# High-Throughput Asset Pricing

**Abstract**

We use empirical Bayes (EB) to mine data on 140,000 long-short strategies constructed from accounting ratios, past returns, and ticker symbols. This "high-throughput asset pricing" produces out-of-sample performance comparable to strategies in top finance journals. But unlike the published strategies, the data-mined strategies are free of look-ahead bias. EB predicts that high returns are concentrated in accounting strategies, small stocks, and pre-2004 samples, consistent with limited attention theories. The intuition is seen in the cross-sectional distribution of t-stats, which is far from the null for equal-weighted accounting strategies. High-throughput methods provide a rigorous, unbiased method for documenting asset pricing facts.

**JEL Classification**: G0, G1, C1

**Keywords**: stock market predictability, stock market anomalies, p-hacking, multiple testing

# 1 Introduction

Data mining refers to searching data for interesting patterns. This search leads to data mining bias, if many patterns are just chance results, as is surely the case with stock return data. To address this problem, the asset pricing literature recommends restricting the search to patterns consistent with theory (Cochrane (2005) and Harvey (2017)). However, recent empirical evidence finds this method is ineffective, even for theories published in top finance journals (Chen, Lopez-Lira, and Zimmermann (2022)).

We offer a different solution. Instead of mining data less, we recommend mining data *rigorously*. Rigorous data mining means conditioning interesting results on the fact that they come from searching through data. This conditioning can be achieved using empirical Bayes (Robbins (1956), Efron and Morris (1973), and Efron (2012)). Rigorous data mining also means that the search should be systematic, as is commonly done in high-throughput biology and chemistry (Yang et al. (2021)). Ironically, systematic search implies that asset pricing should involve *more* data mining, not less.

We use empirical Bayes (EB) to mine for out-of-sample returns among 140,000 long-short trading strategies. The trading strategies are constructed from systematically searching data on accounting ratios, past returns, and stock tickers. Through this "high-throughput asset pricing," we construct a portfolio with out-of-sample returns that are comparable to the returns from the best journals in finance.

Our data-mined portfolio is the simple average of the top 1% of strategies, based on EB-predicted Sharpe ratios. It earns out-of-sample returns of 5.7% per year over the 1983-2020 sample, compared to the mean return across 200 published strategies in the Chen and Zimmermann (2022) dataset of 5.9% per year. But unlike the published strategies, which were selected with knowledge of stock return patterns that occurred in the 1980s and 1990s, our strategies can be constructed using only information available in real time. Thus, unlike the published strategies, our data mined strategies are arguably free of look-ahead bias.

The composition of the top 1% portfolio provides insights into the nature of return predictability. 91.0% of strategies in this portfolio are equal-weighted accounting ratio strategies. Almost all of the remainder are equal-weighted past-return strategies. Moreover, the returns of the top 1% strategy are concentrated in the pre-2004 data. These facts are consistent with the theory that predictabil-

ity is largely due to limited attention and the slow incorporation of information into stock prices (Peng (2005); Chordia, Subrahmanyam, and Tong (2014)).

Other facts shed light on the drivers of the recent decline in cross-sectional predictability. We find that the returns of the top 5% and top 10% of portfolios are also concentrated in the pre-2004 data. These strategies are enormous in number: the top 5% consists of 6,305 strategies, and the top 10% consists of 12,610. As many of these strategies are unlikely to be found in academic journals, this suggests that the key driver of the recent declines in predictability is improvements in information technology (Chordia, Subrahmanyam, and Tong (2014)), rather than investors learning from academic publications (McLean and Pontiff (2016)). Consistent with this idea, we find that the top 20 strategies according to predicted Sharpe ratios using data available in 1993 have themes rarely seen in academic journals, like mortgage debt, growth in interest expense, and depreciation. Themes that were popular in academia in 1993, like book-to-market, momentum, and sales growth are missing from this list.

Overall, high-throughput asset pricing provides not only a method for dealing with look-ahead bias, but also a more rigorous method for documenting asset pricing facts. We post our strategy returns and code publicly, and encourage future researchers to use these methods.

Unlike many big data methods, our EB formulas provide a transparent intuition. In essence, our EB formulas measure the distance between the empirical t-stat distribution and the standard normal null. Ticker-based strategies have t-stats that are extremely close to the null, implying no predictability. In contrast, equal-weighted accounting t-stats are too fat tailed to be consistent with the null, implying strong predictability. Thus, just by visually inspecting the t-stat distributions, one can see where predictability is concentrated.

EB provides highly accurate predictions in pre-2004 data. We construct 120 portfolio tests using the 140,000 data-mined strategies, and compare EB-predicted returns with out-of-sample returns. In almost all of the 120 portfolios, the EB predictions are within 2 standard errors of the out-of-sample mean.

Post-2004, EB has more difficulty with accuracy, though it still captures broad patterns in out-of-sample returns. Compared to pre-2004, predicted returns are closer to zero, and only equal-weighted accounting strategies show notable predicted returns. However, out-of-sample returns are even closer to zero than predicted. This difficulty might be expected given that the rise of information tech-

nology around 2004, which likely led to a structural break in predictability (Chordia, Subrahmanyam, and Tong (2014) and Kim, Ivkovich, and Muravyev (2021)). Our EB predictions are constructed using a simple 20-year rolling window, and thus fail to account for this break. This difficulty suggests that a smart data miner armed with theory might have understood the implications of the internet, and could perhaps have performed much better than our theory-free EB mining process.

We also illustrate how improper use of multiple testing statistics can lead to poor data mining results. We demonstrate this possibility using Harvey, Liu, and Zhu's (2016) recommended methodology for false discovery control. Harvey et al. recommend applying Benjamini and Yekutieli's (2001) Theorem 1.3 to construct a t-stat hurdle that controls the FDR at the 1% level. Nearly all of our 140,000 trading strategies fail to meet this hurdle, suggesting that there are no interesting patterns in this data. But in fact, simple out-of-sample tests show there are thousands of strategies with notable out-of-sample returns. In contrast, the Storey (2002) FDR control, recommended in Barras, Scaillet, and Wermers (2010), captures the majority of notable portfolios.

Fortunately, this error can be avoided by rigorously studying the statistics. According to Benjamini and Yekutieli (2001), their Theorem 1.3 is "very often unneeded, and yields too conservative of a procedure." This negative sentiment is echoed in Efron's (2012) textbook on large scale inference. In contrast, the EB methods we use are recommended for settings like ours in Chapter 1 of Efron (2012), as well as Chapters 6 and 7 of Efron and Hastie (2016).[1]

## 1.1 Related Literature

We add to Yan and Zheng (2017) and Chen, Lopez-Lira, and Zimmermann (2022), who document that mining accounting data can produce substantial out-of-sample returns. Accounting data is important: Chen et al. find that mining ticker variables leads to out of sample returns of approximately zero. Thus, one needs a method for identifying the predictive power of accounting data in real time. Our empirical Bayes formulas provide one such method.

The literature on multiple testing in asset pricing features disagreement on both the methods that should be used and the empirical extent of multiple test-

[1] A brief explanation of why Benjamini and Yekutieli (2001) Theorem 1.3 is excessively conservative is found in Section 2.5 of Chen (2023).

ing problems. Chen and Zimmermann (2020); Chen and Velikov (2022); and Jensen, Kelly, and Pedersen (2023) recommend empirical Bayes shrinkage. In contrast, Harvey, Liu, and Zhu (2016); Harvey and Liu (2020); and Chordia, Goyal, and Saretto (2020) recommend conservative false discovery controls, much more conservative than the FDR methods in Barras, Scaillet, and Wermers (2010). We show how empirical Bayes shrinkage and the recommended method from Barras, Scaillet, and Wermers (2010) leads to much more accurate inferences compared to the recommended FDR method in Harvey, Liu, and Zhu (2016).

In contrast to the intuition that simplicity is a virtue, we find that studying an enormous number of potential predictors leads to insights about the nature of return predictability. A similar theme is found in Kelly, Malamud, and Zhou (2024) and Didisheim et al. (2023), who illustrate the "virtue of complexity" in the modeling of expected returns.

# 2 Data and Methods

We describe the data (Section 2.1) and how we rigorously mine it (Sections 2.2-2.3).

## 2.1 Data on 140,000 Trading Strategies

Table 1 describes our data-mined strategies. The strategies are either based on accounting ratios, past returns, or tickers. Accounting ratio strategies are taken from Chen, Lopez-Lira, and Zimmermann (2022).[2] The past return and ticker strategies are inspired by Yan and Zheng (2017) and Harvey (2017), respectively, but we generate our own strategies in order to ensure that the number of strategies is comparable across data sources and to ensure that each type of strategy consists of many distinct strategies.[3]

[Table 1, Overview of Trading Strategies, about here]

A key feature of these strategies is that they are *not* selected based on having notable historical returns. Instead, they are constructed to systematically explore various types of data. So unlike most datasets in asset pricing (e.g. Ken

[2] We are grateful to that the authors make their data publicly available.

[3] Results that mine data following Yan and Zheng (2017) and Harvey (2017) are similar and can be found in the first draft of our paper on arxiv.org or via our github site.

French's size- and B/M-sorted portfolios; Chen and Zimmermann (2022)), ours is arguably free of data mining bias. Indeed, Table 1 shows that the median sample mean return is close to zero for all sets of strategies.

In high-throughput research, the median measurement is relatively unimportant. What matters is that the extreme measurements show promise for, say, a pharmaceutical intervention or cancer prediction. The extreme measurements in Table 1 suggest that accounting and past return data show promise for predicting returns. These data lead to mean returns that can exceed 5 percent per year in absolute value.

For further details on the strategy definitions, see Appendix A or our github site.

## 2.2 Empirical Bayes Overview

The 140,000 strategies in Table 1 contain the potential for significant data mining bias. To understand the bias, decompose the sample mean return on strategy $i$ as follows:

$$\tilde{r}_i = \tilde{\mu}_i + \tilde{\varepsilon}_i \tag{1}$$

where $\tilde{r}_i$ is the sample mean return, $\tilde{\mu}_i \equiv E(\tilde{r}_i)$ and $\tilde{\varepsilon}_i \equiv \tilde{r}_i - E(\tilde{r}_i)$.

Data mining involves selecting $i$ which has a large $\tilde{r}_i$. To formalize this, suppose $r \gg 0$ (e.g. $r = 10\%$ per year), and we search for a strategy $i^*$ that satisfies $\tilde{r}_{i^*} = r$. Naively, one might think that $r$ is an unbiased estimate of $\tilde{\mu}_{i^*}$. However, it is in general upward biased:

$$\begin{aligned} r &= E(\tilde{r}_{i^*} | \tilde{r}_{i^*} = r) \\ &= E\left(\tilde{\mu}_{i^*} | \tilde{r}_{i^*} = r\right) + E(\tilde{\varepsilon}_{i^*} | \tilde{r}_{i^*} = r) \\ &> E\left(\tilde{\mu}_{i^*} | \tilde{r}_{i^*} = r\right). \end{aligned} \tag{2}$$

Selecting for large $\tilde{r}_i$ selects for large $\tilde{\varepsilon}_i$, leading to $E(\tilde{\varepsilon}_{i^*} | \tilde{r}_{i^*} = r) > 0$, and the bias in Equation (2).

To data mine rigorously, one should estimate and remove the bias term $E(\tilde{\varepsilon}_{i^*} | \tilde{r}_{i^*} = r)$. This bias is just a conditional expectation, so it can be computed using Bayes rule given a probability model.

Let $\Omega$ represent a vector of parameters for the probability model behind

$\{\tilde{r}_i, \tilde{\mu}_i\}_{i=1}^N$, where $N$ is the total number of strategies. One can remove the bias by computing

$$E\left(\tilde{\mu}_{i^*} | \tilde{r}_{i^*} = r; \hat{\Omega}\right) \equiv r - E\left(\tilde{\varepsilon}_{i^*} | \tilde{r}_{i^*} = r, \hat{\Omega}\right) \tag{3}$$

where $\hat{\Omega}$ is a consistent (frequentist) estimate of the probability model parameters. This method, of applying frequentist estimates to Bayesian formulas is known as "empirical Bayes" (Robbins (1956) and Efron and Morris (1973)).

For our main results, we use a slight generalization of Equation (3):

$$\hat{\mu}(r,t) \equiv E\left(\tilde{\mu}_i | \tilde{r}_i = r, \tilde{t}_i = t; \hat{\Omega}\right), \tag{4}$$

where $\tilde{t}_i$ is the (random) t-stat for strategy $i$ and $t$ is a constant. We use Equation (4) to search 140,000 long-short strategies for large expected returns. We will not use economic theory to determine the probability model, and thus our search is largely atheoretical. However, we recognize the bias that comes from such a search (Equation (2)), and carefully correct for it. Thus, we describe our methods as "rigorous data mining."

## 2.3 Probability Model and Estimation

We model the t-stat on strategy $i$ as follows:

$$\tilde{t}_i = \tilde{\theta}_i + \tilde{\delta}_i \tag{5}$$

$$\tilde{\delta}_i \sim \text{Normal}(0,1), \tag{6}$$

where $\tilde{t}_i \equiv \tilde{r}_i / \text{SE}_i$, $\text{SE}_i$ is the standard error of $\tilde{r}_i$, and $\tilde{\theta}_i \equiv \tilde{\mu}_i / \text{SE}_i$ is the standardized expected return. Equations (5) and (6) can be derived from dividing Equation (1) by $\text{SE}_i$ and assuming the CLT holds. For simplicity, we assume $\text{SE}_i$ is known.

We then model the standardized expected return as a mixture of normals:

$$\tilde{\theta}_i \sim \begin{cases} \text{Normal}\left(\mu_{\theta,a}, \sigma^2_{\theta,a}\right) & \text{with prob } \lambda_a \\ \text{Normal}\left(\mu_{\theta,b}, \sigma^2_{\theta,b}\right) & \text{otherwise} \end{cases}. \tag{7}$$

Mixture normals are parsimonious, easy to understand, and yet allow for skewness and fat tails. We then estimate $\Omega \equiv \left[\mu_{\theta,a}, \sigma^2_{\theta,a}, \mu_{\theta,b}, \sigma^2_{\theta,b}, \lambda_a\right]$ using quasi-

maximum likelihood. Plugging in the estimate $\hat{\Omega}$ into Bayes rule (Equation (4)) yields our expected return estimates.

We compute Bayes rule using the distr package (Ruckdeschel et al. (2006)) for generality.[4] But closed form solutions for mixture normal setting can be found in Section 3.3 and Appendix B.2.

The quasi-likelihood is also computed using distr. Estimation of $\Omega$ is done using nloptr (Johnson (2007)). For further details see Appendix B or our github site.

We estimate the models using the past 20 years of long-short returns, separately for each year spanning 1983-2019 and for each strategy "family" and for each year. There are six families of strategies formed by crossing the three types of signals (accounting, past returns, tickers) with two portfolio formation methods (equal-weighted and value-weighted).

# 3 Performance of the Best Data Mined Strategies

We show that rigorous data mining leads to research-like out-of-sample returns (Section 3.1) and take a look at which kinds of strategies are identified by data mining (Section 3.2). We also provide intuition for how rigorous mining works (Section 3.3).

## 3.1 Out-of-Sample Returns

Can rigorous data mining generate out-of-sample returns? To answer this question, we examine portfolios that take advantage of the most extreme EB predictions.

Each year, we sign strategies to have positive EB predicted returns, and then form portfolios that equally-weight strategies in the top $X$% of predicted Sharpe ratios. The predicted Sharpe ratio is defined as the EB predicted return (Equation (4)) divided by the in-sample return volatility. We examine $X$ = 1, 5, and 10. For comparison, we also examine a portfolio that equally-weighs published strategies from the Chen and Zimmermann (2022) dataset.

Table 2 shows the result. The top 1% of data-mined strategies perform simi-

[4] We expect to use other parametric assumptions for Equation (7) in future revisions for robustness.

larly to strategies published in top finance journals. Over the full 1983-2020 sample, the top 1% portfolio earns 5.70% per year, compared to the 5.88% return from published strategies. The Sharpe ratio from data mining is smaller, at 1.46 vs 2.03 for published strategies. However, unlike the data-mined strategies, which are formed using only information available in real-time, the published strategies contain look-ahead bias. Indeed, if we focus on strategies in top journals that were published pre-2004, the performance is very similar to the data-mined strategies in terms of either mean returns or Sharpe ratios.

[Table 2, Returns of Long-Short Portfolios Data-Mined, about here]

The data-mined returns are robust. The top 5% and top 10% of data-mined strategies also perform well and are statistically significant, indicating that the performance of the top 1% is not driven by outliers. Out-of-sample performance is also seen in both the 1983-2004 and 2005-2020 subsamples. And across all samples, the performance of data mining is comparable to the performance of strategies from finance journals.

Figure 1 takes a closer look by plotting the value of $1 invested in each portfolio over time. The top 1% data-mined portfolio has similar performance to published strategies throughout the figure. All portfolios show little relatively little cyclicality during the recessions of 1991, 2009, and 2020. Indeed, the returns are fairly consistent throughout the chart, with an important caveat.

[Figure 1, Cumulative Long-Short Returns, about here]

The caveat is that returns are concentrated in the pre-2004 sample. This is seen in the flattening of the solid line in Figure 1 around 2004 and in the middle and lower panels of Table 2. The top 1% portfolio returns 8.17% per year from 1983-2005, compared to just 2.03% from 2005-2020. A similar decay is seen across all portfolios, both data-mined and academic.

Overall, we find that one can find long-short returns comparable to those from the best journals in finance, just by mining data, with little thought about the underlying economics. Moreover, rigorous data mining can discriminate between data sources that have no information about future returns, like stock market tickers, from data that is rich in information, like accounting ratios. Unlike the published strategy returns, our returns can be found using only information available in real-time. These results show that high-throughput methods provide a bias-free approach to studying stock market predictability. Our

strategy returns and code are public, and we encourage future researchers to use these methods.

## 3.2 The Composition of the Top 1%

Table 3 takes a closer look at the top 1% strategies produced by rigorous data mining. Panel A shows that 91.0% of the top 1% come from the equal-weighted accounting family and 8.6% come from equal-weighted past returns. The other strategy families comprise a negligible part of the top 1%. Ticker strategies are completely absent from the top 1%.

[Table 3, Description of Top 1% Data-Mined Strategies, about here]

Taken with Table 2, these results show that cross-sectional predictability is concentrated in accounting data, small stocks, and pre-2004 samples. These stylized facts offer a parsimonious description of the "factor zoo." Theories that wish to capture the big picture of cross-sectional predictability should be consistent these facts. For example, slow diffusion of economic information is consistent, as this diffusion would be especially slow in small stocks and before the internet era. In this way, high throughput asset pricing provides a way to not only identify out-of-sample returns, but to also provide insight into the underlying economics.

Panel B shows that many of the top 1% strategies are quite far from the predictors noted in the academic literature. In 1993, academics were focused on predictors like book-to-market, 12-month momentum, and sales growth (Fama and French (1992); Jegadeesh and Titman (1993); Lakonishok, Shleifer, and Vishny (1994)). None of these predictors are in the top 20 strategies based on predicted Sharpe ratios from rigorous data mining. Instead, the common themes from data mining include shorting stocks with high or growing debt, as well as buying stocks with high depreciation, depletion, and amortization. Another theme is buying stocks with high returns in quarters $t$ minus 17 and 18.

Based on textbook risk-based or behavioral asset pricing, one might expect that these data-mined predictors will average zero returns out-of-sample. But this is not the case. The realized Sharpe ratios for these strategies in the 10 years after 1993 averages around 1.0 ("SR OOS" column).

### 3.3 Shrinkage Intuition

Unlike many big data and machine learning methods, empirical Bayes has a transparent intuition. The intuition can be seen in a special case of the prediction Equation (4). If $\tilde{\theta}_i \sim \text{Normal}\left(0, \sigma_\theta^2\right)$, we have

$$\hat{\mu}(r,t) = \left[1 - \frac{1}{\widehat{\text{Var}}\left(\tilde{t}_i\right)}\right] r, \tag{8}$$

where $\widehat{\text{Var}}\left(\tilde{t}_i\right)$ is an estimate of the cross-strategy variance of t-stats.

This expression says that rigorous mining involves shrinking sample mean returns $\tilde{r}_i$ toward zero at a rate of $\frac{1}{\widehat{\text{Var}}\left(\tilde{t}_i\right)}$. $\widehat{\text{Var}}\left(\tilde{t}_i\right)$ measures how far the data are from the null of $\tilde{t}_i \sim \text{Normal}(0,1)$. If there is no predictability, then $\tilde{t}_i \sim \text{Normal}(0,1)$, $\widehat{\text{Var}}\left(\tilde{t}_i\right) \approx 1$, and all $\tilde{r}_i$ are shrunk to zero. But if data are far from the null, then a large $\tilde{r}_i$ is a signal of large $\tilde{\mu}_i$—even if $\tilde{r}_i$ is found from searching tens of thousands of strategies, unguided by economic theory.

Figure 2 shows that equal-weighted accounting strategies (upper left) are far from the null using data from 1964 to 1983. Equal-weighted past return strategies (middle left) also show a notable deviation. In contrast, the other strategy families are quite close to the null. Indeed, for both families of ticker-based strategies, the null is a very good fit for the data.

[Figure 2, Distribution of t-stats in 1983, about here]

Accordingly, Equation (8) implies that the bulk of the high return strategies will be found in equal-weighted accounting and equal-weighted past-return strategies. This intuition is consistent with Panel A of Table 3, which shows that the vast majority of the best data-mined strategies come from these families.

Compared to data available in 1983, all strategy families are closer to the null using data from 1985-2004, as seen in Figure 3. All value-weighted families are very close to the null, implying that predictability in large stocks is essentially gone. The long left tail in equal-weighted past return strategies also disappears. Only equal-weighted accounting strategies are visually far from the null. These results imply that predictability is concentrated in the earlier part of the sample.

[Figure 3,Distribution of t-stats in 2004, about here]

The intuition in Figures 2 and 3 is so simple that one might even skip the quasi-maximum likelihood estimation. Just looking at these charts, and the dis-

tance between the data and the null, one can already tell that predictability is concentrated in small stocks, accounting data, and the earlier sample. That is, one can already tell where predictability is concentrated, if one understands the intuition in Equation (8).

# 4 Empirical Bayes Prediction Accuracy Across the Cross-Section

This section takes a closer look at the EB predictions and accuracy. We see when and where EB predictions are successful and when they struggle.

## 4.1 EB Prediction Accuracy 1983-2004

To examine accuracy, we use out-of-sample portfolio sorts. For each year and each strategy family, we form 20 portfolios by sorting strategies into equal-sized groups based on the past 20 years of mean returns. We then predict the mean returns for each portfolio by averaging the EB predictions (Equation (4)), which are also based on the past 20 years of data. Finally, we form a portfolio that equally-weighs strategies in each group and hold for one year (the "out-of-sample" periods).

Figure 4 shows the in-sample, predicted, and out-of-sample returns for each portfolio, averaged the out-of-sample periods from 1983 to 2004. For all six families, there are sizable in-sample returns (dashed line) in the extreme in-sample groups. For accounting strategies, in-sample returns are as extreme as -11% per year. A naive read of this result is that one can flip the long and short legs and find +11% returns out-of-sample. Past return strategies see a similar ±10 percent return in the extreme groups. Even ticker-based strategies show in-sample long-short returns of up to 4 percent per year.

[Figure 4, Empirical Bayes Predictions 1983-2004, about here]

However, the predicted returns are typically much closer to zero. In fact, for both ticker-based strategy families, the predicted return (solid line) is almost exactly zero for all 40 in-sample groups. This result is intuitive given how close the ticker t-stats are to the null of no predictability (Figure 2). This closeness implies that the extreme returns can be entirely accounted for by luck, and so

shrinkage should be 100% (Equation (8)). Significant shrinkage is also seen in value-weighted accounting strategies (top right panel). Rigorous data mining recommends that the extreme returns of around -8% and +9% (dashed line) be shrunk down to about -3% and +2 (solid line), respectively.

Rigorous mining predicts much higher returns in equal-weighted accounting strategies (upper left panel). For these strategies, the predicted returns are actually not far from the in-sample return. This result is consistent with Chen and Zimmermann (2020), who find shrinkage of only 12% for published anomalies, which are largely equal-weighted and based on accounting variables. Predictability is also seen in both families of past return strategies.

These predictions are borne out in out-of-sample returns (markers with error bars). The first group of EW accounting strategies returns -8 percent per year out-of-sample from 1983-2004, almost exactly the same as the EB prediction. Similar accuracy is seen throughout all 120 bins in Figure 4.

These results show that rigorous data mining offers economic insights that are difficult to derive from theory. While theories of slow information diffusion may tell you that predictability is concentrated in small stocks, accounting signals, and pre-2004 data, they are unlikely tell you how much predictability there is. In contrast, empirical Bayes provides quantitative, accurate estimates of the precise amount of predictability.

## 4.2 EB Prediction Accuracy 2004-2020

We split our OOS tests in the mid-2000s, motivated by the idea that there was likely a structural break during this period due to the rise of information technology (Chordia, Subrahmanyam, and Tong (2014)). Comparing the distribution of t-stats available in 1983 vs 2004 supports the idea that the structure of financial markets changed (see Section 3.3).

[Figure 5, Empirical Bayes Predictions 2004-2020, about here]

This structural change can be seen by comparing Figure 5 (EB predictions 2004-2020) to Figure 4 (EB predictions 1983-2004). In all panels, the predicted returns shift closer to zero post-2004. Most notably, the predictability that was present in past return strategies pre-2004 is largely gone. Consistent with these predictions, the past return portfolios show a flat or even negative relationship between out-of-sample and in-sample returns post-2004. A similar weakening of

EB predictions and flattening of out-of-sample returns is seen in the accounting VW family.

An exception to this pattern is the family of equal-weighted accounting ratio strategies (top left). In this chart, the shrinkage is still relatively small, with EB predictions implying returns as extreme as -9 percent per year. This prediction and others in this panel miss the mark: the out-of-sample returns are much closer to zero throughout this panel.

This poor accuracy is natural given the fact that the estimations use a rolling window consisting of the past 20 years of data. This fixed window implies that, for much of the period 2004-2020, our estimates rely on data from a time when accounting statements needed to be retrieved by traditional (snail) mail for investors without special access to the SEC reading room (Bowles et al. (2023)).

This result implies an important role for economic theory: when structural breaks occur, there is no way for data mining to provide a clear understanding of the economy, no matter how rigorously the mining is done. Theory is sometimes used this way in economics and finance, but this is typically not the case. Instead, theory is typically used to understand patterns found in long samples of data, spanning many decades. In our view, the future of theory is bright for theorists who study structural breaks, even in the era of big data. Indeed, a smart data miner armed with theory might have understood the implications of the internet for stock return predictability, and could perhaps have performed much better than our theory-free EB mining process.

# 5 Comparison with False Discovery Controls

Our main analysis corrects for data mining bias using empirical Bayes shrinkage, following Chen and Zimmermann (2020); Chen and Velikov (2022); and Jensen, Kelly, and Pedersen (2023). An alternative approach is to use false discovery controls, following Harvey, Liu, and Zhu (2016); Harvey and Liu (2020); and Chordia, Goyal, and Saretto (2020). This section examines how our results would differ if we use this alternative.

## 5.1 Harvey, Liu, and Zhu (2016)'s Multiple Testing Controls

Harvey, Liu, and Zhu (2016) (HLZ) recommend using Benjamini and Yekutieli's (2001) Theorem 1.3 to control for multiple testing. This theorem provides an estimate of the t-stat hurdle required to ensure an FDR below $q^*$, where $q^*$ is selected by the researcher. HLZ recommend $q^* = 1\%$, though they also examine $q^* = 5\%$.

We describe this as the "HLZ method" because the original paper that proves this theorem does not recommend using it. Benjamini and Yekutieli (2001) describe their theorem as "very often unneeded, and yields too conservative of a procedure." In his textbook on large scale inference, Efron (2012) agrees, stating that the theorem represents a "severe penalty" and is "not really necessary." Moreover, the statistics literature uses the "BY algorithm" to refer to Benjamini and Yekutieli (2005), which is an entirely different procedure.

It is important to examine the HLZ method, since it is arguably the most popular multiple testing control in finance. Several followups to the influential HLZ paper use this method, including Harvey and Liu (2020) and Chordia, Goyal, and Saretto (2020); and Jensen, Kelly, and Pedersen (2023).

We implement HLZ's recommendation as follows: for each year and each strategy family, we solve

$$h_{\text{HLZ},q^*} \equiv \min_{h>0} \left\{ h : \left[ \frac{\Pr(|t_1| > h | \theta_1 = 0)}{\text{Share of } |\tilde{t}_i| > h} \right] \pi_{\text{BY1.3}} \leq q^* \right\} \tag{9}$$

where

$$\pi_{\text{BY1.3}} \equiv \sum_{i=1}^{N} \frac{1}{i} \tag{10}$$

and $N$ is the number of strategies in the year-family. If there is no $h > 0$ that satisfies the constraint, we set $h_{HLZ,q^*}$ = the maximum $|\tilde{t}_i| + 1$ (we do not reject the null for any strategy). Benjamini and Yekutieli's (2001) Theorem 1.3 proves that this algorithm implies a false discovery rate $\leq q^*$, though the bulk of the paper studies Theorem 1.2, which uses $\pi_{\text{BY1.3}} = 1$ instead of Equation (10).

We evaluate HLZ's recommendation using out-of-sample portfolio sorts, as in Section 3. For each year and each strategy family, we sort strategies into 20 groups based on the in-sample t-statistic. We then form portfolios that equally weight strategies in each group and hold for one year.

Figure 6 shows the result. It shows the time-series average of the out-of-sample portfolio returns, as well as the time-series average of $h_{HLZ,q^*}$, for $q^* =$ 1% or 5%. Only one of the 120 portfolios survives HLZ's recommended $q^* = 1\%$ multiple testing control (the most negative t-stat in the equal-weighted accounting family). This portfolio has notable out-of-sample returns of -6% per year and is highly statistically significant. So in this way, one can judge HLZ's recommendation as successful.

[Figure 6, Multiple Testing Following HLZ, about here]

However, HLZ's recommendation misses out on numerous portfolios with notable and statistically significant out-of-sample returns. The roughly 10 accounting strategies with significant returns would be missed by applying HLZ's recommendation. It would also lead to missing the past return portfolio with 8 percent out-of-sample returns. Overall, the HLZ method does not effectively separate strategies with high out-of-sample returns from those with low out-of-sample returns. In contrast, Figures 4 and 5 show that empirical Bayes shrinkage does a good job forecasting returns throughout the cross-section.

## 5.2 Storey (2002)'s FDR Control

Harvey, Liu, and Zhu (2016)'s recommendation is an extremely conservative variant of the Benjamini and Hochberg (1995) method. Much of the statistics literature goes in the opposite direction, modifying Benjamini and Hochberg (1995) to be more aggressive. Indeed, finance papers that came before HLZ emphasized more aggressive FDR controls.

For example, Barras, Scaillet, and Wermers (2010) recommend the Storey (2002) FDR control, which can be written as

$$h_{\text{Storey},q^*} \equiv \min_{h>0} \left\{ h : \left[ \frac{\Pr\left(|\tilde{t}_1| > h | \tilde{\theta}_1 = 0\right)}{\text{Share of } |\tilde{t}_i| > h} \right] \pi_{\text{Storey}} \leq q^* \right\} \tag{11}$$

where

$$\pi_{\text{Storey}} = \frac{\text{Share of } |\tilde{t}_i| \leq 1.0}{\Pr\left(|\tilde{t}_1| \leq 1.0 | \tilde{\theta}_1 = 0\right)} = \frac{\text{Share of } |\tilde{t}_i| \leq 1.0}{0.68} \tag{12}$$

and the cutoff of 1.0 is selected for ease of interpretation. Like HLZ's preferred

method (Equations (11)-(12)), Equations (11)-(12) amount to modifying the Benjamini and Hochberg (1995) with a constant factor. But while HLZ's constant factor of $\sum_{i=1}^{N} \frac{1}{i} \approx \log N \gg 1$ leads to a much more conservative hurdle, $\pi_{\text{Storey}} \leq 1.0$ leads to a more aggressive approach. In particular, Equation (12) says $\pi_{\text{Storey}}$ is found by dividing the share of $|\tilde{t}_i| \leq 1.0$ in the data by the classical "one sigma" area of 68%. This ratio amounts to a conservative estimate of the probability that $\tilde{\theta}_i = 0$,[5] and this approach to sharpening Benjamini and Hochberg (1995) is found in many other FDR controls (Benjamini and Hochberg (2000), Efron, Tibshirani, et al. (2001), Genovese, Roeder, and Wasserman (2006), and Benjamini, Krieger, and Yekutieli (2006)).

Also unlike HLZ's choice of $q^* = 1\%$, the statistics literature tends to prefer using $q^* = 5\%$ or 10% in applications from genetics to functional imaging (Efron (2012) and Benjamini (2020)). We argue that an even larger $q^*$ is preferred for cross-sectional asset pricing, since the consequences of a false discovery are much smaller compared to medical research. We examine $q^* = 20\%$, implying an investor which is satisfied as long as more than 80% of her trading strategies are true discoveries.

Figure 7 evaluates the effectiveness of using $q^* = 10\%$ and 20% using out-of-sample portfolio sorts. We use the same evaluation as we did for HLZ's method. But the results are quite different. Both versions of the Storey method do a good job of separating low from high out-of-sample returns. The majority of portfolios with economically meaningful returns are declared as discoveries using either choice of $q^*$. The $q^* = 20\%$ hurdle (dashed line) captures more of the economically notable portfolios.

[Figure 7, Multiple Testing Following Storey, about here]

Both hurdles, however, miss out on the past return portfolio which generates a notable return of about 3% per year over the full sample (equal or value weighted). This miss is likely related to the symmetry built into FDR methods. In contrast, EB shrinkage can handle these asymmetries, and predicts high returns for these portfolios, at least pre-2004 (Figure 4). More broadly, EB shrinkage allows for closer connection with the economics more generally. One can use the

[5] To see this, start with the law of total probability

$$\Pr\left(|\tilde{t}_i| \leq 1.0\right) = \Pr\left(|\tilde{t}_i| \leq 1.0 | \tilde{\theta}_i = 0\right) \Pr\left(\tilde{\theta}_i = 0\right) + \Pr\left(|\tilde{t}_i| \leq 1.0 | \tilde{\theta}_i \neq 0\right) \Pr\left(\tilde{\theta}_i \neq 0\right),$$

Solve for $\Pr\left(\tilde{\theta}_i = 0\right)$, and note that $\Pr\left(|\tilde{t}_i| \leq 1.0 | \tilde{\theta}_i \neq 0\right) \Pr\left(\tilde{\theta}_i \neq 0\right) \geq 0$.

predicted returns and Sharpe ratios as inputs for portfolio selection and asset pricing, neither of which is possible using FDR methods.

# 6 Conclusion

We show that a solution to data mining bias is to mine data rigorously. We systematically search data on accounting variables, past returns, and ticker symbols for large out-of-sample returns. We adjust large in-sample returns for data mining by conditioning our estimates on the search. This conditioning leads to accurate predictions of out-of-sample returns throughout the cross-section of strategies. Through this "high-throughput asset pricing," we find mean returns comparable to those found in top finance journals. But unlike the published strategies, ours are arguably free of data mining bias.

High-throughput asset pricing finds that returns are concentrated in accounting strategies, small stocks, and pre-2004. These facts are consistent with the theory that anomaly returns are due to mispricing and the slow diffusion of information. While these results could potentially be gleaned from a deep read of the anomalies literature, our methods provide a scientific method for documenting these stylized facts.

Data mined forecasts are less accurate post-2004, likely because the rise of the internet that led to a structural break in financial markets. This idea suggests that an important role for theory is to both understand when structural breaks should occur and provide guidance on how to deal with them.

# A Data Handling Details

## A.1 60,000 Accounting Ratio Strategies

We examine 60,000 accounting ratio strategies constructed by Chen, Lopez-Lira, and Zimmermann (2022). Inspired by Yan and Zheng (2017), Chen et al. construct 30,000 accounting ratio signals as follows. Let $X$ be one of 240 accounting variables from Compustat (+ CRSP market equity) and $Y$ be one of 65 accounting of these 240 variables that is positive for at least 25% of firms in 1963. Apply two transformations: $X/Y$ and $\Delta X/\text{lag}Y$ to get $240 \times 65 \times 2 \approx 30,000$ signals. Then form equal-weighted and value-weighted long-short decile strategies, leading to 60,000 strategies.

These strategies are downloaded from Andrew Chen's website. We are grateful to the others for making their data public.

## A.2 38,000 Past Return-Based Strategies

Inspired by Yan and Zheng (2017) , we construct past-return strategies as follows: Choose 4 quarters out of the past 20 quarters. Compute the first four central moments using the returns in these quarters. This leads to $\binom{20}{4} \times 4 = 19,380$ signals.

Add to this the return over any of the past 20 quarters, as well as the mean return over the past 2 and past 3 quarters. This adds $20 + 2$ signals, for a total of $19,380 + 22 = 19,402$ signals.

Finally, form equal-weighted and value-weighted long-short decile strategies.

We chose this approach, rather than the approach in Yan and Zheng (2017) for three reasons. The first is that we want to have a strategy list that is comparable in length to the length of our accounting ratio strategies. Yan and Zheng's method leads to "only" 4,080 signals. The second is that, while Yan and Zheng's methods are inspired by momentum and short-run reversal, we want to ensure that our methods do not incorporate knowledge that would come from reading finance publications. Last, we chose to reduce the amount of overlap across the different signals, which should lead to better properties of our EB estimator.

Earlier versions of our paper use Yan and Zheng's method and found similar results. These results can be found at arxiv.org.

### A.3 38,000 Ticker-Based Strategies

Inspired by Harvey (2017), we sort stocks into 20 groups based on the alphabetical order of the first ticker symbol. We then long any two of those groups and short two. Repeat using the 2nd, 3rd, and 4th ticker symbols. This yields $\binom{20}{4} \times 4 = 19{,}380$ long-short portfolios.

We chose not to follow Harvey (2017)'s approach in order to have a similar number of strategies as our accounting-based strategies. Harvey's method leads to "only" 6,000 ticker-based strategies.

Earlier versions of our paper used Harvey's method and found similar results. These results can be found at arxiv.org.

## B Theory and Estimation Details

### B.1 Estimation Details

We construct the quasi-likelihood using the distr package in R (Ruckdeschel et al. (2006)) and optimize using the BOBYQA algorithm in the nloptr package (Johnson (2007)). BOBYQA is a derivative-free bound-constrained optimization based on quadratic approximations of the objective.

We also use distr to compute the prediction formula (Equation (4)). To ensure numerical stability, we split the integrals into many smaller parts.

### B.2 Closed-form Prediction Formulas

This section provides details closed forms for the prediction Equation (4) under the mixture normal assumption. Our results actually uses distr to numerically compute these values, but these formulas are helpful for intuition.

Add some notation for describing the mixture:

$$\tilde{Z} = \begin{cases} a & \text{with prob } \lambda_a \\ b & \text{with prob } \lambda_b = 1 - \lambda_a \end{cases}$$

The prediction of $\tilde{\theta}_i$ conditional on $\tilde{t}_i = t$ and $\tilde{Z} = z$ can be found using

$$\hat{\theta}(t,z) \equiv E\left(\tilde{\theta}_i | \tilde{t}_i = t, \tilde{Z} = z\right) \tag{13}$$

$$= \hat{\mu}_{\theta,z} + \frac{\text{Cov}\left(\tilde{\theta}_i, \tilde{t}_i | \tilde{Z} = z\right)}{\widehat{\text{Var}}\left(\tilde{t}_i | \tilde{Z} = z\right)}\left(t - \hat{\mu}_{\theta,z}\right) \tag{14}$$

$$= \hat{\mu}_{\theta,z} + \frac{\widehat{\text{Var}}\left(\tilde{t}_i | \tilde{Z} = z\right) - \text{Cov}\left(\delta_i, \tilde{t}_i || \tilde{Z} = z\right)}{\widehat{\text{Var}}\left(\tilde{t}_i | \tilde{Z} = z\right)}\left(t - \hat{\mu}_{\theta,z}\right) \tag{15}$$

$$= \hat{\mu}_{\theta,z} + \left[1 - \frac{1}{\widehat{\text{Var}}\left(\tilde{t}_i | \tilde{Z} = z\right)}\right]\left(t - \hat{\mu}_{\theta,z}\right) \tag{16}$$

Then average across the possible values of $\tilde{Z}$ (use iterated expectations):

$$\hat{\theta}(t) = \sum_{z \in (a,b)} \Pr\left(\tilde{Z} = z | \tilde{t}_i = t\right) \hat{\theta}(t,z)$$

where the probabilities are found using Bayes formula

$$\Pr\left(\tilde{Z} = z | \tilde{t}_i = t\right) = \frac{f_{\text{Norm}}\left(t | \hat{\mu}_{\theta,z}, \widehat{\text{Var}}\left(\tilde{t}_i | \tilde{Z} = z\right)\right) \lambda_z}{\sum_{z' \in \{a,b\}} f_{\text{Norm}}\left(t | \hat{\mu}_{\theta,z'}, \widehat{\text{Var}}\left(\tilde{t}_i | \tilde{Z} = z'\right)\right) \lambda_{z'}}.$$

and $f_{\text{Norm}}\left(t | \hat{\mu}_{\theta,z}, \widehat{\text{Var}}\left(\tilde{t}_i | \tilde{Z} = z\right)\right)$ is a normal density function with mean $\hat{\mu}_{\theta,z}$ and variance $\widehat{\text{Var}}\left(\tilde{t}_i | \tilde{Z} = z\right)$.

To recover the bias-adjusted prediction about returns, just rescale:

$$\hat{\mu}\left(\tilde{r}_i, \tilde{t}_i\right) \equiv \hat{\theta}\left(\tilde{t}_i\right) \frac{\tilde{r}_i}{\tilde{t}_i}. \tag{17}$$

# References


Barras, Laurent, Olivier Scaillet, and Russ Wermers (2010). "False discoveries in mutual fund performance: Measuring luck in estimated alphas". In: *The journal of finance* 65.1, pp. 179–216.

Benjamini, Yoav (2020). "Selective inference: The silent killer of replicability". In.

Benjamini, Yoav and Yosef Hochberg (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the Royal statistical society: series B (Methodological)* 57.1, pp. 289–300.

— (2000). "On the adaptive control of the false discovery rate in multiple testing with independent statistics". In: *Journal of educational and Behavioral Statistics* 25.1, pp. 60–83.

Benjamini, Yoav, Abba M Krieger, and Daniel Yekutieli (2006). "Adaptive linear step-up procedures that control the false discovery rate". In: *Biometrika* 93.3, pp. 491–507.

Benjamini, Yoav and Daniel Yekutieli (2001). "The control of the false discovery rate in multiple testing under dependency". In: *Annals of statistics*, pp. 1165–1188.

— (2005). "False discovery rate–adjusted multiple confidence intervals for selected parameters". In: *Journal of the American Statistical Association* 100.469, pp. 71–81.

Bowles, Boone et al. (2023). "Anomaly time". In: *Available at SSRN 3069026*.

Chen, Andrew Y (2023). "Do t-Statistic Hurdles Need to be Raised?" In: *arXiv preprint arXiv:2204.10275*.

Chen, Andrew Y, Alejandro Lopez-Lira, and Tom Zimmermann (2022). "Peer-reviewed theory does not help predict the cross-section of stock returns". In: *arXiv preprint arXiv:2212.10317*.

Chen, Andrew Y and Mihail Velikov (2022). "Zeroing in on the Expected Returns of Anomalies". In: *Journal of Financial and Quantitative Analysis*.

Chen, Andrew Y and Tom Zimmermann (2020). "Publication bias and the cross-section of stock returns". In: *The Review of Asset Pricing Studies* 10.2, pp. 249–289.

— (2022). "Open Source Cross Sectional Asset Pricing". In: *Critical Finance Review*.

Chordia, Tarun, Amit Goyal, and Alessio Saretto (2020). "Anomalies and false rejections". In: *The Review of Financial Studies* 33.5, pp. 2134–2179.

Chordia, Tarun, Avanidhar Subrahmanyam, and Qing Tong (2014). "Have capital market anomalies attenuated in the recent era of high liquidity and trading activity?" In: *Journal of Accounting and Economics* 58.1, pp. 41–58.

Cochrane, John H (2005). "The risk and return of venture capital". In: *Journal of financial economics* 75.1, pp. 3–52.

Didisheim, Antoine et al. (2023). *Complexity in factor pricing models*. Tech. rep. National Bureau of Economic Research.

Efron, B and T Hastie (2016). *Computer age statistical inference: Data mining, inference and prediction*. Cambridge: Cambridge University Press.

Efron, Bradley (2012). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*. Vol. 1. Cambridge University Press.

Efron, Bradley and Carl Morris (1973). "Stein's estimation rule and its competitors-an empirical Bayes approach". In: *Journal of the American Statistical Association* 68.341, pp. 117–130.

Efron, Bradley, Robert Tibshirani, et al. (2001). "Empirical Bayes analysis of a microarray experiment". In: *Journal of the American statistical association* 96.456, pp. 1151–1160.

Fama, Eugene F and Kenneth R French (1992). "The cross-section of expected stock returns". In: *the Journal of Finance* 47.2, pp. 427–465.

Genovese, Christopher R, Kathryn Roeder, and Larry Wasserman (2006). "False discovery control with p-value weighting". In: *Biometrika* 93.3, pp. 509–524.

Harvey, Campbell R (2017). "Presidential address: The scientific outlook in financial economics". In: *The Journal of Finance* 72.4, pp. 1399–1440.

Harvey, Campbell R and Yan Liu (2020). "False (and missed) discoveries in financial economics". In: *The Journal of Finance* 75.5, pp. 2503–2553.

Harvey, Campbell R, Yan Liu, and Heqing Zhu (2016). "... and the cross-section of expected returns". In: *The Review of Financial Studies* 29.1, pp. 5–68.

Jegadeesh, Narasimhan and Sheridan Titman (1993). "Returns to buying winners and selling losers: Implications for stock market efficiency". In: *The Journal of finance* 48.1, pp. 65–91.

Jensen, Theis Ingerslev, Bryan Kelly, and Lasse Heje Pedersen (2023). "Is there a replication crisis in finance?" In: *The Journal of Finance* 78.5, pp. 2465–2518.

Johnson, Steven G. (2007). *The NLopt nonlinear-optimization package*. `https://github.com/stevengj/nlopt`.

Kelly, Bryan, Semyon Malamud, and Kangying Zhou (2024). "The virtue of complexity in return prediction". In: *The Journal of Finance* 79.1, pp. 459–503.

Kim, Yong Hyuck, Zoran Ivkovich, and Dmitriy Muravyev (2021). "Causal Effect of Information Costs on Asset Pricing Anomalies". In: *Available at SSRN 3921785*.

Lakonishok, Josef, Andrei Shleifer, and Robert W Vishny (1994). "Contrarian investment, extrapolation, and risk". In: *The journal of finance* 49.5, pp. 1541–1578.

McLean, R David and Jeffrey Pontiff (2016). "Does academic research destroy stock return predictability?" In: *The Journal of Finance* 71.1, pp. 5–32.

Peng, Lin (2005). "Learning with information capacity constraints". In: *Journal of Financial and Quantitative Analysis* 40.2, pp. 307–329.

Robbins, Herbert (1956). "An Empirical Bayes Approach to Statistics". In: *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics* 3.1.

Ruckdeschel, P. et al. (May 2006). "S4 Classes for Distributions". English. In: *R News* 6.2, pp. 2–6.

Storey, John D (2002). "A direct approach to false discovery rates". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 64.3, pp. 479–498.

Yan, Xuemin Sterling and Lingling Zheng (2017). "Fundamental analysis and the cross-section of stock returns: A data-mining approach". In: *The Review of Financial Studies* 30.4, pp. 1382–1423.

Yang, Liangliang et al. (2021). "High-throughput methods in the discovery and study of biomaterials and materiobiology". In: *Chemical reviews* 121.8, pp. 4561–4677.

**Figure 1: Cumulative Long-Short Returns from Rigorous Data-Mining.** We use empirical Bayes to mine 140,000 long-short strategies for large out-of-sample returns. Each year, we sign strategies to have positive returns based on EB predictions and then form portfolios that equal-weight strategies in the top $X$% of predicted Sharpe ratios based on Equation (4). We hold for one year and repeat. **Interpretation:** Rigorous data mining generates notable out-of-sample performance. Returns experience a break around the early-2000s, around the time when internet access became widespread.

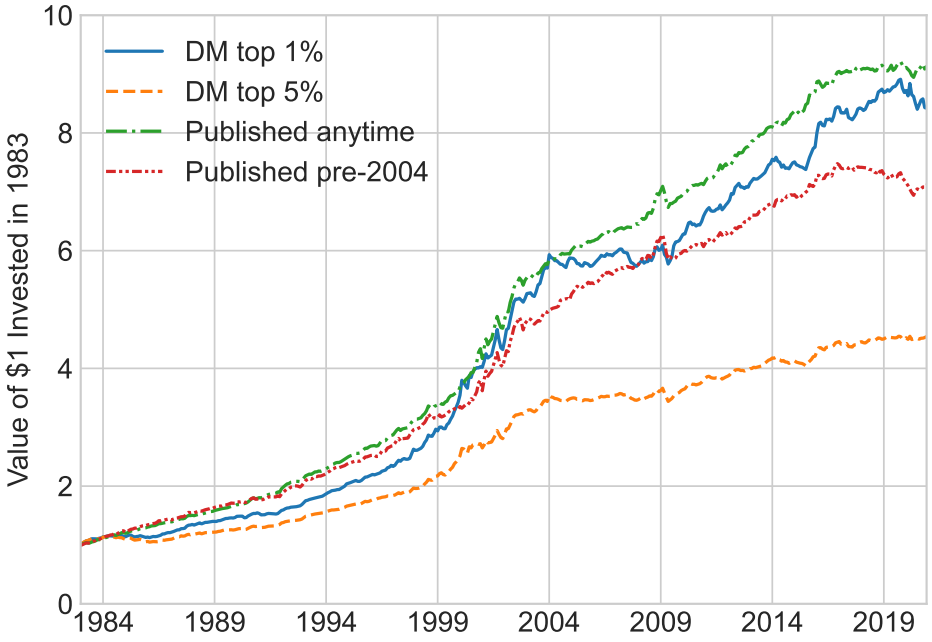

**Figure 2: Distribution of t-stats from long-short deciles strategies: 1983.** "Data" are t-stats testing the null of expected return = 0 from 1964-1983 for 140,000 trading strategies (Table 1). "Model" is Equations (5)-(7). "Null" is a standard normal. "EW" and "VW" are equal- and value-weighting, respectively. **Interpretation:** Equal-weighted accounting and equal-weighted past return strategies are far from the null, indicating true predictability. The models fit the data well.

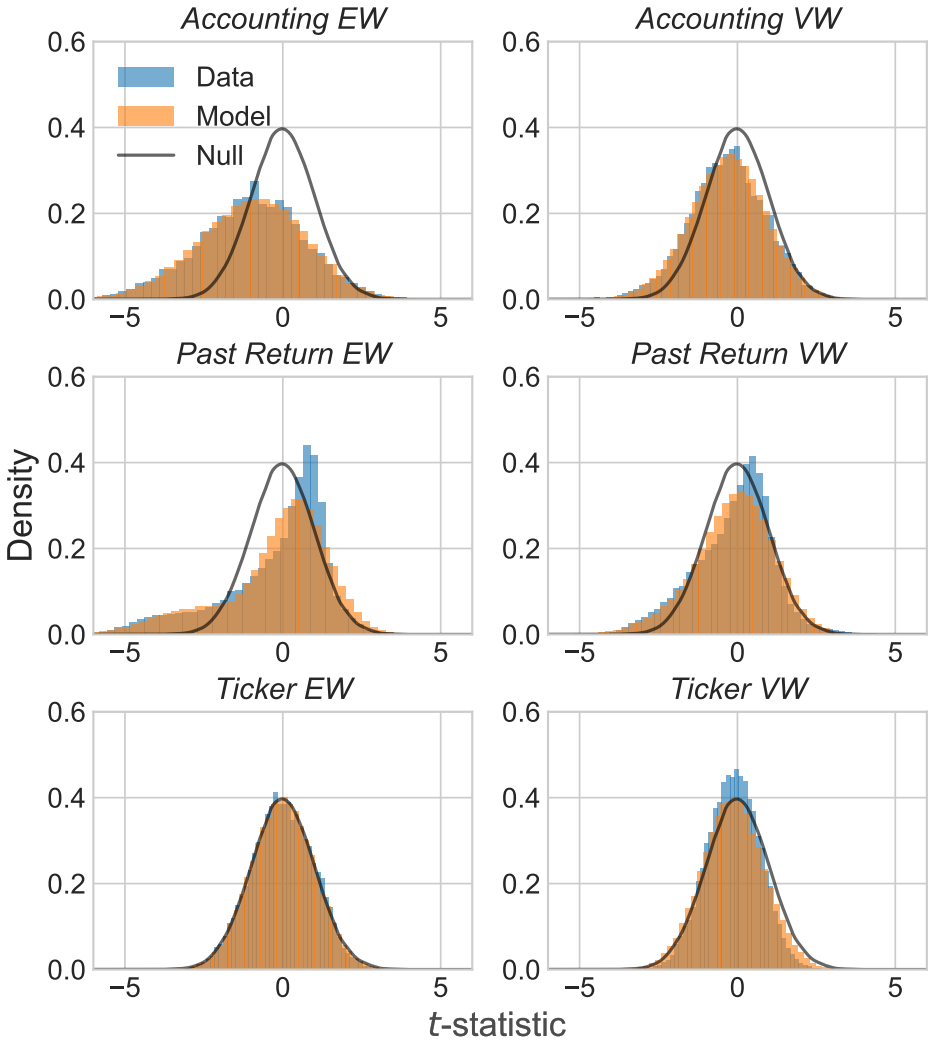

**Figure 3: Distribution of t-stats from long-short deciles strategies: 2004.** "Data" are t-stats testing the null of expected return = 0 from 1985-2004 for 140,000 trading strategies (Table 1). "Model" is Equations (5)-(7). "Null" is a standard normal. "EW" and "VW" are equal- and value-weighting, respectively. **Interpretation:** Compared to 1983 (Figure 2), t-stats from 2004 are much closer to the null, indicating diminished predictability. Equal-weighted accounting strategies still show true predictability, however.

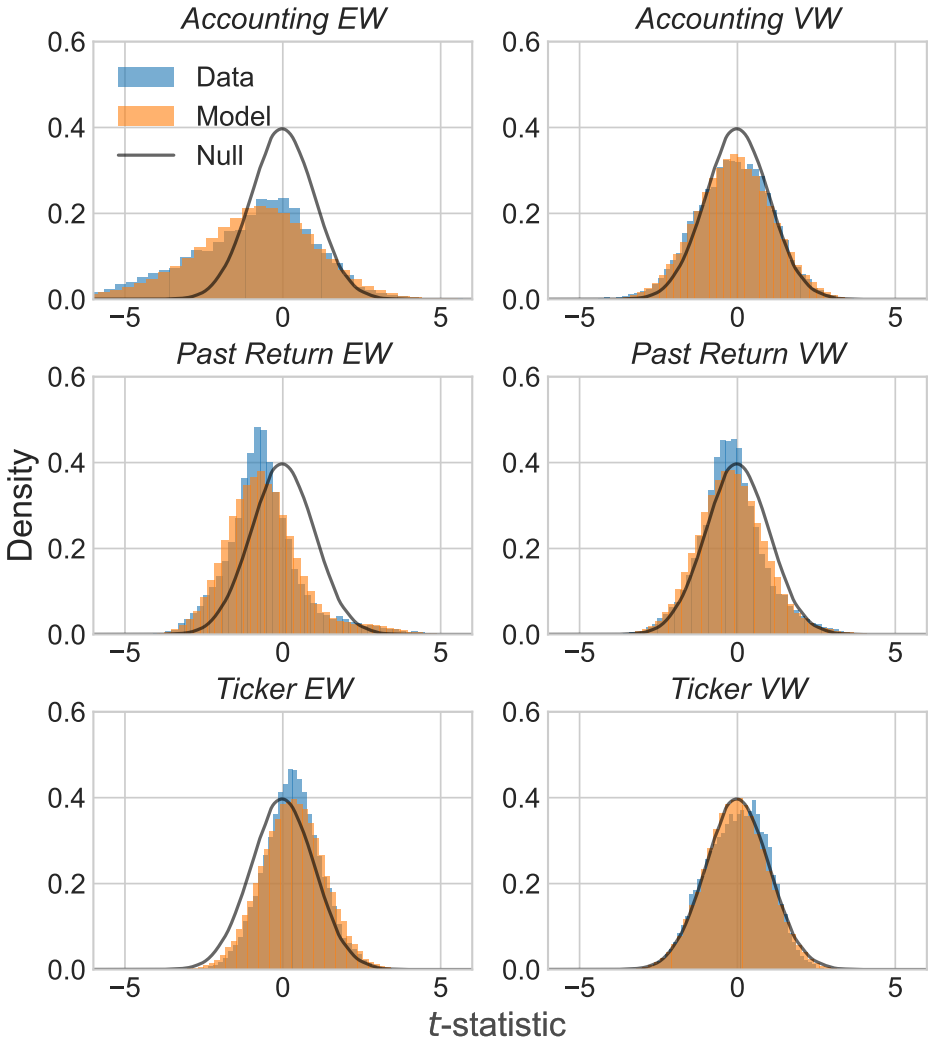

**Figure 4: Empirical Bayes Predictions and Out-of-Sample Returns: 1983-2004.** For each year and each family of strategies, we sort strategies into 20 groups based on the past 20 years of returns ("In-Samp") and predict returns using Bayes rule (Equation (3), "Predicted"). We form equal-weighted portfolios of strategies in each group and hold for one year ("OOS," error bars are two standard errors). **Interpretation:** Pre-2004, empirical Bayes shrinkage provides accurate forecasts of out-of-sample returns, unlike using the naive rule of in-sample return = out-of-sample return. Rigorous data mining removes data mining bias.

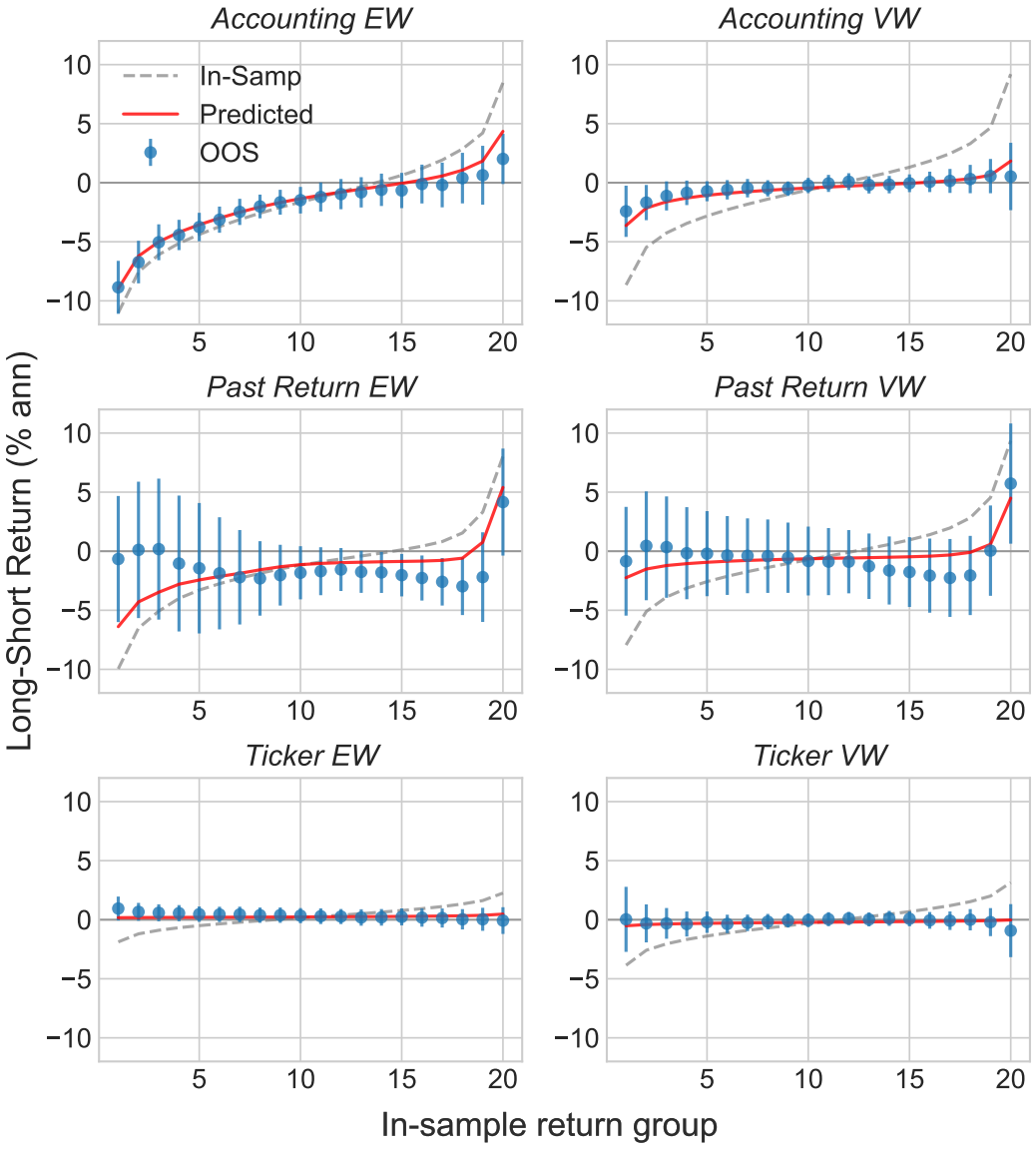

**Figure 5: Empirical Bayes Predictions and Out-of-Sample Returns: 2004-2020.** For each year and each family of strategies, we sort strategies into 20 groups based on the past 20 years of returns ("In-Samp") and predict returns using Bayes rule (Equation (3), "Predicted"). We form equal-weighted portfolios of strategies in each group and hold for one year ("OOS," error bars are two standard errors). **Interpretation:** Compared with pre-2004 (Figure 4), post-2004 predicted returns are closer to zero. Out-of-sample returns are even closer to zero, consistent with a structural break in predictability around 2004.

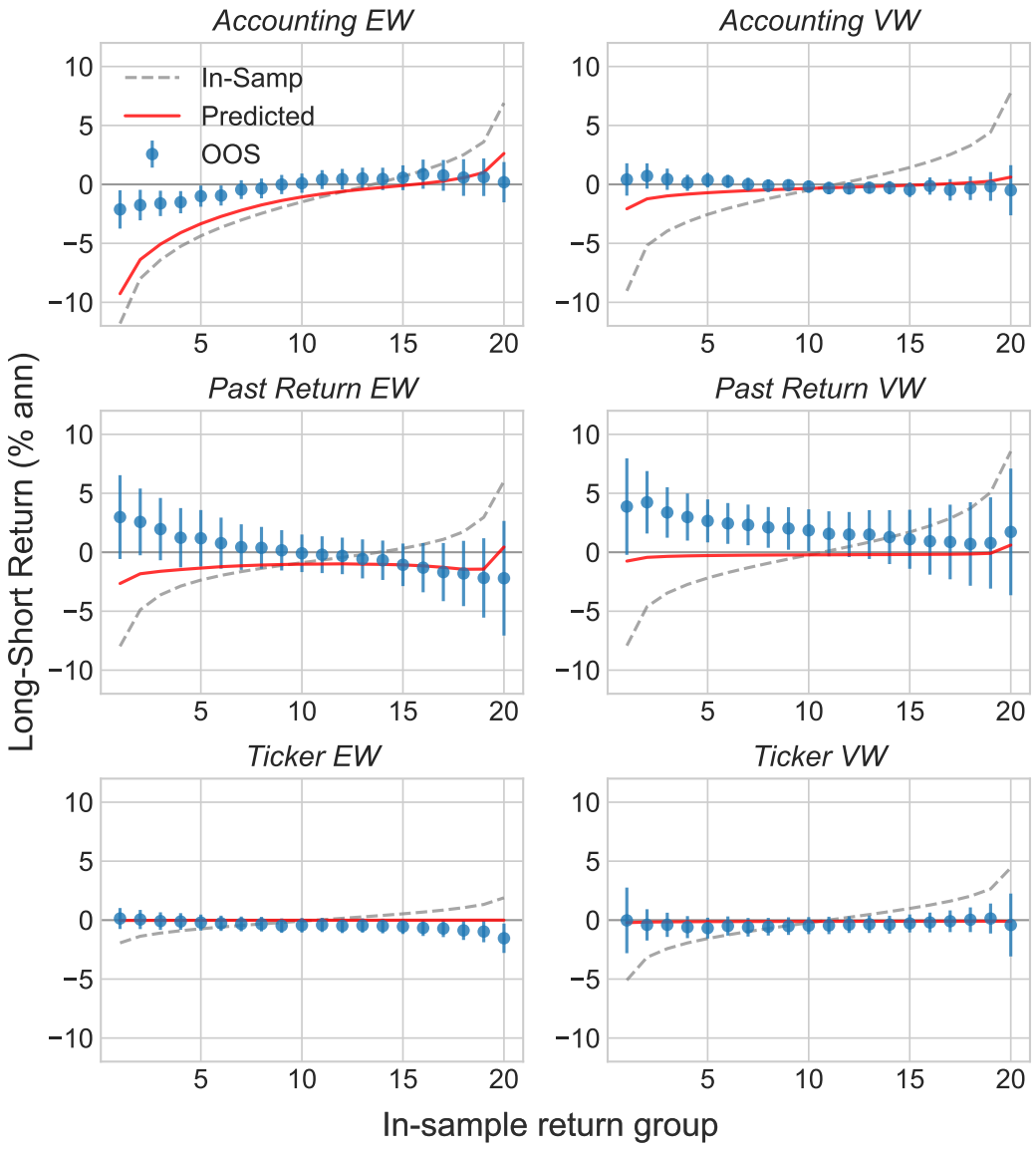

**Figure 6: Multiple Testing Controls Following Harvey, Liu, and Zhu (2016):** For each year and each strategy family, we calculate t-stat hurdles (vertical lines) following Harvey, Liu, and Zhu (2016), who recommend using Benjamini and Yekutieli (2001), Theorem 1.3 with $q^* = 1\%$ or $q^* = 5\%$. We compare with out-of-sample returns of the strategies sorted into 20 bins based on in-sample t-statistics (markers). Hurdles, in-sample t-stats, and out-of-sample returns are calculated each year from 1983-2020, and then averaged across years. Error bars are two standard errors. **Interpretation:** Most strategies with substantial out-of-sample returns fail to pass the hurdles recommended by Harvey, Liu, and Zhu (2016). These hurdles are excessively conservative.

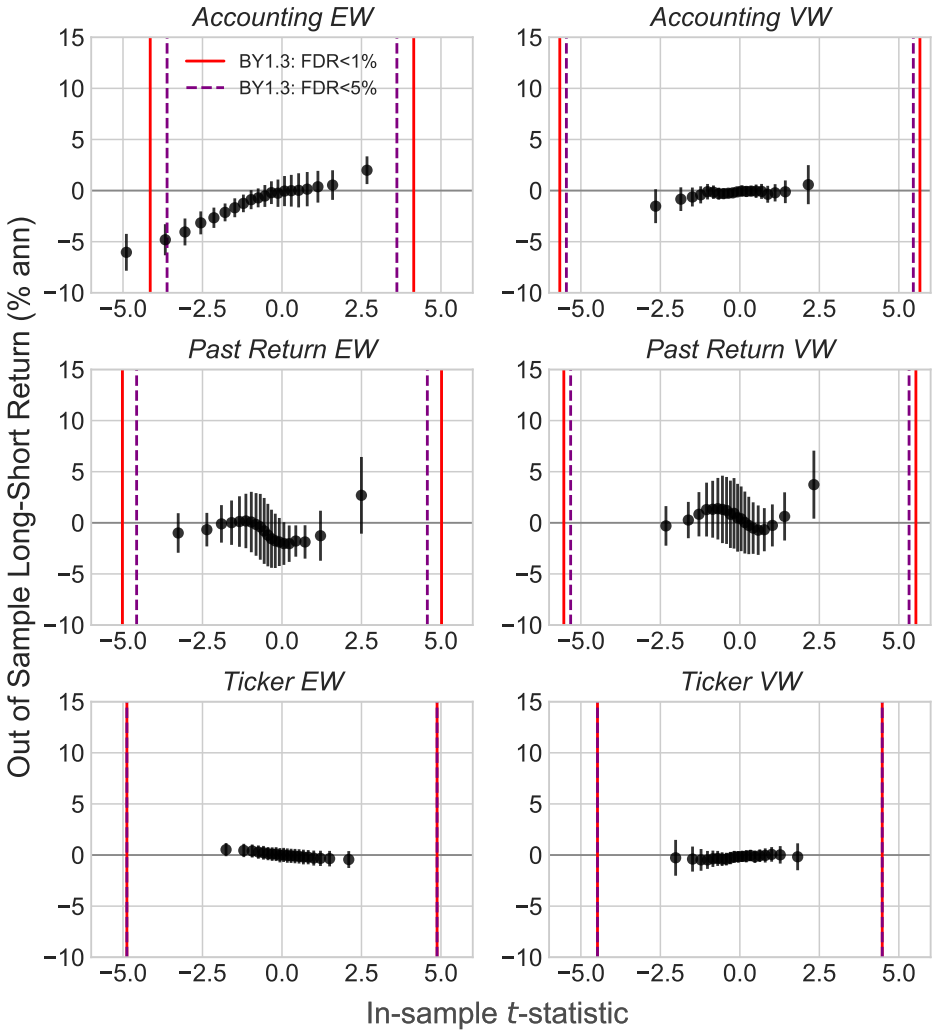

**Figure 7: Multiple Testing Controls Following Storey (2002):** For each year and each strategy family, we calculate t-stat hurdles (vertical lines) following Storey (2002) (Equations (11)-(12)), using $q^* = 10\%$ based on the genetics literature (Efron (2012); Benjamini (2020)) and $q^* = 20\%$ because false discoveries are less of a concern in cross-sectional predictability. We compare with out-of-sample returns of the strategies sorted into 20 bins based on in-sample t-statistics (markers). Hurdles, in-sample t-stats, and out-of-sample returns are calculated each year from 1983-2020, and then averaged across years. **Interpretation:** The Storey FDR control does a fairly good job of separating low from high out-of-sample returns.

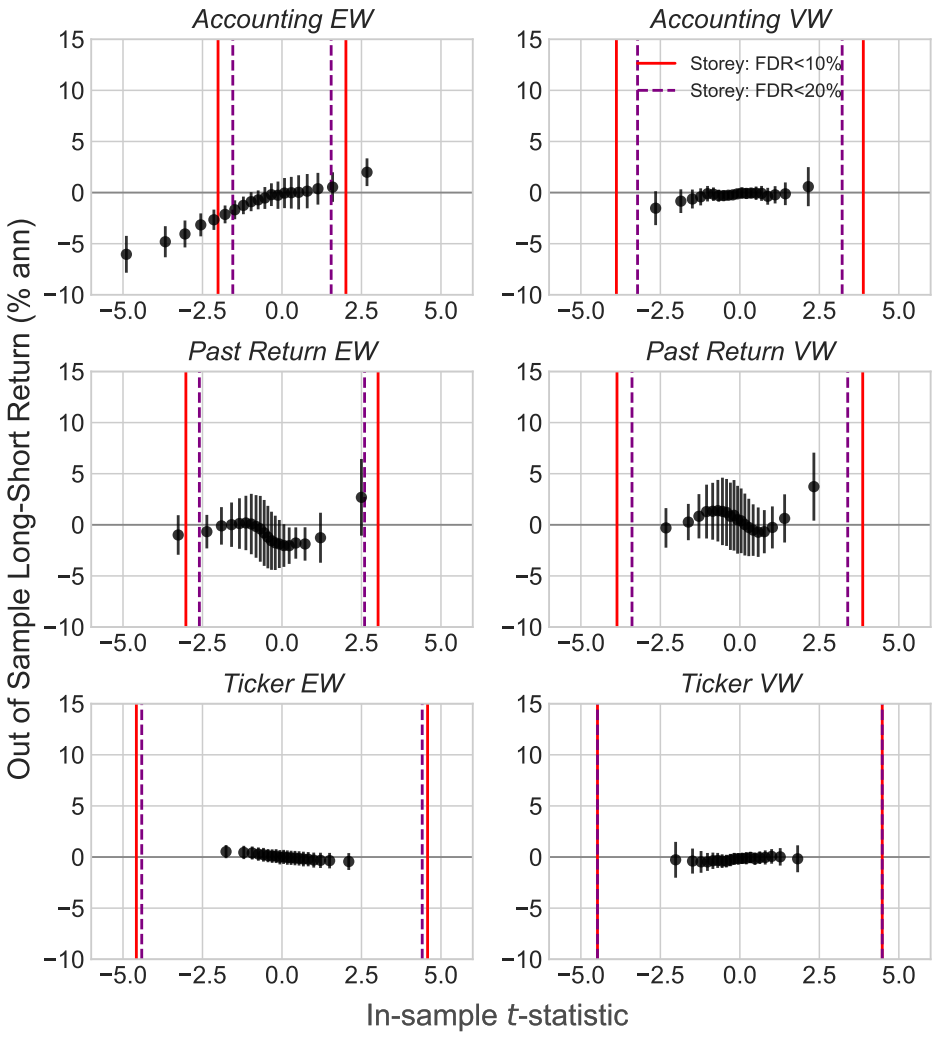

**Table 1: Overview of 140,000 Long-Short Strategies**

Table describes the 136,192 strategies used throughout the paper. Data and code for these strategies are posted publicly. **Interpretation:** Unlike datasets of published strategies (e.g. Chen and Zimmermann (2022)), these strategies are arguably constructed without data-mining bias.

| Panel A: Accounting Strategies | | | | |
|---|---|---|---|---|
| Description: Make ratios from 242 accounting variables by (1) dividing one variable by another and (2) taking first differences and then dividing. Long / short the extreme deciles. Data is from Chen, Lopez-Lira, and Zimmermann (2022). | | | | |
| | # strategies | Mean Return (% ann) | | |
| | | 5 pctile | 50 pctile | 95 pctile |
| EW | 29,314 | -7.0 | -1.1 | 3.7 |
| VW | 29,314 | -4.5 | -0.4 | 3.9 |
| Panel B: Past Return Strategies | | | | |
| Description: Choose 4 quarters out of the past 20 and compute one of the first four central moments, yielding $\binom{20}{4} \times 4 = 19,380$ signals. Add the return over any of the past 20 quarters and mean returns over the past 2 and past 3 quarters to arrive at 19,402 signals. Long / short the extreme deciles. | | | | |
| | # strategies | Mean Return (% ann) | | |
| | | 5 pctile | 50 pctile | 95 pctile |
| EW | 19,402 | -5.3 | -0.4 | 2.1 |
| VW | 19,402 | -3.4 | 0.1 | 4.3 |
| Panel C: Ticker Strategies | | | | |
| Descrption: Sort stocks into 20 groups based on alphabetical order of the first ticker symbol. Long two of those groups and short two. Repeat using the 2nd, 3rd, and 4th ticker symbols. This yields $\binom{20}{4} \times 4 = 19,380$ long-short portfolios. | | | | |
| | # strategies | Mean Return (% ann) | | |
| | | 5 pctile | 50 pctile | 95 pctile |
| EW | 19,380 | -0.9 | 0.0 | 0.8 |
| VW | 19,380 | -2.2 | -0.2 | 1.6 |

**Table 2: Returns of Long-Short Portfolios Data-Mined with Empirical Bayes**

We use empirical Bayes to mine 140,000 long-short strategies for large out-of-sample returns. Each year, we sign strategies to have positive returns based on EB predictions and then form portfolios that equal-weight strategies in the top $X$% of predicted Sharpe ratios based on Equation (4). We hold for one year and repeat. 'Pub Anytime' is a portfolio that equally weights strategies from Chen and Zimmermann (2022). 'Pub Pre-2004' equally weighs strategies published before 2004. **Interpretation:** Rigorous data mining generates out-of-sample returns comparable to those from the best journals in finance, even if the data includes signals with zero out-of-sample mean returns, like ticker-sorted portfolios.

| | Num Strats Combined | Mean Return (% ann) | $t$-stat | Sharpe Ratio (ann) |
|---|---|---|---|---|
| 1983-2020 | | | | |
| DM Top 1% | 1278 | 5.70 | 9.00 | 1.46 |
| DM Top 5% | 6389 | 4.03 | 8.27 | 1.34 |
| DM Top 10% | 12777 | 2.77 | 7.16 | 1.16 |
| Pub Anytime | 203 | 5.88 | 12.54 | 2.03 |
| Pub Pre-2004 | 82 | 5.23 | 9.57 | 1.55 |
| 1983-2004 | | | | |
| DM Top 1% | 1262 | 8.17 | 8.84 | 1.88 |
| DM Top 5% | 6305 | 5.74 | 7.82 | 1.67 |
| DM Top 10% | 12610 | 4.20 | 7.22 | 1.54 |
| Pub Anytime | 201 | 8.18 | 11.84 | 2.52 |
| Pub Pre-2004 | 81 | 7.56 | 9.20 | 1.96 |
| 2005-2020 | | | | |
| DM Top 1% | 1301 | 2.29 | 3.10 | 0.78 |
| DM Top 5% | 6504 | 1.68 | 3.22 | 0.80 |
| DM Top 10% | 13007 | 0.81 | 1.95 | 0.49 |
| Pub Anytime | 207 | 2.71 | 5.44 | 1.36 |
| Pub Pre-2004 | 82 | 2.03 | 3.59 | 0.90 |

**Table 3: Description of the Top 1% Data-Mined Strategies**

Panel A shows the fraction of strategies that comes from each signal family, pooled across all sample years. Panel B lists the definitions of the strategies with highest predicted Sharpe Ratios (SR pred) using data from 1974-1993. SR OOS is the realized Sharpe ratio 1994-2003. Sharpe ratios are annual. **Interpretation:** The top 1% strategies are largely equal-weighted accounting strategies. Equal-weighted past return strategies comprise a non-trivial minority. The top 20 strategies are distant from strategies in the academic literature yet they perform well out-of-sample.

| Panel A: Average Fraction of Signals in the Top 1% | | | | | |
|---|---|---|---|---|---|
| Acct EW | Acct VW | Past Ret EW | Past Ret VW | Ticker EW | Ticker VW |
| 91.0% | 0.3% | 8.6% | 0.1% | 0.0% | 0.0% |

Panel B: Top 20 Strategies in 1993 based on Signed Predicted Sharpe Ratio

| Rank | SR Pred | SR OOS | Signal Family | Signal Name |
|---|---|---|---|---|
| 1 | 1.56 | 1.32 | Acct EW | - ΔInterest paid net / Lag(Common equity) |
| 2 | 1.51 | 0.84 | Acct EW | - Debt due in 2nd year / Depr, depl & amort |
| 3 | 1.43 | 0.92 | Acct EW | - Debt mortgages & other sec / Sales |
| 4 | 1.37 | 1.60 | Acct EW | - Debt mortgages & other sec / Depr, depl & amort |
| 5 | 1.37 | 1.64 | Acct EW | - ΔInterest paid net / Lag(Stockholders equity) |
| 6 | 1.35 | 0.54 | Past Ret EW | + Return in quarters $t$ minus 5, 9, 17, and 18 |
| 7 | 1.35 | 0.68 | Acct EW | - Debt due in 3rd year / Depr, depl, and amort |
| 8 | 1.35 | 0.69 | Acct EW | - Debt mortgages & other sec / Cost of goods sold |
| 9 | 1.34 | 1.00 | Acct EW | - ΔInterest paid net / Lag(Inventories) |
| 10 | 1.33 | 0.62 | Acct EW | - Debt mortgages & other sec / Operating expenses |
| 11 | 1.33 | 0.47 | Past Ret EW | + Return in quarters $t$ minus 17 |
| 12 | 1.32 | 0.62 | Past Ret EW | + Return in quarters $t$ minus 9, 17, 18 and 19 |
| 13 | 1.30 | 0.43 | Acct EW | - ΔLiabilities / Lag(Depr & amort) |
| 14 | 1.29 | 0.44 | Past Ret EW | + Return in quarters $t$ minus 9, 13, 17, and 18 |
| 15 | 1.29 | 1.24 | Acct EW | - ΔInterest paid net / Lag(Equity liquidation value) |
| 16 | 1.29 | 0.68 | Past Ret EW | + Return in quarters $t$ minus 3, 9, 17, and 18 |
| 17 | 1.25 | 0.49 | Acct EW | - Debt due in 4th year / Depr, depl & amort |
| 18 | 1.25 | 1.01 | Acct EW | - Stock issuance / Gross profit |
| 19 | 1.25 | 0.55 | Acct EW | - Debt due in 2nd year / Depr & amort |
| 20 | 1.24 | 1.51 | Acct EW | - ΔLiabilities / Lag(Depr, depl & amort) |