Backtesting

Campbell R. Harvey

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

Yan Liu*

Duke University, Durham, NC 27708 USA

Current version: November 22, 2013

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.

Keywords: Sharpe ratio, Multiple tests, Backtest, Haircut

* 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 Marcos De Prado, Bernhard Scherer and Scott Linn.

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 an independent 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, 2013), 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 an independent test, it may not be if we take multiple tests into account. We proceed to calculate a p-value that appropriately reflects the multiple testing. To do this, we need to make an assumption on the number of previous tests. For example, Harvey, Liu and Zhu (HLZ, 2013) document that at least 314 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 an adjusted Sharpe ratio. The percentage difference between the original Sharpe ratio and the adjusted Sharpe ratio is the "haircut".

The Sharpe ratio that obtains as a result of the multiple testing has the following interpretation. It is the Sharpe ratio that would have resulted from an independent 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 should not be used. 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. However, 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 framework. We present results for three frameworks as well as the average of the methods. Finally, some judgment is needed setting 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.

2 Method

2.1 Independent Tests and Sharpe Ratio

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 if 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 return is 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, 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 given our assumption.

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 snooping 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 snooping biases are taken into account.[4] Profits from trading strategies that use cross-sectional

[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).

equity characteristics involve substantial statistical biases.[5] The return predictability of many previously documented variables is shown to be spurious once more advanced 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 snooping. 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 snooping 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 independent test:

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

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 tens or even hundreds of strategies and only choose to present the most profitable one. In the latter case, the p-value for the independent 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-value, p^M, for observing a maximal t-statistic that is at least as large as the observed t-ratio is

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

[5] See Leamer (1978), Lo and MacKinlay (1990), Fama (1991), Schwert (2003). A recent paper by McLean and Pontiff (2013) shows a significant degradation of performance of identified anomalies after publication.

[6] See Welch and Goyal (2004).

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

When $N = 1$ (independent test) and $p^I = 0.05$, $p^M = 0.05$, so there is no multiple testing adjustment. If $N = 10$ and we observe a strategy with $p^I = 0.05$, $p^M = 0.401$, 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 independent test. Multiple testing greatly reduces the statistical significance of independent test. Hence, p^M is the adjusted p-value after data snooping 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 independent strategies.

By equating the p-value of an independent test to p^M, we obtain the defining equation for the multiple testing adjusted Sharpe ratio $\widehat{SR}^{adj}$:

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

Since p^M is larger than p^I, $\widehat{SR}^{adj}$ 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 8.0×10^{-4} for an independent 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%.

This simple example illustrates the gist of our approach. When there is multiple testing, the usual p-value p^I for independent 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.[9]

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. Let N_r denote the total number of false discoveries, 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:

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

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 independent testing. 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 more lenient than procedures that control FWER. Essentially, FWER is designed to prevent even one error. FDR controls the error rate.[10]

[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).

[9] HLZ focus on the multiple testing adjusted threshold p-value and t-ratio, e.g., a threshold t-ratio of 3.5 at 5% significance. We focus on the entire sequence of adjusted p-values.

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

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[11] adjusts each p-value equally. It inflates the original p-value by the number of tests M:

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

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.

Holm's method[12] relies on the sequence of p-values and adjusts each p-value by:

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

Starting from the smallest p-value, Holm's method allows us to sequentially build up the adjusted p-value sequence. For example, suppose we observe $M = 3$ strategies and the ordered p-value sequence is $(0.02, 0.05, 0.20)$. To assess the significance of the first strategy, $p_{(1)}^{Holm} = 3p_{(1)} = 0.06$. This is identical to the level prescribed by Bonferroni. If our cutoff is 0.10, then this strategy is significant. The second strategy yields $p_{(2)}^{Holm} = \max[3p_{(1)}, 2p_{(2)}] = 2p_{(2)} = 0.10$, which is smaller than Bonferroni implied p-value ($p_{(2)}^{Bonferroni} = 3p_{(2)} = 0.15$). Given a cutoff of 0.10, this strategy is not significant. Finally, the least significant strategy yields $p_{(3)}^{Holm} = \max[3p_{(1)}, 2p_{(2)}, p_{(3)}] = p_{(3)} = 0.20$, which is again smaller than the one prescribed by Bonferroni ($p_{(3)}^{Bonferroni} = 3p_{(3)} = 0.60$). With 0.10 as the cutoff, this strategy is again not significant.

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 independent tests, and given a 0.10 significance level, we would "discover" two 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, $p_{(i)}^{Holm} \leq p_{(i)}^{Bonferroni}$ for any i.[13] Therefore, Bonferroni's method is tougher because it

[11] For the statistical 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).

[12] 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.

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

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 (parts failures), 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[14] defines the adjusted p-values sequentially:

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

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. Using the previous example, suppose we observe $M = 3$ strategies ($c(M) = 1.83$) and the ordered p-value sequence is $(0.02, 0.05, 0.20)$. To assess the significance of the three strategies, we start from the least significant one. BHY sets $p_{(3)}^{BHY}$ at 0.20, the same as the original value of $p_{(3)}$. For the second strategy, BHY yields $p_{(2)}^{BHY} = \min[p_{(3)}^{BHY}, \frac{3 \times 1.83}{2} p_{(2)}] = 0.14$. Finally, for the most significant strategy, $p_{(1)}^{BHY} = \min[p_{(2)}^{BHY}, \frac{3 \times 1.83}{1} p_{(1)}] = 0.11$. Notice that BHY adjusted p-value sequence $(0.11, 0.14, 0.20)$ is different from both Holm adjusted p-value sequence $(0.06, 0.10, 0.20)$ and Bonferroni adjusted p-value sequence $(0.06, 0.15, 0.60)$.

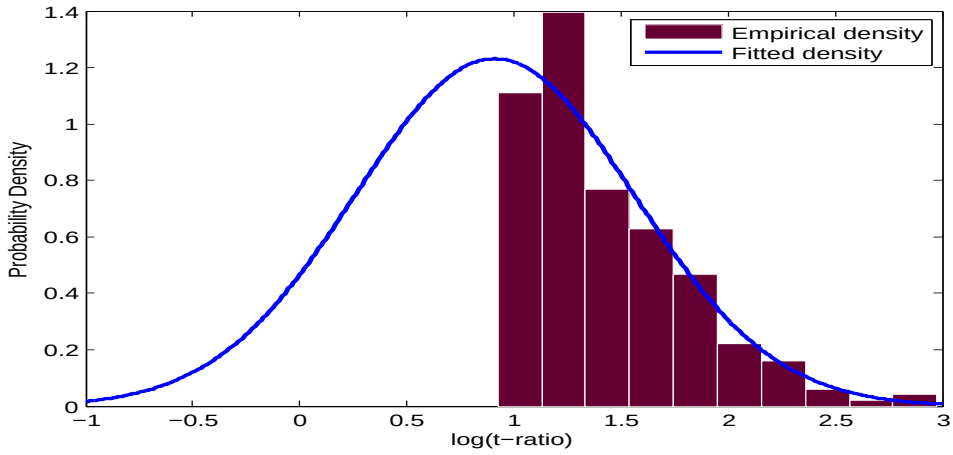
Hypothesis tests based on the adjusted p-values guarantee that the *false discovery rate* (FDR) does not exceed the pre-specified significance level. The constant $c(M)$ controls the generality of the test. In the original work by Benjamini and Hochberg (1995), $c(M)$ is set equal to one and the test works when p-values are independent or positively dependent. With our choice of $c(M)$, the test works under arbitrary dependence structure for the test statistics.

The three multiple testing procedures provide adjusted p-values that control for data snooping. 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

[14]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).

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 snooping?

Figure 1: **Fitted and Empirical densities for the Log t-ratios of Strategies**



For both Holm and BHY, we need the empirical distribution of p-values for strategies that have been tried so far. We use the mixture distribution from HLZ[15] Figure 1 shows the empirical and fitted densities for the log t-statistics of strategies.[16] Given an estimate of the number of alternative strategies, we bootstrap from this mixture distribution. In particular, assuming N strategies have been explored, we sample N p-values from the mixture distribution.[17] We then calculate the adjusted p-value for an investment strategy based on this sample of p-values. We repeat the random sampling many times, each time generating a new adjusted p-value. The median of these adjusted p-values is taken as the overall estimate of the adjusted p-value for the investment strategy.

[15]Assuming their sample of strategies fully covers published strategies that have a t-statistic above 2.5 and employing a truncated likelihood framework, HLZ estimate the underlying t-ratio distribution for all tried strategies. They estimate that the log t-statistics of explored strategy returns follow a normal distribution with mean 0.90 and standard deviation 0.66. We use this normal distribution truncated at log(2.5) to model log t-statistics that are below log(2.5). For log t-statistics that are above log(2.5), we use the empirical t-ratio distribution in HLZ. In sum, the mixture distribution is composed of a log normal distribution that is truncated at log(2.5) and an empirical distribution for t-ratios that are above 2.5.

[16]The empirical density (shaded area) has an area of one, as required by a probability density plot. The area under the fitted density, however, is multiplied by two to highlight how the right tail of the fitted density matches that of the empirical density.

[17]To make sure that the p-value sample covers p-values for significant strategies, we sample $N \times r$ p-values with replacement from the empirical distribution and sample $N \times (1 - r)$ p-values independently from the truncated normal distribution, where $r = 49\%$ is HLZ's estimate of the proportion of strategies with a t-ratio above 2.5 among all explored strategies.

3.4 Multiple Testing and Cross-validation

Recent works by De Prado and his coauthors also consider the ex-post data mining issue for standard backtests.[18] 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 one while the expected out of sample Sharpe ratio is zero. The phenomenon is analogous to the regression overfitting problem when in-sample superior models often perform 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 work 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 works are different in many ways. First, the objectives of analysis are different. Our work focuses on identifying the group of strategies that generate non-zero returns while their work evaluates the relative performance of a certain strategy that is fitted in-sample. As a result, a truly significant factor that earns a nonzero return can still be highly significant after our multiple adjustment even if all the other factors have even larger t-stats, whereas in their framework it will likely have a PBO larger than 50% (i.e., overfitting probability that is larger than 50%) because 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 (2013) needs only test statistics. In contrast, their work relies heavily on the time-series of each individual strategy. While data intensive, in the De Prado approach, it is not necessary to make assumptions regarding the data generating process for the returns. As such, their approach is closer to the machine learning literature and ours is closer to the econometrics literature.

[18] See Bailey et al. (2013a,b) and De Prado (2013).

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 tests have 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.[19] 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 data.

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 (missing discoveries) errors due to data splitting.[20] 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 based on 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 again identify 10 promising strategies. But 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 Type I and Type II errors. While a longer in-sample period reduces the chance of committing a Type II error (i.e., missing observations), it inevitably increases the chance of committing a Type I error (i.e., false discoveries) in the OOS test.

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 identify the "true" discoveries on the full data. This would involve making a more stringent cutoff for test statistics.

[19] See De Prado (2013) for a similar argument.

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

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, 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 test by looking at the intersection of survivors. We leave the exact solution to future research.

4 Applications

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 hedging 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.[21] None of these strategies reflect transaction costs and as such the Sharpe ratios are clearly somewhat overstated.

Two important ingredients to the Sharpe ratio adjustment are the initial value of the Sharpe ratio and the number of trials. To highlight the impact of these two inputs, we focus on the simplest independent case as in Section 2. In this case, the multiple testing p-value p^M and the independent testing p-value p^I are linked through Equation (4). When p^I is small, this relation is approximately the same as in Bonferroni'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 shows the summary statistics for these strategies. Among these strategies, the strategy based on E/P is 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

[21] 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).

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 independent hypothesis test that the mean strategy return is zero (t-stat $= \widehat{SR} \times \sqrt{T/12}$); p^I and p^M report the p-value for independent and multiple test, respectively; $\widehat{SR}^{adj}$ reports the Bonferroni adjusted Sharpe ratio; $\widehat{hc}$ reports the haircut for the adjusted Sharpe ratio ($\widehat{hc} = (\widehat{SR} - \widehat{SR}^{adj})/\widehat{SR}$).

Strategy	Mean(%) (monthly)	Std.(%) (monthly)	$\widehat{SR}$ (annual)	t-stat	p^I	p^M	$\widehat{SR}^{adj}$ (annual)	$\widehat{hc}$
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%

independent 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.[22] We 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.

[22]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 (independent test) is one 50th of the p-value for a t-statistic of 6.64 (multiple testing adjusted t-statistic), i.e., $p^M/p^I \approx 50$.

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,[23] or, equivalently, the p-value p^I. Assuming that N other strategies have been tried, we draw N t-statistics from the mixture distribution as in HLZ. 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^I. 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 2 shows the original vs. adjusted Sharpe ratios and Figure 3 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 (> 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 discovery rather than the absolute number.

[23] 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 $= \widehat{SR} \times \sqrt{T}$.

Figure 2: **Original vs. Adjusted Sharpe Ratios**

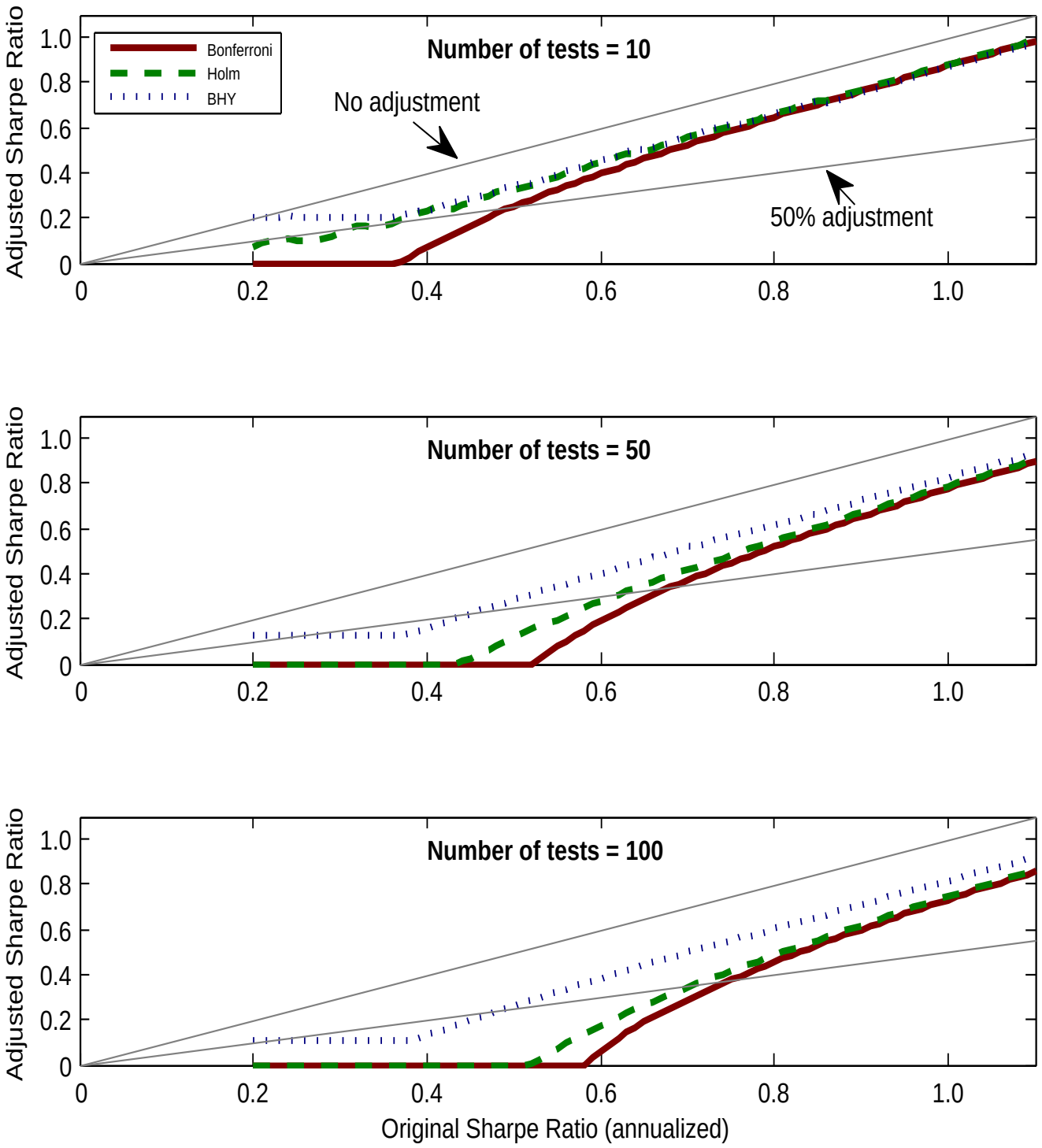
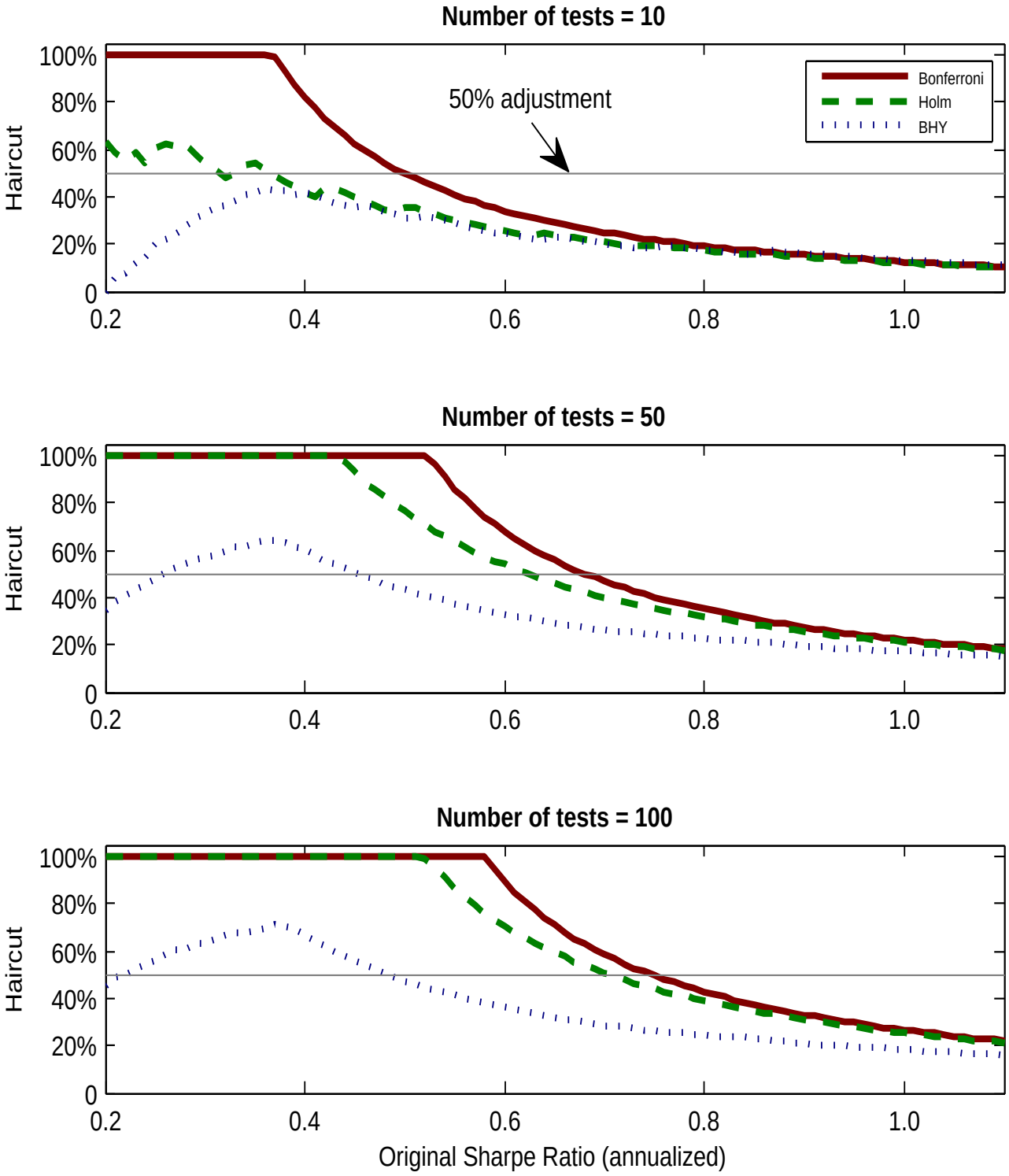


Figure 3: **Haircuts**



4.3 Adjusted VaR

Our framework allows us to adjust the backtested performance of an investment strategy. Due to multiple testing, we adjust the backtested empirical distribution of a strategy either by shifting its mean to the left and/or by inflating its variance, both of which contribute to a reduction in the Sharpe ratio. Our approach also allows us to consider the modification of other risk measures. We illustrate this by adjusting VaR (Value at Risk), a widely used measure for tail risks.[24]

We define $VaR(\alpha)$ 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:

$$\frac{VaR(\alpha)}{\sigma} = SR - z_\alpha, \tag{6}$$

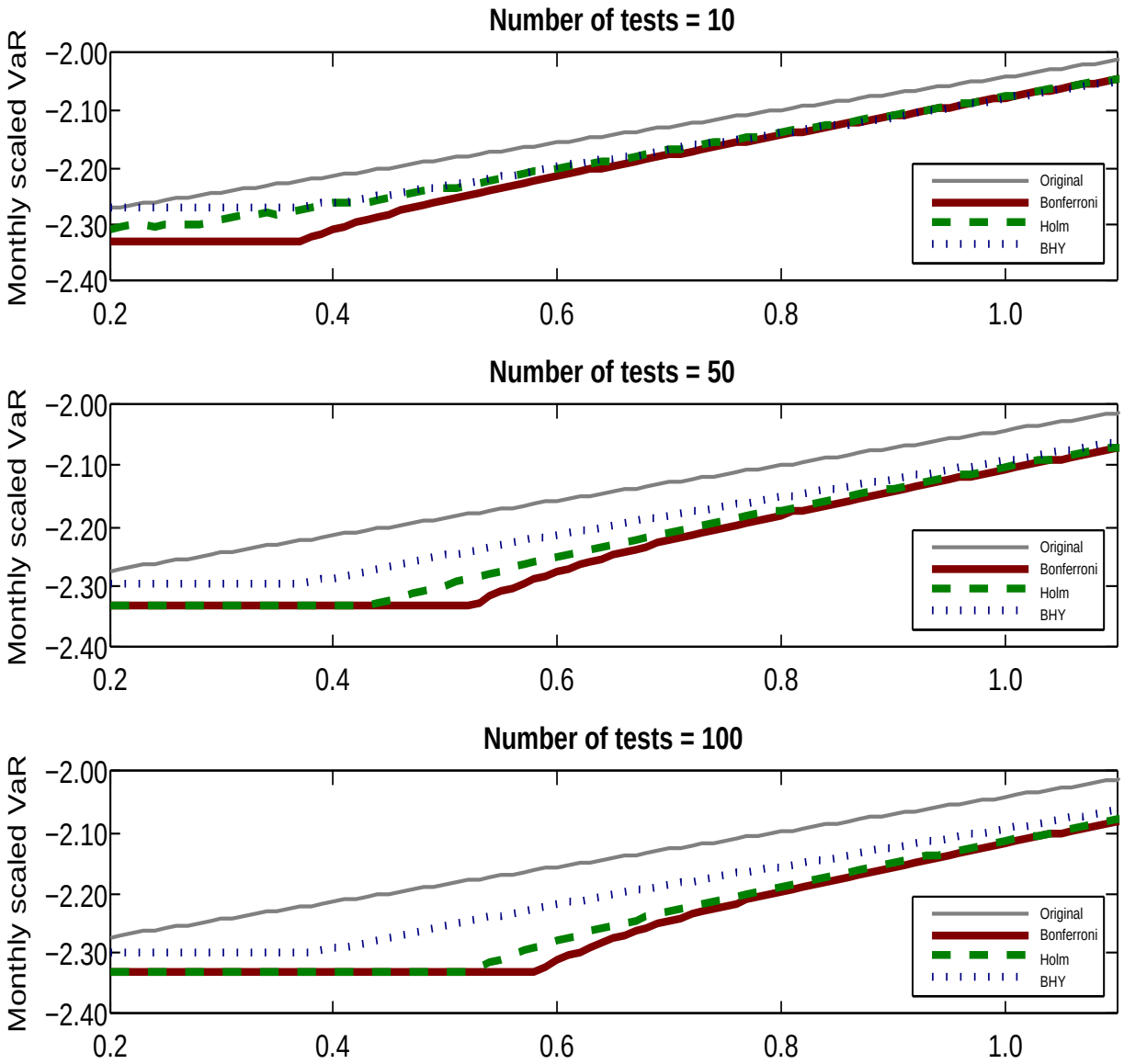
where z_α is the z-score for the $(1-\alpha)$-th percentile of a standard normal distribution and σ is the standard deviation of the return.[25] The same relationship holds for the adjusted VaR, i.e., $\frac{\widehat{VaR}(\alpha)}{\hat{\sigma}} = \widehat{SR} - z_\alpha$, where $\hat{\sigma}$ is the volatility for the adjusted returns. Due to multiple testing, the adjusted Sharpe ratio $\widehat{SR}$ is always smaller than the original Sharpe ratio SR. This implies that the adjusted VaR, scaled by the volatility for the adjusted returns, is more negative than the original VaR/σ.

Figure 4 shows the original and the adjusted VaR, scaled by their respective volatility. Since VaR is mainly interesting over relatively short investment horizons, we focus on monthly observations. The decline in the VaR/σ is substantial. For instance, when there are one hundred tests and $\alpha = 1\%$, the original VaR/σ is -1.73 for a Sharpe ratio of 0.6. The Sharpe ratio shrinks to essentially zero due to multiple testing and implies an adjusted VaR/σ of -2.33. If backtesting does not inflate return volatility, i.e., $\sigma = \hat{\sigma}$, the adjusted VaR is smaller than the original VaR by 0.6 of the return volatility.

[24] For returns that are skewed or heavy-tailed, the Sharpe ratio is a misleading measure of performance.

[25] Instead of the absolute VaR, we focus on the volatility scaled VaR as it is a function of the Sharpe ratio only.

Figure 4: $\dfrac{\mathbf{VaR}}{\sigma}$ (1%)



5 Conclusions

We provide a real time evaluation method for determining the significance of a candidate trading strategy. Our method explicitly takes into account that hundreds 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 for a t-ratio of 2. Second, we determine what the p-value should be explicitly recognizing the multiple tests that preceded the particular investment strategy. Third, based on this new p-value, we transform the corresponding t-ratio back to a Sharpe ratio. The lower Sharpe ratio explicitly takes the multiple testing or data snooping into a account. Our method is readily applied to popular risk metrics, like Value at Risk (VaR). If data mining inflates Sharpe ratios, it makes sense that VaR metrics are understated. We show how to adjust the VaR for multiple tests.

There are many caveats to our method. We do not observe the entire history of tests. 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 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) 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 strategies, out of sample tests are not available. Our method allows for decision to be made, in real time, on the viability of a particular strategy.

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 Lopez de Prado and Qiji Jim Zhu, 2013a, Pseudo-mathematics and financial charlatanism: The effects of backtest overfitting on out-of-sample, *Working Paper, Lawrence Berkeley National Laboratory.*

Bailey, David, Jonathan Borwein, Marcos Lopez de Prado and Qiji Jim Zhu, 2013b, The probability of back-test overfitting, *Working Paper, Lawrence Berkeley National Laboratory.*

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

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, *Working Paper, AQR.*

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, 2013, ... and the cross-section of expected returns, *Working Paper, Duke University.* Available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2249314.

Harvey, Campbell R. and Yan Liu, 2013, Multiple Testing in Economics, *Working Paper, Duke University.*

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.*

McLean, R. David and Jeffrey Pontiff, 2013, Does academic research destroy stock return predictability? *Working Paper, University of Alberta.*

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.

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

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.*

Appendix: The Program

We make the code and data for our calculations publicly available at http://faculty.fuqua.duke.edu/~charvey/ backtesting. The Matlab function allows the user to specify key parameters for our procedure and investigate the impact on the Sharpe ratio. The function *SR_adj_multests* has seven 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.[26] 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 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 is corrected for autocorrelation.[27] The sixth input asks for the autocorrelation of the returns if the Sharpe ratio has not been corrected for autocorrelation.[28] Lastly, the seventh input is the number of tests that are assumed.

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. With these information, the input vector for the program is

$$\text{Input vector} = [3, 120, 1, 1, 0, 0.1, 100]'.$$

Passing this input vector to *SR_adj_multests*, the function generates a sequence of outputs, as shown in Figures 4 and 5. For the intermediate outputs in Figure 4, the program summarizes return characteristics by showing an annualized, autocorrelation corrected Sharpe ratio of 0.912 that is based on 120 month of observations. For the final outputs in Figure 5, the program generates adjusted p-values, adjusted 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 0.612 and the associated haircut is 33.0%.

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

[27] 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 autocorrelation corrected and "0" denotes otherwise.

[28] We follow Lo (2002) to adjust Sharpe ratios for autocorrelations.

Figure 5: **Intermediate outputs**

```
Intermediate Outputs:
Annualized Sharpe ratio = 0.912;
Based on 120 monthly observations.
```

Figure 6: **Final outputs**

```
Final Outputs:
Bonferroni adjustment:
Adjusted p-value = 0.465;
Adjusted Sharpe ratio = 0.232;
Haircut = 74.6%.

Holm adjustment:
Adjusted p-value = 0.274;
Adjusted Sharpe ratio = 0.347;
Haircut = 61.9%.

BHY adjustment:
Adjusted p-value = 0.055;
Adjusted Sharpe ratio = 0.612;
Haircut = 33.0%.

Average adjustment:
Adjusted p-value = 0.265;
Adjusted Sharpe ratio = 0.354;
Haircut = 61.2%.
```