An Evaluation of Alternative Multiple Testing Methods for Finance Applications[*]

Campbell R. Harvey

Duke University, Durham, NC 27708 USA
National Bureau of Economic Research, Cambridge, MA 02138 USA

Yan Liu

Purdue University, West Lafayette, IN 47907 USA

Alessio Saretto

University of Texas at Dallas, Richardson, TX 75080 USA

October 2019

Abstract

In almost every area of empirical finance, researchers are confronted with multiple tests. One high-profile example is the identification of investment managers who outperform. Many beat their benchmarks purely by luck. Multiple testing methods are designed to control for luck. Factor selection is another glaring case in which multiple tests are performed, but other numerous applications do not get as much attention. One important example is a simple regression model in which five variables are tested. In this case, because five variables are tried, a t-statistic of 2.0 is not enough to establish significance. Our paper provides a guide to various multiple testing methods and details a number of applications. We provide simulation evidence on the relative performance of different methods across a variety of testing environments. The goal of our paper is to provide a menu that researchers can choose from to improve inference in financial economics.

JEL Classification: G0, G1, G2, G3, G5, M4, C1, C5

Keywords: Multiple hypothesis testing, False rejections, False discovery rate, False non-discovery rate, False omission rate, Family-wise error rate, Data mining, Data snooping, Type I error, Type II error, False discovery control, Luck, Test power.

[*] We appreciate the comments of the Editor, Jeff Pontiff and two anonymous referees. Kay Jaitly provided editorial assistance.

1. Introduction

Suppose we observe a manager who has outperformed the market for 10 years in a row. With a track record like this, most would consider the manager skilled. Indeed, a simple p-value would be less than 0.001. That p-value fails to take into account, however, the multiplicity of managers. Suppose there are 10,000 managers and all are unskilled (think of them as flipping a coin to lever up or down the market portfolio). By random chance, we would expect to observe 9 managers with 10 consecutive years of outperformance. As a result, we cannot use simple p-values. We need to make an adjustment to deal with the multiple testing problem.

Curiously, for many years the multiple testing problem was largely ignored in empirical finance. For example, it was routine to consider a t-statistic of 2.0 as the threshold for statistical significance, but in the last few years these views have changed. Researchers working in areas where multiple testing problems are acute have had to modify their empirical approaches to control for multiple testing. But many open questions remain, some of which we address in our paper.

Our paper has three objectives. First, we provide a catalogue of different approaches that address the multiple hypothesis testing problem. Second, given the large number of techniques, we provide researchers with suggestions as to which method is best for the particular application at hand. Third, we argue that multiple hypothesis testing is a problem that touches almost all areas of empirical financial economics — not just fund and factor selection.

We have all seen papers published in both corporate finance and investment finance top journals that feature cascading tables in which variables are tried one by one and the final column includes all variables. Often asterisks in the table denote significance with three asterisks representing "significance" at the 1% level. The significance levels denoted by the asterisks are false because they fail to take multiple testing into account. In contrast to the current literature on fund and factor selection, these misleading tables are still routinely published in top journals.

Recognizing that most tests are not single tests is an important first step. For example, if we declare all managers who produce an excess return with a t-statistic greater than 2.0 as "skilled," the false-positive error rate will be massive, far higher than 5%. Each of the multiple testing methods will increase the t-statistic threshold and reduce the false positives. One question remains: What is the false positive rate after implementing the multiple testing method? Another consideration is that as we increase the t-statistic we reduce the false positives but also increase the false negatives. Thinking about the false negatives, or missed discoveries, is important because they are linked to the power of the research method to uncover true discoveries.

Finally, all of the multiple testing methods we consider are statistically based and fall under the rubric of traditional hypothesis testing. As such, they fail to take prior beliefs into account.

We make the case that in many testing situations economically motivated prior beliefs should not be ignored.

Our paper is organized as follows. In Section 2, we detail the various procedures that control for multiple testing and provide some applications. We provide computer code for each of the procedures in the appendix. Section 2.4 reviews the applications put forth in the finance literature. We then construct a simulation environment that allows us to explain the relevant challenges faced in applying multiple procedures in Section 4. Section 5 discusses how to optimize control of both false positive and negative rejections. We offer an intuitive guide on how to select an MHT procedure in Section 6. Some concluding remarks are offered in the final section.

2. Procedures that address the multiple hypotheses testing problem

A simple way to motivate the multiple hypothesis problem is to consider a situation in which a researcher explores M hypotheses using a particular data set. For example, a researcher might try to explain variable Y with M different candidate regressors, $X_1, X_2, \ldots, X_M$. A number of these hypotheses, M_0, are true under the null, H_0, while M_1 are false. The analysis of the data leads the researcher to reject P of the total M hypotheses using a traditional hypothesis testing (THT) threshold (e.g., t-statistic threshold of 1.96 or p-value of 5%). In our example, the researcher finds "significant" coefficients for P of the $X_i, i = 1, \ldots, M$ variables that were tried:

	H_0 not rejected	H_0 rejected	Total
H_0 True	TN	FP	M_0
H_0 False	FN	TP	M_1
	$N = M - P$	$P = FP + TP$	M

FP (i.e., false positive) of the P hypotheses are falsely rejected (i.e., the FP hypotheses are in fact true under the null, that is, the variables are deemed significant when they are not), and therefore FP/P is the false discovery proportion (FDP).[1] At the same time, FN (i.e., false negative) hypotheses are not rejected (variables are deemed not significant when they are), although they should be because they are false under the null, and hence FN/N is the false nondiscovery proportion (FNDP), sometimes also called the False Omission Rate. As M grows, so does FP, to the point at which the probability of making at least one false discovery grows exponentially. For example, if only one hypothesis is tested at the conventional significance level of 5%, then the probability of making a false discovery is 5% and the probability of not making a false discovery is 95%. If we test two hypotheses that are true under the null, and the corresponding tests statistics are independent,

[1] When there are no rejections (i.e., P = 0) we set FDP to 0.

then the probability of not making two false discoveries is the product of the probability of not making a false discovery for each of the two hypotheses, $(1-5\%)^2$. The probability of making a false discovery is then the complement of this quantity and equal to $1-(1-5\%)^2$. As the number of hypotheses being tested increases, the probability of making at least one false discovery for the family of tests also increases. For example, if the researcher is testing 100 different hypotheses, she will almost certainly make at least one false discovery (i.e., $1-(1-5\%)^{100}=99.4\%$).

Multiple hypothesis testing (MHT) methods are designed to control false rejections (or discoveries) by either limiting the probability that a certain number of false rejections (or the proportion relative to the total number of rejections) occur. In this section, we discuss the three main approaches and review the relevant literature.

For ease of notation an individual null hypothesis is labeled as H_m. We order hypotheses according to the magnitude of the associated test statistic: $t_1 \geq t_2 \geq \ldots \geq t_M$, or equivalently the p-values $p_1 \leq p_2 \leq \ldots \leq p_M$.[2]

2.1 Controlling the family-wise error rate (FWER)

The strictest implementation in MHT is to try to avoid any false discoveries. The FWER is defined as the probability of rejecting more than one of the true null hypotheses:

$$\text{FWER} = \text{Prob}\;(\text{FP} \geq 1).$$

A testing procedure controls the FWER at a significance level α if FWER $\leq \alpha$.[3] Five main approaches are available to control the FWER: (i) Bonferroni (1936), (ii) Holm (1979), (iii) bootstrap reality check (BRC) method of White (2000), (iv) StepM approach of Romano and Wolf (2005), and (v) the stepwise model selection (SMS) method of Harvey and Liu (2019b). The Bonferroni and Holm procedures asymptotically control the FWER under particular conditions and tend to do poorly in extreme situations,[4] such as when tests exhibit negative correlation.

The BRC, StepM, and SMS procedures allow for any arbitrary dependence in the data. FWER control is strict, as it allows at most only one false positive. As the number M of hypotheses being tested increases, these methods become more and more stringent (i.e., lead to very high thresholds for rejection). In part to alleviate this concern, the concept of FWER can also be extended to allow for control of any arbitrary number, k, of false rejections, the so called k-FWER.

[2] The methods we consider, which include both simple adjustments that require p-values as the inputs and bootstrap-based approaches, are agnostic about how p-values are generated for a particular testing problem. Researchers should choose tests that are deemed appropriate for their particular application, whether they are in-sample or out-of-sample.

[3] We use α to denote the significance level, which should not be confused with risk-adjusted fund performance.

[4] In particular, the asymptotic control of the FWER by Bonferroni and Holm works under the condition that the distribution of p-values under the null are stochastically dominated by the uniform distribution.

2.1.1 Bonferroni method

Bonferroni (1936) developed a single-step procedure in which all test statistics are compared to a single critical value. The Bonferroni method rejects H_m if $p_m \leq \alpha^* = \alpha/M$. Thus, the critical threshold for p-values is adjusted from α to α/M. Equivalently, the p-value obtained under the single test method is multiplied by the number of tests.

Let us explain. Assume that each hypothesis is tested at some significance level α^*. The probability that each hypothesis is erroneously rejected, when considered individually, is also α^*. For the entire family of M hypotheses, the expected number of rejections is $E(FP) = M \times \alpha^*$. Because Prob$(FP \geq 1) \leq E(FP)$, FWER control at level α is guaranteed in Bonferroni if one sets $\alpha^* = \alpha/M$. Hence, the Bonferroni method is the easiest to implement. The researcher simply divides the desired significance level by the number of tests and sets this as the Bonferonni threshold.

When a very large number of hypotheses are considered, the critical thresholds become very small and thus the only hypotheses that are rejected will have extremely small (large) p-values (t-statistics). Given how strict the Bonferroni method is, it will produce a very large FNDP (i.e., the method has very low power in discovering effects for which the null should be correctly rejected).

Application: 1. Bonferroni
Let us assume that a researcher considers 10 different hypotheses. In the data sample available to the researcher, all of these hypotheses are statistically significant when evaluated against the traditional hypothesis threshold (i.e., t-statistics threshold of 1.96 or p-value of 5%).

Hypothesis	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
t-statistic	1.94	-2.23	2.65	-2.01	2.51	2.70	-4.01	3.64	-2.66	4.43
p-value (%)	5.24	2.57	0.80	4.44	1.21	0.69	0.01	0.03	0.78	0.00

The Bonferroni procedure is the simplest to apply as we adjust the significance level by dividing by the number of tests considered. Since we have 10, the adjusted threshold will be 0.5% = 5%/10. Accordingly, only the nulls of hypotheses 7, 8, and 10 should be rejected.

2.1.2 Holm method

Holm (1979) introduced a stepwise procedure to control FWER, in which null hypothesis H_m is rejected at level α if $p_m \leq \alpha/(M - m + 1)$ for $m = 1, \ldots, M$. The Holm method differs from the Bonferroni method in that it is sequential (i.e., it follows a stepwise procedure). Whereas all tests in the Bonferroni method are evaluated against a constant significant threshold (i.e., α/M), in the Holm method the researcher starts by evaluating the most significant hypothesis. A rejection of the corresponding null will lead the researcher to examine the second most significant hypothesis and the procedure will stop when the researcher fails to reject a null. Importantly, in moving from the most significant tests to the less significant ones, the critical threshold keeps decreasing. In fact, the threshold is identical to that of the Bonferroni method for the most significant hypothesis ($m = 1$). For the second most significant hypothesis, the threshold is smaller at $\alpha/(M-1)$, and so on. Thus, Holm often rejects more hypotheses than Bonferroni. Despite being more lenient, the method suffers from the drawback that FWER control is guaranteed only under the assumption of independence of the individual p-values.

Application: 2. Holm

The Holm (step-down) procedure starts from the most significant hypotheses and compares them with decreasing thresholds. The procedure thus requires sorting the hypotheses based on their p-value ordered from the most significant to the least significant.

Hypothesis	(10)	(7)	(8)	(6)	(9)	(3)	(5)	(2)	(4)	(1)
t-statistic	4.43	-4.01	3.64	2.70	-2.66	2.65	2.51	-2.23	-2.01	1.94
p-value(%)	0.00	0.01	0.03	0.69	0.78	0.80	1.21	2.57	4.44	5.24
Holm threshold(%)	0.50	0.56	0.62	0.71	0.83	1.00	1.25	1.67	2.50	5.00

The procedure stops when it reaches a hypothesis that is not significant. Remember that the thresholds start at 5%/10, when there are 10 hypotheses, and keep increasing. The threshold corresponding to the second hypothesis is 5%/9, the third 5%/8 and so on so forth. In our case, the procedure stops when it arrives at the seventh (ordered) hypothesis (number 5).

2.1.3 Bootstrap reality check

The bootstrap reality check (BRC), which is based on White (2000), is a procedure designed to control the FWER.[5] To implement the method, the researcher first bootstraps the data using a resampling procedure (such as, for example, the stationary bootstrap developed by Politis and Romano, 1994). White (2000) does not impose the null when bootstrapping the data. Instead, after bootstrapping the data, in each bootstrap sample the statistic of interest (e.g., intercept from a factor model regression) is adjusted by removing the corresponding statistic in the data. The studentization is then obtained by dividing the adjusted statistic by the standard error computed in each bootstrap sample. For example, if the statistic under consideration is the intercept from a factor model regression, the bootstrap standard error would be the standard error of the intercept

[5] Notably, methods such as BRC do not apply to the study of nested models. This is not a concern for applications to mutual fund performance, but can affect, for example, the application of BRC to the cross-comparisons of nested factor models. This is because under the null that the smaller model is as good as the big model, the difference in performance between the two models will be asymptotically identical. This means that the standard error of the difference between the two models' performance converges to zero, creating a degenerate distribution for standardized test statistics that reflect differences in the models' performances.

from the regression in that bootstrap sample. For simplicity, we refer to the studentized bootstrap statistic as the bootstrap t-statistic.

For each bootstrapped iteration b, the researcher then calculates the highest (absolute) bootstrap t-statistic across all hypotheses, or $t_{\max}^{(b)}$. The BRC critical value is computed as the $(1-\alpha)$ empirical percentile of the B values $t_{\max}^{(1)}, t_{\max}^{(2)}, \ldots, t_{\max}^{(B)}$ (where B is the number of bootstrap iterations). BRC can be viewed as an improvement over the Bonferroni method in that it produces a less conservative critical value. In economic terms, the BRC method was developed to allow inference of the maximal test statistic (i.e., the most significant among all statistics examined). The idea is that the max statistic dominates other percentiles in terms of test power in rejecting the null hypothesis. As with all methods that rely on resampling, BRC does not require any particular assumption about the dependence structure of the test statistics. Chernozhukov et al. (2019) provide an extension to the BRC procedure in cases where the number of hypotheses being tested increases asymptotically with the sample size.[6]

Application: 3. BRC

Suppose a researcher is interested in evaluating the performance of the best of $1,000$ fund managers. Say the best performing fund had a average risk-adjusted realized return of 1.2% per month over the past 20 years, with a t-statistic of 3.6. The BRC procedure is designed to consider a case of this type. First, bootstrap the returns of each fund, say 5,000 times. To preserve the cross-sectional and time-series properties all fund returns and factors would be resampled using the same blocks (as for example in the stationary bootstrap of Politis and Romano). For each fund, compute the average return and its standard error in each of the bootstrap samples. If the risk adjustment method involves a factor model, then fund by fund compute the intercepts and their respective standard errors in each bootstrap sample. For each fund and bootstrap sample, calculate the studentized bootstrap statistic (i.e., the bootstrap t-statistic) as the difference between the bootstrap average return (or factor model intercept) and the corresponding quantity calculated in the original data, divided by the intercept standard error in the bootstrap sample. Organize the bootstrap t- statistics in a matrix that has 1,000 rows (funds) and 5,000 columns (bootstrap samples). Sorts each column of the matrix in descending order, and consider the first row which contains the maximum t-statistics across all bootstrap samples. The BRC procedure simply designs the threshold such as the $(1-5\%)^{\text{th}}$ percentile of the maximum statistics. Suppose the 95^{th} percentile of the max statistic is 3.5. Also, assume the best performing manager produced a risk adjusted alpha with a t-statistic of 3.6. Under a single test, this t-statistic is highly significant with a p-value$< 0.01\%$. In the bootstrap which controls for multiple testing, the manager's $t = 3.6$ exceeds what we could obtain by pure luck ($t = 3.5$) — but not by much.

[6] See Martin and Nagel (2019) for a similar setup where the number of candidate characteristics grows with time.

2.1.4 StepM method

Romano and Wolf (2005) proposed a stepwise adaptation to the BRC method that increases the number of hypotheses being rejected, thus achieving more power. The procedure requires the following steps:

1. After bootstrapping each of the M hypotheses, compute the set of maximum t-statistics, $t_{\max}^{(1)}, t_{\max}^{(2)}, \ldots, t_{\max}^{(B)}$, and the relative critical value c_1, similar to BRC, as the $(1-\alpha)$ empirical percentile. Applying the c_1 threshold, determine the hypotheses for which the null can be rejected. If there are P_1 hypotheses for which $t_m \geq c_1$, then there are $M - P_1$ strategies remaining with t-statistics ordered as $t_{P_1+1}, t_{P_1+2}, \ldots, t_M$.

2. Repeat the previous step on the set of remaining $M - M_1$ strategies, obtain a new threshold, c_2, and a new set of strategies that have not yet been rejected, $M - P_1 - P_2$ (if there are any).

3. Stop when there are no further rejected hypotheses (i.e., $M_j > 0$ and M_{j+1} is empty).

The StepM critical value is the critical value of the last step, c_j. The overall number of rejections is simply the sum of all those that are rejected at each step, and that are therefore rejected by c_j. In summary, the StepM procedure is a sequence of BRCs that constructs more and more lenient thresholds until no more hypotheses can be rejected. Similar to BRC, StepM is valid under general dependence structure of the test statistics.

Application: 4. StepM
The StepM method extends the BRC method by implementing a stepwise procedure. The goal is not to just form a statistical judgment about the best performing fund of our previous example, but to identify what other funds might be outperforming. The first step of the procedure is the same as in BRC. Different from BRC, the procedure determines how many of the original fund's t-statistics are above the maximal threshold of 3.5. Say that there are five such funds. Now eliminate the five rows corresponding to those funds from the large matrix of bootstrapped t-statistics, thus leaving a $(995 \times 5,000)$ matrix. Sort again each column and consider the first row (i.e., the row that stores the maximum bootstrapped t-statistics) and compute again the the $(1-5\%)^{\text{th}}$ percentile. Suppose that value is 3.4. Now go back and consider all the 995 funds that are left in the consideration set, and find how many have a t-statistic larger than 3.4. If there are any, then the procedure continues by repeating the previous step (i.e., eliminate the funds that are significant, and recompute the $(1-5\%)^{\text{th}}$ percentile of the largest bootstrapped t-statistics). If no new fund is found to be above the new threshold, the procedure stops. After a few iterations, in total the procedure identifies 12 funds that are deemed to have outperformed.

2.1.5 Stepwise model selection

In financial economics, researchers oftentimes seek to find relations between a response variable Y and several candidate explanatory variables $X_1, X_2, \ldots, X_M$. Harvey and Liu (2019b) presented a stepwise model selection (SMS) method that shows how to adapt the ideas in the BRC and the StepM methods to select explanatory variables that survive the multiple testing threshold. In the first step of SMS, each explanatory variable (e.g., X_i) is orthogonalized with respect to Y such that the orthogonalized variable (denoted as $\hat{X}_i$) has zero explanatory power (i.e., the null hypothesis) for Y. Orthogonalization is context specific. For example, Harvey and Liu show how to impose the null hypothesis in panel regression-based factor tests. As another example, if a zero R-squared is deemed the null, X_i can be projected onto Y, and the projection residuals constitute $\hat{X}_i$.

Using the orthogonalized variables and Y, and following a bootstrap method similar to that used in the BRC and StepM methods, Harvey and Liu obtain the bootstrapped iterations of the maximum test statistic (i.e., $t_{max}^{(1)}, t_{max}^{(2)}, \ldots, t_{max}^{(B)}$). Here, again, the maximum may carry different interpretations in different setups. In factor tests where the reduction in the intercept is regarded as evidence against the null, the maximum reduction in intercepts (or scaled intercepts) can be used as the test statistic. In predictive regressions, the maximum statistic usually refers to the maximum adjusted R-squared. After obtaining the bootstrapped maximum statistics, a decision is made by comparing the maximum test statistic that is associated with the original explanatory variables with the empirical percentiles of the bootstrapped maximum statistics. A rejection is made if the former is higher (i.e., more extreme) than the $(1 - \alpha)$ percentile of the bootstrapped maximum statistics, where α is the significance level. Otherwise, a decision to terminate the procedure is made and all variables are declared insignificant.

If a rejection occurs for say the j-th variable, this variable is now assumed to be true. The SMS method includes X_j in the model and continues to search among the remaining $M - 1$ variables. The procedure is repeated for the remaining $M - 1$ variables but this time, the remaining variables are orthogonalized against *both* Y and X_j. The idea is to obtain modified versions of these variables such that they do not provide information for Y on top of what X_j already provides. For factor tests, Harvey and Liu show how to achieve this. For predictive regressions, Y can be projected onto X_j, the projection residuals obtained, and then each of the remaining variables is projected onto the projection residuals to obtain the orthogonalized variables. The orthogonalized variables are then bootstrapped to obtain the empirical distribution of the maximum statistic (among these $M - 1$ variables), conditional on the inclusion of X_j in the baseline model. The actual maximum statistic among the $M - 1$ variables (again conditional on the inclusion of X_j) is next compared to this empirical distribution to test the significance of the remaining $M - 1$ variables. Harvey and Liu continue in this fashion to build up the final model.

Note that the SMS procedure is sequential in nature and is different from StepM. Once a useful variable is identified, Harvey and Liu's method adds it to the benchmark model, which changes the incremental contribution of each of the remaining factors. As such, test statistics are changing at each step of the SMS sequential procedure. In contrast, test statistics are fixed in the StepM method. Given this, Harvey and Liu control the FWER (in selecting an additional factor) at each step of their approach. Unlike StepM, SMS does not have a global control over FWER (in fact, a global FWER is not defined by Harvey and Liu).

The SMS procedure is related to the Model Confidence Set (MCS) proposed by Hansen et al. (2011), and recently applied by Groenborg et al. (2010), that conducts a series of pairwise comparisons among a set possible alternative tests. Similar to SMS, the MCS procedure is also sequential. However, different from SMS, which is path-dependent and includes significant factors into the baseline model at each step of its procedure, MCS puts all models on equal footing.

Application: 5. SMS
Suppose ten variables $(X_1, X_2, \ldots, X_{10})$ are tried to explain a response variable Y. First, orthogonalize each of the variables against Y and obtain the orthogonalized variables $\hat{X}_1$, $\hat{X}_2$, ..., $\hat{X}_{10}$. The orthogonalized variables achieve the null hypothesis (which is specific to the application) in sample — each variable has zero explanatory power of Y after the adjustment. Then bootstrap the time periods to obtain the bootstrapped iterations of the maximum test statistic, $t_{max}^{(1)}, t_{max}^{(2)}, \ldots, t_{max}^{(B)}$ (test statistic is again context specific; see discussion above), where B is the number of bootstrapped iterations. Empirical percentiles of the maximum statistics are then compared with the maximum statistic for the original ten variables to make inference. Suppose the maximum statistic is achieved by X_7 and it exceeds the $(1 - \alpha)$ percentile of the bootstrapped maximum statistics, where α is the significance level. Select X_7 into the model and continue to test the remaining nine variables. Orthogonalize these variables with respect to both X_7 and Y (see the discussion above on how this step of orthogonalization can be done). Bootstrap the time periods again to generate the bootstrapped iterations of the maximum statistics that evaluate the incremental contribution of the best one among the orthogonalized variables conditional on X_7. Empirical percentiles of the bootstrapped maximum statistics are then compared to the maximum statistic among the original nine variables conditional on X_7. Make decisions on the best one among the remaining nine variables. Continue in this fashion to sequentially build up the model.

2.1.6 *k*-FWER

Romano and Wolf (2007) presented an interesting expansion of the FWER control, that considers control of the probability of rejecting at least k, as opposed to one, of the true null hypotheses (i.e., the k-FWER):

$$\text{Prob (FP} \geq \text{k)} \leq \alpha.$$

A number of testing procedures control for k-FWER. Lehmann and Romano (2005) presented extensions of the Bonferroni single-step and Holm step-down procedures to allow control of k-FWER. Korn et al. (2004) proposed a procedure that is valid in the context of a multivariate permutation model, in which data and hypotheses are reshuffled (resampled without replacement); this procedure is useful in cases when hypotheses can be tested with different datasets.

Here we discuss only the k-StepM method of Romano and Wolf, which is assumption free and allows any form of dependence in the data. This procedure is implemented as follows:

1. Start by bootstrapping the set of M test statistics B times. Calculate the k-highest (absolute) t-statistic in each bootstrap sample, $t^{(b)}_{k\text{-max}}$. The critical value c_1 is computed as the $(1-\alpha)$ percentile of the bootstrap values $t^{(1)}_{k\text{-max}}, t^{(2)}_{k\text{-max}}, \ldots, t^{(B)}_{k\text{-max}}$. Assume there P_1 strategies for which $t_m \geq c_1$ and they are, therefore, rejected in this step. Let us call this set `Reject`. $M - P_1$ strategies remain with t-statistics ordered as $t_{P_1+1}, t_{P_1+2}, \ldots, t_M$.

2. Consider all $\binom{M_1}{k-1}$ possible ways of choosing $k-1$ strategies from the set `Reject`. Form the union of each of these sets with the set of remaining $M - P_1$ strategies (i.e., `Remain`). For each union set, compute the $(1-\alpha)$ empirical percentile of the k-highest bootstrap values $t^{(1)}_{k\text{-max}}, t^{(2)}_{k\text{-max}}, \ldots, t^{(B)}_{k\text{-max}}$. The critical value c_2 is the highest among all the $\binom{M_1}{k-1}$ possible $(1-\alpha)$ percentiles. Assume that there are P_2 strategies for which $t_m \geq c_2$ and they are, therefore, rejected in this step. Overall, $M - P_1 - P_2$ strategies remain with t-statistics ordered as $t_{P_2+1}, t_{P_2+2}, \ldots, t_M$.

3. Repeat the previous step and stop when there are no further rejected hypotheses (i.e., $M_j > 0$ and M_{j+1} is empty).

The overall critical value fro the procedure is the critical value computed in the last step. One critical aspect of the k-StepM procedure is the fact that, at each step, the critical value is calculated by re-considering some of the strategies that have already been rejected. The logic for doing so is that, in each step, the set of strategies that have not yet been rejected is the natural candidate for the set of strategies that are true under the null. However, the procedure is supposed to allow for (up to) $k-1$ false rejections. Since the procedure has not terminated, fewer than k (say $k-1$) discoveries have been falsely rejected. Because we do not know which were falsely rejected, Romano et al. (2008) suggest to consider all the possible combinations of $k-1$ rejected strategies.

Application: 5. k-StepM

The k-StepM further extends the BRC and StepM procedures in two significant ways. First, it allows one to derive a threshold starting from the k^{th} largest fund, as opposed to the highest performing fund. Let's say that we set the k parameter to 3. That means that at the first step of the procedure we would again sort the matrix of bootstrapped t-statistics but consider the 3^{rd} row (i.e., the row that stores the third-largest t-statistics) to construct the threshold. As before we would compute the $(1-5\%)^{\text{th}}$ percentile and obtain a value of, say, 3.57 (instead of 3.6 as in Application 3). Comparing the original fund data, we identify 6 funds with larger t-statistics. Now comes the second main difference relative to the StepM procedure. Instead of forming the new threshold (i.e., for the second step of the procedure) based on the 994 remaining funds, we would consider all the combinations (15) of $k-1=2$ among the 6 funds (i.e., 6 choose 2) and consider them once again. In practice, we obtain 15 sets of 996 = 994 + 2 strategies that we are now considering. For each of the 15 sets, we obtain a new matrix of bootstrapped t-statistics that has 5,000 columns and 996 rows. For each of these 15 sets, we compute a threshold as the $(1-5\%)^{\text{th}}$ percentile of the third largest t-statistics (i.e., the 3^{rd} row of the new matrix) and save it. The threshold for the second step of the procedure is the maximum of these 15 possible thresholds. Let us say that it is 3.56, and that means that 12 additional funds are selected, bringing the total to 18. Each further step of the procedure would repeat step two, but increase the number of consideration sets. In our case, at the third step there would be 18 choose 2 sets to consider (153 in total). After a few iterations, in the end, the procedure identifies a total of 24 funds, and is therefore more lenient than the StepM.

2.2 Controlling the false discovery proportion (FDP)

A direct extension of the k-FWER is the idea of controlling the proportion of false discoveries. In essence, instead of trying to control the number FP at some number k, the goal could be to control the ratio of FP to the total number of rejections, FP/P, to some level γ. The basic idea is to find k so that $\gamma = k/P$.

Formally, FDP control at proportion γ and level α is guaranteed when

$$\text{Prob}\ (\text{FDP} \geq \gamma) \leq \alpha.$$

Lehmann and Romano (2005), Romano and Shaikh (2006), and Romano and Wolf (2007) showed how to implement procedures in which, essentially, the number k, corresponding to some k-FWER control, such that $k/P \leq \gamma$ is found. If $\gamma = 0$, FDP control in this procedure reduces to the control of the FWER.

Genovese and Wasserman (2006) and Dudoit et al. (2004) offered other procedures that implement FDP control. Farcomeni (2008) provided a comprehensive review.

We review here the FDP-StepM procedure introduced by Romano and Wolf (2007) and Romano et al. (2008) (RSW). As mentioned previously, this method is directly linked to the procedure introduced to control k-FWER by Romano and Wolf (2007). Rather than pre-specifying k, the

RSW procedure is a recursive method in the sense that it starts by finding how many hypotheses would be rejected by imposing a 1-FWER control. If one additional rejection still keeps the FDP (i.e., $1/(R+1)$) below γ, the algorithm relaxes the constraint by imposing a 2-FWER (i.e., k-FWER with $k = 2$) control, and so on. The algorithm eventually stops at k-FWER when the FDP crosses the threshold γ.

2.2.1 FDP-StepM method

The procedure works as follows:

1. Let $j = 1$ be the index that defines the step progression, so that $k_j = j = 1$.

2. Apply the k_j-StepM method and denote by P_j the number of hypotheses rejected.

3. If $P_j < k_j/\gamma - 1$, then stop. Otherwise, let $j = j + 1$, $k_j = k_{j-1} + 1$, and return to step 2.

The FDP-StepM procedure is a sequence of k-StepM. Assume $\gamma = 10\%$. Apply the 1-StepM (i.e., k-StepM with $k = 1$) method and reject P_1 strategies. Because 1-StepM controls FWER, most likely only one false rejection has been allowed. Consider now whether the strategy H_{P_1+1}, the next most significant strategy, could be rejected. If the null of H_{P_1+1} is true, the false discovery proportion would become $1/(P_1 + 1)$. Thus we would reject H_{P_1+1} only if $1/(P_1 + 1) < 0.1$, which is true if $P_1 > 9$. Thus the procedure would stop at the first step only if $P_1 < 9$. Alternatively, the procedure would continue by applying the 2-StepM method.

The goal of the FDP-StepM procedure is to guarantee that the realized FDP in each of many possible applications that ask the same question but use different data is below a certain threshold. Alternatively, we can think of the RSW method as trying to force the tail of the distribution of FDPs (i.e., Prob(FDP $\geq \gamma$)) to behave in a certain manner (i.e., Prob $< \alpha$).

Application: 6. FDP-StepM

The FDP-StepM procedure builds on the k-StepM and StepM procedures. Suppose we run the StepM procedure (i.e., 1-StepM) and find 12 outperforming funds. Say that we are trying to control the probability that FDP is no larger than 10% (i.e., $\gamma = 10\%$) at the 5% level. In the first step, the procedure is allowing at the very most 1 false discovery with 5% probability (i.e., the 1-StepM is controlling FWER at 5%). After the first step, the procedure stops only if the number of outperforming funds (12) is smaller than $k/\gamma - 1$. Since $k/\gamma - 1 = 1/0.1 - 1 = 9$, the procedure continues and considers $k = 2$, runs the 2-StepM procedure and finds 9 more outperforming funds, bringing the total to 21. The procedure does not stop at the second step either as 21 is larger than $k/\gamma - 1 = 2/0.1 - 1 = 19$. At the third step the procedure stops because the 3-StepM finds a total of 24 outperforming funds (5 more than the second step), and that is below $k/\gamma - 1 = 3/0.1 - 1 = 29$.

2.3 Controlling the false discovery rate (FDR)

As opposed to controlling the probability that FDP is less than or equal to a threshold, the researcher might find more appealing to control the average realized FDP (across different applications). A multiple testing method controls FDR at level δ if

$$\mathrm{FDR} \equiv \mathrm{E}(\mathrm{FDP}) \leq \delta,$$

where δ is the desired tolerance level (in the same spirit as γ and α for FDP control). The two main FDR methods are those of Benjamini and Hochberg (1995) and its extension by Benjamini and Yekutieli (2001).

The procedure introduced by Benjamini and Hochberg (1995) (BH) works in a stepwise fashion. Order the individual p-values from the smallest to largest. Define

$$j^* = \max\left\{ j : p_j \leq \frac{j \times \delta}{M} \right\},$$

as the the rank order of the least significant p-value that corresponds to a rejected hypothesis. In contrast to the Holm method that starts with the most significant hypothesis, BH is a step-up method: it evaluates hypotheses moving up from the least significant test.

We give the intuition next. Say that j is the j-th most significant hypothesis with p-value of p_j. Rejecting all hypotheses up to the j-th one, a researcher should expect $M \times p_j$ false rejections. In other words, $(M \times p_j)/j$ is the realized FDR. The procedure starts from the least significant hypothesis and stops at the j^* so that the FDR is less than δ.

Application: 7. BH
The Benjamini and Hochberg procedure starts from the least significant hypothesis and works its way up. The threshold is calculated in a way that is increasing starting from the less significant hypothesis. When there are 10 hypotheses, the threshold is equal to the chosen FDR control (e.g., 5%) for the least significant hypothesis. For the ninth hypothesis the threshold is 9/10 × 5%; it is 8/10 × 5% for the eighth most significant and so on.

Hypothesis	(10)	(7)	(8)	(6)	(9)	(3)	(5)	(2)	(4)	(1)
p-value	0.00	0.01	0.03	0.69	0.78	0.80	1.21	2.57	4.44	5.24
BH threshold	0.50	1.00	1.50	2.00	2.50	3.00	3.50	4.00	4.50	5.00

The procedure stops when it finds the first significant hypothesis, in our case number (4), thus leaving nine significant hypotheses.

The procedure proposed by Benjamini and Hochberg (1995) controls FDR under the assumption that test statistics are independent. Benjamini and Yekutieli (2001) propose an alternative procedure (BY) that controls FDR and allows for more general dependence in tests statistics. To implement BY, replace the definition of j^* in BH with:

$$j^* = \max\left\{j : p_j \leq \frac{j \times \delta}{M \times C_M}\right\},$$

$C_M = \sum_{i=1}^{M} 1/i \approx \log(M) + 0.5$. The BY procedure is stricter (and less powerful) than BH.

Application: 8. BY
Similarly to BH, the Benjamini and Yekuteli procedure starts from the least significant procedure. The p-value thresholds are a scaled version of those from the BH procedure. The scaling factor is constant across all hypothesis: when there are 10 hypotheses it is equal to $\sum_{i=1}^{10} 1/j = 2.92$. Therefore, the threshold for the tenth hypothesis becomes $5\%/2.92 = 1.71$.

Hypothesis	(10)	(7)	(8)	(6)	(9)	(3)	(5)	(2)	(4)	(1)
p-value	0.00	0.01	0.03	0.69	0.78	0.80	1.21	2.57	4.44	5.24
BY threshold	0.17	0.34	0.51	0.68	0.85	1.02	1.19	1.37	1.54	1.71

In the above case, the procedure stops when it arrives at the sixth hypothesis, number (3). Note that because BY is a step-up procedure and it stops with hypothesis number (3), hypothesis number (6) is still among those for which the null is rejected even if its p-value is below the corresponding BY threshold.

Important extensions of the FDR control result from the work published by Storey (2002, 2003) and Storey et al. (2004), who introduced the positive FDR (pFDR) as a way to correct for cases in which a small or zero number of hypotheses might be rejected. The pFDR is particularly useful because it can be characterized as a Bayesian posterior probability in cases when the data generating process is a mixture distribution.

In practical terms, Storey (2002) suggests replacing M, the total number of strategies used by Benjamini and Hochberg (1995), with an estimate M_0 of the number of true null hypotheses, as given by

$$M_0 = \frac{\#\{p_j > \lambda\}}{1 - \lambda},$$

where $\lambda \in (0,\, 1)$ is a user-specified parameter. Thus, Storey's procedure rejects all hypotheses for which

$$j^* = \max\left\{j : p_j \leq \frac{j \times \delta}{M_0}\right\}.$$

Applying Storey correction is not straightforward because he provides little guidance in how to pick λ. Storey proposes a heuristic to select an "optimal" λ. In the applications, for example, by Barras et al. (2010) and Harvey and Liu (2019b), however, different values of λ produce very similar results.

Differently from procedures that rely on resampling, BH and BY are computationally very easy to implement. One important drawback is, however, that FDR only controls the average FDP, and in a given dataset the realized FDP could be very far from the tolerance level δ (i.e., the realized standard deviation of FDP might be large). Thus in applying a procedure that controls FDR to one dataset, a researcher must be careful not to assume that FDP control is also achieved.

Table 1: Frequently used acronym terms

Acronym	Definition		
FP	False positive. Hypotheses rejected by mistake (Type I Error)		
FN	False negative. Hypotheses not rejected by mistake (Type II Error)		
FDP	FP / P where P is the number of rejected hypotheses		
FNDP	FN / N where N is the number of non-rejected hypotheses		
Acronym	Control	Definition	Typical parameter choice
FWER	Family wise error rate	Prob (FP $\geq$ 1)$\leq \alpha$	$\alpha = 5\%$
FDP	False discovery proportion	Prob (FDP$\geq \gamma$) $\leq \alpha$	$\gamma = 5\%$, $\alpha = \{5\%, 10\%\}$
FDR	False discovery rate	E(FDP)$\leq \delta$	$\delta = \{1\%, 5\%\}$
Acronym	Paper	Control	Assumptions
BRC	Bootstrap reality check, White (2000)	FWER	free
StepM	Romano and Wolf (2005)	FWER	free
SMS	Stepwise model selection, Harvey and Liu (2019b)	FWER	free
RSW	Romano et al. (2008)	FDP	free
BH	Benjamini and Hochberg (1995)	FDR	independent tests
BY	Benjamini and Yekutieli (2001)	FDR	positive correlation

Note: The table lists the most frequently used acronyms and their meanings.

2.4 Recent developments

Recent developments in MHT are aimed at incorporating prior knowledge into multi-step procedures. Benjamini and Hochberg (1997) formalized the idea of weighing each rejection by a positive constant (potentially different for each rejection). A number of ways are available for introducing weights, such as directly in a similar fashion as the Holm procedure or indirectly by modifying the objective function that is optimized by the procedure (e.g., think of a modified FDR in which the expectation of a false rejection is weighted so that more important rejections carry a larger weight). This idea has been operationally developed into several procedures that differ largely depending on the field in which they are applied (i.e., spatial cluster analysis or genome-wide association studies). We review here only a few potentially interesting cases that could be applied in finance.

Benjamini and Bogomolov (2014) proposed reducing the number of hypotheses that are considered by eliminating groups of hypotheses that are likely true under the null.[7] The procedure consists of two steps: in the first step, the researcher divides hypotheses into groups and determines

[7] The idea behind Benjamini and Bogomolov (2014) is related to papers that show how power can be improved by reducing the number of moments that are been tested; see e.g., Hansen (2005), Andrews and Soares (2010), and Romano et al. (2014).

whether each group is worth considering; an in the second step, groups that are not interesting are eliminated. The FDR procedure is then adjusted for the fact that some groups have been discarded, and the procedure is then applied to the remaining hypotheses. The logic behind this procedure is that a researcher who has sufficient *a priori* knowledge to separate hypotheses that are sufficiently different (i.e., non-exchangeable) should be able to eliminate the uninteresting cases. In turn, this increases the probability of finding hypotheses that are interesting. Eventually, the procedure can be viewed as a way to impose weights on different groups, so that the groups viewed as noncredible receive a weight of zero. Barber and Ramdas (2017); Ramdas et al. (2019) provided an algorithm that allows for generalized partitions of hypotheses under FDR control.

Basu et al. (2018) introduced a procedure in which weights are assigned to all hypotheses tested, not just those that are falsely rejected. Their objective function is defined as the maximization of the expected weighted true positives (i.e., that should be rejected) subject to a constraint that the weighted FDR is controlled at some level.

In the same spirit, but with a completely different approach, Harvey and Liu (2019a) proposed a procedure that flexibly estimates both Type I and Type II error rates. Their procedure allows for the calibration of the true Type I error rate for a given data set and multiple testing procedure, which may be very different from the nominal size of the test. It also permits the consideration of the trade-off between Type I and Type II errors by taking into account the potential differential costs of the two types of errors. We discuss this procedure in Section 5.

2.5 Bayesian approaches to multiple hypothesis testing

Whereas in this paper we mainly consider the frequentist approaches to MHT, Bayesian methods have been developed to address the multiple testing issue. One key feature of the Bayesian approach, which is different from the frequentist approach that relies on the introduction of overall error rates such as FWER or FDR, is that the Bayesian posterior probabilities automatically adjust for simultaneous tests of many factors or managers.

Two main strands of research apply Bayesian methods to MHT. The first strand, such as the work of Efron and Tibshirani (2002) and Storey (2003)), uses the Bayesian or empirical-Bayes argument to justify existing frequentist approaches such as the FDR. The other strand, such as the work of Jefferys and Berger (1992), Scott and Berger (2006), and Scott (2009), develops full-blown Bayesian models to address test multiplicity.

For finance applications, the recent work of Barillas and Shanken (2018), Chib et al. (2019), and Bryzgalova et al. (2019) explores the full-Bayesian approach to the selection of useful risk factors in explaining portfolio returns. Harvey et al. (2019) have proposed a Bayesian MHT to adjust existing factors for publication bias and multiple testing.

Several hurdles exist for successfully applying Bayesian methods to the MHT problem. First, the construction of the estimation framework is specific to the research question as well as the data. Bayesian solutions are also much more challenging to obtain compared with the aforementioned frequentist methods. Finally, test correlations are more difficult to incorporate for Bayesian approaches than for certain frequentist methods, such as bootstrap-based adjustment methods. Given these concerns, in this paper we primarily focus on the frequentist approach, however, our emphasis should not be misinterpreted. Despite the challenges in implementing the Bayesian approach, this area of research is very promising. Indeed, later in the paper we will introduce a hybrid approach that requires a minimal amount of prior information to be injected into a frequentist framework.

3. Finance applications

The idea that finance applications might be exposed to the MHT problem is not new. Lo and MacKinlay (1990) made a first attempt to quantify the data-snooping bias in asset pricing applications, and others, such as Shanken (1990) and Ferson and Harvey (1999), have also incorporated MHT-adjusted statistics in their analyses.[8]

Recently, a number of researchers have published papers in which they have applied some of these methods to address questions about either fund performance or the statistical significance of trading strategies constructed by sorting stocks based on a certain signal (i.e., anomalies and factors). The rationale is as follows. For fund evaluation, the interest is in whether funds outperform their benchmarks — in the short run or in the long run — since many funds are benchmarked against the same factors, and by and large, managers have access to the same tradable assets. Therefore, when evaluating a single fund, it is important to take into account that all other funds also need to be evaluated.

In the case of anomalies, the problem is complicated by the fact that not only are many strategies evaluated at the same time (similar to fund returns), but each "new" strategy is often analyzed in isolation. Nonetheless, many long-short portfolios have been studied, although possibly at different points in time. Thus, multiple hypotheses have been tested.

The application of MHT to these two lines of inquiry faces different challenges. On the one hand, when studying fund returns, a researcher has a fixed number of testable hypotheses (i.e., one for each fund) which can be tested on time-series of different and sometimes short samples, thus often leading to a power problem. Even more daunting, funds that appear to be successful often exhibit a sizable correlation in returns, possibly because they herd their trades (Wermers, 1999).

On the other hand, anomalies and factors have much longer time series and are only weakly correlated. The application of MHT is vexed, however, by the complication that researchers do not

[8] See also, Sullivan et al. (1999), Boudoukh et al. (2007), Patton and Timmermann (2010), Bajgrowicz and Scaillet (2012), Bajgrowicz et al. (2016), and Qu et al. (2019).

observe the entire distribution of all long-short portfolios that have been studied, but only those that are associated with a statistically significant (i.e., at conventional thresholds) return (whether that is an abnormal return or a risk premium).

3.1 Fund performance

Barras et al. (2010) applied the FDR procedure of Storey (2002) to quantify the empirical null (i.e., proportion of mutual funds that have zero net alpha). They found a very small percentage (i.e., 0.8%) of fund managers who exhibit some skill. The Storey procedure delivers a robust measure of the FDR for applications in which the true data-generating process is a mixture distribution, under independence of the test statistics, and the assumption that p-values corresponding to true null hypotheses are large. Both assumptions have been under scrutiny in the finance literature.

In the original Storey model, discoveries can only arise from one of two distributions. Introducing a third distribution (i.e., negative, zero, and positive alpha funds) complicates the number of possible classification errors. Ferson and Chen (2019) showed that modifying the classification scheme to allow bad (good) funds to be classified as unskilled (skilled) leads to the estimate of a much smaller percentage of funds as having zero alpha (and many more having negative alphas).

Andrikogiannopoulou and Papakonstantinou (2019) and Barras et al. (2019) found that low power and parameter uncertainty are largely responsible for overestimating the empirical null and for underestimating the proportion of funds with positive alphas. In essence, because individual funds' alpha estimates are very noisy and the fund return time series are all different lengths, there is no guarantee that the empirical distribution of p-values behaves according to the hypothesis made by Storey. That seems to invalidate the second assumption needed for the Storey procedure (i.e., that p-values corresponding to true null hypotheses are large) and casts some doubt on the blanket application of the procedure. To mitigate these problems, Harvey and Liu (2019a) proposed a new procedure (discussed in Section 5) that can aid in situations when power is a concern by balancing the control of false discoveries and false nondiscoveries (i.e., funds with true alpha that go undetected).

The independence assumption is also problematic, because many funds with large realized alphas also exhibit relatively high correlation in their trading strategies (Wermers, 1999). Instead of relying on an MHT procedure that allows for arbitrary dependence in the data, Giglio et al. (2018) proposed recasting the asset pricing tests that quantify abnormal performance so that they are independent, thus allowing the straightforward application of the FDR procedure of Benjamini and Hochberg (1995). Giglio et al. (2018) apply their method to hedge fund returns and found that approximately one-quarter of the statistically significant alphas in a traditional hypothesis framework are likely false.

Another avenue of literature attempts to answer the basic question of whether any fund possesses skill by controlling for multiple testing using a bootstrap procedure, such as proposed by Kosowski et al. (2006) and Fama and French (2010), and as applied recently Blake et al. (2017) and Yan and Zheng (2017).

The idea is to bootstrap the cross-section of fund returns under the null hypothesis that no fund outperforms. Bootstrapping is done by Fama and French over the time dimension while keeping the cross-section intact. This helps preserve the cross-sectional dependence in the data. Inference involves comparing particular percentiles of the ordered statistics in the actual data to the bootstrapped distributions, much in the same way as White (2000) conducted inference on the max statistic.

While both intuitive and simple to use for controlling cross-sectional dependence in the data, recently Harvey and Liu (2019a) called for a more thorough examination of test power for the bootstrap approach. Due to the unbalanced nature of fund histories, funds with a small number of observations may be sampled poorly in the bootstrapped samples, leading to extreme values of the test statistics that distort the tails of the cross-sectional distribution of test statistics. This distortion in turn leads to biased inference on the maximum test statistic, causing researchers to reject less frequently than they should. This lack of power means that many outperforming funds are falsely classified as underperforming.

3.2 Long-short portfolios

An anomaly is a trading strategy whose return cannot be explained by a factor model. Factors are large portfolios that can explain some of the variability in expected stock returns. Before the Fama and French (1993) three-factor model, size and book-to-market were anomalies, but now they are believed by many to be risk factors. We do not make such a distinction for the purpose of our paper. In very simplified terms, the abnormal return on the trading strategy is “equivalent” to its risk premium (if that strategy becomes a factor), and we are concerned with the t-statistic or p-value of the estimate.

Harvey et al. (2016) conducted one of the first meta-studies in empirical asset pricing by collecting the reported t-statistics from circulated papers that introduced a factor or an anomaly. They showed that many factors are likely false even when MHT adjustments are applied exclusively to the set of published discoveries. Harvey et al. (2016) recognized that if all the possibilities that were likely attempted by researchers could be accessed (i.e., the file drawer problem or publication bias, as Harvey, 2017, described), MHT corrections would be even stricter and lead to a higher proportion of false discoveries.[9] The evidence offered by Chordia et al. (2019), which was based on over two million trading strategies, is consistent with this observation.

[9] McLean and Pontiff (2016) showed that after a paper’s publication the strength of the published effect is greatly diminished, which is consistent with the idea that some of the discoveries are false or exaggerated.

Because factors are generally constructed using the entire available CRSP and Compustat databases, power is less of a problem here: almost all factor return realizations have more than 500 observations at the monthly frequency. Correlation among factors might play a role, although the average reported absolute pairwise correlation is not very large (Green et al., 2017, report an average of 9%.).

The application of MHT to factors is particularly interesting given that hundreds of academic papers propose factors and over 1,000 investment products implement these ideas. Given the severity of the multiple testing problem, most of these factor discoveries are likely false and most of the investment products are unlikely to outperform. A growing literature and many other applications are providing insight on multiple testing problems; see, for example, the work of Engelberg et al. (2019) on time-series returns prediction and of Mitton (2019), who deals with a number of problems in empirical corporate finance.

3.3 Other settings

The MHT problem exists in many other settings, interestingly, even in situations when exogenous variation in the data is used to bolster the interpretation of an effect as a causal relationship. For example, List et al. (2019) show how to adapt the k-StepM procedure of Romano and Wolf (2010) to control the k-FWER in the context of experimental economics when many outcomes are examined in the context of the same experiment.

Davidson et al. (2019) studied the MHT problem in a context of which the same experimental design (i.e., the passage of business combination laws and the Regulation SHO pilot) is used by many researchers to investigate different outcomes. Using the StepM method of Romano and Wolf (2005), they showed that subsequent application of the same experimental design increases the probability of false discoveries.

3.4 Challenges in applying MHT methods in finance

We face three key challenges when applying MHT adjustments in financial applications. First, the most appropriate correction method depends on both the problem and the particular data at hand. Second, we do not know all the tests that have been tried — only the ones made public as a result of publication. Finally, even if the first two problems are resolved, we do not know the specific false discovery rates of any of the procedures.

4. Simulation

We highlight the properties of different MHT procedures in a simulated environment where we can exactly control power, noise, correlation, and publication bias. While we focus on these metrics in our simulation framework, readers can mimic our setup to study other issues such as unbalanced panel, heterogeneous and time-varying volatility, and non-normality in the data generating process.

4.1 Illustrative example

Let us consider 200 time series of monthly returns. We will call them “portfolios” with the implicit understanding that they could be the monthly excess returns of 200 funds or of 200 long-short portfolios. For our purpose, the difference is irrelevant. Each month the return of each portfolio is generated by a factor structure,

$$r_{p,t} = \alpha_p + \beta_p \times F_t + \epsilon_{p,t},$$

where F_t is the realization of the benchmark risk factor (i.e., it could be more than one risk factor) and β_p is the portfolio exposure to the factor. ϵ represents normally distributed noise and has mean zero and variance σ_ϵ. Possibly, the variance-covariance matrix of all P noise terms is nondiagonal (i.e., there might be some correlation structure).

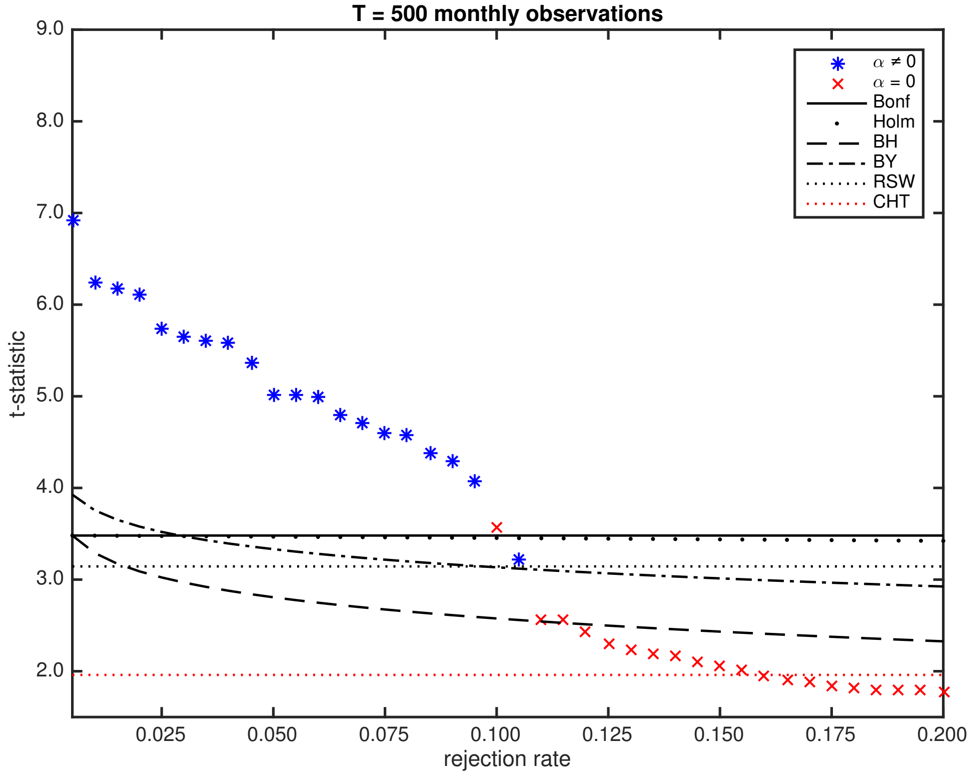
We inject an α_p of 1% per month for 5% of the portfolios and an α_p of -1% per month for 5% of the portfolios. In this way $\pi = 10\%$ of the portfolios have a non-zero alpha. The remaining 90% have zero alpha. Each β_p is also drawn from a normal distribution with mean zero and variance $(10\%)^2$.

After generating a time series of returns for each portfolio and for the factor, we run a time-series regression to recover estimates of α_p and β_p, thus mimicking what researchers do with real data. We are interested in the estimate of α_p and in testing the hypothesis that such estimate is different from zero (i.e., the null is that the true α_p is zero). Aside from the many important and sometimes troubling complications of interpreting them (see Harvey, 2017), hypotheses will be evaluated by comparing t-statistics or p-values to some threshold. If the p-value (absolute value of the t-statistic) is lower (larger) than the threshold, the corresponding null hypothesis is rejected.

Under the null hypothesis, each portfolio true α_p is zero, however some of the hypotheses will be rejected if conventional thresholds are applied. For example, we would expect 5% of the estimated alphas to have a t-statistic larger than 1.64, and 5% to be lower than -1.64. Rejecting the associated nulls would lead the researcher to make false discoveries because we know that under the null all alphas are zero. The problem is more complicated when some of the portfolio's true alphas are in fact different from zero. In that case, we could have some false discoveries and some false nondiscoveries.

We start by illustrating how the main MHT methods work with a simple scenario and then try to explain what will change when we change our baseline assumptions. We simulate portfolios at monthly frequency. Let us say that the distribution of each noise term is normal (i.e., $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$, where $\sigma_\epsilon = 5\%$ per month). We also assume the factor return is $F \sim \mathcal{N}(0.6\%, 4.3\%^2)$, which is similar to the market return characteristics. Besides what is provided by the common factor, no other source of dependence is considered (i.e., ϵ_p are serially and cross-sectionally independent). In total, we have 20 portfolios with $\alpha_p \neq 0$, or 10%. After generating the data, we run 200 regressions and collect the alpha t-statistics.

Figure 1: MHT thresholds



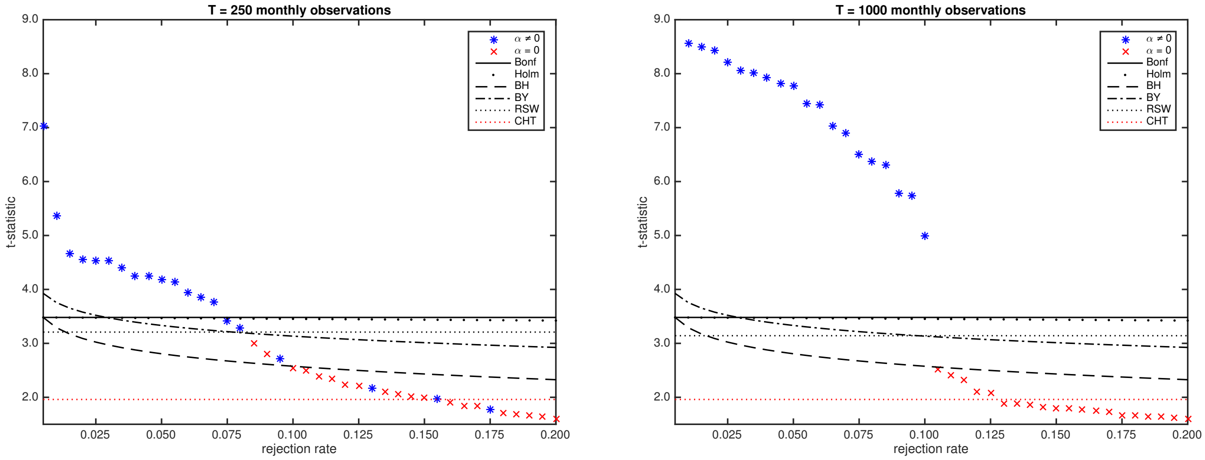
Note: The figure plots the 40 most significant t-statistics from our illustrative example against THT (i.e., t-statistic threshold of 1.96) and MHT thresholds. Blue stars denote hypotheses that are true under the alternative; red x symbols denote hypotheses that are true under the null of no abnormal performance. 200 strategies are simulated for 500 months.

We present a typical situation out of one simulation in Figure 1. Each star and each x-symbol on the plot marks one hypothesis' absolute t-statistic. The plot shows the 40 hypotheses with the highest t-statistics in decreasing order. The stars correspond to hypotheses for which the null is not true (i.e., $\alpha_p \neq 0$) and the x-symbol corresponds to situations for which the null is true ($\alpha_p = 0$). Lines that cross the chart from left to right correspond to thresholds: MHT in black and traditional hypothesis testing (THT) in red. We compare MHT procedures to THT at the conventional statistical significance level of 5%. For MHT procedures, FWER and FDR are also controlled at 5%. In the case of the Romano-Shaikh-Wolf (RSW) method we control the probability that the FDP is less than 5% at 10%.

The Bonferroni and RSW procedures produce a constant threshold. The Holm, BH, and BY methods instead produce threshold functions that are decreasing (i.e., remember these are stepwise procedures), so they will produce lower thresholds if many hypotheses have very large t-statistics. Hypotheses above a threshold line are rejected.

In a perfect situation, we should reject all 20 hypotheses corresponding to the blue stars, while avoiding rejecting any of the red x symbols. In this particular instance, the Bonferroni and Holm procedures correctly reject 19 hypotheses and lead to one false rejection (i.e., FDP of 5.2% = 1 / 19). The BY and RSW procedures correctly reject one further hypothesis. The BH procedure, which is the least stringent, rejects two more null hypotheses than BY and RSW, both of them wrong. Thus, the BH procedure realizes an FDP of 13.4% = 3/23. Remember that BH controls FDR = E(FDP) at 5%, so that over many simulations we would expect the average FDP to be below 5%. In this particular simulation, the control was not achieved but the requirement is focused on the average over many simulations. In contrast, a traditional threshold of 1.96 would lead to 31 total rejections, 11 of which are false (i.e., FDP of 35.4% = 13/31).

Figure 2: MHT thresholds and power



Note: The figure plots the 40 most significant t-statistics from our illustrative example against THT and MHT thresholds. Blue stars denote hypotheses that are true under the alternative; red x symbols denote hypotheses that are true under the null of no abnormal performance. The left panel considers the case where the tests are underpowered (i.e., the time-series of observed returns are short at T= 250 observations). The right panel considered the case were tests are at full power (i.e., T = 1,000).

Increasing the number of observations, which leads to increased power, has a big effect on the MHT procedure. In Figure 2 we report two other possible instances, one characterized by a relatively small number of observations, which could decrease power (i.e., $T = 250$ observations), and one with a relatively large number of observations, which will likely have more power (i.e., $T = 1,000$ observations). When T is small, only the Bonferroni and Holm procedures that control the probability of making one false discovery do not reject any false hypotheses. Remember that because the Bonferroni and Holm procedures control FWER (i.e., at 5%), if we repeated the simulation many times, in not more than 5% of the cases the two procedures should lead to a rejection of one false hypothesis. In general, a small T (hence, likely low power) means that all procedures will fail to reject some hypotheses that should be rejected (BH, which is the most lenient procedure, fails to reject three while THT fails to reject one). When T is large, there is enough power to precisely estimate alphas, and all MHT procedures become very accurate and reject all hypotheses that should be rejected. Since all procedures have some tolerance for false rejections some false discoveries might occur. This is essentially the argument advanced by Andrikogiannopoulou and Papakonstantinou (2019) and Harvey and Liu (2019b).

4.2 Simulated economies

We highlight more formally how the application of MHT procedures is affected by the assumptions that we make by expanding our simulation exercise. We generate 1,000 economies wherein we observe 5,000 portfolios (instead of 200). In order to apply the RSW procedure, each economy is bootstrapped 1,000 times using a stationary block bootstrap (see Politis and Romano, 1994). If it is not otherwise specified, we keep the number of monthly observations T to 500, because that is the typical range found in most finance papers (500 monthly observations cover approximately 40 years during which the Compustat database is free of survivorship bias). In particular, we are interested in examining the three problematic assumptions that are all challenges for most of the literature that we reviewed in Section 3: power, correlation structure, and publication bias.

4.2.1 Sample size, signal to noise, and power

In a traditional setting, the power of a statistical test is the probability of rejecting the null hypothesis when the null is not true. Thus, the power of the test depends on the size of the effect (i.e., the signal to noise ratio), size of the sample, and statistical significance level. In a multiple testing context, power is related to the probability of rejecting all null hypotheses that are false. Therefore, a procedure has maximum power if the false nondiscovery proportion (i.e., FNDP = FP/P) is zero.

Variation in the power of a MHT technique will not only affect FNDP, but also the t-statistic thresholds that the procedures produce and therefore the overall proportion of rejections. Two effects are simultaneously at play. On the one hand, similar to THT, higher power reduces the probability of not rejecting an hypothesis that should be rejected, for any t-statistic threshold. On the other hand, most MHT procedures are adaptive to the data (see Chordia et al., 2019). If power is high as a result of a large number of observations or because the signal-to-noise ratio is high, then it is easier to separate hypotheses with a true non-zero alpha from hypotheses with a true zero alpha. If this is the case, most MHT procedures will require a low threshold to control (the proportion of) false positives, which will largely be driven by the distribution of zero alphas. On the other hand, if power is low then the distributions of zero and non-zero alpha hypotheses are very close. Given that MHT procedures focus on controlling the false positives, they will produce a high t-statistic threshold. Moreover, because the researcher choses the level at which to control false positives, the false positive control does not change with power, however, the realized FDP might slightly change.

We analyze the impact of power on MHT procedures by considering variation in sample size, holding the signal-to-noise ratio fixed, and by considering variation in the signal-to-noise ratio, holding the sample size fixed.[10]

Table 2 shows the averages and standard deviations for t-statistic thresholds, rejection rates, FDP, and FNDP for three different sample sizes: 250, 500, and 1,000 monthly observations. [11] FDP (i.e., FP/P) and FNDP (i.e., FN/N) here are the realized proportion of false discoveries and nondiscoveries in each of the 1,000 simulated economies. Thus, the realized FDR is just the average of the FDP across the 1,000 simulations.[12] We tabulate results for the Bonferroni and Holm procedures that control FWER at a 5% level, for the Benjamini-Hochberg (BH) and Benjamini-Yekuteli (BY) procedures that control FDR at 5%, and for the Romano-Shaikh-Wolf (RSW) procedure that controls the probability that FDP is larger than 5% at the 5% and 10% levels, respectively.[13]

Table 3 shows the averages and standard deviations for t-statistic thresholds, rejection rates, FDP, and FNDP for three cases in which we vary the signal-to-noise ratio by changing the magnitude of non-zero alphas (abnormal returns) at 0.5%, 1%, and 1.5% per month.

[10] MHT procedure asymptotic properties are derived for $T \to \infty$. With the exception of the Bonferroni and Holm methods, we observe a small gain in efficiency measured by the standard deviations of FDP and FNDP by increasing the number of hypotheses under consideration. See for example Chordia et al. (2019).

[11] We do not consider the case in which the sample is unbalanced. Please refer to Andrikogiannopoulou and Papakonstantinou (2019) and Harvey and Liu (2019a), who both consider the effect of heterogeneity in the length of returns time series.

[12] Please note that FDR is defined as the expected value of FDP. We tabulate the average FDP across 1,000 simulations. We refer to this quantity as the realized FDR.

[13] One might wonder how a researcher can chose the parameters that define the FDP control in RSW, Prob(FDP$>\gamma) < \alpha$. γ is the analog of δ in the BH or BY procedures that controls FDR, so it often set to 5%. The choice of α is less obvious, as it controls the tail behavior of the distribution of FDPs. Genovese and Wasserman (2006) offered an intuitive way to think about the combined effect of γ and α. They showed that a threshold that guarantees FDP control also produces an FDR control such that FDR $< \gamma + (1-\gamma) \times \alpha$. In other words, suppose that $\gamma = 5\%$ and $\alpha = 10\%$, the FDP control guarantees that the FDR is lower than 14.5% in that particular application. The logic behind this FDP-FDR correspondence is as follows. Suppose that FDP control holds with equalities (i.e., Prob(FDP $\geq \gamma) \leq \alpha$). In 90% of the cases the FDP is equal to or lower than 5%, and in the remaining 10% of cases is higher and, in fact, as high as 100%. Let us consider this extreme scenario in which case the FDR = E[FDP] = 5% × 90%

Table 2: Impact of Sample Size on MHT procedures

T	Bonf	Holm	BH	BY	RSW		T	Bonf	Holm	BH	BY	RSW	
					5%	10%						5%	10%
	E[t-statistic Thresholds]							std[t-statistic Thresholds]					
250	4.42	4.41	2.87	3.62	3.31	3.18	250	0.00	0.00	0.03	0.04	0.06	0.05
500	4.42	4.40	2.79	3.46	3.03	2.97	500	0.00	0.00	0.00	0.01	0.02	0.01
1,000	4.42	4.39	2.79	3.45	2.97	2.92	1,000	0.00	0.00	0.00	0.00	0.01	0.01
	E[Rejections](%)							std[Rejections](%)					
250	2.22	2.23	8.31	5.32	6.57	7.06	250	0.56	0.57	0.64	0.89	0.87	0.81
500	7.82	7.87	10.38	9.61	10.05	10.13	500	0.61	0.61	0.12	0.20	0.11	0.11
1,000	10.01	10.01	10.48	10.06	10.28	10.32	1,000	0.01	0.01	0.10	0.03	0.08	0.09
	E[FDP](%)							std[FDP](%)					
250	0.11	0.11	4.92	0.74	1.56	2.23	250	0.31	0.31	1.17	0.56	0.79	0.85
500	0.02	0.02	4.55	0.50	2.19	2.67	500	0.06	0.06	0.97	0.28	0.74	0.77
1,000	0.01	0.01	4.55	0.54	2.69	3.12	1,000	0.05	0.05	0.91	0.32	0.77	0.84
	E[FNDP](%)							std[FNDP](%)					
250	75.60	75.45	21.01	47.20	35.35	30.97	250	5.53	5.57	5.88	8.58	8.43	7.75
500	22.04	21.53	0.90	4.39	1.70	1.43	500	5.17	5.16	0.62	1.94	0.97	0.84
1,000	0.17	0.16	0.00	0.00	0.00	0.00	1,000	0.21	0.20	0.00	0.03	0.00	0.00

Note: The table reports averages and standard deviations for thresholds, rejections rates, false discovery proportion (FDP = FP/P) and false nondiscovery proportion (FNDP = FN/N) for simulations where the length of the sample size, T, varies between 250 and 1,000 monthly observations. We tabulate results for the Bonferroni and Holm procedures that control FWER at $\alpha = 5\%$ level, Benjamini-Hochberg and Benjamini-Yekuteli that control FDR at $\delta = 5\%$, and the Romano-Shaihk-Wolf procedure that control the probability that FDP is larger than $\alpha = 5\%$ at $\gamma = 5\%$ or 10% level respectively. We conduct 1,000 simulations of 5,000 portfolios.

The underlying mechanism is the same in both situations: increasing the sample size or increasing the magnitude of alphas improves the precision with which alphas are estimated. As a consequence, the procedures are more easily able to separate hypotheses that should be rejected from those that should not be, with the effect being stronger for procedures that are more adaptive (i.e., BH and RSW — see Chordia et al., 2019). In practical terms, with the exception of the Bonferroni method, which only depends on the number of tests, more power leads to lower t-statistic thresholds. For example, the t-statistic threshold of BY decreases from 3.62 to 3.45 going from 250 to 1,000 observations (see Table 2). The change is more dramatic, 4.14 to 3.45, when the non-zero alpha increases from 0.5% to 1.5% per month (see Table 3). Note that when power is very high, the variation in t-statistic thresholds across simulations becomes very small (i.e., standard deviation equal to or smaller than 0.1).

Changes in sample size and alpha only marginally affect the ability of the procedures to avoid false rejections; with the exception of RSW, FDRs tend to slightly decrease. For example, the FDR of BH decreases from 4.92% to 4.55% when sample size increases from 250 to 1,000 because longer

+ 100% × 10% = 14.5%. If the FDP control holds with strict inequalities, it is then guaranteed that FDR < 14.5%. As α increases, the implicit FDR also increases. A common choice of α is either 5% or 10%.

Table 3: Impact of signal to noise ratio on MHT procedures

alpha	Bonf	Holm	BH	BY	RSW		alpha	Bonf	Holm	BH	BY	RSW	
					5%	10%						5%	10%
	E[Thresholds t-statistic]							std[Thresholds t-statistic]					
0.5	4.42	4.42	3.18	4.14	4.00	3.74	0.5	0.00	0.00	0.05	0.09	0.16	0.16
1.0	4.42	4.40	2.79	3.46	3.03	2.97	1.0	0.00	0.00	0.00	0.01	0.02	0.01
1.5	4.42	4.39	2.79	3.45	3.03	2.96	1.5	0.00	0.00	0.00	0.00	0.01	0.01
	E[Rejections] (%)							std[Rejections] (%)					
0.5	0.37	0.37	2.98	0.77	1.10	1.55	0.5	0.08	0.08	0.50	0.24	0.49	0.51
1.0	7.75	7.78	10.41	9.62	10.08	10.15	1.0	0.48	0.46	0.12	0.18	0.12	0.11
1.5	10.02	10.02	10.49	10.06	10.23	10.29	1.5	0.01	0.01	0.10	0.03	0.07	0.08
	E[FDP] (%)							std[FDP] (%)					
0.5	0.27	0.27	4.88	0.70	0.94	1.54	0.5	1.32	1.32	1.76	1.49	1.51	1.57
1.0	0.01	0.01	4.74	0.58	2.34	2.84	1.0	0.06	0.06	0.98	0.35	0.73	0.79
1.5	0.01	0.01	4.68	0.55	2.28	2.81	1.5	0.05	0.05	0.92	0.33	0.67	0.75
	E[FNDP] (%)							std[FNDP] (%)					
0.5	96.36	96.35	71.62	93.30	91.31	86.63	0.5	0.99	1.00	4.70	2.40	4.16	5.20
1.0	22.23	21.73	0.88	4.37	1.62	1.37	1.0	4.78	4.75	0.57	1.72	0.88	0.78
1.5	0.05	0.04	0.00	0.00	0.00	0.00	1.5	0.09	0.09	0.00	0.01	0.00	0.00

Note: The table reports averages and standard deviations for thresholds, rejections rates, false discovery proportion (FDP = FP/P) and false nondiscovery proportion (FNDP = FN/N) for simulations where we vary the magnitude of the portfolios abnormal returns (alpha = α_p) from 0.5 to 1.5 percent per month. This is equivalent to increasing the signal to noise ratio. We tabulate results for the Bonferroni and Holm procedures that control FWER at $\alpha = 5\%$ level, Benjamini-Hochberg and Benjamini-Yekuteli that control FDR at $\delta = 5\%$, and the Romano-Shaihk-Wolf procedure that control the probability that FDP is larger than $\alpha = 5\%$ at $\gamma = 5\%$ or 10% level respectively. We conduct 1,000 simulations of 5,000 portfolios for 500 months.

time-series reduce sampling noise and thus reduce the chance that a zero-alpha is estimated to be significantly positive or negative in any sample. An increase in the magnitude of the non-zero-alpha has a smaller effect on the average FDP than it has on the standard deviation, which more than halves for all procedure when alpha increases from 0.5% to 1.5%. Different from the other methods, and especially when alphas increase substantially (from 0.5% to 1.5% per month) an increase in power increases the FDR of the RSW procedure. For example, when significance is set to 5% the average FDP increases from 0.94% to 2.38% when alphas increase from 0.5% to 1.5% per month. The change in FDP is on par with a very large change in the threshold that decreases from 4.00 to 3.03. Note, however, that the control in RSW, Prob(FDP>5%), also decreases from 1.2% to 0.2%, which is consistent with the large decrease in the standard deviation of FDP.

As shown in both Tables 2 and 3, the biggest impact of increasing power is on the FNDP for which we observe a decrease in both average and standard deviation when moving from low to high power for all procedures. For example, the FNDP of BH decreases from 21.0% to essentially zero when the sample size increases from 250 to 1,000 monthly observations. Our results about the large effect of sample size on power are directly related to the finding of Andrikogiannopoulou and

Papakonstantinou (2019), who highlighted the shortcomings of the FDR procedure implemented by Barras et al. (2010). Similarly, Harvey and Liu (2019b) focused on the signal-to-noise ratio.

4.2.2 Correlation

Test correlations may have a substantial impact on the performance of multiple testing methods. We propose a simple simulation framework to illustrate the impact of test correlations. Importantly, we make our framework generic to finance applications so interested readers can replicate our framework to perform simulation studies on their own data.

For alphas of the data-generating process and factor F, we follow our setup described in the previous section. We inject 5% of strategies with an α_p of 1% and 5% of strategies with an α_p of -1%, and $F \sim \mathcal{N}(0.6\%, 4.3\%^2)$. To study test correlations, we introduce another factor G that affects all portfolios in the cross-section and is independent of F. We assume that the loading $\beta_{p,G}$ is randomly and normally distributed in the cross section with mean μ_{β_G} and standard deviation of σ_{β_G}. We fix $G \sim \mathcal{N}(0, 2\%^2)$. We can think of G as an unobserved factor that induces correlation in the simple residuals (and the alphas). Finally, similar to our previous setup, we also add idiosyncratic risk that is independent of both F and G to each portfolio. Thus the return of a portfolio can be characterized by the following equation:

$$r_{p,t} = \alpha_p + \beta_{p,F} \times F_t + \beta_{p,G} \times G_t + \epsilon_{p,t},$$

with correlation among portfolio returns stemming from two sources: exposure to the observed factor F_t and exposure to the unobserved factor G_t. The second term is useful in producing a random correlation structure in the unobservable part of the portfolio returns. The correlation between two portfolios will therefore be

$$corr(r_{p,t}, r_{q,t}) = corr(\beta_{p,F} \times F_t + \beta_{p,G} \times G_t \ , \ \beta_{q,F} \times F_t + \beta_{q,G} \times G_t).$$

For compatibility with our previous simulation setup on test power, we fix the total amount of unobservable risk as

$$Var(\beta_{p,G} \times G_t + \epsilon_{p,t}) = (\sigma^2_{\beta_G} + \mu^2_{\beta_G})\sigma^2_G + \sigma^2_\epsilon = 2.5\%^2.$$

Table 4 shows our simulation results. Several patterns emerge from the table. Focusing on BH and BY, the false discovery rate in general decreases when correlations are on average high. For example, when the average correlation increases from $\mu_{b_G} = 0$ to $\mu_{b_G} = 0.7$, the expected FDP goes down from around 4.6% to about 3.3% for BH. Correspondingly, the expected false non-discovery rate increases. On the other hand, dispersion in correlation also seems to have a significant impact on error rates. For example, when $\mu_{b_G} = 0.7$ and the dispersion in correlation increases from $\sigma_{b_G} = 0.1$ to $\sigma_{b_G} = 0.8$, the expected FDP goes down from 4.21% to 3.01%.

Table 4 also shows the substantial variation in the variance of the FDP, which is oftentimes not taken into account when evaluating the performance of multiple testing methods. An increase in the dispersion of correlations substantially increases the standard deviation of the FDP. For instance, at $\mu_{b_G} = 0$, and when σ_{b_G} increases from 0.1 to 0.7, the standard deviation for FDP increases from 0.95% to 6.66%. This large variation in FDP implies that although the mean FDP may be well bounded below the significance level across repeated samples, the probability for the realized FDP for a particular sample to exceed the significance level is high. As a result, if uncertainty in FDP is a concern, it may be necessary to adjust the multiple testing threshold to reduce this uncertainty, which usually implies an increase in the threshold. For example, as shown in Table 4, methods with more stringent thresholds (e.g., Bonferroni and Holm) have a much smaller variation in FDP.

Overall, we have established the sensitivity of the performance of multiple testing methods to test correlations, as generated by residual correlations in our simulation setup. Importantly, our simulation framework is generic with the latent factor assumption particularly relevant to financial applications. Researchers can follow our setup to evaluate multiple testing methods when applied to their data. In particular, they can run a first-stage regression analysis to single out the residuals. Next, they can perform a factor analysis of these residuals to extract the common factors that affect the dependence in the residuals. Lastly, simulations can be run based on the estimated residual factor structure to study the impact of test correlations on multiple testing methods.

Table 4: Impact of Correlation on MHT procedures

μ_{β_G}	σ_{β_G}	correlation			Bonf	Holm	BH	BY	RSW		Bonf	Holm	BH	BY	RSW	
		mean	25^{th}	75^{th}					5%	10%					5%	10%
					E[t-statistic Thresholds]						std[t-statistic Thresholds]					
0.0	0.0	0.00	-0.03	0.03	4.42	4.40	2.79	3.46	3.03	2.97	0.00	0.00	0.00	0.01	0.02	0.01
0.0	0.8	0.00	-0.18	0.18	4.42	4.40	2.80	3.48	3.15	2.97	0.00	0.00	0.03	0.01	0.06	0.05
0.0	0.1	0.00	-0.03	0.03	4.42	4.40	2.79	3.46	3.04	2.97	0.00	0.00	0.00	0.00	0.02	0.02
0.5	0.8	0.11	-0.14	0.40	4.42	4.40	2.81	3.48	3.13	2.93	0.00	0.00	0.04	0.01	0.07	0.06
0.5	0.1	0.15	0.11	0.20	4.42	4.40	2.79	3.46	3.17	3.03	0.00	0.00	0.01	0.01	0.05	0.04
0.7	0.8	0.27	-0.07	0.68	4.42	4.40	2.81	3.49	3.05	2.83	0.00	0.00	0.07	0.01	0.10	0.11
0.7	0.1	0.35	0.31	0.39	4.42	4.40	2.79	3.47	3.24	3.04	0.00	0.00	0.03	0.01	0.08	0.06
					E[Rejections] (%)						std[Rejections] (%)					
0.0	0.0	0.00	-0.03	0.03	7.82	7.87	10.38	9.61	10.05	10.13	0.61	0.61	0.12	0.20	0.11	0.11
0.0	0.8	0.00	-0.18	0.18	7.53	7.56	10.17	9.15	9.55	9.83	0.34	0.31	1.15	0.26	0.25	0.58
0.0	0.1	0.00	-0.03	0.03	7.72	7.72	10.39	9.62	10.04	10.12	0.27	0.30	0.12	0.17	0.12	0.11
0.5	0.8	0.11	-0.14	0.40	7.87	7.89	10.15	8.93	9.39	9.80	0.33	0.33	1.96	0.27	0.39	1.18
0.5	0.1	0.15	0.11	0.20	7.90	7.96	10.42	9.56	9.90	10.07	0.88	0.90	0.48	0.26	0.15	0.21
0.7	0.8	0.27	-0.07	0.68	7.83	7.86	10.09	8.81	9.68	10.14	0.47	0.45	5.18	0.39	4.25	5.91
0.7	0.1	0.35	0.31	0.39	6.35	6.37	10.44	9.39	9.77	10.06	0.37	0.36	1.00	0.42	0.28	0.41
					E[FDP] (%)						std[FDP] (%)					
0.0	0.0	0.00	-0.03	0.03	0.02	0.02	4.55	0.50	2.19	2.67	0.06	0.06	0.97	0.28	0.74	0.77
0.0	0.8	0.00	-0.18	0.18	0.01	0.01	4.76	0.48	1.62	2.91	0.05	0.06	7.34	0.93	2.70	5.06
0.0	0.1	0.00	-0.03	0.03	0.01	0.02	4.62	0.56	2.14	2.67	0.06	0.06	0.98	0.35	0.69	0.77
0.5	0.8	0.11	-0.14	0.40	0.01	0.01	5.10	0.51	1.80	3.57	0.04	0.04	9.90	1.15	4.10	7.78
0.5	0.1	0.15	0.11	0.20	0.01	0.01	4.74	0.58	1.50	2.40	0.04	0.04	4.47	0.74	1.78	2.67
0.7	0.8	0.27	-0.07	0.68	0.01	0.01	3.01	0.43	1.78	2.89	0.04	0.06	9.10	2.41	8.36	9.11
0.7	0.1	0.35	0.31	0.39	0.01	0.01	4.21	0.37	1.00	2.01	0.05	0.08	7.22	0.90	2.45	4.25
					E[FNDP] (%)						std[FNDP] (%)					
0.0	0.0	0.00	-0.03	0.03	22.04	21.53	0.90	4.39	1.70	1.43	5.17	5.16	0.62	1.94	0.97	0.84
0.0	0.8	0.00	-0.18	0.18	21.27	20.97	3.90	8.76	6.09	4.87	3.33	3.33	2.37	2.99	3.00	2.66
0.0	0.1	0.00	-0.03	0.03	22.14	21.63	0.95	4.40	1.75	1.49	4.43	4.36	0.55	1.55	0.88	0.78
0.5	0.8	0.11	-0.14	0.40	21.59	21.36	5.50	10.72	7.86	6.38	2.67	2.70	2.96	3.21	3.29	3.08
0.5	0.1	0.15	0.11	0.20	22.08	21.57	0.98	4.71	2.50	1.79	5.35	5.35	0.92	2.88	1.94	1.47
0.7	0.8	0.27	-0.07	0.68	23.46	23.26	6.58	12.51	8.41	6.68	2.74	2.76	2.76	2.59	2.74	2.75
0.7	0.1	0.35	0.31	0.39	22.27	21.83	0.91	4.68	2.96	1.79	7.26	7.28	1.29	4.15	3.14	2.21

Note: The table reports averages and standard deviations for thresholds, rejections rates, false discovery proportion (FDP = FP/P) and false nondiscovery proportion (FNDP = FN/N) where we vary the parameters that define the correlation structure in the simulated data, μ_{β_G} and σ_{β_G}. We report average, top, and bottom quartiles of the distribution of pairwise correlations. We tabulate results for the Bonferroni and Holm procedures that control FWER at $\alpha = 5\%$ level, Benjamini-Hochberg and Benjamini-Yekuteli that control FDR at $\delta = 5\%$, and the Romano-Shaikh-Wolf procedure that control the probability that FDP is larger than $\alpha = 5\%$ at $\gamma = 5\%$ or 10% level respectively. We conduct 1,000 simulations of 5,000 portfolios over 500 months.

4.2.3 Publication bias

For simplicity, we refer to publication bias as every situation in which we only observe a subset of the tested hypotheses that are made public. Given that journals usually prefer to publish papers with "significant" results (Harvey, 2017), researchers are likely only to circulate (and submit) papers that contain some significant results. Thus, we can assume that publication bias truncates the distribution of observable t-statistics at some level, for example the traditional threshold of 1.96. In traditional hypothesis testing, publication bias has no effect on the researcher's inference. The t-statistic thresholds only depend on the confidence level desired by the researcher and are not a function of the data. Moreover, the truncation level is usually close to the statistical threshold accepted for significance, so the overall evaluation of one or many hypotheses remains unaltered.

In MHT, however, t-statistic thresholds are a function of the data. Therefore, the researcher's overall inference about the hypotheses under consideration is affected when only a portion of the tests is observable. In that sense, the bias induced by the publication process depends on two quantities: the proportion of true effects in the population (i.e., the proportion π of non-zero alpha strategies) and the cut-off (i.e., the minimum t-statistic required for a paper to be publishable).

We start by showing how different assumptions about the proportion of non-zero alphas affect the properties of MHT procedures. We give a brief intuition first. Let us first consider a hypothetical scenario in which the population only contains signals that generate long-short portfolios with non-zero alphas. In other words, only anomalies exist in this world. Should a researcher consider doing an MHT adjustment? The answer would appear to be "no". After all, every strategy has a true non-zero alpha, so the t-statistic cutoff can be set to any arbitrary low number. The researcher would instead greatly benefit from implementing an MHT adjustment when π is very low. In the extreme situation when the proportion of non-zero alphas is zero, some strategies would still have some conventionally significant alphas by luck, and MHT would likely remove most if not all of them. In other words, the benefit of implementing MHT is inversely related to the proportion of true effects in the population.

In Table 5 we report simulation results for different scenarios in which we vary π from 5% to 80%. To simplify further, we simulate portfolios for 500 months under the assumption of zero correlation.

Table 5 shows that thresholds are decreasing with π (with the obvious exception of the Bonferroni method, which only depends on the number of tests). The intuition is similar to the one presented in Section 4.2.1 in which we discussed power: If there are more non-zero alphas, it is easier to separate them from zero alphas, and hence the thresholds decrease. What happens to FDP and FNDP depends on the procedure. While all procedures decrease FNDP when π is larger, RSW does so more pronouncedly. That, however, comes at the expense of slightly increasing the average FDP. The increase is internally consistent as the procedure maintains the probabilistic control (of

Table 5: Impact of π on MHT procedures

π	Bonf	Holm	BH	BY	RSW		π	Bonf	Holm	BH	BY	RSW	
					5%	10%						5%	10%
	E[t-statistic Thresholds]							std[t-statistic Thresholds]					
5	4.26	4.42	3.03	3.68	3.31	3.23	5	0.00	0.01	0.02	0.03	0.02	0.02
10	4.26	4.41	2.81	3.48	3.03	2.97	10	0.00	0.01	0.01	0.02	0.02	0.02
20	4.26	4.38	2.57	3.28	2.72	2.66	20	0.00	0.00	0.01	0.01	0.02	0.01
30	4.26	4.36	2.43	3.16	2.50	2.44	30	0.00	0.00	0.01	0.01	0.02	0.01
40	4.26	4.33	2.32	3.07	2.30	2.25	40	0.00	0.01	0.01	0.01	0.01	0.01
60	4.26	4.27	2.17	2.95	1.90	1.85	60	0.00	0.01	0.01	0.01	0.01	0.01
80	4.26	4.17	2.06	2.86	1.21	1.16	80	0.00	0.02	0.01	0.01	0.03	0.03
	E[Rejections] (%)							std[Rejections] (%)					
5	3.85	3.85	5.18	4.72	4.94	4.99	5	0.34	0.34	0.09	0.13	0.09	0.09
10	7.84	7.92	10.40	9.62	10.06	10.14	10	0.46	0.46	0.12	0.19	0.12	0.11
20	15.58	15.70	20.77	19.51	20.40	20.53	20	1.06	1.01	0.15	0.24	0.14	0.15
30	24.07	24.53	31.02	29.46	30.78	30.95	30	0.94	0.97	0.17	0.28	0.17	0.18
40	31.68	32.18	41.16	39.44	41.21	41.40	40	2.24	2.01	0.16	0.29	0.18	0.19
60	46.87	49.71	61.17	59.36	62.31	62.55	60	2.85	2.82	0.17	0.33	0.25	0.25
80	67.21	68.82	80.73	79.28	84.56	84.92	80	1.05	1.07	0.14	0.36	0.43	0.46
	E[FDP] (%)							std[FDP] (%)					
5	0.02	0.02	5.03	0.56	1.93	2.48	5	0.10	0.10	1.37	0.52	0.85	0.93
10	0.01	0.01	4.68	0.54	2.23	2.78	10	0.06	0.06	0.91	0.34	0.67	0.71
20	0.01	0.01	4.17	0.48	2.68	3.19	20	0.03	0.03	0.62	0.22	0.51	0.55
30	0.00	0.00	3.59	0.40	2.91	3.38	30	0.02	0.02	0.48	0.16	0.44	0.49
40	0.00	0.00	3.04	0.35	3.13	3.56	40	0.01	0.01	0.36	0.13	0.39	0.43
60	0.00	0.00	2.04	0.23	3.75	4.13	60	0.01	0.01	0.25	0.09	0.39	0.38
80	0.00	0.00	1.00	0.12	5.40	5.80	80	0.00	0.00	0.15	0.05	0.48	0.51
	E[FNDP] (%)							std[FNDP] (%)					
5	22.55	22.29	1.56	6.48	3.20	2.70	5	5.16	5.17	0.98	2.56	1.61	1.44
10	22.04	21.53	0.90	4.40	1.66	1.39	10	4.94	4.92	0.59	1.83	0.89	0.78
20	22.20	21.13	0.49	2.92	0.75	0.65	20	4.66	4.60	0.31	1.16	0.42	0.38
30	21.93	20.28	0.31	2.18	0.39	0.33	30	4.71	4.63	0.21	0.92	0.25	0.23
40	22.13	19.77	0.22	1.75	0.21	0.18	40	4.52	4.36	0.14	0.72	0.14	0.12
60	22.00	17.95	0.13	1.29	0.06	0.05	60	4.51	4.28	0.10	0.55	0.05	0.05
80	21.74	15.32	0.09	1.01	0.00	0.00	80	4.63	4.26	0.07	0.45	0.01	0.01

Note: The table reports averages and standard deviations for thresholds, rejections rates, false discovery proportion (FDP = FP/P) and false nondiscovery proportion (FNDP = FN/N) where we vary the proportion of strategies with non-zero alpha (risk-adjusted returns), π, from 5 to 80%. We tabulate results for the Bonferroni and Holm procedures that control FWER at $\alpha = 5\%$ level, Benjamini-Hochberg and Benjamini-Yekuteli that control FDR at $\delta = 5\%$, and the Romano-Shaikh-Wolf procedure that control the probability that FDP is larger than $\alpha = 5\%$ at $\gamma = 5\%$ or 10% level respectively. We conduct 1,000 simulations of 5,000 portfolios for 500 months.

false positives) within a tolerance of 5% or 10%. When it comes to BH and BY, an increase in π is associated not only with a decrease in thresholds, but also in FDP. When $\pi = 80\%$, the t-statistic thresholds are very close to the conventional 1.96 for BH, and are even below that for RSW.[14]

[14] Remember that RSW is based on a bootstrap, so when π is really large, the strategies that determine the threshold are those that have zero alpha, but that in sample have relatively high t-statistics. In a bootstrap, however, their t-statistics will be pretty low and, hence, the low overall thresholds for the entire procedure.

We now examine how the application of MHT is impacted by truncation in the sample. The proportion of non-zero alphas that appeared in the truncated sample is increasing with the cutoff (i.e., the t-statistic level that determines which tests are observable). For example in our simulation, even when $\pi = 10\%$, a sample that is truncated at a t-statistic level of 1.64 (i.e., 10% significance under THT) will be composed of 72% of strategies that have non-zero alphas. At a truncation level of 1.96, the proportion of non-zero alphas is higher than 85%. When $\pi = 30\%$, the proportion of non-zero alphas in the truncated sample is around 90%. Note, however, that the truncated sample is not equivalent to a sample that has the equivalent proportion of non-zero alphas, but for which we can observe all tests. Because the sample is truncated, the tests that are observable and correspond to zero alphas have large t-statistics. At the same time, we do not observe any with very low t-statistics. Thus, adaptive MHT procedures will be "lead" into thinking that the data-generating process has more non-zero alphas than it actually does and hence produce a relatively low t-statistic threshold. Because the proportion of zero-alphas is still considerable, the FDRs will be very large, and in fact much larger than the tolerance. In this sense, procedures that are more adaptive (i.e., BH and RSW) are more valuable when there is no bias in the distribution, exactly because they adapt to the data, but they are more vulnerable to biases (when the distribution is truncated).

In Table 6 we report average t-statistic thresholds and FDPs for two cutoff levels (i.e., 1.64 and 1.96) and proportions of non-zero alphas in the population from 5% to 30%. We observe that, despite relatively low t-statistic thresholds, the realized FDRs of the adaptive procedures are quite large. Even BY, which is very conservative, can have a realized FDR above the chosen tolerance; for example, when $\pi = 5\%$ and the sample is truncated at 1.95, the average FDP is 7.62%, when it should be below 5%. The problem is more acute for BH and RSW. For example, when $\pi = 30\%$ and the cutoff point is 1.64 BH reaches the lowest realized FDR at 9.57%, which is almost twice as large as the tolerance (i.e., 5%).

The cutoff at which the sample is truncated also has an effect on thresholds and FDPs, but such effect is not the same for all procedures. For example, realized FDRs for BH and BY are higher when the cutoff is 1.96 than when it is 1.64, for each level of π. For RSW the exact opposite is true.

We provide an easier visualization of the publication bias in Table 7. For different levels of π and cut-offs, we tabulate the ratio of average t-statistic thresholds and FDP in the truncated samples relative to the counterparts from the entire distributions. A ratio equal to one indicates no bias. A ratio lower (higher) than one means the threshold from the truncated sample is lower (higher) than that from the entire distribution.

The bias in t-statistic thresholds decreases as π increases. When $\pi = 1\%$, the most adaptive procedures (BH and RSW) produce thresholds from the truncated distributions that are close to 50−60% of the corresponding thresholds from the full distribution. The ratios are close to 80%

Table 6: MHT in a truncated sample

π	π_{tr}	Bonf	Holm	BH	BY	RSW		Bonf	Holm	BH	BY	RSW	
					5%	10%						5%	10%
		t-statistic cut-off: 1.64											
		E[t-statistic Thresholds]						E[FDP] (%)					
5	53	3.81	3.92	2.24	3.05	1.59	1.35	0.15	0.22	32.90	4.83	56.75	65.73
10	72	3.88	3.91	2.12	2.92	1.34	1.35	0.05	0.09	23.83	3.33	47.72	47.72
20	81	3.97	3.90	2.04	2.84	1.57	1.55	0.02	0.05	14.38	1.91	28.84	28.84
30	87	4.04	3.90	2.01	2.82	1.26	1.19	0.01	0.03	9.57	1.21	19.24	19.24
		t-statistic cut-off: 1.96											
		E[t-statistic Thresholds]						E[FDP] (%)					
5	70	3.71	3.75	1.96	2.89	1.46	1.52	0.23	0.41	48.95	7.62	48.98	48.98
10	85	3.81	3.75	1.96	2.82	1.57	1.44	0.07	0.18	31.28	4.54	31.19	31.19
20	90	3.93	3.75	1.96	2.78	0.85	0.68	0.02	0.09	16.78	2.30	16.71	16.71
30	93	4.01	3.76	1.96	2.78	0.61	1.53	0.01	0.05	10.57	1.35	10.56	10.56

Note: The table reports the average t-statistic thresholds and the false discovery proportion (FDP = FP/P) computed in the truncated sample of alphas with a t-statistic larger than 1.64 and 1.96, respectively. π refers to the proportion of non-zero alphas (risk-adjusted returns) in the population. π_{tr} refers to the proportion of non-zero alphas in the truncated sample. We tabulate results for the Bonferroni and Holm procedures that control FWER at $\alpha = 5\%$ level, Benjamini-Hochberg and Benjamini-Yekuteli that control FDR at $\delta = 5\%$, and the Romano-Shaikh-Wolf procedure that control the probability that FDP is larger than $\alpha = 5\%$ at $\gamma = 5\%$ or 10% level respectively. We conduct 1,000 simulations of 5,000 portfolios for 500 months and vary the proportion of strategies with non-zero alphas, π, from 5 to 30%.

for the remaining procedures: Bonferroni, Holm, and BY. With the exception of RSW, the ratios increase (i.e., bias decreases) when π gets larger. In the case of RSW, the bias disappears when π is large enough (around 45% — data not tabulated), but the pattern is non-monotonic: it first increases and then decreases.

The bias in realized FDRs measures the failure of the procedure to control false positives in the truncated samples; remember the realized FDRs are always below the control when there is no truncation. The low thresholds produce large FDRs because they reject all the hypotheses that correspond to a zero alpha, but have a t-statistic larger than 1.64 or 1.96. For example, when $\pi = 5\%$ and the cutoff is 1.96, the FDR of BY in the truncated distribution is 7.6%, whereas it is only 0.5% if all calculations are based in the entire distribution.

In summary, the biases produced by considering only observable hypotheses and ignoring the file drawer problem are substantial, especially in the (likely) scenario there are only a small proportion of non-zero alphas compared to the large number of strategies tested.

Thus, in situations for which the sample of tests is truncated, a researcher must attempt to recover the missing part of the distribution before applying a MHT procedure to compute t-thresholds and determine the size of the false discovery problem. Harvey et al. (2016), Chen (2019), and Chordia et al. (2019) strove to do this. The results we present in this paper help explain why the findings in these cited papers appear to support diametrically opposite conclusions. Harvey et al. (2016)

Table 7: Publication bias and MHT procedures

π	Bonf	Holm	BH	BY	RSW		π	Bonf	Holm	BH	BY	RSW	
					5%	10%						5%	10%
t-statistic cut-off: 1.64													
	E[$\mathcal{T}_{\text{truncated}}$] / E[$\mathcal{T}_{\text{entire distribution}}$]							E[FDP$_{\text{truncated}}$] / E[FDP$_{\text{entire distribution}}$]					
5	0.87	0.85	0.65	0.78	0.44	0.47	5	12.20	19.72	9.73	13.65	25.41	19.76
10	0.89	0.85	0.70	0.81	0.52	0.49	10	5.34	13.30	6.69	8.34	13.96	11.21
20	0.92	0.86	0.76	0.85	0.31	0.25	20	3.81	12.89	4.03	4.80	6.24	5.23
30	0.94	0.86	0.81	0.88	0.24	0.63	30	3.42	12.04	2.94	3.40	3.63	3.13
40	0.95	0.87	0.85	0.90	0.85	0.87	40	2.19	12.95	2.31	2.51	2.25	1.98
t-statistic cut-off: 1.96													
	E[$\mathcal{T}_{\text{truncated}}$] / E[$\mathcal{T}_{\text{entire distribution}}$]							E[FDP$_{\text{truncated}}$] / E[FDP$_{\text{entire distribution}}$]					
5	0.89	0.89	0.74	0.83	0.48	0.42	5	8.12	10.31	6.54	8.65	29.45	26.52
10	0.91	0.89	0.76	0.84	0.44	0.46	10	4.03	6.77	5.09	6.11	21.36	17.15
20	0.93	0.89	0.79	0.87	0.58	0.58	20	3.32	7.52	3.45	3.98	10.77	9.03
30	0.95	0.89	0.83	0.89	0.51	0.49	30	3.10	6.99	2.66	3.04	6.62	5.70
40	0.96	0.90	0.86	0.91	0.40	0.37	40	2.13	7.92	2.16	2.34	4.22	3.71

Note: The table reports the ratio of average t-statistic thresholds (i.e., E[$\mathcal{T}$]) and FDP computed on the truncated sample of alphas with a t-statistic larger than 1.64 and 1.96, respectively, and the entire distribution. π refers to the proportion of non-zero alphas in the population. π_{tr} refers to the proportion of non-zero alphas in the truncated sample. We tabulate results for the Bonferroni and Holm procedures that control FWER at $\alpha = 5\%$ level, Benjamini-Hochberg and Benjamini-Yekuteli that control FDR at $\delta = 5\%$, and the Romano-Shaikh-Wolf procedure that control the probability that FDP is larger than $\alpha = 5\%$ at $\gamma = 5\%$ or 10% level respectively. We conduct 1,000 simulations of 5,000 portfolios for 500 months and vary the proportion of strategies with non-zero alpha, π, from 5 to 40%.

and Chordia et al. (2019) produce very large thresholds from MHT procedures. Chen (2019) find almost no evidence that an adjustment, relative to THT, is needed.

The last paper estimates several models in which trading strategies are generated, but only some are observable. The findings show very low proportions of falsely rejected hypotheses (i.e., at conventional levels) among the trading strategies studied. For example, using data on 156 equal-weighted portfolios of published strategies, Chen estimates many versions of a model in which results that have larger t-statistics are more likely to be publishable. Each model differs in the assumptions Chen makes about the correlation structure in the data, the functional form of the selection criteria that determines which strategies are publishable, or the distribution of expected returns. To obtain distributions for the parameters of interest Chen bootstraps the 156 equal-weighted portfolios' returns. He estimates an average π (which in his model is $1 - p_0$) of 55% with very large dispersion around the average and concludes that thresholds do not necessarily need to be increased from conventional levels (i.e., 1.96 at 5% significance). His argument seems to be that identification of π from published strategies is so difficult that it is impossible to definitely conclude π is small enough to need an MHT adjustment.

Two aspects of Chen deserve attention. First, the argument that learning about π from published strategies is difficult is worth considering. For example, Chen and Zimmermann (2019) found that a model of publication bias implies a very modest attenuation of 12% in the average returns of 156 trading strategies. Using an out of sample test McLean and Pontiff (2016) found a much larger post discovery attenuation in return (i.e., 58% in total, although 32% appears to be related to post-publication trading activity) among 97 trading strategies.[15] Second, the estimation of a unconditional (relative to publication) large proportion of non-zero alphas seems economically implausible. It would seem that other observable data could be used to guide a researcher's prior about what π might be. For example, the finance literature suggests that professional investors rarely produce abnormally profitable returns (see for example, Fama and French, 2010; Barras et al., 2010; Busse et al., 2010; Ferson and Chen, 2019). Linnainmaa and Roberts (2018) found that most of the published strategies that they consider do not show any abnormal return in the period of time before that studied by the original authors. In turn, that suggests that the rate of profitable strategies should be relatively low (i.e., small π), in which case the publication bias will be very large, and large MHT adjustments will be necessary to control the number of false discoveries. As Table 5 shows, a very large π implies a relatively low threshold. At $\pi = 60\%$, thresholds produce by BH, which is the method Chen (2019) adopts, are already getting close to 2.0, even before taking into account correlations and any bias induced by the publication process. Moreover, in the truncated samples, such a low threshold will produced very large FDRs, as we show in Table 6.

The truncation model Chen (2019) uses is more complicated than what we present here and follows the model presented by Harvey et al. (2016). In comparison to Harvey et al. (2016), who mainly use BY to show how a researcher can threshold a truncated sample of published discoveries, Chen applies his estimation to a sample of equal-weighted returns from a time series of actual trading strategies known to have significant returns. Even if his estimation procedure is meant to retrieve the truncated part of the distribution, it only learns from the biased sample. There are alternatives. For example, Chordia et al. (2019) estimated a small proportion of informative signals (i.e., about 2%) by mixing information about published strategies with a large set of strategies generated from the data and by allowing many distributional assumptions about strategies returns and signals realizations. In contrast to Chen, the estimation in Chordia et al. (2019) is allowed to "learn" not only from published strategies, but also from the distribution of all value-weighted long-short portfolios that a researcher could construct from the data. Because the latter contains mostly strategies that are not statistically exceptional, it provides a more precise description of the distribution of zero-alpha strategies than what could be inferred from a set of published strategies that are largely significant.

[15] Jacobs and Muller (2019) examined 241 trading strategies in international markets and found no attenuation bias.

5. Optimizing the false discovery control

The simulation exercise we present here, which investigates the impact of publication bias, highlights the importance of π in driving the MHT outcome. When π is small and most strategies in the population are false (as is the case in many prominent problems in finance such as factor research or fund management), the application of MHT is crucial to control the false discovery rate. In contrast, when π is believed to be large — so a large number of strategies possess non-zero but likely small alphas — we do not need MHT and conventional single-test THT thresholds or even ones that are lower may suffice. In this sense, what matters for the application of MHT is the belief in π.

Harvey and Liu (2019a) have proposed a new framework that allows a researcher to incorporate a prior belief about π. Their framework features a two-step bootstrap procedure. For a given π, the first round of bootstrap perturbs the data to isolate π of strategies that are likely true. Conditional on these isolated strategies, a second round of bootstrapping is run to evaluate different MHT techniques. They then average across the iterations of the first round of bootstrap to measure the overall performance of various MHT techniques.

When researchers have a strong prior on π, Harvey and Liu have provided an intuitive approach to incorporate such priors into the MHT procedure. When some uncertainty exists around the prior, they show that the outcome is robust to the prior specification, as long as the specified prior is not very far away from the true value of π.

Harvey and Liu have contributed to the MHT literature in four aspects. First, their technique is able to calibrate the performance of various MHT techniques when applied to a particular data set. Second, their technique can be used directly (i.e., as an alternative to any of the MHT methods highlighted in this paper). Their method helps achieve the desired false discovery rate for a particular data set and hence maximizes test power. In contrast, existing MHT approaches, although theoretically sound, may generate an error rate that is far from the desired level. For example, the BH approach may exceed the pre-specified significance level because its assumption of test independence may be violated, whereas the BY approach may be too conservative because of the way it controls the false discovery rate under arbitrary dependence assumption.

The bottom line is that we do not know the relative performance of different approaches for a given data set. An important advantage of the Harvey and Liu (2019a) method is that the optimal thresholds are determined conditional on the particular data set. Third, in most research in finance, Type I and Type II errors are treated symmetrically. For many important decisions, however, there are differential costs of Type I and Type II errors. The Harvey and Liu framework allows the researcher to specify the relative costs of these two types of errors (inter-error cost variation) within the bootstrap framework. Finally, it is straightforward to modify their technique to allow for intra-error variation; that is, not all Type I errors have the same cost. In most statistical frameworks, however, errors are counted in a binary fashion (e.g., FDR) irrespective of the magnitude of the error.

6. Which technique should you choose?

From a practical point of view the choice of a specific multiple testing adjustment is not easy or obvious. It is much different than switching on a technique to control for, say, possible heteroskedasicity in regression errors. This is due to the fact that the procedures presented above differ in fundamental ways because they implement different controls (i.e., FWER versus FDR versus FDP), and rely on different assumptions about dependence in the data/tests. A practical guideline that includes a recommendation for any situation is challenging, if not impossible.

However, we can provide some general suggestions. The number of tests, N, plays a critical role in the choice of the type of control. When N is very large, FWER control becomes very strict. For example, if a research is only testing 10 hypotheses, it is reasonable to use a procedure that controls for FWER. The MHT adjustment in this case allows at most one false positive. Since the researcher is testing 10 hypotheses, one false rejection would lead to an FDP of 10%, which is not too strict. There are many procedures that control FWER. The choice among those will depend on which assumptions the researcher is willing to accept. Something as simple as Bonferroni, though, might be enough to rule out hypotheses that are only marginally significant.

Another important consideration is the dependence in the data. Situations in which tests are clearly not independent should be addressed in the context of a MHT technique that relies on resampling, and that can therefore guarantee the desired control of false rejections under arbitrary dependence in the data. For example, the StepM procedure of Romano and Wolf (2005) and the FDP-StepM of Romano and Wolf (2007) would be a natural candidate to control FWER and FDP, respectively.

In the case where researchers are able to analyze completely different samples, as for example when the same hypothesis is tested in different countries or sample periods, the adoption of FDR control, which is generally speaking less strict than FDP, might be adequate. Remember that FDR is the expected value of FDP, which one could read, for example, as an expectation across different samples.

Finally, if there is uncertainty about the effectiveness of any of the above multiple testing adjustments, researchers should consider implementing Harvey and Liu's (2019a) method. They present a double bootstrap approach that delivers a set Type I error rate (e.g., 5%) in multiple testing applications. Their method is data dependent so the cutoffs will differ conditional on the particular data at hand. In finding the cutoff that delivers a 5% FDR, they show that the test power is optimized. Harvey and Liu's (2019a) method also allows the researcher to inject their prior (or a range of priors) on the proportion of hypotheses that are true. This prior information impacts the thresholds in an intuitive way — a higher prior on the proportion of positives in the sample leads to a lower threshold. Finally, in contrast to other methods discussed in this paper that focus on Type I errors, Harvey and Liu's method allows the researcher to develop a decision framework that assigns differential costs of Type I and Type II errors.

7. Conclusions

For many years, the multiple testing problem was largely ignored in both finance and economics research. As such, a large number of papers were published with discoveries that barely cleared the two-standard error hurdle. As such, it is likely that many of those published findings are false because the test statistics did not control for multiple testing. A hurdle of two standard errors leads to a massive false discovery rate when so many variables are tried. Moreover, as computer power increases and new datasets become available that provide even more possible variables, the multiple hypothesis problem will get worse.

The two sub-areas of finance most impacted by multiple testing are the fund evaluation and the factor/anomaly literature. With thousands of fund managers and investment products, it is now obvious to most researchers that what appears to be outperformance could naturally occur purely by luck. Both of these subfields have embraced the necessity of allowing for multiple testing corrections to determine appropriate statistical thresholds.

There are, however, many choices of techniques ranging from a simple Bonferroni adjustment to a more complicated bootstrapping procedure. Which one should a researcher choose? Indeed, each of the procedures we detail raises the threshold from the traditional hypothesis single test. Hence, any adjustment reduces the false positive rate compared to a single test by increasing the threshold for significance. But how do these procedures perform across different scenarios? If the single-test threshold produced a false discovery rate of 50% when we target 5%, it is useful to know if the multiple testing correction reduces the false discovery rate to, say, 40%, 7%, or 0.7%. In the first case, the improvement is economically meaningful, but 40% is far from our target. The 7% case is very close to the target and likely acceptable. But, in the 0.7% case the procedure is far too harsh and will lead to many missed discoveries (a high false omission rate).

We provide a guide to an array of different procedures and provide applications for each one (as well as the computer code to generate the corrections). Our empirical work focuses on simulations based on six different methods: Bonferonni (1936), Holm (1979), Benjamini and Hochberg (1995), Benjamini and Yekuteli (2001), and two versions of Romano and Wolf (2007). Our simulations show the impact of power (manifested by changing the number of observations or varying the signal-to-noise ratio) on the different techniques. We then show how the relative performance of the different techniques varies with the correlation among the strategies. Finally, we detail the impact of publication bias (we observe only the published strategies, not everything that has been tried).

Finally, we highlight the method of Harvey and Liu (2019a) that allows the researcher to optimize the Type I error threshold at the desired level depending on the particular data set. This technique has the advantage of hitting the desired level of false discovery and hence maximizing test power. That is, if the FDR is only 0.7% as in the preceding example, power will be much lower than a technique that delivers an FDR of 5%. Implementing this procedure comes at some cost because the research needs to inject a prior — or a range of priors — regarding the overall proportion of true strategies. Fortunately, in most finance applications this proportion is small. It is not easy to outperform the market and it is certainly not easy to discover a new risk factor.

While much of our discussion is focused on some of the most egregious research problems for which adjustments are absolutely essential, multiple testing has a much broader footprint with examples in almost every category of finance research. In corporate finance, researchers test a host of variables to explain variation in payout, leverage, cash holdings, and profitability — and each paper has a multiple hypothesis testing problem. Behavioral research has its own multiple testing problem. Consider the proliferation of biases that are proposed in the behavioral economics literature: the number is analogous to the so-called factor zoo in asset pricing. A number of studies use a single natural experiment to explain many effects which, again, leads to a multiple testing problem. Fortunately, on-going research is making progress in each of these areas.

Most importantly, multiple testing touches almost every published empirical paper. Any regression that shows various specifications in attempting to explain a single variable suffers from a multiple hypothesis testing problem. It is time we both recognize this and improve our inference by deploying multiple testing corrections. Our paper is, hopefully, a step towards improving inference.

References

Andrews, D., and G. Soares. 2010. Inference for parameters defined by moment inequalities using generalized moment selection. *Econometrica* 78:119–157.

Andrikogiannopoulou, A., and F. Papakonstantinou. 2019. Reassessing false discoveries in mutual fund performance: Skill, luck, or lack of power? *Journal of Finance* Forthcoming.

Bajgrowicz, P., and O. Scaillet. 2012. Technical trading revisited: False discoveries, persistence tests, and transaction costs. *Journal of Financial Economics* 106:473–491.

Bajgrowicz, P., O. Scaillet, and A. Treccani. 2016. Jumps in high-frequency data: Spurious detections, dynamics, and news. *Management Science* 62:2198–2217.

Barber, R. F., and A. Ramdas. 2017. The p-filter: multilayer false discovery rate control for grouped hypotheses. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79:1247–1268.

Barillas, F., and J. Shanken. 2018. Comparing asset pricing models. *Journal of Finance* 73:715–754.

Barras, L., O. Scaillet, and R. Wermers. 2010. False discoveries in mutual fund performance: measuring luck in estimated alphas. *Journal of Finance* 65:179–216.

———. 2019. Reassessing false discoveries in mutual fund performance: Skill, luck, or lack of power? A reply. Working paper.

Basu, P., T. Cai, K. Das, and W. Sun. 2018. Weighted false discovery rate control in large-scale multiple testing. *Journal of the American Statistical Association* 113:1172–1183.

Benjamini, Y., and M. Bogomolov. 2014. Selective inference on multiple families of hypotheses. *Journal of the Royal Statistical Society, Series B* 76:297–318.

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

———. 1997. Multiple hypotheses testing with weights. *Scandinavian Journal of Statistics* 24:407–418.

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

Blake, D., T. Caulfield, C. Ioannidis, and I. Tonks. 2017. New evidence on mutual fund performance: A comparison of alternative bootstrap methods. *Journal of Financial and Quantitative Analysis* 52:1279–1299.

Bonferroni, C. E. 1936. *Teoria Statistica Delle Classi e Calcolo Delle Probabilità*. Libreria Internazionale Seeber.

Boudoukh, J., R. Michaely, M. Richardson, and M. R. Roberts. 2007. On the importance of measuring payout yield: Implications for empirical asset pricing. *The Journal of Finance* 62:877–915.

Bryzgalova, S., J. Hung, and C. Julliard. 2019. Bayesian solutions for the factor zoo. Working Paper.

Busse, J. A., A. Goyal, and S. Wahal. 2010. Performance and persistence in institutional investment management. *Journal of Finance* 65:765–790.

Chen, A. 2019. Do t-stat hurdles need to be raised? Identification of publication bias in the cross-section of stock returns. Working Paper.

Chen, A., and T. Zimmermann. 2019. Publication bias and the cross-section of stock returns. *Review of Asset Pricing Studies* Forthcoming.

Chernozhukov, V., D. Chetverikov, and K. Kato. 2019. Inference on causal and structural parameters using many moment inequalities. *Review of Economic Studies* Forthcoming.

Chib, S., X. Zeng, and L. Zhao. 2019. On comparing asset pricing models. *Journal of Finance* Forthcoming.

Chordia, T., A. Goyal, and A. Saretto. 2019. Anomalies and false rejections. *Review of Financial Studies* Forthcoming.

Davidson, H., M. Samadi, M. Ringgenberg, and I. Werner. 2019. Reusing natural experiments. SMU working paper.

Dudoit, S., M. J. van der Laan, and K. S. Pollard. 2004. Multiple testing. Part I. Single-step procedures for control of general Type I error rates. *Statistical Applications in Genetics and Molecular Biology* 3:1–69.

Efron, B., and R. Tibshirani. 2002. Empirical Bayes methods and false discovery rates for microarrays. *Genetic Epidemiology* 23:70–86.

Engelberg, J., D. R. McLean, J. Pontiff, and M. C. Ringgenberg. 2019. Are cross-sectional predictors good market-level predictors? Working Paper.

Fama, E. F., and K. R. French. 1993. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics* 33:3–56.

———. 2010. Luck versus skill in the cross-section of mutual fund returns. *The Journal of Finance* 65:1915–1947.

Farcomeni, A. 2008. A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Statistical Methods in Medical Research* 17:347–388.

Ferson, W., and Y. Chen. 2019. How many good and bad funds are there, really? In C. Lee (ed.), *Handbook of Financial Economics, Mathematics, Statistics and Technology*, chap. 108. World Scientific Press.

Ferson, W. E., and C. R. Harvey. 1999. Conditioning variables and the cross-section of stock returns. *Journal of Finance* 54:1325–1360.

Genovese, C. R., and L. Wasserman. 2006. Exceedance control of the false discovery proportion. *Journal of the American Statistical Association* 101:1408–1417.

Giglio, S., L. Yuan, and D. Xiu. 2018. Thousands of alpha tests. Working paper.

Green, J., J. R. M. Hand, and X. F. Zhang. 2017. The characteristics that provide independent information about average U.S. monthly stock returns. *Review of Financial Studies* 30:4389–4436.

Groenborg, N., A. Lunde, A. Timmermann, and R. Wermers. 2010. Picking funds with confidence. *Journal of Financial Economics* Forthcoming.

Hansen, P., A. Lunde, and J. Nason. 2011. The model confidence set. *Econometrica* 79:453–497.

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

Harvey, C. R. 2017. The scientific outlook in financial economics. *Journal of Finance* 72:1399–1440.

Harvey, C. R., and Y. Liu. 2019a. False (and missed) discoveries in financial economics. *Journal of Finance* Forthcoming.

———. 2019b. Lucky factors. Working Paper.

Harvey, C. R., Y. Liu, N. Polson, and J. Xu. 2019. Revisiting semi-strong market efficiency. Working Paper.

Harvey, C. R., Y. Liu, and H. Zhu. 2016. ... And the cross-section of expected returns. *Review of Financial Studies* 29:5–68.

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

Jacobs, H., and S. Muller. 2019. Anomalies across the globe: Once public, no longer existent? *Journal of Financial Economics* Forthcoming.

Jefferys, W. H., and J. O. Berger. 1992. Ockham's razor and Bayesian analysis. *American Scientist* 80:64–72.

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

Kosowski, R., A. Timmermann, R. Wermers, and H. White. 2006. Can mutual fund "stars" really pick stocks? New evidence from a bootstrap analysis. *The Journal of Finance* 65:2551–2595.

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

Linnainmaa, J. T., and M. Roberts. 2018. The history of the cross section of stock returns. *Review of Financial Studies* 31:2606–2649.

List, J., A. Shaikh, and Y. Xu. 2019. Multiple hypothesis testing in experimental economics. *Experimental Economics* Forthcoming.

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

Martin, I., and S. Nagel. 2019. Market efficiency in the age of big data. Working paper.

McLean, R. D., and J. Pontiff. 2016. Does academic research destroy stock return predictability? *Journal of Finance* 71:5–32.

Mitton, T. 2019. Corporate finance p-hacking. Working Paper.

Patton, A. J., and A. Timmermann. 2010. Monotonicity in asset returns: New tests with applications to the term structure, the CAPM, and portfolio sorts. *Journal of Financial Economics* 98:605 – 625.

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

Qu, R., A. Timmermann, and Y. Zhu. 2019. Do any economists have superior forecasting skills? UCSD working paper.

Ramdas, A., R. F. Barber, M. J. Wainwright, and M. I. Jordan. 2019. A unified treatment of multiple testing with prior knowledge using the p-filter. *Annals of Statistics* forthcoming.

Romano, J., A. Shaikh, and M. Wolf. 2014. A practical two-step method for testing moment inequalities. *Econometrica* 82:1979–2002.

Romano, J. P., and A. Shaikh. 2006. Stepup procedures for control of generalizations of the familywise error rate. *Annals of Statistics* 34:1850–1873.

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

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

———. 2007. Control of generalized error rates in multiple testing. *Annals of Statistics* 35:1378–1408.

———. 2010. Balanced control of generalized error rates. *Annals of Statistics* 38:598–633.

Scott, J. G. 2009. Nonparametric Bayesian multiple testing for longitudinal performance stratification. *Annals of Applied Statistics* 3:16551674.

Scott, J. G., and J. O. Berger. 2006. An exploration of aspects of Bayesian multiple testing. *Journal of Statistic Planning and Inference* 136:21442162.

Shanken, J. 1990. Intertemporal asset pricing: An Empirical Investigation. *Journal of Econometrics* 45:99–120.

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

———. 2003. The positive false discovery rate: A bayesian interpretation and the q-value. *Annals of Statistics* 6:2013–2035.

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

Sullivan, R., A. Timmermann, and H. White. 1999. Data-snooping, technical trading rule performance, and the bootstrap. *Journal of Finance* 54:1647–1691.

Wermers, R. 1999. Mutual fund herding and the impact on stock prices. *Journal of Finance* 55:581–622.

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

Yan, X. S., and L. Zheng. 2017. Fundamental analysis and the cross-section of stock returns: A data-mining approach. *Review of Financial Studies* 30:1382–1423.

A. Matlab code

We review here the basic structure of Matlab code that can be used to construct MHT adjusted thresholds. We specify a few variables as general inputs as follows: tstat indicates a vector of p-statistics with dimension M×1, where each element corresponds to the t-statistic of a certain hypothesis. boot indicates a matrix of dimension M×B that contains B bootstrapped t-statistics. We fix statistical significance at level alpha, the FDP tolerance at the level gamma0, and FDR tolerance at the level delta. It might be necessary to transform t-statistics into p-values.

```
% transform tstat into pvalue
ind = (tstat≤0);
pvalue = NaN(length(Data,1);
pvalue(∼∼ind) = 2*normcdf(tstat(∼∼ind),0,1);
pvalue(∼∼ind) = 2*(1-normcdf(tstat(∼ind),0,1));
```

Please note that note have chosen to transform t-statistics into p-values by means of the normal distribution. Readers should be careful about situations where this transformation could be inaccurate, and might want to use another distribution. Moreover, adjustments are necessary when considering one-sided tests.

A.1 Bonferroni

The implementation of the Bonferroni adjustment is the simplest among the MHT procedures. The procedure only has one step. The threshold is just adjusted scaling the desired statistical level by number of tests.

```
function [Tthreshold, NumberRejected] = Bonf(pvalue,alpha)
% determine how many hypotheses are in the set
N = length(pvalue);
% determine adjusted threshold
threshold = alpha / M;
% find significant hypotheses
NumberRejected= sum(pvalue ≤ threshold);
% convert pvalue into tstat
Tthreshold = -norminv(threshold/2,0,1);
```

A.2 Holm

The Holm procedure is a stepwise procedure that starts by checking the most significant hypothesis against an adjusted threshold equal to that imposed by Bonferroni. As it moves down to less significant hypotheses, the threshold becomes smaller.

```
function [Tthreshold, NumberRejected] = Holm(pvalue,alpha)
% determine how many hypotheses are in the set
M = length(pvalue);
% sort tests from most to least significant
```

```
sorted_pvalue = sort(pvalue);
% compute the adjusted sorted pvalue that can be compared directly with alpha
threshold = alpha /(M+1-(1:M));
% find significant hypotheses
% j_star represents the position of the first hypothesis that is rejected in the set
of ordered hypotheses
% starting from the most significant and working down
for j = 1:M
   if sorted_pvalue(j) > threshold(j)
      j_star = j-1;
      break
   end
   if ~isempty(j_star)
      NumberRejected = j_star;
      % convert pvalue into tstat
      Tthreshold = -norminv(sorted_pvalue(j_star)/2,0,1);
   else
      NumberRejected = M;
      % convert pvalue into tstat
      Tthreshold = -norminv(max(sorted_pvalue)/2,0,1);
   end
end
```

A.3 k-StepM

The k-StepM procedure controls the k-FWER. It is a stepwise procedure that that compares test obtained from the data to the distribution of the $(1-\alpha)$ percentile of the bootstrap distribution of k^{th}-max.

```
function [Tthreshold, NumberRejected] = kStepM(tstat,tstat_boot,alpha,k,Nk)
% Nk is a parameter to limit the size of the Consider set
SigLevel = 100*(1-alpha);
tstat = sort(tstat,'descend');
% 1st iteration
[M,B] = size(tstat_boot);
IndexAll = (1:M)';
Z1temp = sort(tstat_boot,'descend');
CrtVal = prctile(Z1temp(k,:),SigLevel);
IndexRejected = find(tstat≥CrtVal);
NumberRejected = length(IndexRejected);
% Main loop
while 1
   if NumberRejected(end) == 0 || sum(NumberRejected) ≥ M-1
      break
   end
   IndexRemaining = setdiff(IndexAll,IndexRejected);
   % the original Romano and Wolf procedure considers (k-1) combinations
   % from the set of hypotheses that have been eliminated (the reject set).
   % if the number of hypotheses is very large, loop is numerically cumbersome.
   % hence we insert a max controlled by the free parameter Nk;
   K = nchoosek(IndexRejected,min(k-1,Nk);
   Z2temp = size(K,1); cj = NaN(Z2temp,1);
   for i = 1:Z2temp
      I = K(i,:);
      Consider = [I'; IndexRemaining];
      Z1temp = sort(tstat_boot(Consider,:),'descend');
      cj(i) = prctile(Z1temp(k,:),SigLevel);
   end
   CrtVal = [CrtVal; max(cj)];
   IndexRejected = find(tstat≥CrtVal(end));
   IndexRejected_new = find(tstat≥CrtVal(end) & tstat<CrtVal(end-1));
   NumberRejected = [NumberRejected; length(IndexRejected_new)];
end
% convert pvalue into tstat
Tthreshold = CrtVal(end);
```

A.4 FDP-StepM

The FDP-StepM procedure of Romano, Shaikh, and Wolf extends the use of the k-StepM to control for probability that FDP is below a certain threshold. The procedure keeps increasing the k order of the k-StepM until it reaches the desired FDP control.

```
function [Tthreshold,NumberRejected] = FDPStepM(tstat,tstat_boot,alpha,gamma0,Nk)
k = 1;
[CrtVal,NumberRejected] = kStepM(tstat_data,tstat_boot,alpha,k,Nk);
% The following is a little heuristic to speed up the process
k = floor(sum(NumberRejected)*Gam);
while 1
   if sum(NumberRejected) < (k/gamma0-1)
      break
   end
   % here we insert a little heuristic to speed up the process, so that k increases
   % to the lower bound of the constraint, as opposed to increasing by 1.
   k = max(floor(sum(NumberRejected)*gamma0),k+1);
   [Tthreshold ,NumberRejected] = kStepM(tstat_data,tstat_boot,alpha,k,Nk);
end
```

A.5 Benjamini and Hochberg

Benjamini and Hockberg provide a stepwise procedure that controls FDR at percentage δ. The procedure starts from the least significant hypothesis and moves its way up until it finds one significant hypothesis.

```
function [Tthreshold,NumberRejected] = BH(pvalue,delta)
M = length(pvalue);
% sort tests from most to least significant
sorted_pvalue = sort(pvalue);
% compute the threshold for each sorted hypothesis
threshold = delta .*((1:M)/M);
% find significant hypotheses
% j_star represents the position of the first hypothesis that is rejected in the set
of ordered hypotheses
% starting from the lest significant and working up
for j = M:-1:1
   if sorted_pvalue(j) < threshold(j)
      j_star = j;
      break
   end
end
if ~isempty(j_star)
   NumberRejected = j_star;
   % convert pvalue into tstat
   Tthreshold = -norminv(sorted_pvalue(j_star)/2,0,1);
else
   NumberRejected = M;
   % convert pvalue into tstat
   Tthreshold = -norminv(max(sorted_pvalue)/2,0,1);
end
```

A.6 Benjamini and Yekuteli

Benjamini and Yekuteli provide a stepwise procedure that controls FDR at percentage δ. The procedure starts from the least significant hypothesis and moves its way up until it finds one significant hypothesis.

```
function [Tthreshold,NumberRejected] = BY(pvalue,delta)
M = length(pvalue);
% sort tests from most to least significant
sorted_pvalue = sort(pvalue);
% compute the threshold for each sorted hypothesis
CM = log(M) + 0.5;
threshold = delta .*((1:M)/(M*CM));
```

```
% find significant hypotheses
% j_star represents the position of the first hypothesis that is rejected in the set
of ordered hypotheses
% starting from the lest significant and working up
for j = M:-1:1
   if sorted_pvalue(j) < threshold(j)
      j_star = j;
      break
   end
end
if ~isempty(j_star)
NumberRejected = j_star;
% convert pvalue into tstat
   Tthreshold = -norminv(sorted_pvalue(j_star)/2,0,1);
else
   NumberRejected = M;
   % convert pvalue into tstat
   Tthreshold = -norminv(max(sorted_pvalue)/2,0,1);
end
```