Econometric Theory, **24**, 2008, 404–447. Printed in the United States of America.
DOI: 10.1017/S0266466608080171

FORMALIZED DATA SNOOPING BASED ON GENERALIZED ERROR RATES

JOSEPH P. ROMANO
Stanford University

AZEEM M. SHAIKH
University of Chicago

MICHAEL WOLF
University of Zurich

It is common in econometric applications that several hypothesis tests are carried out simultaneously. The problem then becomes how to decide which hypotheses to reject, accounting for the multitude of tests. The classical approach is to control the familywise error rate (FWE), which is the probability of one or more false rejections. But when the number of hypotheses under consideration is large, control of the FWE can become too demanding. As a result, the number of false hypotheses rejected may be small or even zero. This suggests replacing control of the FWE by a more liberal measure. To this end, we review a number of recent proposals from the statistical literature. We briefly discuss how these procedures apply to the general problem of model selection. A simulation study and two empirical applications illustrate the methods.

1. INTRODUCTION

Much empirical research in economics and finance inevitably involves data snooping. The problem arises when several hypothesis tests are carried out simultaneously and one has to decide which hypotheses to reject. One common scenario is the comparison of many strategies (such as investment strategies) to a common benchmark (such as a market index); here, a particular hypothesis test specifies whether a particular strategy outperforms the benchmark or not. Another common scenario is multiple regression models; here, a particular hypothesis test specifies whether a particular regression coefficient is equal to a prespecified value or not.

Economists have long been aware of the dangers resulting from data snooping; see, for example, White (2000), Hansen (2005), Romano and Wolf (2005b),

We thank three anonymous referees for helpful comments that have led to an improved presentation of the paper. The research of the third author has been partially supported by the Spanish Ministry of Science and Technology and FEDER, Grant BMF2003-03324. Address correspondence to Joseph P. Romano, Department of Statistics, Stanford University, Stanford, CA 94305, USA; e-mail: romano@Stanford.edu.

  0266-4666/08 $15.00

and the references therein. The standard approach to account for data snooping is to control (asymptotically) the *familywise error rate* (FWE), which is the probability of making one or more false rejections; see, for example, Westfall and Young (1993).[1] However, this criterion can be too strict when the number of hypotheses under consideration is very large. As a result, it can be become very difficult (or impossible) to make true rejections. In other words, controlling the FWE can be playing it too safe.

When the number of hypotheses is very large and the ability to make true rejections is a main concern, it has been suggested that the researcher relax control of the FWE. In this paper, we discuss and review three proposals to this end. The first proposal is to control the probability of making k or more false rejections, for some integer k greater than or equal to one, which is called the k-FWE.[2] The remaining proposals are based on 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 at all). The second proposal is to control $E(\text{FDP})$, the expected value of the FDP, which is called the *false discovery rate* (FDR). The third proposal is to control $P\{\text{FDP} > \gamma\}$, where γ is a small, user-defined number. In particular, the goal is to construct methods that satisfy $P\{\text{FDP} > \gamma\} \leq \alpha$. Usually $\alpha = 0.05$ or $\alpha = 0.1$; the special case $\alpha = 0.5$ yields control of the median FDP. Although the three proposals are different, they share a common philosophy: by allowing a small number or a small (expected) proportion of false rejections one can improve one's chances of making true rejections and perhaps greatly so. In other words, the price to pay can be small compared to the benefits to reap.

This paper reviews various methods that have been suggested for control of the three criteria previously mentioned, including some very recent multiple testing procedures that account for the dependence structure of the individual test statistics. Part of our contribution is to present the methods in a unified context, allowing an applied researcher to grasp the concepts quickly, rather than having to read and digest the numerous underlying original papers. We also demonstrate, by means of some simulations and two empirical applications, how competing multiple testing procedures compare when applied to data. A previous review paper discussing multiple testing methods is Dudoit, Shaffer, and Boldrick (2003). However, our paper emphasizes more recent methodology and focuses on applications in econometrics and finance rather than microarray experiments.

The remainder of the paper is organized as follows. Section 2 describes the model and the formal inference problem. Section 3 reviews various methods to control the FWE. Section 4 presents various methods to control the k-FWE. Section 5 reviews the method of Benjamini and Hochberg (1995) for control of the FDR. Section 6 presents various methods to control $P\{\text{FDP} > \gamma\}$. Section 7 discusses applications of generalized error rates to the problem of model selection. Section 8 sheds some light on finite-sample performance of the discussed methods via a simulation study. Section 9 provides two empirical appli-

cations. Finally, Section 10 concludes. An Appendix contains some details concerning bootstrap implementation.

2. NOTATION AND PROBLEM FORMULATION

One observes a data matrix $x_{t,l}$ with $1 \leq t \leq T$ and $1 \leq l \leq L$. The data are generated from some underlying probability mechanism P, which is unknown. The row index t corresponds to distinct observations, and there are T of them. In our asymptotic framework, T will tend to infinity. The number of columns L is fixed. For compactness, we introduce the following notation: X_T denotes the complete $T \times L$ data matrix, $X_{t,\cdot}^{(T)}$ is the $L \times 1$ vector that corresponds to the tth row of X_T, and $X_{\cdot,l}^{(T)}$ is the $T \times 1$ vector that corresponds to the lth column of X_T.

Interest focuses on a parameter vector $\theta = \theta(P)$ of dimension S, that is, $\theta = (\theta_1, \ldots, \theta_S)'$. The individual hypotheses concern the elements of θ and can be all one-sided of the form

$$H_s : \theta_s \leq \theta_{0,s} \quad \text{vs. } H_s' : \theta_s > \theta_{0,s}, \tag{1}$$

or they can be all two-sided of the form

$$H_s : \theta_s = \theta_{0,s} \quad \text{vs. } H_s' : \theta_s \neq \theta_{0,s}. \tag{2}$$

For each hypothesis H_s, $1 \leq s \leq S$, one computes a test statistic $w_{T,s}$ from the data matrix X_T. In some instances, we will also consider studentized test statistics $z_{T,s} = w_{T,s}/\hat{\sigma}_{T,s}$, where the standard error $\hat{\sigma}_{T,s}$ estimates the standard deviation of $w_{T,s}$ and is also computed from X_T. In what follows, we often call $w_{T,s}$ a "basic" test statistic to distinguish it from the studentized statistic $z_{T,s}$. We now introduce some compact notation: the $S \times 1$ vector W_T collects the individual basic test statistics $w_{T,s}$; the $S \times 1$ vector Z_T collects the individual studentized test statistics $z_{T,s}$.

A multiple testing method yields a decision concerning each individual testing problem by either rejecting H_s or not. In an ideal world, one would like to reject all those hypotheses that are false. In a realistic world, and given a finite amount of data, this cannot be achieved with certainty. At this point, we vaguely define our goal as making as many true rejections as possible while not making "too many" false rejections. Different notions of accounting for data snooping entertain different views of what constitutes too many false rejections.

We next describe two broad examples where data snooping arises naturally by putting them into the preceding framework.

Example 2.1 (Comparing several strategies to a common benchmark)

Consider S strategies (such as investment strategies) that are compared to a common benchmark (such as a market index). The data matrix X_T has $L = S + 1$ columns: the first S columns record the individual strategies, and the last column records the benchmark. The goal is to decide which strategies out-

perform the benchmark. Here the individual parameters are defined so that $\theta_s \leq 0$ if and only if the sth strategy does not outperform the benchmark. One then is in the one-sided setup (1) with $\theta_{0,s} = 0$ for $s = 1, \ldots, S$.

Example 2.1(a) (Absolute performance of investment strategies)

Historical returns of investment strategy s, such as a particular mutual fund or a particular trading strategy, are recorded in $X^{(T)}_{\cdot,s}$. Historical returns of a benchmark, such as a stock index or a buy-and-hold strategy, are recorded in $X^{(T)}_{\cdot,S+1}$. Depending on preference, these can be "real" returns or log returns; also, returns may be recorded in excess of the risk-free rate if desired. Let μ_s denote the population mean of the return for strategy s. Based on an absolute criterion, strategy s beats the benchmark if $\mu_s > \mu_{S+1}$. Therefore, we define $\theta_s = \mu_s - \mu_{S+1}$. Using the notation

$$\bar{x}_{T,s} = \frac{1}{T} \sum_{t=1}^{T} x_{t,s}$$

a natural basic test statistic is

$$w_{T,s} = \bar{x}_{T,s} - \bar{x}_{T,S+1}. \tag{3}$$

Often, a studentized statistic is preferable and is given by

$$z_{T,s} = \frac{\bar{x}_{T,s} - \bar{x}_{T,S+1}}{\hat{\sigma}_{T,s}}, \tag{4}$$

where $\hat{\sigma}_{T,s}$ is an estimator of the standard deviation of $\bar{x}_{T,s} - \bar{x}_{T,S+1}$.

Example 2.1(b) (Relative performance of investment strategies)

The basic setup is as in the previous example, but now consider a risk-adjusted comparison of the investment strategies, based on the respective Sharpe ratios. With μ_s again denoting the mean of the return of strategy s and with σ_s denoting its standard deviation, the corresponding Sharpe ratio is defined as $SR_s = \mu_s/\sigma_s$.[3] An investment strategy is now said to outperform the benchmark if its Sharpe ratio is higher than the one of the benchmark. Therefore, we define $\theta_s = SR_s - SR_{S+1}$. Let

$$s_{T,s} = \sqrt{\frac{1}{T-1} \sum_{t=1}^{T} (x_{t,s} - \bar{x}_{T,s})^2}.$$

Then, a natural basic test statistic is

$$w_{T,s} = \frac{\bar{x}_{T,s}}{s_{T,s}} - \frac{\bar{x}_{T,S+1}}{s_{T,S+1}}. \tag{5}$$

Again, a preferred statistic might be obtained by dividing by an estimate of the standard deviation of this difference.

Example 2.1(c) (CAPM alpha)

Historical returns of investment strategy s, in excess of the risk-free rate, are recorded in $X^{(T)}_{\cdot,s}$. Historical returns of a market proxy, in excess of the risk-free rate, are recorded in $X^{(T)}_{\cdot,S+1}$. For each strategy s, a simple time series regression

$$x_{t,s} = \alpha_s + \beta_s x_{t,S+1} + \epsilon_{t,s} \tag{6}$$

is estimated by ordinary least squares (OLS). If the capital asset pricing model (CAPM) holds, all intercepts α_s are equal to zero.[4] So, the parameter of interest here is $\theta_s = \alpha_s$. Because the CAPM may be violated in practice, a financial adviser might want to identify investment strategies that have a positive α_s. Hence, a basic test statistic would be

$$w_{T,s} = \hat{\alpha}_{T,s}. \tag{7}$$

Again, it can be advantageous to studentize.

Example 2.2 (Multiple regression)

Consider the multiple regression model

$$y_t = \theta_1 x_{1,t} + \cdots + \theta_H x_{H,t} + \epsilon_t \qquad t = 1, \ldots, T.$$

The data matrix X_T has $L = H + 1$ columns: the first H columns record the explanatory variables, whereas the last column records the response variable, letting $x_{H+1,t} = y_t$. Of interest are $S \leq H$ of the regression coefficients. Without loss of generality, assume that the explanatory variables are ordered in such a way that the coefficients of interest correspond to the first S coefficients, and so $\theta = (\theta_1, \ldots, \theta_S)'$. One typically is in the two-sided setup (2) where the prespecified values $\theta_{0,s}$ depend on the context, but at times the one-sided setup (1) can be more appropriate.

In much applied research, all the regression coefficients are of interest—except possibly an intercept if it is included in the regression—and one would like to decide which of them are different from zero. This corresponds to the two-sided setup (2) where $S = H$—or $S = H - 1$ in the case of an included intercept whose coefficient is not of interest—and $\theta_{0,s} = 0$ for $s = 1, \ldots, S$.

Let $\hat{\theta}_T$ denote an estimator of θ computed from the data matrix X_T, using OLS or feasible generalized least squares (FGLS), say. Then the "basic" test statistic for H_s is simply $w_{T,s} = \hat{\theta}_{T,s}$. The proper choice of the standard error $\hat{\sigma}_{T,s}$ for studentization depends on the context. In the simplest case, it can be the usual OLS standard error. More generally, a standard error that is robust against heteroskedasticity and/or autocorrelation might be required; for example, see White (2001).

For testing an individual hypothesis H_s based on a studentized test statistic $z_{T,s}$, one can typically compute an approximate p-value by invoking asymptotic standard normality. For example, for testing a two-sided hypothesis $H_s: \theta_s = \theta_{0,s}$, one might compute $\hat{p}_{T,s} = 2 \times (1 - \Phi(|z_{T,s}|))$, where $\Phi(\cdot)$ is the standard normal cumulative distribution function (cdf). Of course, one may also appeal to other techniques that rely on approximating the null distribution of $|z_{T,s}|$, such as bootstrapping, subsampling, permutation tests, empirical likelihood, or Edgeworth approximations. In any case, $\hat{p}_{T,s}$ is a marginal p-value in the sense that the test that rejects H_s if $\hat{p}_{T,s} \leq \alpha$ has asymptotic rejection probability α if H_s is true. It follows that if all S null hypotheses are true, and we reject H_s whenever $\hat{p}_{T,s} \leq \alpha$, the expected number of false rejections is $S \times \alpha$ (asymptotically). For example, if $S = 1{,}000$ and $\alpha = 0.05$, the expected number of false rejections is 50 (asymptotically) when all null hypotheses are true. Such an approach is too liberal and does not account for the multitude of tests under study. In the remainder of the paper, we consider various measures of error control that attempt to control false rejections by accounting for the fact that S tests are being carried out simultaneously.

3. METHODS CONTROLLING THE FWE

The usual approach to dealing with data snooping is to try to avoid *any* false rejections. That is, one seeks to control the FWE. The FWE is defined as the probability of rejecting at least one of the true null hypotheses. More specifically, if P is the true probability mechanism, let $I_0 = I_0(P) \subset \{1, \ldots, S\}$ denote the indices of the set of true hypotheses, that is,

$$s \in I_0 \quad \text{if and only if} \begin{cases} \theta_s \leq \theta_{0,s} & \text{in setup (1)} \\ \theta_s = \theta_{0,s} & \text{in setup (2).} \end{cases}$$

The FWE is the probability under P that any H_s with $s \in I_0$ is rejected:[5]

$$\text{FWE}_P = P\{\text{Reject at least one } H_s : s \in I_0(P)\}.$$

If all the individual null hypotheses are false, the FWE is equal to zero by definition.

Control of the FWE requires that, for any P, the FWE be no bigger than α, at least asymptotically. Because this must hold for any P, it must hold not just when all null hypotheses are true (which is called *weak* control) but also when some are true and some are false (which is called *strong* control). As remarked by Dudoit et al. (2003), this distinction is often ignored. The remainder of the paper equates control of the FWE with strong control and similarly for the control of the k-FWE and FDP discussed in later sections. A multiple testing method is said to control the FWE at level α if $\text{FWE}_P \leq \alpha$ for any sample size T and any P. A multiple testing method is said to control the FWE asymptotically at

level α if $\limsup_{T\to\infty} \mathrm{FWE}_P \leq \alpha$ for any P. Methods that control the FWE in finite samples typically can only be derived in special circumstances; see Hochberg and Tamhane (1987) in the context of parametric models and Romano and Wolf (2005a) in the context of semiparametric models and permutation setups.

3.1. The Bonferroni Method

The most familiar multiple testing method for controlling the FWE is the Bonferroni method. For each null hypothesis H_s, one computes an individual p-value $\hat{p}_{T,s}$. The Bonferroni method at level α rejects H_s if $\hat{p}_{T,s} \leq \alpha/S$ and therefore is very simple to apply. Because all p-values are compared to a single critical value, the Bonferroni method is an example of a *single-step* procedure. The disadvantage of the Bonferroni method is that it is in general conservative, resulting in a loss of power.

The Bonferroni method controls the FWE if the distribution of each p-value corresponding to a true null hypothesis is stochastically dominated by the uniform (0, 1) distribution, that is,

$$H_s \text{ true} \Rightarrow P\{\hat{p}_{T,s} \leq u\} \leq u \quad \text{for any } u \in (0,1). \tag{8}$$

The Bonferroni method asymptotically controls the FWE if the distribution of each p-value corresponding to a true null hypothesis is stochastically dominated by the uniform (0, 1) distribution asymptotically, that is,

$$H_s \text{ true} \Rightarrow \limsup_{T\to\infty} P\{\hat{p}_{T,s} \leq u\} \leq u \quad \text{for any } u \in (0,1). \tag{9}$$

3.2. The Holm Method

An improvement over Bonferroni is due to Holm (1979), and it works in a *stepwise* fashion as follows. The individual p-values are ordered from smallest to largest: $\hat{p}_{T,(1)} \leq \hat{p}_{T,(2)} \leq \cdots \leq \hat{p}_{T,(S)}$ with their corresponding null hypotheses labeled accordingly: $H_{(1)}, H_{(2)}, \ldots, H_{(S)}$. Then, $H_{(s)}$ is rejected at level α if $\hat{p}_{T,(j)} \leq \alpha/(S - j + 1)$ for $j = 1, \ldots, s$. In comparison with the Bonferroni method, the criterion for the smallest p-value is equally strict, α/S, but it becomes less and less strict for the larger p-values. Hence, the Holm method will typically reject more hypotheses and is more powerful than the Bonferroni method. On the other hand, the Holm method (asymptotically) controls the FWE under exactly the same condition as the Bonferroni method.

The Holm method starts with examining the most significant hypothesis, corresponding to the smallest p-value, and then moves "down" to the less significant hypotheses. Such stepwise methods are called *stepdown* methods. Different in nature are *stepup* methods, which start by examining the least significant hypothesis, corresponding to the largest p-value, and then move "up" to the more significant hypotheses. An example is the stepwise method of Benjamini and Hochberg (1995); see Section 5.

Although its improvement over Bonferroni can be substantial, the Holm method can also be very conservative. The reason for the conservativeness of the Bonferroni and the Holm methods is that they do not take into account the dependence structure of the individual p-values. Loosely speaking, they achieve control of the FWE by assuming a worst-case dependence structure. If the true dependence structure could (asymptotically) be accounted for, one should be able to (asymptotically) control the FWE but at the same time increase power.[6] In many economic or financial applications, the individual test statistics are jointly dependent. It is therefore important to account for the underlying dependence structure to avoid being overly conservative.

3.3. The Bootstrap Reality Check and the StepM Method

White (2000), in the context of Example 2.1, proposes the bootstrap reality check (BRC). The BRC estimates the sampling distribution of $\max_{1 \leq s \leq S}(w_{T,s} - \theta_s)$, implicitly taking into account the dependence structure of the individual test statistics. Let s_{max} denote the index of the strategy with the largest statistic $w_{T,s}$. The BRC decides whether or not to reject $H_{s_{max}}$ at level α, asymptotically controlling the FWE. It therefore addresses the question of whether the strategy that appears "best" in the observed data really beats the benchmark.[7] On the other hand, it does not attempt to identify as many outperforming strategies as possible.

Hansen (2005) offers some improvements over the BRC. First, his method reduces the influence of "irrelevant" strategies, meaning strategies that "significantly" underperform the benchmark. Second, he proposes the use of studentized test statistics $z_{T,s}$ instead of basic test statistics $w_{T,s}$. However, the method of Hansen (2005) also only addresses the question of whether the strategy that appears best in the observed data really beats the benchmark.

Romano and Wolf (2005b), also in the context of Example 2.1, address the problem of detecting as many outperforming strategies as possible. Often, this will be the relevant problem. For example, if a bank wants to invest money in trading strategies that outperform a benchmark, it is preferable to build a portfolio of several strategies rather than fully invest in the best strategy only. Hence, the goal is to identify the universe of all outperforming strategies for maximum diversification. The stepwise multiple testing (StepM) method of Romano and Wolf (2005b) improves upon the single-step BRC of White (2000) very much in the way that the stepwise Holm method improves upon the single-step Bonferroni method: in terms of being able to detect more outperforming strategies, one is afforded a free lunch. Like the Holm method, the StepM method is of the *stepdown* nature; that is, it starts by examining the most significant hypothesis. Although Romano and Wolf (2005b) develop their StepM method in the context of Example 2.1, it is straightforward to adapt it to the generic multiple testing problems (1) and (2). For details, see Section 4.3.3.

3.4. Use of the Bootstrap

The StepM method and its extensions discussed subsequently are based on the inversion of multiple confidence regions. This can be considered an "indirect" testing method. In the special case of testing a single hypothesis, it corresponds to constructing a confidence interval for the parameter under test and rejecting the null hypothesis if the null value is not contained in the interval. A "direct" testing method, on the other hand, rejects the null hypothesis if a test statistic exceeds a suitable critical value.

Because the StepM method is based on the bootstrap, the indirect testing approach has several advantages. First, in the independent and identically distributed (i.i.d.) case, one can simply resample from the observed data rather than from a distribution that obeys null hypothesis constraints. More generally, one can resample from an estimated model that mimics the underlying probability mechanism, without imposing any null constraints. Second, one can dispense with the technical condition of subset pivotality that is assumed in Westfall and Young (1993) but that is quite restrictive.

Resampling the data was previously suggested by Pollard and van der Laan (2003a) and then generalized by Dudoit, van der Laan, and Pollard (2004a). By recomputing the test statistics from the resampled data and subtracting the values of the original test statistics, they arrive at what they term the *null-value shifted distribution* of the test statistics. It turns out that this is actually equivalent to inverting bootstrap multiple confidence regions. A quite general theory of stepdown methods based on the bootstrap is given in Romano and Wolf (2005a) and is powerful enough to supply both finite-sample and asymptotic results.

A modification to the null-value shifted distribution of Pollard and van der Laan (2003b) and Dudoit et al. (2004a) is proposed by van der Laan and Hubbard (2005). Here, the marginal null distribution of any test statistic can be transformed from the bootstrap distribution to a known marginal null distribution, such as $N(0,1)$ in the context of testing univariate means, while maintaining the multivariate dependence structure.

4. METHODS CONTROLLING THE *k*-FWE

By relaxing the strict FWE criterion one will be able to reject more false hypotheses. This section presents the alternative criterion of controlling the k-FWE. The k-FWE is defined as the probability of rejecting at least k of the true null hypotheses. As before, if P is the true probability mechanism, let $I_0 = I_0(P) \subset \{1, \ldots, S\}$ denote the indices of the set of true hypotheses. The k-FWE is the probability under P that any k or more of the H_s with $s \in I_0$ are rejected:

$$k\text{-FWE}_P = P\{\text{Reject at least } k \text{ of the } H_s : s \in I_0\}.$$

In the case where at least $S - k + 1$ of the individual null hypotheses are false, the k-FWE is equal to zero by definition.

A multiple testing method is said to control the k-FWE at level α if $k\text{-FWE}_P \leq \alpha$ for any sample size T and for any P. A multiple testing method is said to control the FWE asymptotically at level α if $\limsup_{T\to\infty} k\text{-FWE}_P \leq \alpha$ for any P. Methods that control the k-FWE in finite samples typically can only be derived in special circumstances; see Romano and Wolf (2007).

We now describe how the various methods of Section 3 can be generalized to achieve (asymptotic) control of the k-FWE. Of course, because our goal is to reject as many false hypotheses as possible, attention will focus on the generalization of the StepM method.

4.1. Generalization of the Bonferroni Method

The generalized Bonferroni method is due to Hommel and Hoffman (1988) and Lehmann and Romano (2005) and is based on the individual p-values. The method rejects H_s if $\hat{p}_{T,s} \leq k\alpha/S$. It is easy to see that potentially many more hypotheses will be rejected compared to the original Bonferroni method. Indeed, the cutoff value for the individual p-values is k times as large.

If condition (8) holds, then this method controls the k-FWE. If condition (9) holds, then this method asymptotically controls the k-FWE.

4.2. Generalization of the Holm Method

The individual p-values are ordered from smallest to largest, $\hat{p}_{T,(1)} \leq \hat{p}_{T,(2)} \leq \cdots \leq \hat{p}_{T,(S)}$, with their corresponding null hypotheses labeled accordingly, $H_{(1)}, H_{(2)}, \ldots, H_{(S)}$. Then $H_{(s)}$ is rejected at level α if $\hat{p}_{T,(j)} \leq \alpha_j$ for $j = 1, \ldots, s$, where[8]

$$\alpha_j = \begin{cases} \dfrac{k\alpha}{S} & \text{for } j \leq k \\ \dfrac{k\alpha}{S+k-j} & \text{for } j > k. \end{cases}$$

This modification is also due to Hommel and Hoffman (1988) and Lehmann and Romano (2005). It is easy to see that this stepwise method is more powerful than the single-step generalized Bonferroni method. On the other hand, the sufficient conditions for control and asymptotic control, respectively, of the k-FWE are identical.

4.3. Generalization of the StepM Method

We now describe how to generalize the StepM method of Romano and Wolf (2005b) to achieve asymptotic control of the k-FWE. We begin by discussing

the one-sided setup (1) and then describe the necessary modifications for the two-sided setup (2).

4.3.1. Basic Method. We detail the method in the context of using basic test statistics $w_{T,s}$ and discuss the extension to the studentized case later on. Begin by relabeling the strategies according to the size of the individual test statistics, from largest to smallest. Label r_1 corresponds to the largest test statistic and label r_S to the smallest one, so that $w_{T,r_1} \geq w_{T,r_2} \geq \cdots \geq w_{T,r_S}$.

Some further notation is required. Suppose that $\{y_s : s \in K\}$ is a collection of real numbers indexed by a finite set K having $|K|$ elements. Then, for $k \leq |K|$, $k\text{-max}_{s\in K}(y_s)$ is used to denote the kth largest value of the y_s with $s \in K$. So, if the elements y_s, $s \in K$, are ordered as

$$y_{(1)} \leq \cdots \leq y_{(|K|)},$$

then

$$k\text{-max}_{s\in K}(y_s) = y_{(|K|-k+1)}.$$

Further, for any $K \subset \{1, \ldots, S\}$, define

$$c_K(1-\alpha, k, P) = \inf\{x : P\{k\text{-max}_{s\in K}(w_{T,r_s} - \theta_{r_s}) \leq x\} \geq 1-\alpha\}; \tag{10}$$

that is, $c_K(1-\alpha, k, P)$ is the smallest $1-\alpha$ quantile of the sampling distribution under P of $k\text{-max}_{s\in K}(w_{T,r_s} - \theta_{r_s})$.

In the first step of the procedure, we construct a rectangular joint region[9] for the vector $(\theta_{r_1}, \ldots, \theta_{r_S})'$ of the form

$$[w_{T,r_1} - c_1, \infty) \times \cdots \times [w_{T,r_S} - c_1, \infty). \tag{11}$$

Individual decision are then carried out in the following manner: reject H_{r_s} if $\theta_{0,r_s} \notin [w_{T,r_s} - c_1, \infty)$, for $s = 1, \ldots, S$. Equivalently, reject H_{r_s} if $w_{T,r_s} - \theta_{0,r_s} > c_1$, for $s = 1, \ldots, S$.

How should the value c_1 in the construction of the joint region (11) be chosen? Let $\tilde{K}$ denote the index set that corresponds to the relabeled true hypotheses, that is,

$$s \in \tilde{K} \Leftrightarrow r_s \in I_0.$$

Ideally, one would take $c_1 = c_{\tilde{K}}(1-\alpha, k, P)$, because this choice yields control of the k-FWE. To see why, assume without loss of generality that at least k hypotheses are true; otherwise, there is nothing to show. Then, with $c_1 = c_{\tilde{K}}(1-\alpha, k, P)$,

$$
\begin{aligned}
k\text{-FWE}_P &= P\{\text{Reject at least } k \text{ of the } H_s : s \in I_0\} \\
&= P\{\text{Reject at least } k \text{ of the } H_{r_s} : s \in \tilde{K}\} \\
&= P\{k\text{-max}_{s\in\tilde{K}}(w_{T,r_s} - \theta_{r_s}) > c_{\tilde{K}}(1-\alpha, k, P)\} \\
&\le \alpha \quad (\text{by definition of } c_{\tilde{K}}(1-\alpha, k, P)).
\end{aligned}
$$

Unfortunately, the ideal choice $c_1 = c_{\tilde{K}}(1-\alpha, k, P)$ is not available for two reasons. First, the set $\tilde{K}$ is unknown. Second, the probability mechanism P is unknown. The solution to the first problem is to replace $\tilde{K}$ by $\{1, \ldots, S\}$. Because $\tilde{K} \subset \{1, \ldots, S\}$, it follows that $c_{\tilde{K}}(1-\alpha, k, P) \le c_{\{1,\ldots,S\}}(1-\alpha, k, P)$, and so the k-FWE is still controlled. The solution to the second problem is to replace P by an estimate $\hat{P}_T$, that is, to apply the bootstrap. The choice of $\hat{P}_T$ depends on context; see Appendix B of Romano and Wolf (2005b) for details. The cost of replacing P by $\hat{P}_T$ is that control of the k-FWE is weakened to asymptotic control of the k-FWE. Combining the two solutions yields the choice $\hat{c}_1 = c_{\{1,\ldots,S\}}(1-\alpha, k, \hat{P}_T)$. And then any hypothesis H_{r_s} for which $w_{T,r_s} - \theta_{0,r_s} > \hat{c}_1$ is rejected.

By continuing after the first step, more hypotheses can be rejected. Romano and Wolf (2007) show that this increase in power does not come at the expense of sacrificing asymptotic control of the k-FWE. Denote by R_1 the number of rejections in the first step. If $R_1 < k$, stop, because it is plausible that all rejected hypotheses are true. On the other hand, by controlling the k-FWE, if $R_1 \ge k$, we can be confident that some of the rejected hypotheses are false. This knowledge will now lead to smaller joint regions in subsequent steps and hence to potentially further rejections, without sacrificing control of the k-FWE. So if $R_1 \ge k$, continue with the second step and construct a rectangular joint region for the vector $(\theta_{r_{R_1+1}}, \ldots, \theta_{r_S})'$ of the form

$$
[w_{T,r_{R_1+1}} - c_2, \infty) \times \cdots \times [w_{T,r_S} - c_2, \infty). \tag{12}
$$

Individual decisions are carried out analogously to before: reject H_{r_s} if $\theta_{0,r_s} \notin [w_{T,r_s} - c_2, \infty)$, for $s = R_1 + 1, \ldots, S$.

How should the value c_2 in the joint region construction (12) be chosen? Again, the ideal choice $c_2 = c_{\tilde{K}}(1-\alpha, k, P)$ is not available because $\tilde{K}$ and P are unknown. Crucially, instead of replacing $\tilde{K}$ by $\{1, \ldots, S\}$, we can use information from the first step to arrive at a *smaller* value. Namely, if P were known, this value would be given by

$$
c_2 = \max\{c_K(1-\alpha, k, P) : K = I \cup \{R_1 + 1, \ldots, S\}, I \subset \{1, \ldots, R_1\}, |I| = k - 1\},
$$

which is the maximum of a set of $\binom{R_1}{k-1}$ quantiles. The index set of any given quantile corresponds to all the hypotheses not rejected plus $k-1$ out of the R_1 hypotheses that were rejected in the first step, and then one takes the largest such quantile for c_2. The intuition here is as follows. To ensure control of the

k-FWE in the second step, c_2 must satisfy $c_2 \geq c_{\tilde{K}}(1-\alpha, k, P)$. Assuming that k-FWE control was achieved in the first step, it is conceivable that up to $k-1$ true hypotheses have been rejected so far. But, of course, we cannot know which of the rejected hypotheses might be true. So, to play it safe, one must look at all possible combinations of $k-1$ rejected hypotheses, always together with the not rejected hypotheses, and then take the largest of the resulting quantiles. Again, P is unknown, and so c_2 is not available in practice. Replacing P by $\hat{P}_T$ yields the estimate

$$\hat{c}_2 = \max\{c_K(1-\alpha, k, \hat{P}_T) : K = I \cup \{R_1+1, \ldots, S\},$$
$$I \subset \{1, \ldots, R_1\}, |I| = k-1\}.$$

If no further hypotheses are rejected in the second step, stop. Otherwise, continue in this stepwise fashion until no more rejections occur. The following algorithm summarizes the procedure.

ALGORITHM 4.1 (Basic k-StepM method for one-sided setup).

1. *Relabel the strategies in descending order of the test statistics $w_{T,s}$: strategy r_1 corresponds to the largest test statistic and strategy r_S to the smallest.*
2. *For $1 \leq s \leq S$, if $\theta_{0,r_s} \notin [w_{T,r_s} - \hat{c}_1, \infty)$, reject the null hypothesis H_{r_s}. Here*

$$\hat{c}_1 = c_{\{1,\ldots,S\}}(1-\alpha, k, \hat{P}_T).$$

3. *Denote by R_1 the number of hypotheses rejected. If $R_1 < k$, stop; otherwise let $j = 2$.*
4. *For $R_{j-1}+1 \leq s \leq S$, if $\theta_{0,r_s} \notin [w_{T,r_s} - \hat{c}_j, \infty)$, reject the null hypothesis H_{r_s}. Here*

$$\hat{c}_j = \max\{c_K(1-\alpha, k, \hat{P}_T) : K = I \cup \{R_{j-1}+1, \ldots, S\},$$
$$I \subset \{1, \ldots, R_{j-1}\}, |I| = k-1\}. \quad \textbf{(13)}$$

5. *(a) If no further hypotheses are rejected, stop.*
   (b) Otherwise, denote by R_j the number of all hypotheses rejected so far and, afterward, let $j = j+1$. Then return to step 4.

The computation of the constants $\hat{c}_j$ via the bootstrap is detailed in Algorithm A.1 in the Appendix. Let $J_T(P)$ denote the sampling distribution under P of $\sqrt{T}(W_T - \theta)$ and let $J_T(\hat{P}_T)$ denote the sampling distribution under $\hat{P}_T$ of $\sqrt{T}(W_T^* - \hat{\theta}_T)$. Here, $\hat{\theta}_T$ is an estimate of θ based on $\hat{P}_T$.[10] Romano and Wolf (2007) show that a sufficient condition for the basic k-StepM method to control the k-FWE asymptotically is the following.

Assumption 4.1. Let P denote the true probability mechanism and let $\hat{P}_T$ denote an estimate of P based on the data X_T. Assume that $J_T(P)$ converges in distribution to a limit distribution $J(P)$, which is continuous. Further assume that $J_T(\hat{P}_T)$ consistently estimates this limit distribution: $\rho(J_T(\hat{P}_T), J(P)) \to 0$ in probability for any metric ρ metrizing weak convergence.

We now describe how the basic StepM method is modified for the two-sided setup (2). The crux is that the multivariate rectangular joint regions are now the Cartesian products of two-sided intervals rather than one-sided intervals.

To this end, for any $K \subset \{1, \ldots, S\}$, define

$$c_{K,|\cdot|}(1-\alpha, k, P) = \inf\{x : P\{k\text{-max}_{s\in K}|w_{T,r_s} - \theta_{r_s}| \le x\} \ge 1-\alpha\}. \tag{14}$$

That is, $c_{K,|\cdot|}(1-\alpha, k, P)$ is the smallest $1-\alpha$ quantile of the two-sided sampling distribution under P of $k\text{-max}_{s\in K}|w_{T,r_s} - \theta_{r_s}|$.

The following algorithm describes the stepwise method in the two-sided setup.

ALGORITHM 4.2 (Basic k-StepM method for two-sided setup).

1. *Relabel the strategies in descending order of the absolute test statistics $|w_{T,s}|$: strategy r_1 corresponds to the largest absolute test statistic and strategy r_S to the smallest.*
2. *For $1 \le s \le S$, if $\theta_{0,r_s} \notin [w_{T,r_s} \pm \hat{c}_{1,|\cdot|}]$, reject the null hypothesis H_{r_s}. Here*

$$\hat{c}_{1,|\cdot|} = c_{\{1,\ldots,S\},|\cdot|}(1-\alpha, k, \hat{P}_T).$$

3. *Denote by R_1 the number of hypotheses rejected. If $R_1 < k$, stop; otherwise let $j = 2$.*
4. *For $R_{j-1} + 1 \le s \le S$, if $\theta_{0,r_s} \notin [w_{T,r_s} \pm \hat{c}_{j,|\cdot|}]$, reject the null hypothesis H_{r_s}. Here*

$$\hat{c}_{j,|\cdot|} = \max\{c_{K,|\cdot|}(1-\alpha, k, \hat{P}_T) : K = I \cup \{R_{j-1}+1, \ldots, S\},$$
$$I \subset \{1, \ldots, R_{j-1}\}, |I| = k-1\}. \tag{15}$$

5. *(a) If no further hypotheses are rejected, stop.*
   (b) Otherwise, denote by R_j the number of all hypotheses rejected so far and, afterward, let $j = j + 1$. Then return to step 4.

The computation of the constants $\hat{c}_{j,|\cdot|}$ via the bootstrap is detailed in Algorithm A.2 in the Appendix. A sufficient condition for the basic k-StepM method to asymptotically control the k-FWE in the two-sided setup is also given by Assumption 4.1.

4.3.2. Studentized Method. We now describe how to modify the k-StepM method when studentized test statistics are used instead. Ample motivation for

the desirability of studentization in the context of FWE control is provided by Hansen (2005) and Romano and Wolf (2005b). Their reasons carry over to k-FWE control.

Again, begin with the one-sided setup (1). Analogously to (10), define

$$d_K(1-\alpha,k,P) = \inf\{x : P\{k\text{-max}([w_{T,r_s} - \theta_{r_s}]/\hat{\sigma}_{T,r_s} : s \in K) \le x\} \ge 1-\alpha\}. \tag{16}$$

Our stepwise method is then summarized by the following algorithm.

ALGORITHM 4.3 (Studentized k-StepM method for one-sided setup).

1. *Relabel the strategies in descending order of the studentized test statistics $z_{T,s}$: strategy r_1 corresponds to the largest test statistic and strategy r_S to the smallest.*
2. *For $1 \le s \le S$, if $\theta_{0,r_S} \notin [w_{T,r_s} - \hat{\sigma}_{T,r_s}\hat{d}_1, \infty)$, reject the null hypothesis H_{r_s}. Here*

   $$\hat{d}_1 = d_{\{1,\ldots,S\}}(1-\alpha,k,\hat{P}_T).$$

3. *Denote by R_1 the number of hypotheses rejected. If $R_1 < k$, stop; otherwise let $j = 2$.*
4. *For $R_{j-1} + 1 \le s \le S$, if $\theta_{0,r_s} \notin [w_{T,r_s} - \hat{\sigma}_{T,r_s}\hat{d}_j, \infty)$, reject the null hypothesis H_{r_s}. Here*

   $$\hat{d}_j = \max\{d_K(1-\alpha,k,\hat{P}_T) : K = I \cup \{R_{j-1}+1,\ldots,S\},$$
   $$I \subset \{1,\ldots,R_{j-1}\}, |I| = k-1\}. \tag{17}$$

5. *(a) If no further hypotheses are rejected, stop.*
   (b) Otherwise, denote by R_j the number of all hypotheses rejected so far and, afterward, let $j = j+1$. Then return to step 4.

The computation of the constants $\hat{d}_j$ via the bootstrap is detailed in Algorithm A.3 in the Appendix.

The modification to the two-sided setup (2) is now quite obvious. Analogously to (16), define

$$d_{K,|\cdot|}(1-\alpha,k,P) = \inf\{x : P\{k\text{-max}(|w_{T,r_s} - \theta_{r_s}|/\hat{\sigma}_{T,r_s} : s \in K) \le x\} \ge 1-\alpha\}. \tag{18}$$

The algorithm can then be summarized as follows.

ALGORITHM 4.4. (Studentized k-StepM method for two-sided setup).

1. *Relabel the strategies in descending order of the absolute studentized test statistics $|z_{T,s}|$: strategy r_1 corresponds to the largest absolute studentized test statistic and strategy r_S to the smallest.*
2. *For $1 \leq s \leq S$, if $\theta_{0,r_s} \notin [w_{T,r_s} \pm \hat{\sigma}_{T,r_s} \hat{d}_{1,|\cdot|}]$, reject the null hypothesis H_{r_s}. Here*

$$\hat{d}_{1,|\cdot|} = d_{\{1,\ldots,S\},|\cdot|}(1-\alpha, k, \hat{P}_T).$$

3. *Denote by R_1 the number of hypotheses rejected. If $R_1 < k$, stop; otherwise let $j = 2$.*
4. *For $R_{j-1} + 1 \leq s \leq S$, if $\theta_{0,r_s} \notin [w_{T,r_s} \pm \hat{\sigma}_{T,r_s} \hat{d}_{j,|\cdot|}]$, reject the null hypothesis H_{r_s}. Here*

$$\hat{d}_{j,|\cdot|} = \max\{d_{K,|\cdot|}(1-\alpha, k, \hat{P}_T) : K = I \cup \{R_{j-1}+1, \ldots, S\},$$
$$I \subset \{1, \ldots, R_{j-1}\}, |I| = k-1\}. \quad \textbf{(19)}$$

5. *(a) If no further hypotheses are rejected, stop.*
   (b) Otherwise, denote by R_j the number of all hypotheses rejected so far and, afterward, let $j = j + 1$. Then return to step 4.

The computation of the constants $\hat{d}_{j,|\cdot|}$ via the bootstrap is detailed in Algorithm A.4 in the Appendix.

A slightly stronger version of Assumption 4.1 is needed to prove the validity of the studentized method. Again, let X_T^* denote a data matrix generated from probability mechanism $\hat{P}_T$. The basic test statistics computed from X_T^* are denoted by $w_{T,s}^*$. Their corresponding standard errors, also computed from X_T^*, are denoted by $\hat{\sigma}_{T,s}^*$. Romano and Wolf (2007) do not explicitly discuss the case of studentized statistics. However, it is straightforward to show that a sufficient condition for the studentized k-StepM method to control the k-FWE asymptotically, both in the one-sided and the two-sided setup, is the following.

Assumption 4.2. In addition to Assumption 4.1, assume the following condition. For each $1 \leq s \leq S$, both $\sqrt{T}\hat{\sigma}_{T,s}$ and $\sqrt{T}\hat{\sigma}_{T,s}^*$ converge to a (common) positive constant $\sigma_s = \sigma_s(P)$ in probability under P.

Remark 4.1 (Operative method). The computation of the constants $\hat{c}_j$, $\hat{c}_{j,|\cdot|}$, $\hat{d}_j$, and $\hat{d}_{j,|\cdot|}$ in (13), (15), (17), and (19), respectively, may be very expensive if $\binom{R_{j-1}}{k-1}$ is large. In such cases, we suggest the following shortcut. Pick a user-defined number N_{max}, say, $N_{max} = 50$, and let N^* be the largest integer for which $\binom{N^*}{k-1} \leq N_{max}$. The constant $\hat{c}_j$, say, is then computed as

$$\hat{c}_j = \max\{\hat{c}_K(1-\alpha, k, \hat{P}_T) : K = I \cup \{R_{j-1}+1, \ldots, S\},$$
$$I \subset \{R_j - N^* + 1, \ldots, R_j\}, |I| = k-1\}$$

and similarly for the constants $\hat{c}_{j,|\cdot|}$, $\hat{d}_j$, and $\hat{d}_{j,|\cdot|}$. That is, we maximize over subsets I not necessarily of the entire index set of previously rejected hypotheses but only of the index set corresponding to the N^* least significant hypotheses rejected so far. Note that this shortcut does not affect the asymptotic control of the k-FWE even if $N_{max} = 1$ is chosen, resulting in $N^* = k - 1$ and

$$\hat{c}_j = \hat{c}_{\{R_{j-1}-k,\ldots,S\}}(1 - \alpha, k, \hat{P}_T).$$

Nevertheless, in the interest of better k-FWE control in finite samples, we suggest choosing N_{max} as large as possible, subject to computational feasibility.

Remark 4.2. All methods presented in this section can be modified in the following sense while still preserving (asymptotic) control of the k-FWE: reject the $k - 1$ most significant hypotheses no matter what. This means sort the hypotheses in the order of either ascending p-values or descending test statistics to get $H_{r_1}, \ldots, H_{r_S}$; then reject $H_{r_1}, \ldots, H_{r_{k-1}}$ irrespective of the data. Let R denote the number of rejections made by the multiple testing method (before modification). If $R < k$, then the modified method will reject $k - 1$ hypotheses, which is a potentially greater number. If $R \geq k$, then the modified method will reject R hypotheses, that is, the same number. However, it is counterintuitive to reject hypotheses irrespective of the data, and certainly we would also impose the minimal requirement to not reject any hypothesis when the corresponding marginal p-value exceeds α.

Remark 4.3. The use of the k-max statistic was already suggested by Dudoit et al. (2004a) in the construction of a single-step procedure. Our methods here can be seen as stepdown improvements over such single-step procedures.

4.3.3. The StepM Method. Naturally, the StepM method of Romano and Wolf (2005b) can be considered a special case of the k-StepM method by choosing $k = 1$. However, it should be pointed out that the computations are much simplified compared to the case $k > 1$. The reason is that if some hypotheses are rejected in the first step of the StepM method, then for the computation of the values $\hat{c}_j$ and $\hat{d}_j$, $j = 2, 3, \ldots$, one may assume that all hypotheses rejected so far are false.[11] As a result, in the jth step, one does not have to compute the maximum of a set of $\binom{R_{j-1}}{k-1}$ estimated quantiles but rather only a single estimated quantile.

The following algorithm is the simplified version of Algorithm 4.1 for the special case $k = 1$. The simplified versions of Algorithms 4.2–4.4 are analogous.

ALGORITHM 4.5 (Basic StepM method for one-sided setup).

1. *Relabel the strategies in descending order of the test statistics $w_{T,s}$: strategy r_1 corresponds to the largest test statistic and strategy r_S to the smallest.*

2. *For* $1 \le s \le S$, *if* $\theta_{0,r_s} \notin [w_{T,r_s} - \hat{c}_1, \infty)$, *reject the null hypothesis* H_{r_s}. *Here*

   $$\hat{c}_1 = c_{\{1,\ldots,S\}}(1-\alpha, 1, \hat{P}_T).$$

3. *Denote by* R_1 *the number of hypotheses rejected. If* $R_1 = 0$, *stop; otherwise let* $j = 2$.
4. *For* $R_{j-1} + 1 \le s \le S$, *if* $\theta_{0,r_s} \notin [w_{T,r_s} - \hat{c}_j, \infty)$, *reject the null hypothesis* H_{r_s}. *Here*

   $$\hat{c}_j = c_{\{R_{j-1}+1,\ldots,S\}}(1-\alpha, 1, \hat{P}_T).$$

5. *(a) If no further hypotheses are rejected, stop.*
   (b) Otherwise, denote by R_j *the number of all hypotheses rejected so far and, afterward, let* $j = j + 1$. *Then return to step 4.*

4.4. Further Methods

An alternative approach to control the k-FWE is proposed by van der Laan, Dudoit, and Pollard (2004). It begins with an initial procedure that controls the 1-FWE (i.e., the usual FWE) and then rejects in addition the $k-1$ most significant hypotheses not rejected so far. They term this an *augmentation procedure*, because the 1-FWE rejection set is augmented by the $k-1$ next most significant hypotheses to arrive at the k-FWER rejection set. However, this procedure is generally less powerful than the k-StepM method we propose, because it does not take full advantage of the generalized error measure as it relies too heavily on FWE control; for some simulation evidence see Romano and Wolf (2007).

5. METHODS CONTROLLING THE FDR

In many applications, one might be willing to tolerate a larger number of false rejections if there are a larger number of total rejections. In other words, one might be willing to tolerate a certain (small) proportion of false rejections out of the total rejections. This suggests basing error control on the *false discovery proportion* (FDP). Let F be the number of false rejections made by a multiple testing method and let R be the total number of rejections. Then the FDP is defined as follows:

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

Benjamini and Hochberg (1995) propose controlling $E_P(\text{FDP})$, the expected value under P of the FDP, which they termed the *false discovery rate* (FDR). A

multiple testing method is said to control the FDR at level γ if $\mathrm{FDR}_P = E_P(\mathrm{FDP}) \leq \gamma$ for any sample size T and for any P. A multiple testing method is said to control the FDR asymptotically at level γ if $\limsup_{T\to\infty} \mathrm{FDR}_P \leq \gamma$ for any P. Methods that control the FDR in finite samples typically can only be derived in special circumstances.

The stepwise method of Benjamini and Hochberg (1995) is based on individual p-values. The p-values are ordered from smallest to largest, $\hat{p}_{T,(1)} \leq \hat{p}_{T,(2)} \leq \cdots \leq \hat{p}_{T,(S)}$, with their corresponding null hypotheses labeled accordingly, $H_{(1)}, H_{(2)}, \ldots, H_{(S)}$. Then define

$$j^* = \max\{j : \hat{p}_{T,(j)} \leq \gamma_j\} \quad \text{where } \gamma_j = \frac{j}{S}\gamma \tag{20}$$

and reject $H_{(1)}, \ldots, H_{(j^*)}$. If no such j exists, reject no hypotheses. This is an example of a *stepup* method. It starts with examining the least significant hypothesis, $H_{(S)}$, and then moves "up" to the more significant hypotheses if $\hat{p}_{T,(S)} > \gamma$.

Benjamini and Hochberg (1995) prove that their method controls the FDR if condition (8) holds and, in addition, the p-values are mutually independent. Benjamini and Yekutieli (2001) show that independence can be replaced by a more general "positive regression dependency"; see their paper for the exact definition. As a result, it can be proved that, under the dependence condition of Benjamini and Yekutieli (2001), the method of Benjamini and Hochberg (1995) asymptotically controls the FDR if condition (9) holds. On the other hand, (asymptotic) control of the Benjamini and Hochberg (1995) method under an arbitrary dependence structure of the p-values does not hold in general. Benjamini and Yekutieli (2001) show that this more general control can be achieved if the constants γ_j in (20) are replaced by

$$\gamma_j = \frac{j}{S}\frac{\gamma}{C_S} \quad \text{where } C_k = \sum_{s=1}^{k} \frac{1}{s}.$$

Note that $C_S \approx \log(S) + 0.5$ and so this method can have much less power than the original Benjamini and Hochberg (1995) method. For example, when $S = 1{,}000$, then $C_S = 7.49$.

Even when the sufficient condition of Benjamini and Yekutieli (2001) holds, which includes independence as a special case, the method of Benjamini and Hochberg (1995) is conservative in the following sense. Let S_0 denote the number of true null hypotheses, that is, $S_0 = |I(P_0)|$. Then it can be shown that $\mathrm{FDR}_P \leq (S_0/S)\alpha$. So unless $S_0 = S$, power could be improved by replacing the critical constants γ_j in (20) by

$$\gamma_j = \frac{j}{S_0}\gamma.$$

Of course, S_0 is unknown in practice. But there exist several approaches in the literature for estimating S_0. For example, Storey (2002) suggests the following estimator:

$$\hat{S}_0 = \frac{\#\{\hat{p}_{T,j} > \lambda\}}{1 - \lambda}, \tag{21}$$

where $\lambda \in (0,1)$ is a user-specified parameter. The reasoning behind this estimator is as follows. As long as each test has reasonable power, then most of the large p-values should correspond to true null hypotheses. Therefore, one would expect about $S_0(1 - \lambda)$ of the p-values to lie in the interval $(\lambda, 1]$, assuming that the p-values corresponding to the true null hypotheses have approximately a uniform $[0,1]$ distribution. Having estimated S_0, one then applies the Benjamini and Hochberg (1995) procedure with the critical constants γ_j in (20) replaced by

$$\hat{\gamma}_j = \frac{j}{\hat{S}_0}\,\gamma. \tag{22}$$

Storey, Taylor, and Siegmund (2004) study the validity of this "power-improved" FDR procedure when the estimator of S_0 is given by (21). They prove strong control under a weak dependence condition on the individual p-values. This condition includes independence, dependence within blocks, and mixing-type situations. It is, however, stronger than the dependence condition of Benjamini and Yekutieli (2001); for example, it does not allow for a constant correlation across all p-values. Related work is given in Genovese and Wasserman (2004). For another approach to estimating S_0, see Benjamini and Hochberg (2000); however, they do not prove asymptotic strong control of the resulting power-improved FDR procedure.

Finally, there exists a feature with this particular generalized error rate that is often ignored. The FDR is the mean of the FDP, that is, a central tendency of the sampling distribution of the FDP. Therefore, even if the FDR is controlled at level γ, in a given application, the realized FDP could be quite far away from γ. How likely this is depends on the sampling variability of the FDP, which is unknown in practice.[12] Korn, Troendle, McShane, and Simon (2004) provide simulations to shed some light on this issue; also see Section 8.

Remark 5.1 (Positive false discovery rate). The false discovery rate can be rewritten as

$$\mathrm{FDR}_P = E\left[\frac{F}{R}\,1\{R > 0\}\right] = E\left[\frac{F}{R}\,|\,R > 0\right] P(R > 0).$$

In words, it can be expressed as the expectation of the ratio F/R conditional on the fact that there is at least one rejection times the probability of at least one

rejection. As an alternative, Storey (2002, 2003) suggests replacement of the FDR by only the first factor in this product, which he terms the *positive false discovery rate* (pFDP):

$$\text{pFDR}_P = E\left[\frac{F}{R} \,|\, R > 0\right].$$

The pFDR enjoys a number of attractive properties. For example, when the test statistics come from a random mixture of the null distribution and the alternative distribution, the pFDR can be expressed as a simple Bayesian posterior probability. Also, it has a natural connection to classification theory. However, it is not possible to control the pFDR strongly, that is, to achieve (even asymptotically) $\text{pFDR}_P \leq \alpha$ for all P. The reason is that $\text{pFDR}_P = 1$ for all P such that all null hypotheses are true. An application of the pFDR criterion instead involves estimating pFDR_P for the P at hand, say, with the end of improving multiple testing procedures; see Storey (2002). We will not focus further on this alternative error rate in the remainder of the paper.

6. METHODS CONTROLLING THE FDP

Often, one would like to be able to make a statement concerning the realized FDP in a given application. Concretely, one would like to control the FDP in the sense that $P\{\text{FDP} > \gamma\} \leq \alpha$ where $\gamma \in [0,1)$ is a user-defined number. Typical values are $\gamma = 0.05$ and $\gamma = 0.1$; the choice $\gamma = 0$ corresponds to control of the FWE.

A multiple testing method is said to control the FDP at level α if $P\{\text{FDP} > \gamma\} \leq \alpha$ for any sample size T and for any P. A multiple testing method is said to control the FDP asymptotically at level α if $\limsup_{T\to\infty} P\{\text{FDP} > \gamma\} \leq \alpha$ for any P. Methods that control the FDP in finite samples typically can only be derived in special circumstances.

We now describe how some of the methods of Section 3 can be generalized to achieve (asymptotic) control of the FDP. Of course, because our goal is to reject as many false hypotheses as possible, in the end we shall recommend the generalization of the StepM method.

6.1. Generalization of the Holm Method

Lehmann and Romano (2005) develop a stepdown method based on individual p-values. The p-values are ordered from smallest to largest, $\hat{p}_{T,(1)} \leq \hat{p}_{T,(2)} \leq \cdots \leq \hat{p}_{T,(S)}$, with their corresponding null hypotheses labeled accordingly, $H_{(1)}, H_{(2)}, \ldots, H_{(S)}$. Then $H_{(s)}$ is rejected at level α if $\hat{p}_{T,(j)} \leq \alpha_j$ for $j = 1, \ldots, s$, where

$$\alpha_j = \frac{(\lfloor \gamma j \rfloor + 1)\alpha}{S + \lfloor \gamma j \rfloor + 1 - j}.$$

Here, for a real number x, $\lfloor x \rfloor$ denotes the greatest integer that is smaller than or equal to x.

It can be proved that this method provides asymptotic control of the FDP if condition (9) holds. Moreover, this method provides finite-sample control of the FDP if condition (8) holds and the p-values are independent, or at least positively dependent in a certain sense; see Lehmann and Romano (2005). Lehmann and Romano (2005) also show that if one modifies this method by replacing α_j by

$$\alpha'_j = \frac{\alpha_j}{C_{\lfloor \gamma S \rfloor + 1}} \quad \text{where } C_k = \sum_{s=1}^{k} \frac{1}{s}$$

then the resulting stepdown procedure controls the FDP under no dependence assumptions on the p-values. This method has since been improved by Romano and Shaikh (2006a) in that the constant $C_{\lfloor \gamma S \rfloor + 1}$ has been replaced by a smaller one, while still maintaining finite-sample control under assumption (8) and asymptotic control under assumption (9). A similar stepup procedure is derived in Romano and Shaikh (2006b).

6.2. Generalization of the StepM Method

The crux of our procedure is to sequentially apply the k-StepM method, employing $k = 1, 2, 3, \ldots$, until a stopping rule indicates termination. The appropriate variant of the k-StepM method is dictated by the nature of the multiple testing problem, one-sided versus two-sided, and the choice of test statistics, basic versus studentized. For example, the one-sided setup (1) in combination with studentized test statistics calls for Algorithm 4.3.

To develop the idea, consider controlling $P\{\text{FDP} > 0.1\}$. We start out by applying the 1-StepM method, that is, by controlling the FWE. Denote by N_1 the number of hypotheses rejected. Because of the FWE control, one can be confident that no false rejection has occurred and that, in return, the FDP has been controlled. Consider now rejecting $H_{(N_1+1)}$, the next most significant hypothesis. Of course, if $H_{(N_1+1)}$ is false, there is nothing to worry about, and so suppose that $H_{(N_1+1)}$ is true. Assuming FWE control in the first step, the FDP upon rejection of $H_{(N_1+1)}$ then becomes $1/(N_1 + 1)$, which is greater than 0.1 if and only if $N_1 < 9$. So if $N_1 \geq 9$ we can reject one true hypothesis and still avoid FDP > 0.1. This suggests stopping if $N_1 < 9$ and otherwise applying the 2-StepM method, which, by design, should not reject more than one true hypothesis. Denote the total number of hypotheses rejected by the 2-StepM method by N_2. Reasoning similarly to before, if $N_2 < 19$, we stop, and other-

wise we apply the 3-StepM method. This procedure is continued until $N_j < 10j - 1$ at some point.

The following algorithm describes the method for arbitrary γ.

ALGORITHM 6.1 (FDP-StepM method).

1. *Let* $j = 1$ *and let* $k_1 = 1$.
2. *Apply the* k_j*-StepM method and denote by* N_j *the number of hypotheses rejected.*
3. *(a) If* $N_j < k_j/\gamma - 1$, *stop.*
   (b) Otherwise, let $j = j + 1$ *and, afterward, let* $k_j = k_{j-1} + 1$. *Then return to step 2.*

Romano and Wolf (2007) show that a sufficient condition for the FDP-StepM method to provide asymptotic control of the FDP is Assumption 4.1 in the case where the underlying k-StepM method uses basic test statistics. Similarly, it can be proved that a sufficient condition for the FDP-StepM method to provide asymptotic control of the FDP is Assumption 4.2 in the case where the underlying k-StepM method uses studentized test statistics.

6.3. Further Methods

An alternative approach to controlling the FDP is proposed by van der Laan et al. (2004) and Dudoit, van der Laan, and Pollard (2004b). It begins with an initial procedure that controls the 1-FWE (i.e., the usual FWE). Let R denote the number of rejections by the 1-FWE procedure. Then the proposal rejects in addition the D next most significant hypotheses where D is the largest integer that satisfies

$$\frac{D}{D+R} \leq \gamma.$$

This is also an *augmentation procedure* because the 1-FWE rejection set is suitably augmented by the next most significant hypotheses to arrive at the FDP rejection set. However, this procedure is generally less powerful than the FDP-StepM method we propose; for some simulation evidence see Romano and Wolf (2007). A further new method based on an empirical Bayes approach is given in van der Laan, Birkner, and Hubbard (2005).

6.4. Controlling the Median FDP

As an alternative to controlling the FDR, which is the expected value of the FDP, we propose controlling the median of the FDP. Obviously, if one achieves $P\{\text{FDP} > \gamma\} \leq 0.5$, then the median FDP is bounded above by γ. So choosing

$\alpha = 0.5$ for the methods in this section asymptotically controls the median FDP under an arbitrary dependence structure of the p-values (or test statistics).

Although, in this sense, controlling the median FDP is more generally valid than controlling the FDR by the method of Benjamini and Hochberg (1995), it should be pointed out that it is a less stringent measure and, therefore, potentially less useful in applications.

First, if all hypotheses are true, controlling the FDR also controls the FWE in the sense that $\text{FWE}_P \leq \gamma$. On the other hand, assuming that $\gamma > 0$, controlling the median FDP only achieves $\text{FWE}_P \leq 0.5$.

Second, if the FDR is controlled at level $\gamma = 0.1$, say, then the sampling distribution of the FDP must necessarily be quite concentrated around 0.1, given the lower bound of zero for the FDP. In particular, there cannot be a lot of mass at values very much greater than 0.1. On the other hand, control of the median FDP is achieved as long as there is at least probability mass 0.5 below 0.1 for the sampling distribution of the FDP. In particular, this allows for a lot of mass at values very much greater than 0.1 (in principle, up to mass 0.5 at the point 1). As a result, the chance of the realized FDP greatly exceeding 0.1 can be much bigger when controlling the median FDP compared to controlling the FDR. We will examine this issue to some extent in Section 8.

7. APPLICATIONS TO MODEL SELECTION

This section briefly discusses how control of generalized error rates can apply to the problem of model selection. In fact, the term *model selection* is rather vague and can mean different things depending on context. Therefore, we consider various notions.

White (2000) studies the problem of comparing a large number of (forecast) models to a common (forecast) benchmark. In this context, model selection is the challenge of deciding which models are superior to the benchmark. Therefore, in this context, model selection becomes a special case of Example 2.1 by interpreting (forecast) models as strategies. White (2000) proposes control of the FWE, but when the number of strategies is very large this criterion can be too strict and a generalized error rate may be more appropriate. Some empirical applications based on the FWE when the number of strategies is very large are as follows. Sullivan, Timmermann, and White (1999), White (2000), and Sullivan, White, and Golomb (2001) all fail to find any outperforming strategies when comparing a large number, S, of trading strategies against the benchmark of "buy and hold." The numbers of trading strategies considered are $S = 7,846$, $S = 3,654$, and $S = 9,452$, respectively. Hansen (2005) fails to find any outperforming strategies when comparing $S = 3,304$ strategies to forecast inflation against the benchmark of "last period's inflation." On the other hand, when he restricts attention to a smaller universe of $S = 352$ strategies, some outperformers are detected. It appears that when the number of strategies is in the thousands, controlling the FWE becomes too stringent.

The task of constructing an optimal forecast provides another notion. Imagine that several forecasting strategies are available to forecast a quantity of interest. As described in Timmermann (2006, Sect. 6): (i) choosing the lone strategy with the best track record is often a bad idea; (ii) simple forecasting schemes, such as equal-weighting various strategies, are hard to beat; and (iii) trimming off the worst strategies is often required. In this context, model selection is the challenge of identifying the worst strategies. A sensible approach is as follows. First, one labels those strategies as the worst strategies that *underperform* a suitable benchmark.[13] Second, one is now back again in Example 2.1 except that the individual parameters need to be defined in such a way that $\theta_s \leq 0$ if and only if the sth strategy does not *underperform* the benchmark. Typically, this can be achieved by defining the parameters according to Example 2.1 and then reversing their signs.

In many empirical applications, a large-dimensional regression model is estimated, and the question becomes which explanatory variables are the important ones. In this context, model selection is the challenge of identifying the nonzero regression coefficients; see Example 2.2. An unfortunate common practice is identification based on individual p-values, ignoring the multiple testing problem altogether.[14] As a result, one typically identifies too many variables as important. For example, if there are 100 variables under test, all of which are unimportant, then, based on comparing individual p-values to the level $\alpha = 5\%$, one would expect to falsely identify five variables as important. On the other hand, dealing with the multitude of tests by applying the FWE can be too strict, especially when the number of explanatory variables is very large. As a result, one may easily overlook important variables. A sensible solution is therefore to employ a suitable generalized error rate, such as controlling the (median) FDP. Note that the estimated regression coefficients may depend on each other in a way that violates the positive regression dependency assumption and so the validity of the FDR procedure of Benjamini and Hochberg (1995) is not guaranteed.

Related to the model selection notion of the previous paragraph, though more complex, is the problem of determining which explanatory variables to keep in a final model, say, for prediction purposes. This problem is commonly known as "subset selection," and many popular techniques exist, such as pretesting methods, stepwise selection (forward or backward), the application of information criteria such as the Akaike information criterion (AIC) or the Bayesian information criterion (BIC), and principal components regression. See Draper and Smith (1998) and Hastie, Tibshirani, and Friedman (2001) for details. An explicit use of tests of multivariate parameters as a means of consistent variable selection can be found in Pötscher (1983) and Bauer, Pötscher, and Hackl (1988). Another popular technique for subset selection is general-to-specific modeling; see Campos, Ericsson, and Hendry (2005) for an introduction. As a part of the procedure, individual variables are kept in the model based on so-called simplification tests where individual p-values are compared to a (com-

mon) significance level α. The choice of this level appears as much of an art as a science. For example, Krolzig and Hendry (2001) discuss how it is advantageous to choose a small level α when there are many irrelevant explanatory variables. However, they do not address the question of how one is to know whether this is the case in practice. The optimal level α for the individual tests depends not only on the number of explanatory variables, which is known, but also on the number of them that are irrelevant and the dependence structure of the regression coefficient estimates, both of which are unknown. Therefore, a viable alternative may be to consider the simplification tests as a multiple testing problem in conjunction with a generalized error rate such as controlling the (median) FDP. Such an approach can implicitly account both for the number of irrelevant variables and the dependence structure of the regression coefficient estimates.

Jensen and Cohen (2000) discuss multiple comparisons in induction algorithms. In this context, model selection is the challenge of deciding which variables to include in an artificial intelligence (AI) model for prediction and classification purposes. They describe how a procedure ignoring the multiple testing problem leads to undesirable effects such as overfitting, that is, the inclusion of too many variables in the model. Control of a generalized error rate may therefore be desirable. Moreover, Jensen and Cohen (2000) show in some simulations that multiple testing procedures that do not account for the dependence structure of the test statistics, such as Bonferroni, can work well when the dependence structure is absent or weak but work poorly when the dependence structure is noticeable. Hence, it is desirable to employ a procedure that accounts for the dependence structure.

Abramovich and Benjamini (1996) and Abramovich, Benjamini, Donoho, and Johnstone (2005) study the problem of recovering an S-dimensional vector observed in white noise, where S is large and the vector is known to be sparse. Abramovich et al. (2005) discuss various definitions of sparseness, the most intuitive being the proportion of the nonzero entries of the vector. In this context, model selection is the challenge of deciding which entries are nonzero to optimally estimate the vector.[15] A suggested solution is to consider the problem as a suitable multiple testing problem where the individual hypotheses test whether the entries of the vector are zero or nonzero (and so the hypotheses are two-sided). Then the FDR criterion is employed to account for the multitude of tests. Abramovich et al. (2005) show that this approach based on the FDR enjoys optimality properties, but their asymptotic framework is somewhat restrictive.[16] In addition, the error terms are assumed independent of each other. As an alternative, one might control the (median) FDP instead.

Recently, Buena, Wegkamp, and Auguste (2006) show how testing using FDR control can be used to produce consistent variable selection even in high-dimensional models. Of course, other measures of error control can similarly be exploited.

We also mention a notion of model selection that does not fit into our framework. Again, imagine that several forecasting strategies are available to fore-

cast a quantity of interest. Now, the question is which of those strategies is the "best." In this context, model selection is the challenge of identifying the best model out of a universe of candidate models. Needless to say, given a finite amount of data, the best model cannot determined with certainty. Therefore, the suggested solution consists of constructing a *model confidence set*, that is, a data-dependent collection of models that will contain the best model with a prespecified probability, at least asymptotically. For related work see Shimodaira (1998), Hansen, Lunde, and Nason (2003), Hansen, Lunde, and Nason (2005), and the references therein.

Although the preceding discussion reveals that multiple hypothesis testing methods may be useful in the model building process, the problem of inference for parameters of a data-based model is crucial and challenging. For recent entries to the literature on inference after model selection, see Shen, Huang, and Ye (2004) and Kabaila and Leeb (2006) and the references in these works.

8. SIMULATION STUDY

This section presents a small simulation study in the context of testing population means. We generate independent random vectors $X_1, \ldots, X_T$ from an S-dimensional multivariate normal distribution with mean vector $\theta = (\theta_1, \ldots, \theta_S)'$, where $T = 100$ and $S = 500$. The null hypotheses are $H_s: \theta_s \leq 0$ and the alternative hypotheses are $H_s: \theta_s > 0$, and so we are in the one-sided setup (1). The studentized test statistics are $z_{T,s} = w_{T,s}/\hat{\sigma}_{T,s}$, where

$$w_{T,s} = \bar{X}_{\cdot,s} = \frac{1}{T}\sum_{t=1}^{T} X_{t,s} \quad \text{and} \quad \hat{\sigma}^2_{T,s} = \frac{1}{T(T-1)}\sum_{t=1}^{T}(X_{t,s} - \bar{X}_{\cdot,s})^2.$$

The individual means θ_s are equal to either 0 or 0.25. The number of means equal to 0.25 is 0, 100, 200, or 400. The covariance matrix is of the common correlation structure: $\sigma_{s,s} = 1$ and $\sigma_{s,j} = \rho$ for $s \neq j$. We consider the values $\rho = 0$ and $\rho = 0.5$. Other specifications of the covariance matrix do not lead to results that are qualitatively different; see Romano and Wolf (2007).

We include the following multiple testing procedures in the study. The value of k is $k = 10$. The nominal level is $\alpha = 0.05$, unless indicated otherwise.

- (StepM) The studentized StepM construction of Romano and Wolf (2005b).
- (k-gH) The k-FWE generalized Holm procedure described in Section 4.2, where the individual p-values are derived from $z_{T,s} \sim t_{T-1}$ under $\theta_s = 0$.
- (k-StepM) The studentized k-StepM construction described in Section 4.3. This procedure is based on the operative method with $N_{max} = 50$; see Remark 4.1.
- (FDP-LR$_{0.1}$) The FDP procedure of Lehmann and Romano (2005) with $\gamma = 0.1$ described in Section 6.1.

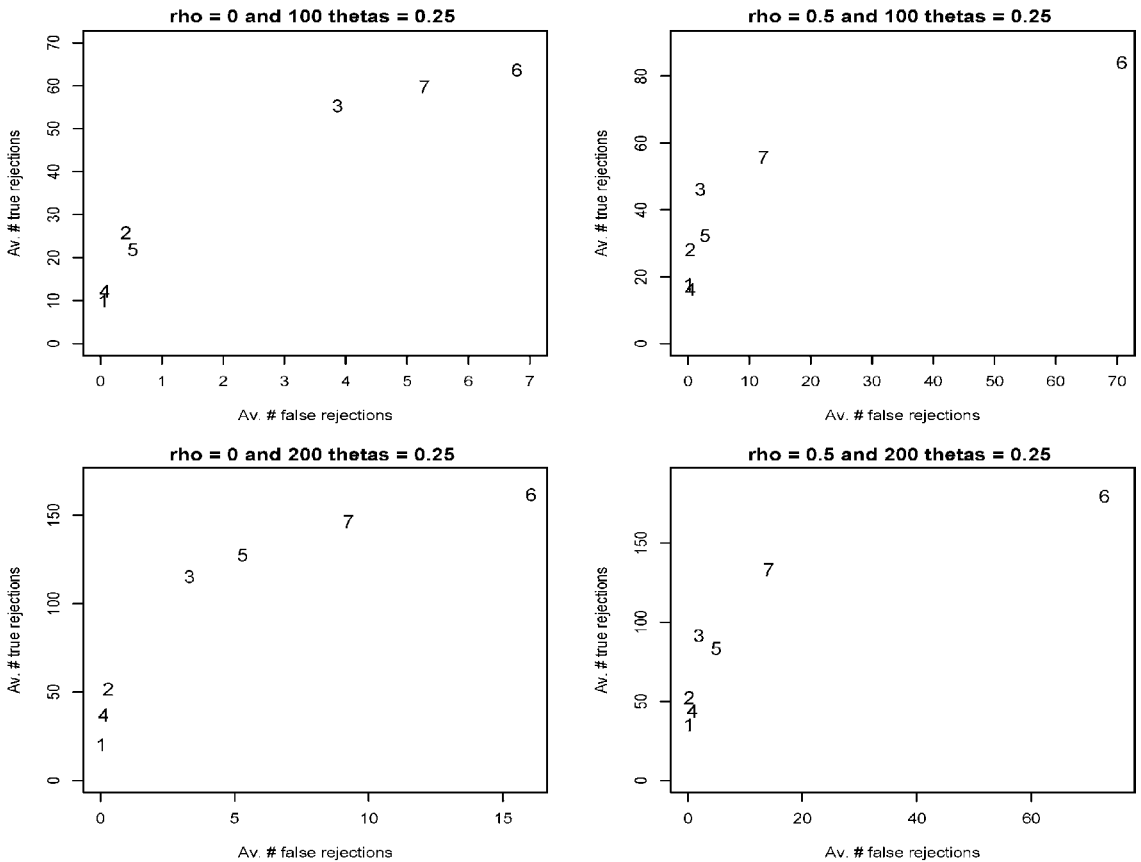
- (FDP-StepM$_{0.1}$) The studentized FDP-StepM construction described in Section 6.2 with $\gamma = 0.1$.
- (FDP-StepM$_{0.1}^{Med}$) The studentized FDP-StepM construction described in Section 6.2 with $\gamma = 0.1$ but nominal level $\alpha = 0.5$. Therefore, this procedure asymptotically controls the median FDP to be bounded above by $\gamma = 0.1$.
- (FDR-BH$_{0.1}$) The FDR construction of Benjamini and Hochberg (1995) described in Section 5 with $\gamma = 0.1$.

The performance criteria are (i) the empirical k-FWEs and FDPs, compared to the nominal level $\alpha = 0.05$ (or $\alpha = 0.5$ for the method controlling the median FDP), and the empirical FDRs and (ii) the average number of false hypotheses rejected. The results are presented in Table 1. They can be summarized as follows.

- All methods provide satisfactory finite-sample control of their respective k-FWE, FDP, or FDR criteria.
- By controlling a generalized error rate, the power is often much improved compared to FWE control.
- The methods that implicitly account for the dependence structure of the test statistics are more powerful than the worst case methods based on individual p-values: 10-StepM is more powerful than 10-gH and FDP-StepM$_{0.1}$ is more powerful than FDP-LR$_{0.1}$.
- The methods controlling a central tendency of the FDP are more powerful than the methods controlling $P\{\text{FDP} > 0.1\}$: FDP-StepM$_{0.1}^{Med}$ and FDR-BH$_{0.1}$ are more powerful than FDP-LR$_{0.1}$ and FDP-StepM$_{0.1}$.

By design, the increase in power that one is afforded by controlling a generalized error rate comes at the expense of relaxing the strict FWE criterion. As a result, the expected number of false rejections typically also increases. This relationship is depicted in Figure 1, where for various scenarios the average number of true rejections is plotted against the average number of false rejections. In these scatter plots, each method is represented by number, where the numbers correspond to the order of the methods in Table 1. Not surprisingly, the relationship is generally increasing and concave with StepM being in the lower left corner and FDP-StepM$_{0.1}^{Med}$ being in the upper right corner.

Recall the discussion of Section 6.4 where some virtues of controlling the FDR versus controlling the median FDP were mentioned. To examine this issue, we look at the sampling distribution of the FDP when the median FDP and the FDR are controlled. Figure 2 summarizes the distribution of the realized FDPs for various scenarios via box plots. It can be seen that, although median FDP control and FDR control are achieved, the variation of the sampling distributions is considerable, especially for the case of common correlation $\rho = 0.5$. As a result, the realized FDP may well be quite above $\gamma = 0.1$. This feature is more pronounced for control of the median FDP, especially when $\rho = 0.5$.


rho = 0 and 100 thetas = 0.25
Av. # true rejections
Av. # false rejections
rho = 0.5 and 100 thetas = 0.25
Av. # true rejections
Av. # false rejections
rho = 0 and 200 thetas = 0.25
Av. # true rejections
Av. # false rejections
rho = 0.5 and 200 thetas = 0.25
Av. # true rejections
Av. # false rejections

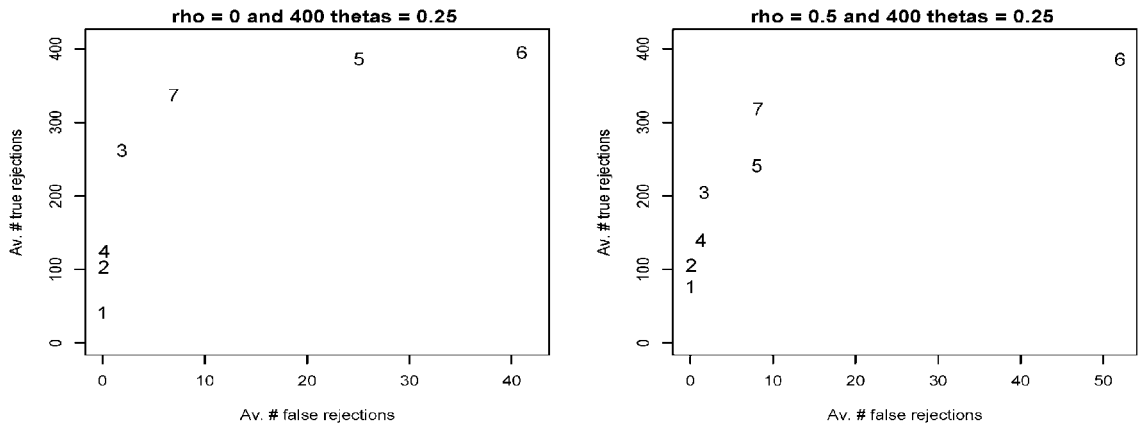


FIGURE 1. Scatter plots of average number of true rejections against average number of false rejections for various scenarios. The numbers correspond to the order of the methods in Table 1. That is, 1 corresponds to StepM, 2 corresponds to 10-gH, . . . , and 7 corresponds to FDR-$BH_{0.1}$.

TABLE 1. Empirical FWEs, FDPs, and FDRs (in the rows "Control" and expressed as %) and average number of false hypotheses rejected (in the rows "Rejected") for various methods, with $T = 100$ and $S = 500$. The nominal level is $\alpha = 5\%$, apart from the second to last column where it is $\alpha = 50\%$. The number of repetitions is 5,000 when all $\theta_s = 0$ and 2,000 for all other scenarios, and the number of bootstrap resamples is $M = 200$.

	StepM	10-gH	10-StepM	FDP-LR$_{0.1}$	FDP-StepM$_{0.1}$	FDP-StepM$_{0.1}^{Med}$	FDR-BH$_{0.1}$
Common correlation: $\rho = 0$							
All $\theta_i = 0$							
Control	5.4	0.0	1.6	5.0	5.4	55.4	10.5
Rejected	0.0	0.0	0.0	0.0	0.0	0.0	0.0
One hundred $\theta_i = 0.25$							
Control	3.5	0.0	0.9	0.9	2.1	43.5	7.9
Rejected	9.9	25.8	55.2	12.1	22.4	63.5	59.6
Two hundred $\theta_i = 0.25$							
Control	3.1	0.0	0.4	0.0	0.2	33.7	6.0
Rejected	20.3	51.6	115.1	36.9	127.7	161.7	146.2
Four hundred $\theta_i = 0.25$							
Control	1.2	0.0	0.0	0.0	1.1	32.1	2.0
Rejected	41.1	102.9	261.0	124.4	385.5	394.8	336.3
Common correlation: $\rho = 0.5$							
All $\theta_i = 0$							
Control	5.6	0.9	5.3	2.2	5.5	52.3	5.0
Rejected	0.0	0.0	0.0	0.0	0.0	0.0	0.0
One hundred $\theta_i = 0.25$							
Control	4.8	0.6	5.2	1.1	5.3	48.9	6.3
Rejected	16.9	27.0	44.4	15.2	30.2	83.6	53.8
Two hundred $\theta_i = 0.25$							
Control	3.9	0.3	4.9	0.6	5.3	49.9	5.3
Rejected	35.3	52.5	92.0	44.0	83.7	179.5	134.0
Four hundred $\theta_i = 0.25$							
Control	3.5	0.1	5.3	0.3	5.3	51.1	2.0
Rejected	77.3	106.0	203.1	139.8	238.4	385.3	316.9

9. EMPIRICAL APPLICATIONS

9.1. Hedge Fund Evaluation

The data set we consider is similar to one in Romano and Wolf (2005b). The difference is that we focus on a shorter time horizon, thereby increasing the number of funds under study. Our universe consists of all hedge funds in the Center for International Securities and Derivatives Markets (CISDM) database that have a complete return history from 01/1994 until 12/2003.

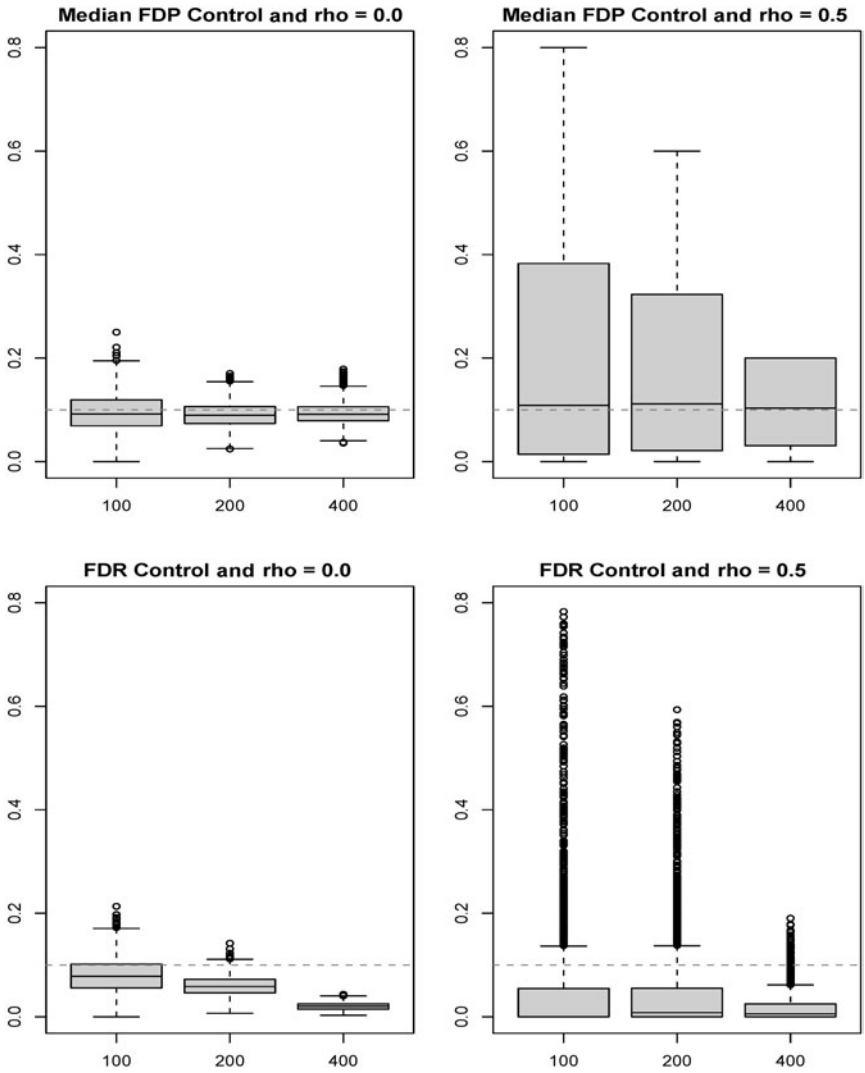


FIGURE 2. Box plots of realized FDPs for various scenarios. The upper part is for control of the median FDP, and the lower part is for control of the FDR. The labels on the x-axis—100, 200, and 400—denote the number of false hypotheses. The horizontal dashed line indicates $\gamma = 0.1$.

All returns are net of management and incentive fees, and so they are the returns obtained by the investors. As in Romano and Wolf (2005b), we benchmark the funds against the risk-free rate,[17] and all returns are log returns. So we are in the situation of Example 2.1(a). It is well known that hedge fund

returns, unlike mutual fund returns, tend to exhibit nonnegligible serial correlations; see, for example, Lo (2002) and Kat (2003). Accordingly, one has to account for this time series nature to obtain valid inference. Studentization for the original data uses a kernel variance estimator based on the prewhitened quadratic spectral (QS) kernel and the corresponding automatic choice of bandwidth of Andrews and Monahan (1992). The bootstrap method is the circular block bootstrap, based on $M = 5{,}000$ repetitions. The studentization in the bootstrap world uses the corresponding "natural" variance estimator; for details, see Götze and Künsch (1996) and Romano and Wolf (2006). The block sizes for the circular bootstrap are chosen via Algorithm A.5. The semiparametric model $\tilde{P}_T$ used in this algorithm is a VAR(1) model in conjunction with bootstrapping the residuals.[18]

There are 210 funds in the CISDM database with a complete return history from 01/1994 until 12/2003, and the number of monthly observations is $T = 120$. However, one fund is deleted from the universe because of a highly unusual return distribution, and so the number of funds included in the study is $S = 209$ in the end. (Fund 154, Paradigm Master Fund, reported one unusually large negative return; see Figure 3. As a result, it unduly dominates the bootstrap sampling distribution of the largest studentized test statistics z^*_{120,r_1}; see Figure 4.) Table 2 lists the 10 largest basic and studentized test statistics, together with the corresponding hedge funds. Similar to the analysis of Romano and Wolf (2005b), the two lists are almost completely disjoint; only the fund JMG Capital Partners appears in both lists.

We now use the various multiple testing methods to identify hedge funds that outperform the risk-free rate, starting with the the Holm procedure and its generalizations and also the FDR procedure of Benjamini and Hochberg (1995), all of which are based on individual p-values only. The p-values are obtained by the studentized circular block bootstrap, which corresponds to applying the StepM method to each single strategy, that is, the special case $S = 1$. The block sizes for the circular block bootstrap are chosen, individually for each fund, via Algorithm A.5 in the Appendix for the special case $S = 1$. The semiparametric model $\tilde{P}_T$ used in this algorithm is an AR(1) model in conjunction with bootstrapping the residuals.[19] The results are displayed in the left half of Table 3.

Next, we turn to the studentized StepM method and its generalizations.[20] The block sizes for the circular block bootstrap are chosen via Algorithm A.5. The semiparametric model $\tilde{P}_T$ used in this algorithm is a VAR(1) model in conjunction with bootstrapping the residuals.[21] The k-StepM procedures are based on the operative method using $N_{max} = 100$; see Remark 4.1. The results are displayed in the right half of Table 3.

Not surprisingly, the results are comparable to those of the simulation study. First, when a generalized error rate is controlled, the number of rejected hypotheses can greatly increase. For example, for the nominal level of $\alpha = 0.1$, whereas the (1-)StepM method rejects 16 hypotheses, the 2-StepM method rejects 29 hypotheses. Second, the methods that implicitly account for the dependence

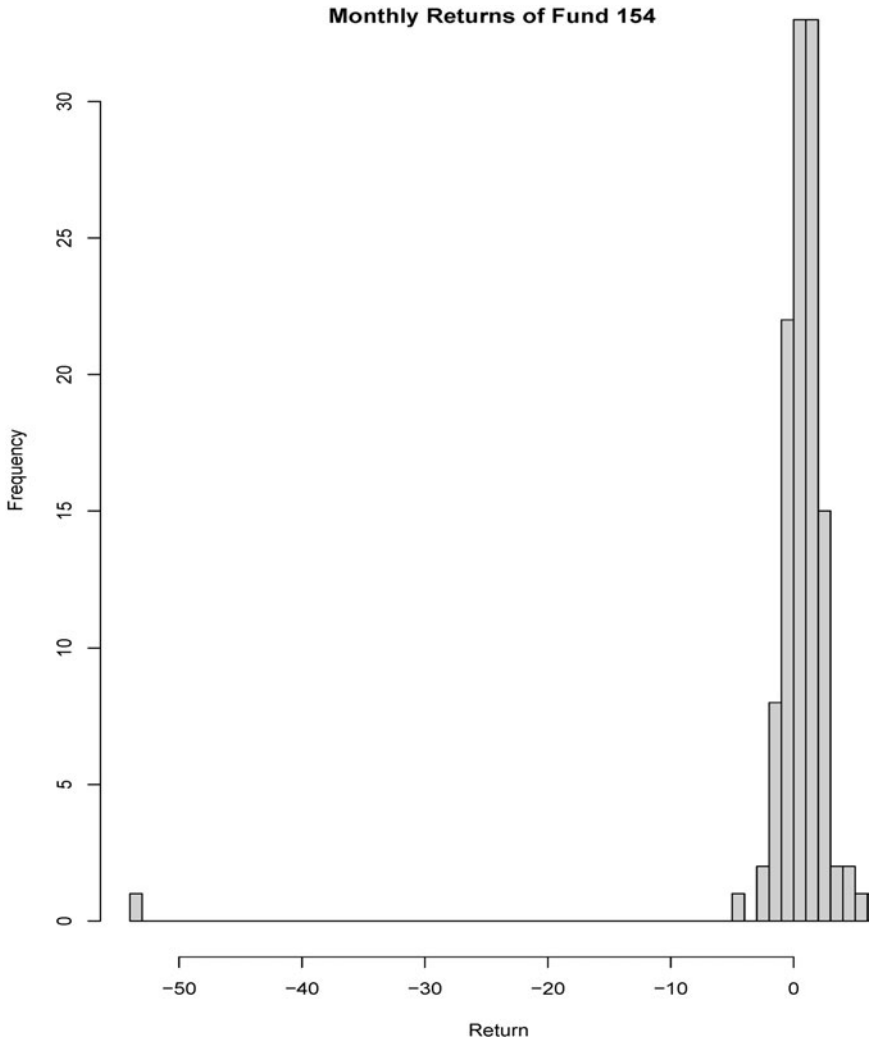


FIGURE 3. Histogram of the monthly log returns of fund 154. In 08/1995 the fund, Paradigm Master Fund, reported a return of −53.77%, resulting in a tremendous outlier to the left.

structure of the test statistics reject more hypotheses than the methods based on individual p-values. For example, for the nominal level of $\alpha = 0.1$, whereas the FDP-$LR_{0.1}$ method rejects 22 hypotheses, the FDP-$StepM_{0.1}$ method rejects 36 hypotheses. Third, the methods controlling a central tendency of the FDP are the ones that reject the most hypotheses.

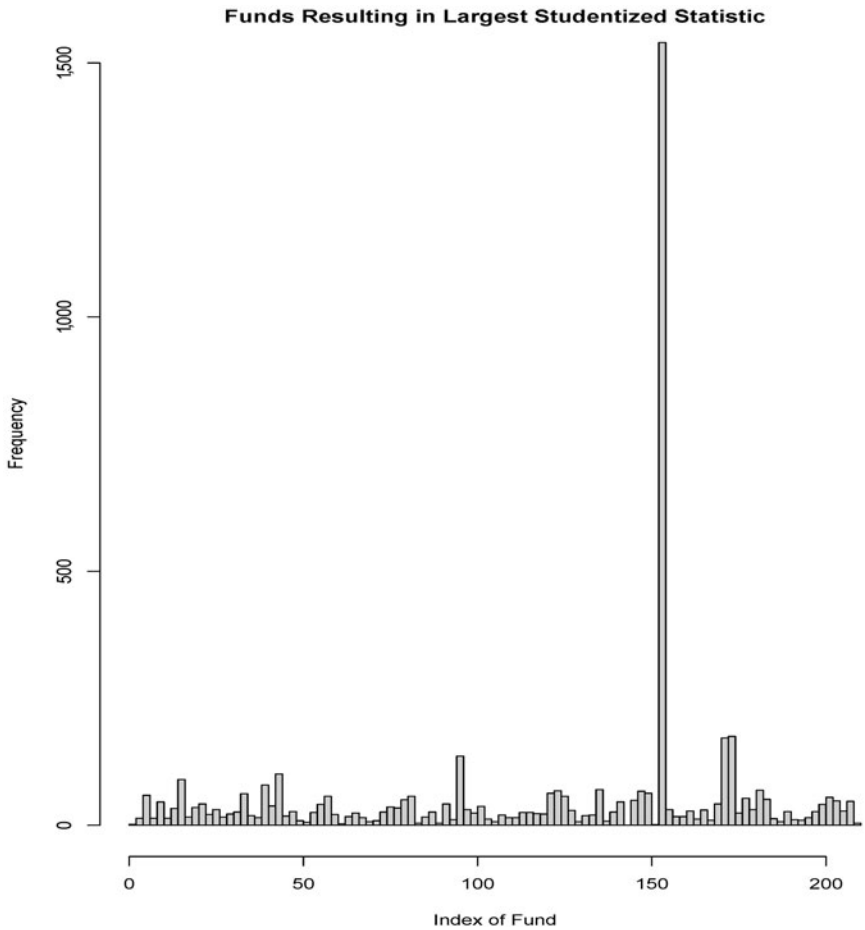


FIGURE 4. Histogram of the fund index that corresponds to the largest studentized statistic z^*_{120,r_1} in $M = 5{,}000$ bootstrap repetitions. Fund 154, Paradigm Master Fund, corresponds to the largest studentized statistic disproportionally often.

Remark 9.1. The number of respective rejections of the augmentation methods of van der Laan et al. (2004) easily can be computed from the algorithms described in Sections 4.3 and 6.2. For example, if the StepM method is used as the initial procedure to control the 1-FWE, then their method to control the 3-FWE at level $\alpha = 0.1$ results in 19 rejections (as opposed to our 33 rejections). And their method to control the FDP with $\gamma = 0.1$ at level $\alpha = 0.1$ results in 17 rejections (as opposed to our 36 rejections).

TABLE 2. The 10 largest basic and studentized test statistics, together with the corresponding hedge funds, in our empirical application. The return unit is 1%.

$w_{T,s}$	Fund	$z_{T,s}$	Fund
1.70	Caduceus Capital	13.65	Coast Enhanced Income
1.67	Libra Fund	9.74	Market Neutral Median
1.48	FBR Weston	8.64	Univest (B)
1.37	Needham Emerging Growth	8.06	JMG Capital Partners
1.34	Westcliff Hedged Strategy	7.77	Market Neutral Long/Short Median
1.31	Spinner Global Technology	6.32	Arden Advisers
1.24	FBR Ashton	6.29	Millennium Partners
1.23	JMG Capital Partners	6.18	Black Diamond Partners
1.21	Bricoleur Partners	6.03	Gabelli Associates
1.20	Emerging Value Opportunities	5.53	Arden International Capital

9.2. Multiple Regression

In empirical work, it is quite common to estimate large-dimensional regression models and to then ask which are the "important" variables. The habitual practice is to assess importance via the individual *t*-statistics or, equivalently, via the individual *p*-values without taking into account the multitude of tests. Consequently, as discussed earlier, typically too many variables will be identified as important.

As an example, we consider a Mincer regression where log wages are regressed on a large number of explanatory variables. The data consist of a random sample of $T = 4,975$ people from the Austrian Social Security database on 08/10/2001. The explanatory variables include a dummy for gender, a dummy for blue collar (vs. white collar), age, age squared, work experience,

TABLE 3. Number of outperforming funds identified

Procedure	$\alpha = 0.05$	$\alpha = 0.1$	α not defined	Procedure	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.5$
Holm	10	13		StepM	11	16	
2-gH	13	20		2-StepM	17	29	
3-gH	16	22		3-StepM	29	33	
4-gH	20	24		4-StepM	29	36	
FDP-$LR_{0.1}$	13	22		FDP-$StepM_{0.1}$	17	36	
FDR-$BH_{0.1}$			101	FDP-$StepM_{0.1}^{Med}$			127
Naive	102	130		Naive	102	130	

TABLE 4. Number of important variables identified

Procedure	$\alpha = 0.05$	$\alpha = 0.1$	α not defined	Procedure	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.5$
Holm	0	0		StepM	5	6	
2-gH	0	9		2-StepM	7	8	
3-gH	9	11		3-StepM	10	11	
5-gH	9	11		5-StepM	12	12	
10-gH	11	14		10-StepM	15	17	
FDP-$LR_{0.1}$	0	0		FDP-$StepM_{0.1}$	5	6	
$FDR_{0.1}$			16	FDP-$StepM_{0.1}^{Med}$			12
Naive	23	33		Naive	23	33	

work experience squared, time at current company, time at current company squared, state dummies, industry dummies, and state-industry interaction dummies, in addition to an intercept. The total number of explanatory variables is $S = 291$.

We now use the various multiple testing methods to identify the important variables, starting with the Holm procedure and its generalizations and also the FDR procedure of Benjamini and Hochberg (1995), all of which are based on individual p-values only. The p-values are obtained by the wild bootstrap to account for possible heteroskedasticity. To generate the resampled errors, we use the two-point distribution; see (6.21) in Davison and Hinkley (1997). Standard errors both in the real world and in the bootstrap world are computed via the well-known White estimator. The White estimator uses the modified residuals rather than the raw residuals because the former have equal variance; see page 271 in Davison and Hinkley (1997). The results are displayed in the left half of Table 4.

Next, we turn to the studentized StepM method and its generalizations. The k-StepM procedures are based on the operative method using $N_{max} = 100$; see Remark 4.1. The results are displayed in the right half of Table 4. The findings, in terms of comparing the various error rates, are similar to those of Sect. 9.1.

10. CONCLUSIONS

The problem of testing multiple hypotheses is ubiquitous in econometric applications. Unfortunately, this problem very often simply is ignored. As a result, too many true null hypotheses will be rejected. The classical approach to account for the multitude of hypotheses under test is to control the familywise error rate (FWE), defined as the probability of falsely rejecting even one true hypothesis. But when the number of hypotheses is very large, this criterion can become too stringent. As a result, potentially very few false hypotheses will be rejected.

This paper has reviewed various generalized error rates. They are more liberal than the FWE yet still account for the multitude of tests by allowing for a small number or a small (expected) proportion of true hypotheses among all rejected hypotheses. Some simulations and two empirical applications have demonstrated that in this way many more false hypotheses can be rejected compared to control of the FWE.

As a special emphasis, we have presented some very recent multiple testing procedures that implicitly account for the dependence structure of the individual test statistics via an application of the bootstrap. The advantage over traditional multiple testing procedures based on individual p-values alone is that the number of false hypotheses rejected often increases, whereas the control of the generalized error rates is not sacrificed. This advantage has also been highlighted via simulations and two empirical applications. The disadvantage is the increased computational cost, but because of the availability of fast computers this is less and less of a concern.

We have discussed further how the control of generalized error rates can apply to various notions of model selection.

NOTES

1. We use the compact terminology of *false rejection* to denote the rejection of a true null hypothesis. Similarly, the terminology *true rejection* denotes the rejection of a false null hypothesis. A false rejection is sometimes termed a *false discovery*.

2. The 1-FWE is simply the usual FWE.

3. The definition of a Sharpe ratio is often based on returns in excess of the risk-free rate. But for certain applications, such as long-short investment strategies, it can be more suitable to base it on the nominal returns.

4. We trust that there is no possible confusion between a CAPM alpha α_s and the level α of multiple testing methods.

5. To show its dependence on P, we may write $\text{FWE} = \text{FWE}_P$.

6. By "power" we mean loosely speaking the ability to detect false hypotheses. Of course, several specific notions exist, such as the probability of rejecting at least one false hypothesis or the probability of rejecting all false hypotheses. In the remainder of the paper, we will mean by "power" the expected number of false hypotheses rejected, which is equivalent to the concept of average power.

7. Equivalently, it addresses the question of whether there are any strategies at all that beat the benchmark.

8. The α_j depend also on S and k, but this dependence is suppressed in the notation.

9. This region could also be called a *generalized confidence region* in that we do not seek to contain all the parameters with probability $1 - \alpha$, but instead seek to contain all, except at most $k - 1$ of them, with probability $1 - \alpha$.

10. Usually, one can take $\hat{\theta}_T = \theta(\hat{P}_T)$.

11. If a true hypothesis has been rejected so far, then the FWE criterion has already been violated, and, therefore, the rejection of further true hypotheses will not do any additional harm.

12. Obviously, some very crude bounds could be obtained using Markov's inequality or variants thereof.

13. For example, when forecasting inflation, a suitable, simple-minded benchmark might be last period's inflation.

14. For example, it is common to provide tables where the important explanatory variables are identified via asterisks; one asterisk if significant at level 10%, two asterisks if significant at level 5%, and three asterisks if significant at level 1%, where the levels are for individual tests always.

15. Here optimality is defined in an asymptotic minimax sense; see Abramovich et al. (2005) for details.

16. For instance, they assume that the sparsity tends to zero, and so the limiting model for the vector is that of a "black object" (where all entries are equal to zero).

17. The risk-free rate is a simple and widely accepted benchmark. But, of course, our methods also apply to alternative benchmarks such as hedge fund indices or multifactor hedge fund benchmarks; for example, see Kosowski, Naik, and Teo (2005).

18. To account for leftover dependence not captured by the VAR(1) model, we use the stationary bootstrap with average block size $b = 5$ for bootstrapping the residuals.

19. To account for leftover dependence not captured by the AR(1) model, we use the stationary bootstrap with average block size $b = 5$ for bootstrapping the residuals.

20. Similar to the analysis of Romano and Wolf (2005b), the basic StepM method does not detect a single outperforming fund, so it is not pursued further.

21. To account for leftover dependence not captured by the VAR(1) model, we use the stationary bootstrap with average block size $b = 5$ for bootstrapping the residuals.

REFERENCES

Abramovich, F. & Y. Benjamini (1996) Adaptive thresholding of wavelet coefficients. *Computational Statistics & Data Analysis* 22, 351–361.

Abramovich, F., Y. Benjamini, D.L. Donoho, & I.M. Johnstone (2005) Adapting to Unknown Sparsity by Controlling the False Discovery Rate. *Annals of Statistics*, forthcoming. Working paper available at http://arxiv.org/PS-cache/math/pdf/0505/0505374.pdf.

Andrews, D.W.K. & J.C. Monahan (1992) An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator. *Econometrica* 60, 953–966.

Bauer, P., B.M. Pötscher, & P. Hackl (1988) Model selection by multiple test procedures. *Statistics* 19, 39–44.

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

Benjamini, Y. & Y. Hochberg (2000) On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics* 25, 60–83.

Benjamini, Y. & D. Yekutieli (2001) The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics* 29, 1165–1188.

Buena, F., M.H. Wegkamp, & A. Auguste (2006) Consistent variable selection in high dimensional regression via multiple testing. *Journal of Statistical Planning and Inference*, forthcoming.

Campos, J., N.R. Ericsson, & D.F. Hendry (2005) *General-to-Specific Modelling*. Edward Elgar.

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

Draper, N.R. & H. Smith (1998) *Applied Regression Analysis*, 3rd ed. Wiley.

Dudoit, S., J.P. Shaffer, & J.C. Boldrick (2003) Multiple hypothesis testing in microarray experiments. *Statistical Science* 18, 71–103.

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

Dudoit, S., M.J. van der Laan, & K.S. Pollard (2004b) Multiple testing, part III: Procedures for control of the generalized family-wise error rate and proportion of false positives. Working paper 171, U.C. Berkeley Division of Biostatistics. Available at http://www.bepress.com/ucbbiostat/paper171/.

Efron, B. (1979) Bootstrap methods: Another look at the jackknife. *Annals of Statistics* 7, 1–26.

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

Götze, F. & H.R. Künsch (1996) Second order correctness of the blockwise bootstrap for stationary observations. *Annals of Statistics* 24, 1914–1933.

Hansen, P.R. (2005) A test for superior predictive ability. *Journal of Business & Economics Statistics* 23, 365–380.

Hansen, P.R., A. Lunde, & J.M. Nason (2003) Choosing the best volatility models: The model confidence set approach. *Oxford Bulletin of Economics and Statistics* 65, 839–861.

Hansen, P.R., A. Lunde, & J.M. Nason (2005) Model Confidence Sets for Forecasting Models. Working paper 2005-7, Federal Reserve Bank of Atlanta. Available at http://ssrn.com/abstract=522382.

Hastie, T.J., R. Tibshirani, & J.H. Friedman (2001) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

Hochberg, Y. & A. Tamhane (1987) *Multiple Comparison Procedures*. Wiley.

Holm, S. (1979) A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6, 65–70.

Hommel, G. & T. Hoffman (1988) Controlled uncertainty. In P. Bauer, G. Hommel, & E. Sonnemann (eds.), *Multiple Hypothesis Testing*, pp. 154–161. Springer.

Jensen, D.D. & P.R. Cohen (2000) Multiple comparisons in induction algorithms. *Machine Learning* 38, 309–338.

Kabaila, P. & H. Leeb (2006) On the large-sample minimal coverage probability of confidence intervals after model selection. *Journal of the American Statistical Association* 101, 619–629.

Kat, H.M. (2003) 10 Things Investors Should Know about Hedge Funds. AIRC Working paper 0015, Cass Business School, City University. Available at http://www.cass.city.ac.uk/airc/papers.html.

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

Kosowski, R., N.Y. Naik, & M. Teo (2005) Is Stellar Hedge Fund Performance for Real? Working paper HF-018, Centre for Hedge Fund Research and Education, London Business School.

Krolzig, H.-M. & D.F. Hendry (2001) Computer automation of general-to-specific selection procedures. *Journal of Economic Dynamics & Control* 25, 831–866.

Lahiri, S.N. (1992) Edgeworth correction by "moving block" bootstrap for stationary and nonstationary data. In R. LePage & L. Billard (eds.), *Exploring the Limits of Bootstrap*, pp. 183–214. Wiley.

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

Lo, A.W. (2002) The statistics of Sharpe ratios. *Financial Analysts Journal* 58, 36–52.

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

Pollard, K.S. & M.J. van der Laan (2003b) Resampling-Based Multiple Testing: Asymptotic Control of Type i Error and Applications to Gene Expression Data. Working paper 121, U.C. Berkeley Division of Biostatistics. Available at http://www.bepress.com/ucbbiostat/paper121/.

Pötscher, B.M. (1983) Order estimation in ARMA models by Lagrange multiplier tests. *Annals of Statistics* 11, 872–885.

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

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

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

Romano, J.P. & M. Wolf (2005b) Stepwise multiple testing as formalized data snooping. *Econometrica* 73, 1237–1282.

Romano, J.P. & M. Wolf (2006) Improved nonparametric confidence intervals in time series regressions. *Journal of Nonparametric Statistics* 18, 199–214.

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

Shen, X., H. Huang, & J. Ye (2004) Inference after model selection. *Journal of the American Statistical Association* 99, 751–762.

Shimodaira, H. (1998) An application of multiple comparison techniques to model selection. *Annals of the Institute of Statistical Mathematics* 50, 1–13.

Storey, J.D. (2002) A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B* 64, 479–498.

Storey, J.D. (2003) The positive false discovery rate: A Bayesian interpretation and the *q*-value. *Annals of Statistics* 31, 2013–2035.

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

Sullivan, R., A. Timmermann, & H.L. White (1999) Data snooping, technical trading rule performance, and the bootstrap. *Journal of Finance* 54, 1647–1692.

Sullivan, R., H.L. White, & B. Golomb (2001) Dangers of data mining: The case of calendar effects in stock returns. *Journal of Econometrics* 105, 249–286.

Timmermann, A. (2006) Forecast combinations. In G. Elliott, C.W.J. Granger, & A. Timmermann (eds.), *Handbook of Economic Forecasting*, vol. 1, pp. 135–196. North-Holland.

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

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

van der Laan, M.J. & A.E. Hubbard (2005) Quantile-Function Based Null Distributions in Resampling Based Multiple Testing. Working paper 198, U.C. Berkeley Division of Biostatistics. Available at http://www.bepress.com/ucbbiostat/paper198/.

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

White, H.L. (2000) A reality check for data snooping. *Econometrica* 68, 1097–1126.

White, H.L. (2001) *Asymptotic Theory for Econometricians*, rev. ed. Academic Press.

APPENDIX A: Use of the Bootstrap

This Appendix details how to compute the constants $\hat{c}_j$, $\hat{c}_{j,|\cdot|}$, $\hat{d}_j$, and $\hat{d}_{j,|\cdot|}$ in Algorithms 4.1, 4.2, 4.3, and 4.4, respectively, via the bootstrap. At first, a proper choice of the estimator $\hat{P}_T$ of the underlying probability mechanism P must be made. (One can implicitly define $\hat{P}_T$ by describing how a bootstrap data matrix X_T^* is generated from $\hat{P}_T$.) This choice depends on the context. If the data $X_{1,\cdot}^{(T)}, \ldots, X_{T,\cdot}^{(T)}$ are i.i.d., one should choose the Efron (1979) bootstrap; if they constitute a time series, one should choose the moving blocks bootstrap, the circular blocks bootstrap, or the stationary bootstrap. These various

bootstrap methods are detailed in Appendix B of Romano and Wolf (2005b). In any case, we use the notation $\hat{\theta}_T$ for a suitable parameter vector corresponding to the bootstrap law.

ALGORITHM A.1 (Computation of the $\hat{c}_j$ via the bootstrap).

1. *The labels $r_1, \ldots, r_S$ and the numerical values of $R_1, R_2, \ldots$ are given in Algorithm 4.1.*
2. *Generate M bootstrap data matrices $X_T^{*,1}, \ldots, X_T^{*,M}$. (One should use $M \geq 1{,}000$ in practice.)*
3. *From each bootstrap data matrix $X_T^{*,m}$, $1 \leq m \leq M$, compute the individual test statistics $w_{T,1}^{*,m}, \ldots, w_{T,S}^{*,m}$.*
4. *(a) For $1 \leq m \leq M$, and any needed K, compute $kmax_{T,K}^{*,m} = k\text{-max}_{s \in K}(w_{T,r_s}^{*,m} - \hat{\theta}_{T,r_s})$.*
   (b) Compute $c_K(1 - \alpha, k, \hat{P}_T)$ as the $1 - \alpha$ empirical quantile of the M values $kmax_{T,K}^{*,1}, \ldots, kmax_{T,K}^{*,M}$.
5. *If $j = 1$, $\hat{c}_1 = c_{\{1,\ldots,S\}}(1 - \alpha, k, \hat{P}_T)$.*
   If $j > 1$, $\hat{c}_j = \max\{c_K(1 - \alpha, k, \hat{P}_T) : K = I \cup \{R_{j-1} + 1, \ldots, S\},\ I \subset \{1, \ldots, R_{j-1}\}, |I| = k - 1\}$.

ALGORITHM A.2 (Computation of the $\hat{c}_{j,|\cdot|}$ via the bootstrap).

1. *The labels $r_1, \ldots, r_S$ and the numerical values of $R_1, R_2, \ldots$ are given in Algorithm 4.2.*
2. *Generate M bootstrap data matrices $X_T^{*,1}, \ldots, X_T^{*,M}$. (One should use $M \geq 1{,}000$ in practice.)*
3. *From each bootstrap data matrix $X_T^{*,m}$, $1 \leq m \leq M$, compute the individual test statistics $w_{T,1}^{*,m}, \ldots, w_{T,S}^{*,m}$.*
4. *(a) For $1 \leq m \leq M$, and any needed K, compute $kmax_{T,K,|\cdot|}^{*,m} = k\text{-max}_{s \in K}|w_{T,r_s}^{*,m} - \hat{\theta}_{T,r_s}|$.*
   (b) Compute $c_{K,|\cdot|}(1 - \alpha, k, \hat{P}_T)$ as the $1 - \alpha$ empirical quantile of the M values $kmax_{T,K,|\cdot|}^{*,1}, \ldots, kmax_{T,K,|\cdot|}^{*,M}$.
5. *If $j = 1$, $\hat{c}_{1,|\cdot|} = c_{\{1,\ldots,S\},|\cdot|}(1 - \alpha, k, \hat{P}_T)$.*
   If $j > 1$, $\hat{c}_{j,|\cdot|} = \max\{c_{K,|\cdot|}(1 - \alpha, k, \hat{P}_T) : K = I \cup \{R_{j-1} + 1, \ldots, S\},\ I \subset \{1, \ldots, R_{j-1}\}, |I| = k - 1\}$.

ALGORITHM A.3 (Computation of the $\hat{d}_j$ via the bootstrap).

1. *The labels $r_1, \ldots, r_S$ and the numerical values of $R_1, R_2, \ldots$ are given in Algorithm 4.3.*
2. *Generate M bootstrap data matrices $X_T^{*,1}, \ldots, X_T^{*,M}$. (One should use $M \geq 1{,}000$ in practice.)*
3. *From each bootstrap data matrix $X_T^{*,m}$, $1 \leq m \leq M$, compute the individual test statistics $w_{T,1}^{*,m}, \ldots, w_{T,S}^{*,m}$. Also, compute the corresponding standard errors $\hat{\sigma}_{T,1}^{*,m}, \ldots, \hat{\sigma}_{T,S}^{*,m}$.*
4. *(a) For $1 \leq m \leq M$, and any needed K, compute $kmax_{T,K}^{*,m} = k\text{-max}_{s \in K}([w_{T,r_s}^{*,m} - \hat{\theta}_{T,r_s}]/\hat{\sigma}_{T,r_s}^{*,m})$.*
   (b) Compute $d_K(1 - \alpha, k, \hat{P}_T)$ as the $1 - \alpha$ empirical quantile of the M values $kmax_{T,K}^{*,1}, \ldots, kmax_{T,K}^{*,M}$.
5. *If $j = 1$, $\hat{d}_1 = d_{\{1,\ldots,S\}}(1 - \alpha, k, \hat{P}_T)$.*
   If $j > 1$, $\hat{d}_j = \max\{d_K(1 - \alpha, k, \hat{P}_T) : K = I \cup \{R_{j-1} + 1, \ldots, S\},\ I \subset \{1, \ldots, R_{j-1}\}, |I| = k - 1\}$.

ALGORITHM A.4. (Computation of the $\hat{d}_{j,|\cdot|}$ via the bootstrap).

1. *The labels $r_1, \ldots, r_S$ and the numerical values of $R_1, R_2, \ldots$ are given in Algorithm 4.4.*
2. *Generate M bootstrap data matrices $X_T^{*,1}, \ldots, X_T^{*,M}$. (One should use $M \geq 1{,}000$ in practice.)*
3. *From each bootstrap data matrix $X_T^{*,m}$, $1 \leq m \leq M$, compute the individual test statistics $w_{T,1}^{*,m}, \ldots, w_{T,S}^{*,m}$. Also, compute the corresponding standard errors $\hat{\sigma}_{T,1}^{*,m}, \ldots, \hat{\sigma}_{T,S}^{*,m}$.*
4. *(a) For $1 \leq m \leq M$, and any needed K, compute $kmax_{T,K,|\cdot|}^{*,m} = k\text{-max}_{s \in K}(|w_{T,r_s}^{*,m} - \hat{\theta}_{T,r_s}|/\hat{\sigma}_{T,r_s}^{*,m})$.*
   (b) Compute $d_{K,|\cdot|}(1-\alpha, k, \hat{P}_T)$ as the $1-\alpha$ empirical quantile of the M values $kmax_{T,K,|\cdot|}^{*,1}, \ldots, kmax_{T,K,|\cdot|}^{*,M}$.
5. *If $j = 1$, $\hat{d}_{1,|\cdot|} = d_{\{1,\ldots,S\},|\cdot|}(1-\alpha, k, \hat{P}_T)$.*
   If $j > 1$, $\hat{d}_{j,|\cdot|} = \max\{d_{K,|\cdot|}(1-\alpha, k, \hat{P}_T) : K = I \cup \{R_{j-1}+1, \ldots, S\},\ I \subset \{1, \ldots, R_{j-1}\}, |I| = k-1\}$.

Remark A.1. For convenience, one can typically use w_{T,r_k} in place of $\hat{\theta}_{T,r_k}$ in step 4(a) of Algorithms A.1–A.4. Indeed, the two quantities are the same under the following conditions: (1) $w_{T,k}$ is a linear statistic; (2) $\theta_k = \mathrm{E}(w_{T,k})$; and (3) $\hat{P}_T$ is based on Efron's bootstrap, the circular blocks bootstrap, or the stationary bootstrap. Even if conditions (1) and (2) are met, w_{T,r_k} and $\hat{\theta}_{T,r_k}$ are not the same if $\hat{P}_T$ is based on the moving blocks bootstrap because of "edge" effects; see Appendix B of Romano and Wolf (2005b). On the other hand, the substitution of w_{T,r_k} for $\hat{\theta}_{T,r_k}$ does not affect in general the consistency of the bootstrap approximation. Lahiri (1992) discusses this subtle point for the special case of time series data and w_{T,r_k} being the sample mean. He shows that centering by $\hat{\theta}_{T,r_k}$ provides second-order refinements but is not necessary for first-order consistency.

When a time series bootstrap is used, then the choice of the (average) block size becomes an important practical problem. The method we propose here to choose a block size for an application of the *k*-StepM procedure is a generalization of Algorithm 7.1 of Romano and Wolf (2005b), who only deal with the StepM procedure.

Consider the first step of the *k*-StepM procedure. The goal is to construct a generalized joint confidence region for the parameter vector θ with nominal coverage probability of $1-\alpha$. Here, importantly, "coverage probability" stands for the probability of containing at least $S-k+1$ elements of θ.

ALGORITHM A.5 (Choice of block size).

1. *Fit a semiparametric model $\tilde{P}_T$ to the observed data X_T.*
2. *Fix a selection of reasonable block sizes b.*
3. *Generate M data sets $\tilde{X}_T^1, \ldots, \tilde{X}_T^M$ according to $\tilde{P}_T$.*
4. *For each data set $\tilde{X}_T^m$, $m = 1, \ldots, M$, and for each block size b, compute a generalized joint confidence region $GJCR_{m,b}$ for θ.*
5. *Compute $\hat{g}(b) = \#\{\text{At least } S-k+1 \text{ elements of } \theta(\tilde{P}_T) \in GJCR_{m,b}\}/M$.*
6. *Find the value of $\tilde{b}$ that minimizes $|\hat{g}(b) - (1-\alpha)|$ and use this value $\tilde{b}$.*

Algorithm A.5 is based on the first step of the k-StepM method. Because the general k-StepM method, for $k > 1$, does not discard any hypotheses in subsequent steps—in contrast to the StepM method—we recommend continuing to use the chosen value $\tilde{b}$ throughout. If, on the other hand, the operative method of Remark 4.1 is used, then at a given subsequent step some hypotheses may already have been discarded. In that case, one can apply Algorithm A.5 to the subset of θ that corresponds to the nondiscarded hypotheses.