Backtesting

Campbell R. Harvey

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

Yan Liu*

Texas A&M University, College Station, TX 77843 USA

Current version: July 28, 2015

Abstract

When evaluating a trading strategy, it is routine to discount the Sharpe ratio from a historical backtest. The reason is simple: there is inevitable data mining by both the researcher and by other researchers in the past. Our paper provides a statistical framework that systematically accounts for these multiple tests. We propose a method to determine the appropriate haircut for any given reported Sharpe ratio. We also provide a profit hurdle that any strategy needs to achieve in order to be deemed "significant".

Keywords: Sharpe ratio, Multiple tests, Backtest, Haircut, Data mining, Overfitting, Data snooping, VaR, Value at Risk, Out-of-sample tests, Trading strategies, Minimum profit hurdle.

* First posted to SSRN, October 25, 2013. Send correspondence to: Campbell R. Harvey, Fuqua School of Business, Duke University, Durham, NC 27708. Phone: +1 919.660.7768, E-mail: cam.harvey@duke.edu . We appreciate the comments of Scott Linn, Marcos López de Prado, Bernhard Scherer, Christian Walder, Nico Weinert and the seminar participants at the Inquire Europe UK, Man-AHL, APG, CPPIB, RA, as well as Wharton Jacobs-Levy, CQA, and SQA.

1 Introduction

A common practice in evaluating backtests of trading strategies is to discount the reported Sharpe ratios by 50%. There are good economic and statistical reasons for reducing the Sharpe ratios. The discount is a result of data mining. This mining may manifest itself by academic researchers searching for asset pricing factors to explain the behavior of equity returns or by researchers at firms that specialize in quantitative equity strategies trying to develop profitable systematic strategies.

The 50% haircut is only a rule of thumb. The goal of our paper is to develop an analytical way to determine the magnitude of the haircut.

Our framework relies on the statistical concept of multiple testing. Suppose you have some new data, Y, and you propose that variable X explains Y. Your statistical analysis finds a significant relation between Y and X with a t-ratio of 2.0 which has a probability value of 0.05. We refer to this as a single test. Now consider the same researcher trying to explain Y with variables $X_1, X_2, \ldots, X_{100}$. In this case, you cannot use the same criteria for significance. You expect by chance that some of these variables will produce t-ratios of 2.0 or higher. What is an appropriate cut-off for statistical significance?

In Harvey and Liu (HL, 2015), we present three approaches to multiple testing. We answer the question in the above example. The t-ratio is generally higher as the number of tests (or X variables) increases.

Consider a summary of our method. Any given strategy produces a Sharpe ratio. We transform the Sharpe ratio into a t-ratio. Suppose that t-ratio is 3.0. While a t-ratio of 3.0 is highly significant in a single test, it may not be if we take multiple tests into account. We proceed to calculate a p-value that appropriately reflects multiple testing. To do this, we need to make an assumption on the number of previous tests. For example, Harvey, Liu and Zhu (HLZ, 2015) document that at least 316 factors have been tested in the quest to explain the cross-sectional patterns in equity returns. Suppose the adjusted p-value is 0.05. We then calculate an adjusted t-ratio which, in this case, is 2.0. With this new t-ratio, we determine a new Sharpe ratio. The percentage difference between the original Sharpe ratio and the new Sharpe ratio is the "haircut".

The haircut Sharpe ratio that obtains as a result of multiple testing has the following interpretation. It is the Sharpe ratio that would have resulted from a single test, that is, a single measured correlation of Y and X.

We argue that it is a serious mistake to use the rule of thumb 50% haircut. Our results show that the multiple testing haircut is nonlinear. The highest Sharpe ratios are only moderately penalized while the marginal Sharpe ratios are heavily penalized.

This makes economic sense. The marginal Sharpe ratio strategies should be thrown out. The strategies with very high Sharpe ratios are probably true discoveries. In these cases, a 50% haircut is too punitive.

Our method does have a number of caveats – some of which apply to any use of the Sharpe ratio. First, high observed Sharpe ratios could be the results of non-normal returns, for instance an option-like strategy with high ex ante negative skew. In this case, Sharpe ratios need to be viewed in the context of the skew. Dealing with these non-normalities is the subject of future research. Second, Sharpe ratios do not necessarily control for risk. That is, the volatility of the strategy may not reflect the true risk. Importantly, our method also applies to Information ratios which use residuals from factor models. Third, it is necessary in the multiple testing framework to take a stand on what qualifies as the appropriate significance level, e.g., is it 0.10 or 0.05? Fourth, a choice needs to made on the multiple testing method. We present results for three methods as well as the average of the methods. Finally, some judgment is needed specifying the number of tests.

Given choices (3)-(5), it is important to determine the robustness of the haircuts to changes in these inputs. We provide a program at:
http://faculty.fuqua.duke.edu/∼charvey/backtesting
that allows the user to vary the key parameters to investigate the impact on the haircuts. We also provide a program that determines the minimal level of profitability for a trading strategy to be considered "significant".

2 Method

2.1 Single Tests and Sharpe Ratios

Let r_t denote the realized return for an investment strategy between time $t-1$ and t. The investment strategy involves zero initial investment so that r_t measures the net gain/loss. Such a strategy can be a long-short strategy, i.e., $r_t = R_t^L - R_t^S$ where R_t^L and R_t^S are the gross investment returns for the long and short position, respectively. It can also be a traditional stock and bond strategy for which investors borrow and invest in a risky equity portfolio.

To evaluate whether an investment strategy can generate "true" profits and maintain those profits in the future, we form a statistical test to see if the expected excess returns are different from zero. Since investors can always switch their positions in the long-short strategy, we focus on a two-sided alternative hypothesis. In other words, in so far as the long-short strategy can generate a mean return that is significantly different from zero (either positive or negative), we think of it as a profitable strategy. To test this hypothesis, we first construct key sample statistics. Given a sample of

historical returns $(r_1, r_2, \ldots, r_T)$, let $\hat{\mu}$ denote the mean and $\hat{\sigma}$ the standard deviation. A t-statistic is constructed to test the null hypothesis that the average return is zero:

$$t\text{-ratio} = \frac{\hat{\mu}}{\hat{\sigma}/\sqrt{T}}. \tag{1}$$

Under the assumption that returns are i.i.d. normal,[1] the t-statistic follows a t-distribution with $T-1$ degrees of freedom under the null hypothesis. We can follow standard hypothesis testing procedures to assess the statistical significance of the investment strategy.

The Sharpe ratio — one of the most commonly used summary statistics in finance — is linked to the t-statistic in a simple manner. Given $\hat{\mu}$ and $\hat{\sigma}$, the Sharpe ratio $(\widehat{SR})$ is defined as

$$\widehat{SR} = \frac{\hat{\mu}}{\hat{\sigma}}, \tag{2}$$

which, based on Equation (1), is simply t-ratio$/\sqrt{T}$.[2] Therefore, for a fixed T, a higher Sharpe ratio implies a higher t-statistic, which in turn implies a higher significance level (lower p-value) for the investment strategy. This equivalence between the Sharpe ratio and the t-statistic, among many other reasons, justifies the use of Sharpe ratio as an appropriate measure of the attractiveness of an investment strategy under our assumptions.

2.2 Sharpe Ratio Adjustment under Multiple Tests

Despite its widespread use, the Sharpe ratio for a particular investment strategy can be misleading.[3] This is due to the extensive data mining by the finance profession. Since academics, financial practitioners and individual investors all have a keen interest in finding lucrative investment strategies from the limited historical data, it is not surprising for them to "discover" a few strategies that appear to be very profitable. This data mining issue is well recognized by both the finance and the science literature. In finance, many well-established empirical "abnormalities" (e.g, certain technical trading rules, calendar effects, etc.) are overturned once data mining biases are taken into account.[4] Profits from trading strategies that use cross-sectional eq-

[1] Without the normality assumption, the t-statistic becomes asymptotically normally distributed based on the Central Limit Theorem.

[2] Lower frequency Sharpe ratios can be calculated straightforwardly assuming higher frequency returns are independent. For instance, if $\hat{\mu}$ and $\hat{\sigma}$ denote the mean and volatility of monthly returns, respectively, then the annual Sharpe ratio equals $12\hat{\mu}/\sqrt{12}\hat{\sigma} = \sqrt{12}\hat{\mu}/\hat{\sigma}$.

[3] It can also be misleading if returns are not i.i.d. (for example, non-normality and/or autocorrelation) or if the volatility does not reflect the risk.

[4] See Sullivan, Timmermann and White (1999, 2001) and White (2000).

uity characteristics involve substantial statistical biases.[5] The return predictability of many previously documented variables is shown to be spurious once appropriate statistical tests are performed.[6] In medical research, it is well-known that discoveries tend to be exaggerated.[7] This phenomenon is termed the "winner's curse" in medical science: the scientist who makes the discovery in a small study is cursed by finding an inflated effect.

Given the widespread use of the Sharpe ratio, we provide a probability based multiple testing framework to adjust the conventional ratio for data mining. To illustrate the basic idea, we give a simple example in which all tests are assumed to be independent. This example is closely related to the literature on data mining biases. However, we are able to generalize important quantities in this example using a multiple testing framework. This generalization is key to our approach as it allows us to study the more realistic case when different strategy returns are correlated.

To begin with, we calculate the p-value for the single test:

$$\begin{aligned} p^S &= Pr(|r| > t\text{-ratio}) \\ &= Pr(|r| > \widehat{SR} \cdot \sqrt{T}), \end{aligned} \tag{3}$$

where r denotes a random variable that follows a t-distribution with $T-1$ degrees of freedom. This p-value might make sense if researchers are strongly motivated by an economic theory and directly construct empirical proxies to test the implications of the theory. It does not make sense if researchers have explored hundreds or even thousands of strategies and only choose to present the most profitable one. In the latter case, the p-value for the single test may greatly overstate the true statistical significance.

To quantitatively evaluate this overstatement, we assume that researchers have tried N strategies and choose to present the most profitable (largest Sharpe ratio) one. Additionally, we assume (for now) that the test statistics for these N strategies are independent. Under these simplifying assumptions and under the null hypothesis that none of these strategies can generate non-zero returns, the multiple testing p-

[5] See Leamer (1978), Lo and MacKinlay (1990), Fama (1991), and Schwert (2003). A recent paper by McLean and Pontiff (2015) shows a significant degradation of performance of identified anomalies after publication.

[6] See Welch and Goyal (2004).

[7] See Button et al. (2013).

value, p^M, for observing a maximal t-statistic that is at least as large as the observed t-ratio is

$$\begin{aligned} p^M &= Pr(\max\{|r_i|, i = 1, \ldots, N\} > t\text{-ratio}) \\ &= 1 - \prod_{i=1}^{N} Pr(|r_i| \leq t\text{-ratio}) \\ &= 1 - (1 - p^S)^N. \end{aligned} \tag{4}$$

When $N = 1$ (single test) and $p^S = 0.05$, $p^M = 0.05$, so there is no multiple testing adjustment. If $N = 10$ and we observe a strategy with $p^S = 0.05$, $p^M = 0.401$, implying a probability of about 40% in finding an investment strategy that generates a t-statistic that is at least as large as the observed t-ratio, much larger than the 5% probability for single test. Multiple testing greatly reduces the statistical significance of single test. Hence, p^M is the adjusted p-value after data mining is taken into account. It reflects the likelihood of finding a strategy that is at least as profitable as the observed strategy after searching through N individual strategies.

By equating the p-value of a single test to p^M, we obtain the defining equation for the multiple testing adjusted (haircut) Sharpe ratio $\widehat{HSR}$:

$$p^M = Pr(|r| > \widehat{HSR} \cdot \sqrt{T}). \tag{5}$$

Since p^M is larger than p^S, $\widehat{HSR}$ will be smaller than $\widehat{SR}$. For instance, assuming there are twenty years of monthly returns ($T = 240$), an annual Sharpe ratio of 0.75 yields a p-value of 0.0008 for a single test. When $N = 200$, $p^M = 0.15$, implying an adjusted annual Sharpe ratio of 0.32 through Equation (5). Hence, multiple testing with 200 tests reduces the original Sharpe ratio by approximately 60% (=(0.75-0.32)/0.75).

This simple example illustrates the gist of our approach. When there is multiple testing, the usual p-value p^S for single test no longer reflects the statistical significance of the strategy. The multiple testing adjusted p-value p^M, on the other hand, is the more appropriate measure. When the test statistics are dependent, however, the approach in the example is no longer applicable as p^M generally depends on the joint distribution of the N test statistics. For this more realistic case, we build on the work of HLZ to provide a multiple testing framework to find the appropriate p-value adjustment.

3 Multiple Testing Framework

When more than one hypothesis is tested, false rejections of the null hypotheses are more likely to occur, i.e., we incorrectly "discover" a profitable trading strategy. Multiple testing methods are designed to limit such occurrences. Multiple testing methods can be broadly divided into two categories: one controls the *family-wise error rate* and the other controls the *false-discovery rate*.[8] Following HLZ, we present three multiple testing procedures.

3.1 Type I Error

We first introduce two definitions of Type I error in a multiple testing framework. Assume that M hypotheses are tested and their p-values are $(p_1, p_2, \ldots, p_M)$. Among these M hypotheses, R are rejected. These R rejected hypotheses correspond to R discoveries, including both true discoveries and false discoveries (remember the null hypothesis is no skill). Let N_r denote the total number of false discoveries (also known as "false positive"), i.e., strategies incorrectly classified as profitable. Then the *family-wise error rate* (FWER) calculates the probability of making at least one false discovery:

$$\text{FWER} = Pr(N_r \geq 1).$$

Instead of studying the total number of false rejections, i.e., profitable strategies that turn out to be unprofitable, an alternative definition — the *false discovery rate* — focuses on the proportion of false rejections. Let the *false discovery proportion* (FDP) be the proportion of false rejections (measured related to total number of rejections, R):

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

Then the *false discovery rate* (FDR) is defined as:

$$\text{FDR} = E[FDP].$$

Both FWER and FDR are generalizations of the Type I error probability in a single test. Comparing the two definitions, procedures that control FDR allow the number of false discoveries to grow proportionally with the total number of tests and are thus

[8] For the literature on the *family-wise error rate*, see Holm (1979), Hochberg (1988) and Hommel (1988). For the literature on the *false-discovery rate*, see Benjamini and Hochberg (1995), Benjamini and Liu (1999), Benjamini and Yekutieli (2001), Storey (2003) and Sarkar and Guo (2009).

more lenient than procedures that control FWER. Essentially, FWER is designed to prevent even one error. FDR controls the error rate.[9]

3.2 *P*-value Adjustment under FWER

We order the p-values in ascending orders, i.e., $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(M)}$ and let the associated null hypotheses be $H_{(1)}, H_{(2)}, \ldots, H_{(M)}$.

Bonferroni's method adjusts each p-value equally. It inflates the original p-value by the number of tests M:[10]

$$Bonferroni: \quad p_{(i)}^{Bonferroni} = \min[Mp_{(i)}, 1], i = 1, \ldots, M.$$

For example, if we observe $M = 10$ strategies and one of them has a p-value of 0.05, Bonferroni would say the more appropriate p-value is $Mp = 0.50$ and hence the strategy is not significant at 5%. For a more concrete example that we will use throughout this section, suppose we observe $M = 6$ strategies and the ordered p-value sequence is $(0.005, 0.009, 0.0128, 0.0135, 0.045, 0.06)$. Five strategies would be deemed "significant" under single tests. Bonferroni suggests that the adjusted p-value sequence is $(0.03, 0.054, 0.0768, 0.081, 0.270, 0.36)$. Therefore, only the first strategy is significant under Bonferroni.

Holm's method relies on the sequence of p-values and adjusts each p-value by:[11]

$$Holm: \quad p_{(i)}^{Holm} = \min[\max_{j \leq i}\{(M - j + 1)p_{(j)}\}, 1], i = 1, \ldots, M.$$

Starting from the smallest p-value, Holm's method allows us to sequentially build up the adjusted p-value sequence. Using the previous example, the Holm adjusted p-value for the first strategy is $p_{(1)}^{Holm} = 6p_{(1)} = 0.03$, which is identical to the level prescribed by Bonferroni. Under 5% significance, this strategy is significant. The second strategy yields $p_{(2)}^{Holm} = \max[6p_{(1)}, 5p_{(2)}] = 5p_{(2)} = 0.045$, which is smaller than the Bonferroni implied p-value. Given a cutoff of 5% and different from what Bonferroni concludes, this strategy is significant. Similarly, the next four adjusted p-values are calculated as $p_{(3)}^{Holm} = \max[6p_{(1)}, 5p_{(2)}, 4p_{(3)}] = 4p_{(3)} = 0.0512$, $p_{(4)}^{Holm} = \max[6p_{(1)}, 5p_{(2)}, 4p_{(3)}, 3p_{(4)}] = 4p_{(3)} = 0.0512$, $p_{(5)}^{Holm} = \max[6p_{(1)}, 5p_{(2)}, 4p_{(3)}, 3p_{(4)}, 2p_{(5)}] =$

[9] For more details on FWER and FDR, see HLZ.

[10] For the statistics literature on Bonferroni's method, see Schweder and Spjotvoll (1982) and Hochberg and Benjamini (1990). For the applications of Bonferroni's method in finance, see Shanken (1990), Ferson and Harvey (1999), Boudoukh et al. (2007) and Patton and Timmermann (2010).

[11] For the literature on Holm's procedure and its extensions, see Holm (1979) and Hochberg (1988). Holland, Basu and Sun (2010) emphasize the importance of Holm's method in accounting research.

$2p_{(5)} = 0.09$, $p^{Holm}_{(6)} = \max[6p_{(1)}, 5p_{(2)}, 4p_{(3)}, 3p_{(4)}, 2p_{(5)}, p_{(6)}] = 2p_{(5)} = 0.09$, making none significant at 5% level. Therefore, the first two strategies are found to be significant under Holm.

Comparing the multiple testing adjusted p-values to a given significance level, we can make a statistical inference for each of these hypotheses. If we made the mistake of assuming single tests, and given a 5% significance level, we would "discover" four factors. In multiple testing, both Bonferroni's and Holm's adjustment guarantee that the *family-wise error rate* (FWER) in making such inferences does not exceed the pre-specified significance level. Comparing these two adjustments, $p^{Holm}_{(i)} \leq p^{Bonferroni}_{(i)}$ for any i.[12] Therefore, Bonferroni's method is tougher because it inflates the original p-values more than Holm's method. Consequently, the adjusted Sharpe ratios under Bonferroni will be smaller than those under Holm. Importantly, both of these procedures are designed to eliminate all false discoveries no matter how many tests for a given significance level. While this type of approach seems appropriate for a space mission (catastrophic consequence of a part failing), asset managers may be willing to accept the fact that the number of false discoveries will increase with the number of tests.

3.3 *P*-value Adjustment under FDR

Benjamini, Hochberg and Yekutieli (*BHY*)'s procedure defines the adjusted p-values sequentially:[13]

$$BHY: \quad p^{BHY}_{(i)} = \begin{cases} p_{(M)} & \text{if } i = M, \\ \min[p^{BHY}_{(i+1)}, \frac{M \times c(M)}{i} p_{(i)}] & \text{if } i \leq M-1, \end{cases}$$

where $c(M) = \sum_{j=1}^{M} \frac{1}{j}$. In contrast to Holm's method, BHY starts from the largest p-value and defines the adjusted p-value sequence through pairwise comparisons. Again using the previous example, we first calculate the normalizing constant as $c(M) = \sum_{j=1}^{6} \frac{1}{j} = 2.45$. To assess the significance of the four strategies, we start from the least significant one. BHY sets $p^{BHY}_{(6)}$ at 0.06, the same as the original value of $p_{(6)}$. For the fifth strategy, BHY yields $p^{BHY}_{(5)} = \min[p^{BHY}_{(6)}, \frac{6 \times 2.45}{5} p_{(5)}] = p^{BHY}_{(6)} = 0.06$. For the fourth strategy, BHY yields $p^{BHY}_{(4)} = \min[p^{BHY}_{(5)}, \frac{6 \times 2.45}{4} p_{(4)}] = \frac{6 \times 2.45}{4} p_{(4)} = 0.0496$. Similarly, the first three adjusted p-values are sequentially calculated as $p^{BHY}_{(3)} = \min[p^{BHY}_{(4)}, \frac{6 \times 2.45}{3} p_{(3)}] = p^{BHY}_{(4)} = 0.0496$, $p^{BHY}_{(2)} = \min[p^{BHY}_{(3)}, \frac{6 \times 2.45}{2} p_{(2)}] = p^{BHY}_{(3)} =$

[12] See Holm (1979) for the proof.

[13] For the statistical literature on BHY's method, see Benjamini and Hochberg (1995), Benjamini and Yekutieli (2001), Sarkar (2002) and Storey (2003). For the applications of methods that control the *false discovery rate* in finance, see Barras, Scaillet and Wermers (2010), Bajgrowicz and Scaillet (2012) and Kosowski, Timmermann, White and Wermers (2006).

0.0496 and $p_{(1)}^{BHY} = \min[p_{(2)}^{BHY}, \frac{6\times 2.45}{1} p_{(1)}] = p_{(2)}^{BHY} = 0.0496$. Therefore, the BHY adjusted p-value sequence is (0.0496, 0.0496, 0.0496, 0.0496, 0.06, 0.06), making the first four strategies significant at 5% level. Based on our example, BHY leads to two more discoveries compared to Holm and Holm leads to one more discovery compared to Bonferroni.

Hypothesis tests based on the adjusted p-values guarantee that the *false discovery rate* (FDR) does not exceed the pre-specified significance level. The constant $c(M)$ controls the generality of the test. In the original work by Benjamini and Hochberg (1995), $c(M)$ is set equal to one and the test works when p-values are independent or positively dependent. We adopt the choice in Benjamini and Yekutieli (2001) by setting $c(M)$ equal to $\sum_{j=1}^{M} \frac{1}{j}$. This allows our test to work under arbitrary dependency for the test statistics.

The three multiple testing procedures provide adjusted p-values that control for data mining. Based on these p-values, we transform the corresponding t-ratios into Sharpe ratios. In essence, our Sharpe ratio adjustment method aims to answer the following question: if the multiple testing adjusted p-value reflects the genuine statistical significance for an investment strategy, what is the equivalent single test Sharpe ratio that one should assign to such a strategy *as if* there were no data mining?

For both Holm and BHY, we need the empirical distribution of p-values for strategies that have been tried so far. We use the structural model estimate from HLZ. The model is based on the performance data for more than 300 risk factors that have been documented by the academic literature. However, a direct multiple testing adjustment based on these data is problematic for two reasons. First, we do not observe all the strategies that have been tried. Indeed, thousands more could have been tried and ignoring these would materially affect our results on the haircut Sharpe ratio. Second, strategy returns are correlated. Correlation affects multiple testing in that it effectively reduces the number of independent tests. Taking these two concerns into account, HLZ propose a new method to estimate the underlying distribution for factor returns. We use this distribution to make Sharpe ratio adjustment for a new strategy.

3.4 Multiple Testing and Cross-Validation

Recent important papers by López de Prado and his coauthors also consider the ex post data mining issue for standard backtests.[14] Due to data mining, they show theoretically that only seven trials are needed to obtain a spurious two-year long backtest that has an in-sample realized Sharpe ratio over 1.0 while the expected out of sample Sharpe ratio is zero. The phenomenon is analogous to the regression overfitting problem when models found to be superior in in-sample test often perform

[14] See Bailey et al. (2014, 2015) and López de Prado (2013).

poorly out-of-sample and is thus termed backtest overfitting. To quantify the degreee of backtest overfitting, they propose the calculation of the *probability of backtest overfitting* (PBO) that measures the relative performance of a particular backtest among a basket of strategies using cross-validation techniques.

Their research shares a common theme with our study. We both attempt to evaluate the performance of an investment strategy in relation to other available strategies. Their method computes the chance for a particular strategy to outperform the median of the pool of alternative strategies. In contrast, our work adjusts the statistical significance for each individual strategy so that the overall proportion of spurious strategies is controlled.

Despite these similar themes, our research is different in many ways. First, the objectives of analyses are different. Our research focuses on identifying the group of strategies that generate non-zero returns while López de Prado evaluates the relative performance of a certain strategy that is fit in-sample. For example, consider a case when where there are a group of factors that are all true. The one with the smallest t-ratio, although dominated by other factors in terms of t-ratios, may still be declared significant in our multiple testing framework. In contrast, it will rarely be considered in the PBO framework as it is dominated by other more significant strategies. Second, our method is based on a single test statistic that summarizes a strategy's performance over the entire sample whereas their method divides and joins the entire sample in numerous ways, each way corresponding to an artificial "hold-out" periods. Our method is therefore more in line with the statistics literature on multiple testing while their work is more related to out-of-sample testing and cross-validation. Third, the extended statistical framework in Harvey and Liu (2015) needs only test statistics. In contrast, their work relies heavily on the time-series of each individual strategy. While data intensive, in the López de Prado approach, it is not necessary to make assumptions regarding the data generating process for returns. As such, their approach is closer to the machine learning literature and ours is closer to the econometrics literature. Finally, the PBO method assesses whether a strategy selection process is prone to overfitting. It is not linked to any particular performance statistics. We primarily focus on Sharpe ratios as they are directly linked to t-statistics and thus p-values, which are the required inputs for multiple testing adjustment. Our framework can be easily generalized to incorporate other performance statistics as long as they also have probabilistic interpretations.

3.5 In-Sample Multiple Testing vs. Out-of-Sample Validation

Our multiple testing adjustment is based on in-sample (IS) backtests. In practice, out-of-sample (OOS) tests are routinely used to select among many strategies.

Despite its popularity, OOS testing has several limitations. First, an OOS test may not be truly "out-of-sample". A researcher tries a strategy. After running an OOS test, she finds that the strategy fails. She then revises the strategy and tries again, hoping it would work this time. This trial and error approach is not truly OOS, but it is hard for outsiders to tell. Second, an OOS test, like any other test in statistics, only works in a probabilistic sense. In other words, a success for an OOS test can be due to luck for both the in-sample selection and the out-of-sample testing. Third, given the researcher has experienced the data, there is no true OOS that uses historical data.[15] This is especially the case when the trading strategy involves economic variables. No matter how you construct the OOS test, it is not truly OOS because you know what happened in the economy.

Another important issue with the OOS method, which our multiple testing procedure can potentially help solve, is the tradeoff between Type I (false discoveries) and Type II (missed discoveries) errors due to data splitting.[16] In holding some data out, researchers increase the chance of missing "true" discoveries for the shortened in-sample data. For instance, suppose we have 1,000 observations. Splitting the sample in half and estimating 100 different strategies in-sample, i.e., based on 500 observations, suppose we identify 10 strategies that look promising (in-sample tests). We then take these 10 strategies to the OOS tests and find that two strategies "work". Note that, in this process, we might have missed, say, three strategies after the first step IS tests due to bad luck in the short IS period. These "true" discoveries are lost because they never get to the second step OOS tests.

Instead of the 50-50 split, now suppose we use a 90-10 data split. Suppose we identify 15 promising strategies. Among the strategies are two of the three "true" discoveries that we missed when we had a shorter in-sample period. While this is good, unfortunately, we have only 100 observations held out for the OOS exercise and it will be difficult to separate the "good" from the "bad". At its core, the OOS exercise faces a tradeoff between in-sample and out-of-sample testing power. While a longer in-sample period leads to a more powerful test and this reduces the chance of committing a Type II error (i.e., missing true discoveries), the shorter out-of-sample period provides too little information to truly discriminate among the factors that are found significant in-sample.

So how does our research fit? First, one should be very cautious of OOS tests because it is hard to construct a true OOS test. The alternative is to apply our multiple testing framework to the full data to identify the "true" discoveries. This involves making a more stringent cutoff for test statistics.

Another, and in our opinion, more promising framework, is to merge the two methods. Ideally, we want the strategies to pass both the OOS test on split data and the multiple test on the entire data. The problem is how to deal with the "true" discoveries that are missed if the in-sample data is too short. As a tentative solution,

[15] See López de Prado (2013) for a similar argument.

[16] See Hansen and Timmermann (2012) for a discussion on sample splitting for univariate tests.

we can first run the IS tests with a lenient cutoff (e.g., p-value = 0.2) and use the OOS tests to see which strategy survives. At the same time, we can run multiple testing for the full data. We then combine the IS/OOS test and the multiple tests by looking at the intersection of survivors. We leave the details of this approach to future research.

4 Applications

To show how to adjust Sharpe ratios for multiple testing, we first use an example to illustrate how Bonferroni's adjustment works under the assumption that test statistics are independent. We next relax the independence assumption and use the model in HLZ to adjust the Sharpe ratio for a new strategy. One salient feature of the model in HLZ is that it allows dependency in test statistics. We show in the appendix on how to apply the framework in HLZ to Sharpe ratio adjustment.

4.1 Three Strategies

To illustrate how the Sharpe ratio adjustment works, we begin with three investment strategies that have appeared in the literature. All of these strategies are zero cost hedge portfolios that simultaneously take long and short positions of the cross-section of the U.S. equities. The strategies are: the earnings-to-price ratio (E/P), momentum (MOM) and the betting-against-beta factor (BAB, Frazzini and Pedersen (2013)). These strategies cover three distinct types of investment styles (i.e., value (E/P), trend following (MOM) and potential distortions induced by leverage (BAB)) and generate a range of Sharpe ratios.[17] None of these strategies reflect transaction costs and as such the Sharpe ratios (and t-statistics) are overstated and should be considered "before costs" Sharpe ratios.

Two important ingredients to the Sharpe ratio adjustment are the initial values of the Sharpe ratios and the number of trials. To highlight the impact of these two inputs, we focus on the simplest independent case as in Section 2. With independence, the multiple testing p-value p^M and the single test p-value p^S are linked through Equation (4). When p^S is small, this relation is approximately the same as in Bon-

[17]For E/P, we construct an investment strategy that takes a long position in the top decile (highest E/P) and a short position in the bottom decile (lowest E/P) of the cross-section of E/P sorted portfolios. For MOM, we construct an investment strategy that takes a long position in the top decile (past winners) and a short position in the bottom decile (past losers) of the cross-section of portfolios sorted by past returns. Both the data for E/P and MOM are obtained from Ken French's on-line data library for the period from July 1963 to December 2012. For BAB, return statistics are extracted from Table IV of Frazzini and Pedersen (2013).

ferroni's adjustment. Hence, the multiple testing adjustment we use for this example can be thought of as a special case of Bonferroni's adjustment.

Table 1: **Multiple Testing Adjustment for Three Investment Strategies**

Summary statistics for three investment strategies: E/P, MOM and BAB (betting-against-beta, Frazzini and Pedersen (2013)). "Mean" and "Std." report the monthly mean and standard deviation of returns, respectively; $\widehat{SR}$ reports the annualized Sharpe ratio; "t-stat" reports the t-statistic for the single hypothesis test that the mean strategy return is zero (t-stat $= \widehat{SR} \times \sqrt{T/12}$); p^S and p^M report the p-value for single and multiple test, respectively; $\widehat{HSR}$ reports the Bonferroni adjusted Sharpe ratio; $\widehat{hc}$ reports the percentage haircut for the adjusted Sharpe ratio ($\widehat{hc} = (\widehat{SR} - \widehat{HSR})/\widehat{SR}$).

Strategy	Mean(%) (monthly)	Std.(%) (monthly)	$\widehat{SR}$ (annual)	t-stat	p^S (single)	p^M (multiple)	$\widehat{HSR}$ (annual)	$\widehat{hc}$ (haircut)
Panel A: N = 10								
E/P	0.43	3.47	0.43	2.99	2.88×10^{-3}	2.85×10^{-2}	0.31	26.6%
MOM	1.36	7.03	0.67	4.70	3.20×10^{-6}	3.20×10^{-5}	0.60	10.9%
BAB	0.70	3.09	0.78	7.29	6.29×10^{-13}	6.29×10^{-12}	0.74	4.6%
Panel B: N = 50								
E/P	0.43	3.47	0.43	2.99	2.88×10^{-3}	1.35×10^{-1}	0.21	50.0%
MOM	1.36	7.03	0.67	4.70	3.20×10^{-6}	1.60×10^{-5}	0.54	19.2%
BAB	0.70	3.09	0.78	7.29	6.29×10^{-13}	3.14×10^{-11}	0.72	7.9%
Panel C: N = 100								
E/P	0.43	3.47	0.43	2.99	2.88×10^{-3}	2.51×10^{-1}	0.16	61.6%
MOM	1.36	7.03	0.67	4.70	3.20×10^{-6}	1.60×10^{-5}	0.51	23.0%
BAB	0.70	3.09	0.78	7.29	6.29×10^{-13}	6.29×10^{-11}	0.71	9.3%

Table 1 shows the summary statistics for these strategies. Among these strategies, the strategy based on E/P is the least profitable as measured by the Sharpe ratio. It has an average monthly return of 0.43% and a monthly standard deviation of 3.47%. The corresponding annual Sharpe ratio is $0.43 (= (0.43\% \times \sqrt{12})/3.47\%)$. The p-value for single test is 0.003, comfortably exceeding a 5% benchmark. However, when multiple testing is taken into account and assuming that there are ten trials, the multiple testing p-value increases to 0.029. The haircut ($\widehat{hc}$), which captures the percentage change in the Sharpe ratio, is about 27%. When there are more trials, the haircut is even larger.

Sharpe ratio adjustment depends on the initial value of the Sharpe ratio. Across the three investment strategies, the Sharpe ratio ranges from 0.43 (E/P) to 0.78 (BAB). The haircut is not uniform across different initial Sharpe ratio levels. For instance, when the number of trials is 50, the haircut is almost 50% for the least profitable E/P strategy but only 7.9% for the most profitable BAB strategy.[18] We

[18]Mathematically, this happens because the p-value is very sensitive to the t-statistic when the t-statistic is large. In our example, when $N = 50$ and for BAB, the p-value for a t-statistic of 7.29 (single test) is one 50th of the p-value for a t-statistic of 6.64 (multiple testing adjusted t-statistic), i.e., $p^M/p^S \approx 50$.

believe this non-uniform feature of our Sharpe ratio adjustment procedure is economically sensible since it allows us to discount mediocre Sharpe ratios harshly while keeping the exceptional ones relatively intact.

4.2 Sharpe Ratio Adjustment for a New Strategy

Given the population of investment strategies that have been published, we now show how to adjust the Sharpe ratio of a new investment strategy. Consider a new strategy that generates a Sharpe ratio of $\widehat{SR}$ in T periods,[19] or, equivalently, the p-value p^S. Assuming that N other strategies have been tried, we draw N t-statistics from the model in HLZ. Additional details are described in the Appendix. These $N+1$ p-values are then adjusted using the aforementioned three multiple testing procedures. In particular, we obtain the adjusted p-value p^M for p^S. To take the uncertainty in drawing N t-statistics into account, we repeat the above procedure many times to generate a sample of p^M's. The median of this sample is taken as the final multiple testing adjusted p-value. This p-value is then transformed back into a Sharpe ratio — the multiple testing adjusted Sharpe ratio. Figure 1 shows the original vs. haircut Sharpe ratios and Figure 2 shows the corresponding haircut.

First, as previously discussed, the haircuts depend on the levels of the Sharpe ratios. Across the three types of multiple testing adjustment and different numbers of tests, the haircut is almost always above and sometimes much larger than 50% when the annualized Sharpe ratio is under 0.4. On the other hand, when the Sharpe ratio is greater than 1.0, the haircut is at most 25%. This shows the 50% rule of thumb discount for the Sharpe ratio is inappropriate: 50% is too lenient for relatively small Sharpe ratios (< 0.4) and too harsh for large ones (> 1.0). This nonlinear feature of the Sharpe ratio adjustment makes economic sense. Marginal strategies are heavily penalized because they are likely false "discoveries".

Second, the three adjustment methods imply different magnitudes of haircuts. Given the theoretical objectives that these methods try to control (i.e., *family-wise error rate* (FWER) vs. *false discovery rate* (FDR)), we should divide the three adjustments into two groups: Bonferroni and Holm as one group and BHY as the other group. Comparing Bonferroni and Holm's method, we see that Holm's method implies a smaller haircut than Bonferroni's method. This is consistent with our previous discussion on Holm's adjustment being less aggressive than Bonferroni's adjustment. However, the difference is relatively small (compared to the difference between Bonferroni and BHY), especially when the number of tests is large. The haircuts under BHY, on the other hand, are usually a lot smaller than those under Bonferroni and Holm when the Sharpe ratio is small (< 0.4). For large Sharpe ratios

[19] Assuming T is in months, if $\widehat{SR}$ is an annualized Sharpe ratio, t-stat $= \widehat{SR} \times \sqrt{T/12}$; if $\widehat{SR}$ is a monthly Sharpe ratio, t-stat $= \widehat{SR} \times \sqrt{T}$.

(> 1.0), however, the haircuts under BHY are consistent with those under Bonferroni and Holm.

In the end, we would advocate the BHY method. The FWER seems appropriate for applications where there is a severe consequence of a false discovery. In financial applications, it seems reasonable to control for the rate of false discoveries rather than the absolute number.

Figure 1: **Original vs. Haircut Sharpe Ratios**

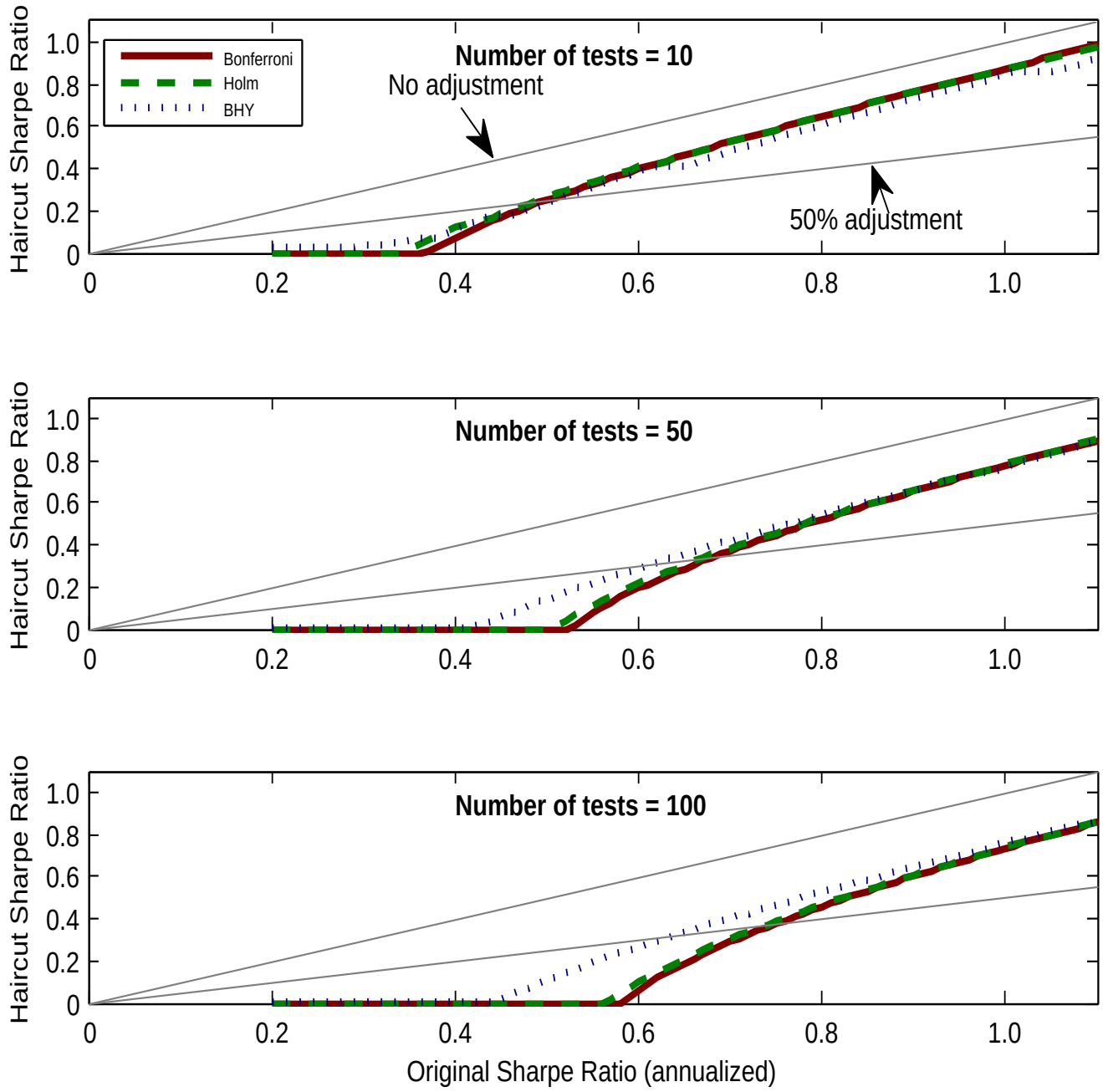
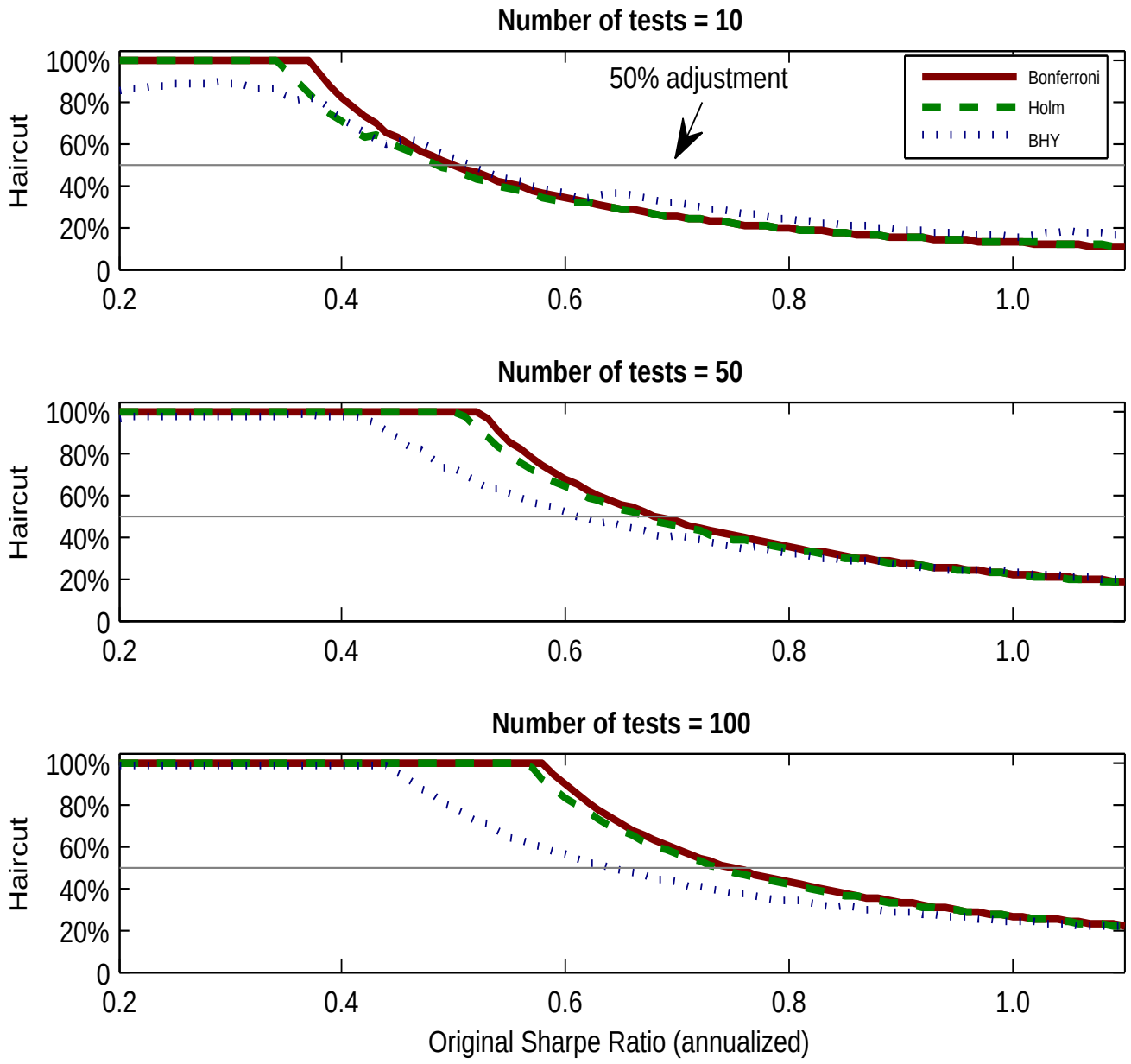


Figure 2: **Haircuts**



4.3 Minimum Profitability for Proposed Trading Strategies

There is another way to pose the problem. Given an agreed upon level of significance, such as 0.05, what is the minimum average monthly return that a proposed strategy needs to exceed? Our framework is ideally suited to answer this question.

The answer to the question depends on a number of inputs. We need to measure the volatility of the strategy. The number of observations is also a critical input.[20] Finally, we need to take a stand on the number of tests that have been conducted.

Table 2 presents an example. Here we consider four different sample sizes: 120, 240, 480 and 1,000 and three different levels of annualized volatility: 5%, 10% and 15%. We then assume the total number of tests is 300. To generate the table, we first find the threshold t-ratios based on the multiple testing adjustment methods provided in the previous section and then transform these t-ratios into mean returns based on the formula in Equation (1).

Table 2 shows the large differences between the return hurdles for single testing and multiple testing. For example, in Panel B (240 observations) and 10% volatility, the minimum required average monthly return for a single test is 0.365% per month or 4.4% annually. However, for BHY, the return hurdle is much higher, 0.616% per month or 7.4% on an annual basis. Appendix A.2 details the program that we use to generate these return hurdles and provides an Internet address to download the program.

5 Conclusions

There are many considerations involved in the evaluation of a trading strategy. The set of criteria may include the strategy's economic foundation, Sharpe ratio, level of significance, drawdown, consistency, diversification, recent performance, etc. We provide a real time evaluation method for determining the significance of a candidate trading strategy. Our method explicitly takes into account that hundreds if not thousands of strategies have been proposed and tested in the past. Given these multiple tests, inference needs to be recalibrated.

Our method follows the following steps. First, we transform the Sharpe ratio into a t-ratio and determine its probability value, e.g., 0.05. Second, we determine what the appropriate p-value should be explicitly recognizing the multiple tests that preceded the discovery of this particular investment strategy. Third, based on this new p-value, we transform the corresponding t-ratio back to a Sharpe ratio. The new measure which we call the haircut Sharpe ratio takes multiple testing or data

[20] The number of observations is also central to converting a Sharpe ratio to a t-statistic.

mining into a account. Our method is readily applied to other popular risk metrics, like Value at Risk (VaR).[21]

Our method is ideally suited to determine minimum profitability hurdles for proposed strategies. We provide open access code where the inputs are the desired level of significance, the number of observations, the strategy volatility as well as the assumed number of tests. The output is the minimum average monthly return that the proposed strategy needs to exceed.

There are many caveats to our method. We do not observe the entire history of tests and, as such, we need to use judgement on an important input — the number of tests — for our method. In addition, we use Sharpe ratios as our starting point. Our method is not applicable insofar as the Sharpe ratio is not the appropriate measure (e.g., non-linearities in the trading strategy or the variance not being a complete measure of risk).

Of course, true out-of-sample test of a particular strategy (not a "holdout" sample of historical data) is a cleaner way to evaluate the viability of a strategy. For some strategies, models can be tested on "new" (previously unpublished) data or even on different (uncorrelated) markets. However, for the majority of trading strategies, true out of sample tests are not available. Our method allows for decisions to be made, in real time, on the viability of a proposed strategy.

[21] Let $VaR(\alpha)$ of a return series to be the α-th percentile of the return distribution. Assuming that returns are approximately normally distributed, it can be shown that VaR is related to Sharpe ratio by $\frac{VaR(\alpha)}{\sigma} = SR - z_\alpha$, where z_α is the z-score for the $(1-\alpha)$-th percentile of a standard normal distribution and σ is the standard deviation of the return. Multiple testing adjusted Sharpe ratios can then be used to adjust VaR's. As with the Sharpe ratio, if non-normalities exist, these features need to be reflected in the VaR.

Table 2: **Minimum Profitability Hurdles**

Average monthly return hurdles under single and multiple tests. At 5% significance, the table shows the minimum average monthly return for a strategy to be significant at 5% with 300 tests. All numbers are in percentage terms. See Appendix for the link to the program.

	Annualized volatility		
	$\sigma = 5\%$	$\sigma = 10\%$	$\sigma = 15\%$
Panel A: Observations = 120			
Single	0.258	0.516	0.775
Bonferroni	0.496	0.992	1.488
Holm	0.486	0.972	1.459
BHY	0.435	0.871	1.305
Panel B: Observations = 240			
Single	0.183	0.365	0.548
Bonferroni	0.351	0.702	1.052
Holm	0.344	0.688	1.031
BHY	0.307	0.616	0.923
Panel C: Observations = 480			
Single	0.129	0.258	0.387
Bonferroni	0.248	0.496	0.744
Holm	0.243	0.486	0.729
BHY	0.217	0.435	0.651
Panel D: Observations = 1000			
Single	0.089	0.179	0.268
Bonferroni	0.172	0.344	0.516
Holm	0.169	0.337	0.505
BHY	0.151	0.302	0.452

References

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

Barras, Laurent, Oliver Scaillet and Russ Wermers, 2010, False discoveries in mutual fund performance: Measuring luck in estimated alphas, *Journal of Finance 65, 179-216.*

Benjamini, Yoav and Daniel Yekutieli, 2001, The control of the false discovery rate in multiple testing under dependency, *Annals of Statistics 29, 1165-1188.*

Benjamini, Yoav and Wei Liu, 1999, A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence, *Journal of Statistical Planning and Inference 82, 163-170.*

Benjamini, Yoav and Yosef Hochberg, 1995, Controlling the false discovery rate: A practical and powerful approach to multiple testing, *Journal of the Royal Statistical Socitey. Series B 57, 289-300.*

Boudoukh, Jacob, Roni Michaely, Matthew Richardson and Michael R. Roberts, 2007, On the importance of measuring payout yield: implications for empirical asset pricing, *Journal of Finance 62, 877-915.*

Button, Katherine, John Ioannidis, Brian Nosek, Jonathan Flint, Emma Robinson and Marcus Munafò, 2013, Power failure: why small sample size undermines the reliability of neuroscience, *Nature Reviews Neuroscience 14, 365-376.*

Bailey, David, Jonathan Borwein, Marcos López de Prado and Qiji Jim Zhu, 2014, Pseudo-mathematics and financial charlatanism: The effects of backtest overfitting on out-of-sample performance, *Notices of the American Mathematical Society, May 2014, 458-471.*

Bailey, David, Jonathan Borwein, Marcos López de Prado and Qiji Jim Zhu, 2015, The probability of backtest overfitting, *Journal of Computational Finance, Forthcoming.*

Fama, Eugene F., 1991, Efficient capital markets: II, *Journal of Finance 46, 1575-1617.*

Fama, Eugene F. and Kenneth R. French, 1992, The cross-section of expected stock returns, *Journal of Finance 47, 427-465.*

Ferson, Wayne E. and Campbell R. Harvey, 1999, Conditioning variables and the cross section of stock returns, *Journal of Finance 54, 1325-1360.*

Frazzini, Andrea and Lasse Heje Pedersen, 2013, Betting against beta, *Journal of Financial Economics 111, 1-25.*

Hansen, Peter Reinhard and Allan Timmermann, 2012, Choice of sample split in out-of-sample forecast evaluation, *Working Paper, Stanford University.*

Harvey, Campbell R., Yan Liu and Heqing Zhu, 2015, ... and the cross-section of expected returns, *Review of Financial Studies, Forthcoming.*

Harvey, Campbell R. and Yan Liu, 2015, Multiple Testing in Economics, Available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2358214

Hochberg, Yosef, 1988, A sharper Bonferroni procedure for multiple tests of significance, *Biometrika 75, 800-802.*

Hochberg, Yosef and Benjamini, Y., 1990, More powerful procedures for multiple significance testing, *Statistics in Medicine 9, 811-818.*

Hochberg, Yosef and Tamhane, Ajit, 1987, Multiple comparison procedures, *John Wiley & Sons.*

Holland, Burt, Sudipta Basu and Fang Sun, 2010, Neglect of multiplicity when testing families of related hypotheses, *Working Paper, Temple University.*

Holm, Sture, 1979, A simple sequentially rejective multiple test procedure, *Scandinavian Journal of Statistics 6, 65-70.*

Hommel, G., 1988, A stagewise rejective multiple test procedure based on a modified Bonferroni test, *Biometrika 75, 383-386.*

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.*

Leamer, Edward E., 1978, Specification searches: Ad hoc inference with nonexperimental data, *New York: John Wiley & Sons.*

Lo, Andrew W., 2002, The statistics of Sharpe ratios, *Financial Analysts Journal 58, 36-52.*

Lo, Andrew W. and Jiang Wang, 2006, Trading volume: Implications of an intertemporal capital asset pricing model, *Journal of Finance 61, 2805-2840.*

López de Prado, Marcos, 2013, What to look for in a backtest, *Working Paper, Lawrence Berkeley National Laboratory.*

McLean, R. David and Jeffrey Pontiff, 2015, Does academic research destroy stock return predictability? *Journal of Finance, Forthcoming.*

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

Sarkar, Sanat K. and Wenge Guo, 2009, On a generalized false discovery rate, *The Annals of Statistics 37, 1545-1565.*

Schweder, T. and E. Spjotvoll, 1982, Plots of p-values to evaluate many tests simultaneously, *Biometrika 69, 439-502.*

Schwert, G. William, 2003, Anomalies and market efficiency, *Handbook of the Economics of Finance*, edited by G.M. Constantinides, M. Haris and R. Stulz, Elsevier Science B.V., 939-974.

Shanken, Jay, 1990, Intertemporal asset pricing: An empirical investigation, *Journal of Econometrics 45, 99-120.*

Sharpe, W.F., 1966, Mutual fund performance, *Journal of Business 39, 119-138.*

Storey, John D., 2003, The positive false discovery ratee: A Bayesian interpretation and the q-value, *The Annals of Statistics 31, 2013-2035.*

Sullivan, Ryan, Allan Timmermann and Halbert White, 1999, Data-snooping, technical trading rule performance, and the Bootstrap, *Journal of Finance 54, 1647-1691.*

Sullivan, Ryan, Allan Timmermann and Halbert White, 2001, Dangers of data mining: The case of calender effects in stock returns, *Journal of Econometrics 105, 249-286.*

Welch, Ivo and Amit Goyal, 2008, A comprehensive look at the empirical performance of equity premium prediction, *Review of Financial Studies 21, 1455-1508.*

White, Halbert, 2000, A reality check for data snooping, *Econometrica 68, 1097-1126.*

A Programs

We make the code and data for our calculations publicly available at:

http://faculty.fuqua.duke.edu/∼charvey/backtesting

A.1 Haircut Sharpe Ratios

The Matlab function *Haircut_SR* allows the user to specify key parameters to make Sharpe ratio adjustments and calculate the corresponding haircuts. It has eight inputs that provide summary statistics for a return series of an investment strategy and the number of tests that are allowed for. The first input is the sampling frequency for the return series. Five options (daily, weekly, monthly, quarterly and annually) are available.[22] The second input is the number of observations in terms of the sampling frequency provided in the first step. The third input is the Sharpe ratio of the strategy returns. It can either be annualized or based on the sampling frequency provided in the first step; it can also be autocorrelation corrected or not. Subsequently, the fourth input asks if the Sharpe ratio is annualized and the fifth input asks if the Sharpe ratio has been corrected for autocorrelation.[23] The sixth input asks for the autocorrelation of the returns if the Sharpe ratio has not been corrected for autocorrelation.[24] The seventh input is the number of tests that are assumed. Lastly, the eighth input is the assumed average level of correlation among strategy returns.

To give an example of how the program works, suppose that we have an investment strategy that generates an annualized Sharpe ratio of 1.0 over 120 months. The Sharpe ratio is not autocorrelation corrected and the monthly autocorrelation coefficient is 0.1. We allow for 100 tests in multiple testing and assume the average level

[22] We use number one, two, three, four and five to indicate daily, weekly, monthly, quarterly and annually sampled returns, respectively.

[23] For the fourth input, "1" denotes a Sharpe ratio that is annualized and "0" denotes otherwise. For the fifth input, "1" denotes a Sharpe ratio that is not autocorrelation corrected and "0" denotes otherwise.

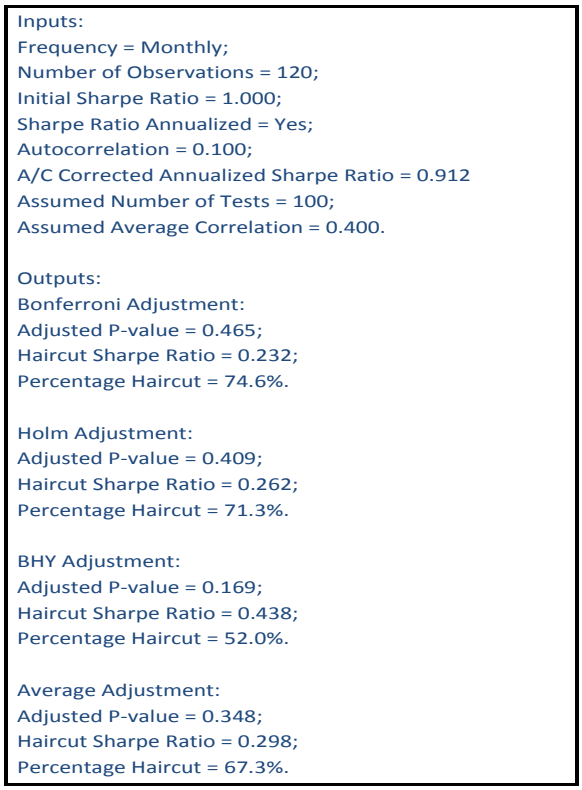
[24] We follow Lo (2002) to adjust Sharpe ratios for autocorrelation.

of correlation is 0.4 among strategy returns. With this information, the input vector for the program is

$$\text{Input vector} = \begin{bmatrix} \text{D/W/M/Q/A=1,2,3,4,5} \\ \text{\# of obs} \\ \text{Sharpe ratio} \\ \text{SR annualized? (1=Yes)} \\ \text{AC correction needed? (0=Yes)} \\ \text{AC level} \\ \text{\# of tests assumed} \\ \text{Average correlation assumed} \end{bmatrix} = \begin{bmatrix} 3 \\ 120 \\ 1 \\ 1 \\ 1 \\ 0.1 \\ 100 \\ 0.4 \end{bmatrix}.$$

Passing this input vector to *Haircut_SR*, the function generates a sequence of outputs, as shown in Figures A.1. The program summarizes the return characteristics by showing an annualized, autocorrelation corrected Sharpe ratio of 0.912 as well as the other data provided by the user. The program output includes adjusted p-values, haircut Sharpe ratios and the haircuts involved for these adjustments under a variety of adjustment methods. For instance, under BHY, the adjusted annualized Sharpe ratio is 0.444 and the associated haircut is 51.3%.

Figure A.1: **Program Outputs**



A.2 Profit Hurdles

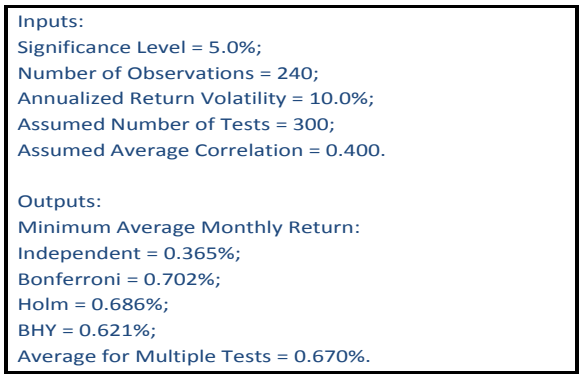
The Matlab function *Profit_Hurdle* allows the user to calculate the required mean return for a strategy at a given level of significance. It has five inputs. The first input is the user specified significance level. The second input is the number of monthly observations for the strategy. The third input is the annualized return volatility of the strategy. The fourth input is the number of tests that are assumed. Lastly, the fifth input is the assumed average level of correlation among strategy returns. The program does not allow for any autocorrelation in the strategy returns.

To give an example of how the program works, suppose we are interested in the required return for a strategy that covers 20 years and has an annual volatility of 10%. In addition, we allow for 300 tests and specify the significance level to be 5%. Finally, we assume that the average correlation among strategy returns is 0.4. With these specifications, the input vector for the program is

$$\text{Input vector} = \begin{bmatrix} \text{Significance level} \\ \text{\# of obs} \\ \text{Annualized return volatility} \\ \text{\# of tests assumed} \\ \text{Average correlation assumed} \end{bmatrix} = \begin{bmatrix} 0.05 \\ 240 \\ 0.1 \\ 300 \\ 0.4 \end{bmatrix}.$$

Passing the input vector to *Profit_Hurdle*, the function generates a sequence of outputs, as shown in Figure A.2. The program summarizes the data provided by the user. The program output includes return hurdles for a variety of adjustment methods. For instance, the adjusted return hurdle under BHY is 0.621% per month and the average multiple testing return hurdle is 0.670% per month.

Figure A.2: **Program Outputs**



B Correlation Adjustment

We use the model estimated in HLZ to provide correlation adjustment when tests are correlated.

HLZ study 316 strategies that have been documented by the academic literature. They propose a structural model to capture the underlying distribution for trading strategies. Two key features mark their model. First, there is publication bias so not all tried factors make it to publication. Second, tests may be correlated and this affects multiple testing adjustment. Taking these two conerns into account, HLZ postulate a mixture distribution for strategy returns. With probability p_0, a strategy has a mean return of zero and therefore comes from the null distribution. With probability $1-p_0$, a strategy has a nonzero mean and therefore comes from the alternative distribution. To capture the heterogeneity in strategy mean returns, HLZ assume that the mean returns for true strategies are drawn from an exponential distribution with a mean of λ. After fixing the mean returns, HLZ assume that the innovations in returns follow a Normal distribution with a mean of zero and a standard deviation of $\sigma = 15\%$ (heterogeneity in return standard deviations are captured by the heterogeneity in return means). Importantly, return innovations are correlated in the cross-section and are captured by the pairwise correlation ρ. At a certain level of correlation (ρ) between strategy returns and by matching the model implied t-ratio quantiles with the observed t-ratio quantiles, HLZ estimate the probability (p_0), the total number of trials (M), and (λ). They show that both p_0 and M are increasing as the level of correlation rises.

We use the model estimates in HLZ to approximate the underlying distribution for strategy returns. The relevant parameters for our application are ρ, p_0 and λ. HLZ provide five sets of estimates, corresponding to five levels of correlation (i.e., $\rho = 0, 0.2, 0.4, 0.6$ and 0.8). Table B.1 shows the model estimates in HLZ.

For our application, at a user specified level of ρ, we use linear interpolation to generate the parameter estimates for p_0 and λ. For example, if the user specifies $\rho = 0.3$, then the parameter estimates for p_0 and λ would be:

$$\begin{aligned} p_0(0.3) &= 0.5 \times p_0(0.2) + 0.5 \times p_0(0.4) = 0.5 \times 0.444 + 0.5 \times 0.485 = 0.465, \\ \lambda(0.3) &= 0.5 \times \lambda(0.2) + 0.5 \times \lambda(0.4) = 0.5 \times 0.555 + 0.5 \times 0.554 = 0.555, \end{aligned}$$

where $p_0(\rho)$ and $\lambda(\rho)$ denote the estimate for p_0 and λ when the correlation is set at ρ, respectively. When ρ is higher than 0.8, we interpolate based on $\rho = 0.6$ and $\rho = 0.8$, that is:

$$\begin{aligned} p_0(\rho) &= \frac{0.8-\rho}{0.2} \times p_0(0.6) + \frac{\rho-0.6}{0.2} \times p_0(0.8), \\ \lambda(\rho) &= \frac{0.8-\rho}{0.2} \times \lambda(0.6) + \frac{\rho-0.6}{0.2} \times \lambda(0.8). \end{aligned}$$

Table B.1: **Model Parameter Estimates in HLZ**

HLZ model estimates. ρ is the correlation coefficient between two strategy returns in the same period. p_0 is the probability of having a strategy that has a mean of zero. λ is the mean parameter of the exponential distribution for the monthly means of the true factors.

ρ	p_0	$\lambda(\%)$
0	0.396	0.550
0.2	0.444	0.555
0.4	0.485	0.554
0.6	0.601	0.555
0.8	0.840	0.560

When ρ is not specified, we use the preferred estimates in HLZ, i.e., $\rho = 0.2$.

The user specifies the value for ρ. We take the following steps to obtain the multiple testing adjusted Sharpe ratios:

I. We obtain the estimate for p_0 and λ using the aforementioned linear interpolation.

II. The user calculates the Sharpe ratio ($\widehat{SR}$) of the new strategy that is under consideration and specifies how many alternative strategies have been tried (N).

III. With these parameter specifications ($\rho, p_0, \lambda, \widehat{SR}, N$), we run B (=5,000) sets of simulations to find the haircut Sharpe ratio. The following steps describe the steps of the simulations:

   a. For each set of simulation, we draw N strategies based on the model in HLZ that is parameterized by ρ, p_0, and λ. In particular, with probability p_0, the strategy mean is drawn as zero; with probability $1-p_0$, the strategy mean is drawn from an exponential distribution with mean λ. The return innovations are contemporaneously correlated with a correlation coefficient of ρ and are assumed to be uncorrelated over time. All strategy returns have a volatility of $\sigma = 15\%$.

   b. We calculate the p-values for the N simulated return series and use the three multiple testing adjustment procedures described in the main text to calculate the adjusted p-value for the new strategy.

c. We take the median p-value across the B sets of simulations as the final adjusted p-value. Lastly, we convert this p-value into the haircut Sharpe ratio $\widehat{HSR}$.

Intuitively, a larger p_0 implies more flukes among the strategies and thus a higher haircut. A larger λ means that true strategies have higher means and are thus more significant. As a result, the haircut is smaller. Our model allows one to calculate exactly what level of haircut is needed for a specification of p_0 and λ.