Centre de Referència en Economia Analítica

Barcelona Economics Working Paper Series

Working Paper nº 17

Stewise Multiple Testing as Formalized Data Snooping

Joseph P. Romano and Michael Wolf

October 2003

Stepwise Multiple Testing as Formalized Data Snooping

Joseph P. Romano
Department of Statistics
Sequoia Hall
Stanford University
Stanford, CA 94305
U.S.A

Michael Wolf *
Department of Economics and Business
Universitat Pompeu Fabra
Ramon Trias Fargas, 25–27
08005 Barcelona
Spain

Barcelona Economics WP # 17

October 2003

Abstract

It is common in econometric applications that several hypothesis tests are carried out at the same time. The problem then becomes how to decide which hypotheses to reject, accounting for the multitude of tests. In this paper, we suggest a stepwise multiple testing procedure which asymptotically controls the familywise error rate at a desired level. Compared to related single-step methods, our procedure is more powerful in the sense that it often will reject more false hypotheses. In addition, we advocate the use of studentization when it is feasible. Unlike some stepwise methods, our method implicitly captures the joint dependence structure of the test statistics, which results in increased ability to detect alternative hypotheses. We prove our method asymptotically controls the familywise error rate under minimal assumptions. We present our methodology in the context of comparing several strategies to a common benchmark and deciding which strategies actually beat the benchmark. However, our ideas can easily be extended and/or modified to other contexts, such as making inference for the individual regression coefficients in a multiple regression framework. Some simulation studies show the improvements of our methods over previous proposals. We also provide an application to a set of real data.

KEY WORDS: Bootstrap, data snooping, familywise error, multiple testing, step-down method.

JEL CLASSIFICATION NOS: C12, C14, C52.

*Research of the second author supported by the Spanish Ministry of Science and Technology and FEDER, grant BMF2003-03324, and by the Barcelona Economics Program of CREA.

"If you can do an experiment in one day, then in 10 days you can test 10 ideas, and maybe one of the 10 will be right. Then you've got it made."

– Solomon H. Snyder

1 Introduction

Much empirical research in economics and finance inevitably involves data snooping. Unlike the physical sciences, it is typically impossible to design replicable experiments. As a consequence, existing data sets are analyzed not once but repeatedly. Often, many strategies are evaluated on a single data set to determine which strategy is 'best' or, more generally, which strategies are 'better' than a certain benchmark. A benchmark can be fixed or random. An example of a fixed benchmark is the problem of determining whether a certain trading strategy has a positive CAPM alpha (so the benchmark is zero).[1] An example of a random benchmark is the problem of determining whether a trading strategy beats a specific investment, such as a stock index. If many strategies are evaluated, some are bound to appear superior to the benchmark by chance alone, even if in reality they are all equally good or inferior. This effect is known as data snooping (or data mining).

Economists have long been aware of the dangers of data snooping. For example, see Cowles (1933), Leamer (1983), Lovell (1983), Lo and MacKinley (1990), and Diebold (2000). However, in the context of comparing several strategies to a benchmark, little has been suggested on how to properly account for the effects of data snooping. A notable exception is White (2000). The aim of this work is to determine whether the strategy that is best in the available sample indeed beats the benchmark, after accounting for data snooping. White (2000) coins his technique the Bootstrap Reality Check (BRC). Often one would like to identify further strategies that beat the benchmark, in case such strategies exist, apart from the one that is best in the sample. While the specific BRC algorithm of White (2000) does not address this question, it could be modified to do so. The main contribution of our paper is to provide a method that goes beyond the BRC: it can identify strategies that beat the benchmark which are not detected by the BRC. This is achieved by a *stepwise* multiple testing method, where the modified BRC would correspond to the first step. But further strategies that beat the benchmark can be detected in subsequent steps, while maintaining control of the familywise error rate. So the method we propose is more powerful than the BRC.

To motivate our contribution, consider the example of a large number of actively managed mutual funds that aim to outperform the S&P 500 index, which plays the role of the benchmark. In this context, a mutual fund would outperform the S&P 500 index if its returns had at the same time a higher expected value and an equal (or lower) standard deviation. Certain forms of the efficient market hypothesis imply that no mutual fund can actually outperform the S&P 500 index (assuming that the S&P 500 index is taken as a proxy for the 'market'). A financial economist interested in the validity of certain forms of the efficient market hypothesis would therefore ask: "Is there *any* mutual fund which outperforms the S&P 500 index?". This financial economist is served well by the BRC as proposed by White (2000). On the other

[1] See Example 2.3 for a definition of the CAPM alpha.

hand, a financial advisor might be looking for mutual funds to recommend to a client. If the client's benchmark is the S&P 500 index, the financial advisor will ask: "*Which* mutual funds outperform the S&P 500 index?". In this case, the 'original' BRC is not adequate, though the modified BRC would be. The method we propose would be even more useful to the financial advisor, since it can detect more outperforming mutual funds than the modified BRC.

As a second contribution, we propose the use of studentization to improve size and power properties in finite samples. Studentization is not always feasible, but when it is we argue that it should be incorporated and we give several good reasons for doing so.

We seek to control the chance that even one true hypothesis is incorrectly rejected. Statisticians often refer to this chance as the familywise error rate (FWE); see Westfall and Young (1993). An alternative approach would be to seek to control the false discovery rate (FDR); see Benjamini and Hochberg (1995). The FDR is defined as the expected proportion of rejected hypotheses (i.e., strategies identified as beating the benchmark) that are actually true (i.e., do not beat the benchmark). The FDR approach is less strict than the FWE approach and will, generally, 'discover' a greater number of strategies beating the benchmark. But a certain proportion of these discoveries are, by design, expected to be false ones. Which approach is more suitable depends on the application and/or the preferences of the researcher. Future research will be devoted to use of a FDR framework in order to identify strategies that beat a benchmark.

The remainder of the paper is organized as follows. Section 2 describes the model, the formal inference problem, and some existing methods. Section 3 presents our stepwise method. Section 4 discusses modifications when studentization is used. Section 5 lists several possible extensions. Section 6 proposes how to choose the bootstrap block size in the context of time series data. Section 7 sheds some light on finite-sample performance via a simulation study. Section 8 provides an application to real data. Section 9 concludes. An appendix contains proofs of mathematical results, an overview of the most important bootstrap methods, and some power considerations for studentization.

2 Notation and Problem Formulation

2.1 Notation and Some Examples

One observes a data matrix $x_{t,k}$ with $1 \leq t \leq T$ and $1 \leq k \leq K+1$. The data is generated from some underlying probability mechanism P which is unknown. The row index t corresponds to distinct observations, and there are T of them. In our asymptotic framework, T will tend to infinity. The column index k corresponds to strategies, and there is a fixed number K of them. The final column, $K+1$, is reserved for the benchmark. To keep the notation unique, we include the benchmark in the data matrix even if it is nonstochastic. For compactness, we introduce the following notation: X_T denotes the complete $T \times (K+1)$ data matrix; $X_{t,\cdot}^{(T)}$ is the $(K+1) \times 1$ vector that corresponds to the t-th row of X_T; and $X_{\cdot,k}^{(T)}$ is the $T \times 1$ vector that corresponds to the k-th column of X_T.

For each strategy k, $1 \leq k \leq K$, one computes a test statistic $w_{T,k}$ that measures the 'performance' of the strategy relative to the benchmark. We assume that $w_{T,k}$ is a (measurable) function of $X_{\cdot,k}^{(T)}$ and $X_{\cdot,K+1}^{(T)}$ only. Each statistic $w_{T,k}$ tests a univariate parameter θ_k. We assume that this parameter is defined in such a way that $\theta_k \leq 0$ under the null hypothesis that strategy k does not beat the benchmark. In some instances, we will also consider studentized test statistics $z_{T,k} = w_{T,k}/\hat{\sigma}_{T,k}$, where $\hat{\sigma}_{T,k}$ estimates the standard deviation of $w_{T,k}$. In the sequel, we often call $w_{T,k}$ a 'basic' test statistic to distinguish it from the studentized statistic $z_{T,k}$. To introduce some compact notation: the $K \times 1$ vector θ collects the individual parameters of interest θ_k; the $K \times 1$ vector W_T collects the individual basic test statistics $w_{T,k}$; and the $K \times 1$ vector Z_T collects the individual studentized test statistics $z_{T,k}$.

We proceed by giving some relevant examples where several strategies are compared to a benchmark, giving rise to data snooping.

Example 2.1 (Absolute Performance of Investment Strategies) Historic returns of investment strategy k, say a particular mutual fund or a particular trading strategy, are recorded in $X_{\cdot,k}^{(T)}$. Historic returns of a benchmark, say a stock index or a buy-and-hold strategy, are recorded in $X_{\cdot,K+1}^{(T)}$. Depending on preference, these can be 'real' returns or log returns; also, returns may be recorded in excess of the risk free rate if desired. Let μ_k denote the population mean of the returns for strategy k. Based on an absolute criterion, strategy k beats the benchmark if $\mu_k > \mu_{K+1}$. Therefore, we define $\theta_k = \mu_k - \mu_{K+1}$. Using the notation

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

a natural basic test statistic is

$$w_{T,k} = \bar{x}_{T,k} - \bar{x}_{T,K+1} \tag{1}$$

As we will argue later on, a studentized statistic is preferable and given by

$$z_{T,k} = \frac{\bar{x}_{T,k} - \bar{x}_{T,K+1}}{\hat{\sigma}_{T,k}} \tag{2}$$

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

Example 2.2 (Relative Performance of Investment Strategies) The basic setup is as in the previous example. But now consider a risk-adjusted comparison of the investment strategies, based on the respective Sharpe ratios. With μ_k again denoting the mean of the returns of strategy k and with σ_k denoting their standard deviation, the corresponding Sharpe ratio is defined as $SR_k = \mu_k/\sigma_k$.[2] An investment strategy is now said to outperform the benchmark if its Sharpe Ratio is higher than the one of the benchmark. Therefore, we define $\theta_k = SR_k - SR_{K+1}$. Let

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

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

Then a natural basic test statistic is

$$w_{T,k} = \frac{\bar{x}_{T,k}}{s_{T,k}} - \frac{\bar{x}_{T,K+1}}{s_{T,K+1}} \tag{3}$$

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

Example 2.3 (CAPM alpha) Historic returns of investment strategy k, in excess of the risk-free rate, are recorded in $X^{(T)}_{\cdot,k}$. Historic returns of a market proxy, in excess of the risk-free rate, are recorded in $X^{(T)}_{\cdot,K+1}$. For each strategy k, a simple time series regression

$$x_{t,k} = \alpha_k + \beta_k x_{t,K+1} + \epsilon_{t,k}$$

is estimated by ordinary least squares (OLS). If the CAPM model holds, all intercepts α_k are equal to zero.[3] So the parameter of interest here is α_k instead of the generic θ_k. Since the CAPM model may be violated in practice, a financial advisor might attempt to identify the investment strategies which have a positive α_k. Hence, an obvious basic test statistic would be

$$w_{T,k} = \hat{\alpha}_{T,k} \tag{4}$$

Again, it can be advantageous to studentize by dividing by an estimated standard deviation of $\hat{\alpha}_{T,k}$:

$$z_{T,k} = \frac{\hat{\alpha}_{T,k}}{\hat{\sigma}_{T,k}} \tag{5}$$

Note the slight abuse of notation in this example. The vector $X^{(T)}_{\cdot,K+1}$ contains the excess returns of the market proxy, which are needed to estimate the CAPM regressions. On the other hand, the benchmark for the α_k is simply zero.

Example 2.4 (Value-at-Risk) An investment portfolio is held over time. At a given time t, the goal is to estimate the λ quantile of the conditional distribution of the portfolio return over the next period. (Here, conditional means on the past return history of the portfolio.) This quantile is generally known as the Value-at-Risk (VaR) at level λ. Common numbers for λ in practice are 1% and 5%. Many strategies to estimate the VaR exist. For a general reference, see Jorion (2000). An industry standard for VaR estimation is the well known GARCH(1,1) model. To list only a few alternative strategies: more complex GARCH models (such as GARCH(2,2), asymmetric GARCH, EGARCH, etc.), stochastic volatility, historic simulation, RiskMetrics, and extreme value theory. For a description of the various models see Bao et al. (2001) for example.[4] Most evaluation schemes of VaR estimates are simply based on 0-1 variables. In this sense, $x_{t,k} = 0$ if the return on the investment portfolio at time t exceeded the corresponding VaR estimate computed by strategy k. Otherwise $x_{t,k} = 1$. Sometimes these $x_{t,k}$ are called

[3] We trust there is no possible confusion between a CAPM alpha α_k and the level α of multiple testing methods discussed later on.

[4] An exhaustive listing of relevant papers on VaR can be found at http://www.eco.fundp.ac.be/cerefim/reources_fichiers/var.htm.

'hit variables'. Obviously, a sensible VaR strategy should produce a 'hit rate' $\bar{x}_{T,k}$ that is close to the nominal level λ. Hence, one possible test statistic would be

$$w_{T,k} = |\bar{x}_{T,K+1} - \lambda| - |\bar{x}_{T,k} - \lambda| \tag{6}$$

On the other hand, a sensible VaR strategy aims to produce a hit variable that is uncorrelated over time. Let $\mathrm{LB}_{T,k}$ denote a Ljung-Box statistic measuring autocorrelation, based on a fixed number of sample autocorrelations, applied to the time series vector $X^{(T)}_{\cdot,k}$. Then an alternative test statistic would be given by

$$w_{T,k} = \mathrm{LB}_{T,K+1} - \mathrm{LB}_{T,k} \tag{7}$$

One might even think of combining the two statistics in an appropriate way to simultaneously examine the hit rates and autocorrelations. For further evaluation schemes of VaR techniques, see Bao et al. (2001) again.

2.2 Problem Formulation

For a given strategy k, consider the individual testing problem

$$H_{0,k} \colon \theta_k \leq 0 \quad \text{vs.} \quad H_{1,k} \colon \theta_k > 0$$

Note that the parameters θ_k are allowed to vary freely of each other.[5] A multiple testing method will yield a decision concerning each testing problem by either rejecting $H_{0,k}$ or not. In an ideal world, we reject $H_{0,k}$ exactly for those strategies for which $\theta_k > 0$. In a realistic world, and given a finite amount of data, this usually cannot be achieved with certainty. In order to prevent us from declaring true null hypotheses to be false, we seek to control the familywise error rate (FWE). The FWE is defined as the probability of rejecting at least one of the true null hypotheses. More specifically, if P is the true probability mechanism, let $I_0 = I_0(P) \subset \{1, \ldots, K\}$ denote the indices of the set of true hypotheses, that is, $k \in I_0$ if and only if $\theta_k \leq 0$. The FWE is the probability under P that any $H_{0,k}$ with $k \in I_0$ is rejected:

$$\mathrm{FWE} = \mathrm{Prob}_P\{\text{Reject at least one } H_{0,k} : k \in I_0(P)\}$$

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

We require a method that, for any P, has FWE is no bigger than α, at least asymptotically. In particular, this constraint must hold for all possible configurations of true and false null hypotheses, that is, we demand *strong* control of the FWE. A method that only controls the FWE when all K null hypotheses are true is said to have *weak* control of the FWE. As remarked by Dudoit et al. (2002), this distinction is often ignored. Indeed, White (2000) only proves weak control of the FWE for his method. The remainder of the paper equates control of the FWE with strong control of the FWE.

A multiple testing method is said to control the FWE at level α if, for the given sample

[5] Holm (1979) coins this the *free combinations* condition.

size T, FWE $\leq \alpha$, for any P. A multiple testing method is said to asymptotically control the FWE at level α, if $\limsup_{T\to\infty}$ FWE $\leq \alpha$, for any P. Methods that control the FWE in finite sample can typically only be derived in special circumstances, or they suffer from lack of power because they do not incorporate the dependence structure of the test statistics. We therefore seek to control the FWE asymptotically, while trying to achieve high power at the same time.

Several well-known methods that (asymptotically) control the FWE exist. The problem is that they often have low power. What is the meaning of 'power' in a multiple testing framework? Unfortunately, there is no unique definition as in the context of a single hypothesis test. Some possible notions of power are:

- 'Global' power: the probability of rejecting all false null hypotheses.
- 'Minimal' power: the probability of rejecting at least one false null hypothesis.
- 'Average' power: the average of the individual probabilities of rejecting each false null hypothesis.

Of course, one can think of further notions. Once a given notion has been agreed upon, one can study whether a particular method is more powerful than another method in this specific sense. In rare instances, a particular method (say method 1) can be 'universally' more powerful than another method (say method 2). This happens, if for any false null hypothesis, the probability of rejecting is as large or larger for method 1 compared to method 2, and strictly larger for at least one false null hypothesis.

2.3 Existing Methods

The most familiar multiple testing method for controlling the FWE is the Bonferroni method. It works as follows. For each null hypothesis $H_{0,k}$, one computes an individual P-value $p_{T,k}$. How this P-value is computed depends on the context. It is assumed that if $H_{0,k}$ is true, the distribution of $p_{T,k}$ is Uniform (0,1), at least asymptotically.[6] The Bonferroni method at level α rejects $H_{0,k}$ if $p_{T,k} < \alpha/K$. If the null distribution of each $p_{T,k}$ is (asymptotically) uniform (0,1), then the Bonferroni method (asymptotically) controls the FWE at level α. The disadvantage of the Bonferroni method is that it is in general conservative: the FWE is in general (asymptotically) strictly less than α.[7] Indeed, it can be overly conservative, meaning that the FWE can (asymptotically) be very close to zero, which results in low power.

Actually, there exists a simple method which (asymptotically) controls the FWE at level α but is 'universally' more powerful than the Bonferroni method. This procedure is due to Holm (1979) and works as follows. The individual P-values are ordered from smallest to largest: $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(K)}$ with their corresponding null hypotheses labeled accordingly: $H_{0,(1)}$,$H_{0,(2)}$, $\ldots$, $H_{0,(K)}$. Then $H_{0,(k)}$ is rejected at level α if $p_{(j)} < \alpha/(K-j+1)$ for $j = 1, \ldots, k$. In comparison with the Bonferroni method, the criterion for the smallest P-value

[6] Actually, the following weaker assumption would be sufficient: If $H_{0,k}$ is true, then $\text{Prob}_P(p_{T,k} \leq x) \leq x$, at least asymptotically.

[7] If we say the FWE is asymptotically less than α, we mean $\limsup_{T\to\infty}$ FWE $< \alpha$.

is equally strict, α/K, but it becomes less and less strict for the larger P-values. This explains the 'universal' improvement in power. While its improvement can be substantial, the Holm method can also be very conservative.

The reason for the conservativeness of the Bonferroni and the Holm methods is that they do not take into account the dependence structure of the individual P-values. Loosely speaking, they achieve control of the FWE by assuming a worst-case dependence structure. If the true dependence structure could be accounted for, one should be able to (asymptotically) control the FWE but at the same time increase power. To illustrate, take the extreme case of perfect dependence, where all P-values are identical. In this case, one should reject $H_{0,k}$ if $p_{T,k} < \alpha$. This (asymptotically) controls the FWE but obviously is 'universally' more powerful than both the Bonferroni and Holm methods.

In many economic or financial applications, the individual test statistics are jointly dependent. Often, the dependence is positive. It is therefore important to account for the underlying dependence structure in order to avoid being overly conservative. A partial solution, for our purposes, is provided by White (2000) who coins his method the bootstrap reality check (BRC). The BRC estimates the asymptotic distribution of $\max_{1\leq k\leq K}(w_{T,k} - \theta_k)$, taking into account the correlation structure of the individual test statistics. Let k_{max} denote the index of strategy with the largest statistic w_k. The BRC decides whether or not to reject $H_{0,k_{max}}$ at level α, asymptotically controlling the FWE. It therefore answers the question whether the strategy that appears 'best' in the observed data really beats the benchmark. However, it does not attempt to identify all strategies that do. The method we present in the next section does just this. In addition, we argue that by studentizing the (individual) test statistics, in situations where studentization is feasible, one can hope to improve certain size and power properties in finite sample. This represents a second enhancement of White's (2000) approach.

To learn about further methods that control the FWE, the reader is referred to Westfall and Young (1993) as a general reference.

3 Stepwise Multiple Testing Method

Our goal is to identify all strategies for which $\theta_k > 0$. We do this by considering individual hypothesis tests

$$H_{0,k}\colon \theta_k \leq 0 \quad \text{vs.} \quad H_{1,k}\colon \theta_k > 0$$

A decision rule results in acceptance or rejection of each null hypothesis. The individual decisions are supposed to be taken in a manner that asymptotically controls the FWE at a given level α. At the same time, we want to reject as many false hypotheses as possible in finite sample.

We describe our method in the context of using basic test statistics $w_{T,k}$. The extension to the studentized case is straightforward and will be discussed later on. The method begins by relabeling the strategies according to the size of the individual test statistics, from largest to smallest. Label r_1 corresponds to the largest test statistic and label r_K to the smallest one, so that $w_{r_1} \geq w_{r_2} \geq \ldots \geq w_{r_K}$. Then the individual decisions are taken in a *stepwise*

manner.[8] In a first step, we construct a rectangular joint confidence region for the vector $\theta = (\theta_{r_1}, \ldots, \theta_{r_K})^T$ with asymptotic joint coverage probability $1-\alpha$. The confidence region is of the form

$$[w_{r_1} - c_1, \infty) \times \ldots \times [w_{r_K} - c_1, \infty) \tag{8}$$

where the common value c_1 is chosen in such as way as to ensure the proper joint coverage probability. It is not immediately clear how to achieve this in practice. Part of our contribution is describing a data-dependent way to choose c_1 in practice; details are below. If a particular individual confidence interval $[w_{r_k} - c_1, \infty)$ does not contain zero, the corresponding null hypothesis H_{0,r_k} is rejected.

If the above joint confidence region (8) has asymptotic joint coverage probability $1-\alpha$, this method asymptotically controls the FWE at level α. The method of White (2000) corresponds to computing the confidence interval $[w_{r_1} - c_1, \infty)$ only, resulting in a decision on H_{0,r_1} alone. However, his method can be easily modified to be equivalent to our first step. The critical advantage or our method is that we do not stop after the first step, unless no hypothesis is rejected. Say we reject the first K_1 relabeled hypotheses in this first step. Then $K - K_1$ hypotheses remain, corresponding to the labels r_{K_1+1} until r_K. In a second step, we construct a rectangular joint confidence region for the vector $(\theta_{r_{K_1+1}}, \ldots, \theta_{r_K})^T$ with, again, asymptotic joint coverage probability $1-\alpha$. The new confidence region is of the form

$$[w_{r_{K_1+1}} - c_{K_1+1}, \infty) \times \ldots \times [w_{r_K} - c_{K_1+1}, \infty) \tag{9}$$

where the common constant c_{K_1+1} is chosen in such a way as to ensure the proper joint coverage probability. Again, if a particular individual confidence interval $[w_{r_k} - c_{K_1+1}, \infty)$ does not contain zero, the corresponding null hypothesis H_{0,r_k} is rejected. This stepwise process is then repeated until no further hypothesis is rejected. Not stopping after the first step will, in general, reject more null hypotheses.[9] The stepwise procedure is therefore more powerful than the single-step method.[10] Nevertheless, the stepwise procedure still asymptotically controls the FWE at level α. The proof is in Theorem 3.1.

How should the value c_1 in the joint confidence region construction (8) be chosen? Ideally, one would take the $1-\alpha$ quantile of the sampling distribution of $\max_{1\le k\le K}(w_{T,r_k} - \theta_{r_k})$. This is the sampling distribution of the maximum of the individual differences "test statistic minus true parameter". Concretely, the corresponding quantile is defined as

$$c_1(1-\alpha, P) = \inf\{x : \text{Prob}_P\{\max_{1\le k\le K}(w_{T,r_k} - \theta_{r_k}) \le x\} \ge 1-\alpha\}$$

The ideal choice of c_{K_1+1}, c_{K_2+1}, and so on in the subsequent steps would be analogous. For example, the ideal $c_{r_{K_1+1}}$ for (9) would be the $1-\alpha$ quantile of the sampling distribution of

[8] Our stepwise method is a *step-down* method, since we start with the null hypothesis corresponding to the largest test statistic. The Holm method is also a step-down method. It starts with the null hypothesis corresponding to the smallest P-value, which in return corresponds to the largest test statistic. Stepwise methods that start with the null hypothesis corresponding to the smallest test statistics are called *step-up* methods; e.g., see Dunnett and Tamhane (1992).

[9] The reason is that $c_{K_1+1} < c_1$ in general.

[10] Indeed, its improvement in power is analogous to the improvement in power of the Holm method over the Bonferroni method.

$\max_{K_1+1\leq k\leq K}(w_{T,r_k}-\theta_{r_k})$ defined as

$$c_{K_1+1}(1-\alpha,P)=\inf\{x:\text{Prob}_P\{\max_{K_1+1\leq k\leq K}(w_{T,r_k}-\theta_{r_k})\leq x\}\geq 1-\alpha\}$$

The problem is that P is unknown in practice and therefore the ideal quantiles cannot be computed. The feasible solution is to replace P by an estimate $\hat{P}_T$. For an estimate $\hat{P}_T$ and any number $1\leq\tilde{K}\leq K$, define

$$c_{\tilde{K}}(1-\alpha,\hat{P}_T)=\inf\{x:\text{Prob}_{\hat{P}_T}\{\max_{\tilde{K}\leq k\leq K}(w^*_{T,r_k}-\theta^*_{T,r_k})\leq x\}\geq 1-\alpha\} \tag{10}$$

Here, the notation w^*_{T,r_k} makes clear that we mean the sampling distribution of the test statistics under $\hat{P}_T$ rather than under P; and the notation θ^*_{T,r_k} makes clear that the true parameters are those of $\hat{P}_T$ rather than those of P.[11] We can summarize our stepwise method by the following algorithm. The algorithm is based on a generic estimate $\hat{P}_T$ of P. Specific choices of this estimate, based on the bootstrap, are discussed below.

Algorithm 3.1 (Basic StepM Method)

1. Relabel the strategies in descending order of the test statistics $w_{T,k}$: strategy r_1 corresponds to the largest test statistic and strategy r_K to the smallest one.
2. Set $i=1$ and $K_1=0$.
3. For $K_i+1\leq k\leq K$, if zero is not contained in $[w_{T,r_k}-c_{K_i+1}(1-\alpha,\hat{P}_T),\infty)$, reject the null hypothesis H_{0,r_k}.
4. (a) If no null hypothesis is rejected, stop.
   (b) Otherwise, let $i=i+1$ and denote by K_i the number of all null hypotheses rejected so far. Then return to step 3.

To present our main theorem in a compact and general fashion, we make use of the following high-level assumption. Several scenarios where this assumption is satisfied will be detailed below. Introduce the following notation. $J_T(P)$ denotes the sampling distribution under P of $\sqrt{T}(W_T-\theta)$; and $J_T(\hat{P}_T)$ denotes the sampling distribution under $\hat{P}_T$ of $\sqrt{T}(W^*_T-\theta^*)$.

Assumption 3.1 *Let P denote the true probability mechanism and let $\hat{P}_T$ denote an estimate of P based on the data X_T. Assume that $J_T(P)$ converges in distribution to a nondegenerate limit distribution $J(P)$, which is continuous. Further assume that $J_T(\hat{P}_T)$ consistently estimates this limit distribution: $\rho(J_T(\hat{P}_T),J(P))\to 0$ in probability for any metric ρ metrizing weak convergence.*

[11] We implicitly assume here that, with probability one, $\hat{P}_T$ will belong to a class of distributions for which the parameter vector θ is well-defined. This holds in all of the examples in this paper.

Theorem 3.1 *Suppose Assumption 3.1 holds. Then the following statements concerning Algorithm 3.1 are true.*

(i) If $\theta_k > 0$, then the null hypothesis $H_{0,k}$ will be rejected with probability tending to one, as $T \to \infty$.

(ii) The method asymptotically controls the FWE at level α, that is, $limsup_T Prob_P\{Reject\ any\ true\ null\ hypothesis\} \leq \alpha$.

Theorem 3.1 is related to Algorithm 2.8 of Westfall and Young (1993). Our result is more flexible in the sense that we do not require their *subset pivotality* condition (see Section 2.2). Furthermore, in the context of this paper, our result is easier to apply in practice for two reasons. First, it is based on the K individual test statistics. In contrast, Algorithm 2.8 of Westfall and Young (1993) is based on the K individual P-values, which would require an extra round of computation. Second, the quantiles $c_{K_i+1}(1-\alpha, \hat{P}_T)$ are computed 'directly' from the estimated distribution $\hat{P}_T$. There is no need to impose certain null hypotheses constraints as in Algorithm 2.8 of Westfall and Young (1993).

We proceed by listing some fairly flexible scenarios where Assumption 3.1 is satisfied and Theorem 3.1 applies. The list is not meant to be exhaustive.

Scenario 3.1 (Smooth Function Model with I.I.D. Data) Consider the case of independent and identically distributed (i.i.d.) data $X_{t,\cdot}$. In the general 'smooth function' model of Hall (1992), the test statistic $w_{T,k}$ is a smooth function of certain sample moments of $X_{\cdot,k}^{(T)}$ and $X_{\cdot,K+1}^{(T)}$; and the parameter θ_k is the same function applied to the corresponding population moments. Examples that fit into this framework are given by (1), (3), and (4). If the smooth function model applies and appropriate moment conditions hold, then $\sqrt{T}(W_T - \theta)$ converges in distribution to a multivariate normal distribution with mean zero and some covariance matrix Ω. As shown by Hall (1992), one can use the i.i.d. bootstrap of Efron (1979) to consistently estimate this limiting normal distribution, that is, $\hat{P}_T$ is simply the empirical distribution of the observed data.[12]

Scenario 3.2 (Smooth Function Model with Time Series Data) Consider the case of strictly stationary time series data $X_{t,\cdot}$. The smooth function model is defined as before and the same examples (1), (3), and (4) apply; an additional example now is (7). Under moment and mixing conditions on the underlying process, $\sqrt{T}(W_T - \theta)$ again converges in distribution to a multivariate normal distribution with mean zero and some covariance matrix Ω; e.g., see White (2001). Obviously, in the time series case, the limiting covariance matrix Ω not only depends on the distribution of $X_{t,\cdot}$ but it also depends on the underlying dependence structure. The consistent estimation of the limiting distribution now requires a time series bootstrap. Künsch (1989) gives conditions under which the block bootstrap can be used; Politis and Romano (1992) show that the same conditions guarantee consistency of the circular block bootstrap; Politis and Romano (1994) give conditions under which the stationary bootstrap can be used.

[12] Hall (1992) also shows that the bootstrap approximation can be better than a normal approximation of the type $N(0, \hat{\Omega}_T)$ when the limiting covariance matrix Ω can be estimated consistently, which is not always the case.

Test statistics not covered immediately by the smooth function model can often be accommodated with some additional effort. For example, consider the test statistic (6) which involves the non-differentiable absolute value function. It is reasonable to assume that a VaR method is not quite perfect, so that the true hit rates $\mathrm{E}(x_{t,k})$ are not exactly equal to the nominal level λ. In this case Scenario 3.2 asymptotically applies. Depending upon whether $\mathrm{E}(x_{t,k}) - \lambda$ is positive or negative, the absolute value in $|\bar{x}_{T,k} - \lambda|$ can asymptotically be treated as multiplying the difference $\bar{x}_{T,k} - \lambda$ by 1 or by -1, respectively. Hence, the smooth function model applies and a time series bootstrap can be used to consistently estimate the limiting distribution of $\sqrt{T}(W_T - \theta)$. On the other hand, a problem arises if one of the $\mathrm{E}(x_{t,k})$ is exactly equal to λ and hence $\mathrm{E}(x_{t,k}) - \lambda$ is exactly equal to zero. The bootstrap now has difficulties: the parameter $|\mathrm{E}(x_{t,k}) - \lambda|$ lies on the boundary of its parameter space, the interval $[0, \infty)$ and the absolute value function is nondifferentiable. This results in inconsistency of the bootstrap; see Shao and Tu (1995, Section 3.6). In this situation the subsampling method could be used to consistently estimate the limiting distribution of $\sqrt{T}(W_T - \theta)$. Subsampling is known to work under weaker conditions than the bootstrap and would apply in this particular example; see Politis et al. (1999).

Scenario 3.3 (Strategies that Depend on Estimated Parameters) Consider the case where strategy k depends on a parameter vector β_k. In case β_k is unknown, it is estimated from the data. Denote the corresponding estimator by $\hat{\beta}_{T,k}$. Denote the value of the test statistic for strategy k, as a function of the estimated parameter vector $\hat{\beta}_{T,k}$, by $w_{T,k}(\hat{\beta}_{T,k})$. Further, let $W_T(\hat{\beta}_T)$ denote the $K \times 1$ vector collecting these individual test statistics. White (2000), in the context of a stationary time series, gives conditions under which $\sqrt{T}(W_T(\hat{\beta}_T) - \theta)$ converges to a limiting normal distribution with mean zero and some covariance matrix Ω. He also demonstrates that the stationary bootstrap can be used to consistently estimate this limiting distribution. Alternatively, the moving blocks bootstrap or the circular blocks bootstrap can be used. Note that a direct application of our Algorithm 3.1 would use the sampling distribution of $\sqrt{T}(W_T^*(\hat{\beta}_T^*) - \theta_T^*)$ under $\hat{P}_T$. That is, the β_k would be re-estimated based on data X_T^* generated from $\hat{P}_T$. But White (2000) shows that, under certain regularity conditions, it is actually sufficient to use the sampling distribution of $\sqrt{T}(W_T^*(\hat{\beta}_T) - \theta_T^*)$ under $\hat{P}_T$. Hence, in this case it is not really necessary to re-estimate the β_k parameters. Details are in White (2000).

For concreteness, we now describe how to compute the $c_{K_1+1}(1-\alpha, \hat{P}_T)$ Algorithm 3.1. In what follows $T \times (K+1)$ pseudo data matrices X^* are generated by a generic bootstrap method. In this context, $\hat{P}_T$ denotes the bootstrap data generating mechanism. The true parameter vector corresponding to $\hat{P}_T$ is denoted by θ_T^*. The specific choice of bootstrap method depends, of course, on the context. For the reader not completely familiar with the variety of bootstrap methods that do exist, we describe the most important ones in Appendix B.

Algorithm 3.2 (Computation of the $c_{K_i+1}(1-\alpha, \hat{P}_T)$ via the Bootstrap)

1. The labels $r_1, \ldots, r_K$ and the numerical values of $K_1, K_2, \ldots$ are from Algorithm 3.1.
2. Generate J bootstrap data matrices $X_T^{*,1}, \ldots, X_T^{*,J}$. We recommend to use $J \geq 1,000$ in practice.
3. From each bootstrap data matrix $X_T^{*,j}$, compute the individual test statistics $w_{T,1}^{*,j}, \ldots, w_{T,K}^{*,j}$.
4. Set $i = 1$.
5. (a) For $1 \leq j \leq J$, compute $max_{T,K_i+1}^{*,j} = \max_{K_i+1 \leq k \leq K}(w_{T,r_k}^{*,j} - \theta_{T,r_k}^*)$.
   (b) Compute $c_{K_i+1}(1-\alpha, \hat{P}_T)$ as the $1-\alpha$ empirical quantile of the J values $max_{T,K_i+1}^{*,1}, \ldots, max_{T,K_i+1}^{*,J}$.
6. Let $i = i + 1$ and return to step 5.

Remark 3.1 For convenience, one can typically use w_{T,r_k} in place of θ_{T,r_k}^* in step 5.(a) of the algorithm. Indeed, the two are the same under the following following conditions: (1) $w_{T,k}$ is a linear statistic; (2) $\theta_k = \mathrm{E}(w_{T,k})$; and (3) $\hat{P}_T$ is based on Efron's bootstrap, the circular blocks bootstrap, or the stationary bootstrap. Even if conditions (1) and (2) are met, w_{T,r_k} and θ_{T,r_k}^* are not the same if $\hat{P}_T$ is based on the moving blocks bootstrap due to 'edge' effects; see Appendix B. On the other hand, the substitution of w_{T,r_k} for θ_{T,r_k}^* does in general not effect the consistency of the bootstrap approximation and Theorem 3.1 continues to hold. Lahiri (1992) discusses this subtle point for the special case of time series data and w_{T,r_k} being the sample mean. He shows that centering by θ_{T,r_k}^* provides second-order refinements but is not necessary for first-order consistency.

4 Studentized Stepwise Multiple Testing Method

This section argues that the use of studentized test statistics, when feasible, is preferred. We first present the general method and then give three good reasons for its use.

4.1 Description of Method

An individual test statistic is now of the form $z_{T,k} = w_{T,k}/\hat{\sigma}_{T,k}$, where $\hat{\sigma}_{T,k}$ estimates the standard deviation of $w_{T,k}$. Typically, one would choose $\hat{\sigma}_{T,k}$ in such a way that the asymptotic variance of $z_{T,k}$ is equal to one. But this is actually not required for Theorem 4.1 to hold. Our stepwise method is analogous to the case of basic test statistics but slightly more complex due to the studentization. Again, $\hat{P}_T$ is an estimate of the underlying probability mechanism P based on the data X_T. Let X_T^* denote data generated from $\hat{P}_T$ and let $w_{T,k}^*$ denote a test statistic $w_{T,k}$ computed from X_T^*. Then $\hat{\sigma}_{T,k}^*$ denotes the estimated standard deviation of $w_{T,k}^*$ based on the data X_T^*.[13] We need an analog of the quantile (10) for the studentized method.

[13] Since $\hat{P}_T$ is completely specified, one actually knows the true standard deviation of $w_{T,k}^*$. However, the bootstrap mimics the real world, where standard deviation of $w_{T,k}$ is unknown, by estimating this standard deviation from the data.

It is given by

$$d_{\tilde{K}}(1-\alpha, \hat{P}_T) = \inf\{x : \text{Prob}_{\hat{P}_T}\{\max_{\tilde{K} \le k \le K} (w^*_{T,r_k} - \theta^*_{T,r_k})/\hat{\sigma}^*_{T,r_k} \le x\} \ge 1-\alpha\} \tag{11}$$

Our stepwise studentized method can now be summarized by the following algorithm.

Algorithm 4.1 (Studentized StepM Method)

1. Relabel the strategies in descending order of the test statistics $z_{T,k}$: strategy r_1 corresponds to the largest test statistic and strategy r_K to the smallest one.
2. Set $i = 1$ and $K_1 = 0$.
3. For $K_i + 1 \le k \le K$, if zero is not contained in $[w_{T,r_k} - \hat{\sigma}_{T,r_k} d_{K_i+1}(1-\alpha, \hat{P}_T), \infty)$, reject the null hypothesis H_{0,r_k}.
4. (a) If no null hypothesis is rejected, stop.
   (b) Otherwise, let $i = i+1$ and denote by K_i the number of all null hypotheses rejected so far. Then return to step 3.

A stronger version of Assumption 3.1 is needed to prove the validity of the studentized method.

Assumption 4.1 *In addition to Assumption 3.1, assume the following condition. For each k, both $\sqrt{T}\hat{\sigma}_{T,k}$ and $\sqrt{T}\hat{\sigma}^*_{T,k}$ converge to a (common) positive constant σ_k in probability.*

Theorem 4.1 *Suppose Assumption 4.1 holds. Then the following statements concerning Algorithm 4.1 are true.*

(i) If $\theta_k > 0$, then the null hypothesis $H_{0,k}$ will be rejected with probability tending to one, as $T \to \infty$.

(ii) The method asymptotically controls the FWE at level α, that is, $limsup_T Prob_P\{Reject\ any\ true\ null\ hypothesis\} \le \alpha$.

Assumption 4.1 is stricter than Assumption 3.1. Nevertheless, it covers many interesting cases. Under certain moment and mixing conditions (for the time series case), Scenarios 3.1 and 3.2 generally apply. Hall (1992) shows that a studentized version of Efron's (1979) bootstrap consistently estimates the limiting distribution of studentized statistics in the framework of Scenario 3.1. Götze and Künsch (1996) demonstrate that a studentized version of the moving blocks bootstrap consistently estimates the limiting distribution of studentized statistics in the framework of Scenario 3.2. Note that their arguments immediately apply to the circular bootstrap as well. By similar techniques the validity of a studentized version of the stationary bootstrap can be established. Relevant examples of practical interest are given by (2) and (5). Examples where less obvious studentized test statistics exist are given by (6) and (7).

For concreteness, we now describe how to compute the $d_{K_1+1}(1-\alpha, \hat{P}_T)$ in Algorithm 4.1. Again, $T \times (K+1)$ pseudo data matrices X^*_T are generated by a generic bootstrap method.

Algorithm 4.2 (Computation of the $d_{K_i+1}(1-\alpha, \hat{P}_T)$ via the Bootstrap)

1. The labels $r_1, \ldots, r_K$ and the numerical values of $K_1, K_2, \ldots$ are from Algorithm 4.1.
2. Generate J bootstrap data matrices $X_T^{*,1}, \ldots, X_T^{*,J}$. We recommend to use $J \geq 1,000$ in practice.
3. From each bootstrap data matrix $X_T^{*,j}$, compute the individual test statistics $w_{T,1}^{*,j}, \ldots, w_{T,K}^{*,j}$. Also, compute the corresponding estimated standard deviations $\hat{\sigma}_{T,1}^{*,j}, \ldots, \hat{\sigma}_{T,K}^{*,j}$.
4. Set $i = 1$.
5. (a) For $1 \leq j \leq J$, compute $max_{T,K_i+1}^{*,j} = \max_{K_i+1 \leq k \leq K}(w_{T,r_k}^{*,j} - \theta_{T,r_k}^{*})/\hat{\sigma}_{T,r_k}^{*,j}$.
   (b) Compute $d_{K_i+1}(1-\alpha, \hat{P}_T)$ as the $1-\alpha$ empirical quantile of the J values $max_{T,K_i+1}^{*,1}, \ldots, max_{T,K_i+1}^{*,J}$.
6. Let $i = i + 1$ and return to step 5.

Remark 3.1 applies here in spirit.

How to studentize properly depends on the context. In the case of i.i.d. data there is usually an obvious 'formula' for $\hat{\sigma}_{T,k}$, which is applied to the data matrix X_T. To give an example, the formula for $\hat{\sigma}_{T,k}$ corresponding to the test statistic (1) based on i.i.d. data is given by

$$\hat{\sigma}_{T,k} = \sqrt{\frac{\sum_{t=1}^{T}(x_{t,k} - x_{t,K+1} - \bar{x}_{T,k} + \bar{x}_{T,K+1})^2}{T-1}} \tag{12}$$

In the Efron bootstrap world, the value of $\hat{\sigma}_{T,k}^{*}$ is then obtained by applying the same formula to the bootstrap data matrix X_T^{*}. Things get more complex in the case of stationary time series data. There no longer exists a simple formula to compute $\hat{\sigma}_{T,k}$ from X_T. Instead, one typically uses a kernel variance estimator that can be described by a certain algorithm; e.g., see Andrews (1991) and Andrews and Monahan (1992). In principle, $\hat{\sigma}_{T,k}^{*}$ can be obtained by applying the same algorithm to the bootstrap data matrix X_T^{*}. When X_T^{*} is obtained by the moving blocks bootstrap or the circular blocks bootstrap, Götze and Künsch (1996) suggest to use a 'natural' variance estimator $\hat{\sigma}_{T,k}^{*}$. This is due to the two facts that (1) these two methods generate a bootstrap data sequence by concatenating blocks of data of a fixed size and that (2) the individual blocks are selected independently of each other. For the sake of space, we refer the interested reader to Götze and Künsch (1996) and Romano and Wolf (2003) to learn more about 'natural' block bootstrap variance estimators.

4.2 Reasons for Studentization

We now provide three reasons for making the additional effort of studentization.

The first reason is power. The studentized method is not uniformly more powerful than the basic method. However, it performs better for several reasonable definitions of power. Details can be found in Appendix C.

The second reason is level (or size). Consider for the moment the case of a single null hypothesis $H_{0,k}$ of interest. Under certain regularity conditions, it is well-known that (1) bootstrap confidence intervals based on studentized statistics provide asymptotic refinements in terms of coverage level; and that (2) bootstrap tests based on studentized test statistics provide asymptotic refinements in terms of level. The underlying theory is provided by Hall (1992) for the case of i.i.d. data and by Götze and Künsch (1996) for the case of stationary data. The common theme is that one should use asymptotically pivotal (test) statistics in bootstrapping. This is only partially satisfied for our studentized multiple testing method, since we studentize the test statistics *individually*. Hence, the limiting *joint* distribution is not free of unknown population parameters. Such a limiting joint distribution could be obtained by a joint studentization, taking also into account the covariances of the individual test statistics $w_{T,k}$. However, this would no longer result in the rectangular joint confidence regions which are the basis for our stepwise testing method. A joint studentization is not feasible for our purposes. While individual studentization cannot be proven to result in asymptotic refinements in terms of the level, there is still hope that it leads to finite sample improvements, which might show up in simulation studies; see Section 7.

The third reason is individual coverage probabilities. As a by-product, the first step of our multiple testing method yields a joint confidence region for the parameter vector θ. The basic method yields the following region

$$[w_{T,r_1} - c_1(1-\alpha, \hat{P}_T), \infty) \times \ldots \times [w_{T,r_K} - c_1(1-\alpha, \hat{P}_T), \infty) \tag{13}$$

The studentized method yields the following region

$$[w_{T,r_1} - \hat{\sigma}_{T,r_1} d_1(1-\alpha, \hat{P}_T), \infty) \times \ldots \times [w_{T,r_K} - \hat{\sigma}_{T,r_K} d_1(1-\alpha, \hat{P}_T), \infty) \tag{14}$$

If the sample size T is large, both regions (13) and (14) have joint coverage probability of about $1-\alpha$. But they are distinct as far as the individual coverage probabilities for the θ_k values are concerned. Assume that the test statistics $w_{T,k}$ have different standard deviations, which happens in many applications. Say w_{r_1} has a smaller standard deviation than w_{r_2}. Then the confidence interval for θ_{r_1} derived from (13) will typically have a larger (individual) coverage probability compared to the confidence interval for θ_{r_2}. This is not the case for (14) where, thanks to studentization, the individual coverage probabilities are comparable and hence the individual confidence intervals are 'balanced'. The latter is clearly a desirable property; see Beran (1988). Indeed, we make a decision concerning H_{0,r_k} by inverting a confidence interval for θ_{r_k}. Balanced confidence intervals result in a balanced power 'distribution' among the individual hypotheses. Unbalanced confidence intervals, obtained from basic test statistics, distribute the power unevenly among the individual hypotheses.

To sum up, when the standard deviations of the basic test statistics $w_{T,k}$ are different, the $w_{T,k}$ live on different scales. Comparing one basic test statistic to another is then like comparing apples to oranges. If one wants to compare apples to apples, one should use the studentized test statistics $z_{T,k}$.[14]

[14] Alternatively, one could compare individual P-values. But this becomes more involved in practice.

5 Possible Extensions

The aim of this paper is to introduce a new multiple testing methodology based on *stepwise* joint confidence regions. For sake of brevity and succinctness, we have presented the methodology in a compact yet rather flexible framework. This section briefly lists several possible extensions. The details are left for future research.

In our setup, the individual null hypotheses $H_{k,0}$ are one-sided. This makes sense because we want to test whether individual strategies *improve* upon a benchmark, rather than whether their performance is just *different* from the benchmark. Nevertheless, for other multiple testing problems two-sided tests can be more appropriate; for example, see the multiple regression example of the next paragraph. If two-sided tests are preferred, our methods can be easily adapted. Instead of one-sided joint confidence regions, one would construct two-sided joint confidence regions. To give an example, the first-step region based on simple test statistics would look as follows

$$[w_{T,r_1} \pm c_{1,|\cdot|}(1-\alpha, \hat{P}_T)] \times \ldots \times [w_{T,r_K} \pm c_{1,|\cdot|}(1-\alpha, \hat{P}_T)]$$

Here, $c_{1,|\cdot|}(1-\alpha, \hat{P}_T)$ estimates the $1-\alpha$ quantile of the *two-sided* sampling distribution under P of $\max_{1\leq k\leq K} |w_{T,r_k} - \theta_{r_k}|$. The corresponding modifications of Algorithms 3.1 and 3.2 are straightforward. Note that in the modified Algorithm 3.1, the strategies would have to relabeled in descending order of the $|w_{T,k}|$ values instead of the $w_{T,k}$ values.

Since our focus is on comparing a number of strategies to a common benchmark, we assume that a test statistic $w_{T,k}$ is a function of the vectors $X^{(T)}_{\cdot,k}$ and $X^{(T)}_{\cdot,K+1}$ only, where $X^{(T)}_{\cdot,K+1}$ corresponds to the benchmark. This assumption is not crucial for our multiple testing methods. Take the example of a multiple regression model with regression parameters $\theta_1, \theta_2, \ldots, \theta_K$. The individual null hypotheses are of the form $H_{0,k}$: $\theta_k = \theta_{0,k}$ for some constants $\theta_{0,k}$. The alternatives can be (all) one-sided or (all) two-sided. Note that there is no benchmark here, so the last column of the $T \times (K+1)$ data matrix X_T would correspond to the response variable while the first K columns would respond to the explanatory variables. In this setting, $w_{T,k} = \hat{\theta}_{T,k}$, where the estimation might be done by OLS say. Obviously, $w_{T,k}$ will be a function of the entire data matrix now. Still, our multiple testing methods can be applied to this setting and the modifications are minor: one rejects H_{0,r_k} if θ_{0,r_k}, rather than 0, is not contained in a confidence interval for θ_{r_k}.

We assume the usual $\sqrt{T}$ convergence, meaning that $\sqrt{T}(W_T - \theta)$ has a nondegenerate limiting distribution. In nonstandard situations, the rate of convergence can be another function of T instead of the square root. In these instances, the bootstrap often fails to consistently estimate the limiting distribution. But if this happens, one can use the subsampling method instead; see Politis et al. (1999) for a general reference. Our multiple testing methods can be modified for the use of subsampling instead of the bootstrap. Examples where the rate of convergence is $T^{1/3}$ can be found in Rodríguez-Poo et al. (2001).[15] An example where the rate of convergence is T can be found in Gonzalo and Wolf (2003).

[15] This paper focuses on the use of subsampling for testing purposes. But the modifications for the construction of confidence intervals are straightforward.

6 Choice of Block Sizes

If the data sequence is a stationary time series, one needs to use a time series bootstrap. Each possible choice – the moving blocks bootstrap, the circular blocks bootstrap, or the stationary bootstrap – involves the problem of choosing the block size b in practice. (When the stationary bootstrap is used, we denote by b the expected block size.) Asymptotic requirements on b include $b \to \infty$ and $b/T \to 0$ as $T \to \infty$, which is of practical help. In this section, we give concrete advice on how to select b in a data-dependent fashion. Note that the block size b has to be chosen 'from scratch' in each step of our stepwise multiple testing methods, and the individual choices may well be different.

Consider the ith step of a stepwise procedure. The goal is to construct a joint confidence region for the vector $(\theta_{r_{K_i+1}}, \ldots, \theta_{r_K})'$ with nominal coverage probability of $1-\alpha$. The actual coverage probability in finite sample, denoted by $1-\lambda$, is generally not exactly equal to $1-\alpha$. Moreover, conditional on P and T, we can think of the actual coverage probability as a function of the block size b. This function $g : b \to 1-\lambda$ was coined the *calibration* function by Loh (1987). The idea is now to adjust the 'input' b in order to obtain the actual coverage probability close to the desired one. If $g(\cdot)$ was known, so would be the optimal adjustment, that is, the optimal choice of b. Indeed, one should find $\tilde{b}$ that minimizes $|g(b)-(1-\alpha)|$ and use the value $\tilde{b}$ as the block size in practice; note that $|g(b)-(1-\alpha)| = 0$ may not always have a solution.

Unfortunately, the function $g(\cdot)$ depends on the underlying probability mechanism P and is unknown. We therefore propose a method to estimate $g(\cdot)$. The idea is that in principle we could simulate $g(\cdot)$ if P were known by generating data of size T according to P and by computing joint confidence regions for $(\theta_{r_{K_i+1}}, \ldots, \theta_{r_K})'$ for a number of different block sizes b. This process is then repeated many times and for a given b one estimates $g(b)$ as the fraction of the corresponding intervals that contain the true parameter. The method we propose is identical except that P is replaced by a semi-parametric estimate $\tilde{P}_T$. For compact notation, define $\theta_{K_i}^{(r)} = (\theta_{r_{K_i+1}}, \ldots, \theta_{r_K})'$.

Algorithm 6.1 (Choice of Block Sizes)

1. The labels $r_1, \ldots, r_K$ and the numerical values $K_1, K_2, \ldots$ are from Algorithm 3.1 if the basic method is used or from Algorithm 4.1 if the studentized method is used, respectively.
2. Fit a semi-parametric model $\tilde{P}_T$ to the observed data X_T.
3. Fix a selection of reasonable block sizes b.
4. Generate M data sets $\tilde{X}_T^1, \ldots, \tilde{X}_T^M$ according to $\tilde{P}_T$.
5. Set $i = 1$.
6. For each data set $\tilde{X}_T^m$, $m = 1, \ldots, M$, and for each block size b, compute a joint confidence region $\mathrm{CI}_{m,b}$ for $\theta_{K_i}^{(r)}$
7. Compute $\hat{g}(b) = \#\{\theta_{K_i}^{(r)}(\tilde{P}_T) \in \mathrm{CI}_{m,b}\}/M$.

8. Find the value of $\tilde{b}$ that minimizes $|\hat{g}(b) - (1-\alpha)|$ and use this value $\tilde{b}$ in the construction of the ith joint confidence region.

9. Let $i = i + 1$ and return to step 6.

Several remarks concerning this algorithm are in order.

Remark 6.1 The motivation of fitting a semi-parametric model $\tilde{P}_T$ to P is that such models do not involve a block size of their own. In general, we suggest to use a low-order vector autoregressive (VAR) model. While such a model will usually be misspecified, its role can be compared to the role of a semi-parametric model in the prewhitening process for prewhitened kernel variance estimation; e.g. see Andrews and Monahan (1992). Even if the model is misspecified, it should contain some valuable information on the dependence structure of the true mechanism P that can be exploited to estimate $g(\cdot)$.

Remark 6.2 The method for choosing the block sizes is computationally expensive. To estimate $g(b)$, a total of M joint confidence regions have to be computed, and each joint confidence region is based on J bootstrap samples. Hence, a total of JM bootstrap samples will have to be generated and processed.

Remark 6.3 Algorithm 6.1 provides a reasonable method to select the block sizes in a practical application. We do not claim any asymptotic optimality properties. On the other hand, in the simpler setting of constructing confidence intervals for a single parameter of interest, Romano and Wolf (2003) find that this algorithm works very well in a simulation study.

7 Simulation Study

The goal of this section is to shed some light on the finite sample performance of our methods by means of a simulation study. It should be pointed out that any data generating process (DGP) has a large number of input variables, including: the number of observations T, the number of strategies K, the number of false hypotheses, the numerical values of the parameters θ_k, the dependence structure across strategies, and the dependence structure over time (in case of time series data). An exhaustive study is clearly beyond the scope of this paper and our conclusions will necessarily be limited. The main interest is to see how the multi-step method compares to the single-step method and to judge the effect of studentization. Performance criteria are the empirical FWE and the (average) number of false hypotheses that are rejected. To save space, we only report results for the nominal level $\alpha = 0.1$.[16]

To keep the computational burden manageable, we consider the simplest case of comparing the population mean of a strategy to that of the benchmark, as in Example 2.1. Simulation results that are not reported show that when the standard deviations of all strategies are the same, and the data are i.i.d., the basic and the studentized methods perform nearly identically. Hence, we only report results for scenarios where the standard deviations are not identical.

[16] The results for $\alpha = 0.05$ are similar and available from the authors upon request.

7.1 I.I.D. Data

We start with observations that are i.i.d. over time. The number of observations is $T = 100$ and there are $K = 40$ strategies. A basic test statistic is given by (1) and a studentized test statistic is given by (2). The studentized statistic uses the formula (12). The bootstrap method is Efron's bootstrap. The number of bootstrap repetitions is $J = 200$ due to the computational expense of the simulation study. The number of DGP repetitions in each scenario is 2,000.

The distribution of the observation $X_{t,\cdot}$ is jointly normal. There is common correlation between the individual strategies and also between strategies and the benchmark. This common correlation is either equal to 0 or equal to 0.5. The mean of the benchmark is always 1. We consider two scenarios.

In the first class of DGPs, there are four cases as far as the means of the strategies are concerned: all means are equal to 1; six of the means are equal to 1.4 and the remaining ones are equal to 1; twenty of the means are equal to 1.4 and the remaining ones are equal to 1; all forty means are equal to 1.4. The standard deviation of the benchmark is always equal to 1. As far as the standard deviations of the strategies are concerned, half of them are equal to 1 and the other half are equal to 2. Note that the strategies that have the same mean as the benchmark always have half their standard deviations equal to 1 and the other half equal to 2; the same for the strategies with means greater than that of the benchmark. The results are reported in Table 1. The (strong) control of the FWE is satisfactory for all methods (single-step vs. multi-step and basic vs. studentized). When comparing the average number of false hypotheses rejected, one observes: (i) the multi-step method improves over the single-step method; (ii) the studentized method improves significantly over the basic method; (iii) the single-step basic method—that is, the modified White (2000) approach—performs worst in all scenarios. Finally, the bootstrap successfully captures the dependence structure across strategies. When the cross correlation is 0.5 as opposed to 0, a larger number of false hypotheses are rejected on average.

In the second class of DGPs, the strategies that are superior to the benchmark have their means evenly distributed between 1 and 4. Again there are four cases: all means are equal to 1; six of the means are bigger than 1 and the remaining ones are equal to 1; twenty of the means are bigger than 1 and the remaining ones are equal to 1; all forty means are bigger than 1. For example, when six of the means are bigger than 1, those are 1.5, 2, 2.5, 3.0, 3.5 and 4.0. When twenty of the means are bigger than 1, those are 1.15, 1.30, ..., 3.85, 4.0. For any strategy, the standard deviation is 2 times the corresponding mean. For example, the standard deviation of a strategy with mean 1 is 2; the standard deviation of a strategy with mean 1.5 is 3; and so on. The results are reported in Table 2. The (strong) control of the FWE is satisfactory for all methods (single-step vs. multi-step and basic vs. studentized). When comparing the average number of false hypotheses rejected, one observes: (i) the multi-step method improves significantly over the single-step method; (ii) the studentized method improves over the basic method for the single-step approach, however it is somewhat worse than the basic method for the multi-step approach; (iii) the single-step basic method—that is, the modified White (2000) approach—performs worst in five out of six scenarios. Finally, the bootstrap successfully captures the dependence structure across strategies. When the cross

correlation is 0.5 as opposed to 0, a larger number of false hypotheses are rejected on average.

7.2 Time Series Data

The main modification with respect to the previous DGPs is that now the observations are not i.i.d. but rather a multivariate normal stationary time series. Marginally, each vector $X_{k,\cdot}$ is a AR(1) process with autoregressive coefficient $\rho = 0.6$. The number of observations is increased to $T = 200$ to make up for the dependence over time. A basic test statistic is given by (1) and a studentized test statistic is given by (2). The studentized statistic uses a prewhitended kernel variance estimator based on the QS kernel and the corresponding automatic choice of bandwidth of Andrews and Monahan (1992). The bootstrap method is the circular block bootstrap. The studentization in the bootstrap world uses the corresponding 'natural' variance estimator; for details, see Götze and Künsch (1996) or Romano and Wolf (2003). The number of bootstrap repetitions is $J = 200$ due to the computational expense of the simulation study. The number of DGP repetitions in each scenario is 2,000.

The choice of the block size is an important practical problem in applying a block bootstrap. Unfortunately, the data-dependent Algorithm 6.1 is computationally too expensive to be incorporated in our simulation study. (This would not be a problem in a practical application where only one data set has to processed, instead of several thousand as in a simulation study.) We therefore found the 'reasonable' block sizes $b = 20$ for the basic method and $b = 15$ the studentized method, respectively, by trial and error. Given that a variant of Algorithm 6.1 is seen to perform very well in a less computer intensive simulation study of Romano and Wolf (2003)[17], we are quite confident that would also perform well in the context of multiple testing. We cannot offer any simulation evidence to this end, however.

The first class of DGPs is similar to the i.i.d. case, except that the strategy means greater than 1 are equal to 1.6 rather than 1.4. The results are reported in Table 3.

The second class of DGPs is similar to the i.i.d. case, except that the strategy means bigger than 1 are evenly distributed between 1 and 7 rather than between 1 and 4. The results are reported in Table 4.

Contrary to the findings for i.i.d. data, the basic method does not provide a satisfactory control of the FWE in finite sample and is too liberal. (This is not because of the choice of block size $b = 20$ but was observed for all other block sizes we tried as well.) On the other hand, the studentized method does a good job of controlling the FWE. Again, the multi-step method does in general reject more false hypotheses compared to the single-step method and the magnitude of the improvement depends on the underlying probability mechanism.

[17] Their simulation study is for confidence intervals for a single regression coefficient, which is much faster to implement compared to a multiple testing method.

8 Empirical Application

This section provides an application to real data, using Example 2.3. It is quite common in financial econometrics to estimate CAPM alphas based on a time series of the past 120 monthly return data. We use monthly returns from 12/1992 until 12/2002, provided by DataStream. The market proxy is the S&P 500 index and the 'strategies' are the $K = 100$ largest stocks, as measured by their market value in 12/2002, with a complete 10 year return history. The CAPM model for each stock is estimated via OLS. A basic test statistic is given by (4). A studentized test statistic is given by (5). Studentization uses a kernel variance estimator based on the prewhitened QS kernel and the corresponding automatic choice of bandwidth of Andrews and Monahan (1992). The bootstrap method is the circular block bootstrap. The studentization in the bootstrap world uses the corresponding 'natural' variance estimator; for details, see Götze and Künsch (1996) or Romano and Wolf (2003). Given the well-known low autocorrelation of monthly stock returns, we employ a relatively small block size of $b = 5$. The number of bootstrap repetitions is $J = 1,000$.

Table 5 lists the ten largest basic test statistics together with the corresponding stocks. Table 6 lists the ten largest studentized test statistics together with the corresponding stocks. Not surprisingly, the two lists of stocks are quite different. Once the magnitude of the uncertainty about the basic test statistics is taking into account through studentization, the order of the test statistics changes.

We now use the various multiple testing methods to identify stocks with a positive CAPM alpha, asymptotically controlling the FWE at level 0.1. The basic single-step method, that is, the modified version of White (2000), identifies the stocks corresponding to the three largest basic statistics: AOL Time Warner, Qualcomm, and Dell Computer. The basic multi-step method further identifies Oracle and Clear Chl. Comms. (both in the second step). On the other hand, the studentized method identifies the stocks corresponding to the six largest studentized statistics: Kohls, Citigroup, Clear Chl. Comms., AOL Time Warner, MBNA Corp., and Fifth Third Bancorp.. All of these are identified in the first step, and no further stocks are identified in subsequent steps.

9 Conclusion

In this paper, we advocated a *stepwise* multiple testing method in the context of comparing several strategies to a common benchmark. To account for the undesirable effects of data snooping, our method asymptotically controls for the familywise error rate (FWE). Loosely speaking, the FWE is defined as the probability of falsely rejecting one or more of the true null hypotheses. Our proposal extends the bootstrap reality check (BCR) of White (2000). The way it was originally presented, the BCR only addresses whether the strategy that appears 'best' in sample actually beats the benchmark, asymptotically controlling for the FWE. But the BCR can easily be modified to potentially identify several strategies that do so. Our stepwise method would regard this modified BCR as the first step. The crucial difference is that if some hypotheses are rejected in this first step, our method does not stop there and potentially will

reject further hypotheses in subsequent steps. Therefore, our method is more powerful without sacrificing the asymptotic control of the FWE. To decide which hypotheses to reject in a given step, we construct a joint confidence region for the set of parameters pertaining to the set of null hypotheses not rejected in previous steps. This joint confidence region is determined by an appropriate bootstrap, depending upon whether the observed data are i.i.d. or a time series.

In addition, we proposed the use of studentization in situations when it is feasible. There are several reasons why we prefer studentization, one of them being that it results in a more even distribution of power among the individual tests. We also showed that, for several sensible definitions of power, it is more powerful compared to not studentizing.

It is important to point out that our ideas can be generalized. For example, we focused on comparing several strategies to a common benchmark. But there are alternative contexts where multiple testing, and hence data snooping, occurs. One instance is simultaneous inference for individual regression coefficients in a multiple regression framework. With suitable modifications, our stepwise testing method can be employed in such alternative contexts. To give another example, the bootstrap may not result in asymptotic control of the FWE in non-standard situations, such as when the rate of convergence is different from the square root of the sample size. In many of such situations one can then use a stepwise method based on subsampling rather than on the bootstrap.

Some simulation studies investigated finite-sample performance. Of course, stepwise methods reject more false hypotheses than their single-step counterparts. Our simulations show that the actual size of the improvement depends on the underlying probability mechanism—for example, through the number of false null hypotheses, their respective magnitudes, etc.—and can range from negligible to dramatic. On the other hand, the studentized stepwise method can be less powerful or more powerful than the non-studentized (or 'basic') stepwise method, depending on the underlying mechanism. We still advocate the use of studentization: (i) the underlying mechanism is unknown in practice, so one cannot find whether studentizing is more powerful or not; (ii) but studentizing always results in a more even (or 'balanced') distribution of power among the individual hypotheses, which is a desirable property. In addition, the use of studentization appears particularly important in the context of time series data. Our simulations show that non-studentized (or 'basic') method can fail to control the FWE in finite samples when there is notable dependence over time; the studentized method does much better.

A Proofs of Mathematical Results

We begin by stating two lemmas. The first one is quite obvious.

Lemma A.1 *Suppose that Assumption 3.1 holds. Let L_T denote a random variable with distribution $J_T(P)$ and let L denote a random variable with distribution $J(P)$. Let $I = \{i_1, \ldots, i_m\}$ be a subset of $\{1, \ldots, K\}$. Denote by $L(I)$ the corresponding subset of L, that is, $L(I) = (L_{i_1}, \ldots, L_{i_m})'$. Analogously, denote by $L_T(I)$ the corresponding subset of L_T, that is, $L_T(I) = (L_{T,i_1}, \ldots, L_{T,i_m})'$.*

Then for any subset I of $\{1, \ldots, K\}$, $L_T(I)$ converges in distribution to $L(I)$.

Lemma A.2 *Suppose that Assumption 3.1 holds. Let $I = \{i_1, \ldots, i_m\}$ be a subset of $\{1, \ldots, K\}$. Define $L(I)$ and $L_T(I)$ as in Lemma A.1 before and use analogous definitions for $W_T(I)$ and $\theta(I)$. Also, define*

$$c_I(1-\alpha, \hat{P}_T) = \inf\{x : Prob_{\hat{P}_T}\{\max_{k \in I}(w^*_{T,k} - \theta^*_{T,k}) \leq x\} \geq 1-\alpha\} \tag{15}$$

Then

$$[w_{i_1} - c_I(1-\alpha, \hat{P}_T), \infty) \times \ldots \times [w_{i_m} - c_I(1-\alpha, \hat{P}_T), \infty) \tag{16}$$

is a joint confidence region (JCR) for $(\theta_{i_1}, \ldots, \theta_{i_m})'$ with asymptotic coverage probability of $1-\alpha$.

Proof To start out, note that

$$\begin{aligned} \text{Prob}_P\{(\theta_{i_1}, \ldots, \theta_{i_m})' \in \text{JCR (16)}\} &= \text{Prob}_P\{\max(W_T(I) - \theta(I)) \leq c_I(1-\alpha, \hat{P}_T)\} \\ &= \text{Prob}_P\{\max \sqrt{T}(W_T(I) - \theta(I)) \leq \sqrt{T} c_I(1-\alpha, \hat{P}_T)\} \end{aligned}$$

By Assumption 3.1, Lemma A.1, and the continuous mapping theorem, $\max L_T(I)$ converges weakly to $\max L(I)$, whose distribution is continuous. Our notation implies that the sampling distribution under P of $\max \sqrt{T}(W_T(I) - \theta(I))$ is identical to the distribution of $\max L_T(I)$, so it converges weakly to $\max L(I)$. By similar reasoning, also the sampling distribution under $\hat{P}_T$ of $\max \sqrt{T}(W^*_T(I) - \theta^*(I))$ converges weakly to $\max L(I)$. The proof that

$$\text{Prob}_P\{\max \sqrt{T}(W_T(I) - \theta(I)) \leq \sqrt{T} c_{(I)}(1-\alpha, \hat{P}_T)\} \to 1-\alpha$$

is now very similar to the proof of Theorem 1 of Beran (1984). ■

Proof of Theorem 3.1 We start with the proof of (i). Assume that $\theta_k > 0$. Assumption 3.1 and definition (10) imply that $\sqrt{T} c_1(1-\alpha, \hat{P}_T)$ is stochastically bounded. So $c_1(1-\alpha, \hat{P}_T)$ converges to zero in probability. By Assumption 3.1 and Lemma A.1, $\sqrt{T}(w_{T,k} - \theta_k)$, converges weakly. So $w_{T,k}$ converges to θ_k in probability. These two convergence results imply that, with probability tending to one, $w_{T,k} - c_1(1-\alpha, \hat{P}_T)$ will be greater than $\theta_k/2$, resulting in the rejection of $H_{k,0}$ in the first step.

We now turn to the proof of (ii). The result trivially holds in case all null hypotheses $H_{k,0}$ are false. So assume at least one of them is true. Let $I_0 = I_0(P) \subset \{1, \ldots, K\}$ denote the indices of the set of true hypotheses; that is, $k \in I_0$ if and only if $\theta_k \leq 0$. Denote the number of true hypotheses by m and let $I_0 = \{i_1, \ldots, i_m\}$. Part (i) implies that, with probability tending to one, all false hypotheses will be rejected in the first step. Since $c_{I_0}(1-\alpha, \hat{P}_T) \leq c_1(1-\alpha, \hat{P}_T)$, where $c_{I_0}(1-\alpha, \hat{P}_T)$ is defined analogously to (15), we therefore have

$$\begin{aligned}
\limsup_{T\to\infty} \text{FWE} &\leq \limsup_{T\to\infty} \text{Prob}_P\{0 \notin [w_{T,k} - c_{I_0}(1-\alpha, \hat{P}_T), \infty) \text{ for at least one } k \in I_0\} \\
&\leq \limsup_{T\to\infty} \text{Prob}_P\{\theta_k \notin [w_{T,k} - c_{I_0}(1-\alpha, \hat{P}_T), \infty) \text{ for at least one } k \in I_0\} \\
&= 1 - \liminf_{T\to\infty} \text{Prob}_P\{\theta(I_0) \in [w_{T,i_1} - c_{I_0}(1-\alpha, \hat{P}_T), \infty) \times \ldots \times [w_{T,i_m} - c_{I_0}(1-\alpha, \hat{P}_T), \infty)\} \\
&\leq 1 - (1-\alpha) \quad \text{(by Lemma A.2)} \\
&= \alpha
\end{aligned}$$

This proves the control of the FWE at level α.[18] ■

Proof of Theorem 4.1 The proof is very similar to the proof of Theorem 3.1 and hence it is omitted. ■

B Overview of Bootstrap Methods

For readers not completely familiar with the variety of bootstrap methods that do exist, we now briefly describe the most important ones. To recall our notation, the observed data matrix is X, which can be 'decomposed' in the observed data sequence $X_{1,\cdot}, X_{2,\cdot}, \ldots X_{T,\cdot}$. When the data are i.i.d, the order of this sequence is of no importance. When the data is a time series, the order is crucial.

Bootstrap B.1 (Efron's Bootstrap)
The bootstrap of Efron (1979) is appropriate when the data are i.i.d.. The method generates random indices $t_1^*, t_2^*, \ldots, t_T^*$ i.i.d. from the discrete uniform distribution on the set $\{1, 2, \ldots, T\}$. The bootstrap sequence is then given by $X_{1,\cdot}^*, X_{2,\cdot}^*, \ldots X_{T,\cdot}^* = X_{t_1^*,\cdot}, X_{t_2^*,\cdot}, \ldots, X_{t_T^*,\cdot}$. The corresponding $T \times (K+1)$ bootstrap data matrix is denoted by X_T^*. The probability mechanism generating a X_T^* is denoted by $\hat{P}_T$.

Bootstrap B.2 (Moving Blocks Bootstrap)
The moving blocks bootstrap of Künsch (1989) is appropriate when the data sequence is a stationary time series. It generates a bootstrap sequence by concatenating blocks of data which are resampled from the original series. A particular block $B_{t,b}$ is defined by its starting index t and by its length or block size b, that is, $B_{t,b} = \{X_{t,\cdot}, X_{t+1,\cdot} \ldots, X_{t+b-1,\cdot}\}$. The moving blocks bootstrap selects a fixed block size $1 < b < T$. It then chooses random starting

[18] Since the argument does not assume that all K null hypotheses are true, we have indeed proven strong control of the FWE.

indices $t_1^*, t_2^*, \ldots, t_l^*$ i.i.d. from the uniform distribution on the set $\{1, 2, \ldots, T - b + 1\}$, where l is the smallest integer for which $l \times b \geq T$. The thus selected blocks are concatenated as $\{B_{t_1^*,b}, B_{t_2^*,b}, \ldots, B_{t_l^*,b}\}$. If $l \times b > T$, the sequence is truncated at length T to obtain the bootstrap sequence $X_{1,\cdot}^*, X_{2,\cdot}^*, \ldots X_{T,\cdot}^*$. The corresponding $T \times (K+1)$ bootstrap data matrix is denoted by X_T^*. The probability mechanism generating a X_T^* is denoted by $\hat{P}_T$.

Bootstrap B.3 (Circular Blocks Bootstrap)
The circular blocks bootstrap of Politis and Romano (1992) is appropriate when the data sequence is a stationary time series. It generates a bootstrap sequence by concatenating blocks of data which are resampled from the original series. The difference with respect to the moving blocks bootstrap is that the original data are 'wrapped' into a 'circle' in the sense of $X_{T+1,\cdot} = X_{1,\cdot}, X_{T+2,\cdot} = X_{2,\cdot}, \ldots$. As before, a particular block $B_{t,b}$ is defined by its starting index t and by its block size b. The circular blocks bootstrap selects a fixed block size $1 < b < T$. It then chooses random starting indices $t_1^*, t_2^*, \ldots, t_l^*$ i.i.d. from the uniform distribution on the set $\{1, 2, \ldots, T\}$, where l is the smallest integer for which $l \times b \geq T$. The thus selected blocks are concatenated as $\{B_{t_1^*,b}, B_{t_2^*,b}, \ldots, B_{t_l^*,b}\}$. If $l \times b > T$, the sequence is truncated at length T to obtain the bootstrap sequence $X_{1,\cdot}^*, X_{2,\cdot}^*, \ldots X_{T,\cdot}^*$. The corresponding $T \times (K+1)$ bootstrap data matrix is denoted by X_T^*. The probability mechanism generating a X_T^* is denoted by $\hat{P}_T$.

The motivation of this scheme is as follows. The moving blocks bootstrap displays certain 'edge effects'. For example, the data points $X_{1,\cdot}$ and $X_{T,\cdot}$ of the original series are less likely to end up in a particular bootstrap sequence than the data points in the middle of the series. This is because they appear in one of the data blocks only, whereas a 'middle' data point appears in b of the blocks. By wrapping up the data in a circle, each data point appears in b of the blocks. Hence, the edge effects disappear.

Bootstrap B.4 (Stationary Bootstrap)
The stationary bootstrap of Politis and Romano (1994) is appropriate when the data sequence is a stationary time series. It generates a bootstrap sequence by concatenating blocks of data which are resampled from the original series. As does the circular blocks bootstrap, it wraps the original data into a circle to avoid edge effects. The difference between it and the two previous methods is that the block sizes are of random lengths. As before, a particular block $B_{t,b}$ is defined by its starting index t and by its block size b. The stationary bootstrap chooses random starting indices $t_1^*, t_2^*, t_2^*, \ldots$ i.i.d. from the discrete uniform distribution on the set $\{1, 2, \ldots, T\}$. Independently, it chooses random block sizes $b_1^*, b_2^*, \ldots$ i.i.d. from a geometric distribution with parameter $0 < q < 1/T$.[19] The thus selected blocks are concatenated as $\{B_{t_1^*,b_1^*}, B_{t_2^*,b_2^*}, \ldots\}$ until a sequence of length greater than or equal to T is generated. The sequence is then truncated at length T to obtain the bootstrap sequence $X_{1,\cdot}^*, X_{2,\cdot}^*, \ldots X_{T,\cdot}^*$. The corresponding $T \times (K+1)$ bootstrap data matrix is denoted by X_T^*. The probability mechanism generating a X_T^* is denoted by $\hat{P}_T$.

The motivation of this scheme is as follows. If the underlying data series is stationary, it might be desirable for the bootstrap series to be stationary as well. This not true, however, for the moving blocks bootstrap and the circular blocks bootstrap. The intuition is that

[19] So the average block size is given by $1/q$.

stationarity is 'lost' where the blocks of fixed size are pieced together. Politis and Romano (1994) show that if the blocks have random sizes from a geometric distribution, then the resulting bootstrap series is indeed stationary (conditional on the observed data). There is also some evidence to the fact the dependence on the model parameter q is not as pronounced as the dependence on the model parameter b in the two previous methods.

C Some Power Considerations

We assume a stylized and tractable model which allows us to make exact power calculations. In particular, we consider the limiting model of Scenarios 3.1 and 3.2.

Our simple setup specifies that $K = 2$ and that $w_k \sim N(\theta_k, \sigma_k^2)$, with σ_k known, for $k = 1, 2$.[20] (The subscript T in $w_{T,k}$ is suppressed for convenience.) In addition, the setup specifies a joint normal distribution for $(w_1, w_2)'$. Thus, the results in this section will hold approximately for quite general models where the limiting distribution is normal. As in the rest of the paper, an individual null hypothesis is of the form $H_{0,k}$: $\theta_k \leq 0$. We analyze power for the first step of our stepwise methods. The basic method is equivalent to the following scheme:

$$\text{Reject } \mathrm{H}_{0,k} \text{ if } w_k > c \quad \text{where } c \text{ satisfies:} \quad \mathrm{Prob}_{0,0}\{\max w_k > c\} = \alpha \tag{17}$$

Here, the notation $\mathrm{Prob}_{0,0}$ is shorthand for $\mathrm{Prob}_{\theta_1=0,\theta_2=0}$. The studentized method is equivalent to the following scheme:

$$\text{Reject } \mathrm{H}_{0,k} \text{ if } w_k/\sigma_k > d \quad \text{where } d \text{ satisfies:} \quad \mathrm{Prob}_{0,0}\{\max w_k/\sigma_k > d\} = \alpha \tag{18}$$

To get going, we assume that w_1 and w_2 are independent of each other. Let $\Phi(\cdot)$ the cumulative distribution function of the standard normal distribution. Then the constant c in (17) satisfies

$$\Phi\left(\frac{c}{\sigma_1}\right)\Phi\left(\frac{c}{\sigma_2}\right) = 1 - \alpha \tag{19}$$

and the constant d in (18) satisfies

$$\Phi^2(d) = 1 - \alpha \quad \text{so} \quad d = \Phi^{-1}(\sqrt{1-\alpha}) \tag{20}$$

The first notion of power we consider is the 'worst' power over the set $\{(\theta_1, \theta_2) : \theta_k > 0 \text{ for some } k\}$. A proper definition of this worst power is

$$\inf_{\epsilon>0} \inf_{\{(\theta_1,\theta_2):\max\theta_k \geq \epsilon\}} \text{Power at } (\theta_1, \theta_2) \tag{21}$$

Obviously, this infimum is the minimum of the two powers at $(-\infty, 0)$ and at $(0, -\infty)$.[21] The

[20] The argument generalizes easily for $K > 2$.

[21] The power at $(-\infty, 0)$ denotes the limit of the power at $(0, \theta_2)$ as θ_2 tends to $-\infty$; and analogously for the power at $(-\infty, 0)$.

basic method yields

$$\text{Prob}_{(-\infty,0)}\{\max w_k > c\} = \text{Prob}_{\theta_2=0}\{w_2 > c\} = 1 - \Phi\left(\frac{c}{\sigma_2}\right)$$

and

$$\text{Prob}_{(0,-\infty)}\{\max w_k > c\} = \text{Prob}_{\theta_1=0}\{w_1 > c\} = 1 - \Phi\left(\frac{c}{\sigma_1}\right)$$

The studentized method yields

$$\text{Prob}_{(-\infty,0)}\{\max w_k/\sigma_k > c\} = \text{Prob}_{(0,-\infty)}\{\max w_k/\sigma_k > c\} = 1 - \Phi(d) = 1 - \sqrt{1-\alpha}$$

To demonstrate that the worst power is smaller for the basic method, we are therefore left to show that

$$\min\left(1 - \Phi\left(\frac{c}{\sigma_k}\right)\right) \leq 1 - \sqrt{1-\alpha}$$

or, equivalently, that

$$\max \Phi\left(\frac{c}{\sigma_k}\right) \geq \sqrt{1-\alpha}$$

But this last inequality follows from (19), and it is strict unless $\sigma_1 = \sigma_2$. Note that even for the studentized method the worst power is equal to $1 - \sqrt{1-\alpha}$ and therefore strictly less than α. Hence, both the basic and the studentized method are biased, but the worst bias is smaller for the studentized method.

We continue to assume that w_1 and w_2 are independent. But now we consider the worst power against alternatives in the class $C_\delta = \{(\theta_1, \theta_2) : \theta_k = \sigma_k\delta \text{ for some } k\}$, where δ is a positive number. Obviously, the worst power is the minimum of the two powers at $(-\infty, \sigma_2\delta)$ and at $(\sigma_1\delta, -\infty)$. The basic method yields

$$\text{Prob}_{(-\infty,\sigma_2\delta)}\{\max w_k > c\} = \text{Prob}_{\theta_2=\sigma_2\delta}\{w_2 > c\} = 1 - \Phi\left(\frac{c - \sigma 2\delta}{\sigma_2}\right) = 1 - \Phi\left(\frac{c}{\sigma_2} - \delta\right)$$

and

$$\text{Prob}_{(\sigma_1\delta,-\infty)}\{\max w_k > c\} = \text{Prob}_{\theta_1=\sigma_1\delta}\{w_1 > c\} = 1 - \Phi\left(\frac{c - \sigma 1\delta}{\sigma_1}\right) = 1 - \Phi\left(\frac{c}{\sigma_1} - \delta\right)$$

The studentized method yields

$$\text{Prob}_{(-\infty,\sigma_2\delta)}\{\max w_k/\sigma_k > c\} = \text{Prob}_{(\sigma_1\delta,-\infty)}\{\max w_k/\sigma_k > c\} = 1 - \Phi(d - \delta)$$

To demonstrate that the worst power is smaller for the basic method, we are therefore left to show that

$$\max \Phi\left(\frac{c}{\sigma_k} - \delta\right) \geq \Phi(d - \delta) \tag{22}$$

This is true if $c/\sigma_k \geq d$ for some k. But assume the latter relation is false, that is, $c/\sigma_k < d$

for both k. This would imply that

$$\Phi\left(\frac{c}{\sigma_1}\right)\Phi\left(\frac{c}{\sigma_2}\right) < \Phi^2(d) = 1-\alpha$$

resulting in a violation of (19). Hence, inequality (22) holds; and it is strict unless $\sigma_1 = \sigma_2$. So, unless $\sigma_1 = \sigma_2$, the worst power over C_δ of the basic method is strictly smaller than the worst power of the studentized method.

Next, we consider correlated test statistics, with $\rho = Cor(w_1, w_2)$. We claim that also in this case the basic method has a smaller worst power (21) than the studentized method. As before, the infimum in (21) is the minimum of the two powers at $(-\infty, 0)$ and at $(0, -\infty)$. For the basic method, we get

$$\min\left(\text{Prob}_{\theta_1=0}\{w_1 > c\}, \text{Prob}_{\theta_2=0}\{w_2 > c\}\right) = \min\left(\text{Prob}\left\{\sigma_1 z_1 > c\right\}, \text{Prob}\left\{\sigma_2 z_2 > c\right\}\right)$$

where z_1 and z_2 are two standard normal variables with correlation ρ. For the studentized method, we get

$$\min\left(\text{Prob}_{\theta_1=0}\{w_1/\sigma_1 > d\}, \text{Prob}_{\theta_2=0}\{w_2/\sigma_2 > d\}\right) = \text{Prob}\{z_1 > d\}$$

We are therefore again left to show that $c/\sigma_k \geq d$ for some k. But assume the latter relation is false, that is, $c/\sigma_k < d$ for both k. Also assume without loss of generality that $\sigma_1 \leq \sigma_2$. Then

$$\begin{aligned}
\text{Prob}_{0,0}\{\max w_k > c\} &= \text{Prob}\{\max \sigma_k z_k > c\} \\
&= \text{Prob}\{\max(\sigma_k/\sigma_1) z_k > c/\sigma_1\} \\
&\geq \text{Prob}\{\max z_k > c/\sigma_1\} \\
&> \text{Prob}\{\max z_k > d\} \\
&= \text{Prob}_{0,0}\{\max w_k/\sigma_k > d\} \\
&= \alpha \quad \text{(by (18))}
\end{aligned}$$

resulting in a violation of (17). Hence, the infimum in (21) for the basic method is smaller than or equal to the infimum for the studentized method. And again, unless $\sigma_1 = \sigma_2$, the infimum for the basic method is strictly smaller.

We have just demonstrated that also in the case of correlated test statistics, $c/\sigma_k \geq d$ for some k. Hence, by the reasoning leading up to (22), also in the case of correlated test statistics, the worst power over C_δ of the basic method is smaller than the worst power of the studentized method. And it is strictly smaller unless $\sigma_1 = \sigma_2$.

References

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

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

Bao, Y., Lee, T.-H., and Saltoglu, B. (2001). Evaluating predictive performance of value-at-risk models in emerging markets: A reality check. Technical report, Department of Economics, University of California at Riverside. Available at `http://www.economics.ucr.edu/papers/2001papers.html`.

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

Beran, R. (1984). Bootstrap methods in statistics. *Jahresberichte des Deutschen Mathematischen Vereins*, 86:14–30.

Beran, R. (1988). Balanced simultaneous confidence sets. *Journal of the American Statistical Association*, 83:679–686.

Cowles, A. (1933). Can stock market forecasters forecast? *Econometrica*, 1:309–324.

Diebold, F. X. (2000). *Elements of Forecasting*. South-Western College Publishing, Cincinnati, Ohio, second edition.

Dudoit, S., Shaffer, J., and Boldrick, J. (2002). Multiple hypothesis testing in microarray experiments. Technical report, Division of Biostatistics, U.C. Berkeley. Available at `http://www.bepress.com/ucbbiostat/paper110/`.

Dunnett, C. W. and Tamhane, A. C. (1992). A step-up multiple test procedure. *Journal of the American Statistical Association*, 87:162–170.

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

Gonzalo, J. and Wolf, M. (2003). Subsampling inference in threshold autoregressive models. Technical report, Department of Economics, Universitat Pompeu Fabra. Available at `http://www.econ.upf.es/~wolf/preprints.html`.

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

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer, New York.

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

Jorion, P. (2000). *Value at Risk: The New Benchmark for Managing Financial Risk*. McGraw-Hill, second edition.

Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *Annals of Statistics*, 17:1217–1241.

Lahiri, S. N. (1992). Edgeworth correction by 'moving block' bootstrap for stationary and nonstationary data. In LePage, R. and Billard, L., editors, *Exploring the Limits of Bootstrap*, pages 183–214. John Wiley, New York.

Leamer, E. (1983). Let's take the con out of econometrics. *American Economic Review*, 73:31–43.

Lo, A. and MacKinley, C. (1990). Data snooping biases in tests of financial asset pricing models. *Review of Financial Studies*, 3:431–468.

Loh, W. Y. (1987). Calibrating confidence coefficients. *Journal of the American Statistical Association*, 82:155–162.

Lovell, M. C. (1983). Data mining. *Review of Economics and Statistics*, 65(1):1–12.

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

Politis, D. N. and Romano, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical Association*, 89:1303–1313.

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

Rodríguez-Poo, J., Delgado, M., and Wolf, M. (2001). Subsampling inference in cube root asymptotics with an application to Manski's maximum score estimator. *Economics Letters*, 73:241–250.

Romano, J. P. and Wolf, M. (2003). Improved nonparametric confidence intervals in time series regressions. Technical report, Department of Economics, Universitat Pompeu Fabra. Available at `http://www.econ.upf.es/~wolf/preprints.html`.

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

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

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

White, H. L. (2001). *Asymptotic Theory for Econometricians*. Academic Press, New York, second edition.

Table 1: Empirical FWEs and average number of false hypotheses rejected. The nominal level is $\alpha = 10\%$. Observations are i.i.d., the number of observations is $T = 100$, and the number of strategies is $K = 40$. The mean of the benchmark is 1; the strategy means are 1 or 1.4. The standard deviation of the benchmark is 1; half of the strategy standard deviations are 1, the other half is 2. The number of repetitions is 2,000 per scenario.

All strategy means = 1, cross correlation = 0				
Method	FWE (single)	FWE (multi)	Rejected (single)	Rejected (multi)
Basic	10.6	10.6	0.0	0.0
Stud	10.5	10.5	0.0	0.0
All strategy means = 1, cross correlation = 0.5				
Method	FWE (single)	FWE (multi)	Rejected (single)	Rejected (multi)
Basic	11.0	11.0	0.0	0.0
Stud	11.1	11.1	0.0	0.0
Six strategy means = 1.4, cross correlation = 0				
Method	FWE (single)	FWE (multi)	Rejected (single)	Rejected (multi)
Basic	9.4	9.9	1.1	1.2
Stud	10.1	10.7	2.2	2.2
Six strategy means = 1.4, cross correlation = 0.5				
Method	FWE (single)	FWE (multi)	Rejected (single)	Rejected (multi)
Basic	9.8	10.1	2.6	2.7
Stud	9.5	10.0	3.8	3.9
Twenty strategy means = 1.4, cross correlation = 0				
Method	FWE (single)	FWE (multi)	Rejected (single)	Rejected (multi)
Basic	6.0	7.7	3.8	4.2
Stud	6.8	8.3	7.4	7.8
Twenty strategy means = 1.4, cross correlation = 0.5				
Method	FWE (single)	FWE (multi)	Rejected (single)	Rejected (multi)
Basic	6.3	8.7	8.6	9.6
Stud	6.6	9.0	12.6	13.2
Forty strategy means = 1.4, cross correlation = 0				
Method	FWE (single)	FWE (multi)	Rejected (single)	Rejected (multi)
Basic	0.0	0.0	7.5	10.0
Stud	0.0	0.0	14.8	17.1
Forty strategy means = 1.4, cross correlation = 0.5				
Method	FWE (single)	FWE (multi)	Rejected (single)	Rejected (multi)
Basic	0.0	0.0	17.2	23.3
Stud	0.0	0.0	25.2	29.4

Table 2: Empirical FWEs and average number of false hypotheses rejected. The nominal level is $\alpha = 10\%$. Observations are i.i.d., the number of observations is $T = 100$, and the number of strategies is $K = 40$. The mean of the benchmark is 1; the strategy means that are bigger than 1 are equally spaced between 1 and 4. The standard deviation of the benchmark is 2; the standard deviation of a strategy is 2 times its mean. The number of repetitions is 2,000 per scenario.

All strategy means = 1, cross correlation = 0				
Method	FWE (single)	FWE (multi)	Rejected (single)	Rejected (multi)
Basic	11.0	11.0	0.0	0.0
Stud	10.3	10.3	0.0	0.0
All strategy means = 1, cross correlation = 0.5				
Method	FWE (single)	FWE (multi)	Rejected (single)	Rejected (multi)
Basic	11.1	11.1	0.0	0.0
Stud	11.1	11.1	0.0	0.0
Six strategy means greater than 1, cross correlation = 0				
Method	FWE (single)	FWE (multi)	Rejected (single)	Rejected (multi)
Basic	0.0	9.0	3.6	4.7
Stud	8.3	9.4	3.3	3.5
Six strategy means greater than 1, cross correlation = 0.5				
Method	FWE (single)	FWE (multi)	Rejected (single)	Rejected (multi)
Basic	0.0	9.3	4.1	5.3
Stud	8.5	10.0	4.3	4.4
Twenty strategy means greater than 1, cross correlation = 0				
Method	FWE (single)	FWE (multi)	Rejected (single)	Rejected (multi)
Basic	0.0	6.3	9.0	13.7
Stud	4.9	7.8	9.6	10.5
Twenty strategy means greater than 1, cross correlation = 0.5				
Method	FWE (single)	FWE (multi)	Rejected (single)	Rejected (multi)
Basic	0.0	8.4	11.0	16.3
Stud	5.5	8.8	13.1	13.9
Forty strategy means greater than 1, cross correlation = 0				
Method	FWE (single)	FWE (multi)	Rejected (single)	Rejected (multi)
Basic	0.0	0.0	15.3	24.5
Stud	0.0	0.0	18.1	21.5
Forty strategy means greater than 1, cross correlation = 0.5				
Method	FWE (single)	FWE (multi)	Rejected (single)	Rejected (multi)
Basic	0.0	0.0	19.7	31.5
Stud	0.0	0.0	25.4	29.0

Table 3: Empirical FWEs and average number of false hypotheses rejected. The nominal level is $\alpha = 10\%$. Observations are a multivariate time series, the number of observations is $T = 200$, and the number of strategies is $K = 40$. The mean of the benchmark is 1; the strategy means are 1 or 1.6. The standard deviation of the benchmark is 1; half of the strategy standard deviations are 1, the other half is 2. The number of repetitions is 2,000 per scenario.

All strategy means = 1, cross correlation = 0				
Method	FWE (single)	FWE (multi)	Rejected (single)	Rejected (multi)
Basic	15.7	15.7	0.0	0.0
Stud	5.8	5.8	0.0	0.0
All strategy means = 1, cross correlation = 0.5				
Method	FWE (single)	FWE (multi)	Rejected (single)	Rejected (multi)
Basic	16.3	16.3	0.0	0.0
Stud	5.2	5.2	0.0	0.0
Six strategy means = 1.3, cross correlation = 0				
Method	FWE (single)	FWE (multi)	Rejected (single)	Rejected (multi)
Basic	14.7	15.5	1.8	1.9
Stud	5.0	5.4	1.8	1.8
Six strategy means = 1.3, cross correlation = 0.5				
Method	FWE (single)	FWE (multi)	Rejected (single)	Rejected (multi)
Basic	15.6	16.8	3.7	3.8
Stud	6.8	7.5	3.3	3.4
Twenty strategy means = 1.3, cross correlation = 0				
Method	FWE (single)	FWE (multi)	Rejected (single)	Rejected (multi)
Basic	9.4	12.7	6.1	6.8
Stud	3.7	5.0	5.9	6.3
Twenty strategy means = 1.3, cross correlation = 0.5				
Method	FWE (single)	FWE (multi)	Rejected (single)	Rejected (multi)
Basic	11.3	16.0	12.3	13.3
Stud	4.3	6.8	11.2	12.0
Forty strategy means = 1.3, cross correlation = 0				
Method	FWE (single)	FWE (multi)	Rejected (single)	Rejected (multi)
Basic	0.0	0.0	12.5	16.8
Stud	0.0	0.0	11.6	14.3
Forty strategy means = 1.3, cross correlation = 0.5				
Method	FWE (single)	FWE (multi)	Rejected (single)	Rejected (multi)
Basic	0.0	0.0	24.3	30.2
Stud	0.0	0.0	22.3	27.9

Table 4: Empirical FWEs and average number of false hypotheses rejected. The nominal level is $\alpha = 10\%$. Observations are a multivariate time series the number of observations is $T = 200$, and the number of strategies is $K = 40$. The mean of the benchmark is 1; the strategy means that are bigger than 1 are equally spaced between 1 and 7. The standard deviation of the benchmark is 2; the standard deviation of a strategy is 2 times its mean. The number of repetitions is 2,000 per scenario.

All strategy means = 1, cross correlation = 0				
Method	FWE (single)	FWE (multi)	Rejected (single)	Rejected (multi)
Basic	15.1	15.1	0.0	0.0
Stud	7.4	7.4	0.0	0.0
All strategy means = 1, cross correlation = 0.5				
Method	FWE (single)	FWE (multi)	Rejected (single)	Rejected (multi)
Basic	17.9	17.9	0.0	0.0
Stud	7.4	7.4	0.0	0.0
Six strategy means greater than 1, cross correlation = 0				
Method	FWE (single)	FWE (multi)	Rejected (single)	Rejected (multi)
Basic	0.0	12.4	3.4	4.9
Stud	5.5	6.0	2.0	2.1
Six strategy means greater than 1, cross correlation = 0.5				
Method	FWE (single)	FWE (multi)	Rejected (single)	Rejected (multi)
Basic	0.0	13.0	3.8	5.4
Stud	4.5	5.3	2.5	2.6
Twenty strategy means greater than 1, cross correlation = 0				
Method	FWE (single)	FWE (multi)	Rejected (single)	Rejected (multi)
Basic	0.0	6.1	8.0	13.3
Stud	2.7	3.5	5.2	5.9
Twenty strategy means greater than 1, cross correlation = 0.5				
Method	FWE (single)	FWE (multi)	Rejected (single)	Rejected (multi)
Basic	0.0	12.0	9.5	15.8
Stud	2.3	4.1	7.5	8.5
Forty strategy means greater than 1, cross correlation = 0				
Method	FWE (single)	FWE (multi)	Rejected (single)	Rejected (multi)
Basic	0.0	0.0	13.0	22.1
Stud	0.0	0.0	9.4	11.5
Forty strategy means greater than 1, cross correlation = 0.5				
Method	FWE (single)	FWE (multi)	Rejected (single)	Rejected (multi)
Basic	0.0	0.0	16.5	29.4
Stud	0.0	0.0	14.9	19.3

Table 5: The ten largest basic test statistics $\hat{\alpha}_{T,k}$ and the corresponding stocks in our empirical application. The return unit is 1 percent.

$\hat{\alpha}_{T,k}$	Stock
4.03	AOL Time Warner
3.80	Qualcomm
3.44	Dell Computer
2.67	Oracle
2.65	Clear Chl. Comms.
2.24	Applied Mats.
2.12	Cisco Systems
2.06	Lowe's Cos.
2.02	Kohls
1.87	Forest Labs.

Table 6: The ten largest studentized test statistics $\hat{\alpha}_{T,k}/\hat{\sigma}_{T,k}$ and the corresponding stocks in our empirical application. The return unit is 1 percent.

$\hat{\alpha}_{T,k}/\hat{\sigma}_{T,k}$	Stock
3.98	Kohls
3.08	Citigroup
2.96	Clear Chl. Comms.
2.87	AOL Time Warner
2.83	MBNA Corp.
2.77	Fifth Third Bancorp.
2.59	Wells Fargo & Co
2.52	Anheuser-Busch
2.51	Dell Computer
2.51	Amer.Intl.Gp.