

ESTIMATING THE ANOMALY BASERATE*

Alex Chinco[†], Andreas Neuhierl[‡] and Michael Weber[§]

June 12, 2019

Abstract

The academic literature contains literally hundreds of variables that seem to predict the cross-section of expected returns. This so-called ‘anomaly zoo’ has caused many to question whether researchers are using the right tests for statistical significance. But, here’s the thing: even if a researcher is using the right tests, he will still be drawing the wrong conclusions from his analysis if he is starting out with the wrong priors—i.e., if he is starting out with incorrect beliefs about the ex ante probability of discovering a tradable anomaly prior to seeing any test results.

So, what are the right priors to start out with? What is the correct anomaly baserate?

We propose a new statistical approach to answer this question. The key insight is that, under certain conditions, there’s a one-to-one mapping between the ex ante probability of discovering a tradable anomaly and the best-fit tuning parameter in a penalized regression. When we apply our new statistical approach to the cross-section of monthly returns, we find that the anomaly baserate has fluctuated substantially since the start of our sample in May 1973. The ex ante probability of discovering a tradable anomaly was much higher in 2003 than in 1990. As a proof of concept, we construct a trading strategy that invests in previously discovered predictors and show that adjusting this strategy to account for the prevailing anomaly baserate boosts its performance.

JEL CLASSIFICATION: C12, C52, G11

KEYWORDS: Return Predictability, Data Mining, Penalized Regression

*We would like to thank Justin Birru, Svetlana Bryzgalova, Zhi Da, Xavier Gabaix, Christian Julliard, Ralph Koijen, Yan Liu, Walt Pohl, Tarun Ramadorai, and Andrea Tamoni for extremely helpful comments and suggestions. This paper has also benefited greatly from presentations at the University of Illinois, the MFA meetings, AQR Asset-Management Institute’s Academic Symposium, the Future of Financial Information Conference, and the 5th BI-SHoF Conference. Bianca He provided excellent research assistance. Weber also gratefully acknowledges financial support from the University of Chicago, the Fama Research Fund, and the Fama-Miller Center.

Current Version: <http://www.alexchinco.com/anomaly-baserate.pdf>

[†]University of Illinois, Gies College of Business. alexchinco@gmail.com

[‡]University of Notre Dame, Mendoza College of Business. aneuhier@nd.edu

[§]University of Chicago, Booth School of Business and the NBER. michael.weber@chicagobooth.edu

1 Introduction

Imagine you are a financial economist sitting at your weekly research seminar. This week’s speaker is describing a new variable, X_n , that seems to predict the cross-section of expected returns. He has regressed the excess returns of each stock, R_n , on lagged values of X_n :

$$R_n = \hat{\mu} + \hat{\beta} \cdot X_n + \hat{\varepsilon}_n \quad \text{for stocks } n = 1, \dots, N + 1. \quad (1)$$

The speaker has standardized his new variable, X_n , to have zero mean and unit variance. So, in the regression specification above, $\hat{\mu}$ is the mean excess return in the current month, $\hat{\beta}$ is an estimated slope coefficient, and $\hat{\varepsilon}_n$ is the regression residual for the n th stock. On the current slide, he is showing an estimated $\hat{\beta} > 0$ that is statistically significant at the 5% level.

On one hand, this could be an interesting finding. Predictive regressions and trading-strategy returns are two sides of the same coin (Fama, 1976; Pedersen, 2015). You can interpret the estimated $\hat{\beta}$ in Equation (1) as the realized return to a zero-cost portfolio that is long stocks with high X_n values last month and short stocks with low X_n values:

$$\hat{\beta} \stackrel{\text{def}}{=} \frac{\text{Cov}[R_n, X_n]}{\text{Var}[X_n]} = \frac{1}{N} \cdot \sum_n (R_n - \hat{\mu}) \cdot (X_n - 0). \quad (2)$$

Thus, an estimated $\hat{\beta} > 0$ implies both that stocks with high predictor values last month, $(X_n - 0) > 0$, tended to have high excess returns this month, $(R_n - \hat{\mu}) > 0$, and also that it was profitable to trade on X_n in the previous month. And, since traders should immediately exploit (and thereby eliminate) such an arbitrage opportunity, a positive estimate for the slope coefficient could present a puzzle. It might suggest a gap in our economic understanding.

But, on the other hand, the seminar speaker’s estimated slope coefficient is just that... an estimate. An estimated $\hat{\beta} > 0$ implies that it *was* profitable to trade on X_n in the previous month; but, what you really want to know is whether it *will be* profitable to trade on X_n in the future. This subtle change in verb tense makes a world of difference. “With the combination of unreported tests, lack of adjustment for multiple tests, and direct and indirect p -hacking, many of the results being published will fail to hold up in the future (Harvey, 2017).” Thus, even if the seminar speaker estimates a positive and significant slope coefficient, X_n might not represent a tradable anomaly. The estimated $\hat{\beta} > 0$ might just be a fluke, a chance event.

You need to figure out the probability that X_n represents a tradable anomaly going forward given the speaker’s statistically significant estimate using historical data, $\Pr[\textit{anom} \mid \textit{signif}]$. That’s the inference problem you face. And, Bayes’ theorem gives the recipe for solving it:

$$\Pr[\textit{anom} \mid \textit{signif}] = \left(\frac{\Pr[\textit{signif} \mid \textit{anom}]}{\Pr[\textit{signif}]} \right) \times \Pr[\textit{anom}]. \quad (3)$$

Equation (3) says you can compute the probability that X_n is a tradable anomaly in the future given the speaker’s statistically significant in-sample estimate, $\Pr[\textit{anom} \mid \textit{signif}]$, by multiplying the ex-ante probability that X_n is a tradable anomaly, $\Pr[\textit{anom}]$, times the relative increase in this baserate due to the speaker’s statistically significant estimate, $\left(\frac{\Pr[\textit{signif} \mid \textit{anom}]}{\Pr[\textit{signif}]}\right)$.

The current paper is motivated by two observations about the inference problem you face as an audience member in this seminar. The first is that financial econometrics is almost entirely concerned with the first term on the right-hand side of Equation (3), $\left(\frac{\Pr[\textit{signif} \mid \textit{anom}]}{\Pr[\textit{signif}]}\right)$. This ‘Bayes factor’ captures the weight of evidence presented by the speaker. If his results are extremely compelling, then the Bayes factor will be much larger than one. Your posterior beliefs walking out of the seminar room will be large relative to your prior beliefs walking in, $\frac{\Pr[\textit{anom} \mid \textit{signif}]}{\Pr[\textit{anom}]} \gg 1$. By contrast, if the speaker’s results are weak, then the Bayes factor will be close to one. Sitting through the seminar will not change your priors much, $\frac{\Pr[\textit{anom} \mid \textit{signif}]}{\Pr[\textit{anom}]} \approx 1$. When your colleagues ask the speaker ‘How did you cluster your standard errors?’ and ‘Which p -value cutoff did you use?’, it is this term that they are inquiring about.

The second observation is that the Bayes factor is not the only term on the right-hand side of Equation (3). The Bayes factor represents a proportional gain. So, your prior beliefs walking into the seminar room, $\Pr[\textit{anom}]$, are going to have a huge impact on your posterior beliefs walking out. If you entered the room with a wildly inaccurate prior, then you will draw the wrong conclusions from the speaker’s analysis even if the speaker himself uses all of the right econometric techniques. And, in the same way that there are good reasons to worry about researchers applying the wrong Bayes factor, there are good reasons to worry about researchers using the wrong priors. Before the seminar even started, you already knew that a couple of your colleagues would leave the seminar room skeptical of the speaker’s results no matter what he said. Likewise, you also already knew which of your colleagues would be most receptive to the speaker’s findings. Both sets of priors cannot be right.

As a working financial economist, you have seen many different speakers present evidence on various other cross-sectional predictors over the course of your career. Some of these predictors were in fact tradable anomalies. Others were not. When you walked into the seminar room today, your prior beliefs about the anomaly baserate—i.e., the ex-ante probability that X_n would represent a tradable anomaly—were informed by this past experience. The question is: ‘How?’ What is the right way to convert your past experience with other predictors into a working prior for next time around? This is a question of great practical importance. But, it is also a question that is outside the scope of the existing academic literature. It requires modeling the anomaly-discovery process rather than measurement/sampling error.

With these ideas in mind, we propose a way to estimate the prevailing anomaly baserate by running penalized regressions and searching for the best-fit tuning parameter. Here is the

intuition behind our statistical approach. We start by estimating a separate Ridge regression for each previously discovered predictor in a given month. If the true slope coefficients for different predictors are drawn from a common distribution each month, then the best-fit tuning parameter in these Ridge regressions will be a decreasing function of the volatility of this common prior distribution. So, we invert the best-fit tuning parameter for each predictor to get a estimates of the prior volatility, take the average of these estimates each month, and then use this time series to make out-of-sample forecasts for the anomaly baserate.

Statistical Approach. We begin by fleshing out the details of our statistical approach. The standard way to estimate the slope coefficient, $\hat{\beta}$, associated with X_n is to use an ordinary least squares (OLS) regression. This means solving the optimization problem below:

$$\min_{\beta} \left\{ \frac{1}{N} \cdot \sum_n (R_n - \hat{\mu} - \beta \cdot X_n)^2 \right\}. \quad (4)$$

Notice that the solution to this optimization problem yields the expression for $\hat{\beta}$ in Equation (2) when X_n has zero mean and unit variance, $\frac{1}{N+1} \cdot \sum_n X_n = 0$ and $\frac{1}{N} \cdot \sum_n X_n^2 = 1$.

But, several recent papers have traveled an alternative route (DeMiguel, Garlappi, Nogales, and Uppal, 2009; Bryzgalova, 2017; Feng, Giglio, and Xiu, 2017; Freyberger, Neuhierl, and Weber, 2017; Chinco, Clark-Joseph, and Ye, 2018; Kozak, Nagel, and Santosh, 2018). Instead of using an OLS regression to estimate $\hat{\beta}$, these papers use a penalized regression, which means solving a modified version of the optimization problem in Equation (4):

$$\min_{\beta} \left\{ \frac{1}{N} \cdot \sum_n (R_n - \hat{\mu} - \beta \cdot X_n)^2 + \lambda \cdot |\beta|^q \right\}. \quad (5)$$

Above, $\lambda \geq 0$ is known as the ‘tuning parameter’, and $q \geq 0$ governs the shape of the penalty function. If $q = 0$, then this penalized regression is the same as the Bayesian information criteria (BIC; Schwarz, 1978). If $q = 1$, then it is the same as the LASSO (Tibshirani, 1996). And, if $q = 2$, then it is the same as a Ridge regression (Hoerl and Kennard, 1970).

The key insight behind our statistical approach is that the modified optimization problem in Equation (5) has an alternative Bayesian interpretation. The penalty function, $\lambda \cdot |\beta|^q$, can be thought of as the effect of incorporating your prior beliefs about the distribution of the true slope coefficient, β^* . Think about β^* as the excess return you would realize if you held the same zero-cost portfolio described in Equation (2) next month. To see where this Bayesian interpretation comes from, imagine that the cross-section of excess returns is governed by the following data-generating process with $\varepsilon_n^* \stackrel{\text{iid}}{\sim} \text{Normal}[0, N \cdot se^2]$:

$$R_n = \mu^* + \beta^* \cdot X_n + \varepsilon_n^*.$$

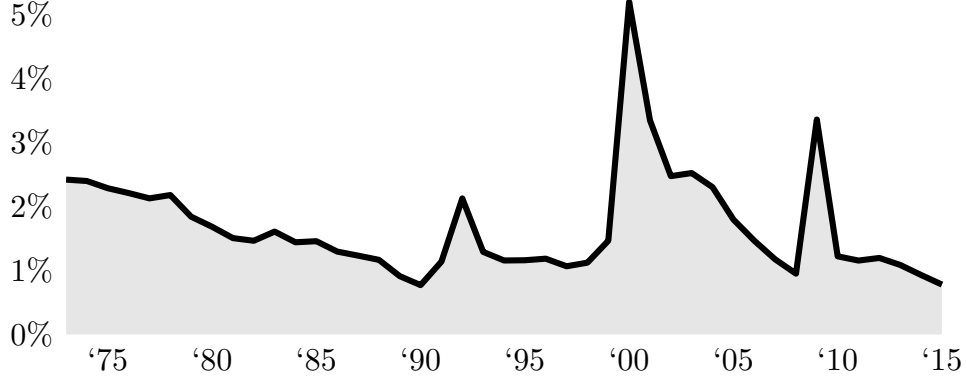


Figure 1. Forecasted \bar{v}_t . Average forecasted prior volatility each year. For each predictor i discovered prior to month t , we first estimate the in-sample parameter $\hat{v}_{i,t'}$ for all months $t' < t$ using the procedure described in Proposition 2.2. We then make a one-month-ahead forecast for month t by fitting an AR(3) model to squared values in months $\{t - 60, \dots, t - 1\}$. Finally, we combine the forecasts for all previously discovered predictors to compute \bar{v}_t . Units: % per month. Sample Period: 1973-2015.

In this setting, if the true slope coefficient is also drawn from a normal distribution, $\beta^* \sim \text{Normal}[0, \sigma^2]$, then the log likelihood of the true slope coefficient taking on a particular value, $\beta^* = \beta$, given the observed data, $\{R_1, \dots, R_{N+1}\}$ and $\{X_1, \dots, X_{N+1}\}$, can be written as

$$-\log \Pr[\beta] = \frac{1}{2 \cdot (N \cdot se^2)} \cdot \sum_n (R_n - \hat{\mu} - \beta \cdot X_n)^2 + \frac{1}{2 \cdot \sigma^2} \cdot (\beta - 0)^2 + \dots \quad (6)$$

where the “ \dots ” represents constants that do not depend on the choice of β .

In other words, when $q = 2$ there is a direct one-for-one correspondence between the functional form of the log likelihood in Equation (6) and the Ridge-regression optimization problem specified in Equation (5). This match implies that using a Ridge regression is equivalent to Bayesian updating when the true slope coefficient, β^* , is drawn from a normal distribution—i.e., when a researcher has Gaussian priors—and the tuning parameter is chosen so that $\lambda = se^2/\sigma^2$. What is more, since the most likely estimate for the size of the cross-sectional slope coefficient will also deliver the lowest prediction error, this match implies that we can learn about the prior distribution for β^* —i.e., about the prevailing anomaly baserate—by finding the best-fit tuning parameter and then inverting the formula for λ to obtain an estimate, v^2 , for the prior variance, σ^2 . We will write the forecasted value of this estimator each month as \bar{v}_t . And, this forecast will represent our best guess about the anomaly baserate based on our experience with other previously discovered predictors.

Estimation Results. We next apply this statistical approach to analyze the cross-section of monthly returns from May 1973 to May 2015. Our analysis considers a collection of 83 different predictors found in the academic literature, $i = 1, \dots, 83$. We refer to the i th

predictor as an anomaly in month t if it is expected to have absolute returns in excess of some tradable threshold in month t . We recognize that the predictive power of some of these variables might not stem from a market error. This is perfectly fine. It could just as well come from some as-yet undiscovered risk factor. Our statistical approach works either way. We refer to strong predictors as tradable anomalies for convenience.

To get an estimate for $\Pr[\textit{anom}]$ in month t , we first run univariate Ridge regressions of excess returns on the lagged value of all $I \geq 2$ predictors discovered before t , month by month. We then take the best-fit tuning parameter for each predictor to get an estimate for the variance of the true slope coefficients, \hat{v}_i^2 , using the formula $\sigma^2 = se_i^2/\lambda_i$. If there are five predictors discovered prior to month t , then we compute five different in-sample estimates of $\hat{v}_{i,t}^2$ in month t . Next, we fit a time-series process to past values of these predictor-specific estimates to forecast the predictor variance in the following month, $\bar{v}_{i,t}^2$. We average these $I \geq 2$ forecasts for all previously-discovered predictors to arrive at \bar{v}_t , which we plot in Figure 1. If we assume the true slope coefficients, $\beta_{i,t}^*$, are drawn from a common normal distribution each month,¹ we can use this estimate to form a prior how likely it is that a newly discovered $(I + 1)$ st predictor in period t will have returns larger than a threshold, $\Pr[|\beta_{I+1,t}^*| > \textit{threshold}] = 2 \cdot \Phi[-\textit{threshold}/\bar{v}_t]$, where $\Phi[\cdot]$ represents the standard normal CDF. Note that it is not essential for us to assume normality; we discuss each step of the estimation process in more detail below.

Trading Strategy. Finally, we show how to use our forecasts for each predictor, $\bar{v}_{i,t}^2$, to combine these predictors into a single trading strategy. As pointed out in Lewellen (2015), there is substantial time-series variation in Fama-MacBeth slope coefficients for each predictor. So, the question is: how should you interpret the evidence for each previously discovered predictor given the prevailing anomaly baserate? We perform this exercise for two reasons. First, this exercise represents a useful real-world application. Just think about hitting **Ctrl+H** and replacing ‘seminar speaker’ with ‘quant researcher’ on the first two pages. Second, the success of this predictor-combination strategy gives evidence that our statistical approach to estimating the anomaly baserate is economically meaningful.

We start with a benchmark trading strategy that explicitly does not account for the anomaly baserate. The strategy holds an equal-weighted portfolio of all previously discovered predictors that have a forecasted return in month t higher than 1% per month. So, for example, if there were 5 predictors discovered prior to month t with forecasted returns higher than 1% per month, then the strategy would invest 1/5th of its assets in each of these 5 predictors. We use the 1% per month performance threshold to capture the idea that, in order for a predictor to be tradable, it must generate excess returns high enough to cover its

¹We discuss the economic motivation for this assumption in Section 2.1 below.

implementation costs. This benchmark only generates returns of 0.29% per month net of the 1% performance threshold during our sample period and has an annualized out-of-sample Sharpe ratio of 0.22.

We then adjust the benchmark strategy to account for the anomaly baserate. If the forecasted $\bar{v}_{i',t}$ was small for all other predictors $i' \neq i$ in month t and the realized returns for all predictors were drawn from a common distribution, then the i th predictor's true magnitude is likely small as well, $\beta_{i,t}^* \approx 0$. In which case, we should be more reluctant to trade on a large forecasted $\bar{\beta}_{i,t} \gg 0$. Following this logic, our baserate-adjusted strategy holds an equal-weighted portfolio of all predictors that still have one-month-ahead return forecasts higher than 1% per month after adjusting for the prevailing \bar{v}_t computed using all other previously discovered predictors $i' \neq i$. This strategy delivers excess returns of 0.68% per month net of the 1% threshold and has an annualized out-of-sample Sharpe ratio of 0.50.

1.1 Related Literature

We add to a large and still growing literature on cross-sectional return predictability in high dimensions. This literature is organized around three main topics.

Data Mining. The first topic is data mining (Lo and MacKinlay, 1990; Ferson, Sarkissian, and Simin, 1999; Sullivan, Timmermann, and White, 1999; White, 2000). Publication requires statistically significant results. So, there is an incentive for researchers to fiddle with a regression specification until they achieve statistical significance. For some predictors, this is impossible. But, for others, a little fiddling can flip results from marginally insignificant to strongly significant. If this is more likely to happen with spurious predictors, then data mining will increase the proportion of statistically significant predictors that are not tradable anomalies. You should adjust the Bayes factor you are using so that your posteriors respond less to statistically significant results.

And, there has been a recent string of influential papers on data mining in financial economics (Barras, Scaillet, and Wermers, 2010; Bajgrowicz and Scaillet, 2012; McLean and Pontiff, 2016; Harvey, Liu, and Zhu, 2016; Yan and Zheng, 2017; Harvey and Liu, 2018c,b,a; Linnainmaa and Roberts, 2018). Many researchers have spent a lot of time thinking about data mining. So, it is important to emphasize that this is not a paper about data mining. We are doing something different. We are proposing a way for estimating the anomaly baserate, $\Pr[anom]$, not for adjusting the Bayes factor, $\left(\frac{\Pr[signif | anom]}{\Pr[signif]}\right)$. But, this is not an either/or situation. You should use the above papers to adjust your Bayes factor for data mining when interpreting the seminar speaker's evidence. Then, you should apply this data-mining adjusted Bayes factor to the prevailing anomaly baserate, which we show how to estimate in this paper.

Factor Structure. The second topic is the factor structure of predictors. Even if there are many statistically significant predictors, it still might be possible to summarize the information in all these predictors using a few well-chosen factors? The goal here is to simplify investors' lives by collapsing the anomaly zoo down into a few manageable variables. Ideally, once you condition on these variables, there would be no incremental value in considering any other predictors when forecasting returns. And, researchers typically try to accomplish this goal by applying some form of principal-component analysis (Kelly, Pruitt, and Su, 2017; Green, Hand, and Zhang, 2017; Kelly, Pruitt, and Su, 2018; Lettau and Pelger, 2018).

If you are interested in forecasting returns, then it would be really useful to summarize the information in the anomaly zoo with a few manageable variables. But, we are not primarily interested in forecasting returns. Our goal is to learn whether a particular predictor is a tradable anomaly; our goal is to learn about the underlying structure of the market. And, forecasting an important outcome is not the same thing as learning about the underlying structure. Summarizing the forecasting power in all statistically significant predictors with a few well-chosen factors does not help you learn which individual predictors are tradable anomalies. These are two very interesting but logically distinct goals. By analogy, I can forecast who will be a good public speaker with a single summary variable: Were they a member of the college debate team? But, this summary variable does not reveal anything about the underlying structure of the brain, which makes speech possible.

Penalized Regressions. Finally, many recent papers have used penalized-regression procedures to solve asset-pricing problems (DeMiguel, Garlappi, Nogales, and Uppal, 2009; Bryzgalova, 2017; Feng, Giglio, and Xiu, 2017; Freyberger, Neuhierl, and Weber, 2017; Ledoit and Wolf, 2017; Chinco, Clark-Joseph, and Ye, 2018; Kozak, Nagel, and Santosh, 2018). These papers impose penalty functions for reasons related to both of the topics mentioned above. For example, the primary reason for imposing a penalty function in Freyberger, Neuhierl, and Weber (2017) and Chinco, Clark-Joseph, and Ye (2018) is to rule out weak predictors. Whereas, Bryzgalova (2017) and Kozak, Nagel, and Santosh (2018) mainly impose a penalty function to look for an economically motivated factor structure. Other papers motivate their analysis with some combination of these two concerns.

We are using a penalized-regression with an entirely different goal in mind: estimating the anomaly baserate. If there were only one or two predictors to choose from, then you would not need to use a penalized regression in these earlier papers. By contrast, it is important to use the correct anomaly baserate regardless of the number of candidate predictors. Thus, while we are using the same toolkit, we are using this toolkit to solve a different kind of problem, a kind of problem that exists even in the absence of the anomaly zoo. To our knowledge, this is the first instance where insights from machine learning are being used to shape our

understanding of a more fundamental asset-pricing problem, a problem that would exist even in a low-dimensional setting.

2 Statistical Approach

Let's return to the inference problem described on the opening two pages. You are a financial economist sitting at your weekly research seminar. The seminar speaker is showing that it was profitable to trade on a new variable, X_n , in the past. You want to figure out how likely it is that it will be profitable to trade on X_n in the future. To do this correctly, you need to start out with the right anomaly base rate and then appropriately update this prior given the seminar speaker's evidence. Your financial-econometrics training tells you how to appropriately update your prior beliefs. This section describes our statistical approach to estimating the anomaly base rate. You have seen evidence on $I \geq 2$ cross-sectional predictors in past seminars. Some of these predictors turned out to be tradable anomalies; others were spurious. You want to use your past experience with these other predictors to inform your ex-ante beliefs about what to expect from the current seminar speaker's presentation.

2.1 Inference Problem

We begin by defining the inference problem you face in more detail.

Statistical Framework. Suppose there are $N + 1$ stocks in the market, which are indexed by $n = 1, \dots, N + 1$. And, let R_n denote the excess return of stock n in the current month. In the analysis below, we are going to use the convention that you have already seen the first $I \geq 2$ predictors. And, you want to use this information to construct a prior for use in evaluating the seminar speaker's evidence about the $(I + 1)$ st predictor.

Assume that each stock's excess return this month is related to lagged values of the i th predictor as follows:

$$R_n = \mu^* + \beta_i^* \cdot X_{n,i} + \varepsilon_{n,i}^* \quad \text{for predictors } i = 1, \dots, I + 1. \quad (7)$$

In the data-generating process above, $X_{n,i}$ is the lagged value of the i th predictor for stock n , the parameter μ^* is the average excess return in the current month, the parameter β_i^* denotes the true slope coefficient associated with the i th predictor, and $\varepsilon_{n,i}^* \stackrel{\text{iid}}{\sim} \text{Normal}[0, N \cdot se_i^2]$ denotes the residual excess return for stock n . In other words, $\varepsilon_{n,i}^*$ represents the portion of stock n 's excess return that is not explained by the i th predictor. These residuals might come from either of two sources: idiosyncratic shocks that are fundamentally random or the effects of other cross-sectional predictors $i' \neq i$. Note that these other cross-sectional predictors may be as-yet undiscovered—i.e., it might be the case that $i' \notin \{1, \dots, I + 1\}$.

There are two technical details that are worth mentioning about the data-generating process defined in Equation (7). First, we define $\text{Var}[\varepsilon_{n,i}^*] = N \cdot se_i^2$ rather than just se_i^2 so that the parameter se_i represents the standard error when using only the i th predictor to explain the cross-section of excess returns. Lower values of se_i^2 mean that the i th predictor does a better job of fitting the observed data in sample. And, you can estimate se_i using an OLS regression:

$$N \cdot \widehat{se}_i^2 = \frac{1}{N-1} \cdot \sum_n (R_n - \hat{\mu} - \hat{\beta}_i \cdot X_{n,i})^2.$$

Second, we normalize the lagged values of each predictor to have zero mean and unit variance, $\frac{1}{N+1} \cdot \sum_n X_{n,i} = 0$ and $\frac{1}{N} \cdot \sum_n (X_{n,i} - 0)^2 = 1$. This normalization ensures that our estimates of $\hat{\beta}_i$ are comparable across predictors. To see why this is important, consider the following example. Suppose that the first predictor corresponds to a portfolio that is long/short the top/bottom deciles while the second predictor corresponds to a portfolio that is long/short the top/bottom quintiles. The first predictor will have a variance of $\text{Var}[X_{n,1}] = \frac{1}{10} \cdot (+1)^2 + \frac{1}{10} \cdot (-1)^2 = 1/5$; whereas, the second predictor will have a variance of $\text{Var}[X_{n,2}] = \frac{1}{5} \cdot (+1)^2 + \frac{1}{5} \cdot (-1)^2 = 2/5$. Thus, since $\hat{\beta}_i \stackrel{\text{def}}{=} \frac{\text{Cov}[R_n, X_{n,i}]}{\text{Var}[X_{n,i}]}$, we should expect that $\hat{\beta}_1 > \hat{\beta}_2$ even if both predictors have equal predictive power.

Discovery Process. You want to inform your beliefs about the current seminar speaker's predictor based on your experience with $I \geq 2$ other predictors that you have seen in past research seminars, then it must be the case that these $(I + 1)$ variables all share something in common. If each predictor were its own special butterfly, then none of your past experience with the first I predictors would be helpful when constructing priors for use with the $(I + 1)$ st predictor. But, how exactly should we expect the $(I + 1)$ cross-sectional predictors to be related? The answer is far from obvious.

It seems unlikely that there is a single, concise, unified explanation for why each variable $X_{n,i}$ does or does not predict the cross-section of expected returns. They are all so different. Cross-sectional predictability might be the result of any number of different limits-to-arbitrage models (Barberis and Thaler, 2003; Gromb and Vayanos, 2010). There is no shortage of risk-based explanations to choose from either (Fama and French, 1996). And, different predictors can also be based on entirely different data sources. Some predictors only involve past market data (e.g., medium-term momentum; Jegadeesh and Titman, 1993). Others use only accounting data (e.g., investment growth; Titman, Wei, and Xie, 2004). Why would there be a single explanation for the predictability associated with both medium-term momentum and investment growth?

So, rather than modeling the reason why each individual variable either does or does not predict the cross-section of expected returns, we instead assume that the strength of each

predictor is drawn from a common distribution:

$$\beta_i^* \stackrel{\text{iid}}{\sim} \text{Normal}[0, \sigma^2]. \quad (8)$$

In other words, we represent the required commonality across predictors with a statistical model for the anomaly discovery process. The key assumption embedded in Equation (8) is that there is a single parameter σ^2 governing the range of strengths for the predictors that financial economists discover. It is not essential to assume normality and independence across predictors; we show how to relax them in Appendix B. The mean-zero assumption is also not crucial, and in our empirical analysis we will be using predictors with both positive and negative β_i s.

The motivation for this assumption is that, even if the economic mechanism for why $X_{n,i}$ and $X_{n,i'}$ predict the cross-section of expected returns are completely different, these variables were still discovered by financial economists using a similar mental toolkit. We have all attended the same PhD programs. We have taken the same courses. We have all been trained by the same advisors. We have all been given access to the same data sources. So, while some researchers are better at searching for new predictors than others, everyone's search process is constrained by the same inputs. Notice that even though the economic rationales for using medium-term momentum and investment growth to predict the cross-section of expected returns are quite different, Sheridan Titman was a co-author on both papers. And, it seems plausible that the strength of these predictors is governed by a common distribution.

The idea of trying to estimate the volatility of cross-sectional predictors in spite of the fact that the economic mechanism behind each predictor is quite different is similar in spirit to the idea of estimating stock-market volatility (Andersen, Bollerslev, Diebold, and Labys, 2003) or an uncertainty index (Baker, Bloom, and Davis, 2016). We certainly do not think that there is a single unified explanation for all stock-market fluctuations. And, almost by definition, uncertainty can rise for any number of reasons. If your first reaction to reading this subsection was that it would be really interesting to have a more detailed prior-volatility model, then we are in complete agreement. However, to the best of our knowledge, we are first to point out the need for one. We view this as a key contribution of the paper. So, we think it is best to start with something simple.

Inference Problem. If the strength of the i th predictor is drawn from the distribution in Equation (8), then the parameter σ controls the typical size of cross-sectional predictors in the market. A larger value of σ means that researchers are more likely to discover strong predictors. But, what constitutes a tradable anomaly? When exactly is a predictor strong

enough to be considered tradable? We call the i th predictor a tradable anomaly if its true predictive power exceeds some minimum performance threshold, $\mathbf{1}[\beta_i^* > \text{threshold}]$ for some $\text{threshold} \geq 0$. There are multiple ways to think about this minimum performance threshold. You can think of it as coming from trading costs (Novy-Marx and Velikov, 2015). Or, you can think about it more broadly as any form of implementation costs. If you are running a trading desk, then how strong does a predictor need to be before you start redeploying scarce resources so that you can start trading on it?

In our empirical analysis, we will typically set $\text{threshold} = 1\%$ per month for a trading strategy that is long/short the top/bottom deciles. But, the exact point estimate is not critically important. All that matters is that there is some threshold determining whether or not a predictor is a tradable anomaly. This assumption implies that calibrating your expectations for what today’s seminar speaker will say about the $(I + 1)$ st predictor is tantamount to learning about the prevailing value for σ .

Proposition 2.1 (Inference Problem). *Suppose that there exists some threshold > 0 . The anomaly baserate for use with the $(I + 1)$ st predictor is given by*

$$\Pr[|\beta_{I+1}^*| > \text{threshold}] = 2 \cdot \Phi[-\text{threshold}/\sigma]$$

where $\Phi[\cdot]$ represents the standard normal CDF.

Suppose the seminar speaker is presenting evidence that the $(I + 1)$ st candidate predictor is very strong, $|\hat{\beta}_{I+1}| \gg \text{threshold}$. If your past experience with the other $I \geq 2$ predictors tells you that $\sigma \gg 0$, then this result is quite plausible. It is quite common for a researcher to draw a β_i^* far from zero. That is what it means for $\sigma \gg 0$. By contrast, if your past experience suggests that $\sigma \approx 0$, then the seminar speaker’s result will strike you as highly implausible. You will rationally discount the seminar speaker’s evidence, not because he did anything wrong, but because you think his result is very unlikely to begin with. If you believe that $\sigma \approx 0$, then you believe that tradable anomalies are quite rare. In fact, if your past experience dictates that $\sigma = 0$, then you will not even bother showing up to the seminar this week. There is nothing that the seminar speaking could possibly say to convince you that $\beta_{I+1}^* \neq 0$. This is known as having ‘dogmatic priors’.

2.2 Econometric Estimator

Having defined the inference problem, we now describe how to generate an estimator for σ , which we will denote by v , based on your past experience with $I > 1$ other predictors by searching for the best-fit tuning parameter in a penalized regression.

Ridge Regression. We study a penalized-regression procedure known as the Ridge regression (Hoerl and Kennard, 1970). A Ridge regression combines a standard OLS regression with a quadratic penalty that shrinks OLS-regression coefficients towards zero. Estimating the slope coefficient $\hat{\beta}_i[\lambda]$ for the i th predictor using a Ridge regression means solving the following optimization problem:

$$\hat{\beta}_i[\lambda] \stackrel{\text{def}}{=} \arg \min_{\beta} \left\{ \frac{1}{N} \cdot \sum_n (R_n - \hat{\mu} - \beta \cdot X_{n,i})^2 + \lambda \cdot \beta^2 \right\}. \quad (9)$$

Above, $\lambda \cdot \beta^2$ represents the quadratic penalty, and $\lambda \geq 0$ is known as the ‘tuning parameter’.

To create an estimate for the prior volatility that you should use when learning about the $(I+1)$ st predictor, we first estimate separate Ridge regressions for each predictor $i \in \{1, \dots, I\}$ that you have had past experience with. And, because we do this separately for each predictor, it is possible to analytically solve for the resulting Ridge-regression slope coefficient:

$$\hat{\beta}_i[\lambda] = \left(\frac{1}{1+\lambda} \right) \cdot \hat{\beta}_i. \quad (10)$$

The formula above indicates that when $\lambda = 0$, the penalty function disappears and the results of a Ridge regression coincide with those of a standard OLS regression, $\hat{\beta}_i = \hat{\beta}_i[0]$. But, for all choices of $\lambda > 0$, the quadratic penalty in Equation (9) will shrink the standard OLS-regression coefficient toward zero, with larger choices of λ resulting in more shrinkage.

Bayesian Interpretation. The key insight motivating our statistical approach is that there is a Bayesian interpretation for the shrinkage imposed by λ . You can interpret it as the effect of incorporating your prior beliefs about the distribution of the true coefficient β^* when these coefficient values are also drawn from a normal distribution. The negative log likelihood of the true slope coefficient taking on a particular value, $\beta_i^* = \beta$, given the realized data, $\{R_1, \dots, R_{N+1}\}$ and $\{X_{1,i}, \dots, X_{N+1,i}\}$, corresponds to

$$\begin{aligned} -\log \Pr[\beta] &= \frac{1}{2 \cdot (N \cdot se_i^2)} \cdot \sum_n (R_n - \hat{\mu} - \beta \cdot X_{n,i})^2 + \frac{1}{2 \cdot \sigma^2} \cdot (\beta - 0)^2 + \dots \\ &= \frac{1}{2 \cdot se_i^2} \cdot \left\{ \frac{1}{N} \cdot \sum_n (R_n - \hat{\mu} - \beta \cdot X_{n,i})^2 + \frac{se_i^2}{\sigma^2} \cdot \beta^2 \right\} + \dots \end{aligned} \quad (11)$$

where the “ \dots ” represents constants that do not depend on the choice of β .

Thus, estimating a Ridge regression—i.e., solving the optimization problem in Equation (9)—is the same as finding the most likely estimate for β_i^* given your prior beliefs and the observed data—i.e., minimizing the negative posterior log likelihood in Equation (11)—when using the appropriate tuning parameter:

$$\lambda_i = se_i^2 / \sigma^2.$$

And, since the fully Bayesian estimate will have the lowest prediction error on average, it stands to reason that we can use each of the $I > 1$ predictors that you have had past experience with to learn about the true value of σ by searching for the best-fit tuning parameter λ . Note that it is also possible to estimate the mean as well as the variance of the prior distribution for β^* . We focus on the prior variance because there is no ex-ante reason to suspect the mean to vary over time since predictors have no preferred sign. And, in unreported results, we confirm that this intuition holds true in the data.

Empirical/Objective Bayes. The idea of using frequentist statistics describing past outcomes to inform your prior beliefs for use in future Bayesian analysis goes by the name of ‘empirical Bayes’ (Robbins, 1956) or ‘objective Bayesian analysis’ (Berger, 2006) in the statistics literature. For other applications of these ideas in the finance literature, see Frost and Savarino (1986), Karolyi (1993), and Harvey and Liu (2018a). We use the Ridge regression in our main analysis because we are assuming in our stylized model of the anomaly-discovery process that the true predictor strengths are drawn from a normal distribution (see Equation 8). Using a different penalized-regression procedure would mean making a different assumption about the distribution of predictor strengths. For example, Park and Casella (2008) shows that using the LASSO is tantamount to adopting a Laplace prior. But, the basic insight underpinning both empirical and objective Bayesian thinking is the same. “Any sensible estimator is Bayesian for some prior (Diaconis and Skyrms, 2017).”

In-Sample Overfitting. How exactly should you go about constructing an estimator for σ , though? One approach you could take would be to choose the tuning parameter $\lambda_i > 0$ that best fits the data you have observed for each cross-sectional predictor and then invert the formula $\lambda_i = se_i^2/\sigma^2$ to solve for σ . Let $\text{Err}_i[\lambda]$ denote the Ridge regression’s in-sample prediction error when using a particular value of λ given the realized cross-section of excess returns in particular month and lagged values of the i th predictor, $\{R_1, \dots, R_{N+1}\}$ and $\{X_{1,i}, \dots, X_{N+1,i}\}$:

$$\text{Err}_i[\lambda] \stackrel{\text{def}}{=} \text{E} \left[(R_n - \hat{\mu} - \hat{\beta}_i[\lambda] \cdot X_{n,i})^2 \right]. \quad (12)$$

This error is often called ‘training error’ (Hastie, Tibshirani, and Friedman, 2001).

The lemma below shows that this approach is too naïve. No matter what the true value of σ^2 is, the training error associated with the Ridge regression will always be minimized by setting $\lambda_i = 0$. In other words, an OLS regression will always outperform a Ridge regression in sample according to this metric.

Lemma 2.2 (In-Sample Overfitting). *Let $\text{E}[\cdot]$ denote an expectations operator evaluated with respect to realizations of β_i^* . If \tilde{v}_i^2 denotes the parameter estimate with the minimum in-sample*

prediction error for the i th predictor,

$$\tilde{v}_i^2 \stackrel{\text{def}}{=} \arg \min_{v^2 > 0} \left\{ \text{Err}_i[se_i^2/v^2] \right\},$$

then we have that $E[\tilde{v}_i^2] = \infty$ regardless of which $\sigma^2 > 0$ was used to generate the data.

Minimizing the training error requires you to fine-tune the slope coefficient to explain variation in excess returns coming from in-sample noise. And, this is easiest when there is no penalty for doing so—i.e., when $\lambda_i = 0$; or equivalently, when $\tilde{v}_i^2 = \infty$. Put another way, the ratio $\lambda_i = se_i^2/\sigma^2$ governs the relative likelihood that a statistically significant estimate for the slope coefficient, $|\hat{\beta}_i| \gg 0$, is due to in-sample overfitting rather than the existence of an honest-to-goodness tradable anomaly, $\beta_i^* \neq 0$. But, the estimator \tilde{v}_i^2 does not reflect this comparison. It only reveals whether the in-sample fit is good; it does not tell you anything about the origins of this good performance.

Econometric Estimator. If we want a consistent estimator for σ^2 , then we need to adjust the metric used in Lemma (2.2) so that it does not reward in-sample overfitting. One way to do this would be to look at out-of-sample prediction errors—i.e., to estimate σ^2 via cross-validation (Stone, 1974). But, since we specified a data-generating process in the previous subsection, we can actually do better. We can directly adjust the training error for the expected amount of in-sample overfitting given σ^2 . Large values of σ^2 will result in lots of in-sample overfitting, so these choices should receive a large penalty. Whereas, values of σ^2 that are close to zero will result in very little in-sample overfitting, so these choices should only be penalized a little.

Proposition 2.2 (Econometric Estimator). *Let $E[\cdot]$ denote an expectations operator evaluated with respect to realizations of β_i^* . If v_i^2 denotes the parameter estimate with the minimum in-sample prediction error subject to an overfitting penalty for the i th predictor,*

$$v_i^2 \stackrel{\text{def}}{=} \arg \min_{v^2 > 0} \left\{ \text{Err}_i[se_i^2/v^2] + 2 \cdot \left(\frac{1}{1+se_i^2/v^2} \right) \cdot se_i^2 \right\}, \quad (13)$$

then for all $\sigma^2 > 0$ we have that $E[v_i^2] = \sigma^2$.

Note that this econometric estimator follows from the same basic intuition as many other information-theoretic model-selection criteria, such as the Akaike information criterion (AIC Akaike, 1974), which minimize a mean squared-error loss function plus an additional penalty proportional to the number of degrees of freedom, df , in the model times the noise variance, $\text{Var}[\varepsilon_{n,i}]$. In fact, one way to derive the penalty function in Equation (13) is to note that the effective degrees of freedom in a univariate Ridge regression is given by $1/(1 + \lambda)$. Thus, since

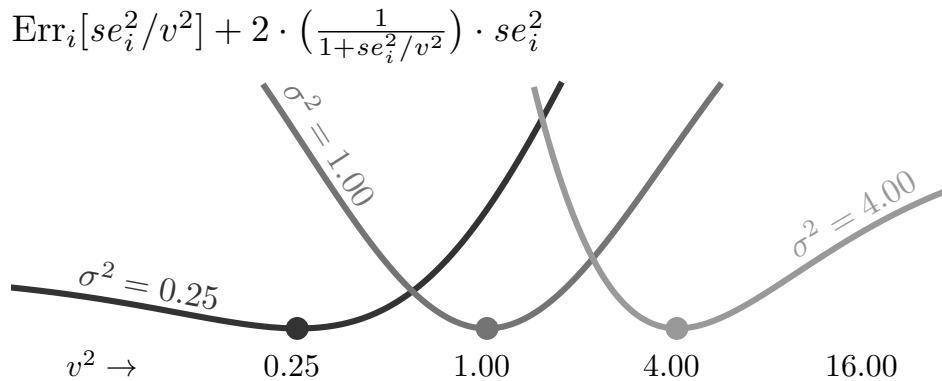


Figure 2. Numerical Simulation. Results for 1,000 simulations of a market with $N + 1 = 1,000$ stocks. For each simulation, we generate a new realization of the cross-section of excess returns using the parameters $\mu^* = 0$, $\beta_i^* \stackrel{\text{iid}}{\sim} \text{Normal}[0, \sigma^2]$, and $\varepsilon_{n,i}^* \stackrel{\text{iid}}{\sim} \text{Normal}[0, N \cdot 1.00\%^2]$. Solid lines: value of the objective function within the curly braces in Equation (13) averaged across all simulation at each value of $v^2 \in (0, 16]$. Different lines correspond to different values of $\sigma^2 \in \{0.25\%, 1.00\%, 4.00\%\}$. Large dot: v^2 values that minimize each curve, which correspond to the output of the estimator defined in Proposition 2.2. x -axis: input to objective function, $v^2 > 0$, on a log scale.

$\text{Var}[\varepsilon_{n,i}] = N \cdot se_i^2$, the Akaike penalty function, $2 \cdot (df/N) \times \text{Var}[\varepsilon_{n,i}]$, reduces to the one above when $\lambda_i = se_i^2/\sigma^2$. By the same logic, Stone (1977) shows that using Equation (13) to estimate σ^2 will deliver estimates that are asymptotically equivalent to those of cross-validation.

Numerical Simulation. Before taking it to the data, we first verify the econometric estimator from Proposition 2.2's performance using numerical simulations. We study three different anomaly-discovery regimes, $\sigma^2 \in \{0.25\%, 1.00\%, 4.00\%\}$. $\sigma^2 = 4.00\%$ denotes a regime where tradable anomalies are likely; $\sigma^2 = 0.25\%$ denotes a regime where tradable anomalies are unlikely; and, $\sigma^2 = 1.00\%$ denotes a regime somewhere in between. For each regime, we run 1,000 simulations of a market with $N + 1 = 1,000$ stocks. For each simulation, we generate a cross-section of excess returns using the data-generating process in Equation (7) with parameters $\mu^* = 0$ and $se_i^2 = 1.00\%$. The solid lines in Figure 2 report the value of the objective function within the curly braces in Equation (13) at each input parameter $v^2 > 0$ averaged across all simulation. The figure shows that, when we simulate data using a particular value of σ^2 , the objective function in Proposition 2.2 is minimized at this value, which means that our econometric estimator is able to recover the parameter value from simulated data.

Forecasting vs. Learning. We have just outlined a univariate approach to estimating σ^2 . In essence, we are asking: 'How can we separately use the information in each of the first $I > 1$ predictors to learn about σ^2 ?' We are not looking for the best combination of predictors. But, that is not to say looking for the best combination of predictors is wrong. It just depends on

what you are after. Taming the factor zoo (Feng, Giglio, and Xiu, 2017) will mean different things to different people.

If all you care about is making good forecasts, then taming the factor zoo will mean looking for the best combination of predictors. You should use something like principle-component analysis to collapse the information in all $I > 1$ predictors down into a single forecasting variable (Kelly, Pruitt, and Su, 2017). But, if you want to use these same $I > 1$ predictors to learn about the market’s underlying structure, then you do not want to combine predictors. This would throw away valuable information and make it harder for you to learn.

Consider an example. Driving a car is a complicated activity; there are lots of different factors involved. It has taken quite a while for engineers to teach a computer how to do it.² But, it is nevertheless easy to forecast who will be a safe driver. You can collapse all the information in the zoo of driving-related factors into a single variable: $\text{IsTeenager} \in \{True, False\}$ (Jonah, 1986). If you are an insurance underwriter trying to forecast future claims, this single variable effectively tames the factor zoo. But, if you are a Google engineer, it does not. You cannot create a self-driving car by *not installing* a 17-year-old operating system.

3 Estimation Results

Having described our new statistical approach, we next apply this approach to recover an estimate for σ each month. We use data the cross-section of excess returns each month, and we study a collection of 83 different predictors that were published in the academic literature some time after May 1973.

3.1 Data Description

We begin by describing the data and variables used in our analysis.

Data Sources. We study the cross-section of monthly returns from May 1973 to May 2015 for each U.S. stock traded on either the NYSE, Amex, or NASDAQ. These data comes from the Center for Research in Security Prices (CRSP) monthly stock file. To make sure that our results are not being driven by small illiquid stocks, we exclude any stock with a price below \$1 at the end of the previous month. We use balance-sheet data from the Standard and Poor’s Compustat database. All items are taken from the fiscal year ending in calendar year $y - 1$ for estimation starting in June of year y until May of year $y + 1$ predicting returns from July of year y until June of year $y + 1$. To alleviate a potential survivorship bias due to back-filling, we require that a firm has at least two years of Compustat data for it to be included in our sample. Let N_t denote the number of stocks in our sample in month t .

²Wired. 12/13/2018. *The Wired Guide to Self-Driving Cars.*

Return Predictors. We use a collection of 83 different cross-sectional predictors that were first documented in the academic literature sometime on or after May 1973. We list each predictor along with its publication date in Tables 1a, 1b, and 1c. Let \mathcal{I}_t denote the set of predictors discovered prior to month t :

$$\mathcal{I}_t \stackrel{\text{def}}{=} \left\{ i \in \mathcal{I} : \text{publication date for } i\text{th predictor} < t \right\}.$$

And, let $I_t \stackrel{\text{def}}{=} |\mathcal{I}_t|$ denote the number of predictors discovered prior to month t . So, for example, looking at the first four rows of Table 1a, we have $I_{\text{Apr}73} = 0$, $I_{\text{May}73} = 1$, \dots , $I_{\text{May}77} = 1$, $I_{\text{Jun}77} = 2$, \dots , $I_{\text{May}79} = 2$, and $I_{\text{Jun}79} = 4$.

Slope Coefficients. We compute the realized returns to a zero-cost strategy based on each previously predictor $i \in \mathcal{I}_t$ by running a separate cross-sectional OLS-regression in each month after normalizing the predictor to have zero mean and unit variance:

$$R_{n,t} = \hat{\mu}_t + \hat{\beta}_{i,t} \cdot X_{n,i,t-1} + \hat{\varepsilon}_{n,i,t}.$$

$R_{n,t}$ is the excess return of the n th stock in month t , $\hat{\mu}_t$ is the cross-sectional average excess return for all stocks in our sample during month t , $X_{n,i,t-1}$ is the value of the i th predictor for stock n in the previous month, $\hat{\beta}_{i,t}$ is the OLS-regression coefficient for the i th predictor in month t , and $\hat{\varepsilon}_{n,i,t}$ is the regression residual.

Table 2 provides summary statistics describing these realized returns. There are two things about this table that are worth pointing out. First, contrarian predictors, such as the long-term reversals captured by the predictor ‘*Ret, 36-13*’ in row five, will result in estimated values that are negative on average. This is consistent with the modeling assumption that the true $\beta_{i,t}^*$ values are drawn from a mean-zero normal distribution. When we incorporate these sorts of predictors in a trading strategy, we will always trade in the appropriate direction.

Second, our estimates for each predictor’s $\hat{\beta}_{i,t}$ tend to be smaller than the ones reported in the original papers. This is because researchers typically report the excess returns to sorted high-minus-low portfolios, which is tantamount to running a cross-sectional regression where $\text{Var}[X_{n,i,t-1}] < 1$. For example, going long the top 10% of stocks and short the bottom 10% of stocks corresponds to $\text{Var}[X_{n,i,t-1}] = \frac{1}{10} \cdot (+1)^2 + \frac{1}{10} \cdot (-1)^2 = 1/5$. And, since $\hat{\beta}_{i,t} = \text{Cov}[R_{n,t}, X_{n,i,t-1}] / \text{Var}[X_{n,i,t-1}]$, this approach would result in point estimates that are five times larger than ours. We will adjust for this difference when we consider the effects of implementation costs before combining predictors into a single strategy in Section 4 below.

Name	P. Date	Description
1. <i>BetaSq</i>	1973-05	Rolling CAPM beta, squared
2. <i>Earn/Share</i>	1977-06	Earnings per share
3. <i>Debt/Price</i>	1979-06	Debt to price
4. <i>Divd/Price</i>	1979-06	Dividend to price
5. <i>Mcap</i>	1981-03	Market capitalization, prev. fiscal year
6. <i>Earn/Price</i>	1982-08	Earnings to price
7. <i>Ret, 36-12</i>	1985-07	Cum. return, months $[-36, -12)$
8. <i>AvgSpread</i>	1986-12	Mean bid-ask spread
9. <i>Assets/Mcap</i>	1988-07	Assets to market cap
10. <i>Levrg</i>	1988-06	Leverage
11. <i>Levrg/Price</i>	1988-06	Leverage to price
12. <i>Sales/Cash</i>	1989-11	Sales to cash
13. <i>LtCF</i>	1989-11	Long-term cash flow
14. <i>CurrRatio</i>	1989-11	Current ratio
15. $\% \Delta \text{CurrRatio}$	1989-11	Perc. change in current ratio
16. $\% \Delta \text{QuickRatio}$	1989-11	Perc. change in quick ratio
17. $\% \Delta [\text{Sales}/\text{Invtry}]$	1989-11	Perc. change in sales to inventory
18. <i>QuickRatio</i>	1989-11	Quick ratio
19. <i>Sales/Invtry</i>	1989-11	Sales to inventory
20. <i>Sales/Recv</i>	1989-11	Sales to receivables
21. <i>Ret, 1-0</i>	1990-07	Return, month $[-1, 0)$
22. <i>Ret, 12-1</i>	1990-07	Cum. return, months $[-12, -1)$
23. <i>BkVal</i>	1992-06	Book value
24. <i>MonthlyMcap</i>	1992-06	Market cap, prev. month
25. <i>Sales/Price</i>	1992-06	Sales to price
26. $\% \Delta [\text{Deprc}/\text{PP\&E}]$	1992-09	Perc. change in depreciation to PP&E
27. $D\&A/\text{Assets}$	1992-09	D&A to assets
28. <i>Deprc/PP&E</i>	1992-09	Depreciation to PP&E

Table 1a. List of Predictors. List of variables that predict the cross-section of expected returns that have been documented in the academic literature sometime on or after May 1973. Predictors are constructed using data from CRSP and Compustat. Name: The name for the predictor used throughout this paper. P. Date: The month of publication for the first academic paper about each predictor. Description: A description of how predictor is constructed for each stock.

	Name	P. Date	Description
29.	<i>Ret</i> , 6-1	1993-03	Cum. return, months $[-6, -1]$
30.	$\% \Delta Sales$	1994-12	Perc. change in sales
31.	<i>OpAccr</i>	1996-07	Operating accruals
32.	<i>CapitalTOver</i>	1996-07	Capital turnover
33.	<i>RetOnEquity</i>	1996-07	Return on equity
34.	<i>KaplanZingales</i>	1997-02	Kaplan-Zingales index
35.	$\% \Delta [\Delta Sales / \Delta Invtry]$	1997-04	Perc. change in $\Delta Sales$ to $\Delta Inventory$
36.	$\% \Delta [\Delta Sales / \Delta Recv]$	1997-04	Perc. change in $\Delta Sales$ to $\Delta Receivables$
37.	$\% \Delta [\Delta Sales / \Delta XG\&A]$	1997-04	Perc. change in $\Delta Sales$ to $\Delta XG\&A$
38.	$\% \Delta [\Delta GrMgn / \Delta Sales]$	1998-01	Perc. change in $\Delta Gross\ margin$ to $\Delta Sales$
39.	<i>LagTOver</i>	1998-09	Lagged turnover
40.	$Adj[BkVal/Mcap]$	2000-02	Ind. adjusted book-to-market ratio
41.	<i>AdjMcap</i>	2000-02	Ind. adjusted market cap
42.	<i>SdTOver</i>	2001-01	Std. deviation of daily turnover
43.	<i>AdvertRate/Ret</i>	2001-12	Advertising expense rate to returns
44.	$R\&D/Mcap$	2001-12	R&D to market cap
45.	$R\&D/Sales$	2001-12	R&D to sales
46.	<i>Advert/Mcap</i>	2001-12	Advertising expense to market cap
47.	$\Delta Invtry/Assets$	2002-06	Inventory changes to assets
48.	$OpCF/Price$	2004-04	Operating cash flow to price
49.	$Invmt/Lag[AugInvmt]$	2004-12	Investment to trailing 3 years average
50.	$NetOpAssets/Sales$	2004-12	Net operating assets to lagged sales
51.	$\% \Delta BkVal$	2005-09	Perc. change in book value
52.	$\% \Delta LtDebt$	2005-09	Perc. change in long-term debt
53.	<i>Price-52WkHi</i>	2005-11	Closeness to previous 52-week high
54.	<i>IdioVol</i>	2006-02	Idiosyncratic volatility
55.	<i>TotVol</i>	2006-02	Total volatility
56.	$\varepsilon \% \Delta Mcap$	2006-08	Residual perc. change in market cap

Table 1b. List of Predictors, Ctd. List of variables that predict the cross-section of expected returns that have been documented in the academic literature sometime on or after May 1973. Predictors are constructed using data from CRSP and Compustat. Name: The name for the predictor used throughout this paper. P. Date: The month of publication for the first academic paper about each predictor. Description: A description of how predictor is constructed for each stock.

	Name	P. Date	Description
57.	<i>NetExtnlFin/Assets</i>	2006-10	Net external financing to assets
58.	<i>Beta</i>	2006-11	Rolling CAPM beta
59.	<i>NetPO/Price</i>	2007-04	Net payouts to price
60.	<i>PO/Price</i>	2007-04	Payouts to price
61.	<i>NetPO</i>	2007-04	Net payout ratio
62.	<i>RetOnInvstCap</i>	2007-06	Return on invested capital
63.	<i>%ΔShares</i>	2008-04	Perc. change in shares outstanding
64.	<i>ProfMgn</i>	2008-05	Profit margin
65.	<i>AdjProfMgn</i>	2008-05	Ind. adjusted profit margin
66.	<i>RetOnOpAssets</i>	2008-05	Return on net operating assets
67.	<i>AssetTOver</i>	2008-05	Asset turnover
68.	<i>AdjAssets</i>	2008-05	Ind. adjusted total assets
69.	<i>%ΔInvmtX</i>	2008-07	Perc. change in investments (Xing)
70.	<i>%ΔInvmt</i>	2008-08	Perc. change in investments
71.	<i>ΔAdjShares</i>	2008-08	Change in split-adjusted shares outstanding
72.	<i>RetOnCash</i>	2009-01	Return on cash
73.	<i>Tangibility</i>	2009-04	Asset tangibility
74.	<i>ΔAdjTOver</i>	2009-10	Change in market-adjusted turnover
75.	<i>UnexplVlm</i>	2009-10	Standardized unexplained volume
76.	<i>RetOnAssets</i>	2010-05	Return on assets
77.	<i>OpLevrg</i>	2011-01	Operating leverage
78.	<i>MaxRet</i>	2011-01	Max monthly return during prev. year
79.	<i>FreeCF</i>	2011-05	Free cash flow
80.	<i>R&Dcapital</i>	2011-09	R&D capital
81.	<i>%ΔInvtry</i>	2012-01	Perc. change in inventory
82.	<i>Ret, 12-6</i>	2012-03	Cum. return, months [-12, -6)
83.	<i>CashHldgs</i>	2012-04	Cash holdings

Table 1c. List of Predictors, Ctd. List of variables that predict the cross-section of expected returns that have been documented in the academic literature sometime on or after May 1973. Predictors are constructed using data from CRSP and Compustat. Name: The name for the predictor used throughout this paper. P. Date: The month of publication for the first academic paper about each predictor. Description: A description of how predictor is constructed for each stock.

		Avg	Sd			Avg	Sd		Avg	Sd	
<i>BetaSq</i>		-0.05	2.29	<i>Ret, 6-1</i>		0.09	2.52	<i>NetExtnFin/Assets</i>	-	-0.11	0.58
<i>Earn/Share</i>		-0.02	2.31	<i>%ΔSales</i>		-0.22	0.87	<i>Beta</i>	~	0.06	1.74
<i>Debt/Price</i>		0.06	1.45	<i>OpAccr</i>		-0.11	0.72	<i>NetPO/Price</i>	~	0.06	1.30
<i>Divd/Price</i>		0.04	1.77	<i>CapitalTOver</i>		0.06	0.97	<i>PO/Price</i>	~	0.04	1.07
<i>Mcap</i>		-0.22	1.79	<i>RetOnEquity</i>		-0.07	2.16	<i>NetPO</i>	-	0.00	0.55
<i>Earn/Price</i>		0.05	2.11	<i>KaplanZingales</i>		-0.11	0.94	<i>RetOnInvstCap</i>	~	0.04	1.27
<i>Ret, 36-12</i>		-0.30	1.78	<i>%Δ[ΔSales/ΔInvtry]</i>		0.06	0.49	<i>%ΔShares</i>	-	0.11	0.39
<i>AvgSpread</i>		0.17	1.91	<i>%Δ[ΔSales/ΔRecv]</i>		0.08	0.54	<i>ProfMgn</i>	~	-0.18	1.52
<i>Assets/Mcap</i>		0.21	1.85	<i>%Δ[ΔSales/ΔXG&A]</i>		-0.06	0.57	<i>AdjProfMgn</i>	-	0.13	1.00
<i>Levrg</i>		-0.09	1.38	<i>%Δ[ΔGrMgn/ΔSales]</i>		0.06	0.60	<i>RetOnOpAssets</i>	~	0.02	0.82
<i>Levrg/Price</i>		0.01	1.59	<i>LagTOver</i>		0.11	2.50	<i>AssetTOver</i>	~	0.23	1.01
<i>Sales/Cash</i>		0.02	1.40	<i>Adj[BkVal/Mcap]</i>		0.29	0.93	<i>AdjAssets</i>	-	0.11	0.56
<i>LtCF</i>		0.10	1.97	<i>AdjMcap</i>		-0.11	1.44	<i>%ΔInvmtX</i>	-	-0.10	0.49
<i>CurrRatio</i>		0.02	1.28	<i>SdTOver</i>		0.10	1.85	<i>%ΔInvmt</i>	~	-0.28	0.88
<i>%ΔCurrRatio</i>		-0.14	0.60	<i>AdvertRate/Ret</i>		0.22	1.01	<i>ΔAdjShares</i>	~	-0.07	1.10
<i>%ΔQuickRatio</i>		-0.12	0.60	<i>R&D/Mcap</i>		0.24	1.73	<i>RetOnCash</i>	-	-0.01	0.90
<i>%Δ[Sales/Invtry]</i>		0.09	0.44	<i>R&D/Sales</i>		-0.06	2.02	<i>Tangibility</i>	-	-0.10	0.79
<i>QuickRatio</i>		0.03	1.56	<i>Advert/Mcap</i>		0.40	1.24	<i>ΔAdjTOver</i>	-	0.21	0.38
<i>Sales/Invtry</i>		0.01	0.87	<i>ΔInvtry/Assets</i>		-0.11	0.54	<i>UnexplVlm</i>	-	0.30	0.56
<i>Sales/Recv</i>		0.06	0.73	<i>OpCF/Price</i>		0.22	1.12	<i>RetOnAssets</i>	~	0.06	1.05
<i>Ret, 1-0</i>		-0.47	2.19	<i>Invmt/Lag[AvgInvmt]</i>		0.01	0.47	<i>OpLevrg</i>	-	0.06	0.55
<i>Ret, 12-1</i>		0.12	2.57	<i>NetOpAssets/Sales</i>		0.00	0.89	<i>MaxRet</i>	~	-0.23	1.15
<i>BkVal</i>		0.33	1.44	<i>%ΔBkVal</i>		-0.10	0.77	<i>FreeCF</i>	-	0.18	1.01
<i>MonthlyMcap</i>		-0.31	2.11	<i>%ΔLtDebt</i>		-0.09	0.43	<i>R&Dcapital</i>	-	0.09	1.30
<i>Sales/Price</i>		0.34	1.52	<i>Price-52WkHi</i>		-0.11	2.43	<i>%ΔInvtry</i>	-	-0.04	0.44
<i>%Δ[Deprc/PP&E]</i>		-0.01	0.48	<i>Idio Vol</i>		-0.10	1.79	<i>Ret, 12-6</i>	-	0.14	0.74
<i>D&A/Assets</i>		0.23	1.24	<i>TotVol</i>		-0.10	1.89	<i>CashHldgs</i>	-	0.02	0.98
<i>Deprc/PP&E</i>		0.09	1.80	<i>ε%ΔMcap</i>		-0.12	1.27				

Table 2. Estimated $\hat{\beta}_{i,t}$. Summary statistics describing the realized returns of zero-cost portfolios based on each predictor as defined in Equation (2). Sample Period: May 1973 to May 2015. Units: % per month. Sparkline Plots: time series for each predictor on a common scale, $-3\% < \hat{\beta}_{i,t} < 3\%$. Every time series ends in May 2015. Predictors discovered later have shorter sparkline plots. Red indicated negative returns.

		Avg	Sd			Avg	Sd		Avg	Sd
<i>BetaSq</i>		2.16	1.26	<i>Ret</i> , 6-1		2.31	2.84	<i>NetExtnlFin/Assets</i>	0.52	0.17
<i>Earn/Share</i>		1.48	0.92	<i>%ΔSales</i>		0.57	0.34	<i>Beta</i>	1.13	0.62
<i>Debt/Price</i>		0.92	0.68	<i>OpAccr</i>		0.43	0.25	<i>NetPO/Price</i>	0.84	0.26
<i>Divd/Price</i>		1.20	0.66	<i>CapitalTOver</i>		0.69	0.25	<i>PO/Price</i>	0.72	0.20
<i>Mcap</i>		1.21	0.53	<i>RetOnEquity</i>		1.35	0.84	<i>NetPO</i>	0.27	0.10
<i>Earn/Price</i>		1.34	0.85	<i>KaplanZingales</i>		0.58	0.33	<i>RetOnInvstCap</i>	0.84	0.27
<i>Ret</i> , 36-12		1.10	0.59	$\% \Delta [\Delta Sales / \Delta Invtry]$		0.24	0.13	$\% \Delta Shares$	0.22	0.06
<i>AvgSpread</i>		1.23	0.47	$\% \Delta [\Delta Sales / \Delta Recv]$		0.27	0.14	<i>ProfMgn</i>	1.04	0.47
<i>Assets/Mcap</i>		1.22	0.87	$\% \Delta [\Delta Sales / \Delta XG\&A]$		0.34	0.09	<i>AdjProfMgn</i>	0.62	0.37
<i>Levrg</i>		0.84	0.69	$\% \Delta [\Delta GrMgn / \Delta Sales]$		0.36	0.18	<i>RetOnOpAssets</i>	0.51	0.19
<i>Levrg/Price</i>		1.01	0.76	<i>LagTOver</i>		1.66	1.13	<i>AssetTOver</i>	0.66	0.37
<i>Sales/Cash</i>		0.84	0.66	$Adj [BkVal / Mcap]$		0.62	0.35	<i>AdjAssets</i>	0.35	0.13
<i>LtCF</i>		1.18	1.02	<i>AdjMcap</i>		0.79	0.83	$\% \Delta InvmtX$	0.31	0.12
<i>CurrRatio</i>		0.74	0.63	<i>SdTOver</i>		1.22	0.81	$\% \Delta Invmt$	0.56	0.17
$\% \Delta CurrRatio$		0.30	0.22	<i>AdvertRate/Ret</i>		0.58	0.27	$\Delta AdjShares$	0.74	0.31
$\% \Delta QuickRatio$		0.30	0.23	$R\&D / Mcap$		1.10	0.56	<i>RetOnCash</i>	0.60	0.23
$\% \Delta [Sales / Invtry]$		0.22	0.12	$R\&D / Sales$		1.23	0.72	<i>Tangibility</i>	0.50	0.21
<i>QuickRatio</i>		0.90	0.81	<i>Advert/Mcap</i>		0.83	0.37	$\Delta AdjTOver$	0.24	0.11
<i>Sales/Invtry</i>		0.53	0.25	$\Delta Invtry / Assets$		0.33	0.07	<i>UnexplVlm</i>	0.38	0.07
<i>Sales/Recv</i>		0.48	0.24	<i>OpCF/Price</i>		0.73	0.31	<i>RetOnAssets</i>	0.71	0.10
<i>Ret</i> , 1-0		1.19	0.82	$Invmt / Lag [AvgInvmt]$		0.27	0.13	<i>OpLevrg</i>	0.36	0.05
<i>Ret</i> , 12-1		1.49	0.94	$NetOpAssets / Sales$		0.60	0.23	<i>MaxRet</i>	0.86	0.15
<i>BkVal</i>		0.99	0.61	$\% \Delta BkVal$		0.42	0.22	<i>FreeCF</i>	0.70	0.16
<i>MonthlyMcap</i>		1.29	0.58	$\% \Delta LtDebt$		0.23	0.09	$R\&Dcapital$	0.83	0.21
<i>Sales/Price</i>		1.00	0.67	<i>Price-52WkHi</i>		1.61	0.94	$\% \Delta Invtry$	0.22	0.04
$\% \Delta [Deprc / PP\&E]$		0.23	0.14	<i>IdioVol</i>		1.23	0.55	<i>Ret</i> , 12-6	0.54	0.19
$D\&A / Assets$		0.83	0.44	<i>TotVol</i>		1.30	0.51	<i>CashHldgs</i>	0.69	0.19
<i>Deprc/PP&E</i>		1.03	0.92	$\varepsilon \% \Delta Mcap$		0.91	0.42			

Table 3. Forecasted $\bar{v}_{i,t}$. Summary statistics describing the one-month-ahead forecasts for $\bar{v}_{i,t}$. We first estimate the in-sample parameter $\hat{v}_{i,t}^2$ separately each month using the procedure described in Proposition 2.2. We then make one-month-ahead forecasts by fitting an AR(3) model to squared values in months $\{t - 60, \dots, t - 1\}$. Sample Period: May 1973 to May 2015. Units: % per month. Sparkline Plots: time series of each predictor-specific forecasts on a common scale, $0\% < \bar{v}_{i,t} < 6\%$. Every time series ends in May 2015. Predictors discovered later have shorter sparkline plots.

3.2 Anomaly Baserate

We first use this data to create estimates $\hat{v}_{i,t}^2$ for σ^2 each month for all previously discovered predictors, $i \in \mathcal{I}_t$. We then use the time series associated with each predictor to make one-month-ahead forecasts $\bar{v}_{i,t}^2$ for use when evaluating the next predictor you encounter.

Predictor-Specific Estimates. We start by creating separate estimates for σ^2 using every previously discovered predictor in a given month. For each month t such that $i \in \mathcal{I}_t$, we first solve the optimization problem outlined in Proposition 2.2 separately in each month $t' \in \{t-60, \dots, t-1\}$. This generates a 60-month time series of $\hat{v}_{i,t'}^2$ values prior to each forecast month t . We will then denote the one-month-ahead forecasted value by $\bar{v}_{i,t}^2$ to distinguish it from an in-sample estimate. And, we begin making forecasts for each predictor in the month they were first published.

To make each forecast, we fit an AR(3) model to these 60 observations prior to month t :

$$\hat{v}_{i,t'}^2 = \check{a}_i + \sum_{\ell=1}^3 \check{b}_{i,\ell} \cdot \hat{v}_{i,t'-\ell}^2 + \check{e}_{i,t'} \quad \text{for months } t' = (t-60), \dots, (t-1).$$

Above, \check{a}_i and $\{\check{b}_{i,\ell}\}_{\ell=1,2,3}$ denote the coefficients associated with the i th predictor, and $\check{e}_{i,t'}$ represents the regression residual in month t' . Note that these coefficients will be different for each forecast date t ; we have just suppressed the t subscripts for clarity. We then compute one-month-ahead forecasts for the prior variance by applying these estimated coefficients to the final three months of data prior to month t :

$$\bar{v}_{i,t}^2 \stackrel{\text{def}}{=} \mathbf{E}_{t-1}[\hat{v}_{i,t}^2] = \check{a}_i + \sum_{\ell=1}^3 \check{b}_{i,\ell} \cdot \hat{v}_{i,t-\ell}^2.$$

Table 3 provides summary statistics describing these forecasted values $\bar{v}_{i,t}$. The sparkline plots represent the time series for each predictor $i = 1, \dots, 83$ on a common scale, $0\% < \bar{v}_{i,t} < 6\%$. Every time series ends in May 2015, so shorter sparkline plots correspond to predictors that were discovered later in our sample period. The sparkline plots provide evidence that predictor strengths were drawn from a common distribution. Predictors in our sample are constructed in very different ways using entirely different datasets—e.g., compare ‘*Sales/Cash*’ and ‘*Ret, 12-1*’. Nevertheless, notice that these two predictors generate similar point estimates for $\bar{v}_{i,t}$ each month. And, this is true even though, when you look at Table 2, the $\hat{\beta}_{i,t}$ time-series for each of these predictors is very different.

Combining Results. If you only had past experience with one predictor, $i = 1$, then you could only use the forecasted value of $\bar{v}_{1,t}^2$ to inform your prior beliefs about the likelihood of encountering a tradable anomaly. But, if you have seen $I_t > 1$ different cross-sectional predictors in past research seminars, then you can combine these various signals about σ^2 to


	#Obs	Avg	Sd	1%	25%	50%	75%	99%
	506	1.63	0.95	0.74	1.12	1.36	2.05	6.05

Table 4. Forecasted \bar{v}_t . Summary statistics describing the sample average of the rolling one-month-ahead forecasts for $\bar{v}_{i,t}$. We first estimate the in-sample parameter $\hat{v}_{i,t}$ separately for each previously discovered predictor each month as described in Proposition 2.2. We then make our one-month-ahead forecasts by fitting an AR(3) model to squared values in months $\{t - 60, \dots, t - 1\}$. Finally, we combine these forecasts for all previously discovered predictors to compute \bar{v}_t as in Equation (14). Units: % per month. Sample Period: May 1973 to May 2015. Sparkline Plot: aggregate time series during sample period. It corresponds to Figure 1 in miniature.

generate a more precise forecast:

$$\bar{v}_t^2 \stackrel{\text{def}}{=} \frac{1}{I_t} \cdot \sum_{i \in \mathcal{I}_t} \bar{v}_{i,t}^2. \quad (14)$$

This combined forecast, \bar{v}_t , represents your best guess about which prior variance of predictor strengths to use when evaluating predictors in month t —i.e., your best guess about the true value σ^2 to use in Equation (8)—based on your past experience with I_t other predictors. So, as the number of previously discovered predictors grows, we will be getting more and more signals about the anomaly baserate that we should be using going forward. By analogy, a seasoned researcher who has seen many candidate predictors come and go will have more signals with which to inform his beliefs than a junior researcher with less experience.

Table 4 provides summary statistics describing this combined forecast each month, \bar{v}_t . And, Figure 1 plots the \bar{v}_t time-series. The yearly average forecasted value for the prior volatility peaked in the year 2000 at 5.20% per month, and it reached its low point in 1990 at 0.77% per month. While Figure 1 reveals a large spike in \bar{v}_t at the time of the DotCom crash, this variable is not just a crash indicator. For example, note that there is no spike in predictor volatility around the 1987 crash. We estimate that $\bar{v}_{\text{Oct87}} = 1.17\%$ per month, which is much lower than the average value over the entire sample period, 1.63% per month.

Macroeconomic Correlations. Of course, there are numerous macroeconomic variables that forecast returns. So, to make sure that we are not just repackaging and rebranding one of these existing variables, we run a suite of regressions. Specifically, we regress our forecast for \bar{v}_t in month t on lagged values \bar{v}_{t-1} , \bar{v}_{t-2} , and \bar{v}_{t-3} as well as values of other macroeconomic variables: the level of the VIX index, realized volatility, a smoothed time-series of log GDP growth, the term spread, the NBER recession indicator, and the value-weighted market return. Table 5 provides summary statistics for each of these macroeconomic variables. Note that we do not have data on the VIX or the term spread for our entire sample period.

	VIX_t	$RVol_t$	$dGDP_t$	$tSpd_t$	$Rcsn_t$	$R_{Mkt,t}$
	(1)	(2)	(3)	(4)	(5)	(6)
Avg	19.89	14.34	1.53	1.81	0.14	0.95
Sd	7.59	8.40	0.94	1.15	0.35	4.59
#Obs	306	506	506	506	402	506

Table 5. Macroeconomic Variables. Summary statistics for macroeconomic forecasting variables. Sample Period: May 1973 to May 2015. VIX_t : level of the VIX index. $RVol_t$: realized volatility on value-weighted market index. $dGDP_t$: smoothed value of log GDP growth. $tSpd_t$: the term spread. $Rcsn_t$: NBER indicator for whether a recession is taking place. $R_{Mkt,t}$: return on the CRSP value-weighted market index.

We then estimate the relationship between each of these forecasting variables and our estimate of \bar{v}_t using regressions of the form

$$\bar{v}_t = \hat{a} + \sum_{\ell=1}^3 \hat{b}_\ell \cdot \bar{v}_{t-\ell} + \hat{\mathbf{c}}^\top \mathbf{Z}_t + \hat{e}_t \quad (15)$$

where \mathbf{Z}_t denotes a vector of macroeconomic variables in month t . Table 6 reports the results of these regression specifications. Each column reports the results of a separate regression. The table reveals that, while our forecasts for the prior volatility of predictor size are sometimes related to these well-known macroeconomic variables, they are certainly not subsumed by them. For example, the three lags of \bar{v}_t explain 55% of the variation whereas the VIX explains only 2%. And, once other macroeconomic variables are included, the VIX is no longer significant. In other words, \bar{v}_t is not just a proxy for the VIX. We are going to be looking at volatility-managed portfolios. But, instead of looking at the volatility of returns like in Moreira and Muir (2017), we are looking at the volatility of your priors.

4 Trading Strategy

In the last part of the analysis, we show how to use the forecasted value for σ^2 created using each predictor, $\bar{v}_{i,t}^2$, to combine different predictors into a single trading strategy. We start with a benchmark strategy that does not consider the prevailing anomaly baserate. Then, we show how using the anomaly baserate boosts this strategy's performance.

4.1 Portfolio Construction

Here is how we construct both the benchmark and the baserate-adjusted trading strategies that we are interested in.

Dependent Variable: \bar{v}_t											
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
<i>Const</i>	0.41*** (0.06)	1.14*** (0.18)	1.32*** (0.08)	1.64*** (0.04)	1.46*** (0.08)	1.50*** (0.10)	1.60*** (0.05)	0.83*** (0.19)	0.66** (0.32)	0.29** (0.14)	0.36 (0.23)
\bar{v}_{t-1}	0.57*** (0.04)									0.56*** (0.05)	0.55*** (0.06)
\bar{v}_{t-2}	0.35*** (0.05)									0.34*** (0.05)	0.33*** (0.06)
\bar{v}_{t-3}	-0.18*** (0.04)									-0.19*** (0.05)	-0.19*** (0.06)
<i>VIX</i> _{<i>t</i>}		0.02** (0.01)							-0.03 (0.02)		-0.01 (0.01)
<i>RVol</i> _{<i>t</i>}			2.14*** (0.49)					3.27*** (0.66)	6.53*** (1.42)	1.34*** (0.47)	2.48** (1.05)
<i>R</i> _{Mkt,<i>t</i>}				-0.67 (0.92)				1.56 (1.20)	1.44 (1.58)	0.72 (0.84)	0.77 (1.14)
<i>dGDP</i> _{<i>t</i>}					11.43** (4.47)			15.53* (8.45)	40.06*** (12.97)	-1.07 (5.97)	0.80 (9.60)
<i>tSpd</i> _{<i>t</i>}						0.01 (0.04)		-0.00 (0.04)	0.01 (0.05)	-0.01 (0.03)	-0.02 (0.04)
<i>Rcsn</i> _{<i>t</i>}							0.18 (0.12)	-0.01 (0.19)	0.08 (0.25)	-0.09 (0.13)	-0.12 (0.18)
<i>Adj R</i> ²	0.55	0.02	0.03	-0.00	0.01	-0.00	0.00	0.05	0.09	0.54	0.53
#Obs	503	306	506	506	506	402	506	402	306	402	306

Table 6. Macroeconomic Correlations. Relationship between the forecasted \bar{v}_t in month t and six macroeconomic variables. Sample Period: May 1973 to May 2015. *VIX*_{*t*}: level of the VIX index. *RVol*_{*t*}: realized volatility on value-weighted market index. *dGDP*_{*t*}: smoothed value of log GDP growth. *tSpd*_{*t*}: the term spread. *Rcsn*_{*t*}: NBER indicator for whether a recession is taking place. *R*_{Mkt,*t*}: return on the CRSP value-weighted market index. Each column represents the results of a separate regression of the form described in Equation (15). Numbers in parentheses are standard errors. Significance: * = 10%, ** = 5%, and *** = 1%.

Predictor-Specific Returns. Both strategies are going to hold positions in predictor-specific zero-cost portfolios that mirror cross-sectional regression coefficients. The realized return for the portfolio associated with the i th predictor in month t is just the estimate for $\hat{\beta}_{i,t}$:

$$R_{i,t} \stackrel{\text{def}}{=} \hat{\beta}_{i,t} = \frac{1}{N_t} \cdot \sum_n (R_{n,t} - \hat{\mu}_t) \cdot X_{n,i,t-1}. \quad (16)$$

In other words, $R_{i,t}$ is the realized return to a zero-cost portfolio that is long stocks which had high $X_{n,i,t-1}$ values and short stocks which had low $X_{n,i,t-1}$ values. In Equation (16), $\hat{\mu}_t$ denotes the mean excess return of all stocks in month t .

Of course, as emphasized in the introduction, it is next month's realized returns that matter for any trading strategy. So, to make a one-month-ahead forecast of $\hat{\beta}_{i,t}$, we first fit an AR(3) model to the previous five years of monthly data, $t' \in \{(t-60), \dots, (t-1)\}$:

$$\hat{\beta}_{i,t'} = \check{a}_i + \sum_{\ell=1}^3 \check{b}_{i,\ell} \cdot \hat{\beta}_{i,t'-\ell} + \check{e}_{i,t'} \quad \text{for months } t' = (t-60), \dots, (t-1). \quad (17)$$

Above, \check{a}_i and $\{\check{b}_{i,\ell}\}_{\ell=1,2,3}$ denote estimated coefficients, and $\check{e}_{i,t'}$ represents the regression residual in month t' . Note that these coefficients will be different for each forecast date t ; we have just suppressed the t subscripts for clarity. We then compute one-month-ahead forecasts by applying these estimated coefficients to the final three months of data prior to month t :

$$\bar{\beta}_{i,t} \stackrel{\text{def}}{=} \mathbf{E}_{t-1}[\hat{\beta}_{i,t}] = \check{a}_i + \sum_{\ell=1}^3 \check{b}_{i,\ell} \cdot \hat{\beta}_{i,t-\ell}.$$

If the resulting forecast is very different from zero, $|\bar{\beta}_{i,t}| \gg 0$, then we say that the i th predictor is a strong signal. By contrast, if $|\bar{\beta}_{i,t}| \approx 0$, then we say it is a weak signal.

Benchmark Strategy. Our benchmark strategy only uses these one-month-ahead forecasts to decide whether or not to invest in the i th predictor in month t . It explicitly does not take into consideration any information about the anomaly baserate. It only looks at the data; it does not consider the ex-ante probability that the i th predictor is a tradable anomaly. Let \mathcal{A}_t denote the set of 'active' predictors for the benchmark strategy in month t :

$$\mathcal{A}_t \stackrel{\text{def}}{=} \left\{ i \in \mathcal{I}_t : |\bar{\beta}_{i,t}| > \text{threshold} \right\}.$$

This is the collection of previously discovered anomalies whose past performance exceeds some minimum threshold. In the analysis below, we are going to set the minimum performance threshold to 1% per month. And, the dotted line in Figure 3 reports the number of active predictors each month when using this 1%-per-month threshold level.

To account for the fact that some predictors are contrarian—e.g., the predictor 'Ret,

1-0', which represents short-run reversals, delivers negative returns on average—we define an indicator variable $direction_i \in \{-1, +1\}$ that flips the sign of the benchmark portfolio's holdings for all contrarian strategies. So, since there are return reversals at the one-month horizon and momentum at the 12-month horizon, we have that $direction_{Ret, 1-0} = -1$ while $direction_{Ret, 12-1} = +1$. Note that this direction indicator is defined only once at the time the predictor is discovered.

The benchmark strategy holds an equal-weighted position in all active predictors, $i \in \mathcal{A}_t$, in month t . Thus, its raw returns are given by:

$$R_{\mathcal{A}_t, t} \stackrel{\text{def}}{=} \frac{1}{|\mathcal{A}_t|} \cdot \sum_{i \in \mathcal{A}_t} R_{i, t} \cdot direction_i. \quad (18)$$

If you want these returns to reflect the minimum performance threshold, you can also compute an analogous 'net' return:

$$\frac{1}{|\mathcal{A}_t|} \cdot \sum_{i \in \mathcal{A}_t} (R_{i, t} \cdot direction_i - threshold).$$

We will study both kinds of returns in various contexts in the analysis below. If a strategy has positive raw returns but negative net returns, then it is not something that is tradable. The strategy is generating phantom returns that will likely disappear once you start trading.

Baserate Adjustment. Now, suppose that you wanted to adjust the holdings of this benchmark strategy to account for the prevailing anomaly baserate. Imagine that you forecast a large positive realized return in month t for the i th predictor, $\bar{\beta}_{i, t} \gg 0$. If the forecasted value of $\bar{v}_{i', t}^2$ was small in month t for all other predictors $i' \neq i$ and the realized returns for all predictors $i \in \mathcal{I}_t$ were drawn from a common distribution, then the i th predictor's true magnitude is likely small as well, $\beta_{i, t}^* \approx 0$. And, even though your forecasted value represents a strong signal, you should still be reluctant to trade on it.

But, exactly how reluctant? How much should we revise next month's forecast? We can return to the statistical framework outlined in Section 2 to answer this question. Suppose that your forecast for the i th predictor's realized returns next month is a noisy signal about the true realized returns next month, $\bar{\beta}_{i, t} \sim \text{Normal}[\beta_{i, t}^*, N \cdot se_i^2]$, where $se_i^2 > 0$ represents the typical size of your forecast error for the i th predictor. Then, if your prior beliefs are that realized returns for each predictor are drawn from a Normal distribution each month with prior variance $\sigma^2 > 0$, $\beta_{i, t}^* \sim \text{Normal}[0, \sigma^2]$, then your best guess about the realized returns for the i th predictor in month t after calculating your return forecast would be:

$$E[R_{i, t} | \bar{\beta}_{i, t}, se_i^2, \sigma^2] = (1 + se_i^2 / \sigma^2)^{-1} \times \bar{\beta}_{i, t}.$$

This formula is just an application Bayesian-normal learning. And, you could use it to revise our forecast for next month's realized returns if you also had forecasts for se_i^2 and σ^2 .

Forecasting se_i^2 in month t is easy enough. You can fit an AR(3) model to the squared residuals from the $\hat{\beta}_{i,t}$ -forecasting regression in Equation (17) in each 60-month window:

$$\check{e}_{i,t'}^2 = \hat{\eta}_i + \sum_{\ell=1}^3 \hat{\theta}_{i,\ell} \cdot \check{e}_{i,t'-\ell}^2 + \hat{\omega}_{i,t'} \quad \text{for months } t' = (t - 60), \dots, (t - 1).$$

Above, $\hat{\eta}_i$ and $\{\hat{\theta}_{i,\ell}\}_{\ell=1,2,3}$ denote coefficients associated with the i th predictor for use when forecasting month t , and $\hat{\omega}_{i,t'}$ represents the regression residual in month t' . Note that these coefficients will be different for each forecast date t ; we have just suppressed the t subscripts for clarity. You can then forecast the standard error in the following month by applying these estimated coefficients to the final three months of data prior to month t :

$$\bar{se}_{i,t}^2 \stackrel{\text{def}}{=} E_{t-1}[\check{e}_{i,t}^2] = \hat{\eta}_i + \sum_{\ell=1}^3 \hat{\theta}_{i,\ell} \cdot \check{e}_{i,t-\ell}^2.$$

Computing a forecast of prior variance for use when evaluating the i th predictor in month t requires a bit more subtlety. You do not just want to use the average value of $\bar{v}_{i,t}^2$ for all previously discovered predictors, $i \in \mathcal{I}_t$, because the i th predictor itself is a previously discovered predictor. You cannot use data about the i th predictor to inform the prior you use when evaluating this data. So, when adjusting the forecast of the i th predictor, we compute the average prior-variance forecast for all other previously discovered predictors $i' \in \mathcal{I}_t \setminus i$:

$$\bar{v}_{-i,t}^2 \stackrel{\text{def}}{=} \frac{1}{I_t - 1} \cdot \sum_{i' \in \mathcal{I}_t \setminus i} \bar{v}_{i',t}^2.$$

This means that you are going to be using a slightly different prior volatility when adjusting the forecast of each predictor:

$$\bar{\beta}_{i,t}^\sigma = \left(1 + \bar{se}_{i,t}^2 / \bar{v}_{-i,t}^2\right)^{-1} \times \bar{\beta}_{i,t}.$$

It also means that you cannot calculate the baserate adjustment unless you have past experience with at least two other predictors. In other words, you need $I_t \geq 3$. For this reason, even though our data sample starts in May 1973, all of the trading-strategy results in this section will compare the benchmark strategy to a baserate-adjusted strategy starting in June 1979, the publication date of the third predictor in Table 1a.

Economic Interpretation. We have constructed the benchmark strategy defined in Equation (18) so that it does not require any information about the estimated values of $\bar{v}_{i,t}^2$. But, that does not mean this strategy is agnostic about what the anomaly baserate actually is. In

Dependent Variable: $R_{i,t}$						
	1979-2015		1979-1997		1998-2015	
	(1)	(2)	(3)	(4)	(5)	(6)
$Const$	0.55*** (0.06)	0.60*** (0.06)	0.41*** (0.11)	0.46*** (0.11)	0.56*** (0.07)	0.60*** (0.07)
$\bar{\beta}_{i,t}$	0.23*** (0.02)		0.42*** (0.04)		0.21*** (0.02)	
$\bar{\beta}_{i,t}^\sigma$		0.30*** (0.03)		0.65*** (0.07)		0.27*** (0.03)
$Adj. R^2$	0.01	0.01	0.03	0.03	0.01	0.01
$\#Obs$	15,330		3,103		12,227	

Table 7. Predictive Power. Panel regressions showing that you can boost your predictive power by adjusting your return forecasts for the prevailing anomaly baserate. Columns (1) and (2) use a monthly panel data set that starts in June 1979, ends in May 2015, and includes one observation each month for each previously discovered predictor, $i \in \mathcal{I}_t$. So, there were 4 observations in June 1979 and 83 observations in May 2015. Columns (3)-(6) study sub-samples of the same data set. Columns (1), (3), and (5) report the estimated coefficients from regressing each predictor’s realized returns on the raw return forecast, $R_{i,t} = \hat{a} + \hat{b} \cdot \bar{\beta}_{i,t} + \hat{e}_{i,t}$. Columns (2), (4), and (6) report analogous results using the baserate-adjusted forecast, $R_{i,t} = \hat{a} + \hat{b} \cdot \bar{\beta}_{i,t}^\sigma + \hat{e}_{i,t}$. $Const$ has units of % per month, and the slope coefficients are dimensionless. Numbers in parentheses are standard errors. Significance: * = 10%, ** = 5%, and *** = 1%.

fact, the benchmark strategy makes a strong implicit assumption about the anomaly baserate. By using the benchmark strategy, you are implicitly saying that you think extremely strong cross-sectional predictors are very common:

$$\lim_{\bar{v}_{-i,t}^2 \rightarrow \infty} \bar{\beta}_{i,t}^\sigma = \bar{\beta}_{i,t}.$$

By contrast, if your past experience tells you that there are no cross-sectional predictors, then you will discard all evidence to the contrary:

$$\lim_{\bar{v}_{-i,t}^2 \rightarrow 0} \bar{\beta}_{i,t}^\sigma = 0.$$

This is the sense in which believing that $\sigma^2 = 0$ is equivalent to having dogmatic priors. You have to take a stand on your priors in order to draw an inference from a test result. You cannot punt on this issue. And, ignoring the problem is a really strong stance.

Predictive Power. What is more, before looking at any trading-strategy returns, we can run

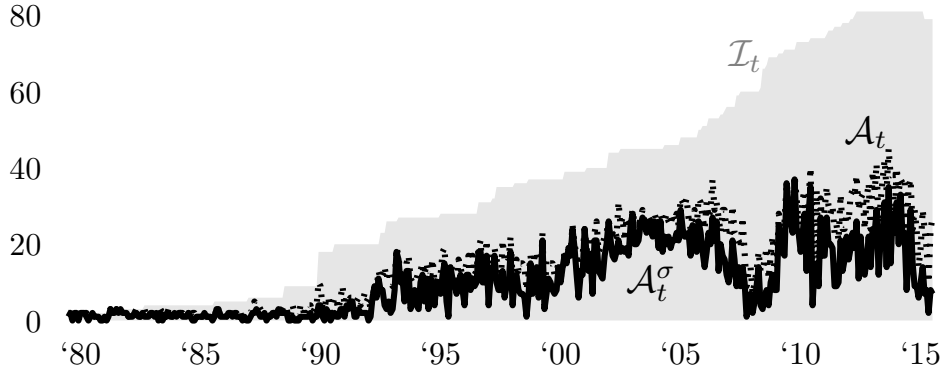


Figure 3. Number of Predictors. Shaded Region: number of previously discovered predictors, \mathcal{I}_t . Dotted Line: number of predictors used in the benchmark strategy in month t , \mathcal{A}_t , when the minimum-performance threshold is set to 1% per month. Solid Line: number of predictors used in the baserate-adjusted strategy in month t , \mathcal{A}_t^σ , when using the same threshold. Sample Period: June 1979 to May 2015.

a simple panel regression to show that it helps to adjust your predictor-specific forecasts for the prevailing anomaly baserate. Consider a panel data set with predictor \times month observations. The data set starts in June 1979 and ends in May 2015. And, it includes one observation each month for each previously discovered predictor, $i \in \mathcal{I}_t$. So, there were 4 observations in June 1979 and 83 observations in May 2015. We first use this data to regress the realized returns for each predictor i in month t on the benchmark forecasts, $\bar{\beta}_{i,t}$:

$$R_{i,t} = \hat{a} + \hat{b} \cdot \bar{\beta}_{i,t} + \hat{e}_{i,t}.$$

Then, we run the exact same regression, only this time using the baserate-adjusted forecasts, $\bar{\beta}_{i,t}^\sigma$, as the right-hand-side variable:

$$R_{i,t} = \hat{a} + \hat{b} \cdot \bar{\beta}_{i,t}^\sigma + \hat{e}_{i,t}.$$

Table 7 reports the results from these two regressions. The first two columns show that adjusting your one-month-ahead return forecasts for the anomaly baserate boosts your predictive power from $\hat{b} = 0.23$ to $\hat{b} = 0.30$. In other words, shrinking your return forecasts towards zero improves your performance. In the remaining four columns, we split our sample period in half and show that this result holds in both sample periods. The amount of predictability is higher on average for each predictor earlier in our sample period. And, this is exactly what you would expect if many of the recently discovered predictors that constitute the anomaly zoo were spurious. But, adjusting your forecasts for the prevailing anomaly baserate always results in a statistically significant increase in predictive power.

	\mathcal{A}_t	\mathcal{A}_t^σ		\mathcal{A}_t	\mathcal{A}_t^σ		\mathcal{A}_t	\mathcal{A}_t^σ			
<i>BetaSq</i>			<i>Ret, 6-1</i>		0.30	0.19	<i>NetExtnlFin/Assets</i>		0.00	0.00	
<i>Earn/Share</i>			<i>%ΔSales</i>		0.82	0.51	<i>Beta</i>		0.06	0.00	
<i>Debt/Price</i>		0.40	0.10	<i>OpAccr</i>		0.61	0.39	<i>NetPO/Price</i>		0.30	0.00
<i>Divd/Price</i>		0.63	0.11	<i>CapitalTOver</i>		0.22	0.08	<i>PO/Price</i>		0.36	0.01
<i>Mcap</i>		0.83	0.59	<i>RetOnEquity</i>		0.94	0.76	<i>NetPO</i>		0.00	0.00
<i>Earn/Price</i>		0.16	0.14	<i>KaplanZingales</i>		0.03	0.01	<i>RetOnInvstCap</i>		0.72	0.00
<i>Ret, 36-12</i>		0.75	0.69	$\% \Delta[\Delta Sales/\Delta Invtry]$		0.11	0.03	$\% \Delta Shares$		0.00	0.00
<i>AvgSpread</i>		0.07	0.00	$\% \Delta[\Delta Sales/\Delta Recv]$		0.21	0.17	<i>ProfMgn</i>		0.92	0.00
<i>Assets/Mcap</i>		0.60	0.32	$\% \Delta[\Delta Sales/\Delta XG\&A]$		0.00	0.00	<i>AdjProfMgn</i>		0.01	0.00
<i>Levrg</i>		0.01	0.00	$\% \Delta[\Delta GrMgn/\Delta Sales]$		0.00	0.00	<i>RetOnOpAssets</i>		0.17	0.00
<i>Levrg/Price</i>		0.27	0.06	<i>LagTOver</i>		0.20	0.00	<i>AssetTOver</i>		0.84	0.17
<i>Sales/Cash</i>		0.08	0.05	$Adj[BkVal/Mcap]$		0.92	0.91	<i>AdjAssets</i>		0.00	0.00
<i>LtCF</i>		0.01	0.00	<i>AdjMcap</i>		0.87	0.84	$\% \Delta InvmtX$		0.15	0.00
<i>CurrRatio</i>		0.08	0.02	<i>SdTOver</i>		0.29	0.09	$\% \Delta Invmt$		0.87	0.58
$\% \Delta CurrRatio$		0.36	0.32	<i>AdvertRate/Ret</i>		0.89	0.81	$\Delta AdjShares$		0.00	0.00
$\% \Delta QuickRatio$		0.00	0.00	$R\&D/Mcap$		0.91	0.42	<i>RetOnCash</i>		0.00	0.00
$\% \Delta [Sales/Invtry]$		0.05	0.00	$R\&D/Sales$		0.28	0.06	<i>Tangibility</i>		0.00	0.00
<i>QuickRatio</i>		0.09	0.03	<i>Advert/Mcap</i>		0.98	0.95	$\Delta AdjTOver$		0.00	0.00
<i>Sales/Invtry</i>		0.02	0.00	$\Delta Invtry/Assets$		0.44	0.39	<i>UnexplVlm</i>		0.98	0.86
<i>Sales/Recv</i>		0.29	0.16	<i>OpCF/Price</i>		0.08	0.00	<i>RetOnAssets</i>		1.00	0.47
<i>Ret, 1-0</i>		0.89	0.60	$Invmt/Lag[AvgInvmt]$		0.17	0.03	<i>OpLevrg</i>		1.00	0.61
<i>Ret, 12-1</i>		0.17	0.03	$NetOpAssets/Sales$		0.17	0.03	<i>MaxRet</i>		0.83	0.00
<i>BkVal</i>		0.81	0.48	$\% \Delta BkVal$		0.87	0.73	<i>FreeCF</i>		0.86	0.00
<i>MonthlyMcap</i>		1.00	0.34	$\% \Delta LtDebt$		0.31	0.23	$R\&Dcapital$		0.96	0.00
<i>Sales/Price</i>		0.93	0.34	<i>Price-52WkHi</i>		0.87	0.00	$\% \Delta Invtry$		0.76	0.55
$\% \Delta [Deprc/PP\&E]$		0.00	0.00	<i>IdioVol</i>		0.87	0.01	<i>Ret, 12-6</i>		0.68	0.00
$D\&A/Assets$		0.78	0.67	<i>TotVol</i>		0.86	0.00	<i>CashHldgs</i>		0.08	0.00
<i>Deprc/PP&E</i>		0.52	0.09	$\varepsilon\% \Delta Mcap$		0.53	0.21				

Table 8. Predictor Selection. \mathcal{A}_t : fraction of all post-discovery months that predictor was held by benchmark strategy. \mathcal{A}_t^σ : same fraction for baserate-adjusted strategy. Sparkline plots: x -axis is time in months. All time series end in May 2015. Predictors discovered later have shorter sparkline plots. If predictor held by benchmark strategy, then there is a bar on the bottom half of the plot; if it is held by the baserate-adjusted strategy, then there is a bar on the top half. Sample period: June 1979 to May 2015.

Baserate-Adjusted Strategy. We modify the benchmark trading strategy to account for changes in the anomaly baserate by choosing which predictors to invest in based on $\bar{\beta}_{i,t}^\sigma$ rather than $\bar{\beta}_{i,t}$. Let \mathcal{A}_t^σ denote the set of active predictors for the baserate-adjusted strategy in month t :

$$\mathcal{A}_t^\sigma \stackrel{\text{def}}{=} \left\{ i \in \mathcal{I}_t : |\bar{\beta}_{i,t}^\sigma| > \text{threshold} \right\}.$$

The solid line in Figure 3 reports the number of active predictors each month for the baserate-adjusted strategy when the threshold is set to 1% per month. Because the baserate-adjusted strategy is using the same return forecasts as the benchmark strategy only revised toward zero, the black line will always lie below the dotted line. Let $R_{\mathcal{A}_t^\sigma}$ denote the realized return of the benchmark-adjusted strategy at time t :

$$R_{\mathcal{A}_t^\sigma} \stackrel{\text{def}}{=} \frac{1}{|\mathcal{A}_t^\sigma|} \cdot \sum_{i \in \mathcal{A}_t^\sigma} R_{i,t} \cdot \text{direction}_i.$$

Note that, because they are defined in analogous ways, if the baserate-adjusted strategy outperforms the benchmark-adjusted strategy, then it must be because the predictors that make up the gap between the dotted and solid lines in Figure 3 had poor returns.

Predictor Selection. Table 8 takes a more detailed look at which predictors each strategy is holding and when. The column labeled \mathcal{A}_t reports the fraction of all months following discovery that a predictor was held in the benchmark strategy; whereas, the column labeled \mathcal{A}_t^σ reports the same fraction but for the baserate-adjusted strategy. And, because the baserate-adjusted strategy is based on return forecasts that have been revised toward zero, the fraction in the \mathcal{A}_t^σ column will always be smaller.

The sparkline plots then report precisely which months each predictor was held by each strategy. The x -axis in each figure represents time in months. Every time series ends in May 2015. So, predictors discovered later have shorter sparkline plots. Months in which a predictor was held by the benchmark strategy are denoted by vertical bars on the bottom half of each sparkline plot; whereas, months in which a predictor was held by the baserate-adjusted strategy are denoted by vertical bars on the top half of each sparkline plot. Thus, a full bar represents a month in which the predictor was held by both strategies, a half bar represents a month in which the predictor was only held by the benchmark strategy, and an empty space represents a month in which neither strategy invested in the predictor.

4.2 Realized Returns

We find that the baserate-adjusted trading strategy outperforms the benchmark trading strategy in a way that is statistically significant, economically large, and robust to controlling for risk-factor exposures. And, we now explore this fact in several different ways.

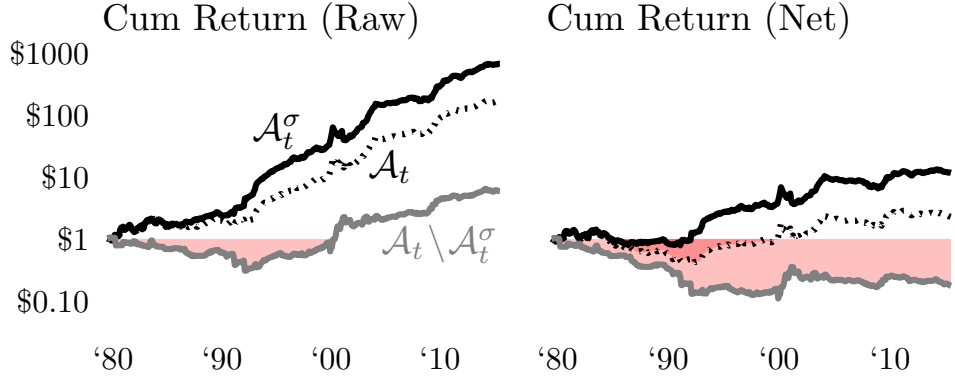


Figure 4. Cumulative Returns. Solid black Line, \mathcal{A}_t^σ : current value of portfolio that started with \$1 in June 1979 and continually reinvested its holdings in the baserate-adjusted strategy. Dotted black Line, \mathcal{A}_t : current value of same initial \$1 portfolio when investing in the benchmark strategy. Solid gray Line, $\mathcal{A}_t \setminus \mathcal{A}_t^\sigma$: current value of same initial \$1 portfolio when investing in only those predictors that are held by the benchmark strategy but not the baserate-adjusted strategy. Left panel: cumulative raw returns. Right panel: cumulative net returns after subtracting off the 1% per month performance threshold. Red shading: regions where your portfolio would have been worth less than \$1 when following a particular strategy. Sample Period: June 1979 to May 2015.

Cumulative Returns. To quantify the performance boost that comes from incorporating information about the anomaly baserate, we start by computing the cumulative returns to investing \$1 in both the benchmark strategy and the baserate-adjusted strategy starting in June 1979. Figure 4 reports the total amount of money that you would have in your account in month t if you followed either strategy, continually reinvesting any capital gains along the way. First, look at the left panel, which shows results using raw returns. By May 2015 the solid black line, which represents the wealth generated by the baserate-adjusted strategy is more than four times as high as the dotted black line, which represents the wealth generated by the benchmark strategy. When looking at raw returns, following the baserate-adjusted strategy turns \$1 in June 1979 into \$671.03 in May 2015; whereas, following the benchmark strategy leaves you with only \$161.23 in May 2015. The right panel shows analogous results using net returns based on the 1%-per-month performance threshold. After accounting for this implementation cost, the benchmark strategy actually loses money for the first half of the sample period. And, if we look at the predictors held by the benchmark strategy but not the baserate-adjusted strategy—i.e., the solid gray line—we see that a strategy composed of only these predictors would lose money throughout the sample period on net.

Performance Metrics. In Table 9, we explore the difference in performance between the benchmark and baserate-adjusted strategies in more detail. We report both the mean and

	Benchmark \mathcal{A}_t (1)	Adjusted \mathcal{A}_t^σ (2)	Difference $\mathcal{A}_t \setminus \mathcal{A}_t^\sigma$ (3)
$E[R_{s,t}^{\text{Raw}}]$	1.28	1.62	0.56
$E[R_{s,t}]$	0.29	0.68	-0.25
$\text{Sd}[R_{s,t}]$	4.50	4.75	5.36
$\text{Skew}[R_{s,t}]$	0.61	0.96	-0.16
$\text{Kurt}[R_{s,t}]$	4.14	7.83	12.56
SR_s	0.22	0.50	-0.16
$\text{Max}[DD_s]$	73.81	46.73	89.43
$\text{Pr}[Inv_s]$	0.99	0.94	0.81

Table 9. Performance Metrics. Performance statistics for the excess returns to three different trading strategies. Column (1): benchmark strategy, \mathcal{A}_t . Column (2): baserate-adjusted strategy, \mathcal{A}_t^σ . Column (3): strategy that invests in predictors held by the benchmark strategy but not the baserate-adjusted strategy, $\mathcal{A}_t \setminus \mathcal{A}_t^\sigma$. All return statistics quoted in % per month. $E[R_{s,t}^{\text{Raw}}]$: mean raw monthly return. $E[R_{s,t}]$: mean net monthly return. $\text{Sd}[R_{s,t}]$: standard deviation of net monthly returns. $\text{Skew}[R_{s,t}]$: skewness of net monthly returns. $\text{Kurt}[R_{s,t}]$: kurtosis of net month returns. SR_s : annualized Sharpe ratio using net monthly returns. $\text{Max}[DD_s]$: maximum drawdown for strategy based on net returns units of in % decline from peak. $\text{Pr}[Inv_s]$: fraction of all 433 months during the sample period in which a strategy is invested in at least one predictor. Sample period: June 1979 to May 2015.

standard deviation of each strategy’s monthly returns as well as their skewness and kurtosis. We also report the annualized Sharpe ratio, the maximum cumulative drawdown in percent, and the percent of all months in which each strategy is invested in at least one predictor. The first column reveals that, after accounting for implementation costs, the annualized Sharpe ratio of the benchmark strategy is only 0.22 during our sample period. By contrast, when we look at the second column, we can see that the baserate-adjusted strategy has an annualized Sharpe ratio that is more than twice as high, 0.50. And, as you would expect, this difference is due to systematically dropping predictors from the portfolio that only seem to have strong signals. The third column shows that the net returns of a trading strategy that only invests in predictors that are held by the benchmark strategy but not by the benchmark-adjusted strategy are negative. The baserate-adjusted strategy has more positive skewness, slightly higher kurtosis, and a much lower maximum drawdown compared to the benchmark strategy.

Abnormal Returns. We next show that the performance of the baserate-adjusted strategy is not just the result of exposure to market risk. To show this, we run a time-series regression of the net excess returns to the baserate-adjusted strategy on the excess returns to the

(a) Summary Statistics

	Avg	Sd	SR
$R_{\text{Mkt},t}$	0.65	4.49	0.50

(b) Regression Results

	Benchmark \mathcal{A}_t		Adjusted \mathcal{A}_t^σ		Difference $\mathcal{A}_t \setminus \mathcal{A}_t^\sigma$	
	(1)	(2)	(3)	(4)	(5)	(6)
<i>Const</i>	0.29 (0.27)	0.27 (0.26)	0.68** (0.27)	0.67** (0.27)	-0.25 (0.27)	-0.23 (0.27)
$R_{\text{Mkt},t}$		0.04 (0.07)		0.02 (0.06)		-0.04 (0.07)
<i>Adj. R</i> ²		0.00		0.00		0.00

Table 10. Abnormal Returns. Net abnormal returns relative to the market model for three different trading strategies: benchmark strategy, \mathcal{A}_t ; baserate-adjusted strategy, \mathcal{A}_t^σ ; set-difference strategy that invests in predictors held by the benchmark strategy but not the baserate-adjusted strategy, $\mathcal{A}_t \setminus \mathcal{A}_t^\sigma$. $R_{\text{Mkt},t}$: excess return on the value-weighted market portfolio. **(a) Summary Statistics.** Mean and standard deviation of the excess return on the market in units of % per month as well as the annualized Sharpe ratio. **(b) Regression Results.** Each column reports the results of a separate time-series regression with the net excess returns of a particular strategy as the left-hand side variable. *Const* has units of % per month, and the slope coefficients are dimensionless. Numbers in parentheses are Newey-West standard errors. Statistical significance: * = 10%, ** = 5%, and *** = 1%. Sample period: June 1979 to May 2015. All regressions involve 433 monthly observations.

value-weighted market:

$$R_{\mathcal{A}_t^\sigma,t} = \hat{a} + \hat{b} \cdot R_{\text{Mkt},t} + \hat{e}_t.$$

In the equation above, $R_{\text{Mkt},t}$ is the excess return on the market in month t , \hat{a} is the abnormal return to the baserate-adjusted strategy, \hat{b} is the slope coefficient from this time-series regression, and \hat{e}_t is the regression residual. These data come from Kenneth French’s website.³ Column (4) in Table 10 shows that market-risk exposure does not account for the baserate-adjusted trading strategy’s good performance. The average net return of the baserate-adjusted strategy is 0.68% per month; and, after accounting for market-risk exposure, the net abnormal returns to this strategy are 0.67% per month. There is hardly any difference. The columns (2) and (6) in Table 10 replicate the same analysis using the net returns of the benchmark

³See http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.

Dependent Variable: $R_{i,Post}$						
	(1)	(2)	(3)	(4)	(5)	(6)
$Const$	-0.35 (4.27)	-1.73 (3.58)	4.06 (6.92)	5.35 (5.67)	4.32 (4.76)	1.97 (5.18)
$\mathbf{1}[i \in \mathcal{A}_{t_0(i)}^\sigma]$	15.07*** (4.52)	12.22*** (2.94)	15.87*** (4.69)	12.52** (4.81)		18.77*** (6.42)
$\bar{\beta}_{i,t_0(i)}$		1.02*** (0.29)		1.40** (0.57)		
$\bar{s}e_{i,t_0(i)}$			-0.89 (0.94)	-1.53* (0.84)		
$\mathbf{1}[i \in \mathcal{A}_{t_0(i)}^{E[\bar{v}_t]}]$					6.48 (4.95)	-6.37 (7.69)
$Adj. R^2$	0.18	0.18	0.16	0.18	0.00	0.17

Table 11. New Predictors. Evidence that the anomaly baserate is helpful when evaluating new cross-sectional predictors. Each column reports results of a separate cross-sectional regression with 46 observations, one for each predictor i held by the benchmark strategy in month following discovery, $t_0(i)$. $R_{i,Post}$: annualized return to predictor-specific strategy associated with i th predictor defined in Equation (16) during 10 years immediately following publication. $Const$: intercept term; units of % per year. $\mathbf{1}[i \in \mathcal{A}_{t_0(i)}^\sigma]$: indicator variable for whether i th predictor was held by baserate-adjusted strategy in month $t_0(i)$; units of % per year. $\bar{\beta}_{i,t_0(i)}$: forecasted return of the i th predictor in month $t_0(i)$. $\bar{s}e_{i,t_0(i)}$: forecasted standard error of the i th predictor in month $t_0(i)$. $\mathbf{1}[i \in \mathcal{A}_{t_0(i)}^{E[\bar{v}_t]}]$: indicator variable for whether i th predictor was held by time-invariant version of baserate-adjusted strategy in month $t_0(i)$; units of % per year. Numbers in parentheses are Newey-West standard errors using the optimal number of lags. Statistical significance: * = 10%, ** = 5%, and *** = 1%. Sample period: June 1979 to May 2015.

and set-difference strategies as the left-hand-side variable. While the net excess returns to each of these strategies is much lower than that of the baserate-adjusted strategy, neither strategy had excess returns that are very correlated with the market.

New Predictors. Finally, we motivated this paper by discussing the problem faced by a researcher who is trying to evaluate statistical evidence concerning a *new* cross-sectional predictor. However, the trading strategy defined in Equation (18) holds positions in both brand new and previously discovered predictors. e.g., it might invest based on a firm’s investment growth at any time on or after December 2004 when Titman, Wei, and Xie (2004) published a paper documenting this predictor. So, maybe the improved performance of the baserate-adjusted strategy relative to the benchmark strategy is only due to different positions in previously discovered predictors? In the last part of our analysis, we show this is not the case.

First, we compute the realized returns of each predictor-specific trading strategy held by

the benchmark strategy in the 10 years immediately after its publication. Let $R_{i,Post}$ denote the annualized return to the predictor-specific strategy associated with the i th predictor defined in Equation (16) during the 10 years immediately following its publication date as given in Tables 1a, 1b, and 1c. And, let $t_0(i)$ denote the month immediately following the publication of the i th predictor. e.g., for investment growth as defined in Titman, Wei, and Xie (2004), $t_0(i) = Jan2005$. There are 81 predictors for which we can forecast the prior variance in month $t_0(i)$. And, 46 of these 81 possible predictors have forecasted returns in excess of 1% for month $t_0(i)$ and are thus held by the benchmark strategy, $\sum_{i=3}^{83} \mathbf{1}[i \in \mathcal{A}_{t_0(i)}] = 46$.

Then, we regress the post-publication returns of each predictor held by the benchmark strategy on a variable indicating whether or not the predictor would also have been held by the baserate-adjusted strategy:

$$R_{i,Post} = \hat{a} + \hat{b} \cdot \mathbf{1}[i \in \mathcal{A}_{t_0(i)}^\sigma] + \hat{e}_i \quad \text{for predictors } i \in \mathcal{A}_{t_0(i)}. \quad (19)$$

Column (1) of Table 11 reports the results of this cross-sectional regression, which contains one observation for each of the 46 predictors held by the benchmark strategy immediately following their initial discovery. We estimate that $\hat{b} = 15.07\%$ per year, and this suggests that incorporating information about the prevailing anomaly baserate is useful when evaluating new predictors. The realized returns over the next 10 years of the 27 predictors which had baserate-adjusted return forecasts in excess of 1% per month, $\mathbf{1}[i \in \mathcal{A}_{t_0(i)}^\sigma] = 1$, were 15.07% per year higher than those of the remaining 19 predictors.

The remaining columns in Table 11 provide supporting evidence that this result is not mechanical. This evidence comes in two flavors. In Columns (2), (3), and (4), we start by showing that adjusting each predictor's $\bar{\beta}_i$ for the prevailing anomaly baserate is doing more than just selecting the strongest in-sample predictors. We document that the estimated \hat{b} remains positive and statistically significant even after we add the level and the standard error of each predictor's forecasted return as right-hand-side variables in Equation (19). The results in these columns also suggest that other popular factor-timing strategies such as the valuation spread (Baba Yara, Boons, and Tamoni, 2018) do not subsume the predictive power of the baserate. The columns that include \bar{e}_i as a right-hand-side variable also indicate that our baserate-adjusted strategy is different from a strategy that only trades predictors with sufficiently large t -stats. In Columns (5) and (6), we then directly investigate whether or not the results are driven by time-series variation in the prevailing anomaly baserate. We do this by replacing the indicator variable in Equation (19), which is defined using the prevailing prior volatility, with an analogous indicator variable, which is defined using the time-series average of the aggregate prior volatility reported in Table 4, $E[\bar{v}_t] = 1.63\%$ per month. This new

indicator variable, $\mathbf{1}[i \in \mathcal{A}_{t_0(i)}^{E[\bar{v}_t]}]$, shrinks each predictor’s return forecast by a constant amount that is correct on average. Columns (5) and (6) show that this correct-on-average formulation is nowhere nearly as helpful when it comes to evaluating new predictors. Accounting for time-series variation in the anomaly baserate is key to our results.

5 Conclusion

The academic literature contains hundreds of statistically significant cross-sectional predictors. The existence of this so-called “anomaly zoo” has caused many to question whether researchers are using the right tests for statistical significance. But, here is the thing: even if researchers are using the right tests, they will still be drawing the wrong conclusions from their econometric analysis if they are starting out with the wrong priors—i.e., if they are starting out with incorrect beliefs about the probability of discovering an anomaly. And yet, the academic literature offers no guidance about how to estimate this anomaly baserate. Motivated by this logical gap, we provide a way to estimate the anomaly baserate by running penalized regressions and searching for the best-fit tuning parameter.

Machine-learning techniques, such as the Ridge regression and LASSO, have become popular of late in the asset-pricing literature. And, in the past, researchers have used these tools to answer canonical asset-pricing questions in high-dimensional settings. This paper is different. We are not using machine-learning techniques as off-the-shelf statistical tools. Instead, we are using the Bayesian interpretation of the best-fit tuning parameter to shed light on an economic object, the anomaly baserate, that is of first-order importance regardless of the dimensionality of the problem. Thus, while we use the same machine-learning toolkit as earlier papers, we are using this toolkit to solve a qualitatively different kind of problem, a kind of problem that exists even in the absence of the anomaly zoo that popularized the machine-learning techniques that we use to solve it.

References

- Akaike, H. (1974). A new look at statistical model identification. *IEEE Transactions on Automatic Control*.
- Andersen, T., T. Bollerslev, F. Diebold, and P. Labys (2003). Modeling and forecasting realized volatility. *Econometrica*.
- Baba Yara, F., M. Boons, and A. Tamoni (2018). Value-return predictability across asset classes and commonalities in risk premia. *Working Paper*.
- Bajgrowicz, P. and O. Scaillet (2012). Technical trading revisited: False discoveries, persistence tests, and transaction costs. *Journal of Financial Economics*.
- Baker, S., N. Bloom, and S. Davis (2016). Measuring economic-policy uncertainty. *Quarterly Journal of Economics*.
- Barberis, N. and R. Thaler (2003). A survey of behavioral finance. *Handbook of Financial Economics*.
- Barras, L., O. Scaillet, and R. Wermers (2010). False discoveries in mutual-fund performance: Measuring luck in estimated alphas. *Journal of Finance*.
- Berger, J. (2006). The case for objective Bayesian analysis. *Bayesian Analysis*.
- Bryzgalova, S. (2017). Spurious factors in linear asset-pricing models. *Working Paper*.
- Chinco, A., A. Clark-Joseph, and M. Ye (2018). Sparse signals in the cross-section of returns. *Journal of Finance*.
- DeMiguel, V., L. Garlappi, F. Nogales, and R. Uppal (2009). A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms. *Management Science*.
- Diaconis, P. and B. Skyrms (2017). *Ten great ideas about chance*. Princeton University Press.
- Fama, E. (1976). *Foundations of finance: Portfolio decisions and securities prices*. Basic Books (AZ).
- Fama, E. and K. French (1996). Multi-factor explanations of asset-pricing anomalies. *Journal of Finance*.
- Feng, G., S. Giglio, and D. Xiu (2017). Taming the factor zoo. *Working Paper*.
- Ferson, W., S. Sarkissian, and T. Simin (1999). The alpha-factor asset-pricing model: A parable. *Journal of Financial Markets*.
- Freyberger, J., A. Neuhierl, and M. Weber (2017). Dissecting characteristics non-parametrically. *Working Paper*.

- Frost, P. and J. Savarino (1986). An empirical-Bayes approach to efficient portfolio selection. *Journal of Financial and Quantitative Analysis*.
- Green, J., J. Hand, and F. Zhang (2017). The characteristics that provide independent information about average US monthly stock returns. *Review of Financial Studies*.
- Gromb, D. and D. Vayanos (2010). Limits of arbitrage. *Annual Review of Financial Economics*.
- Harvey, C. (2017). Presidential address: The scientific outlook in financial economics. *Journal of Finance*.
- Harvey, C. and Y. Liu (2018a). Detecting repeatable performance. *Review of Financial Studies*.
- Harvey, C. and Y. Liu (2018b). False (and missed) discoveries in financial economics. *Working Paper*.
- Harvey, C. and Y. Liu (2018c). Lucky factors. *Working Paper*.
- Harvey, C., Y. Liu, and H. Zhu (2016). ... and the cross-section of expected returns. *Review of Financial Studies*.
- Hastie, T., R. Tibshirani, and J. Friedman (2001). *The elements of statistical learning: Data mining, inference, and prediction*. Springer Science & Business Media.
- Hoerl, A. and R. Kennard (1970). Ridge regression: Applications to non-orthogonal problems. *Technometrics*.
- Jegadeesh, N. and S. Titman (1993). Returns to buying winners and selling losers: Implications for stock-market efficiency. *Journal of Finance*.
- Jonah, B. (1986). Accident risk and risk-taking behaviour among young drivers. *Accident Analysis & Prevention*.
- Karolyi, A. (1993). A Bayesian approach to modeling stock-return volatility for option valuation. *Journal of Financial and Quantitative Analysis*.
- Kelly, B., S. Pruitt, and Y. Su (2017). Instrumented principal-component analysis. *Working Paper*.
- Kelly, B., S. Pruitt, and Y. Su (2018). Characteristics are covariances: A unified model of risk and return. *Journal of Financial Economics*.
- Kozak, S., S. Nagel, and S. Santosh (2018). Shrinking the cross-section. *Journal of Financial Economics*.
- Ledoit, O. and M. Wolf (2017). Non-linear shrinkage of the covariance matrix for portfolio selection: Markowitz meets Goldilocks. *Review of Financial Studies*.

- Lettau, M. and M. Pelger (2018). Estimating latent asset-pricing factors. *Working Paper*.
- Lewellen, J. (2015). The cross-section of expected stock returns. *Critical Finance Review*.
- Linnainmaa, J. and M. Roberts (2018). The history of the cross-section of stock returns. *Review of Financial Studies*.
- Lo, A. and C. MacKinlay (1990). Data-snooping biases in tests of financial asset-pricing models. *Review of Financial Studies*.
- McLean, D. and J. Pontiff (2016). Does academic research destroy stock-return predictability? *Journal of Finance*.
- Moreira, A. and T. Muir (2017). Volatility-managed portfolios. *Journal of Finance*.
- Novy-Marx, R. and M. Velikov (2015). A taxonomy of anomalies and their trading costs. *Review of Financial Studies*.
- Park, T. and G. Casella (2008). The Bayesian LASSO. *Journal of the American Statistical Association*.
- Pedersen, L. (2015). *Efficiently inefficient: How smart money invests and market prices are determined*. Princeton University Press.
- Robbins, H. (1956). An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical and Statistical Probability*.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*.
- Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society*.
- Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike’s criterion. *Journal of the Royal Statistical Society*.
- Sullivan, R., A. Timmermann, and H. White (1999). Data-snooping, technical trading-rule performance, and the bootstrap. *Journal of Finance*.
- Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society*.
- Titman, S., J. Wei, and F. Xie (2004). Capital investments and stock returns. *Journal of Financial and Quantitative Analysis*.
- White, H. (2000). A reality check for data snooping. *Econometrica*.
- Yan, X. and L. Zheng (2017). Fundamental analysis and the cross-section of stock returns: A data-mining approach. *Review of Financial Studies*.
- Zou, H., T. Hastie, and R. Tibshirani (2007). On the “degrees of freedom” of the LASSO. *Annals of Statistics*.

A Technical Appendix

Proof (Proposition 2.1). The result follows from properties of the normal distribution. If $z \sim \text{Normal}[0, \sigma^2]$, then for any $\omega > 0$ we have that:

$$\begin{aligned}\Pr[z > \omega] &= \Pr[z < -\omega] \\ &= \Phi[-\omega/\sigma].\end{aligned}$$

Thus, we have $\Pr[|\beta_{I+1}^*| > \text{threshold}] = 2 \cdot \Pr[\beta_{I+1}^* < -\text{threshold}] = 2 \cdot \Phi[-\text{threshold}/\sigma]$. \square

Derivation (Equation 10). Optimizing Equation (9) results in the first-order condition:

$$\begin{aligned}0 &= -2 \cdot \frac{1}{N} \cdot \sum_n (R_n - \hat{\mu} - \beta \cdot X_{n,i}) \cdot X_{n,i} + 2 \cdot \lambda \cdot \beta \\ &= -\frac{1}{N} \cdot \sum_n (R_n - \hat{\mu}) \cdot X_{n,i} + \beta \cdot \frac{1}{N} \cdot \sum_n X_{n,i}^2 + \lambda \cdot \beta \\ &= -\hat{\beta}_i + (1 + \lambda) \cdot \beta.\end{aligned}$$

Solving for β yields the desired result. \square

Proof (Lemma 2.2). Let $S[v^2]$ denote the shrinkage in a Ridge regression with $\lambda_i = se_i^2/v^2$:

$$S[v^2] = v^2/(v^2 + se_i^2).$$

The Ridge-regression optimization problem can then be expressed as:

$$\min_{v^2 > 0} \left\{ \mathbb{E} \left[(R_n - \hat{\mu} - S[v^2] \cdot \hat{\beta}_i \cdot X_{n,i})^2 \right] \right\}.$$

This optimization problem results in the following first-order condition:

$$\begin{aligned}0 &= 2 \cdot \mathbb{E} \left[(R_n - \hat{\mu} - S[v^2] \cdot \hat{\beta}_i \cdot X_{n,i}) \cdot (S'[v^2] \cdot \hat{\beta}_i \cdot X_{n,i}) \right] \\ &= 2 \cdot S'[v^2] \cdot (\hat{\beta}_i \cdot \mathbb{E}[(R_n - \hat{\mu}) \cdot X_{n,i}] - S[v^2] \cdot \mathbb{E}[(\hat{\beta}_i \cdot X_{n,i})^2]) \\ &= 2 \cdot S'[v^2] \cdot (\hat{\beta}_i^2 - S[v^2] \cdot \hat{\beta}_i^2) \\ &= 2 \cdot S'[v^2] \cdot (1 - S[v^2]) \cdot \hat{\beta}_i^2 \\ &= 2 \cdot \frac{se_i^4}{(v^2 + se_i^2)^3} \cdot (\hat{\beta}_i)^2.\end{aligned}$$

Taking the expectation with respect to realizations of the true slope coefficient yields:

$$0 = \mathbb{E} \left[2 \cdot \frac{se_i^4}{(v^2 + se_i^2)^3} \cdot \hat{\beta}_i^2 \right] = 2 \cdot \frac{se_i^4}{(v^2 + se_i^2)^3} \cdot (\sigma^2 + se_i^2).$$

The only way to satisfy this first-order condition is to choose $v^2 = \infty$. \square

Proof (Proposition 2.2). Suppose we add a correction term, $C[v^2]$, to the training error in Equation (12) to undo this in-sample overfitting. The objective function would then become:

$$\min_{v^2 > 0} \left\{ \mathbb{E} \left[(R_n - \hat{\mu} - S[v^2] \cdot \hat{\beta}_i \cdot X_{n,i})^2 \right] + C[v^2] \right\}.$$

Our goal is to find a functional form for $C[v^2]$ that yields an unbiased estimate of $v^2 = \sigma^2$.

Note that this corrected optimization problem yields the following first-order condition:

$$\begin{aligned} 0 &= 2 \cdot \mathbb{E} \left[(R_n - \hat{\mu} - S[v^2] \cdot \hat{\beta}_i \cdot X_{n,i}) \cdot S'[v^2] \cdot \hat{\beta}_i \cdot X_{n,i} \right] - C'[v^2] \\ &= 2 \cdot S'[v^2] \cdot (1 - S[v^2]) \cdot \hat{\beta}_i^2 - C'[v^2] \\ &= 2 \cdot \frac{se_i^4}{(v^2 + se_i^2)^3} \cdot \hat{\beta}_i^2 - C'[v^2]. \end{aligned}$$

And, taking the expectation of this corrected first-order condition with respect to realizations of the true slope coefficient yields:

$$0 = \mathbb{E} \left[2 \cdot \frac{se_i^4}{(v^2 + se_i^2)^3} \cdot \hat{\beta}_i^2 \right] - C'[v^2] = 2 \cdot \frac{se_i^4}{(v^2 + se_i^2)^3} \cdot (\sigma^2 + se_i^2) - C'[v^2].$$

By inspection, we see that choosing any $C[v^2]$ with the following first derivative,

$$\begin{aligned} C'[v^2] &= 2 \cdot \frac{se_i^4}{(v^2 + se_i^2)^3} \cdot (v^2 + se_i^2) \\ &= 2 \cdot se_i^4 \cdot (v^2 + se_i^2)^{-2}, \end{aligned}$$

will result in a minimum at $v^2 = \sigma^2$. Thus, by appropriately choosing the constant of integration, we can arrive at the desired result:

$$C[v^2] = 2 \cdot \left(\frac{1}{1 + se_i^2/v^2} \right) \cdot se_i^2.$$

□

B Distributional Assumptions

The statistical approach described in Section 2 models the anomaly-discovery process as independent draws from a normal distribution. The key assumption is that the strength of cross-sectional predictors is drawn from a common distribution. The assumptions of independence and normality are not key. And, we now show why.

B.1 Independence

To see why the assumption of independent draws is not crucial, note that Proposition 2.2 shows that $E[v_i^2] = \sigma^2$ for any $i \in \mathcal{I}_t$. If we relax the independence assumption, then the worst thing that could happen would be that the realized value of β_i^* is the same for all $i \in \mathcal{I}_t$. In this extreme case, we would effectively only have one signal about σ^2 . But, we could still use this lone signal, $v_1^2 = v_2^2 = \dots = v_{I_t}^2$. Thus, relaxing the independence assumption only affects the precision of our estimate for σ^2 .

B.2 Normality

To see why the assumption of normality is not crucial, let's consider an alternative setting where the true slope coefficients are instead drawn from a Laplace distribution:

$$\beta_i^* \stackrel{\text{iid}}{\sim} \text{Laplace}[\sqrt{2}/\sigma].$$

The probability density function of this Laplace distribution is given by $\text{pdf}[\beta] = \frac{1}{\sigma\sqrt{2}} \cdot e^{-\frac{\sqrt{2}}{\sigma}|\beta|}$, which implies that the mean and variance of the resulting draws are the same as in the original normally distributed case: $E[\beta_i^*] = 0$ and $\text{Var}[\beta_i^*] = \sigma^2$. We now show that, even though the true slope coefficients are being drawn from a different prior distribution, you can apply the exact same logic to estimate the anomaly baserate.

Inference Problem. If the true slope coefficients are drawn from a Laplace distribution, then the functional form of our inference problem will change slightly. Now, the negative log likelihood of the true slope coefficient taking on a particular value, $\beta_i^* = \beta$, given the realized cross-section of returns and lagged predictor values, $\{R_1, \dots, R_{N+1}\}$ and $\{X_{1,i}, \dots, X_{N+1,i}\}$, will correspond to

$$\begin{aligned} -\log \Pr[\beta] &= \frac{1}{2 \cdot (N \cdot se_i^2)} \cdot \sum_n (R_n - \hat{\mu} - \beta \cdot X_{n,i})^2 + \frac{\sqrt{2}}{\sigma} \times |\beta| + \dots \\ &= \frac{1}{2 \cdot se_i^2} \cdot \left\{ \frac{1}{N} \cdot \sum_n (R_n - \hat{\mu} - \beta \cdot X_{n,i})^2 + \sqrt{8} \cdot \frac{se_i^2}{\sigma} \times |\beta| \right\} + \dots \end{aligned}$$

where the “...” represents constants that do not depend on the choice of β . This inference problem suggests using a different penalized-regression procedure than before—i.e., a procedure with an absolute-value penalty rather than a quadratic penalty like a Ridge regression.

Bayesian LASSO. The least absolute shrinkage and selection operator (the LASSO; Tibshirani, 1996) is just such a penalized-regression procedure. Estimating the LASSO involves solving the optimization problem below:

$$\hat{\beta}_i[\lambda] \stackrel{\text{def}}{=} \arg \min_{\beta} \left\{ \frac{1}{N} \cdot \sum_n (R_n - \hat{\mu} - \beta \cdot X_{n,i})^2 + \lambda \cdot |\beta| \right\}.$$

Note that this is just the optimization problem given in Equation (5) when $q = 1$. What is more, when there is only one predictor that has been standardized to have zero mean and unit variance, it is possible to characterize the solution to this optimization problem analytically:

$$\hat{\beta}_i[\lambda] = \text{Sign}[\hat{\beta}_i] \cdot (|\hat{\beta}_i| - \lambda)_+.$$

Thus, as pointed out in Park and Casella (2008), the LASSO's absolute-value penalty can be interpreted as the effect of imposing Laplace priors on an inference problem when the tuning parameter is chosen as follows:

$$\lambda_i = \sqrt{8} \cdot se_i^2 / \sigma.$$

Econometric Estimator. The proposition below shows that, if the true slope coefficients are drawn from a Laplace distribution instead of a normal distribution, then we can learn about the anomaly baserate by studying the best-fit tuning parameter in the LASSO instead of a Ridge regression. Different prior distribution. Different penalized-regression procedure. Same underlying approach.

Proposition B.2 (Econometric Estimator, The LASSO). *Let $E[\cdot]$ denote an expectations operator evaluated with respect to realizations of β_i^* drawn from a Laplace distribution. If v_i denotes the parameter estimate with the minimum in-sample prediction error subject to an overfitting penalty for the i th predictor,*

$$v_i \stackrel{\text{def}}{=} \arg \min_{v>0} \left\{ \text{Err}_i[se_i^2/v^2] + 2 \cdot \mathbf{1}[|\hat{\beta}_i| > \sqrt{8} \cdot se_i^2/v] \cdot se_i^2 \right\},$$

then for all $\sigma > 0$ we have that $E[v_i] = \sigma$.

Proof (Proposition B.2). The $2 \cdot \mathbf{1}[|\hat{\beta}_i| > \sqrt{8} \cdot se_i^2/v] \cdot se_i^2$ term in Proposition B.2 is an information-criterion penalty. This sort of penalty takes the form $2 \cdot (df/N) \times \text{Var}[\varepsilon_{n,i}]$ where df represents the estimator's degrees of freedom. Zou, Hastie, and Tibshirani (2007) proves that the number of non-zero slope coefficients is an unbiased estimator for the degrees of freedom when using the LASSO:

$$\Pr[|\hat{\beta}_i| > \lambda] = \text{df}[\lambda].$$

Thus, since $\text{Var}[\varepsilon_{n,i}] = N \cdot se_i^2$, the generalized information-criterion penalty reduces to the one above when $\lambda_i = \sqrt{8} \cdot se_i^2/v$. □