Anomalies and Multiple Hypothesis Testing: Evidence from Two Million Trading Strategies

Tarun Chordia **Amit Goyal** **Alessio Saretto***

May 2018

Abstract

We construct a large laboratory of over two million trading strategies, which we obtain by data-mining the two most commonly used datasets in finance (i.e., CRSP and COMPUSTAT). We use this very large sample for three purposes. First, we evaluate the properties of multiple hypothesis testing methods when applied to financial data. We find that only adaptive methods should be employed in finance applications (i.e., FDP-SetpM). Second, we provide an optimal thresholding for applications that evaluate trading strategies. Third, we quantify the proportion of false discoveries due to the failure to take into account testing a multitude of hypotheses. Our estimates for the proportion of lucky discoveries is over 90%, which is considerably larger than previously reported.

*Tarun Chordia is from Emory University, Amit Goyal is from Swisss Finance Institute at the University of Lausanne, and Alessio Saretto is from University of Texas at Dallas. We thank Hank Bessembinder, Ing-Haw Cheng, Campbell Harvey, Ohad Kadan, Yan Liu, Kalle Rinne, Olivier Scaillet, Peter Westfall, Michael Wolf, Lu Zhang and seminar participants at the 6th Luxembourg Asset Management Summit, TAU Finance Conference 2017, 2017 FRA conference, Inquire Europe Autumn Seminar 2017, Lone Star Conference, Caltech, Case Western, CUHK, Texas Tech, Tinbergen Institute Amsterdam, University of San Diego, UNSW, and the University of Texas at Dallas for helpful discussions. All errors are our own.

An increasingly large body of literature studies the profitability of trading strategies based on signals obtained from publicly available information. Researchers are currently tracking a number of strategies well in excess of 300 and new papers keep adding to that list.[1] A recent paper by Yan and Zheng (2017) finds that a very large number of trading signals (more than two thousand) "exhibit genuine predictive power."

In his presidential address, Harvey (2017) questions the performance of these strategies due to a number of possible problems with the way in which these strategies are discovered and evaluated. One particular problem is related to the fact that test procedures are often evaluated in isolation. Harvey and Liu (2014, 2015) and Harvey, Liu, and Zhu (2015) advocate the use of multiple hypothesis testing (MHT) as one tool that finance researcher should employ to limit the number of false discoveries. The essential idea is that, when studying the entire distribution of trading strategies, one has to account for the fact that some strategies' performance will appear exceptional by luck, thus leading to some false rejections of the null hypothesis of no outperformance.

Applying MHT, Harvey, Liu, and Zhu (2015) advocate the use of a threshold of three for t-statistics (as opposed to the traditional 1.96) and find that the abnormal performance of as many as 50% of the 316 strategies they consider could be due to luck. Applying a threshold of three, Hou, Xue, and Zhang (2017) also find that as many as 58% of the strategies that survive their replication effort could be due to failure to account for MHT.

Harvey, Liu, and Zhu (2015) note that optimal thresholding for MHT requires the researcher to have a good sense of the entire distribution of tests, not only those that end up in the right tail (i.e., the ones for which the null was rejected and made it to a circulated paper). Yet, virtually all papers on this subject look at the sample of only strategies contained in published or working papers. In contrast, we construct a very large sample of approximately two and half million trading strategies obtained by data-mining the information contained in the two most commonly used datasets in finance, viz. CRSP and COMPUSTAT.

[1] For example, Harvey, Liu, and Zhu (2015) examine 316 strategies, Green, Hand, and Zhang (2013) study over 300, and Hou, Xue, and Zhang (2017) study 447.

Armed with our very large sample of strategies, we evaluate the ability of MHT to conform to finance data. Our paper has three main objectives. First, using a data-informed Monte Carlo analysis, we highlight the properties of several MHT methods. Second, we provide an optimized threshold of t-statistics for finance applications that rely on CRSP and COMPUSTAT data. Third, we compute the proportion of false discoveries within our sample.

The main feature of our study that enables us to achieve our objectives is our procedure for generating trading signals. Our strategy yields a comprehensive set of trading strategies, some of which have been studied and published as well as some that have been studied but not published (likely because they do not surpass traditionally accepted statistical hurdles), and those that have yet to be studied (likely because their economic foundation is not immediately justifiable or simply because researchers have not thought about them). By considering strategies without filtering on their ex-post significance, and by not relying on published anomalies, our large-scale exercise allows us to avoid data snooping. Our results are robust to the inclusion of small stocks, various sample definitions, and the application of different methods and factor models to adjust for the risk of the strategies.

Our paper relies extensively on MHT for conducting tests. The statistics and economics literature has proposed a variety of ways for controlling the number of null hypothesis that are erroneously rejected in testing multiple hypotheses. We consider the three most common approaches: family-wise error rate (FWER), false discovery ratio (FDR), and false discovery proportion (FDP). FWER controls the probability of making more than one false rejection, FDP controls the probability of a user-specified proportion of false rejections in a given sample, while FDR controls the expected (across different samples) proportion of false rejections. We concentrate on a total of five methods: two that control FWER (i.e., Bonferroni and Holm); two that control FDR (i.e., BH and BHY); and one that controls FDP (i.e., FDP-StepM).[2]

[2]The BH procedure is from Benjamini and Hochberg (1995). The BHY procedure is a combination of Benjamini and Hochberg, and Benjamini and Yekutieli (2001). FDP-StepM was developed by Romano and

We first perform a Monte Carlo simulation that allows us to highlight the properties of the various MHT methods applied to financial data. It is important to remember that MHT methods were developed for applications in different fields. In genomics, for example, after adjusting for MHT it is not uncommon to reject only strategies with p-values of the order of 5×10^{-7} (see, The Wellcome Trust Case Control Consortium, 2007). This is not only dictated by the need to be conservative in the medical profession, but also by the fact that statistical relations are stronger in other fields than they are in finance and economics. Since the signal-to-noise ratio is probably very different in financial data, one might expect very different size and power properties for the various MHT tests.

Our Monte Carlo experiment produces a cross-section of strategies with features matching those of the actual empirical distribution. We conduct various experiments differing in the fraction of true alternatives, the strength of abnormal returns under the alternative distribution, correlation between strategies, number of strategies, and number of simulation. We find that, while all MHT methods have good size properties, they differ in their power and their ability to adapt to situations where the proportion of true alternatives might be high. In particular, we find that only one of the FDR methods (the BH procedure) and the FDP method (the FDP-StepM procedure) are reliable in terms of power properties; the other MHT methods have substantially low power (due to very high critical values that they impose on the data). We also find that the magnitude of critical value (and the corresponding rejection rates) are primarily dictated by the signal-to-noise ratio (magnitude of true alphas relative to return volatility) as well as the fraction of true rejections. The BH and FDP methods are adaptive (but the other MHT methods are not) in the sense that they produce lower critical values if the true number of rejections in the data is high. Thus, our Monte Carlo experiment suggests that when dealing with as much variability in the data as that contained in CRSP and COMPUSTAT, one should be selective about the choice of MHT methods.

Wolf (2007).

Our Monte Carlo experiment also allows us to address the concern about the use of our large sample of over two million strategies in real data. Of course, we do not mean to imply that researchers are looking at the set of strategies that we use in our experiment. In fact, the set of strategies that finance researchers look at (or will potentially look at) is much smaller. This raises the concern about whether there might be statistical biases introduced by our very large set of strategies. Out Monte Carlo experiment shows that, while the thresholds do increase mechanically for some MHT methods, the BH and the FDP methods do not lead to a mechanical increase in statistical thresholds. Therefore, even if researchers will not consider all the strategies that we study, our results are still of general interest since our experiment relies on the basic variability of information contained in the CRSP and COMPUSTAT datasets.

Armed with an understanding of the properties of MHT methods, we move on to analyze the actual data. We calculate two measures of risk-adjusted performance for each of our strategies. First, we construct a long-short portfolio based on the top and bottom decile of each signal's distribution. We then compute portfolio alphas using the Fama and French (2015) five factor model augmented with the Carhart (1997) momentum factor. Second, we calculate the Fama and MacBeth (1973, henceforth FM), coefficient for each signal following the methodology proposed by Brennan, Chordia, and Subrahmanyam (1998).

We find a relatively large discrepancy in thresholds and rejection rates across the five models and for different evaluation measures (i.e., alphas and FM coefficients). Following what we learn from the Monte Carlo simulation we focus on FDR-BH and FDP-StepM methods. Differently from the simulation, these two methods produce very different results in the actual data; FDR-BH rejects a higher proportions of nulls than does FDP-StepM.

Besides the conceptual distinction in what they are trying to control, the FDR and FDP methods also differ in their underlying assumptions. For our purposes, an important assumption is that of non-zero correlation between strategies. Trading strategies are not independent of each other, as there is cross-correlation in stock returns across different firms

and in the information used to construct the signals, not only across different firms but also within a particular firm (i.e., total assets and profitability are not independent). The FDR methods make strong assumptions about the correlation structure of strategies whereas the FDP method delivers statistical cutoffs that account for the cross-correlations present in the data. Therefore, we rely on the FDP method more heavily.

Imposing a tolerance of 5% of false discoveries (false discovery proportion) and a significance level of 5%, we find that the critical value for alpha t-statistic (t_α) is 3.79 while that for FM coefficient t-statistic (t_λ) is 3.12. While these critical values are quite a bit higher than the conventional levels, they are not far from the suggestion of Harvey, Liu, and Zhu (2015) to use a critical value of three. Our higher thresholds are due to our choice of a different MHT method, our sample of over two million strategies vis-à-vis 316 strategies in Harvey, Liu, and Zhu, and the fact that we fully account for dependence in the data. At these thresholds, 2.67% of strategies have significant alphas and 16.31% have significant FM coefficients. The smaller critical values for t_λ than those for t_α are due to the fact that the cross-strategy distribution of the former has longer tails (i.e., the standard deviation of the distribution of t_λ is equal to 1.93, while the standard deviation of t_α is 1.82).

Comparing the rejection rates obtained from MHT to the rejection rates obtained from classical single hypothesis testing (CHT), which rejects any hypothesis with a t-statistic higher than 1.96, gives a lower bound for the magnitude of false discoveries (i.e., MHT methods also allow for some false discoveries to happen). Under CHT we reject the null hypothesis in about 30% of the cases for both alpha and FM coefficient t-statistics.

We conclude that, in our experiment, the great majority of the discoveries (i.e., rejections of the null of no predictability) that are made by relying on CHT and without accounting for the very large number of strategies that are never made public, are very likely false. In the case of alphas, that percentage can be as high as $91\% (= 1 - 2.67/30.36)$, while the problem is less severe for FM coefficients, although it could still be as high as $59\% (= 1 - 16.31/39.33)$.

In order to gauge some consistency between performance measures we ask of a trading

signal to not only generate a high long-short portfolio alpha but also to explain the broader cross-section of returns in a regression setting. Eliminating strategies that have statistically significant t_α but insignificant t_λ, or vice-versa, drastically reduces the number of successful strategies to 806 (i.e., 0.04% of the total) under MHT and to 33,881 (i.e., 1.62% of the total) under CHT. The lower bound on the proportion of false discoveries remains very high, north of 95%. Notably, the very large proportion of lucky discoveries is constant across many ways of classifying the trading strategies according to their strength: average return, alpha, Sharpe ratio or information ratio.

The very high proportion of false discoveries (relative to classical hypothesis testing) is also persistent across many experimental and modeling choices: use of different factor models to adjust the strategies returns; different combinations of controls in the FM regression; inclusion of small stocks; and number of strategies that we consider. This robustness across many specification lends credibility to the idea that our results should generalize to different sets of trading strategies that are constructed using different datasets.

Our paper echoes the increasing skepticism about the validity of many research findings in a variety of fields. While the findings on the lack of replicability in medical research by Ioannidis (2005) are widely cited, the economics profession has also made an effort to tackle this problem. Leamer (1978, 1983) famously complains about specification searches in empirical research and asks researchers to take the 'con' out of econometrics. Dewald, Thursby, and Anderson (1986), McCullogh and Vinod (2003), and Chang and Li (2017) also report disappointing results from replication of economics papers. The use of replication in finance is less widespread with Hou, Xue, and Zhang (2017) being a notable recent exception.

Our paper also joins the list of the growing finance literature that studies the proliferation of discoveries of abnormally profitable trading strategies and/or pricing factors and its relation to data-snooping biases in finance. See Lo and MacKinlay (1990) and MacKinlay (1995) for early work emphasizing statistical biases in hypothesis testing. The question of whether the profitability of published strategies survives the test of time is studied in Schwert

(2003), Chordia, Subrahmanyam, and Tong (2014), McLean and Pontiff (2015), Linnainmaa and Roberts (2016), and Hou, Xue, and Zhang (2017). Towards the turn of the century, more formal statistical approaches were developed and applied to the problem of evaluating multiple strategies (see, for example, Sullivan, Timmermann, and White (1999), White (2000), and Romano and Wolf (2005)). The MHT approach has been more recently applied to financial settings in Barras, Scaillet, and Wermers (2010), Harvey, Liu, and Zhu (2015), and emphasized in the presidential address of Harvey (2017).

Our paper is also closely related to Yan and Zheng (2017). Both papers share the goal of evaluating a broader universe of strategies than just the published ones. Despite inevitable differences in sample construction etc., we find very similar results in our sample of two million strategies that Yan and Zheng find in their sample of around 18,000 strategies. In particular, using the same bootstrap experiment that Yand and Zheng use, we reject all percentiles of t-statistics below 40 and above 60. Such a large rejection rate suggests substantial miss-pricing in the market, a viewpoint adopted by Yan and Zheng. Our conclusions about market efficiency differ markedly from theirs for two main reasons. First is our use of formal statistical approaches to MHT rather than the heuristic-based bootstrapped approach. In fact, we show through simulation that bootstrap-based methods tend to substantially over-reject the null hypothesis while the size and power properties of MHT methods are markedly better. Second, we show that economic and statistical considerations play a large role in restricting the set of statistically significant strategies.

1 Data and trading strategies

Monthly returns and prices are obtained from CRSP. Annual accounting data come from the merged CRSP/COMPUSTAT files. We collect all items included in the balance sheet, the income statement, the cash-flow statement, and other miscellaneous items for the years 1972 to 2015. We choose 1972 as the beginning of our sample as it corresponds to the first

year of trading on Nasdaq that dramatically increased the number of stocks in the CRSP dataset. All our results are robust to beginning the sample in 1963, which is the first date on which the COMPUSTAT data are not affected by backfilling bias. Following convention, we set a six-month lag between the end of the fiscal year and the availability of accounting information.

We impose several filters on the data to obtain our basic sample. First, we include only common stocks with CRSP share codes of 10 or 11. Second, we require that data for each variable be available for at least 300 firms each month for at least 30 years during the sample period. Third, in FM (1973) regressions described later, we require that data be available for all independent variables (including the variable of interest) for at least 300 firms each month for at least 30 years during the sample period. Fourth, at portfolio formation at the end of June of each year (exact procedure described later), we require stocks to have a price higher than three dollars and market capitalization to be higher than the bottom twentieth percent of the NYSE capitalization. The last filter ensures that we eliminate micro-cap stocks alleviating concerns about transaction costs as well those about generalizability and relevance (Novy-Marx and Velikov (2016) and Hou, Xue, and Zhang (2017)).

There are 156 variables that clear our filters and can be used to develop trading signals. The list of these variables is provided in Appendix Table A1. We refer to these variables as *Levels*. We also construct *Growth rates* from one year to the next for these variables. Since it is common in the literature to construct ratios of different variables we also compute all possible combinations of ratios of two levels, denoted *Ratios of two*, and ratios of any two growth rates, denoted *Ratios of growth rates*. Finally, we also compute all possible combinations that can be expressed as a ratio between the difference of two variables to a third variable (i.e., $(x_1 - x_2)/x_3$). We refer to this last group as *Ratios of three*. We obtain a total of 2,385,778 possible signals.

We evaluate trading signals by estimating abnormal performance of the hedge portfolios using a factor model and by evaluating the ability of the signal in explaining the cross-section

of firms' abnormal returns.

1.1 Hedge portfolios

We sort firms into value-weighted deciles on June 30 of each year and rebalance these portfolios annually. The first portfolio formation is June 1973 and the last formation date is June 2015. We require a minimum of 30 stocks in each decile (300 stocks in total) in a month to consider that month as having a valid return. The signal is considered to have generated a valid portfolio if there are at least 360 months of valid returns. We consider long-short portfolios only. Thus, we compute a hedge portfolio return that is long in decile ten and short in decile one. Since we do not know ex-ante which of the two extreme portfolios has the largest average return, our hedge portfolios can have either positive or negative average returns. Obviously, it is always possible to obtain a positive average return for a hedge portfolio that has a negative average return by taking the opposite positions. For expositional convenience, we decide not to force average returns to be positive.

Our benchmark evaluation factor model is composed of the five factors in Fama and French (2015) plus the momentum factor. The five factors are the market, size, value, investment, and profitability factors. For each trading strategy, we run a time-series regression of the corresponding hedge portfolio returns on the six factors and obtain the alpha as well as its heteroskedasticty-adjusted t-statistic, t_α.

1.2 Fama-MacBeth regressions

Given that the alphas of the long-short portfolio effectively consider the efficacy of the strategy in only 20% of the sample, we also evaluate a signal's ability to predict returns in the cross-section of stocks using FM regressions. In particular, we evaluate the ability of the signal to explain stock returns by estimating the following cross-sectional regression each month:

$$R_{it} - \widehat{\beta}_i F_t = \lambda_{0t} + \lambda_{1t} X_{it-1} + \lambda_{2t} Z_{it-1} + e_{it}, \tag{1}$$

where X is the variable that represents the signal and Z's are control variables. We use the most commonly used control variables, namely size (i.e., the natural logarithm of the firm's market capitalization), natural logarithm of the book-to-market ratio, past one-month and 11-month return (skipping the most recent month), asset growth, and profitability ratio. Book-to-market is calculated following Fama and French (1992) while asset growth and profitability are calculated following Fama and French (2015). We risk-adjust the returns on the left-hand-side of equation (1) following Brennan, Chordia, and Subrahmanyam (1998). We use the same six-factor model used to calculate hedge portfolio alphas, and calculate full-sample betas $\widehat{\beta}$ for each stock. We require at least 60 months of valid returns to estimate the time-series regression. All right-hand-side variables are winsorized at the 1st and 99th percentile in FM regressions.

In estimating the cross-sectional regressions, we require a minimum of 300 stocks in a month. Finally, we require a minimum of 360 valid monthly cross-sectional estimates during the sample period to calculate a valid λ_1 coefficient for a signal. Thus, we calculate the FM coefficient λ_1 as well its heteroskedasticty-adjusted t-statistic (t_λ). Given that we require a valid beta for each stock and data on additional control variables, the data requirements for the FM regressions are slightly more stringent than those for portfolio formation.

2 Strategy performance

In this section we discuss the statistical properties of the signals and the trading strategy returns. We analyze raw returns and Sharpe ratios in Section 2.1, and abnormal returns and regression coefficients in Section 2.2.

2.1 Raw returns and Sharpe ratios

Table 1 reports summary statistics of raw returns on the hedge portfolios. We report cross-sectional means, medians, standard deviation, minimum, and maximum across portfolios.

These statistics are reported for the sample of all portfolios as well as the sub-sample of portfolios formed by the different trading signals (i.e., ratio of two, ratio of three, etc.). We report monthly average returns, t-statistics for returns, and monthly Sharpe ratios in Panel C. We also report the number and percentage of portfolios that cross specific thresholds.

We report these results for two different samples of stocks. Panel A uses only the stocks filtered by size and price as described in the previous section while Panel B uses all stocks. We report results for different kinds of strategies in each panel but Panel B does not include strategies 'Ratios of three.'[3]

Panel A shows that the cross-sectional mean and median average return of the portfolios are close to zero. The cross-sectional standard deviation of returns at 0.17% coupled with the fact that we have over two million portfolios implies that there are many portfolios with very large absolute returns. For example, there are 20,34 portfolios with absolute average monthly return greater than 0.5%. A large number of portfolios also have average return t-statistics that exceed conventional statistical significance levels. 129,689 (26,800) portfolios have average return t-statistics larger than 1.96 (2.57) (in absolute value); although, as expected, this represents only about 5% (1%) of the total number of portfolios. The economic importance of these portfolios is also very impressive as many portfolios have monthly Sharpe ratios higher than the historical market Sharpe ratio (approximately 0.116), with one portfolio having a Sharpe ratio higher than 0.232. These facts, while not perhaps surprising, are, nevertheless, interesting because they are obtained despite the stringent rules that affect the composition of our universe of stocks and signals (e.g., we eliminate stocks that are in the bottom quintile of the NYSE size distribution and that have prices below three dollars).

As is to be expected, the dispersion in the performance of strategies is largest in the subset of strategies 'Ratios of three.' The most profitable and statistically significant returns come from this group. The largest absolute average return is 1.07 per cent per month, and the largest absolute t-statistic is 5.41.

[3]The reason for excluding the large set of strategies for the sample of all stocks is computational time.

In order to examine the tails of the distribution, we list the top 50 strategies by average returns, return t-statistic, and Sharpe ratio in Tables **??**, **??**, and **??**, respectively. Most of the strategies in the tails are new and appear unrelated to existing anomalies (as it should be, since we control for the well-known anomalies in the factor models and regressions). For example, the most profitable strategy in terms of raw returns is the ratio of the difference between Capital surplus-share premium reserve (CAPS) and Cash and cash equivalent increase/decrease (CHECH) to advertising expense (XAD). This strategy has an average return of -1.07 per cent per month with a t-statistic of -4.40.

Panel B considers the sample of all stocks and reports results only for the subset of 12,239 strategies. Thus, while we use fewer strategies than those in Panel A, we use more stocks including small stocks. Fama and French (2008) show that anomalies are more pronounced amongst these stocks. Therefore, the net effect of these two opposing forces on the cross-sectional distribution of returns is not clear a priori. Looking at results in Panel B, we find that the extremes of returns and t-statistics are slightly lower in Panel B than those in Panel A. 1,197 (500) strategies have average return t-statistics higher than 1.96 (2.57); this represents 10% (4%) of the total number of strategies. Thus, this sample of strategies and stocks indicates more rejections of null for average return judged by conventional thresholds.

2.2 Abnormal returns and Fama-MacBeth regression coefficients

We next compute abnormal returns for our strategies using various factor models. We use five different factor models: CAPM, FF3, FF6, BS, and HXZ. CAPM one-factor model uses the market factor. FF3 is the Fama and French (2015) three-factor model. FF6 is the Fama and French (2015) five-factor model augmented with the momentum factor. BS is the Barillas and Shanken (2015) six-factor model. HXZ is the Hou, Xue and Zhang (2015) q-model augmented with the momentum factor. We calculate alphas corresponding to each factor model. We also run corresponding FM regressions where the betas on the left-hand-side and the additional control variables Z on the right-hand-side of equation (1) correspond

to the factor model used. Thus, we do not include any other control when risk adjusting stock returns CAPM. We include size and book-to-market when adjusting stock returns using the FF3 model. In all the other cases, we include size, book-to-market, profitability, asset growth, and one- and twelve-month lagged returns. We report t-statistics on alphas and FM coefficients in Table 2.

We also calculate these statistics for different sample of stocks and different subsample of strategies. Panels A and B show results for the subset of strategies that does not include 'Ratios of three' (there are 12,239 such strategies) while Panel C uses all strategies. Panels A and C use only stocks filtered by size and price, as described in Section 1, while Panel B includes all stocks (including stocks that have prices below three dollars).

The distribution of alpha and FM t-statistics in Table 2 reveals even more exceptional performance of strategies than that in raw returns of Table 1. For example, Panel A shows that 29.59% (16.44%) of FF6 model alpha t-statistics are higher than 1.96 (2.57). The fraction of FM t-statistics (in the lower half of Panel A) that cross these conventional statistical thresholds is also high. For example, 13.65% (6.12%) of FM coefficient t-statistics when risk-adjusting using FF6 model are higher than 1.96 (2.57).

Looking at alpha t-statistics, amongst different factor models, CAPM generates the fewest rejections while BS model generates the highest rejections of null of zero alpha. Rejection rates for FM t-statistics are relatively similar across different factor models except for CAPM that, in contrast to alpha t-statistics, generates the highest number of rejections. This is partly due to the fact that the right-hand-side control variables are the same (size, book-to-market, profitability, asset growth, and one- and twelve-month lagged returns) in FF6, BS, and HXZ specification, and even the FF3 specification uses two of these controls (size and book-to-market). The CAPM specification uses no control and generates the most rejections of the null for the FM coefficients.

Panel B expands the sample to all stocks. The extremes of t-statistics are higher in the sample of all stocks than those in the sample of stocks filtered by size and price. However,

the fraction of alpha t-statistics that cross the conventional thresholds is relatively similar to that in Panel A. The fraction of FM t-statistics that cross the threshold of 1.96 or 2.57 is slightly higher in Panel B than that in Panel A. Once again, CAPM generates the fewest (the highest) rejections of alpha (FM) t-statistics while BS model generates the highest rejections of alpha t-statistics. The relative similarity of results in Panels A and B suggests that our results later in the paper are not going to be overly sensitive to the sample of stocks that we use in our experiment.

In Panel C of Table 2 we report on all strategies. Unsurprisingly, the extremes of the distribution of t-statistics are higher than those in Panel A. At the same time, the fraction of alpha t-statistics that cross the thresholds is similar to all other panels. In contrast, the fraction of FM t-statistics that cross the thresholds is much higher in Panel C than that in the other two panels. This is partly due to higher cross-sectional standard deviation of these t-statistics in the sample of all strategies. This means that we expect to see differences later in the paper when looking at FM t-statistics across different sets of strategies.

Figure 1 depicts the histograms for the average return, six-factor alpha, the Sharpe ratio and the t-statistics for the average return, the six-factor alpha and the FM coefficients.[4] The distributions are generally centered around zero and seem normally distributed. The support for the distributions is consistent with the standard deviations in Tables 1 and 2. For instance, the Sharpe ratio has the lowest standard deviation of 0.04 while the FM coefficient t_λ has the highest standard deviation and this is reflected in the empirical distributions of Figure 1. Note that the distributions of t_α and t_λ are fat-tailed, consistent with the large number of rejections of the null in Panel A of Table 2.

It is not too surprising that, among a sample of around 12,00 strategies or in the sample of two and a half million strategies, we uncover *some* strategies in the tails that appear exceptional. However, the fact that we find almost 30% of the strategies to appear exceptional casts some doubt on rejection rates based on CHT. We start addressing these doubts in the

[4]Note that the x-axis is different for the different histograms.

next section, where we present the bootstrap approach of Yan and Zheng (2017).

3 Bootstrap approach

Recently, Yan and Zheng (2017) use a bootstrap approach to analyze multiple trading strategies generated through a procedure similar to ours. This approach, inspired by Kosowski, Timmermann, Wermers, and White (2006) and Fama and French (2010), relies on bootstrapping the cross-section of fund returns through time thereby preserving the cross-sectional dependence structure in strategy returns and ultimately their alpha estimates.

In this section we present the results of the bootstrap approach (i.e., YZ bootstrap, henceforth) to our set of trading strategies. We follow Yan and Zheng (2017) and construct bootstrap distributions of the alphas and their t-statistics under the null hypothesis that the alphas are zero. To bootstrap under the null, we first subtract the six-factor alpha (i.e., FF6) from the monthly portfolio returns. Each bootstrap run is a random sample (with replacement) of the alpha-adjusted returns and the factors over 522 months of the sample period 1972 to 2015. To preserve the cross-sectional correlation we apply the same bootstrap draw to all portfolios and to the factors. To preserve possible autocorrelation in the return structure, we construct the stationary bootstrap of Politis and Romano (1994) by drawing random blocks with an average length of six months. Due to the computational constraints imposed by the large scale of our exercise we limit the exercise to 1,000 bootstrap samples as opposed to the 10,000 runs implemented by Yan and Zheng.

For each bootstrap run we obtain the portfolio alphas and their t-statistics under the null of zero alpha. Following Yan and Zheng (2017), we then compare the percentiles of the t-statistics from the actual data sample to the corresponding percentiles in the bootstrap samples (i.e., the collection of x^{th} percentile from each bootstrap run). We focus on t-statistics rather than on the coefficients themselves because t-statistics control for the precision of coefficients and are advocated by, for example, Romano, Shaikh, and Wolf (2008).

Table 3 documents selected percentiles of the t-statistics from the actual distribution (Data). We also report the fraction of iterations where the bootstrapped percentile is bigger than the actual t-statistic for that percentile (% Boot) for percentiles above 50 and the fraction of iterations where the bootstrapped percentile is smaller than the actual t-statistic for that percentile for percentiles below 50. Yan and Zheng (2017) refer to this fraction as the p-value of the selected pereentile while Fama and French (2010) refer to this fraction as the likelihood.

Consider the 99th percentile. The actual alpha t_α from the data is 4.03 while the average (across iterations) bootstrap t_α under the null is 2.35 (not reported in the table). In the the collection of 99th percentiles from each bootstrap run, we do not find any bootstrapped t_α larger than 4.03. Similar observations apply to other percentiles implying that, relative to bootstrap distribution under the null of zero alpha, the extreme of the distributions of actual t-statistics in the data are atypical.

We conduct a similar experiment for FM coefficients (i.e., again using the FF6 factor to compute the risk-adjusted return, and all six controls on the right hand side of the regressions). In particular, for each signal variable we start by subtracting the average from the time-series of λ_{1t} coefficients from equation (1), thus obtaining a time-series of adjusted coefficients under the null of no explanatory power. We then bootstrap 1,000 times the time-series of pseudo coefficients and calculate the means and t-statistics for each bootstrap iteration. Finally, for each percentile of interest we collect the corresponding quantity from each bootstrap cross-sectional distribution of FM coefficients. We then compare the t_λ based on the data to the corresponding bootstrap quantities in the same way as we do for the t_α. We report the comparisons in the right panel of Table 3. We find very similar patterns than those observed for alphas. Consider, for example, the 95th percentile of the actual t_λ, which is equal to 3.77. The distribution of the corresponding bootstrap percentiles has an average of 1.64 (not reported in the table). No bootstrapped 95th percentile of t_λ is larger than 3.77. Therefore, the very large values of t_λ observed in the data appear atypical when compared

to their bootstrap distributions.

Overall, we find that actual t-statistics of percentiles in the tails do not appear to be drawn from the same percentile's distribution generated from zero alpha or zero FM coefficients. Yan and Zheng (2017) obtain very similar results based on their sample of 18,000 strategies. It is a little surprising that the bootstrap method "rejects" so many strategies — we find that the likelihood/p-value of t-statistics appearing from the null is close to zero for *all* percentiles below 40 and above 60. These rejection rates are even higher than those reported using classical thresholds. For example, Panel A of Table 2 shows rejection rates of *only* 16.44% for t_α and 25.19% for t_λ at the high threshold of 2.57 for a significance level of 1%.

We conduct a Monte Carlo study in Section 5 to to shed some light on the properties of the YZ bootstrap method, as well as of the MHT techniques that we present in the next section.

Alessio: ***We should get rid of the following.*** However, it is important at this stage to recognize some important limitations of the bootstrap method. Although the cross-section of alphas does provide some information about luck versus skill (i.e., true versus false null hypotheses), it does not inform us about the relative proportion of true versus false rejections of the null. As illustrated by Barras, Scaillet, and Wermers (2010), this is particularly true of the tails of the distribution. For example, if one observes that 16% of the t-statistics are above the threshold for a significance level of 1% in a two tailed test, then one can infer that there are some strategies that do beat the benchmark. However, one still cannot infer how many of these strategies represents a true discovery (i.e., for which the null should be rejected) without knowing the proportion of strategies that have truly no alpha but were lucky in generating abnormal performance in the sample (i.e., false positives). In other words, comparing the data to the bootstrap is a useful first diagnostic but one needs a formal MHT approach to the problem of assessing the proportion of outperforming strategies.

4 Multiple hypotheses testing

Classical single hypothesis testing (CHT) uses a significance level α to control Type I error (discovery of false positives). In multiple hypothesis testing (MHT), using α to test each individual hypothesis does not control the overall probability of false positives.[5] For instance, consider a number test statistics which are true under the null and independent of each other. If we set the significance level at 5%, the rate of Type I error (i.e., the probability of making at least one false discovery) is $1 - 0.95^{10} = 40\%$ in testing ten hypotheses and over 99% in testing 100 hypotheses.

There are three broad approaches in the statistics literature to deal with this problem: family-wise error rate (FWER), false discovery rate (FDR), and false discovery proportion (FDP). In this section, we describe these approaches and provide details on their implementation.

We are interested in testing the performance of trading strategies by analyzing the abnormal returns generated by M signals. The test statistic is either t_α or t_λ (equivalently the p-values). The null hypothesis corresponding to each strategy is labeled as H_m. For ease of notation, we will relabel the strategies and order them from the best (highest t-statistic) to the worst (lowest t-statistic). In other words, it is assumed that $t_1 \geq t_2 \geq \ldots \geq t_M$, or equivalently the p-values $p_1 \leq p_2 \leq \ldots \leq p_M$. Some of the methods used in this section use a bootstrap procedure which is the same as that described in the previous section.

4.1 FWER

The strictest idea in MHT is to try to avoid any false rejections. This translates to controlling the FWER, which is defined as the probability of rejecting even one of the true null hypotheses:

$$\text{FWER} = \text{Prob}\{\text{Reject even one true null hypothesis}\}.$$

[5]The use of symbol α to denote both the significance level as well as the abnormal returns from a factor model is standard. We hope that this does not cause any confusion and the usage is clear from the context.

Thus, FWER measures the probability of even one false discovery, i.e., rejecting even one true null hypothesis (type I error). A testing method is said to control the FWER at a significance level α if FWER $\leq \alpha$. There are many approaches to controlling FWER.

4.1.1 Bonferroni method

The Bonferroni method, at level α, rejects H_m if $p_m \leq \alpha/M$. The Bonferroni method is a single-step procedure because all p-values are compared to a single critical value. This critical p-value is equal to α/M. For a very large number of strategies, this leads to an extremely small (large) critical p-value (t-statistic). While widely used for its simplicity, the biggest disadvantage of the Bonferroni method is that it is very conservative and leads to a loss of power. One of the main reasons for the lack of power is that the Bonferroni method implicitly treats all test statistics as independent and, consequently, ignores the cross-correlations that are bound to be present in most financial applications.

4.1.2 Holm method

This is a stepwise method based on Holm (1979) and works as follows. The null hypothesis H_i is rejected at level α if $p_i \leq \alpha/(M - i + 1)$ for $i = 1, \ldots, M$. In comparison with the Bonferroni method, the criterion for the smallest p-value is equally strict at α/M but it becomes less and less strict for larger p-values. Thus, the Holm method will typically reject more hypotheses and is more powerful than the Bonferroni method. However, because it also does not take into account the dependence structure of the individual p-values, the Holm method is also very conservative.

4.1.3 Bootstrap reality check

Bootstrap reality check (BRC) is based on White (2000). The idea is to estimate the sampling distribution of the largest test statistic taking into account the dependence structure of the individual test statistics, thereby asymptotically controlling FWER.

The implementation of the method proceeds as follows. Bootstrap the data using procedure described in Section 3. For each bootstrapped iteration b, calculate the highest (absolute) t-statistic across all strategies and call it $t^{(b)}_{\max}$, where the superscript b is used to clarify that these t-statistics come from the bootstrap. The critical value is computed as the $(1-\alpha)$ empirical percentile of B bootstrap iterations values $t^{(1)}_{\max}, t^{(2)}_{\max}, \ldots, t^{(B)}_{\max}$.

Statistically speaking, BRC can be viewed as a method that improves upon Bonferroni by using the bootstrap to get a less conservative critical value. From an economic point of view, BRC addresses the question of whether the strategy that appears the best in the observed data really beats the benchmark. However, BRC method does not attempt to identify as many outperforming strategies as possible.

4.1.4 StepM method

This method, based on Romano and Wolf (2005) addresses the problem of detecting as many out-performing strategies as possible. The stepwise StepM method is an improvement over the single-step BRC method in very much the same way as the stepwise Holm method improves upon the single-step Bonferroni method. The implementation of this procedure proceeds as follows:

1. Consider the set of all M strategies. For each cross-sectional bootstrap iteration, compute the maximum t-statistic, thus obtaining the set $t^{(1)}_{\max}, t^{(2)}_{\max}, \ldots, t^{(B)}_{\max}$. Then compute the critical value c_1 as the $(1-\alpha)$ empirical percentile of the set of maximal t-statistics, as in BRC method. Apply now the c_1 threshold to the set of original t-statistics and determine the number of strategies for which the null can be rejected. Say that there are M_1 strategies, for which $t_m \geq c_1$. We have now $M - M_1$ strategies remaining with t-statistics ordered as $t_{M_1+1}, t_{M_1+2}, \ldots, t_M$.

2. Consider the set of remaining $M - M_1$ strategies. For each bootstrapped iteration b, calculate the highest (absolute) t-statistic across all remaining strategies. To avoid

cluttering up the notation, we will use the same symbols as before and call the maximal t-statistics of the b bootstrap iteration across the $M - M_1$ remaining strategies as $t^{(b)}_{\max}$. The critical value c_2 is computed as the $(1-\alpha)$ empirical percentile of B bootstrap iterations values $t^{(1)}_{\max}, t^{(2)}_{\max}, \ldots, t^{(B)}_{\max}$. Say that there are M_2 strategies, for which $t_m \geq c_2$, and are, therefore, rejected in this step. After this step, $M - M_1 - M_2$ strategies remain with t-statistics ordered as $t_{M_1+M_2+1}, t_{M_1+M_2+2}, \ldots, t_M$.

3. Repeat the procedure until there are no further strategies that are rejected. The StepM critical value for the entire procedure is equal to the critical value of the last step and the number of strategies that are rejected is equal to the sum of the number of strategies that are rejected in each step.

Like the Holm method, the StepM method is a stepdown method that starts by examining the most significant strategies. The main advantage of the method is that, because it relies on bootstrap, it is valid under arbitrary correlation structure of the test statistics. As mentioned before, this method will detect many more out-performing strategies than the Bonferroni method or the BRC approach.

It is easy to see that the BRC approach amounts to only step one of the above procedure, namely computing only the critical value c_1. By continuing the method after the first step, more false null hypotheses can be rejected. Moreover, since typically $c_1 > c_2 > \ldots$, the critical value in StepM method is less conservative than that in BRC approach. Nevertheless, the StepM procedure still asymptotically controls FWER at significance level α.

4.2 *k*-FWER

By relaxing the strict FWER criterion, one can reject more false hypotheses. For instance, k-FWER is defined as the probability of rejecting at least k of the true null hypotheses:

$$k\text{-FWER} = \text{Prob}\{\text{Reject at least } k \text{ of the true null hypothesis}\}.$$

A testing method is said to control for k-FWER at a significance level α if k-FWER $\leq \alpha$. Testing methods such as Bonferroni and Holm, discussed earlier, can be generalized for k-FWER testing. Please refer to Romano, Shaikh, and Wolf (2008) for further details. Here we discuss only the extension of the StepM method which is known as the k-StepM method.

4.2.1 *k*-StepM method

The implementation of this procedure proceeds as follows:

1. Consider the set of all M strategies. For each bootstrapped iteration b, calculate the k-highest (absolute) t-statistic across all strategies and call it $t^{(b)}_{k\text{-max}}$, where the superscript b is used to clarify that these t-statistics come from the bootstrap. Compute the critical value c_1 as the $(1-\alpha)$ empirical percentile of B bootstrap iterations values $t^{(1)}_{k\text{-max}}, t^{(2)}_{k\text{-max}}, \ldots, t^{(B)}_{k\text{-max}}$. Say that there are M_1 strategies, for which $t_m \geq c_1$, and are, therefore, rejected in this step. After this step, $M - M_1$ strategies remain with t-statistics ordered as $t_{M_1+1}, t_{M_1+2}, \ldots, t_M$. Apart from the use of k-max instead of max, this step is identical to the first step of StepM procedure.

2. Consider the set of remaining $M - M_1$ strategies. Call this set `Remain`. Also consider a number $k-1$ of strategies from the set of already rejected strategies. Call this set `Reject`. Now consider the union of these two sets, `Consider` = `Remain` $\cup$ `Reject`. For each bootstrapped iteration b, calculate the k-highest (absolute) t-statistic across all strategies in the set `Consider` and call it $t^{(b)}_{k\text{-max}}$. Compute the $(1-\alpha)$ empirical percentile of B bootstrap iterations values $t^{(1)}_{k\text{-max}}, t^{(2)}_{k\text{-max}}, \ldots, t^{(B)}_{k\text{-max}}$. This empirical percentile will depend on which $k-1$ strategies were included in the set `Reject`. Given that there are $\binom{M_1}{k-1}$ possible ways of choosing $k-1$ strategies from a set of M_1 strategies, the critical value c_2 is computed as the maximum across all these permutations. Say that there are M_2 strategies, for which $t_m \geq c_2$, and are, therefore, rejected in this step. After this step, $M - M_1 - M_2$ strategies remain with t-statistics ordered as $t_{M_2+1}, t_{M_2+2}, \ldots, t_M$.

3. Repeat the procedure until there are no further strategies that are rejected. The critical value of the procedure is equal to the critical value of the last step and the number of strategies that are rejected is equal to the sum of the number of strategies that are rejected in each step.

The key innovation in the k-StepM procedure is in the inclusion of (some of the) rejected strategies while calculating subsequent critical values (c_2 and thereafter). The intuition is as follows. Remember that ideally we want to calculate the empirical critical value from the set of strategies that are true under the null hypothesis. This set is unknown in practice. However, we can use the results of the first step to arrive at this set. The set `Remain` of remaining strategies that have not (yet) been rejected is an obvious candidate for strategies that are true under the null. If we are in the second step of the procedure, it stands to reason that the first step was not able to control k-FWER. In other words, less than k true null hypotheses were rejected in the first step. Let's say that number is in fact $k-1$. Obviously, we do not know with precision which $k-1$ true nulls have been rejected among the many strategies rejected in the first step. Therefore, to be cautious, Romano, Shaikh, and Wolf (2008) suggest looking at all possible combinations of $k-1$ rejected hypotheses from the set `Reject`.

4.3 False Discovery Ratio (FDR)

In many applications, we are willing to tolerate a larger number of false rejections if there are a large number of total rejections. In other words, rather than controlling for the "number" of false rejections, one can control for the "proportion" of false rejections or the False Discovery Proportion (FDP). FDR measures and controls the expected FDP among all discoveries. More formally, a multiple testing method is said to control FDR at level δ if FDR $\equiv$ E(FDP) $\leq \delta$. The level δ is a user-defined parameter which should not be confused with a significance level α. Since FDR is already an expectation, controlling for FDR does not need additional specification of probabilistic significance level. Nevertheless, the literature often uses δ and

α interchangeably. It is to be noted though that choosing false discovery ratio δ in FDR methods to be the same as the significance level α in FWER methods would imply that the FDR methods are more lenient than the FWER methods as FDR tolerates a larger number of false rejections. Harvey, Liu, and Zhu (2016) explore δ of both 5% and 1%.

One of the earliest methods to controlling FDR is by Benjamini and Hochberg (1995) and proceeds in a stepwise fashion as follows. Assuming as before that the individual p-values are ordered from the smallest to largest, and defining:

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

one rejects all hypotheses $H_1, H_2, \ldots, H_{j^*}$ (i.e., j^* is the index of the largest p-value among all hypotheses that are rejected). This is a step-up method that starts with examining the least significant hypothesis and moves up to more significant test statistics. We label this method as BH method in the rest of the paper.

Benjamini and Hochberg (1995) show that their method controls FDR if the p-values are mutually independent. Benjamini and Yekutieli (2001) show that a more general control of FDR under a more arbitrary dependence structure of p-values can be achieved by replacing the definition of j^* with:

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

where the constant $C_M = \sum_{i=1}^{M} 1/i \approx \log(M) + 0.5$. However, the Benjamini and Yekutieli method is less powerful than that of Benjamini and Hochberg. We label this method as BHY method in the rest of the paper.

4.4 False Discovery Proportion (FDP)

One caveat with FDR is that it is designed to control only the central tendency of the sampling distribution of FDP. In a given application, the realized FDP could still be far away from the level δ. Therefore, FDR's application is better suited for cases where a researcher

can analyze a large number of data sets thus allowing one to make confidence statements about the realized average FDP across the various data sets. Since our application of MHT is based on a single dataset, it is more appropriate to use a method that directly controls the FDP.[6]

A multiple testing method is said to control FDP at proportion γ and level α if Prob(FDP $> \gamma) \leq \alpha$. Lehman and Romano (2005) and Romano and Shaikh (2006) develop extensions of the Holm method for FDP control. Here we discuss only the extension of the StepM procedure developed by Romano and Wolf (2007).

4.4.1 FDP-StepM method

The StepM procedure for control of FDP is as follows:

1. Let $j = 1$ and $k_1 = 1$.

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

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

The FDP-StepM method is, thus, a sequence of k-StepM procedures. The intuition of applying an increasing series of k's is as follows. Consider controlling FDP at proportion $\gamma = 10\%$. We start by applying the 1-StepM method. Denote by M_1 the number of strategies rejected at this stage. Since the basic 1-StepM procedure controls for FWER, we can be confident that no false rejections have occurred so far, which in turn also implies that FDP has also been controlled. Consider now the issue of rejecting the strategy H_{M_1+1}, the next most significant strategy (recall that StepM is a stepdown procedure).

Rejection of H_{M_1+1}, if the null of this strategy is true, renders the false discovery proportion to be equal to $1/(M_1 + 1)$. Since we are willing to tolerate 10% of false rejections, we would be willing to tolerate rejecting this strategy if $1/(M_1 + 1) < 0.1$ which is true if $M_1 > 9$. Thus if $M_1 < 9$ the procedure would stop at the first step. Alternatively, if $M_1 > 9$,

[6]We thank Michael Wolf for explaining this important difference to us.

the procedure would continue with the 2-StepM method, which by design should not reject more than one true hypothesis.

Besides the fact that the FDP-StepM method allows the researcher to directly control FDP, one other big advantage of this method for us is that it accounts for generalized correlation structure in the data and, therefore, in the individual p-values. Such cross-correlation arises from two sources. First, different trading strategies rely on firm level data that are economically related through the balance sheet, the income statement, or market assessment of such data. Therefore, the trading signals are not independent. Second, even if the signals were truly independent, they are still applied to a common set of stock returns that co-move in time because of aggregate forces.

Thus, it is important to use methods that do not rely on restrictive assumptions about cross-correlations but are able to take into account the actual cross-correlations present in the data to deliver more precise critical values. For these reasons (and for reasons discussed earlier regarding appropriateness to our setting), we dedicate more attention to the FDP method.

5 Monte Carlo simulations

In this section we perform a Monte Carlo simulation to assess the performance of various MHT methods as well as the relative performance of the YZ bootstrap method of Yan and Zheng (2017). We simulate time-series of returns for $T = 500$ periods (i.e., months) for $N = 10,000$ strategies based on the empirical distribution of factors, factor sensitivities, and time-series residuals obtained from our sample.

The data generating process for returns to strategy p is as follows:

$$R_{pt} = \alpha_p + \beta_p' F_t + \epsilon_{pt}.$$

We simulate a six-factor model, as in the actual data, that mimics the statistical properties

of the five Fama and French (2015) factors augmented with the momentum factor.

Each month, we draw the factor returns, F_t, from a multivariate normal distribution with means and covariance matrix matched to that from the actual distribution of factor returns. For each strategy p, we draw the 6×1 vector of betas, β_p, from a multivariate normal distribution with means and covariance matrix matched to that of the cross-sectional distribution of betas from the actual data. The $N \times 1$ vector α is populated by zeros and by a fraction f of non-zero values. To better represent the actual data, we inject into each simulation, non-zero α strategies of different magnitudes. In particular, $f/2$ of the strategies are assigned a negative α and the other $f/2$ are assigned a positive value. Non-zero αs are equal to either 0.5% or 1.0%. The residuals ϵ_t are drawn from a multivariate normal distribution with mean zero and a $N \times N$ covariance matrix Σ_ϵ. The diagonal variance elements of Σ_ϵ are drawn from a normal distribution with mean and standard deviation matched to that of the cross-sectional distribution of variances of residuals from the actual data (to avoid negative values we winsorize the variances at the minimum value in the empirical distribution). The average monthly standard deviation of residuals in the data is approximately 2.8%, with minimum and maximum values of 1.2% and 7.2%, respectively. The off-diagonal covariance elements of Σ_ϵ are obtained by imposing a constant correlation ρ. Given that the average pairwise correlation between strategy returns in the real data is 3%, we report simulation results for ρ equal to 0%, 3%, and 6%.

For each simulation run, we generate the data, estimate alphas, and calculate rejection rates from the YZ bootstrap and the MHT methods. We simulate the economy $S = 1,000$ times and within each simulation run we resample $B = 1,000$ times (i.e., used to determine the YZ bootstrap and the FDP-StepM method).

5.1 Results for YZ bootstrap method

We start by discussing simulation results of the bootstrap method of Yan and Zheng (2017). In order to simplify exposition we convert the simulated population and bootstrap distri-

butions by taking the absolute value of the t-statistics. Note that this is equivalent to considering the distribution of p-values. For each percentile between 65 and 99, we tabulate the average across simulations of the frequency of cases in which the bootstrap t-statistics is larger than the corresponding population t-statistic. Yan and Zheng refer to this quantity as the percentile p-value. Since there is no formal statistical control in the bootstrap method, we assume that a t-statistic at a percentile will be "rejected" if this p-value is lower than 0.05.

With this definition in mind, we observe in Table 4 that the bootstrap method has very good size properties when the alphas across all the strategies are zero. In other words, one would not reject the null hypothesis of zero at any percentile of the distribution when the actual alpha is indeed zero. However, when 5% of the strategies have non-zero alphas, the inference changes substantially. The bootstrap method tends to over-reject. For example, with the non-zero alpha set to 0.5% and with zero correlation in residuals, the p-value is lower than 0.05 for all percentiles from 68 and higher.

The degree of over-rejection decreases with an increase in correlation due to an increase in the standard errors. It can be shown that the standard deviation of the distribution of any percentile's t-statistic increases proportionally to the square root of correlation (see Owen and Steck (1962)). Rejections also increase as the magnitude of alpha increases. This is to be expected as higher alphas are easier to detect and reject. The fewest rejections occur with alpha set to 0.5% and correlation to 6%, when only percentiles 86 and higher have p-values lower than 0.05.

The implied overall "rejection rate" is calculated as the lowest percentile that gets rejected with p-value lower than 0.05. Table 4, thus, implies a rejection rate of 32% with $\alpha = 0.5\%/\rho = 0$ and 15% with $\alpha = 1\%/\rho = 6\%$ from the bootstrap method versus the true rejection rates of only 5%. We conclude that the YZ bootstrap method has very poor power properties and substantially over-rejects in a wide variety of scenarios.

5.2 Results for MHT methods

We next turn our attention to size and power properties of MHT methods. It is useful to remember that MHT methods were developed for applications in different fields. In genomics, for example, after adjusting for MHT it is not uncommon to reject only strategies with p-values of the order of 5×10^{-7} (see, The Wellcome Trust Case Control Consortium, 2007). The American Association for the Advancement of Science writing about the discovery of Higgs boson (2012) notes that "in particle physics, the 5σ criterion has become a convention to claim discovery but should not be interpreted literally." Indeed, the motivation of the 5σ detection threshold is not to keep the false detection rate below 1 in 3.4 million tests. Rather it is an attempt to account for concerns associated with multiple testing, calibration, and/or systematic errors, and statistical error rates that are not well calibrated due to general model misspecification.

These very high thresholds in other fields are not only dictated by the need to be conservative, but also by the fact that underlying relations (natural/biological) are stronger in other fields than they are in finance and economics. Since the signal-to-noise ratio is probably very different in financial data than in the data from natural sciences, one might expect very different size and power properties for the various MHT tests applied to finance.

Table 5 presents the rejection rates for the MHT methods including Bonferroni and Holm for FWER, BH and BHY for FDR, and FDP-StepM method. The results are presented for different values of f, α and ρ. We adopt a 5% statistical significant level, and when necessary a 5% ratio or proportion of false discoveries. We present the statistical thresholds as well as rejection rates for each experiment.

Panel A presents the basic size and power properties. We fix the number N of strategies at 10,000. Consider the scenario where all alphas are equal to zero. We expect zero rejections for all MHT methods and the simulation results confirm this. The thresholds are also similar for all methods. Thus, all MHT methods considered in this paper have very good size properties.

In the second and third part of Panel A of Table 5, we keep the fraction f of non-zero

alphas at 5% but progressively increase the magnitude of alpha. In interpreting the rejection rates in these panels, it is important to bear in mind the statistical uncertainty regarding the "estimated" alphas. Even though the population alpha for some strategies may be non-zero, the estimated alpha will have a sampling distribution. For instance, in the case where the true alpha is 0.5%, given the choice of the standard deviation of residuals and the number of months, the average t-statistic of the estimated alpha should be 3.7. However, the estimated alpha t-statistic will have a distribution around (about) 3.7. It is possible that t-statistics of estimated alphas are too low to cross the statistical thresholds. In other words, if the signal is not high enough, then the noise in returns might make it difficult to detect true out-performance. Thus, the power of the tests will depend on the signal-to-noise ratio.

Consider now the case where we keep the population α at 0.5%. The thresholds from FWER methods are high at around 4.4 (and higher than the case where $f = 0$). Correspondingly, the rejection rates are low at around 1.8% regardless of the correlation in residuals. This is to be expected as FWER methods have low power as they enforce a strict control of even one false discovery. FDR methods have lower thresholds (and correspondingly higher rejection rates). But, there is a disparity in the results for BH and BHY methods. Thresholds for BH and BHY are around 3.1 and 3.8, and rejection rates are around 3.5% and 2.4%, respectively. Thus, while FDR methods are better than FWER methods, the BHY method has much lower power than the BH method. The power properties of FDP method are somewhere in between those of BHY and BH methods as FDP method rejects around 2.9% of strategies. At the same time, none of the MHT methods reach the maximum power (rejection rates of 5%) because the signal (magnitude of α) is not strong enough.

When we increase the magnitude of α to 1.0%, all methods display improved power. The thresholds for FWER methods barely change (the threshold for Bonferroni depends only on the number of strategies) but interestingly, the thresholds for FDR and FDP decline a bit vis-à-vis the case for $\alpha = 0.5\%$ suggesting that these methods are adaptive to the properties of the data. The FDP method is the most powerful method reaching rejection rates of close

to 5%. The BH method slightly overshoots but, nevertheless, maintains rejection rates of close to 5.2%. Overall, we conclude from Panel A that some MHT methods (BH and FDP) have better power properties than the other methods.

In subsequent panels of Table 5, we explore whether the MHT methods are adaptive to the properties of the data. In Panel B, we vary the proportion f of non-zero alphas from 5% to 50%. We keep N at 10,000 and also fix $\rho = 0$ and $\alpha = 0.5\%$. The goal is not only to check the power of the MHT tests but also to analyze the statistical thresholds. For instance, in the extreme case of $f = 100\%$ with zero correlation, simulating many different strategies is akin to simulating one strategy many times. In that case, MHT is the same as single hypothesis testing and the thresholds should converge to the conventional threshold of 1.96. More generally, we expect to see a decline in thresholds as f increases. Panel B shows that this is the case for the FDR and the FDP method but not for the other methods. For example, the thresholds decline from 3.46 to 2.21 (from 3.14 to 2.30) for FDP (BH) method as f increases from 5% to 50%. In contrast, the thresholds barely move for FWER methods (obviously, they are not expected to change at all for Bonferroni). The corresponding rejection rates are very low for FWER and BHY methods but similar for BH and FDP methods. This experiment, therefore, informs us that the BH and FDP methods are adaptive to the data. If the data have many true rejections, these methods will 'discover' this fact and impose lower statistical thresholds. In other words, one of the main determinants of the statistical threshold is the fraction of true rejections of the null. Echoing the results from Panel A, some MHT methods (BH and FDP) have much better power properties than the other methods.

In Panel C of Table 5, we increase the number of strategies N up to one million. We fix $f = 5\%$, $\rho = 0$ and $\alpha = 0.5\%$. The goal here is to check whether thresholds and rejection rates depend on the number of strategies. We find, as expected, that the thresholds are dependent on the number of strategies for the FWER methods (with correspondingly low rejection rates as N increases). However, the thresholds are not very dependent on N for FDR and FDP methods. The threshold and the rejection rates stabilize as soon as we

have around 50,000 strategies. Recall that all MHT methods' (indeed any method's) power depends on the signal-to-noise ratio. While increasing N does not directly translate into increase in this ratio, nevertheless, increasing N reduces the sampling error resulting in slight improvement in power. Therefore, we conclude from this panel that (a) having a large number of strategies does not impose any bias on MHT tests conducted with FDR and FDP procedures, and (b) there is slight power advantage to having more strategies than fewer strategies.

We conduct several other robustness checks and report only the results of two experiments here (without corresponding numbers in a table). First, we increase the number of time periods from 500 to 5,000. While such a long sample period is unrealistic, our goal is to assess whether abstracting from potential small sample issues leads to better performance of the MHT methods. We find that all rejection rates for MHT go towards 5% even for the case when the population value of $\alpha = 0.5\%$. For instance, the rejection rate of the FDP method increases from 2.87% with $T = 500$ to 4.62% with $T = 5,000$ (for the case when $\alpha = 0.5\%/\rho = 6\%$). Therefore, the fact that rejection rates of MHT methods are sometimes lower than 5% is driven by a low signal-to-noise ratio. Second, we increased the number of bootstrap runs from 1,000 to 10,000. We conduct this experiment to check whether the choice of $B = 1,000$, which we will adopt in the actual empirical implementation, leads to any biases. We find that the rejection rates for the FDP method that relies on bootstrapped samples are fairly unchanged regardless of the number of bootstraps.

To summarize, we find that BH and FDP methods have much better power properties than other MHT methods. We also find that the main determinants of thresholds are the signal-to-noise ratio in the data (the magnitude of α relative to volatility of returns) and the 'quality' of the data (the fraction of true rejections). The 'quantity' of data (number of strategies analyzed) has not only a minimal impact on rejection rates but also leads to no statistical bias in inferences.

6 Statistical and economic hurdles

6.1 Adjusted confidence levels

As detailed in Section 4, all MHT methods essentially consist of adjustments to the threshold p-value or t-statistic associated with a desired level of significance. We use significance level of 5% and 5% for the ratio and proportions of false discoveries. We tabulate the t-statistic thresholds and the fraction of rejections corresponding to this threshold in Table 6. Even though the previous section documents the difference in power properties for different MHT methods, we continue to report the results for all MHT methods in this table.

Consider Panel A where we use few strategies (excluding strategies 'Ratios of three') and stocks filtered by size and price. The FWER thresholds for both alpha and FM are high at 4.61 and the rejection rates are also low. For instance, only 0.20% percent of FF6 alphas have t-statistic higher than 4.61. Thresholds are also high for FDR-BHY and FDP-StepM methods but the rejection rates are a bit higher than those for FWER methods. For example, the threshold for FF6 alpha is 3.91 under FDP-StepM and 1.94% of alphas cross this threshold. As expected, by construction, FDR-BH method imposes lower thresholds, and rejects more, than the FDR-BHY method.

There are also differences across factor models. Focusing on the FDP-StepM method, surprisingly CAPM generates the least rejections while the BS model generates more than 28% rejections of the null (recall from Table 2 that cross-sectional standard deviation of BS alphas is the highest amongst all factor models). The rejection rates are more uniform across FM t-statistics. As discussed earlier in Section 2, this is partly due to the fact that the right-hand-side control variables are the same (size, book-to-market, profitability, asset growth, and one- and twelve-month lagged returns) in FF6, BS, and HXZ specification, and even the FF3 specification uses two of these controls (size and book-to-market). The CAPM specification uses no control and generates the most rejections of the null for the FM coefficients.

It is also interesting to note the contrast between the simulation results in Section 5.2 and the results here using real data. The simulation results pointed to FDR-BHY method as being more strict (and having poorer power properties) than the FDP-StepM method. We see in Table 6, however, that sometimes the thresholds are higher for the FDP-StepM method in the real data. This is partly explained by the more complicated correlation structure of return residuals in the real data (than in the data generating process in the simulation). The FDP-StepM method 'utilizes' these correlations, thereby reaching the same thresholds as the FDR-BHY method in some instances.

Panel B of Table 6 uses the same set of strategies as Panel A but we now use all stocks. It is widely known that anomalies are stronger in small-cap stocks (see, Fama and French (2008)). We, therefore, expect to see more rejections in this panel than those in Panel A. The results support this prior. For instance, rejection rate for FF6 alpha using the FDP-StepM method is 3.25% in Panel B vis-à-vis 1.94% in Panel A. As we had noted in the previous section on Monte Carlo experiments, some of the MHT methods (in particular the FDR and FDP methods) are adaptive. In particular, the rejection rates are higher and thresholds are lower if the fraction of true rejections is higher in the data. The results in Panel B reiterate this adaptive feature of our MHT tests. The thresholds are lower and rejection rates are higher in Panel B than those in Panel A. While, obviously, the fraction of true rejections is unknown to us in the real data, the results do suggest that there are more rejections of the null when we include all stocks in our analysis, consistent with our prior intuition.

Panel C expands the set of strategies to all strategies but keeps only the stocks filtered by size and price. We expect the FWER thresholds to increase mechanically. Our Monte Carlo experiments have shown that increasing N per se does not lead to an increase in threshold and/or loss of power (as long as the fraction of true rejections is kept constant). In other words, there is no statistical bias introduced by the use of a larger set of strategies. We find that ...

6.2 Economic hurdles

It is possible that some of the strategies that pass even the stricter statistical thresholds are just lucky. Although our MHT procedures are designed to guard against luck in the discovery process, some false discoveries may still slip through the net. In fact, both the FDR and the FDP methods tolerate a certain fraction of false discoveries. We would, therefore, like to consider strategies that are not only statistically significant but are also economically meaningful and relevant.

We impose additional consistency requirements and economic restrictions on the strategies that survive statistical thresholds. First, we require consistency between results obtained by studying portfolio returns and those derived from FM regressions. As discussed in Section 1.2, there are advantages and disadvantages to both portfolio sorts and regressions. We would like a trading signal to not only generate a high long-short portfolio alpha but also to explain the broader cross-section of returns in a regression setting. Therefore, we reject strategies that have statistically significant t_α but insignificant t_λ or vice-versa. Imposing this filter drastically reduces the number of strategies (we report exact numbers in the next subsection).

Second, we consider the economic magnitudes of these remaining strategies. Recall that our statistical hurdles are based on t-statistics. Since, there is a close relation between the magnitude of alpha and its t-statistic, the strategies that survive our statistical hurdles are also invariably strategies that have large alphas. For example, strategies for which both alpha and FM t-statistics are above the FDP-StepM critical values at five per cent significance and proportion have an average alpha of 0.72% per month (in absolute value). The use of alpha as an absolute indication of performance presents some difficulties. First, any value chosen as the threshold to define whether a risk-adjusted return is large enough would be largely subjective. Second, alphas do not reflect the actual trading profits realized by the strategy. For this reason, we opt for another metric that is often used in performance evaluation, that of Sharpe ratio.

The motivation for the choice comes from MacKinlay (1995), who argues that risk-based explanations for the rejections of the null hypothesis result in Sharpe ratios that are bounded while non-risk explanations would result in unbounded Sharpe ratios. MacKinlay (1995) suggests that a reference value for the bounds that separate trading strategies (between risk and non-risk based) could be taken as a multiple of the market Sharpe ratio. Following his suggestion, we relate the strategy's Sharpe ratio to the Sharpe ratio of the market (SRM). We use various cutoffs from half to 1.5× the SRM. For the entire sample, the monthly SRM is 0.116, corresponding to an annualized SRM of 0.4.

6.3 Lucky rejections

Comparing rejection rates using single and multiple hypothesis testing gives us an idea of lucky rejections. For example, consider the set of few strategies on stocks filtered by size and price. Panel A of Table 2 showed that 1,882 strategies have CAPM alpha t-statistic greater than 1.96. Table 6 showed that 1.04% (=127) of the strategies have t-statistic greater than 3.47, the threshold imposed by FDP-BH. This means that 93% ($= 1 - 127/1,882$) of the rejections using CHT are likely false/lucky. We tabulate these proportions of lucky rejections in Table 7.

Each panel of Table 7 is divided into three parts. The first part shows the number of strategies that cross the conventional 1.96 threshold from CHT. We report these numbers separately for strategies that cross the threshold for alpha, FM, and both alpha and FM (the number of strategies that cross the threshold for only alpha or FM is the same as that reported in Table 2). As discussed in previous subsection, the strategies that cross both the thresholds are further stratified based on their Sharpe ratio. The second and third parts of each panel show the proportion of false rejections using FDR-BH and FDP-StepM methods, respectively. We choose to focus only on these two MHT methods based on their better power properties (inferences are not different when using the other FDR-BHY method).

We start by discussing Panel A which presents the results for few strategies for stocks

filtered by size and price. Consider first the intersection of the set of strategies that cross the threshold for both alpha and FM. For example, 1,882 (3,502) strategies have CAPM alpha (FM) t-statistic greater than 1.96. The intersection of these two sets gives us only 864 strategies (7.06% of the total number of strategies). Thus, economic considerations play an important role in restricting the set of statistically significant strategies to an economically feasible set even using CHT. However, only 63 of these strategies have Sharpe ratios larger than the market in the full sample, further showing the economic limitations of statistically significant strategies.

The second part of Panel A shows the proportion of lucky rejections using the thresholds from FDR-BH. Focusing only on rejections of alpha and/or FM, we find that these proportions range from 51% to 98%. Imposing the economic constraint of consistency between alpha and FM coefficients, the proportion of lucky rejections ranges from 77% (for the BS model) to 98% (for the FF3 model). Imposing the additional economic constraint of Sharpe ratio at least that of the market, the lucky rejections range from 62% to 100%.

The last part of Panel A shows the proportion of lucky rejections using the FDP-StepM method. Since this method is more stringent (has a higher threshold) than the FDR-BH method, the fraction of rejections is lower, and correspondingly the fraction of lucky rejections is higher than that in the FDR-BH method. The proportion of lucky rejections ranges from 77% to 100%. In fact, imposing consistency between alpha and FM coefficient gives the proportion of lucky rejections to be close to 100% for CAPM, FF3, FF6, and HXZ models (and 93% for the remaining BS model). Hardly any strategy that has Sharpe ratio higher than that of the market is classified as 'true' rejection using our MHT methods.

Panel B considers the same set of strategies but expands the sample to all stocks. Recall from Table 2 and Table 6 that the number of rejections in this scenario is higher for both CHT and MHT. The top half of Panel B shows that imposing the consistency of alpha and FM again shrinks the set of feasible strategies. For example, considering the FF6 model, the intersection of 3,532 alpha rejections and 2,127 FM coefficient rejections is a set of 703

strategies, out of which 48 strategies have Sharpe ratio higher than that of the market.

The second and the third part of Panel B shows that the proportion of lucky rejections is similar to that in Panel A. Depending on the MHT method and the factor model, the proportion of lucky rejections ranges from 74% to 98%. The proportion is high even amongst strategies with Sharpe ratio higher than that of the market. To summarize, even though the absolute number of rejections (both CHT and MHT) using all stocks is higher than that using filtered stocks, the proportion of lucky rejections does not depend on the sample of stocks.

Panel C expands the set of strategies to all strategies. We find that ...

In summary, we find that between 75% and 99% of discoveries using CHT are probably lucky. The proportion of lucky discoveries is lower using the less stringent FDR-BH method and higher using the more stringent FDP-StepM method. It is also useful to recall that the FDP-StepM relies on less assumptions about the correlations across residuals (in fact, uses the correlations in the data in deriving the statistics) and also has better power properties than the FDR-BH method. Thus, if we properly account for the statistical properties of the data-generating process and use the FDP-StepM approach, the proportion of lucky discoveries is close to 98% and, thus, much higher than that reported by Harvey, Liu, and Zhu (2015) and Hou, Xue, and Zhang (2017).

7 Conclusion

We consider all firm-level accounting variables from Compustat with sufficient data along with market variables from CRSP and constructs almost two and a half million trading strategies from these variables. We examine alphas from the long-short decile portfolios as well as the FM coefficients on these variables. The traditional statistics show a large number of rejections of the null of no profitability. However, using the proper statistical hurdles based on multiple hypothesis testing, we find far fewer rejections of the null.

More importantly, we focus on the economic significance of the strategies that survive the statistical hurdles. We require the strategy to not only have a significant alpha but also a significant FM coefficient. In addition, we require that the Sharpe ratio of the strategy exceed that of the market. With these additional economic hurdles we are left with only a handful of significant strategies. The proportion of lucky rejections is close to 98% meaning that a vast majority of findings reported in the literature are likely false. Our results also suggest that markets are quite efficient after all.

References

American Association for the Advancement of Science, 2012, The Higgs Boson, *Science* 338, 1558–1559.

Barillas, Francisco and Jay Shanken, 2017, Comparing Asset Pricing Models, forthcoming *Journal of Finance*.

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

Bajgrowicz, Pierre, and Olivier Scaillet, 2012, Technical Trading Revisited: False Discoveries, Persistence Tests, and Transaction Costs, *Journal of Financial Economics* 106, 473–491.

Benjamini, Yoav, and Yosef 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.

Benjamini, Yoav, and Daniel Yekutieli, 2001, The Control of the False Discovery Rate in Multiple Testing under Dependency, *Annals of Statistics* 29, 1165–1188.

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

Chang, Andrew C., and Phillip Li, 2017, Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals Say "Often Not," forthcoming *Critical Finance Review*.

Dewald, William G., Jerry G. Thursby, and Richard G. Anderson, 1986, Replication in Empirical Economics: The Journal of Money, Credit, and Banking Project, *American Economic Review* 76, 587–630.

Fama, Eugene F., 1970, Efficient Capital Markets: A Review of Theory and Empirical Work, *Journal of Finance* 25, 383–417.

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

Fama, Eugene F., and Kenneth R. French, 2008, Dissecting Anomalies, *Journal of Finance* 63, 1653–1678.

Fama, Eugene F., and Kenneth R. French, 2010, Luck Versus Skill in the Cross-Section of Mutual Fund Returns, *Journal of Finance* 65, 1915–1947.

Fama, Eugene F., and Kenneth R. French, 2015, A Five-Factor Asset Pricing Model, *Journal of Financial Economics* 116, 1–22.

Green, Jeremiah, John R. M. Hand, and X. Frank Zhang, 2013, The Supraview of Return Predictive Signals, *Review of Accounting Studies* 18, 692–730.

Harvey, Campbell R., 2017, The Scientific Outlook in Financial Economics, *Journal of Finance* 72, 1399–1440.

Harvey, Campbell R., and Yan Liu, 2014, Evaluating Trading Strategies, *Journal of Portfolio Management* 40, 108–118.

Harvey, Campbell R., and Yan Liu, 2015, Backtesting, *Journal of Portfolio Management* 42, 13–28.

Harvey, Campbell R., and Yan Liu, 2016, Lucky Factors, Working paper.

Harvey, Campbell R., Yan Liu, and Heqing Zhu, 2015, ... and the Cross-Section of Expected Returns, *Review of Financial Studies* 29, 5–68.

Holm, Sture, 1979, A Simple Sequentially Rejective Multiple Test Procedure, *Scandinavian Journal of Statistics* 6, 65–70.

Hou, Kewei, Chen Xue, and Lu Zhang, 2015, Digesting Anomalies: An Investment Approach, *Review of Financial Studies* 28, 650–705.

Hou, Kewei, Chen Xue, and Lu Zhang, 2017, Replicating Anomalies, NBER Working Paper 23394.

Ioannidis, John P. A., 2005, Why Most Published Research Findings Are False, *PLoS Medicine* 2, 696–701.

Kosowski, Robert, Allan Timmermann, Russ Wermers, and Hal White, 2006, Can Mutual Fund "Stars" Really Pick Stocks? New Evidence from a Bootstrap Analysis, *Journal of Finance* 61, 2551–2595.

Lehmann, Eric L., and Joseph P. Romano, 2005, Generalizations of the Familywise Error Rate, *Annals of Statistics* 33, 1138–1154.

Leamer, Edward E., 1978, *Specification Searches*, Wiley, New York.

Leamer, Edward E., 1983, Let's Take the Con Out of Econometrics, *American Economic Review* 73, 31–43.

Lewellen, Jonathan, Stefan Nagel, and Jay Shanken, 2010, A Skeptical Appraisal of Asset Pricing Tests, *Journal of Financial Economics* 96, 175–194.

Linnainmaa, Juhani T., and Michael Roberts, 2016, The History of the Cross-Section of Stock Returns, forthcoming *Review of Financial Studies*.

Lo, Andrew W., and A. Craig MacKinlay, 1990, Data-Snooping Biases in Tests of Financial Asset Pricing Models, *Review of Financial Studies* 3, 431–467.

MacKinlay, A. Craig, 1995, Multifactor Models do not Explain Deviations From the CAPM, *Journal of Financial Economics* 38, 3–28.

McCullough, B. D., and H. D. Vinod, 2003, Verifying the Solution from a Nonlinear Solver: A Case Study, *American Economic Review* 93, 873–892.

McLean, R. David, and Jeffrey Pontiff, 2016, Does Academic Research Destroy Stock Return Predictability?, *Journal of Finance* 71, 5–32.

Novy-Marx, Robert, and Mihail Velikov, 2016, A Taxonomy of Anomalies and Their Trading Costs, *Review of Financial Studies* 29, 104–147.

Owen, D. B., and G. P. Steck, 1962, Moments of Order Statistics from the Equicorrelated Multivariate Normal Distribution, *Annals of Mathematical Statistics* 33, 1286–1291.

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

Romano, Joseph P., Azeem M. Shaikh, and Michael Wolf, 2008, Formalized Data Snooping Based On Generalized Error Rates, *Econometric Theory* 24, 404–447.

Romano, Joseph P., and Azeem M. Shaikh, 2006, Stepup Procedures for Control of Generalizations of the Familywise Error Rate, *Annals of Statistics* 34, 1850–1873.

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

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

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

Sullivan, Ryan, Allan Timmermann, and Halbert White, 1999, Data-Snooping, Technical Trading Rule Performance, and the Bootstrap, *Journal of Finance* 54, 1647–1691.

Yan, Xuemin (Sterling), and Lingling Zheng, 2017, Fundamental Analysis and the Cross-Section of Stock Returns: A Data-Mining Approach, *Review of Financial Studies* 30, 1382–1423.

Wellcome Trust Case Control Consortium, 2007, Genome-Wide Association Study of 14,000 Cases of Seven Common Diseases and 3,000 Shared Controls, *Nature* 447, 661–678.

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

Figure 1: Empirical distributions of portfolios returns

We construct trading strategies as described in the text. The figure shows cross-sectional histograms for average returns, alphas, Sharpe ratios, average return t-statistics, alpha t-statistics, and Fama-MacBeth regression coefficients. Alphas are computed relative to the Fama and French (2015) five-factor model augmented with a momentum factor. All returns and alphas are reported in monthly percentages. The sample period is 1972 to 2015.

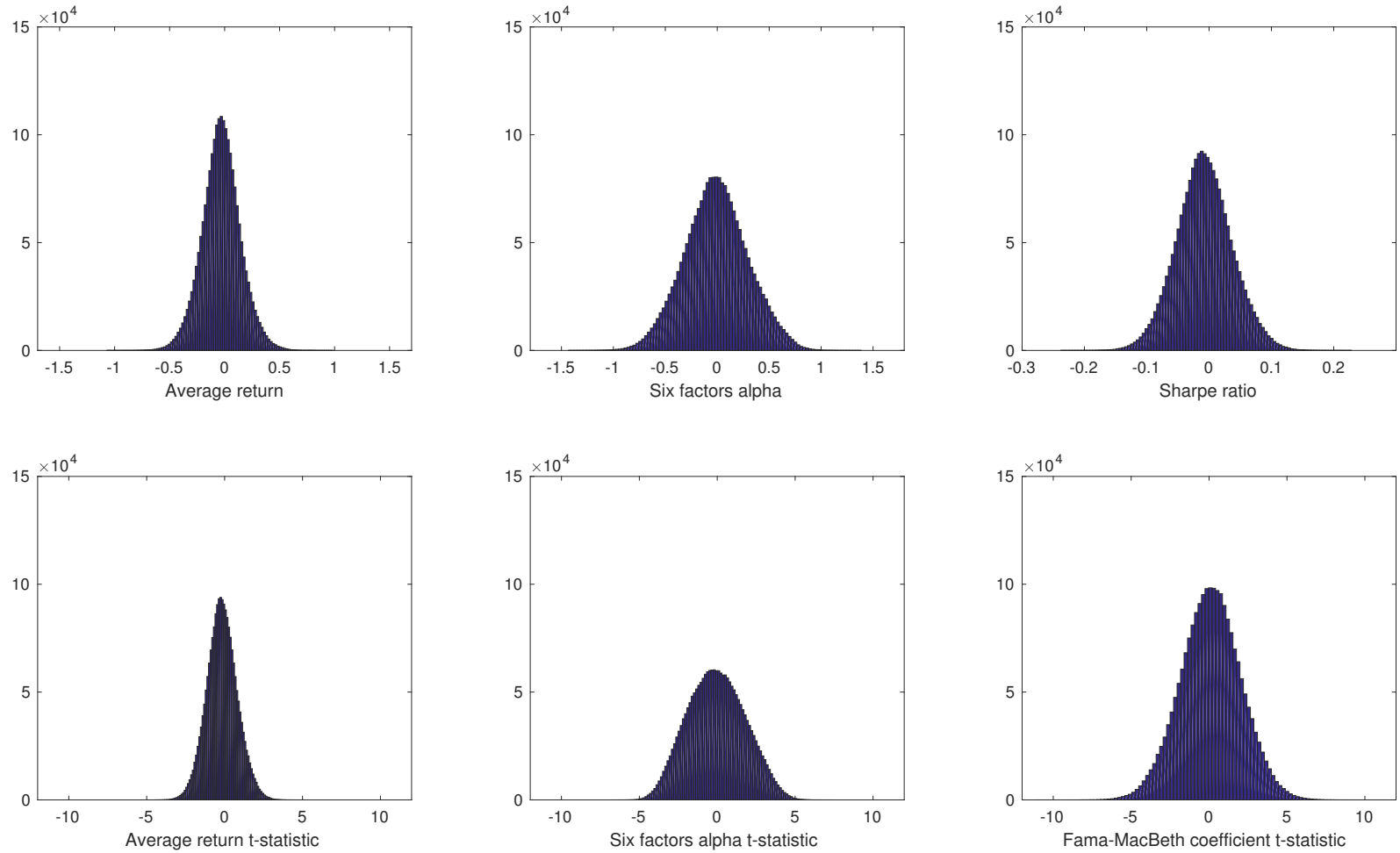

Table 1: Descriptive statistics of portfolio raw returns on trading strategies
We construct trading strategies as described in the text. This table reports the cross-sectional mean, median, standard deviation, minimum, and maximum of the monthly average return, t-statistic, and monthly Sharpe ratio. Panel A uses stocks that pass the size and price filters (please see text for details) while Pane; B uses all stocks. All returns are reported in percentages. We also report the number and percentage of strategies that cross specific thresholds in each panel. The sample period is 1972 to 2015.

Panel A: Stocks filtered by size and price										
	Average return									
	N	Mean	Median	Std	Min	Max	$\lvert ret \rvert > 0.5\%$		$\lvert ret \rvert > 1.0\%$	
							#	%	#	%
Levels	168	−0.03	−0.05	0.15	−0.34	0.62	3	1.79	0	0.00
Growth rates	142	−0.16	−0.15	0.21	−0.68	0.48	7	4.93	0	0.00
Ratios of two	11,929	−0.02	−0.02	0.17	−0.78	0.77	103	0.86	0	0.00
Ratios of three	2,373,539	−0.03	−0.03	0.17	−1.07	0.99	19,050	0.80	4	0.00
	Average return t-statistic									
	N	Mean	Median	Std	Min	Max	$\lvert t_\mu \rvert > 1.96$		$\lvert t_\mu \rvert > 2.57$	
							#	%	#	%
Levels	168	−0.31	−0.33	0.87	−2.69	2.46	10	5.95	3	1.79
Growth rates	142	−1.07	−1.08	1.35	−4.14	3.58	43	30.28	23	16.20
Ratios of two	11,929	−0.10	−0.13	0.98	−4.31	3.77	552	4.63	118	0.99
Ratios of three	2,373,539	−0.16	−0.18	0.98	−5.41	5.26	119,883	5.05	24,879	1.05
	Sharpe ratio									
	N	Mean	Median	Std	Min	Max	$\lvert SR \rvert > 0.116$		$\lvert SR \rvert > 0.232$	
							#	%	#	%
Levels	168	−0.01	−0.01	0.04	−0.12	0.11	2	1.19	0	0.00
Growth rates	142	−0.05	−0.05	0.06	−0.18	0.19	22	15.49	0	0.00
Ratios of two	11,929	−0.00	−0.01	0.04	−0.19	0.17	148	1.24	0	0.00
Ratios of three	2,373,539	−0.01	−0.01	0.04	−0.24	0.23	26,900	1.13	1	0.00

Panel B: All stocks

Average return

	N	Mean	Median	Std	Min	Max	|ret| > 0.5%		|ret| > 1.0%	
							#	%	#	%
Levels	168	0.06	−0.02	0.30	−0.79	1.11	18	10.71	2	1.19
Growth rates	142	−0.19	−0.21	0.21	−0.67	0.43	13	9.15	0	0.00
Ratios of two	11,929	0.01	−0.01	0.24	−1.03	1.22	627	5.26	25	0.21

Average return t-statistic

	N	Mean	Median	Std	Min	Max	$\|t_\mu\| > 1.96$		$\|t_\mu\| > 2.57$	
							#	%	#	%
Levels	168	0.08	−0.12	1.36	−2.64	4.00	29	17.26	14	8.33
Growth rates	142	−1.22	−1.20	1.36	−4.81	3.34	37	26.06	26	18.31
Ratios of two	11,929	−0.01	−0.08	1.20	−4.97	4.72	1,131	9.48	460	3.86

Sharpe ratio

	N	Mean	Median	Std	Min	Max	|SR| > 0.116		|SR| > 0.232	
							#	%	#	%
Levels	168	0.00	−0.01	0.06	−0.12	0.18	12	7.14	0	0.00
Growth rates	142	−0.05	−0.05	0.06	−0.21	0.17	26	18.31	0	0.00
Ratios of two	11,929	−0.00	−0.00	0.05	−0.23	0.21	460	3.86	0	0.00

Table 2: Descriptive statistics of abnormal returns and regression coefficients t-statistics

This table reports the cross-sectional mean, median, standard deviation, minimum, and maximum of alpha and Fama-MacBeth t-statistics for different samples and different factor models. Panels A and B exclude the strategies denoted 'Ratios of three' and use 12,239 strategies only while Panel C uses all 2,385,778 strategies. Panels A and C use only stocks filtered by size and price (please see text for exact description of filters) while Panel B uses all stocks. The CAPM uses the market factor. FF3 is the Fama and French (1993) three-factor model. FF6 is the Fama and French (2015) five-factor model augmented with the momentum factor. BS is the Barillas and Shanken (2015) six-factor model. HXZ is the Hou, Xue and Zhang (2015) q-model augmented with the momentum factor. Each panel shows alpha t-statistics in the top half and Fama-MacBeth t-statistics in the bottom half. In Fama-MacBeth regressions, we do not include any other control when risk adjusting stock returns CAPM. We include size and book-to-market when adjusting stock returns using the FF3 model. In all the other cases, we include size, book-to-market, profitability, asset growth, and one- and twelve-month lagged returns. All right-hand-side variables are winsorized at the 1st and 99th percentile in Fama-MacBeth regressions. We also report the number and percentage of strategies that cross specific thresholds in each panel. The sample period is 1972 to 2015.

	Mean	Median	Std	Min	Max	$\lvert t\rvert > 1.96$		$\lvert t\rvert > 2.57$	
						#	%	#	%
Panel A: Few strategies, Stocks filtered by size and price									
	Alpha t-statistics								
CAPM	−0.08	−0.05	1.37	−5.46	4.15	1,882	15.38	780	6.37
FF3	−0.36	−0.39	1.54	−5.30	4.86	2,701	22.07	1,195	9.76
FF6	−0.58	−0.65	1.71	−4.96	5.71	3,621	29.59	2,012	16.44
BS	−0.79	−0.84	2.27	−6.51	7.43	6,122	44.16	4,316	31.13
HXZ	−0.56	−0.60	1.68	−5.05	5.02	4,089	29.50	2,215	15.98
	Fama-MacBeth t-statistics								
CAPM	0.24	0.33	1.81	−7.33	5.79	3,502	28.61	1,948	15.92
FF3	0.31	0.34	1.30	−6.45	5.45	1,631	13.33	696	5.69
FF6	0.34	0.38	1.31	−6.40	5.78	1,671	13.65	749	6.12
BS	0.34	0.36	1.33	−6.66	6.25	1,935	14.00	894	6.47
HXZ	0.30	0.33	1.25	−6.69	5.62	1,696	12.27	683	4.94

	Mean	Median	Std	Min	Max	$\lvert t\rvert > 1.96$		$\lvert t\rvert > 2.57$	
						#	%	#	%
Panel B: Few strategies, All stocks									
	Alpha t-statistics								
CAPM	0.08	−0.01	1.62	−6.07	5.43	2,616	22.18	1,414	11.99
FF3	−0.14	−0.24	1.84	−6.03	7.53	3,561	30.20	1,982	16.81
FF6	−0.38	−0.33	1.77	−5.47	6.06	3,532	29.95	1,995	16.92
BS	−0.66	−0.65	2.27	−6.74	7.87	4,968	42.13	3,529	29.93
HXZ	−0.43	−0.37	1.77	−5.83	5.49	3,607	30.59	2,098	17.79
	Fama-MacBeth t-statistics								
CAPM	−0.21	−0.21	1.81	−11.26	10.36	2,545	21.58	1,516	12.86
FF3	0.21	0.19	1.48	−10.53	6.33	2,051	17.39	1,033	8.76
FF6	0.26	0.24	1.49	−10.78	6.58	2,127	18.04	1,080	9.16
BS	0.25	0.22	1.52	−11.03	6.60	2,230	18.91	1,160	9.84
HXZ	0.20	0.18	1.49	−11.03	6.20	2,091	17.73	1,062	9.01
Panel C: All strategies, Stocks filtered by size and price									
	Alpha t-statistics								
CAPM	−0.37	−0.45	1.40	−5.74	6.78	434,302	18.20	163,436	6.85
FF3	−0.41	−0.44	1.46	−6.09	6.85	485,426	20.34	213,577	8.95
FF6	−0.05	−0.08	1.82	−6.75	7.36	724,442	30.36	401,271	16.82
BS	−0.09	−0.12	2.41	−7.94	7.73	1,085,859	45.51	760,742	31.88
HXZ	−0.15	−0.14	1.72	−6.33	6.51	659,498	27.64	342,001	14.33
	Fama-MacBeth t-statistics								
CAPM	0.04	0.06	1.82	−7.55	6.93	686,718	28.78	382,160	16.02
FF3	−0.03	−0.02	1.56	−7.68	7.81	478,836	20.07	248,639	10.42
FF6	0.17	0.18	2.25	−11.68	10.93	938,357	39.33	623,755	26.15
BS	0.17	0.18	2.43	−11.27	11.35	999,646	41.90	693,572	29.07
HXZ	0.17	0.18	2.42	−12.50	11.06	997,478	41.81	693,143	29.05

Table 3: Bootstraped distributions of t-statistics

The table reports results of the bootstrap method of Yan and Zheng (2017) described in Section 3. We use stocks filtered by size and price, consider all 2,385,778 strategies, and use only the FF6 model ((Fama and French (2015) five-factor model augmented with the momentum factor) for calculating alphas and in Fama-MacBeth regressions. We run 1,000 bootstraps preserving cross-correlation between strategy returns and factors (please see the text for further details). For each percentile (i.e., each row in the table), we report the percentile of the actual t-statistics (Data) and the percentage of times when the t-statistics in the bootstrap distribution are below the actual t-statistic for percentiles one to 50, and the percentage of times when the t-statistics in the bootstrap distribution are above the actual t-statistic for percentiles 51 to 100 (% Boot). We report results for both alpha t-statistic (t_α) and Fama-MacBeth t-statistic (t_λ). The sample period is 1972 to 2015.

	t_α		t_λ	
Percentile	Data	% Boot	Data	% Boot
0.5	−4.15	0.00	−5.49	0.00
1.0	−3.85	0.00	−4.97	0.00
2.5	−3.38	0.00	−4.20	0.00
5.0	−2.94	0.00	−3.52	0.00
10.0	−2.38	0.00	−2.72	0.00
20.0	−1.63	0.00	−1.71	0.10
30.0	−1.05	0.10	−0.99	0.60
40.0	−0.55	2.50	−0.39	6.20
50.0	−0.08	26.50	0.17	0.00
60.0	0.41	15.90	0.73	0.00
70.0	0.92	2.20	1.34	0.00
80.0	1.53	0.10	2.05	0.00
90.0	2.36	0.00	3.01	0.00
95.0	3.00	0.00	3.77	0.00
97.5	3.50	0.00	4.44	0.00
99.0	4.03	0.00	5.27	0.00
99.5	4.36	0.00	5.81	0.00

Table 4: Simulations for the bootstrap of Yan and Zheng (2017)

Data are generated as described in the text. We show the results from $S = 1,000$ simulations of $N = 10,000$ strategies and $T = 500$ months. Each strategy is bootstrapped $B = 1,000$ times. f is the fraction of non-zero alphas, α is the absolute value of alpha (in percent per month), and ρ is the constant (percent) correlation amongst residuals. Since the distribution of alpha is symmetric, we focus on the absolute value of the t-statistics. For each percentile, we report the average (across simulations) of the frequency of cases in which the bootstrapped t-statistics are larger than the corresponding t-statistic in that simulation.

f	0	0	0	5	5	5	5	5	5
α	0.0	0.0	0.0	0.5	0.5	0.5	1.0	1.0	1.0
ρ	0	3	6	0	3	6	0	3	6
Prct.	Fraction of bootstrapped t-statistics above the actual t-statistic								
65	49.19	49.41	49.69	5.70	9.06	14.63	5.15	8.44	14.00
66	49.22	49.35	49.72	5.52	8.82	14.31	4.92	8.15	13.65
67	49.01	49.31	49.78	5.25	8.45	14.02	4.68	7.79	13.34
68	49.05	49.36	49.82	4.99	8.13	13.65	4.41	7.46	12.96
69	49.19	49.43	49.78	4.67	7.77	13.28	4.11	7.11	12.57
70	49.22	49.36	49.85	4.39	7.46	12.87	3.82	6.75	12.15
71	49.28	49.21	49.90	4.07	7.16	12.53	3.50	6.46	11.80
72	49.38	49.23	49.88	3.78	6.83	12.16	3.23	6.12	11.40
73	49.31	49.24	49.76	3.48	6.47	11.75	2.93	5.73	10.95
74	49.18	49.36	49.71	3.22	6.07	11.37	2.67	5.33	10.54
75	48.94	49.44	49.83	2.94	5.69	10.88	2.39	4.95	10.01
76	48.92	49.44	49.79	2.66	5.31	10.37	2.13	4.56	9.51
77	48.81	49.40	49.80	2.41	4.89	9.88	1.90	4.15	8.99
78	48.83	49.35	49.85	2.12	4.46	9.38	1.64	3.75	8.46
79	48.82	49.19	49.71	1.86	4.09	8.84	1.38	3.38	7.92
80	48.71	49.13	49.67	1.60	3.71	8.27	1.17	3.00	7.33
81	48.82	49.03	49.67	1.36	3.31	7.76	0.95	2.64	6.79
82	48.71	48.97	49.61	1.13	2.95	7.16	0.75	2.29	6.21
83	48.63	49.00	49.47	0.92	2.57	6.58	0.58	1.92	5.61
84	48.63	49.00	49.33	0.73	2.18	5.96	0.43	1.58	5.01
85	48.65	48.92	49.45	0.55	1.82	5.33	0.30	1.25	4.38
86	48.61	48.88	49.41	0.39	1.47	4.71	0.19	0.96	3.76
87	48.57	48.84	49.36	0.27	1.14	4.09	0.12	0.70	3.14
88	48.51	48.84	49.36	0.17	0.86	3.46	0.06	0.47	2.54
89	48.58	48.65	49.21	0.10	0.60	2.82	0.02	0.29	1.93
90	48.43	48.59	49.14	0.05	0.39	2.19	0.01	0.16	1.36
91	48.30	48.54	49.09	0.02	0.22	1.58	0.00	0.07	0.84
92	48.30	48.82	49.17	0.00	0.11	1.03	0.00	0.02	0.43
93	48.23	48.68	49.17	0.00	0.04	0.57	0.00	0.00	0.14
94	48.05	48.58	49.11	0.00	0.01	0.23	0.00	0.00	0.01
95	47.94	48.34	49.02	0.00	0.00	0.05	0.00	0.00	0.00
96	47.73	48.17	48.70	0.00	0.00	0.00	0.00	0.00	0.00
97	47.69	48.12	48.73	0.00	0.00	0.00	0.00	0.00	0.00
98	47.30	47.71	48.20	0.00	0.00	0.00	0.00	0.00	0.00
99	47.75	47.62	48.40	0.00	0.00	0.00	0.00	0.00	0.00

Table 5: MHT properties in simulations

Data are generated as described in the text. We show the results from $S = 1,000$ simulations of N strategies and $T = 500$ months. If required by the statistical method, each strategy is bootstrapped $B = 1,000$ times. f is the fraction of non-zero alphas, α is the absolute value of alpha, and ρ is the constant correlation amongst residuals. Bonf(erroni) and Holm control for FWER; BH and BHY control for FDR; and StepM controls for FDP. Panel A describes the basic size and power properties for different values of $f, \alpha,$ and ρ (keeping N fixed at 10,000). Panel B describes the adaptive properties by varying f (keeping $N, \alpha,$ and ρ fixed). Panel C describes the adaptive properties by varying N (keeping $f, \alpha,$ and ρ fixed). We use 5% for the ratio and proportions of false discoveries in the FDR and FDP methods. All rejection rates are for a significance level of 5%.

Panel A: Basic properties (N = 10,000)							
			FWER		FDR		FDP
f	α	ρ	Bonf	Holm	BH	BHY	StepM
			Thresholds				
0	0.0	0	3.88	3.88	3.88	3.88	3.87
0	0.0	3	3.87	3.87	3.88	3.87	3.86
0	0.0	6	3.85	3.85	3.86	3.85	3.84
5	0.5	0	4.44	4.43	3.14	3.83	3.46
5	0.5	3	4.44	4.43	3.14	3.83	3.47
5	0.5	6	4.44	4.43	3.14	3.83	3.50
5	1.0	0	4.46	4.45	3.03	3.71	3.26
5	1.0	3	4.46	4.45	3.04	3.71	3.27
5	1.0	6	4.46	4.45	3.04	3.71	3.30
			Rejections rates				
0	0.0	0	0.00	0.00	0.00	0.00	0.00
0	0.0	3	0.00	0.00	0.00	0.00	0.00
0	0.0	6	0.00	0.00	0.00	0.00	0.00
5	0.5	0	1.76	1.77	3.45	2.44	2.94
5	0.5	3	1.76	1.77	3.45	2.44	2.92
5	0.5	6	1.77	1.77	3.46	2.44	2.87
5	1.0	0	4.59	4.60	5.19	4.86	5.04
5	1.0	3	4.59	4.60	5.20	4.86	5.03
5	1.0	6	4.59	4.60	5.19	4.86	5.01

	FWER		FDR		FDP
	Bonf	Holm	BH	BHY	StepM
Panel B: Adaptive properties by varying f ($N = 10{,}000, \rho = 0, \alpha = 0.5\%$)					
f	Thresholds				
5	4.44	4.43	3.14	3.83	3.46
10	4.43	4.42	3.63	2.87	3.15
15	4.42	4.41	2.76	3.50	2.96
25	4.42	4.40	2.57	3.34	2.70
50	4.42	4.38	2.30	3.12	2.21
	Rejection rates				
5	1.76	1.77	3.45	2.44	2.94
10	3.54	3.56	5.37	7.61	6.69
15	5.33	5.36	11.71	8.51	10.77
25	8.89	8.99	20.49	15.20	19.57
50	17.81	18.24	43.20	33.15	44.23
Panel C: Adaptive properties by varying N ($f = 5\%, \rho = 0, \alpha = 0.5\%$)					
N	Thresholds				
1,000	4.14	4.14	3.19	3.84	3.63
10,000	4.44	4.43	3.14	3.83	3.46
50,000	4.89	4.89	3.13	3.88	3.38
100,000	5.03	5.03	3.13	3.90	3.37
500,000	5.33	5.33	3.13	3.93	3.37
1,000,000	5.45	5.45	3.13	3.94	3.37
	Rejections rates				
1,000	2.08	2.09	3.41	2.44	2.71
10,000	1.76	1.77	3.45	2.44	2.94
50,000	1.37	1.37	3.47	2.38	3.06
100,000	1.27	1.27	3.46	2.36	3.07
500,000	1.08	1.08	3.47	2.32	3.08
1,000,000	1.00	1.00	3.47	2.31	3.08

Table 6: MHT thresholds and rejection rates

The table shows alpha and Fama-MacBeth statistical thresholds adjusted for multiple hypothesis testing, as well as the percent of strategies rejected. We report FWER (Bonferroni and Holm), FDR (BH and BHY), and FDP (StepM) adjusted thresholds ('Thresh') and rejections ('%'). The numbers are reported for significance level of 5% and we use 5% for the ratio and proportions of false discoveries. Panels A and B exclude the strategies denoted 'Ratios of three' and use 12,239 strategies only while Panel C uses all 2,385,778 strategies. Panels A and C use only stocks filtered by size and price (please see text for exact description of filters) while Panel B uses all stocks. The CAPM uses the market factor. FF3 is the Fama and French (1993) three-factor model. FF6 is the Fama and French (2015) five-factor model augmented with the momentum factor. BS is the Barillas and Shanken (2015) six-factor model. HXZ is the Hou, Xue and Zhang (2015) q-model augmented with the momentum factor. Each panel shows alpha t-statistics in the top half and Fama-MacBeth t-statistics in the bottom half. The sample period is 1972 to 2015.

	FWER				FDR				FDP	
	Bonferroni		Holm		BH		BHY		StepM	
	Thresh	%	Thresh	%	Thresh	%	Thresh	%	Thresh	%
Panel A: Few strategies, Stocks filtered by size and price										
	Alpha t-statistic									
CAPM	4.61	0.04	4.90	0.04	3.47	1.04	4.90	0.04	4.53	0.06
FF3	4.61	0.14	4.61	0.14	3.02	5.08	4.27	0.41	3.54	4.03
FF6	4.61	0.20	4.63	0.20	2.69	14.45	3.89	1.99	3.91	1.94
BS	4.61	5.04	4.60	5.11	2.35	37.48	3.30	19.33	2.81	28.14
HXZ	4.61	0.24	4.61	0.24	2.68	14.83	3.96	1.47	4.73	0.11
	Fama-MacBeth t-statistic									
CAPM	4.61	1.16	4.61	1.17	2.70	13.80	3.72	3.95	3.32	6.46
FF3	4.61	0.20	4.62	0.20	3.31	1.86	4.25	0.44	4.34	0.35
FF6	4.61	0.27	4.61	0.27	3.25	2.34	4.20	0.53	4.34	0.42
BS	4.61	0.40	4.61	0.40	3.19	2.88	4.11	0.85	4.04	0.88
HXZ	4.61	0.11	4.61	0.11	3.40	1.36	4.60	0.14	4.61	0.12

	FWER				FDR				FDP	
	Bonferroni		Holm		BH		BHY		StepM	
	Thresh	%	Thresh	%	Thresh	%	Thresh	%	Thresh	%
Panel B: Few strategies, All stocks										
	Alpha t-statistic									
CAPM	4.61	0.41	4.62	0.41	2.86	8.56	3.93	1.75	3.57	3.15
FF3	4.61	0.77	4.61	0.77	2.69	14.40	3.76	3.39	3.52	4.54
FF6	4.61	0.27	4.62	0.27	2.69	14.49	3.83	2.62	3.70	3.25
BS	4.61	4.62	4.60	4.64	2.41	31.55	3.33	17.07	2.94	22.80
HXZ	4.61	0.33	4.61	0.33	2.66	15.43	3.85	2.35	3.91	2.11
	Fama-MacBeth t-statistic									
CAPM	4.61	2.39	4.61	2.39	2.81	9.92	3.68	4.65	3.43	5.83
FF3	4.61	0.47	4.61	0.47	3.01	5.18	4.02	1.19	3.79	1.76
FF6	4.61	0.56	4.61	0.56	2.99	5.60	3.95	1.58	3.60	2.68
BS	4.61	0.57	4.61	0.57	2.97	6.01	3.94	1.67	3.61	2.70
HXZ	4.61	0.41	4.61	0.41	3.01	5.19	3.97	1.42	3.72	2.07
Panel C: All strategies, Stocks filtered by size and price										
	Alpha t-statistic									
CAPM	5.62	0.000	5.62	0.000	3.66	0.503	6.04	0.000	4.35	0.052
FF3	5.62	0.002	5.61	0.002	3.12	3.579	4.83	0.042	4.28	0.232
FF6	5.62	0.022	5.60	0.022	2.69	14.393	4.05	1.572	3.79	2.675
BS	5.62	0.917	5.60	0.920	2.36	36.411	3.47	15.916	2.76	27.609
HXZ	5.62	0.002	5.61	0.002	2.78	10.995	4.27	0.584	3.85	1.647
	Fama-MacBeth t-statistic									
CAPM	5.62	0.085	5.62	0.085	2.72	13.146	3.92	2.760	4.11	2.056
FF3	5.62	0.064	5.62	0.064	2.96	6.214	4.10	1.250	3.95	1.598
FF6	5.62	1.059	5.61	1.062	2.47	27.148	3.57	10.755	3.12	16.312
BS	5.62	2.206	5.60	2.214	2.41	32.225	3.48	15.495	3.08	19.950
HXZ	5.62	2.035	5.60	2.043	2.41	32.172	3.48	15.293	3.08	19.884

Table 7: Proportion of lucky rejections

The table reports the number of strategies that cross statistical thresholds from classical hypothesis testing as well as proportion of lucky rejections out of these rejections accounted for multiple hypothesis testing. The thresholds are 1.96 for classical hypothesis testing and those given by Table 6 for multiple hypothesis testing methods. We use only FDR-BH and FDP-StepM methods of multiple hypothesis testing. For example, in the sample of few stratgies on stocks filtered by size and price, 1,882 strategies have CAPM alpha t-statistic greater than 1.96; 127 ($= 1.04\% \times 12,239$) strategies have this alpha larger than 3.47 threshold for FDR-BH (numbers from Table 6); and, therefore, the proportion of lucky rejections is $1 - 127/1,882 = 0.93$. The columns under the heading 'Both Alpha and FM' show the number/proportion of strategies that cross the statistical threshold for both alpha and Fama-MacBeth t-statistic. These strategies are further classified for various levels of economic significance which are determined by comparing the level of the absolute value of the strategy's Sharpe ratio to various targets determined by the market Sharpe ratio (SRM, the market Sharpe ratio for the entire sample is 0.116). These strategies are stratified into four groups: between 0 and SRM/2; between SRM/2 and SRM; between SRM and 1.5×SRM; and more than 1.5×SRM. Panels A and B exclude the strategies denoted 'Ratios of three' and use 12,239 strategies only while Panel C uses all 2,385,778 strategies. Panels A and C use only stocks filtered by size and price (please see text for exact description of filters) while Panel B uses all stocks. The CAPM uses the market factor. FF3 is the Fama and French (1993) three-factor model. FF6 is the Fama and French (2015) five-factor model augmented with the momentum factor. BS is the Barillas and Shanken (2015) six-factor model. HXZ is the Hou, Xue and Zhang (2015) q-model augmented with the momentum factor. Each panel shows alpha t-statistics in the top half and Fama-MacBeth t-statistics in the bottom half. The sample period is 1972 to 2015.

	Alpha	FM	Both Alpha and FM				
			All	0 to SRM/2	SRM/2 to SRM	SRM to 1.5×SRM	More than 1.5×SRM
Panel A: Few strategies, Stocks filtered by size and price							
	Number of rejections by classical hypothesis testing						
CAPM	1,882	3,502	864	333	468	58	5
FF3	2,701	1,631	371	200	136	32	3
FF6	3,621	1,671	568	514	39	15	0
BS	5,652	1,760	961	827	113	21	0
HXZ	3,777	1,551	537	474	45	17	1
	Proportion of lucky rejections after controlling FDR-BH						
CAPM	0.93	0.52	0.94	1.00	0.94	0.62	0.20
FF3	0.77	0.86	0.98	0.99	0.97	0.94	0.67
FF6	0.51	0.83	0.88	0.87	0.92	1.00	—
BS	0.19	0.80	0.77	0.74	0.93	1.00	—
HXZ	0.52	0.89	0.94	0.93	0.96	1.00	1.00
	Proportion of lucky rejections after controlling FDP-StepM						
CAPM	1.00	0.77	1.00	1.00	1.00	1.00	0.20
FF3	0.91	0.97	1.00	1.00	1.00	1.00	1.00
FF6	0.93	0.97	0.99	0.99	1.00	1.00	—
BS	0.39	0.94	0.93	0.93	0.98	1.00	—
HXZ	1.00	0.99	1.00	1.00	1.00	1.00	1.00

	Alpha	FM	Both Alpha and FM				
			All	0 to SRM/2	SRM/2 to SRM	SRM to 1.5×SRM	More than 1.5×SRM
Panel B: Few strategies, All stocks							
	Number of rejections by classical hypothesis testing						
CAPM	2,616	2,545	900	209	567	116	8
FF3	3,561	2,051	691	374	211	95	11
FF6	3,532	2,127	703	541	114	43	5
BS	4,968	2,230	990	821	141	24	4
HXZ	3,607	2,091	643	492	101	42	8
	Proportion of lucky rejections after controlling FDR-BH						
CAPM	0.60	0.52	0.75	0.92	0.73	0.53	0.50
FF3	0.51	0.69	0.83	0.91	0.74	0.71	0.82
FF6	0.50	0.67	0.84	0.86	0.82	0.65	0.80
BS	0.22	0.67	0.74	0.73	0.83	0.62	0.75
HXZ	0.48	0.70	0.83	0.86	0.78	0.60	0.88
	Proportion of lucky rejections after controlling FDP-StepM						
CAPM	0.84	0.71	0.93	1.00	0.95	0.77	0.75
FF3	0.83	0.89	0.97	0.98	0.95	0.94	0.91
FF6	0.88	0.84	0.98	0.99	0.99	0.91	0.80
BS	0.44	0.85	0.91	0.91	0.94	0.92	0.75
HXZ	0.93	0.88	0.98	0.99	0.99	0.93	0.88
Panel C: All strategies, Stocks filtered by size and price							
	Number of rejections by classical hypothesis testing						
CAPM	434,302	686,718	184,325	79,855	94,293	9,988	189
FF3	485,426	478,836	108,533	48,189	52,522	7,646	176
FF6	724,442	938,357	300,275	253,787	43,130	3,217	141
BS	1,085,859	999,646	476,018	403,408	68,984	3,507	119
HXZ	659,498	997,478	285,210	238,656	42,866	3,534	154
	Proportion of lucky rejections after controlling FDR-BH						
CAPM	0.97	0.53	0.98	1.00	0.97	0.85	0.45
FF3	0.82	0.67	0.94	0.97	0.93	0.84	0.56
FF6	0.51	0.28	0.65	0.65	0.64	0.63	0.59
BS	0.20	0.23	0.37	0.36	0.40	0.45	0.35
HXZ	0.60	0.23	0.70	0.71	0.65	0.60	0.63
	Proportion of lucky rejections after controlling FDP-StepM						
CAPM	1.00	0.92	1.00	1.00	1.00	0.99	0.84
FF3	0.99	0.92	1.00	1.00	1.00	0.99	0.84
FF6	0.91	0.59	0.96	0.97	0.94	0.93	0.96
BS	0.38	0.50	0.68	0.67	0.71	0.75	0.69
HXZ	0.94	0.51	0.98	0.98	0.94	0.94	0.97

Table A1: Basic variables used to construct trading strategies

#	Short	Long	#	Short	Long
1	aco	Current Assets Other Total	61	esub	Equity in Earnings Unconsolidated Subsidiaries
2	acox	Current Assets Other Sundry	62	esubc	Equity in Net Loss Earnings
3	act	Current Assets Total	63	fca	Foreign Exchange Income (Loss)
4	ao	Assets Other	64	fopo	Funds from Operations Other
5	aox	Assets Other Sundry	65	gp	Gross Profit (Loss)
6	ap	Accounts Payable Trade	66	ib	Income Before Extraordinary Items
7	aqc	Acquisitions	67	ibadj	Income Before Extraordinary Items Adjusted for Common Stock Equivalents
8	aqi	Acquisitions Income Contribution	68	ibc	Income Before Extraordinary Items (Cash Flow)
9	aqs	Acquisitions Sales Contribution	69	ibcom	Income Before Extraordinary Items Available for Common
10	at	Assets Total	70	icapt	Invested Capital Total
11	caps	Capital Surplus-Share Premium Reserve	71	idit	Interest and Related Income Total
12	capx	Capital Expenditures	72	intan	Intangible Assets Total
13	capxv	Capital Expend Property, Plant and Equipment Schd V	73	intc	Interest Capitalized
14	ceq	Common-Ordinary Equity Total	74	invfg	Inventories Finished Goods
15	ceql	Common Equity Liquidation Value	75	invrm	Inventories Raw Materials
16	ceqt	Common Equity Tangible	76	invt	Inventories Total
17	ch	Cash	77	invwip	Inventories Work In Process
18	che	Cash and Short-Term Investments	78	itcb	Investment Tax Credit (Balance Sheet)
19	chech	Cash and Cash Equivalents Increase-(Decrease)	79	itci	Investment Tax Credit (Income Account)
20	cogs	Cost of Goods Sold	80	ivaeq	Investment and Advances Equity
21	cshfd	Common Shares Used to Calc Earnings Per Share Fully Diluted	81	ivao	Investment and Advances Other
22	csho	Common Shares Outstanding	82	ivch	Increase in Investments
23	cshpri	Common Shares Used to Calculate Earnings Per Share Basic	83	ivst	Short-Term Investments Total
24	cshr	Common-Ordinary Shareholders	84	lco	Current Liabilities Other Total
25	cstk	Common-Ordinary Stock (Capital)	85	lcox	Current Liabilities Other Sundry
26	cstkcv	Common Stock-Carrying Value	86	lct	Current Liabilities Total
27	dc	Deferred Charges	87	lifr	LIFO Reserve
28	dclo	Debt Capitalized Lease Obligations	88	lifrp	LIFO Reserve Prior
29	dcpstk	Convertible Debt and Preferred Stock	89	lo	Liabilities Other Total
30	dcvsr	Debt Senior Convertible	90	lse	Liabilities and Stockholders Equity Total
31	dcvsub	Debt Subordinated Convertible	91	lt	Liabilities Total
32	dcvt	Debt Convertible	92	mib	Noncontrolling Interest (Balance Sheet)
33	dd	Debt Debentures	93	mibt	Noncontrolling Interests Total Balance Sheet
34	dd1	Long-Term Debt Due in One Year	94	mii	Noncontrolling Interest (Income Account)
35	dd2	Debt Due in 2nd Year	95	mrc1	Rental Commitments Minimum 1st Year
36	dd3	Debt Due in 3rd Year	96	mrc2	Rental Commitments Minimum 2nd Year
37	dd4	Debt Due in 4th Year	97	mrc3	Rental Commitments Minimum 3rd Year
38	dd5	Debt Due in 5th Year	98	mrc4	Rental Commitments Minimum 4th Year
39	dlc	Debt in Current Liabilities Total	99	mrc5	Rental Commitments Minimum 5th Year
40	dltis	Long-Term Debt Issuance	100	mrct	Rental Commitments Minimum 5 Year Total
41	dlto	Other Long-term Debt	101	msa	Marketable Securities Adjustment
42	dltp	Long-Term Debt Tied to Prime	102	ni	Net Income (Loss)
43	dltt	Long-Term Debt Total	103	niadj	Net Income Adjusted for Common-Ordinary Stock (Capital) Equivalents
44	dm	Debt Mortgages Other Secured	104	nopi	Nonoperating Income (Expense)
45	dn	Debt Notes	105	nopio	Nonoperating Income (Expense) Other
46	do	Discontinued Operations	106	np	Notes Payable Short-Term Borrowings
47	dp	Depreciation and Amortization	107	ob	Order Backlog
48	dpact	Depreciation, Depletion and Amortization (Accumulated)	108	oiadp	Operating Income After Depreciation
49	dpc	Depreciation and Amortization (Cash Flow)	109	oibdp	Operating Income Before Depreciation
50	dpvieb	Depreciation (Accumulated) Ending Balance (Schedule VI)	110	pi	Pretax Income
51	ds	Debt-Subordinated	111	ppegt	Property, Plant and Equipment Total (Gross)
52	dv	Cash Dividends (Cash Flow)	112	ppent	Property, Plant and Equipment Total (Net)
53	dvc	Dividends Common-Ordinary	113	ppeveb	Property, Plant, and Equipment Ending Balance (Schedule V)
54	dvp	Dividends Preferred-Preference	114	prstkc	Purchase of Common and Preferred Stock
55	dvt	Dividends Total	115	pstkc	Preferred Stock Convertible
56	ebit	Earnings Before Interest and Taxes	116	pstkl	Preferred Stock Liquidating Value
57	ebitda	Earnings Before Interest	117	pstkn	Preferred-Preference Stock Nonredeemable
58	emp	Employees	118	pstkr	Preferred-Preference Stock Redeemable
59	epsfx	Earnings Per Share (Diluted) Excluding Extraordinary Items	119	pstkrv	Preferred Stock Redemption Value
60	epspx	Earnings Per Share (Basic) Excluding Extraordinary Items	120	rea	Retained Earnings Restatement

#	Short	Long	#	Short	Long
121	reajo	Retained Earnings Other Adjustments	145	txs	Income Taxes State
122	recco	Receivables Current Other	146	txt	Income Taxes Total
123	recd	Receivables Estimated Doubtful	147	txw	Excise Taxes
124	rect	Receivables Total	148	wcap	Working Capital (Balance Sheet)
125	recta	Retained Earnings Cumulative Translation Adjustment	149	xacc	Accrued Expenses
126	rectr	Receivables Trade	150	xad	Advertising Expense
127	reuna	Retained Earnings Unadjusted	151	xido	Extraordinary Items and Discontinued Operations
128	revt	Revenue Total	152	xidoc	Extraordinary Items and Discontinued Operations (Cash Flow)
129	sale	Sales-Turnover (Net)	153	xint	Interest and Related Expense Total
130	seq	Stockholders Equity Parent	154	xlr	Staff Expense Total
131	sppe	Sale of Property	155	xopr	Operating Expenses Total
132	sstk	Sale of Common and Preferred Stock	156	xpp	Prepaid Expenses
133	tlcf	Tax Loss Carry Forward	157	xpr	Pension and Retirement Expense
134	tstk	Treasury Stock Total (All Capital)	158	xrd	Research and Development Expense
135	tstkc	Treasury Stock Common	159	xrdp	Research Development Prior
136	tstkn	Treasury Stock Number of Common Shares	160	xrent	Rental Expense
137	txc	Income Taxes Current	161	xsga	Selling, General and Administrative Expense
138	txdb	Deferred Taxes (Balance Sheet)	162	ret3	3m Past Return
139	txdi	Income Taxes Deferred	163	ret6	6m Past Return
140	txditc	Deferred Taxes and Investment Tax Credit	164	ret9	9m Past Return
141	txfed	Income Taxes Federal	165	ret12	1y Past Return
142	txfo	Income Taxes Foreign	166	price	Price
143	txp	Income Taxes Payable	167	turn	Turnover
144	txr	Income Tax Refund	168	vol	1y Return Volatility