The Annals of Statistics
2010, Vol. 38, No. 1, 598–633
DOI: 10.1214/09-AOS734

BALANCED CONTROL OF GENERALIZED ERROR RATES

BY JOSEPH P. ROMANO[1] AND MICHAEL WOLF[2]

Stanford University and University of Zurich

Consider the problem of testing s hypotheses simultaneously. In this paper, we derive methods which control the generalized family-wise error rate given by the probability of k or more false rejections, abbreviated k-FWER. We derive both single-step and step-down 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 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.

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 while at the same time incorporating *balance*. In particular, we construct computationally feasible methods which provide control (at least asymptotically) of some generalized notions of the *family-wise error rate*. However, the theory also applies to exact finite sample control in certain situations. At the same time, 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.

In this sense, we provide a synthesis of our previous work [31] and the works of [5, 6]. Given the rising popularity and importance of generalized error rates, due to applications where a very large of hypotheses are tested at the same time, such a synthesis appears both timely and useful. In [31], we presented computer-intensive methods to control generalized notions of the family-wise error rate, obtaining both

Received July 2008; revised July 2009.

[1] Supported by NSF Grant DMS-07-07085.

[2] Supported by the Swiss National Science Foundation (NCCR FINRISK, Module A3) and the Spanish Ministry of Science and Technology and FEDER (Grant MTM2006-05650).

AMS 2000 subject classifications. Primary 62J15; secondary 62G10.

Key words and phrases. Bootstrap, false discovery proportion, generalized family-wise error rate, multiple testing, step-down procedure.

asymptotic and finite-sample results, but these methods do not generally incorporate balance. Previously, [5, 6] constructed balanced simultaneous confidence regions for s real-valued parameters $\theta_1(P), \dots, \theta_s(P)$; tests of the s hypotheses $H_i : \theta_i(P) = 0$ can then be constructed by rejecting any H_i for which 0 is not in the confidence region. By the usual duality of tests and confidence regions, such a procedure then controls the traditional family-wise error rate. Moreover, this method can be viewed as a single-step procedure. One of the main goals of this paper is to generalize [5, 6]'s construction to other generalized error rates and at the same time to provide a step-down improvement by allowing the possibility of further rejections (even for the family-wise error rate). Of course, we want the constructions to be computationally feasible, to offer control of the given error measure, and to provide balance.

The paper is organized as follows. Section 2 provides some overview and motivation. In Section 3, we review [5]'s construction of balanced simultaneous confidence regions and then generalize this construction to accommodate control of the *generalized family-wise error rate* (k-FWER). These methods are single-step methods, in that individual test statistics are compared to their respective critical values simultaneously. In Section 4, we show that, if we apply critical values that have a monotonicity property, then the basic problem of constructing a valid step-down 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 I 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 5. In summary, step-down 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 6. Section 7 briefly discusses control of the average number of false rejections. Results for control of the *false discovery proportion* (FDP) are obtained in Section 8. A simulation study is presented in Section 9. All proofs are collected in the Appendix.

2. Overview and motivation.

2.1. *Problem at hand.* 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 $i = 1, \dots, s$. 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 *family-wise 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 (probabilistically) control false rejections less severely in hopes of better power.

2.2. *Various error rates*. First, we shall consider the k-FWER, the probability of rejecting at least k true null hypotheses, where k is some integer. For testing $H_i : P \in \omega_i$, $i = 1, \dots, s$, 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 the following.

$$k\text{-FWER}_P = P\{\text{reject at least } k \text{ hypotheses } H_i : i \in I(P)\}. \tag{2.1}$$

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

$$k\text{-FWER}_P \le \alpha \qquad \text{for all } P. \tag{2.2}$$

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

If (2.2) were required to hold only when $I(P) = \{1, \dots, s\}$, then control for such P is called weak control. We are requiring (2.2) to hold for all P, and so $I(P)$ can be a general subset of $\{1, \dots, s\}$. Control for general P is called strong control, which is the desired form of control here and throughout the literature on multiple testing; e.g., see [9] for a related discussion. The notion of weak control is not useful for practical and theoretical reasons, and will not be used in the remainder of this paper.

A related measure of error control is the average number of false rejections, also known as the *per-family error rate* (PFER). To this end, let F denote the number of true null hypotheses rejected. Control of the average number of false rejections at level k just means

$$E_P(F) \le k \qquad \text{for all } P. \tag{2.3}$$

More generally, the integer k could be replaced by some real-valued $\lambda \in (0, \infty)$. Such a measure of error control was suggested in [34]. Note that with this definition of F, one can write

$$k\text{-FWER}_P = P\{F > k - 1\}.$$

In many applications, it may not be obvious what value of k for the *number* of false rejections should be chosen. Instead, it often is more natural to focus on the *proportion* of false rejections out of all rejections. Therefore, we will also consider 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. Analogously to the number of false rejections, F, we may wish to control two different things. On the one hand, the probability that the FDP exceeds some given bound; and on the other hand, the expected value of the FDP.

As the FDP is a proportion, a bound on it should lie in the unit interval. Hence, for a specified $\gamma \in [0, 1)$, consider the probability $P\{\text{FDP} > \gamma\}$. (We now use γ as opposed to the λ of PFER control to further stress that $\gamma < 1$ necessarily.) Probabilistic control of the FDP requires that this probability be $\leq \alpha$ for all P; that is

$$P\{\text{FDP} > \gamma\} \leq \alpha \qquad \text{for all } P. \tag{2.4}$$

Thus, control of the FDP means more fully that we are controlling the tail probability that the FDP exceeds a given value in the sense of (2.4). Note that the choice $\gamma = 0$, for arbitrary α, results in control of the FWER at level α. This follows because the event $\{\text{FDP} > 0\}$ is equivalent to the event $\{F > 0\}$.

Arguably, a more consistent terminology, in the spirit of [40], would be as follows. For a specified $\gamma \in [0, 1)$, define the γ-TPFDP as the γ tail probability of the FDP; that is, $\gamma\text{-TPFDP} = P\{\text{FDP} > \gamma\}$. Then, control of the γ-TPFDP requires that

$$\gamma\text{-TPFDP} \leq \alpha \qquad \text{for all } P.$$

However, we already used the terminology of (probabilistic) control of the FDP in the previous works [19] and [31]; so we prefer to stick with it. Throughout this manuscript, the shorthand language "control of the FDP" will always refer to (2.4), exactly or sometimes just asymptotically.

Alternative terminology for (2.4) include "rate ceiling confidence thresholds" by [12] and "controlling the tail probability for the proportion of false positives among the rejected hypotheses (TPPFP)" by [40].

A related measure of error control is the expected value of the FDP, known as the *false discovery rate* (FDR). Control of the FDR requires that, for a specified $\gamma \in [0, 1)$,

$$E_P(\text{FDP}) \leq \gamma \qquad \text{for all } P.$$

An interesting feature of FDR control is that it results in *weak* control of the FWER at level $\alpha = \gamma$. That is, when all null hypotheses are true, control of the FDR for a given value $\gamma \in [0, 1)$ implies $P(F > 0) \leq \gamma$. The FDR as an error rate dates back to [1] and, up to now at least, is more popular than the (probabilistic) control of the FDP.

However, an important distinction between controlling an exceedance probability and controlling an expected value, respectively, of an underlying random variable should be pointed out. In doing so, we restrict attention to the FDP as the underlying random variable. In this case, control of an exceedance probability corresponds to (probabilistic) control of the FDP while control of the expected value corresponds to control of the FDR. (The distinction is analogous concerning F as the underlying random variable, with k-FWER and PFER as the two resulting error rates.)

When controlling an exceedance probability, one can make meaningful statements about the realized random variable, the FDP. In particular, (probabilistic) control of the FDP allows one to be $1-\alpha$ confident that the realized FDP is $\leq\gamma$. On the other hand, control of the FDR does not allow one to make any useful statements on the realized FDR; though some very crude declarations based on the Markov inequality are possible; see [19]. Put differently, even if $E_P(\text{FDP}) \leq \gamma$, the probability of the FDP exceeding γ can actually be quite large. For some related discussion and simulation studies, see [17, 29, 36] and [37]. Unfortunately, this important point is still underappreciated. Indeed, it is quite common that researchers apply methods to control the FDR but then interpret their results as if they had (probabilistically) controlled the FDP instead.

2.3. *Single-step vs. stepwise methods.* In single-step methods, individual test statistics are compared to their critical values simultaneously, and after this simultaneous "joint" comparison, the method stops. Often there is only one common critical value, but this does not need to be the case. More generally, the critical value for the ith test statistic may depend on i. As an example, consider the Bonferroni method with $T_{n,i} = -\hat{p}_{n,i}$, where $\hat{p}_{n,i}$ is an individual p-value for H_i. It is a single-step method with common critical value $-\alpha/s$, that is, H_i is rejected iff $T_{n,i} \geq -\alpha/s$ or, equivalently, iff $\hat{p}_{n,i} \leq \alpha/s$. More generally, the weighted Bonferroni method is a single-step method with the ith critical value given by $-w_i\alpha/s$, where the constants w_i reflect the "importance" of the individual hypotheses, satisfying $w_i \geq 0$ and $\sum w_i = 1$.

Often, single-step methods can be improved in terms of power via stepwise methods, while nevertheless maintaining control of the desired error rate. Stepdown methods start with a single-step method but then continue by possibly rejecting further hypotheses in subsequent steps. This is achieved by decreasing the critical values for the remaining hypotheses depending on the hypotheses already rejected in previous steps. As soon as no further hypotheses are rejected anymore, the method stops. As an example, consider the Holm method of [15], which is a step-down improvement over Bonferroni and can be formulated as follows. Let R_j denote the total number of rejected hypotheses in the previous $j-1$ steps, where $j > 1$. Then the (common) critical value in the jth step becomes $-\alpha/(s-R_j)$. (This formulation is different from the standard description but is equivalent.)

Such stepwise methods which improve upon single-step methods by possible rejecting “less significant” hypotheses in subsequent steps are called step-down methods. Intuitively, this is because such methods start with the most significant hypotheses, having the largest test statistics, and then “step down” to further examine the remaining hypotheses having smaller test statistics. All resampling methods that we have proposed in previous work—such as [30, 31] and [28]—as well as the ones developed in this paper are step-down methods.

In contrast, there also exist stepup methods that start with the least significant hypotheses, having the smallest test statistics, and then “step up” to further examine the remaining hypotheses having larger test statistics. The crucial difference is that, at any given step, the question is whether to reject all remaining hypotheses or not. And so the hypotheses “sorted out” in previous steps correspond to not rejected hypotheses rather than rejected hypotheses, as in step-down methods. A prominent example is the FDR controlling method of [1]. Like Bonferroni it also uses $T_{n,i} = -\hat{p}_{n,i}$. The first step checks whether all $T_{n,i} \geq -\alpha$ or, equivalently, whether $\min_i p_{n,i} \leq \alpha$. If yes, all hypotheses are rejected. Otherwise, the hypothesis corresponding to the largest test statistic is discarded. In the second step, all remaining test statistics are compared to $-\alpha \cdot (s-1)/s$ and so on.

2.4. *Previous work and introduction of balance.* Recently, there have been many new proposals which control generalized error rates that are less stringent than the FWER. A notable such technique is the FDR controlling method of [1]. Additional methods that control the FDR are given in [2, 3, 32] and [35], among others. Asymptotic procedures that control the FDP (and the FDR) in the framework of a random effects mixture model are studied in [12]. These ideas are extended in [21], where in the context of random fields, the number of null hypotheses is uncountable. Methods that control both the k-FWER and the FDP are given in [17]; 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 [16, 19, 26] and [27]. Alternative methods for control of the k-FWER and FDP are given in [39] and [40]. Control of the false discovery rate via resampling is considered in [8, 28] and [42].

In this paper, building upon our previous works [30, 31] where balanced was not emphasized, we employ resampling and subsampling techniques to achieve our goals and do not require the use of the subset pivotality condition of [41]. 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.

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 contribute equally to error control, as argued by [5, 24] and [38].

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 [5, 6]. 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, as such a scheme would be computationally very expensive. Instead, 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).

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 $k \le |K|$, the $k\text{-max}(y_i : i \in K)$ 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 $y_{(1)} \le \cdots \le y_{(|K|)}$, then $k\text{-max}(y_i : i \in K) = y_{(|K|-k+1)}$.

3. Balanced (generalized) simultaneous confidence regions. Throughout this section, the integer k is fixed. We first review and then generalize Beran's [5] 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. 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, but we will not pursue this further.

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

$$\{\theta_i : \tau_n|\hat{\theta}_{n,i} - \theta_i| \le c_{n,i}(\gamma, P)\} \tag{3.1}$$

has coverage probability γ. Continuity of $H_{n,i}(P)$ is only assumed here for convenience and is certainly not required in our asymptotic results. Of course, the interval (3.1) 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 $K \subseteq \{1, \dots, s\}$ denote an arbitrary subset of $\{1, \dots, s\}$. We would like to make joint inferences for the parameters $\theta_i(P)$ simultaneously for $i \in K$. The case where $K = \{1, \dots, s\}$ is especially important, but the general case is required for our step-down multiple testing method presented later. Then the probability of the event

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

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 $f_{n,K}(\gamma, P) \ge 1 - \alpha$. 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 [5] 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 [7]. 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 [18]. The procedure is to replace P by $\hat{Q}_n$ in (3.1). Specifically, Beran proposes the set of intervals

$$\{\theta_i : \tau_n|\hat{\theta}_{n,i} - \theta_i| \le c_{n,i}(\gamma, \hat{Q}_n)\} = \{\theta_i : \tau_n|\hat{\theta}_{n,i} - \theta_i| \le H_{n,i}^{-1}(\gamma, \hat{Q}_n)\}, \tag{3.2}$$

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 family-wise 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

$$\{\tau_n|\hat{\theta}_{n,i} - \theta_i(P)| \le c_{n,i}(\gamma, P) \text{ for all but at most } (k-1) \text{ of the } i \in K\}. \tag{3.3}$$

To this end, the previous event (3.3) can be rewritten as

$$\{H_{n,i}(\tau_n|\hat{\theta}_{n,i} - \theta_i(P)|, P) \le \gamma \text{ for all but at most } (k-1) \text{ of the } i \in K\} \tag{3.4}$$

or

$$\{k\text{-max}(H_{n,i}(\tau_n|\hat{\theta}_{n,i} - \theta_i(P)|, P), i \in K) \le \gamma\}. \tag{3.5}$$

Let $f_{n,K}(\gamma, k, P)$ denote the probability under P of the event in (3.3)–(3.5), 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

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

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

$$\gamma_{n,K}(\alpha, k, \hat{Q}_n) = L_{n,K}^{-1}(1 - \alpha, k, \hat{Q}_n). \tag{3.6}$$

Combining (3.2) and (3.6) yields the joint *generalized* confidence region

$$\{(\theta_i, i \in K) : \tau_n|\hat{\theta}_{n,i} - \theta_i| \le H_{n,i}^{-1}(L_{n,K}^{-1}(1 - \alpha, k, \hat{Q}_n), \hat{Q}_n)\}. \tag{3.7}$$

Under fairly weak conditions, this simultaneous generalized confidence region covers all $\theta_i(P)$ 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 depend on i asymptotically.

REMARK 3.1 [Calculating (3.7)]. In order to calculate (3.7), 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 $b = 1, \dots, B$, draw a sample of size n from $\hat{Q}_n$ and let $\hat{\theta}_{n,i}^*(b)$ be the estimate of $\theta_i(P)$. 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 $\le x$; this leads to a corresponding approximation to the quantile function $H_{n,i}^{-1}(\cdot, \hat{Q}_n)$. 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 $\le x$; its largest $1 - \alpha$ quantile is a simulation-based approximation of $L_{n,K}^{-1}(1 - \alpha, k, \hat{Q}_n)$.

As [5] 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 its respective parameter 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 $\overset{L}{\to}$ and $\overset{P}{\to}$ will denote convergence in law (or in distribution) and convergence in probability, respectively. For $K \subseteq \{1, \dots, s\}$, 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 $K = \{i\}$. Typically, the joint distribution of the estimators tends to an asymptotic limit, which is stated formally in the following assumption.

ASSUMPTION B1. $J_{n,\{1,\dots,s\}}(P) \overset{L}{\to} J_{\{1,\dots,s\}}(P)$.

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 $K \subseteq \{1, \dots, s\}$, $L_{n,K}(k, P)$ has a continuous limiting distribution $L_K(k, P)$; see Lemma A.1 in the Appendix.

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 $J_{\{1,\dots,s\}}(P)$ 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 fact that it implies that $L_K(k, P)$ has a continuous and strictly increasing c.d.f. on its interval of support; see Corollary A.1 in the Appendix.

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 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 $i \in \{1, \dots, s\}$. Specifically, consider the following assumption.

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

$$\rho(J_{n,\{1,\dots,s\}}(P), J_{n,\{1,\dots,s\}}(\hat{Q}_n)) \overset{P}{\to} 0.$$

Assumption 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 [13, 33] and Chapter 1 of [22]. 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 (3.7). The result here requires fewer assumptions than [5]. In particular, we can dispense with his Assumption 4 in view of our above Lemma A.2. Moreover, our result will apply toward control of the k-FWER for general k (while the results in [5] only apply to $k = 1$).

THEOREM 3.1. *Suppose data is generated from P satisfying Assumptions B1–B3. Let $\hat{Q}_n$ be an estimator of P satisfying Assumption B4. Fix $K \subseteq \{1, \dots, s\}$ and a positive integer k. Consider the joint confidence region given by (3.7), with the marginal interval $\hat{C}_{n,i}$ for $\theta_i(P)$ with $i \in K$ expressed as*

$$\hat{C}_{n,i} \equiv \hat{\theta}_{n,i} \pm \tau_n^{-1} H_{n,i}^{-1}\big(L_{n,K}^{-1}(1-\alpha, k, \hat{Q}_n), \hat{Q}_n\big). \tag{3.8}$$

(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*

$$\lim_{n\to\infty} P\{\theta_i(P) \in \hat{C}_{n,i}\} = \gamma \qquad \textit{independent of } i, \tag{3.9}$$

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

A value of 0 for $\theta_i(P)$ falls outside the region (3.8) if and only if

$$|\tau_n \hat{\theta}_{n,i}| > H_{n,i}^{-1}\big(L_{n,K}^{-1}(1-\alpha, k, \hat{Q}_n), \hat{Q}_n\big). \tag{3.10}$$

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

COROLLARY 3.1. *Assume the conditions of Theorem 3.1. For testing the multiple hypotheses $H_i : \theta_i(P) = 0$, consider the procedure which rejects H_i if (3.10) holds with $K = \{1, \dots, s\}$. Then:*

(i) $$\lim_{n\to\infty} k\text{-FWER}_P \le \alpha.$$

(ii) *Moreover,*

$$\lim_{n\to\infty} P\{\textit{reject } H_i\} = 1 - L_K^{-1}(1-\alpha, k, P) \tag{3.11}$$

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

$$\hat{p}_{n,i} \equiv 1 - H_{n,i}(|\tau_n \hat{\theta}_{n,i}|, \hat{Q}_n). \tag{3.12}$$

If balance were not imposed, as in [31], 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. This rules out potential "inconsistencies" of the unbalanced method of [31] where it can happen that, say, H_1 is rejected while H_2 is not, even though $\hat{p}_{n,2} < \hat{p}_{n,1}$. For example, such a situation can arise when the standard deviation of $\hat{\theta}_{n,1}$ is larger than the standard deviation of $\hat{\theta}_{n,2}$.

As previously mentioned, the choice of $\hat{Q}_n$ should reflect the underlying P. We will later also consider a subsampling approach in Section 5.2. 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 [30]. The finite sample results also extend to step-down methods considered later, using ideas developed in Section 4.

REMARK 3.2 (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 $\tau_n |\hat{\theta}_{n,i} - \theta_i(P)|/\hat{\sigma}_{n,i}$. As in [5], 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

$$k\text{-max}\big(H_{n,i}\big(R_{n,i}\big(X^{(n)}, \theta_i(P)\big)\big), i \in K\big).$$

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

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

in the generalization of (3.7). 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 3.3 (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.

REMARK 3.4 (Relationship to studentization). As argued by [5], 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 (3.12), 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 (3.8), 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-hand 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 [5] 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.

It is also possible to obtain marginal p-values adjusted for multiplicity. The ith adjusted p-value is the smallest value of the overall significance level α for which H_i can be rejected. Generally, it could be found indirectly by "trial and error," which would be rather cumbersome. It can, however, also be found directly by changing the inequality in (3.10) to an equality and solving for α. This results in

$$\begin{aligned} \hat{p}_{n,i}^{\text{adjust}} &= 1 - L_{n,\{1,\dots,s\}}(H_{n,i}(|\tau_n \hat{\theta}_{n,i}|, k, \hat{Q}_n)) \\ &= 1 - L_{n,\{1,\dots,s\}}(1-\hat{p}_{n,i}, k, \hat{Q}_n), \end{aligned} \tag{3.13}$$

where $\hat{p}_{n,i}$ is the unadjusted marginal p-value given in (3.12).

4. 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 3, 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 step-down 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

$H_1, \dots, H_s$. Any hypothesis H_i is rejected if $T_{n,i} > c_{n,\{1,\dots,s\},i}(1-\alpha,k)$. 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 step-down methods developed in [31] use a common critical value at each stage of the algorithm, 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 3.1 is to let $c_{n,K,i}(1-\alpha,k)$ to be the right-hand side of (3.10), but other choices are possible as well.

ALGORITHM 4.1 (Generic stepdown method for control of the k-FWER).

1. Let $A_1 = \{1,\dots,s\}$. If $T_{n,i} \le \hat{c}_{n,A_1,i}(1-\alpha,k)$ 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.

⋮

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.

⋮

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 step-down method will control the k-FWER.

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 $K \supseteq I$,

$$\hat{c}_{n,K,i}(1-\alpha, k) \geq \hat{c}_{n,I,i}(1-\alpha, k). \tag{4.1}$$

Under the monotonicity assumption (4.1), we will show that k-FWER control of a step-down procedure is reduced to that of a single-step method. Thus, the construction of a step-down 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 4.1. *Consider Algorithm* 4.1 *with critical values* $\hat{c}_{n,K,i}(1-\alpha, k)$ *satisfying* (4.1).

(i) *Then*

$$k\text{-FWER}_P \leq P\{T_{n,i} > \hat{c}_{n,I(P),i} \textit{ for all but at most } k-1 \textit{ of } i \in I(P)\}. \tag{4.2}$$

(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-hand side of* (4.2) *is* $\leq \alpha$ (*in finite samples or asymptotically*), *then* k-FWER$_P \leq \alpha$ (*in finite samples or asymptotically*).

The monotonicity assumption (4.1) cannot be removed, as shown in Example 2.1 of [30] in the case $k = 1$; an analogous construction works for general k. The general resampling constructions we describe later will inherently satisfy (4.1). 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 [30] and for general k in [31], but without the requirement of balance. However, we see how the argument generalizes given Theorem 4.1. We also observe that Theorem 4.1 applies to certain semiparametric problems where permutation and randomization tests apply. Such a setting is discussed in [17], 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 4.2 (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 Section 5.1, 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, $\hat{p}_{n,(1)} \le \cdots \le \hat{p}_{n,(s)}$. Denote by $\{r_1, \dots, r_s\}$ the permutation of $\{1, \dots, s\}$ which yields this ordering; that is, $\hat{p}_{n,(1)} = \hat{p}_{n,r_1}, \dots, \hat{p}_{n,(s)} = \hat{p}_{n,r_s}$. Accordingly, let $H_{(1)} = H_{r_1}, \dots, H_{(s)} = H_{r_s}$. Then, $H_{(1)}$ is the most significant and $H_{(s)}$ is the least significant hypothesis. The algorithm now is analogous to Algorithm 4.1. The only difference is that in any step $j > 1$, the critical value $\hat{d}_{n,A_j,i}(1-\alpha, k)$ is replaced by the critical value

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

5. Asymptotic results on k-FWER control.

The main goal of this section is to show how Theorem 4.1 can be used to construct step-down 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 4.1, 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. This is natural because, essentially, our multiple testing methods are based on inverting (generalized) balanced simultaneous confidence regions, extending the well-known duality between confidence intervals and hypotheses tests for univariate parameters to the multivariate case. Such an inversion can be viewed as a two-stage procedure. In the first stage, one computes a (generalized) simultaneous confidence region. In the second stage, one carries out the individual tests concerning the H_i by checking whether 0 is contained in the implied confidence interval for $\theta_i(P)$. Therefore, the null constraints are completely irrelevant in the first stage and come only into play in the second stage. Since the bootstrap is used in the first stage only, one can, therefore, simply resample "from the data."

Alternatively, hypothesis test constructions that do obey the constraints imposed by the null hypothesis, as discussed in [4] and [25], 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 [14], and it is enforced throughout [41]. However, this direct approach of obtaining critical values by resampling from a null distribution requires the subset pivotality condition of Section 2.2 of [41]. In particular, this condition specifies that the joint distribution of any subvector of test statistics is not affected by the truth or falsehood of the hypotheses corresponding to test statistics not included in this subvector. This allows one to always resample from a "global" null distribution where all individual null hypotheses are true. While this condition holds in many applications of interest, under weak regularity conditions, such as testing the elements of a multivariate mean vector or testing the elements of a multivariate regression coefficients vector, there also exist practically relevant counterexamples, such as testing the elements of a correlation matrix; see Example 4.1 of [30].

Our indirect approach based on resampling from the data avoids the subset pivotality condition and is thereby more generally valid than "direct" approaches as in [41]. Related schemes have been suggested previously by [23] and [10]. They derive a null distribution for the test statistics by resampling from the data combined with a transformation via recentering using the estimated parameters from the observed data rather than the null parameters (and potentially also rescaling using the standard errors from the observed data). In this way, they are also able to dispense with the subset pivotality condition.

We shall consider two concrete applications of Theorem 4.1, the first based on the bootstrap and the second based on subsampling.

5.1. *A bootstrap construction.*

We now apply Theorem 4.1 to develop an asymptotically valid approach based on the bootstrap. As in Section 3, we specialize to the case where hypothesis H_i is specified by $\{P : \theta_i(P) = 0\}$ for some real-valued parameter $\theta_i(P)$. Implicitly, the alternatives are two-sided, but the one-sided case can be similarly handled. Recalling the notation of Section 3, suppose $\hat{\theta}_{n,i}$ is an estimate of $\theta_i(P)$. 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 3.1 suggests using Algorithm 4.1 with critical values

$$\hat{c}_{n,K,i}(1-\alpha, k) = H_{n,i}^{-1}\big(L_{n,K}^{-1}(1-\alpha, k, \hat{Q}_n), \hat{Q}_n\big). \tag{5.1}$$

Note that, regardless of asymptotic behavior, the monotonicity assumption (4.1) is always satisfied for the choice (5.1). Indeed, whenever $I \subseteq K$, we must show

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

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

$$L_{n,I}^{-1}(1-\alpha, k, \hat{Q}_n) \le L_{n,K}^{-1}(1-\alpha, k, \hat{Q}_n). \tag{5.2}$$

But, for any Q and $I \subseteq K$, the left-hand side of (5.2) is the $1-\alpha$ quantile under Q of the k-max of $|I|$ variables, while the right-hand side of (5.2) 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 4.1 immediately reduces the problem of step-down control to that of single-step control, which was already obtained in Corollary 3.1. The following result is an improvement over Corollary 3.1 in that more rejections are possible, while maintaining asymptotic control of the k-FWER.

COROLLARY 5.1. *Under the setup and conditions of Corollary* 3.1, *consider Algorithm* 4.1 *with critical values given by* (5.1).

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

(ii) $\lim_{n\to\infty} P\{\textit{reject } H_i\}$ *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 step-down 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* III *error* [*i.e., of declaring* $\theta_i(P)$ *positive when it is negative or declaring it negative when it is positive*] *tends to* 0.

Compare this balanced bootstrap method to the unbalanced bootstrap method in Section 3 of [31]. Results (i), (iii) and (iv) also hold for our earlier method but, crucially, result (ii) generally does not hold. The key difference is that our earlier method can be considered a *common cut-off* method, since the critical value (or the cut-off) does not depend on i, that is, $\hat{c}_{n,K,i}(1-\alpha, k) = \hat{c}_{n,K}(1-\alpha, k)$ for all i; and this common critical value is obtained as the $1 - \alpha$ quantile of a suitable *joint* distribution. In contrast, the critical values of our balanced bootstrap method depend on i. Instead what is common now is the quantile $L_{n,K}^{-1}(1 - \alpha, k, \hat{Q}_n)$ used for all *marginal* distributions $H_{n,i}(\cdot, \hat{Q}_n)$. Therefore, the new method can be considered a *common quantile* method.

So far, the bootstrap construction has been based on Algorithm 4.1. But, asymptotic control of the k-FWER is also achieved by the computationally less expensive streamlined Algorithm 4.2.

COROLLARY 5.2. *The statements of Corollary* 5.1 *continue to hold if Algorithm* 4.1 *is replaced by Algorithm* 4.2.

REMARK 5.1 (Operative method). While the streamlined Algorithm 4.2 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 4.1. On the other hand, its computational burden can be too high. As a feasible compromise, we suggest an operative method that retains some of the desirable properties of the generic algorithm. 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 4.1, a critical value is then computed as follows:

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

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. The philosophy of this operative method is to be as close as possible to the generic Algorithm 4.1, given the limitation to the computational burden expressed by $N_{\max}$. Actually, the streamlined algorithm is a special case of the operative method when $N_{\max} = 1$ is chosen, resulting in $M = k - 1$.

REMARK 5.2 (Asymptotic sharpness). The $\limsup_{n\to\infty}$ in Corollary 5.1(i) can actually be 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 joint distribution $H_i(|Y_i|, P)$, as defined through (A.1) and (A.2), is exchangeable, then equality holds. Nevertheless, the step-down 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 the k-FWER of the single-step procedure is given by (A.11) with $K = \{1, \dots, s\}$, while the asymptotic expression for the step-down 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 5.1(i) is always an equality for the streamlined method of Algorithm 4.2.

5.2. *A general subsampling construction.*

In this subsection, we sketch an alternative construction of critical values in our step-down procedure by using subsampling. As in the bootstrap approach of Section 5.1, 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 $X_1, \dots, X_n$ 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 $Y_1, \dots, Y_{N_n}$ be equal to the $N_n := \binom{n}{b}$ subsets of $\{X_1, \dots, X_n\}$, ordered in any fashion. Let $\hat{\theta}_{b,i}^{(a)}$ be equal to the statistic $\hat{\theta}_{n,i}$ evaluated at the data set Y_a, for $a = 1, \dots, N_n$. The subsampling estimator of $H_{n,i}(\cdot, P)$

is then given by

$$\hat{H}_{n,i}(x) = \frac{1}{N_n} \sum_a I\{\tau_b |\hat{\theta}_{b,i}^{(a)} - \hat{\theta}_{n,i}| \le x\}. \tag{5.3}$$

We also define

$$\hat{L}_{n,K}(x,k) = \frac{1}{N_n} \sum_a I\{k\text{-max}(\hat{H}_{n,i}(\tau_b |\hat{\theta}_{b,i}^{(a)} - \hat{\theta}_{n,i}|)) \le x\}. \tag{5.4}$$

If we replace the bootstrap estimators by these subsampling estimators, we can prove a result analogous to Theorem 3.1, while removing Assumption B4.

THEOREM 5.1. *Suppose data is generated from P satisfying Assumptions B1–B3. Fix $K \subseteq \{1,\dots,s\}$ 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*

$$\tilde{C}_{n,i} \equiv \hat{\theta}_{n,i} \pm \tau_n^{-1} \hat{H}_{n,i}^{-1}(\hat{L}_{n,K}^{-1}(1-\alpha,k)). \tag{5.5}$$

(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*

$$\lim_{n\to\infty} P\{\theta_i(P) \in \tilde{C}_{n,i}\} = \gamma \qquad \textit{independent of } i, \tag{5.6}$$

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 3.1, except that the uniform convergence of the subsampling estimators (in probability) is proved by the now standard arguments for subsampling; see Chapter 2 of [22]. Thus, the result also generalizes quite easily; for example, in a stationary time series model, one only considers subsamples of consecutive observations; see Chapter 3 of [22].

REMARK 5.3 (Effects of centering). 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 [22], 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 [28].

In the case $k = 1$, not centering the subsampled values can be advantageous in that it results in more powerful procedure. For example, suppose $(X_1, Y_1), \dots, (X_n, Y_n)$ 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. $\sim N(0, 1)$. 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_i's will asymptotically dominate those based on the X_i's. A smaller critical value then implies greater power.

We can also provide a step-down improvement by applying the step-down Algorithm 4.1 with the critical values

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

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

$$\hat{c}_{n,K,i}(1 - \alpha, k) \geq \hat{c}_{n,I,i}(1 - \alpha, k). \tag{5.7}$$

This simple observation together with Theorems 4.1 and 5.1 immediately yields an asymptotic improvement. The details are left to the reader.

6. 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 $H_1, \dots, H_s$ 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

$$P\{\hat{p}_{n,i} \le u\} \le u \qquad \text{for all } u,\ P \in \omega_i. \tag{6.1}$$

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 $\hat{p}_{n,i} \le w_i \alpha$. 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 construct such weighted procedures which control the k-FWER, and at the same time provide a step-down improvement.

THEOREM 6.1. *Consider the problem of testing $H_1, \dots, H_s$ with marginal p-values satisfying* (6.1). *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 $\hat{p}_{n,i} \le w_i k\alpha$ controls the k-FWER*; *that is*

$$k\text{-FWER}_P \le \alpha. \tag{6.2}$$

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 step-down procedure using Algorithm* 4.1 *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 (6.2).

The computational application of Algorithm 4.1 is straightforward. The algorithm can be translated as follows. First, reject any H_i whose corresponding p-value $\hat{p}_{n,i}$ satisfies $\hat{p}_{n,i} \le w_i k\alpha$; 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 regardless.) 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.

7. Control of average number of false rejections.

In this section, we briefly consider control of the average number of false rejections, also known as the *per-family error rate* (PFER); see (2.3). Suppose p-values $\hat{p}_{n,i}$ are available for testing H_i, so that (6.1) holds. As is well known, the procedure which rejects H_i if $\hat{p}_{n,i} \le \lambda/s$ satisfies (2.3). More generally, and analogous to Theorem 6.1, the following is true.

THEOREM 7.1. *Consider the problem of testing $H_1, \dots, H_s$ with marginal p-values satisfying* (6.1). *Assume w_i are known weights with $\sum_{i=1}^{s} w_i = 1$. Then the single-step procedure which rejects H_i if $\hat{p}_{n,i} \le w_i \lambda$ controls the average number of false rejections*; *that is*, (2.3) *holds. Moreover, if $\hat{p}_{n,i}$ has a uniform* (0, 1) *distribution whenever H_i is true and if $w_i \lambda$ 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 step-down improvement is not possible. To see why, suppose $w_i = 1/s$, all H_i are true, and $\hat{p}_{n,i}$ has a uniform (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 (2.3). (Note it is asymptotically possible to provide a step-down 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 (3.12). Of course, by linearity of expectation, no further modification of the procedure is needed to take into account the dependence between the test statistics.

8. Asymptotic results on FDP control. In many 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:

$$\mathrm{FDP} = \begin{cases} \dfrac{F}{R}, & \text{if } R > 0, \\ 0, & \text{if } R = 0. \end{cases}$$

A multiple testing procedure is said to (probabilistically) control the FDP at level α if, for the given sample size n, $P\{\mathrm{FDP} > \gamma\} \le \alpha$, for all P. A multiple testing procedure is said to asymptotically control the FDP at level α, if $\limsup_n P\{\mathrm{FDP} > \gamma\} \le \alpha$, for all P. Our focus will be on procedures that provide asymptotic control.

The approach we propose is analogous to the one already presented in [31]. It is built upon an underlying procedure that (asymptotically) controls the k-FWER for any fixed $k \ge 1$. We then sequentially apply this k-FWER procedure for $k = 1, 2, \dots$ until a stopping rule indicates termination. In the end, we reject all hypotheses that were rejected in the last round before stopping. This leads to the following algorithm; see [31] for some corresponding motivation and intuition.

ALGORITHM 8.1 (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.

This algorithm is similar to the proposal of [17] for FDP control which is, however, restricted to a multivariate permutation model. The proposal of [17] 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 [31], asymptotic control of Algorithm 8.1 is established when using an unbalanced bootstrap or subsampling approach for the underlying k-FWER procedure, with simulations showing good finite sample control. The following theorem establishes the corresponding result if one uses instead 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 analogously. The proofs are very similar to the unbalanced cases established in [31].

THEOREM 8.1. *Consider the setup of Corollary* 5.1. *Fix P satisfying Assumptions* B1–B3. *Let $\hat{Q}_n$ be an estimate of P satisfying Assumption* B4. *Employ the step-down procedure of Algorithm* 4.1 *with $\hat{c}_{n,K,i}(1-\alpha, k)$ as the underlying k-FWER procedure. Then the following statements concerning Algorithm* 8.1 *are true*:

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

(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.*

9. Simulation study. This section presents a small simulation study in the context of testing population means. We generate random vectors $X_1, \dots, X_n$ from an s-dimensional multivariate normal distribution with mean vector $\theta(P) = (\theta_1(P), \dots, \theta_s(P))$, 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

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

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 $\rho = 0.0$ and 0.5.

The goal is to compare the balanced bootstrap procedures of this paper with the stepwise bootstrap procedures of [31] based on the maximum test statistic. For the latter procedures, the individual test statistics $T_{n,i}$ are either basic (i.e., nonstudentized) or studentized, that is,

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

The abbreviations for the included procedures are as follows.

- (k-max T$^{\text{bas}}$). The bootstrap k-FWER procedure of [31] with $T_{n,i}^{\text{bas}}$.
- (k-max T$^{\text{stud}}$). The bootstrap k-FWER procedure of [31] with $T_{n,i}^{\text{stud}}$.
- (k-bal$^{\text{bas}}$). The balanced bootstrap k-FWER procedure of Section 5.1 with $\tau_n|\hat{\theta}_{n,i}|$.
- (k-bal$^{\text{stud}}$). A balanced bootstrap k-FWER procedure analogous to Section 5.1 but with studentized roots $\tau_n|\hat{\theta}_{n,i}|/\hat{\sigma}_{n,i}$; see Remarks 3.2 and 3.4.
- (FDP-max T$^{\text{bas}}$). The bootstrap FDP procedure of [31] with $T_{n,i}^{\text{bas}}$.
- (FDP-max T$^{\text{stud}}$). The bootstrap FDP procedure of [31] with $T_{n,i}^{\text{stud}}$.
- (FDP-bal$^{\text{bas}}$). The balanced bootstrap FDP procedure of Section 8 with $\tau_n|\hat{\theta}_{n,i}|$.
- (FDP-bal$^{\text{stud}}$). A balanced bootstrap FDP procedure analogous to Section 8 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 Section 19.3 of [11]. The computational burden we can handle corresponds to $B = 10{,}000$. For that reason, we pick 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 5.1.

The values of k for k-FWER control we consider are $k = 1$ and 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

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

(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 5000 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-max $\mathrm{T}^{\mathrm{bas}}$ results in asymptotically unbalanced inference. Due to studentization, k-max $\mathrm{T}^{\mathrm{stud}}$ is invariant to the $\sigma_{i,i}$ and yields asymptotically balanced inference. This is reflected in the empirical imbalances which are always larger for k-max $\mathrm{T}^{\mathrm{bas}}$, and sometimes much larger.
- If balance is applied to the basic method, resulting in k-bal$^{\mathrm{bas}}$, then the empirical imbalances become comparable to k-max $\mathrm{T}^{\mathrm{stud}}$. On the other hand, if balance is applied to the studentized method, resulting in k-bal$^{\mathrm{stud}}$, no meaningful further improvement over k-max $\mathrm{T}^{\mathrm{stud}}$ is achieved.
- Both k-max $\mathrm{T}^{\mathrm{bas}}$ and k-max $\mathrm{T}^{\mathrm{stud}}$ achieve satisfactory control of the k-FWER. However, k-max $\mathrm{T}^{\mathrm{bas}}$ is always less powerful compared to k-max $\mathrm{T}^{\mathrm{stud}}$.
- k-bal$^{\mathrm{bas}}$ can be anticonservative, especially when all null hypotheses are true; see top "Control" row, third columns in both panels ($\rho = 0$ and $\rho = 0.5$) of the Table 1. However, it should be pointed out that the worst results happen under very stringent type I error control conditions: all null hypotheses true, mutual independence and FWER control (which in this case is equivalent to FDP control), based on nonstudentized test statistics with unequal population variances. The results appear satisfactory in other settings of particular interest, that is, when significant fractions of the null hypotheses are expected to be false. Furthermore, consistent with our asymptotic theory, the results improve when the sample size increases; see Remark 9.1 below.
- k-bal$^{\mathrm{bas}}$ is somewhat more powerful compared to k-max $\mathrm{T}^{\mathrm{stud}}$. But this not surprising given the previous observation of anticonservativeness. On the other hand, k-bal$^{\mathrm{stud}}$ performs very similarly compared to k-max $\mathrm{T}^{\mathrm{stud}}$ 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 9.1. Ceteris paribus, the finite-sample control of k-bal$^{\mathrm{bas}}$ 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). So the improvement shows up already under very stringent type I error control conditions (FWER control, all null hypotheses true, and mutual independence).

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* 5000 *per scenario and the number of bootstrap resamples is* $B = 10{,}000$. *Both the empirical error rates and imbalances are expressed in percentages. There are three sections corresponding to control of FWER*, 3-*FWER and FDP. In each section*, *the order of the methods*, *from left to right*, *is given by* $\max \mathrm{T}^{\mathrm{bas}}$, $\max \mathrm{T}^{\mathrm{stud}}$, $\mathrm{bal}^{\mathrm{bas}}$ *and* $\mathrm{bal}^{\mathrm{stud}}$

	Common correlation: $\rho = 0$											
	FWER control				**3-FWER control**				**FDP control**			
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	38.5	38.5
Common correlation: $\rho = 0.5$												
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

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 $\theta_i(P) = 0$, and common correlation $\rho = 0$, increasing the number of bootstrap resamples from $B = 10{,}000$ to $B = 50{,}000$ made virtually no difference.

REMARK 9.2. Further simulation results and comparisons are reported in [31], but the methods there were not necessarily balanced. In particular, the methods of [31] are compared to those of [19, 39] and [40].

10. Concluding remarks. We have shown how computationally feasible step-down methods can be constructed to control generalized error rates in multiple testing, with special emphasis on procedures which are appropriately balanced. This emphasis is certainly of practical relevance. Multiple testing methods which are not balanced can lead to "contradictions" when applied to sets of data. For example, it can happen that a certain hypothesis gets rejected by the multiple testing method while another one does not, even though the first hypothesis is associated with a larger unadjusted marginal p-value than the second hypothesis. Balanced procedures automatically rule out this possibility.

Various measures of error control have been considered, with emphasis on control of the k-FWER, average number of false rejections, and 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). All of these generalized error rates relax the traditional FWER and in return lead to improved ability to reject false null hypotheses. They are of special interest and importance when the number of hypotheses under test is large, which happens more and more frequently. Moreover, improved power is gained not only by relaxing the given measure of error control, but also by using resampling. Indeed, resampling methods implicitly account for the dependence structure between the test statistics, leading to improved power compared to methods based on the individual p-values alone. To the best of our knowledge, there have been no previous proposals combining resampling and the imposition of balance to control generalized error rates.

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 being tested is large, while at the same time satisfying the constraints of balance and error control.

Future work will examine the actual order of errors in our asymptotic approximations, both pointwise and uniformly with respect to the underlying probability mechanism P. Moreover, an asymptotic framework in which the number of tests gets large with the sample size will be studied as well. Finally, we would like to develop weighted methods analogous to that studied in Theorem 6.1 which also employ resampling to account for the dependence between tests.

APPENDIX: PROOFS AND AUXILIARY RESULTS

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

$$k\text{-max}\big(H_i(|Y_i|, P), i \in K\big), \tag{A.1}$$

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

$$H_i(x, P) = J_i(x, P) - J_i(-x, P). \tag{A.2}$$

PROOF. Fix P and let

$$Y_{n,i} = \tau_n[\hat{\theta}_{n,i} - \theta_i(P)]. \tag{A.3}$$

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

$$k\text{-max}\big(H_{n,i}(|Y^*_{n,i}|, P), i \in K\big) \tag{A.4}$$

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 (A.2). By Pòlya'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 (A.4) and

$$k\text{-max}\big(H_i(|Y^*_{n,i}|, P), i \in K\big) \tag{A.5}$$

tends to 0. But, by continuity of the $H_i(\cdot, P)$ and the k-max function, we have that (A.5) 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

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

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

LEMMA A.2. *Let $X = (X_1, \dots, X_s)$ 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 $Y \equiv k$-max$(h_i(X_1), \dots, h_s(X_s))$ has a continuous and strictly increasing c.d.f. on its interval of support.*

PROOF. 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 $x = (x_1, \dots, x_s) \in \operatorname{supp}(F)$ such that $k\text{-max}(h_1(x_1), \dots, h_s(x_s)) \leq a$ and $x' \in \operatorname{supp}(F)$ such that $k\text{-max}(h_1(x_1'), \dots, h_s(x_s')) \geq b$. Consider the set

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

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

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

which yields the desired contradiction. □

Even in the case in which $s = k = 1$, both assumptions in Lemma A.2 are necessary to conclude that the distribution of Y is continuous and strictly increasing. Therefore, the assumptions used in Lemma A.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 $|X_i| \sim H_i]$. The last example is most pertinent to this paper. Also, note that the lemma continues to hold if the k-max function is replaced by any continuous function which returns one of its arguments.

COROLLARY A.1. *Assume Assumptions* B1–B3. *Then $L_K(k, P)$ has a continuous and strictly increasing c.d.f. on its interval of support.*

PROOF. For ease of notation, the proof is presented in the case $K = \{1, \dots, s\}$ (with no loss of generality). Recall the limiting distribution of $L_K(k, P)$ can be represented by the distribution of (A.1). The assumptions of Lemma A.2 are satisfied. Indeed, suppose $(X_1, \dots, X_s)$ has distribution $J_{\{1,\dots,s\}}(P)$. 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 THEOREM 3.1. By Lemma A.1,

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

Moreover, by Corollary A.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 [20] that

$$L_{n,K}^{-1}(1-\alpha, k, P) \to \gamma_K(1-\alpha, k, P). \tag{A.6}$$

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

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

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

$$L_{n,K}^{-1}(1-\alpha, k, \hat{Q}_n) \overset{P}{\to} \gamma_K(1-\alpha, k, P). \tag{A.7}$$

Then

$$\begin{aligned} &P\{\theta_i \in \hat{C}_{n,i} \text{ except for at most } k-1 \text{ of the } i \in K\} \\ &\quad = P\bigl\{k\text{-max}\bigl(H_{n,i}\bigl(\tau_n|\hat{\theta}_{n,i} - \theta_i(P)|, \hat{Q}_n\bigr), i \in K\bigr) \\ &\qquad\qquad \le L_{n,K}^{-1}(1-\alpha, k, \hat{Q}_n)\bigr\}. \end{aligned} \tag{A.8}$$

But, by Assumption B2, Pòlya's theorem, and a subsequence argument,

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

where $H_i(x, P) = J_i(x, P) - J_i(-x, P)$. So, the random variable on the left-hand side of the inequality in (A.8) is

$$k\text{-max}\bigl(H_i\bigl(\tau_n|\hat{\theta}_{n,i} - \theta_i(P)|, P\bigr), i \in K\bigr) + o_P(1). \tag{A.9}$$

To examine the limiting distributional behavior of (A.9), let $Y_{n,i} = \tau_n[\hat{\theta}_{n,i} - \theta_i(P)]$. By the almost sure representation theorem, we can assume there exist versions $Y_{n,i}^*$ with $(Y_{n,1}, \dots, Y_{n,s})$ having the same distribution as $(Y_{n,1}^*, \dots, Y_{n,s}^*)$ such that $Y_{n,i}^* \to Y_i$ almost surely, for all i, where $(Y_1, \dots, Y_n)$ has distribution $J_{\{1,\dots,s\}}(P)$. It follows that (A.9) converges in distribution to the distribution of k-max$(|Y_i|, i \in K)$, which is exactly $L_K(\cdot, k, P)$. We can now apply Slutsky's theorem to evaluate (A.8) to conclude its limiting probability is

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

To prove (ii),

$$\begin{aligned} &P\{\theta_i(P) \in \hat{C}_{n,i}\} \\ \text{(A.10)} \qquad &= P\{\tau_n|\hat{\theta}_{n,i} - \theta_i(P)| \le H_{n,i}^{-1}(L_{n,K}^{-1}(1-\alpha, k, \hat{Q}_n), \hat{Q}_n)\} \\ &= P\{H_{n,i}(\tau_n|\hat{\theta}_{n,i} - \theta_i(P), \hat{Q}_n) \le L_{n,K}^{-1}(1-\alpha, k, \hat{Q}_n)\}. \end{aligned}$$

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-hand side of (A.10) tends in probability to $\gamma_K(1-\alpha, k, P)$, the result follows by Slutsky's theorem. □

PROOF OF COROLLARY 3.1. Using the arguments as in the proof of Theorem 3.1, 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

$$\text{(A.11)} \qquad \lim_{n\to\infty} k\text{-FWER}_P = P\{k\text{-max}(J_i(|Y_i|, P), i \in I(P)) > L_K^{-1}(1-\alpha, k, P)\}$$

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

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

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 4.1. Assume $|I(P)| \ge k$, 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 5.1. The proofs of parts (i) and (ii) follow from the arguments preceding the corollary. The proofs of parts (iii) and (iv) are very similar to the proofs of parts (iii) and (iv) of Theorem 3.2 in [31]. □

PROOF OF THEOREM 6.1. 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,

$$\begin{aligned} k\text{-FWER}_P = P\{F \geq k\} &\leq \frac{E(F)}{k} \\ &= \frac{1}{k} E\Big(\sum_{i \in I(P)} I\{\hat{p}_n \leq w_i k\alpha\}\Big) \\ &= \frac{1}{k} \sum_{i \in I(P)} P\{\hat{p}_n \leq w_i k\alpha\} \\ &\leq \frac{1}{k} \sum_{i \in I(P)} w_i k\alpha = \alpha \sum_{i \in I(P)} w_i \leq \alpha. \end{aligned}$$

To prove (ii), the result follows from Theorem 4.1 once we verify the monotonicity condition (4.1). But to show that monotonicity holds, let $I \subseteq K$. Then

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

PROOF OF THEOREM 7.1. Let F be the number of false rejections and let $I(P)$ denote the set of true null hypotheses. Then

$$\begin{aligned} E_P(F) &= E\Big[\sum_{i \in I(P)} I\{\hat{p}_{n,i} \leq w_i \lambda\}\Big] \\ &= \sum_{i \in I(P)} P\{\hat{p}_{n,i} \leq w_i \lambda\} \\ &\leq \lambda \cdot \sum_{i \in I(P)} w_i \leq \lambda. \end{aligned}$$

The second statement is trivial. □

Acknowledgments. We would like to thank the coeditor, the Associate Editor and two anonymous referees for their careful reviews of this manuscript which have led to an improved presentation.

REFERENCES

[1] BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392

[2] BENJAMINI, Y., KRIEGER, A. M. and YEKUTIELI, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika* **93** 491–507. MR2261438

[3] BENJAMINI, Y. and YEKUTIELI, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29** 1165–1188. MR1869245

[4] BERAN, R. (1986). Simulated power functions. *Ann. Statist.* **14** 151–173. MR0829560

[5] BERAN, R. (1988). Balanced simultaneous confidence sets. *J. Amer. Statist. Assoc.* **83** 679–686. MR0963795

[6] BERAN, R. (1988). Prepivoting test statistics: A bootstrap view of asymptotic refinements. *J. Amer. Statist. Assoc.* **83** 687–697. MR0963796

[7] DAVISON, A. C. and HINKLEY, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge Univ. Press, Cambridge. MR1478673

[8] DUDOIT, S., GILBERT, H. and VAN DER LAAN, M. J. (2008). Resampling-based empirical Bayes multiple testing procedures for controlling generalized tail probability and expected value error rates: Focus on the false discovery rate and simulation study. *Biom. J.* **50** 716–744.

[9] DUDOIT, S., SHAFFER, J. P. and BOLDRICK, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statist. Sci.* **18** 71–103. MR1997066

[10] DUDOIT, S., VAN DER LAAN, M. J. and POLLARD, K. S. (2004). Multiple testing. I. Single-step procedures for control of general type I error rates. *Stat. Appl. Genet. Mol. Biol.* **3** 71. Available at http://www.bepress.com/sagmb/vol3/iss1/art13. MR2101462

[11] EFRON, B. and TIBSHIRANI, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York. MR1270903

[12] GENOVESE, C. R. and WASSERMAN, L. (2004). A stochastic process approach to false discovery control. *Ann. Statist.* **32** 1035–1061. MR2065197

[13] HALL, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer, New York. MR1145237

[14] HALL, P. and WILSON, S. (1991). Two guidelines for bootstrap hypothesis testing. *Biometrics* **47** 757–762. MR1132543

[15] HOLM, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Statist.* **6** 65–70. MR0538597

[16] HOMMEL, G. and HOFFMAN, T. (1988). Controlled uncertainty. In *Multiple Hyptheses Testing* (P. Bauer, G. Hommel and E. Sonnemann, eds.) 154–161. Springer, Heidelberg.

[17] 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. *J. Statist. Plann. Inference* **124** 379–398. MR2080371

[18] LAHIRI, S. N. (2003). *Resampling Methods for Dependent Data*. Springer, New York. MR2001447

[19] LEHMANN, E. L. and ROMANO, J. P. (2005). Generalizations of the family-wise error rate. *Ann. Statist.* **33** 1138–1154. MR2195631

[20] LEHMANN, E. L. and ROMANO, J. P. (2005). *Testing Statistical Hypotheses*, 3rd ed. Springer, New York. MR2135927

[21] PERONE PACIFICO, M., GENOVESE, C. R., VERDINELLI, I. and WASSERMAN, L. (2004). False discovery control for random fields. *J. Amer. Statist. Assoc.* **99** 1002–1014. MR2109490

[22] POLITIS, D. N., ROMANO, J. P. and WOLF, M. (1999). *Subsampling*. Springer, New York. MR1707286

[23] 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* 3–9.

[24] ROGERS, J. and HSU, J. (2001). Multiple comparisons of biodiversity. *Biom. J.* **43** 617–625. MR1863493

[25] ROMANO, J. P. (1988). A bootstrap revival of some nonparametric distance tests. *J. Amer. Statist. Assoc.* **83** 698–708. MR0963797

[26] ROMANO, J. P. and SHAIKH, A. M. (2006). On step-down control of the false discovery proportion. In *2nd Lehmann Symposium—Optimality* (J. Rojo, ed.). *Institute of Mathematical Statistics Lecture Notes—Monograph Series*. **49** 33–50. Inst. Math. Statist., Beachwood, OH. MR2337829

[27] ROMANO, J. P. and SHAIKH, A. M. (2006). Stepup procedures for control of generalizations of the family-wise error rate. *Ann. Statist.* **34** 1850–1873. MR2283720

[28] ROMANO, J. P., SHAIKH, A. M. and WOLF, M. (2008). Control of the false discovery rate under dependence using the bootstrap and subsampling (with discussion). *Test* **17** 417–442. MR2470085

[29] ROMANO, J. P., SHAIKH, A. M. and WOLF, M. (2008). Formalized data snooping based on generalized error rates. *Econometric Theory* **24** 404–447. MR2422863

[30] ROMANO, J. P. and WOLF, M. (2005). Exact and approximate step-down methods for multiple hypothesis testing. *J. Amer. Statist. Assoc.* **100** 94–108. MR2156821

[31] ROMANO, J. P. and WOLF, M. (2007). Control of generalized error rates in multiple testing. *Ann. Statist.* **35** 1378–1408. MR2351090

[32] SARKAR, S. K. (2002). Some results on false discovery rate in stepwise multiple testing procedures. *Ann. Statist.* **30** 239–257. MR1892663

[33] SHAO, J. and TU, D. (1995). *The Jackknife and the Bootstrap*. Springer, New York. MR1351010

[34] SPJØTVOLL, E. (1972). On the optimality of some multiple comparison procedures. *Ann. Math. Statist.* **43** 398–411. MR0301871

[35] 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. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **66** 187–205. MR2035766

[36] TROENDLE, J. F. (2000). Stepwise normal theory test procedures controlling the false discovery rate. *J. Statist. Plann. Inference* **84** 139–158. MR1747501

[37] TROENDLE, J. F. (2008). Comment on "Control of the false discovery rate under dependence using the bootstrap and subsampling," by J. Romano, A. Shaikh and M. Wolf. *Test* **17** 456–457.

[38] TU, W. and ZHOU, X. (2000). Pairwise comparison of the means of skewed data. *J. Statist. Plann. Inference* **88** 59–74. MR1767559

[39] 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. *Stat. Appl. Genet. Mol. Biol.* **4** 32. Available at http://www.bepress.com/sagmb/vol4/iss1/art29/. MR2170445

[40] 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. *Stat. Appl. Genet. Mol. Biol.* **3** 27. Available at http://www.bepress.com/sagmb/vol3/iss1/art15/. MR2101464

[41] WESTFALL, P. H. and YOUNG, S. S. (1993). *Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment*. Wiley, New York.

[42] YEKUTIELI, D. and BENJAMINI, Y. (1999). Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *J. Statist. Plann. Inference* **82** 171–196. MR1736442

DEPARTMENTS OF ECONOMICS
AND STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305-4065
USA
E-MAIL: romano@stanford.edu

INSTITUTE FOR EMPIRICAL RESEARCH
IN ECONOMICS
UNIVERSITY OF ZURICH
CH-8006 ZURICH
SWITZERLAND
E-MAIL: mwolf@iew.uzh.ch