

# Exact and Approximate Stepdown Methods For Multiple Hypothesis Testing

March 2004

## Abstract

Consider the problem of testing  $k$  hypotheses simultaneously. In this paper, we discuss finite and large sample theory of stepdown methods that provide control of the familywise error rate (FWE). In order to improve upon the Bonferroni method or Holm's (1979) stepdown method, Westfall and Young (1993) make effective use of resampling to construct stepdown methods that implicitly estimate the dependence structure of the test statistics. However, their methods depend on an assumption called subset pivotality. The goal of this paper is to construct general stepdown methods that do not require such an assumption. In order to accomplish this, we take a close look at what makes stepdown procedures work, and a key component is a monotonicity requirement of critical values. By imposing monotonicity on estimated critical values (which is not an assumption on the model but an assumption on the method), it is shown how to construct stepdown tests that can be applied in a stagewise fashion so that at most  $k$  tests need to be computed. Moreover, at each stage, an intersection test that controls the usual probability of a Type 1 error is calculated, which allows us to draw upon an enormous resampling literature as a general means of test construction. In addition, it is possible to carry this out with the same set of resamples (or subsamples) for each of the intersection tests.

KEY WORDS: Bootstrap, Familywise Error Rate, Multiple Testing, Permutation Test, Randomization Test, Stepdown Procedure, Subsampling.

# 1 Introduction

The main point of this paper is to show how computer-intensive methods can be used to construct asymptotically valid tests of multiple hypotheses under very weak conditions. As in the case of single testing, bootstrap and other resampling methods offer viable nonparametric alternatives to constructing tests that do require normality or other parametric assumptions. The treatise by Westfall and Young (1993) takes good advantage of resampling to estimate the joint distributions of multiple test statistics in order to construct valid and more efficient multiple testing methods. However, their methods rely heavily on the assumption of subset pivotality. Thus, the main goal of this paper is to show how to construct valid stepdown methods that do not require this assumption, while still being computationally feasible.

Consider the problem of testing hypotheses  $H_1, \dots, H_k$ . Suppose corresponding  $p$ -values  $\hat{p}_1, \dots, \hat{p}_k$  are available. A starting point for a general method that controls the familywise error rate (FWE) is the Bonferroni method, which rejects any  $H_j$  for which  $\hat{p}_j \leq \alpha/k$ . Holm (1979) improves this single-stage procedure by the following stepdown method. Order the  $p$ -values as

$$\hat{p}_{(1)} \leq \hat{p}_{(2)} \leq \dots \leq \hat{p}_{(k)} ,$$

and let  $H_{(1)}, \dots, H_{(k)}$  denote the corresponding hypotheses. If  $\hat{p}_{(1)} \leq \alpha/k$ , then reject  $H_{(1)}$ ; otherwise, accept all hypotheses and stop. If continuing, reject  $H_{(2)}$  if  $\hat{p}_{(2)} \leq \alpha/(k-1)$ ; otherwise, stop testing and accept all remaining hypotheses. And so on. Then, hypotheses  $H_{(1)}, \dots, H_{(r)}$  are rejected if  $\hat{p}_{(j)} \leq \alpha/(k-j+1)$  for  $j = 1, \dots, r$  and the remaining are accepted if  $\hat{p}_{(r+1)} > \alpha/(k-r)$ . This procedure holds under arbitrary dependence on the joint distribution of  $p$ -values. As shown in Westfall and Young (1993), the Holm procedure can be improved by incorporating or estimating the dependence structure into the algorithm.

In Section 2, we discuss stepdown methods that control the familywise error rate in finite samples. Such methods proceed stagewise by testing an intersection hypothesis at each stage.. That is, like the Holm method, once a hypothesis is rejected, the testing of the remaining hypotheses is accomplished as if the remaining hypotheses were a new family of joint hypotheses to be tested. Moreover, the decision to reject an hypothesis at a given stage only depends on the outcome of the intersection test for that stage.

However, one cannot always achieve strong control in such a simple manner. By understanding the limitations of this approach in finite samples, we can then see why an asymptotic approach will be valid under fairly weak assumptions. It turns out that a simple monotonicity condition for theoretical critical values allows for some immediate results.

For any  $K \subset \{1, \dots, k\}$ , let  $H_K$  denote the hypothesis that all  $H_j$  with  $j \in K$  are true. The closure method of Marcus et al. (1976) allows one to construct methods that control the FWE if one knows how to test each intersection hypothesis  $H_K$ . However, in general, this might require the construction of  $2^k - 1$  tests. The constructions studied here only require order  $k$  number of tests. In fact, the monotonicity assumptions we invoke can be viewed as justification of an order  $k$  stagewise application of closure. (In some cases, shortcuts to applying the closure

method are known. For example, Westfall et al. (2001) show how to apply closure to Fisher combination tests with only  $k^2$  evaluations.) A further advantage of our constructions are that they lead to consonant multiple testing procedures in the sense of Hommel (1986): if the intersection hypothesis  $H_K$  is rejected, then necessarily at least one of the hypotheses  $H_j$  with  $j \in K$  will be rejected. This property is appealing but does not always hold for the closure method of Marcus et al. (1976).

In general, we suppose that rejection of a test of  $H_j$  is based on large values of a test statistic  $T_{n,j}$ . (To be consistent with later notation, the  $n$  is used for asymptotic purposes and typically refers to sample size.) Of course, if a  $p$ -value  $\hat{p}_j$  is available for testing  $H_j$ , one possibility is to take  $T_{n,j} = 1 - \hat{p}_j$ . Then, we restrict attention to tests that reject an intersection hypothesis  $H_K$  when  $\max\{T_{n,j} : j \in K\}$  is large. In some problems where a monotonicity condition holds (distinct from the monotonicity assumption here), Lehmann et al. (2003) show that such stepwise procedures are optimal in a maximin sense. In other situations, it may be better to consider other test statistics that combine the individual test statistics in a more powerful way. A related issue is one of balance; see Remark 4.5. At this time, our primary goal is to show how stepdown procedures can be constructed quite generally without having to assume subset pivotality, but still controlling the FWE.

In Section 3, we show that, if we estimate critical values that have a monotonicity property, then the basic problem of constructing a valid multiple test procedure can be reduced to the problem of sequentially constructing critical values for (at most order  $k$ ) single tests. This then allows us to directly apply what we know about tests based on permutation and randomization distributions. Similarly, we can apply bootstrap and subsampling methods as well, which is done in Section 4.

In Sections 5 and 6, we present two small simulation studies and an empirical application, respectively. All proofs are collected in an appendix.

As previously remarked, the closure method of Marcus et al. (1976) allows one in principle to reduce the problem of constructing a valid multiple test procedure which controls the FWE to the problem of constructing a single test which controls the usual probability of a Type 1 error; however, the number of tests that must be calculated increases exponentially with  $k$ . In general, if one wishes to calculate a bootstrap test for each intersection hypothesis, there is not only the computational issue of constructing a large number of tests, but also the question of an appropriate resampling mechanism which obeys the null hypothesis for each intersection hypothesis (unless one has a strong assumption like subset pivotality). Our methods are both computationally feasible and avoid the need for a distinct resampling mechanism for each hypothesis. Thus, this work is a sustained essay designed to produce computationally feasible general test constructions that control the FWE by effective reduction to the problem of construction of single tests that control the usual probability of a type 1 error. This then allows us to draw upon an enormous resampling literature.

## 2 Nonasymptotic Results

Suppose data  $X$  is generated from some unknown probability distribution  $P$ . In anticipation of asymptotic results, we may write  $X = X^{(n)}$ , where  $n$  typically refers to the sample size. A model assumes that  $P$  belongs to a certain family of probability distributions  $\Omega$ , though we make no rigid requirements for  $\Omega$ . Indeed,  $\Omega$  may be a nonparametric model, a parametric model, or a semiparametric model.

Consider the problem of simultaneously testing a hypothesis  $H_j$  against  $H_j^c$ , for  $j = 1, \dots, k$ . Of course, a hypothesis  $H_j$  can be viewed as a subset,  $\omega_j$ , of  $\Omega$ , in which case the hypothesis  $H_j$  is equivalent to  $P \in \omega_j$  and  $H_j^c$  is equivalent to  $P \notin \omega_j$ . For any subset  $K \subset \{1, \dots, k\}$ , let  $H_K = \bigcap_{j \in K} H_j$  be the hypothesis that  $P \in \bigcap_{j \in K} \omega_j$ .

Suppose that a test of the individual hypothesis  $H_j$  is based on a test statistic  $T_{n,j}$ , with large values indicating evidence against the  $H_j$ . For an individual hypothesis, numerous approaches exist to approximate a critical value, such as those based on classical likelihood theory, bootstrap tests, Edgeworth expansions, permutation tests, etc. The main problem addressed in the present work is to construct a procedure that controls the familywise error rate (FWE). Recall that the familywise error rate is the probability of rejecting at least one true null hypothesis. More specifically, if  $P$  is the true probability mechanism, let  $I = I(P) \subset \{1, \dots, k\}$  denote the indices of the set of true hypotheses; that is,  $j \in I$  if and only if  $P \in \omega_j$ . The FWE is the probability under  $P$  that any  $H_j$  with  $j \in I$  is rejected. To show its dependence on  $P$ , we may write  $\text{FWE} = \text{FWE}_P$ . We require that any procedure satisfy that the familywise error rate be no bigger than  $\alpha$  (at least asymptotically). Furthermore, this constraint must hold for all possible configurations of true and null hypotheses; that is, we demand strong control of the FWE. A procedure that only controls the FWE when all  $k$  null hypotheses are true is said to have weak control of the FWE. As remarked by Dudoit et al. (2003), this distinction is often ignored.

For any subset  $K$  of  $\{1, \dots, k\}$ , let  $c_{n,K}(\alpha, P)$  denote an  $\alpha$ -quantile of the distribution of  $\max_{j \in K} T_{n,j}$  under  $P$ . Concretely,

$$c_{n,K}(\alpha, P) = \inf\{x : P\{\max_{j \in K} T_{n,j} \leq x\} \geq \alpha\}. \quad (1)$$

For testing the intersection hypothesis  $H_K$ , it is only required to approximate a critical value for  $P \in \bigcap_{j \in K} \omega_j$ . Because there may be many such  $P$ , we define

$$c_{n,K}(1 - \alpha) = \sup\{c_{n,K}(1 - \alpha, P) : P \in \bigcap_{j \in K} \omega_j\}. \quad (2)$$

In order to define  $c_{n,K}(\alpha, P)$ , we implicitly assumed  $\bigcap_{j=1}^k \omega_j$  is not empty. At this point, we acknowledge that calculating these constants may be formidable in some problems (which is why we later turn to approximate or asymptotic methods).

Let

$$T_{n,r_1} \geq T_{n,r_2} \geq \dots \geq T_{n,r_k} \quad (3)$$

denote the observed ordered test statistics, and let  $H_{r_1}, H_{r_2}, \dots, H_{r_k}$  be the corresponding hypotheses.

Stepdown procedures begin by testing the joint null hypothesis  $H_{\{1, \dots, k\}}$  that all hypotheses are true. This hypothesis is rejected if  $T_{n, r_1}$  is large. If it is not large, accept all hypotheses; otherwise, reject the hypothesis corresponding to the largest test statistic. Once a hypothesis is rejected, remove it and test the remaining hypotheses by rejecting for large values of the maximum of the remaining test statistics, and so on. Thus, at any step, one tests an intersection hypothesis, and an ideal situation would be to proceed at any step without regard to previous rejections, in the sense that once a hypothesis is rejected, the remaining hypotheses are treated as a new family and testing for this new family proceeds independent of past decisions in such a way that rejecting one of the remaining hypotheses is based solely on the rejection of the next intersection test calculated. Because the Holm procedure (discussed later in Example 2.4) works in this way, one might hope that one can generally test the intersection hypothesis at any step by treating only those hypotheses that remain. Forgetting about whether or not such an approach generally yields strong control for the time being, we consider the following conceptual algorithm, which proceeds in stages by testing intersection hypotheses.

**Algorithm 2.1 (Idealized Stepdown Method)**

1. Let  $K_1 = \{1, \dots, k\}$ . If  $T_{n, r_1} \leq c_{n, K_1}(1 - \alpha)$ , then accept all hypotheses and stop; otherwise, reject  $H_{r_1}$  and continue.
2. Let  $K_2$  be the indices of the hypotheses not previously rejected. If  $T_{n, r_2} \leq c_{n, K_2}(1 - \alpha)$ , then accept all remaining hypotheses and stop; otherwise, reject  $H_{r_2}$  and continue.
- ⋮
- j. Let  $K_j$  be the indices of the hypotheses not previously rejected. If  $T_{n, r_j} \leq c_{n, K_j}(1 - \alpha)$ , then accept all remaining hypotheses and stop; otherwise, reject  $H_{r_j}$  and continue.
- ⋮
- k. If  $T_{n, r_k} \leq c_{n, K_k}(1 - \alpha)$ , then accept  $H_{r_k}$ ; otherwise, reject  $H_{r_k}$ .

The above algorithm is an idealization for two reasons: the critical values may be impossible to compute and, without restriction, there is no general reason why such a stepwise approach strongly controls the FWE. The determination of conditions where the algorithm leads to strong control will help us understand the limitations of a stepdown approach as well as understand how such a general approach can at least approximately work in large samples. First, we present an example to show that some condition is required to exhibit strong control.

**Example 2.1** Suppose  $T_{n,1}$  and  $T_{n,2}$  are independent and normally distributed, with  $T_{n,1} \sim N(\theta_1, (1 + \theta_2)^{2a})$  and  $T_{n,2} \sim N(\theta_2, (1 + \theta_2)^{-2a})$ , where  $\theta_1 \geq 0$  and  $\theta_2 \geq 0$ . (The index  $n$

plays no role here, but we retain it for consistent notation.) Here,  $q$  is a suitable positive constant, chosen to be large. Also, let  $\Phi(\cdot)$  denote the standard normal cumulative distribution function. The hypothesis  $H_j$  specifies  $\theta_j = 0$  while  $H'_j$  specifies  $\theta_j > 0$ . Therefore, the first step of Algorithm 2.1 is to reject the overall joint hypothesis  $\theta_1 = \theta_2 = 0$  for large values of  $\max(T_{n,1}, T_{n,2})$  when  $T_{n,1}$  and  $T_{n,2}$  are i.i.d.  $N(0, 1)$ . Specifically, accept both hypotheses if

$$\max(T_{n,1}, T_{n,2}) \leq c(1 - \alpha) \equiv \Phi^{-1}(\sqrt{1 - \alpha}) ;$$

otherwise, reject the hypothesis corresponding to the larger  $T_{n,j}$ . Such a procedure exhibits weak control but not strong control. For example, the probability of rejecting the  $H_1$  at the first step when  $\theta_1 = 0$  and  $\theta_2 = c(1 - \alpha)/2$  satisfies

$$P_{0, \theta_2} \{T_{n,1} > c(1 - \alpha), T_{n,1} > T_{n,2}\} \rightarrow 1/2$$

as  $q \rightarrow \infty$ . So, if  $\alpha < 1/2$ , for some large enough but fixed  $q$ , the probability of incorrectly declaring  $H_1$  to be false is greater than  $\alpha$ . Incidentally, this also provides an example of a single-step procedure which exhibits weak control but not strong control. (Single-step procedures are those where hypotheses are rejected on the basis of a single critical value; see Westfall and Young (1993).)

Therefore, in order to prove strong control, some condition is required. Consider the following monotonicity assumption: for  $I \subset K$ ,

$$c_{n,K}(1 - \alpha) \geq c_{n,I}(1 - \alpha) . \tag{4}$$

The condition (4) can be expected to hold in many situations because the left hand side is based on computing the  $1 - \alpha$  quantile of the maximum of  $|K|$  variables, while the right hand side is based on the maximum of  $|I| \leq |K|$  variables (though one must be careful and realize that the quantiles are computed under possibly different  $P$ , which is why some condition is required).

**Theorem 2.1** *Let  $P$  denote the true distribution generating the data. Assume  $\bigcap_{j=1}^k \omega_j$  is not empty.*

(i) *Assume for any  $K$  containing  $I(P)$ ,*

$$c_{n,K}(1 - \alpha) \geq c_{n,I(P)}(1 - \alpha) . \tag{5}$$

*Then, the probability that Algorithm 2.1 rejects any  $j \in I(P)$  is  $\leq \alpha$ ; that is,  $FWE_P \leq \alpha$ .*

(ii) *Strong control persists if, in Algorithm 2.1, the critical constants  $c_{n,K_j}(1 - \alpha)$  are replaced by  $d_{n,K_j}(1 - \alpha)$  which satisfy*

$$d_{n,K_j}(1 - \alpha) \geq c_{n,K_j}(1 - \alpha) . \tag{6}$$

(iii) Moreover, the condition (5) may be removed if the  $d_{n,K_j}(1 - \alpha)$  satisfy

$$d_{n,K}(1 - \alpha) \geq d_{n,I(P)}(1 - \alpha) \quad (7)$$

for any  $K \supset I(P)$ .

**Remark 2.1** Under weak assumptions, one can show the sup over  $P$  of the probability (under  $P$ ) that Algorithm 2.1 rejects any  $j \in I(P)$  is equal to  $\alpha$ . It then follows that the critical values cannot be made smaller, in hopes of increasing the ability to detect false hypotheses, without violating the strong control of the FWE. (However, this does not negate the possibility of smaller random critical values, as long as they are not smaller with probability one.)

**Example 2.2 (Assumption of subset pivotality)** Assumptions stronger than (5) have been used. Suppose, for example, that for every subset  $K \subset \{1, \dots, k\}$ , there exists a distribution  $P_K$  which satisfies

$$c_{n,K}(1 - \alpha, P) \leq c_{n,K}(1 - \alpha, P_K) \quad (8)$$

for all  $P$  such that  $I(P) \supset K$ . Such a  $P_K$  may be referred to being least favorable among distributions  $P$  such that  $P \in \bigcap_{j \in K} \omega_j$ . (For example, if  $H_j$  corresponds to a parameter  $\theta_j \leq 0$ , then intuition suggests a least favorable configuration should correspond to  $\theta_j = 0$ .)

In addition, assume the subset pivotality condition of Westfall and Young (1993); that is, assume there exists a  $P_0$  with  $I(P_0) = \{1, \dots, k\}$  such that the joint distribution of  $\{T_{n,j} : j \in I(P_K)\}$  under  $P_K$  is the same as the distribution of  $\{T_{n,j} : j \in I(P_K)\}$  under  $P_0$ . This condition says the (joint) distribution of the test statistics used for testing the hypotheses  $H_j$ ,  $j \in I(P_K)$  is unaffected by the truth or falsehood of the remaining hypotheses (and therefore we assume all hypotheses are true by calculating the distribution of the maximum under  $P_0$ ). It follows that, in step  $j$  of Algorithm 2.1,

$$c_{n,K_j}(1 - \alpha) = c_{n,K_j}(1 - \alpha, P_{K_j}) = c_{n,K_j}(1 - \alpha, P_0) = c_{n,K_j}(1 - \alpha) ; \quad (9)$$

the outer equalities in (9) follow by the assumption (8) and the middle equality follows by the subset pivotality condition. Therefore, in Algorithm 2.1, we can replace  $c_{n,K_j}(1 - \alpha)$  by  $c_{n,K_j}(1 - \alpha, P_0)$ , which in principle is known because it is the  $1 - \alpha$  quantile of the distribution of  $\max(T_{n,j} : j \in K_j)$  under  $P_0$ , and  $P_0$  is some fixed (least favorable) distribution. At the very least, this quantile may be simulated.

The asymptotic behavior of stepwise procedures is considered in Finner and Roters (1998), and they recognize the importance of monotonicity for the validity of stepwise procedures. However, they also suppose the existence of a single least favorable  $P_0$  for all configurations of true hypotheses, which then guarantees monotonicity of critical values for stepdown procedures. As previously seen, such assumptions do not hold generally.

**Example 2.3** To exhibit an example where condition (5) holds, but subset pivotality does not, suppose that  $T_{n,1}$  and  $T_{n,2}$  are independent, normally distributed, with  $T_{n,1} \sim N(\theta_1, 1/(1+\theta_2^2))$  and  $T_{n,2} \sim N(\theta_2, 1/(1+\theta_1^2))$ . The hypothesis  $H_j$  specifies  $\theta_j = 0$  while the alternative  $H'_j$  specifies  $\theta_j > 0$ . Then, it is easy to check that, with  $K_1 = \{1, 2\}$ ,

$$c_{n,K_1}(1-\alpha) = \Phi^{-1}(\sqrt{1-\alpha}) > \Phi^{-1}(1-\alpha) = c_{n,\{j\}}(1-\alpha) .$$

Therefore, (5) holds, but subset pivotality fails.

**Example 2.4 (The Holm Procedure)** Suppose  $-T_{n,j} \equiv \hat{p}_{n,j}$  is a  $p$ -value for testing  $H_j$ ; that is, assume the distribution of  $\hat{p}_{n,j}$  is Uniform on  $(0, 1)$  when  $H_j$  is true. Note that this assumption is much weaker than subset pivotality (if  $k > 1$ ) because we are only making an assumption about the one-dimensional marginal distribution of the  $p$ -value statistic. Furthermore, we may assume the weaker condition

$$P\{\hat{p}_{n,j} \leq x\} \leq x$$

for any  $x \in (0, 1)$  and any  $P \in \omega_j$ . If  $I(P) \supset K$ , the usual argument using the Bonferroni inequality yields

$$c_{n,K}(1-\alpha, P) \leq -\alpha/|K| ,$$

which is independent of  $P$ , and so

$$c_{n,K}(1-\alpha) \leq -\alpha/|K| . \tag{10}$$

It is easy to construct joint distributions for which this is attained, and so we have equality here if the family  $\Omega$  is so large that it includes all possible joint distributions for the  $p$ -values. In such case, we have equality in (10) and so the condition (5) is satisfied. Of course, even if the model is not so large, this procedure has strong control. Simply, let  $d_{n,K}(1-\alpha) = -\alpha/|K|$ , and strong control follows by Theorem 2.1(iii).

Part (iii) of Theorem 2.1 points toward a more general method that has strong control even when (5) is violated, and that can be much less conservative than the Holm procedure.

**Corollary 2.1** Assume  $\bigcap_{j=1}^k \omega_j$  is not empty. Let

$$c_{n,K_j}^*(1-\alpha) = \max\{c_{n,K}(1-\alpha) : K \subset K_j\} . \tag{11}$$

Then, if you replace  $c_{n,K_j}(1-\alpha)$  by  $c_{n,K_j}^*(1-\alpha)$  in Algorithm 2.1, strong control holds.

Corollary 2.1 is simply the closure principle of Marcus et al. (1976); also see Hommel (1986) and Theorem 4.1 of Hochberg and Tamhane (1987). Thus, in order to have a valid stepdown procedure, one must not only consider the critical value  $c_{n,K}(1-\alpha)$  when testing an intersection hypothesis  $H_K$ , one must also compute all  $c_{n,I}(1-\alpha)$  for  $I \subset K$ .

Finally, one can remove the assumption that  $\bigcap_{j=1}^k \omega_j$  is not empty as follows.

**Theorem 2.2** *Let*

$$\tilde{c}_{n,K_j}(1 - \alpha) = \max\{c_{n,K}(1 - \alpha) : K \subset K_j \text{ and } \bigcap_{j \in K} \omega_j \neq \emptyset\} . \quad (12)$$

(i) *If you replace  $c_{n,K_j}(1 - \alpha)$  by  $\tilde{c}_{n,K_j}(1 - \alpha)$  in Algorithm 2.1, strong control holds.*

(ii) *Strong control persists if, in Algorithm 2.1, the critical constants  $c_{n,K_j}(1 - \alpha)$  are replaced by  $\tilde{d}_{n,K_j}(1 - \alpha)$  which satisfy*

$$\tilde{d}_{n,K_j}(1 - \alpha) \geq \tilde{c}_{n,K_j}(1 - \alpha) . \quad (13)$$

Theorem 2.2 shows that the Holm method applies with no assumptions, that is, the assumption that all hypotheses can be true need not hold.

### 3 Random Critical Values and Randomization Tests

#### 3.1 Preliminaries and a Basic Inequality

In general, the critical values used in Algorithm 2.1 are the smallest constants possible without violating the FWE. As a simple example, suppose  $X_j$ ,  $j = 1, \dots, k$ , are independent  $N(\theta_j, 1)$ , with the  $\theta_j$  varying freely. The null hypothesis  $H_j$  specifies  $\theta_j \leq 0$ . Then,

$$c_{n,K}(1 - \alpha) = \Phi^{-1}[(1 - \alpha)^{(1/|K|)}] .$$

Suppose  $c$  is a constant and  $c < c_{n,K}(1 - \alpha)$  for some subset  $K$ . As  $\theta_j \rightarrow \infty$  for  $j \notin K$  and  $\theta_j = 0$  for  $j \in K$ , the probability of a type 1 error tends to

$$P\{\max_{j \in K} X_j > c\} > P\{\max_{j \in K} X_j > c_{n,K}(1 - \alpha)\} = \alpha .$$

Of course, if the  $\theta_j$  are bounded, the argument fails, but typically such assumptions are not made.

However, the above only applies to nonrandom critical values and leaves open the possibility that critical values can be estimated, and therefore be random. That is, if we replace  $c_{n,K}(1 - \alpha)$  by some estimate  $\hat{c}_{n,K}(1 - \alpha)$ , it can sometimes be smaller than  $c_{n,K}(1 - \alpha)$  as long as it is not with probability one. Of course, this is the typical case where critical values need to be estimated, such as by the bootstrap in the next section. In this section, we focus on the use of permutation and randomization tests that replace the idealized critical values by estimated ones, while still retaining finite sample control of the FWE.

One simple way to deal with permutation and randomization tests is to define critical values conditional on an appropriate  $\sigma$ -field, and then the monotonicity assumptions of the previous section would then turn into monotonicity assumptions for the conditional critical values. (For

example, in the context of comparing two samples, everything would be conditional on the values of the combined sample, and this would directly lead to permutation tests.)

For the sake of increased generality, we instead proceed as follows. Suppose the  $c_{n,K}(1 - \alpha)$  in Algorithm 2.1 are replaced by estimates  $\hat{c}_{n,K}(1 - \alpha)$ . These could be obtained by a permutation test if it applies, but for the moment their construction is left unspecified. However, we will assume two things. First, we will replace the monotonicity assumption (5) by monotonicity of the estimated critical values; that is, for any  $K \supset I(P)$ ,

$$\hat{c}_{n,K}(1 - \alpha) \geq \hat{c}_{n,I(P)}(1 - \alpha) . \quad (14)$$

We then also require that, if  $\hat{c}_{n,K}(1 - \alpha)$  is used to test the intersection hypothesis  $H_K$ , then it is level  $\alpha$  when  $K = I(P)$ ; that is,

$$P\{\max(T_{n,j} : j \in I(P)) > \hat{c}_{n,I(P)}(1 - \alpha)\} \leq \alpha . \quad (15)$$

We will show the basic inequality that the  $FWE_P$  is bounded above by left side of (15). This will then show that, if we can construct monotone critical values such that each intersection test is level  $\alpha$ , then the stepdown procedure controls the FWE. Thus, the construction of a stepdown procedure is effectively reduced to construction of single tests, as long as the monotonicity assumption holds. (Also, note the monotonicity assumption for the critical values, which is something we can essentially enforce because they only depend on the data, can hold even if the corresponding nonrandom ones are not monotone.) Note that here (and in the rest of the paper) we no longer need to assume  $\bigcap_{j=1}^k \omega_j$  is not empty.

**Theorem 3.1** *Let  $P$  denote the true distribution generating the data. Consider Algorithm 2.1 with  $c_{n,K}(1 - \alpha)$  replaced by estimates  $\hat{c}_{n,K}(1 - \alpha)$  satisfying (14).*

(i) *Then,*

$$FWE_P \leq P\{\max(T_{n,j} : j \in I(P)) > \hat{c}_{n,I(P)}(1 - \alpha)\} . \quad (16)$$

(ii) *Therefore, if the critical values also satisfy (15), then  $FWE_P \leq \alpha$ .*

### 3.2 Permutation and Randomization Tests

Before applying Theorem 3.1, we first review a general construction of a randomization test in the context of a single test. Our setup is framed in terms of a population model, but similar results are possible in terms of a randomization model (as in Section 3.1.7 of Westfall and Young (1993)).

Based on data  $X$  taking values in a sample space  $\mathcal{X}$ , it is desired to test the null hypothesis  $H$  that the underlying probability law  $P$  generating  $X$  belongs to a certain family  $\omega$  of distributions. Let  $\mathbf{G}$  be a finite group of transformations  $g$  of  $\mathcal{X}$  onto itself. The following assumption, which we will call the *randomization hypothesis*, allows for a general test construction.

**The Randomization Hypothesis** The null hypothesis implies that the distribution of  $X$  is invariant under the transformations in  $\mathbf{G}$ ; that is, for every  $g$  in  $\mathbf{G}$ ,  $gX$  and  $X$  have the same distribution whenever  $X$  has distribution  $P$  in  $\omega$ .

As an example, consider testing the equality of distributions based on two independent samples  $(Y_1, \dots, Y_m)$  and  $(Z_1, \dots, Z_n)$ . Under the null hypothesis that the samples are generated from the same probability law, the observations can be permuted or assigned at random to either of the two groups, and the distribution of the permuted samples is the same as the distribution of the original samples. In this example, and more generally when the randomization hypothesis holds, the following construction of a randomization test applies.

Let  $T(X)$  be any real-valued test statistic for testing  $H$ . Suppose the group  $\mathbf{G}$  has  $M$  elements. Given  $X = x$ , let

$$T^{(1)}(x) \leq T^{(2)}(x) \leq \dots \leq T^{(M)}(x)$$

be the values of  $T(gx)$  as  $g$  varies in  $\mathbf{G}$ , ordered from smallest to largest. Fix a nominal level  $\alpha$ ,  $0 < \alpha < 1$ , and let  $m$  be defined by

$$m = M - [M\alpha] , \tag{17}$$

where  $[M\alpha]$  denotes the largest integer less than or equal to  $M\alpha$ . Let  $M^+(x)$  and  $M^0(x)$  be the number of values  $T^{(j)}(x)$  ( $j = 1, \dots, M$ ) which are greater than  $T^{(m)}(x)$  and equal to  $T^{(m)}(x)$ , respectively. Set

$$a(x) = \frac{M\alpha - M^+(x)}{M^0(x)} .$$

Define the randomization test function  $\phi(X)$  to be equal to 1,  $a(X)$ , or 0 according to whether  $T(X) > T^{(m)}(X)$ ,  $T(X) = T^{(m)}(X)$ , or  $T(X) < T^{(m)}(X)$ , respectively.

Under the randomization hypothesis, Hoeffding (1952) shows this construction produces a test that is exact level  $\alpha$ , and this result is true for *any* choice of test statistic  $T$ . Note that this test is possibly a randomized test if  $M\alpha$  is not an integer of there are ties in the ordered values. Alternatively, if one prefers not to randomize, the slightly conservative but *nonrandomized* test that rejects if  $T(X) > T^{(m)}(X)$  is level  $\alpha$ .

For any  $x \in \mathcal{X}$ , let  $\mathbf{G}^x$  denote the  $\mathbf{G}$ -orbit of  $x$ ; that is,

$$\mathbf{G}^x = \{gx : g \in \mathbf{G}\} .$$

These orbits partition the sample space. Then, under the randomization hypothesis, it can be shown that the conditional distribution of  $X$  given  $X \in \mathbf{G}^x$  is uniform on  $\mathbf{G}^x$ .

In general, one can define a  $p$ -value  $\hat{p}$  of a randomization test by

$$\hat{p} = \frac{1}{M} \sum_g I\{T(gX) \geq T(X)\} . \tag{18}$$

It is easily shown that  $\hat{p}$  satisfies, under the null hypothesis,

$$P\{\hat{p} \leq u\} \leq u \quad \text{for all } 0 \leq u \leq 1 . \quad (19)$$

Therefore, the *nonrandomized* test that rejects when  $\hat{p} \leq \alpha$  is level  $\alpha$ .

Because  $\mathbf{G}$  may be large, one may resort to a stochastic approximation to construct the randomization test, for example, by randomly sampling transformations  $g$  from  $\mathbf{G}$  with or without replacement. In the former case, for example, suppose  $g_1, \dots, g_{B-1}$  are i.i.d. and uniformly distributed on  $\mathbf{G}$ . Let

$$\tilde{p} = \frac{1}{B} \left[ 1 + \sum_{i=1}^{B-1} I\{T(g_i X) \geq T(X)\} \right] . \quad (20)$$

Then, it can be shown that, under the randomization hypothesis,

$$P\{\tilde{p} \leq u\} \leq u \quad \text{for all } 0 \leq u \leq 1 , \quad (21)$$

where this probability reflects variation in both  $X$  and the sampling of the  $g_i$ . Note that (21) holds for any  $B$ , and so the test that rejects when  $\tilde{p} \leq \alpha$  is level  $\alpha$  even when a stochastic approximation is employed. Of course, the larger the value of  $B$ , the closer  $\hat{p}$  and  $\tilde{p}$  are to each other; in fact,  $\hat{p} - \tilde{p} \rightarrow 0$  in probability as  $B \rightarrow \infty$ . The argument for (20) is based on the following simple fact.

**Lemma 3.1** *Suppose  $Y_1, \dots, Y_B$  are exchangeable real-valued random variables; that is, their joint distribution is invariant under permutations. Let  $\tilde{q}$  be defined by*

$$\tilde{q} = \frac{1}{B} \left[ 1 + \sum_{i=1}^{B-1} I\{Y_i \geq Y_B\} \right] .$$

*Then,  $P\{\tilde{q} \leq u\} \leq u$  for all  $0 \leq u \leq 1$ .*

We now return to the multiple testing problem. Assume  $\mathbf{G}_K$  is a group of transformations for which the randomization hypothesis holds for  $H_K$ . Then, we can apply the above construction to test the single intersection hypothesis  $H_K$  based on the test statistic

$$T_{n,K} = \max(T_{n,j} : j \in K) \quad (22)$$

and reject  $H_K$  when

$$T_{n,K}(X) > T_{n,K}^{(|\mathbf{G}_K| - \lceil |\mathbf{G}_K| \alpha \rceil)}(X) .$$

If we further specialize to the case where  $\mathbf{G}_K = \mathbf{G}$ , so that the same  $\mathbf{G}$  applies to all intersection hypotheses, then we can verify the monotonicity assumption for the critical values. Set  $m_\alpha = |\mathbf{G}| - \lceil |\mathbf{G}| \alpha \rceil$ . Then, for any  $g \in \mathbf{G}$  and  $I \subset K$ ,

$$\max(T_{n,j}(gX) : j \in K) \geq \max(T_{n,j}(gX) : j \in I) , \quad (23)$$

and so as  $g$  varies, the  $m_\alpha$ th largest value of the left side of (23) is at least as large as the  $m_\alpha$ th largest value of the right side.

Consequently, the critical values

$$\hat{c}_{n,K}(1 - \alpha) = T_{n,K}^{(m_\alpha)}, \quad (24)$$

satisfy the monotonicity requirement of Theorem 3.1. Moreover, by the general randomization construction of a single test, the test that rejects  $H_K$  when  $T_K \geq T_{n,K}^{(m_\alpha)}$  is level  $\alpha$ . Therefore, the following is true.

**Corollary 3.1** *Suppose the randomization hypothesis holds for a group  $\mathbf{G}$  when testing any intersection hypothesis  $H_K$ . Then, the stepdown method with critical values given by (24) controls the FWE.*

Equivalently, in analogy with (18), we can compute  $p$ -values for testing  $H_K$  via

$$\hat{p}_{n,K} = \frac{1}{M} \sum_g I\{T_{n,K}(gX) \geq T_{n,K}(X)\}, \quad (25)$$

and at stage  $j$  where we are testing an intersection hypothesis, say  $H_K$ , reject if  $\hat{p}_{n,K} \leq \alpha$ .

Alternatively, we can approximate these  $p$ -values and still retain the level of the test. In analogy with (20), randomly sample  $g_1, \dots, g_{B-1}$  from  $\mathbf{G}$  and let

$$\tilde{p}_{n,K} = \frac{1}{B} \left[ 1 + \sum_{i=1}^{B-1} I\{T_{n,K}(g_i X) \geq T_{n,K}(X)\} \right]. \quad (26)$$

By an almost identical argument, we have the following.

**Corollary 3.2** *Suppose the randomization hypothesis holds for a group  $\mathbf{G}$  when testing any intersection hypothesis  $H_K$ . Consider the stepdown method which rejects  $K_j$  at stage  $j$  if  $\tilde{p}_{n,K_j} \leq \alpha$ . Then,  $FWE_P \leq \alpha$ .*

**Remark 3.1** In the above corollaries, we have worked with the randomization construction using nonrandomized tests. A similar result would hold if we permit randomization.

**Example 3.1 (Two Sample Problem With  $k$  Variables)** Suppose  $Y_1, \dots, Y_{n_Y}$  is a sample of  $n_Y$  independent observations from a probability distribution  $P_Y$  and  $Z_1, \dots, Z_{n_Z}$  is a sample of  $n_Z$  observations from  $P_Z$ . Here,  $P_Y$  and  $P_Z$  are probability distributions on  $\mathbf{R}^k$ , with  $j$ th components denoted  $P_{Y,j}$  and  $P_{Z,j}$ , respectively. The hypothesis  $H_j$  asserts  $P_{Y,j} = P_{Z,j}$  and we wish to test these  $k$  hypotheses based on  $X = (Y_1, \dots, Y_{n_Y}, Z_1, \dots, Z_{n_Z})$ . Also, let  $Y_{i,j}$  denote the  $j$ th component of  $Y_i$  and  $Z_{i,j}$  denote the  $j$ th component of  $Z_i$ . As in Troendle (1995), we assume a semiparametric model. In particular, assume  $P_Y$  and  $P_Z$  are governed

by a family of probability distributions  $Q_\theta$  indexed by  $\theta = (\theta_1, \dots, \theta_k) \in \mathbf{R}^k$  (and assumed identifiable), so that  $P_Y$  has law  $Q(\theta_Y)$  and  $P_Z$  has law  $Q(\theta_Z)$ . For concreteness, one may think of  $\theta$  as being the mean vector, though this assumption is not necessary. Now,  $H_j$  can be viewed as testing  $\theta_{Y,j} = \theta_{Z,j}$ . Note that the randomization construction does not need to assume knowledge of the form of  $Q$  (just as a single two-sample permutation test in a shift model does not need to know the form of the underlying distribution under the null hypothesis).

Let  $n = n_Y + n_Z$ , and for  $x = (x_1, \dots, x_n) \in \mathbf{R}^n$ , let  $gx \in \mathbf{R}^n$  be defined by  $(x_{\pi(1)}, \dots, x_{\pi(n)})$ , where  $(\pi(1), \dots, \pi(n))$  is a permutation of  $(1, 2, \dots, n)$ . Let  $\mathbf{G}$  be the collection of all such  $g$  so that  $M = n!$ . Under the hypothesis  $P_Y = P_Z$ ,  $gX$  and  $X$  have the same distribution for any  $g$  in  $\mathbf{G}$ .

Unfortunately, this  $\mathbf{G}$  does not apply to any subset of the hypotheses, because  $gX$  and  $X$  need not have the same distribution if only a subcollection of the hypotheses are true. However, we just need a slight generalization to cover the example. Suppose that the test statistic  $T_{n,j}$  used to test  $H_j$  only depends on the  $j$ th components of the observations, namely  $Y_{i,j}$ ,  $i = 1, \dots, n_Y$  and  $Z_{i,j}$ ,  $i = 1, \dots, n_Z$ ; this is a weak assumption indeed. In fact, let  $X_K$  be the data set consisting of the components  $Y_{i,j}$  and  $Z_{i,j}$  as  $j$  varies only in  $K$ . The simple but important point here is that, for this reduced data set, the randomization hypothesis holds. Specifically, under the null hypothesis  $\theta_{Y,j} = \theta_{Z,j}$  for  $j \in K$ ,  $X_K$  and  $gX_K$  have the same distribution (though  $X$  and  $gX$  need not). Also, for any  $g \in \mathbf{G}$ ,  $T_{n,j}(gX)$  and  $T_{n,j}(X)$  have the same distribution under  $H_j$ , and similarly for any  $K \subset \{1, \dots, k\}$ ,  $T_{n,K}(gX)$  and  $T_{n,K}(X)$  have the same distribution under  $H_K$ .

Then, because the same  $\mathbf{G}$  applies in this manner for all  $K$ , the critical values from the randomization test are monotone, just as in (23). Moreover, each intersection hypothesis can be tested by an exact level  $\alpha$  randomization test (since inference for  $H_K$  is based only on  $X_K$ ). Therefore, essentially the same argument leading to Corollaries 3.1 and 3.2 applies. In particular, even if we need to resort to approximate randomization tests at each stage, but as long as we sample the same set of  $g_i$  from  $\mathbf{G}$ , the resulting procedure retains its finite sample property of controlling the FWE. In contrast, Troendle (1995) concludes asymptotic control.

**Remark 3.2** It is interesting to study the behavior of randomization procedures if the model is such that the randomization hypothesis does not hold. For example, in Example 3.1, suppose we are just interested in testing the hypothesis  $H'_j$  that the mean of  $P_{Y,j}$  is the mean of  $P_{Z,j}$  (assumed to exist). Then, the randomization test construction of this section fails because the randomization hypothesis need not hold. However, since the randomization procedure has monotone critical values (as this is only a property of how the data is used), Theorem 3.1(i) applies. Therefore, one can again reduce the problem of studying control of the FWE to that of controlling the level of a single intersection hypothesis. But the problem of controlling the level of a single test when the randomization hypothesis fails is studied in Romano (1990) and so similar methods can be used here, with the hope of at least proving asymptotic control. Alternatively, the more general resampling approaches of Section 4 can be employed; the

comparison of randomization and bootstrap tests is studied in Romano (1989) and it is shown they are often quite close, at least when the randomization hypothesis holds.

**Example 3.2 (Problem of Multiple Treatments)** Consider the one-way anova model. We are given  $k + 1$  independent samples, with the  $j$ th sample having  $n_j$  i.i.d. observations  $X_{i,j}$ ,  $i = 1, \dots, n_j$ . Suppose  $X_{i,j}$  has distribution  $P_j$ . The problem is to test the hypotheses of  $k$  treatments with a control; that is,  $H_j : P_j = P_{k+1}$ . (Alternatively, we can test all pairs of distributions, but the issues are much the same, so we illustrate them with the slightly easier setup.) Under the joint null hypothesis, we can randomly assign all  $n = \sum_j n_j$  observations to any of the groups; that is, the group  $\mathbf{G}$  consists of all permutations of the data. However, if only a subset of the hypotheses are true, this group is not valid. A simple remedy is to permute only within subsets; that is, to test any subset hypothesis  $H_K$ , only consider those permutations that permute observations within the sample  $X_{i,k+1}$  and the samples  $X_{i,j}$  with  $j \in K$ . Therefore, one computes a critical value by  $\hat{c}_{n,K}(1 - \alpha)$  by the randomization test with the group  $\mathbf{G}_K$  of permutations within samples  $j \in K$  and  $j = k + 1$ . Unfortunately, this does not lead to monotonicity of critical values, and the previous results do not apply. But, there is an analogue of Corollary 2.1, if one is willing to compute critical values for all subset hypotheses; that is, replace  $\hat{c}_{n,K_j}(1 - \alpha)$  by

$$\hat{c}_{n,K_j}^*(1 - \alpha) = \max\{\hat{c}_{n,K}(1 - \alpha) : K \subset K_j\} .$$

On the other hand, this can be computationally prohibitive. Such issues are raised by Petrondas and Gabriel (1983) (although the problem was not framed in terms of a monotonicity requirement). However, we will shortly see that the lack of monotonicity of critical values is only a finite sample concern; see Example 4.2.

## 4 Asymptotic Results

The main goal of this section is to construct asymptotically valid stepdown procedures that hold under very weak assumptions, even when the monotonicity condition of Theorem 2.1 fails. The assumptions are identical to the weakest assumptions available for the construction of asymptotically valid tests of a single hypothesis, which are used in many resampling schemes, and so one cannot expect to improve them without improving the now well-developed theory of resampling methods for testing a single hypothesis.

Of course, Corollary 2.1 reminds us that it may be possible to construct a test that controls the FWE if we are willing and able to compute critical values for all possible  $2^k - 1$  nontrivial intersection hypotheses. If each such test is computed by a bootstrap or resampling method, the number of computations could get quite large for even moderate  $k$ . Not only will we provide weak conditions, but we will consider a method that only requires *one* set of bootstrap resamples, as well as a method based on *one* set of subsamples.

In order to accomplish this without having to invoke an assumption like subset pivotality, we will consider resampling schemes that do *not* obey the constraints of the null hypothesis. Schemes that do obey the constraints of the null hypothesis, as discussed in Beran (1986) and Romano (1988), are based on the idea that the critical value should be obtained under the null hypothesis and so the resampling scheme should reflect the constraints of the null hypothesis. This idea is even advocated as a principle in Hall and Wilson (1991), and it is enforced throughout Westfall and Young (1993). While appealing, it is by no means the only approach toward inference in hypothesis testing. Indeed, the well-known explicit duality between tests and confidence intervals means that if you can construct good or valid confidence intervals, then you can construct good or valid tests, and conversely. But, resampling the empirical distribution to construct a confidence interval for a single parameter can produce very desirable intervals, which would then translate into desirable tests. The same holds for simultaneous confidence sets and multiple tests.

That is not to say that the approach of obeying the null constraints is less appealing. It is, however, often more difficult to apply, and it is unlikely that one resampling scheme obeying the constraints of all hypotheses would work in general in the multiple testing framework. An alternative approach would be to resample from a different distribution at each step, obeying the constraints of the null hypotheses imposed at each step. This approach would probably succeed in a fair amount of generality, but even so, two problems would remain. First, it may be difficult to determine the appropriate resampling scheme for testing each subset hypothesis. Second, even if one knew how to resample at each stage, there is increased computation. Our approach avoids these complications.

Before embarking on the general theory, a motivating example is presented to fix ideas.

**Example 4.1 (Testing Correlations)** Suppose  $X_1, \dots, X_n$  are i.i.d. random vectors in  $\mathbb{R}^s$ , so that  $X_i = (X_{i,1}, \dots, X_{i,s})$ . Assume  $E|X_{i,j}|^2 < \infty$  and  $Var(X_{i,j}) > 0$ , so that the correlation between  $X_{1,i}$  and  $X_{1,j}$ , namely  $\rho_{i,j}$  is well-defined. Let  $H_{i,j}$  denote the hypothesis that  $\rho_{i,j} = 0$ , so that the multiple testing problem consists in testing all  $k = \binom{s}{2}$  pairwise correlations. Also let  $T_{n,i,j}$  denote the ordinary sample correlation between variables  $i$  and  $j$ . (Note that we are indexing hypotheses and test statistics now by 2 indices  $i$  and  $j$ .) As noted by Westfall and Young (1993), Example 2.2, p.43, subset pivotality fails here. For example, using results of Aitken (1969) Aitken (1971), if  $s = 3$ ,  $H_{1,2}$  and  $H_{1,3}$  are true but  $H_{2,3}$  is false, the joint limiting distribution of  $n^{1/2}(T_{n,1,2}, T_{n,1,3})$  is bivariate normal with means zero, variances one, and correlation  $\rho_{2,3}$ . As acknowledged by Westfall and Young (1993), their methods fail to address this problem (even asymptotically).

## 4.1 General Results.

We now develop some asymptotic theory. For any  $K \subset \{1, \dots, k\}$ , let  $G_{n,K}(P)$  be the joint distribution of  $T_{n,j}$ ,  $j \in K$  under  $P$ , with corresponding joint c.d.f.  $G_{n,K}(x, P)$ ,  $x \in \mathbb{R}^{|K|}$ .

Also, let  $H_{n,K}(P)$  denote the distribution of  $\max\{T_{n,j} : j \in K\}$  under  $P$ . As in the previous section,  $c_{n,K}(1 - \alpha, P)$  denotes a  $1 - \alpha$  quantile of  $H_{n,K}(P)$ . Also, the symbols  $\xrightarrow{L}$  and  $\xrightarrow{P}$  will denote convergence in law (or distribution) and convergence in probability, respectively.

Typically, the asymptotic behavior of  $G_{n,I(P)}(P)$  is governed by one of the following two possibilities. Either it has a nondegenerate limiting distribution, or it converges weakly to a nondegenerate constant vector (possibly with some components  $-\infty$ ). Actually, this has nothing to do with the fact that we are studying joint distributions of multiple test statistics. For example, suppose we are testing a population mean  $\mu(P)$  is  $\leq 0$  versus  $> 0$  based on an i.i.d. sample  $X_1, \dots, X_n$  from  $P$ , assumed to have a finite nonzero variance  $\sigma^2(P)$ . Consider the test statistic  $T_n = n^{-1/2} \sum_i X_i$ . If  $\mu(P) = 0$ , then  $T_n \xrightarrow{L} N(0, \sigma^2(P))$ . On the other hand, if  $\mu(P) < 0$ , then  $T_n$  converges in probability to  $-\infty$ . Alternatively, if the test statistic is  $T'_n = \max(0, T_n)$ , then if  $\mu(P) = 0$ ,  $T'_n$  converges in distribution to  $\max(0, \sigma(P)Z)$ , where  $Z \sim N(0, 1)$ . But, under  $\mu(P) < 0$ ,  $T'_n$  converges in probability to 0. Note, the two cases exhaust all possibilities under the null hypothesis. On the other hand, for the two-sided problem of testing  $\mu(P) = 0$  versus  $\mu(P) \neq 0$  based on  $|n^{-1/2} \sum_i X_i|$ , a nondegenerate limit law exists under the null hypothesis, and this exhausts all possibilities under the null hypothesis (under the assumption of a finite positive variance).

Formally, we will distinguish between the following assumptions, which are only imposed when  $K = I(P)$  is the set of true hypotheses.

**Assumption A1** Under  $P$ , the joint distribution of the test statistics  $T_{n,j}$ ,  $j \in I(P)$ , has a limiting distribution; that is,

$$G_{n,I(P)}(P) \xrightarrow{L} G_{I(P)}(P) . \quad (27)$$

This implies that, under  $P$ ,  $\max\{T_{n,j} : j \in I(P)\}$  has a limiting distribution, say  $H_{I(P)}(P)$ , with limiting c.d.f.  $H_{I(P)}(x, P)$ . We will assume further that

$$H_{I(P)}(x, P) \quad \text{is continuous and strictly increasing at } x = c_{I(P)}(1 - \alpha, P) . \quad (28)$$

Note that the continuity condition in (28) is satisfied if the  $|I(P)|$  univariate marginal distributions of  $G_{I(P)}(P)$  are continuous. Also, the strictly increasing assumption can be weakened as well, but it holds in all known examples where the continuity assumption holds, as typical limit distributions are of the Gaussian, Chi-squared, etc. type. Actually, the strictly increasing assumption can be removed entirely (see Remark 1.2.1 of Politis et al. (1999)).

**Assumption A2** Under  $P$ ,  $G_{n,I(P)}(P)$  converges weakly to a point mass at  $d = d(P)$ , where  $d = (d_1(P), \dots, d_{|I(P)|}(P))$  is a vector of  $|I(P)|$  components. (In the case where  $d_j(P) = -\infty$ , we mean  $T_{n,j}$  converges in probability under  $P$  to  $-\infty$ .)

Now, we prove a basic result that can be applied to several resampling or asymptotic methods to approximate critical values. Consider the stepdown method presented in Algorithm 2.1 with  $c_{n,K}(1 - \alpha)$  replaced by some estimates  $\hat{c}_{n,K}(1 - \alpha)$ . We will consider some concrete choices later.

**Theorem 4.1** (i) Fix  $P$  and suppose Assumption A1 holds, so that (27) and (28) hold. Assume the estimated critical values  $\hat{c}_{n,K}(1 - \alpha)$  satisfy: for any  $K \supset I(P)$ , the estimates  $\hat{c}_{n,K}(1 - \alpha)$  are bounded below by  $c_{I(P)}(1 - \alpha)$ ; by this we mean, for any  $\epsilon > 0$

$$\hat{c}_{n,K}(1 - \alpha) \geq c_{I(P)}(1 - \alpha) - \epsilon \quad \text{with probability} \rightarrow 1 . \quad (29)$$

Then,  $\limsup_n FWE_P \leq \alpha$ .

(ii) Fix  $P$  and suppose Assumption A1 holds. Assume the estimated critical values are monotone in the sense that

$$\hat{c}_{n,K}(1 - \alpha) \geq \hat{c}_{n,I}(1 - \alpha) \quad \text{whenever } I \subset K . \quad (30)$$

Then, (29) holds for all  $K \supset I(P)$  if it holds in the special case  $K = I(P)$ . Therefore, if Assumption A1 and the monotonicity condition (30) hold, and for any  $\epsilon > 0$

$$\hat{c}_{n,I(P)}(1 - \alpha) \geq c_{I(P)}(1 - \alpha) - \epsilon \quad \text{with probability} \rightarrow 1 , \quad (31)$$

then  $\limsup_n FWE_P \leq \alpha$ .

(iii) Fix  $P$  and suppose Assumption A2 holds. Also, assume the monotonicity condition (30). If, for some  $\epsilon > 0$ ,

$$\hat{c}_{n,I(P)}(1 - \alpha) > \max\{d_j(P) : j \in I(P)\} + \epsilon \quad \text{with probability} \rightarrow 1 , \quad (32)$$

then  $\limsup_n FWE_P = 0$ .

Note that Assumption A1 implies

$$c_{n,I(P)}(1 - \alpha) \rightarrow c_{I(P)}(1 - \alpha) \quad \text{as } n \rightarrow \infty .$$

In part (i) of Theorem 4.1, we replace the monotonicity requirement of Theorem 3.1 by a weak asymptotic monotonicity requirement (29).

In general, the point of Theorem 4.1 is that  $\limsup_n FWE_P \leq \alpha$  regardless of whether the convergence of the null hypotheses satisfies Assumption A1 or Assumption A2, at least under reasonable behavior of the estimated critical values. Moreover, the monotonicity condition (30) assumed in parts (ii) and (iii) will be shown to hold generally for some construction based on the bootstrap and subsampling. Therefore, the crux of proving strong control requires that the estimated critical values satisfy (31); that is, the critical value for testing the intersection hypothesis  $H_{I(P)}$  is consistent in that it leads to a test that asymptotically controls the probability of a Type 1 error. In other words, the problem is essentially reduced to the problem of estimating the critical value for a single (intersection) test without having to worry about the multiple testing issue of controlling the FWE. Thus, the problem of controlling the FWE is reduced to the problem of controlling the Type 1 error of a single test. This will be further clarified for specific choices of estimates of the critical values.

Before applying Theorem 4.1 (ii), (iii), which assumes monotonicity of critical values, we demonstrate consistency without the assumption of monotonicity. In this regard, a simple alternative to Theorem 4.1 (i) is the following.

**Theorem 4.2** Fix  $P$  and suppose Assumption A1 holds. Suppose (29) holds for  $K = I(P)$ , that is, for any  $\epsilon > 0$

$$\hat{c}_{n,I(P)}(1 - \alpha) \geq c_{I(P)}(1 - \alpha) - \epsilon \quad \text{with probability} \rightarrow 1 . \quad (33)$$

Further, suppose the test is consistent in the sense that, for any hypothesis  $H_j$  with  $j \notin I(P)$ , the probability of rejecting  $H_j$  by the stepdown procedure tends to one. This happens, for example, if the critical values  $\hat{c}_{n,K}$  are bounded in probability while  $T_{n,j} \rightarrow \infty$  if  $j \notin I(P)$ . Then,  $\limsup_n FWE_P \leq \alpha$ .

**Example 4.2 (Example 3.2, revisited)** In the setup of Example 3.2, suppose the observations are real-valued, and consider a test of  $H_j$  based on

$$T_{n,j} = n^{1/2} |\bar{X}_j - \bar{X}_{k+1}| ,$$

where  $\bar{X}_j = n_j^{-1} \sum_i X_{i,j}$ . Suppose we use the permutation test where at stage  $j$  for testing  $H_{K_j}$ , only permutations of observations  $X_{i,j}$  with  $j \in K$  and  $X_{i,k+1}$  are used. Assume  $n_j/n \rightarrow \lambda_j \in (0, 1)$ . Let  $\mu(P_j)$  denote the true mean of  $P_j$ , assumed to exist; also assume the variance of  $P_i$  is finite. Then, Theorem 4.2 applies to any  $P$  for which, if  $j \notin I(P)$ ,  $\mu(P_j) \neq \mu(P_{k+1})$  (which, of course, is not the same as  $P_j \neq P_{k+1}$ ). Indeed,  $T_{n,j} \rightarrow \infty$  in probability then. Also, using arguments as in Romano (1990),  $\hat{c}_{n,K}(1 - \alpha)$  is bounded in probability for any  $K$ , because asymptotically it behaves like the  $1 - \alpha$  quantile of the maximum of  $|K|$  normal variables. Therefore, asymptotic control of the FWE persists. However, if the distributions differ but the means are the same, the test statistic should be designed to capture arbitrary differences in distribution, such as a two-sample Kolmogorov Smirnov test statistic (unless one really wants to pick up just differences in the mean, but then the null hypothesis should reflect this.)

## 4.2 A Bootstrap Construction

We now specialize a bit and will develop a concrete construction based on the bootstrap. For now, we suppose hypothesis  $H_j$  is specified by  $\{P : \theta_j(P) \leq 0\}$  for some real-valued parameter  $\theta_j$ . Suppose  $\hat{\theta}_{n,j}$  is an estimate of  $\theta_j$ . Also, let  $T_{n,j} = \tau_n \hat{\theta}_{n,j}$  for some nonnegative (nonrandom) sequence  $\tau_n \rightarrow \infty$ . The sequence  $\tau_n$  is introduced for asymptotic purposes so that a limiting distribution for  $\tau_n \hat{\theta}_{n,j}$  exists when  $\theta_j(P) = 0$ .

**Remark 4.1** Typically,  $\tau_n = n^{1/2}$ . Also, it is possible to let  $\tau_n$  vary with the hypothesis  $j$ . Extensions to cases where  $\tau_n$  depends on  $P$  are also possible, using ideas in Bertail et al. (1999).

The bootstrap method relies on its ability to approximate the joint distribution of  $\{\tau_n[\hat{\theta}_{n,j} - \theta_j(P)] : j \in K\}$ , whose distribution we denote by  $J_{n,K}(P)$ . We will assume the normalized estimates satisfy the following.

**Assumption B1(i)**  $J_{n,I(P)}(P) \xrightarrow{L} J_{I(P)}(P)$ , a nondegenerate limit law.

Let  $L_{n,K}(P)$  denote the distribution under  $P$  of  $\max\{\tau_n[\hat{\theta}_{n,j} - \theta_j(P)] : j \in K\}$ , with corresponding distribution function  $L_{n,K}(x, P)$  and  $\alpha$ -quantile

$$b_{n,K}(\alpha, P) = \inf\{x : L_{n,K}(x, P) \geq \alpha\} .$$

Assumption B1 implies  $L_{n,I(P)}(P)$  has a limiting distribution  $L_{I(P)}(P)$ .

We will further assume

**Assumption B1(ii)**  $L_{J(P)}(P)$  is continuous and strictly increasing on its support.

Under Assumption B1, it follows that, when  $K = J(P)$ ,

$$b_{n,K}(1 - \alpha, P) \rightarrow b_K(1 - \alpha, P) , \quad (34)$$

where  $b_K(\alpha, P)$  is the  $\alpha$ -quantile of the limiting distribution  $L_K(P)$ .

Assume B1 holds. If  $P$  satisfies at least one  $\theta_j(P)$  is exactly 0, then A1 holds. On the other hand, if  $P$  satisfies all  $\theta_j(P) < 0$  among the  $\theta_j(P)$  which are  $\leq 0$ , then A2 holds. Indeed, if  $\tau_n(\hat{\theta}_{n,j} - \theta_j(P))$  converges to a limit law and  $\tau_n\theta_j(P) \rightarrow -\infty$ , then  $\tau_n\hat{\theta}_{n,j} \rightarrow -\infty$  in probability.

Let  $\hat{Q}_n$  be some estimate of  $P$ . For i.i.d. data,  $\hat{Q}_n$  is typically taken to be the empirical distribution, or possibly a smoothed version. For time series or data-dependent situations, block bootstrap methods should be employed; see Lahiri (2003). Then, a nominal  $1 - \alpha$  level bootstrap confidence region for the subset of parameters  $\{\theta_j(P) : j \in K\}$  is given by

$$\begin{aligned} & \{(\theta_j : j \in K) : \max_{j \in K} \tau_n[\hat{\theta}_{n,j} - \theta_j] \leq b_{n,K}(1 - \alpha, \hat{Q}_n)\} \\ & = \{(\theta_j : j \in K) : \theta_j \geq \hat{\theta}_{n,j} - \tau_n^{-1}b_{n,K}(1 - \alpha, \hat{Q}_n)\} . \end{aligned}$$

So a value of 0 for  $\theta_j(P)$  falls outside the region iff  $\tau_n\hat{\theta}_{n,j} > b_{n,K}(1 - \alpha, \hat{Q}_n)$ . By the usual duality of confidence sets and hypothesis tests, this suggests the use of the critical value

$$\hat{c}_{n,K}(1 - \alpha) = b_{n,K}(1 - \alpha, \hat{Q}_n) , \quad (35)$$

at least if the bootstrap is a valid asymptotic approach for confidence region construction.

Note that, regardless of asymptotic behavior, the monotonicity assumption (30) is always satisfied for the choice (35). Indeed, for any  $Q$  and if  $I \subset K$ ,  $b_{n,I}(1 - \alpha, Q)$  is the  $1 - \alpha$  quantile under  $Q$  of the maximum of  $|I|$  variables, while  $b_{n,K}(1 - \alpha, Q)$  is the  $1 - \alpha$  quantile of these same  $|I|$  variables together with  $|K| - |I|$  variables.

Therefore, in order to apply Theorem 4.1 to conclude  $\limsup_n \text{FWE}_P \leq \alpha$ , it is now only necessary to study the asymptotic behavior of  $b_{n,K}(1 - \alpha, \hat{Q}_n)$  in the case  $K = J(P)$ . For this, we further assume the usual conditions for bootstrap consistency when testing the *single* hypothesis that  $\theta_j(P) \leq 0$  for all  $j \in J(P)$ ; that is, we assume the bootstrap consistently

estimates the joint distribution of  $\tau_n[\hat{\theta}_{n,j} - \theta_j(P)]$  for  $j \in I(P)$ . Specifically, consider the following.

**Assumption B2** For any metric  $\rho$  metrizing weak convergence on  $\mathbb{R}^{|J(P)|}$ ,

$$\rho\left(J_{n,I(P)}(P), J_{n,I(P)}(\hat{Q}_n)\right) \xrightarrow{P} 0.$$

**Theorem 4.3** Fix  $P$  satisfying assumption B1. Let  $\hat{Q}_n$  be an estimate of  $P$  satisfying B2. Consider the stepdown method in Algorithm 2.1 with  $c_{n,K}(1 - \alpha)$  replaced by  $b_{n,K}(1 - \alpha, \hat{Q}_n)$ .

(i) Then,  $\limsup_n FWE_P \leq \alpha$ .

(ii) Suppose  $B_1$  and  $B_2$  hold when  $I(P)$  is replaced by any subset  $K$ . If  $P$  is such that  $j \notin I(P)$ , i.e.  $H_j$  is false and  $\theta_j(P) > 0$ , then the probability that the stepdown method rejects  $H_j$  tends to one.

**Example 4.3 (Continuation of Example 4.1)** The analysis of sample correlations is a special case of the smooth function model studied in Hall (1992), and the bootstrap approach is valid for such models.

**Remark 4.2** The main reason why the bootstrap works here can be traced to the simple result Theorem 3.1. The bootstrap approach, by resampling from a fixed distribution, generates monotone critical values. Therefore, since we know how to construct valid bootstrap tests for each intersection hypothesis, this leads to valid multiple tests. But we learn more. If we use a bootstrap approach such that each intersection test has a rejection probability equal to  $\alpha + O(\epsilon_n)$ , then we also can deduce  $\limsup_n FWE_P \leq \alpha + O(\epsilon_n)$ . In other words, if a bootstrap method has good performance for the construction of single tests, then this translates into good performance of the bootstrap for constructing stepdown multiple tests.

**Remark 4.3** The bootstrap can also give dramatic finite-sample gains by accommodating non-normalities, even when the test statistics are independent; for example, see Westfall and Young (1993, page 162) and Westfall and Wolfinger (1997).

**Remark 4.4** Typically, the asymptotic behavior of a test procedure when  $P$  is true will satisfy that it is consistent in the sense that all false hypotheses will be rejected with probability tending to one (as is the case under Theorem 4.3). However, one can also study the behavior of procedures against contiguous alternatives so that not all false hypotheses are rejected with probability tending to one under such sequences. But, of course, if alternative hypotheses are in some sense close to their respective null hypotheses, then the procedures will typically reject even fewer hypotheses, and so the limiting probability of any false rejection under a sequence of contiguous alternatives should then be bounded by  $\alpha$ .

**Remark 4.5** In addition to constructing tests that control the FWE, one typically would like to choose test statistics that lead to procedures that are balanced in the sense that all tests have about the same power. As argued by Beran (1988a), Tu and Zhou (2000), and Rogers and Hsu (2001), balance can be desirable. Alternatively, lack of balance may be desirable so that certain tests are given more weight; see Westfall and Young (1993, page 162) and Westfall and Wolfinger (1997). While the goal of this paper has been the evaluation of significance while maintaining strong control based on given test statistics, achieving balance is best handled by appropriate choice of test statistics. For example, transforming test statistics to  $p$ -values and then using the negative  $p$ -values as the basic statistics will lead to better balance. Quite generally, Beran's pre-pivoting transformation can lead to balance; see Beran (1988a, 1988b). The assumptions of our theorem must then hold for the transformed test statistics. Alternatively, balance can sometimes be achieved by studentization. The construction developed in this subsection can be extended to the case of studentized test statistics. The details are straightforward and left to the reader.

We now briefly consider the two-sided case. Suppose  $H_j$  specifies  $\theta_j(P) = 0$  against the alternative  $\theta_j(P) \neq 0$ . Let  $L'_{n,K}(P)$  denote the distribution under  $P$  of  $\max\{\tau_n|\hat{\theta}_{n,j} - \theta_j(P)| : j \in K\}$  with corresponding distribution function  $L'_{n,K}(x, P)$  and  $\alpha$ -quantile

$$b'_{n,K}(\alpha, P) = \inf\{x : L'_{n,K}(x, P) \geq \alpha\}.$$

Accordingly,  $L'_K(P)$  denotes the limiting distribution of  $L'_{n,K}(P)$ . Finally, let  $T'_{n,j} = \tau_n|\hat{\theta}_{n,j}|$ .

**Theorem 4.4** *Fix  $P$  satisfying assumption B1, but with  $L_{I(P)}(P)$  in B1(ii) replaced by  $L'_{I(P)}(P)$ . Let  $\hat{Q}_n$  be an estimate of  $P$  satisfying B2. Consider the stepdown method in Algorithm 2.1 using the test statistics  $T'_{n,j}$  and with  $c_{n,K}(1 - \alpha)$  replaced by  $b'_{n,K}(1 - \alpha, \hat{Q}_n)$ .*

(i) *Then,  $\limsup_n FWE_P \leq \alpha$ .*

(ii) *Suppose  $B_1$  and  $B_2$  hold when  $I(P)$  is replaced by any subset  $K$  (and  $L$  replaced by  $L'$ ). If  $P$  is such that  $j \notin I(P)$ , i.e.  $H_j$  is false and  $\theta_j(P) \neq 0$ , then the probability that the stepdown method rejects  $H_j$  tends to one.*

(iii) *Moreover, if the above algorithm rejects  $H_j$  and it is declared that  $\theta_j > 0$  when  $\hat{\theta}_{n,j} > 0$ , the the probability of making a Type 3 error (i.e. of declaring  $\theta_j(P)$  is positive when it is negative or declaring it negative when it is positive) tends to 0.*

An alternative approach to the two-sided case is to balance the tails of the bootstrap distribution of the original estimates (without the absolute values) separately. An analogous result would hold. The comparison of these approaches in the case of a single test is made in Hall (1992).

The result (iii) shows that the directional error is asymptotically negligible. It would be more interesting to obtain both finite sample results, as well as studying the behavior of the

directional error under contiguous alternatives so that the problem is no longer asymptotically degenerate; future work will consider these problems. For references to the literature on controlling the directional error as well as some finite sample results, see Finner (1999).

### 4.3 A Subsampling Construction

In this section, we present an alternative construction that applies under weaker conditions than the bootstrap. We now assume that we have available an i.i.d. sample  $X_1, \dots, X_n$  from  $P$ , and  $T_{n,j} = T_{n,j}(X_1, \dots, X_n)$  is the test statistic we wish to use for testing  $H_j$ . To describe the test construction, fix a positive integer  $b \leq n$  let  $Y_1, \dots, Y_{N_n}$  be equal to the  $N_n = \binom{n}{b}$  subsets of  $\{X_1, \dots, X_n\}$ , ordered in any fashion. Let  $T_{b,j}^{(i)}$  be equal to the statistic  $T_{b,j}$  evaluated at the data set  $Y_i$ . Then, for any subset  $K \subset \{1, \dots, k\}$ , the joint distribution of  $(T_{n,j} : j \in K)$  can be approximated by the empirical distribution of the  $\binom{n}{b}$  values  $(T_{b,j}^{(i)} : j \in K)$ . In other words, for  $x \in \mathbf{R}^k$ , the true joint c.d.f. of the test statistics evaluated at  $x$ ,

$$G_{n,\{1,\dots,k\}}(x, P) = P\{T_{n,1} \leq x_1, \dots, T_{n,k} \leq x_k\}$$

is estimated by the subsampling distribution

$$\hat{G}_{n,\{1,\dots,k\}}(x) = \binom{n}{b}^{-1} \sum_i I\{T_{b,1}^{(i)} \leq x_1, \dots, T_{b,k}^{(i)} \leq x_k\}. \quad (36)$$

Note that the marginal distribution of any subset  $K \subset \{1, \dots, k\}$ ,  $G_{n,K}(P)$ , is then approximated by the marginal distribution induced by (36) on that subset of variables. So,  $\hat{G}_{n,K}$  refers to the empirical distribution of the values  $(T_{n,j}^{(i)} : j \in K)$ . (In essence, one only has to estimate one joint sampling distribution for all the test statistics because this then induces that of any subset, even though we are not assuming anything like subset pivotality).

Similarly, the estimate of the whole joint distribution of test statistics induces an estimate for the distribution of the maximum of test statistics. Specifically,  $H_{n,K}(P)$  is estimated by the empirical distribution  $\hat{H}_{n,K}(x)$  of the values  $\max(T_{n,j}^{(i)} : j \in K)$ ; that is,

$$\hat{H}_{n,K}(x) = \binom{n}{b}^{-1} \sum_i I\{\max(T_{b,j}^{(i)} : j \in K) \leq x\}.$$

Also, let

$$\hat{c}_{n,K}(1 - \alpha) = \inf\{x : \hat{H}_{n,K}(x) \geq 1 - \alpha\}$$

denote the estimated  $1 - \alpha$  quantile of the maximum of test statistics  $T_{n,j}$  with  $j \in K$ .

Note the monotonicity of the critical values: for  $I \subset K$

$$\hat{c}_{n,K}(1 - \alpha) \geq \hat{c}_{n,I}(1 - \alpha); \quad (37)$$

and so the monotonicity assumption in Theorem 4.1 holds (and also compare with (4)).

This leads us to consider the idealized stepdown algorithm with  $c_{n,K}(1 - \alpha, P)$  replaced by the estimates  $\hat{c}_{n,K}(1 - \alpha)$ . The following result proves consistency and strong control of this subsampling approach. Note, in particular, that Assumption B2 is not needed here at all, a reflection of the fact that the bootstrap requires much stronger conditions for consistency; see Politis et al. (1999). Also notice that we do not even need to assume that there exists a  $P$  for which all hypotheses are true.

**Theorem 4.5** *Suppose Assumption A1 holds. Let  $b/n \rightarrow 0$ ,  $\tau_b/\tau_n \rightarrow 0$ , and  $b \rightarrow \infty$ .*

(i). *The subsampling approximation satisfies*

$$\rho \left( \hat{G}_{n,I(P)}, G_{n,I(P)}(P) \right) \xrightarrow{P} 0, \quad (38)$$

for any metric  $\rho$  metrizing weak convergence on  $\mathbb{R}^{|I(P)|}$ .

(ii) *The subsampling critical values satisfy*

$$\hat{c}_{n,I(P)}(1 - \alpha) \xrightarrow{P} c_{I(P)}(1 - \alpha). \quad (39)$$

(iii). *Therefore, using Algorithm 2.1 with  $c_{n,K}(1 - \alpha, P)$  replaced by the estimates  $\hat{c}_{n,K}(1 - \alpha)$  results in  $\limsup_n FWE_P \leq \alpha$ .*

**Example 4.4 (Cube root asymptotics)** Kim and Pollard (1990) show that a general class of  $M$ -estimators converge at rate  $\tau_n = n^{1/3}$  to a non-normal limiting distribution. As a result, inconsistency of the bootstrap typically follows. Rodríguez-Poo et al. (2001) demonstrate the consistency of the subsampling method for constructing hypothesis tests for a single null hypothesis. By similar arguments, the validity of the subsampling construction of Theorem 4.5 in the context of cube root asymptotics can be established.

The above approach can be extended to dependent data. For example, if the data form a stationary sequence, we would only consider the  $n - b + 1$  subsamples of the form  $(X_i, X_{i+1}, \dots, X_{i+b-1})$ . Generalizations for nonstationary time series, random fields, and point processes are further treated in Politis et al. (1999).

## 5 Two Simulation Studies

### 5.1 Testing Means

This section presents a small simulation study in the context of testing population means. We generate random vectors  $X_1, \dots, X_{100}$  from a  $k$ -dimensional multivariate normal distribution with mean vector  $\theta = (\theta_1, \dots, \theta_k)'$ . The values of  $k$  are  $k = 10$  and  $k = 40$ . Each null hypothesis is  $H_j: \theta_j \leq 0$  and each alternative hypothesis is one-sided. We apply the stepdown

bootstrap construction of Subsection 4.2, resampling from the empirical distribution. In the spirit of Remark 4.5 we employ the studentized test statistics  $T_{n,j} = \sqrt{n}\bar{X}_j/s_j$ , where  $\bar{X}_j$  and  $s_j$  are the usual sample average and sample standard deviation of  $j$ th sample, respectively. In addition, we also include the Holm method in the study. The nominal FWE levels are  $\alpha = 0.05$  and  $\alpha = 0.1$ . Performance criteria are the empirical FWE and the (average) number of false hypotheses that are rejected.

We consider three scenarios for the mean vector of the multivariate normal distribution. In the first scenario, all means  $\theta_j$  are equal to 0. In the second scenario, half of the means are equal to 0 and the other half are equal to 0.25. In the third scenario, all means are equal to 0.25.

We consider three scenarios for the covariance matrix of the multivariate normal distribution. In the first scenario, the covariance matrix is the identity matrix. In the second scenario, all the variances are equal to 1 and all the correlations are equal to  $\rho = 0.5$ . In the third scenario, all the variances are equal to 1 and all the correlations are equal to  $\rho = 0.9$ . We would expect that our stepwise method performs similarly to Holm in the first scenario but rejects more false null hypotheses in the latter two scenarios.

Tables 1 and 2 report the results based on 10,000 repetitions. The number of bootstrap resamples is  $B = 1,000$ . The results demonstrate the good control of the FWE in finite sample and the increased power of the stepdown method compared to the Holm method in case of a positive common correlation  $\rho$ .

Note that the FWE control of the Holm method for the case  $\rho = 0$  somewhat deteriorates when the number of hypotheses tested increases from  $k = 10$  to  $k = 40$  while this does not happen with the stepdown method. The reason for this behavior of the Holm method is that individual  $p$ -values are computed using the asymptotic standard normal distribution of the  $t$ -statistic under the null. Since the true distribution under the null is  $t_{99}$ , the  $p$ -values are somewhat anticonservative in finite sample and when  $k$  increases from 10 to 40 this effect is apparently magnified. (Of course, using the  $t_{99}$  distribution instead to compute individual  $p$ -values would correspond to knowing the parametric nature of the underlying probability mechanism, which is not realistic.)

Further note that relative advantage in terms of power of the stepdown method does not seem to diminish when the number of hypotheses tested increases from  $k = 10$  to  $k = 40$ . For example, consider the case where  $\alpha = 0.1$  and  $\rho = 0.9$ . When  $k = 10$  the stepdown method on average rejects about 20% more false hypotheses compared to the Holm method. When  $k = 40$  the improvement is about 50% instead. Of course, this is an observation restricted to the particular data generating mechanism used in our simulation study and should not be interpreted as a general theoretical statement.

## 5.2 Testing Correlations

This section presents a small simulation study in the context of Example 4.1. We generate random vectors  $X_1, \dots, X_n$  from a 10-dimensional multivariate normal distribution. Hence, there are a total of  $k = \binom{10}{2} = 45$  pairwise correlations to test. The values for the sample size are  $n = 50$  and  $n = 100$ . Each null hypothesis is  $H_{i,j}: \rho_{i,j} = 0$  and each alternative hypothesis is two-sided. We apply the stepdown bootstrap construction of Subsection 4.2, resampling from the empirical distribution. The test statistics are given by  $T_{n,i,j} = \sqrt{n}\hat{\rho}_{i,j}$ , where  $\hat{\rho}_{i,j}$  is the usual sample correlation between the  $i$ th and the  $j$ th sample. As a special case, we also look at the corresponding single-step method. The nominal FWE levels are  $\alpha = 0.05$  and  $\alpha = 0.1$ . Performance criteria are the empirical FWE and the (average) number of false hypotheses that are rejected.

We consider three scenarios. In the first scenario, all correlations are equal to 0. In the second scenario, all  $\rho_{1,j}$  are equal to 0.3, for  $j = 2, \dots, 10$ , and the remaining correlations are equal to 0. In the third scenario, all correlations are equal to 0.3.

Table 3 reports the results based on 10,000 repetitions. The number of bootstrap resamples is  $B = 1,000$ . The results demonstrate the good control of the FWE in finite sample and the increased power of the stepdown method compared to the single-step method.

## 6 Empirical Application

Westfall and Young (1993, Example 6.4) apply a multiple testing method for 10 pairwise correlations. Each null hypothesis is that corresponding pairwise population correlation is equal to zero; and each alternative hypothesis is two-sided. The reader is referred to their Example 6.4 for the details of the real data set. Westfall and Young (1993) carry out a bootstrap multiple test under the assumption of complete independence. As they admit, this is a conservative approach in general. Instead we apply the stepdown bootstrap construction of Subsection 4.2, resampling from the empirical distribution. For each null hypothesis, the stepdown construction yields an adjusted  $p$ -value: it is given by the smallest FWE-level  $\alpha$  at which the construction rejects this particular hypothesis.

Table 4 compares the adjusted  $p$ -values of Westfall and Young (1993) to ours. The conservativeness of the Westfall and Young (1993) method can be clearly appreciated.

## 7 Concluding Remarks.

We have shown how computationally feasible stepdown methods can be constructed to control the FWE in a fair amount of generality. Further study is needed to study the control of directional errors. Further work will focus on a similar treatment for stepup procedures. We

also would like to extend our results to show how resampling can be used to estimate the dependence structure of the test statistics in order to obtain improved methods that control the false discovery rate of Benjamini and Hochberg (1995). Some results are obtained in Benjamini and Yekutieli (2001), but they also assume the subset pivotality condition. By extending our work, we hope to avoid such conditions.

## A Proofs

### Proof of Theorem 2.1

Consider the event that a true hypothesis is rejected, so that for some  $j \in I(P)$ , hypothesis  $H_j$  is rejected. Let  $\hat{j}$  be the (random) smallest index  $j$  in the algorithm where this occurs, so that

$$T_{n,r_{\hat{j}}} > c_{n,K_{\hat{j}}}(1 - \alpha) . \quad (40)$$

Since  $K_{\hat{j}} \supset I(P)$ , assumption (5) implies

$$c_{n,K_{\hat{j}}}(1 - \alpha) \geq c_{n,I(P)}(1 - \alpha) \geq c_{n,I(P)}(1 - \alpha, P)$$

and so

$$T_{n,r_{\hat{j}}} > c_{n,I(P)}(1 - \alpha, P) .$$

Furthermore, by definition of  $\hat{j}$ ,

$$T_{n,r_{\hat{j}}} = \max(T_{n,j}, j \in K_{\hat{j}}) = \max(T_{n,j}, j \in I(P)) ,$$

and so the event that a false rejection occurs under  $P$  implies the event

$$\max(T_{n,j}, j \in I(P)) > c_{n,I(P)}(1 - \alpha, P) . \quad (41)$$

Therefore, the probability of a Type 1 error is bounded above by the probability of the event (41), which by definition has probability bounded above by  $\alpha$ . The proof of (ii) is obvious because the procedure becomes more conservative. The proof of (iii) holds by the proof of (i) upon replacing the constants  $c_{n,K_{\hat{j}}}(1 - \alpha)$  by  $d_{n,K_{\hat{j}}}(1 - \alpha)$ . ■

### Proof of Corollary 2.1

We verify the conditions for  $d_{n,K_j}(1 - \alpha)$  when  $d_{n,K_j}(1 - \alpha) = c_{n,K_j}^*(1 - \alpha)$  in Theorem 2.1 (ii) and (iii). Clearly,

$$c_{n,K}^*(1 - \alpha) \geq c_{n,I}(1 - \alpha) .$$

Also, for  $K \supset I(P)$ ,

$$c_{n,K}^*(1 - \alpha) = \max\{c_{n,J}(1 - \alpha) : J \subset K\} \geq \max\{c_{n,J}(1 - \alpha) : J \subset I(P)\} = c_{n,I(P)}^*(1 - \alpha) ,$$

and so (7) holds. ■

### Proof of Theorem 2.2

To prove (i), let  $\hat{j}$  be the smallest (random) index  $j$  such that  $T_{n,r_{\hat{j}}} > \tilde{c}_{n,K_{\hat{j}}}(1 - \alpha)$ . But,  $K_{\hat{j}} \supset I(P)$  and so

$$\tilde{c}_{n,K_{\hat{j}}}(1 - \alpha) \geq \tilde{c}_{n,I(P)}(1 - \alpha) \geq c_{n,I(P)}(1 - \alpha, P) .$$

So, the event that a false rejection occurs under  $P$  implies the event

$$\max(T_{n,j}, j \in I(P)) > c_{n,I(P)}(1 - \alpha, P) , \quad (42)$$

which has probability bounded by  $\alpha$ . The proof of (ii) is obvious because the procedure becomes more conservative. ■

### Proof of Theorem 3.1

As in the argument of Theorem 2.1, the event a false rejection occurs is the event

$$\max\{T_{n,j} : j \in I(P)\} > \hat{c}_{n,K_{\hat{j}}}(1 - \alpha) , \quad (43)$$

where  $\hat{j}$  is the smallest (random) index where a false rejection occurs. Since  $K_{\hat{j}} \supset I(P)$ ,

$$\hat{c}_{n,K_{\hat{j}}}(1 - \alpha) \geq \hat{c}_{n,I(P)}(1 - \alpha) \quad (44)$$

and so (i) follows. Part (ii) follows immediately from (i). ■

### Proof of Theorem 4.1

As in the proofs of Theorems 2.1 and 3.1, namely (43), it suffices to show

$$\limsup_n P\{\max\{T_{n,j} : j \in I(P)\} > \hat{c}_{n,K_{\hat{j}}}(1 - \alpha)\} \leq \alpha .$$

But assumption (29) implies

$$\hat{c}_{n,K_{\hat{j}}}(1 - \alpha) \geq c_{I(P)}(1 - \alpha) - \epsilon \quad \text{with probability} \rightarrow 1 .$$

Therefore, using Assumption A1, the limit superior of the probability of a false rejection is bounded above by

$$\limsup_n FWE_P \leq P\{\max(T_j, j \in I(P)) > c_{I(P)}(1 - \alpha) - \epsilon\} ,$$

where  $(T_j, j \in I(P))$  denote variables whose joint distribution is  $G_{I(P)}(P)$ . But letting  $\epsilon \rightarrow 0$ , the right side of the last expression becomes

$$1 - H_{I(P)}(c_{I(P)}(1 - \alpha), P) = 1 - (1 - \alpha) = \alpha .$$

To prove (ii), since (29) holds when  $K = I(P)$ , then it must hold for any  $K$  containing  $I(P)$ , by assumption (30).

To prove (iii), the probability of false rejection, i.e. the event (43), is again bounded by the probability of the event

$$\max\{T_{n,j} : j \in I(P)\} > \hat{c}_{n,I(P)}(1 - \alpha) ,$$

which converges to 0 by Assumption A2 and (32). ■

### Proof of Theorem 4.2

Following the proof of Theorem 4.1 (i), the random index  $\hat{j}$  is equal to  $k - |I(P)| + 1$  with probability tending to one, and this index is no longer random; that is, with probability tending to one, we first reject all false hypotheses and then commit a false rejection when we get to the stage where we are testing the  $|I(P)|$  true hypotheses. But then, Assumption A1 and (33) allow us to conclude control of the FWE. ■

### Proof of Theorem 4.3

To prove (i), fix  $P$  and assume  $\theta_j(P) = 0$  for at least one  $j \in I(P)$ . Then, by the comments leading up to the statement of the theorem, the conditions of Theorem 4.1 (ii) are satisfied if we can verify

$$b_{n,I(P)}(1 - \alpha, \hat{Q}_n) \xrightarrow{P} c_{I(P)}(1 - \alpha) .$$

But by the Continuous Mapping Theorem, the assumption B2 implies

$$\rho_1 \left( L_{n,I(P)}(P), L_{n,I(P)}(\hat{Q}_n) \right) \xrightarrow{P} 0 ,$$

where  $\rho_1$  is any metric metrizing weak convergence on  $\mathbb{R}$ . Furthermore,  $L_{n,I(P)}(P)$  converges weakly to a continuous limit law by Assumption B1, and so

$$b_{n,I(P)}(1 - \alpha, \hat{Q}_n) \rightarrow b_{I(P)}(1 - \alpha, P)$$

and

$$b_{n,I(P)}(1 - \alpha, P) \xrightarrow{P} b_{I(P)}(1 - \alpha, P) .$$

So it suffices to show

$$\liminf b_{n,I(P)}(1 - \alpha, P) \rightarrow c_{I(P)}(1 - \alpha, P) . \tag{45}$$

But, for  $\theta_j(P) \leq 0$ ,

$$\tau_n[\hat{\theta}_{n,j} - \theta_j(P)] \geq \tau_n \hat{\theta}_{n,j} = T_n ,$$

which implies

$$b_{n,I(P)}(1 - \alpha, P) \geq c_{n,I(P)}(1 - \alpha, P) .$$

But, the right term converges to  $c_{I(P)}(1 - \alpha, P)$ , and so (45) follows.

Next, assume  $P$  has  $\theta_j(P) < 0$  for all  $j \in I(P)$ . Then, we just need to verify the conditions of Theorem 4.1 (iii). All that is left to verify is, for some  $\epsilon > 0$ ,

$$b_{n,I(P)}(1 - \alpha, \hat{Q}_n) > \max\{d_j(P) : j \in I(P)\} + \epsilon$$

with probability tending to one. But, the right side here is  $-\infty$  (for any finite  $\epsilon$ ), so it just suffices to verify the left side is  $O_P(1)$ . But, by B2, it suffices to show  $b_{n,I(P)}(1 - \alpha, P)$  is bounded away from  $-\infty$ , which follows by (34).

To prove (ii), the assumptions imply that, for any  $K \supset I(P)$ ,

$$b_{n,K}(1 - \alpha, \hat{Q}_n) \xrightarrow{P} b_K(1 - \alpha, P) < \infty .$$

But,

$$\max(T_{n,j} : j \in K) \geq T_{n,j} = \tau_n \hat{\theta}_{n,j} \xrightarrow{P} \infty ,$$

since  $\hat{\theta}_{n,j} \xrightarrow{P} \theta_j(P) > 0$  and  $\tau_n \rightarrow \infty$ . Therefore, with probability tending to one, for any  $K \supset I(P)$ ,

$$\max(T_{n,j} : j \in K) \geq b_{n,K}(1 - \alpha, \hat{Q}_n) ,$$

and the result follows. ■

**Proof of Theorem 4.4** The proof is completely analogous to the proof of Theorem 4.3. The only additional fact needed to prove (iii) is that, when  $\theta_j(P) > 0$ ,  $\tau_n \hat{\theta}_{n,j} > 0$  with probability tending to one, and similarly for  $\theta_j(P) < 0$ . Indeed, assumption  $B_1(i)$  implies  $\tau_n(\hat{\theta}_{n,j} - \theta_j(P))$  has a limiting distribution, which implies  $\tau_n \hat{\theta}_{n,j} \xrightarrow{P} \infty$  if  $\theta_j(P) > 0$ . ■

#### Proof of Theorem 4.5

The proof of (i) is the essential subsampling argument, which derives from (36) being a U-statistic; see Politis et al. (1999), Theorem 2.6.1, where one statistic is treated, but the argument is extendable to the simultaneous estimation of the joint distribution. The result (ii) follows as well. To verify (iii), apply Theorem 4.1 (ii). The monotonicity requirement follows by (37) and (31) follows by (ii). ■

## References

- Aitken, M. (1969). Some tests for correlation matrices. *Biometrika*, 56:443–446.
- Aitken, M. (1971). Correction to ‘some tests for correlation matrices’. *Biometrika*, 58:245.
- 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(1):289–300.
- Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29(4):1165–1188.
- Beran, R. (1986). Simulated power functions. *Annals of Statistics*, 14:151–173.
- Beran, R. (1988a). Balanced simultaneous confidence sets. *Journal of the American Statistical Association*, 83:679–686.
- Beran, R. (1988b). Prepivoting test statistics: a bootstrap view of asymptotic refinements. *Journal of the American Statistical Association*, 83:687–697.
- Bertail, P., Politis, D., and Romano, J. (1999). On subsampling estimators with unknown rate of convergence. *Journal of the American Statistical Association*, 94:569–579.
- Dudoit, S., Shaffer, J., and Boldrick, J. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18:71–103.
- Finner, H. (1999). Stepwise multiple test procedures and control of directional errors. *Annals of Statistics*, 27:274–289.
- Finner, H. and Roters, M. (1998). Asymptotic comparison of step-down and step-up multiple test procedures based on exchangeable test statistics. *Annals of Statistics*, 26:505–524.
- Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer, New York.
- Hall, P. and Wilson, S. (1991). Two guidelines for bootstrap hypothesis testing. *Biometrics*, 47:757–762.
- Hochberg, Y. and Tamhane, A. (1987). *Multiple Comparison Procedures*. Wiley, New York.
- Hoeffding, W. (1952). The large-sample power of tests based on permutations of observations. *Annals of Mathematical Statistics*, 23:169–192.
- Hommel, G. (1986). Multiple test procedures for arbitrary dependence structures. *Metrika*, 33:321–336.
- Kim, J. and Pollard, D. B. (1990). Cube root asymptotics. *Annals of Statistics*, 18:191–219.
- Lahiri, S. (2003). *Resampling Methods for Dependent Data*. Springer, New York.

- Lehmann, E., Romano, J. P., and Shaffer, J. (2003). On optimality of stepdown and stepup multiple test procedures. Technical Report 2003-12, Department of Statistics, Stanford University.
- Marcus, R., Peritz, E., and Gabriel, K. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63:655–660.
- Petrondas, D. and Gabriel, K. (1983). Multiple comparisons by rerandomization tests. *Journal of the American Statistical Association*, 78(384):949–957.
- Politis, D. N., Romano, J. P., and Wolf, M. (1999). *Subsampling*. Springer, New York.
- Rodríguez-Poo, J., Delgado, M., and Wolf, M. (2001). Subsampling inference in cube root asymptotics with an application to Manski’s maximum score estimator. *Economics Letters*, 73:241–250.
- Rogers, J. and Hsu, J. (2001). Multiple comparisons of biodiversity. *Biometrical Journal*, 43:617–625.
- Romano, J. (1988). A bootstrap revival of some nonparametric distance tests. *Journal of the American Statistical Association*, 83(403):698–708.
- Romano, J. (1989). Bootstrap and randomization tests of some nonparametric hypotheses. *Annals of Statistics*, 17:141–159.
- Romano, J. (1990). On the behavior of randomization tests without a group invariance assumption. *Journal of the American Statistical Association*, 85(411):686–692.
- Troendle, J. (1995). A stepwise resampling method of multiple hypothesis testing. *Journal of the American Statistical Association*, 90:370–378.
- Tu, W. and Zhou, X. (2000). Pairwise comparison of the means of skewed data. *Journal of Statistical Planning and Inference*, 88:59–74.
- Westfall, P. H. and Wolfinger, R. D. (1997). Multiple tests with discrete distributions. *The American Statistician*, 51:3–8.
- Westfall, P. H. and Young, S. S. (1993). *Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment*. John Wiley, New York.
- Westfall, P. H., Zaykin, D. V., and Young, S. S. (2001). Multiple tests for genetic effects in association studies. In Looney, S., editor, *Methods in Molecular Biology: Biostatistical Methods*, volume 184, pages 143–168. Humana Press, Toloway, NJ.

Table 1: Empirical FWEs and average number of false hypotheses rejected for both the Holm method and the general stepdown construction of Subsection 4.2. The nominal levels are  $\alpha = 5\%$  and  $\alpha = 10\%$ . Observations are i.i.d. multivariate normal, the dimension is  $k = 10$ , and the number of observations is  $n = 100$ . The number of repetitions is 10,000 per scenario and the number of bootstrap resamples is  $B = 1,000$ .

Level $\alpha$	FWE (Holm)	FWE (stepdown)	Rejected (Holm)	Rejected (stepdown)
All $\theta_j = 0$ , all $\rho = 0$				
5	5.8	5.1	0.0	0.0
10	10.5	9.8	0.0	0.0
All $\theta_j = 0$ , all $\rho = 0.5$				
5	3.9	5.2	0.0	0.0
10	7.2	10.3	0.0	0.0
All $\theta_j = 0$ , all $\rho = 0.9$				
5	1.9	5.2	0.0	0.0
10	3.3	10.1	0.0	0.0
Half of the $\theta_j = 0.25$ , all $\rho = 0$				
5	4.1	3.6	2.5	2.5
10	8.2	7.9	3.0	3.1
Half of the $\theta_j = 0.25$ , all $\rho = 0.5$				
5	3.8	5.1	2.5	2.7
10	7.4	10.0	3.1	3.3
Half of the $\theta_j = 0.25$ , all $\rho = 0.9$				
5	2.5	5.0	2.5	3.4
10	4.5	10.0	3.0	3.9
All $\theta_j = 0.25$ , all $\rho = 0$				
5	0.0	0.0	5.8	5.8
10	0.0	0.0	7.3	7.3
All $\theta_j = 0.25$ , all $\rho = 0.5$				
5	0.0	0.0	5.8	6.1
10	0.0	0.0	7.0	7.5
All $\theta_j = 0.25$ , all $\rho = 0.9$				
5	0.0	0.0	5.7	7.1
10	0.0	0.0	6.8	8.2

Table 2: Empirical FWEs and average number of false hypotheses rejected for both the Holm method and the general stepdown construction of Subsection 4.2. The nominal levels are  $\alpha = 5\%$  and  $\alpha = 10\%$ . Observations are i.i.d. multivariate normal, the dimension is  $k = 40$ , and the number of observations is  $n = 100$ . The number of repetitions is 10,000 per scenario and the number of bootstrap resamples is  $B = 1,000$ .

Level $\alpha$	FWE (Holm)	FWE (stepdown)	Rejected (Holm)	Rejected (stepdown)
All $\theta_j = 0$ , all $\rho = 0$				
5	6.1	4.8	0.0	0.0
10	11.6	10.0	0.0	0.0
All $\theta_j = 0$ , all $\rho = 0.5$				
5	3.8	4.9	0.0	0.0
10	6.4	10.0	0.0	0.0
All $\theta_j = 0$ , all $\rho = 0.9$				
5	1.0	5.0	0.0	0.0
10	1.6	9.9	0.0	0.0
Half of the $\theta_j = 0.25$ , all $\rho = 0$				
5	3.4	3.2	6.4	6.3
10	6.9	6.7	8.2	8.1
Half of the $\theta_j = 0.25$ , all $\rho = 0.5$				
5	3.5	4.6	6.6	7.6
10	6.3	9.4	8.3	10.0
Half of the $\theta_j = 0.25$ , all $\rho = 0.9$				
5	1.1	5.3	6.7	12.0
10	2.2	10.1	8.4	14.6
All $\theta_j = 0.25$ , all $\rho = 0$				
5	0.0	0.0	14.1	13.5
10	0.0	0.0	18.4	18.1
All $\theta_j = 0.25$ , all $\rho = 0.5$				
5	0.0	0.0	15.3	17.5
10	0.0	0.0	19.6	23.2
All $\theta_j = 0.25$ , all $\rho = 0.9$				
5	0.0	0.0	15.6	25.1
10	0.0	0.0	19.2	30.3

Table 3: Empirical FWEs and average number of false hypotheses rejected for both the single-step construction and the general stepdown construction of Subsection 4.2. The nominal levels are  $\alpha = 5\%$  and  $\alpha = 10\%$ . Observations are i.i.d. multivariate normal, the number of observations is  $n = 50$  and  $n = 100$ , and the number of pairwise correlations is  $k = 45$ . The number of repetitions is 10,000 per scenario and the number of bootstrap resamples is  $B = 1,000$ .

Level $\alpha$	FWE (single-step)	FWE (stepdown)	Rejected (single-step)	Rejected (stepdown)
All $\rho_{i,j} = 0, n = 50$				
5	3.5	3.5	0.0	0.0
10	8.2	8.2	0.0	0.0
All $\rho_{1,j} = 0.3$ and remaining $\rho_{i,j} = 0, n = 50$				
5	3.0	3.0	0.96	0.97
10	7.6	7.6	1.5	1.6
All $\rho_{i,j} = 0.3, n = 50$				
5	0.0	0.0	6.6	7.1
10	0.0	0.0	10.2	11.2
All $\rho_{i,j} = 0, n = 100$				
5	4.4	4.4	0.0	0.0
10	9.7	9.7	0.0	0.0
All $\rho_{1,j} = 0.3$ and remaining $\rho_{i,j} = 0, n = 100$				
5	3.9	4.0	3.7	3.8
10	8.7	9.0	4.5	4.6
All $\rho_{i,j} = 0.3, n = 100$				
5	0.0	0.0	21.9	25.4
10	0.0	0.0	26.5	30.9

Table 4: Sample correlations and  $p$ -values for the data of Example 6.4 of Westfall and Young (1993). ‘W-Y  $p$ -value’ denotes the adjusted  $p$ -value of Westfall and Young; ‘Step  $p$ -value’ denotes the adjusted bootstrap  $p$ -value obtained from the stepdown construction of Subsection 4.2 (based on  $B = 5,000$  bootstrap resamples).

Variables	Sample correlation	Raw $p$ -value	W-Y $p$ -value	Step $p$ -value
(SATdev, % Black)	-0.5089	.0002	.0019	.0016
(Salary, Crime)	0.4902	.0003	.0030	.0028
(% Black, Crime)	0.4844	.0004	.0036	.0034
(SATdev, S/T Ratio)	-0.3864	.0061	.0404	.0346
(SATdev, Crime)	-0.3033	.0341	.1843	.1483
(S/T Ratio, Crime)	0.2290	.1135	.4485	.3921
(S/T Ratio, % Black)	0.1732	.2341	.6474	.5986
(SATdev, Salary)	0.0980	.5030	.8753	.8572
(Salary, % Black)	-0.0354	.8090	.9641	.9645
(S/T Ratio, Salary)	0.0045	.9754	.9759	.9761