Improved stepdown methods for asymptotic control of generalized error rates

Gray Calhoun*

2015-04-27

Abstract

This paper proposes new stepdown methods for testing multiple hypotheses and constructing confidence intervals while controlling the Familywise Error Rate and other generalized error rates. One method is a refinement of Romano and Wolf's StepM (2005, *Econometrica*) that also removes inequalities that fall outside any $n^{-1/2}$-neighborhood of binding; it has the advantage that the threshold construction is incorporated into the stepdown procedure so it accounts for the number of total hypotheses (leading to better size control than some alternative methods) and excludes more nonbinding inequalities (leading to higher power). This method can also be used to test multiple inequality hypotheses simultaneously and construct confidence intervals for partially identified parameters. The paper presents methods for controlling the k-familywise error rate and the False Discovery Proportion for families of one and two-sided hypotheses as well. The paper also provides Monte Carlo evidence that the methods perform well in finite samples and demonstrates their application in an empirical analysis of hedge fund returns.

JEL Classification Codes: C12, C52

Keywords: Multiple testing, bootstrap, Familywise Error Rate, False Discovery Proportion, moment inequalities

*Iowa State University, email: «gcalhoun@iastate.edu», web: «http://gray.clhn.org». This paper is still a work in progress. Please let me know if you find errors, either by email or by opening a new issue at «https://git.ece.iastate.edu/gcalhoun/stepdown-paper/issues». I would like to thank (without implication) Helle Bunzel, Jianqing Fan, Brent Kreider, Joseph Romano, Elie Tamer, Michael Wolf, participants of the Annual Meetings of the Midwest Econometrics Group, Joint Statistical Meetings, and NBER Summer Institute workshop on Forecasting and Empirical Methods in Macroeconomics and Finance, and especially the anonymous referees who handled this paper for helpful suggestions and comments.

1 Introduction

This paper develops improvements to sequential procedures for testing multiple hypotheses. Many existing procedures can lose power when some of the individual null hypotheses hold with parameter values that are far from the alternative—the multiplicity correction is then unnecessarily large and decreases the procedure's power to reject other, false, hypotheses. The canonical example of this problem is testing many one-sided hypotheses (see Hansen, 2005, as well as Andrews, 2012, and Hirano and Porter, 2012, for recent assessments of these issues), but our main contribution is for statistics that control other generalized error rates—the k-Familywise Error Rate (k-FWE) and the False Discovery Proportion (FDP)—and we present settings where this issue also arises for two-sided hypotheses.

For concreteness, suppose that each hypothesis s is of the form $\theta_s \in \Theta_0$ vs. $\theta_s \in \Theta_a$ and has corresponding test statistic T_s; T_s rejects if it is above some critical value q and the procedure controls the *familywise error rate* (FWE) at level α if

$$\Pr[T_s > q \text{ for at least one } s \text{ such that } \theta_s \in \Theta_0] \leq \alpha \tag{1}$$

(we focus on the FWE in this example for simplicity but will present results for other error rates later in the paper; see Section 3). FWE control is stronger than control of the size of the composite hypothesis $\theta_s \in \Theta_0$ for all s, since it must hold for any arrangement of θ_s. This stronger concept is essential if a researcher wants to interpret individual rejections ($T_s > q$) as evidence against the individual hypotheses ($\theta_s \in \Theta_0$).

"Single-step" procedures construct a critical value q_1 to control FWE at α and then reject all of the individual hypotheses with $T_s > q_1$. A sequential procedure (as in Holm, 1979) continues from there by constructing a second critical value q_2 to control FWE at α for the family of hypotheses left after the first step, $\{s : T_s \leq q_1\}$, and then rejects all of the hypotheses with $T_s > q_2$. The sequential procedure then continues in the same way, constructing each critical value to control FWE over the remaining hypotheses, until it stops rejecting at, say, the jth step. Somewhat surprisingly, using $q = q_j$ in (1) typically controls the FWE at α. (Subject to natural restrictions on the test procedure, of course; see Holm, 1979, Romano and Wolf, 2005a,b, and Goeman and Solari, 2010, among others.)

This paper's method works by identifying subsets of the null, $\Theta' \subset \Theta_0$, where the probability that T_s rejects if $\theta \in \Theta'$ is negligible. Instead of testing $\theta_s \in \Theta_0$ against $\theta_s \in \Theta_a$

at level α in each step, we also simultaneously test $\theta_s \in \Theta_0 \setminus \Theta'$ vs. $\theta_s \in \Theta'$ at level ϵ (an arbitrarily small positive quantity).We remove hypothesis s if either test rejects and then proceed sequentially. As ϵ converges to zero, the FWE of this procedure converges to α. Of course, at the end we only reject those hypotheses that were determined to be in Θ_a (i.e. T_s is greater than the last q_j), but removing the additional hypotheses in Θ' during the sequential process can increase the method's power, sometimes dramatically, resulting in more rejections. Each step of our procedure is similar to the Bonferroni correction proposed by McCloskey (2012) and Romano et al. (2012), but extending these results to $\epsilon \approx 0$ can be very useful in practice, especially for more complicated error measures.

Section 2 uses this principle to improve Romano and Wolf's (2005a) *StepM* procedure and increase its power for families of one-sided hypotheses. The StepM, like White's (2000) *Bootstrap Reality Check* (BRC) and Hansen's (2005) test of *Superior Predictive Ability* (SPA), uses the bootstrap to approximate the joint distribution of the test statistics for each hypothesis, and so obtains higher power than methods that assume a worst-case dependence structure (Holm, 1979, for example) and more general validity than those that assume a convenient dependence structure. Romano and Wolf (2005a) improve on White (2000) and Hansen (2005) by incorporating an iterative stepdown method as described above; White (2000) and Hansen (2005) propose single step procedures. Our refinement amounts to using a heavily asymmetric two-sided version of the StepM and removing hypotheses far from the boundary between the null and the alternative in either direction, but then, after the sequential procedure stops, only rejecting the hypotheses that violate the null. This refinement is similar to existing procedures—Hansen (2005) proposes discarding the hypotheses with corresponding t-statistics below $-\sqrt{2 \log \log n}$ before using the BRC for the null hypotheses $\theta_s \leq 0$, a threshold motivated by the Law of the Iterated Logarithm, and Hsu et al. (2010) propose the same procedure for the StepM—but our threshold accounts for the number of hypotheses, giving it better size control in finite samples. Simulations presented in Section 4 show that Hansen's (2005) and Hsu et al.'s (2010) test can overreject in practice.

Section 2 also shows how to apply this procedure to test composite null hypotheses with several inequality restrictions (Corollary 1), and how to apply that result to the partial identification problem considered by Imbens and Manski (2004) (Remark 6). More simulations presented in Section 4 show that our procedure has roughly equal power to Andrews and Barwick's (2012a) preferred statistic (their AQLR) and to McCloskey's

(2012) and Romano et al.'s (2012) procedures to detect at least one violation of the inequalities. As mentioned earlier, our procedure (as well as McCloskey's, 2012, and Romano et al.'s, 2012) has the advantage of also controlling FWE, so the individual rejections can be taken as evidence against the individual hypotheses—in contrast, Andrews and Barwick's (2012a) AQLR only tells the researcher that one or more of the inequalities does not hold, but not which one. Our method has the further advantage that it will typically reject more of the individual false hypotheses than McCloskey's (2012) and Romano et al.'s (2012), even though the probabilities of rejecting the composite null hypothesis are roughly equal.

Section 3 applies the same concepts to procedures that control other generalized error rates, namely the k-FWE and the FDP. These error rates can be used when FWE is too demanding a measure to be useful. In such situations, the researcher may be willing to allow for a few false rejections (k-FWE), or allow for a known percentage of the total rejections to be false (FDP, but see Section 3 for formal definitions of these terms). We show that the same ideas apply as before and present refinements to Romano and Wolf's (2007) k-StepM—a sequential procedure designed to control these error rates—that can require substantially fewer calculations while maintaining uniform control of their error rates. In addition to corrections for one-sided testing, we also present new restrictions that are implied by the error rates themselves and apply to families of two-sided tests as well.

Sections 2 and 3 lay out our theory as described above. Section 4 presents Monte Carlo simulations studying the behavior of our procedure and several competing methods in finite samples. Section 5 uses our new statistics to study hedge fund performance from 1994 to 2012, and Section 6 concludes.

2 Testing families of one-sided hypotheses with FWE control

Consider the following environment. Suppose that there are S null hypotheses $H_s : \theta_s \leq 0$ against the alternatives $H_s' : \theta_s > 0$, let $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_S)$ be an estimator of θ, and let $I = \{s \in \{1, \ldots, S\} : \theta_s \leq 0\}$ index the true null hypotheses. A critical value q that

controls FWE at level α satisfies

$$\Pr[\hat{\theta}_s > q \text{ for at least one } s \in I] \leq \alpha. \tag{2}$$

(We will deal with studentized statistics in the actual results, but (2) is presented with unstudentized statistics for simplicity.) This is a more stringent criterion than controlling the probability that (2) holds only when $I = \{1, \ldots, S\}$, which would be the focus if this were a test of the composite null hypothesis.[1]

We will derive our results under the following high-level assumption.

Assumption 1. *For any sequence of parameter values* $\{\theta_n\}$, $\sqrt{n}(\hat{\theta} - \theta_n) \to^d N(0, V)$ *where* V *is positive semi-definite with uniformly positive diagonal elements and* $(\hat{v}_1^2, \ldots, \hat{v}_S^2)' \to^p \text{diag}(V)$. *Moreover, let* $\hat{F}_n$ *be an estimator of the distribution of*

$$(\sqrt{n}(\hat{\theta}_1 - \theta_{1n})/\hat{v}_1, \ldots \sqrt{n}(\hat{\theta}_S - \theta_{Sn})/\hat{v}_S))$$

such that $\hat{F}_n \to^d N(0, W)$, *where* W *is the asymptotic correlation matrix of* $\sqrt{n}\hat{\theta}$.

In the main text of the paper, we suppress the dependence of θ on n to simplify the notation, but we will make that dependence explicit in the proofs. Typically the distribution $\hat{F}$ can be estimated with the bootstrap, and it will be useful to define a random vector $\hat{\psi}^*$ that is distributed as $\hat{F}_n$. We assume asymptotic normality to simplify the presentation and the proofs, but it is not essential. Moreover, we work with studentized statistics to improve the procedure's performance (see Hansen, 2005, and Romano and Wolf, 2005a,b, 2010, among many others) but that assumption can be relaxed.

Algorithm 1 presents our approach, a variation of the StepM, for generating q. Theorem 1 then shows that this value of q controls the FWE in the sense of (2).

Algorithm 1 (StepM variation for one-sided tests)**.** *Set* $M_0 = \{1, \ldots, S\}$, $\alpha \in (0, 1)$, *and* $\epsilon \in (0, \alpha)$. *Repeat the listed steps for each* $j = 1, 2, \ldots$ *and stop when* $M_j = M_{j-1}$ *or* $M_j = \emptyset$.

1. *Set* p_j *to be the* ϵ *quantile of the distribution of* $\min_{s \in M_{j-1}} \hat{\psi}_s^*$.
2. *Set* q_j *to be the* $1 - \alpha$ *quantile of the distribution of* $\max_{s \in M_{j-1}} \hat{\psi}_s^*$.
3. *Set* $M_j = \{s : \sqrt{n}\,\hat{\theta}_s/\hat{v}_s \in [p_j, q_j]\}$.

[1] This less stringent criterion is often referred to as *weak control of the familywise error rate*.

Let q be the last q_j when the algorithm stops.

Remark 1. An informal description of the algorithm may be useful. This procedure begins by testing the *wrong* null hypotheses, $\theta_s \geq 0$, at an arbitrarily small level ϵ. The critical value for this test is p_1 and this critical value accounts for multiplicity—when more hypotheses are considered, p_1 will tend to move further away from zero. The hypotheses rejected at this stage, corresponding to statistics with $\sqrt{n}\hat{\theta}_s/\hat{v}_s < p_1$, are not rejected by the algorithm, since the parameter estimates satisfy the correct null hypothesis. But they are so far from the boundary between the null and the alternative that including them further would degrade the power of the test procedure without improving its size, so they are set aside and ignored in future steps.

In the second step, the procedure tests the correct null hypotheses, $\theta_s \leq 0$, at the correct level, α, for the remaining parameters. The critical value for this second test is q_1 and, as before, this critical value accounts for multiplicity.[2] Hypotheses rejected at this stage are rejected by the algorithm and are set aside.

In the third step, the set of parameters under consideration is shrunk: only parameters that were not rejected by either of the first two steps will be considered in the future; the rest are temporarily set aside. If the first two steps had no rejections, the procedure ends without rejecting any hypotheses. If either of the first two steps rejected one or more hypotheses, the procedure continues to try to reject more hypotheses by repeating steps 1 and 2 over the smaller subset of remaining hypotheses.

The algorithm stops when it has stopped removing hypotheses in either step 1 or step 2, or when there are no hypotheses remaining. After stopping, adds all of the parameters removed at any point in step 1 to the set of "accepted" hypotheses, and rejects the hypotheses that were rejected at any point by step 2.

Remark 2. Readers familiar with the StepM may notice that we have omitted several steps related to sorting and reordering the test statistics. Sorting the statistics can be important for efficiently implementing the StepM, but it is not necessary for Romano and Wolf's (2005a) theoretical results or for ours. We have omitted these steps to simplify the presentation here, but interested readers should look to our computer code for more efficient implementations of the procedures described in this paper.

[2] In fact, this critical value *overcorrects* for multiplicity, since it does account for the hypotheses removed in step 1. This is not a problem, though, because the procedure is iterative, and the critical values will be updated in the next iteration.

Theorem 1 establishes that the q produced by Algorithm 1 asymptotically controls FWE when ϵ is small.

Theorem 1 (FWE control for one-sided hypotheses)**.** *Suppose Assumption 1 holds and choose $\alpha \in (0,1)$. For any $\epsilon < \alpha$, let $q(\epsilon)$ denote the last q_j in Algorithm 1. Then*

$$\limsup_{\substack{n\to\infty \\ \epsilon\to 0}} \sup_{\theta\in\mathbb{R}^S} \Pr_\theta[\sqrt{n}\,\hat{\theta}_s/\hat{v}_s > q(\epsilon) \text{ for at least one } s \text{ such that } \theta_s \leq 0] \leq \alpha. \tag{3}$$

Moreover, if ϵ_δ is a sequence of random variables s.t. $\epsilon_\delta \to^p 0$ as $\delta \to 0$ then

$$\limsup_{\substack{n\to\infty \\ \delta\to 0}} \sup_{\theta\in\mathbb{R}^S} \Pr_\theta[\sqrt{n}\,\hat{\theta}_s/\hat{v}_s > q(\epsilon_\delta) \text{ for at least one } s \text{ such that } \theta_s \leq 0] \leq \alpha. \tag{4}$$

Remark 3. To implement this method, we must set ϵ. If the quantiles are estimated with a bootstrap, ϵ can be set arbitrarily small by using the minimum of the bootstrap replications for p_j:

$$p_j = \min_{b=1,\ldots,B} \min_{s\in M_{j-1}} \hat{\psi}^*_{bs}$$

where $\hat{\psi}^*_{bs}$ is the sth element of the vector $\hat{\psi}^*_b$ and $\hat{\psi}^*_1, \ldots, \hat{\psi}^*_B$ are the bootstrap replications. This construction implies that $\epsilon \to^p 0$ as $B \to \infty$ and so $p_j \to -\infty$ in probability (slowly) as $n \to \infty$ and $B \to^p \infty$. (And is why we explicitly allow ϵ to be random.)

This method of setting p_j is used in the empirical section and Monte Carlo simulations later in the paper.

Remark 4. One could use larger values of ϵ by setting each q_j to be the $1-\alpha+\epsilon$ quantile of the distribution of $\max_{s\in M_{j-1}} \hat{\psi}^*_s$. There are advantages to either approach — explicitly allowing ϵ to remain positive in the limit would allow us to derive the value of ϵ that gives the most expected rejections. But using the $1-\alpha+\epsilon$ quantile raises the possibility that, in some applications, the correction for one-sided tests may reject fewer hypotheses than a naive implementation of the uncorrected StepM — i.e., if step 1 of the algorithm does not remove any hypotheses, then testing at $1-\alpha+\epsilon$ in the second step will always have less power than a test that skips the first step and tests at $1-\alpha$ in the second step. Choosing ϵ to be nearly zero removes this risk.

In the more complicated algorithms presented later in the paper, we will also find that choosing ϵ to be nearly zero makes the procedures much simpler. Moreover, we

present our main results for the $1-\alpha$ quantile to emphasize that ϵ should be very small in practice — small enough that the $1-\alpha+\epsilon$ quantile and the $1-\alpha$ quantile are essentially the same. Otherwise our approach can have size distortions in finite samples. We set $\epsilon = 0$ in all of the computations in this paper, which follows the recommendation in Remark 3 because the "0th quantile" of a vector returns its smallest element in many statistical packages, including R (R Development Core Team, 2012), the package used here.

Remark 5. This procedure differs from Romano and Wolf's (2005a) StepM in the p_j term—if we set $p_j = -\infty$ they are the same. This term fills the same role as Hansen's (2005) and Hsu et al.'s (2010) threshold, and if we set $p_j = -\sqrt{2\log\log n}$ our algorithm becomes Hsu et al.'s (2010). Even though Hsu et al.'s (2010) threshold explicitly depends on n and diverges to $-\infty$ as n grows, p_j will typically be substantially farther from zero than $-\sqrt{2\log\log n}$ because it explicitly accounts for the number of hypotheses (and $\sqrt{\log\log n}$ grows very slowly). This has size implications that can cause Hansen's (2005) and Hsu et al.'s (2010) statistics to overreject, as shown in Section 4.

Although the focus of this paper is on testing many individual hypotheses, the algorithm in Theorem 1 also provides an attractive test statistic for joint tests of several inequality restrictions. Corollary 1 formalizes this application.

Corollary 1 (Testing composite one-sided hypotheses)**.** *Under the assumptions of Theorem 1,*

$$\limsup_{\substack{n\to\infty \\ \epsilon\to 0}} \sup_{\theta\leq 0} \Pr_\theta[\max_s \sqrt{n}\,\hat{\theta}_s/\hat{\nu}_s > q(\epsilon)] \leq \alpha \tag{5}$$

and

$$\limsup_{\substack{n\to\infty \\ \delta\to 0}} \sup_{\theta\leq 0} \Pr_\theta[\max_s \sqrt{n}\,\hat{\theta}_s/\hat{\nu}_s > q(\epsilon_\delta)] \leq \alpha. \tag{6}$$

For both results, $\theta \leq 0$ holds element by element.

Remark 6. Theorem 1 and Corollary 1 apply to many settings where the parameter of interest is only partially identified. As an example, consider Imbens and Manski's (2004) missing data problem: (Y_i, W_i) is an i.i.d. sequence for $i = 1,\ldots,n$; W_i is Bernoulli; and Y_i is bounded between 0 and 1 a.s. and is observed only when $W_i = 1$. The parameter of

interest is $\mathrm{E}\,Y_i$ which must satisfy

$$\begin{aligned}\mathrm{E}\,Y_i &\geq \mathrm{E}(Y_i \mid W_i = 1)\Pr[W_i = 1] \\ \mathrm{E}\,Y_i &\leq \mathrm{E}(Y_i \mid W_i = 1)\Pr[W_i = 1] + (1 - \Pr[W_i = 1]).\end{aligned} \tag{7}$$

All of the quantities in (7) can be estimated from the data; the lower bound comes from setting $\mathrm{E}(Y_i \mid W_i = 0) = 0$ and the upper bound from $\mathrm{E}(Y_i \mid W_i = 0) = 1$. Note that $\mathrm{E}\,Y_i$ can not be estimated consistently without further assumptions on the distribution of Y_i given $W_i = 0$, assumptions that may be unrealistic if individuals self-select into the data set, but researchers can still estimate valid confidence intervals and conduct hypothesis tests without such assumptions.

To use Corollary 1 to test $\mathrm{E}\,Y_i = \mu_0$ for some value μ_0, we can define

$$\begin{aligned}\theta_1 &\equiv \mathrm{E}(Y_i \mid W_i = 1)\Pr[W_i = 1] - \mu_0 \\ \theta_2 &\equiv \mu_0 - \mathrm{E}(Y_i \mid W_i = 1)\Pr[W_i = 1] - (1 - \Pr[W_i = 1]),\end{aligned} \tag{8}$$

so (7) becomes $(\theta_1, \theta_2) \leq (0, 0)$. Also define

$$\begin{aligned}\hat{\theta}_1 &= (1/n)\sum_{i=1}^{n} Y_i\, 1\{W_i = 1\} - \mu_0 \\ \hat{\theta}_2 &= \mu_0 - (1/n)\sum_{i=1}^{n} Y_i\, 1\{W_i = 1\} - \Big(1 - (1/n)\sum_{i=1}^{n} 1\{W_i = 1\}\Big).\end{aligned} \tag{9}$$

Assuming $\Pr[W_i = 1]$ is bounded away from zero (as do Imbens and Manski, 2004) and standard moment and dependence conditions, each $\sqrt{n}(\hat{\theta}_i - \theta_i)$ is asymptotically normal under the null and Corollary 1 applies, even if $\Pr[W_i = 1]$ is near 1 (addressing the concern raised by Stoye, 2009).

Confidence intervals for $\mathrm{E}\,Y_i$ can be constructed by inverting these hypothesis tests as usual. Notice that, when the gap between $\hat{\theta}_1$ and $\hat{\theta}_2$ is large, the confidence interval will be based on the distribution of only the closest $\hat{\theta}_i$ since the other inequality will be rejected with very high probability in the first stage, but when the gap is small the interval will use the distributions of both estimators, so our approach matches the key features of Imbens and Manski's (2004) statistic. Our approach has the additional advantage that it can be trivially extended to multivariate Y_i—let $\hat{\theta}_1$ and $\hat{\theta}_2$ be vectors and let (9) apply to each of their elements.

Remark 7. Note that the algorithm should continue as long as it removes inequalities because the corresponding statistic is lower than p_j; removing these inequalities lowers the next upper bound, q_{j+1}. Obviously, there is no need to continue the algorithm once an individual statistic is greater than q_j when testing the composite null. However, if the researcher wants to interpret the individual rejections as well, continuing to reject as many hypotheses as possible (while still controlling FWE) is probably appropriate. See the next remark as well.

Remark 8. Our Monte Carlo section, Section 4, shows that our procedure has comparable power to tests designed specifically for the composite null hypothesis (as studied by, for example, Andrews and Barwick, 2012a). Although there are theoretical reasons to believe that those dedicated tests may have a power advantage in principle, there is another reason to prefer tests that control FWE, even if they suffer a slight power disadvantage: interpretation of the results. We can interpret the individual rejections to learn which inequalities are violated if the test controls FWE, but not if it only controls size for the composite null. If the test recommended by Andrews and Barwick (2012a) rejects we do not learn which inequalities fail, but if our test rejects, we do.

3 Tests that control generalized error rates for families of one-sided and two-sided hypotheses

In some applications, tests that control FWE lack sufficient power and it may be appropriate to control a weaker measure of the error rate. In this section, we show how to apply the principles of the previous section to stepdown methods that control two such measures, the k-FWE and the False Discovery Proportion. We first consider k-FWE, a straightforward extension of the FWE. A critical value that controls k-FWE at level α satisfies

$$\Pr[\hat{\theta}_s \geq q \text{ for at least } k \text{ values of } s \text{ such that } \theta_s \leq 0] \leq \alpha \tag{10}$$

for one-sided tests or

$$\Pr[|\hat{\theta}_s| \geq q \text{ for at least } k \text{ values of } s \text{ such that } \theta_s = 0] \leq \alpha \tag{11}$$

for two-sided tests. (As before, (10) and (11) use unstudentized statistics for simplicity, but our results will use studentized statistics for improved performance.)

Stepdown procedures that control k-FWE face some new difficulties. By design, they continue to run after rejecting true hypotheses, so each step after the first operates under the assumption that some true hypotheses have been rejected, but fewer than k (meaning that the previous steps did not violate (10) or (11)). In Romano and Wolf's (2007) k-StepM procedure, separate critical values are generated using every subset that contains $k-1$ of the rejected hypotheses, and then the most conservative (largest) of those critical values is used for that step of the test. Even if k is relatively small (5 or 6) taking these combinations can be computationally costly.

Our algorithm improves on the k-StepM by ignoring the statistics so large that they would occur with negligibly small probability under the null. It also partitions the alternative space, further restricting the combinations of $k-1$ elements that must be calculated, by estimating the distribution of the ith largest $\hat{\theta}_s$ under the null, for every $i = 1, \ldots, k$. Intuitively, if both $\hat{\theta}_1$ and $\hat{\theta}_2$ are larger than the upper bound for the second-largest test statistic, combinations that include both $s = 1$ and $s = 2$ can be ignored.

First we present an algorithm for two-sided tests. Define the k-max operator to return the kth largest of its arguments and let $\#A$ denote the number of elements in a set A.

Algorithm 2 (k-StepM variation for two-sided tests). *Set* $M_0 = \{1, \ldots, S\}$, $R_0 = \{\emptyset\}$, $\alpha \in (0,1)$, *and* $\epsilon_i \in (0, \alpha)$ *for* $i = 1, \ldots, k-1$. *Repeat the following steps for each* $j = 1, 2, \ldots$ *and stop when* $M_j = \emptyset$ *or* $(M_j, N_{1j}, \ldots, N_{k-1,j}) = (M_{j-1}, N_{1,j-1}, \ldots, N_{k-1,j-1})$.

1. *Set* $r_{ij} = \max_{I \in R_{j-1}} \rho_{ij}(I)$ *for* $i = 1, \ldots, k-1$, *where* $\rho_{ij}(I)$ *is the* $1-\epsilon_i$ *quantile of the distribution of* i-$\max_{s \in M_{j-1} \cup I} |\hat{\psi}_s^*|$.

2. *Set* $q_j = \max_{I \in R_{j-1}} q_{Ij}$, *where* q_{Ij} *is the* $1-\alpha$ *quantile of the distribution of* k-$\max_{s \in M_{j-1} \cup I} |\hat{\psi}_s^*|$.

3. *Set* $M_j = \{s : |\sqrt{n}\,\hat{\theta}_s / \hat{v}_s| \leq q_j\}$,

$$N_{ij} = \begin{cases} \{s : |\sqrt{n}\,\hat{\theta}_s / \hat{v}_s| \in (r_{i+1,j}, r_{1j}]\} & i = 1, \ldots, k-2 \\ \{s : |\sqrt{n}\,\hat{\theta}_s / \hat{v}_s| \in (q_j, r_{1j}]\} & i = k-1 \end{cases}$$

and $R_j = \{I \subset N_{k-1,j} : \#(I \cap N_{ij}) \leq i \text{ for } i = 1, \ldots, k-1\}$.

Let q be the last q_j when the algorithm stops.

Theorem 2 establishes that this algorithm is valid for very small values of ϵ.

Theorem 2 (k-FWE control for two-sided hypotheses)**.** *Suppose Assumption 1 holds and choose $\alpha \in (0,1)$. For any $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_{k-1})'$, let $q(\boldsymbol{\epsilon})$ denote the last q_j produced by Algorithm 2. Then*

$$\limsup_{\substack{n\to\infty \\ \epsilon\to 0}} \sup_{\theta\in\mathbb{R}^S} \Pr_\theta[|\sqrt{n}\,\hat{\theta}_s/\hat{v}_s| > q(\boldsymbol{\epsilon}) \text{ for at least } k \text{ values of } s \text{ such that } \theta_s = 0] \leq \alpha. \quad (12)$$

Moreover, if $\boldsymbol{\epsilon}_\delta$ is a sequence of random variables such that $\boldsymbol{\epsilon}_\delta \to^p 0$ as $\delta \to 0$ then

$$\limsup_{\substack{n\to\infty \\ \delta\to 0}} \sup_{\theta\in\mathbb{R}^S} \Pr_\theta[|\sqrt{n}\,\hat{\theta}_s/\hat{v}_s| > q(\boldsymbol{\epsilon}_\delta) \text{ for at least } k \text{ values of } s \text{ such that } \theta_s = 0] \leq \alpha. \quad (13)$$

Remark 9. As in Remark 3, if $\hat{F}_n$ is estimated with a bootstrap, we can often set r_{ij} as

$$r_{ij} = \begin{cases} \max_{b=1,\ldots,B} \max_{s\in M_0} i\text{-max}\,|\hat{\psi}^*_{bs}| & j = 1,\ i = 1,\ldots,k-1 \\ \max_{b=1,\ldots,B} \max_{I\in R_{j-1}} i\text{-max}_{s\in M_{j-1}\cup I}\,|\hat{\psi}^*_{bs}| & j > 1,\ i = 1,\ldots,k-1 \end{cases}$$

where $\hat{\psi}^*_{1s}, \ldots, \hat{\psi}^*_{Bs}$ are the bootstrap replications of the test statistic for the sth hypothesis. Again, this is the procedure we recommend in practice.

As with Algorithm 1, we could allow the ϵ_i to remain positive in the limit by using the $1-\alpha+\sum_i \epsilon_i$ quantile for q_j.

Remark 10. As in most sequential algorithms, it is sufficient to show that each step of the algorithm controls the error rate only when all of the previous steps have already done so. So, for $j > 1$, we can assume that fewer than k true null hypotheses have been rejected in the previous steps. In Romano and Wolf's (2007) original proof, the outer maximum corresponding to our step 2 is taken over all sets I of size $k-1$, where $I \subset \{s : |\sqrt{n}\,\hat{\theta}_s/\hat{v}_s| > q_{j-1}\}$ (in our notation). The set of these I is potentially much larger than our R_{j-1}, reducing power and lengthening computational time.

We can use a smaller set because R_{j-1} removes the combinations that only occur with negligible probability under the null. For example, r_{1j} is essentially an upper bound

on $\max|\sqrt{n}\,\hat{\theta}_s/\hat{v}_s|$ under the null and r_{2j} is essentially an upper bound on the second largest $|\sqrt{n}\,\hat{\theta}_s/\hat{v}_s|$ under the null. So at most one true hypothesis can have its test statistic between $r_{1,j-1}$ and $r_{2,j-1}$ and we can ignore all combinations that include the indices of two or more statistics between $r_{1,j-1}$ and $r_{2,j-1}$. The justification for the other bounds is the same.

Typically, the most useful restriction will be that statistics greater than r_{i1} can be rejected and ignored—none of the sets in R_{j-1} contain the indices of those statistics.

Remark 11. We can also compare this approach to the streamlined algorithm proposed by Romano and Wolf (2007) (their Algorithm 2.2) which restricts R_j even further. In our notation, they propose using R_j that contains a single set with the indices of the $k-1$ smallest test statistics. This choice of R_j will increase power and decrease computational costs even further, but is valid asymptotically only if the parameters are far from zero under the alternative. If some of the parameters are local alternatives, there is no guarantee that their approach will generate a valid critical value, but ours will.

This algorithm can of course be modified for one-sided tests by adding the threshold p_j used in Theorem 1 (i.e. excluding those statistics that are too far below zero). Algorithm 3 presents this result.

Algorithm 3 (k-StepM variation for one-sided tests)**.** *Set $M_0 = \{1,\ldots,S\}$, $R_0 = \{\emptyset\}$, and $\alpha \in (0,1)$, and $\epsilon_i \in (0,\alpha)$ for $i = 1,\ldots,k$. Repeat the following steps for each $j = 1,2,\ldots$ and stop when $M_j = \emptyset$ or $(M_j, N_{1j},\ldots,N_{k-1,j}) = (M_{j-1}, N_{1,j-1},\ldots,N_{k-1,j-1})$.*

1. *Set $r_{ij} = \max_{I\in R_{j-1}} \rho_{ij}(I)$ for $i = 1,\ldots,k-1$, where $\rho_{ij}(I)$ is the $1-\epsilon_i$ quantile of the distribution of i-$\max_{s\in M_{j-1}\cup I}|\hat{\psi}^*_s|$.*

2. *Set $p_j = \min_{I\in R_{j-1}} p_{Ij}$, where p_{Ij} is the ϵ_k quantile of the distribution of $\min_{s\in M_{j-1}\cup I}\hat{\psi}^*_s$.*

3. *Set $q_j = \max_{I\in R_{j-1}} q_{Ij}$, where q_{Ij} is the $1-\alpha$ quantile of the distribution of k-$\max_{s\in M_{j-1}\cup I}|\hat{\psi}^*_s|$.*

4. *Set $M_j = \{s : \sqrt{n}\,\hat{\theta}_s/\hat{v}_s \in [p_j, q_j]\}$,*

$$N_{ij} = \begin{cases} \{s : \sqrt{n}\,\hat{\theta}_s/\hat{v}_s \in (r_{i+1,j}, r_{1j}]\} & i = 1,\ldots,k-2 \\ \{s : \sqrt{n}\,\hat{\theta}_s/\hat{v}_s \in (q_j, r_{1j}]\} & i = k-1 \end{cases}$$

and $R_j = \{I \subset N_{k-1,j} : \#(I \cap N_{ij}) \leq i \text{ for } i = 1, \ldots, k-1\}$.

Let q *be the last* q_j *when the algorithm stops.*

Theorem 3 (k-FWE control for one-sided hypotheses)**.** *Suppose Assumption 1 holds and choose* $\alpha \in (0,1)$*. For any* $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_k)'$*, let* $q(\boldsymbol{\epsilon})$ *denote the last* q_j *produced by Algorithm 3. Then*

$$\limsup_{\substack{n\to\infty \\ \boldsymbol{\epsilon}\to 0}} \sup_{\theta \in \mathbb{R}^S} \Pr_\theta[\sqrt{n}\,\hat{\theta}_s/\hat{v}_s > q(\boldsymbol{\epsilon}) \text{ for at least } k \text{ values of } s \text{ such that } \theta_s \leq 0] \leq \alpha. \quad (14)$$

If $\boldsymbol{\epsilon}_\delta$ *is a sequence of random vectors s.t.* $\boldsymbol{\epsilon}_\delta \to^p 0$ *as* $\delta \to 0$ *then*

$$\limsup_{\substack{n\to\infty \\ \delta\to 0}} \sup_{\theta \in \mathbb{R}^S} \Pr_\theta[\sqrt{n}\,\hat{\theta}_s/\hat{v}_s > q(\boldsymbol{\epsilon}) \text{ for at least } k \text{ values of } s \text{ such that } \theta_s \leq 0] \leq \alpha. \quad (15)$$

Remark 12. Note that the parameters estimated to be far below the binding inequality are removed and do not enter as elements of $I \subset R_j$ or M_j. Just as before, this modification increases the test's power.

Remark 13. If $\hat{F}_n$ is estimated with a bootstrap, we can often use

$$p_j = \begin{cases} \min_{b=1,\ldots,B} \min_{s \in M_0} \hat{\psi}^*_s & j = 1 \\ \min_{b=1,\ldots,B} \min_{I \in R_{j-1}} \min_{s \in M_{j-1} \cup I} \hat{\psi}^*_s & j > 1 \end{cases}$$

and

$$r_{ij} = \begin{cases} \max_{b=1,\ldots,B} i\text{-max}_{s \in M_0} \hat{\psi}^*_{bs} & j = 1,\ i = 1, \ldots, k-1 \\ \max_{b=1,\ldots,B} i\text{-max}_{I \in R_{j-1}} i\text{-max}_{s \in M_{j-1} \cup I} \hat{\psi}^*_{bs} & j > 1,\ i = 1, \ldots, k-1 \end{cases}$$

where, again, $\hat{\psi}^*_{1s}, \ldots, \hat{\psi}^*_{Bs}$ are the bootstrap replications of the test statistic for the sth hypothesis (also see Remarks 3 and 9).

We now turn to the second generalized error rate, the *False Discovery Proportion*

(FDP). A critical value q controls FDP at level α if it satisfies

$$\Pr\left[\frac{\#\{s : |\hat{\theta}_s| \geq q \text{ and } \theta_s = 0\}}{\max(1, \#\{s : |\hat{\theta}_s| \geq q\})} > \gamma\right] \leq \alpha \tag{16}$$

for two-sided tests or

$$\Pr\left[\frac{\#\{s : \hat{\theta}_s \geq q \text{ and } \theta_s \leq 0\}}{\max(1, \#\{s : \hat{\theta}_s \geq q\})} > \gamma\right] \leq \alpha \tag{17}$$

for one-sided tests, for γ determined by the researcher in advance; i.e. it controls the probability that a predetermined percentage of the rejections are incorrect. As shown by Lehmann and Romano (2005), procedures that control k-FWE can be used to build procedures that control FDP. Suppose that a test that controls k-FWE at level α rejects N hypotheses. If $N > k/\gamma$, then $\Pr[k/N > \gamma] \leq \alpha$ and FDP is controlled at level α as well. So one can proceed sequentially in k, starting with $k = 1$, then 2, etc., stopping when $N \leq k/\gamma$. Corollary 2 demonstrates how to extend the Algorithms 2 and 3 to this application.

Corollary 2 (FDP control)**.** *Suppose Assumption 1 holds and take $\alpha, \gamma \in (0, 1)$.*

1. *(One-sided hypotheses): Apply Algorithm 3 sequentially at level α with $k = 1, 2, \ldots$ producing a sequence of critical values $q_k(\boldsymbol{\epsilon})$, and stop at the first k with*

$$k/\gamma \geq \#\{s : \sqrt{n}\,\hat{\theta}_s/\hat{v}_s > q_k(\boldsymbol{\epsilon})\}. \tag{18}$$

   Let $q(\boldsymbol{\epsilon})$ denote the last $q_k(\boldsymbol{\epsilon})$. Then

$$\limsup_{\substack{n\to\infty \\ \boldsymbol{\epsilon}\to 0}} \sup_{\theta\in\mathbb{R}^n} \Pr_\theta\left[\frac{\#\{s : \sqrt{n}\,\hat{\theta}_s/\hat{v}_s > q(\boldsymbol{\epsilon}) \text{ and } \theta_s \leq 0\}}{\max(1, \#\{s : \sqrt{n}\,\hat{\theta}_s/\hat{v}_s > q(\boldsymbol{\epsilon})\})} > \gamma\right] \leq \alpha. \tag{19}$$

   If $\boldsymbol{\epsilon}_\delta$ is a sequence of random variables s.t. $\boldsymbol{\epsilon}_\delta \to^p 0$ as $\delta \to 0$ then

$$\limsup_{\substack{n\to\infty \\ \delta\to 0}} \sup_{\theta\in\mathbb{R}^n} \Pr_\theta\left[\frac{\#\{s : \sqrt{n}\,\hat{\theta}_s/\hat{v}_s > q(\boldsymbol{\epsilon}_\delta) \text{ and } \theta_s \leq 0\}}{\max(1, \#\{s : \sqrt{n}\,\hat{\theta}_s/\hat{v}_s > q(\boldsymbol{\epsilon}_\delta)\})} > \gamma\right] \leq \alpha \tag{20}$$

   as well.

2. *(Two-sided hypotheses): Apply Algorithm 2 sequentially at level α with $k = 1, 2, \ldots$*

producing a sequence of critical values $q_k(\epsilon)$, and stop at the first k with

$$k/\gamma \geq \#\{s : |\sqrt{n}\,\hat{\theta}_s/\hat{v}_s| > q_k(\epsilon)\}. \tag{21}$$

Let $q(\epsilon)$ denote the last $q_k(\epsilon)$. Then

$$\limsup_{\substack{n\to\infty \\ \epsilon\to 0}} \sup_{\theta\in\mathbb{R}^n} \Pr_\theta\left[\frac{\#\{s : |\sqrt{n}\,\hat{\theta}_s/\hat{v}_s| > q(\epsilon) \text{ and } \theta_s = 0\}}{\max(1, \#\{s : |\sqrt{n}\,\hat{\theta}_s/\hat{v}_s| > q(\epsilon)\})} > \gamma\right] \leq \alpha. \tag{22}$$

If ϵ_δ is a sequence of random variables s.t. $\epsilon_\delta \to^p 0$ as $\delta \to 0$ then

$$\limsup_{\substack{n\to\infty \\ \delta\to 0}} \sup_{\theta\in\mathbb{R}^n} \Pr_\theta\left[\frac{\#\{s : |\sqrt{n}\,\hat{\theta}_s/\hat{v}_s| > q(\epsilon_\delta) \text{ and } \theta_s = 0\}}{\max(1, \#\{s : |\sqrt{n}\,\hat{\theta}_s/\hat{v}_s| > q(\epsilon_\delta)\})} > \gamma\right] \leq \alpha. \tag{23}$$

Remark 14. The computational improvements of our algorithm are especially important when controlling FDP since k grows. To further reduce computational costs, step $k+1$ can be started where step k left off: if $r'_1, \ldots, r'_{k-1}$ and p' denote the last values of $r_{1j}, \ldots, r_{k-1,j}$ and p_j at step k, we can set $r_{i1} = r'_i$ and $p_1 = p'$ for step $k+1$.

Remark 15. Note that steps can sometimes be skipped: if

$$(k+m)/\gamma < \#\{s : |\sqrt{n}\hat{\theta}_s/\hat{v}_s| > q_k\}$$

then we can go immediately to $k+m+1$ instead of $k+1$.

Remark 16. If the computational costs become overwhelming, the algorithm can be stopped early. It still controls FDP at the prespecified levels, but obviously sacrifices some power. Since the computational costs grow with the number of hypotheses rejected, this scenario will come into play when many hypotheses have already been rejected and the loss of power may be acceptable.

4 Monte Carlo evidence

For a sense of the finite sample performance of our tests we present simulations for several different DGPs based loosely on Romano and Wolf's (2005a) Monte Carlo design. We study the performance of three of our procedures: the StepM modification for one-sided tests derived in Theorem 1, the k-StepM modification for one-sided tests described

in Theorem 3, and the test of composite one-sided nulls described in Corollary 1. All of these simulations were programmed in R (R Development Core Team, 2012) and use the MASS (Venables and Ripley, 2002), xtable (Dahl, 2012), dbframe (Calhoun, 2010), RSQLite (James, 2012), R.Matlab (Bengtsson, 2013), and Combinations (Temple Lang, 2010) packages.

The Monte Carlo design is fairly basic: $\theta_s = \mathrm{E}X_{s,t} - \mathrm{E}Y_t$ for $s = 1, \ldots, S$. For design 1, $s = 2$ and

$$\begin{pmatrix} X_{1t} \\ X_{2t} \\ Y_t \end{pmatrix} \sim i.i.d.N\left(\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 2 & 0 & 1 \\ 0 & 2 & 1 \\ 1 & 1 & 1 \end{pmatrix} \right). \tag{24}$$

This covariance structure ensures that $\bar{X}_{1,\cdot} - \bar{Y}$ and $\bar{X}_{2,\cdot} - \bar{Y}$ are perfectly negatively correlated and this design is used to study size distortions in these procedures. This DGP also mimics the partial-identification setting of Imbens and Manski (2004) — see Remark 6. For designs 2–6, $S = 40$ and

$$(X_{1,t}, \ldots, X_{40,t}, Y_t) \sim N(\mu, \mathrm{diag}(1, 2, 1, 2, \ldots, 2, 1)). \tag{25}$$

And for designs 7–9, $S = 4$ and

$$(X_{1,t}, \ldots, X_{4,t}, Y_t) \sim N(\mu, I). \tag{26}$$

Designs 1–6 are used for the multiple-testing simulations and designs 7–9 are used for the composite null hypotheses (one of the comparison methods, Andrews and Barwick's, 2012a, is computationally infeasible for 40 inequalities, so we drop the number). The mean, μ, is determined by the DGP; $\mathrm{E}Y_t = 1$ for all of the simulations, so $\mathrm{E}X_{i,t}$ takes on different values. Table 1 presents these different possible values.

The first Monte Carlo compares procedures that control FWE. It studies the size and power of Romano and Wolf's (2005a) StepM, Hsu et al.'s (2010) Step-SPA, and the our refinement of the StepM. The Step-SPA is a variation of Romano and Wolf's (2005a) StepM that initially discards the null hypotheses s for which $\hat{\theta}_s / \hat{\sigma}_s \leq -\sqrt{2 \log \log n}$, a threshold suggested by Hansen's (2005). All results are based on 1000 simulations and critical values are estimated using the i.i.d. bootstrap with 999 bootstrap samples, and use DGP designs 1–6. The lower thresholds, p_j, are set as the minimum of the

bootstrap replications as suggested in Remark 3. The tests have nominal FWE of 5%, but simulations at 10% and 1% show similar patterns.

The second Monte Carlo compares procedures that control k-FWE with $k = 3$; it uses Romano and Wolf's (2007) k-StepM and our refinement presented in Theorem 2. The simulations use DGP designs 2–6 (the same as the first Monte Carlo, except that design 1 has fewer than k total hypotheses and is dropped) and the results are again based on 1000 simulations using an i.i.d bootstrap with 999 bootstrap samples, and the thresholds are set as in Remark 13 The nominal k-FWE is 5%.

The third Monte Carlo studies tests of composite null hypotheses: Andrews and Barwick's (2012a) AQLR and McCloskey's (2012) and Romano et al.'s (2012) Bonferroni-based procedures in addition to our method described in Corollary 1; Andrews and Barwick (2012a) conduct an extensive simulation study in which they demonstrate that the AQLR performs better than many other recent tests of the composite null; see their paper for further discussion and comparisons. The Bonferroni-based method is a two-step procedure: it first conducts a one-sided test of the null $\theta_s \geq 0$ for each s at level $\alpha/10$, then constructs the $\alpha \cdot 9/10$ one-sided critical value for the null $\theta_s \leq 0$ for all s not rejected in the first-stage test. The procedure rejects if any of the test statistics are greater than this critical value, which can be approximated through the bootstrap. These simulations use DGP designs 7–9 for computational feasibility; the results are based on 1000 simulations and our method and the Bonferroni procedure both use 999 bootstrap samples. The AQLR is implemented using Matlab code provided by Andrews and Barwick (2012a,b) with their recommended settings, which is called from R via R.Matlab (Bengtsson, 2013). The nominal size for these tests is 5% and the lower threshold is again set as in Remark 3.

Table 2 presents the results for the FWE experiment and strongly supports this paper's new approach. Both the StepM and our refinement control FWE reliably, but the Step-SPA overrejects when there is a small number of equal-performing models—in simulation 1 it overrejects by almost 5 percentage points for 50 observations and 2.3 percentage points for 100 observations (note that the Step-SPA is equivalent to Hansen's original SPA in this experiment since all of the null hypotheses are true). For DGPs with no under-performing models (DGPs 1–3, and 5) our new procedure performs essentially the same as the StepM in terms of FWE and power. When some models under-perform (DGPs 4 and 6), the new method identifies more incorrect null hypotheses than the original. For example, in DGP 4 with 100 observations, Romano and Wolf's (2005a)

test finds on average 2.1 false hypotheses while this paper's test finds 4.1 out of 6, a substantial improvement; for 50 observations, the StepM finds 0.7 false hypotheses on average and this paper's test finds 2.0. Our method and the Step-SPA have basically the same power, but our method avoids over-rejecting when the inequalities bind.

Table 3 presents results for the k-FWE experiment, which again favor our approach. Our method and the k-StepM control the k-FWE at essentially identical (and correct) rates and when none of the models underperform the methods have almost identical power. But when some models do underperform, our method correctly rejects substantially more hypotheses. For example, in DGP 6 with 100 observations, our method rejects 4 more statistics (16.4 vs. 12.5) with identical control of k-FWE. The relative performance in other DGPs is similar.

Finally, Table 4 presents results for the size experiment. Here all of the methods perform about the same. All have estimated size slightly less than nominal size, but without cause for concern. And all of the statistics have nearly identical power when there are false hypotheses. As mentioned in Remark 8, an advantage of our statistic (and the Bonferroni-based statistic) is that researchers are justified in interpreting the individual statistic-by-statistic test results, while the AQLR does not. These simulations indicate that the power loss from taking this approach may be very small, the numbers are virtually identical, which makes our statistic more attractive.

Taken collectively, these simulations show that our improvements lead to substantially more powerful tests in the multiple testing scenario that they were designed for, and also perform roughly as well as specialized (and complicated) statistics like the AQLR for testing composite null hypotheses.

5 An analysis of hedge fund performance

To demonstrate this paper's new approach on a real dataset, we conduct an empirical study of hedge funds similar to Romano and Wolf (2005a) and Romano et al. (2008), but controlling the FDP. It is well known that accounting for data-snooping is particularly important when analyzing investment strategies since many strategies can appear profitable by pure chance and mistakes can be costly, so this is a natural setting to apply tests that correct for multiple hypotheses (see especially Lo and MacKinlay, 1990, White, 2000, Sullivan et al., 2001, and Kosowski et al., 2006).

The CISDM database reports monthly net returns for a large number of active and

closed hedge funds from January, 1994 to December, 2011. Our empirical exercise will determine which of the funds outperformed the risk-free rate over this time period. If r_{it} is the log return of fund i in period t and r_{ft} the log return of the risk-free rate, we test the family of null hypotheses

$$(1/n)\sum_{i=1}^{n} \mathrm{E}(r_{it} - r_{ft}) \leq 0$$

using Algorithm 1.

It is plausible that under-performing hedge funds are more likely to close than high performers (see, e.g., Amin and Kat, 2003, and Capocci and Hübner, 2004). To try to mitigate this sort of survivorship bias, we include every fund in the CISDM database that was active at the beginning of the sample, January, 1994. There are 196 active funds that meet this criterion, 126 closed funds, and 10 hedge fund indices constructed by the CISDM. Many of the funds (146 of the 196 active funds) are missing observations, but most are only missing a few. Figure 1 plots a histogram of the number of missing observations for the funds that are missing one or more observations.

It is possible that returns are reported selectively and the missing returns have lower mean than the reported returns, but a full examination of this problem is beyond the scope of this analysis. The individual test statistics for each fund are constructed by pasting together the non-missing observations and estimating the mean and variance of the new series; hedge fund returns can exhibit serial correlation (Lo, 2002, and Getmansky et al., 2004) so the standard error is calculated with a prewhitened QS kernel with Andrews and Monahan's (1992) automatic bandwidth calculation (see Andrews, 1991, as well). The statistics' distribution is estimated with a Circular Block Bootstrap with block length 15 (Politis and Romano, 1992)—the empirical results were essentially the same across a range of block lengths; we sample missing observations too and calculate the bootstrap average return using only the non-missing observations; the bootstrap variance is estimated using the natural variance estimator corresponding to the block length (Gotze and Kunsch, 1996). The lower thresholds, p_j, are set as in Remark 3 to be the smallest statistic across the bootstrap replications. We use R (R Development Core Team, 2012, version 2.14.1) as well as the tikzDevice (Sharpsteen and Bracken, 2012), Combinations (Temple Lang, 2010), dbframe (Calhoun, 2010) and sandwich (Zeileis, 2004, version 2.2-9) packages for the analysis.

Tables 5 and 6 present results; the funds listed were determined to outperform the risk-

free rate, controlling the FDP at 5% with $\gamma = 0.1$, and we calculate the corresponding confidence intervals for each of those funds. These confidence intervals have been estimated by testing the null hypotheses

$$(1/n)\sum_{i=1}^{n} \mathrm{E}(r_{it} - r_{ft}) \leq c$$

for successively larger values of c using Algorithm 3, and the lower bound for each fund is the last value of c that the algorithm rejects; consequently there is at most a 5% probability that the true expected returns above the risk-free rate lie outside these intervals for 10% or more of the funds.

Table 5 lists the active outperforming funds and Table 5 lists the outperforming funds that have been closed; the first column of each table is the lower bound of the confidence interval, the second column is the average excess return over the time period, and the last column is the studentized return. Two of the CISDM indices have significantly outperformed the risk free rate over this time horizon, but the rest of the listings are individual funds. We also applied the one-sided StepM to these funds, and the fourteen funds that were found to outperform the risk-free rate by that measure are marked with an asterisk. Our FDP procedure found 27 funds to outperform the risk-free rate, demonstrating that there can be substantial gains in power from relaxing FWE control.

6 Conclusion

This paper proposes simple modifications of existing stepdown procedures that increase power and reduce computational costs. The underlying idea—find and exclude events that occur with arbitrarily small probability in sequential testing—has other potential applications as well. Our simulation evidence indicates that the increase in power can be substantial.

Appendix A: Proofs of main results

Proof of Theorem 1. We will present the proof of (3) only, as the proof of (4) is essentially identical. Let $\{\theta_n\}$ be any sequence of vectors in $\mathbb{R}^S$ and $\{\epsilon_n\}_n$ a sequence of positive numbers that converges to zero as $n \to \infty$, where ϵ_n is used in place of ϵ in the theorem's

statement. Then there exists a subsequence $\{n(m)\}_m$ of $\{n\}$ such that the limit of

$$\Pr_{\theta_{n(m)}}[\sqrt{n(m)}\,\theta_s > q \text{ for at least one } s \text{ such that } \theta_{n(m),s} \leq 0] \tag{27}$$

exists as $m \to \infty$; call this limit β. There also exists a further subsequence $\{n(m(\ell)\}_\ell$ such that each element of $\{\sqrt{n(m(\ell))}\,\theta_{n(m(\ell))}\}$ either converges to a finite limit or diverges to $\pm\infty$. To reduce the notational clutter, we'll write $n(m(\ell))$ as n_ℓ, $\epsilon_{n(m(\ell))}$ as ϵ_ℓ, and $\Pr_{\theta_{n_\ell}}$ as $\Pr_\ell$ for the rest of the proof. It suffices to prove that $\beta \leq \alpha$ for any such subsequence.

Define two subsets of $\{1,\ldots,S\}$:

$$I_1 \equiv \{s : -\infty < \lim_{\ell\to\infty} \sqrt{n_\ell}\,\theta_{n_\ell,s} \leq 0\}$$
$$I_2 \equiv \{s : \lim_{\ell\to\infty} \sqrt{n_\ell}\,\theta_{n_\ell,s} = -\infty\}.$$

We can assume that $I_1 \cup I_2$ is nonempty (otherwise $\beta = 0$ for this subsequence and the result is trivial). Moreover,

$$\lim_{\ell\to\infty} \Pr_\ell[\max_{s\in I_1\cup I_2} \hat{\psi}_s > q] \leq \lim_{\ell\to\infty} \Pr_\ell[\max_{s\in I_1} \hat{\psi}_s > q] + \lim_{\ell\to\infty} \Pr_\ell[\max_{s\in I_2} \hat{\psi}_s > q], \tag{28}$$

where $\hat{\psi}_s \equiv \sqrt{n_\ell}\,\hat{\theta}_{n_\ell,s}/\hat{v}_s$, so it suffices to prove that

$$\lim_{\ell\to\infty} \Pr_\ell[\max_{s\in I_1} \hat{\psi}_s > q] \leq \alpha \tag{29}$$

and

$$\lim_{\ell\to\infty} \Pr_\ell[\max_{s\in I_2} \hat{\psi}_s > q] = 0 \tag{30}$$

and we can assume for the rest of the proof that neither I_1 nor I_2 are empty.

Start with the obvious inequality

$$\Pr_\ell[\max_{s\in I_1} \hat{\psi}_s > q] \leq \Pr_\ell[\max_{s\in I_1} \hat{\psi}_s > q \text{ or } \min_{s\in I_1} \hat{\psi}_s < p]; \tag{31}$$

it suffices to bound the lim sup of the larger quantity. Consider the event on the right side of (31) and let j be the first step in the algorithm where one of these inequalities

holds, i.e.

$$\max_{s\in I_1}\hat{\psi}_s > q_j \text{ or } \min_{s\in I_1}\hat{\psi}_s < p_j$$

but

$$\max_{s\in I_1}\hat{\psi}_s \leq q_g \text{ and } \min_{s\in I_1}\hat{\psi}_s \geq p_g \quad \text{for all } g < j.$$

We know (by construction of j) that $I_1 \subset M_{j-1}$ almost surely, and so $p_j \leq p'$, and $q_j \geq q'$ almost surely where p' and q' are the ϵ_ℓ quantile of the distribution of $\min_{s\in I_1}\hat{\psi}^*_s$ and the $1-\alpha$ quantile of the distribution of $\max_{s\in I_1}\hat{\psi}^*_s$ respectively. Consequently,

$$\begin{aligned}\Pr_\ell[\max_{s\in I_1}\hat{\psi}_s > q \text{ or } \min_{s\in I_1}\hat{\psi}_s < p] &\leq \Pr_\ell[\max_{s\in I_1}\hat{\psi}_s > q' \text{ or } \min_{s\in I_1}\hat{\psi}_s < p'] \\ &\leq \Pr_\ell[\max_{s\in I_1}\hat{\psi}_s > q'] + \Pr_\ell[\min_{s\in I_1}\hat{\psi}_s < p']\end{aligned} \tag{32}$$

and

$$\Pr_\ell[\max_{s\in I_2}\hat{\psi}_s > q] \leq \Pr_\ell[\max_{s\in I_2}\hat{\psi}_s > q']. \tag{33}$$

Finally, consistency of $\hat{F}_n$ for the limiting distribution of $\hat{\psi}$ ensures that

$$\lim_{\ell\to\infty}\Pr_\ell[\max_{s\in I_1}\hat{\psi}_s > q'] \leq \alpha, \tag{34}$$

$$\Pr_\ell[\min_{s\in I_1}\hat{\psi}_s < p'] \to 0 \text{ as } \ell \to \infty \tag{35}$$

and

$$\Pr_\ell[\max_{s\in I_2}\hat{\psi}_s > q'] \to 0 \text{ as } \ell \to \infty \tag{36}$$

completing the proof. □

Proof of Theorem 2. We will only present the proof of (12), since the proof of (13) is essentially identical. As in the proof of Theorem 1, let $\{\theta_n\}$ be any sequence of vectors in $\mathbb{R}^S$ and $\{\epsilon_n\}_n$ a sequence of positive numbers that converges to zero as $n \to \infty$, where ϵ_n is used in place of ϵ in the theorem's statement, and let $\{n_\ell\}_\ell$ and $\{\epsilon_\ell\}_\ell$ be subsequences

such that, as $\ell \to \infty$,

$$\Pr_{\theta_{n_\ell}}[|\sqrt{n_\ell}\ \hat{\theta}_s/\hat{v}_s| > q \text{ for at least } k \text{ values of } s \text{ such that } \theta_{\ell,s} = 0] \to \beta \tag{37}$$

and each element of $\{\sqrt{n_\ell}\ \theta_{n_\ell}\}$ either converges to a finite limit or diverges to $\pm\infty$. It suffices to prove that $\beta \le \alpha$ for any such subsequence.

Define $\hat{\psi}_s = \sqrt{n_\ell}\ \hat{\theta}_{n_\ell,s}/\hat{v}_s$ and write $\Pr_{\theta_{n_\ell}}$ as $\Pr_\ell$ for the rest of the proof to further simplify notation. Define a subset of $\{1,\ldots,S\}$:

$$I_1 \equiv \{s : \lim_{\ell\to\infty} \sqrt{n_\ell}\ \theta_{n_\ell,s} = 0\}.$$

We can assume that I_1 has k or more elements and it suffices to prove that

$$\lim_{\ell\to\infty} \Pr_\ell[k\text{-}\max_{s\in I_1} |\hat{\psi}_s| > q] \le \alpha. \tag{38}$$

Note that

$$\Pr_\ell[k\text{-}\max_{s\in I_1} |\hat{\psi}_s| > q] \le \Pr_\ell[k\text{-}\max_{s\in I_1} |\hat{\psi}_s| > q \text{ or } i\text{-}\max_{s\in I_1} |\hat{\psi}_s| > r_{i+1} \text{ for at least one } i = 1,\ldots,k-1]; \tag{39}$$

where each r_i denotes the last r_{ij}, so it suffices to bound the lim sup of the larger quantity. Let j be the first step in the algorithm where one of these inequalities holds, so

$$k\text{-}\max_{s\in I_1} |\hat{\psi}_s| > q_j \text{ or } i\text{-}\max_{s\in I_1} |\hat{\psi}_s| > r_{i+1,j} \text{ for at least one } i = 1,\ldots,k-1$$

but

$$k\text{-}\max_{s\in I_1} |\hat{\psi}_s| \le q_g \text{ and } i\text{-}\max_{s\in I_1} |\hat{\psi}_s| \le r_{i+1,g},\ i = 1,\ldots,k-1 \quad \text{for all } g < j.$$

Then

$$i\text{-}\max_{s\in I_1} |\hat{\psi}_s| > \max_{I\in R_{j-1}} r_{iI} \tag{40}$$

for some $i \in \{1, \ldots, k-1\}$ or

$$k\text{-}\max_{s \in I_1} |\hat{\psi}_s| > \max_{I \in R_{j-1}} q_I \tag{41}$$

must hold a.s., with r_{iI} the $1-\epsilon_\ell$ quantile of the distribution of $i\text{-}\max_{s \in M_{j-1} \cup I} \hat{\psi}_s^*$ and q_I the $1-\alpha$ quantile of $k\text{-}\max_{s \in M_{j-1} \cup I} \hat{\psi}_s^*$. We know (by construction of j) that $I_1 \subset M_{j-1} \cup I$ almost surely for at least one $I \in R_{j-1}$, and so

$$\max_{I \in R_{j-1}} r_{iI} \geq r_i', \tag{42}$$

and

$$\max_{I \in R_{j-1}} q_I \geq q' \tag{43}$$

almost surely where each r_i' is the $1-\epsilon_\ell$ quantile of the distribution of $i\text{-}\max_{s \in I_1} |\hat{\psi}_s^*|$ and q' is the $1-\alpha$ quantile of the distribution of $k\text{-}\max_{s \in I_1} |\hat{\psi}_s^*|$. Consequently,

$$\begin{aligned} &\Pr_\ell[k\text{-}\max_{s \in I_1} |\hat{\psi}_s| > q \text{ or } i\text{-}\max_{s \in I_1} |\hat{\psi}_s| > r_i \text{ for at least one } i] \\ &\qquad \leq \Pr_\ell[k\text{-}\max_{s \in I_1} |\hat{\psi}_s| > q'] + \sum_{i=1}^{k-1} \Pr_\ell[i\text{-}\max_{s \in I_1} |\hat{\psi}_s| > r_i'] \end{aligned} \tag{44}$$

and consistency of $\hat{F}_n$ ensures that

$$\lim_{\ell \to \infty} \Pr_\ell[k\text{-}\max_{s \in I_1} |\hat{\psi}_s| > q'] \leq \alpha \tag{45}$$

and

$$\sum_{i=1}^{k-1} \Pr_\ell[i\text{-}\max_{s \in I_1} |\hat{\psi}_s| > r_i'] \to 0 \text{ as } \ell \to \infty \tag{46}$$

completing the proof. □

Proof of Theorem 3. This proof is a straightforward combination of the arguments for Theorems 1 and 2 and is omitted. □

References

G. S. Amin and H. M. Kat. Welcome to the dark side: hedge fund attrition and survivorship bias over the period 1994-2001. *The Journal of Alternative Investments*, 6(1):57–73, 2003.

D. W. Andrews. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, 59(3):817–858, 1991.

D. W. Andrews. Similar-on-the-boundary tests for moment inequalities exist, but have poor power. Discussion Paper 1815R, Cowles Foundation, 2012.

D. W. Andrews and P. J. Barwick. Inference for parameters defined by moment inequalities: A recommended moment selection procedure. *Econometrica*, 80(6):2805–2826, 2012a.

D. W. Andrews and P. J. Barwick. Supplement to 'Inference for parameters defined by moment inequalities: A recommended moment selection procedure'. *Econometrica Supplemental Material*, 80(6), 2012b.

D. W. Andrews and J. C. Monahan. An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator. *Econometrica*, 60(4):953–966, 1992.

H. Bengtsson. *R.matlab: Read and write of MAT files together with R-to-MATLAB connectivity*, 2013. R package version 2.0.1, available at http://cran.r-project.org/package=R.matlab.

G. Calhoun. *dbframe: An R to SQL interface*, 2010. R package version 0.3.3.

D. Capocci and G. Hübner. Analysis of hedge fund performance. *Journal of Empirical Finance*, 11(1):55–89, 2004.

D. B. Dahl. *xtable: Export tables to LaTeX or HTML*, 2012. R package version 1.7-0, available at http://cran.r-project.org/package=xtable.

M. Getmansky, A. W. Lo, and I. Makarov. An econometric model of serial correlation and illiquidity in hedge fund returns. *Journal of Financial Economics*, 74(3):529–609, 2004.

J. J. Goeman and A. Solari. The sequential rejection principle of familywise error control. *The Annals of Statistics*, 38(6):3782–3810, 2010.

F. Gotze and H. Kunsch. Second-order correctness of the blockwise bootstrap for stationary observations. *The Annals of Statistics*, 24(5):1914–1933, 1996.

P. R. Hansen. A test for superior predictive ability. *Journal of Business and Economic Statistics*, 23(4):365–380, 2005.

K. Hirano and J. R. Porter. Impossibility results for nondifferentiable functionals. *Econometrica*, 80(4):1769–1790, 2012.

S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

P.-H. Hsu, Y.-C. Hsu, and C.-M. Kuan. Testing the predictive ability of technical analysis using a new stepwise test without data snooping bias. *Journal of Empirical Finance*, 17(3):471–484, 2010.

G. W. Imbens and C. F. Manski. Confidence intervals for partially identified parameters. *Econometrica*, 72(6):1845–1857, 2004.

D. A. James. *RSQLite: SQLite interface for R*, 2012. R package version 0.11.2.

R. Kosowski, A. Timmermann, R. Wermers, and H. White. Can mutual fund "stars" really pick stocks? new evidence from a bootstrap analysis. *The Journal of Finance*, 61(6): 2551–2595, 2006.

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

A. Lo and A. MacKinlay. Data-snooping biases in tests of financial asset pricing models. *Review of Financial Studies*, 3(3):431–467, 1990.

A. W. Lo. The statistics of sharpe ratios. *Financial Analysts Journal*, 58(4):36–52, 2002.

A. McCloskey. Bonferroni-based size-correction for nonstandard testing problems. Working Papers 2012-16, Brown University, Department of Economics, 2012.

D. N. Politis and J. P. Romano. A circular block resampling procedure for stationary data. In R. Page and R. LePage, editors, *Exploring the Limits of Bootstrap*, pages 263–270. John Wiley, New York, 1992.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, http://www.r-project.org/, Vienna, Austria, 2012.

J. P. Romano and M. Wolf. Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4):1237–1282, 2005a.

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

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

J. P. Romano and M. Wolf. Balanced control of generalized error rates. *The Annals of Statistics*, 38(1):598–633, 2010.

J. P. Romano, A. M. Shaikh, and M. Wolf. Formalized data snooping based on generalized error rates. *Econometric Theory*, 24(2):404–447, 2008.

J. P. Romano, A. M. Shaikh, and M. Wolf. A simple two-step method for testing moment inequalities with an application to inference in partially identified models. Working Paper, 2012.

C. Sharpsteen and C. Bracken. *tikzDevice: A Device for R Graphics Output in PGF/TikZ Format*, 2012. R package version 0.6.3/r49.

J. Stoye. More on confidence intervals for partially identified parameters. *Econometrica*, 77(4):1299–1315, 2009.

R. Sullivan, A. Timmermann, and H. White. Dangers of data mining: The case of calendar effects in stock returns. *Journal of Econometrics*, 105(1):249–286, 2001.

D. Temple Lang. *Combinations: Compute the combinations of choosing r items from n elements.*, 2010. R package version 0.2-0.

W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, 4th edition, 2002. R package version 7.3.22.

H. White. A reality check for data snooping. *Econometrica*, 68(5):1097–1126, 2000.

A. Zeileis. Econometric computing with HC and HAC covariance matrix estimators. *Journal of Statistical Software*, 11(10):1–17, 2004. R package version 2.2-10.

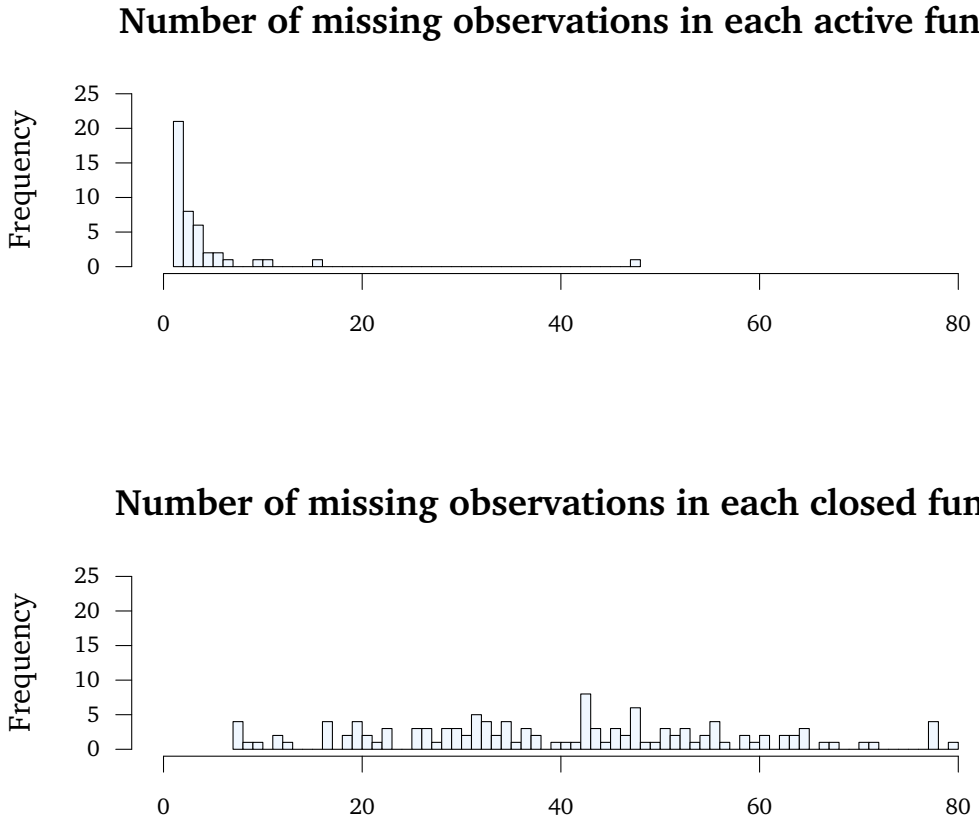

Figure 1: Histogram of missing observations: the horizontal axis depicts the number of observations reported as missing and the vertical axis is the number of funds with that many observations missing.

DGP	S	Out-performance	Under-performance	Equal Performance
1	2			$\mu_1 = \mu_2 = 1$
2	40			$\mu_1 = \cdots = \mu_{40} = 1$
3	40	$\mu_1 = \cdots = \mu_6 = 1.4$		$\mu_7 = \cdots = \mu_{40} = 1$
4	40	$\mu_1 = \cdots = \mu_6 = 1.4$	$\mu_7 = \cdots = \mu_{40} = -1$	
5	40	$\mu_1 = \cdots = \mu_{20} = 1.4$		$\mu_{21} = \cdots = \mu_{40} = 1$
6	40	$\mu_1 = \cdots = \mu_{20} = 1.4$	$\mu_{21} = \cdots = \mu_{40} = -1$	
7	4			$\mu_1 = \cdots = \mu_4 = 1$
8	4	$\mu_1 = \mu_2 = 1.4$		$\mu_3 = \mu_4 = 1$
9	4	$\mu_1 = \mu_2 = 1.4$	$\mu_3 = \mu_4 = -1$	

Table 1: Parameters for Monte Carlo experiments.

		Familywise error rate (%)			Average # discoveries			
# Obs.	Type	Ours	StepM	SPA	Ours	StepM	SPA	# False
50	1	5.0	5.0	9.8				0
	2	5.1	5.1	5.1				0
	3	1.6	1.6	1.6	0.8	0.8	0.8	6
	4	0.0	0.0	0.0	2.0	0.7	2.0	6
	5	3.2	3.2	3.2	2.7	2.7	2.7	20
	6	0.0	0.0	0.0	4.3	2.8	4.4	20
100	1	4.7	4.7	7.3				0
	2	4.6	4.6	4.6				0
	3	1.7	1.7	1.7	2.2	2.2	2.2	6
	4	0.0	0.0	0.0	4.1	2.1	4.1	6
	5	4.7	4.7	4.7	7.5	7.5	7.6	20
	6	0.0	0.0	0.0	10.7	7.7	10.7	20

Table 2: Results of the first Monte Carlo experiment—control of FWE. The columns under the heading "Familywise error rate (%)" present results for our refinement of the StepM (under "Ours"), Romano and Wolf's (2005a) original StepM ("StepM") and Hsu et al.'s (2010) Step-SPA ("SPA"). The columns under the heading "Average # discoveries" follow the same naming convention. The column "# False" lists the number of false hypotheses for that DGP for convenience. These results are based on 1000 simulations for each DGP using 999 bootstrap replications and the nominal FWE is 5%.

		k-familywise error rate (%)		Average # discoveries		
# Obs.	Type	Ours	k-StepM	Ours	k-StepM	# False
50	2	4.1	4.1			0
	3	0.3	0.3	2.2	2.2	6
	4	0.0	0.0	4.2	1.7	6
	5	3.6	3.6	6.4	6.4	20
	6	0.0	0.0	10.3	6.6	20
100	2	4.6	4.6			0
	3	0.2	0.1	3.9	3.9	6
	4	0.3	0.0	5.6	3.5	6
	5	4.4	4.3	12.3	12.3	20
	6	0.0	0.0	16.4	12.5	20

Table 3: Results of second Monte Carlo experiment—control of k-FWE with $k = 3$. The columns under the heading "k-familywise error rate (%)" present results for our extension of the k-StepM ("Ours") and Romano and Wolf's (2005a) original StepM ("k-StepM"). The columns under the heading "Average # discoveries" follow the same naming convention. The column "# False" lists the number of false hypotheses for that DGP for convenience. These results are based on 1000 simulations for each DGP using 999 bootstrap replications and the nominal k-FWE is 5%.

		Size (%)			Power (%)			
# Obs.	Type	Ours	Bon.	AQLR	Ours	Bon.	AQLR	# False
100	1	4.7	4	4.4				0
	8	4.7	4	4.5				0
	9				86.9	85.3	86.9	2
	10				91.6	90.8	91.7	2

Table 4: Results of third Monte Carlo experiment—control of size when testing composites of one-sided hypotheses. The columns under the heading "Size (%)" present results for our method in Corollary 1 ("Ours"), McCloskey's (2012) and Romano et al.'s (2012) Bonferroni-based procedure ("Bon."), and Andrews and Barwick's (2012a) AQLR ("AQLR"). The columns under the heading "Power (%)" follow the same naming convention. The column "# False" lists the number of false hypotheses for that DGP for convenience. These results are based on 1000 simulations for each DGP using 999 bootstrap replications and the nominal size is 5%.

	LB (%)	Avg. (%)	t-stat
Stenham Trading Portfolio Inc	0.05	5.34	3.37
Libra Fund LP	0.51	20.11	3.58
Otter Creek Partners I LP	0.51	7.53	3.93
Longfellow Merger Arbitrage	0.72	4.41	4.42
CISDM merger.arbitrage *	0.79	4.65	4.45
GAM Trading USD	0.80	6.24	4.24
Loeb Arbitrage Fund L.C.	0.86	7.50	4.17
High Sierra Partners I	0.93	9.71	4.09
TIG Arbitrage Associates Ltd *	1.44	4.46	5.45
Gabelli Associates Limited *	1.44	4.77	5.29
Momentum AssetMaster I USD *	1.87	6.65	5.69
CISDM equity.market.neutral *	1.88	4.42	7.55
Equity Income Partners LP *	2.43	5.00	8.45
Bryn Mawr Capital, L.P. *	3.55	7.68	8.05
Millennium International Ltd *	4.41	11.33	7.10
Millennium USA LP Fund *	4.48	11.45	7.13

Table 5: Active funds that outperform the risk-free rate over 1994–2011 (5% FDP for $\gamma = 0.1$). The column labeled "LB (%)" lists the lower bound on the expected excess return of the corresponding confidence interval for each fund. The column labeled "Avg. (%)" lists the funds' average return over the time period. And the column labeled "t-stat" lists the test statistic for the one-sided hypothesis test. Data are from the CISDM database which includes 332 active and closed funds. Funds marked with an asterisk (*) were found to outperform the risk-free rate when controlling the FWE at 5% as well.

	LB (%)	Avg. (%)	t-stat
Archstone Partners (Onshore & Offshore)	0.51	6.64	3.64
Balboa LP	0.51	7.90	3.61
Haberman Value Fund	0.51	5.57	3.67
Key Group Investors LP	0.51	5.50	3.82
Geewax Domestic Long/Short Alpha	0.78	8.68	4.06
Prism Partners I L.P. (Domestic)	0.81	7.82	4.12
Coast Enhanced Income Fund II Ltd *	1.87	2.49	14.87
Rainbow Fund Ltd *	1.87	7.17	5.51
Ultra Distressed Securities Fund LP *	1.87	7.97	5.31
Fairfield Sentry Ltd *	3.86	5.54	14.31
Greenwich Sentry LP *	4.32	6.28	13.88

Table 6: Closed funds that outperform the risk-free rate over 1994–2011 (5% FDP for $\gamma = 0.1$). The column labeled "LB (%)" lists the lower bound on the expected excess return of the corresponding confidence interval for each fund. The column labeled "Avg. (%)" lists the funds' average return over the time period. And the column labeled "t-stat" lists the test statistic for the one-sided hypothesis test. Data are from the CISDM database which includes 332 active and closed funds. Funds marked with an asterisk (*) were found to outperform the risk-free rate when controlling the FWE at 5% as well.