Balanced control of generalized error rates

Romano, Joseph P ; Wolf, Michael

DOI: https://doi.org/10.1214/09-AOS734

Institute for Empirical Research in Economics
University of Zurich

Working Paper Series
ISSN 1424-0459

Working Paper No. 379

Balanced Control of Generalized Error Rates

Joseph P. Romano and Michael Wolf

July 2008

Balanced Control of Generalized Error Rates

Joseph P. Romano
Depts. of Economics and Statistics
Stanford University
romano@stanford.edu

Michael Wolf
Inst. for Empirical Research in Economics
University of Zurich
mwolf@iew.uzh.ch

July 2008

Abstract

Consider the problem of testing s hypotheses simultaneously. In this paper, we derive methods which control the generalized familywise error rate given by the probability of k or more false rejections, abbreviated k-FWER. We derive both single-step and stepdown procedures that control the k-FWER in finite samples or asymptotically, depending on the situation. Moreover, the procedures are asymptotically balanced in an appropriate sense. We briefly consider control of the average number of false rejections. Additionally, we consider the false discovery proportion (FDP), defined as the number of false rejections divided by the total number of rejections (and defined to be 0 if there are no rejections). Here, the goal is to construct methods which satisfy, for given γ and α, $P\{\text{FDP} > \gamma\} \leq \alpha$, at least asymptotically. Special attention attention is paid to the construction of methods which implicitly take into account the dependence structure of the individual test statistics in order to further increase the ability to detect false null hypotheses. A general resampling and subsampling approach is presented which achieves these objectives, at least asymptotically.

KEY WORDS: Bootstrap, False Discovery Proportion, Generalized familywise error rate, Multiple Testing, Stepdown procedure.

JEL CLASSIFICATION NOS: C12, C14.

ACKNOWLEDGMENTS: Special thanks to Azeem Shaikh for helpful discussions. The research of the first author has been supported by the National Science Foundation, grant DMS 0707085. The research of the second author has been supported by by the Spanish Ministry of Science and Technology and FEDER, grant MTM2006-05650.

1 Introduction

The main goal 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. In particular, we construct computationally feasible methods which provide control (at least asymptotically) of some generalized notions of the familywise error rate. However, the theory also applies to exact finite sample control in certain situations. Moreover, explicit attention is paid to the construction of methods that are balanced, which roughly means that individual hypotheses are treated fairly in the allocation of overall error measure.

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 Ω, though we make no rigid requirements for Ω; it may be a parametric, semiparametric or a nonparametric model.

Consider the problem of simultaneously testing a hypothesis H_i against H_i', for $i = 1, \ldots, s$. Of course, a hypothesis H_i can be viewed as a subset, ω_i, of Ω, in which case the hypothesis H_i is equivalent to $P \in \omega_i$ and H_i' is equivalent to $P \notin \omega_i$. We also assume a test of the individual hypothesis H_i is based on a test statistic $T_{n,i}$, with large values indicating evidence against H_i.

The classical approach to dealing with the multiplicity problem is to restrict attention to procedures that control the probability of one or more false rejections, which is called the familywise error rate (FWER). But, safeguards against false rejections are not the only concern of multiple testing procedures. Corresponding to the power of a single test, one must also consider the ability of a procedure to detect departures from the null hypotheses. When the number of tests, s, is large, such as in genomics studies, control of the FWER at conventional levels becomes so stringent that individual departures from the null hypotheses have little chance of being detected. For this reason, we shall consider alternatives to the FWER that control false rejections less severely in hopes of better power.

First, we shall consider the k-FWER, the probability of rejecting at least k true null hypotheses. More formally, suppose data X is available from some model $P \in \Omega$. A general hypothesis H can be viewed as a subset ω of Ω. For testing $H_i : P \in \omega_i$, $i = 1, \ldots, s$, let $I(P)$ denote the set of true null hypotheses when P is the true probability distribution; that is, $i \in I(P)$ if and only if $P \in \omega_i$. Then, the k-FWER, which depends on P, is defined to be

$$k\text{-FWER}_P = P\{\text{reject at least } k \text{ hypotheses } H_i : i \in I(P)\} \ . \tag{1}$$

Control of the k-FWER requires that k-FWER $\leq \alpha$ for all P; that is,

$$k\text{-FWER}_P \leq \alpha \quad \text{for all } P \ . \tag{2}$$

Evidently, the case $k = 1$ reduces to control of the usual FWER.

We will also consider control of the *false discovery proportion* (FDP), defined as the total number of false rejections divided by the total number of rejections (and equal to 0 if there are no rejections). Given a user specified value $\gamma \in [0, 1)$, the measure of error control we wish to control is $P\{\text{FDP} > \gamma\}$; thus, we wish to construct methods satisfying

$$P\{\text{FDP} > \gamma\} \leq \alpha \quad \text{for all } P \ . \tag{3}$$

We will derive methods where this holds (at least asymptotically). Evidently, control of the FDP with $\gamma = 0$ reduces to the usual FWER. Control of the false discovery rate (FDR) requires that $E_P(\text{FDP}) \leq \gamma$.

Another measure of error control is the average number of false rejections. That is, for a given multiple testing procedure, let F denote the number of true null hypotheses rejected. Control of the average number of false rejections at level λ just means

$$E_P(F) \leq \lambda \quad \text{for all } P \ . \tag{4}$$

Note that λ need not be restricted to $(0, 1)$. Such a measure of error control was suggested in Spjøtvoll (1972).

Recently, there have been many new methods which control generalized error rates that are less stringent than the FWER. A notable such technique is the FDR controlling method of Benjamini and Hochberg (1995). Additional methods that control the FDR are given in Benjamini and Yekutieli (2001), Sarkar (2002), Storey et al. (2004), and Benjamini et al. (2006), among others. Asymptotic procedures that control the FDP (and the FDR) in the framework of a random effects mixture model are studied in Genovese and Wasserman (2004). These ideas are extended in Perone Pacifico et al. (2004), where in the context of random fields, the number of null hypotheses is uncountable. Methods that control both the k-FWER and FDP are given in Korn et al. (2004); they provide some justification for their methods, but they are limited to a multivariate permutation model. Stepwise methods based on p-values having finite sample validity are obtained in Hommel and Hoffman (1988), Lehmann and Romano (2005a), Romano and Shaikh (2006a), and Romano and Shaikh (2006b). Alternative methods of control of the k-FWER and FDP are given in van der Laan et al. (2004) and van der Laan et al. (2005). Building upon our work in Romano and Wolf (2005) and Romano and Wolf

(2007), we employ resampling and subsampling to achieve our goals and do not require the use of the subset pivotality condition of Westfall and Young (1993). The virtue of utilizing computer-intensive methods is that one can construct more powerful procedures by implicitly or explicitly taking into account the joint distribution of the test statistics. In addition, we construct procedures which are balanced, in a sense to be described later. Control of the false discovery rate via resampling is considered in Romano et al. (2007).

In general, we suppose that rejection of H_i is based on large values of a test statistic $T_{n,i}$ (with the subscript n used for asymptotic purposes). If a p-value $\hat{p}_{n,i}$ is available for testing H_i, one can take $T_{n,i} = -\hat{p}_{n,i}$. Typically, one would like to choose test statistics which lead to procedures that are balanced in the sense that all tests have about the same power and contribute equally to error control, as argued by Beran (1988a), Tu and Zhou (2000), and Rogers and Hsu (2001). Achieving balance can often be handled by appropriate choice of test statistics. For example, using p-values as the basic statistics will lead to better balance of error control. Quite generally, Beran's prepivoting transformation can lead to balance; see Beran (1988a) and Beran (1988b). Alternatively, balance can sometimes be achieved by studentization. However, if studentization or transforming a test statistic to a p-value is accomplished by resampling, we would not want to have to employ an iterated resampling scheme to obtain overall error control. (Indeed, we would not want to have to bootstrap the distribution of the minimum, or more generally the kth ordered p-value if the individual p-values were first obtained via resampling, because this would require an iterative computation at each stage of a stepwise algorithm.) Nevertheless, in order to avoid such heavy computational schemes, one of the main contributions here is that we can obtain balance and error control via resampling without resorting to an iterated bootstrap (and use the same set of resamples or subsamples at each stage).

In Section 2, we review Beran's (1988a) construction of balanced simultaneous confidence regions, which can be inverted to construct multiple tests of hypotheses which control the usual familywise error rate. We then generalize this construction to accomodate control of the k-FWER. These methods are single-step methods, in that individual test statistics are compared to their respective critical values simultaneously. (Note that the critical values used for testing H_i can depend on i in contrast to typical single-step methods, such as a Bonferroni method or a method based on the maximum test statistic. However, we still call them single-step in contrast to the stepdown methods that we will also consider; stepdown methods start with a single-step method but allow possible further rejections by changing the critical values depending on the hypotheses already rejected.) In Section 3, we show that, if we apply critical values that have a monotonicity property, then the basic problem of constructing a valid

stepdown multiple test procedure that controls the k-FWER can be reduced to the easier problem of constructing single-step methods which control the k-FWER. In particular, if finite sample methods which offer control of the Type 1 error are available for each of the individual tests, then this will immediately translate into control of the k-FWER. Otherwise, we can apply bootstrap and subsampling methods to achieve asymptotic control, as described in Section 4. In summary, stepdown improvements of the single-step method are presented. We also present a generalized Bonferroni type of method which has finite sample control of the k-FWER in Section 5. Section 6 briefly discusses control of the average number of false rejections. Results for control of the FDP are obtained in Section 7. A simulation study is presented in Section 8. All proofs are collected in an appendix.

Some further notation which is used throughout the paper is required. Suppose $\{y_i : i \in K\}$ is a collection of real numbers indexed by a finite set K having $|K|$ elements. Then, for $k \leq |K|$, the k-max$(y_i : i \in K)$ is used to denote the kth largest value of the y_i with $i \in K$. So, if the elements y_i, $i \in K$, are ordered as $y_{(1)} \leq \cdots \leq y_{(|K|)}$, then k-max$(y_i : i \in K) = y_{(|K|-k+1)}$.

2 Balanced Simultaneous Confidence Regions

Throughout this section, k is fixed. We first review and then generalize Beran's (1988a) construction of simultaneous confidence regions as a building block. For now, assume hypothesis H_i is concerned with a test of a real-valued parameter $\theta_i(P)$. Specifically, H_i specifies $P \in \omega_i$, where

$$\omega_i = \{P : \ \theta_i(P) = 0\} \ .$$

Let $\hat{\theta}_{n,i}$ be some estimate of $\theta_i(P)$. Tests of a particular H_i (without regard to multiplicity) can be constructed by the usual duality between tests and confidence intervals, if one knows or can estimate the sampling distribution of $\hat{\theta}_{n,i} - \theta_i(P)$ under P. Let $J_{n,i}(P)$ denote the sampling distribution of $\tau_n[\hat{\theta}_{n,i} - \theta_i(P)]$ under P, with $J_{n,i}(\cdot, P)$ denoting the corresponding left-continuous cumulative distribution. (As in Beran, the purpose of working with a left-continuous c.d.f. F is that it satisfies $\{x : x \leq F^{-1}(\gamma)\}$ is equivalent to $\{x : F(x) \leq \gamma\}$, if $F^{-1}(\gamma)$ is taken to be the largest γ quantile of F. Such definitions are not crucial for asymptotic results, and we can, for example, just as well work with smallest γ quantiles or anything in between the two.) The nonrandom sequence τ_n is introduced for asymptotic purposes so that a nondegenerate limiting distribution for $J_{n,i}(\cdot, P)$ exists. (Note that it is possible to let τ_n vary with the hypothesis i. Extensions to cases where τ_n depends on P are also possible, using ideas in Politis et al., 1999, Chapter 8.)

Also, let $H_{n,i}(\cdot, P)$ denote the c.d.f. of $\tau_n|\hat{\theta}_{n,i} - \theta_i(P)|$ under P. Let $c_{n,i}(\gamma, P)$ denote the largest γ quantile of $H_{n,i}(\cdot, P)$. Then, assuming continuity of $H_{n,i}(\cdot, P)$, the confidence interval

$$\{\theta_i : \tau_n|\hat{\theta}_{n,i} - \theta_i| \leq c_{n,i}(\gamma, P)\} \tag{5}$$

has coverage probability γ. (Note that continuity of $H_{n,i}(P)$ is only assumed here for convenience and is certainly not required in our asymptotic results.) Of course, this interval is generally unavailable because $c_{n,i}(\gamma, P)$ is unknown, as it depends on P. However, even if these critical values were available, we would like to make a statement about the simultaneous coverage of the intervals.

To this end, let $K \subseteq \{1, \ldots, s\}$ denote an arbitrary subset of $\{1, \ldots, s\}$. We would like to make joint inferences for the parameters $\theta_i(P)$ simultaneously for $i \in K$. (Of course, the case where $K = \{1, \ldots, s\}$ is especially important, but the general case is required for our stepdown multiple testing method presented later.) Then, the probability of the event

$$\{\tau_n|\hat{\theta}_{n,i} - \theta_i(P)| \leq c_{n,i}(\gamma, P) \text{ for all } i \in K\}$$

is some function of γ and P, say $f_{n,K}(\gamma, P)$. Again, for the moment, ignoring the fact that P is unknown, the idea for constructing a simultaneous confidence region for the set of parameters $\{\theta_i(P) : i \in K\}$ is to vary γ so that this last expression is equal to $1 - \alpha$. Thus, we choose γ so that $f_{n,K}(\gamma, P) = 1 - \alpha$, or more formally the infimum over all γ such that $f_{n,K}(\gamma, P) \geq 1 - \alpha$. Suppose $\gamma_{n,K}(\alpha, P)$ is such that

$$f_{n,K}(\gamma_{n,K}(\alpha, P), P) = 1 - \alpha \ .$$

Then, in addition to the simultaneous coverage statement, each marginal interval for a particular $\theta_i(P)$ has coverage probability $\gamma_{n,K}(\alpha, P)$, which is independent of i. That each interval covers its corresponding parameter with the same probability is the property of *balance*.

Beran's (1988a) asymptotic solution to the construction of balanced simultaneous confidence regions is to utilize the bootstrap. That is, let $\hat{Q}_n$ be some estimate of P. For i.i.d. data, in the absence of a parametric model for P, $\hat{Q}_n$ is typically taken to be the empirical distribution of the observed data, or possibly a smoothed version (i.e., nonparametric bootstrap); on the other hand, if a parametric model for P is assumed, then $\hat{Q}_n$ should be based on this model (i.e., parametric bootstrap); see Davison and Hinkley (1997). For time series or data-dependent situations, bootstrap methods that can capture the underlying dependence structure should be employed, such as block bootstraps, sieve bootstraps, or Markov bootstraps; see Lahiri (2003). The procedure is to replace P by $\hat{Q}_n$ in (5). Specifically, Beran

proposes the set of intervals

$$\{\theta_i : \tau_n|\hat{\theta}_{n,i} - \theta_i| \leq c_{n,i}(\gamma, \hat{Q}_n)\} = \{\theta_i : \tau_n|\hat{\theta}_{n,i} - \theta_i| \leq H_{n,i}^{-1}(\gamma, \hat{Q}_n)\} \ , \tag{6}$$

where γ is chosen to be $\gamma_{n,K}(\alpha, \hat{Q}_n)$. Under appropriate regularity conditions, these intervals simultaneously contain the true parameters $\{\theta_i(P) : \ i \in K\}$ with limiting probability $1 - \alpha$ and are asymptotically balanced.

Of course, simultaneous confidence regions for $\{\theta_i(P) : i \in K\}$ of nominal level $1 - \alpha$ can be used to construct tests of the hypotheses $H_i, i \in K$, by rejecting any H_i for which 0 is not included in the confidence interval for $\theta_i(P)$. Such a procedure would control the familywise error rate at nominal level α. However, our current goal is to control the k-FWER. Therefore, we now generalize Beran's construction. It is now required to approximate the probability of the event

$$\{\tau_n|\hat{\theta}_{n,i} - \theta_i(P)| \leq c_{n,i}(\gamma, P) \ \text{ for all but at most } (k-1) \text{ of the } i \in K\} \ . \tag{7}$$

To this end, the previous event (7) can be rewritten as

$$\{H_{n,i}(\tau_n|\hat{\theta}_{n,i} - \theta_i(P)|, P) \leq \gamma \ \text{ for all but at most } (k-1) \text{ of the } i \in K\} \ , \tag{8}$$

or

$$\{k\text{-max}(H_{n,i}(\tau_n|\hat{\theta}_{n,i} - \theta_i(P)|, P), i \in K) \leq \gamma\} \ . \tag{9}$$

Let $f_{n,K}(\gamma, k, P)$ denote the probability under P of the event in (7)–(9), and let $\gamma_{n,K}(\alpha, k, P)$ denote the value of γ such that $f_{n,K}(\gamma, k, P) = 1 - \alpha$, or more precisely the infimum over all γ such that

$$f_{n,K}(\gamma, k, P) \geq 1 - \alpha \ .$$

Then, the solution γ of the previous equation can be represented as the $1 - \alpha$ quantile of the distribution of

$$k\text{-max}(H_{n,i}(\tau_n|\hat{\theta}_{n,i} - \theta_i(P)|, P), i \in K)$$

under P, which we denote by $L_{n,K}(k, P)$.

A bootstrap choice for the level γ can be represented as

$$\gamma_{n,K}(\alpha, k, \hat{Q}_n) = L_{n,K}^{-1}(1 - \alpha, k, \hat{Q}_n) \ . \tag{10}$$

Combining (6) and (10) yields the joint confidence region

$$\{(\theta_i, i \in K) : \ \tau_n|\hat{\theta}_{n,i} - \theta_i| \leq H_{n,i}^{-1}(L_{n,K}^{-1}(1 - \alpha, k, \hat{Q}_n), \hat{Q}_n)\} \ . \tag{11}$$

Under fairly general conditions, this simultaneous confidence region covers all θ_i with $i \in K$, except for at most $k-1$ of them, with limiting probability $1-\alpha$. Moreover, the intervals are asymptotically balanced in the sense that the probability that $\theta_i(P)$ is covered does not asymptotically depend on i.

Remark 2.1 (Calculating (11)) In order to calculate (11), we usually resort to an approximation by simulation. However, only one set of resamples is needed, and nested simulations are not required in order to derive asymptotic results. To describe the algorithm in a little detail, for $b = 1, \ldots, B$, draw a sample of size n from $\hat{Q}_n$ and let $\hat{\theta}^*_{n,i}(b)$ be the estimate of θ_i. Then, $H_{n,i}(x, \hat{Q}_n)$ can be approximated by the proportion of times the values $\tau_n|\hat{\theta}^*_{n,i}(b) - \hat{\theta}_{n,i}|$ are $\leq x$; this leads to a corresponding approximation to the quantile function $H^{-1}_{n,i}(\cdot, \hat{Q}_n)$. Next, $L_{n,K}(x, k, \hat{Q}_n)$ is estimated by the proportion of times the values k-max$(H_{n,i}(\tau_n|\hat{\theta}^*_{n,i}(b) - \hat{\theta}_{n,i}|, \hat{Q}_n), i \in K)$ are $\leq x$; its largest $1-\alpha$ quantile is a simulation-based approximation of $L^{-1}_{n,K}(1-\alpha, k, \hat{Q}_n)$.

As Beran argued in the case $k = 1$, this construction can reproduce some classical solutions in certain parametric models. Moreover, the construction implicitly studentizes the individual estimators, so that each marginal interval covers with the same probability. However, outside certain parametric models or permutation models, the solution is only approximate. In order to describe the asymptotic behavior of the above quantities, we introduce some notation and assumptions. The symbols $\xrightarrow{L}$ and $\xrightarrow{P}$ will denote convergence in law (or distribution) and convergence in probability, respectively. For $K \subseteq \{1, \ldots, s\}$, let $J_{n,K}(P)$ denote the joint distribution of $\{\tau_n[\hat{\theta}_{n,i} - \theta_i(P)],\ i \in K\}$. So, $J_{n,\{i\}}(P) = J_{n,i}(P)$ for a singleton subset $\{i\} \subseteq K$. Typically, the joint distribution of the estimators tends to an asymptotic limit, which is stated formally in the following assumption.

Assumption B1 $J_{n,\{1,\ldots,s\}}(P) \xrightarrow{L} J_{\{1,\ldots,s\}}(P)$.

For a reasonable asymptotic theory, the asymptotic distribution should be nondegenerate, and so we will also use the following assumption.

Assumption B2 $J_i(P)$ has a continuous distribution function for all i.

Assumptions B1 and B2 imply that, for every $K \subseteq \{1, \ldots, s\}$, $L_{n,K}(k, P)$ has a continuous limiting distribution $L_K(k, P)$.

Lemma 2.1 *Suppose Assumptions B1 and B2 hold. Then, for every $K \subseteq \{1, \ldots, s\}$, $L_{n,K}(k, P)$ has a continuous limiting distribution $L_K(k, P)$, which can be represented as the distribution of*

$$k\text{-}max(H_i(|Y_i|, P), i \in K) \ , \tag{12}$$

where $(Y_1, \ldots, Y_s)$ has distribution $J_{\{1,\ldots,s\}}(P)$ and

$$H_i(x, P) = J_i(x, P) - J_i(-x, P) \ . \tag{13}$$

Under an additional mild assumption, we can show that this limiting distribution is strictly increasing on its support, which will prove quite useful. This additional assumption is the following.

Assumption B3 The support of the limiting distribution $J_{\{1,\ldots,s\}}(P)$ is connected.

Assumption B3 is indeed very weak. It holds whenever the joint limiting distribution is multivariate Gaussian, as long as the diagonal entries of the covariance matrix are nonzero. In particular, this covariance matrix may even be singular (which happens in some simultaneous inference problems; e.g., pairwise comparisons of means). The utility of Assumption B3 derives from the following lemma and its corollary.

Lemma 2.2 *Let $X = (X_1, \ldots, X_s)$ be a random vector on $\mathbb{R}^s$ with multivariate distribution F. Suppose that the support of the distribution F, denoted supp(F), is connected. Let h_i be continuous with $h_i(X_i)$ having a continuous distribution. Then, $Y \equiv k\text{-}max(h_i(X_1), \ldots, h_s(X_s))$ has a continuous and strictly increasing c.d.f. on its interval of support.*

Remark 2.2 Note that even in the case in which $s = k = 1$, both assumptions in Lemma 2.2 are necessary to conclude that the distribution of Y is continuous and strictly increasing. Therefore, the assumptions used in Lemma 2.2 seem as weak as possible. Note that the assumption that $h_i(X_i)$ has a continuous distribution follows if X_i has a continuous distribution and h_i is the identity function ($h_i(x) = x$), the absolute value function ($h_i(x) = (|x|)$, the distribution function of X_i ($h_i(x) = F_i(x)$ where $X_i \sim F_i$) , or the distribution function of $|X_i|$ evaluated at $|X_i|$ ($h_i(x) = H_i(|x|)$, where $H_i(\cdot)$ is the distribution of $|X_i|$. The last example is most pertinent to this paper. Also, note that the lemma holds if the k-max function is replaced by any continuous function which returns one of its arguments.

Corollary 2.1 *Assume B1–B3. Then, $L_K(k, P)$ has a continuous and strictly increasing c.d.f. on its interval of support.*

Finally, in order to show asymptotic validity of the bootstrap, we need a further assumption on the behavior of the estimator $\hat{Q}_n$ of P. For this, we further assume the usual conditions for bootstrap consistency when testing the *single* hypothesis that $\theta_i(P) = 0$ for all $i \in I(P)$; that is, we assume the bootstrap consistently estimates the joint distribution of $\tau_n[\hat{\theta}_{n,i} - \theta_i(P)]$ for $i \in \{1, \ldots, s\}$. Specifically, consider the following assumption.

Assumption B4 For any metric ρ metrizing weak convergence on $\mathbb{R}^s$,

$$\rho\left(J_{n,\{1,\ldots,s\}}(P), J_{n,\{1,\ldots,s\}}(\hat{Q}_n)\right) \xrightarrow{P} 0 \ .$$

Assumptions B4 is quite standard in the bootstrap literature, and readily holds for general classes of statistics, such as estimators which are smooth functions of means, U-statistics, L-statistics, estimators which are differentiable functions of the empirical process, etc.; see Hall (1992), Shao and Tu (1995) and Chapter 1 of Politis et al. (1999). Thus, our results apply to a wide range of problems. Under these assumptions, the following theorem proves asymptotic control of the k-FWER of our bootstrap method based on the simultaneous intervals (11). The result here requires fewer assumptions than Beran (1988a). In particular, we can dispense with his Assumption 4 in view of our above Lemma 2.2. Moreover, our result will apply toward control of the k-FWER for general k (while Beran's results only apply to $k = 1$).

Theorem 2.1 *Suppose data is generated from P satisfying Assumptions B1–B3. Let $\hat{Q}_n$ be an estimator of P satisfying Assumption B4. Fix $K \subseteq \{1, \ldots, s\}$ and a positive integer k. Consider the joint confidence region given by (11), with the marginal interval $\hat{C}_{n,i}$ for $\theta_i(P)$ with $i \in K$ expressed as*

$$\hat{C}_{n,i} \equiv \hat{\theta}_{n,i} \pm \tau_n^{-1} H_{n,i}^{-1}(L_{n,K}^{-1}(1-\alpha, k, \hat{Q}_n), \hat{Q}_n) \ . \tag{14}$$

(i) For $i \in K$, the intervals $\hat{C}_{n,i}$ simultaneously cover all the corresponding true parameter values $\theta_i(P)$, except for at most $k-1$ of them, with asymptotic probability $1-\alpha$.

(ii) The intervals $\hat{C}_{n,i}$ are balanced in the sense that

$$\lim_{n\to\infty} P\{\theta_i(P) \in \hat{C}_{n,i}\} = \gamma, \quad \text{independent of } i \ , \tag{15}$$

where $\gamma = \gamma_K(1-\alpha, k, P)$ is the unique $1-\alpha$ quantile of the limiting distribution $L_K(k, P)$.

Remark 2.3 (Other resampling schemes) As previously mentioned, the choice of $\hat{Q}_n$ should reflect the underlying P. We will later also consider a subsampling approach in Section 4.2. In

some cases where permutation methodology is applicable, one can obtain exact finite sample results as well. (Computationally, one can achieve this feasibly without an iterative scheme because the set of permutations of a permutation is exactly the set of all permutations; in contrast, the set of bootstrap samples from a bootstrap sample itself is not the same as the set of all bootstrap samples from the original data.) To see how this is done in the case $k = 1$; see Romano and Wolf (2005). The finite sample results also extend to stepdown methods considered later, using ideas developed in Section 3.

Remark 2.4 (Planned Imbalance) The argument can be generalized to weighting schemes if some parameters are more important than others. That is, if it is desired to have the individual parameters $\theta_i(P)$ covered with probability proportional to some fixed weights w_i, then the argument can be adapted to accomplish this.

Remark 2.5 (General Roots) If standard errors $\hat{\sigma}_{n,i}$ of the scaled estimators $\tau_n\hat{\theta}_{n,i}$ are available, it usually makes sense (especially from a higher-order asymptotic viewpoint) to base inference on the (estimated) distributions of the studentized roots $\tau_n|\hat{\theta}_{n,i} - \theta_i(P)|/\hat{\sigma}_{n,i}$. In general, as in Beran, we allow for general roots as follows. Based on data x_n from P, let $R_{n,i}(x_n, \theta_i(P))$ be a real-valued function of the sample and $\theta_i(P)$, with c.d.f. $H_{n,i}(\cdot, P)$. (We use the same notation as we did for the special case when $R_{n,i}(x_n, \theta_i(P)) = \tau_n|\hat{\theta}_{n,i} - \theta_i(P)|$.) Then, let $L_{n,K}(\cdot, k, P)$ denote the distribution of

$$k\text{-max}\left(H_{n,i}(R_{n,i}(x_n, \theta_i(P))), i \in K\right) \ .$$

The bootstrap replaces P by $\hat{Q}_n$, leading to the joint confidence region

$$\{(\theta_i, i \in K) : \ R_{n,i}(x_n, \theta_i) \leq H_{n,i}^{-1}(L_{n,K}^{-1}(1-\alpha, k, \hat{Q}_n), \hat{Q}_n)\} \ ,$$

in generalization of (11). For example, if we consider the 'one-sided' roots $R_{n,i} = \tau_n(\hat{\theta}_{n,i} - \theta_i(P))$, then the construction leads to simultaneous one-sided confidence intervals. Alternatively, if standard errors are available, we could also consider the 'one-sided' studentized roots $R_{n,i} = \tau_n[\hat{\theta}_{n,i} - \theta_i(P)]/\hat{\sigma}_{n,i}$ to obtain simultaneous one-sided confidence intervals.

Remark 2.6 (Balance in the tails) So far, balance is achieved with respect to the marginal coverage probability of each interval. The construction can easily be modified if it is also desired to have balance in the tails of each marginal interval as well. A simple way to do this is by considering the "one-sided" roots explained in the previous remark at level $1 - \alpha/2$, and then the negative of these roots at level $1 - \alpha/2$; combine them to obtain balance in the tails as well as balance of marginal coverage.

A value of 0 for $\theta_i(P)$ falls outside the region (14) if and only if

$$|\tau_n \hat{\theta}_{n,i}| > H_{n,i}^{-1}(L_{n,K}^{-1}(1-\alpha, k, \hat{Q}_n), \hat{Q}_n) \ . \tag{16}$$

By design, there exists a duality between confidence sets contructed and control of the k-FWER, so the following holds.

Corollary 2.2 *Assume the conditions of Theorem 2.1. For testing the multiple hypotheses $H_i : \ \theta_i(P) = 0$, consider the procedure which rejects H_i if (16) holds with $K = \{1, \ldots, s\}$. Then,*

(i)

$$\lim_{n\to\infty} k\text{-}FWER_P \leq \alpha \ .$$

(ii) Moreover,

$$\lim_{n\to\infty} P\{\text{reject } H_i\} = 1 - L_K^{-1}(1-\alpha, k, P) \tag{17}$$

exists and is independent of i for $i \in I(P)$, i.e., the error allocation is asymptotically balanced.

Note that, for testing H_i alone, a marginal (unadjusted) p-value can be obtained by

$$\hat{p}_{n,i} \equiv 1 - H_{n,i}(|\tau_n \hat{\theta}_{n,i}|, \hat{Q}_n) \ . \tag{18}$$

If balance were not imposed as in Romano and Wolf (2007), then the larger $|\hat{\theta}_{n,i}|$, the more significant H_i; that is, tests are essentially ordered by the values of $|\hat{\theta}_{n,i}|$. By imposing balance, tests are now ordered by the ordering of p-values.

Remark 2.7 (Relationship to studentization) As argued by Beran (1988a), the construction implicitly accounts for the variation in i of the estimates $\hat{\theta}_{n,i}$ and is asymptotically equivalent to studentization. Note that in the expression for the marginal p-value $\hat{p}_{n,i}$ given in (18), the transformation $H_{n,i}(\cdot, \hat{Q}_n)$ is essentially Beran's prepivoting transformation, and has the effect of putting all the test statistics on a common scale. Indeed by (14), the multiple testing procedure rejects an H_i if

$$H_{n,i}(|\tau_n \hat{\theta}_{n,i}|, \hat{Q}_n) > L_{n,K}^{-1}(1-\alpha, k, \hat{Q}_n) \ ,$$

where the right side now does not depend on i. In general, if one can studentize an estimator or convert it to a p-value, balance will (asymptotically) be achieved. However, if resampling

is required to do so, then a nested level of resampling may be required to assess overall error control. The approach here and in Beran (1988a) allows one to accomplish both without having to compute iterative bootstraps. Certainly, one can apply the above methodology to studentized roots in hopes of better balance in finite samples.

3 Stepdown Methods that Control the k-FWER

We now return to the general setup. Test statistics $T_{n,i}$ are available to test H_i. Given a single-step method, such as the resampling method discussed in Section 2, we will show how a stepdown improvement may be obtained. Suppose we have in mind critical values $\hat{c}_{n,K,i}(1-\alpha, k)$ which could be used to control the k-FWER at level α when testing the multiple hypotheses H_i with $i \in K$; that is, such a single step procedure would reject H_i if $T_{n,i} > c_{n,K,i}(1-\alpha, k)$.

A stepdown method begins by first applying a single-step method, but then additional hypotheses may be rejected after this first stage by proceeding in a stepwise fashion, which we now describe. Begin by testing all null hypotheses $H_1, \ldots, H_s$. Any hypothesis H_i is rejected if $T_{n,i} > c_{n,\{1,\ldots,s\},i}(1-\alpha, k)$. If there are no rejections, then stop. If there are rejections, let A_2 be the set of hypotheses not yet rejected. Then, we compare $T_{n,i}$ for $i \in A_2$ with smaller critical values than used in the first stage, leading to the possibility of further rejections.

In the algorithm below, the critical constants $\hat{c}_{n,K,i}(1-\alpha, k)$ may be fixed or random, but the reader should have in mind that they should be designed to control the k-FWER when testing H_i with $i \in K$. Note that, in comparison, the stepdown methods developed in Romano and Wolf (2007) use a common critical value at each stage of the algorthm (which does not depend on i). Of course, it is vital to allow these critical values to depend on i if balance is desirable (and the test statistics are not studentized or already balanced). A particular choice we will study later and suggested by Corollary 2.2 is to let $c_{n,K,i}(1-\alpha, k)$ to be the right hand side of (16), but other choices are possible as well.

Algorithm 3.1 (Generic Stepdown Method for Control of the k-FWER)

1. Let $A_1 = \{1, \ldots, s\}$. If $T_{n,i} \leq \hat{c}_{n,A_1,i}(1-\alpha, k)$ for all i, then accept all hypotheses and stop; otherwise, reject any H_i for which $T_{n,i} > \hat{c}_{n,A_1,i}(1-\alpha, k)$ and continue.

2. Let R_2 be the indices i of hypotheses H_i previously rejected, and let A_2 be the indices of the the remaining hypotheses. If $|R_2| < k$, then stop. Otherwise, reject any H_i with

$i \in A_2$ if $T_{n,i} > \hat{d}_{n,A_2,i}(1-\alpha, k)$, where

$$\hat{d}_{n,A_2,i}(1-\alpha, k) = \max_{I \subseteq R_2,\ |I|=k-1} \{\hat{c}_{n,K,i}(1-\alpha, k) : K = A_2 \cup I\} \ .$$

If there are no further rejections, stop.

$\vdots$

j. Let R_j be the indices i of hypotheses H_i previously rejected, and let A_j be the indices of the remaining hypotheses. Let

$$\hat{d}_{n,A_j,i}(1-\alpha, k) = \max_{I \subseteq R_j,\ |I|=k-1} \{\hat{c}_{n,K,i}(1-\alpha, k) : K = A_j \cup I\} \ .$$

Then, reject any H_i with $i \in A_j$ satisfying $T_{n,i} > \hat{d}_{n,A_j,i}(1-\alpha, k)$. If there are no further rejections, stop.

$\vdots$

And so on.

Note that, in the case $k = 1$, once a hypothesis is removed, it no longer enters into the algorithm. However, for $k > 1$, the algorithm becomes more complex. The reason is that, for control of the k-FWER, we must acknowledge that when we consider a set of hypotheses not previously rejected, we may have gotten to that stage by rejecting true null hypotheses, but hopefully at most $k-1$ of them. Since we do not know which of the hypotheses rejected thus far are true or false, we must maximize over subsets including some of those rejected, but at most $k-1$ among the previously rejected ones. Our main point will be that, if we can control the k-FWER at any stage of the algorithm, then the stepdown method will control the k-FWER.

Remark 3.1 (Modified Generic Stepdown Method for Control of the k-FWER) One can modify the above algorithm or any method that controls the k-FWER as follows. If the method rejects at least $k-1$ hypotheses, no modification is applied; otherwise, reject the $k-1$ most significant hypotheses (where most significant is determined by marginal or unadjusted p-values). This would not change control of the k-FWER. However, we do not generally promote this modification, because hypotheses can be rejected without compelling evidence (that is, even if they have large unadjusted p-values).

In order to prove such an algorithm controls the k-FWER for a suitable choice of critical values $\hat{c}_{n,K,i}(1-\alpha,k)$, we assume monotonicity of the estimated critical values; that is, for any $K \supseteq I$,

$$\hat{c}_{n,K,i}(1-\alpha,k) \geq \hat{c}_{n,I,i}(1-\alpha,k) \ . \tag{19}$$

Under the monotonicity assumption (19), we will show that k-FWER control of a stepdown procedure is reduced to that of a single-step method. Thus, the construction of a stepdown procedure is effectively reduced to construction of single tests, as long as the monotonicity assumption holds (and it *always* does for specific choices studied later).

Theorem 3.1 *Consider Algorithm 3.1 with critical values $\hat{c}_{n,K,i}(1-\alpha,k)$ satisfying (19).*

(i) Then, k-FWER$_P \leq$

$$P\{T_{n,i} > \hat{c}_{n,I(P),i} \text{ for all but at most } k-1 \text{ of } i \in I(P)\} \ . \tag{20}$$

(ii) Therefore, if the critical values $\hat{c}_{n,I(P),i}$ control the k-FWER as a single-step procedure in the sense that the right side of (20) is $\leq \alpha$ (in finite samples or asymptotically), then k-FWER$_P \leq \alpha$ (in finite samples or asymptotically).

The monotonicity assumption (19) cannot be removed, as shown in Example 2.1 of Romano and Wolf (2005) in the case $k = 1$; an analogous construction works for general k. The general resampling constructions we describe later will inherently satisfy (19). When testing multiple hypotheses, it seems natural that the critical values should satisfy the monotonicity condition, because larger critical values should be required when testing more hypotheses rather than a smaller subset of them.

Our main goal will be to employ resampling methods to calculate critical values, which can account for the dependence structure of the test statistics. This was accomplished in the case $k = 1$ by Romano and Wolf (2005) and for general k in Romano and Wolf (2007), but without the requirement of balance. However, we see how the argument generalizes given Theorem 3.1. We also observe that Theorem 3.1 applies to certain semiparametric problems where permutation and randomization tests apply. Such a setting is discussed in Korn et al. (2004), though the requirement of balanced was not addressed.

Outside some parametric models, application of the Generic Stepdown Method can be computationally intensive, so we will also consider the following more streamlined algorithm. The basic idea is that at any stage, when testing whether or not to include further rejections,

we need only look at the hypotheses not previously rejected together with the $k-1$ hypotheses that are least significant among those previously rejected. So, we avoid maximizing over all subsets of size $k-1$ of previously rejected hypotheses and just look at the least significant $k-1$ rejections. The arguments for such a procedure will be asymptotic.

Algorithm 3.2 (Streamlined Stepdown Method for Control of the k-FWER) We assume the existence of generic marginal p-values $\hat{p}_{n,i}$ for testing the individual hypotheses H_i. How they are computed depends on the context in general; for example, in the bootstrap approach detailed in Subsection 4.1, one can use $\hat{p}_{n,i} = 1 - H_{n,i}(\tau_n|\hat{\theta}_{n,i}|, \hat{Q}_n)$. The ordering of these p-values determines an ordering of the hypotheses in terms of their significance. To this end, order the p-values in ascending order, $\hat{p}_{n,(1)} \leq \ldots \leq \hat{p}_{n,(s)}$. Denote by $\{r_1, \ldots, r_s\}$ the permutation of $\{1, \ldots, s\}$ which yields this ordering; that is, $\hat{p}_{n,(1)} = \hat{p}_{n,r_1}, \ldots, \hat{p}_{n,(s)} = \hat{p}_{n,r_s}$. Accordingly, let $H_{(1)} = H_{r_1}, \ldots, H_{(s)} = H_{r_s}$. Then, $H_{(1)}$ is the most significant and $H_{(s)}$ is the least significant hypothesis. The algorithm now is analogous to Algorithm 3.1. The only difference is that in any step $j > 1$, the critical value

$$\hat{d}_{n,A_j,i}(1-\alpha, k) = \max_{I \subseteq R_j,\ |I|=k-1} \{\hat{c}_{n,K,i}(1-\alpha, k) : K = A_j \cup I\}$$

is replaced by the critical value

$$\tilde{d}_{n,A_j,i}(1-\alpha, k) = \hat{c}_{n,K,i}(1-\alpha, k) \quad \text{where} \quad K = \{r_{|R_j|-k+2}, r_{|R_j|-k+1}, \ldots, r_s\} \ .$$

4 Asymptotic Results on k-FWER Control

The main goal of this section is to show how Theorem 3.1 can be used to construct stepdown procedures that asymptotically control the k-FWER under very weak assumptions. The use of resampling techniques will be a key ingredient. The methods constructed will be based on Algorithm 3.1, and so potentially many tests are constructed in a stepwise fashion. However, a key feature is that the methods will only require *one* set of resamples for all of the tests, whether they are bootstrap samples or subsamples.

In order to accomplish this, we will consider resampling schemes that do *not* obey the null hypothesis constraints. Such schemes have been suggested previously by Pollard and van der Laan (2003) and Dudoit et al. (2004), and have the benefit of avoiding the subset pivotality condition of Westfall and Young (1993). Hypothesis test constructions that do obey the constraints imposed by 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. In some problems, the *subset pivotality* condition of Westfall and Young (1993) holds, and so the same null distribution can be used at each step. However, this condition does not hold in general; for instance, see Example 4.1 of Romano and Wolf (2005). To obtain a more general construction, we exploit the well-known explicit duality between tests and confidence intervals; so, if one can construct good or valid confidence intervals, then one can construct good or valid tests, and conversely. The same holds for simultaneous confidence sets and multiple tests.

We shall consider two concrete applications of Theorem 3.1, the first based on the bootstrap and the second based on subsampling.

4.1 A Bootstrap Construction

We now apply Theorem 3.1 to develop an asymptotically valid approach based on the bootstrap. As in Section 2, we specialize to the case where hypothesis H_i is specified by $\{P:\ \theta_i(P) = 0\}$ for some real-valued parameter θ_i. Implicitly, the alternatives are two-sided, but the one-sided case can be similarly handled. Recalling the notation of Section 2, suppose $\hat{\theta}_{n,i}$ is an estimate of θ_i. Also, $T_{n,i} = \tau_n|\hat{\theta}_{n,i}|$ for some nonnegative (nonrandom) sequence $\tau_n \to \infty$.

The duality between simultaneous confidence sets and multiple hypothesis tests already exploited in Corollary 2.2 suggests using Algorithm 3.1 with critical values

$$\hat{c}_{n,K,i}(1-\alpha, k) = H_{n,i}^{-1}(L_{n,K}^{-1}(1-\alpha, k, \hat{Q}_n), \hat{Q}_n)\ . \tag{21}$$

Note that, regardless of asymptotic behavior, the monotonicity assumption (19) is always satisfied for the choice (21). Indeed, whenever $I \subseteq K$, we must show

$$H_{n,i}^{-1}(L_{n,I}^{-1}(1-\alpha, k, \hat{Q}_n), \hat{Q}_n) \leq H_{n,i}^{-1}(L_{n,K}^{-1}(1-\alpha, k, \hat{Q}_n), \hat{Q}_n)\ ,$$

or equivalently (applying $H_{n,i}(\cdot, \hat{Q}_n)$ to both sides),

$$L_{n,I}^{-1}(1-\alpha, k, \hat{Q}_n) \leq L_{n,K}^{-1}(1-\alpha, k, \hat{Q}_n)\ . \tag{22}$$

But, for any Q and $I \subseteq K$, the left side of (22) is the $1-\alpha$ quantile under Q of the k-max of $|I|$ variables, while the right side of (22) is the $1-\alpha$ quantile of the k-max of these same $|I|$ variables together with additional $|K|-|I|$ variables. This simple observation together with Theorem 3.1

immediately reduces the problem of stepdown control to that of single-step control, which was already obtained in Corollary 2.2. The following result is an improvement over Corollary 2.2 in that more rejections are possible, while maintaining control of the k-FWER.

Corollary 4.1 *Under the setup and conditions of Corollary 2.2, consider Algorithm 3.1 with critical values given by (21).*

(i) Then, $\limsup_{n\to\infty}$ k*-FWER*$_P \leq \alpha$.

(ii) $\lim_{n\to\infty} P\{\text{reject } H_i\}$ *exists and is independent of* $i \in I(P)$.

(iii) If P *is such that* $i \notin I(P)$, *i.e.,* H_i *is false and* $\theta_i(P) \neq 0$, *then the probability that the stepdown method rejects* H_i *tends to one.*

(iv) Moreover, if the procedure rejects H_i *and it is declared that* $\theta_i(P) > 0$ *when* $\hat{\theta}_{n,i} > 0$, *and vice versa, then the probability of making a Type 3 error (i.e., of declaring* $\theta_i(P)$ *positive when it is negative or declaring it negative when it is positive) tends to 0.*

So far, the bootstrap construction has been based on Algorithm 3.1. But, asymptotic control of the k-FWER is also achieved by the computationally less expensive streamlined Algorithm 3.2.

Corollary 4.2 *The statements of Corollary 4.1 continue to hold if Algorithm 3.1 is replaced by Algorithm 3.2.*

Remark 4.1 (Operative Method) While the streamlined Algorithm 3.2 also results in asymptotic control of the k-FWER, finite sample considerations provide some motivation to base the bootstrap construction on the more conservative generic Algorithm 3.1. On the other hand, its computational burden can be very high. To compute a critical value $\hat{d}_{n,A_j,i}(1-\alpha,k)$ in the jth step, one has to evaluate $N_j = \binom{R_j}{k-1}$ quantiles $\hat{c}_{n,K,i}(1-\alpha,k)$ in order to then take the largest one of those. Depending on R_j and k, this number N_j may be very large. Therefore, we now suggest an operative method that retains some of the desirable properties of Algorithm 3.1 while remaining always computationally feasible. The suggestion is as follows. Pick a user specified number N_{max}, say $N_{max} = 50$, and let M be the largest integer for which $\binom{M}{k-1} \leq N_{max}$. In step j of Algorithm 3.1, a critical value is then computed as follows:

$$\hat{d}_{n,A_j,i}(1-\alpha,k) = \max_{I\subseteq\{r_{\max\{1,|R_j|-M+1\}},\ldots,r_{|R_j|}\},\ |I|=k-1} \{\hat{c}_{n,K,i}(1-\alpha,k) : K = A_j \cup I\} \ .$$

That is, we maximize over subsets I not necessarily of the entire index set R_j of previously rejected hypotheses but only of the index set corresponding to the M least significant hypotheses rejected so far. (Of course, when $M \geq |R_j|$, we maximize over all subsets $I \subseteq R_j$ of size $k-1$.) The philosophy of this operative method is to be as close as possible to the generic Algorithm 3.1, given the limitation to the computational burden expressed by N_{max}. Finally, note that the streamlined algorithm is a special case of the operative method when $N_{max} = 1$ is chosen, resulting in $M = k - 1$.

Remark 4.2 (Asymptotic sharpness) The $\limsup_{n\to\infty}$ in Corollary 4.1 (i) can actually replaced by a $\lim_{n\to\infty}$. Moreover, in the case $k = 1$, the inequality is an equality. For $k > 1$, the limiting value may be less than α. However, if the limiting distribution of the estimators is exchangeable, then equality holds. Nevertheless, the stepdown method represents a strict improvement over the single step method in that it leads to at least as many rejections, and the effect shows up asymptotically. Indeed, the limiting expression for k-FWER of the single-step procedure is given by (17) with $K = \{1, \ldots, s\}$, while the asymptotic expression for the stepdown procedure replaces $L_K^{-1}(1-\alpha, k, P)$ with the generally smaller value $L_{K_0}^{-1}(1-\alpha, k, P)$, where $K_0 \subseteq K$ is given by the set of true hypotheses $I(P)$ together with at most $k-1$ other indices. (Of course, the value will not strictly decrease if there are less than k hypotheses which are false.) The limiting value should be near α if $I(P)$ is large in comparison with k, because $L_{I(P)}^{-1}(1-\alpha, k, P)$ should be close to $L_{K_0}^{-1}(1-\alpha, k, P)$.

On the other hand, the inequality in Corollary 4.1 (i) is always an equality for the streamlined method of Algorithm 3.2.

4.2 A General Subsampling Construction

In this subsection, we sketch an alternative construction of critical values in our stepdown procedure by using subsampling. As in the bootstrap approach of Subsection 4.1, we assume H_i is concerned with the test of a parameter θ_i, but this can be generalized. Quite generally, the approach based on subsampling will hold under weaker asymptotic conditions than required for the bootstrap.

We now detail the general subsampling construction in the case of n i.i.d. observations $X_1, \ldots, X_n$ from P. The previous bootstrap estimators $H_{n,i}(\cdot, \hat{Q}_n)$ and $L_{n,K}(\cdot, k, \hat{Q}_n)$ are replaced by subsampling estimators as follows. Fix a positive integer $b < n$ and let $Y_1, \ldots, Y_{N_n}$ be equal to the $N_n := \binom{n}{b}$ subsets of $\{X_1, \ldots, X_n\}$, ordered in any fashion. Let $\hat{\theta}_{b,i}^{(a)}$ be equal to the statistic $\hat{\theta}_{n,i}$ evaluated at the data set Y_a, for $a = 1, \ldots, N_n$. The subsampling estimator

of $H_{n,i}(\cdot, P)$ is then given by

$$\hat{H}_{n,i}(x) = \frac{1}{N_n} \sum_a I\{\tau_b |\hat{\theta}_{b,i}^{(a)} - \hat{\theta}_{n,i}| \le x\} \ . \tag{23}$$

We also define

$$\hat{L}_{n,K}(x,k) = \frac{1}{N_n} \sum_a I\{k\text{-max}(\hat{H}_{n,i}(\tau_b |\hat{\theta}_{b,i}^{(a)} - \hat{\theta}_{n,i}|) \le x\} \ . \tag{24}$$

If we replace the bootstrap estimators by these subsampling estimators, we can prove a result analogous to Theorem 2.1, while removing the assumption B4.

Theorem 4.1 *Suppose data is generated from P satisfying Assumptions B1–B3. Fix $K \subseteq \{1, \ldots, s\}$ and a positive integer k. Let $b \to \infty$, $b/n \to 0$ and $\tau_b/\tau_n \to 0$. Consider the joint confidence region rectangle, with marginal intervals $\tilde{C}_{n,i}$ for $\theta_i(P)$ with $i \in K$ expressed as*

$$\tilde{C}_{n,i} \equiv \hat{\theta}_{n,i} \pm \tau_n^{-1} \hat{H}_{n,i}^{-1}(\hat{L}_{n,K}^{-1}(1-\alpha, k)) \ . \tag{25}$$

(i) For $i \in K$, the intervals $\tilde{C}_{n,i}$, simultaneously cover all the corresponding true parameter values $\theta_i(P)$, except for at most $k-1$ of them, with asymptotic probability $1-\alpha$.

(ii) The intervals $\tilde{C}_{n,i}$ are balanced in the sense that

$$\lim_{n\to\infty} P\{\theta_i(P) \in \tilde{C}_{n,i}\} = \gamma, \quad \text{independent of } i \ , \tag{26}$$

where $\gamma = \gamma_K(1-\alpha, k, P)$ is the unique $1-\alpha$ quantile of the limiting distribution $L_K(k, P)$.

The proof is analogous to the proof of Theorem 2.1, except that the uniform convergence of the subsampling estimators (in probability) is proved by the now standard arguments for subsampling; see Politis et al. (1999, Chapter 2). Thus, the result also generalizes quite easily; for example, in a stationary time series model, one only considers subsamples of consecutive observations; see Politis et al. (1999, Chapter 3). ■

Remark 4.3 For testing a single hypothesis H_i, $\tau_n|\hat{\theta}_{n,i}|$ is compared to the $1-\alpha$ quantile of the subsampling distribution based on the N_n values $\tau_b|\hat{\theta}_{b,i}^{(a)} - \hat{\theta}_n|$. Another possibility is to not "center" the subsampling values by instead using the N_n values of $\tau_b|\hat{\theta}_{b,i}^{(a)}|$. In fact, both approaches are asymptotically equivalent under the null hypothesis and under contiguous alternatives, at least when $k = 1$. The former approach more closely matches the bootstrap approach introduced earlier. The latter approach makes it easier to reject hypotheses because

the critical value is generally smaller. In Section 2.6 of Politis, et. al. (1999), the latter approach was used, as it generalizes easily to other types of hypotheses (such as when using a Kolmogorov-Smirnov type of statistic). When testing many hypotheses, the two approaches are not asymptotically equivalent because, if one does not "center", the subsampling critical value does not settle down against a fixed alternative. (This is not an issue with one hypothesis because the test statistic would then be growing at an even faster rate.) As a consequence, if one does not center when considering multiple hypotheses at once, the subsampled values for the test statistics corresponding to false null hypotheses will tend to be much larger than those corresponding to true hypotheses, and the result is that the estimate $\hat{L}_{n,K}(\cdot, k)$ will be too large if $k > 1$, and will negate the effects of utilizing a weaker measure of error control. For purposes of k-FWER control with $k > 1$, we recommend centering the subsampling distribution. However, we also note that sometimes there are advantages to not doing so, as in the control of the false discovery rate considered in Romano et al. (2007).

In the case $k = 1$, not centering the subsampled values can be advantageous in that it results in more powerful procedure. For example, suppose $(X_1, Y_1), \ldots, (X_n, Y_n)$ is a sample of n i.i.d. observations with $X_i \sim N(\theta_1, 1)$, $Y_i \sim N(\theta_2, 1)$ and X_i independent of Y_i. If, for example, $\theta_1 < 0$ and $\theta_2 > 0$, then the centered subsampling approach (as well as the bootstrap) will be based on a single-step critical value which behaves asymptotically like the $1-\alpha$ quantile of $\max(Z_1, Z_2)$, where the Z_i are i.i.d. $N(0, 1)$. On the other hand, if subsampling is used with no centering, then the single-step critical value will behave asymptotically like $z_{1-\alpha}$ because the subsampled averages of the Y_is will asymptotically dominate those based on the X_is. A smaller critical value then implies greater power.

We can also provide a stepdown improvement by applying the stepdown Algorithm 3.1 with the critical values

$$\hat{c}_{n,I,i}(1-\alpha, i) = \hat{H}_{n,i}^{-1}(\hat{L}_{n,K}^{-1}(1-\alpha, k)) \ .$$

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

$$\hat{c}_{n,K,i}(1-\alpha, k) \geq \hat{c}_{n,I,i}(1-\alpha, k) \ . \tag{27}$$

This simple observation together with Theorems 3.1 and 4.1 immediately yields an asymptotic improvement. The details are left to the reader.

5 Planned Imbalance and Weighted Control of k-FWER

Lack of balance is especially undesirable if hypotheses which we would like to treat equally are treated unequally. However, sometimes lack of balance is desirable, if it is handled appropriately. For example, if the various hypotheses are not equally important, we might want to control for rejection error by allocating different weights to the hypotheses.

Consider the general setting of testing hypotheses $H_1, \ldots, H_s$ based on data X from P, where H_i specifies $P \in \omega_i$. Assume $\hat{p}_{n,i}$ is a p-value for testing H_i in the sense

$$P\{\hat{p}_{n,i} \leq u\} \leq u \quad \text{for all } u \, , P \in \omega_i \, . \tag{28}$$

Suppose H_i is given weight w_i, where $\sum_i w_i = 1$. For example, the weighted Bonferroni method rejects any H_i such that $\hat{p}_{n,i} \leq w_i \alpha$ This controls the usual FWER with $k = 1$. (Note that hypotheses with larger weights w_i are given more importance.) In this section, we show how to contructed such weighted procedures which control the k-FWER, and at the same time provide a stepdown improvement.

Theorem 5.1 *Consider the problem of testing $H_1, \ldots, H_s$ with marginal p-values satisfying (28). Assume w_i are known weights with $\sum_{i=1}^{s} w_i = 1$.*

(i) (Weighted generalized Bonferroni)
The single-step procedure which rejects H_i if $\hat{p}_{n,i} \leq w_i k \alpha$ controls the k-FWER; that is

$$k\text{-}FWER_P \leq \alpha \, . \tag{29}$$

Moreover, if $\hat{p}_{n,i}$ has a uniform $(0,1)$ distribution whenever H_i is true, then $P\{H_i$ is rejected$\} = w_i k \alpha \propto w_i$.

(ii) (Weighted generalized Holm)
The stepdown procedure using Algorithm 3.1 with $T_{n,i} = -\hat{p}_{n,i}$ and

$$\hat{c}_{n,K}(1-\alpha, k) = -\frac{w_i}{\sum_{j \in K} w_j} k\alpha$$

also satisfies (29).

The computational application of Algorithm 3.1 is straightforward. The algorithm can be translated as follows. First, reject any H_i whose corresponding p-value $\hat{p}_{n,i}$ satisfies $\hat{p}_{n,i} \leq w_i k \alpha$; that is, apply the single-step procedure. If there are fewer than k rejections, then stop. (Of

course, there is the possibility of allowing up to $k-1$ rejections.) If there are k or more rejections, we can next test the remaining p-values as follows. Let A be the indices of hypotheses not yet rejected and let $s_A = \sum_{j \in A} w_j$. Let R be the indices of hypotheses already rejected, and let s_R be the sum of the $k-1$ largest values among w_j with $j \in R$. Compare $\hat{p}_{n,i}$ with $w_i k\alpha/(s_A + s_R)$. If there are no further rejections, then stop; otherwise, continue in the same fashion after updating both A and R.

6 Control of Average Number of False Rejections

In this section, we briefly consider control of the expected number of false rejections; see (4). Suppose p-values $\hat{p}_{n,i}$ are available for testing H_i, so that (28) holds. As is well-known, the procedure which rejects H_i if $\hat{p}_{n,i} \leq \lambda/s$ satisfies (4). More generally, and analogous to Theorem 5.1, the following is true.

Theorem 6.1 *Consider the problem of testing $H_1, \ldots, H_s$ with marginal p-values satisfying (28). Assume w_i are known weights with $\sum_{i=1}^{s} w_i = 1$. Then, the single-step procedure which rejects H_i if $\hat{p}_{n,i} \leq w_i\lambda$ controls the average number of false rejections; that is, (4) holds. Moreover, if $\hat{p}_{n,i}$ has a uniform $(0,1)$ distribution whenever H_i is true and $w_i\lambda$ is ≤ 1, then $P\{H_i \text{ is rejected}\} = w_i\lambda \propto w_i$.*

For finite-sample control of the average number of false rejections, a stepdown improvement is not possible. To see why, suppose $w_i = 1/s$, all H_i are true, and $\hat{p}_{n,i}$ has a $U(0,1)$ distribution. Then, the expected number of false rejections of the above procedure is exactly λ. If the possibility of further rejections were allowed, then the average number of false rejections must necessarily increase, which would violate error control given by (4). (Note it is asymptotically possible to provide a stepdown improvement, but this is not pursued here. For example, with $w_i = 1/s$, one could attempt to estimate or bound the number of true null hypotheses by $\hat{I}$ and then replace the critical value λ/s with $\lambda/\hat{I}$.)

If exact p-values are not available, one can use subsampling or the bootstrap, as in (18). Of course, by linearity of expectation, no further modification of the procedure is needed to take into account the dependence of the test statistics.

7 Asymptotic Results on FDP Control

In some applications, one might be willing to tolerate a certain small fraction of false rejections out of the total rejections. This leads to control based on the *false discovery proportion* (FDP). Let F be the number of false rejections made by a multiple testing procedure and let R be the total number of rejections. Then the FDP is defined as follows:

$$\text{FDP} = \begin{cases} \frac{F}{R} & \text{if } R > 0 \\ 0 & \text{if } R = 0 \end{cases}$$

A multiple testing procedure is said to control the FDP at level α if, for the given sample size n, $P\{\text{FDP} > \gamma\} \leq \alpha$, for all P. A multiple testing procedure is said to asymptotically control the FDP at level α, if $\limsup_n P\{\text{FDP} > \gamma\} \leq \alpha$, for all P. Our focus will be on procedures that provide asymptotic control. Notice that a procedure satisfying $P\{\text{FDP} > \gamma\} \leq 0.5$ guarantees that the median of the FDP is $\leq \gamma$. The main goal of this section is to construct a method which provides asymptotic control of the FDP.

The approach we propose is built upon an underlying procedure that (asymptotically) controls the k-FWER for any fixed $k \geq 1$. We then sequentially apply this k-FWER procedure for $k = 1, 2, \ldots$ until a stopping rule indicates termination. In the end, we reject all hypotheses that were rejected in the last round of applying the k-FWER procedure.

To develop the idea, consider controlling $P\{\text{FDP} > 0.1\}$. We start out by applying the 1-FWER procedure, that is, by (asymptotically) controlling the FWER. Denote by N_1 the number of hypotheses rejected. Due to the FWER control, one can be confident that no false rejection has occurred and that, in return, the FDP has been controlled. Consider now rejecting $H_{r_{N_1+1}}$, the next most significant hypothesis. Of course, if $H_{r_{N_1+1}}$ is false, there is nothing to worry about, so suppose $H_{r_{N_1+1}}$ is true. In case the FWER was controlled successfully in the first step, the FDP upon rejection of $H_{r_{N_1+1}}$ then becomes $1/(N_1 + 1)$, which is greater than 0.1 if and only if $N_1 < 9$. So if $N_1 \geq 9$ we can reject one true hypothesis and still avoid FDP > 0.1. This suggests to stop if $N_1 < 9$ and otherwise to apply the 2-FWER procedure which, by design, controls the probability of making two or more false rejections. Denote the total number of hypotheses rejected by the 2-FWER base procedure by N_2. Reasoning similarly to before, if $N_2 < 19$, we stop and otherwise we apply the 3-FWER procedure. If N_j denotes the total number of hypotheses rejected by the j-FWER procedure, the stepdown method is continued until $N_j < 10j - 1$, at which point termination incurs.

The following algorithm summarizes the method for arbitrary γ.

Algorithm 7.1 (Generic Method for Control of the FDP)

1. Let $j = 1$ and let $k_1 = 1$.
2. Apply the k_j-FWER procedure and denote by N_j the number of hypotheses it rejects.
3. (a) If $N_j < k_j/\gamma - 1$, stop and reject all hypotheses rejected by the k_j-FWER procedure.
   (b) Otherwise, let $j = j + 1$ and then $k_j = k_{j-1} + 1$. Return to step 2.

Note that the algorithm does not specify the underlying k-FWER procedure. However, in order to reject as many false hypotheses as possible while maintaining (asymptotic) control of the FDP, we suggest to employ a stepdown procedure which accounts for the dependence structure of the test statistics $T_{n,i}$. Algorithm 7.1 is similar to the proposal of Korn et al. (2004) for FDP control which is, however, restricted to a multivariate permutation model. The proposal of Korn et al. (2004) is heuristic in the sense that they cannot guarantee finite sample nor asymptotic control of the FDP even if the permutation hypothesis is valid. In Romano and Wolf (2007), asymptotic control of Algorithm 7.1 is established when using a bootstrap or subsampling approach for the underlying k-FWER procedure, with simulations showing good finite sample control. The theorem establishes the corresponding result if one uses a balanced k-FWER controlling procedure. The result covers a general bootstrap construction where the individual tests are two-sided and concern univariate parameters $\theta_i(P)$. The bootstrap construction for one-sided tests and the more general subsampling construction can be handled similarly. The proofs are very similar to the unbalanced cases established in Romano and Wolf (2007).

Theorem 7.1 *Consider the setup of Corollary 4.1. Fix P satisfying Assumptions B1–B3. Let $\hat{Q}_n$ be an estimate of P satisfying B4. Employ the stepdown procedure of Algorithm 3.1 with $\hat{c}_{n,K,i}(1-\alpha, k)$ as the underlying k-FWER procedure. Then the following statements concerning Algorithm 7.1 are true.*

(i) $\limsup_{n\to\infty} P\{FDP > \gamma\} \leq \alpha$.

(ii) If P is such that $i \notin I(P)$, i.e., H_i is false and $\theta_i(P) \neq 0$, then the probability that the method rejects H_i tends to one.

Remark 7.1 The theorem remains valid if the stepdown bootstrap k-FWER procedure is based on the operative method of Remark 4.1 or even the streamlined Algorithm 3.2 instead of the generic Algorithm 3.1. But, again, in view of finite sample performance, we suggest the use of the generic Algorithm 3.1 if feasible or at least the use of the operative method.

8 Simulation Study

This section presents a small simulation study in the context of testing population means. We generate random vectors $X_1, \ldots, X_n$ from an s-dimensional multivariate normal distribution with mean vector $\theta = (\theta_1, \ldots, \theta_s)$, where $n = 100$ and $s = 40$. The null hypotheses are $H_i : \theta_i(P) = 0$ and the alternative hypotheses are $H_i : \theta_i(P) \neq 0$. define

$$\bar{X}_{n,i,\cdot} = \frac{1}{n}\sum_{j=1}^{n} X_{i,j} \quad \text{and} \quad \hat{\sigma}^2_{n,i} = \frac{1}{n-1}\sum_{j=1}^{n}(X_{i,j} - \bar{X}_{n,i,\cdot})^2 \ .$$

Then we use $\hat{\theta}_{n,i} = \bar{X}_{n,i,\cdot}$ and $\tau_n = \sqrt{n}$.

The individual means $\theta_i(P)$ are equal to either 0 or 0.4. The number of means equal to 0.4 is 0, 10, 20, or 40. Denote the elements of the covariance matrix by $\sigma_{i,j}$. Then half of the $\sigma_{i,i}$ are equal to 1 while the other half are equal to 4. This is done in a way such that both the 'null' variables and the 'alternative' variables have half of their variances equal to 1 and the other half equal to 4. The correlation ρ is constant; that is, $\sigma_{i,j}/\sqrt{\sigma_{i,i}\sigma_{j,j}} = \rho$ for all $i \neq j$. We employ $\rho = 0.0, 0.5$.

The goal is to compare the balanced bootstrap procedures of this paper with the stepwise bootstrap procedures of Romano and Wolf (2007) based on the maximum test statistic. For the latter procedures, the individual test statistics $T_{n,i}$ are either basic (i.e., non-studentized) or studentized, that is

$$T^{bas}_{n,i} = \tau_n|\hat{\theta}_{n,i}| \qquad \text{or} \qquad T^{stud}_{n,i} = \tau_n|\hat{\theta}_{n,i}|/\hat{\sigma}_{n,i}$$

The abbreviations for the included procedures are as follows.

- (**k-maxT^{bas}**) The bootstrap k-FWER procedure of Romano and Wolf (2007) with $T^{bas}_{n,i}$.
- (**k-maxT^{stud}**) The bootstrap k-FWER procedure of Romano and Wolf (2007) with $T^{stud}_{n,i}$.
- (**k-bal^{bas}**) The balanced bootstrap k-FWER procedure of Subsection 4.1 with $\tau_n|\hat{\theta}_{n,i}|$.
- (**k-bal^{stud}**) A balanced bootstrap k-FWER procedure analogous to Subsection 4.1 but with studentized roots $\tau_n|\hat{\theta}_{n,i}|/\hat{\sigma}_{n,i}$; see Remarks 2.5 and 2.7.
- (**FDP-maxT^{bas}**) The bootstrap FDP procedure of Romano and Wolf (2007) with $T^{bas}_{n,i}$.
- (**FDP-maxT^{stud}**) The bootstrap FDP procedure of Romano and Wolf (2007) with $T^{stud}_{n,i}$.
- (**FDP-bal^{bas}**) The balanced bootstrap FDP procedure of Section 7 with $\tau_n|\hat{\theta}_{n,i}|$.

- **(FDP-bal^{stud})** A balanced bootstrap k-FWER procedure analogous to Section 7 but with $\tau_n|\hat{\theta}_{n,i}|/\hat{\sigma}_{n,i}$.

In order to properly estimate an appropriate quantile, one must employ a large number of bootstrap resamples, denoted by B. In effect, one needs to construct individual confidence intervals at level γ, where γ is close to one. *Ceteris paribus*, γ increases with the number of hypotheses. To make the point, assume it is known that the individual estimators $\hat{\theta}_{n,i}$ are independent of each other. In this case, γ is given by $\gamma = (1-\alpha)^{1/s}$. The larger γ, the larger should be B; see Efron and Tibshirani (1993, Section 19.3). The computational burden we can handle corresponds to $B = 10,000$. For that reason we chosse the relatively small value of $s = 40$ individual hypotheses. Furthermore, we use $\alpha = 0.1$ rather than $\alpha = 0.05$. The value of N_{max} for the operative method is $N_{max} = 50$; see Remark 4.1.

The values of k for k-FWER control we consider are $k = 1, 3$. The latter value is relatively small, since $s = 40$ is relatively small. For the same reason, we have to chose the value of γ for FDP control relatively large, or the differences between control of the 1-FWER and control of the FDP would hardly show up. Therefore, we use $\gamma = 0.2$.

The performance criteria are (i) the various empirical error rates, compared to the nominal level $\alpha = 0.1$; (ii) the average number of false hypotheses rejected; and (iii) the empirical imbalance. The latter is defined as the difference between the maximal and the minimal empirical rejection probabilities over all true null hypotheses. In other words, if the empirical rejection probability of null hypothesis H_i is denoted by $e.r.p._i$, then the empirical imbalance is defined as

$$\max_{i \in I(P)} e.r.p._i - \min_{i \in I(P)} e.r.p._i \ .$$

(When all null hypotheses are false, this measure is not defined.) Note that due to sampling error, the empirical imbalance will typically be positive even if a procedure is perfectly balanced. The performance criteria are computed from 5,000 repetitions in each scenario. For every repetition (i.e., every simulated data set), the same set of B = 10,000 bootstrap resamples is shared by all procedures.

The results are presented in Table 1 and can be summarized as follows.

- Because the $\sigma_{i,i}$ are different, k-maxT^{bas} results in asymptotically unbalanced inference. Due to studentization, k-maxT^{stud} is invariant to the $\sigma_{i,i}$ and yields asymptotically balanced inference. This is reflected in the empirical balances which are always larger for k-maxT^{bas}, and sometimes much larger.

- If balance is applied to the basic method, resulting in k-bal^{bas}, then the empirical balances become comparable to k-maxT^{stud}. On the other hand, if balance is applied to the studentized method, resulting in k-bal^{stud}, no meaningful further improvement over k-maxT^{stud} is achieved.

- Both k-maxT^{bas} and k-maxT^{stud} achieve satisfactory control of the k-FWER. However, k-maxT^{bas} is always less powerful compared to k-maxT^{stud}.

- k-bal^{bas} is somewhat liberal, which explains its somewhat larger power compared to k-maxT^{stud}. On the other hand, k-bal^{stud} performs very similarly compared to k-maxT^{stud} both in terms of k-FWER control and power.

- The comparisons are similar with respect to FDP control as opposed to k-FWER control.

Remark 8.1 *Ceteris paribus*, the finite-sample control of k-bal^{bas} improves with both k and n. Some evidence for the former claim can be seen in Table 1. Unfortunately, running a complete simulation study with a large n is computationally too expensive. But we considered the case of all $\theta_i = 0$ and common correlation $\rho = 0$, and increased the sample size from $n = 100$ to $n = 400$. The empirical controls improve from 13.4% to 10.7% (for 1-FWER and FDP) and from 11.6% to 10.2% (for 3-FWER).

In addition, it is also advisable to choose the number of bootstrap resamples, B, as large as possible, given the computational resources. But at least for the scenario with $n = 100$, all $\theta_i = 0$, and common correlation $\rho = 0$, increasing the number of bootstrap resamples from $B = 10,000$ to $B = 50,000$ made virtually no difference.

9 Concluding Remarks

We have shown how computationally feasible stepdown methods can be constructed to control generalized error rates in multiple testing. On the one hand, we have considered the k-FWER, which is defined as the probability of making k or more false rejections. This concept would be appropriate when a given number of false rejections can be tolerated. On the other hand, we have also considered the FDP, which is the ratio of false rejections out of the total number of rejections (and defined to be zero when there are no rejections). This concept would be appropriate when a certain proportion of false rejections can be tolerated. Some simulations have shown that these less strict methods can reject many more false hypotheses compared to the traditional FWER control, especially when the number of hypotheses under test is large.

Our stepdown methods (asymptotically) account for the dependence structure across test statistics. As a result, they are more powerful than the generalized Holm stepdown methods of Hommel and Hoffman (1988) and Lehmann and Romano (2005a), which are based on individual p-values and designed to handle a 'worst case' dependence structure. An alternative approach that also accounts for the dependence structure across test statistics is the augmentation approach of van der Laan et al. (2004). However, simulations show their methods are noticeably less powerful, especially when the number of hypotheses under test is large. The empirical Bayes method of van der Laan et al. (2005) can sometimes be more powerful than our bootstrap approach for FDP control. However, it also can be quite liberal and it does not offer asymptotic control of the FDP when all null hypotheses are true. Overall, our methods for control of the k-FWER and FDP appear competitive with or outperform currently available methods.

A Proofs

PROOF OF LEMMA 2.1: Fix P and let

$$Y_{n,i} = \tau_n[\hat{\theta}_{n,i} - \theta_i(P)] \ . \tag{30}$$

By the Almost Sure Representation Theorem, we can assume there exist versions $Y^*_{n,i}$ such that $(Y^*_{n,1}, \ldots, Y^*_{n,s})$ has the same distribution as $(Y_{n,1}, \ldots, Y_{n,s})$ and $Y^*_{n,i} \to Y_i$ almost surely for every i. We must show that

$$k\text{-max}\left(H_{n,i}(|Y^*_{n,i}|, P), i \in K\right) \tag{31}$$

has a limiting distribution. But, since $J_{n,i}(P)$ has a continuous limiting distribution with c.d.f. $J_i(\cdot, P)$, then by the Continuous Mapping Theorem, $H_{n,i}(P)$ has a limiting distribution $H_i(P)$ with c.d.f. given by (13). By Polya's Theorem, $H_{n,i}(x, P) \to H_i(x, P)$ uniformly in x. Therefore, by continuity of the k-max function, the difference between (31) and

$$k\text{-max}\left(H_i(|Y^*_{n,i}|, P), i \in K\right) \tag{32}$$

tends to 0. But, by continuity of the $H_i(\cdot, P)$ and the k-max function, we have that (32) tends almost surely to k-max$(H_i(|Y_i|, P), i \in K)$, and hence in distribution as well.

To show that this limiting distribution is continuous, note that

$$P\{k\text{-max}(H_i(|Y_i|, P), i \in K) = x\} \leq \sum_{i \in K} P\{H_i(|Y_i|, P) = x\} = 0 \ ,$$

because, for every i, $H_i(|Y_i|, P)$ has the uniform distribution on $(0, 1)$. ■

PROOF OF LEMMA 2.2: To see that the c.d.f. of Y is continuous, simply note that

$$P\{Y = x\} \leq \sum_{1 \leq i \leq s} P\{h_i(X_i) = x\} = 0 \ ,$$

where the final equality follows from the assumption that $h_i(X_i)$ has a continuous distribution. To see that the c.d.f. of Y is strictly increasing, suppose by way of contradiction that there exists $a < b$ such that $P\{Y \in (a, b)\} = 0$, but $P\{Y \leq a\} > 0$ and $P\{Y \geq b\} > 0$. Thus, there exists $x = (x_1, \ldots, x_s) \in \text{supp}(F)$ such that $k\text{-max}(h_1(x_1), \ldots, h_s(x_s)) \leq a$ and $x' \in \text{supp}(F)$ such that $k\text{-max}(h_1(x'_1), \ldots, h_s(x'_s)) \geq b$. Consider the set

$$A_{a,b} = \{x \in \text{supp}(X) : a < k\text{-max}(h_1(x_1), \ldots, h_s(x_s)) < b\} \ .$$

By continuity of the k-max function and assumption (ii), $A_{a,b}$ is non-empty. Moreover, again by continuity of the k-max function $A_{a,b}$ must contain an open subset of supp(F) (relative to the topology on supp(X)). It therefore follows by the definition of supp(X) that

$$P\{X \in A_{a,b}\} = P\{k\text{-max}(h_1(X_1), \ldots, h_s(X_s)) \in (a,b)\} > 0 \ ,$$

which yields the desired contradiction. ■

Proof of Corollary 2.1. For ease of notation, the proof is presented in the case $K = \{1, \ldots, s\}$ (with no loss of generality). Recall the limiting distribution of $L_K(k, P)$ can be represented by the distribution of (12). The assumptions of Lemma 2.2 are satisfied. Indeed, suppose $(X_1, \ldots, X_s)$ has distribution $J_{\{1,\ldots,s\}}(P)$. Take $h_i(x) = H_i(|x|, P)$. Note that $H_i(|X_i|, P)$ has the uniform (0,1) distribution, which is continuous. The connectedness assumption holds by Assumption B3. ■

Proof of 2.1. By Lemma 2.1,

$$L_{n,K}(\cdot, k, P) \xrightarrow{L} L_K(\cdot, k, P) \ .$$

Moreover, by Corollary 2.1 we can conclude that the c.d.f. $L_K(\cdot, k, P)$ is continuous and strictly increasing with unique inverse function

$$\gamma_K(1-\alpha, k, P) = L_K^{-1}(1-\alpha, k, P) \ .$$

It follows by Lemma 11.2.1 of Lehmann and Romano (2005b) that

$$L_{n,K}^{-1}(1-\alpha, k, P) \to \gamma_K(1-\alpha, k, P) \ . \tag{33}$$

But, we can apply the identical argument to get a triangular array convergence result simply by replacing P by a sequence P_n; it follows that for any sequence $\{P_n\}$ satisfying

$$\rho(J_{n,K}(P_n), J_K(P)) \to 0 \ ,$$

we have

$$L_{n,K}(k, P_n) \xrightarrow{L} L_K(k, P)$$

and

$$L_{n,K}^{-1}(1-\alpha, k, P_n) \to \gamma_K(1-\alpha, k, P) \ .$$

But, by virtue of Assumption B4 and a subsequence argument, it follows that

$$L_{n,K}^{-1}(1-\alpha, k, \hat{Q}_n) \xrightarrow{P} \gamma_K(1-\alpha, k, P) \ . \tag{34}$$

Then,

$$P\{\theta_i \in \hat{C}_{n,i} \text{ except for at most } k-1 \text{ of the } i \in K\} =$$

$$P\{k\text{-max}(H_{n,i}(\tau_n|\hat{\theta}_{n,i} - \theta_i(P)|, \hat{Q}_n), i \in K) \leq L_{n,K}^{-1}(1-\alpha, k, \hat{Q}_n)\} \ . \tag{35}$$

But, by Assumption B2, Polya's Theorem and a subsequence argument,

$$\sup |H_{n,i}(x, \hat{Q}_n) - H_i(x, P)| \xrightarrow{P} 0 \ ,$$

where $H_i(x, P) = J_i(x, P) - J_i(-x, P)$. So, the random variable on the left side of the inequality in (35) is

$$k\text{-max}(H_i(\tau_n|\hat{\theta}_{n,i} - \theta_i(P)|, P), i \in K) + o_P(1) \ . \tag{36}$$

To examine the limiting distributional behavior of (36), let $Y_{n,i} = \tau_n[\hat{\theta}_{n,i} - \theta_i(P)]$. By the Almost Sure Representation Theorem, we can assume there exists versions $Y_{n,i}^*$ with $(Y_{n,1}, \ldots, Y_{n,s})$ having the same distribution as $(Y_{n,1}^*, \ldots, Y_{n,s}^*)$ such that $Y_{n,i}^* \to Y_i$ almost surely, for all i, where $(Y_1, \ldots, Y_n)$ has distribution $J_{\{1,\ldots,s\}}(P)$. It follows that (36) converges in distribution to the distribution of $k\text{-max}(|Y_i|, i \in K)$, which is exactly $L_K(\cdot, k, P)$. We can now apply Slutsky's Theorem to evaluate (35) to conclude its limiting probability is

$$P\{k\text{-max}(|Y_i|, i \in K) \leq \gamma_K(1-\alpha, k, P)\} = 1 - \alpha \ .$$

To prove (ii),

$$P\{\theta_i(P) \in \hat{C}_{n,i}\} = P\{\tau_n|\hat{\theta}_{n,i} - \theta_i(P)| \leq H_{n,i}^{-1}(L_{n,K}^{-1}(1-\alpha, k, \hat{Q}_n), \hat{Q}_n)\}$$

$$= P\{H_{n,i}(\tau_n|\hat{\theta}_{n,i} - \theta_i(P), \hat{Q}_n) \leq L_{n,K}^{-1}(1-\alpha, k, \hat{Q}_n)\} \ . \tag{37}$$

But, a similar argument to the above by invoking the Almost Sure Representation Theorem (taking $K = \{i\}$) gives that

$$H_{n,i}(\tau_n|\hat{\theta}_{n,i} - \theta_i(P)|, \hat{Q}_n) \xrightarrow{L} H_i(|Y_i|, P)$$

which is uniform $U(0,1)$. Since the right side of (37) tends in probability to $\gamma_K(1-\alpha, k, P)$, the result follows by Slutsky's Theorem. ■

Proof of Corollary 2.2. Using the arguments as in the proof of Theorem 2.1, we can calculate an exact limiting expression (rather than just the bound α). If $(Y_1, \ldots, Y_s)$ is a random vector with distribution $J_{\{1,\ldots,s\}}(P)$, then

$$\lim_{n\to\infty} k\text{-FWER}_P = P\{k\text{-max}(J_i(|Y_i|, P), i \in I(P)) > L_K^{-1}(1-\alpha, k, P)\} \ ,$$

with $K = \{1, \ldots, s\}$. The previous expression is exactly α if $K = I(P)$, but since we always have

$$L_{I(P)}^{-1}(1-\alpha, k, P)\} \leq L_K^{-1}(1-\alpha, k, P)\} \ ,$$

the inequality in the corollary follows. To prove (ii), we can calculate

$$\begin{aligned} \lim_{n\to\infty} P\{\text{reject } H_i\} &= P\{J_i(|Y_i|, P) > L_K^{-1}(1-\alpha, k, P)\} \\ &= P\{U_i > L_K^{-1}(1-\alpha, k, P)\} \ , \end{aligned}$$

where $U_i \sim U(0,1)$, and the result follows. ■

PROOF OF THEOREM 3.1 Assume $|I(P)| \geq k$, or there is nothing to prove. Consider the event that at least k true null hypotheses are rejected. Let $\hat{j}$ be the smallest (random) index j in the algorithm where this occurs, so that at least k of the $T_{n,i}$ with $i \in I(P)$ satisfy

$$T_{n,i} > \hat{d}_{n,A_{\hat{j}},i}(1-\alpha, k) \ .$$

By definition of $\hat{j}$ (now fixed), $I(P) \subseteq A_{\hat{j}} \cup I_0$, where I_0 is some set of indices satisfying $I_0 \subseteq R_{\hat{j}}$ and $|I_0| = k-1$. Let L be any set of indices of false null hypotheses which satisfy $A_{\hat{j}} \cup I_0 = I(P) \cup L$. Since $\hat{d}_{n,A_{\hat{j}},i}(1-\alpha, k)$ is defined by taking the maximum over sets I of $\hat{c}_{n,K,i}(1-\alpha, k)$ with $K = A_{\hat{j}} \cup I$ as I varies over indices satisfying $I \subseteq R_{\hat{j}}$ and $|I| = k-1$, it follows that $\hat{d}_{n,A_{\hat{j}},i}(1-\alpha, k) \geq \hat{c}_{n,I(P)\cup L,i}(1-\alpha, k)$. By the monotonicity assumption,

$$\hat{c}_{n,I(P)\cup L,i}(1-\alpha, k) \geq \hat{c}_{n,I(P),i}(1-\alpha, k) \ .$$

To summarize, the event that at least k true null hypotheses are rejected implies that at least k of the $T_{n,i}$ with $i \in I(P)$ satisfy

$$T_{n,i} > \hat{c}_{n,I(P),i}(1-\alpha, k)$$

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

PROOF OF COROLLARY 4.1 The proofs of parts (i)–(ii) follow from the arguments preceding the corollary. The proofs of parts (iii)–(iv) are very similar to the proofs of parts (iii)-(iv) of Theorem 3.2 in Romano and Wolf (2007). ■

PROOF OF THEOREM 5.1. To prove (i), let F be the number of false rejections and let $I(P)$ denote the set of true null hypotheses. Then, using Markov's inequality,

$$k\text{-FWER}_P = P\{F \geq k\} \leq \frac{E(F)}{k}$$

$$= \frac{1}{k} E\left(\sum_{i \in I(P)} I\{\hat{p}_n \leq w_i k\alpha\} \right) = \frac{1}{k} \sum_{i \in I(P)} P\{\hat{p}_n \leq w_i k\alpha\}$$

$$\leq \frac{1}{k} \sum_{i \in I(P)} w_i k\alpha = \alpha \sum_{i \in I(P)} w_i \leq \alpha \ .$$

To prove (ii), the result follows from Theorem 3.1 once we verify the monotonicity condition (19). But to show that monotonicity holds, let $I \subseteq K$. Then,

$$\hat{c}_{n,I,i}(1-\alpha,k) = -\frac{w_i}{\sum_{j \in I} w_j} k\alpha \leq -\frac{w_i}{\sum_{j \in K} w_j} k\alpha = \hat{c}_{n,I,i}(1-\alpha,k) \ . \blacksquare$$

Proof of Theorem 6.1.. Let F be the number of false rejections and let $I(P)$ denote the set of true null hypotheses. Then,

$$E_P(F) = E\left[\sum_{i \in I(P)} I\{\hat{p}_{n,i} \leq w_i \lambda\} \right] = \sum_{i \in I(P)} P\{\hat{p}_{n,i} \leq w_i \lambda\} \leq \lambda \sum_{i \in I(P)} w_i \leq \lambda \ .$$

The second statement is trivial. ■

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57(1):289–300.

Benjamini, Y., Krieger, A. M., and Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, 93(3):491–507.

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.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Application*. Cambridge University Press, Cambridge.

Dudoit, S., van der Laan, M. J., and Pollard, K. S. (2004). Multiple testing. Part I. Single-step procedures for control of general type I error rates. *Statistical Applications in Genetics and Molecular Biology*, 3(1):Article 13. Available at `http://www.bepress.com/sagmb/vol3/iss1/art13`.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York.

Genovese, C. R. and Wasserman, L. (2004). A stochastic process approach to false discovery control. *Annals of Statistics*, 32(3):1035–1061.

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.

Hommel, G. and Hoffman, T. (1988). Controlled uncertainty. In Bauer, P., Hommel, G., and Sonnemann, E., editors, *Multiple Hypthesis Testing*, pages 154–161. Springer, Heidelberg.

Korn, E. L., Troendle, J. F., McShane, L. M., and Simon, R. (2004). Controlling the number of false discoveries: application to high-dimensional genomic data. *Journal of Statistical Planning and Inference*, 124(2):379–398.

Lahiri, S. N. (2003). *Resampling Methods for Dependent Data*. Springer, New York.

Lehmann, E. L. and Romano, J. P. (2005a). Generalizations of the familywise error rate. *Annals of Statistics*, 33(3):1138–1154.

Lehmann, E. L. and Romano, J. P. (2005b). *Testing Statistical Hypotheses*. Springer, New York, third edition.

Perone Pacifico, M., Genovese, C. R., Verdinelli, I., and Wasserman, L. (2004). False discovery control for random fields. *Journal of the American Statistical Association*, 99(468):1002–1014.

Politis, D. N., Romano, J. P., and Wolf, M. (1999). *Subsampling*. Springer, New York.

Pollard, K. S. and van der Laan, M. J. (2003). Multiple testing for gene expression data: an investigation of null distributions with consequences for the permutation test. In *Proceedings of the 2003 International MultiConference in Computer Science and Engineering*, METMBS'03 Conference, pages 3–9.

Rogers, J. and Hsu, J. (2001). Multiple comparisons of biodiversity. *Biometrical Journal*, 43:617–625.

Romano, J. P. (1988). A bootstrap revival of some nonparametric distance tests. *Journal of the American Statistical Association*, 83(403):698–708.

Romano, J. P. and Shaikh, A. M. (2006a). On stepdown control of the false discovery proportion. In Rojo, J., editor, *IMS Lecture Notes—Monograph Series, 2nd Lehmann Symposium—Optimality*, pages 33–50.

Romano, J. P. and Shaikh, A. M. (2006b). Stepup procedures for control of generalizations of the familywise error rate. *Annals of Statistics*, 34(4):1850–1873.

Romano, J. P., Shaikh, A. M., and Wolf, M. (2007). Control of the false discovery rate under dependence using the bootstrap and subsampling. Working Paper 337, IEW, University of Zurich. Available at `http://www.iew.uzh.ch/publications/wp_en.html`.

Romano, J. P. and Wolf, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, 100(469):94–108.

Romano, J. P. and Wolf, M. (2007). Control of generalized error rates in multiple testing. *Annals of Statistics*, 35(4):1378–1408.

Sarkar, S. K. (2002). Some results on false discovery rate in stepwise multiple testing procedures. *Annals of Statistics*, 30(1):239–257.

Shao, J. and Tu, D. (1995). *The Jackknife and the Bootstrap*. Springer, New York.

Spjøtvoll, E. (1972). On the optimality of some multiple comparison procedures. *Annals of Mathematical Statistics*, 43:398–411.

Storey, J. D., Taylor, J. E., and Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach. *Journal of the Royal Statistical Society, Series B*, 66(1):187–205.

Tu, W. and Zhou, X. (2000). Pairwise comparison of the means of skewed data. *Journal of Statistical Planning and Inference*, 88:59–74.

van der Laan, M. J., Birkner, M. D., and Hubbard, A. E. (2005). Empirical bayes and resampling based multiple testing procedure controlling tail probability of the proportion of false positives. *Statistical Applications in Genetics and Molecular Biology*, 4(1):Article 29. Available at `http://www.bepress.com/sagmb/vol4/iss1/art29/`.

van der Laan, M. J., Dudoit, S., and Pollard, K. S. (2004). Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. *Statistical Applications in Genetics and Molecular Biology*, 3(1):Article 15. Available at `http://www.bepress.com/sagmb/vol3/iss1/art15/`.

Westfall, P. H. and Young, S. S. (1993). *Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment*. John Wiley, New York.

Table 1: Empirical FWERs and FDPs (in the rows 'Control'); average number of false hypotheses rejected (in the rows 'Rejected'); and empirical imbalances (in the rows 'Imbalance'), for various procedures, with $n = 100$ and $s = 40$. The nominal level is $\alpha = 10\%$. The number of repetitions is 5,000 per scenario and the number of bootstrap resamples is $B = 10,000$. Both The empirical error rates and imbalances are expressed in percentages. Table 2 explains which procedures correspond to which columns.

Common correlation: $\rho = 0$

	1	2	3	4	5	6	7	8	9	10	11	12
All $\theta_i = 0$												
Control	9.4	9.4	13.4	9.7	9.7	8.9	11.6	8.8	9.4	9.4	13.4	9.7
Rejected	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Imbalance	0.7	0.3	0.4	0.3	11.0	1.3	1.3	1.2	0.7	0.3	0.4	0.3
Ten $\theta_i = 0.4$												
Control	8.1	9.9	13.2	10.0	6.9	7.3	9.3	7.3	7.9	6.5	8.5	6.5
Rejected	1.4	4.9	5.2	4.9	5.7	6.9	7.0	6.9	1.4	5.7	6.9	5.7
Imbalance	0.8	0.3	0.3	0.3	7.1	1.1	1.3	1.0	0.8	0.7	0.7	0.7
Twenty $\theta_i = 0.4$												
Control	5.4	6.2	8.6	6.5	4.2	5.4	6.7	5.3	4.1	3.3	4.5	3.3
Rejected	2.9	10.1	10.6	10.1	12.1	14.3	14.5	14.3	4.4	14.8	15.0	14.8
Imbalance	0.8	0.2	0.3	0.3	7.5	1.4	1.5	1.4	2.4	1.2	1.3	1.1
All $\theta_i = 0.4$												
Control	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Rejected	6.4	21.8	22.7	21.8	30.6	33.1	33.5	33.1	29.4	38.5 1	38.5	38.5

Common correlation: $\rho = 0.5$

	1	2	3	4	5	6	7	8	9	10	11	12
All $\theta_i = 0$												
Control	10.2	10.2	12.8	10.4	11.5	10.6	12.3	10.7	10.2	10.2	12.8	10.4
Rejected	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Imbalance	1.2	0.4	0.4	0.4	12.7	1.6	1.6	1.6	1.2	0.4	0.4	0.4
Ten $\theta_i = 0.4$												
Control	8.9	8.7	11.3	8.8	8.3	8.7	9.7	8.7	8.6	8.1	9.5	8.1
Rejected	1.9	5.4	5.6	5.4	5.1	6.8	6.9	6.8	2.3	6.0	6.2	6.0
Imbalance	1.2	0.3	0.4	0.3	5.5	0.8	0.9	0.8	2.6	0.6	0.7	0.5
Twenty $\theta_i = 0.4$												
Control	7.4	7.9	9.9	8.0	7.4	8.5	9.5	8.6	8.3	7.5	8.4	7.5
Rejected	4.2	11.0	11.4	11.0	10.5	13.8	14.0	13.8	6.9	13.7	14.1	13.7
Imbalance	1.4	0.3	0.4	0.3	6.5	0.7	0.8	0.7	6.6	0.9	0.8	0.9
All $\theta_i = 0.4$												
Control	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Rejected	11.2	24.1	24.9	24.1	25.6	31.4	31.7	31.4	21.6	34.9	35.1	34.9

Table 2: Column ordering of the procedures in Table 1.

1. 1-maxT^{bas}
2. 1-maxT^{stud}
3. 1-bal^{bas}
4. 1-bal^{stud}
5. 3-maxT^{bas}
6. 3-maxT^{stud}
7. 3-bal^{bas}
8. 3-bal^{stud}
9. FDP-maxT^{bas}
10. FDP-maxT^{stud}
11. FDP-bal^{bas}
12. FDP-bal^{stud}