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

October 2003; this revision February 2005

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. This paper suggests a stepwise multiple testing procedure which asymptotically controls the familywise error rate at a desired level. Compared to related single-step methods, the 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, the method implicitly captures the joint dependence structure of the test statistics, which results in increased ability to detect alternative hypotheses. We prove asymptotic control of the familywise error rate under minimal assumptions. The methodology is presented 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, stepwise method.

JEL CLASSIFICATION NOS: C12, C14, C52.

[*] Research supported by National Science Foundation grant DMS 010392.

[†] Research supported by the Spanish Ministry of Science and Technology and FEDER, grant BMF2003-03324, and by the Barcelona Economics Program of CREA.

We thank the co-editor and three anonymous referees for helpful comments that have led to an improved presentation of the paper. We have benefited from discussions with Peter Hansen, Olivier Ledoit, and seminar participants at the European Central Bank, Hong Kong University of Science and Technology, UCLA, Universidad de Zaragoza, Universität Mannheim, Universität Zürich, and Universitat Pompeu Fabra. All remaining errors remain ours.

"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. For example, in the problem of determining whether a certain trading strategy has a positive CAPM alpha, the benchmark is fixed at zero.[1] On the other hand, in the problem of determining whether a trading strategy beats a specific investment, such as a stock index, the benchmark is usually random. 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), among others. However, in the context of comparing several strategies to a benchmark, little has been suggested 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. The concept to account for data mining is the (asymptotic) control of the familywise error rate (FWE). The FWE is defined as the probability of incorrectly identifying at least one strategy as superior.[2]

White (2000) coins his technique the Bootstrap Reality Check (BRC). Often one would like to identify further outperforming strategies, 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 but 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. Further outperforming strategies can be detected in subsequent steps, while maintaining control of the FWE. So the method we propose is more powerful than the BRC.

To motivate our main contribution, consider the following three exemplary persons who would benefit from the more powerful stepwise method. First, a trader who backtests several quantitative trading ideas on historical data and wants to know how many of these are worth launching for real; then the benchmark is whichever benchmark the trader is subjected to. Second, a CEO of a multi-strategy mutual fund family who has to choose which individual portfolio managers to promote by comparing them with the market index. Third, the manager of a fund of hedge funds who has to choose which individual hedge fund he wants to invest his clients' capital in, by benchmarking them against the risk-free rate.

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

[2] This means at least one strategy that in truth is as good as or inferior to the benchmark will get identified as superior to the benchmark by the statistical method.

The challenge of constructing an 'optimal' forecast provides another motivation. Imagine several different forecasting strategies are available to forecast a quantity of interest. As described in Timmermann (2006, Chapter 6): (i) choosing the (lone) strategy with the best track record is often a bad idea; (ii) simple forecasting schemes, such as equal-weighting various strategies, are hard to beat; and (iii) trimming off the worst strategies is often required. Accordingly, a sensible approach would be to identify (hopefully) all strategies that *underperfom* a simple-minded benchmark[3] and to then use the equal-weighted average of the remaining strategies for out-of-sample forecasts. (Obviously, methods that can identify outperforming strategies can also be modified to identify underperforming strategies.[4])

As a second contribution, we propose the use of studentization to improve level 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.

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 briefly discusses alternatives to controlling the FWE. Section 7 proposes how to choose the bootstrap block size in the context of time series data. Section 8 sheds some light on finite-sample performance via a simulation study. Section 9 provides an application to real data. Section 10 concludes. An appendix contains proofs of mathematical results, an overview of the most important bootstrap methods, some power considerations for studentization, and a brief discussion of multiple testing versus joint testing.

2 Notation and Problem Formulation

2.1 Notation and Some Examples

One observes a data matrix $x_{t,s}$ with $1 \leq t \leq T$ and $1 \leq s \leq S+1$. 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 s corresponds to strategies, and there is a fixed number S of them. The final column, $S+1$, is reserved for the benchmark. 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 $T \times (S+1)$ data matrix; $X_{t,\cdot}^{(T)}$ is the $(S+1) \times 1$ vector that corresponds to the tth row of X_T; and $X_{\cdot,s}^{(T)}$ is the $T \times 1$ vector that corresponds to the sth column of X_T.

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

[3] For example, when forecasting inflation the simple-minded benchmark might be the current inflation.

[4] The ability to detect as many underperforming strategies as possible would also be useful to a CEO of a multi-strategy mutual fund company who has to choose which individual portfolio managers to fire.

collects the individual parameters of interest θ_s; the $S \times 1$ vector W_T collects the individual basic test statistics $w_{T,s}$; and the $S \times 1$ vector Z_T collects the individual studentized test statistics $z_{T,s}$.

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) Historical returns of investment strategy s, say a particular mutual fund or a particular trading strategy, are recorded in $X^{(T)}_{\cdot,s}$. Historical returns of a benchmark, say a stock index or a buy-and-hold strategy, are recorded in $X^{(T)}_{\cdot,S+1}$. Depending on preference, these can be 'real' returns or log returns; also, returns may be recorded in excess of the risk free rate if desired. Let μ_s denote the population mean of the return for strategy s. Based on an absolute criterion, strategy s beats the benchmark if $\mu_s > \mu_{S+1}$. Therefore, we define $\theta_s = \mu_s - \mu_{S+1}$. Using the notation

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

a natural basic test statistic is

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

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

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

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

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_s again denoting the mean of the return of strategy s and with σ_s denoting its standard deviation, the corresponding Sharpe ratio is defined as $SR_s = \mu_s/\sigma_s$.[5] An investment strategy is now said to outperform the benchmark if its Sharpe ratio is higher than the one of the benchmark. Therefore, we define $\theta_s = SR_s - SR_{S+1}$. Let

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

Then a natural basic test statistic is

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

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

Example 2.3 (CAPM alpha) Historical returns of investment strategy s, in excess of the risk-free rate, are recorded in $X^{(T)}_{\cdot,s}$. Historical returns of a market proxy, in excess of the risk-free rate, are

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

recorded in $X^{(T)}_{\cdot,S+1}$. For each strategy s, a simple time series regression

$$x_{t,s} = \alpha_s + \beta_s x_{t,S+1} + \epsilon_{t,s} \tag{4}$$

is estimated by ordinary least squares (OLS). If the CAPM holds, all intercepts α_s are equal to zero.[6] So the parameter of interest here is $\theta_s = \alpha_s$. Since the CAPM may be violated in practice, a financial advisor might want to identify investment strategies which have a positive α_s. Hence, an obvious basic test statistic would be

$$w_{T,s} = \hat{\alpha}_{T,s} \tag{5}$$

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

$$z_{T,s} = \frac{\hat{\alpha}_{T,s}}{\hat{\sigma}_{T,s}} \tag{6}$$

2.2 Problem Formulation

It is assumed that depending on the underlying probability mechanism P, the parameter $\theta_s = \theta_s(P)$ either satisfies it is ≤ 0 or not. So, the parameter θ_s can really be viewed as a functional of the unknown P. For a given strategy s, consider the individual testing problem

$$H_s \colon \theta_s \leq 0 \quad \text{vs.} \quad H'_s \colon \theta_s > 0$$

A multiple testing method yields a decision concerning each individual testing problem by either rejecting H_s or not.[7] In an ideal world, one would reject H_s exactly for those strategies for which $\theta_s > 0$. 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 control of 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 $I_0 = I_0(P) \subset \{1, \ldots, S\}$ denote the indices of the set of true hypotheses; that is, $s \in I_0$ if and only if $\theta_s \leq 0$. The FWE is the probability under P that any H_s with $s \in I_0$ is rejected:[8]

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

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_P no bigger than α, at least asymptotically. In particular, this constraint must hold for all P, and therefore regardless of which hypotheses are true and which are false. That is, we demand *strong* control of the FWE. A method that only controls the FWE for a probability mechanism P such that all S null hypotheses are true is said to have *weak* control of the FWE. As remarked by Dudoit et al. (2003), 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 size T, $\text{FWE}_P \leq \alpha$, for any P. A multiple testing method is said to asymptotically control the FWE at

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

[7] This is related to, but distinct from, the problem of joint testing; see Appendix D for a brief discussion.

[8] To show its dependence on P, we may write FWE = FWE_P.

level α, if $\limsup_T \mathrm{FWE}_P \leq \alpha$, for any P. Methods that control the FWE in finite samples 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 control of 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 testing a single hypothesis. Some possible notions of power are:

- 'Minimal' power: the probability of rejecting at least one false null hypothesis. Since our goal is to reject as many false null hypotheses as possible, rather than just rejecting at least one of them, this notion is not suitable for our purposes. Indeed, if we adopted this notion, then the stepwise method we will present would not improve upon the BRC of White (2000).
- 'Global' power: the probability of rejecting all false null hypotheses. Arguably, this notion is too strict for our purposes. While we aim to reject as many false null hypotheses as possible, we do not necessarily consider it a failure to miss a single one of them.
- 'Average' power: the average of the individual probabilities of rejecting each false null hypothesis. This is equivalent to the expected number of false null hypotheses that will be rejected. Therefore, we consider it the most appropriate notion for our purposes.
- The expected proportion of false null hypotheses that will be rejected.
- The probability of rejecting at least $\gamma\,100\%$ of the false null hypotheses, where $\gamma \in (0,1]$ is a user-specified number.

For the sake of argument, when we use statements like "more powerful" in the remainder of the paper we mean in the sense of better average power. But these statements would also apply to any other reasonable notion of power that increases in the number of false hypotheses rejected. (Only with the notion of minimal power, which is not suitable for our purposes, there is no difference between our stepwise method and the BRC.)

A special case in comparing the power of two multiple testing methods, say methods 1 and 2, arises in the following scenario: by design, method 1 rejects all hypotheses rejected by method 2 and possibly some further ones. It then trivially follows that method 1 is more powerful than method 2.

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_s, one computes an individual p-value $\hat{p}_{T,s}$. It is assumed that if H_s is true, the distribution of $\hat{p}_{T,s}$ is Uniform (0,1), at least asymptotically.[9] The Bonferroni method at level α rejects H_s if $\hat{p}_{T,s} < \alpha/S$. If the null distribution of each $\hat{p}_{T,s}$ 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, which can result in low power.

[9] Actually, the following weaker assumption would be sufficient: If H_s is true, then $\mathrm{Prob}_P(\hat{p}_{T,s} \leq x) \leq x$, at least asymptotically.

Actually, there exists a simple method which (asymptotically) controls the FWE at level α but is more powerful than the Bonferroni method. This *stepwise* procedure is due to Holm (1979) and works as follows. The individual p-values are ordered from smallest to largest: $\hat{p}_{T,(1)} \leq \hat{p}_{T,(2)} \leq \cdots \leq \hat{p}_{T,(S)}$ with their corresponding null hypotheses labeled accordingly: $H_{(1)}$,$H_{(2)}$, ..., $H_{(S)}$. Then $H_{(s)}$ is rejected at level α if $\hat{p}_{T,(j)} < \alpha/(S-j+1)$ for all $j = 1, \ldots, s$. In comparison with the Bonferroni method, the criterion for the smallest p-value is equally strict, α/S, but it becomes less and less strict for larger p-values. This explains the improvement in power. Still, the Holm method can be quite 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_s if $\hat{p}_{T,s} < \alpha$. This (asymptotically) controls the FWE but obviously is 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 $\max_{1 \leq s \leq S}(w_{T,s} - \theta_s)$, implicitly accounting for the dependence structure of the individual test statistics. Let s_{max} denote the index of strategy with the largest statistic $w_{T,s}$. The BRC decides whether or not to reject $H_{s_{max}}$ at level α, asymptotically controlling the FWE. It therefore addresses the question whether the strategy that appears 'best' in the observed data really beats the benchmark.[10] However, it does not attempt to identify as many outperforming strategies as possible. The method we present in the next section does just that. In addition, we argue that by studentizing the test statistics, in situations where studentization is feasible, one can hope to improve size and certain power properties in finite samples. This represents a second enhancement of White's (2000) approach.

Hansen (2004) offers some improvements over the BRC; in addition, see Hansen (2003). First, his method reduces the influence of 'irrelevant' strategies, meaning strategies that 'significantly' underperform the benchmark. Second, he also proposes the use of studentized test statistics $z_{T,s}$ instead of basic test statistics $w_{T,s}$. However, like the BRC, the method of Hansen (2004) 'only' addresses the question whether the strategy that appears 'best' in the observed data really beats the benchmark.

3 Stepwise Multiple Testing Method

Our goal is to identify as many strategies as possible for which $\theta_s > 0$. We do this by considering individual hypothesis tests

$$H_s \colon \theta_s \leq 0 \quad \text{vs.} \quad H_s' \colon \theta_s > 0$$

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.

[10] Equivalently, it addresses the question whether there are any strategies at all that beat the benchmark.

We describe our method in the context of using basic test statistics $w_{T,s}$. 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_S to the smallest one, so that $w_{T,r_1} \geq w_{T,r_2} \geq \ldots \geq w_{T,r_S}$. Then the individual decisions are taken in a *stepwise* manner.[11] In the first step, we construct a rectangular joint confidence region for the vector $(\theta_{r_1}, \ldots, \theta_{r_S})'$ with nominal joint coverage probability $1 - \alpha$. The confidence region is of the form

$$[w_{T,r_1} - c_1, \infty) \times \ldots \times [w_{T,r_S} - c_1, \infty) \tag{7}$$

where the common value c_1 is chosen in such as way as to ensure the proper joint (asymptotic) 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 $[w_{T,r_s} - c_1, \infty)$ does not contain zero, the corresponding null hypothesis H_{r_s} is rejected.

If the above joint confidence region (7) 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 $[w_{T,r_1} - c_1, \infty)$ only, resulting in a decision on H_{r_1} alone. However, his method can be easily modified to be equivalent to our first step.[12] The critical advantage of our method is that we do not stop after the first step, unless no hypothesis is rejected. Suppose we reject the first R_1 relabeled hypotheses in this step one. Then $S - R_1$ hypotheses remain, corresponding to the labels $r_{R_1+1}, \ldots, r_S$. In the second step, we construct a rectangular joint confidence region for the vector $(\theta_{r_{R_1+1}}, \ldots, \theta_{r_S})'$ with, again, nominal joint coverage probability $1 - \alpha$. The new confidence region is of the form

$$[w_{T,r_{R_1+1}} - c_2, \infty) \times \ldots \times [w_{T,r_S} - c_2, \infty) \tag{8}$$

where the common constant c_2 is chosen in such a way as to ensure the proper joint (asymptotic) coverage probability. Again, if a particular individual confidence interval $[w_{T,r_s} - c_2, \infty)$ does not contain zero, the corresponding null hypothesis H_{r_s} is rejected. This stepwise process is then repeated until no further hypotheses are rejected. By continuing after the first step, more false hypotheses can be rejected.[13] The stepwise procedure is therefore more powerful than the single-step method. Nevertheless, the stepwise procedure still asymptotically controls the FWE at level α; the proof is in Theorem 3.1. Hence, our stepwise multiple testing (StepM) procedure improves upon the single-step BRC of White (2000) very much in the way that the stepwise Holm method improves upon the single-step Bonferroni method.

Remark 3.1 By design, the StepM procedure rejects all hypotheses that the BRC rejects and potentially some more. One consequence is that often more false null hypotheses are rejected. Clearly, this is an advantage, resulting in improved power. However, another consequence is that more true null hypotheses can be rejected as well. Even so, the main point here is that the resulting

[11] 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).

[12] Since the method of White (2000) amounts to computing the constant c_1, it has the potential to identify further outperforming strategies, apart from the one that appears best in sample. Namely, the method rejects all null hypotheses H_{r_s} for which $[w_{T,r_s} - c_1, \infty)$ does not contain 0.

[13] The reason is that $c_1 > c_2 > c_3 > \ldots$ typically.

procedure can greatly increase the chance of rejecting false hypotheses while still controlling the FWE at a prescribed (small) level. Thus, our improvement is in the same sense in which the Holm procedure is an improvement over the Bonferroni procedure, which is well-accepted and documented in the literature. The BRC can be viewed as a procedure to improve upon Bonferroni by using the bootstrap to get a less conservative critical value. In the same way, our procedure improves upon the Holm procedure by using the bootstrap to (implicitly) estimate the dependence structure of the test statistics to achieve greater power. Table 1 summarizes the characteristics of the various procedures. While all of them (asymptotically) control the FWE, power increases (i) in each column going down and (ii) in each row going from left to right.

Table 1: Characteristics of various procedures that asymptotically control the FWE.

	Handles Worst-Case Dependence	Accounts for True Dependence Structure
Single-Step	Bonferroni	White (2000), Hansen (2004)
Stepwise	Holm (1979)	Our stepwise procedure

How should the value c_1 in the joint confidence region construction (7) be chosen? Ideally, one would take the $1-\alpha$ quantile of the sampling distribution of $\max_{1 \le s \le S}(w_{T,r_s} - \theta_{r_s})$. This is the sampling distribution of the maximum of the individual differences "test statistic minus true parameter". Concretely, the corresponding quantile is defined as

$$c_1 \equiv c_1(1-\alpha, P) = \inf\{x : \text{Prob}_P\{\max_{1 \le s \le S}(w_{T,r_s} - \theta_{r_s}) \le x\} \ge 1-\alpha\}$$

The ideal choice of c_2, c_3, and so on in the subsequent steps would be analogous. For example, the ideal c_2 for (8) would be the $1-\alpha$ quantile of the sampling distribution of $\max_{R_1+1 \le s \le S}(w_{T,r_s} - \theta_{r_s})$ defined as

$$c_2 \equiv c_2(1-\alpha, P) = \inf\{x : \text{Prob}_P\{\max_{R_1+1 \le s \le S}(w_{T,r_s} - \theta_{r_s}) \le x\} \ge 1-\alpha\}$$

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 $j \ge 1$, let R_{j-1} denote the number of hypotheses rejected in the first $j-1$ steps (with $R_0 \equiv 0$) and define

$$\hat{c}_j \equiv c_j(1-\alpha, \hat{P}_T) = \inf\{x : \text{Prob}_{\hat{P}_T}\{\max_{\tilde{R}_{j-1}+1 \le s \le S}(w^*_{T,r_s} - \theta^*_{T,r_s}) \le x\} \ge 1-\alpha\} \tag{9}$$

Here the notation w^*_{T,r_s} 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_s} makes clear that the true parameters are those of $\hat{P}_T$ rather than those of P, that is, $\theta^*_T = \theta(\hat{P}_T)$.[14] 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.

[14] 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.

Algorithm 3.1 (Basic StepM Method)

1. Relabel the strategies in descending order of the test statistics $w_{T,s}$: strategy r_1 corresponds to the largest test statistic and strategy r_S to the smallest one.
2. Set $j = 1$ and $R_0 = 0$.
3. For $R_{j-1} + 1 \leq s \leq S$, if $0 \notin [w_{T,r_s} - \hat{c}_j, \infty)$, reject the null hypothesis H_{r_s}.
4. (a) If no (further) null hypotheses are rejected, stop.

   (b) Otherwise, denote by R_j the total number of hypotheses rejected so far and, afterwards, let $j = j + 1$. 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 $\sqrt{T}(W_T^* - \theta_T^*)$.

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 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.*

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

(i) If $\theta_s > 0$, then the null hypothesis H_s will be rejected with probability tending to one, as $T \to \infty$.

(ii) The method asymptotically controls the FWE at level α; that is, $\lim_T FWE_P \leq \alpha$.

(iii) Assume in addition that the limiting distribution $J(P)$ in Assumption 3.1 has a density that is positive everywhere.[15] Then the limiting probability in (ii) is equal to α iff there exists at least one θ_s with $\theta_s = 0$ and no θ_s with $\theta_s < 0$.

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).[16] Furthermore, in the context of this paper, our result is easier to apply in practice for two reasons. First, it is based on the S individual test statistics. In contrast, Algorithm 2.8 of Westfall and Young (1993) is based on the S individual p-values, which would require an extra round of computation. Second, the quantiles $\hat{c}_j$ 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).

[15] This additional assumption is very weak and holds, for example, in the case of a limiting multivariate normal distribution with nonsingular covariance matrix.

[16] For instance, this condition is violated, even asymptotically, when carrying out individual tests on the correlations of a joint correlation matrix, but our methods apply.

Remark 3.2 Part (iii) of the Theorem shows that it is not possible to have a limiting FWE exactly equal to α in general. Indeed, this can only be achieved if all the nonpositive θ_s values are exactly equal to zero. If there exists at least one negative θ_s value, then the FWE is asymptotically bounded away from α. (On the other hand, if all the θ_s values are positive than the limiting FWE is trivially equal to zero.) In contrast, a similar result[17] for BRC of White (2000) establishes that its limiting FWE is equal to α iff *all* the θ_s values are equal to 0. The impossibility of achieving a limiting FWE exactly equal to α in general has nothing to do with the problem of multiple testing and/or the application of the bootstrap. Instead, it occurs generally even when testing a single composite null hypothesis for which the rejection probability depends on the exact value of the parameter in the null hypothesis parameter space. Take the simple example of $X \sim N(\theta, 1)$ and testing $H : \theta \leq 0$ vs. $H' : \theta > 0$. The universally most powerful (UMP) test rejects H at nominal level $\alpha = 0.05$ iff $X > 1.645$. But the actual rejection probability, under the null, is strictly less than α unless θ lies on the boundary, that is, $\theta = 0$. For example, if $\theta = -0.5$, then the actual rejection probability equals 0.016. Finally, when the individual tests are two-sided, namely $H_s : \theta_s = 0$ vs. $H'_s : \theta_s \neq 0$, then the limiting FWE of our stepwise method is indeed equal to α, unless all θ_s are nonzero (in which case it is not possible to incorrectly reject a null hypothesis). On the other hand, the limiting FWE of the BRC is again strictly less than α, unless all θ_s are equal to zero.

Remark 3.3 Our framework assumes that the probability mechanism P is fixed. In particular, the parameters $\theta_s > 0$ are fixed. Asymptotically, according to Theorem 3.1 (i), if $\theta_s > 0$, then H_s will be rejected with probability tending to one. Alternatively, one can also study the behavior of multiple testing methods under contiguous (or local) alternatives $\theta_{T,s} \to 0$, so that not all false hypotheses are rejected with probability tending to one. For example, one can consider sequences $\theta_{T,s} = h_s/\sqrt{T}$, with $h_s > 0$ fixed. However, evidently, if alternative hypotheses are in some sense closer to their respective null hypothesis, then the methods will typically reject even fewer hypotheses. In other words, the probability of rejecting any set of hypotheses is smaller (asymptotically), whether they are true or false. And so the limiting probability of rejecting any true hypotheses (i.e., the FWE) under a sequence of contiguous alternatives will be bounded above by α, thus part (ii) of the Theorem continues to hold. On the other hand, part (iii) no longer holds. The existence of local alternatives generally causes the limiting FWE to be bounded away from α.

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 $X_{t,\cdot}^{(T)}$, $1 \leq t \leq T$. In the 'smooth function' model of Hall (1992), the test statistic $w_{T,s}$ is a smooth function of certain sample moments of $X_{\cdot,s}^{(T)}$ and $X_{\cdot,S+1}^{(T)}$, and the parameter θ_s is the same function applied to the corresponding population moments. Examples that fit into this framework are given by (1), (3), and (5). 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.[18]

[17] The corresponding proof is analogous to the proof of part (iii) of the Theorem 3.1 and left to the reader.

[18] 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.

Scenario 3.2 (Smooth Function Model with Time Series Data) Consider the case of strictly stationary time series data $X_{t,\cdot}^{(T)}$, $1 \leq t \leq T$. The smooth function model is defined as before and examples (1), (3), and (5) apply. Under moment and mixing conditions on the underlying process, $\sqrt{T}(W_T - \theta)$ converges in distribution to a multivariate normal distribution with mean zero and some covariance matrix Ω; e.g., see White (2001). In the time series case, the limiting covariance matrix Ω not only depends on the marginal distribution of $X_{t,\cdot}^{(T)}$ but it also depends on the underlying dependence structure over time. 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; also see Gonçalves and de Jong (2003).

Test statistics not covered immediately by the smooth function model can often be accommodated with some additional effort. In many cases where the bootstrap is known to fail[19], the subsampling method can 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; see Politis et al. (1999).

Scenario 3.3 (Strategies that Depend on Estimated Parameters) Consider the case where strategy s depends on a parameter vector β_s. In case β_s is unknown, it is estimated from the data. Denote the corresponding estimator by $\hat{\beta}_{T,s}$. Denote the value of the test statistic for strategy s, as a function of the estimated parameter vector $\hat{\beta}_{T,s}$, by $w_{T,s}(\hat{\beta}_{T,s})$. Further, let $W_T(\hat{\beta}_T)$ denote the $S \times 1$ 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_s 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_s parameters, at least for first-order asymptotic consistency. Details are in White (2000).

For concreteness, we now describe how to compute the $\hat{c}_j$ in Algorithm 3.1 via the bootstrap.[20] In what follows, pseudo data matrices X_T^* are generated by a generic bootstrap mechanism, denoted by $\hat{P}_T$. The true parameter vector corresponding to $\hat{P}_T$ is denoted by $\theta_T^* = \theta(\hat{P}_T)$. The specific choice of bootstrap method depends 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 $\hat{c}_j$ via the Bootstrap)

1. The labels $r_1, \ldots, r_S$ and the numerical values of $R_0, R_1 \ldots$ are given in Algorithm 3.1.

[19] For example, this can happen when the true parameter lies on the boundary of the parameter space; see Shao and Tu (1995) and Andrews (2000).

[20] Of course, one could use alternative methods to compute the $\hat{c}_j$, such as based on a limiting normal distribution in conjunction with a consistently estimated covariance matrix.

2. Generate M bootstrap data matrices $X_T^{*,1}, \ldots, X_T^{*,M}$. (One should use $M \geq 1,000$ in practice.)

3. From each bootstrap data matrix $X_T^{*,m}$, $1 \leq m \leq M$, compute the individual test statistics $w_{T,1}^{*,m}, \ldots, w_{T,S}^{*,m}$.

4. (a) For $1 \leq m \leq M$, compute $max_{T,j}^{*,m} = \max_{R_{j-1}+1 \leq s \leq S}(w_{T,r_s}^{*,m} - \theta_{T,r_s}^*)$.

   (b) Compute $\hat{c}_j$ as the $1-\alpha$ empirical quantile of the M values $max_{T,j}^{*,1}, \ldots, max_{T,j}^{*,M}$.

Remark 3.4 For convenience, one can typically use w_{T,r_s} in place of θ_{T,r_s}^* in step 4(a) of the algorithm. Indeed, the two are the same under the following conditions: (1) $w_{T,s}$ is a linear statistic; (2) $\theta_s = \mathrm{E}(w_{T,s})$; 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_s} and θ_{T,r_s}^* 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_s} for θ_{T,r_s}^* does in general not affect 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_s} being the sample mean. He shows that centering by θ_{T,r_s}^* provides second-order refinements but it is not necessary for first-order consistency.

Remark 3.5 A main point of our paper is that, to avoid making parametric assumptions, we use the bootstrap to approximate critical values. However, for testing one-sided hypotheses in some parametric models, the stepwise procedures we propose enjoy certain optimality properties; see Lehmann et al. (2005). (Of course, in such cases the critical values are derived from the underlying parametric model then.)

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 $z_{T,s} = w_{T,s}/\hat{\sigma}_{T,s}$, where $\hat{\sigma}_{T,s}$ estimates the standard deviation of $w_{T,s}$. Typically, one would choose $\hat{\sigma}_{T,s}$ in such a way that the asymptotic variance of $z_{T,s}$ is equal to one, but this is actually not required for Theorem 4.1 to hold. The 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 a data matrix generated from $\hat{P}_T$, let $w_{T,s}^*$ denote a basic test statistic computed from X_T^*, and let $\hat{\sigma}_{T,s}^*$ denote the estimated standard deviation of $w_{T,s}^*$ computed from X_T^*.[21] We need an analogue of the quantile (9) for the studentized method. It is given by

$$\hat{d}_j \equiv d_j(1-\alpha, \hat{P}_T) = \inf\{x : \mathrm{Prob}_{\hat{P}_T}\{\max_{R_{j-1}+1 \leq s \leq S}(w_{T,r_s}^* - \theta_{T,r_s}^*)/\hat{\sigma}_{T,r_s}^* \leq x\} \geq 1-\alpha\} \tag{10}$$

[21] Since $\hat{P}_T$ is completely specified, one actually knows the true standard deviation of $w_{T,s}^*$. However, the bootstrap mimics the real world, where standard deviation of $w_{T,s}$ is unknown, by estimating this standard deviation from the data. Hansen (2004) uses $\hat{\sigma}_{T,s}^* = \hat{\sigma}_{T,s}$. While this results in first-order consistency, it is preferable to compute $\hat{\sigma}_{T,s}^*$ from the bootstrap data; see Hall (1992).

Algorithm 4.1 (Studentized StepM Method)

1. Relabel the strategies in descending order of the test statistics $z_{T,s}$: strategy r_1 corresponds to the largest test statistic and strategy r_S to the smallest one.
2. Set $j = 1$ and $R_0 = 0$.
3. For $R_{j-1} + 1 \leq s \leq S$, if $0 \notin [w_{T,r_s} - \hat{\sigma}_{T,r_s}\hat{d}_j, \infty)$, reject the null hypothesis H_{r_s}.
4. (a) If no (further) null hypotheses are rejected, stop.
   (b) Otherwise, denote by R_j the total number of hypotheses rejected so far and, afterwards, let $j = j + 1$. Then return to step 3.

Assumption 4.1 *In addition to Assumption 3.1, assume the following condition. For each* $1 \leq s \leq S$, *both* $\sqrt{T}\hat{\sigma}_{T,s}$ *and* $\sqrt{T}\hat{\sigma}^*_{T,s}$ *converge to a (common) positive constant* σ_s *in probability.*

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

(i) If $\theta_s > 0$, *then the null hypothesis* H_s *will be rejected with probability tending to one, as* $T \to \infty$.

(ii) The method asymptotically controls the FWE at level α; *that is,* $\lim_T FWE_P \leq \alpha$.

(iii) Assume in addition that the limiting distribution $J(P)$ *in Assumption 3.1 has a density that is positive everywhere. Then the limiting probability in (ii) is equal to* α *iff there exists at least one* θ_s *with* $\theta_s = 0$ *and no* θ_s *with* $\theta_s < 0$.

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 (6).

For concreteness, we now describe how to compute the $\hat{d}_j$ in Algorithm 4.1 via the bootstrap. Again, pseudo data matrices X^*_T are generated by a generic bootstrap method.

Algorithm 4.2 (Computation of the $\hat{d}_j$ via the Bootstrap)

1. The labels $r_1, \ldots, r_S$ and the numerical values of $R_0, R_1 \ldots$ are given in Algorithm 4.1.
2. Generate M bootstrap data matrices $X^{*,1}_T, \ldots, X^{*,M}_T$. (One should use $M \geq 1{,}000$ in practice.)
3. From each bootstrap data matrix $X^{*,m}_T$, $1 \leq m \leq M$, compute the individual test statistics $w^{*,m}_{T,1}, \ldots, w^{*,m}_{T,S}$. Also, compute the corresponding standard errors $\hat{\sigma}^{*,m}_{T,1}, \ldots, \hat{\sigma}^{*,m}_{T,S}$.

4. (a) For $1 \leq m \leq M$, compute $max_{T,j}^{*,m} = \max_{R_{j-1}+1 \leq s \leq S} (w_{T,r_s}^{*,m} - \theta_{T,r_s}^{*}) / \hat{\sigma}_{T,r_s}^{*,m}$.

   (b) Compute $\hat{d}_j$ as the $1 - \alpha$ empirical quantile of the M values $max_{T,j}^{*,1}, \ldots, max_{T,j}^{*,M}$.

Remark 3.4 applies here as well.

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

$$\hat{\sigma}_{T,s} = \sqrt{\frac{\sum_{t=1}^{T} (x_{t,s} - x_{t,S+1} - \bar{x}_{T,s} + \bar{x}_{T,S+1})^2}{T-1}} \tag{11}$$

In the Efron bootstrap world, the value of $\hat{\sigma}_{T,s}^{*}$ 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 $\hat{\sigma}_{T,s}$ 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, $\hat{\sigma}_{T,s}^{*}$ 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 $\hat{\sigma}_{T,s}^{*}$. 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 'universally' 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. Consider for the moment the case of a single null hypothesis H_s 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 $w_{T,s}$. 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, it might still lead to finite sample improvements; see Section 8.

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

$$[w_{T,r_1} - \hat{c}_1, \infty) \times \ldots \times [w_{T,r_S} - \hat{c}_1, \infty) \tag{12}$$

The studentized method yields the following region

$$[w_{T,r_1} - \hat{\sigma}_{T,r_1}\hat{d}_1, \infty) \times \ldots \times [w_{T,r_S} - \hat{\sigma}_{T,r_S}\hat{d}_1, \infty) \tag{13}$$

If the sample size T is large, both regions (12) and (13) have joint coverage probability of about $1-\alpha$. But they are distinct as far as the individual coverage probabilities for the θ_{r_s} values are concerned. Assume that the test statistics $w_{T,s}$ have different standard deviations, which happens in many applications. Say w_{T,r_1} has a smaller standard deviation than w_{T,r_2}. Then the confidence interval for θ_{r_1} derived from (12) will typically have a larger (individual) coverage probability compared to the confidence interval for θ_{r_2}. This is not the case for (13) 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_{r_s} by inverting a confidence interval for θ_{r_s}. 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 $w_{T,s}$ are different, the $w_{T,s}$ 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 $z_{T,s}$.[22]

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.

In our setup, the individual null hypotheses H_s 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

$$[w_{T,r_1} \pm \hat{c}_{1,|\cdot|}] \times \ldots \times [w_{T,r_S} \pm \hat{c}_{1,|\cdot|}]$$

Here $\hat{c}_{1,|\cdot|}$ estimates the $1-\alpha$ quantile of the sampling distribution of $\max_{1 \le s \le S} |w_{T,r_s} - \theta_{r_s}|$. 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 $|w_{T,s}|$ values instead of the $w_{T,s}$ values; analogous for the modification of Algorithm 3.2

Since our focus is on comparing a number of strategies to a common benchmark, we assume that a test statistic $w_{T,s}$ is a function of the vectors $X^{(T)}_{\cdot,s}$ and $X^{(T)}_{\cdot,S+1}$ only, where $X^{(T)}_{\cdot,S+1}$ corresponds to the

[22] Alternatively, one could compare individual p-values, but this becomes more involved.

benchmark. This assumption is not crucial for our multiple testing methods. Take the example of a multiple regression model with regression parameters $\theta_1, \theta_2, \ldots, \theta_S$. The individual null hypotheses are of the form H_s: $\theta_s = \theta_{0,s}$ for some constants $\theta_{0,s}$. The alternatives can be (all) one-sided or (all) two-sided. Note that there is no benchmark here, so the last column of the $T \times (S + 1)$ data matrix X_T would correspond to the response variable while the first S columns would respond to the explanatory variables. In this setting, $w_{T,s} = \hat{\theta}_{T,s}$, where the estimation might be done by OLS say. Obviously, $w_{T,s}$ is now a function of the entire data matrix. Still, our multiple testing methods can be applied to this setting and the modifications are minor: one rejects H_{r_s} if θ_{0,r_s}, rather than zero, is not contained in a confidence interval for θ_{r_s}.

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 Delgado et al. (2001).[23] An example where the rate of convergence is T can be found in Gonzalo and Wolf (2005).

6 Alternatives to FWE Control

In this paper, we propose (asymptotic) FWE control to account for data snooping, which is the common approach. However, for certain applications, FWE control may be too strict. In particular, when the number of hypotheses is very large, it can become very difficult to reject false hypotheses. Therefore, it may be appropriate to relax control of the FWE in order to increase power. We briefly discuss three alternative proposals to this end.

The first proposal is to control the probability of making k or more false rejections, which is called the k-FWE. Here k is some integer greater than one. The second proposal is based on the false discovery proportion (FDP), defined by the number of false rejections divided by the total number of rejections. (And defined to be zero if there are no rejections at all.) In particular, one might want to control $\text{Prob}_P\{\text{FDP} > \gamma\}$, where γ is a small, user-defined number. The third proposal is to control $E(\text{FDP})$, the expected value of the FDP, which is called the false discovery rate (FDR). While different in their approaches, these three proposals share the same philosophy. By allowing a small number or (expected) fraction of false rejections, one can improve one's chances to reject false hypotheses, and perhaps greatly so.

Lehmann and Romano (2005) propose stepwise methods for controlling the k-FWE and $\text{Prob}_P\{\text{FDP} > \gamma\}$, based on individual p-values. Their methods assume a 'worst-case' dependence structure of the p-values and can therefore be viewed as generalizations of the Holm method. Current research is devoted to incorporate the dependence structure of p-values and/or test statistics in such methods in order to improve power.

Benjamini and Hochberg (1995) propose a stepwise method for controlling the FDR, based on individual p-values. However, they make the very strong assumption that the p-values are independent of each other. Benjamini and Yekutieli (2001) show that the method of Benjamini and Hochberg

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

(1995) remains valid under certain types of dependence. The problem of controlling the FDR under arbitrary dependence structures remains an open research question. For some applications of the method of Benjamini and Hochberg (1995) to econometric problems and related discussions, see Williams (2003).

7 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. (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 little practical help. In this section, we give concrete advice on how to select b in a data-dependent fashion. The method we propose, in the simpler context of constructing a confidence interval for a univariate parameter, appears in Romano and Wolf (2003), but we state it again here for completeness. 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 jth step of a stepwise procedure. The goal is to construct a joint confidence region for the vector $(\theta_{r_{R_{j-1}+1}}, \ldots, \theta_{r_S})'$ with nominal coverage probability of $1 - \alpha$. The actual coverage probability in finite samples, 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. More specifically, the solution is to 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 $(\theta_{r_{R_{j-1}+1}}, \ldots, \theta_{r_S})'$ 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 vector. The method we propose is identical except that P is replaced by a semiparametric estimate $\tilde{P}_T$. For compact notation, define $\theta^{(r)}_{R_{j-1}} = (\theta_{r_{R_{j-1}+1}}, \ldots, \theta_{r_S})'$.

Algorithm 7.1 (Choice of Block Sizes)

1. The labels $r_1, \ldots, r_S$ and the numerical values $R_0, R_1, \ldots$ are given in Algorithm 3.1 if the basic method is used or in Algorithm 4.1 if the studentized method is used, respectively.

2. Fit a semiparametric model $\tilde{P}_T$ to the observed data X_T.

3. Fix a selection of reasonable block sizes b.

4. Generate M data sets $\tilde{X}^1_T, \ldots, \tilde{X}^M_T$ according to $\tilde{P}_T$.

5. For each data set $\tilde{X}^m_T$, $m = 1, \ldots, M$, and for each block size b, compute a joint confidence region $\text{JCR}_{m,b}$ for $\theta^{(r)}_{R_{j-1}}$.

6. Compute $\hat{g}(b) = \#\{\theta^{(r)}_{R_{j-1}}(\tilde{P}_T) \in \text{JCR}_{m,b}\}/M$.

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

Remark 7.1 The motivation of fitting a semiparametric 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 semiparametric 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 7.2 Algorithm 7.1 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 context of constructing a confidence interval for a univariate parameter, Romano and Wolf (2003) find that this algorithm works very well in a simulation study.

Remark 7.3 We have suggested the use of the subsampling method in nonstandard situations where the bootstrap fails. Arguably, the choice of a good block size is then even more crucial compared to the application of a block bootstrap. A calibration method similar to Algorithm 7.1 can also be used with subsampling. For some simulation evidence that this approach yields good finite sample performance in general, see Delgado et al. (2001), Giersbergen (2002), Choi (2005), and Gonzalo and Wolf (2005).

8 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 S, the number of false hypotheses, the numerical values of the parameters θ_s, 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 stepwise 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, only results for the nominal level $\alpha = 0.1$ are reported.[24] We consider the simplest case of comparing the population mean of a strategy to that of the benchmark, as in Example 2.1.

8.1 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 $S = 40$ strategies. A basic test statistic is given by (1) and a studentized test statistic

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

is given by (2). The studentized statistic uses the formula (11). The bootstrap method is Efron's bootstrap. The number of bootstrap repetitions is $M = 200$ due to the computational expense of the simulation study. The number of DGP repetitions in each scenario is 5,000.

The distribution of the observation $X_{t,\cdot}^T$ is jointly normal. We consider two cases for the joint correlation matrix. In the first case, there is a common correlation ρ between the individual strategies and also between strategies and the benchmark; we use $\rho = 0$ and $\rho = 0.5$. In the second case, we split the strategies into two groups of size 20 each. All strategies are uncorrelated with the benchmark. Within groups, there is a common correlation of $\rho_1 = 0.5$. Across groups, there is a common correlation of $\rho_2 = -0.2$. The mean of the benchmark is always equal to 1.

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 2. The control of the FWE is satisfactory for all methods (single-step vs. stepwise and basic vs. studentized). When comparing the average number of false hypotheses rejected, one observes: (i) the stepwise method improves upon the single-step method; (ii) the studentized method improves significantly upon the basic method. Finally, the bootstrap successfully captures the dependence structure across strategies. When the correlation matrix differs from the identity, more false hypotheses are rejected.

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 3. The control of the FWE is satisfactory for all methods (single-step vs. stepwise and basic vs. studentized). When comparing the average number of false hypotheses rejected, one observes: (i) the stepwise method improves significantly upon the single-step method; (ii) the studentized method improves upon the basic method for the single-step approach, however it is worse than the basic method for the stepwise approach. Finally, the bootstrap successfully captures the dependence structure across strategies. When the correlation matrix differs from the identity, more false hypotheses are rejected.

In addition, we provide FWE-corrected results for the average number of false hypotheses rejected. To this end we adjust the nominal FWE level of the single-step methods (basic and studentized) by trial and error such that their empirical FWEs match those of the corresponding stepwise methods. The results are reported in Tables 4 and 5 (for the two classes of DGPs). It can be seen that when not all null hypotheses are false the FWE-corrected single-step methods perform very similarly now to their stepwise counterparts.[25] Therefore, the power gain of the stepwise methods can basically be explained by their ability to bring the empirical FWE closer to the nominal one in general. This

[25] When all null hypotheses are false then the FWE is equal to zero for all methods and all nominal levels α by definition, so it is not clear how to carry out a FWE correction is this case.

finding is certainly of academic interest. On the other hand, a FWE-corrected single-step method is not feasible in practice, since the proper adjustment of the nominal level would be unknown. Our simulation show that, depending on the DGP, sometimes no adjustment is required at all while at other times the adjustment can be tremendous, with nominal levels over 70% required!

8.2 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 $X^T_{\cdot,s}$ is a AR(1) process with autoregressive coefficient $\vartheta = 0.6$. In addition, we only consider the case of a common correlation $\rho = 0$ and $\rho = 0.5$ for the joint correlation matrix of a $X^T_{t,\cdot}$ vector. 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 $M = 200$ 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 7.1 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 $b = 15$ for the studentized method, respectively, by trial and error. Given that a variant of Algorithm 7.1 is seen to perform very well in a less computer intensive simulation study of Romano and Wolf (2003), we are quite confident that it 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 6. The second class of DGPs is similar to the i.i.d. case, except that the strategy means greater than 1 are evenly distributed between 1 and 7 rather than between 1 and 4. The results are reported in Table 7.

Contrary to the findings for i.i.d. data, the basic method does not provide a satisfactory control of the FWE in finite samples 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 stepwise 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.

9 Empirical Application

We consider the challenge of performance analysis when a large number of investment managers are being evaluated. In the words of Grinold and Kahn (2000, page 479): "The fundamental goal of performance analysis is to separate skill from luck. But, how do you tell them apart? In a population of 1,000 investment managers, about 5 percent, or 50, should have exceptional performance by chance alone. None of the successful managers will admit to being lucky; all of the unsuccessful managers will cite bad luck."

Our universe consists of all hedge funds in the CISDM data base that have a complete return history from 01/1992 until 03/2004. There are $S = 105$ such funds and the number of monthly observations is $T = 147$. All returns are net of management and incentive fees, that is, they are the returns obtained by the investors. As is standard in the hedge fund industry, we benchmark the funds against the riskfree rate[26], and all returns are log returns. So we are in the general situation of Example 2.1: a basic test statistic is given by (1); and a studentized test statistic is given by (2). It is well known that hedge fund returns, unlike mutual fund returns, tend to exhibit non-negligible serial correlations; for example, see Lo (2002) and Kat (2003). Indeed, the median first-order autocorrelation of the 105 funds in our universe is 0.172. Accordingly, one has to account for this time series nature in order to obtain valid inference. Studentization for the original data 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, based on $M = 5,000$ repetitions. The studentization in the bootstrap world uses the corresponding 'natural' variance estimator; for details, see Götze and Künsch (1996) or Romano and Wolf (2003). The block sizes for the circular bootstrap are chosen via Algorithm 7.1. The semi-parametric model $\tilde{P}_T$ used in this algorithm is a VAR(1) model in conjunction with bootstrapping the residuals.[27]

Table 8 lists the ten largest basic and studentized test statistics, together with the corresponding hedge funds. While one expects the two lists to be different, it is striking that they are completely disjoint. However, this result can be explained by the fact hedge funds apply very different investment strategies and, in contrast to mutual funds, can be leveraged in addition. Therefore, many funds that achieve a high average return do so at the expense of a (relatively) high risk, measured by the standard deviation. Once the magnitude of the uncertainty about the basic test statistics is taking into account through studentization, the order of the test statistics changes. The studentized list presents the more 'fair' ranking, since it accounts for the varying estimation uncertainty.

We now use the various multiple testing methods to identify hedge funds that outperform the riskfree rate, asymptotically controlling the FWE at level 0.05. The basic method does not identify a single fund. The studentized method identifies six funds in the first step and an additional seventh fund in the second step. The failure of the basic method to identify any outperformers can be attributed to the highly varying risk level across funds. The upper part of Figure 1 shows a scatterplot of the standard errors $\hat{\sigma}_{147,s}$ against the basic test statistics $w_{147,s}$. The ratio of the largest standard error to the smallest one equals $1.057/0.0477 = 22.2$! As a result the high risk hedge funds dominate the $\hat{c}_j$ values of the basic method. If the high risk funds corresponded to the funds with the largest basic test statistics $w_{147,s}$, then some outperformers might still be detected. However, as can be seen from the scatterplot, this is not the case; for example, the fund with the largest standard error actually yields a negative basic test statistic. (The lower part of Figure 1 displays the cumulative wealth in excess of the riskfree rate over the investment period of $T = 147$ months for the three funds with the highest $w_{147,s}$, $z_{147,s}$, and $\hat{\sigma}_{147,s}$ statistics, respectively.) On the other hand, the studentized method is robust in this sense because it accounts for the varying risk levels across funds

[26] The riskfree rate is a simple and widely accepted benchmark. But, of course, our methods also apply to alternative benchmarks such as hedge fund indices or multi-factor hedge fund benchmarks; for example, see Kosowski et al. (2005).

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

via studentization. To look at this issue in some more detail: If the five funds with a standard error $\hat{\sigma}_{147,s}$ above 0.8 are deleted from the sample, then the $\hat{c}_1$ value of the basic method decreases dramatically from 2.12 to 1.48. As a result, the fund with the largest $w_{147,s}$ statistic, Libra Fund, is now identified as an outperformer. In contrast, the $\hat{d}_1$ value of the studentized method decreases only slightly from 5.25 to 5.18 and the total number of identified funds remains unchanged at seven.[28]

As a final remark, when the return data are mistakenly analyzed as i.i.d. data, then the studentized method identifies 34 outperforming funds while the basic method still does not identify a single fund.

10 Conclusion

This paper advocates 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 the familywise error rate (FWE), 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 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 it potentially will reject further hypotheses in subsequent steps. This results in improved power, 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 method, 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 nonstandard 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 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

[28] Needless to say, deleting strategies from a sample based on their standard errors is an ad-hoc method that is not recommended in practice.

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 the 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 $\{1, \ldots, S\}$. 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 $\{1, \ldots, S\}$, $L_T(I)$ 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*

$$\hat{c}_I \equiv c_I(1-\alpha, \hat{P}_T) = \inf\{x : Prob_{\hat{P}_T}\{\max_{s \in I}(w^*_{T,s} - \theta^*_{T,s}) \leq x\} \geq 1 - \alpha\} \tag{14}$$

Then

$$[w_{T,i_1} - \hat{c}_I, \infty) \times \ldots \times [w_{T,i_m} - \hat{c}_I, \infty) \tag{15}$$

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

$$\begin{aligned} \text{Prob}_P\{(\theta_{i_1}, \ldots, \theta_{i_m})' \in \text{JCR (15)}\} &= \text{Prob}_P\{\max(W_T(I) - \theta(I)) \leq \hat{c}_I\} \\ &= \text{Prob}_P\{\max \sqrt{T}(W_T(I) - \theta(I)) \leq \sqrt{T}\hat{c}_I\} \end{aligned}$$

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 analogous reasoning, the sampling distribution under $\hat{P}_T$ of $\max \sqrt{T}(W^*_T(I) - \theta^*_T(I))$ also converges weakly to $\max L(I)$. The proof that

$$\text{Prob}_P\{\max \sqrt{T}(W_T(I) - \theta(I)) \leq \sqrt{T}\hat{c}_I\} \to 1 - \alpha$$

is now similar to the proof of Theorem 1 of Beran (1984). *Q.E.D.*

Proof of Theorem 3.1 We start with the proof of (i). Assume that $\theta_s > 0$. Assumption 3.1 and definition (9) imply that $\sqrt{T}\hat{c}_1$ is stochastically bounded. So $\hat{c}_1$ converges to zero in probability. By Assumption 3.1 and Lemma A.1, $\sqrt{T}(w_{T,s} - \theta_s)$, converges weakly. So $w_{T,s}$ converges to θ_s in probability. These two convergence results imply that, with probability tending to one, $w_{T,s} - \hat{c}_1$ will be greater than $\theta_s/2$, resulting in the rejection of H_s in the first step.

We now turn to the proof of (ii). The result trivially holds in case all null hypotheses H_s are false. So assume at least one of them is true. Let $I_0 = I_0(P) \subset \{1, \ldots, S\}$ denote the indices of the set of true hypotheses; that is, $s \in I_0$ if and only if $\theta_s \leq 0$. 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 $\hat{c}_{I_0} \leq \hat{c}_1$, where $\hat{c}_{I_0}$ is defined analogously to (14), we therefore have

$$\begin{aligned}
\lim_T \text{FWE}_P &= \lim_T \text{Prob}_P\{0 \notin [w_{T,s} - \hat{c}_{I_0}, \infty) \text{ for at least one } s \in I_0\} \\
&\leq \lim_T \text{Prob}_P\{\theta_s \notin [w_{T,s} - \hat{c}_{I_0}, \infty) \text{ for at least one } s \in I_0\} \qquad (16) \\
&= 1 - \lim_T \text{Prob}_P\{\theta(I_0) \in [w_{T,i_1} - \hat{c}_{I_0}, \infty) \times \ldots \times [w_{T,i_m} - \hat{c}_{I_0}, \infty)\} \\
&= 1 - (1 - \alpha) \quad \text{(by Lemma A.2)} \\
&= \alpha.
\end{aligned}$$

This proves the control of the FWE at level α. Since the argument does not assume that all S null hypotheses are true, we have indeed proven strong control of the FWE.

To prove (iii), we claim that, under the additional assumption made, the inequality (16) is strict iff at least one of the $\theta_s \in I_0$ is less than 0. Obviously, we have equality in (16) when all the $\theta_s \in I_0$ are equal to zero. So assume there exists at least one $\theta_s \in I_0$ that is strictly less than 0. Without loss of generality, assume $\theta_{i_1} < 0$ then. Adopt the notation of Lemma A.2. Since $J(P)$ has strictly positive density everywhere, the same is true for the distribution of $\max L(I_0)$, which implies that $\max L(I_0)$ has a unique $1-\alpha$ quantile. Call this quantile $\bar{c}_{I_0}$; that is, $\text{Prob}\{\max L(I_0) \leq \bar{c}_{I_0}\} = 1-\alpha$. Lemma A.2, together with the fact that the distribution function of $\max L(I_0)$ is strictly increasing everywhere, imply that $\sqrt{T}\hat{c}_{I_0}$ converges to $\bar{c}_{I_0}$ in probability. Hence,

$\lim_T \text{Prob}_P\{0 \notin [w_{T,s} - \hat{c}_{I_0}, \infty)$ for at least one $s \in I_0\}$

$$\begin{aligned}
&= \lim_T \text{Prob}_P\{\exists s \in I_0 : 0 \notin [w_{T,s} - \hat{c}_{I_0}, \infty)\} \\
&= \lim_T \text{Prob}_P\{\exists s \in I_0 : w_{T,s} > \hat{c}_{I_0}\} \\
&= \lim_T \text{Prob}_P\{\exists s \in I_0 : \sqrt{T}(w_{T,s} - \theta_s) > \sqrt{T}(\hat{c}_{I_0} - \theta_s)\} \\
&\leq \lim_T \text{Prob}_P\{\exists s \in I_0 : \sqrt{T}(w_{T,s} - \theta_s) > \sqrt{T}\hat{c}_{I_0} - \theta_s\} \quad (\text{since } \theta_s \leq 0 \ \forall s \in I_0) \\
&= \lim_T \text{Prob}_P\{\exists s \in I_0 : \sqrt{T}(w_{T,s} - \theta_s) > \bar{c}_{I_0} - \theta_s\} \quad (\text{since } \sqrt{T}\hat{c}_{I_0} \to_P \bar{c}_{I_0}) \\
&= \text{Prob}\{\exists j \in \{1, \ldots, m\} : L_{i_j} > \bar{c}_{I_0} - \theta_{i_j}\} \\
&= \text{Prob}\{L_{i_1} > \bar{c}_{I_0} - \theta_{i_1} \cup \exists j \in \{2, \ldots, m\} : L_{i_j} > \bar{c}_{I_0} - \theta_{i_j}\} \\
&< \text{Prob}\{L_{i_1} > \bar{c}_{I_0} \cup \exists j \in \{2, \ldots, m\} : L_{i_j} > \bar{c}_{I_0} - \theta_{i_j}\} \\
&= \lim_T \text{Prob}_P\{\sqrt{T}(w_{T,i_1} - \theta_{i_1}) > \bar{c}_{I_0} \cup \exists j \in \{2, \ldots, m\} : \sqrt{T}(w_{T,i_j} - \theta_{i_j}) > \bar{c}_{I_0} - \theta_{i_j}\} \\
&\leq \lim_T \text{Prob}_P\{\exists j \in \{1, \ldots, m\} : \sqrt{T}(w_{T,i_j} - \theta_{i_j}) > \bar{c}_{I_0}\} \quad (\text{since } \theta_{i_j} \leq 0 \ \forall j \in \{2, \ldots, m\}) \\
&= \lim_T \text{Prob}_P\{\exists s \in I_0 : \sqrt{T}(w_{T,s} - \theta_s) > \bar{c}_{I_0}\} \\
&= \lim_T \text{Prob}_P\{\exists s \in I_0 : \sqrt{T}(w_{T,s} - \theta_s) > \sqrt{T}\hat{c}_{I_0}\} \quad (\text{since } \sqrt{T}\hat{c}_{I_0} \to_P \bar{c}_{I_0}) \\
&= \lim_T \text{Prob}_P\{\exists s \in I_0 : w_{T,s} - \theta_s > \hat{c}_{I_0}\} \\
&= \lim_T \text{Prob}_P\{\theta_s \notin [w_{T,s} - \hat{c}_{I_0}, \infty) \text{ for at least one } s \in I_0\} \\
&= \alpha.
\end{aligned}$$

The lone strict inequality in this derivation follows from the fact that $L(I_0)$ has strictly positive density everywhere combined with the assumption that $\theta_{i_1} < 0$. *Q.E.D.*

Proof of Theorem 4.1 The proof is very similar to the proof of Theorem 3.1 and hence it is omitted. *Q.E.D.*

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_T, which can be 'decomposed' into the observed data sequence $X_{1,\cdot}^{(T)}, X_{2,\cdot}^{(T)}, \ldots X_{T,\cdot}^{(T)}$. 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 $X_{1,\cdot}^{*,(T)}, X_{2,\cdot}^{*,(T)}, \ldots X_{T,\cdot}^{*,(T)} = X_{t_1^*,\cdot}^{(T)}, X_{t_2^*,\cdot}^{(T)}, \ldots, X_{t_T^*,\cdot}^{(T)}$. The corresponding $T \times (S+1)$ bootstrap data matrix is denoted by X_T^*. The probability mechanism generating X_T^* is denoted by $\hat{P}_T$.

Bootstrap B.2 (Moving Blocks Bootstrap)
The moving blocks bootstrap of Künsch (1989) and Liu and Singh (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. A particular block $B_{t,b}$ is defined by its starting index t and by its length or block size b, that is, $B_{t,b} = \{X_{t,\cdot}^{(T)}, X_{t+1,\cdot}^{(T)} \ldots, X_{t+b-1,\cdot}^{(T)}\}$. The moving 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 $\{1, 2, \ldots, T-b+1\}$, where l is the smallest integer for which $l \times b \geq T$. The 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 $X_{1,\cdot}^{*,(T)}, X_{2,\cdot}^{*,(T)}, \ldots X_{T,\cdot}^{*,(T)}$. The corresponding $T \times (S+1)$ bootstrap data matrix is denoted by X_T^*. The probability mechanism generating 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 $X_{T+1,\cdot}^{(T)} = X_{1,\cdot^{(T)}}, X_{T+2,\cdot}^{(T)} = X_{2,\cdot}^{(T)}$, etc.. 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 $lb \geq T$. The thus selected blocks are concatenated as $\{B_{t_1^*,b}, B_{t_2^*,b}, \ldots, B_{t_l^*,b}\}$. If $lb > T$, the sequence is truncated at length T to obtain the bootstrap sequence $X_{1,\cdot}^{*,(T)}, X_{2,\cdot}^{*,(T)}, \ldots X_{T,\cdot}^{*,(T)}$. The

corresponding $T \times (S+1)$ bootstrap data matrix is denoted by X_T^*. The probability mechanism generating 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 $t_1^*, t_2^*, \ldots$ 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 $0 < q < 1/T$.[29] 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 $X_{1,\cdot}^{*,(T)}, X_{2,\cdot}^{*,(T)}, \ldots X_{T,\cdot}^{*,(T)}$. The corresponding $T \times (S+1)$ bootstrap data matrix is denoted by X_T^*. The probability mechanism generating 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 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 that the dependence on the model parameter q is not as pronounced as the dependence on the model parameter b in the two previous methods.

Remark B.1 According to a claim of Lahiri (1999), in the context of variance estimation, the moving blocks bootstrap can be 'infinitely more efficient' than the stationary bootstrap. However, there is a mistake in the calculations of Lahiri (1999), invalidating his claim. See Politis and White (2004) for a correction.

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 $S = 2$ and that[30]

$$w \sim N\left(\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right)$$

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

[30] The argument generalizes easily for $S > 2$.

with σ_1, σ_2, and ρ known. (The subscript T in w_T is suppressed for convenience.) 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_s: $\theta_s \leq 0$. We analyze power for the first step of our stepwise methods. The basic method is equivalent to the following scheme:

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

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

$$\text{Reject H}_s \text{ if } w_s/\sigma_s > d \quad \text{where } d \text{ satisfies:} \quad \text{Prob}_{0,0}\{\max w_s/\sigma_s > d\} = \alpha \tag{18}$$

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

$$\inf_{\epsilon>0} \inf_{\{(\theta_1,\theta_2):\max\theta_s \geq \epsilon\}} \text{Power at } (\theta_1, \theta_2) \tag{19}$$

Obviously, this infimum is the minimum of the two powers at $(-\infty, 0)$ and at $(0, -\infty)$.[31]

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

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

resulting in a violation of (17). Hence, the infimum in (19) for the basic method is smaller than or equal to the infimum for the studentized method. Unless $\sigma_1 = \sigma_2$, the infimum for the basic method is strictly smaller.

The second notion of power we consider is the worst power against alternatives in the class $C_\delta = \{(\theta_1, \theta_2) : \theta_s = \sigma_s \delta \text{ for some } s\}$, where δ is a positive number. Obviously, the worst power is

[31] 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 $(-\infty, 0)$.

the minimum of the two powers at $(-\infty, \sigma_2\delta)$ and at $(\sigma_1\delta, -\infty)$. The basic method yields

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

and

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

The studentized method yields

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

To demonstrate that the worst power is smaller for the basic method, we must show that

$$\max \Phi\left(\frac{c}{\sigma_s} - \delta\right) \geq \Phi(d - \delta) \tag{20}$$

This is true if $c/\sigma_s \geq d$ for some s, which we already have demonstrated above. Hence, inequality (20) holds; 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.

D Multiple Testing versus Joint Testing

To avoid possible confusion, we briefly discuss the differences between multiple testing and the related problem of joint testing; for a broader discussion see Savin (1984). It is helpful to consider two-sided hypotheses in doing so. The individual hypotheses are of the sort

$$H_s\colon \theta_s = 0 \quad \text{vs.} \quad H_s'\colon \theta_s \neq 0 \qquad \text{for } s = 1, \ldots, S \tag{21}$$

whereas the *joint hypothesis* states

$$H : \theta_s = 0 \;\forall s \quad \text{vs.} \quad H' : \exists s \text{ with } \theta_s \neq 0 \tag{22}$$

In principle, multiple testing is concerned with making individual decisions about the S hypotheses in (21) whereas joint testing is concerned with testing the *single* hypothesis (22). But one typically can use a joint test for multiple testing purposes, and vice versa.

For ease of exposition, consider the following simple parametric setup:

$$w \sim N\left(\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$$

Then the natural joint test rejects H of (22) at significance level $\alpha = 0.05$ iff $w_1^2 + w_2^2 > 5.99$. Scheffé (1959) has shown that this test can be interpreted as an induced test where there are an infinite number of separate null hypotheses of the 'linear combination' form

$$H(a) : a'\theta = a_1\theta_1 + a_2\theta_2 = 0 \quad \text{vs.} \quad H'(a) : a'\theta \neq 0 \qquad \text{with } a'a = 1$$

In particular, this test allows to make decisions about the individual null hypotheses in (21) by choosing $a = (1, 0)'$ or $a = (0, 1)'$. Therefore, the test that rejects H_s iff $w_s^2 > 5.99$, or equivalently iff $|w_s| > 2.45$, $s = 1, 2$, controls the FWE at level $\alpha = 0.05$.

But if the goal is to make individual decisions only about each parameter and not about all possible linear combinations, then the joint test is suboptimal in a multiple testing framework. A more powerful test, which also controls the FWE at level $\alpha = 0.05$, rejects H_s iff $|w_s| > 2.24$, $s = 1, 2$.

A further undesirable feature of the joint test, when applied for multiple testing purposes, is that it does not constitute a *consonant* testing procedure in the sense of Hommel (1986): a rejection of the joint hypothesis H does not necessarily result in the rejection of (at least) one of the individual hypotheses H_s. For example, in the above parametric setup, this happens if the data point $(1.9, 1.9)'$ is observed.

The message is that multiple testing and joint testing are related but distinct problems. While a joint test can, in particular, be used to address a multiple testing problem, it is generally suboptimal to do so, and vice versa.[32]

References

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

Andrews, D. W. K. (2000). Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica*, 68:399–405.

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

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.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29(4):1165–1188.

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.

Choi, I. (2005). Subsampling vector autoregressive tests of linear constraints. *Journal of Econometrics*, 124(1):55–89.

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

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

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

[32] A multiple testing method rejects the joint hypothesis H of (22) iff it rejects at least one of the individual hypotheses H_s in (21).

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

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.

Giersbergen, N. P. A. (2002). Subsampling intervals in (un)stable autoregressive models with stationary covariates. UvA-Econometrics discussion paper 2002/07, Universiteit van Amsterdam.

Gonçalves, S. and de Jong, R. (2003). Consistency of the stationary bootstrap under weak moment conditions. *Economics Letters*, 81:273–278.

Gonzalo, J. and Wolf, M. (2005). Subsampling inference in threshold autoregressive models. *Journal of Econometrics*. Forthcoming. Available at the JoE home page under 'Articles in Press'.

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

Grinold, R. C. and Kahn, R. N. (2000). *Active Portfolio Management*. McGraw-Hill, New York, second edition.

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

Hansen, P. R. (2003). Asymptotic tests of composite hypotheses. Working Paper No. 03-09, Brown University, Department of Economics. Available at `http://ssrn.com/abstract=399761`.

Hansen, P. R. (2004). A test for superior predictive ability. Working Paper No. 01-06, Brown University, Department of Economics. Available at `http://ssrn.com/abstract=264569`.

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

Hommel, G. (1986). Multiple test procedures for arbitrary dependence structures. *Metrika*, 33:321–336.

Kat, H. M. (2003). 10 things investors should know about hedge funds. AIRC Working Paper # 0015, Cass Business School, City University. Available at `http://www.cass.city.ac.uk/airc/papers.html`.

Kosowski, R., Naik, N. Y., and Teo, M. (2005). Is stellar hedge fund performance for real? Working Paper HF-018, Centre for Hedge Fund Research and Education, London Business School.

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.

Lahiri, S. N. (1999). Theoretical comparison of block bootstrap methods. *Annals of Statistics*, 27:386–404.

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

Lehmann, E. L. and Romano, J. P. (2005). Generalizations of the familywise error rate. *Annals of Statistics*, 33. Forthcoming.

Lehmann, E. L., Romano, J. P., and Shaffer, J. P. (2005). On optimality of stepdown and stepup multiple test procedures. *Annals of Statistics*, 33. Forthcoming.

Liu, R. Y. and Singh, K. (1992). Moving blocks jackknife and bootstrap capture weak dependence. In LePage, R. and Billard, L., editors, *Exploring the Limits of Bootstrap*, pages 225–248. John Wiley, New York.

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

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

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

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

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.

Politis, D. N. and White, H. L. (2004). Automatic block-length selection for the dependent bootstrap. *Econometric Reviews*, 23(1):53–70.

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`.

Savin, N. E. (1984). Multiple hypotheses testing. In Griliches, Z. and Intriligator, M. D., editors, *Handbook of Econometrics, Volume II*, pages 827–879. North-Holland, Amsterdam.

Scheffé, H. (1959). *The Analysis of Variance*. John Wiley, New York.

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

Timmermann, A. (2006). Forecast combinations. In Elliott, G., Granger, C. W. J., and Timmermann, A., editors, *Handbook of Economic Forecasting*. North Holland, Amsterdam. Forthcoming. Available at `http://www1.elsevier.com/homepage/sae/hesfor/draft.htm`.

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, revised edition.

Williams, E. (2003). *Essays in Multiple Comparison Testing.* PhD thesis, UCSD, Department of Economics.

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 $S = 40$. 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 5,000 per scenario.

Method	FWE (single)	FWE (step)	Rejected (single)	Rejected (step)
All strategy means = 1, cross correlation $\rho = 0$				
Basic	10.5	10.5	0.0	0.0
Stud	10.4	10.4	0.0	0.0
All strategy means = 1, cross correlation $\rho = 0.5$				
Basic	10.6	10.6	0.0	0.0
Stud	10.6	10.6	0.0	0.0
All strategy means = 1, $\rho_1 = 0.5, \rho_2 = -0.2$				
Basic	10.5	10.5	0.0	0.0
Stud	9.9	9.9	0.0	0.0
Six strategy means = 1.4, cross correlation $\rho = 0$				
Basic	9.7	9.7	1.1	1.2
Stud	9.6	10.1	2.2	2.3
Six strategy means = 1.4, cross correlation $\rho = 0.5$				
Basic	10.0	10.3	2.6	2.7
Stud	9.3	10.1	3.8	3.9
Six strategy means = 1.4, $\rho_1 = 0.5, \rho_2 = -0.2$				
Basic	9.7	10.1	1.4	1.5
Stud	9.7	10.1	2.6	2.6
Twenty strategy means = 1.4, cross correlation $\rho = 0$				
Basic	6.0	7.7	3.7	4.1
Stud	6.7	8.4	7.4	7.8
Twenty strategy means = 1.4, cross correlation $\rho = 0.5$				
Basic	6.1	8.9	8.6	9.6
Stud	6.2	9.4	12.6	13.2
Twenty strategy means = 1.4, $\rho_1 = 0.5, \rho_2 = -0.2$				
Basic	5.7	7.1	4.6	5.3
Stud	5.8	7.3	8.5	9.0
Forty strategy means = 1.4, cross correlation $\rho = 0$				
Basic	0.0	0.0	7.5	10.0
Stud	0.0	0.0	14.7	17.1
Forty strategy means = 1.4, cross correlation $\rho = 0.5$				
Basic	0.0	0.0	17.2	23.2
Stud	0.0	0.0	25.2	29.3
Forty strategy means = 1.4, $\rho_1 = 0.5, \rho_2 = -0.2$				
Basic	0.0	0.0	9.5	12.8
Stud	0.0	0.0	16.9	19.5

Table 3: 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 $S = 40$. 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 5,000 per scenario.

Method	FWE (single)	FWE (step)	Rejected (single)	Rejected (step)
All strategy means = 1, cross correlation $\rho = 0$				
Basic	11.3	11.3	0.0	0.0
Stud	10.4	10.4	0.0	0.0
All strategy means = 1, cross correlation $\rho = 0.5$				
Basic	11.3	11.3	0.0	0.0
Stud	10.4	10.4	0.0	0.0
All strategy means = 1, $\rho_1 = 0.5, \rho_2 = -0.2$				
Basic	10.4	10.4	0.0	0.0
Stud	10.1	10.1	0.0	0.0
Six strategy means greater than 1, cross correlation $\rho = 0$				
Basic	0.0	9.4	3.6	4.7
Stud	8.6	9.8	3.4	3.5
Six strategy means greater than 1, cross correlation $\rho = 0.5$				
Basic	0.0	10.2	4.1	5.3
Stud	8.5	10.1	4.3	4.5
Six strategy means greater than 1, $\rho_1 = 0.5, \rho_2 = -0.2$				
Basic	0.0	9.6	3.8	4.8
Stud	8.6	10.2	3.7	3.8
Twenty strategy means greater than 1, cross correlation $\rho = 0$				
Basic	0.0	6.3	9.0	13.7
Stud	5.3	8.2	9.7	10.6
Twenty strategy means greater than 1, cross correlation $\rho = 0.5$				
Basic	0.0	8.4	11.0	16.3
Stud	5.5	9.3	13.1	13.9
Twenty strategy means greater than 1, $\rho_1 = 0.5, \rho_2 = -0.2$				
Basic	0.0	5.5	9.9	14.4
Stud	5.0	6.7	10.8	11.6
Forty strategy means greater than 1, cross correlation $\rho = 0$				
Basic	0.0	0.0	15.4	24.6
Stud	0.0	0.0	18.1	21.5
Forty strategy means greater than 1, cross correlation $\rho = 0.5$				
Basic	0.0	0.0	19.7	31.5
Stud	0.0	0.0	25.6	29.2
Forty strategy means greater than 1, $\rho_1 = 0.5, \rho_2 = -0.2$				
Basic	0.0	0.0	17.3	26.3
Stud	0.0	0.0	20.1	23.8

Table 4: FWE-corrected average number of false hypotheses rejected. In each case, the nominal level of the single-step method is adjusted so that its empirical FWE matches that of the stepwise method. Observations are i.i.d., the number of observations is $T = 100$, and the number of strategies is $S = 40$. 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 5,000 per scenario.

Method	Nominal level (single)	FWE (both)	Rejected (single)	Rejected (step)
All strategy means = 1, cross correlation $\rho = 0$				
Basic	10.0	10.5	0.0	0.0
Stud	10.0	10.4	0.0	0.0
All strategy means = 1, cross correlation $\rho = 0.5$				
Basic	10.0	10.6	0.0	0.0
Stud	10.0	10.6	0.0	0.0
All strategy means = 1, $\rho_1 = 0.5, \rho_2 = -0.2$				
Basic	10.0	10.5	0.0	0.0
Stud	10.0	9.9	0.0	0.0
Six strategy means = 1.4, cross correlation $\rho = 0$				
Basic	10.0	9.7	1.1	1.2
Stud	10.5	10.1	2.3	2.3
Six strategy means = 1.4, cross correlation $\rho = 0.5$				
Basic	10.3	10.3	2.7	2.7
Stud	10.4	10.1	3.9	3.9
Six strategy means = 1.4, $\rho_1 = 0.5, \rho_2 = -0.2$				
Basic	10.3	10.1	1.5	1.5
Stud	10.3	10.1	2.6	2.6
Twenty strategy means = 1.4, cross correlation $\rho = 0$				
Basic	11.6	7.7	4.1	4.1
Stud	12.2	8.4	7.9	7.8
Twenty strategy means = 1.4, cross correlation $\rho = 0.5$				
Basic	13.2	8.9	9.9	9.6
Stud	13.4	9.4	13.3	13.2
Twenty strategy means = 1.4, $\rho_1 = 0.5, \rho_2 = -0.2$				
Basic	11.5	7.1	4.9	5.3
Stud	11.6	7.3	8.7	9.0
Forty strategy means = 1.4, cross correlation $\rho = 0$				
Basic	10.0	0.0	7.5	10.0
Stud	10.0	0.0	14.7	17.1
Forty strategy means = 1.4, cross correlation $\rho = 0.5$				
Basic	10.0	0.0	17.2	23.2
Stud	10.0	0.0	25.2	29.3
Forty strategy means = 1.4, $\rho_1 = 0.5, \rho_2 = -0.2$				
Basic	10.0	0.0	9.5	12.8
Stud	10.0	0.0	16.9	19.5

Table 5: FWE-corrected average number of false hypotheses rejected. In each case, the nominal level of the single-step method is adjusted so that its empirical FWE matches that of the stepwise method. Observations are i.i.d., the number of observations is $T = 100$, and the number of strategies is $S = 40$. 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 5,000 per scenario.

Method	Nominal level (single)	FWE (both)	Rejected (single)	Rejected (step)
	All strategy means = 1, cross correlation $\rho = 0$			
Basic	10.0	11.3	0.0	0.0
Stud	10.0	10.4	0.0	0.0
	All strategy means = 1, cross correlation $\rho = 0.5$			
Basic	10.0	11.3	0.0	0.0
Stud	10.0	10.4	0.0	0.0
	All strategy means = 1, $\rho_1 = 0.5, \rho_2 = -0.2$			
Basic	10.0	10.4	0.0	0.0
Stud	10.0	10.1	0.0	0.0
	Six strategy means greater than 1, cross correlation $\rho = 0$			
Basic	48.5	9.4	4.7	4.7
Stud	11.4	9.8	3.5	3.5
	Six strategy means greater than 1, cross correlation $\rho = 0.5$			
Basic	51.2	10.2	5.3	5.3
Stud	11.8	10.1	4.5	4.5
	Six strategy means greater than 1, $\rho_1 = 0.5, \rho_2 = -0.2$			
Basic	43.6	9.6	4.8	4.8
Stud	12.4	10.2	3.8	3.8
	Twenty strategy means greater than 1, cross correlation $\rho = 0$			
Basic	77.8	6.3	14.6	13.7
Stud	16.2	8.2	10.7	10.6
	Twenty strategy means greater than 1, cross correlation $\rho = 0.5$			
Basic	73.2	8.4	16.7	16.3
Stud	16.5	9.3	14.0	13.9
	Twenty strategy means greater than 1, $\rho_1 = 0.5, \rho_2 = -0.2$			
Basic	62.7	5.5	14.7	14.4
Stud	13.8	6.7	11.3	11.6
	Forty strategy means greater than 1, cross correlation $\rho = 0$			
Basic	10.0	0.0	15.4	24.6
Stud	10.0	0.0	18.1	21.5
	Forty strategy means greater than 1, cross correlation $\rho = 0.5$			
Basic	10.0	0.0	19.7	31.5
Stud	10.0	0.0	25.6	29.2
	Forty strategy means greater than 1, $\rho_1 = 0.5, \rho_2 = -0.2$			
Basic	10.0	0.0	17.3	26.3
Stud	10.0	0.0	20.1	23.8

Table 6: 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 $S = 40$. 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.

Method	FWE (single)	FWE (step)	Rejected (single)	Rejected (step)
		All strategy means = 1, cross correlation $\rho = 0$		
Basic	15.7	15.7	0.0	0.0
Stud	5.8	5.8	0.0	0.0
		All strategy means = 1, cross correlation $\rho = 0.5$		
Basic	16.3	16.3	0.0	0.0
Stud	5.2	5.2	0.0	0.0
		Six strategy means = 1.6, cross correlation $\rho = 0$		
Basic	14.7	15.5	1.8	1.9
Stud	5.0	5.4	1.8	1.8
		Six strategy means = 1.6, cross correlation $\rho = 0.5$		
Basic	15.6	16.8	3.7	3.8
Stud	6.8	7.5	3.3	3.4
		Twenty strategy means = 1.6, cross correlation $\rho = 0$		
Basic	9.4	12.7	6.1	6.8
Stud	3.7	5.0	5.9	6.3
		Twenty strategy means = 1.6, cross correlation $\rho = 0.5$		
Basic	11.6	16.0	12.3	13.3
Stud	4.3	6.8	11.2	12.0
		Forty strategy means = 1.6, cross correlation $\rho = 0$		
Basic	0.0	0.0	12.5	16.8
Stud	0.0	0.0	11.6	14.3
		Forty strategy means = 1.6, cross correlation $\rho = 0.5$		
Basic	0.0	0.0	24.3	30.2
Stud	0.0	0.0	22.3	27.9

Table 7: 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 $S = 40$. 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.

Method	FWE (single)	FWE (step)	Rejected (single)	Rejected (step)
All strategy means = 1, cross correlation $\rho = 0$				
Basic	15.1	15.1	0.0	0.0
Stud	7.4	7.4	0.0	0.0
All strategy means = 1, cross correlation $\rho = 0.5$				
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 $\rho = 0$				
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 $\rho = 0.5$				
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 $\rho = 0$				
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 $\rho = 0.5$				
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 $\rho = 0$				
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 $\rho = 0.5$				
Basic	0.0	0.0	16.5	29.4
Stud	0.0	0.0	14.9	19.3

Table 8: The ten largest basic and studentized test statistics, together with the corresponding hedge funds, in our empirical application. The return unit is 1 percent. Funds identified in the first step are indicated by the superscript * and funds identified in the second step are indicated by the superscript **.

$\bar{x}_{T,s} - \bar{x}_{T,S+1}$	Fund	$(\bar{x}_{T,s} - \bar{x}_{T,S+1})/\sigma_{T,s}$	Fund
1.70	Libra Fund	10.63	Market Neutral *
1.41	Private Investment Fund	9.26	Market Neutral Arbitrage *
1.36	Agressive Appreciation	8.43	Univest (B) *
1.27	Gamut Investments	6.33	TQA Arbitrage Fund *
1.26	Turnberry Capital	5.48	Event-Driven Risk Arbitrage *
1.14	FBR Weston	5.29	Gabelli Associates *
1.11	Berkshire Partnership	5.24	Elliott Associates **
1.09	Eagle Capital	5.11	Event Driven Median
1.07	York Capital	4.97	Halcyon Fund
1.07	Gabelli Intl.	4.65	Mesirow Arbitrage Trust

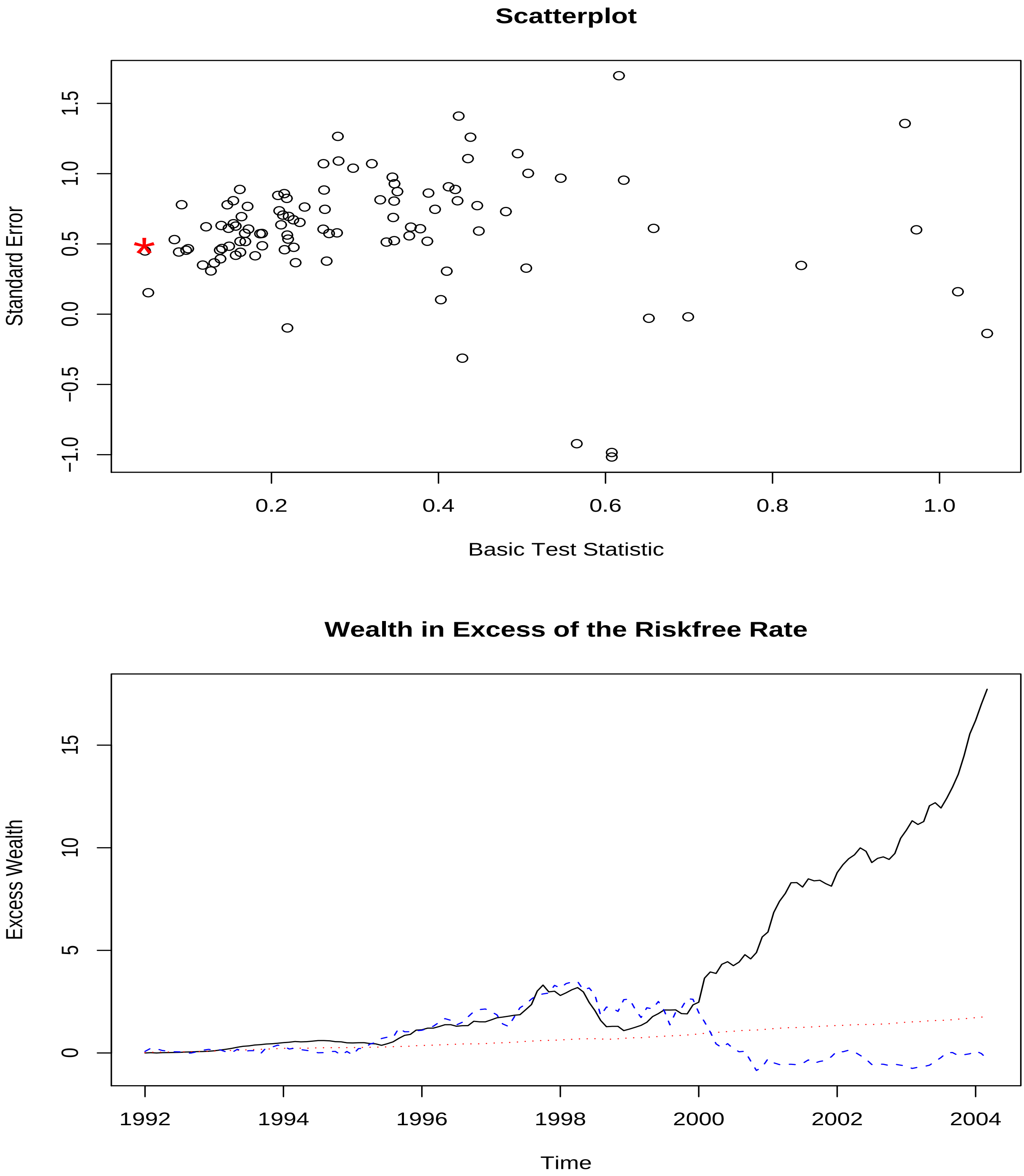

Figure 1: The top part shows a scatter plot of the standard errors $\hat{\sigma}_{147,s}$ against the basic test statistics $w_{147,s}$ in the empirical application of Section 9. The point $(0.0476, 0.5062)$ which corresponds to the largest studentized statistic $z_{147,s} = w_{147,s}/\hat{\sigma}_{147,s}$ is marked by the symbol $*$. The bottom part displays the cumulative wealth in excess of the riskfree rate, given an initial investment of 1, over the investment period of 01/1992 until 03/2004 for three hedge funds: the one with the largest basic test statistic $w_{147,s}$ (solid line), the one with the largest studentized statistic $z_{147,s}$ (dotted line), and the one with the largest standard error $\hat{\sigma}_{147,s}$ (dashed line).