s)$ is multivariate normal with mean vector $(\theta_1, \ldots, \theta_s)$ and known covariance matrix $\Sigma$. Assume $\theta \in \Omega$, with $\Omega$ given by (2). We also assume (without loss of generality) that $\text{Var}(X_i) = 1$.

For testing $H_i : \theta_i = 0$ against $\theta_i > 0$, Proposition 2.1 applies verbatim for general $s$, i.e., the test that rejects $H_i$ if $X_i > z_{1-\alpha}$ is UMP level $\alpha$.

Next, we consider tests of the intersection null hypothesis, which states that $\theta_i = 0$ for all $i$.

**Proposition 2.8** Consider the multivariate location model with mean vector $\theta \in \Omega$ and known nonsingular covariance matrix $\Sigma$, where the parameter space $\Omega$ is given by (2).

(i) For testing $\theta_i = 0$ for all $i$ against the fixed alternative $(\theta'_1, \ldots, \theta'_s)$, the most powerful test rejects for large values of $(\theta')^T \Sigma^{-1} X$, where $X$ is a column vector with transpose $X^T = (X_1, \ldots, X_s)$ and $\theta'$ is a column vector with transpose $(\theta')^T = (\theta'_1, \ldots, \theta'_s)$. In particular, no UMP test exists.

(ii) For testing $\theta_i = 0$ for all $i$ against alternatives $(\theta'_1, \ldots, \theta'_s)$ such that all $\theta'_i$ are equal, a UMP test exists and rejects for large values of the sum of the components of $\Sigma^{-1} X$.

(iii) Assume $\Sigma$ has diagonal elements 1 and off-diagonal elements $\rho$. If we restrict attention in the alternative hypothesis parameter space to $\theta \in \Omega$ such that $\theta_1 = \theta_2 = \cdots = \theta_s$, then a UMP level $\alpha$ test exists and rejects the intersection hypothesis that all $\theta_i = 0$ when $\sum_i X_i > z_{1-\alpha} [1 + s(s-1)\rho^{1/2}]$. 

10
Thus, rejecting the intersection hypothesis for large values of the sum $\sum_i X_i$ is UMP, but only for a very restricted alternative parameter space, and under a strong assumption on $\Sigma$. We now obtain a maximin result that applies to a much larger alternative parameter space. In particular, consider $\omega_1(\theta)$ defined in (5) (except now with the understanding there are $s$ $\theta_i$s).

**Proposition 2.9** Assume $\Sigma$ has diagonal elements $1$ and off-diagonal elements $\rho$. For testing $\theta = (0, 0, \ldots, 0)$ against $\theta \in \omega_1(\theta)$, the test that rejects when $\sum X_i > z_{1-\alpha}[1 + s(s-1)\rho]^{1/2}$ is maximin.

**Remark 2.3** The covariance structure of Proposition 2.9, known as "compound symmetry", is a tractable correlation model that is used in a number of practical situations, e.g. repeated measures ANOVA. Unfortunately, if $\Sigma$ has a different structure, the linear combination $1'\Sigma^{-1}X$ of the $X_i$ (i.e. a sum weighted by the—generally unequal—column totals of $\Sigma^{-1}$) is maximin. Note the similarity of this test statistic, derived here by testing and maximizing power, to O'Brien's (1984, p. 1082) best linear unbiased estimate of the common mean of possibly correlated random variables.

Finally, for two-sided alternatives, an analogous result holds for the test that rejects for large values of $|\sum X_i|$.

### 3 Optimal Consenont Tests

Consider the general situation of testing individual hypotheses $H_1, \ldots, H_s$ based on data $X$. In general, hypothesis $H_i$ is specified by a subset $\omega_i$ of $\Omega$; that is, $H_i$ is true if and only if $\theta \in \omega_i$. Recall the general closure method, based on level $\alpha$ tests of the intersection hypothesis $H_K$ defined in (1). Then, the decision rule that rejects $H_i$ if and only if $H_K$ is rejected for all subsets $K$ for which $\{i\} \subseteq K$ strongly controls the FWE. So, in order for $H_i$ to be deemed significant, every intersection hypothesis which includes $H_i$ must be deemed significant.

The closure method always guarantees that the resulting decision rule is coherent; that is, if $H_f$ implies $H_K$ in the sense that $\omega_f \subset \omega_K$, and $H_f$ is not rejected, then $H_K$ is not rejected. (We may only require this to hold for subsets $I$ and $K$ containing just singletons, but the original hypotheses under test can include intersection hypotheses. If the original family of hypotheses $H_1, \ldots, H_s$ does not include all intersection hypotheses, we may choose to enlarge our family, or just acknowledge that testing the additional intersection hypotheses is merely a useful device to test the hypotheses of interest.)

On the other hand, not all methods generated by the closure method are consonant. Recall that consonant methods satisfy that, if $H_K$ is rejected, then some $H_i$ with $i \in K$ is rejected. To quote Hochberg and Tamhane (1987), "nonconsonance does not imply logical contradictions as
noncoherence does. This is because the failure to reject a hypothesis is not usually interpreted as its acceptance.... Thus whereas coherence is an essential requirement, consonance is only a desirable property.”

However, a nonconsonant or dissonant procedure can leave the statistician in a difficult situation when explaining the results of a study. For example, consider a randomized experiment for testing the efficacy of a drug versus a placebo with two primary endpoints: testing for reduction in headaches and testing for reduction in muscle pain. Suppose $H_1$ postulates the drug is no more effective than the placebo for reduction of headaches and $H_2$ postulates the drug is no more effective than the placebo for reduction of muscle pain. If the joint intersection hypothesis $H_{\{1,2\}}$ is rejected, but the statistician cannot reject either of the individual hypotheses, then compelling evidence has not been established to promote a particular drug indication. The net result is that neither hypothesis can be rejected, even though one might conclude that the drug has some beneficial effect. Lack of consonance makes interpretation awkward.

However, we will argue that, not merely from an interpretive viewpoint, but from a mathematical statistics viewpoint, dissonance is undesirable in that it results in decreased ability to reject false null hypotheses.

The main problem we now address is how to choose the tests of the intersection hypotheses. First, let us consider a classic example.

**Example 3.1 (Two-sided normal means problem)** Suppose $X_i$ are independent with $X_i \sim N(\theta_i, 1)$. Let $s = 2$, so there are only two hypotheses and the parameter space $\Omega$ for $\theta = (\theta_1, \theta_2)$ is the entire plane. Hypothesis $H_i$ specifies $\theta_i = 0$ while the alternative specifies $\theta_i \neq 0$. Based on Proposition 2.5, we are content rejecting $H_i$ if $|X_i| > z_{1-\frac{\alpha}{2}}$. All that remains is to choose a test of the joint intersection hypothesis $H_{\{1,2\}}$. There are two well-known choices here.

(i) **The uniformly most powerful invariant test.** Apply the test that rejects $H_{\{1,2\}}$ if and only if $(X_1, X_2)$ falls in the rejection region $R_\alpha$ given by

$$R_\alpha = \{(X_1, X_2) : X_1^2 + X_2^2 > c_2(1 - \alpha)\} ,$$

where $c_d(1 - \alpha)$ denotes the $1 - \alpha$ quantile of the Chi-squared distribution with $d$ degrees of freedom. This test is also maximin and most stringent (see Section 8.6 of Lehmann and Romano (2005)).

(ii) **Stepdown test based on maximum.** Reject $H_{\{1,2\}}$ if and only if

$$\max(|X_1|, |X_2|) > m_2(1 - \alpha) ,$$

(8)
where $m_1(1 - \alpha)$ is the $1 - \alpha$ quantile of the distribution of $\max(|X_1|, \ldots, |X_s|)$ when the $X_i$ are i.i.d. $N(0, 1)$.

In both cases, the closed testing method begins by testing $H_{\{1,2\}}$. If $H_{\{1,2\}}$ is retained, there are no rejections, but if it is rejected, then $H_i$ is rejected if $|X_i| > z_{1-\frac{\alpha}{2}}$. It is easy to see that

$$z_{1-\frac{\alpha}{2}} < m_2(1 - \alpha) < c_2^{1/2}(1 - \alpha).$$

(9)

The rejection region for the maximin test (i) is the outside of a disc centered at the origin of radius $c_2^{1/2}(1 - \alpha)$, while the rejection region for the stepdown test (ii) is the outside of a square centered at the origin and having side length $2m_2(1 - \alpha)$; see Figure 1. Test (ii) is consonant whereas test (i) is not. It is clearly possible, based on test (i), to reject the intersection hypothesis but not reject either $H_1$ or $H_2$. For example, if $\alpha = 0.05$, then $c_2^{1/2}(0.95) = 2.448$; if $X_1 = X_2 = 1.83$, then $X_1^2 + X_2^2 = 6.698 = 2.588^2$, so $H_{\{1,2\}}$ is rejected but neither $X_i$ satisfies $|X_i| > 1.96$. (See Section 3.2 in Chapter 2 of Hochberg and Tamhane (1987).)

Of course, it does not follow that Test (ii) is preferred merely because it is consonant. The point we wish to make immediately is that, Test (i) can be improved if that goal is to make correct decisions about $H_1$ and $H_2$. To appreciate why, based on Test (i), there are points in the rejection region for testing the intersection hypothesis $H_{\{1,2\}}$ that do not allow for rejection of either $H_1$ or $H_2$. By removing such points from the rejection region when testing $H_{\{1,2\}}$, we can instead include other points in the rejection region that satisfy the constraint that the overall rule be consonant, while still maintaining error control. In this example, all that means is that, for our overall test of $H_{\{1,2\}}$, we restrict attention to tests that have a rejection region in the plane which lies entirely in

$$\{(X_1, X_2) : \max(|X_1|, |X_2|) > z_{1-\frac{\alpha}{2}}\}.$$

Any intersection test satisfying this constraint for testing the intersection hypothesis will result in a consonant procedure when applying the closure method.

To see a concrete way to improve upon Test (i), consider a rejection region $R'_\alpha$ of $H_{\{1,2\}}$ of the form

$$R'_\alpha = \{(X_1, X_2) : X_1^2 + X_2^2 > c'_2(1 - \alpha), \max(|X_1|, |X_2|) > z_{1-\frac{\alpha}{2}}\},$$

(10)

where the critical value $c'_2(1 - \alpha)$ is chosen so that

$$P_{0_{0,0}}(R'_\alpha) = \alpha.$$

Clearly, $c'_2(1 - \alpha) < c_2(1 - \alpha)$, and the resulting test is consonant. For an illustration, see Figure 2.
It follows that, for any $i = 1, 2$, with $\theta_i \neq 0$,

$$P_{\theta_1, \theta_2} \{\text{reject } H_i \text{ using } R_\alpha\} < P_{\theta_1, \theta_2} \{\text{reject } H_i \text{ using } R'_\alpha\}.$$ 

The new consonant test has uniformly greater power at detecting a false null hypothesis $H_i$. Similarly, the new procedure has a uniformly greater chance of detecting both hypotheses as false (if both nulls are false) or at least one false hypothesis (if both nulls are false). In summary, imposing consonance not only makes interpretation easier, but it provides better discriminating ability.

### 3.1 Some Optimality Results

The new consonant procedure, while an improvement over the procedure based on the classical maximin test, does not have any clear optimality property at this point. We now pursue the construction of an optimal choice of the intersection test. In general, the constraint that consonance imposes is easy to identify. The problem of choosing the intersection test satisfying the consonance constraint is now addressed in a more formal way.

The following is a modest generalization of the Neyman-Pearson Lemma, where we now impose the added constraint that the rejection region be restricted to a region $R$ of the sample space.

**Lemma 3.1** Suppose $P_0$ and $P_1$ are two probability distributions with densities $p_0$ and $p_1$ with respect to a dominating measure. Restrict attention to tests $\phi = \phi(X)$ that are level $\alpha$, i.e. $E_0[\phi(X)] \leq \alpha$, and such that $\phi(X) = 0$ if $X \in A$, for some fixed region $A$ in the sample space. Let $R = A^c$ be the complement of $A$. Among such tests, a test that maximizes the power against $P_1$ is given by

$$\phi(x) = \begin{cases} 
1 & \text{if } L(x) > C \text{ and } x \in R \\
\gamma & \text{if } L(x) = C \text{ and } x \in R \\
0 & \text{if } L(x) < 0 \text{ or } x \in A,
\end{cases} \quad (11)$$

where $L(x)$ denotes the usual likelihood ratio $p_1(x)/p_0(x)$ and $C$ and $\gamma$ are chosen to meet the level constraint.

One can provide a partial converse as well, but we leave this to the reader. Note that, if the $p_i$ have the same support and the distribution of the likelihood ratio $L(X)$ is continuous under $P_0$, then the most powerful test is unique with probability one (under either $P_i$).

Next, we construct a maximin test by generalizing Theorem 8.1.1 in Lehmann and Romano (2005), except now we have the added constraint that the rejection region must lie in some fixed set $R$. Denote by $\omega$ the null hypothesis parameter space and by $\omega'$ the alternative hypothesis.
parameter space over which it is desired to maximize the minimum power. So, the goal now is to determine the test that maximizes

$$\inf_{\theta \in \omega} E_\theta[\phi(X)]$$

subject to

$$\sup_{\theta \in \omega} E_\theta[\phi(X)] \leq \alpha$$

and to the constraint that the rejection region must lie entirely in a fixed subset $R$. Let \{\(P_\theta, \theta \in \omega \cup \omega'\)\} be a family of probability distributions over a sample space \((\mathcal{X}, \mathcal{A})\) with densities \(p_\theta = dP_\theta/d\mu\) with respect to a \(\sigma\)-finite measure \(\mu\), and suppose that the densities \(p_\theta(x)\) considered as functions of the two variables \((x, \theta)\) are measurable \((A \times B)\) and \((A \times B')\), where \(B\) and \(B'\) are given \(\sigma\)-fields over \(\omega\) and \(\omega'\). By extending Theorem 8.1.1 in Lehmann and Romano (2005), the following result gives conditions under which a solution of a suitable Bayes problem provides a test with the required properties.

**Theorem 3.1** For any distributions \(\Lambda\) and \(\Lambda'\) over \(B\) and \(B'\), for testing

$$h(x) = \int_{\omega} p_\theta(x) d\Lambda(\theta)$$

against

$$h'(x) = \int_{\omega'} p_\theta(x) d\Lambda'(\theta),$$

let \(\varphi_{\Lambda, \Lambda'}\) be the most powerful among level \(\alpha\) tests \(\phi\) that also satisfy \(\phi(x) = 0\) if \(x \in R^c\). Also, let \(\beta_{\Lambda, \Lambda'}\), be its power against the alternative \(h'\). If \(\Lambda\) and \(\Lambda'\) satisfy

$$\sup_{\omega} E_\theta \varphi_{\Lambda, \Lambda'}(X) \leq \alpha,$$

$$\inf_{\omega} E_\theta \varphi_{\Lambda, \Lambda'}(X) = \beta_{\Lambda, \Lambda'},$$

then:

(i) \(\varphi_{\Lambda, \Lambda'}\) maximizes \(\inf_{\omega} E_\theta \phi(X)\) among all level-\(\alpha\) tests \(\phi(\cdot)\) of the hypothesis \(H : \theta \in \omega\) which also satisfy \(\phi(x) = 0\) if \(x \in R^c\), and it is the unique test with this property if it is the unique most powerful level-\(\alpha\) test among tests that accept on \(R^c\) for testing \(h\) against \(h'\).

(ii) The pair of distributions \(\Lambda, \Lambda'\) is least favorable in the sense that for any other pair \(\nu, \nu'\) we have

$$\beta_{\Lambda, \Lambda'} \leq \beta_{\nu, \nu'}.$$
Corollary 3.1 Let $\Lambda, \Lambda'$ be two probability distributions and $C$ a constant such that

$$\varphi_{\Lambda, \Lambda'}(x) = \begin{cases} 1 & \text{if } \int_{\omega} p_\theta(x) \, d\Lambda'(\theta) > C \int_{\omega} p_\theta(x) \, d\Lambda(\theta), \ x \in R \\ \gamma & \text{if } \int_{\omega} p_\theta(x) \, d\Lambda'(\theta) = C \int_{\omega} p_\theta(x) \, d\Lambda(\theta), \ x \in R \\ 0 & \text{if } \int_{\omega} p_\theta(x) \, d\Lambda'(\theta) < C \int_{\omega} p_\theta(x) \, d\Lambda(\theta), \ x \in R^c \end{cases}$$

(13)

is a size-$\alpha$ test for testing that the density of $X$ is $\int_{\omega} p_\theta(x) \, d\Lambda(\theta)$ and such that

$$\Lambda(\omega_0) = \Lambda'(\omega'_0) = 1,$$

(14)

where

$$\omega_0 = \left\{ \theta : \theta \in \omega \text{ and } E_{\theta}' \varphi_{\Lambda, \Lambda'}(X) = \sup_{\theta' \in \omega} E_{\theta'} \varphi_{\Lambda, \Lambda'}(X) \right\}$$

$$\omega'_0 = \left\{ \theta : \theta \in \omega' \text{ and } E_{\theta} \varphi_{\Lambda, \Lambda'}(X) = \inf_{\theta' \in \omega'} E_{\theta'} \varphi_{\Lambda, \Lambda'}(X) \right\}.$$

Then the conclusions of Theorem 3.1 hold.

3.2 Application to Two-sided Testing

Recall the problem of testing two independent normal means with unit variances, previously considered in Example 3.1. Here, $X_i \sim N(\theta_i, 1)$, and $H_i$ specifies $\theta_i = 0$ against two-sided alternatives. In this case the usual justification for the rejection region $R_\alpha$ for testing $H_{1,2}$ is by invariance with respect to orthogonal transformations of $(X_1, X_2)$, which also leads to the rejection region $R_\alpha$ being maximin against alternatives

$$\{(\theta_1, \theta_2) : \theta_1^2 + \theta_2^2 \geq \epsilon^2\}.$$

However, our multiple testing problem does not possess such invariance, since, for example, a hypothesis of interest $H_1$ specifies $\theta_1 = 0$ (and not some orthogonal transformation of $(\theta_1, \theta_2)$).

We now allow $X_1$ and $X_2$ to be correlated with correlation $\rho$. To determine the intersection test, we find the maximin level $\alpha$ consonant test that maximizes the minimum power against the alternatives

$$\omega_5(\epsilon) = \{(\epsilon, 0), (0, \epsilon), (-\epsilon, 0), (0, -\epsilon)\},$$

as well as

$$\omega_6(\epsilon) = \{(\theta_1, \theta_2) : \text{ at least one } |\theta_i| \geq \epsilon \}.$$

First consider $\omega_5(\epsilon)$. Clearly, the least favorable distribution should give the four points in $\omega_5(\epsilon)$ equal weight. Indeed, Corollary 3.1 is applicable and the optimal intersection test has rejection region of the following form. For fixed $\epsilon$ and $\rho$, let

$$f(t) = \exp[\epsilon t/(1 - \rho^2)].$$

16
Then, large values of the likelihood ratio are equivalent to large values of

\[ T(X_1, X_2) = f(X_1 - \rho X_2) + f(X_2 - \rho X_1) + f(\rho X_2 - X_1) + f(\rho X_1 - X_2); \]

consonance then gives the optimal region as

\[ R_\alpha'' = \{(X_1, X_2) : T(X_1, X_2) > t(1 - \alpha), \max(|X_1|, |X_2|) \geq z_{1-\frac{\alpha}{2}} \}, \quad (17) \]

where the constant \( t(1 - \alpha) \) (which depends on \( \rho \) and \( \epsilon \)) is determined so

\[ P_{0,0}\{R_\alpha''\} = \alpha. \]

In the case \( \rho = 0 \), \( R_\alpha'' \) takes the form

\[ \{(X_1, X_2) : \sum_{i=1}^{2} [\exp(\epsilon X_i) + \exp(-\epsilon X_i)] > t(1 - \alpha), \max(|X_1|, |X_2|) \geq z_{1-\frac{\alpha}{2}} \}, \quad (18) \]

Note that if \( \epsilon \) is small and \( \rho = 0 \), then by Taylor series,

\[ T(X_1, X_2) \approx 4 + 2\epsilon^2(X_1^2 + X_2^2), \]

and so the rejection region, for small \( \epsilon \), is approximately that of \( R_\alpha' \) given by (10). On the other hand, for large \( \epsilon \), the test approximately reduces to the stepdown test based on the maximum given by (8).

Next, we consider the larger region \( \omega_6(\epsilon) \) defined in (16), in the case \( \rho = 0 \). The test function which is the indicator of the region (18) is nondecreasing in each of \( |X_i| \). It follows that the power function of the intersection test with this rejection region is an increasing function of \( |\theta_i| \). Therefore, by Lemma 3.4.2 of Lehmann and Romano (2005), this region is also the optimal consonant, maximin level \( \alpha \) test against alternatives \( \omega_6(\epsilon) \).

### 3.3 Application to Restricted One-sided Testing

Recall the problem considered in Subsection 2.2. Here, \( (X_1, X_2) \) is bivariate normal with unit variances, \( E(X_i) = \theta_i \), and the correlation between \( X_1 \) and \( X_2 \) is known to be \( \rho \). The problem is to test the null hypotheses \( H_1 : \theta_i = 0 \) against the one-sided alternatives \( \theta_i > 0 \). In order to apply the closure method, Proposition 2.3 offers a justification for rejecting for large values of \( X_1 + X_2 \) for the joint intersection test. However, this test is not consonant. Therefore, if we are primarily interested in testing \( H_1 \) and \( H_2 \), we can improve upon this test. Specifically, Corollary 3.1 immediately implies that the optimal, maximin, level \( \alpha \) test against the alternative region \( \omega_1(\epsilon) \) given in (5) has rejection region \( S_\alpha \) of the form:

\[ S_\alpha = \{(X_1, X_2) : X_1 + X_2 > s(1 - \alpha), \max(X_i) > z_{1-\alpha}\}, \quad (19) \]
where the constant $s(1 - \alpha)$ is determined so that under $(\theta_1, \theta_2) = (0, 0)$, the region $S_\alpha$ has probability $\alpha$. The rejection region takes a similar form as without the requirement of consonance, except we now have to impose a restriction on the rejection region so that consonance holds. By doing so, the chance of rejecting any false individual null hypothesis increases when applying the closure method. For an illustration, see Figure 3.

### 3.4 Application to Restricted Two-sided Testing

Consider the setup of Section 2.3. To improve upon Proposition 2.7, we can now determine the optimal consonant, maximin, level $\alpha$ test against $\omega_1^c(e)$, the region defined in (7). It is given by

$$\{(X_1, X_2) : |X_1 + X_2| > r(1 - \alpha), \max(|X_i|) > z_{1 - \frac{\alpha}{2}}\},$$

where the constant $r(1 - \alpha)$ is determined so that the region has probability $\alpha$ under $(\theta_1, \theta_2) = (0, 0)$. Again, the optimal region takes the same form as the one without restricting to consonant tests, but just adds the necessary restriction on the rejection region. For an illustration, see Figure 4.

Again consider the scenario where $(X_1, X_2)$ is bivariate normal with unit variances, $E(X_i) = \theta_i$, and the correlation between $X_1$ and $X_2$ is equal to $\rho$. Table 1 shows the critical values $r(0.90)$, $r(0.95)$ and $r(0.99)$ as functions of $\rho$. We note that, by symmetry, $s(1 - \alpha) = r(1 - 2\alpha)$, so some one-sided critical values can also be derived from the table.

The critical values $r(1 - \alpha)$ in Table 1 were approximated by simulation. To see how, fix $\alpha$. Draw $B$ random samples from the bivariate normal distribution with means 0, unit variances, and correlation $\rho$. Call the $b$th such sample $(X_1^b, X_2^b)$. For a range of values $t$, let $\hat{G}(t)$ denote the proportion of samples where $|\sum_{i=1}^2 X_i^b| \geq t$ and $\max_{i \in \{1, 2\}}(|X_i^b|) > z_{1 - \frac{\alpha}{2}}$. The value of $t$ satisfying $\hat{G}(t) = \alpha$ (at least approximately) serves as the approximation to $r(1 - \alpha)$. This algorithm is motivated by the definition of $r(1 - \alpha)$, but it is somewhat cumbersome, since the proper value of $t$ has to be found by ‘trial and error’. Instead, an equivalent algorithm can be applied that leads to a ‘direct’ approximation of $r(1 - \alpha)$ in a single step. Bivariate normal data are generated as described above. Again, call the $b$th such sample $(X_1^b, X_2^b)$. If $\max_{i \in \{1, 2\}}(|X_i^b|) > z_{1 - \frac{\alpha}{2}}$, let $Y(b) = |\sum_{i=1}^2 X_i^b|$; otherwise, let $Y(b) = 0$. Then, $r(1 - \alpha)$ is approximated as the empirical $1 - \alpha$ quantile of the $B$ values $Y(1), \ldots, Y(B)$. The critical values were obtained from $B = 10^6$ simulations. For an illustration of the effect of the correlation $\rho$, see Figure 5.
<table>
<thead>
<tr>
<th>( \rho )</th>
<th>( r(0.90) )</th>
<th>( r(0.95) )</th>
<th>( r(0.99) )</th>
<th>( \rho )</th>
<th>( r(0.90) )</th>
<th>( r(0.95) )</th>
<th>( r(0.99) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.1</td>
<td>1.804</td>
<td>2.075</td>
<td>2.596</td>
<td>0.0</td>
<td>1.982</td>
<td>2.290</td>
<td>2.878</td>
</tr>
<tr>
<td>-0.2</td>
<td>1.617</td>
<td>1.853</td>
<td>2.313</td>
<td>0.1</td>
<td>2.152</td>
<td>2.498</td>
<td>3.153</td>
</tr>
<tr>
<td>-0.3</td>
<td>1.423</td>
<td>1.622</td>
<td>2.031</td>
<td>0.2</td>
<td>2.314</td>
<td>2.700</td>
<td>3.421</td>
</tr>
<tr>
<td>-0.4</td>
<td>1.228</td>
<td>1.392</td>
<td>1.743</td>
<td>0.3</td>
<td>2.467</td>
<td>2.892</td>
<td>3.687</td>
</tr>
<tr>
<td>-0.5</td>
<td>1.024</td>
<td>1.160</td>
<td>1.452</td>
<td>0.4</td>
<td>2.611</td>
<td>3.071</td>
<td>3.943</td>
</tr>
<tr>
<td>-0.6</td>
<td>0.819</td>
<td>0.930</td>
<td>1.161</td>
<td>0.5</td>
<td>2.746</td>
<td>3.240</td>
<td>4.194</td>
</tr>
<tr>
<td>-0.7</td>
<td>0.614</td>
<td>0.696</td>
<td>0.866</td>
<td>0.6</td>
<td>2.872</td>
<td>3.401</td>
<td>4.427</td>
</tr>
<tr>
<td>-0.8</td>
<td>0.409</td>
<td>0.465</td>
<td>0.581</td>
<td>0.7</td>
<td>2.988</td>
<td>3.548</td>
<td>4.634</td>
</tr>
<tr>
<td>-0.9</td>
<td>0.205</td>
<td>0.231</td>
<td>0.291</td>
<td>0.8</td>
<td>3.095</td>
<td>3.683</td>
<td>4.827</td>
</tr>
<tr>
<td>-1.0</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>1.0</td>
<td>3.290</td>
<td>3.920</td>
<td>5.152</td>
</tr>
</tbody>
</table>

Table 1: The critical values \( r(0.90), r(0.95), \) and \( r(0.99) \) as functions of the correlation \( \rho \).

### 3.5 General \( s \)

The results of the previous sections generalize to the case of \( s \) hypotheses. For example, consider the setting of Section 3.4, but now generalized to testing \( s \) parameters. Let \( (X_1, \ldots, X_s) \) be multivariate normal with known covariance matrix \( \Sigma \), and mean vector \((\theta_1, \ldots, \theta_s)\). The parameter space consists of \( \Omega' \) given by (3). Assume \( \Sigma \) has all off-diagonal elements equal to \( \rho \) and diagonal elements equal to one. The test that rejects for large values of \( \sum X_i \) is maximin, but if used for testing the intersection hypothesis that all \( \theta_i = 0 \), application of the closure method does not result in a consonant procedure. To see how closure leads to an improved multiple testing procedure, first test individual hypotheses \( H_i \) by rejecting \( H_i \) if \( |X_i| > z_{1-\frac{\alpha}{2}} \).

For testing the general intersection hypothesis \( H_K \), which specifies \( \theta_i = 0 \) for \( i \in K \), consider the following test with rejection region

\[
R_{\alpha,K} \equiv \{(X_1, \ldots, X_s) : \sum_{i \in K} X_i > r(1 - \alpha, K), \text{ and} \}
\]

at least one \( H_i, i \in K \) is rejected when applying closure to the family \( \{H_i, i \in K\} \),

where the critical value \( r(1 - \alpha, K) \) is determined so that the above region has probability \( \alpha \) when \( \theta_i = 0 \) for all \( i \). Evidently, the critical values \( r(1 - \alpha, K) \) must be determined inductively, so that in order to determine \( r(1 - \alpha, K) \), we first determine \( r(1 - \alpha, K') \) for all \( K' \subset K \).

The test \( H_K \) is maximin among level \( \alpha \) tests which satisfy the consonant constraint that the rejection region \( R_{\alpha,K} \) must lie in \( \bigcup_{K' \subset K} R_{\alpha,K'} \). Critical values may be approximated by simulation similar to the case \( s = 2 \). Note that \( r(1 - \alpha, K) \) does not depend on \( s \) and, because of the compound symmetry assumption on \( \Sigma \), \( r(1 - \alpha, K) \) only depends on \( K \) through \( |K| \).
Application to the PROactive Clinical Trial

To illustrate the concepts developed here, we use data from PROactive (PROspective pioglitA-zone Clinical Trial In macroVascular Events), a randomized, double-blind clinical trial designed to prospectively investigate the effect of an oral glucose-lowering drug on macrovascular outcomes (Dormandy et al., 2005). The study enrolled $n = 5,238$ patients with type 2 diabetes and evidence of macrovascular disease from 321 treatment centers in 19 European countries. Patients were randomly assigned to either pioglitazone treatment or a placebo, and were allowed to remain on whatever other anti-diabetic medication they were taking at the start of the study—except for other agents in pioglitazone's class—as well as specific cardiovascular and lipid-altering medications. The PROactive study aimed to achieve significance in a primary composite endpoint, the time to first occurrence of any of seven events: death, non-fatal MI (myocardial infarction, including silent MI), stroke, major leg amputation, acute coronary syndrome (ACS), cardiac intervention—including coronary artery bypass graft (CABG) or percutaneous coronary intervention (PCI)—and leg revascularization. A second endpoint was also of interest, and consisted of a subset of the primary events: time to first occurrence among death, non-fatal MI (excluding silent MI) and stroke. Two interim analyses were performed using an alpha spending function, following the methods of Lan and DeMets (1983) and O'Brien and Fleming (1979), which reduced the nominal FWE rate available at the end of the study to 0.044 from the original 0.05.

After completion of the three-year study, the log-rank test (described below in (21)) of the primary endpoint yielded a $p$-value of 0.095, obtained based on an asymptotic Chi-squared approximation. The log-rank test of the principal secondary endpoint yielded a corresponding $p$-value of 0.027. More information on the PROactive trial can be found in the website www.proactive-results.com/index.htm. While Dormandy et al. (2005) claimed a significant outcome, critics such as Freemantle (2005) countered that a secondary endpoint cannot be deemed significant absent a significant outcome in the primary endpoint, an assertion supported by Chi (1998) among others. However, if we view both endpoints as corresponding to hypotheses of equal interest, rather than tiered as "primary" and "secondary"—that is, consider the endpoints to be "co-primary"—significance of the second endpoint could be scientifically assessed through a testing strategy that controls the FWE rate. Would the second (rather than secondary) endpoint have attained significance had the testing methods set forth in this article been applied?

We note that at the time of study design the primary endpoint was defined and recognized as clinically relevant, under the assumption that all vascular beds would be equally affected by the disease state. However, the clinical relevance of the secondary endpoint was also apparent. So a post-hoc design that reclassifies the two endpoints as co-primary is not farfetched, and
is in the spirit of Pocock et al. (1987), who delved into the difficulties and pitfalls behind selecting a single primary endpoint, and entertained use of a global test statistic to gauge the common biological effect of endpoints that measure closely related facets of patient response. It also addresses the concern voiced by Lehmacher et al. (1991) that the clinician is interested in not just a significant global test but which endpoints led to its significance.

The closed family of tests for this example is straightforward. Define $H_{1,2}$ to be the global null hypothesis that neither endpoint exhibits a treatment effect. Then $H_{11}$, or simply $H_1$, is the hypothesis that the first endpoint yields no treatment effect, and $H_2$ is the corresponding hypothesis for the second endpoint. Results from tests of $H_1$ and $H_2$ are already available from the log-rank tests, as outlined above. Thus we proceed to test the intersection hypothesis $H_{1,2}$ by applying the methods of this paper, and note that if our test produces a $p$-value less than or equal to 0.044, the second endpoint can be declared significant—not by ignoring the multiplicity problem, but by proper control of the FWE rate.

4.1 Considerations for Censored Data

Survival data are subject to censoring—an inability to determine an event time due to cessation of observation (e.g., due to end of study or withdrawal or loss to follow-up of a subject). Such data are usually represented by a pair of random variables, one a dichotomous indicator of censoring and the other the time of the event or of censoring, as appropriate. The validity of virtually all inferential survival analysis methods depends on censoring being non-informative. That is, the hazard rate of an event among those in a risk set (not censored) should be the same as it would have been in censored subjects, had censoring not removed them from the risk set. In practical terms, the most frequent potential source of informative censoring may be competing risks. For example, death may be considered to change the risk of other events by making those events impossible. This potential deviation from non-informative censoring is frequently avoided by making the competing risk part of the event definition (e.g., death or non-fatal MI as an endpoint rather than just non-fatal MI). Often such an approach to defining events makes better sense clinically as well. In the multivariate setting there is greater potential for informative censoring, since censoring for one variable may be informative about a different variable. In fact, Wei et al. (1989) require the multivariate event vector to be independent of the censoring vector, conditional on the covariates in the Cox model. We assume, as they did, that censoring in the PROactive study occurred independently of events, noting that death is a component of both endpoints mentioned above.
4.2 Asymptotic Analysis

It is clear from the definitions of the now co-primary endpoints that they are highly correlated. Even before results were published, we would have expected a treatment effect—if there were one—to be apparent in both endpoints, although not necessarily a statistically significant outcome in either endpoint. A thorough search of the literature revealed the scarcity of methods available for testing the global null in this survival setting, motivating this paper. If we wish to find a global test statistic that directs a reasonable amount of power in the direction of our perceived alternative hypothesis, clearly in the region $\omega_1^*(e)$ of (7), our maximin sum test of Proposition 2.7 and our consonant maximin sum test of Section 3.4 are the obvious choices. Since our sum tests are an improvement over the Wei and Lachin (1984) method, we only mention that, according to Lin (1994), the latter test corresponds to the “Wald (Sandwich)” global null test in the default output of SAS procedure PHREG when the marginal distribution method of Wei et al. (1989) is used to fit a two-endpoint model in the proportional hazards (PH) setting. Since the log-rank test statistics are available, our first inclination might be to sum them. But the log-rank test’s connection to the PH model and the ease of using SAS to fit the PH model allowed us to explore an equivalent, more user-friendly option.

The log-rank test is an extension to censored data of the exponential ordered scores test (Lawless 1982) for the equality of lifetime distributions. In the case of two treatment groups, let $t_1 < t_2 < \ldots < t_R$ be the $R$ distinct event times in the entire sample. At time $t_i$ let $n_{ji}$ and $d_{ji}$, for $j = 1, 2$, be the size of the risk set and the number of events, respectively, in group $j$. Letting $n_i = n_{1i} + n_{2i}$ and $d_i = d_{1i} + d_{2i}$, the log-rank test statistic is given by

$$U = \sum_{i=1}^{R} \left[ d_{ji} - \frac{n_{1i}d_i}{n_i} \right]$$

with asymptotic variance given by

$$V = \sum_{i=1}^{R} \frac{d_in_{1i}n_{2i}(n_i - d_i)}{n_i^2(n_i - 1)} .$$

The quotient $U^2/V$ is, under the null hypothesis, asymptotically a Chi-squared random variable with 1 degree of freedom. (Lawless (1982) points out that the log-rank test can be formulated as a test of stratified $2 \times 2$ tables, and indeed earlier in its history was called the Mantel-Haenszel test.) In the case of no ties, so $d_{ij} = 1$ for every $i$, Cox (1972) derived the log-rank test as an efficient score test in a PH regression model with a single binary covariate for treatment group (Lachin 2000). This equivalence with the PH model, which holds approximately if there are relatively few ties, allows us instead to sum the parameter estimates in a simple fit of the WLW marginal model with two endpoints, a relatively simple
task in SAS. Notice that this sum can represent an overall treatment effect, as measured by the PH model, and corresponds to the sum of the logs of the hazard ratios. Wei et al. (1989) showed that these parameter estimates, based on the endpoint-specific partial likelihoods, are approximately normal for large sample sizes.

We used SAS to carry out the analysis, beginning with a corroboration of the log-rank test results for each of the two endpoints. To verify the assumption of proportional hazards, we looked at plots of $\log(-\log(\hat{S}(t_i)))$ vs. $\log(t_i)$, where $\hat{S}(\cdot)$ is the estimated survivor function, overlaid by group, for each endpoint. We also generated Schoenfeld (1982) residual plots, after fitting the WLW model, and found no evidence to contradict the PH assumption.

Let $\beta$ be the vector of $s$ parameters in the marginal (WLW) fit of the PH model, of length $s = 2$ in our example. After a first pass through the model, SAS outputs estimates, $\hat{\beta}_1 = -0.101$ and $\hat{\beta}_2 = -0.172$, of the log hazard ratios corresponding to the two endpoints, with respective standard errors, SE($\hat{\beta}_1$) = 0.061 and SE($\hat{\beta}_2$) = 0.078. The standard errors are based on the robust covariance matrix, as is the correlation between the parameter estimates, $\hat{\rho} = 0.74$. Studentizing the parameter estimates as $X_i = \hat{\beta}_i/SE(\hat{\beta}_i)$ yields $X_1 = -1.667$ and $X_2 = -2.202$. Tests of the individual hypotheses $H_1: \beta_1 = 0$ and $H_2: \beta_2 = 0$ can now be carried out using a standard normal table to find the corresponding $p$-values (against two-sided alternatives) of 0.095 and 0.028, nearly identical to those obtained from the log-rank tests. To apply closure, we test the intersection hypothesis $H_{\{1,2\}}: \beta_1 = \beta_2 = 0$ by forming the test statistic, as in Proposition 2.7, $(X_1 + X_2)/(2 + 2\hat{\rho})^{1/2} = -2.073$. The probability of a larger absolute value under the standard normal is 0.038. Since it is below 0.044, the available $\alpha$, we reject the intersection hypothesis and, by the closure principle, claim that the second endpoint indeed had a significant treatment effect, even after accounting for multiple testing.

Notice that in the absence of a statistical package, Proposition 2.7 can be used whenever an estimate of the correlation and a standard normal table are available. But to let SAS do all the work, we can invoke PHREG a second time, this time with a TEST statement. The TEST statement in PROC PHREG tests hypotheses of the form $c'\beta = 0$ where $c$ is a user-specified vector of coefficients, of length $s$ for $s$ endpoints. Because Proposition 2.7 restricts the parameter space to $\omega(e)$, defined in (7), the unknown parameters must be either all non-negative or all non-positive. If we restrict the coefficients $c_i$ to be all positive or all negative, then $c'\beta = 0$ if and only if $\beta_1 = \beta_2 = \ldots = \beta_s = 0$. So we can use SAS to perform our test, $H_{\{1,2\}}: \beta_1 = \beta_2 = 0$, as long as we set $c_i = [SE(\hat{\beta}_i)]^{-1}$ so as to studentize our test statistics (which we recognize are data-dependent estimates, but the substitution is asymptotically permissible). Studentizing in this manner is not done automatically by the TEST statement, so it must be done manually or with macro coding. SAS computes the $p$-value for $X_{12}^2$, where $X_{12} = (X_1 + X_2)/SE(X_1 + X_2)$, under the Chi-squared distribution with
1 degree of freedom. Applying this technique to the PROactive data results in a global test
$p$-value of 0.038. Not coincidentally, the $p$-value is the same as we obtained in our original—
calculator in hand—application of Proposition 2.7, and in fact, $X_{12} = -2.073$.

But what result would be obtained from the PROactive data if we applied the consonant
maximin sum test of Section 3.4? To calculate the critical value for this test we drew
$B = 50,000$ random samples from the bivariate normal $(0, 0, 1, 1)$ distribution with correlation
0.74, the observed correlation between the WLW parameter estimates. Following the approach
at the end of Section 3.4, we generated the approximate quantile $r(1 - 0.044) = 3.700$. A
linearly interpolated value from Table 1, at $\rho = 0.74$, is roughly 3.768, somewhat far from the
value generated through simulation, as these critical values are quite nonlinear in the inputs,
especially in the level.

The sum of the studentized parameter estimates has absolute value 3.869, larger than the
critical value. The studentized second parameter estimate has absolute value 2.202, larger
than 2.014, the $1 - 0.022$ quantile of the standard normal distribution. Hence, we can again
reject the intersection hypothesis and claim significance of the second endpoint. In fact, the
$p$-value for the PROactive test statistic under the consonant maximin sum test was 0.036.
(Note that, by default, SAS output includes the Wald (Sandwich) test, which is equivalent to
the Wei and Lachin (1984) test, and it yielded an observed significance level of 0.088 under
the Chi-squared distribution with 2 degrees of freedom.)

Our conclusion assumes asymptotic normality of parameter estimates from the PH model,
which, given the amount of data, including events, produced in PROactive, should be well
justified. But we would also like to test the intersection hypothesis with a finite sample, per-
mutation test, so as to validate the results of the asymptotic case and illustrate this alternative
method.

4.3 Permutation Analysis

In general, permutation tests offer validity of error control in finite samples without having
to assume a parametric model (Lehmann and Romano 2005, Section 15.2). Although our
specific situation has a large enough sample size to justify the bivariate normality assumptions
of Proposition 2.7 and Section 3.4, this may not always be the case. To carry out the finite
sample, permutation tests analogous to the normal model tests, we found it more convenient to
base them on the log-rank test statistics. The individual one-sided test statistics, $X_1$ and $X_2$,
respectively, for the two PROactive endpoints are studentized log-rank test statistics, defined
as in Example 4.12 of Davison and Hinkley (1997, page 160), and equivalent to $U/V^{1/2}$ in
the notation of (21). Their observed values are $X_1 = -1.670$ and $X_2 = -2.210$, respectively,
almost identical to the studentized WLW parameter estimates of the previous section.
The permutation method randomly permutes treatment and placebo labels. To be more specific, each of the \( n = 5,238 \) patients is labeled either as ‘1’ if assigned to pioglitazone treatment or as ‘0’ if assigned to the placebo. This results in a \( 5,238 \times 1 \) vector of ones and zeros that is used to carry out the test. This vector is then permuted at random while all the other patient data remain fixed. From this pseudo-data set thereby obtained under the intersection null hypothesis of no treatment effects, the two test statistics are recomputed. And then the process is repeated. If we use a total of \( B \) permutation repetitions, we obtain pseudo-statistics \( X_1^*(1), \ldots, X_1^*(B) \) and \( X_2^*(1), \ldots, X_2^*(B) \). According to (4.11) of Davison and Hinkley (1997, page 141), the two-sided individual \( p \)-values are computed as

\[
\hat{p}_1 = \frac{\#\{|X_1^*(b)| \geq 1.670\} + 1}{B + 1} \quad \text{and} \quad \hat{p}_2 = \frac{\#\{|X_2^*(b)| \geq 2.210\} + 1}{B + 1}.
\]

The results, based on \( B = 49,999 \) permutations, are \( \hat{p}_1 = 0.094 \) and \( \hat{p}_2 = 0.027 \), respectively, in close agreement with the asymptotic normal approximation.

The two-sided test statistic for the intersection hypothesis \( H_{\{1,2\}} \) is given by \( |X_1 + X_2| = | -1.670 - 2.210| = 3.880 \). The \( p \)-value analogous to that in Proposition 2.7 is given by

\[
\frac{\#\{|X_1^*(b) + X_2^*(b)| \geq 3.880\} + 1}{B + 1} = 0.037.
\]

Notice the similarity to the \( p \)-value of 0.038 obtained above under the bivariate normal assumption. So our permutation test analog rejects the intersection hypothesis \( H_{\{1,2\}} \) at significance level \( \alpha = 0.044 \) and, together with \( \hat{p}_2 = 0.027 \), provides further evidence for the significance of the second endpoint in a closed test framework.

The improved \( p \)-value according to Section 3.4 is obtained as follows. First, we compute the \( 1 - 0.044 \) empirical quantiles of the pseudo-statistics \( |X_1^*(1)|, \ldots, |X_1^*(B)| \) and \( |X_2^*(1)|, \ldots, |X_2^*(B)| \). (The use of \( 1 - \alpha \) quantiles instead of \( 1 - \alpha/2 \) reflects the distribution of the absolute value.) The results are 2.001 and 2.012, respectively. The optimal consonant \( p \)-value at level \( \alpha = 0.044 \) is then given by

\[
\frac{\#\{|X_1^*(b) + X_2^*(b)| \geq 3.880 \cap \{|X_1^*(b)| > 2.001 \cup |X_2^*(b)| > 2.012\}\} + 1}{B + 1} = 0.036.
\]

Since 0.036 \( \leq 0.044 \), we reject the intersection hypothesis. The fact that \( 2.210 > 2.012 \) shows that the test statistic associated with the second endpoint is larger than the corresponding \( 1 - \alpha \) critical value. (As in the asymptotic case, the test statistic for the first endpoint falls short of its critical value: 1.670 < 2.001.) Hence the permutation test equivalent to the consonant sum test of Section 3.4 also reaches the conclusion that the second endpoint is significant.
5 Conclusion

Clinical research and other investigations may record several endpoints which each measure the effectiveness of a treatment. Indeed, a recent compilation by the US Food and Drug Administration of opportunities for improvement in clinical development states (US FDA, 2006, page L-9): "In many diseases, more than a single efficacy endpoint may be of importance." These are often positively correlated or mechanistically related (and likely have both characteristics), so that a treatment effect in one endpoint will be associated with similar effects in the others—that is, they exhibit common effect direction. Statistical inference for such study designs should determine whether there is a treatment effect overall, and may additionally identify which endpoints exhibit the effect, all the while controlling the FWE of the hypothesis tests that inform these decisions. In general a closed testing procedure accomplishes all of these objectives by testing intersection hypotheses that address overall effect and individual hypotheses that correspond to each endpoint.

Experimenter who measure multiple endpoints may use a global test statistic, like the sum of studentized treatment differences, to assess treatment effect. They also must choose whether to interpret the global test itself as an overall measure of efficacy versus basing efficacy decisions and conclusions on the various separate endpoints. For example, a global statistic may be useful in intermediate stages of experimentation, when statistical evidence of only an overall treatment effect is required (e.g. clinical Phases Ib and II). By contrast, confirmatory studies that establish effect on particular endpoints (e.g. clinical Phase III) usually must test each of them, controlling the FWE rate.

One of the first treatments of this problem was that of O'Brien (1984), who proposed a test based on the sum of the endpoint measures. He and Pocock et al. (1987) viewed the global test statistic as a meaningful endpoint in its own right. Like Pocock et al. (1987) we extended O'Brien's method beyond ANOVA to any setting in which normally distributed test statistics are available. Also, we established a theoretical grounding to the sum test, complementing its intuitive appeal. Specifically, we proved that the test statistic based on the simple (unweighted) sum provides a maximin level α test against alternatives in the direction of a common effect, when the correlation has a simplified structure useful in a number of practical procedures. If the alternatives can be further constrained and the endpoint means are assumed to be all the same, the test is UMP.

Lehmacher et al. (1991) considered the sum test as an engine for testing within a closed test framework. The straightforward use of the sum test is not consonant in this situation, but we defined a new class of rejection regions in which intersection tests based on sum statistics do achieve consonance. Thus, the procedure inherently guarantees that at least one endpoint will
be declared statistically significant if the intersection tests are statistically significant. These new regions exclude the sub-regions that would violate consonance and enlarge the remaining rejection region while still maintaining the level $\alpha$ constraint. The new, consonant test has a maximin property among tests that control the FWE rate.

Development of this research was motivated by the recently completed PROactive clinical trial, in which the outcomes of two important endpoints have been the source of some controversy in the clinical literature. We applied the new procedures to PROactive data and showed that, had the trial been designed as a closed test, with the two endpoints as co-primary, both the simple sum test and the new consonant sum test would have been able to establish a general treatment effect and identify one of the endpoints as statistically significant.

While the finite sample theoretical justification is based on normally distributed endpoints, the results apply whenever test statistics are asymptotically multivariate normal. In situations where this assumption is suspect, permutation tests may be used to ensure robustness of validity.
A Proofs

PROOF OF 2.1 The result is essentially contained in Example 3.9.2 of Lehmann and Romano (2005). ■

PROOF OF PROPOSITION 2.2. The proof follows immediately from the Neyman Pearson Lemma. ■

PROOF OF PROPOSITION 2.3. Apply Theorem 8.1.1 of Lehmann and Romano (2005) by placing a least favorable distribution over $\omega_1(\epsilon)$ which concentrates on the point $(\epsilon, \epsilon)$. The optimal test for the simple versus simple problem was given in Proposition 2.2 and is the stated maximin test. To show it is indeed maximin, it is enough to observe that the power of this test achieves its lower bound over $\omega_1(\epsilon)$ at $(\epsilon, \epsilon)$. ■

PROOF OF PROPOSITION 2.4. To prove (i), put equal mass on the two points $(\epsilon, 0)$ and $(0, \epsilon)$. For testing $\theta = (0, 0)$ against the simple alternative obtained from the mixture of the two distributions, apply the Neyman Pearson Lemma to conclude that the most powerful test rejects for large values of (6). By symmetry, the power at each of the two points in $\omega_4(\epsilon)$ is the same, and maximinity follows.

To prove (ii), if $\rho = 1$, then $X_1 - X_2$ is a degenerate normal variable equal to 0 under the null and $\epsilon$ or $-\epsilon$ if $\theta \in \omega_4(\epsilon)$. Similarly, if $\rho = -1$, then $X_1 + X_2$ is, with probability one, equal to $\theta_1 + \theta_2$, so rejecting for positive values of $X_1 + X_2$ is perfect.

To prove (iii), it is enough to observe that the test statistic $T$ given in (6) is increasing in each of the components $X_i$. This is seen, for example, by differentiating $T = T(X_1, X_2)$ with respect to $X_i$ and seeing that it is indeed positive if $\rho \leq 0$.

To prove (iv), observe that

$$\lim_{X_i \to \pm \infty} T(X_1, X_2) = +\infty.$$ 

To see (v), expand $T$ in a Taylor series to get

$$T = 2 + \frac{\epsilon}{(1 - \rho^2)(1 - \rho)(X_1 + X_2)} + O(\epsilon^2) \quad \text{as } \epsilon \to 0.$$ 

Hence, the test that rejects for large $T$ is asymptotically equivalent (as $\epsilon \to 0$) to the test that rejects for large values of $X_1 + X_2$. The proof of (vi) follows by taking logs and applying L'Hôpital’s rule. ■

PROOF OF PROPOSITION 2.5. The proof follows by application of Theorem 4.4.1 in Lehmann and Romano (2005). Indeed, for testing $\theta_1 = 0$ against $\theta_1 > 0$, the joint density of $(X_1, X_2)$ can
be written in the canonical form (4.11) of Lehmann and Romano (2005) with $U = X_1 - \rho X_2$
and $T = X_2 - \rho X_1$. A UMPU test rejects for large and small values of $U$, where the critical
values are obtained from the conditional distribution of $U$ given $T$ and are chosen under the
null hypothesis to satisfy the level and unbiasedness constraints. But, the joint distribution of
$(U, T)$ is bivariate normal, and it is easily checked that

$$E_{\theta}(U) = \theta_1 - \rho \theta_2, \quad E_{\theta}(T) = \theta_2 - \rho \theta_1$$

$$Var_{\theta}(U) = Var_{\theta}(T) = 1 - \rho^2$$

and the correlation between $U$ and $T$ is $-\rho$. Therefore, the conditional distribution of $U$
given $T$ is normal with conditional mean

$$\theta_1 - \rho \theta_2 - \rho (T - \theta_2 + \rho \theta_1) = \theta_1 (1 - \rho^2) - \rho T$$

and conditional variance $(1 - \rho^2)^2$. Using this conditional distribution when $\theta_1 = 0$ yields that
the UMPU test rejects for large and small values of

$$(U + \rho T)/(1 - \rho^2) = X_1,$$

i.e., when $|X_1| > z_{1-\frac{\alpha}{2}}$. ■

**Proof of Proposition 2.6.** For this reduced parameter space indexed by $\theta_1$, we have a
one-parameter exponential family. By (4.3) of Lehmann and Romano (2005) with natural
sufficient statistic $T = X_1 + X_2$, the result follows. ■

**Proof of Proposition 2.7.** For the least favorable distribution, put equal mass at $(\epsilon, \epsilon)$ and
$(-\epsilon, -\epsilon)$. For testing the simple null hypothesis that $\theta = (0, 0)$ against the mixture alternative,
the likelihood ratio equals

$$\exp \left[ \frac{X + Y}{1 - \rho} \right] + \exp \left[ -\frac{X + Y}{1 - \rho} \right].$$

This statistic is an even function of $X + Y$ and is monotone increasing in $|X + Y|$. Therefore,
the likelihood ratio test equivalently rejects for large values of $|X + Y|$. To see that the test is
maximin, it is enough to observe that its power against any $\theta \in \omega_1(\epsilon)$ is minimized at the two
points $(\epsilon, \epsilon)$ and $(-\epsilon, -\epsilon)$. ■

**Proof of Proposition 2.8.** The proof is an application of the Neyman-Pearson Lemma. ■

**Proof of Proposition 2.9.** The least favorable distribution concentrates on the single point
$(\epsilon, \ldots, \epsilon)$. Maximinity results because the resulting test against this fixed alternative has an
increasing power function in each of the components $\theta_i$, and therefore the power is minimized over $\omega_1(\epsilon)$ at $(\epsilon, \ldots, \epsilon)$. ■

**Proof of Lemma 3.1** The goal is to maximize $E_{P_\theta}[\phi(X)I\{X \in R\}]$ subject to

$$E_{P_\theta}[\phi(X)I\{X \in R\}] \leq \alpha.$$  

Let $Q_i$ denote the conditional distribution of $X$ given $X \in R$ when $X \sim P_i$. Also, let $\beta_i = P_1\{R\}$. Then, equivalently, the problem is to maximize

$$\frac{1}{\beta_1} E_{Q_1}[\phi(X)]$$

subject to

$$E_{Q_0}[\phi(X)] \leq \alpha / \beta_0,$$

or equivalently maximize $E_{Q_1}[\phi(X)]$ subject to

$$E_{Q_0}[\phi(X)] \leq \alpha' = \alpha / \beta_0.$$  

By the usual Neyman Pearson Lemma, the optimal test rejects for large values of the likelihood ratio $dQ_1(X)/dQ_0(X)$, which is a constant multiple of $L(X)$. ■

**Proof of Theorem 3.1.** (i): If $\varphi^*$ is any other level-\(\alpha\) test of $H$ satisfying $\varphi^*(X) = 0$ if $X \in R^c$, it is also of level $\alpha$ for testing the simple hypothesis that the density of $X$ is $h$; therefore, the power of $\varphi^*$ against $h'$ cannot exceed $\beta_{\Lambda,\Lambda'}$. It follows that

$$\inf_{\omega'} E_{\omega} \varphi^*(X) \leq \int_{\omega'} E_{\omega} \varphi^*(X) d\Lambda'(\theta) \leq \beta_{\Lambda,\Lambda'} = \inf_{\omega'} E_{\omega} \varphi_{\Lambda,\Lambda'}(X),$$

and the second inequality is strict if $\varphi_{\Lambda,\Lambda'}$ is unique.

(ii): Let $\nu$, $\nu'$ be any other distributions over $(\omega, B)$ and $(\omega', B')$, and let

$$g(x) = \int_{\omega} p_\theta(x) d\nu(\theta), \quad g'(x) = \int_{\omega'} p_{\theta}(x) d\nu'(\theta).$$

Since both $\varphi_{\Lambda,\Lambda'}$ and $\varphi_{\nu,\nu'}$ are level-\(\alpha\) tests of the hypothesis that $g(x)$ is the density of $X$, and by definition both satisfy the constraint that their acceptance region must include $R^c$, it follows that

$$\beta_{\nu,\nu'} \geq \int \varphi_{\Lambda,\Lambda'}(x) g'(x) d\mu(x) \geq \inf_{\omega'} E_{\omega} \varphi_{\Lambda,\Lambda'}(X) = \beta_{\Lambda,\Lambda'}.$$  

**Proof of Corollary 3.1.** If $h$, $h'$, and $\beta_{\Lambda,\Lambda'}$ are defined as in Theorem 3.1, then Lemma 3.1 implies that $\varphi_{\Lambda,\Lambda'}$ is a most powerful level-\(\alpha\) test for testing $h$ against $h'$, among tests whose acceptance regions must include $R^c$. But, the assumptions also imply that

$$\sup_{\omega} E_{\omega} \varphi_{\Lambda,\Lambda'}(X) = \int_{\omega} E_{\omega} \varphi_{\Lambda,\Lambda'}(X) d\Lambda(\theta) = \alpha,$$

30
and that

\[ \inf_{\omega'} E_{\theta' \varphi_{A,A'}}(X) = \int_{\omega'} E_{\theta' \varphi_{A,A'}}(X) \, d\Lambda'(\theta) = \beta_{A,A'} \]

Condition (12) is thus satisfied and Theorem 3.1 applies.
References


33


Figure 1: The rejection regions for the two tests of Example 3.1 with nominal level $\alpha = 0.05$. Test (i) rejects for points that fall outside the solid circle with radius 2.448. Test (ii) rejects for points that fall outside the dashed square with length $2 \times 2.234$. For example, the point $(1.83, 1.83)$ leads to a rejection of the intersection hypothesis by Test (i) but not by Test (ii), but when applying closure using Test (i), no individual hypotheses are rejected.
Figure 2: The rejection region of the improved Test(i) of Example 3.1 with nominal level $\alpha = 0.05$. The test rejects for points outside the solid curve. This curve is obtained as the ‘union’ of a circle with radius 2.421 and a square with length $2 \times 1.96$. While the point $(1.83, 1.83)$ leads to a rejection of the intersection hypothesis by Test (i) of Example 3.1, no rejections of individual hypotheses can be made. But, $(2.43, 0)$ leads to rejection of the improved intersection test and rejection of $H_1$. 
Figure 3: The rejection regions for the test of Proposition 2.3 and its improvement of Subsection 3.3 with nominal level \( \alpha = 0.05 \) when the correlation \( \rho = 0 \). The test of Proposition 2.3 rejects for points to the right and above the dashed line with intercept 2.326 and slope \(-1\). The improved test rejects for points to the right and above the solid curve defined by (19) with \( s(0.95) = 1.985 \) and \( z_{1-\alpha} = 1.645 \). For example, the point (1.4, 1.4) leads to a rejection by the test of Proposition 2.3 but not by the improved test. On the other hand, the point (1.9, 0.25) leads to a rejection by the improved test but not by the test of Proposition 2.3.
Figure 4: The rejection regions for the test of Proposition 2.7 and its improvement of Subsection 3.4 with nominal level $\alpha = 0.05$ when the correlation $\rho = 0$. The test of Proposition 2.7 rejects for points outside the dashed band. The improved test rejects for points outside the solid 'band'. For example, the point $(1.6, 1.6)$ leads to a rejection by the test of Proposition 2.7 but not by the improved test. On the other hand, the point $(2.3, 0.2)$ leads to a rejection by the improved test but not by the test of Proposition 2.7.
Figure 5: The effect of the correlation $\rho$ on the rejection region of the improved test of Subsection 3.4 with nominal level $\alpha = 0.05$. When $\rho = 0.5$, the test rejects for points outside the solid 'band'. When $\rho = -0.5$, the test rejects for points outside the dashed 'band'. For example, the point $(-0.5, 2.5)$ leads to a rejection when $\rho = -0.5$ but not when $\rho = 0.5$. 