Journal of Financial Econometrics, 2014, Vol. 12, No. 4, 730–755

A Generalized Stepwise Procedure with Improved Power for Multiple Inequalities Testing

YU-CHIN HSU
Institute of Economics, Academia Sinica

CHUNG-MING KUAN
National Taiwan University

MENG-FENG YEN
Graduate Institute of Banking and Finance, National Cheng Kung University

ABSTRACT

We propose a stepwise test, Step-SPA(k), for multiple inequalities testing. This test is analogous to the Step-SPA test of Hsu, Hsu, and Kuan (2010, J. Empirical Econ., **17**, 471–484) but has asymptotic control of a generalized familywise error rate: the probability of at least k false rejections. This test improves Step-RC(k) of Romano and Wolf (2007, Ann. Stat., **35**, 1378–1408) by avoiding the least favorable configuration used in Step-RC(k). We show that the proposed Step-SPA(k) test is consistent and more powerful than Step-RC(k) under any power notion defined in Romano and Wolf (2005, Econometrica, **73**, 1237–1282). An empirical study on Commodity Trading Advisor fund performance is then provided to illustrate the Step-SPA(k) test. Finally, we extend Step-SPA(k) to a procedure that asymptotically controls the false discovery proportion, the ratio of the number of false rejections over the number of total rejections, and show that it is more powerful than the corresponding procedure proposed by Romano and Wolf (2007, Ann. Stat., **35**, 1378–1408). (*JEL*: C12, C52)

KEYWORDS: data snooping, false discovery proportion, familiywise error rate, least favorable configuration, multiple inequalities testing, Reality Check, SPA test

We would like to thank the editor and two anonymous referees for their constructive comments. We are indebted to Stephen G. Donald, Po-Hsuan Hsu, Robert P. Lieli, Marc Paolella, and Hal White for very helpful discussions on this topic. We also thank Yi-Ting Yeh for excellent research assistance. C.-M. Kuan gratefully acknowledges the research support from National Science Council of Taiwan (NSC97-2415-H-002-217-MY3). Address correspondence to C.-M. Kuan, Department of Finance, National Taiwan University, Taipei 106, Taiwan, or e-mail: ckuan@ntu.edu.tw

doi:10.1093/jjfinec/nbu014

In a multiple hypotheses testing problem, it is often of interest to identify as many false null hypotheses as possible while accounting for the data-snooping effect. For example, among a given set of models such as portfolios, mutual funds, hedge funds or trading rules, one would like to examine if there are some models having superior performance relative to a benchmark. Then, data snooping may arise because, when many models are evaluated *individually*, some are bound to be superior by chance alone even though they are not. To avoid data snooping in multiple hypotheses testing, Romano and Wolf (2005) extend the Reality Check (RC) of White (2000) and propose a stepwise RC test (Step-RC) that is capable of identifying significant models while controlling the familywise error rate (FWER), the probability of at least one false rejection.

When the hypotheses involve inequality constraints, it is well known that Step-RC is conservative because it is based on the least favorable configuration (LFC). As discussed in Hansen (2005), a test based on the LFC may lose power dramatically when many "poor" models are included in the test. To circumvent this problem, Hsu, Hsu, and Kuan (2010) adopt the re-centering method in the "superior predictive ability" (SPA) test of Hansen (2005) that is able to remove those poor models from consideration asymptotically. This, together with the stepwise procedure in Step-RC, leads to a stepwise SPA test (Step-SPA). It has been shown that Step-SPA is more powerful than Step-RC, especially when "poor" models are present; see Hsu, Hsu, and Kuan (2010) for details. Both RC- and SPA-type tests have been widely applied in empirical studies, e.g., Sullivan, Timmermann, and White (1999, 2001), Hansen and Lunde (2005), Hsu and Kuan (2005), Qi and Wu (2006), Yen and Hsu (2010), and Hsu, Kuan, and Yen (2013).

Although Step-SPA is already an improvement over Step-RC, its ability to identify significant models is still limited due to the control of the FWER. When multiple testing involves a large number of hypotheses, incorrectly rejecting a few of them may not be a very serious problem in practice. If so, controlling only one false rejection poses a very stringent criterion. In view of this, one may lower the rejection criterion and hence increase the test power by tolerating more false rejections. Romano and Wolf (2007) thus propose Step-RC(k) that asymptotically controls the FWER(k), the probability of k or more false rejections, where $k \geq 2$. Note that Step-RC(k) is still a conservative test due to its dependence on LFC.

In this article, we contribute to the literature by proposing Step-SPA(k), that is analogous to Step-SPA of Hsu, Hsu, and Kuan (2010) but has asymptotic control of the FWER(k). Just as Step-SPA improves the power of Step-RC, the proposed test is an improvement of Step-RC(k) of Romano and Wolf (2007) because it also employs the re-centering method of Hansen (2005). We show that the proposed Step-SPA(k) is consistent in that it can identify the violated null hypotheses with probability approaching one. It is also shown analytically and by simulations that Step-SPA(k) is more powerful than Step-RC(k) under any power notion defined in Romano and Wolf (2005). An empirical study on Commodity Trading Advisor (CTA) fund performance is then provided to illustrate the proposed test. Finally, we extend Step-SPA(k) to a procedure that asymptotically controls the false discovery

proportion (FDP), the ratio of the number of false rejections over the number of total rejections, and show that it is more powerful than the corresponding procedure proposed by Romano and Wolf (2007).

The paper proceeds as follows. We summarize existing tests for multiple inequalities testing in Section 1. We propose the Step-SPA(k) test and show that it can control the FWER(k) asymptotically and is more powerful than Step-RC(k) in Section 2. The simulation results are reported in Section 3. The empirical results are discussed in Section 4. We then propose a procedure that controls the FDP and proves it power advantage in Section 5. Section 6 concludes the article. All proofs are deferred to Appendix.

1 TESTS WITHOUT DATA SNOOPING BIAS

In this section, we review some existing stepwise tests without the data-snooping bias: Step-RC of Romano and Wolf (2005), Step-SPA of Hsu, Hsu, and Kuan (2010), and Step-RC(k) of Romano and Wolf (2007). For some reviews on the tests without data-snooping bias, we refer to Dudoit, Shaffer, and Boldrick (2003), and Romano, Shaikh, and Wolf (2008).

Let θ_ℓ be a performance measure of model ℓ, $\ell=1,\ldots,m$. For example, θ_ℓ may be the CAPM alpha of the ℓ-th portfolio (mutual fund, hedge fund) or the sample mean of the realized return of the ℓ-th technical trading rule. We are interested in knowing the portfolios (mutual funds, hedge funds) that have a positive CAPM alpha or the trading rules that generate positive mean returns. That is, we want to identify the set: $\mathcal{L}^+ \equiv \{\ell:\ \theta_\ell > 0\}$. This amounts to testing the following inequality constraints:

$$H_0^\ell : \theta_\ell \le 0, \quad \ell = 1,\ldots,m. \tag{1}$$

Data snooping may arise when models are tested individually but without a proper control of the probability of false rejections. Thus, one may find some models with positive θ_ℓ by chance alone, even though they are not. To be specific, suppose there are 100 models that are mutually independent, and we apply a t-test to each model with the significance level 5%. The probability of falsely rejecting at least one correct null hypothesis is $1-(0.95)^{100} \approx 0.994$. It is thus highly likely that an individual test may incorrectly suggest an inferior model to be a significant one. Therefore, an appropriate method that can control such data-snooping bias is needed to avoid spurious inference when many models are examined together.

1.1 Assumptions

Let $\hat{\theta}_n = (\hat{\theta}_{1,n},\ldots,\hat{\theta}_{m,n})'$ be an estimator for $\theta = (\theta_1,\ldots,\theta_m)'$. We first make high-level assumptions on $\hat{\theta}_n$.

Assumption 1: *Assume the following conditions hold.*

1. $\sqrt{n}(\hat{\theta}_n-\theta)\xrightarrow{d}\mathcal{N}(0,\Omega)$, *where* Ω *is the* $m\times m$ *asymptotic covariance matrix of* $\hat{\theta}_n$, *with the* (i,j)*-th element* ω_{ij}. *For some* $\delta>0$, *the diagonal elements are* $\omega_{jj}=\sigma_j^2\geq\delta$, $j=1,\ldots,m$.
2. *There exists a consistent estimator* $\widehat{\Omega}_n$ *for* Ω *whose* (i,j)*-th element is* $\hat{\omega}_{ij,n}$ *such that* $\hat{\omega}_{ij,n}\xrightarrow{p}\omega_{ij}$, $i,j=1,\ldots,m$.
3. $\sqrt{n}\widehat{\Lambda}_n^{-1}(\hat{\theta}_n-\theta)\xrightarrow{d}\mathcal{N}(0,\Xi)$, *where* $\widehat{\Lambda}_n=diag(\hat{\sigma}_{1,n},\ldots,\hat{\sigma}_{m,n})$, $\hat{\sigma}_{j,n}=\sqrt{\hat{\omega}_{jj,n}}$, *and the* (i,j)*-th element of* Ξ *is* $\xi_{ij}=\omega_{ij}/(\sigma_i\sigma_j)$, *and*

$$\widehat{\Xi}_n=\widehat{\Lambda}_n^{-1}\widehat{\Omega}_n\widehat{\Lambda}_n^{-1}\xrightarrow{p}\Xi.$$

This assumption is not restrictive. Assumption 1(i) requires that $\hat{\theta}_n$ is $\sqrt{n}$-consistent and asymptotically normal with the asymptotic covariance matrix Ω. This usually holds under suitable regularity conditions; see, e.g., White (2001) for the results in the context of OLS estimation. Assumption 1(ii) requires a consistent estimator $\widehat{\Omega}$ for Ω, which may be computed as a HAC (heteroskedasticity and autocorrelation consistent) estimator; see Newey and West (1987) or Andrews (1991) for details. Assumption 1(iii) is in fact implied by Assumption 1(i)–(ii); we state it as an assumption here for simplicity.

For $\mathcal{N}(0,\Xi)$ in Assumption 1(iii), we also assume it can be well approximated by a simulated distribution $\Psi_n^u=(\psi_{1,n}^u,\ldots,\psi_{m,n}^u)'$.

Assumption 2: $\Psi_n^u\xrightarrow{d}\mathcal{N}(0,\Xi)$ *conditional on the sample path with probability one.*

There are various methods to obtain Ψ_n^u. One may generate Ψ_n^u by drawing samples from the pseudo random variable $\mathcal{N}(0,\widehat{\Xi}_n)$ that is independent of the sample. Given the consistency of $\widehat{\Xi}_n$, the simulated distribution would satisfy Assumption 2. One may also approximate $\mathcal{N}(0,\Xi)$ by a proper bootstrap method; see, e.g., White (2000), Hansen (2005), Romano and Wolf (2005, 2007), and Hsu, Hsu, and Kuan (2010) among others.

1.2 Step-RC

To account for potential data snooping, one needs to control a proper error measure. A leading measure is

$$\text{FWER}=\text{P}\big[\text{reject at least one true hypothesis}\big]. \tag{2}$$

Romano and Wolf (2005) propose Step-RC that, while controlling the FWER asymptotically, is able to identify many models that significantly deviate from the null hypotheses.

Step-RC proceeds as follows. Let $\widehat{T}_{\ell,n}=\hat{\theta}_{\ell,n}/\hat{\sigma}_{\ell,n}$ be the standardized test statistic for H_0^{ℓ}.[1] For $0<\alpha<1$ and for any subset $K\subseteq\{1,\ldots,m\}$, let $\tilde{c}_{n,K}(\alpha,1)$ be the α-th quantile of $\max\{\psi_j^u : j\in K\}$, where $\{\psi_j^u : j\in K\}$ are the simulated distributions that satisfy Assumption 2. We set

$$\hat{c}_{n,K}(\alpha,1)=\max\{\tilde{c}_{n,K}(\alpha,1),0\}.$$

To implement Step-RC with asymptotic FWER control at α, we re-arrange $\widehat{T}_{\ell,n}$'s in descending order. A top model ℓ would be rejected if $\sqrt{n}\widehat{T}_{\ell,n}$ is greater than $\hat{c}_{n,A_1}(1-\alpha,1)$, where $A_1=\{1,\ldots,m\}$. If none of the null hypotheses is rejected, the process stops; otherwise, we remove $\widehat{T}_{\ell,n}$ of the rejected models from the data. The index set of the remaining models is denoted as A_2 ($A_2\subset A_1$). We then re-calculate the critical values using the remaining data and obtain $\hat{c}_{n,A_2}(1-\alpha,1)$. A top model i would be rejected if $\sqrt{n}\widehat{T}_{i,n}$ is greater than $\hat{c}_{n,A_2}(1-\alpha,1)$. The procedure continues till no more model can be rejected.

It should be emphasized that, given that the null hypothesis (1) involves inequality constraints, RC obtains its null distribution by choosing the LFC, i.e., $\theta_\ell=0$ for $\ell\in A_1$. The distribution at each step of Step-RC is also based on the LFC. The resulting tests are thus conservative in that the asymptotic size can be strictly smaller than the significance level.

1.3 Step-SPA

Hsu, Hsu, and Kuan (2010) propose Step-SPA that improves the power of Step-RC by invoking the re-centering method of Hansen (2005). Let $\{a_n\}$ be a sequence of positive numbers such that $\lim_{n\to\infty}a_n=\infty$ and $\lim_{n\to\infty}n^{-1/2}a_n=0$. For each ℓ, define $\hat{\mu}_{\ell,n}$ as

$$\hat{\mu}_{\ell,n}=\widehat{T}_{\ell,n}\cdot 1\left(\sqrt{n}\widehat{T}_{\ell,n}\le -a_n\right), \tag{3}$$

where $1(\cdot)$ denotes the indicator function. For any subset $K\subseteq\{1,\ldots,m\}$, let $\hat{q}_{n,K}(\alpha,1)=\max\{\tilde{q}_{n,K}(\alpha,1),0\}$ where $\tilde{q}_{n,K}(\alpha,1)$ is the α-th quantile of $\max\{\psi_j^u+\sqrt{n}\hat{\mu}_{j,n} : j\in K\}$. The procedure of Step-SPA is identical to that of Step-RC, except that the RC critical values $\hat{c}_{n,K}(\alpha,1)$ are replaced by the SPA critical values $\hat{q}_{n,K}(\alpha,1)$. Hsu, Hsu, and Kuan (2010) show that Step-SPA is more powerful than Step-RC under any power notion defined in Romano and Wolf (2005) while still controlling the asymptotic FWER well.

The re-centering method works as follows. If θ_k, $k\in A_j$, is strictly less than zero, then one can show that $\hat{\theta}_{k,n}$ will not contribute to the null distribution

[1] Even though the standardized test is not uniformly more powerful than nonstandardized tests (Donald and Hsu 2011), it possesses some advantages; see, e.g., Hansen (2005) and Romano and Wolf (2005).

of $\max_{\ell\in A_j}\{\sqrt{n}\widehat{T}_{\ell,n},0\}$. By adding $\sqrt{n}\hat{\mu}_{k,n}$ that diverges to negative infinity with probability one to the simulated distribution ψ_k^u, one can asymptotically remove the k-th model from consideration so as to lower the critical values and hence improve the power of the test. This method has also been adopted in Donald and Hsu (2013), which is similar to the generalized moment selection method of Andrews and Soares (2010) and Andrews and Shi (2013) and the contact set method of Linton, Song, and Whang (2010). These methods are all used to obtain tests with better power than those based on LFC.

Note that the theory for Step-SPA works as long as a_n satisfies that $\lim_{n\to\infty}a_n=\infty$ and $\lim_{n\to\infty}n^{-1/2}a_n=0$. In this article, we follow Hansen (2005) and Hsu, Hsu, and Kuan (2010) and recommend $a_n=\sqrt{2(\log\log n)}$. Andrews and Soares (2010), on the other hand, recommend $a_n=\sqrt{\log n}$.

1.4 Step-RC(k)

When the number of hypotheses is large, the control of only one false rejection becomes a stringent criterion such that the resulting test has limited ability of identifying false hypotheses in finite samples. The test power may be increased by allowing for more than one false rejection. That is, we may relax the FWER control to the FWER(k) control:

$$\text{FWER}(k)=\text{P}\big[\text{reject at least } k \text{ true hypotheses}\big]. \tag{4}$$

Clearly, when $k=1$, this measure reduces to the FWER given in (2). Step-RC(k) of Romano and Wolf (2007) is a test that achieves the asymptotic control of the FWER(k) and also an improvement of the original Step-RC.

We summarize the procedure of Step-RC(k) below, which is a slightly modified version of Algorithm 2.1 of Romano and Wolf (2007). We need some more notation. Let $\mathcal{Y}\equiv\{y_j|\ j=1,\ldots,J\}$ be a collection of real numbers. Then for $k\le J$, k-max$\{\mathcal{Y}\}$ denotes the k-th largest value of $\mathcal{Y}$. For example, if the elements in $\mathcal{Y}$ are ordered as $y_{(1)}\ge\cdots\ge y_{(J)}$, then k-max$\{\mathcal{Y}\}=y_{(k)}$. For any subset $K\subseteq\{1,\ldots,m\}$, let $\hat{c}_{n,K}(\alpha,k)=\max\{\tilde{c}_{n,K}(\alpha,k),0\}$ where $\tilde{c}_{n,K}(\alpha,k)$ is the α-th quantile of k-max$\{\psi_j^u:j\in K\}$.

Algorithm 1: *(Step-RC(k))*

1. *Re-arrange* $\widehat{T}_{\ell,n}$ *in descending order.*
2. *Let* $A_1=\{1,\ldots,m\}$ *and* $\hat{d}_{n,A_1}(1-\alpha,k)=\hat{c}_{n,A_1}(1-\alpha,k)$. *If*

$$\max\{\sqrt{n}\widehat{T}_{\ell,n}:\ell\in A_1\}\le\hat{d}_{n,A_1}(1-\alpha,k),$$

then accept all hypotheses and stop; otherwise, reject H_ℓ^0 *if* $\sqrt{n}\widehat{T}_{\ell,n}>\hat{d}_{n,A_1}(1-\alpha,k)$ *and continue.*

3. *Let R_2 be the collection of the indices ℓ of the rejected hypotheses H_ℓ^0 in the previous step, and let A_2 be the collection of the indices of the remaining hypotheses. If $|R_2|<k$, then stop; otherwise, let*

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

Reject H_ℓ^0 with $\ell\in A_2$ such that $\sqrt{n}\widehat{T}_{\ell,n}>\hat{d}_{n,A_2}(1-\alpha,k)$. If there is no further rejection, stop; otherwise, go to next step.

4. *Repeat the previous step (with R_2 and A_2 replaced by R_j and A_j, $j\geq 3$) till there is no further rejection.*

Note that when $k>1$, the rejected hypotheses may still stay in the algorithm. The reason is that after the first step, it is possible that some true null hypotheses might have been rejected, but hopefully there are (at most) $k-1$ of them. Because we have no idea which of the rejected hypotheses are true or false, we need to consider all possible subsets of $k-1$ rejected hypotheses in determining the critical values. Once we can control the FWER(k) at each step, the stepwise procedure would also control the FWER(k). It can also be verified that the critical values in the last step of Step-RC(k) is no greater than that of Step-RC. As such, all models rejected by Step-RC will also be rejected by Step-RC(k) but not conversely.

Note also that in the original procedure of Romano and Wolf (2007), the critical value is $\tilde{c}_{n,K}(\alpha,k)$, instead of $\hat{c}_{n,K}(\alpha,k)$. A drawback of their choice is that some hypotheses with non-positive statistics may be rejected, because $\tilde{c}_{n,K}(\alpha,k)$ may be strictly negative with a positive probability. We consider this an undesirable property because a negative statistic should not be viewed as an evidence for the alterative hypothesis. In contrast, Algorithm 1 is based on $\hat{c}_{n,K}(\alpha,k)$ and hence can never reject any hypothesis with a non-positive statistic.

2 PROPOSED STEPWISE TEST

In this section, we introduce Step-SPA(k) and establish its asymptotic properties. As an extension of Step-SPA of Hsu, Hsu, and Kuan (2010), Step-SPA(k) achieves the asymptotic control of the FWER(k). Step-SPA(k) is also an improvement of Step-RC(k) because it avoids the LFC by invoking the re-centering method of Hansen (2005).

For any subset $K\subseteq\{1,\ldots,m\}$, let the $\hat{q}_{n,K}(\alpha,k)=\max\{\tilde{q}_{n,K}(\alpha,k),0\}$ where $\tilde{q}_{n,K}(\alpha,k)$ is the α-th quantile of k-max$\left\{\psi_j^u+\sqrt{n}\hat{\mu}_j:j\in K\right\}$. The algorithm of Step-SPA(k) stated below is virtually the same as Algorithm 1 for Step-RC(k), except that we replace $\hat{c}_{n,K}(\alpha,k)$ with $\hat{q}_{n,K}(\alpha,k)$.

Algorithm 2: *(Step-SPA(k))*

1. *Re-arrange the $\widehat{T}_{\ell,n}$'s in descending order.*

2. *Let* $A_1=\{1,\ldots,m\}$ *and* $\hat{w}_{n,A_1}(1-\alpha,k)=\hat{q}_{n,A_1}(1-\alpha,k)$. *If*

$$\max\{\sqrt{n}\widehat{T}_{\ell,n}:\ell\in A_1\}\leq\hat{w}_{n,A_2}(1-\alpha,k),$$

then accept all hypotheses and stop; otherwise, reject H_ℓ^0 *if* $\sqrt{n}\widehat{T}_{\ell,n}>\hat{w}_{n,A_1}(1-\alpha,k)$ *and continue.*

3. *Let* R_2 *be the collection of the indices* ℓ *of the rejected hypotheses* H_ℓ^0 *in the previous step, and let* A_2 *be the collection of the indices of the remaining hypotheses. If* $|R_2|<k$, *then stop; otherwise, let*

$$\hat{w}_{n,A_2}(1-\alpha,k)=\max_{I\subset R_2,|I|=k-1}\left\{\hat{q}_{n,K}(1-\alpha,k):K=A_2\cup I\right\}.$$

Reject H_ℓ^0 *with* $\ell\in A_2$ *such that* $\sqrt{n}\widehat{T}_{\ell,n}>\hat{w}_{n,A_2}(1-\alpha,k)$. *If there is no further rejection, stop; otherwise, go to next step.*

4. *Repeat the previous step (with* R_2 *and* A_2 *replaced by* R_j *and* A_j, $j\geq 3$*) till there is no further rejection.*

Clearly, Step-SPA(k) reduces to Step-SPA when $k=1$. It is straightforward to see that $\hat{w}_{n,K}(1-\alpha,k)$ satisfies the monotonicity requirement discussed in Romano and Wolf (2007), because by construction, for any $K_1\subseteq K_2$, $\hat{w}_{n,K_1}(\alpha,k)\leq\hat{w}_{n,K_2}(\alpha,k)$. Let $I(P)$ be the set of the indices of the true null hypotheses. In Appendix, we show that Algorithm 2 satisfies the size control requirement in Romano and Wolf (2007), i.e.,

$$\lim_{n\to\infty}\mathrm{P}\left[k\text{-max}\left\{\sqrt{n}\widehat{T}_{\ell,n}:\ell\in I(P)\right\}>\hat{q}_{n,I(P)}(1-\alpha,k)\right]\leq\alpha.$$

Then in the light of Theorem 2.1 of Romano and Wolf (2007), Step-SPA(k) has the asymptotic FWER(k) control. Note that if $\theta_\ell>0$, then $\sqrt{n}\widehat{T}_{\ell,n}\to\infty$ in probability, whereas the critical value $\hat{q}_{n,A_1}(1-\alpha,k)$ is bounded in probability. Thus, any superior model will be rejected in the first step with probability approaching one. This establishes the consistency of Step-SPA(k). These results are summarized below.

Theorem 1: *The following results hold under Assumptions 1 and 2:*

1. *Given the pre-specified level* α, *the asymptotic FWER(k) of Step-SPA(k) will be less than or equal to* α.
2. *The hypothesis* H_0^ℓ *with* $\theta_\ell>0$ *will be rejected by Step-SPA(k) with probability approaching one.*

As discussed in Section 1.4, we want to avoid rejecting any hypothesis with a strictly negative statistic. This would not happen in Step-SPA(k) defined in Algorithm 2 because $\hat{q}_{n,K}(\alpha,k)$ is greater than or equal to zero by definition. On the

other hand, if Step-SPA(k) is based on $\tilde{q}_{n,K}(\alpha,k)$, we can still show that it controls the FWER(k) and is consistent. Yet, Step-SPA(k) may reject a hypothesis with a strictly negative statistic because $\tilde{q}_{n,K}(\alpha,k)$ may be strictly negative even when α is very small.

Key result concerning the power properties of Step-SPA(k) follows.

Theorem 2: *Under Assumptions 1 and 2, Step-SPA(k) defined in Algorithm 2 is more powerful than Step-RC(k) in Algorithm 1 under the notions of power defined in Romano and Wolf (2005).*

Theorem 2 shows that the power of Step-SPA(k) in Algorithm 2 is greater than or equal to that of Step-RC(k) in Algorithm 1. In particular, we note that the null hypotheses rejected by Step-RC(k) will also be rejected by Step-SPA(k). This is so because, by construction, $\tilde{q}_{n,K}(\alpha,k) \leq \tilde{c}_{n,K}(\alpha,k)$. As a result, the critical values for Step-SPA(k), $\hat{q}_{n,K}(\alpha,k)$, must be less than or equal to those for Step-RC(k), $\hat{c}_{n,K}(\alpha,k)$.

3 SIMULATIONS

In this section, we report some simulation results of the proposed Step-SPA(k) test with $k=3$. For comparison, we also compute Step-RC, Step-RC(3), and Step-SPA.

In our simulations, we consider two random variables: $N(\mu,1)$ and $(t(4)/\sqrt{2}+\mu)$, where the latter also has variance 1. For each variable, there are S models (with different μ values), each with n i.i.d. observations. We set $S=100,200,500$ and $n=100,200,500$. This enables us to examine how different tests perform when the number of models is less than, equal to, or greater than the number of observations. These S models may be uncorrelated ($\rho=0$) or correlated ($\rho=0.2,0.4$). For model ℓ, we compute the standardized Step-SPA(3) statistic $\widehat{T}_{\ell,n}$, with the re-centering parameter $a_n=\sqrt{2\log(\log n)}$. The number of bootstraps for computing the critical values is $B=1000$; the number of replications for each simulation is $R=1000$. Note that all tests are based on 5% significance level.

To be more specific about our bootstrap, Ψ_n^u is defined as $\sqrt{n}\widehat{\Lambda}^{-1}(\hat{\theta}_n^b-\hat{\theta}_n)$, where $\hat{\theta}_n^b$ is calculated from each bootstrap sample formed by n random draws with replacement form the original data. Another approach is to calculate the standardized test statistic based on the bootstrap sample: $\sqrt{n}(\widehat{T}_n^b-\widehat{T}_n)$, where $\widehat{T}_n=\widehat{\Lambda}^{-1}\hat{\theta}_n$, and $\widehat{T}_n^b=(\widehat{\Lambda}^b)^{-1}\hat{\theta}_n^b$. To save computation time, we adopt the first method in the simulations. But we use the second method in the empirical study, because it is preferable to calculate $\widehat{\Lambda}^b$ from the bootstrap sample in practice; see Footnote 22 of Romano and Wolf (2005).

We first study the control of the FWER(3) under LFC by setting all models with $\mu=0$. Here we report only the FWER(3) results of Step-RC(3) and Step-SPA(3) in Tables 1 and 2 for models generated from, respectively, normal and $t(4)$ variables. It can be seen that, for models generated from normal random variables with $\rho=0$, these two tests have good control of the FWER(3) when the number of models S is

Table 1 Control of FWER(3) under LFC: Normal random variables with $\mu = 0$

	$S=100$			$S=200$			$S=500$		
	$n=100$	$n=200$	$n=500$	$n=100$	$n=200$	$n=500$	$n=100$	$n=200$	$n=500$
					Model correlation $\rho=0$				
Step-RC(3)	4.5	5.4	4.9	5.4	4.2	5.2	5.5	6.7	5.5
Step-SPA(3)	5.5	6.0	5.4	6.0	4.7	5.5	5.8	7.2	6.1
					Model correlation $\rho=0.2$				
Step-RC(3)	5.0	4.9	5.5	6.2	5.5	5.3	8.2	5.8	4.5
Step-SPA(3)	5.0	4.9	5.5	6.2	5.5	5.3	8.2	5.9	4.5
					Model correlation $\rho=0.4$				
Step-RC(3)	6.5	6.2	4.6	6.3	7.0	5.7	6.7	5.1	7.0
Step-SPA(3)	6.5	6.2	4.6	6.3	7.0	5.7	6.7	5.1	7.0

S is the number of models, n is the number of observations, and ρ is the correlation coefficient between models. Empirical FWER(3)'s are expressed in percentages; the nominal significance level is $\alpha = 5\%$.

Table 2 Control of FWER(3) under LFC: $t(4)$ random variables with $\mu = 0$

	$S=100$			$S=200$			$S=500$		
	$n=100$	$n=200$	$n=500$	$n=100$	$n=200$	$n=500$	$n=100$	$n=200$	$n=500$
					Model correlation $\rho=0$				
Step-RC(3)	3.5	2.9	2.9	2.9	3.1	2.4	3.2	1.9	3.6
Step-SPA(3)	4.0	3.3	3.3	3.2	3.3	2.8	3.2	2.0	3.6
					Model correlation $\rho=0.2$				
Step-RC(3)	4.9	4.6	4.7	4.2	5.6	4.6	4.6	4.2	4.4
Step-SPA(3)	5.0	4.6	4.7	4.2	5.6	4.6	4.6	4.2	4.4
					Model correlation $\rho=0.4$				
Step-RC(3)	4.8	4.3	5.2	5.1	5.9	4.7	5.1	4.8	4.2
Step-SPA(3)	4.8	4.3	5.2	5.1	5.9	4.7	5.1	4.8	4.2

S is the number of models, n is the number of observations, and ρ is the correlation coefficient between models. Empirical FWER(3)'s are expressed in percentages; the nominal significance level is $\alpha = 5\%$.

less than or equal to the number of observations n, yet they tend to over-reject when $S > n$. The control of the FWER(3) is adversely affected by model correlation ($\rho = 0.4$). For models generated from $t(4)$ variables which have fatter tails than $N(0,1)$, both tests have better control of the FWER(3). Although these tests may under-reject when $\rho = 0$, their FWER(3) are quite close to 5% when models are correlated.

Table 3 Average power performance and control of FWER: Normal random variables

	$S=100$			$S=200$			$S=500$		
Model correlation $\rho=0$									
	$n=100$	$n=200$	$n=500$	$n=100$	$n=200$	$n=500$	$n=100$	$n=200$	$n=500$
Step-RC	7.4	22.9	73.0	5.4	17.3	66.7	3.3	12.1	58.1
(FWER)	(1.0)	(0.6)	(0.4)	(0.9)	(0.6)	(0.8)	(1.3)	(1.1)	(0.7)
Step-SPA	12.9	33.4	82.7	9.3	26.0	77.0	5.7	18.7	68.6
(FWER)	(2.5)	(1.9)	(1.8)	(1.8)	(1.6)	(2.9)	(3.8)	(2.5)	(2.4)
Step-RC(3)	25.8	54.3	93.2	18.6	43.9	89.8	11.7	32.5	83.1
(FWER(3))	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.1)	(0.0)	(0.0)	(0.0)
Step-SPA(3)	43.3	75.9	98.5	32.8	64.9	97.2	21.0	50.1	94.1
(FWER (3))	(0.5)	(1.0)	(1.8)	(0.3)	(0.4)	(3.3)	(2.8)	(0.3)	(2.8)
Model correlation $\rho=0.2$									
Step-RC	8.1	24.2	75.1	6.2	18.4	69.0	3.8	13.1	59.7
(FWER)	(0.8)	(0.7)	(0.3)	(0.8)	(0.7)	(0.7)	(1.0)	(0.5)	(0.8)
Step-SPA	13.3	35.0	84.2	10.0	26.8	78.4	6.3	19.4	69.9
(FWER)	(2.8)	(2.0)	(1.6)	(1.9)	(2.0)	(1.7)	(2.7)	(1.1)	(1.9)
Step-RC(3)	21.4	48.0	91.0	15.8	37.5	86.5	10.1	27.4	79.1
(FWER(3))	(0.2)	(0.1)	(0.1)	(0.1)	(0.1)	(0.0)	(0.0)	(0.4)	(0.2)
Step-SPA(3)	36.8	69.3	97.8	27.6	56.6	95.7	17.7	42.5	91.3
(FWER(3))	(1.6)	(1.6)	(2.4)	(1.0)	(1.9)	(4.0)	(1.2)	(1.8)	(3.3)
Model correlation $\rho=0.4$									
Step-RC	9.2	29.1	77.4	7.9	23.1	71.9	5.0	16.8	65.3
(FWER)	(0.7)	(0.6)	(0.9)	(1.7)	(0.9)	(0.5)	(1.5)	(0.7)	(1.0)
Step-SPA	14.6	39.0	85.2	12.0	31.6	80.3	7.7	23.3	73.9
(FWER)	(2.2)	(2.7)	(2.0)	(3.3)	(2.4)	(2.0)	(2.5)	(2.3)	(2.6)
Step-RC(3)	20.8	47.8	89.9	16.7	39.5	86.0	10.8	29.6	79.8
(FWER(3))	(0.6)	(0.1)	(0.6)	(0.9)	(0.6)	(0.1)	(0.7)	(0.6)	(0.7)
Step-SPA(3)	34.5	66.4	97.0	27.0	56.0	94.7	17.4	43.2	90.8
(FWER(3))	(2.2)	(2.8)	(2.2)	(2.9)	(3.2)	(3.4)	(2.5)	(2.6)	(3.8)

S is the number of models, n is the number of observations, and ρ is the correlation coefficient between models. Empirical FWER, FWER(3), and average powers are all expressed in percentages; the nominal significance level is $\alpha=5\%$.

In the power simulations, the models are generated as follows. There are 10% of S models with $\mu=0$, 20% with $\mu>0$ (μ distributed evenly between 0.15 and 0.2), and 70% with $\mu<0$ (μ distributed evenly between 0 and -3).[2] Recall that SPA-type tests, by construction, have better power than RC-type tests when "poor" models are present. We generate a larger proportion of models with negative means so

[2]For example, for $S=100$, there are 20 positive means $(0.1525, 0.155, 0.1575, \ldots, 0.2)$, 10 zero means, and 70 negative means $(-3/70, -6/70, \ldots, -3)$.

Table 4 Average power performance and control of FWER: $t(4)$ random variables

	Model correlation $\rho=0$								
	$S=100$			$S=200$			$S=500$		
	$n=100$	$n=200$	$n=500$	$n=100$	$n=200$	$n=500$	$n=100$	$n=200$	$n=500$
Step-RC	8.7	24.2	73.1	6.0	18.5	66.9	3.8	13.4	58.3
(FWER)	(0.5)	(0.4)	(0.2)	(0.4)	(0.4)	(0.5)	(0.7)	(0.4)	(0.2)
Step-SPA	14.3	34.9	82.6	10.4	27.8	76.8	6.4	20.1	68.6
(FWER)	(1.2)	(1.6)	(2.3)	(1.5)	(1.5)	(1.4)	(2.0)	(0.9)	(0.9)
Step-RC(3)	26.3	53.4	92.7	19.0	44.1	88.6	11.7	32.2	81.6
(FWER(3))	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)
Step-SPA(3)	44.5	74.7	98.3	33.6	65.0	96.7	21.1	49.6	93.1
(FWER (3))	(0.3)	(1.0)	(1.6)	(0.3)	(0.3)	(2.8)	(0.3)	(0.3)	(2.6)
	Model correlation $\rho=0.2$								
Step-RC	9.5	25.4	75.0	6.8	20.9	68.4	4.2	14.6	60.3
(FWER)	(0.6)	(0.5)	(0.5)	(0.6)	(0.4)	(0.5)	(0.6)	(0.6)	(0.3)
Step-SPA	15.0	36.1	83.7	11.1	29.9	77.8	6.9	21.2	70.0
(FWER)	(1.2)	(1.5)	(1.9)	(1.8)	(1.5)	(2.3)	(1.6)	(1.8)	(1.1)
Step-RC(3)	23.2	48.6	90.4	16.8	40.5	85.7	10.5	29.2	78.5
(FWER(3))	(0.0)	(0.0)	(0.0)	(0.0)	(0.1)	(0.0)	(0.0)	(0.3)	(0.2)
Step-SPA(3)	38.8	69.2	97.7	29.1	59.4	95.2	18.1	44.2	90.7
(FWER(3))	(0.6)	(1.5)	(2.2)	(0.6)	(1.4)	(3.2)	(0.5)	(1.7)	(2.6)
	Model correlation $\rho=0.4$								
Step-RC	11.1	28.9	77.9	8.2	25.1	72.2	5.4	17.9	65.1
(FWER)	(0.7)	(0.6)	(0.9)	(0.8)	(0.9)	(1.2)	(0.9)	(1.5)	(0.6)
Step-SPA	16.8	39.1	85.4	12.3	33.6	80.4	8.3	24.6	73.6
(FWER)	(1.7)	(1.4)	(2.3)	(1.5)	(1.8)	(1.9)	(2.0)	(2.4)	(1.7)
Step-RC(3)	23.1	48.4	90.3	17.0	41.3	85.6	11.2	30.6	79.4
(FWER(3))	(0.0)	(0.0)	(0.2)	(0.1)	(0.3)	(0.3)	(0.2)	(0.9)	(0.4)
Step-SPA(3)	36.9	66.4	97.2	27.7	58.0	94.3	18.1	44.0	89.9
(FWER(3))	(1.2)	(1.6)	(2.7)	(1.2)	(2.8)	(3.4)	(2.0)	(3.1)	(3.2)

S is the number of models, n is the number of observations, and ρ is the correlation coefficient between models. Empirical FWER, FWER(3), and average powers are all expressed in percentages; the nominal significance level is $\alpha=5\%$.

as to make the difference between the performance of Step-RC(3) and Step-SPA(3) more obvious. We simulate the average power, global power, and minimum power, as defined in Romano and Wolf (2005). The average powers (the proportion of true rejections) of these tests are summarized in Tables 3 and 4 for models generated from, respectively, normal and $t(4)$ variables. We also report the corresponding FWER (for Step-RC and Step-SPA) and FWER(3) (for Step-RC(3) and Step-SPA(3)) in parentheses in these tables. We omit the other two powers because they do not

help distinguish the performance of different tests. For example, we find that the global powers are typically very small and that the minimal powers are all equal (or close) to one.

We observe the following from our simulation results. First, all tests control the FWER or the FWER(3) well. Second, Step-SPA(3) (Step-RC(3)) has much higher average power than the corresponding Step-SPA (Step-RC) test. This confirms that a test would have better power if it controls the FWER(k) instead of the FWER. Third, as expected, Step-SPA(3) outperforms Step-RC(3) quite remarkably in all cases considered. Fourth, for normal random variables, the average powers of Step-SPA(3) are high, as long as the number of observations is greater than or equal to the number of models. Finally, model correlation has an adverse effect on the average powers of Step-SPA(3) and Step-RC(3). These observations also hold for models generated from $t(4)$. We thus conclude that Step-SPA(3) is to be preferred to Step-RC(3), especially when the number of observations is large relative to the number of models.

4 EMPIRICAL STUDY ON CTA FUNDS

In our empirical study we apply the proposed test to assess the performance of CTA funds, a subset of Macro hedge funds according to the categorization of Hedge Fund Research, Inc.[3] A CTA fund mainly trades futures and forwards in commodities and financial instruments. There are two main strategies employed by CTA funds: systematic and discretionary. A systematic fund uses trading rules based on quantitative variables such as technical indicators, fundamental information and/or macro statistics. A discretionary fund trades based mainly on the past trading experience of the fund manager. The CTA fund family has been under the spotlight of the investment industry since 2008 because of its low correlation with traditional financial assets such as stocks and bonds, and its relatively good performance in 2008, as compared to mutual funds and other hedge funds.[4]

4.1 Data and Models

The monthly data on CTA funds are taken from the Hedge Fund Research (HFR) database, which is a leading database in hedge fund research; see, e.g., Kosowski, Naik, and Teo (2007) and Jagannathan, Malakhov, and Novikov (2010).

[3]Barras, Scaillet, and Wermers (2010) also evaluate mutual fund performance based on the "false discovery rates" approach of Storey (2002). Yet, their procedure, unlike the Step-SPA(k) test, is not readily applicable to check multiple inequalities.

[4]For example, the Barclays CTA index yielded a return of about 14% in 2008, while mutual fund and hedge fund reported, on average, loss of more than 30% and 23%, respectively.

There are 1050 funds during the period of July 1994 to June 2010.[5] Following Kosowski, Naik, and Teo (2007), we also exclude the first 12 months of data in the subsequent analysis, so as to mitigate the incubation bias. Note that certain "tiny" funds, those with assets under management less than $20 million, are also excluded because they are often not available to general investors. There are 315 remaining funds.

To assess fund performance, we employ the CAPM and the other two factor models. The CAPM is:

$$r_t^\ell = \alpha^\ell + \beta^\ell \left(R_{m,t} - R_{f,t}\right) + \varepsilon_t^\ell, \tag{5}$$

where r_t^ℓ is the t-th month return of the ℓ-th fund in excess of $R_{f,t}$, the one-month T-bill rate, and $R_{m,t}$ the t-th month return on the US stock markets, which is the value-weighted return on all NYSE, AMEX, and NASDAQ stocks from the CRSP database. We also consider the K-factor model as:

$$r_t^\ell = \alpha^\ell + \sum_{k=1}^{K} \beta_k^\ell F_{k,t} + \varepsilon_t^\ell, \tag{6}$$

where $F_{k,t}$ denotes the k-th factor. In particular, for the four-factor model obtained by adding the momentum factor of Carhart (1997) to the three-factor model of Fama and French (1993), $F_{k,t}$ are the excess return of the CRSP value-weighted U.S. stock market index ($R_{m,t}$ in (5)), size factor, value factor, and previous one-year momentum.[6] For the five-factor model of Fung and Hsieh (1997, 2001), $F_{k,t}$ are the t-th month return of the lookback straddle on the following five underlying futures markets: bond, currency, commodity, short-term interest rate, and stock index.[7] The performance measure is the t-ratio of the estimated abnormal return α^ℓ in (5) and (6). Note that Kazemi and Li (2009) also consider these three models in evaluating the market timing of CTAs.

4.2 Empirical Results

We apply Step-SPA(k) and Step-RC(k), $k = 1, 2, 3$, to identify outperforming funds from all funds and from two subgroups: discretionary funds and systematic funds.

[5]Most hedge fund database vendors, such HFR, started collecting data in the beginning of 1994; all data before 1994 were backfilled. As a result, the data of funds which were dead before 1994 are not available to the database vendors. To avoid the survivorship bias, we include only the data of live and defunct funds from July 1994.

[6]The monthly return data on all four factors are taken from the data library of Kenneth R. French at http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.

[7]The data on the five factors are available from the data library of David Hsieh at http://faculty.fuqua.duke.edu/~dah7/DataLibrary/TF-Fac.xls, last accessed April 1, 2014.

For every fund in each group, we test its performance based on the t-ratio of the estimated α^{ℓ} in the CAPM, 4-factor model, and 5-factor model. We compute Step-SPA(k) and Step-RC(k) as in our simulations, except that $\hat{\sigma}_{\ell,n}$ in the standardized test statistics are obtained from a pre-whitened HAC-consistent covariance matrix estimate based on the quadratic spectral kernel (Andrews and Monahan 1992) and the critical values are computed using the stationary bootstrap (Politis and Romano 1994). The standardization in the bootstrap is carried out as the second bootstrap method discussed in Section 4. Our statistics and critical values are thus robust to possible serial correlations in data. The expected block length in our stationary bootstrap is 4, and the number of bootstraps is 1000. Note that our results are not affected by other choices of block length, such as 2 and 10.

As CTA funds do not survive a long period of time, we report, as examples, the number of identified funds based on two arbitrarily chosen, 10-year sample periods: July 1996 to June 2006 and July 1998 to June 2008. The summary statistics of the data in these two sample periods are collected in Table 5. It is readily seen that the data in these two samples are skewed to the right and clearly deviate from normality. The testing results based on the period from July 1998 to June 2008 are given in Table 6, in which the upper and lower panels contain the results under the nominal levels FWER(k)=5% and FWER(k)=10%, respectively. Similarly, the testing results based on the period from July 1996 to June 2006 are summarized in Table 7.

It is easy to see from the upper panel of Table 6 that, for a given k, the number of funds identified by Step-SPA(k) is no less than that by Step-RC(k). The power advantage of Step-SPA(k) is more obvious when $k=3$. In particular, Step-SPA(3) is able to identify more outperforming funds from all funds and from systematic funds when the performance measure is based on the 4- and 5-factor models. As there are only 14 discretionary funds, Step-SPA(3) and Step-RC(3) tend to identify the same number of funds. Since the number of identified funds varies across different models, we also report the funds that are identified by all three models. It can be seen that Step-SPA(k) again selects more funds from systematic funds.

Table 5 Summary statistics of the data in two sample periods

	Sample: July 1996–June 2006			Sample: July 1998–June 2008		
Statistic	All funds	Discretionary	Systematic	All funds	Discretionary	Systematic
Mean	0.940	0.920	1.034	0.862	0.830	1.008
Median	0.500	0.520	0.413	0.419	0.400	0.460
Standard dev.	5.187	5.292	4.639	4.858	4.872	4.790
Min	−36.500	−23.330	−36.500	−36.500	−20.540	−36.500
Max	47.100	47.100	44.270	47.100	47.100	44.980
Skewness	0.897	0.906	0.826	0.891	0.837	1.149
Kurtosis	6.262	5.406	12.716	6.806	5.135	14.816
Number of funds	65	54	11	77	63	14

Table 6 The number of funds identified by Step-SPA(k) and Step-RC(k): July 1998–June 2008

Model	Test	All funds			Discretionary			Systematic		
		$k=1$	2	3	$k=1$	2	3	$k=1$	2	3
Nominal FWER(k)=5%										
CAPM	Step-RC(k)	1	8	12	3	5	5	0	9	9
	Step-SPA(k)	1	8	12	3	5	5	0	9	9
4-factor	Step-RC(k)	0	0	0	0	5	7	0	0	3
	Step-SPA(k)	0	0	4	0	5	7	0	1	5
5-factor	Step-RC(k)	4	5	8	1	3	3	4	4	11
	Step-SPA(k)	4	5	10	1	3	3	4	9	16
All 3	Step-RC(k)	0	0	0	0	2	3	0	0	3
models	Step-SPA(k)	0	0	0	0	2	3	0	1	5
Nominal FWER(k)=10%										
CAPM	Step-RC(k)	3	12	14	3	5	13	5	9	12
	Step-SPA(k)	3	12	14	3	5	13	5	10	13
4-factor	Step-RC(k)	0	0	9	0	6	7	0	2	5
	Step-SPA(k)	0	2	9	0	6	9	0	5	8
5-factor	Step-RC(k)	4	14	21	3	3	8	4	16	25
	Step-SPA(k)	4	18	27	3	3	10	5	19	27
All 3	Step-RC(k)	0	0	7	0	3	5	0	2	5
models	Step-SPA(k)	0	1	7	0	3	9	0	5	7

There is a total of 77 funds, in which 14 are discretionary and 63 are systematic.

When FWER(k)=10%, the conclusions are basically the same (lower panel of Table 6), except that Step-SPA(k) with $k=2$ now also shows power advantage over Step-RC(k).

For the results in Table 7, Step-SPA(k) and Step-RC(k) have very similar performance in most cases when FWER(k)=5% (upper panel). Yet when FWER(k)=10%, the power advantages of Step-SPA(k) for $k=2,3$ become obvious. It is also interesting to observe from both tables that the conventional Step-SPA test (i.e., Step-SPA(1)) typically has no power advantage relative to the conventional Step-RC(1) test, because the former does not identify more outperforming funds. This justifies why allowing for more false rejections (i.e., a larger k) in Step-SPA is practically desirable.

As a robustness check, we follow Kosowski et al. (2006) and Carhart (1997) to test if the performance of the identified funds persists over time. To this end, we take every 10 years as one in-sample period and the following year as its out-of-sample period. This results in 6 in- and out-of-sample periods.[8] We construct an equally

[8]The first in-sample period is from July 1994 through June 2004 with the associated out-of-sample period from July 2004 through June 2005. The last in-sample period is from July 1999 through June 2009 with the out-of-sample period from July 2009 through June 2010.

Table 7 The number of funds identified by Step-SPA(k) and Step-RC(k): July 1996–June 2006

		All funds			Discretionary			Systematic		
Model	Test	$k=1$	2	3	$k=1$	2	3	$k=1$	2	3
Nominal FWER(k)=5%										
CAPM	Step-RC(k)	1	1	7	1	5	7	0	4	8
	Step-SPA(k)	1	1	7	1	5	7	0	4	8
4-factor	Step-RC(k)	1	3	3	1	3	7	0	1	1
	Step-SPA(k)	1	3	3	1	3	7	0	1	1
5-factor	Step-RC(k)	1	6	12	1	5	8	0	7	13
	Step-SPA(k)	1	8	13	1	5	8	0	7	13
All 3	Step-RC(k)	1	1	1	1	2	6	0	0	0
models	Step-SPA(k)	1	1	1	1	2	6	0	0	0
Nominal FWER(k)=10%										
CAPM	Step-RC(k)	1	7	12	1	6	8	0	8	13
	Step-SPA(k)	1	8	13	1	6	8	0	9	14
4-factor	Step-RC(k)	1	3	7	2	5	7	1	1	7
	Step-SPA(k)	2	3	7	2	7	9	1	1	7
5-factor	Step-RC(k)	2	12	18	2	5	8	1	12	18
	Step-SPA(k)	2	13	18	2	5	9	1	13	18
All 3	Step-RC(k)	1	1	4	1	4	6	0	0	6
models	Step-SPA(k)	1	1	4	1	5	8	0	0	6

There is a total of 65 funds, in which 11 are discretionary and 54 are systematic.

weighted portfolio from the funds identified from each in-sample period (based on Step-SPA and Step-SPA(3)) and compute its return in the out-of-sample period. A factor model is then estimated using these out-of-sample returns. Carhart (1997) suggests a bootstrap approach to testing the significance of the abnormal return in this factor model. We summarize the out-of-sample results under the nominal level of 10% in Table 8. These testing results support, in general, that the funds identified by Step-SPA(3) continue to produce significant abnormal returns out of sample. For example, for the funds identified from all funds, discretionary funds, and systematic funds by Step-SPA(3) using the 5-factor model, our testing results indicate that the estimated abnormal returns of those portfolios are significant at, respectively, 1%, 1%, and 10% levels.

5 ASYMPTOTIC CONTROL OF FALSE DISCOVERY PROPORTION

A drawback of a test that controls the FWER(k) is that the choice of k does not depend on data. For the cases that a large number of false hypotheses is present, a

Table 8 Persistence test of standardized alpha of equally weighted portfolios based on selected CTA funds

	Funds selected by Step-SPA								
	CAPM			4-factor model			5-factor model		
	All	Disc.	Syst.	All	Disc.	Syst.	All	Disc.	Syst.
alpha	−0.105	0.005	0.877	−0.425	−0.690	−1.342	2.398	1.750	1.922
p-value	0.419	0.974	0.120	0.742	0.696	0.999	<0.0001	<0.0001	0.019
	Funds selected by Step-SPA(3)								
alpha	0.943	2.562	0.687	0.160	2.908	0.641	2.818	2.941	2.666
p-value	0.015	<0.0001	0.387	1.000	0.008	0.022	<0.0001	<0.0001	0.096

alpha denotes regression standardized alpha; *p*-value is bootstrapped *p*-value. The funds for the portfolios are selected by CAPM, 4-factor model, and 5-factor model under FWER(k)=10%.

test that allows for a fixed, small number of false rejections, e.g., FWER(k) with a small k, may still be conservative. This problem can be circumvented by controlling a different error rate, such as FDP. Note that FDP is defined as the ratio of the number of false rejections (F) over the number of total rejections (R):

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

For a given number $0 < \gamma < 1$, a multiple testing procedure is said to asymptotically control the FDP at the level α if $\limsup P[\text{FDP} > \gamma] \leq \alpha$.

The following examples illustrate the relation between FWER(k) and FDP. Letting $\gamma = 0.1$ and $\alpha = 5\%$, suppose that there are 10 superior models in the sample. Assuming that the procedure is consistent in that all superior models will be rejected in the first step with probability approaching one, FDP will then be equal to $F/(F+10)$ asymptotically which is strictly increasing in F. Therefore, $F/(F+10) > 0.1$ if, and only if, $F \geq 2$. In this case, the FDP control is asymptotically equivalent to the FWER(2) control. If there are more, say 100, superior models, then FDP control with $\gamma = 0.1$ would be equivalent to FWER(11). In view of these examples, the FDP control may be interpreted as a data-dependent FWER(k) control, in the sense that k depends on the underlying data-generating process.

A procedure that controls the FDP at the level α may be constructed from a procedure that controls the FWER(k) with k fixed. The following FDP-SPA algorithm is based on Step-SPA(k).

Algorithm 3: *(FDP-SPA)*

1. *Set $k = 1$ and a γ value between 0 and 1.*

2. *Apply Step-SPA(k) with α. Let N_k denote the number of the rejected hypotheses by Step-SPA(k).*
3. *If $N_k < k/\gamma - 1$, stop and reject all hypotheses rejected by Step-SPA(k); otherwise, set $k = k+1$ and return to Step 2.*

In this algorithm, the stopping rule is $N_k < k/\gamma - 1$. Among N_k rejected models, the probability of having $k-1$ or less false rejections is greater than or equal to $1-\alpha$. If we move from k to $k+1$, it is very likely to get one more false rejection, but no true rejection. Then, the FDP becomes $k/(N_k+1)$. When $k/(N_k+1) \leq \gamma$, we can still control the FDP well if we continue to implement Step-SPA($k+1$). In other words, we should stop the procedure when $k/(N_k+1) > \gamma$, which is equivalent to $N_k < k/\gamma - 1$.

To show that FDP-SPA defined in Algorithm 3 controls the FDP at the level α asymptotically, it suffices to show that Step-SPA(k) satisfies the following properties of Romano and Wolf (2007): (i) it is monotonic in k, (ii) it is consistent for all k, and (iii) it asymptotically controls the FWER(k) at the level α for any $k \geq 1$. Theorem 3.1 already establishes the properties (ii) and (iii) for Step-SPA(k). We will show in Appendix the monotonicity of Step-SPA(k), i.e., any hypothesis rejected by Step-SPA(k_1) will be rejected by Step-SPA(k_2) whenever $k_1 < k_2$. We thus have the following result.

Theorem 3: *Suppose Assumptions 1 and 2 hold, then FDP-SPA defined in Algorithm 3 asymptotically controls the FDP at the level α.*

Note that the FDP-RC in Romano and Wolf (2007) is virtually the same as Algorithm 3, except that it replaces Step-SPA(k) with Step-RC(k). Given that Step-SPA(k) is more powerful than Step-RC(k) (Theorem 2), the following result shows the power advantage of FDP-SPA over FDP-RC.

Theorem 4: *Under Assumptions 1 and 2, FDP-SPA is more powerful than the FDP-RC under the notions of power defined in Romano and Wolf (2005).*

6 CONCLUSION

In this article, a stepwise test for multiple inequalities testing that can asymptotically control the FWER(k), the probability of at least k false rejections, is proposed. By allowing for more false rejections, this Step-SPA(k) test is more powerful than Step-SPA of Hsu, Hsu, and Kuan (2010). As the proposed test avoids the LFC in determining the null distribution, it compares favorably with Step-RC(k) of Romano and Wolf (2007), which also controls the FWER(k). The power advantage of the proposed test is demonstrated by our simulations and empirical study on CTA funds. To allow for a flexible number of false rejections, we further extend

the results for Step-SPA(k) to FDP-SPA which asymptotically controls the FDP at the level α. It is shown that, analogous to the power advantage of Step-SPA(k) over Step-RC(k), FDP-SPA is more powerful than FDP-RC of Romano and Wolf (2007). Our results confirm that, comparing with the testing procedures that rely on the LFC, the corresponding procedures based on the re-centering method are to be preferred.

APPENDIX

A: Auxiliary Lemmas

We present several lemmas that will be used to prove Theorem 2. Suppose we re-label the hypotheses in the descending order of $\widehat{T}_{\ell,n}$. For $s=1,\ldots,k$, define $\mathcal{K}_s^k=\{A_1\}$ and for $s=k+1,\ldots,m$, define

$$\mathcal{K}_s^k=\big\{\{s,\ldots,m\}\cup I\big|\ I\subset\{1,\ldots,s-1\},|I|=k-1\big\}.$$

For $s=1,\ldots,m$, define

$$\begin{aligned}\hat{w}_\alpha(s,k)&=\max_{K_s\in\mathcal{K}_s^k}\big\{\hat{q}_{n,K_s}(1-\alpha,k)\big\},\\ \hat{d}_\alpha(s,k)&=\max_{K_s\in\mathcal{K}_s^k}\big\{\hat{c}_{n,K_s}(1-\alpha,k)\big\}.\end{aligned}\tag{A1}$$

Lemma A1: *For any fixed k, $\hat{w}_\alpha(s,k)$ and $\hat{d}_\alpha(s,k)$ are nonincreasing in s and $\hat{w}_\alpha(s,k)\le\hat{d}_\alpha(s,k)$ for all $s=1,\ldots,m$.*

Lemma A2: *Suppose the hypotheses are re-labeled in the descending order of $\widehat{T}_{\ell,n}$ and $\hat{w}_\alpha(s,k)$ and $\hat{d}_\alpha(s,k)$ are defined in (A1). Then H_0^ℓ is rejected by the Step-SPA(k) if and only if $\sqrt{n}\widehat{T}_{s,n}>\hat{w}_\alpha(s,k)$ for all $s=1,\ldots,\ell$. Similarly, H_0^ℓ is rejected by the Step-RC(k) if and only if $\sqrt{n}\widehat{T}_{s,n}>\hat{d}_\alpha(s,k)$ for all $s=1,\ldots,\ell$.*

Lemma A3: *For any fixed s, $\hat{w}_\alpha(s,k)$ and $\hat{d}_\alpha(s,k)$ are nonincreasing in k.*

B: Proof of Theorems

Proof of Theorem 1: For the first part, it is sufficient to show that Algorithm 2 satisfies Theorem 2.1 of Romano and Wolf (2007). By construction, it is obvious that $\hat{q}_{n,K_1}(\alpha,k)\ge\hat{q}_{n,K_2}(\alpha,k)$ for any two subsets of indices K_1 and K_2 with $K_1\supseteq K_2$. Therefore, this is sufficient to show that for any $K\supseteq I(P)$, $\hat{q}_{n,K}(\alpha,k)\ge\hat{q}_{n,I(P)}(\alpha,k)$ where $I(P)$ is the set of the indices of the true null hypotheses. We next show that $P[k\text{-max}\{\sqrt{n}\widehat{T}_{\ell,n}:\ell\in I(P)\}>\hat{q}_{n,I(P)}(1-\alpha,k)]\le\alpha$.

Without loss of generality, let $|I(P)| = \{1, \ldots, I_0\}$ with $I_0 \geq k$. If $I_0 < k$, then there is nothing to show. Furthermore, let $\theta_\ell = 0$ for $\ell = 1, \ldots, k_0$ and $\theta_\ell < 0$ for $\ell = k_0 + 1, \ldots, I_0$. Hence, for $\ell = 1, \ldots, k_0$, $\widehat{T}_{\ell,n} \xrightarrow{p} 0$ and for $\ell = k_0 + 1, \ldots, I_0$, $\widehat{T}_{\ell,n} \xrightarrow{p} \theta_\ell / \sigma_\ell < 0$. First, if $k_0 < k$, then

$$k\text{-max}\left\{\widehat{T}_{\ell,n} : \ell \in I(P)\right\} \xrightarrow{p} \max\left\{\theta_\ell / \sigma_\ell : \ell = k_0 + 1, \ldots, I_0\right\} < 0,$$

which implies that

$$k\text{-max}\left\{\sqrt{n}\widehat{T}_{\ell,n} : \ell \in I(P)\right\} \xrightarrow{p} -\infty.$$

Since the critical value is greater than or equal to zero,

$$P\left[k\text{-max}\left\{\sqrt{n}\widehat{T}_{\ell,n} : \ell \in I(P)\right\} > \hat{q}_{n,I(P)}(1-\alpha, k)\right] = 0 \leq \alpha.$$

Second, if $k_0 \geq k$, then with probability approaching one, the index of the k-th largest element among $\{\widehat{T}_{\ell,n} : \ell \in I(P)\}$ will be in $\{1, \ldots, k_0\}$. Therefore, $k\text{-max}\{\sqrt{n}\widehat{T}_{\ell,n} : \ell \in I(P)\}$ is asymptotically equivalent to $k\text{-max}\{\sqrt{n}\widehat{T}_{\ell,n} : \ell = 1, \ldots, k_0\}$ in that the difference of the two will converge in probability to zero. Also, $k\text{-max}\{\sqrt{n}\widehat{T}_{\ell,n} : \ell = 1, \ldots, k_0\}$ converges in distribution to $k\text{-max}(\mathcal{N}(0, \Xi_0))$, where Ξ_0 is a $k_0 \times k_0$ sub-matrix of Ξ. Therefore, $k\text{-max}(\sqrt{n}\widehat{T}_{\ell,n} : \ell \in I(P)) \xrightarrow{d} k\text{-max}\{\mathcal{N}(0, \Xi_0)\}$, On the other hand, by Lemma 3.3 of Donald and Hsu (2011), we have $\hat{\mu}_{\ell,n} \xrightarrow{p} \theta_\ell / \sigma_\ell$ for all $\ell \in I(P)$. Also, we have $\psi^u_{\ell,n} / \sqrt{n} \xrightarrow{p} 0$ for all $\ell \in I(P)$. As a result, $\psi^u_{\ell,n} / \sqrt{n} + \hat{\mu}_{\ell,n} \xrightarrow{p} \theta_\ell / \sigma_\ell$ for $\ell = 1, \ldots, I_0$ and similarly the index of the k-th largest element among $\{\psi^u_{\ell,n} + \sqrt{n}\hat{\mu}_{\ell,n} : \ell \in I(P)\}$ will be in $\{1, \ldots, k_0\}$. Hence, the $k\text{-max}\{\psi^u_{\ell,n} + \sqrt{n}\hat{\mu}_{\ell,n} : \ell \in I(P)\}$ is asymptotically equivalent to $k\text{-max}\{\psi^u_{\ell,n} + \sqrt{n}\hat{\mu}_{\ell,n} : \ell = 1, \ldots, k_0\}$. Also, by Lemma 3.3 of Donald and Hsu (2011), it is true that $\sqrt{n}\hat{\mu}_{\ell,n} \xrightarrow{p} 0$ for $\ell = 1, \ldots, k_0$. Hence, $k\text{-max}\{\psi^u_{\ell,n} + \sqrt{n}\hat{\mu}_{\ell,n} : \ell = 1, \ldots, k_0\}$ is asymptotically equivalent to $k\text{-max}\{\psi^u_{\ell,n} : \ell = 1, \ldots, k_0\}$ which converges in distribution to $k\text{-max}\{\mathcal{N}(0, \Xi_0)\}$ conditional on sample with probability approaching one. Therefore, $k\text{-max}\{\psi^u_{\ell,n} + \sqrt{n}\hat{\mu}_{\ell,n} : \ell \in I(P)\} \xrightarrow{d} k\text{-max}\{\mathcal{N}(0, \Xi_0)\}$ conditional on sample with probability approaching one. By Lemma A.1 of Romano and Wolf (2007), the distribution of $k\text{-max}\{\mathcal{N}(0, \Xi_0)\}$ is continuous. Let q_0 denote the $(1-\alpha)$-th quantile of $k\text{-max}\{\mathcal{N}(0, \Xi_0)\}$, then $\tilde{q}_{n,I(P)}(1-\alpha, k) \xrightarrow{p} q_0$. Consequently,

$$\begin{aligned}
&\lim_{n\to\infty} \mathrm{P}\left[k\text{-max}\left\{\sqrt{n}\widehat{T}_{\ell,n} : \ell \in I(P) > \hat{q}_{n,I(P)}(1-\alpha, k)\right\}\right] \\
&\quad \leq \lim_{n\to\infty} \mathrm{P}\left[k\text{-max}\left\{\sqrt{n}\widehat{T}_{\ell,n} : \ell \in I(P) > \tilde{q}_{n,I(P)}(1-\alpha, k)\right\}\right] \\
&\quad = \mathrm{P}\left[k\text{-max}\left\{\mathcal{N}(0, \Xi_0)\right\} > q_0\right] \\
&\quad = \alpha.
\end{aligned}$$

To show the second part, note that $\tilde{q}_{n,A_1}(1-\alpha, k)$ converges in probability to the $(1-\alpha)$-th quantile of $k\text{-max}\{\mathcal{N}(0, \Xi)\}$. Hence, $\tilde{q}_{n,A_1}(1-\alpha, k)$ is bounded

above in probability and $\hat{q}_{n,A_1}(1-\alpha,k)=\max\{\tilde{q}_{n,A_1}(1-\alpha,k),0\}$ is bounded above in probability too. On the other hand, if $\theta_\ell>0$, $\widehat{T}_{\ell,n}\xrightarrow{p}\theta_\ell/\sigma_\ell>0$ and it follows that $\sqrt{n}\widehat{T}_{\ell,n}$ diverges to positive infinity with probability approaching one. Consequently, if $\theta_\ell>0$,

$$\lim_{n\to\infty}\mathrm{P}\big[\sqrt{n}\widehat{T}_{\ell,n}>\hat{q}_{n,A_1}(1-\alpha,k)\big]=1.$$

In other words, all false null hypotheses will be rejected in the first step with probability approaching one and this proves the second part. ■

Proof of Theorem 2: Define R_S^k and R_R^k as the collections of indices of the hypotheses being rejected by Step-SPA(k) and Step-RC(k) respectively. To show Theorem 2, it is sufficient to show that $R_R^k\subseteq R_S^k$.

If $R_R^k=\emptyset$, then it is obvious $R_R^k\subseteq R_S^k$. If R_R^k is not empty, then for any $\ell\in R_R^k$, we have $\sqrt{n}\widehat{T}_{s,n}>\hat{d}_\alpha(s,k)$ for all $s=1,\ldots,\ell$. By Lemma A1, $\sqrt{n}\widehat{T}_{s,n}>\hat{d}_\alpha(s,k)\geq\hat{w}_\alpha(s,k)$ for all $s=1,\ldots,\ell$. Then by Lemma A2, $\ell\in R_S^k$ too. Consequently, $R_R^k\subseteq R_S^k$.

Let I_1 denote the collections of the indices of false hypotheses. Let P_S^k and P_R^k denote the average powers of Step-SPA(k) and Step-RC(k) such that

$$P_S^k=\frac{E[|R_S^k\cap I_1|]}{|I_1|},\quad P_R^k=\frac{E[|R_R^k\cap I_1|]}{|I_1|}.$$

If $I_1=\emptyset$, we define $P_S^k=P_R^k=0$. Otherwise, because $R_R^k\subseteq R_S^k$, we have $R_R^k\cap I_1\subseteq R_S^k\cap I_1$ and $|R_S^k\cap I_1|\geq|R_R^k\cap I_1|$ which implies that $P_S^k\geq P_R^k$. The proofs for the cases with other notions of power in Romano and Wolf (2005) are similar and hence omitted.■

Proof of Theorem 3: By Theorem 4.1 in Romano and Wolf (2007), to show Theorem 3, it is sufficient to show that Step-SPA(k) (i) is monotone in k, (ii) is consistent for all k, and (iii) can asymptotically control the FWER(k) at level α for any $k\geq1$. The requirements (ii) and (iii) are shown in Theorem 3.1, so we just need to show the monotonicity of Step-SPA(k). Let R_S^k denote the set of the indices of the rejected hypotheses by Step-SPA(k). We need to show $R_S^k\subseteq R_S^{k+1}$. If $\ell\in R_S^k$, by Lemma A2, $\sqrt{n}\widehat{T}_{s,n}>\hat{w}_\alpha(s,k)$ for all $s=1,\ldots,\ell$. By Lemma A3, it is true that $\sqrt{n}\widehat{T}_{s,n}>\hat{w}_\alpha(s,k)\geq\hat{w}_\alpha(s,k+1)$ for all $s=1,\ldots,\ell$ which implies that $\ell\in R_S^{k+1}$. As a result, $R_S^k\subseteq R_S^{k+1}$ and the monotonicity of Step-SPA(k) holds. ■

Proof of Theorem 4: Let R_S^k and R_R^k denote the sets of the indices of the rejected hypotheses by Step-SPA(k) and Step-RC(k), respectively. Define $\hat{k}_{spa}$ and $\hat{k}_{rc}$ as the numbers of steps proceed in FDP-SPA and FDP-RC, respectively. We claim that $\hat{k}_{rc}\leq\hat{k}_{spa}$. By the definition of $\hat{k}_{rc}$, we have

$$|R_R^k|<\frac{k}{\gamma}-1,\quad\text{if } k=1,\ldots,\hat{k}_{rc}-1,$$

$$|R_R^k|\geq\frac{k}{\gamma}-1,\quad\text{if } k=\hat{k}_{rc}.$$

By the proof of Theorem 3.2, we have $|R_R^k| \le |R_S^k|$ for all k and it follows that

$$|R_S^k| < \frac{k}{\gamma} - 1, \quad \text{if } k = 1, \ldots, \hat{k}_{rc} - 1.$$

As a result, Algorithm 3 will continue till $\hat{k}_{rc}$-th step at least, so $\hat{k}_{rc} \le \hat{k}_{spa}$. Therefore, $R_S^{\hat{k}_{spa}}$ and $R_S^{\hat{k}_{rc}}$ are the sets of the indices of the rejected hypotheses by FDP-SPA and FDP-RC respectively. Consequently, $R_R^{\hat{k}_{rc}} \subseteq R_S^{\hat{k}_{rc}} \subseteq R_S^{\hat{k}_{spa}}$ given $\hat{k}_{rc} \le \hat{k}_{spa}$. This is sufficient to show Theorem 4. ■

C: Proofs of Auxiliary Lemmas

Proof of Lemma A1: If $s \le k$, $\hat{w}_\alpha(s,k) = \hat{q}_{n,A_1}(1-\alpha,k)$, so it is nonincreasing. If $s \ge k+1$, for any $K_s \in \mathcal{K}_s^k$, we have $K_s = \{s, \ldots, m\} \cup I_s$ where $I_s \subseteq \{1, \ldots, s-1\}$ with $I_s = k-1$. If $s-1 \notin I_s$, then define $K_{s-1} = \{s-1, \ldots, m\} \cup I_s$. On the other hand, if $s-1 \in I_s$, let $I_{s-1} \subseteq \{1, \ldots, s-1\}$ and $I_{s-1} = k-1$ such that $I_s \subseteq I_{s-1} \cup \{s-1\}$ and define $K_{s-1} = \{s-1, \ldots, m\} \cup I_{s-1}$. It is obvious that $K_{s-1} \in \mathcal{K}_{s-1}$ in both cases and $K_s \subseteq K_{s-1}$ by construction. As a result, for any $K_s \in \mathcal{K}_s^k$, there exists $K_{s-1} \in \mathcal{K}_{s-1}^k$ such that $K_s \subseteq K_{s-1}$ which implies that $\hat{q}_{n,K_s}(1-\alpha,k) \le \hat{q}_{n,K_{s-1}}(1-\alpha,k)$. It follows that

$$\hat{w}_\alpha(s,k) = \max_{K_s \in \mathcal{K}_s^k} \left\{ \hat{q}_{n,K_s}(1-\alpha,k) \right\} \le \max_{K_{s-1} \in \mathcal{K}_{s-1}^k} \left\{ \hat{q}_{n,K_{s-1}}(1-\alpha,k) \right\} = \hat{w}_\alpha(s-1,k).$$

The argument for $\hat{d}_\alpha(s,k)$ is similar and we omit it.

For the second part, note that $\psi_{\ell,n}^u + \sqrt{n}\hat{\mu}_{\ell,n} \le \psi_{\ell,n}^u$, since $\hat{\mu}_{\ell,n}$ is nonpositive for all $\ell = 1, \ldots, m$. It follows that for any $K \subseteq \{1, \ldots, m\}$,

$$k\text{-max}\left\{ \sqrt{n}\widehat{T}_{\ell,n} + \sqrt{n}\hat{\mu}_{\ell,n} : \ell \in K \right\} \le k\text{-max}\left\{ \hat{\theta}_{\ell,n} : \ell \in K \right\}$$

which implies $\tilde{q}_{n,K}(\alpha,k) \le \tilde{c}_{n,K}(\alpha,k)$ and $\hat{q}_{n,K}(\alpha,k) \le \hat{c}_{n,K}(\alpha,k)$. Therefore, for any s,

$$\hat{w}_\alpha(s,k) = \max_{K_s \in \mathcal{K}_s^k} \left\{ \hat{q}_{n,K_s}(1-\alpha,k) \right\} \le \max_{K_s \in \mathcal{K}_s^k} \left\{ \hat{c}_{n,K_s}(1-\alpha,k) \right\} = \hat{d}_\alpha(s,k),$$

and this completes our proof. ■

Proof of Lemma A2: Suppose $\sqrt{n}\widehat{T}_{s,n} > \hat{w}_\alpha(s,k)$ for all $s = 1, \ldots, \ell$. If $\ell \le k$, then $\sqrt{n}\widehat{T}_{s,n} > \hat{w}_\alpha(1,k)$ for all $s = 1, \ldots, \ell$ because $\hat{w}_\alpha(s,k) = \hat{w}_\alpha(1,k)$. Hence, H_0^ℓ will be rejected in the first step. If $\ell > k$, then in the first step, at least H_j^0 for $j = 1, \ldots, k$ will be rejected. If H_ℓ^0 is also rejected at this stage, then we are done. If not, let k_1 be the number of rejections in the first step. Then we have $k_1 \ge k$ and this procedure will continue to next step. In the second step, $\sqrt{n}\widehat{T}_{k_1+1,n} > \hat{w}_\alpha(k_1+1,k)$, so at least $H_{k_1+1}^0$ will be rejected and the procedure continues. If H_ℓ^0 is also rejected in this step,

then we are done; otherwise, by the same argument, H_ℓ^0 will be rejected in finite steps.

On the other hand, suppose the statement that $\sqrt{n}\widehat{T}_{s,n} > \hat{w}_\alpha(s,k)$ for all $s=1,\ldots,\ell$ is not true and let $k_1 \le \ell$ be the first index that $\sqrt{n}\widehat{T}_{k_1,n} \le \hat{w}_\alpha(k_1,k)$. From the first part, the procedure will continue until the first k_1-1 hypotheses are rejected. Since $\hat{w}_\alpha(s,k)$ is nondecreasing in s, we have $\hat{w}_\alpha(s,k) \ge \hat{w}_\alpha(k_1,k) \ge \sqrt{n}\widehat{T}_{k_1,n}$ for all $s=1,\ldots,k_1-1$, so $H_{k_1}^0$ will not be rejected in previous steps no matter how the procedure proceeds. After first k_1-1 hypotheses are all rejected, we have $\sqrt{n}\widehat{T}_{k_1,n} \le \hat{w}_\alpha(k_1,k)$. Since there is no further rejection in this step, the procedure will stop here and H_ℓ^0 will not be rejected.

The proof for the Step-RC(k) procedure is similar and hence omitted. ■

Proof of Lemma A3: For any fixed s, if $k \ge s$, we have $\hat{w}_\alpha(s,k) = \hat{q}_{n,A_1}(1-\alpha,k)$. Because k-max$\{\mathcal{Y}\} \ge (k+1)$-max$\{\mathcal{Y}\}$ for any finite set of $\mathcal{Y}$, it follows that

$$k\text{-max}\left\{\sqrt{n}\widehat{T}_{\ell,n} + \sqrt{n}\hat{\mu}_{\ell,n} : \ell \in A_1\right\} \ge (k+1)\text{-max}\left\{\sqrt{n}\widehat{T}_{\ell,n} + \sqrt{n}\hat{\mu}_{\ell,n} : \ell \in A_1\right\}.$$

Consequently, it implies that $\hat{q}_{n,A_1}(1-\alpha,k) \ge \hat{q}_{n,A_1}(1-\alpha,k+1)$. If $k<s$, note that for any $K_s^{k+1} \in \mathcal{K}_s^{k+1}$, we have $K_s^{k+1} = I_s^{k+1} \cup \{s,\ldots,m\}$ where $I_s^{k+1} \subseteq \{1,\ldots,s\}$ and $|I_s^{k+1}| = k+1$. Let $K_s^k = I_s^k \cup \{s,\ldots,m\}$ such that $I_s^k \subseteq I_s^{k+1}$ and $|I_s^k| = k$. It is obvious that $K_s^k \in \mathcal{K}_s^k$. Therefore, it is true that

$$k\text{-max}\left\{\sqrt{n}\widehat{T}_{\ell,n} + \sqrt{n}\hat{\mu}_{\ell,n} : \ell \in K_s^k\right\} \ge (k+1)\text{-max}\left\{\sqrt{n}\widehat{T}_{\ell,n} + \sqrt{n}\hat{\mu}_{\ell,n} : \ell \in K_s^{k+1}\right\}.$$

As a result, for any $K_s^{k+1} \in \mathcal{K}_s^{k+1}$, there exists $K_s^k \in \mathcal{K}_s^k$ such that $\hat{q}_{n,K_s^k}(1-\alpha,k) \ge \hat{q}_{n,K_s^{k+1}}(1-\alpha,k+1)$. Consequently,

$$\hat{w}_\alpha(s,k) = \max_{K_s \in \mathcal{K}_s^k}\left\{\hat{q}_{n,K_s}(1-\alpha,k)\right\} \ge \max_{K_s^{k+1} \in \mathcal{K}_s^{k+1}}\left\{\hat{q}_{n,K_s^{k+1}}(1-\alpha,k+1)\right\} = \hat{w}_\alpha(s,k+1).$$

As a result, for any fixed s, $\hat{w}_\alpha(s,k)$ is nonincreasing in k. The argument for $\hat{d}_\alpha(s,k)$ is similar and we omit it. ■

REFERENCES

Andrews, D. W. K. 1991. Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation. *Econometrica* 59: 817–858.

Andrews, D. W. K., and C. J. Monahan. 1992. An Improved Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimator. *Econometrica* 60: 953–966.

Andrews, D. W. K., and X. Shi. 2013. Nonparametric Inference Based on Conditional Moment Inequalities. *Econometrica* 81: 609–666.

Andrews, D. W. K., and G. Soares. 2010. Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection. *Econometrica* 78: 119–157.

Barras, L., O. Scaillet, and R. Wermers. 2010. False Discoveries in Mutual Fund Performance: Measuring Luck in Estimated Alphas. *Journal of Finance* 65: 179–216.

Carhart, M. 1997. On Persistence of Mutual Fund Performance. *Journal of Finance* 52: 57–82.

Dudoit, S., J. P. Shaffer, and J. C. Boldrick. 2003. Multiple Hypothesis Testing in Microarray Experiments. *Statistical Science* 18: 71–103.

Donald, S. G., and Y.-C. Hsu. 2011. A New Test for Linear Inequality Constraints When the Variance Covariance Matrix Depends on the Unknown Parameters. *Economics Letters* 113: 241–243.

Donald, S. G., and Y.-C. Hsu. 2013. Improving the Power of Tests of Stochastic Dominance. *Econometric Reviews* (Forthcoming).

Fama, E. F., and K. R. French. 1993. Common Risk Factors in the Returns on Stocks and Bonds. *Journal of Financial Economics* 33: 3–56.

Fung, W., and D. Hsieh. 1997. Survivorship Bias and Investment Style in the Returns of CTAs. *Journal of Portfolio Management* 24: 30–41.

Fung, W., and D. Hsieh. 2001. The Risk in Hedge Fund Strategies: Theory and Evidence from Trend Followers. *Review of Financial Studies* 14: 313–341.

Hansen, P. R. 2005. A Test for Superior Predictive Ability. *Journal of Business and Economic Statistics* 23: 365–380.

Hansen, P. R., and A. Lunde. 2005. A Forecast Comparison of Volatility Models: Does Anything Beat a GARCH(1,1)? *Journal of Applied Econometrics* 20: 873–889.

Hsu, P.-H., Y.-C. Hsu, and C.-M. Kuan. 2010. Testing The Predictive Ability of Technical Analysis Using a New Stepwise Test without Data-Snooping Bias. *Journal of Empirical Finance* 17: 471–484.

Hsu, P.-H., and C.-M. Kuan. 2005. Reexamining The Profitability of Technical Analysis with Data Snooping Checks. *Journal of Financial Econometrics* 3: 606–628.

Hsu, Y.-L., C.-M. Kuan, and S. M. F. Yen. 2013. "Selecting Top Funds of Hedge Funds Based on Alpha and Other Performance Measures." In Greg N. Gregoriou (ed.), *Reconsidering Funds of Hedge Funds: The Financial Crisis and Best Practices in UCITS, Tail Risk, Performance, and Due Diligence*, pp. 351–366. Amsterdam: Elsevier.

Jagannathan, R., A. Malakhov, and D. Novikov. 2010. Do Hot Hands Exist among Hedge Fund Managers? An Empirical Evaluation. *Journal of Finance* 65: 217–255.

Kazemi, H., and Y. Li. 2009. Market Timing of CTAs: An Examination of Systematic CTAs vs. Discretionary CTAs. *Journal of Futures Markets* 29: 1067–1099

Kosowski, R., Naik, N. Y., and Teo, M. 2007. Do Hedge Funds Deliver Alpha? A Baysian and Bootstrap Analysis. *Journal of Financial Economics* 84: 229–264.

Kosowski, R., A. Timmermann, R. Wermers, and H. White. 2006. Can Mutual Fund "Stars" Really Pick Stocks? New Evidence from a Bootstrap Analysis. *Journal of Finance* 61: 2551–2596.

Linton, O., K. Song, and Y.-J. Whang. 2010. An Improved Bootstrap Test of Stochastic Dominance. *Journal of Econometrics* 154: 186–202.

Newey, W. K., and K. D. West. 1987. A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica* 55: 703–708.

Politis, D. N., and J. P. Romano. 1994. The Stationary Bootstrap. *Journal of the American Statistical Association* 89: 1303–1313.

Qi, M., and Y. Wu. 2006. Technical Trading-rule Profitability, Data Snooping, and Reality Check: Evidence from The Foreign Exchange Market. *Journal of Money, Credit and Banking* 30: 2135–2158.

Romano, J. P., and M. Wolf. 2005. Stepwise Multiple Testing as Formalized Data Snooping. *Econometrica* 73: 1237–1282.

Romano, J. P., and M. Wolf. 2007. Control of Generalized Error Rates in Multiple Testing. *Annals of Statistics* 35: 1378–1408.

Romano, J. P., A. M. Shaikh, and M. Wolf. 2008. Formalized Data Snooping Based on Generalized Error Rates. *Econometric Theory* 24: 404–447.

Storey, J. D. 2002. A Direct Approach to False Discovery Rates. *Journal of the Royal Statistical Society: Series B* 64: 479–498.

Sullivan, R., A. Timmermann, and H. White. 1999. Data-snooping, Technical Trading Rule Performance, and the Bootstrap. *Journal of Finance* 54: 1647–1691.

Sullivan, R., A. Timmermann, and H. White. 2001. Dangers of Data Mining: The Case of Calendar Effects in Stock Returns. *Journal of Econometrics* 105: 249–286.

White, H. 2000. A Reality Check for Data Snooping. *Econometrica* 68: 1097–1126.

White, H. 2001. *Asymptotic Theory for Econometricians*. Revised Edition, San Diego: Academic Press.

Yen, M.-F., and Y.-L. Hsu. 2010. Profitability of Technical Analysis in Financial and Commodity Futures Markets – A Reality Check. *Decision Support Systems* 50: 128–139