ESSAYS ON
PREDICTING AND EXPLAINING THE CROSS SECTION OF STOCK RETURNS

A Dissertation
Submitted to
the Temple University Graduate Board

In Partial Fulfillment
of the Requirements for the Degree
DOCTOR OF PHILOSOPHY

---

by
Xun Zhong
Diploma Date May 2019

Examining Committee Members:

Oleg Rytchkov, Advisory Chair, Department of Finance
Gurdip Bakshi, Department of Finance
Kose John, External Member, Department of Finance, New York University
Lalitha Naveen, Department of Finance
Sudipta Basu, External Examiner, Department of Accounting

ABSTRACT

My dissertation consists of three chapters that study various aspects of stock return predictability.

In the first chapter, I explore the interplay between the aggregation of information about stock returns and p-hacking. P-hacking refers to the practice of trying out various variables and model specifications until the result appears to be statistically significant, that is, the p-value of the test statistic is below a particular threshold. The standard information aggregation techniques exacerbate p-hacking by increasing the probability of the type I error. I propose an aggregation technique, which is a simple modification of 3PRF/PLS, that has an opposite property: the predictability tests applied to the combined predictor become more conservative in the presence of p-hacking. I quantify the advantages of my approach relative to the standard information aggregation techniques by using simulations. As an illustration, I apply the modified 3PRF/PLS to three sets of return predictors proposed in the literature and find that the forecasting ability of combined predictors in two cases cannot be explained by p-hacking.

In the second chapter, I explore whether the stochastic discount factors (SDFs) of five characteristic-based asset pricing models can be explained by a large set of macroeconomic shocks. Characteristic-based factor models are linear models whose risk factors are returns on trading strategies based on firm characteristics. Such models are very popular in finance because of their superior ability to explain the cross-section of expected stock returns, but they are also criticized for their lack of interpretability. Each characteristic-based factor model is uniquely characterized by its SDF. To approximate the SDFs by a comprehensive

set of 131 macroeconomic shocks without overfitting, I employ the elastic net regression, which is a machine learning technique. I find that the best combination of macroeconomic shocks can explain only a relatively small part of the variation in the SDFs, and the whole set of macroeconomic shocks approximates the SDFs not better than only few shocks. My findings suggest that behavioral factors and sentiment are important determinants of asset prices.

The third chapter investigates whether investors efficiently aggregate analysts' earnings forecasts and whether combinations of the forecasts can predict announcement returns. The traditional consensus forecast of earnings used by academics and practitioners is the simple average of all analysts' earnings forecasts (Naive Consensus). However, this measure ignores that there exists a cross-sectional variation in analysts' forecast accuracy and persistence in such accuracy. I propose a consensus that is an accuracy-weighted average of all analysts' earnings forecasts (Smart Consensus). I find that Smart Consensus is a more accurate predictor of firms' earnings per share (EPS) than Naive Consensus. If investors weight forecasts efficiently according to the analysts' forecast accuracy, the market reaction to earnings announcements should be positively related to the difference between firms' reported earnings and Smart Consensus (Smart Surprise) and should be unrelated to the difference between firms' reported earnings and Naive Consensus (Naive Surprise). However, I find that market reaction to earnings announcements is positively related to both measures. Thus, investors do not aggregate forecasts efficiently. In addition, I find that the market reaction to Smart Surprise is stronger in stocks with higher institutional ownership. A trading strategy based on Expectation Gap, which is the difference between Smart and Naive Consensuses, generates positive risk-adjusted returns in the three-day window

around earnings announcements.

ACKNOWLEDGMENTS

First and foremost, I would like to thank my thesis advisors, Professors Oleg Rytchkov, Lalitha Naveen, Gurdip Bakshi, and Kose John for their guidance and support throughout the years of my study at Temple University. I am deeply indebted to my mentor and Dissertation Chair, Professor Oleg Rytchkov for his fundamental role in my Ph.D. education. He coached me on advanced econometric techniques, which paved my way to research in empirical asset pricing. He devoted a lot of effort and time to helping me with my papers by providing invaluable advice, and preparing me for my academic career. I am sincerely grateful for Professor Lalitha Naveen for her tremendous support and professional instructions along the way since my first day in graduate school. I am also very lucky to have Professor Gurdip Bakshi as one of my advisors during my last year at Temple, and I cannot overstate how much I have learned from his knowledge and experience. I am especially thankful to Professor Kose John for his continuous encouragement and invaluable advice.

I would also like to express my deep gratitude to Professors Connie Mao, Lily Li, Yan Li, Francesca Brusa, and Xiaohui Gao Bakshi, for their valuable comments and feedback. I also greatly benefited from the experties of Professor Sudipta Basu and his suggestions.

Last, but not least, I want to give my wholehearted thanks to my parents, Di Zhong and Yan Luo, and my fiancé, Fenglin Luan for their unconditional love and unwavering support. The past five years would have been more stressful without their faith in me.

TABLE OF CONTENTS

Page

ABSTRACT ii

ACKNOWLEDGMENTS v

LIST OF FIGURES ix

LIST OF TABLES ix

CHAPTER

1 INFORMATION AGGREGATION AND P-HACKINFORMATION AGGREGATION AND P-HACKINGING 1

Introduction 1

Time Series Predictability of Market Returns 10

Methodological Framework 10

Return Forecasts and Their Performance Measures 10

Three-Pass Regression Filter (3PRF) 12

Modified Three-Pass Regression Filter (3PRFm) 13

Simulation Analysis 16

Baseline Setting 16

Adding True Signals 23

Real Out-of-Sample Performance of 3PRF and 3PRFm Forecasts . . 25

Alternative Forecast Combination Techniques 28

Example: Predictability of Aggregate Stock Returns by 13 Predictors 32

Example: Predictability of Aggregate Stock Returns by Book-to-Market Ratios 35

Cross-Sectional Predictability of Individual Stock Returns 38

Information Aggregation Techniques 38

Simulation Analysis 43

Example: Aggregation of Information from Real Firm Characteristics . . . 48
Conclusion . . . 53

2 DO CHARACTERISTICS-BASED ASSET PRICING FACTORS REPRESENT MACROECONOMIC SHOCKS? A MACHINE LEARNING ANALYSIS 55
Introduction . . . 55
Empirical Approach . . . 63
Machine Learning . . . 63
Elastic Net . . . 64
Empirical Tuning . . . 67
Stochastic Discount Factors . . . 69
Description of Models and Macroeconomic Shocks . . . 70
Characteristic-Based Asset Pricing Models . . . 71
Macroeconomic Shocks . . . 74
Do Characteristic-based Factor Models Admit Macroeconomic Interpretations? . 80
Elastic Net Regression of SDFs on 131 Macroeconomic Shocks . . . 81
Explain SDFs with Nine Classic Macroeconomic Shocks . . . 85
Elastic Net Regression of SDFs on 131 Macroeconomic Shocks and Their Interaction Terms . . . 89
Do Intermediary Asset Pricing Factors Explain SDFs? . . . 92
Intermediary Asset Pricing Risk Factors . . . 92
Empirical Results . . . 93
Robustness Tests . . . 96
GMM Estimation of SDFs with Alternative Test Assets . . . 96
Construct Macroeconomic Shocks as AR(1) Innovations . . . 101
Conclusion . . . 102

3 DO INVESTORS EFFICIENTLY AGGREGATE ANALYSTS' FORECASTS? . 105
Introduction . . . 105
Data and Variables . . . 113
Sample Selection . . . 113

Naive Consensus and Smart Consensus 115
Naive Surprise, Smart Surprise, and Expectation Gap 117
Earnings Announcement Returns 118
Control Variables 121
Empirical Tests 125
Validity of Smart Consensus 125
Is Smart Consensus More Accurate Than Naive Consensus? 126
Does Expectation Gap Predict Naive Surprise? 130
Do Investors Efficiently Weight Analysts Forecasts? 138
Earnings Announcement Returns in Reaction to Naive and Smart Surprises 139
Subsample Analysis 145
Trading Strategies 148
Conclusion 150

BIBLIOGRAPHY 152

APPENDIX 166

LIST OF FIGURES

Figure Page

1.1 Real Out-of-Sample Performance of 3PRF and 3PRFm 27

2.1 Ten-Fold Cross-Validation 68

2.2 Estimated Monthly SDFs for Five Characteristic-Based Factor Models . . . 76

2.4 Heat Map of Correlations Between Macroeconomic Shocks 79

A1 Variable Selection by Elastic Net 170

LIST OF TABLES

Table Page

1.1 Simulated Quasi-Out-of-Sample Performance of 3PRF and 3PRFm 20

1.2 3PRF and 3PRFm Combinations of True and Spurious Predictors 24

1.3 Simulated Quasi-Out-of-Sample Performance of Alternative Forecast Combination Techniques 31

1.4 Out-of-Sample Predictability of Excess Market Returns by Combinations fo Standard Predictors 34

1.5 Out-of-Sample Predictability of Market Returns by the EW, 3PRF, and 3PRFm Combinations of Book-to-Market Ratios 37

1.6 Simulated Performance of Various Cross-Sectional Information Aggregation Techniques 47

1.7 Cross-Sectional Predictability of Individual Stock Returns 51

2.1 Risks Prices and SDFs for Five Characteristic-Based Factor Models 75

2.2 Elastic Net Regressions of SDFs on 131 Macroeconomic Shocks 83

2.3 SDFs and Nine Classic Macroeconomic Shocks 87

2.4 Explain SDFs with 131 Macroeconomic Shocks and Their Bivariate Interactions 91

2.5 Explain SDFs with Three Sets of Shocks 95

2.6 GMM Estimation of Risks Prices and SDFs with Alternative Test Assets . . 99

2.7 Elastic Net Regressions of Alternative SDFs on 131 Macroeconomic Shocks 100

2.8 Elastic Net Regressions of SDFs on 131 AR(1) Macroeconomic Shocks . . 102

3.1 Summary Statistics of Main Variables 122

3.2 Comparisons for Forecasting Errors of Two Consensus 129

3.3 The Signs of Naive Surprise Conditioning on Expectation Gap 132

3.4 Predict Naive Surprise with Expectation Gap 135

3.5 Explain Earnings Announcement Returns (CR) with Naive and Smart Surprises 141

3.6 Explain Alternative CARs with Naive and Smart Surprises 144

3.7 Explain CR or CAR with Naive and Smart Surprises in Subsamples 147
3.8 Abnormal Returns on A Trading Strategy Based on Expectation Gap 149
A.1 Description of 131 Macroeconomic Shocks 166

CHAPTER 1

INFORMATION AGGREGATION AND P-HACKINFORMATION AGGREGATION AND P-HACKINGING

Introduction

Academic business research heavily relies on statistical methods in general and hypothesis testing in particular. However, often only historical data are available to researchers, and they are destined to recycle the same datasets in multiple tests. As a consequence, researchers intentionally or unintentionally try out various variables and model specifications until the result appears to be statistically significant, that is, the p-value of the test statistic is below a particular threshold. The practice of highlighting statistical results with low p-values and ignoring results with high p-values is often referred to as p-hacking (e.g., Head, Holman, Lanfear, Kahn, and Jennions 2015, Harvey 2017). P-hacking is exacerbated by the publication process, which favors statistically significant results (e.g., Rosenthal 1979, Sterling, Rosenbaum, and Weinkam 1995). By inflating the probability of the type I error, p-hacking undermines the validity of inference and may lead to incorrect conclusions.

One of the areas that is particularly vulnerable to p-hacking is the search for predictors of stock returns (e.g., Foster, Smith, and Whaley 1997; Neuhierl and Schlusche 2011; Harvey, Liu, and Zhu 2016; Bartsch, Dichtl, Drobetz, and Neuhierl 2018).[1] It has been documented

[1] In finance, the cherry-picking of test results with low p-values can also distort the sizes of asset pricing tests (e.g., Lo and MacKinlay 1990), complicate the identification of risk factors (e.g., Ferson, Sarkissian, and Simin 1999; Kogan and Tian 2015), be responsible for the existence of calendar anomalies (e.g., Sullivan, Timmermann, and White 2001) and predictability of future earnings changes (e.g., Basu, Holland, and Sun 2012), make trading rules look more profitable than they actually are (e.g., Sullivan, Timmermann, and White

that almost 20 macroeconomic variables can significantly forecast market returns (e.g., Rapach and Zhou 2013). Moreover, hundreds of firm characteristics demonstrate statistically significant relations to returns on individual stocks in cross section (e.g., Green, Hand, and Zhang 2013; Hou, Xue, and Zhang 2017). However, because the tests of all predictors use the same historical returns, there is a widespread concern that at least some of the identified variables are spurious predictors produced by p-hacking.[2]

Besides the concern about p-hacking, the existence of many variables that arguably predict returns raises an important question: how to aggregate the information about returns contained in them and construct a single most powerful predictor? The aggregation brings about several advantages such as the efficient use of information, stability of forecast quality, and robustness to crashes in individual forecasts (e.g., Timmermann 2006). The aggregation techniques and their applications to return predictors have received much attention in the academic literature.[3] They are also relevant to practitioners, who develop multi-signal strategies that aggregate information from different sources in a single actionable forecast (e.g., Sorensen, Qian, Schoen, and Hua 2004; Grinold 2010; Jacobs and Levy 2014).

1999; Ready 2002; Bajgrowicz and Scaillet 2012; Harvey and Liu 2014), invalidate the standard measures of portfolio performance (e.g., Barras, Scaillet, and Wermers 2010; Bailey and Lopez de Prado 2014; Harvey and Liu 2015), and exacerbate the spurious regression bias (e.g., Ferson, Sarkissian, and Simin 2003).

[2]This possibility has been even discussed in popular press. See, for example, "When use of pseudo-maths adds up to fraud" (*Financial Times*, April 16, 2014) and "Huge returns at low risk? Not so fast" (*The Wall Street Journal*, June 27, 2014).

[3]An incomplete list of studies that aggregate individual firm characteristics includes Ou and Penman (1989), Lev and Thiagarajan (1993), Haugen and Baker (1996), Abarbanell and Bushee (1998), Piotroski (2000), Hanna and Ready (2005), Lewellen (2015), Green, Hand, and Zhang (2017), Light, Maslov, and Rytchkov (2017), Freyberger, Neuhierl, and Weber (2018), Jiang, Tang, and Zhou (2018), and Han, He, Rapach, and Zhou (2018). Stambaugh and Yuan (2017), Kozak, Nagel, and Santosh (2018), and Kelly, Pruitt, and Su (2018) construct new asset pricing models by aggregating information from multiple characteristics. The predictability of market returns by forecast combinations is investigated by Ludvigson and Ng (2007), Welch and Goyal (2008), Rapach, Strauss, and Zhou (2010), Kelly and Pruitt (2013), and Neely, Rapach, Tu, and Zhou (2014), among others.

Unfortunately, the standard information aggregation techniques exacerbate the impact of p-hacking, that is, the predictability tests applied to a combination of variables discovered by p-hacking incorrectly reject the null of no predictability even more often than when applied to the individual predictors. The objective of this chapter is to address this problem and propose an aggregation technique that is affected by p-hacking in the opposite way. When many individual predictors result from p-hacking, the proposed combination of them demonstrates no forecasting power in the predictability tests much more often than other combinations (and, therefore, produces more trustworthy backtesting results) but provides a more precise out-of-sample prediction. Such a property of an aggregate predictor will be called robustness to p-hacking.

Our approach builds upon the three-pass regression filter (3PRF) developed by Kelly and Pruitt (2013,0) in the time series context and modified for the cross-sectional setting by Light, Maslov, and Rytchkov (2017), who refer to it as the partial least squares regression (PLS). At the heart of 3PRF/PLS is the assumption that all individual predictors are related to the target variable (returns) through one or few common latent factors, which are not necessarily the main source of common variation in the predictors. The best predictor of the target variable is a linear combination of the latent factors, and 3PRF/PLS consistently estimates it when the number of time periods and predictors is large. The aggregation technique that we propose is a modification of 3PRF/PLS, and we refer to it as 3PRFm in the time series setting and as PLSm in the cross-sectional setting. In a nutshell, 3PRFm/PLSm prescribes to identify the sign of the relation between each individual predictor and returns in the available sample, invert the signs of the variables that negatively predict returns, and then apply the original 3PRF/PLS procedure. Obviously, 3PRFm/PLSm is easily imple-

mentable and, like 3PRF/PLS, provides consistent estimates of expected returns when the assumptions of 3PRF/PLS are satisfied.

The robustness of 3PRFm/PLSm to p-hacking hinges on two observations. First, if all individual predictors forecast returns with the same sign and have comparable quality, the 3PRF/PLS combination of them is a poor predictor of returns.[4] This is a peculiar finite sample property of 3PRF/PLS, which does not hold for other standard information aggregation techniques. Second, the test statistics of the predictors obtained by p-hacking tend to cluster in the vicinity of critical values that separate the rejection and acceptance regions (e.g., Masicampo and Lalande 2012, Basu and Park 2014, Simonsohn, Nelson, and Simmons 2014). For example, the t-statistics of predictive regression slopes concentrate right above 2 and right below -2 if those are the two-sided test's critical values. Therefore, the predictors produced by p-hacking are likely to have comparable forecasting power, although they may forecast returns with different signs. After inverting the signs of the variables that negatively predict returns, we obtain a set of spurious predictors that fit the first observation. As a result, 3PRF/PLS applied to the modified predictors yields a new variable that is a less precise in-sample forecast than that we would get without the sign inversion. However, spurious predictors are useless out of sample, so the aggregate forecast that downweights them has a higher out-of-sample quality.

The standard way to reduce the impact of p-hacking on statistical inference is to use multiple testing methods, which take into account that many hypotheses have been tested simultaneously by adjusting p-values of the individual hypotheses. Those methods include

[4]The 3PRF forecast is determined by the covariance between individual predictors and measures of their quality (the latter are proportional to the time-series covariances between individual predictors and future returns). The covariances are small and noisy when the dispersion in the predictor quality is low.

the Bonferroni adjustment (e.g., Miller 1981) as well as more sophisticated adjustments that control for the family-wise error rate (e.g., Holm 1979), false discovery rate (e.g., Benjamini and Hochberg 1995), and generalized false discovery rate (e.g., Sarkar 2006). One of the most popular techniques in financial applications is the "bootstrap reality check" proposed by White (2000) and refined by Hansen (2005) and Romano and Wolf (2005). Thus, an alternative to our aggregation technique would be a procedure that first finds truly statistically significant predictors by one of the multiple testing techniques and then aggregates all identified variables using one of the standard aggregation methods. However, compared to 3PRFm/PLSm, such a procedure would have two important limitations.

First, all multiple testing techniques assume that the total number of tested hypotheses (in our context, the total number of all variables tried by a p-hacker as predictors) is known, but practically this number is unavailable and should be estimated.[5] Unfortunately, the estimation is a daunting task, which is prone to two types of errors. On the one hand, it is easy to underestimate the freedom that a p-hacker has while looking for statistically significant predictors. Even though the original dataset may include a limited number of variables, there are numerous ways to create non-linear combinations of them. Moreover, a p-hacker can invoke an almost unlimited number of rules prescribing which variable should be used in the particular moment. On the other hand, it is unreasonable to assume that a p-hacker actually tried all possible predictors, so allowing for a large number of simultaneously tested hypotheses may make the identified variables look as weaker predictors than they really are. In addition, different predictors have been discovered by different researchers,

[5]Examples of such estimation include Sullivan, Timmermann, and White (1999), who characterize the universe of technical trading rules, Sullivan, Timmermann, and White (2001), who characterize the universe of calendar anomalies, and Harvey, Liu, and Zhu (2016) and Yan and Zheng (2017), who characterize the universe of all potentially tested predictors of stock returns.

and most likely those researchers have tried different numbers of potential predictors. This fact further complicates the application of the multiple testing framework. In contrast, our approach does not need any assumptions about the total number of variables tried by a p-hacker. This is a particularly important advantage when the objective is to aggregate information from predictors discovered by different researchers.

Second, as recently pointed out by Novy-Marx (2016), the multiple testing bias, which can be viewed as an outcome of p-hacking, and the overfitting bias, which arises as a result of combining information from multiple predictors by standard techniques, amplify each other. This observation suggests that even if the error produced by multiple testing is controlled for individual predictors when they are selected for aggregation, it still can be amplified by aggregation. In contrast, in our setting the effects of aggregation and p-hacking offset each other, not amplify, so aggregation itself can mitigate the adverse consequences of p-hacking without explicit identification of which individual predictors are truly informative.

To quantify the robustness of 3PRFm/PLSm to p-hacking, we conduct Monte Carlo simulations. Because p-hacking is inherently a finite sample problem, the standard asymptotic analysis is inapplicable, and analytical expressions for distributions of test statistics are unavailable. We separately consider two settings: (i) the time series setting with predictability of market returns by forecast combinations and (ii) the cross-sectional setting with predictability of individual stock returns in the cross section by combinations of firm characteristics. In both settings the simulation parameters closely match the standard empirical moments. To test the null hypothesis of no predictability, we use the Clark and West (2007) t-statistic in the time series analysis and the t-statistics of the Fama-MacBeth

regressions slopes and differences in returns on top and bottom decile portfolios formed by the predictor in the cross-sectional analysis. In the time series setting, we also report the out-of-sample R-squared statistic R^2_{OS}, which measures the usefulness of the forecast out of sample.[6]

First, we simulate variables that are truly unrelated to returns and find the distributions of the test statistics under the null of no predictability. In particular, we confirm that the Clark and West (2007) test has comparable actual and nominal sizes. Second, we generate a fixed number of spurious predictors by imitating p-hacking: to obtain each predictor, we repetitively conduct predictability tests for randomly generated variables and the same sample of returns until the null hypothesis of no predictability is rejected. Finally, using the pool of simulated spurious predictors, we explore the outcomes of the predictability tests applied to various combinations of them. In a separate set of tests, we also generate true signals and explore the performance of 3PLS and 3PRFm when true signals are combined with spurious ones.

Our simulations confirm that when all predictors are spurious, the predictability tests applied to the 3PRFm/PLSm combinations of them cannot reject the null of no predictability. Moreover, in the time series setting R^2_{OS} produced by the 3PRFm combinations of spurious forecasts are consistently negative. Importantly, those results cannot be explained by the inability of 3PRFm/PLSm to aggregate truly informative signals: as the fraction of true predictors in the set of combined variables grows, the null of no predictability is more often rejected, and R^2_{OS} of the 3PRFm forecast increases and becomes positive.

[6]Because spurious predictors are identified in the whole sample, we call the period for which R^2_{OS} is computed the quasi-out-of-sample period. The computation of R^2_{OS} replicates a typical backtesting of predictors.

The results are completely different for the 3PRF/PLS combinations of spurious signals. The tests almost always reject the null of no predictability, and the t-statistics are substantially higher than those of individual signals. We also find that higher critical values proposed in the literature as safeguards against p-hacking (e.g., Harvey, Liu, and Zhu 2016) may work well for individual predictors but can be woefully inappropriate for their aggregates. For example, in the cross-sectional setting the upper 5% quantile of t-statistics for individual spurious predictors is around 3, but it does not fall below 5 when 20 predictors are combined by PLS. The results are qualitatively similar for several other information aggregation techniques such as equal weighting, MSFE-based weighting, discounted MSFE-based weighting, OLS regression, and principal component regression (PCR) in the time series setting and the Fama-MacBeth regression, average percentile rank, and principal component analysis (PCA) in the cross-sectional setting. Overall, our findings highlight that in the presence of p-hacking, the standard predictability tests and even their adjustments for potential p-hacking of individual predictors reject the null of no predictability too often when applied to combinations of predictors.

To assess the practical usefulness of the 3PRF and 3PRFm forecasts, we compare their real out-of-sample behavior. We repeat the analysis using only the estimation period for generating spurious signals, not the whole sample, and find that the 3PRFm forecast on average delivers higher R^2_{OS} than the 3PRF forecast when many signals are spurious. However, it is less efficient and underperforms the 3PRF forecast when many signals are truly informative. This is the standard econometric tradeoff between robustness and efficiency, and our results imply that 3PRFm is valuable not only as a backtesting tool but also as a robust real-time aggregation technique when, as in many practical situations, p-hacking is

a serious concern.

To illustrate the applicability of 3PRFm/PLSm, we consider several empirical settings. First, we investigate whether excess returns on the S&P 500 index are predicted by various combinations of 13 variables individually identified by prior research as informative about future returns. Consistent with Rapach, Strauss, and Zhou (2010), we find that the equal-weighted forecasts, MSFE-weighted forecasts, and discounted MSFE-weighted forecasts produce positive R^2_{OS} for quarterly and monthly returns. In contrast, the 3PRFm forecasts typically have negative R^2_{OS} and the null of no predictability is not rejected, so p-hacking can be responsible for superior performance of the other forecast combinations.

Second, following Kelly and Pruitt (2013), we explore the predictability of returns on the CRSP stock index by combinations of book-to-market ratios of 6, 25, and 100 Fama-French value- and size-sorted portfolios. Note that in this case the predictors have been selected ex ante, not by their predictive ability, and, as a result, p-hacking is a much smaller concern than in the previous example. We empirically confirm this conclusion by demonstrating that R^2_{OS} produced by the 3PRFm combinations of 6 and 25 book-to-market ratios are positive for annual and monthly returns, and the null of no predictability is rejected at 10% level.

Third, we apply various cross-sectional aggregation techniques to firm characteristics that have been identified as predictors of individual stock returns in cross section. Motivated by Green, Hand, and Zhang (2017) and Light, Maslov, and Rytchkov (2017), we consider two sets of characteristics with 90 and 21 variables. We find that in both cases the estimates of expected returns constructed by PLSm can predict returns. Therefore, we can again conclude that at least several characteristics are informative, and it would be incorrect to fully attribute the forecasting power of the other aggregate estimates of expected returns

to p-hacking.

Overall, our results have two major implications. First, our analysis shows that not all information aggregation techniques are born equal. In contrast to the standard techniques, the proposed 3PRFm/PLSm approach provides at least partial safeguard against rejecting the null of no predictability too often in backtesting and delivers better out-of-sample predictions when many individual forecasts are produced by p-hacking. Second, our results urge researchers and practitioners to exercise caution when backtesting the forecasting power of variables that are combinations of several preselected predictors because in such cases the standard tests have incorrect sizes. It should be emphasized that although we study the interaction between information aggregation and p-hacking in the context of predicting stock returns, our methodological contribution and qualitative conclusions are general and apply to any set of forecasts and any target variable.

Time Series Predictability of Market Returns

In this section, we investigate the interplay between forecast combination and p-hacking in the time series setting. The cross-sectional predictability of individual stock returns by aggregate predictors in the presence of p-hacking is explored in a later section.

Methodological Framework

Return Forecasts and Their Performance Measures

The problem of forecasting market returns has the following general formulation. At time t an investor observes historical realizations of market returns r_s, $s = 1,\ldots,t$, and a set of

signals X_s^a, $a = 1,\ldots,A$, which without losing generality are assumed to be demeaned and standardized. The objective of the investor is to construct the best forecast of r_{t+1} that minimizes the expected mean squared forecast error (MSFE). To construct such forecasts $\hat{\mu}_t^a$ based on individual signals, the investor should i) run time series regressions of r_s individually on each signal X_{s-1}^a, $a = 1,\ldots,A$, using the observations $s = 2,\ldots,t$, ii) obtain the intercepts $\hat{\alpha}_t^a$ and slopes $\hat{\beta}_t^a$, and iii) find the fitted value $\hat{\mu}_t^a = \hat{\alpha}_t^a + \hat{\beta}_t^a X_t^a$.

Because the objective of forecasting is to minimize the MSFE, the forecast $\hat{\mu}_t^a$ is better than an alternative forecast $\hat{\mu}_t^{a'}$ if it has a smaller MSFE. A natural benchmark is the naive forecast $\bar{r}_t = \frac{1}{t}\sum_{s=1}^t r_s$. The comparison of the MSFE of the given forecast with that of the naive forecast motivates various forecast performance measures, and we will use two of them.[7] First, we will compute the out-of-sample R^2 statistic R_{OS}^2, which shows the reduction in the MSFE produced by the given forecast $\hat{\mu}_t$ relative to the naive forecast. The out-of-sample R^2 is defined as

$$R_{OS}^2 = 1 - \frac{\sum_{t=k}^{T-1}(r_{t+1} - \hat{\mu}_t)^2}{\sum_{t=k}^{T-1}(r_{t+1} - \bar{r}_t)^2},$$

where T is the total number of periods in the sample. To ensure that the imprecise forecasts from the beginning of the sample have only a limited impact on the performance measure, the computation of R_{OS}^2 starts in the period $k > 1$. The out-of-sample R^2 is widely used in the return predictability literature. Note that R_{OS}^2 is not a formal test statistic; it is only a diagnostic tool that shows the usefulness of the forecast out of sample.

Our second performance measure is the Clark and West (2007) test statistic of the hy-

[7] A comprehensive discussion of forecast evaluation techniques can be found in West (2006).

pothesis that the given forecast $\hat{\mu}_t$ has the same MSFE as the naive forecast $\bar{r}_t$ (the no predictability hypothesis). The test is conducted in several steps. First, using the realized excess returns r_{t+1}, the average excess returns $\bar{r}_t$, and the forecast $\hat{\mu}_t$, we compute

$$d_{t+1} = (r_{t+1} - \bar{r}_t)^2 - [(r_{t+1} - \hat{\mu}_t)^2 - (\bar{r}_t - \hat{\mu}_t)^2]$$

for $t = k, \ldots, T-1$, where k is the first observation in the testing period. The statistic d_{t+1} is dubbed by Clark and West (2007) the *MSPE-adjusted* statistic, and it adjusts the difference in the forecast precisions to the fact that the parameters of the forecast $\hat{\mu}_t$ are estimated. Then, we regress d_{t+1} on a constant and obtain the t-statistic of the intercept, which is the Clark and West (2007) test statistic. Its distribution under the null can be approximated by the standard normal distribution, and the p-value of the test is the probability that the realization of the standard normal random variable is larger than the realized t-statistic. In our analysis, we set the nominal size of the test to 5% and measure the actual size by $\text{Prob}(t > 1.645)$, where 1.645 is the upper 5% quantile of the standard normal distribution.

Three-Pass Regression Filter (3PRF)

In the presence of multiple signals X_t^a, $a = 1, \ldots, A$, the standard objective is to aggregate all of them in a single most informative forecast of returns $\hat{\mu}_t$. One of the powerful aggregation techniques is the three-pass regression filter (3PRF). The econometric foundations of this technique are laid out in Kelly and Pruitt (2015), and it has been used by Kelly and Pruitt (2013), Grover and McCracken (2014), Huang, Jiang, Tu, and Zhou (2015), Schmidt (2016), Lin, Wu, and Zhou (2016), and Haddad, Kozak, and Santosh (2017), among oth-

ers. 3PRF assumes that among latent factors that determine the correlation matrix of the signals, only few can predict the target variable (returns). In the simplest case, there is a single latent factor F_t, which is related to returns and predictors as $r_{t+1} = \beta_0 + \beta_1 F_t + \eta_{t+1}$ and $X_t^a = \phi^a F_t + \varepsilon_t^a$, where η_{t+1} and ε_t^a are orthogonal to F_t. In general, the errors ε_t^a are correlated across predictors and may contain other factors. Under those assumptions, 3PRF prescribes to construct the return forecast at time t in three steps: i) run separate time series regressions of each lagged signal X_{s-1}^a on the realized returns r_s, $s = 2,\ldots,t$ and retain the slopes $\hat{\phi}_t^a$, $a = 1,\ldots,A$, ii) run period-by-period regressions of the signals X_s^a on $\hat{\phi}_t^a$ for $s = 1,\ldots,t$ and retain the slope estimates $\hat{F}_s$, which are the estimates of the latent factor F_s, and iii) run a time series regression of realized returns r_s on $\hat{F}_{s-1}$, $s = 2,\ldots,t$, and use the obtained intercept $\hat{\beta}_{0t}$ and slope $\hat{\beta}_{1t}$ to construct the return forecast as $\hat{\mu}_t = \hat{\beta}_{0t} + \hat{\beta}_{1t}\hat{F}_t$. The multifactor version of the procedure is described in detail by Kelly and Pruitt (2015).

Modified Three-Pass Regression Filter (3PRFm)

The main contribution of this chapter is the modification of the 3PRF procedure dubbed 3PRFm that is robust to p-hacking. The modification is motivated by the following observation: if all individual predictors forecast returns with the same sign and they all have comparable quality, the 3PRF combination of them is a poor predictor of returns. To see the intuition behind this result, recall that the factor estimates $\hat{F}_s$ are obtained in the second step of 3PRF as slopes from the period-by-period regressions of X_s^a on the first step regression slopes $\hat{\phi}_t^a$. Thus, each $\hat{F}_s$ is proportional to the cross-predictor covariance between X_s^a and $\hat{\phi}_t^a$. When the dispersion of $\hat{\phi}_t^a$ is low, the covariances are small, and the 3PRF forecast has poor quality. In the extreme case with identical $\hat{\phi}_t^a$ for all predictors the procedure breaks

down. More formally, the factor estimates $\hat{F}_s$ are computed as

$$\hat{F}_s = \frac{\sum_{a=1}^{A}(\hat{\phi}_t^a - \overline{\hat{\phi}_t})(X_s^a - \bar{X}_s)}{\sum_{a=1}^{A}(\hat{\phi}_t^a - \overline{\hat{\phi}_t})^2} = C_t \left(\sum_{a=1}^{A} w_t^a X_s^a - \frac{1}{A} \sum_{a=1}^{A} X_s^a \right), \tag{1.1}$$

where

$$w_t^a = \frac{\hat{\phi}_t^a}{\sum_{a=1}^{A} \hat{\phi}_t^a}, \qquad C_t = \frac{\sum_{a=1}^{A} \hat{\phi}_t^a}{\sum_{a=1}^{A}(\hat{\phi}_t^a - \overline{\hat{\phi}_t})^2},$$

and $\overline{\hat{\phi}_t}$ and $\bar{X}_s$ are the cross-signal averages of the first-step regression slopes and signals, respectively. If all individual predictors X_s^a are related to returns with the same sign, which without losing generality can be assumed to be positive, then $\hat{\phi}_t^a > 0$, $w_t^a > 0$, and $C_t > 0$ for all signals in all periods. Equation (1.1) implies that the factor estimates can be represented as $F_s = C_t(X_s^W - X_s^E)$, where X_s^W and X_s^E are signal combinations with the weights w_t^a and $1/A$, respectively. Taking into account that the individual signals have been demeaned over time and using equation (1.1), the fitted values from the third step of 3PRF can be written as

$$\hat{\mu}_t = \bar{r}_t + \hat{C}_t(X_t^W - X_t^E), \qquad \hat{C}_t = \frac{\sum_{s=1}^{t-1} r_{s+1}(X_s^W - X_s^E)}{\sum_{s=1}^{t-1}(X_s^W - X_s^E)^2}. \tag{1.2}$$

Thus, the 3PRF forecast is the naive forecast $\bar{r}_t$ corrected by a term that is proportional to the difference $X_t^W - X_t^E$. When the individual predictors forecast returns with the same sign and have similar precisions, the first-step regression slopes $\hat{\phi}_t^a$ are close to each other, and the weights w_t^a cluster around $1/A$. Therefore, X_t^W and X_t^E, which are related to r_{t+1} with the same sign, have comparable forecasting power, and the difference $X_t^W - X_t^E$ has an ambiguous relation to future returns. Equation (1.2) implies that in this case $\hat{\mu}_t$ is the sum of $\bar{r}_t$ and noise, so $\hat{\mu}_t$ is a worse predictor than even the naive forecast.

The described property of 3PRF can be used for constructing aggregates that are robust to p-hacking. We propose a modification of 3PRF dubbed 3PRFm that instead of the original predictors X_s^a uses the standardized time-t individual forecasts $\hat{\mu}_s^a = \hat{\alpha}_t^a + \hat{\beta}_t^a X_s^a$, where $\hat{\alpha}_t^a$ and $\hat{\beta}_t^a$ are the intercept and slope from the regression of r_s on the individual signal X_{s-1}^a. Effectively, 3PRFm inverts the signs of those signals that are negatively related to future returns. When a p-hacker generates spurious individual predictors by running time series regressions of the same realizations of returns on various variables and conducting two-sided t-tests with the typical critical values of -2 and 2, the variables identified as informative are likely to have t-statistics concentrated just above 2 and just below -2. After making the signals positively aligned with returns, their t-statistics become concentrated just above 2. Because the precisions of the slope estimates are almost identical across the signals, the concentration of t-statistics implies low dispersion of the predictive regression slope estimates, and this is exactly the property penalized by 3PRF. Note that our modification of 3PRF does not invalidate its asymptotic properties derived in Kelly and Pruitt (2015), so the 3PRFm forecast converges to the infeasible best forecast as the number of periods and signals increases.

The intuition presented above also implies that the 3PRF and 3PRFm forecasts are particularly different when the numbers of positive and negative spurious predictors of returns are comparable. Indeed, in this case $\hat{\phi}_t^a$, $a = 1, \ldots, A$, are symmetrically distributed around 0, and $\overline{\hat{\phi}_t} \approx 0$ when the number of signals is relatively large (exactly when the 3PRF forecast becomes particularly precise). Therefore, the factor estimates $\hat{F}_s$ obtained in the second

step of 3PRF are

$$\hat{F}_s = \frac{1}{\sum_{a=1}^{A}(\hat{\phi}_t^a - \overline{\hat{\phi}}_t)^2}\left(\sum_{a=1}^{A}\hat{\phi}_t^a X_s^a - A\overline{\hat{\phi}}_t\bar{X}_s\right) \approx \frac{1}{\sum_{a=1}^{A}(\hat{\phi}_t^a)^2}\sum_{a=1}^{A}\hat{\phi}_t^a X_s^a.$$

Because $\hat{\phi}_t^a > 0$ ($\hat{\phi}_t^a < 0$) for those signals that positively (negatively) predict returns, $\hat{\phi}_t^a X_s^a$ are positively associated with future returns for all a. Hence, $\hat{F}_s$ has the unambiguously positive relation to returns even when the signals are spurious, and this relation is inherited by the 3PRF forecast $\hat{\mu}_t = \hat{\beta}_{0t} + \hat{\beta}_{1t}\hat{F}_t$. Therefore, in contrast to 3PRFm, 3PRF is likely to exacerbate p-hacking.

Simulation Analysis

Baseline Setting

To quantitatively evaluate the performance of 3PRF and 3PRFm in various situations, we use Monte Carlo simulations. P-hacking is intrinsically a finite sample problem, so the distributions of test statistics based on p-hacked forecasts cannot be characterized analytically. The simulation analysis complements our theoretical considerations because there is some dispersion even in the quality of spurious signals, and without simulations it is unclear whether this dispersion is strong enough to invalidate our arguments why 3PRFm should be robust to p-hacking. Moreover, quantitative results can be sensitive to the specifics of the setting such as the chosen parameters, sample size, etc.

The main steps of our simulation procedure are as follows. First, we simulate T realizations of returns r_t, $t = 1,\ldots,T$, as random draws from the normal distribution with the mean $\bar{\mu}$ and standard deviation $\bar{\sigma}$. We separately consider three sets of parameters that

correspond to annual returns ($T = 69$, $\bar{\mu} = 0.065$, $\bar{\sigma} = 0.16$), quarterly returns ($T = 276$, $\bar{\mu} = 0.016$, $\bar{\sigma} = 0.08$), and monthly returns ($T = 828$, $\bar{\mu} = 0.005$, $\bar{\sigma} = 0.04$). The parameter T is chosen so that the simulated sample has the same number of observations as the period from January 1947 to December 2015. The parameters $\bar{\mu}$ and $\bar{\sigma}$ are approximately equal to the average excess returns on the S&P 500 index with dividend reinvestment and the standard deviation of excess returns, respectively, in the same time period. The choice of the sample period and calibration of $\bar{\mu}$ and $\bar{\sigma}$ are motivated by our empirical analysis. In the Online Appendix we demonstrate that our conclusions do not change if the simulated returns are replaced with actual realizations of excess returns on the S&P 500 index.

Next, we generate two sets of signals X_t^a, $a = 1,\ldots,A_{max}$, with $A_{max} = 500$. In the first set, all realizations of signals are independent from returns and randomly drawn from the standard normal distribution. We use those signals to simulate the distributions of aggregate forecast statistics under the null of no predictability. The signals that are included in the second set are constructed using a procedure that mimics p-hacking; we refer to those signals as spurious predictors. To obtain each signal, we keep generating candidate signals until the candidate appears to predict the simulated returns in the whole sample. The realizations of the signals are independent over time and drawn from the standard normal distribution.[8] To test the forecasting ability of each signal X_{t-1}^a, we run the OLS regression of r_t on X_{t-1}^a and retain the signal if the absolute value of its t-statistic exceeds 2.[9] The

[8]Note that in this approach the total number of tried variables is random, but the number of the selected variables is fixed. An alternative approach would be to pick predictors with low p-values from a fixed pool of purely random variables, so that the total number of tried variables is fixed, but the number of selected variables is random. However, the latter approach is less appealing because practically it is hard to estimate the total number of variables tried by researchers, but it is easy to calibrate the number of selected variables knowing the number of empirically discovered statistically significant predictors.

[9]The Online Appendix contains the results of simulations with an alternative critical value of 3 and correlated spurious signals.

predictive regression is the main tool for discovering predictors of future returns in the finance literature (e.g., Campbell 1987, Fama and French 1989, Kothari and Shanken 1997, Pontiff and Schall 1998, Baker and Wurgler 2000). For consistency, each retained signal is demeaned and standardized in each moment, although this operation has a negligible effect on our results because spurious signals are independently drawn from the distribution with zero mean and unit variance.

In the baseline setting, we consider the 3PRF and 3PRFm combinations of $A = 15$ signals. This choice is motivated by our empirical analysis, which involves a similar number of empirical predictors of returns. Because the 3PRF and 3PRFm estimates are consistent only when the number of signals is large, the results of our simulations can be affected by the choice of A: a poor performance of an aggregate forecast can be attributed to a small number of signals but not to the robustness of the aggregation technique to p-hacking. To alleviate such concerns, in the Online Appendix we consider alternative specifications of A and demonstrate that our results are insensitive to its choice.

We split the whole sample into an estimation period with $t = 1,\ldots,k$, which is used solely for the estimation of the forecast parameters, and a quasi-out-of-sample period with $t = k+1,\ldots,T$, in which the forecasts are tested. Forecasts in the quasi-out-of-sample period are constructed using a recursive (expanding) estimation window, that is, in the period $t \geq k$ we form the forecast $\hat{\mu}_t$ of r_{t+1} using observations from the periods $s = 1,\ldots,t$. We set k such that the testing period starts with the observation corresponding to January 1981, so the number of quasi-out-of-sample periods is $T_{OS} = 35$ for annual returns, $T_{OS} = 140$ for quarterly returns, and $T_{OS} = 420$ for monthly returns. Note that our analysis is not truly out of sample because the individual predictors are identified by p-hacking in the

whole sample, and we use the term "quasi-out-of-sample" to highlight this fact.[10]

To find the distributions of test statistics for each forecast combination, we need B realizations of them computed from different sets of signals, where B is sufficiently large to ensure the accuracy of the inference. One way to obtain them is to produce B independent copies of A signals. However, for large A and B repeating p-hacking AB times is a computationally intensive procedure, so instead of it we use subsampling. Specifically, having a pool of A_{max} signals, we randomly draw from it with replacement B subsets with A signals each. Practically, we set $B = 1000$, so in the case with $A = 15$ and $A_{max} = 500$ we have $A_{max} \ll AB$. Therefore, subsampling substantially decreases the total number of individual signals required for the analysis. For each set with A signals, we apply 3PRF and 3PRFm, and for both aggregate forecasts compute their test statistics.

We start our analysis with the consideration of the benchmark case in which the 3PRF and 3PRFm forecasts aggregate signals that are independent from returns. We simulate 1000 realizations of the Clark and West (2007) t-statistics and R^2_{OS} for the aggregate forecasts and report the characteristics of the obtained distributions in Panel A of Table 1.1. Also, we present the same characteristics of the distributions obtained by computing the same statistics for the 500 individual signals.

Most importantly, our results confirm the conclusion of Clark and West (2007) who argue that the distribution of the Clark and West (2007) t-statistic can be reasonably approximated by the standard normal distribution, although the test is slightly undersized if the rejection region is determined by the standard normal critical values. Indeed, according

[10]Our procedure replicates a typical backtesting of forecasts. The real out-of-sample performance of the 3PRF and 3PRFm forecasts is also considered in a section, titled "".

Table 1.1. Simulated Quasi-Out-of-Sample Performance of 3PRF and 3PRFm

Each aggregate forecast combines 15 predictors. "Mean(R^2_{OS})" and "$q_{0.05}(R^2_{OS})$" are the mean and upper 5% quantile of the out-of-sample R^2; "Prob($R^2_{OS} < 0$)" is the empirical probability of negative R^2_{OS}. "Mean(t)" and "$q_{0.05}(t)$" are the mean and upper 5% quantile of the Clark and West (2007) test statistic; "Prob($t > 1.645$)" is the empirical probability that the Clark and West (2007) test statistic exceeds 1.645. The sizes of full samples and quasi-out-of-sample periods are $T = 69$ and $T_{OS} = 35$ for annual returns, $T = 276$ and $T_{OS} = 140$ for quarterly returns, and $T = 828$ and $T_{OS} = 420$ for monthly returns. Column Ind contains the statistics of the individual predictors.

Panel A: Random predictors									
	Annual returns			Quarterly returns			Monthly returns		
	Ind	3PRF	3PRFm	Ind	3PRF	3PRFm	Ind	3PRF	3PRFm
Mean(R^2_{OS}),%	-2.41	-20.30	-11.17	-0.48	-5.88	-2.32	-0.16	-2.11	-0.82
$q_{0.05}(R^2_{OS})$,%	4.46	0.80	4.49	0.99	0.37	2.09	0.47	-0.01	0.54
Prob($R^2_{OS} < 0$),%	78.80	93.90	88.90	79.20	94.30	84.10	78.20	95.20	85.10
Mean(t)	-0.28	-0.10	-0.13	-0.35	-0.05	-0.07	-0.30	0.05	-0.09
Prob($t > 1.645$),%	4.20	2.90	3.70	2.60	5.60	6.00	4.00	6.00	5.00
$q_{0.05}(t)$	1.53	1.44	1.43	1.33	1.73	1.79	1.45	1.79	1.64
Panel B: Spurious predictors									
	Annual returns			Quarterly returns			Monthly returns		
	Ind	3PRF	3PRFm	Ind	3PRF	3PRFm	Ind	3PRF	3PRFm
Mean(R^2_{OS}),%	4.53	42.74	-44.76	1.33	15.98	-10.87	0.42	5.62	-3.58
$q_{0.05}(R^2_{OS})$,%	14.17	58.70	-28.90	4.12	23.99	-7.79	1.29	8.96	-2.60
Prob($R^2_{OS} < 0$),%	22.80	0.30	100.00	20.80	0.20	100.00	21.20	0.60	100.00
Mean(t)	1.51	2.96	-2.53	1.52	4.61	-2.98	1.47	4.86	-3.03
Prob($t > 1.645$),%	40.40	99.50	0.00	45.20	99.90	0.00	38.80	99.90	0.00
$q_{0.05}(t)$	2.74	3.82	-1.46	2.66	5.67	-1.82	2.59	6.01	-1.72

to Table 1.1, the averages of the Clark and West (2007) t-statistics produced by individual signals and their aggregates are relatively close to zero. Moreover, the empirical probabilities that the Clark and West (2007) t-statistics exceed 1.645, which are actual sizes of the tests with the 5% nominal size, are slightly below 5% in a shorter sample with annual returns but on average closer to 5% in a longer sample with monthly returns. Accordingly, the simulated upper 5% quantiles of the t-statistics are not far from 1.645.

Panel A of Table 1.1 also shows that for all data frequencies the average R^2_{OS} is negative, but its absolute value decreases with the number of observations. This result holds

for both aggregate and individual predictors and illustrates two facts. First, the forecasts bases on useless signals produce higher MSFE than the naive forecast because the estimation of the forecasts adds noise. Second, the estimation noise becomes weaker as the number of observations increases, and the MSFEs of constructed forecasts approach those of naive forecasts. Consistently with the negative averages, the empirical probability to obtain negative R^2_{OS} is substantially higher than 50%, but in contrast to the mean it does not change with the sample size. Together with the decrease in the R^2_{OS} upper 5% quantile, this observation suggests that the dispersion of R^2_{OS} decreases with the number of observations.

Although the probabilities to falsely reject the null of no predictability are comparable for 3PRF, 3PRFm, and individual forecasts, the averages of their R^2_{OS} are quite different: they are most negative for the 3PRF forecast and least negative for the individual forecasts. This result is strong and robustly holds for all frequencies of returns. It demonstrates that the aggregation itself brings additional noise into the forecasts and increases their MSFEs compared to the MSFE of the naive forecast. However, the 3PRFm forecast is less prone to such noise than the 3PRF forecast.

The results are substantially different when 3PRF and 3PRFm aggregate spurious predictors obtained by p-hacking. Panel B of Table 1.1 demonstrates that for all frequencies of returns the probability that the Clark and West (2007) test applied to the 3PRF forecast rejects the null of no predictability is close to 100%, and it is much larger than its analog for individual signals.[11] Moreover, the 3PRF forecasts have on average high and positive R^2_{OS}, which substantially exceed R^2_{OS} of the individual signals. Thus, being a powerful informa-

[11] We use the predictive regression and the whole sample to find individual spurious predictors, so it is not surprising that the rejection rates of the Clark and West (2007) test applied to those predictors in the quasi-out-of-sample period are significantly lower than 100%.

tion aggregation technique, 3PRF exacerbates the detrimental consequences of p-hacking, and the actual sizes of the predictability tests become much larger than the tests' nominal sizes.

The results are opposite for the 3PRFm forecast. Panel B of Table 1.1 shows that the Clark and West (2007) test never rejects the null of no predictability, so due to p-hacking the actual size of the test is much smaller than its nominal size. Accordingly, for all return frequencies the averages of R^2_{OS} are large and negative, and as implied by reported $\text{Prob}(R^2_{OS} < 0)$, R^2_{OS} is negative for all randomly chosen combinations of the 15 spurious signals. Thus, p-hacking makes the tests of the 3PRFm forecast more conservative, and this fact justifies the use of 3PRFm when p-hacking is a potential concern.

The reported statistics for individual signals provide another interesting observation. Even though their average R^2_{OS} is positive, the individual realizations of R^2_{OS} have a relatively large dispersion, and approximately 20% of them are negative. This result can at least partially explain the findings of Welch and Goyal (2008), who demonstrate that many variables with statistically significant slopes in in-sample predictive regressions have negative R^2_{OS}.

To summarize, our simulations demonstrate that p-hacking distorts the sizes of the predictability tests applied to the 3PRF and 3PRFm forecasts in opposite directions. In particular, there is a very high probability to erroneously detect predictability of the 3PRF forecast when all signals are spurious but the same probability is very low for the 3PRFm forecast.

Adding True Signals

The demonstrated conservatism of the predictability tests applied to the 3PRFm forecast in the presence of p-hacking may indicate poor ability of 3PRFm to aggregate truly informative signals. To alleviate this concern, we repeat the simulation analysis from the previous section, but instead of combining A spurious signals we apply 3PRF and 3PRFm to A_{true} true signals and $A - A_{true}$ spurious signals, where A_{true} varies from 1 to A. For consistency with the previous section, we set $A = 15$. To model true predictability, we decompose realized returns into predictable and unpredictable components so that $r_{t+1} = \mu_t + \varepsilon_{t+1}$, where μ_t and ε_{t+1} are independent from each other and over time. We also assume that $\mu_t \sim \mathcal{N}(\bar{\mu}, \sigma_\mu^2)$, $\varepsilon_{t+1} \sim \mathcal{N}(0, \sigma_\varepsilon^2)$, and neither of them is observable at time t. The true signals X_t^a, which are noisy proxies for μ_t, are generated as $X_t^a = \delta^a \mu_t + u_t^a$, where u_t^a are uncorrelated with μ_t as well as with u_s^b when $s \neq t$ or $b \neq a$. To save space, we consider only the specification with quarterly returns and, as in the previous section, set $T = 276$ and $\bar{\mu} = 0.016$. The parameters σ_μ and σ_ε are calibrated as $\sigma_\mu = 0.03$ and $\sigma_\varepsilon = \sqrt{\bar{\sigma}^2 - \sigma_\mu^2} \approx 0.074$, where $\bar{\sigma} = 0.08$. Thus, the annualized volatility of expected returns is consistent with its estimates in the predictability literature (e.g., Cochrane 2011), and the total volatility of returns is the same as in the previous section. Also, for each signal $u_t^a \sim \mathcal{N}\left(0, 1 - (\delta^a)^2 \sigma_\mu^2\right)$, and the calibration of variance ensures that as before each signal has a unit standard deviation. The loadings δ^a are assumed to be independently drawn from the uniform distribution $U[-1/\sigma_\mu, 1/\sigma_\mu]$. This specification implies that $\sigma(\delta^a \mu_t)$, which is the informativeness of the signal about the predictable component of returns, varies from 0 to 1, and signals can be either positively or negatively related to future returns. For con-

sistency with the generation process of spurious signals, we retain only those true signals that have absolute values of in-sample t-statistics greater than 2.

To assess the predictive power of the 3PRF and 3PRFm forecasts, we again use the signal subsampling procedure. First, we generate $A_{max} = 500$ spurious signals and $A_{max} = 5000$ true signals as described above.[12] Next, for each chosen A_{true}, we draw $B = 1000$ copies of A_{true} signals from the pool of true signals and $A - A_{true}$ signals from the pool of spurious signals. For each set of signals, we construct the 3PRF and 3PRFm forecasts, compute the out-of-sample statistics, and report them in Table 1.2.

Table 1.2. 3PRF and 3PRFm Combinations of True and Spurious Predictors

A_{true} indicates the number of true predictors among the 15 combined predictors. There are $T = 276$ observations in each simulated sample, and the quasi-out-of-sample period contains $T_{OS} = 140$ observations. "Mean(R^2_{OS})" and "$q_{0.05}(R^2_{OS})$" are the mean and upper 5% quantile of the out-of-sample R^2; "Prob($R^2_{OS} < 0$)" is the empirical probability of negative R^2_{OS}. "Mean(t)" and "$q_{0.05}(t)$" are the mean and upper 5% quantile of the Clark and West (2007) test statistic; "Prob($t > 1.645$)" is the empirical probability that the Clark and West (2007) test statistic exceeds 1.645.

A_{true}		1	3	5	7	9	11	13	15
3PRF	Mean(R^2_{OS}),%	20.10	22.75	21.84	19.99	18.70	17.75	16.69	15.93
	$q_{0.05}(R^2_{OS})$,%	29.13	29.44	27.47	24.47	22.42	20.99	19.41	18.45
	Prob($R^2_{OS} < 0$),%	0.10	0.10	0.00	0.00	0.00	0.00	0.00	0.00
	Mean(t)	5.04	5.27	5.15	4.95	4.83	4.71	4.60	4.52
	Prob($t > 1.645$),%	99.90	100.00	100.00	100.00	100.00	100.00	99.90	100.00
	$q_{0.05}(t)$	6.14	6.15	5.87	5.58	5.35	5.17	5.01	4.89
3PRFm	Mean(R^2_{OS}),%	-8.49	-3.83	-1.33	0.35	1.67	2.61	3.11	3.56
	$q_{0.05}(R^2_{OS})$,%	-3.12	1.84	3.59	4.46	5.41	6.46	7.39	8.88
	Prob($R^2_{OS} < 0$),%	99.90	83.70	59.30	38.20	22.30	15.20	12.10	13.30
	Mean(t)	-1.90	-0.04	0.83	1.39	1.76	2.06	2.21	2.37
	Prob($t > 1.645$),%	0.00	9.20	27.40	45.70	63.00	75.80	83.00	84.50
	$q_{0.05}(t)$	0.55	1.89	2.32	2.55	2.80	3.08	3.28	3.52

Table 1.2 shows that the quality of the 3PRFm forecast monotonically increases with

[12] Our simulations show that the distributions of performance metrics are more sensitive to the particular realizations of true signals than to the particular realizations of spurious signals. Therefore, we use a larger pool of true signals in our analysis.

A_{true}. The case with one true signal is very similar to the case with only spurious signals: the Clark and West (2007) t-statistic never exceeds the critical value of 1.645, and almost all realized R^2_{OS} are negative. However, as the proportion of true signals in the set of aggregated signals becomes larger, the average Clark and West (2007) t-statistic increases from -1.90 for $A_{true} = 1$ to 2.37 for $A_{true} = 15$. In the latter case, the null hypothesis that the 3PRFm forecast cannot predict returns is rejected with the probability 84.5%, so the Clark and West (2007) test has a substantial power to reject the null. A similar pattern is observed for average R^2_{OS}: it monotonically increases and approaches 3.56% when all signals are true. In this case the probability to observe negative R^2_{OS} is only 13.30%. Thus, our results confirm that 3PRFm does have power to aggregate true signals and generate a reasonable forecast of future returns.

The simulation results for the 3PRF forecast reveal two patterns. First, all performance metrics are particularly strong when the number of spurious signals is large and tend to decrease with the number of true signals. Thus, the 3PRF forecast is substantially affected by p-hacking. Second, 3PRF delivers higher average R^2_{OS} and the Clark and West (2007) t-statistic than 3PRFm when all signals are truly informative. This is a manifestation of the standard econometric tradeoff between robustness and efficiency: compared to the 3PRF forecast, the 3PRFm forecast is more conservative and robust to p-hacking but less efficient when all signals contain information about future returns.

Real Out-of-Sample Performance of 3PRF and 3PRFm Forecasts

So far our analysis of 3PRF and 3PRFm was not truly out of sample: even though the aggregates of signals at time t were constructed using only the information revealed at

times $s \leq t$, the signals themselves were selected based on their ability to predict returns in the whole sample. In this section, we explore the real out-of-sample performance of the 3PRF and 3PRFm forecasts and their usefulness as aggregation techniques, which is measured by R^2_{OS}. We combine A_{true} true signals and $15 - A_{true}$ spurious signals, where A_{true} varies from 1 to 15. The whole sample again contains $T = 276$ observations, and it is split into the estimation period with $T - T_{OS} = 136$ observations and the out-of-sample period with $T_{OS} = 140$ observations. The true signals are constructed as in the previous section. However, spurious signals are obtained by p-hacking only in the estimation period, and their realizations in the out-of-sample period, which are independently drawn from the standard normal distribution, are useless for predicting returns.

Figure 1.1 plots the simulated cumulative distribution functions (cdfs) of R^2_{OS} for various A_{true} and reveals two major patterns. First, as the number of true signals increases, the cdfs of the 3PRF and 3PRFm forecasts drift to the right. This result is intuitive because a larger number of informative signals makes the aggregates more informative and increases their R^2_{OS}. In particular, when the number of true signals exceeds 7, both the 3PRF and 3PRFm forecasts have positive R^2_{OS} for almost all signal realizations.

Second, the comparison of cdfs corresponding to 3PRF and 3PRFm shows that the relative performance of those techniques depends on the number of true signals. When the majority of signals are spurious, the 3PRFm forecast is more likely to produce positive R^2_{OS} than the 3PRF forecast, which effectively puts higher weights on spurious signals that are useless out of sample. For example, when $A_{true} = 3$, almost all R^2_{OS} produced by 3PRFm are positive, whereas approximately one third of the realizations of R^2_{OS} produced by 3PRF are negative. Moreover, R^2_{OS} of the 3PRF forecasts are more dispersed and can be as low as

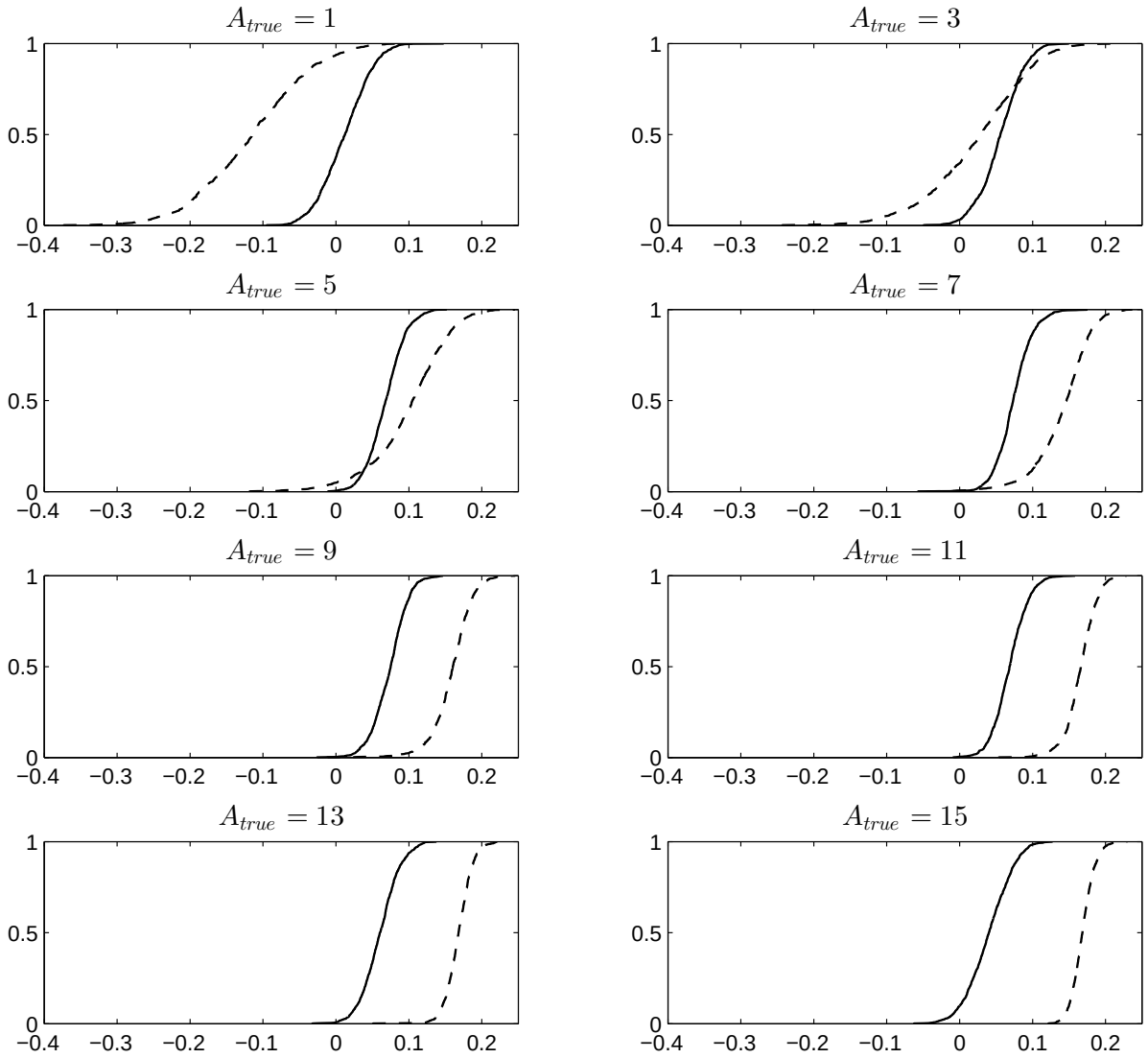

Figure 1.1. Real Out-of-Sample Performance of 3PRF and 3PRFm

The figure plots the cumulative distribution functions of simulated real R^2_{OS} of the 3PRF (dashed line) and 3PRFm (solid line) combinations of A_{true} true predictors and $15 - A_{true}$ spurious predictors. The signal search period contains 136 observations, and the out-of-sample period contains 140 observations.

-10%. The situation is different when the majority of signals are true: although almost all R^2_{OS} of 3PRF and 3PRFm are positive, they are typically larger for 3PRF. Thus, in this case 3PRF is a more efficient aggregation techniques than 3PRFm.

Overall, our out-of-sample results are consistent with the quasi-out-of-sample results from the previous section, and they illustrate the standard econometric tradeoff between robustness and efficiency. The 3PRF forecast is more efficient, but it has high quality only when the individual signals are true. Thus, it is not robust to p-hacking. In contrast, the 3PRFm forecast is less accurate when signals are true, but it is robust to p-hacking and delivers more reliable out-of-sample predictions than the 3PRF forecast when many signals are spurious, as it is likely to be in realistic situations.

Alternative Forecast Combination Techniques

In this section, we consider several alternative forecast combination techniques that have been used for predicting market returns and demonstrate that all of them, in contrast to 3PRFm, are not robust to p-hacking.

Equal-weighted forecasts (EW). The simplest way to combine forecasts is to assign equal weights to them. In this case, $\hat{\mu}_t = \frac{1}{A}\sum_{a=1}^{A}\hat{\mu}_t^a$, where $\hat{\mu}_t^a$ is the individual forecast.

MSFE-weighted forecast (MSFE). Simple averaging of forecasts can be suboptimal when the forecasts have different precisions. An alternative approach is to average forecasts with the weights that are inversely proportional to their past MSFEs. The MSFE-weighted forecast of returns is computed as $\hat{\mu}_t = \sum_{a=1}^{A} w_t^a \hat{\mu}_t^a$, where $w_t^a = (\kappa_t^a)^{-1}/\sum_{i=1}^{A}(\kappa_t^i)^{-1}$ with $\kappa_t^a = \sum_{s=m}^{t-1}(r_{s+1} - \hat{\mu}_s^a)^2$, and a new parameter m is such that $1 < m < t-1$. We choose m so that at least 10 years of data are used for the estimation of the weights w_t^a. Note

that the weights assigned to each individual forecast are updated every period to reflect the forecast's most recent performance. This approach was proposed by Bates and Granger (1969) and used for combining forecasts of excess market returns by Rapach, Strauss, and Zhou (2010).

Discounted MSFE-weighted forecast (DMSFE). This approach is similar to the previous one, but now the weights are computed as $w_t^a = (\kappa_t^a)^{-1}/\sum_{i=1}^{A}(\kappa_t^i)^{-1}$ with $\kappa_t^a = \sum_{s=m}^{t-1}\theta^{t-1-s}(r_{s+1}-\hat{\mu}_s^a)^2$. The parameter $\theta \in (0,1]$ measures how quickly the contribution of the forecast error to the MSFE decreases with time. Following Rapach, Strauss, and Zhou (2010), we set $\theta = 0.9$. Obviously, when $\theta = 1$ the discounted MSFE-weighted forecast coincides with the MSFE-weighted forecast.

OLS regression (OLS). This approach uses the fitted values from the OLS regression of realized returns on lagged signals (e.g., Welch and Goyal 2008; Rapach, Strauss, and Zhou 2010). To construct the forecast at time t from the signals X_t^a, $a = 1,\ldots,A$, we i) run a time series regression of r_s on the lagged signals X_{s-1}^a, $s = 2,\ldots,t$, ii) obtain the estimated intercept $\hat{\alpha}_t$ and slopes $\hat{\beta}_t^a$, and iii) compute the fitted value $\hat{\mu}_t = \hat{\alpha}_t + \sum_{a=1}^{A}\hat{\beta}_t^a X_t^a$. In contrast to the approaches described above, this approach uses the original signals X_t^a, not the individual forecasts $\hat{\mu}_t^a$, and it is applicable only when the number of predictors is relatively small compared to the number of time periods.

Principal component regression (PCR). This technique is particularly popular in the macroeconomic forecasting literature (e.g., Stock and Watson 1999, 2002a,0, Forni, Hallin, Lippi, and Reichlin 2003, Boivin and Ng 2006, Bai and Ng 2008), although it has also found applications in the return predictability literature (e.g., Neely, Rapach, Tu, and Zhou

2014). We limit our analysis to the specification with one principal component and construct the forecast $\hat{\mu}_t$ in three steps: i) using the data from all periods $s \leq t$, define the new signal G_s, $s = 1,\ldots,t$, as the first principal component of the signals X_s^a, ii) run the time series regression of r_s on G_{s-1} for $s = 2,\ldots,t$, and obtain the intercept $\hat{\alpha}_t$ and slope $\hat{\beta}_t$, and iii) compute the fitted value $\hat{\mu}_t = \hat{\alpha}_t + \hat{\beta}_t G_t$. In contrast to the OLS regression, PCR is applicable even when the number of signals is large (e.g., Stock and Watson 2002a).

To assess the sensitivity of the described techniques to p-hacking, we conduct simulations for each technique and report the obtained statistics in Table 1.3, which provides two main observations. First, the Clark and West (2007) tests applied to the considered combinations of spurious signals almost always reject the null of no predictability. The weakest result is demonstrated by PCR applied to monthly signals and returns, but even in this case the rejection rate is 67.2%. Accordingly, average R^2_{OS} for all forecast combinations are positive, and for many combinations R^2_{OS} never becomes negative. The conclusions are consistent across frequencies, although R^2_{OS} decreases with the number of observations in the sample.

Second, combining of forecasts substantially exacerbates p-hacking: the combinations have higher average R^2_{OS} and Clark and West (2007) rejection probabilities than the individual spurious signals. For example, for quarterly returns and individual signals the average R^2_{OS} is only 1.33% and Prob$(t > 1.645)$ is 45.2%, whereas they are above 3.3% and 83.8%, respectively, for all forecast combinations. Although the magnitudes of the effect are likely to depend on statistical properties of spurious signals and returns (volatilities, correlations, autocorrelations, etc.), Table 1.3 indicates that they are unambiguously large.

Table 1.3. Simulated Quasi-Out-of-Sample Performance of Alternative Forecast Combination Techniques

Each technique is applied to 15 spurious predictors. "Mean(R^2_{OS})" and "$q_{0.05}(R^2_{OS})$" are the mean and upper 5% quantile of the out-of-sample R^2; "Prob($R^2_{OS} < 0$)" is the empirical probability of negative R^2_{OS}. "Mean(t)" and "$q_{0.05}(t)$" are the mean and upper 5% quantile of the Clark and West (2007) test statistic; "Prob($t > 1.645$)" is the empirical probability that the Clark and West (2007) test statistic exceeds 1.645. Column Ind contains the statistics of individual spurious forecasts.

Panel A: Annual returns ($T = 69$, $T_{OS} = 35$)						
	Ind	EW	MSFE	DMSFE	OLS	PCR
Mean(R^2_{OS}),%	4.53	12.43	12.25	12.25	26.99	36.55
$q_{0.05}(R^2_{OS})$,%	14.17	15.10	15.05	15.05	48.64	53.21
Prob($R^2_{OS} < 0$),%	22.80	0.00	0.00	0.00	3.00	0.20
Mean(t)	1.51	3.00	2.94	2.95	2.79	2.78
Prob($t > 1.645$),%	40.40	100.00	100.00	100.00	99.20	98.80
$q_{0.05}(t)$	2.74	3.85	3.81	3.82	3.76	3.72
Panel B: Quarterly returns ($T = 276$, $T_{OS} = 140$)						
	Ind	EW	MSFE	DMSFE	OLS	PCR
Mean(R^2_{OS}),%	1.33	3.34	3.34	3.33	14.92	6.93
$q_{0.05}(R^2_{OS})$,%	4.12	4.11	4.12	4.13	23.62	15.94
Prob($R^2_{OS} < 0$),%	20.80	0.00	0.00	0.00	0.10	7.40
Mean(t)	1.52	4.75	4.71	4.70	4.60	2.90
Prob($t > 1.645$),%	45.20	100.00	100.00	100.00	100.00	83.80
$q_{0.05}(t)$	2.66	5.80	5.73	5.71	5.71	4.77
Panel C: Monthly returns ($T = 828$, $T_{OS} = 420$)						
	Ind	EW	MSFE	DMSFE	OLS	PCR
Mean(R^2_{OS}),%	0.42	1.06	1.06	1.06	5.71	1.38
$q_{0.05}(R^2_{OS})$,%	1.29	1.32	1.32	1.31	8.97	4.22
Prob($R^2_{OS} < 0$),%	21.20	0.00	0.00	0.00	0.20	16.00
Mean(t)	1.47	5.03	5.01	5.01	4.96	2.20
Prob($t > 1.645$),%	38.80	100.00	100.00	100.00	100.00	67.20
$q_{0.05}(t)$	2.59	6.07	6.06	6.06	6.01	4.28

Example: Predictability of Aggregate Stock Returns by 13 Predictors

Although there are many studies that document the predictability of aggregate stock returns in sample, the evidence of the out-of-sample predictability is scarce and inconclusive. On the one hand, Welch and Goyal (2008) argue that the most prominent in-sample predictors of returns are incapable of forecasting returns out of sample. On the other hand, Rapach, Strauss, and Zhou (2010) demonstrate that the combinations of individual forecasts deliver statistically and economically significant R^2_{OS}. In this section, we extend the analysis of Rapach, Strauss, and Zhou (2010) by applying 3PRFm to the same set of predictors and exploring whether the obtained combination can forecast returns. The latter would indicate that the predictive ability of at least several variables cannot be attributed to p-hacking. We also combine the individual predictors by 3PRF and the five standard techniques and compare the results with those produced by 3PRFm.

The target variable is the excess market return, which is constructed as the difference between continuously compounded return on the S&P 500 index, including dividends, and the three-month Treasury bill rate. We explore the predictability of returns at the annual, quarterly, and monthly frequencies. To make our results comparable to those in the literature, we employ 13 predictors used by Welch and Goyal (2008) and Rapach, Strauss, and Zhou (2010), which are the dividend-price ratio, dividend yield, earnings-price ratio, stock variance, book-to-market ratio, net equity expansion, Treasury bill rate, long-term bond yield, long-term bond return, default yield spread, default return spread, inflation, and investment-to-capital ratio.[13] The construction of the predictors is described in Welch

[13]We are grateful to Amit Goyal for providing the updated data on his web page http://www.hec.unil.ch/agoyal/. The dividend-payout ratio and term spread have been dropped because they are linear combinations of the other variables. The investment-to-capital ratio is available only at

and Goyal (2008). For consistency with the simulation analysis, each variable in each period is demeaned and standardized before being used in a forecast combination.

Our sample covers the period from January 1947 to December 2015. The choice of 1947 as the first year maximizes the number of predictors available for the whole sample and reduces the number of structural breaks in the predictability of returns (e.g., Chen 2009). The testing period for the out-of-sample performance starts in January 1981.[14] Following Rapach, Strauss, and Zhou (2010), for the MSFE- and discounted MSFE-weighted forecasts we use an additional 10-year holdout out-of-sample period that covers the years from 1971 to 1980 to assess the individual forecast quality and initiate the weights assigned to individual forecasts in their combinations. As in the simulation analysis, we measure the performance of each forecast combination by R^2_{OS} and test its superiority relative to the historical average using the Clark and West (2007) t-statistic.

Table 1.4 reports our results. First, it largely confirms the conclusions of Rapach, Strauss, and Zhou (2010), who find that quarterly returns are predicted by forecast combinations: the EW, MSFE, and DMSFE forecasts have positive R^2_{OS}, and their Clark and West (2007) t-statistics are marginally statistically significant. Note that our out-of-sample period, which spans the years 1981–2015, substantially overlaps with the out-of-sample period 1976–2005, in which according to Rapach, Strauss, and Zhou (2010) the predictability of returns is relatively weak. Our results also show that a similar conclusion holds for monthly returns, although the p-values of the Clark and West (2007) test tend to be higher. However, for the annual returns R^2_{OS} are negative for all forecast combinations except for

the annual and quarterly frequencies and, therefore, excluded from the analysis of monthly returns.

[14] In the Online Appendix we repeat the analysis for alternative out-of-sample periods and demonstrate the robustness of our results.

the EW forecast, for which it is positive but close to zero. Thus, the conclusions of Rapach, Strauss, and Zhou (2010) do not generalize to annual returns.

Table 1.4. Out-of-Sample Predictability of Excess Market Returns by Combinations fo Standard Predictors

The forecasts combine 13 predictors in the cases of annual and quarterly returns and 12 predictors in teh case of monthly returns. The sample is from January 1947 to December 2015; the out-of-sample period starts in January 1981. R^2_{OS} is the empirical out-of-sample R^2; "p-value" is the p-value of the Clark and West (2007) t-statistic.

				Panel A: Annual returns			
	EW	MSFE	DMSFE	OLS	PCR	3PRF	3PRFm
R^2_{OS},%	0.09	-0.60	-1.03	-134.75	-12.98	-19.31	-34.31
p-value	0.38	0.45	0.52	0.84	0.41	0.35	0.86
				Panel B: Quarterly returns			
	EW	MSFE	DMSFE	OLS	PCR	3PRF	3PRFm
R^2_{OS},%	0.88	0.86	0.89	-38.78	-2.69	-1.80	-5.54
p-value	0.10	0.11	0.10	0.38	0.37	0.08	0.45
				Panel C: Monthly returns			
	EW	MSFE	DMSFE	OLS	PCR	3PRF	3PRFm
R^2_{OS},%	0.30	0.30	0.30	-3.42	-0.80	-1.41	0.36
p-value	0.14	0.15	0.16	0.07	0.38	0.08	0.13

Second, Table 1.4 unambiguously documents the failure of the forecasts based on OLS, PCR, and 3PRF to produce positive out-of-sample R^2. For OLS this result is also reported by Welch and Goyal (2008) and Rapach, Strauss, and Zhou (2010), and it is largely attributed to the instability of the estimated regression slopes. We demonstrate that the result extends to PCR and 3PRF, although for them R^2_{OS} is less negative than for OLS.

Third, R^2_{OS} produced by 3PRFm is negative for annual and quarterly returns, and its absolute value is larger than that of R^2_{OS} produced by 3PRF. In both cases the Clark and West (2007) test does not reject the null of no predictability. For monthly returns, R^2_{OS} of the 3PRFm forecast is positive, but the Clark and West (2007) p-value is only 0.13. Thus,

our results do not dismiss the possibility that the documented good performance of the EW, MSFE, and DMSFE forecasts can be explained by p-hacking. However, the failure of the 3PRFm forecast can also result from the forecast instability, and negative R^2_{OS} of the 3PRF forecast suggests that this is likely to be the case. Because of that, our results do not exclude the possibility that one or several considered variables actually contain information about future returns.

Example: Predictability of Aggregate Stock Returns by Book-to-Market Ratios

An alternative set of predictors for aggregate stock returns has been proposed by Kelly and Pruitt (2013), who argue that the cross section of valuation ratios aggregated by 3PRF can reliably forecast market returns out of sample. We explore the robustness of this result to p-hacking by separately applying 3PRFm to log book-to-market ratios of 6, 25, and 100 size- and value-sorted portfolios.

We obtained the data on the portfolios from Ken French's library.[15] All stocks are sorted into portfolios at the end of June according to their book-to-market ratios and sizes. The book-to-market ratio of a portfolio in month t is calculated as the latest available annual portfolio book equity divided by the portfolio market equity at the end of the month. We adopt the standard convention that the annual book equity of year t is available after June of year $t+1$ (e.g., Fama and French 1993).

Like Kelly and Pruitt (2013), we choose our target variables to be annual and monthly log returns including dividends on the value-weighted CRSP index. We have also extended the analysis to quarterly log returns. However, we predict non-overlapping quarterly and

[15] http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html. We are grateful to Ken French for making the data publicly available.

annual returns, whereas Kelly and Pruitt (2013) predict rolling annual returns every month. Our data on book-to-market ratios covers the period from January 1930 to December 2015, so the last predicted annual return is realized in December 2016. Compared to Kelly and Pruitt (2013), our sample includes five additional years. To save space, we report only the performance of the EW, 3PRF, and 3PRFm forecasts, which as before is measured by R^2_{OS} and the Clark and West (2007) p-values.

Overall, the results reported in Table 1.4 broadly support the conclusions of Kelly and Pruitt (2013) and confirm that the 3PRF forecasts produce positive R^2_{OS} in almost all considered specifications. Moreover, in seven out of nine cases the p-values of the Clark and West (2007) tests do not exceed 10%.[16] The weakest performance is demonstrated by the 3PRF forecast based on 100 book-to-market ratios at the quarterly and monthly frequencies.

The most interesting results are produced by the 3PRFm forecasts. Note that Kelly and Pruitt (2013) choose all predictors (book-to-market ratios) ex ante without conditioning on their statistical significance as individual forecasts. As a result, p-hacking is unlikely to affect their results, so the 3PRF and 3PRFm forecasts should have comparable predictive power. In general, Table 1.5 confirms this intuition. It shows that for 6 and 25 predictors and all data frequencies R^2_{OS} of the 3PRF and 3PRFm forecasts are close to each other and often even coincide. Thus, 3PRFm does not identify p-hacking in the cases with 6 and 25 predictors, and the predictability of returns is likely to be real.

The results are different when the returns are predicted by 100 book-to-market ratios. Table 1.5 shows that in this case R^2_{OS} produced by 3PRFm are negative even though their

[16] Our p-values are different from those in Kelly and Pruitt (2013) because for consistency with the rest of the chapter we report the Clark and West (2007) p-values, whereas Kelly and Pruitt (2013) use the "encompassing" forecast test ENC-NEW derived by Clark and McCracken (2001).

Table 1.5. Out-of-Sample Predictability of Market Returns by the EW, 3PRF, and 3PRFm Combinations of Book-to-Market Ratios

The forecasts combinations are constructed from log book-to-market ratios of protfoios formed on size and book-to-market (6 portfolios, 25 portfolios, 100 portfolios). The sample of the ratios is from January 1930 to December 2015; the out-of-sample period starts in January 1980. R^2_{OS} is the empirical out-of-sample R^2; "p-value" is the p-value of the Clark and West (2007) t-statistic.

Panel A: Annual returns									
	6 portfolios			25 portfolios			100 portfolios		
	EW	3PRF	3PRFm	EW	3PRF	3PRFm	EW	3PRF	3PRFm
R^2_{OS},%	-3.09	5.56	5.56	-2.48	6.58	3.93	-1.69	11.65	-12.72
p-value	0.41	0.07	0.07	0.43	0.07	0.08	0.39	0.06	0.20
Panel B: Quarterly returns									
	6 portfolios			25 portfolios			100 portfolios		
	EW	3PRF	3PRFm	EW	3PRF	3PRFm	EW	3PRF	3PRFm
R^2_{OS},%	-1.58	1.51	1.51	-1.33	0.83	0.83	-0.87	-0.81	-3.99
p-value	0.40	0.08	0.08	0.41	0.10	0.10	0.37	0.17	0.26
Panel C: Monthly returns									
	6 portfolios			25 portfolios			100 portfolios		
	EW	3PRF	3PRFm	EW	3PRF	3PRFm	EW	3PRF	3PRFm
R^2_{OS},%	-0.29	0.56	0.56	-0.24	0.40	0.40	-0.13	0.02	-0.47
p-value	0.39	0.08	0.08	0.39	0.10	0.10	0.33	0.14	0.16

3PRF analogs are positive for annual and monthly returns. There are two possible explanations for this finding. First, it can be caused by lower efficiency of the 3PRFm forecast: as we demonstrate by simulations, the inversion of the predictors' signs make the latent factor to be less precisely estimated, and this hurts the quality of the forecast. Second, negative R^2_{OS} can result from higher sensitivity of the 3PRF and 3PRFm forecasts based on 100 variables to small variations in the sample and forecast construction procedure. This interpretation is justified by relatively poor performance of even 3PRF forecasts for quarterly and monthly returns in our sample.

Finally, the results reported for the EW forecast demonstrate that it fails to predict re-

turns in all specifications. As highlighted by Kelly and Pruitt (2013,0), the 3PRF forecasts reveal the common factor in individual predictors that is related to future returns even though it may not be the most important determinant of the correlations between predictors. As a result, the 3PRF forecast is more efficient than many simpler alternatives, such as the EW forecast, which are contaminated by common factors in predictors unrelated to returns.

Cross-Sectional Predictability of Individual Stock Returns

In this section, we explore how information aggregation interacts with p-hacking in the cross-sectional analysis of individual stock returns. Many results of this section have close analogs in the time series analysis and, therefore, are presented with less details. Nevertheless, the cross-sectional information aggregation techniques are not identical to the forecast combination techniques from the time series analysis and deserve a separate discussion.

Information Aggregation Techniques

The problem of predicting the cross section of stock returns has the following general formulation. At time t an investor observes returns r_{is} on stocks $i = 1,\ldots,N$ and firm characteristics X^a_{is}, $a = 1,\ldots,A$, realized in time periods $s = 1,\ldots,t$. Without losing generality, we assume that each characteristic has been cross-sectionally demeaned and standardized in each period. The characteristics are arguably correlated with expected returns in the cross section, that is, $\mathrm{Cov}(r_{is}, X^a_{is-1}) \neq 0$ for any s and a. In general, the covariances can be either positive or negative. Assuming that the relations between characteristics and re-

turns are stable over time, we also introduce aligned characteristics $\mathscr{X}_{is}^{a}$. By definition, they all are positively related to returns and obtained from X_{is}^{a} by flipping the signs of the characteristics that negatively predict returns.[17] The main problem that we focus on is how to construct a predictor of individual stock returns $\hat{\mu}_{it}$ that aggregates information from all characteristics.

The information aggregation technique that is an analog of 3PRF in the cross-sectional setting is the partial least squares (PLS) regression, and it has been applied to predicting the cross-section of stock returns by Light, Maslov, and Rytchkov (2017) and Jiang, Tang, and Zhou (2018). This technique assumes that unobservable expected stock returns μ_{it} are related to observable firm characteristics X_{it}^{a} through only one or few latent factors. In the simplest case with one latent factor, $X_{it}^{a} = \delta_{t}^{a}(\mu_{it} - \bar{\mu}_{t}) + u_{it}^{a}$, where $\bar{\mu}_{t}$ is the cross-sectional average of expected returns, and the errors u_{it}^{a} can be arbitrarily correlated in the cross section and across characteristics. The PLS estimates of expected returns at time t are obtained in several steps: i) at times $2 \leq s \leq t$ cross-sectionally regress the realized returns r_{is} on each individual lagged characteristic X_{is-1}^{a}, ii) average the obtained slopes λ_{s}^{a} over time to get $\bar{\lambda}_{t}^{a} = \frac{1}{t-1}\sum_{s=2}^{t}\lambda_{s}^{a}$, iii) regress the characteristics X_{it}^{a} of each stock i on $\bar{\lambda}_{t}^{a}$, and iv) use the obtained slopes $\hat{\mu}_{it}$ as proxies for expected returns. We refer to Light, Maslov, and Rytchkov (2017) for further details. Note that $\hat{\mu}_{it}$ estimates expected returns only up to a positive multiplicative factor.

We demonstrate that if all individual time-series predictors forecast returns with the same sign and they all have comparable quality, the 3PRF combination of them is a poor

[17] The sign of the relation between a characteristic and future returns is defined as the sign of the slope in the Fama-MacBeth regression of returns on the lagged characteristic.

predictor of returns. A similar but not identical observation holds for PLS: if all characteristics predict returns with the same sign and comparable precisions, the PLS combination of them is related to returns with an opposite sign. To see the intuition behind this fact, consider a setting with three periods labeled by $t-1$, t, and $t+1$ and with the characteristics X^a_{it-1} and X^a_{it} that have positive Fama-MacBeth regression slopes in the whole sample. The latter condition implies that

$$\frac{\sum_{i=1}^{N}(r_{it}-\bar{r}_t)(X^a_{it-1}-\bar{X}^a_{t-1})}{\sum_{i=1}^{N}(X^a_{it-1}-\bar{X}^a_{t-1})^2}+\frac{\sum_{i=1}^{N}(r_{it+1}-\bar{r}_{t+1})(X^a_{it}-\bar{X}^a_{t})}{\sum_{i=1}^{N}(X^a_{it}-\bar{X}^a_{t})^2}=U^a>0,$$

where the bar denotes the cross-sectional average. Because each characteristic is cross-sectionally demeaned and standardized in each period, this condition reduces to

$$\frac{1}{N}\sum_{i=1}^{N} r_{it}X^a_{it-1}+\frac{1}{N}\sum_{i=1}^{N} r_{it+1}X^a_{it}=U^a>0. \tag{1.3}$$

Recall that the first step of PLS is a cross-sectional regression of returns on individual lagged characteristics, and the obtained slopes at time t are

$$\lambda^a_t=\frac{1}{N}\sum_{i=1}^{N} r_{it}X^a_{it-1},$$

where we again use that the characteristics are standardized. Thus, equation (1.3) implies that $\lambda^a_t+\lambda^a_{t+1}=U^a>0$. The PLS estimates of expected returns at time t are the slopes

from the cross-characteristic regressions of X_{it}^a on λ_t^a, that is,

$$\hat{\mu}_{it} = \frac{\sum_{a=1}^{A}(X_{it}^a - \tilde{X}_{it})\lambda_t^a}{\sum_{a=1}^{A}(\lambda_t^a - \tilde{\lambda}_t)^2},$$

where the tilde denotes the cross-characteristic average. Note that by construction $\sum_{i=1}^{N}\hat{\mu}_{it} = 0$. The predictive ability of $\hat{\mu}_{it}$ is naturally measured by the slope in the Fama-MacBeth regression, which in the case with three periods reduces to the slope in the OLS regression of r_{t+1} on $\hat{\mu}_{it}$. The sign of this slope is determined by the sample covariance $\widehat{\mathrm{Cov}}(r_{it+1}, \hat{\mu}_{it})$:

$$\widehat{\mathrm{Cov}}(r_{it+1}, \hat{\mu}_{it}) = \frac{1}{N}\sum_{i=1}^{N} r_{it+1}\hat{\mu}_{it} = \frac{\frac{1}{N}\sum_{a=1}^{A}\sum_{i=1}^{N} r_{it+1}(X_{it}^a - \tilde{X}_{it})\lambda_t^a}{\sum_{a=1}^{A}(\lambda_t^a - \tilde{\lambda}_t)^2} = \frac{\sum_{a=1}^{A}(\lambda_{t+1}^a - \tilde{\lambda}_{t+1})\lambda_t^a}{\sum_{a=1}^{A}(\lambda_t^a - \tilde{\lambda}_t)^2}$$
$$= \frac{\sum_{a=1}^{A}(U^a - \tilde{U} - \lambda_t^a + \tilde{\lambda}_t)\lambda_t^a}{\sum_{a=1}^{A}(\lambda_t^a - \tilde{\lambda}_t)^2} = -1 + \frac{A\tilde{U}}{\sum_{a=1}^{A}(\lambda_t^a - \tilde{\lambda}_t)^2}\left(\sum_{a=1}^{A} w^a\lambda_t^a - \frac{1}{A}\sum_{a=1}^{A}\lambda_t^a\right), \quad (1.4)$$

where $w^a = U^a / \sum_{a=1}^{A} U^a$. Note that the condition $U^a > 0$ implies that $\tilde{U} > 0$ and $w^a > 0$. When the characteristics have comparable predictive power, $w^a \approx 1/A$ and $\sum_{a=1}^{A} w^a\lambda_t^a - \frac{1}{A}\sum_{a=1}^{A}\lambda_t^a \approx 0$. Therefore, equation (1.4) yields $\widehat{\mathrm{Cov}}(r_{it+1}, \hat{\mu}_{it}) \approx -1 < 0$, so despite a positive relation between individual characteristics and returns, $\hat{\mu}_{it}$ negatively predicts returns. Note that when the relations between the characteristics and future returns can be either positive or negative, for a relatively large number of characteristics $\tilde{\lambda}_t \approx 0$ and $\tilde{\lambda}_{t+1} \approx 0$. Using again equation (1.4), we get $\widehat{\mathrm{Cov}}(r_{it+1}, \hat{\mu}_{it}) \approx \sum_{a=1}^{A}\lambda_{t+1}^a\lambda_t^a / \sum_{a=1}^{A}(\lambda_t^a)^2 > 0$ when λ_t^a and λ_{t+1}^a are positively correlated, that is, when the relations between characteristics and returns are stable over time.

The described property of PLS motivates its modification dubbed PLSm that instead of the original characteristics X_{is}^a uses their aligned versions $\mathscr{X}_{is}^a$. PLSm is a cross-sectional

analog of 3PRFm and, as 3PRFm, it is robust to p-hacking. Indeed, spurious predictors tend to have comparable absolute values of the Fama-MacBeth regression slopes that are concentrated just above the critical value of 2. The inversion of the characteristic signs and aggregation of the characteristics by PLS yields a predictor that is likely to be negatively related to returns. Therefore, the predictability tests do not reject the null when the aggregated characteristics are spurious predictors of returns. Note that the inversion of the characteristic signs does not invalidate the consistency of PLS, so PLSm provides consistent estimates of expected returns under the PLS assumptions.

For comparison, along with PLS and PLSm we also consider several other cross-sectional information aggregation techniques.

Fama-MacBeth regression (FM). The most popular approach to aggregating information about future returns from multiple firm characteristics is to use fitted values from the Fama-MacBeth regression of realized returns on lagged characteristics (e.g., Haugen and Baker 1996; Hanna and Ready 2005; Chan, Dimmock, and Lakonishok 2009; Lewellen 2015; Han, Zhou, and Zhu 2016; Green, Hand, and Zhang 2017). To find the expected return $\hat{\mu}_{it}$ on stock i at time t, the approach prescribes to i) run cross-sectional regressions of realized returns r_{is} on a constant and lagged characteristics X^a_{is-1}, $a = 1,\ldots,A$, in each period $s = 2,\ldots,t$, ii) average the obtained intercepts β^0_s and slopes β^a_s to get $\bar{\beta}^a_t = \frac{1}{t-1}\sum_{s=2}^{t}\beta^a_s$, $a = 0,\ldots,A$, and iii) compute $\hat{\mu}_{it}$ as the fitted value $\hat{\mu}_{it} = \bar{\beta}^0_t + \sum_{a=1}^{A}\bar{\beta}^a_t X^a_{it}$.

Average percentile rank (AvPR). Various versions of this approach are used by Piotroski (2000), Stambaugh, Yu, and Yuan (2015), and Cao and Han (2016), among others. We implement the approach in three steps: i) in each period, sort all stocks according to each

characteristic $\mathscr{X}_{it}^a$, $a = 1,\ldots,A$, ii) find their percentile ranks, and iii) use the averages of stocks' percentile ranks as estimates of expected returns. To ensure that high stock ranks correspond to high expected returns, the stocks are sorted by the aligned characteristics. In contrast to the Fama-MacBeth regression approach, the average percentile ranks predict only cross-sectional rankings of returns, not the returns themselves.

Principal component analysis (PCA). This approach prescribes to use the first principal component of the aligned characteristics as the predictor of returns. To construct estimates of expected returns at time t, we i) find the coefficients λ_s^a of the first principal component of the aligned characteristics $\mathscr{X}_{is}^a$, $i = 1,\ldots,N$, in each month s, $s \leq t$, ii) average them over time to get $\bar{\lambda}_t^a = \frac{1}{t}\sum_{s=1}^t \lambda_s^a$, and iii) compute the predictor as $\hat{\mu}_{it} = \sum_{a=1}^A \bar{\lambda}_t^a \mathscr{X}_{it}^a$. Note that as the average percentile rank approach, PCA uses the aligned characteristics and provides only relative estimates of expected returns.

Simulation Analysis

We use simulations to quantitatively explore the forecasting power of the predictors produced by various aggregation techniques when all individual predictors are obtained by p-hacking. The main steps of our cross-sectional simulation procedure are as follows.

First, we simulate a data set of fictitious returns on $N = 5000$ stocks over $T = 432$ time periods. The number of time periods corresponds to the number of months from January 1980 to December 2015, which is the sample used in our empirical analysis. The chosen number of cross-sectional observations is on par with the average number of firms in Compustat after 1980. The return on stock i at time t is generated as $r_{it} = \mu_i + \varepsilon_{it}$, where

the expected returns μ_i are assumed to be constant over time, and the unexpected returns ε_{it} are i.i.d. across stocks and over time. The expected returns μ_i are drawn from the normal distribution with the mean of 1% and the standard deviation of 0.8%. This calibration is in the ballpark of the empirical estimates of expected monthly returns on individual stocks (e.g., Lewellen 2015). The unexpected returns ε_{it} are assumed to be normally distributed with zero mean and standard deviation of 10%, which is on par with the average monthly volatility of individual stock returns.

As in the time series setting, we generate two types of firms characteristics. The characteristics of the first type are independent from returns, and for each stock in each time period they are obtained as random draws from the standard normal distribution. The characteristics of the second type are spurious predictors of returns generated by p-hacking: we keep drawing candidate characteristics from the multivariate standard normal distribution until the candidate appears to be able to predict the cross section of stock returns in the whole sample. In practice, a characteristic is deemed informative if the slope in the Fama-MacBeth regression of returns on the lagged characteristic is statistically significant or if the characteristic can produce portfolio sorts such that the difference in returns on the top and bottom portfolios is statistically significant. To demonstrate the robustness of our conclusions, we use both approaches. In the former approach, the simulated characteristic is retained if the absolute value of its t-statistic in the Fama-MacBeth regression exceeds 2. In the latter approach, we compute the difference in returns on the top and bottom decile portfolios sorted by the characteristic and retain the characteristic if the absolute value of the t-statistic of the difference is above 2. Because all samples are simulated, the portfolios are equal-weighted. Along with the raw simulated characteristics, we also construct aligned

characteristics by inverting the signs of those characteristics that are negatively related to returns. To avoid the look-ahead bias, we determine the signs of the characteristics at time t using only the data from the periods $s \leq t$.

To explore the forecasting power of return predictors that aggregate information from A characteristics, we use the subsampling approach. For consistency with one of our empirical examples considered, we set $A = 20$ and generate $A_{max} = 100$ characteristics. Then, we randomly draw $B = 100$ sets of A characteristics from the obtained pool and aggregate the characteristics from each set.[18] Finally, we test the ability of the constructed variables to predict returns using two statistics. One of them is the t-statistic of the slope in the Fama-MacBeth regression of returns on the constructed predictor. The other is the t-statistic of the average difference in equal-weighted returns on the top and bottom decile portfolios formed by the aggregate predictor (hedge returns). The t-statistics are random variables whose realizations depend on which characteristics are aggregated and we characterize the distributions of the t-statistics by averages, upper 5% quantiles, and probabilities that the statistic is above 1.645. Note that because the individual predictors are aligned and the aggregate predictors are supposed to be positively related to returns, we conduct one-sided tests with 1.645 as the 5% nominal critical value.

Panel A of Table 1.6 reports the results for the benchmark case with purely random individual characteristics. It shows that for the individual characteristics the probabilities that the t-statistics exceed 1.645 are close to 5%, and the upper 5% quantiles of their distributions are close to 1.645. This is not surprising given that the characteristics have been

[18] The total number of spurious characteristics A_{max} and the number of subsampled sets of characteristics B are smaller than in the time series simulation because the cross-sectional simulations, which involve characteristics and returns of multiple stocks, are more computationally intensive than the time series simulations, which involve only one target variable.

randomly drawn from the standard normal distribution. For the aggregate predictors, the same probabilities and quantiles also cluster around 5% and 1.645, respectively. However, the differences are slightly larger, and this fact admits two non-competing explanations. First, it can be produced by non-linearities in the aggregation techniques and the resulting deviation of the test statistic distribution under the null from the standard normal distribution. Second, the differences can be explained by the limited number of data points in the simulated distributions and the ensuing imprecision of the estimated probabilities and quantiles.

Panels B and C of Table 1.6 provide the information on the t-statistics when the individual characteristics are spurious predictors of returns. They clearly demonstrate that irrespectively from whether p-hacking is based on the Fama-MacBeth regression or portfolio sorts, the PLSm predictors constructed from spurious characteristics are negatively related to future returns (the vast majority of the realized t-statistics are negative) and the t-tests never reject the null of no predictability. This result indicates that p-hacking makes the predictability tests applied to PLSm more conservative. Thus, as 3PRFm in the time series framework, PLSm can be viewed as a tool that mitigates the impact of p-hacking on testing of aggregate predictors in the cross-sectional framework.

The results are conspicuously different for the other aggregation techniques. Panels B and C of Table 1.6 demonstrate that the t-tests applied to them almost always reject the null of no predictability. Moreover, the t-statistics are substantially larger than their analogs for the individual characteristics. Thus, the standard information aggregation techniques amplify the distortions produced by p-hacking by making the type I error of the predictability tests even larger than it is for individual spurious characteristics.

Table 1.6. Simulated Performance of Various Cross-Sectional Information Aggregation Techniques

Each aggregation techniques combines 20 predictors. The column Ind reports the statistics of individual aligned predictors. The simulated sample contains 5000 stocks and 432 time periods. The test statistics are the t-statistics of the Fama-MacBeth regression slopes (FM slopes) and t-statistics of the difference in returns on the top and bottom decile portfolios (Hedge returns). "Mean(t)" and "$q_{0.05}(t)$" are the averages and upper 5% quantiles of the t-statistics; "Prob($t > 1.645$)" is the empirical probability that the t-statistic exceeds 1.645.

		Ind	FM	AvPR	PCA	PLS	PLSm
Panel A: Random predictors							
FM slopes	Mean(t)	0.09	-0.14	0.35	0.05	-0.21	-0.23
	Prob($t > 1.645$),%	5.00	3.00	8.00	8.00	3.00	1.00
	$q_{0.05}(t)$	1.68	1.12	1.88	1.69	1.14	1.13
Hedge returns	Mean(t)	0.04	-0.16	0.17	-0.01	-0.16	-0.26
	Prob($t > 1.645$),%	7.00	3.00	8.00	5.00	3.00	1.00
	$q_{0.05}(t)$	1.75	1.33	1.74	1.72	1.21	1.21
Panel B: Spurious predictors preselected by Fama-MacBeth regressions							
		Ind	FM	AvPR	PCA	PLS	PLSm
FM slopes	Mean(t)	2.01	7.85	8.77	7.37	7.73	-1.94
	Prob($t > 1.645$),%	74.00	100.00	100.00	100.00	100.00	0.00
	$q_{0.05}(t)$	2.93	8.77	9.66	8.76	8.62	-0.81
Hedge returns	Mean(t)	1.56	6.11	6.92	5.82	6.01	-1.52
	Prob($t > 1.645$),%	45.00	100.00	100.00	100.00	100.00	0.00
	$q_{0.05}(t)$	2.71	7.08	8.27	7.16	7.31	-0.32
Panel C: Spurious predictors preselected by portfolio sorts							
		Ind	FM	AvPR	PCA	PLS	PLSm
FM slopes	Mean(t)	1.54	5.31	6.14	5.66	5.21	-1.28
	Prob($t > 1.645$),%	44.00	100.00	100.00	100.00	100.00	0.00
	$q_{0.05}(t)$	2.76	6.85	7.18	6.80	6.58	0.00
Hedge returns	Mean(t)	2.04	3.96	4.64	4.51	3.92	-1.05
	Prob($t > 1.645$),%	80.00	99.00	100.00	100.00	100.00	0.00
	$q_{0.05}(t)$	2.82	5.37	5.93	5.96	5.48	0.44

Table 1.6 also shows the upper 5% quantiles of the simulated t-statistics. They are the lowest values of the t-statistics for which the null hypothesis of no predictability can be rejected at the 5% level when there is a possibility that all individual characteristics are selected by p-hacking. For the individual characteristics the quantiles are close to 3, which is the critical value recommended by Harvey, Liu, and Zhu (2016) as a safeguard against p-hacking. In contrast, for the aggregate predictors the quantiles are much higher. In particular, for all information aggregation techniques except for PLSm the quantiles of the Fama-MacBeth t-statistics are above 8 (above 6) when the individual characteristics are preselected by the Fama-MacBeth regression (by portfolio sorts). Thus, for the aggregate predictors the bar is much higher than for the individual characteristics, and the use of 3 as a critical value of the t-statistics for combinations of characteristics is likely to result in incorrect test sizes and unreliable inference.

Example: Aggregation of Information from Real Firm Characteristics

As an empirical example, we consider the aggregation of information from firm characteristics that have been previously identified in the academic literature as cross-sectional predictors of individual stock returns. Although such aggregation is conducted in many studies (e.g., Ou and Penman 1989; Lev and Thiagarajan 1993; Haugen and Baker 1996; Abarbanell and Bushee 1998; Piotroski 2000; Hanna and Ready 2005; Lewellen 2015; Green, Hand, and Zhang 2017; and Light, Maslov, and Rytchkov 2017), the out-of-sample forecasting power of the aggregate predictor can be lower than indicated by the standard in-sample tests because at least some characteristics could have been preselected by p-

hacking.[19]

As standard in the literature, our sample includes only NYSE, AMEX, and NASDAQ firms with common stocks and non-missing market values of equity. The characteristics that we consider are chosen from the pool of 102 variables listed in Green, Hand, and Zhang (2013,0) and computed from the data available in CRSP, Compustat, and I/B/E/S.[20] To ensure that the vast majority of the characteristics are available for at least several firms, we start the sample in January 1980. Following Green, Hand, and Zhang (2017), we drop 8 characteristics that are strongly related to the others and may cause the multicollinearity problem in the Fama-MacBeth regression. Also, we exclude 4 characteristics that are missing for all firms in the beginning of the sample. Using the notations of Green, Hand, and Zhang (2017), the full list of considered 90 variables includes *absacc*, *acc*, *aeavol*, *age*, *agr*, *baspread*, *beta*, *bm*, *bm_ia*, *cash*, *cashdebt*, *cashpr*, cfp, cfp_ia, *chatoia*, *chcsho*, *chempia*, $chfeps$, *chinv*, *chmom*, *chnanalyst*, *chpmia*, *chtx*, *cinvest*, *convind*, *currat*, $depr$, *disp*, *divi*, *divo*, *dy*, *ear*, *egr*, ep, *gma*, *grCAPX*, *grltnoa*, $herf$, *hire*, *idiovol*, *ill*, *indmom*, *invest*, *IPO*, lev, $mom1m$, $mom12m$, $mom36m$, *ms*, *mve*, *mve_ia*, *nanalyst*, *nincr*, $operprof$, *orgcap*, *pchcapx_ia*, *pchcurrat*, *pchdepr*, *pchgm_pchsale*, *pchsale_pchinvt*, *pchsale_pchrect*, *pchsale_pchxsga*, *pchsaleinv*, *pctacc*, *pricedelay*, *ps*, *rd*, *rd_mve*, *rd_sale*, *retvol*, *roaq*, *roavol*, *roeq*, *roic*, *rsup*, *salecash*, *saleinv*, *salerec*, sfe, *sgr*, *sin*, *SP*, *std_dolvol*, *std_turn*, $stdcf$, *sue*, $tang$, tb, *turn*, and *zerotrade*. In each month, all characteristics are winsorized at 1% and 99% of their cross-sectional distributions and standardized to have zero mean

[19]The validity of the latter concern is confirmed by Linnainmaa and Roberts (2018), who extend the sample of anomaly-related characteristics backward in time and demonstrate that many of them fail to predict returns in the earlier periods.

[20]We are grateful to Jeremiah Green for providing the SAS code that constructs the characteristics from Green, Hand, and Zhang (2017) on his web page https://sites.google.com/site/jeremiahrgreenacctg/home.

and unit standard deviation.

Using the standard techniques, we construct five predictors of individual stock returns that aggregate information from multiple firm characteristics. To ensure that all characteristics aggregated by the Fama-MacBeth regression and PCA are available for all firms, we assign zeros (the cross-sectional average of standardized characteristics) to all missing values. As in the simulation analysis, the predictors incorporate only the information available in real time, so our analysis does not suffer from the look-ahead bias. To test the quality of the obtained aggregate predictors, we use the slopes of Fama-MacBeth regressions of returns on the predictors as well as the differences in returns on the top and bottom decile portfolios formed by the predictors. In the portfolio approach, we consider both equal-weighted and value-weighted portfolios because the strength of return predictability is known to be negatively related to firm size (e.g., Fama and French 2008).

Panel A of Table 1.7 shows the performance of various predictors constructed from all 90 characteristics. Overall, the results demonstrate that all information aggregation techniques except PCA consistently produce statistically significant predictors of returns, although the strength of the predictability varies across the techniques. For example, the Fama-MacBeth fitted values and PLS-based predictors have the highest regression slopes, hedge returns, and t-statistics. The hedge returns produced by the Fama-MacBeth fitted values are higher than their analogs in Green, Hand, and Zhang (2017), and this result is likely to be explained by differences in portfolio construction (all-stocks decile breakpoints vs. NYSE decile breakpoints) and sample periods (1980–2015 vs. 1990–2014). A relatively poor performance of PCA is consistent with Light, Maslov, and Rytchkov (2017), who explain it by the existence of common factors in characteristics that are unrelated to

future returns. Also, as expected, the value-weighted hedge returns produced by all techniques except PCA are smaller than their equal-weighted counterparts.

Table 1.7. Cross-Sectional Predictability of Individual Stock Returns

"Slopes" and "t-stats" are slopes and t-statistics from Fama-MacBeth regressions of monthly returns on the aggregate predictors. Hedge returns are the differences in monthly returns on top and bottom decile portfolios formed by the aggregate predictors, and "t-stats" are their t-statistics. The portfolios are either equal-weighted (EW) or value-weighted (VW). The slopes and hedge returns are reported in percentage points. The sample period is from January 1980 to December 2015.

Panel A: 90 characteristics						
		FM	AvPR	PCA	PLS	PLSm
FM slopes	Slopes	0.98	0.51	0.24	0.74	0.60
	t-stats	14.98	5.52	1.92	7.15	6.65
Hedge returns	EW	3.71	1.77	0.60	2.60	2.16
	t-stats	16.09	5.18	1.31	7.20	6.72
	VW	1.50	1.27	1.12	1.73	1.26
	t-stats	5.29	3.61	2.60	4.89	4.16
Panel B: 21 characteristics						
		FM	AvPR	PCA	PLS	PLSm
FM slopes	Slopes	0.79	0.48	0.03	0.69	0.49
	t-stats	10.73	5.25	0.26	6.19	5.37
Hedge returns	EW	2.91	1.77	-0.07	2.51	1.88
	t-stats	11.62	5.63	-0.16	6.64	5.60
	VW	1.45	1.01	0.45	1.91	1.47
	t-stats	5.22	3.17	0.98	5.49	4.46

Most interestingly, Panel A of Table 1.7 shows that the Fama-MacBeth regression slopes and hedge returns produced by the PLSm predictors are positive and statistically significant, although they are slightly smaller than their PLS analogs. For example, the value-weighted PLSm hedge returns are 1.26% (with t-statistic 4.16), whereas the value-weighted hedge returns on the PLS predictor are 1.73% (with t-statistic 4.89). Because the aggregation of spurious predictors by PLSm produces a variable that typically is negatively related to

future returns, this result provides evidence that at least several aggregated predictors are likely to be real and unaffected by p-hacking.

Along with the set of 90 characteristics, we also consider a reduced set of 21 characteristics that largely correspond to the characteristics used in Light, Maslov, and Rytchkov (2017). This set includes size (*mve*), book-to-market ratio (*bm*), earnings-to-price ratio (ep), cash flow-to-price ratio (cfp), market beta ($beta$), momentum ($mom12m$), long-term reversal ($mom36m$), short-term reversal ($mom1m$), idiosyncratic volatility (*idiovol*), maximum daily return over the past month (*maxret*), total asset growth (*agr*), corporate investment (*cinvest*), growth in capital expenditures (*grCAPX*), capital expenditures and inventory (*invest*), accruals (*acc*), change in shares outstanding (*chcsho*), returns on equity (*roeq*), gross profitability (*gma*), leverage (lev), turnover ($turn$), and analysts' forecasts dispersion ($disp$). We aggregate those characteristics by the same techniques as before and check whether the conclusions of Light, Maslov, and Rytchkov (2017) can be attributed to p-hacking.

Overall, the Fama-MacBeth regression slopes and hedge returns produced by predictors combining 21 characteristics and reported in Panel B of Table 1.7 are typically smaller than their counterparts in the case with 90 characteristics. Thus, the decrease in the number of characteristics slightly worsens the performance of the aggregate predictors. Nevertheless, the differences are small, and all techniques except PCA still produce highly statistically significant Fama-MacBeth regression slopes and hedge returns. The magnitudes of hedge returns are consistent with those in Light, Maslov, and Rytchkov (2017). As before, the most important result in Panel B of Table 1.7 is provided by PLSm. As in the case with 90 characteristics, the PLSm-based Fama-MacBeth regression slopes and hedge returns

are smaller than their PLS analogs but still quite large and highly statistically significant. This result indicates that the predictability of stock returns cannot be fully explained by p-hacking.

Conclusion

This chapter conveys two main messages. First, academics and practitioners should be alert when backtesting the predictors of returns (either in cross section or time series) that aggregate pre-selected variables because the outcome can be affected by p-hacking far beyond its impact on the individual predictors. Second, p-hacking can be confronted not only by careful characterization of the universe of potential predictors from which a few successful predictors have been selected but also by using the proposed modification of 3PRF/PLS. Thus, our study highlights both the dark and bright sides of aggregation. On the one hand, the aggregation can be dangerous because it further distorts the sizes of the tests in the presence of p-hacking. On the other hand, simultaneous consideration of multiple predictors helps to alleviate the impact of p-hacking because it opens up the way to decrease the contribution of the predictors that are likely to be selected by p-hacking.

Unfortunately, there is no perfect safeguard against p-hacking, and 3PRFm/PLSm can also be gamed. For example, being aware of the tendency of 3PRFm/PLSm to implicitly penalize predictors with comparable forecasting power, a p-hacker can start aggregating spurious predictors of diverse quality. Nevertheless, 3PRFm/PLSm sets a higher bar for potential predictors and makes p-hacking more difficult. In particular, the diversity in the predictor quality can be achieved either by adding predictors with t-statistics that substan-

tially exceed 2 or that are well below 2. In the first case, the p-hacker should try a larger number of potential predictors because it is much harder to get higher t-statistics by chance. In the second case, the addition of weak predictors dilutes the power of aggregate forecasts and adversely affects their performance in backtesting.

Our study suggests several directions for future research. For example, it could be interesting to further explore the finite sample properties of the 3PRFm/PLSm forecasts in various settings. In particular, it would be insightful to quantify the loss in the efficiency of the 3PRFm/PLSm forecasts compared to the 3PRF/PLS forecasts and investigate the sensitivity of our conclusions to distributional assumptions. Also, in addition to the empirical example presented in the chapter, there are likely to exist many other situations in which the success of 3PRFm/PLSm predictors can indicate that at least several aggregated predictors are not spurious, and we leave the search for such applications of 3PRFm/PLSm to future work.

CHAPTER 2

DO CHARACTERISTICS-BASED ASSET PRICING FACTORS REPRESENT MACROECONOMIC SHOCKS? A MACHINE LEARNING ANALYSIS

Introduction

One of the most important and challenging objectives of asset pricing is the construction of models that can explain the cross section of expected stock returns. The most successful models are characteristic-based models whose factors are returns on trading strategies based on firm characteristics. The most famous model of this type is the Fama-French 3-factor model (Fama and French 1993); more recent examples include the Fama-French 5-factor model (Fama and French 2015), q-factor model (Hou, Xue, and Zhang 2015), and behavioral 3-factor model (Daniel, Hirshleifer, and Sun 2018). The factors in the latest generation of characteristic-based asset pricing models combine information from multiple firm characteristics (e.g., Stambaugh and Yuan 2017, Kozak, Nagel, and Santosh 2017a,0, Kelly, Pruitt, and Su 2018). Although such models are useful for summarizing the main patterns in the cross section of expected stock returns, they have an important limitation: their factors often lack the economic interpretation. In particular, it remains unclear whether expected stock returns are determined by exposures of stock returns to macroeconomic risks or investor sentiment (e.g., Nagel 2013).

The objective of this chapter is to shed new light on characteristic-based asset pricing

models by exploring whether their factors represent macroeconomic shocks. Although my study is not the first attempt in the literature to link characteristic-based factors to macroeconomic risks, it differs from other studies in several important respects. First, my analysis involves a large dataset that includes 131 macroeconomic variables. The simultaneous consideration of a large number of variables associated with different macroeconomic activities decreases the probability that a macroeconomic factor that determines asset prices is omitted from the analysis. In particular, the inability of all macroeconomic shocks to approximate the asset pricing factors would suggest that asset prices cannot be fully explained by exposures to macroeconomic risks.

The standard projection of asset pricing factors onto a large number of macroeconomic variables is likely to produce misleading results because of overfitting. Therefore, I use an elastic net regression, a machine learning technique, to identify the linear combination of macroeconomic factors that is relevant to asset pricing factors not only in sample but also out of sample. The measures of fit produced by the elastic net regression fitted values allow me to robustly quantify the explanatory power of macroeconomic shocks. The application of machine learning techniques to the analysis of economic interpretability of asset pricing factors is a unique contribution of my study to the literature.

The next distinguishing feature of my study is the focus on the models' stochastic discount factors (SDFs) instead of the individual characteristic-related factors. Because all factors in the characteristic-based models are returns on tradable portfolios, the discount factor that is constructed from them is also tradable and uniquely characterizes the model (e.g., Cochrane 2005). In contrast, the decomposition of the SDF into individual factors such as SMB and HML is not unique, and it is possible to find another set of factors with

even different number of factors that would represent exactly the same SDF. The factors used to describe the existing models are likely outcomes of historical accident and may not individually correspond to macroeconomic shocks. Therefore, the macroeconomic content of the SDF might be obscured by how the SDF is decomposed into the individual factors. By interpreting the SDF instead of its constituent components, I avoid the ambiguity in the macroeconomic interpretation of the model and ensure that the results are not affected by the chosen decomposition of the SDF into the individual factors.

Finally, my study is the first to consider five characteristic-based asset pricing models simultaneously, and is thus the most comprehensive study of such models to date. The simultaneous consideration of the models allows me not only to study their individual macroeconomic interpretability but also to compare their macroeconomic content and rank the models accordingly.

I start the analysis by constructing the SDFs of five characteristic-based asset pricing models. The models that I consider are the Fama-French three- and five-factor models developed in Fama and French (1993) and Fama and French (2015), respectively, the q-theory model of Hou, Xue, and Zhang (2015), the mispricing four-factor model constructed by Stambaugh and Yuan (2017), and the behavioral finance three-factor model proposed by Daniel, Hirshleifer, and Sun (2018). For each model, I estimate its SDF assuming that it is a linear combination of the individual factors. Because the characteristic-based factors are tradable, the coefficients of the SDFs, which are often associated with the prices of risk, are uniquely identified under the assumption that the model correctly prices returns on the factor portfolios. This is the weakest requirement on the model that must hold even if the model cannot correctly price all assets in the market. The obtained SDFs summarize the

pricing ability of each model, and I use them as inputs in my empirical analysis.

Next, I build a comprehensive set of macroeconomic shocks. For this purpose, I use 131 macroeconomic variables from FRED-MD, which is a large publicly available macroeconomic data set collected by the Federal Reserve Bank of St. Louis and described in McCracken and Ng (2015). FRED-MD contains information on a broad spectrum of economic activities, and it is designed to be a standard database for the "big data" macroeconomic research. In the main analysis, I construct macroeconomic shocks as first differences of the variables that characterize the state of the economy. In a robustness test, I consider shocks constructed as residuals of the AR(1) processes fitted for each macroeconomic variable.

Because the pricing content of a model is summarized by its SDF, the macroeconomic interpretation of a model can be obtained by regressing its SDF on macroeconomic shocks. In particular, the macroeconomic interpretability of a model is naturally measured by the R^2 of that regression. This procedure is straightforward and easily applicable when the number of considered macroeconomic shocks is small. However, OLS regression works poorly when the number of explanatory variables is large because it overfits the data in the sample but the obtained fitted values cannot predict the realizations of the dependent variable out of sample. In my context, the OLS regression would indicate that SDFs are almost perfectly replicated by macroeconomic shocks in sample, but this relation would likely be spurious and break down out of sample. To confront the overfitting problem, I use an elastic net regression, which is one of the supervised machine learning techniques.[1] The idea of this approach is to impose a regularization on the linear regression parameters that constrains them in sample thereby mitigating the overfitting problem. Practically, the

[1] Hastie, Tibshirani, and Friedman (2016) provides a textbook exposition of machine learning techniques.

elastic net estimates of the linear regression parameters are obtained by minimizing the sum of squared errors augmented with a convex combination of the sum of the parameter absolute values and the sum of the squared parameters. Thus, an elastic net estimation combines the elements of two other popular techniques — the Lasso (it would have only the first regularization term) and the ridge regression (it would have only the second regularization term). Intuitively, the Lasso part promotes the sparsity of the model, whereas the ridge regression part shrinks the coefficients of correlated variables toward each other so that the information contained in them is used more efficiently. To find the regularization parameters ("tune the algorithm"), I use cross-validation, which is another standard tool in machine learning.

To measure the ability of the elastic net regression fitted values to track SDFs, I use the out-of-sample R^2. In particular, to find the mean squared error for the realization of the SDF at time t I use the fitted value whose construction does not involve the time-t observations. Note that by itself R^2 is a statistic whose realizations depend on the sample. To obtain the standard errors of R^2 and conduct statistical inference, I implement the bootstrap procedure that randomly draws time periods with replacement and recomputes the out-of-sample R^2 for all simulated samples.

Results demonstrate that the SDFs of the five characteristic-based factor models do not fully represent macroeconomic shocks: the obtained out-of-sample R^2 are in the range from 3% to 20%. Interestingly, the models vary in terms of their macroeconomic content: the models that arguably admit a risk-based interpretation such as the Fama-French three- and five-factor models have substantially higher out-of-sample R^2 than the behavioral models such as the mispricing four-factor model of Stambaugh and Yuan (2017) and the three-

factor model of Daniel, Hirshleifer, and Sun (2018) (20% and 15%, respectively, versus 5% and 3%, respectively). The differences in the out-of-sample R^2 are not only large but also statistically significant at the 5% level. This result suggests that the behavioral factors and sentiment are important determinants of asset prices.

Next, I examine whether all macroeconomic shocks are needed to maximize the out-of-sample R^2 and find that the relation between macroeconomic shocks and SDF is sparse. In particular, three shocks, the innovations in the log industrial productivity, log dividend yield on S&P 500, and term spread, produce an out-of-sample R^2 that is comparable to the out-of-sample R^2 produced by all 131 macroeconomic shocks. Notably, the three shocks represent macroeconomic fundamentals, equity market, and bond market.

My approach can not only quantify the relation between SDFs and macroeconomic shocks but also inform whether a particular shock provides additional information about the SDF compared to all other macroeconomic shocks. In particular, I can test whether specific theory-motivated factors contain unique information about SDFs. As an illustration, I consider an intermediary asset pricing model (e.g., Adrian, Etula, and Muir 2014, He, Kelly, and Manela 2017) and investigate the explanatory power of a shock to the intermediary capital ratio. I find that the intermediary asset pricing factors explain the variation in SDFs with the out-of-sample R^2 that ranges from 4% to 30%. Nevertheless, the inclusion of 131 macroeconomic shocks significantly increases the out-of-sample R^2 for all characteristic-base factor models, so the intermediary asset pricing factors do not capture all pricing information contained in the macroeconomic factors. In contrast, adding intermediary asset pricing factors to 131 macroeconomic shocks increases the out-of-sample R^2 only in the case of the Fama-French three-factor model.

This chapter is closely related to the literature that aims to economically interpret characteristic-based asset pricing factors. Those studies are largely motivated by the Intertemporal CAPM (ICAPM) of Merton (1973), which predicts that the priced factors should be innovations in the state variables that describe investment opportunities. For example, Liew and Vassalou (2000) document that HML and SMB predict GDP growth; Vassalou (2003) shows that news about future GDP growth subsume the pricing power of the HML and SMB factors; Petkova (2006) relate HML and SMB to innovations in several variables that predict investment opportunities. The conclusions regarding the momentum factor are more controversial. Griffin, Ji, and Martin (2003) demonstrate that momentum returns are unrelated to macroeconomic factors in 17 markets. In contrast, Liu and Zhang (2008) find that the growth rate of industrial production explains more than half of momentum profits. Aretz, Bartram, and Pope (2010) extend that line of research by examining HML, SMB, and momentum factors simultaneously and considering a broader set of six macroeconomic variables. Maio and Santa-Clara (2012) adopt a more general stance on the interpretability of asset pricing models and argue that many multifactor models do not satisfy the restrictions imposed by the ICAPM. That conclusion is challenged by Boons (2016). My study differs from those papers in three ways. First, my objective is to interpret the SDFs of the models, not the individual factors. Second, I simultaneously consider several models including the recent ones whose relations to macroeconomic factors have not been studied at all. Third, I use a much larger set of macroeconomic shocks to explain the asset pricing factors.

My study also belongs to the rapidly growing strand of literature that applies machine learning techniques to various asset pricing problems. In particular, machine learning can

be used to extract information about future returns from a large number of variables and construct optimal portfolios. For example, Freyberger, Neuhierl, and Weber (2018) use grouped Lasso to select firm characteristics that provide incremental information about the cross section of stock returns and combine them nonparametrically. Gu, Kelly, and Xiu (2018) discuss a large variety of machine learning techniques and how they can be used for measuring conditional risk premia on the market and individual stocks. Chinco, Clark-Joseph, and Ye (2017) apply machine learning to improve the predictability of high-frequency returns. DeMiguel, Martin-Utrera, Nogales, and Uppal (2018) exploit the ability of Lasso to identify characteristics that are incrementally important to the construction of optimal portfolios in the presence of transaction costs. Another set of papers apply machine learning to construction of asset pricing models. For example, Kozak, Nagel, and Santosh (2017a) propose an estimation of SDF that incorporates information from a large number of firm characteristics and features using a regularization that is similar to the elastic net regularization. Feng, Giglio, and Xiu (2017) demonstrate how a double-selection Lasso can be used to evaluate the contribution of a new asset pricing factor relative to a high-dimensional set of existing factors. Bryzgalova (2016) develops shrinkage-based estimators of linear factor models that are robust to the presence of spurious factors. Rapach and Zhou (2019) apply unsupervised machine learning on a large number of macro variables to estimate sparse principal components (PCs), and find the selected macroeconomic variables to earn a significant risk premium. Notably, my study is distinct from these papers in two ways. First, I use machine learning techniques to interpret existing asset pricing models, not to propose new models or predict returns. Second, I extract information from macroeconomic shocks, not from firm characteristics or empirical risk factors.

Empirical Approach

This section describes machine learning, with a particular focus on the elastic net with cross-validation. Next, I present the construction of SDFs for a class of characteristic-based factor models. These SDFs are the basis of my empirical investigation.

Machine Learning

Machine learning describes a class of algorithms whose objective is to construct statistical predictions from a large set of predictors.[2] For the purpose of prediction, on the one hand, a high-dimensional model with large number of predictors may be favored over a sparsely parameterized model. This is because a high-dimensional model can lower the probability of omitting the relevant predictors and thus provides a better approximation for the dependent variable of interest. This advantage is especially important when the data generating process underlying the dependent variable is unknown and complex. On the other hand, a high-dimensional model is subject to overfitting in a finite sample: it demonstrates high predicting power in training a model by adjusting to the sample specific patterns, and thus fails to predict the dependent variable reliably in a new dataset. The essence of machine learning is that it manages to fit a high-dimensional model to the data without overfitting.[3]

Machine learning relies on regularization and empirical tuning techniques to prevent overfitting. Regularization mitigates the overfitting problem by constraining the complex-

[2] I use the term "prediction" to describe the process of using x to uncover contemporaneous y as if the value of y is unobserved

[3] More precisely, the machine learning described here is the supervised machine learning. It works in a context where there is a dependent variable y and the goal is to use some independent variables x to predict y. There is another branch of machine learning that is irrelevant to my study, unsupervised machine learning, which works in a context where only x is available and the goal is to characterize or categorize x. Principal component analysis (PCA) is an example of unsupervised machine learning that is most familiar to economists.

ity of a model. The standard way to regularize the model is to augment the loss function, which quantifies the forecast error, with an additional term that will increase with the complexity of a model. Absent regularization, machine learning collapses to a traditional prediction technique. Supplementally, empirical tuning helps identify the optimal level of regularization such that the out-of-sample predictive power of the model is maximized. More specifically, it experiments with a large grid of regularization levels, and picks the one that causes the fitted model to generate the lowest expected loss out-of-sample. Empirical tuning guards against overfitting by emphasizing stable out-of-sample performance to pick the degree of regularization and the predictive model.

Elastic Net

The machine learning technique I use belongs to the class of linear regressions with regularization. Before getting into details, I introduce the following notations.

y_t: the dependent variable of interest with T observations.

$x_{i,t}$: predictive variables indexed by $i = 0, ..., q$. Here $x_{0,t}$ denotes a constant term. q is large and can be greater than T.

β_i (respectively β_i^*): parameters of the model (the true value of the parameters).

$\mathscr{L}$: minimized objective function.

F: predictive model.

$\mathbb{R}[\beta]$: the regularization function of parameters.

In order to predict y_t, linear regressions with regularization estimate a high-dimensional model of the type, $F[x_{i,t}] = \beta_0 + \sum_{i=1}^{q} \beta_i x_{i,t}$, by minimizing an objective function that con-

sists of the sample mean-squared error (MSE) and a term of regularization:

$$\mathscr{L}[\beta;\lambda] = \frac{1}{T}\sum_{t=1}^{T}(y_t - \beta_0 - \sum_{i=1}^{q}\beta_i x_{i,t})^2 + \lambda\,\mathbb{R}[\beta].$$

Here $\mathbb{R}[\beta]$ denotes the regularization, whose level is determined by $\lambda \in (0, +\infty]$. For any given λ, the model is estimated as $\widehat{F}_\lambda = \arg\min_{\{\beta\}} \mathscr{L}[\beta;\lambda]$.

The difference in objective functions between a regularized regression and standard OLS is the presence of the regularization term $\lambda\,\mathbb{R}[\beta]$, which generally increases with model flexibility.

To examine the relations between SDFs and macroeconomic factors, I use a machine learning technique known as elastic net (Zou and Hastie 2005). Essentially, it is a linear regression with regularization that penalizes a large number of non-zero regression slopes as well as substantial heterogeneity among them. Denote the realization of the target variable at time t as yt (it is m_t in my application) and the realizations of N individual predictors as $x_{i,t}$, i = 1,..., q (they are macroeconomic shocks in my application). Each $x_{i,t}$ is assumed to be demeaned and standardized.

The OLS-based predictor of y_t by contemporaneous $x_{i,t}$ is constructed by running an OLS regression of y_t on $x_{i,t}$ using the training sample of size T, obtaining the intercept $\hat{\beta}_0$ and slopes $\hat{\beta}_i$, and computing the fitted values $\hat{y}_t = \hat{\beta}_0 + \sum_{i=1}^{q} x_{i,t}\hat{\beta}_i$. Unfortunately, OLS-based prediction suffers from serious overfitting when q is large, and breaks down when q $>$ T. The elastic net modifies OLS-based prediction by augmenting the loss function of

OLS regression, which is the mean-squared error (MSE), by a regularization term:

$$\mathscr{L}[\beta;\lambda,\alpha] = \underbrace{\frac{1}{T}\sum_{t=1}^{T}(y_t - \beta_0 - \sum_{i=1}^{q}\beta_i x_{i,t})^2}_{\text{mean squared errors}} + \underbrace{\lambda\alpha\sum_{i=1}^{q}|\beta_i|}_{\text{L}^1\text{ penalty}} + \underbrace{\lambda(1-\alpha)\sum_{i=1}^{q}\beta_i^2}_{\text{L}^2\text{ penalty}}. \tag{2.1}$$

Elastic net incorporates both L^1-norm and L^2-norm regularizations, and assigns weights to them with respect to $\alpha \in (0,1]$. The overall level of regularization is governed by λ.

An L^1 penalty sets $\lambda\mathbb{R}[\beta] = \lambda\sum_{i=1}^{q}|\beta_i|$ to imposes an upper limit on the sum of absolute value of slope coefficient. It translates each slope coefficient by $\lambda/2$ towards zero, truncating at zero. Thus, zero coefficients are assigned to weak predictors. As a result, such a decrease in the number of estimated parameters effectively suppresses the variance of estimators at a cost of biasness. Lasso, a linear regression with L^1 penalty, is useful for variable selection and has already found many applications in economics and finance. However, Lasso is likely to make model-selection mistakes in finite samples, particularly when the dimensions of the true model are not sparse (e.g. Kozak, Nagel, and Santosh 2017a) or the predictors are highly correlated (e.g. Tibshirani 1996, Zou and Hastie 2005).[4]

An L^2 penalty sets $\lambda\mathbb{R}[\beta] = \lambda\sum_{i=1}^{q}\beta_i^2$ to impose an upper limit on the sum of squares of slope coefficients. It shrinks all parameters towards zero (and each other) by $1/(1+\lambda)$, without dropping any variable. As a result, the decrease in the magnitudes of estimated parameters suppresses the variance of estimators at a cost of biasness. Ridge is the linear regression with L^2 penalty. It effectively exploits the linear combination of the correlated variables to cancel out the idiosyncratic noise and isolate the common signals. However,

[4]As explained in Zou and Hastie (2005), Lasso completely ignores correlations between predictors and tends to pick one variable and disregard the rest.

Ridge is unable to generate a parsimonious predictive model even when sparsity exists.

There are several reasons why an elastic net is an appropriate tool for my analysis of discount factors. First, in contrast to more flexible machine learning techniques such as regression trees, random forests, and neural networks, it predicts the target variable by linear combinations of individual predictors. Therefore, it preserves the structure of linear factor models. Second, macroeconomic shocks are correlated within several groups, and the ability of an elastic net to efficiently aggregate information from correlated predictors, which is inherited from the ridge regression, ensures that I do not underestimate the ability of macroeconomic shocks to explain discount factors. Third, many macroeconomic shocks are likely irrelevant to asset prices, and the ability of elastic net to discard weak predictors, which is inherited from Lasso, likely reduces the amount of noise in the constructed predictors.

Empirical Tuning

In an elastic net, two regularization parameters, λ and α, need to be optimally chosen by empirical tuning. Cross-validation is an efficient and widely used tuning method. Since various regularizations are compared based on the out-of-sample prediction performance, the main idea of cross-validation is to use a part of the whole sample for model estimation (training) and reserve the other part of the sample, that mimics the "out-of-sample," for model testing (validation).

I use a A ten-fold cross-validation in this chapter. The figure below illustrates the

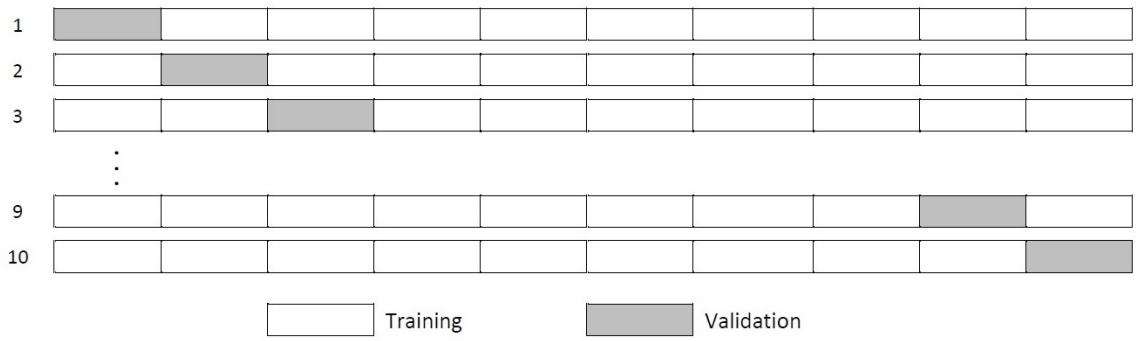

Figure 2.1. Ten-Fold Cross-Validation

process.[5] I split the whole sample into ten parts, shown as the ten equal-sized boxes in a row. Then I take a grid of regularization parameters (λ, α), and iterate the following two-step procedure for $k = 1, 2, ..., 10$. In the k^{th} round (the k^{th} row of the figure), (i) reserve the k^{th} subsample as validation sample, and estimate the models by minimizing all possible loss functions on the rest nine training subsamples (denoted by $-k$): $\widehat{F}^{-k}_{(\lambda,\alpha)} = \arg\min_{\{\beta\}} \mathscr{L}^{-k}[\beta; \lambda, \alpha]$; and (ii) test the estimated models on the k^{th} validation subsample and record the prediction errors as $y_t - \widehat{F}^{-k}_{(\lambda,\alpha)}[x_t]$, $\forall t \in k$.

After the iterations, I have constructed the prediction errors (of all models) over the whole sample of $t = 1, 2, .., T$. These prediction errors are essentially out-of-sample errors, because when any observation is used for testing a model, that observation is left out of the sample for model estimation. The out-of-sample MSE of cross-validation is thus constructed as:

$$MSE^{CV}_{(\lambda,\alpha)} = \frac{1}{T}\sum_{t=1}^{T}(y_t - \widehat{F}^{-k[t]}_{(\lambda,\alpha)}[x_t])^2.$$

Here, $k[t]$ refers to the part of sample that period t belongs to, while $-k[t]$ refers to other nine parts that period t is absent from. The optimal regularization parameters are the (λ, α)

[5]The decision about the number of folds involves the typical trade-off between variance and bias. 5- or 10-fold cross-validation are recommended as a good compromise (Tibshirani 1996).

associated with the lowest out-of-sample MSE and are denoted by (λ^*, α^*).

Once the optimal regularization parameters are found, the loss function and ensuing predictive model are fixed accordingly. Note that since the cross-validation procedure splits the sample randomly, some randomness is brought to the realization of the optimal regularization parameters. Though this randomness leads to variation in model specifications, the predictive power of the model stays stable.

Stochastic Discount Factors

Under the law of one price, there exists a stochastic discount factor m_t such that for excess returns on any asset R_t^e,

$$E(m_t R_t^e) = 0. \tag{2.2}$$

Any asset pricing model is characterized by its SDF, which in general is specified as a function of data and several parameters: $m_t = f(\text{data, parameters})$ (Cochrane 2005). I start my analysis by constructing SDFs of the characteristic-based factor models. Such models are linear factor models whose SDFs have the following specification:

$$m_t = 1 - b'(f_t - \mathbb{E}(f_t)). \tag{2.3}$$

Here f_t is a $p \times 1$ vector of factors formed as excess returns on trading strategies based on firm characteristics and b is a $p \times 1$ vector of slope coefficients known as a vector of risk prices of the factors f_t. Since risk-free rate is not described by a characteristic-based factor model, the mean of the empirical SDF remains unidentified and without loss of generality is normalized to 1.

In order to construct the SDF, risk prices b should be carefully estimated. Since characteristic-based factors f_t are excess returns on tradable portfolios, they can simultaneously serve as the pricing factors of a model and as the assets whose returns must be explained by the model. Therefore, I implement GMM estimation on the moment conditions specified by equation 2.2, using the risk factors f_t as the only test assets R_t^e. The estimated risk prices are essentially:

$$b = \mathbb{E}\left((f_t - \mathbb{E}(f_t))(f_t - \mathbb{E}(f_t))'\right)^{-1}\mathbb{E}(f_t). \tag{2.4}$$

Therefore, the SDFs of characteristic-based factor models can be estimated as

$$\hat{m}_t = 1 - \left(\left(\frac{1}{T}\sum_{t=1}^{T}(f_t - \bar{f})(f_t - \bar{f})'\right)^{-1}\bar{f}\right)'(f_t - \bar{f}), \tag{2.5}$$

where $\bar{f}$ is the sample average of factor realizations. Because all factors are returns on tradable assets, $\hat{m}_t$ is also a return on a tradable asset. Note that although there exist an infinite number of SDFs that identically price all tradable assets in incomplete markets, the projection of any SDF on the space of tradable assets is unique and coincides with m_t. I also estimate b and $\hat{m}_t$ by a GMM estimation with an enlarged set of test assets, and discuss the robustness of the findings.

Description of Models and Macroeconomic Shocks

In this study, I use monthly data on characteristic-based risk factors and macroeconomic shocks from January 1963 to December 2015. This section describes the data source and variable constructions.

Characteristic-Based Asset Pricing Models

Consistent with the treatment in equation (2.3), my empirical work considers the following five characteristic-based factor models:

$$\text{FF3}: \; R^e_{i,t} = \alpha_i + \beta_{i,m}\text{MKT}_t + \beta_{i,s}\text{SMB}_t + \beta_{i,h}\text{HML}_t + u_{i,t}$$

$$\text{FF5}: \; R^e_{i,t} = \alpha_i + \beta_{i,m}\text{MKT}_t + \beta_{i,s}\text{SMB}_t + \beta_{i,h}\text{HML}_t + \beta_{i,r}\text{RMW}_t + \beta_{i,c}\text{CMA}_t + u_{i,t}$$

$$\text{Q4}: \; R^e_{i,t} = \alpha_i + \beta_{i,m}\text{MKT}_t + \beta_{i,d}\text{ME}_t + \beta_{i,k}\text{IA}_t + \beta_{i,g}\text{ROE}_t + u_{i,t}$$

$$\text{M4}: \; R^e_{i,t} = \alpha_i + \beta_{i,m}\text{MKT}_t + \beta_{i,s}\text{SMB}_t + \beta_{i,q}\text{MGMT}_t + \beta_{i,p}\text{PERM}_t + u_{i,t}$$

$$\text{BF3}: \; R^e_{i,t} = \alpha_i + \beta_{i,m}\text{MKT}_t + \beta_{i,f}\text{FIN}_t + \beta_{i,e}\text{PEAD}_t + u_{i,t}.$$

The construction of all the characteristic-based factors are listed below in alphabetical order:

CMA investment factor. This factor is constructed using the 2-by-3 value-weighted portfolios formed on size and investment. It is the average returns on the two conservative investment portfolios minus the average returns on the two aggressive investment portfolios

FIN long-term financing factor. This factor is constructed using the 2-by-3 value-weighted portfolios formed on size and financing characteristics. It is the average returns on the two high financing characteristics portfolios minus the the average returns on the two low financing characteristics portfolios.

HML value factor. This factor is constructed using the value-weighted returns on 18 port-

folios that are sorted 2-by-3-by-3 independently based on size, investment-to-asset, and return-on-equity. It is the average returns on the three small portfolios minus the average returns on the three big portfolios;

IA is an investment factor. This factor is constructed using the 2-by-3-by-3 value-weighted portfolios formed on size, investment-to-assets, and return-on-equity. It is the average returns on the six low investment-to-asset portfolios minus average returns on the six high investment-to-asset portfolios.

ME size factor. This factor is constructed using the value-weighted returns on 18 portfolios that are sorted 2-by-3-by-3 independently based on size, investment-to-asset, and return-on-equity. It is the average returns on the nine small size portfolios minus the average returns on the nine big portfolios.

MGMT management anomaly factor. This factor is constructed using the 2-by-3 value-weighted portfolios formed on size and firms' average ranking based on a cluster of six asset pricing anomalies that represent quantities that firms' managements can affect directly. It is the average returns on the two high management anomalies portfolios minus the average returns on the two low management anomalies portfolios.

MKT market factor. It is the monthly excess returns on the market;

PEAD short-term post earning announcement drift factor. This factor constructed using the 2-by-2 value-weighted portfolios formed on size and the sign of post earnings announcement drift. It is the returns on the two positive post earnings announcement drift portfolios minus the average returns on the two negative post earnings announcement drift portfolios.

PERM performance anomaly factor. This factor is constructed using the 2-by-3 value-weighted portfolios formed on size and firms' average ranking based on a cluster of five asset pricing anomalies that represent quantities that are related to firm performance. It is the average returns on the two high performance anomalies portfolios minus the average returns on the two low performance anomalies portfolios.

ROE profitability factor. This factor is constructed using value-weighted returns on 18 portfolios that are sorted 2-by-3-by-3 independently based on size, investment-to-asset, and return-on-equity. It is the average returns on the six high return-on-equity portfolios minus the average returns on the six low return-on-equity portfolios.

RMW profitability factor. This factor is constructed using the 2-by-3 value-weighted portfolios formed on size and operating profitability. It is the average returns on the two robust operating profitability portfolios minus the average returns on the two weak operating profitability portfolios;

SMB size factor. This factor is constructed using the 2-by-3 value-weighted portfolios formed on size and book-to-market ratio. It is the average returns on the three small portfolios minus the average returns on the three big portfolios;

Data on the factors in FF3 and FF5 are taken from Ken French's data library. Data on the factors in M4 come from Robert Stambaugh's website. Data on the factors in Q4 and BF3 are generously provided by the authors. I use the intersection of the data for all the factors, and consider a sample period from January 1963 to December 2015 for my empirical analysis.

Utilizing the characteristic-based risk factors of the models, I estimate the SDFs fol-

lowing equation (2.5) and present the output in Table 2.1. Panel A presents the estimated risk prices $\hat{b}$ of factors for each model. Panel B presents the annualized standard deviations of the five SDFs FF3 has the least volatile SDF and M4 has most volatile SDF. The mean of $\hat{m}_t$ are unity by construction, and are omitted from the table. Panel C presents the pairwise correlations between SDFs of different models. The five series of SDFs are positively correlated with each other. FF5 and Q4 has the highest correlation, that is, 0.75. Such a high correlation is reasonable, since both models the market, size, profitability and investment factors, though the constructions of the factors are different. FF3 and BF3 have the lowest correlation, that is, 0.27. Such a low correlation is also understandable, considering that FF3 is created to mimic systematic risks, however BF3 is purely motivated by the systematic mispricing of stock market.

In addition, I plot the time series of the SDFs in Figure 2.2, where the the gray shaded areas correspond to the business recessions. All SDFs are counter cyclical, which is consistent with the intuition that required rate of returns are higher over recessionary periods.

Macroeconomic Shocks

To build a comprehensive set of macroeconomic shocks, I use 131 macroeconomic variables from FRED-MD, and then construct the unexpected innovations of them.

FRED-MD is a large publicly available macroeconomic data set collected by the Federal Reserve Bank of St. Louis and described in McCracken and Ng (2015). This data set contains information on a broad spectrum of economic activities and incorporates eight categories of variables: (1) output and incomes, (2) labor market, (3) housing, (4) con-

Table 2.1. Risks Prices and SDFs for Five Characteristic-Based Factor Models

The five characteristics-based asset pricing models are FF3 (Fama and French 1993), FF5 (Fama and French 2015), Q4 (Hou, Xue, and Zhang 2015), M4 (Stambaugh and Yuan 2017) and BF3 (Daniel, Hirshleifer, and Sun 2018). Panel A presents the GMM estimations of risk prices $\hat{b}$ and corresponding standard errors (*s.e.*) for each factor in each model, where the factors are also used as test assets. Panel B presents the annualized standard deviations of the estimated SDFs ($\hat{m}_t = 1 - \hat{b}' f_t$). Panel C presents the pairwise correlations between SDFs of the five characteristic-based factor models. The sample is from 01/1963 to 12/2015. Factors of FF5 are missing from 01/1963 to 06/1963; factors of Q4 are missing from 01/1963 to 12/1966; factors of BF3 are missing from 01/1963 to 06/1972 and in 2015.

Panel A: Prices of risks

	FF3		FF5		Q4		M4		BF3	
	$\hat{b}$	*s.e.*	$\hat{b}$	*s.e.*	$\hat{b}$	*s.e.*	$\hat{b}$	*s.e.*	$\hat{b}$	*s.e.*
MKT	3.24	(1.07)	5.16	(1.18)	5.62	(1.22)	8.50	(1.44)	7.86	(1.54)
SMB	2.15	(1.37)	4.15	(1.52)			7.31	(1.64)		
HML	6.43	(1.54)	-0.17	(2.20)						
RMW			9.68	(2.18)						
CMA			13.07	(3.17)						
ME					5.79	(1.71)				
IA					17.72	(2.92)				
ROE					12.62	(2.16)				
MGMT							16.41	(2.20)		
PERF							7.90	(1.54)		
FIN									10.41	(1.61)
PEAD									21.99	(2.82)

Panel B: Annualized standard deviations of SDFs

	FF3	FF5	Q4	M4	BF3
SD	0.21	0.31	0.44	0.54	0.52

Panel C: Correlations of SDFs

	FF3	FF5	Q4	M4	BF3
FF3	1.00				
FF5	0.62	1.00			
Q4	0.42	0.75	1.00		
M4	0.42	0.70	0.64	1.00	
BF3	0.27	0.38	0.45	0.54	1.00

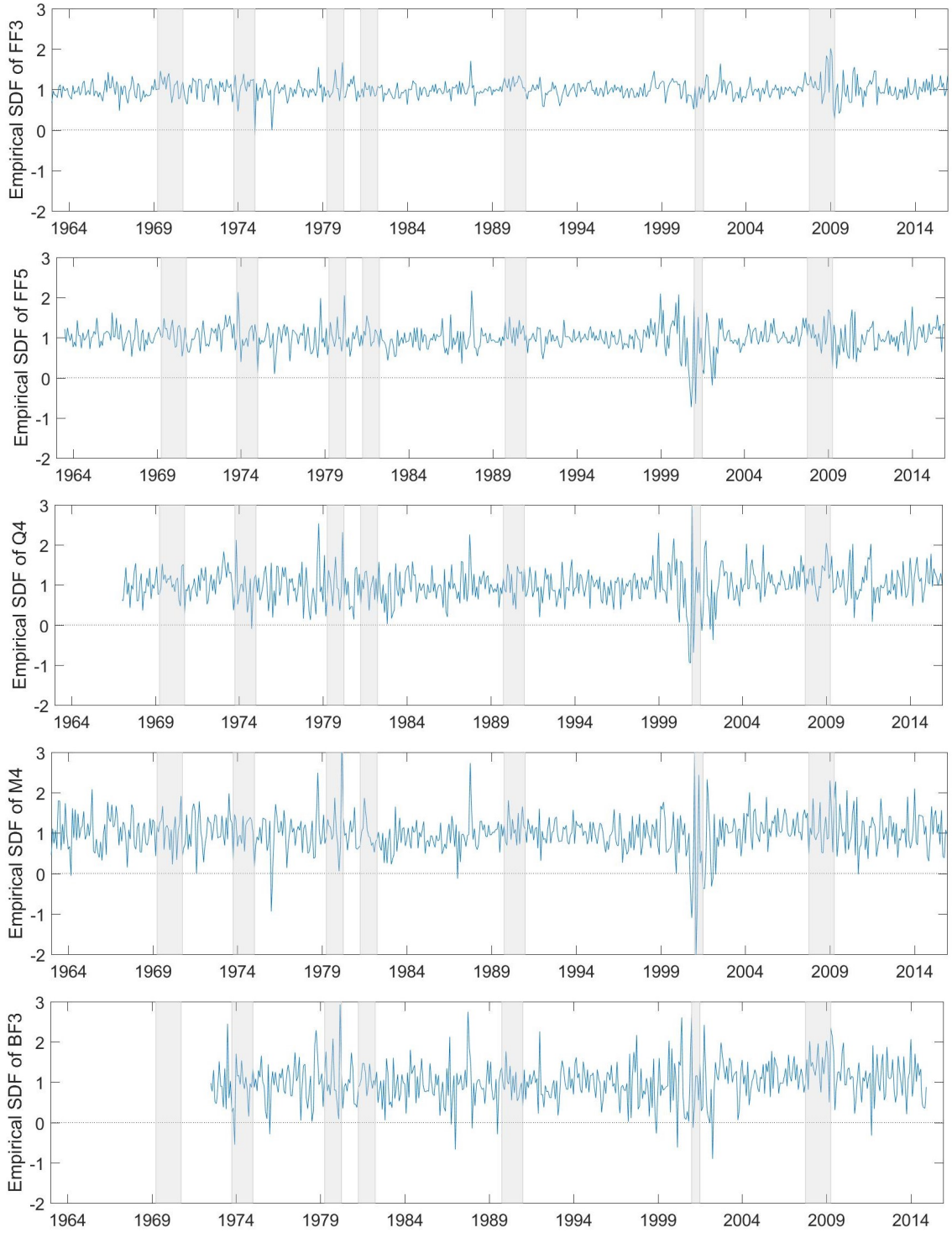

Figure 2.2. Estimated Monthly SDFs for Five Characteristic-Based Factor Models

The five characteristic-based factor models are FF3, FF5, M4, Q4, and BF3. The gray shaded areas correspond to business recessions. The sample is from January 1963 to December 2015.

sumption, orders, and inventories, (5) money and credit, (6) interest and exchange rates, (7) prices, and (8) stock market. FRED-MD is also designed to be a standard data base for the "big data" macroeconomic research.

It shares the same predictive content for economic aggregates as various vintages (e.g., Stock-Waston dataset) that have already been widely used for "big data" research in finance and economics (McCracken and Ng 2015).[6] FRED-MD was first published in March 2015 and has been updated on a monthly basis since then. I use the data vintage updated in May 2016, in which 135 macro series are provided.[7] I form a balanced panel of 131 series dating from January 1963 to December 2015 by dropping four series. If any observation of a series lies ten times the interquartile range away from the median, it is defined as an outlier and is replaced by the truncated value.

Macroeconomic shocks are constructed as unexpected innovations of the 131 macroeconomic series. Unexpected innovations are realizations minus expectations, however, the latter is unobservable and can be estimated in different ways. Since the macroeconomic series is almost stationary, I use its lagged values to proxy the expectations and construct the unexpected innovations as the first-order differences of the macroeconomic series (Maio and Santa-Clara 2012, Aretz, Bartram, and Pope 2010, Chen, Roll, and Ross 1986). I present the names, short descriptions, and data transformations for all macroeconomic shocks in Table A1. Alternatively, one may assume that investors form expectations from the information on all available macroeconomic data and employ a VAR or AR model to approximate the expectations (Boons 2016). In one of my robustness tests, I construct

[6]The data is available at both monthly and quarterly frequency. They can be downloaded from the website https://research.stlouisfed.org/econ/mccracken/fred-databases/

[7]For different vintages, variables were removed or suspended for publish for agreement issues with the data source. See the document titled "fredmdchange" provided by authors for detail.

macroeconomic shocks as the residuals of AR(1) regressions of macroeconomic series.

The set of macroeconomic shocks I build contains a small number of conventional shocks that are frequently studied in empirical asset pricing literatures, as well as many other shocks. For instance, the log consumption growth (Breeden, Gibbons, and Litzenberger 1989), log industrial productivity (Liu and Zhang 2008), and log oil prices (Chen, Roll, and Ross 1986) are the conventional shocks to economic fundamentals and are respectively the series 60, 3, and 110 of my collection (see Table A1). Moreover the innovations in dividend yield of S&P500 index, price-to-earnings ratio of the S&P500 index, term spread, default spread, one-month Treasury Bill rate, and inflations (Boons 2016, Maio and Santa-Clara 2012) are the classic shocks to economic state variables, that predict future investment opportunities, and are respectively series 129, 130, 99, 100, 85, and 113 in my collection.

In Figure 2.4, I show the pairwise unsigned correlations of the macroeconomic shocks in a heat map. On the heat map, the macroeconomic shocks are placed in ascending order of their series numbers (see Table A1) on both axis. Each small square on the heat map corresponds to the absolute value of correlations between two macroeconomic shocks. Higher absolute values of correlations are represented by dark gray or black squares and lower absolute value of correlations are presented by lighter squares.

Note that the small squares on the diagonal line represent correlations of macroeconomic shocks with themselves, which always have the value of one and are colored black. The first observation is that there are several clusters of dark gray areas along the diagonal line, which implies that macroeconomic shocks from the same categories are highly correlated with each other. The second observation is that most of the of-diagonal areas of the heat

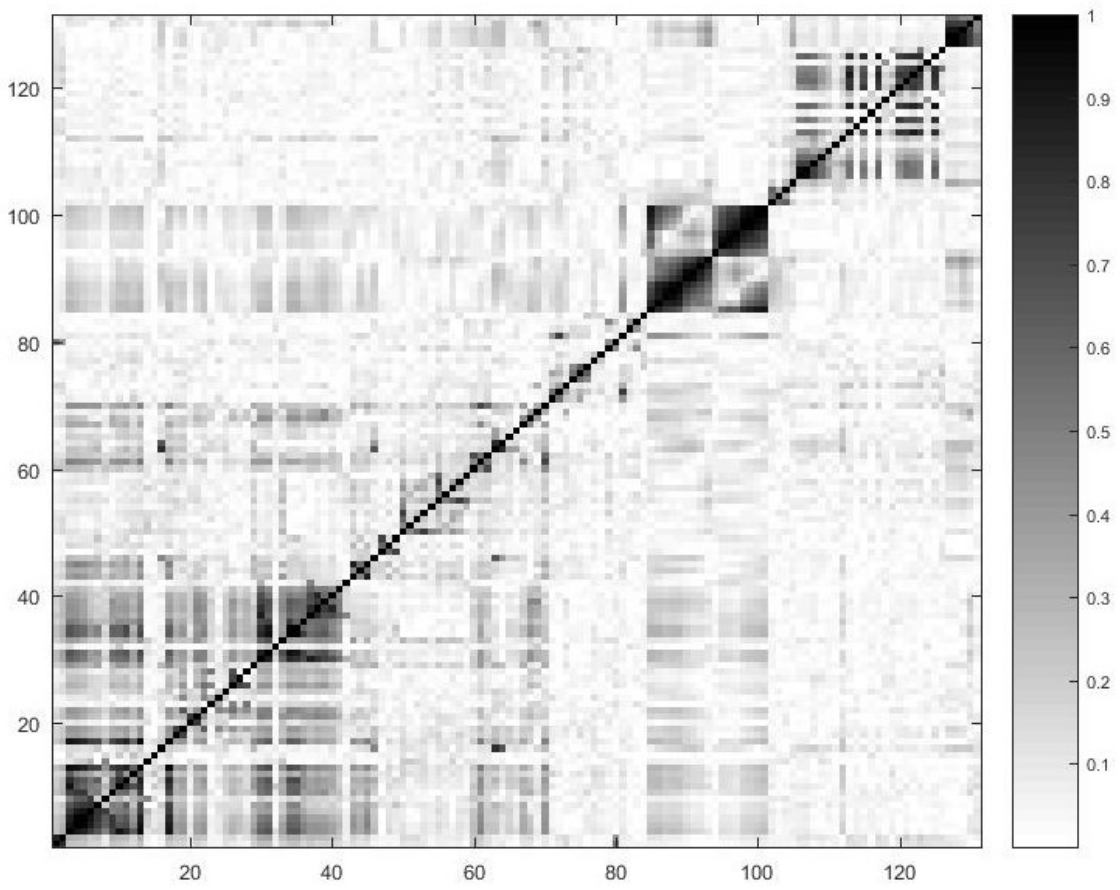

Figure 2.4. Heat Map of Correlations Between Macroeconomic Shocks

The darker color represents a higher correlation. Macroeconomic shocks are constructed as the first-order differences of 131 macroeconomic variables from FRED-MD. In the figure, each macroeconomic shock is represented by its series number, which is listed in Table A1 together with additional information about the name, short description and data transformation. The sample is from January 1963 to December 2015.

map is white or light gray, which implies that correlations between macroeconomic shocks

from different categories are small.

Do Characteristic-based Factor Models Admit Macroeconomic Interpretations?

The main objective of this section is to investigate to what extent the SDFs of characteristic-based factor models are explained by a comprehensive set of macroeconomic shocks. Complementing the main test, I also explore whether the quantified explanatory power can be attributed to a parsimonious set of macroeconomic shocks, as well as the impact of augmenting the original linear predictive model with a large number of nonlinear terms of macroeconomic shocks on explaining SDFs.

The performance metric I use is out-of-sample R^2, denoted by R^2_{os}. It quantifies the ability of macroeconomic shocks to explain the variation in SDFs. R^2_{os} has the following formation:

$$R^2_{os} = 1 - \frac{\sum_{t=1}^{T}(\hat{m}_t - \hat{F}^{-k[t]}[x_t])^2}{\sum_{t=1}^{T}(\hat{m}_t - 1)^2}, \tag{2.6}$$

where $\hat{m}_t$ denotes the estimated SDF and is the dependent variable of interest, and x_t is a vector of macroeconomic shocks. As described 2, the whole sample is split into ten parts, and $-k[t]$ denotes nine parts that period t is absent from.

The numerator of equation (2.6), is the sum of squared out-of-sample prediction errors. Note that, the prediction error for the realization of the SDF at time-t is found by using the fitted value of $\hat{F}^{-k[t]}$, whose construction does not involve the time-t observations. The denominator of equation (2.6) is the variance of the SDF. Using R^2_{os} instead of in-sample R^2 avoids overstating of the macroeconomic interpretability of the SDFs.

By itself R^2_{os} is a statistic that depends on the particular realization of the sample. Be-

cause of the limited number of observations in sample, the computed R^2_{os} night be higher or lower than it would be in an infinite sample. To assess the errors in estimating R^2_{os}, I need B realizations of it from randomized samples, where B is sufficiently large to ensure the accuracy of the statistical inference. I start by simulating $B = 1000$ samples of the SDFs and macroeconomic shocks by randomly drawing periods from the original sample with replacement. Each simulated sample has the same size with the original sample. Note that the original sample of three SDFs (i.e., FF5, Q4, and BF3) contains missing observations. I require the number of missing observations of each simulated SDF to be less than 20% of the total number of observations. Next, I repeat the regression of SDFs on macroeconomic shocks for each of the simulate sample, and construct B R^2_{os} following equation (2.6). In the rest of the chapter, I use $R^2_{os^*}$ and $R^2_{os,B}$ to respectively denote the R^2_{os} obtained from original sample and from bootstrapped sample.

Elastic Net Regression of SDFs on 131 Macroeconomic Shocks

I regress the SDFs on 131 macroeconomic shocks using an elastic net, and quantify the macroeconomic content of the model by R^2_{os} described in equation (2.6). When using an elastic net, the predictive model $\widehat{F}$ is essentially the $\widehat{F}_{(\lambda^*,\alpha^*)}$. Thus, the numerator in equation (2.6) is actually $T \times MSE^{CV}_{(\lambda^*,\alpha^*)}$, that is, the sum of minimum out-of-sample MSE produced by cross-validation.

In the process of bootstrapping the distribution of R^2_{os}, the randomized samples brings about two sources of randomness to the realizations of R^2_{os}: randomness produced by the realizations of optimal regularization parameters (λ^*, α^*), and randomness produced by the fitting the model F into randomized sample under fixed (λ^*, α^*). I eliminate the first

source of randomness to reduce the computational intensity, because repeating ten-fold cross-validation for 1000 simulated samples can be computationally demanding.[8] Particularly, I apply elastic net regression with ten-fold cross-validation for once in the original sample, and fixed down the (λ^*, α^*). Then I construct R^2_{os} for each bootstrap as if (λ^*, α^*) is given.

The output of regressing five SDFs on 131 macroeconomic shocks are presented in Panel A of Table 2.2 which delivers several observations.

First, all five characteristic-based factor models that I consider impart some macroeconomic interpretations. The $R^2_{os^*}$ based on original sample is 20%, 15%, 6%, 5%, 3% respectively for FF3, FF5, Q4, M4, and BF3. The mean and standard deviations of the bootstrapped $R^2_{os,B}$ are presented as "Mean($R^2_{os,B}$)" and "Std($R^2_{os,B}$)" in the table. "Prob($R^2_{os,B} < 0$)" presents the probability of $R^2_{os,B}$ falling below zero, and are smaller than 1% for all models. This indicates that the null hypothesis of no macroeconomic interpretability is rejected at 1% level for each characteristic-based factor model.

Second, the SDFs of the five characteristic-based factor models do not fully represent macroeconomic shocks. Under the null hypothesis that an SDF is fully explained by macroeconomic shocks, R^2_{os} should be equal to 100%. However, because of measurement errors in estimating SDF and constructing macroeconomic shocks, a standard test will likely reject the null even when it is true. Although a formal hypothesis test is not appli-

[8] To identify the optimal regularization parameters, one should search a wide range of possible values to avoid finding a corner solution or local minimum. At the same time, the density of possible numbers over the range should be sufficiently high to ensure a good precision of optimization. In my setting, I consider searching over all combinations of 100 values of possible λ and 11 values of possible α. Upon that, each model needs to be estimated at 10 different training samples in a 10-fold cross-validation. Therefore the algorithm of elastic net has to be implemented 11000 times to accomplish a model training procedure for one SDF in a single bootstrapped sample. Considering that I need to repeat the process for five SDFs and 1000 bootstrapped sample, and the resultant computational demand is extremely high.

Table 2.2. Elastic Net Regressions of SDFs on 131 Macroeconomic Shocks

SDFs of five characteristic-based factor models: FF3, FF5, M4, Q4, and BF3, are the dependent variables. Macroeconomic shocks are estimated as the first-order differences of 131 macroeconomic variables and are the explanatory variables. Panel A presents the out-of-sample R^2 (R^2_{os}) of the regressions. Particularly, "$R^2_{os^*}$" denotes the R^2_{os} obtained from original sample. "Mean($R^2_{os,B}$)" and "Std($R^2_{os,B}$)" represent the mean and standard deviation, respectively, of the R^2_{os} that are estimated on the 1000 bootstrapped samples. "Prob($R^2_{os,B} < 0$)" represents the probability that bootstrapped R^2_{os} falls below zero. Panel B presents the comparisons of R^2_{os} between different characteristic-based factor models. Particularly, "Prob($R^2_{os,B}(X) > R^2_{os,B}(Y)$)" represents the probability that the bootstrapped R^2_{os} of a model X (listed in the first column) is greater than that of a model Y (listed in the first row).The sample is from 01/1963 to 12/2015.

Panel A: R^2_{os}					
	FF3	FF5	Q4	M4	BF3
$R^2_{os^*}$, %	20.02	14.97	6.20	5.47	3.34
Mean($R^2_{os,B}$), %	24.09	19.10	9.74	11.80	8.78
Std($R^2_{os,B}$), %	4.49	3.83	2.88	3.87	3.85
Prob($R^2_{os,B} < 0$), %	0.00	0.00	0.00	0.00	0.60
Panel B: Prob($R^2_{os,B}(X) > R^2_{os,B}(Y)$), %					
Y / X	FF3	FF5	Q4	M4	BF3
FF3	-				
FF5	12.70	-			
Q4	0.10	0.10	-		
M4	1.80	3.50	71.60	-	
BF3	0.50	1.50	39.90	24.10	-

cable without quantifying these measurement errors, it is safe to reject the null considering the extremely small R^2_{os} that I obtain. Particularly, $R^2_{os^*}$ are in the range from 3% to 20%. Moreover, according to the simulations, the R^2_{os} that are two standard deviations above the mean are still small, which range from 12% to 28.6%.

Third, the five characteristic-based factor models vary in terms of their macroeconomic content. I observe that $R^2_{os^*}$ are substantially higher for FF3 and FF5 than M4 and BF3, that is 20% and 15% versus 5% and 3%. This is consistent with the economic motivations of the models. Particularly, the FF3 and FF5 are intended to capture rational systematic

risks and I show that they are rooted in macroeconomic interpretations. On the other hand, M4 and BF3 are proposed to capture systematic mispricing of the stock market and I found them to be less relevant to macroeconomic shocks.

I also conduct statistical comparisons on R^2_{os} for the five models, and present the outcome in Panel B of Table 2.2. The statistic for comparison is "Prob$(R^2_{os,B}(X) > R^2_{os,B}(Y))$", which denotes the probability that $R^2_{os,B}$ of any model X is greater than $R^2_{os,B}$ of another model Y. For instance, the 12.70% at row two and column one shows that $R^2_{os,B}$ of FF5 is 12.70% of the chance greater than that of FF3. I ignore the comparison of models to themselves, therefore the off-diagonal cells of the table is empty. Note that, if Prob$(R^2_{os,B}(X) > R^2_{os,B}(Y))$ is smaller than 5%, the null hypothesis that model X is more macroeconomically interpretable than model Y is rejected at 5% level. Panel B confirms that, FF3 and FF5 are better explained by macroeconomic shocks than the other four models at 5% significance level. The comparisons on R^2_{os} between FF3 and FF5, or among the other three models (i.e., Q4, M4, and BF3) are not statistically significant.

The takeaway is that, while characteristic-based factor models perform well in explaining cross-sectional stock returns and asset pricing anomalies, only a small proportion (maximum of 20%) of their SDFs are explained by macroeconomic shocks. I attribute the rest of the variation in the SDFs to behavioral explanations. Additionally, five characteristic-based factor models are not equally explained by macroeconomic shocks. The models with arguable risk-based interpretations admit more macroeconomic interpretations, while the models that are completely built on market mispricing are less interpretable by macroeconomic shocks.

Explain SDFs with Nine Classic Macroeconomic Shocks

Knowing that 131 macroeconomic shocks explain 3% to 20% of the variation in the models' SDF, I further examine whether they are all needed to maximize the R^2_{os}. Particularly, I consider a parsimonious subset of macroeconomic shocks that are conventionally used in the literature, and measure their ability to track the variation in SDFs.

There are nine classic macroeconomic shocks that are frequently used in the literature of empirical asset pricing. They are (1) growth of log consumption ΔC, (2) growth of log oil prices ΔOP, (3) growth of log industrial productivity ΔIP, (4) innovations to dividend yield of S&P 500, DY, (5) innovations to log price-earnings ratio of S&P 500, PE, (6) innovations to term structure spread TERM, (7) innovations to default spread DS, (8) innovations to one-month Treasury Bill rate RF, and (9) innovations to inflations UI.[9] The first three are conventionally interpreted as shocks on economic fundamentals. The other six are often interpreted as innovations in economic state variables that can predict future investment opportunities.

In this specification, the procedure of implementing regressions and constructing R^2_{os} are different from the procedure in the baseline specification with 131 macroeconomic shocks. When the number of regressors is sufficiently small, machine learning does not have a comparative advantage over standard prediction techniques. Therefore, an OLS regression is adopted in place of an elastic net. Note that a standard OLS does not involve cross-

[9]The definition of TERM and DS in my collection of macroeconomic shocks are different from the conventional definitions. In Chen, Roll, and Ross (1986) and Maio and Santa-Clara (2012), TERM is the yield spread between 10- and 1-year Treasury bond, and DS is the yield spread between BAA and AAA corporate bonds from Moody's. In my collection, TERM is the difference between the ten-year Treasury bond rate and the one-month Federal Fund Rate, and DS is the difference between AAA corporate bonds and the Federal Fund rate.

validation, and I need to obtain the numerator of the R^2_{os} in equation (2.6) by mimicking a cross-validation procedure. Specifically, I split the whole sample into ten parts, then iteratively estimate the OLS on each $9/10$ of data, and construct the squared out-of-sample prediction errors on last tenth.

The results of regressing five SDFs on the nine classic macroeconomic shocks are presented in Table 2.3. Panel A describes the relative importance of each shock in explaining the variation of SDFs. I measure the relative importance of any shock as the rate of decrease in in-sample R^2 when this shock is dropped from the regression of SDFs on the nine shocks.[10] In Panel A, I place the nine classic shocks in descending order of their relative importance. I observe that three factors have substantially higher relative importance than the rests: TERM, DY, and ΔIP. More interestingly, this observations implies that all the characteristic-based factor models sufficiently approximate shocks that represent bond market, equity market, and macroeconomic fundamentals.

Instead of simply regressing SDFs on the full set of nine classic shocks, I run nine different regressions by sequentially adding one shock at a time, following the descending order of their relative importance. For instance, the first regression is a univariate regression of SDF on the most important macroeconomic shock, TERM; the last regression is a regression of SDF on the full set of nine classic shocks. Panel B of Table 2.3 presents the R^2_{os} of the nine regressions.

There are three major observations. First, the "3-shock" regressions, that is, regressing SDFs on TERM, DY, and ΔIP, generate comparable R^2_{os} with the "9-shock" regressions, in

[10] I do not normalize the relative importance measure, therefore they do not sum up to 1. Note that, any 8-shock model is nested in the full model of 9 shocks, the relative importance is non-negative.

Table 2.3. SDFs and Nine Classic Macroeconomic Shocks

SDFs of five characteristic-based factor models: FF3, FF5, M4, Q4, and BF3, are the dependent variables. A subset of nine classic macroeconomic shocks are the explanatory variables. Panel A presents the relative importance of each shock in explaining the SDFs, which is measured as the rate of decrease in in-sample R^2 when the shock is excluded from the regression. The nine macroeconomic shocks are placed in an descending order of their average relative importance across the five SDFs. Panel B presents the R^2_{os} of nine regressions, which consecutively add one shock at a time along the descending order of the shocks' relative importance. For instance, the "3-shock" regression represents regressions of SDFs on TERM, DY, and ΔIP. Presented in Panel C are the estimated coefficients of the nine shocks (left panel) and the corresponding t-values (right panel). The sample is from 06/1963 to 12/2015.

Panel A: Relative importance (%)						Panel B: R^2_{os}, %					
shocks	FF3	FF5	Q4	M4	BF3	# of shocks	FF3	FF5	Q4	M4	BF3
TERM	0.02	24.11	34.41	42.05	28.32	1	15.71	13.87	5.75	4.23	1.21
DY	21.58	9.74	21.02	0.19	12.67	2	19.56	13.96	6.06	5.30	0.77
ΔIP	31.50	6.26	11.78	0.58	6.88	3	25.34	15.45	6.84	5.15	0.81
PE	0.89	4.42	2.67	12.23	0.00	4	23.24	15.05	7.67	6.07	-0.11
RF	6.12	0.18	0.87	0.02	1.91	5	24.80	15.96	6.35	6.29	-2.42
DS	4.43	0.06	1.09	1.54	0.45	6	25.30	14.62	4.84	5.38	-0.15
ΔOP	0.80	0.60	1.09	0.00	3.15	7	25.07	16.17	5.72	4.78	-1.37
ΔC	0.00	0.90	0.49	1.33	0.08	8	24.88	16.07	6.02	4.78	-0.37
UI	0.41	0.00	0.05	0.09	0.29	9	24.28	15.74	4.36	5.33	-2.69

Panel C: Estimated Coefficients

	coefficients					t-values				
	FF3	FF5	Q4	M4	BF3	FF3	FF5	Q4	M4	BF3
TERM	-0.04	1.38	1.63	1.90	1.02	-0.17	4.33	3.47	3.64	1.97
DY	10.35	8.18	11.88	1.20	6.55	3.54	2.25	2.06	0.26	1.25
ΔIP	-12.95	-6.79	-9.19	2.15	-5.01	-4.58	-2.05	-1.78	0.51	-1.01
PE	1.47	3.87	2.99	6.69	-0.06	1.14	2.55	1.29	2.43	-0.02
RF	-0.13	-0.03	0.06	-0.01	-0.06	-3.51	-0.53	0.66	-0.10	-0.58
DS	0.14	-0.02	-0.08	-0.10	-0.04	3.04	-0.29	-0.72	-0.91	-0.30
ΔOP	-1.02	-1.05	-1.54	0.06	1.91	-1.37	-0.93	-0.84	0.03	0.66
ΔC	-0.08	2.23	1.66	2.82	-0.48	-0.06	1.48	0.71	0.96	-0.16
UI	-0.23	0.01	-0.10	-0.13	0.16	-0.87	0.04	-0.16	-0.21	0.23
Cons	1.01	0.99	0.98	0.98	0.99	101.56	66.39	48.49	41.18	37.81

which all nine classic macroeconomic shocks are included. Note that the inclusion of some weakly relevant shocks can impair the R^2_{os}. For instance, the 3-shock regression explain the SDF of BF3 with R^2_{os} of 0.81 %, which drops to -2.69% after including the other six shocks.

Second, a parsimonious regression with three or nine classic macroeconomic shocks can explain the variations in SDFs almost as well as a high-dimensional regression with 131 macroeconomic shocks. Particularly, the R^2_{os} of regressing SDFs on TERM, DY, and ΔIP range from 0.81% to 25.34%. Recall that R^2_{os} of regressing SDF on 131 macroeconomic shocks using elastic net range from 3% to 20%.

Third, SDFs of the five characteristic-based factor models can not be equally explained by the classic macroeconomic shocks. Consistent with the baseline specification with 131 macroeconomic shocks, I observe that FF3 and FF5 have higher R^2_{os} than M4 and BF3 (25% and 15% versus 5% and 1%). Also the SDF of BF3 is interpreted the least by classic macroeconomic shocks, with R^2_{os} close to zero or even negative.

In addition, I also present the coefficients and corresponding t-statistics of the OLS regression with nine classic macroeconomic shocks in Panel C of Table 2.3. The intercepts are approximately equal to one, which is consistent with the fact that the means of SDFs are unity by construction. The coefficient on ΔIP is negative, implying that SDFs are higher when there is a negative shock to the growth rate of industrial productivity. The coefficients on TERM, DY, and PE are positive, implying that SDFs are higher when there is positive shock to the term spread, dividend and price-earning ratio of S&P500. However, DS, ΔOP, ΔC, and UI are not significant predictors of SDFs.

The main take away is that the macroeconomic interpretations of SDFs have sparse

dimensions. Particularly, three classic shocks explain SDFs with comparable R^2_{os} to that of 131 macroeconomic shocks. However, my findings do not imply that those non-classical macroeconomic shocks are irrelevant or useless for explaining characteristic-based factor models. In Figure A1, I plot the frequency at which each of the 131 macroeconomic shock has selected by elastic net to predict SDFs. I observe that elastic net does not always pick TERM, DY, and ΔIP. Instead, combinations of other macroeconomic shocks perform comparably well in explaining SDFs. Moreover, my findings do not imply that the pricing content of macroeconomic shocks has sparse dimension. The SDFs I aim to explain are just approximations for the true SDF. My findings indicate that though a large number of macroeconomic shocks are potential systematic risks and bear nontrivial risk premia, the characteristic-based factor models can only admit a small number of them.

Elastic Net Regression of SDFs on 131 Macroeconomic Shocks and Their Interaction Terms

In this section, I investigate whether the nonlinear forms of macroeconomic shocks help explaining the SDFs of characteristic-based factor models.

On the one hand, the assumption of a linear relationship between the SDFs and macroeconomic shocks is motivated by the convention of developing linear factor models in both empirical and theory asset pricing studies. For example, the classic CCAPM predicts that the SDF is a linear model of aggregate consumption growth; the more recent intermediary asset pricing theory predicts that SDF is a linear model of the market returns and leverage of financial intermediary. On the other hand, one might doubt that the obtained low R^2_{os} could be attributed to the linearity restriction, and a predictive model of macroeconomic

shocks with more flexible function form could have improved explanatory power for SDF. To mitigate this concern, I consider nonlinear terms of macroeconomic shocks and analyze the change in R^2_{os}.

Based on $q = 131$ macroeconomic shocks, I construct $q \times (q-1)/2$ bivariate interactions terms. I augment them to my original 131 macroeconomic shocks and obtain a huge set of 8,646 predictors for SDFs. As number of regressors exceeds the number of time periods (T=636), this is exactly the case where the traditional regression models fail and machine learning can make a a large improvement. I run elastic net regressions of the five SDFs on the 8,646 macroeconomic shocks and interaction terms, and report the results in Table 2.4

Panel A presents the R^2_{os} of elastic net regressions for each of the five characteristic-based factor models. The most important observation is that the greatly enlarging the set of explanatory variables does not improve the explanatory power for the variation in SDFs. Particularly, the R^2_{os} of regressing the five SDFs on 8,646 macroeconomic shocks and interaction terms are in the range of 3% to 20%, which is the same as those with only 131 macroeconomic shocks in Table 2.2. Besides the level of R^2_{os}, the precision of estimated R^2_{os} (i.e., the standard deviations) are also comparable in two sets of regressions. Interestingly, the R^2_{os} of M4 and BF3 become insignificant with more predictors as the probability of negative bootstrapped $R^2_{os,B}$ becomes 13.1% and 33.4%, respectively (which are 0% and 0.6% respectively in Table 2.2).

Panel B shows across model comparisons of R^2_{os}. The numbers in this panel represent the probability that the bootstrapped Ros, b^2 for Model X (presented in the first column)

Table 2.4. Explain SDFs with 131 Macroeconomic Shocks and Their Bivariate Interactions

SDFs of five characteristic-based factor models: FF3, FF5, M4, Q4, and BF3 are the independent variables. The macroeconomic shocks are estimated as the first-order differences of 131 macroeconomic variables. The explanatory variables include the 131 macroeconomic shocks and their bivariate interactions. Panel A presents the out-of-sample R^2 (R^2_{os}) of the regressions. Particularly, "$R^2_{os^*}$" denotes the R^2_{os} obtained from original sample. "Mean($R^2_{os,B}$)" and "Std($R^2_{os,B}$)" represent the mean and standard deviation, respectively, of the R^2_{os} that are estimated on the 1000 bootstrapped samples. "Prob($R^2_{os,B} < 0$)" represents the probability that bootstrapped R^2_{os} falls below zero. Panel B presents the comparisons of R^2_{os} between different characteristic-based factor models. Particularly, "Prob($R^2_{os,B}(X) > R^2_{os,B}(Y)$)" represents the probability that the bootstrapped R^2_{os} of a model X (listed in the first column) is greater than that of a model Y (listed in the first row).The sample is from 01/1963 to 12/2015.

Panel A: R^2_{os} of each model					
	FF3	FF5	Q4	M4	BF3
$R^2_{os^*}$, %	20.43	11.82	4.89	6.52	3.74
Mean($R^2_{os,B}$),%	19.87	12.75	4.62	4.00	1.16
Std($R^2_{os,B}$), %	5.44	3.09	2.59	3.71	2.47
Prob($R^2_{os,B} < 0$), %	0.20	0.00	2.00	13.10	33.40
Panel B: Prob($R^2_{os,B}(X) > R^2_{os,B}(Y)$), %					
Model Y / Model X	FF3	FF5	Q4	M4	BF3
FF3	-				
FF5	5.60	-			
Q4	0.50	0.60	-		
M4	0.40	0.90	44.70	-	
BF3	0.20	0.10	11.90	22.50	-

is greater than that for Model Y (presented in the first row). To avoid proving redundant information, I only represent the left lower half the table, and the omitted numbers in the right upper half of the table are simply the transpose of the left lower half the table subtracted from 100%. The numbers can be read as the p-value of a one-sided t-test of $R^2_{os,b}(X) > R^2_{os,b}(Y)$. Consistent with the regression of SDFs on 131 macroeconomic shocks in Table 2.2, the relative magnitudes of R^2_{os} follow a descending order of FF3 $>$ FF5 $>$ M4 $>$ Q4 $>$ BF3. Specifically, the difference in R^2_{os} between FF3/FF5 with the other three

models are statistically significantly.

The results oft this section demonstrate that relaxing the linearity constrain of the predictive model does not improve the explaining for SDFs. It further confirms the conclusion that though all the characteristics-based factor models I consider admit macroeconomic interpretations, none of them are fully grounded in macroeconomic foundations.

Do Intermediary Asset Pricing Factors Explain SDFs?

My approach can also inform whether a particular theory-based asset pricing factor provides additional information about the SDF compared to all other macroeconomic shocks. To illustrate this application, this section considers an intermediary asset pricing model and investigates the risk factors' unique information about SDFs.

Intermediary Asset Pricing Risk Factors

Standard consumption asset pricing theories consider households as the marginal investors. In contrast, intermediary asset pricing theories consider financial intermediaries as the marginal investors, because they are sufficiently sophisticated and have more active participation in financial market than households. In intermediary asset pricing theories, the pricing kernel is the marginal value of wealth of financial intermediaries.

He, Kelly, and Manela (2017) construct the intermediary pricing model as a linear function of two factors: excess market returns and shocks to financial intermediaries' capital ratio. The intermediary capital ratio is constructed as primary dealers' aggregate market equity over aggregate assets. Further, the shock to intermediary capital ratio is constructed

as the AR(1) innovations to its current value over its lagged value. The excess market returns capture the asset pricing role of persistent productivity shocks that affect the overall fundamentals of the economy. Supplement to that, the shock to intermediary capital ratio captures the soundness of financial intermediaries. Intermediaries have high marginal value of wealth when their capital ratio is low. Therefore, the intermediary factor carries a positive risk premium, meaning that assets paying more in states of the world with a lower intermediary capital ratio have lower expected returns in equilibrium.

The data of intermediary asset pricing factors is provided by Asaf Manela, and is available starting from January 1970. [11] I therefore cut my sample for SDFs and macroeconomic shocks correspondingly, and use monthly data from January 1970 to December 2015 for the empirical analysis in this section.

Empirical Results

To investigate the explanatory power of intermediary asset pricing factors compared to macroeconomic shocks, I regress SDFs separately on three sets of shocks and compare the R^2_{os}.

The first set of shocks consist of two factors in intermediary asset pricing model, namely, excess market returns and shock to intermediary capital ratio, and is denoted by $\chi^{intermediary}$. Note that $\chi^{intermediary}$ only contains two variables, I I regress SDFs on it using standard OLS. R^2_{os} is constructed, as in the regression of SDFs on nine classic macroeconomic shocks, and the distribution of R^2_{os} is simulated by bootstrap. The second set of shocks consists of the 131 macroeconomic shocks, and is denoted by χ^{macro}. The third set of

[11] The factors at annual, monthly and daily frequency is published in Manela's data library: http://apps.olin.wustl.edu/faculty/manela/data.html

shocks consists of all the variables from $\chi^{intermediary}$ and χ^{macro}, and is denoted by χ^{all}. Both χ^{macro} and χ^{all} are high-dimensional, I thus regress SDFs on them using elastic net. Specifically, I follow the baseline regression of SDFs on 131 macroeconomic shocks to construct R^2_{os} and simulate the distributions of R^2_{os}.

I regress five SDFs separately on the three sets of shocks and present the R^2_{os} respectively in Panel A, B, and C of Table 2.5. Comparing Panel A and B, I observe that the explanatory power of two theory-based factors can be comparable with or even higher than a large set of 131 macroeconomic shocks. More specifically, $\chi^{intermediary}$ can explain the variation in SDFs with R^2_{os} range from 4% to 30%. Also $\chi^{intermediary}$ explains the SDF of FF3 with higher R^2_{os} than $\chi^{intermediary}$, which is 30% versus R^2_{os} of 19%. Panel C shows that the combination of intermediary asset pricing factors and 131 macroeconomic shocks can slightly improve the explanation for all SDFs, and the R^2_{os} range from 5% to 33%.

Next, I investigate the additional explanatory power of macroeconomic shocks compared to intermediary asset pricing factors, and present the results in Panel D of Table 2.5. Particularly, I construct the difference of R^2_{os} between the regressions with χ^{all} and $\chi^{intermediary}$ and test the statistical significance of this difference based on the probability of it falling belows zero. According to Panel D, $R^2_{os}(\chi^{all}) - R^2_{os}(\chi^{intermediary})$ is positive across all SDFs and the range is from 2% to 4%. Although the differences of R^2_{os} are small, they are significant at 5% level as their probability of being negative is lower than 5%. This observation implies the intermediary asset pricing factors can not fully subsume the explanatory power of the macroeconomic shocks.

Table 2.5. Explain SDFs with Three Sets of Shocks

SDFs of five characteristic-based factor models: FF3, FF5, M4, Q4, and BF3, are the independent variables. $\chi^{intermediary}$ consists of intermediary asset pricing factor; χ^{macro} consists of 131 macroeconomic shocks; and χ^{all} is a union of the previous two sets of shocks. Panel A, B and C respectively present the R^2_{os} of an OLS regression of SDFs on $\chi^{intermediary}$, and two elastic net regressions of SDFs separately on χ^{macro} and χ^{all}. Particularly, "$R^2_{os^*}$" denotes the R^2_{os} obtained from original sample. "Mean($R^2_{os,B}$)" and "Std($R^2_{os,B}$)" represent the mean and standard deviation, respectively, of the R^2_{os} that are estimated on the 1000 bootstrapped samples. "Prob($R^2_{os,B} < 0$)" represents the probability that bootstrapped R^2_{os} falls below zero. Panel D and E present the comparisons of bootstrapped R^2_{os} between the three specifications. "$\Delta R^2_{os^*}$" denotes the difference of R^2_{os} in the original sample; "Mean($\Delta R^2_{os,B}$)" and "Std($\Delta R^2_{os,B}$)" denote the mean and standard deviations of bootstrapped ΔR^2_{os}; "Prob($\Delta R^2_{os,B} < 0$)" counts the probability that bootstrapped ΔR^2_{os} falls below zero. The sample is from 01/1970 to 12/2015.

Panel A: $R^2_{os}(\chi^{intermediary})$

	FF3	FF5	Q4	M4	BF3
$R^2_{os^*}$	29.75	11.69	5.86	3.51	3.59
Mean($R^2_{os,B}$)	32.06	12.26	6.47	4.34	4.22
Std($R^2_{os,B}$)	4.67	4.63	3.08	3.54	2.95
Prob($R^2_{os,B} < 0$)	0.00	0.00	0.30	6.80	5.00

Panel B: $R^2_{os}(\chi^{macro})$

	FF3	FF5	Q4	M4	BF3
$R^2_{os^*}$	18.75	13.08	6.13	4.18	3.79
Mean($R^2_{os,B}$)	22.37	17.76	9.24	10.18	10.11
Std($R^2_{os,B}$)	4.77	3.98	3.07	3.92	3.63
Prob($R^2_{os,B} < 0$)	0.00	0.00	0.10	0.00	0.20

Panel C:$R^2_{os}(\chi^{all})$

	FF3	FF5	Q4	M4	BF3
$R^2_{os^*}$	33.52	14.71	6.90	3.87	5.11
Mean($R^2_{os,B}$)	39.06	19.57	10.75	10.27	9.99
Std($R^2_{os,B}$)	5.01	4.52	3.30	3.84	3.80
Prob($R^2_{os,B} < 0$)	0.00	0.00	0.00	0.10	0.10

Panel D:$R^2_{os}(\chi^{all}) - R^2_{os}(\chi^{intermediary})$

	FF3	FF5	Q4	M4	BF3
$\Delta R^2_{os^*}$	3.78	3.02	1.04	0.36	1.52
Mean($\Delta R^2_{os^*}$)	7.00	7.30	4.28	5.93	5.77
Std($\Delta R^2_{os,B}$)	3.21	3.39	2.85	3.93	3.34
Prob($\Delta R^2_{os,B} < 0$)	0.70	0.40	4.40	3.40	2.30

Panel E:$R^2_{os}(\chi^{all}) - R^2_{os}(\chi^{macro})$

	FF3	FF5	Q4	M4	BF3
$\Delta R^2_{os^*}$	14.77	1.64	0.77	-0.31	1.33
Mean($\Delta R^2_{os,B}$)	16.70	1.81	1.51	0.09	-0.11
Std($\Delta R^2_{os,B}$)	3.43	2.09	1.86	1.91	2.51
Prob($\Delta R^2_{os,B} < 0$)	0.00	20.40	79.80	49.60	53.20

On the flip side, I also investigate the additional explanatory power of intermediary asset pricing factors compared to the large set of macroeconomic shocks, and present the results in Panel E. Particularly, I construct and test the difference of R^2_{os} between the regressions with χ^{all} and χ^{macro}. According to Panel E, $R^2_{os}(\chi^{all}) - R^2_{os}(\chi^{macro})$ is positive for the SDFs of FF3, FF5, and Q4 and the range is from 1% to 15%. The difference of R^2_{os} between the two specifications is large for FF3 and none of the realizations of it is negative. This observation implies that with the presence of macroeconomic shocks, intermediary asset pricing factors contain additional information only for the SDF of FF3.

The takeaway of this section is that intermediary asset pricing factors cannot subsume the explanatory ability of macroeconomic shocks for SDFs, and they contain additional information only for the SDF of FF3. Note that, though I focus on intermediary asset pricing factors for illustration, my approach can be widely used for all other theory-motivated factors.

Robustness Tests

This section discusses the robustness of my findings with respect to alternative constructions of SDFs and macroeconomic shocks.

GMM Estimation of SDFs with Alternative Test Assets

SDFs of characteristic-based factor models are the basis of my empirical investigation. In the main analysis, I construct the prices of risks and SDF based on the weakest requirement that the model correctly prices the returns on the factor portfolios. Knowing that many other

tradable assets can also be priced by the model, I use them as the test assets and estimate the prices of risks and SDF by GMM (e.g., Jagannathan and Wang 2002). The moment conditions are specified as follows:

$$\begin{aligned} \mathbb{E}[R_t^e(1-b'(f_t-\mu))] &= 0; \\ \mathbb{E}[f_t(1-b'(f_t-\mu))] &= 0. \end{aligned} \tag{2.7}$$

Here R_t^e represents a $n \times 1$ a vector of excess returns on test assets. f_t represents a $p \times 1$ vector of factors, whose sample average across all periods is denoted by μ.

The test assets I consider are the 25 portfolios sorted by size and book-to-market ratios, and 10 portfolios sorted by firm industry. Returns on all the 35 portfolios are from Ken French's data library. This set of tests assets are often used in the literature, however it is an empirical fact that characteristic-based factor models cannot perfectly price all of them. For instance, FF3 is not able to price the extremely small firms. Therefore, I use a 2-stage GMM method following the suggestions of Cochrane (2005). The first-stage picks parameters to minimize a equal-weighted sum of squared errors. The second stage finds the parameters targeting to minimize a linear combination of pricing errors with the smallest sampling variations. Intuitively, 2-stage GMM is like a GLS in linear regression, which assign higher weights on the moments that are fitted more precisely. With a 2-stage GMM, the construction of SDF will will be downgrade the impact of those test assets that are poorly priced by the model and decrease the noise of estimations.

I present the GMM estimations in Table 2.6, which can be considered as a counterpart of Table 2.1. Panel A presents the GMM estimated prices of risks $\hat{b}$ and the respective

standard errors *se*. The GMM estimated prices of risks are close to those of Table 2.1 and are all statistically significant. Panel B presents the annualized standard deviations for the GMM estimated SDFs. Overall, GMM estimations result in more volatile SDFs for all models. For instance, the standard deviations of Q4, M4 and BF3 are doubled after being estimated by GMM. However, the comparative volatility remain the same, that is, FF3 has the least volatile SDF, while M4 and BF3 have the most volatile SDFs. Panel C shows the correlation of GMM estimated SDFs between different models. Since the five SDFs are estimated based on a fixed set of test assets, the correlations become stronger. Particularly, FF3 and FF5 have the highest correlation of 0.88, which was 0.62 under simple construction.

With the GMM estimated SDFs of the five characteristic-based factor models, I regress each of them on 131 macroeconomic shocks using elastic net. The results are presented in Panel A and B of Table 2.7. Consistent with the main results in Table 2.2, Panel A shows that R^2_{os} are positive and significantly at 1% level for all SDFs. Though the magnitudes of $R^2_{os^*}$ are slightly higher than the main results, which is from 7% to 21 % versus 3% to 20%, the conclusion still holds that none of the characteristic-based factor models can be fully represented by macroeconomic shocks. According to Panel B, FF3 and FF5 are still better explained by macroeconomic shocks than M4 and BF3, and the the difference in R^2_{os} are significant at 10% level. This section confirms that my main findings are robustness to the alternative approach to estimate SDFs.

Table 2.6. GMM Estimation of Risks Prices and SDFs with Alternative Test Assets

The five models are FF3, FF5, Q4, M4 and BF3. Panel A presents the two-stage GMM estimated risk prices ($\hat{b}$) and corresponding standard errors (*se*) of each factor in each model, where the factors, 25 size and book-to-market portfolios and 10 industry portfolios are used as the test assets. Panel A lists the five models in the first row and the corresponding factors in the first column. Panel B presents the annualized standard deviations of the five estimated SDFs ($\hat{m}_t = 1 - \hat{b}' f_t$). Panel C presents the pairwise correlations between SDFs of different characteristic-based factor models. The sample is from 01/1963 to 12/2015. Factors of FF5 are missing from 01/1963 to 06/1963; factors of Q4 are missing from 01/1963 to 12/1966; factors of BF3 are missing from 01/1963 to 06/1972 and in 2015.

Panel A: Prices of risks estimated by GMM

	FF3		FF5		Q4		M4		BF3	
	$\hat{b}$	*se*	$\hat{b}$	*se*	$\hat{b}$	*se*	$\hat{b}$	*se*	$\hat{b}$	*se*
MKT	4.72	(1.04)	6.15	(1.11)	7.06	(1.12)	11.30	(1.27)	9.99	(1.31)
SMB	3.82	(1.34)	4.94	(1.39)			9.78	(1.47)		
HML	9.96	(1.44)	5.55	(2.10)						
RMW			6.65	(1.97)						
CMA			8.28	(3.07)						
ME					8.32	(1.44)				
IA					20.52	(2.36)				
ROE					11.78	(1.70)				
MGMT							18.83	(1.84)		
PERF							8.80	(1.25)		
FIN									10.28	(1.36)
PEAD									22.92	(2.35)

Panel B: Annualized standard deviations of SDFs

	FF3	FF5	Q4	M4	BF3
SD	0.38	0.45	0.86	1.14	1.07

Panel C: Correlations of SDFs

	FF3	FF5	Q4	M4	BF3
FF3	1.00				
FF5	0.88	1.00			
Q4	0.55	0.75	1.00		
M4	0.50	0.68	0.70	1.00	
BF3	0.32	0.40	0.45	0.56	1.00

Table 2.7. Elastic Net Regressions of Alternative SDFs on 131 Macroeconomic Shocks

SDFs of five characteristic-based factor models: FF3, FF5, M4, Q4, and BF3, are the independent variables. The SDFs are estimated by a two-stage GMM with the factors, 25 size and book-to-market portfolios and 10 industry portfolios as the test assets. The macroeconomic shocks are estimated the first-order differences of 131 macroeconomic variables and are the explanatory variables. Panel A presents the out-of-sample R^2 (R^2_{os}) of the regressions. Particularly, "$R^2_{os^*}$" denotes the R^2_{os} obtained from original sample. "Mean($R^2_{os,B}$)" and "Std($R^2_{os,B}$)" represent the mean and standard deviation, respectively, of the R^2_{os} that are estimated on the 1000 bootstrapped samples. "Prob($R^2_{os,B} < 0$)" represents the probability that bootstrapped R^2_{os} falls below zero. Panel B presents the comparisons of R^2_{os} between different characteristic-based factor models. Particularly, "Prob($R^2_{os,B}(X) > R^2_{os,B}(Y)$)" represents the probability that the bootstrapped R^2_{os} of a model X (listed in the first column) is greater than that of a model Y (listed in the first row).The sample is from 01/1963 to 12/2015.

Panel A: R^2_{os}					
	FF3	FF5	Q4	M4	BF3
$R^2_{os^*}$, %	19.37	20.88	11.35	10.43	6.63
Mean($R^2_{os,B}$), %	21.94	24.15	15.03	15.41	13.01
Std($R^2_{os,B}$), %	4.47	4.47	3.66	4.01	3.77
Prob($R^2_{os,B} < 0$), %	0.00	0.00	0.00	0.00	0.00
Panel B: Prob($R^2_{os,B}(X) > R^2_{os,B}(Y)$), %					
X \ Y	FF3	FF5	Q4	M4	BF3
FF3	-				
FF5	77.30	-			
Q4	6.40	0.60	-		
M4	8.30	1.90	53.90	-	
BF3	3.90	0.90	30.60	27.90	-

Construct Macroeconomic Shocks as AR(1) Innovations

The macroeconomic shocks are innovations to macroeconomic variables and are potential systematic risks captured by the characteristic-based factor models. This section investigates the sensitivity of my findings to alternative construction of shocks to macroeconomic variables. In the main analysis, I construct macroeconomic shocks as the first-order differences of 131 macro series from FRED-MD, since they were carefully transformed (by taking log or difference) to be almost stationary. However, in many existing studies, macroeconomic shocks more cautiously constructed as the the residuals to a VAR(1) or AR(1) regression on the lagged macroeconomic variables. Considering that there are 131 macro series in my setting, VAR regression is not applicable, I use AR(1) instead. Particularly, I assume that investors rely on the lagged realization of a macro series to form expectations and the residuals to the AR(1) regression captures the unexpected innovations.

I regress the SDFs on the newly constructed 131 macroeconomic shocks using elastic net. The results are presented in Panel A and B of Table 2.8. According to Panel A, the R^2_{os} are close to the main results in Table 2.2. Particularly, R^2_{os} for five characteristic-based factor models has the range from 3 % to 20% and are all significant at 1% level. As shown in Panel B, the comparative macroeconomic interpretability of the five SDFs remain unchanged.

To sum up, macroeconomic shocks constructed as AR(1) innovations of macroeconomic variables are potentially more precise than first-order differences of macroeconomic series. My findings are not sensitive to the alternative constructions of macroeconomic shocks.

Table 2.8. Elastic Net Regressions of SDFs on 131 AR(1) Macroeconomic Shocks

SDFs of five characteristic-based factor models: FF3, FF5, M4, Q4, and BF3, are the independent variables. Macroeconomic shocks are estimated as the AR(1) innovations to 131 macroeconomic variables. Panel A presents the out-of-sample R^2 (R^2_{os}) of the regressions. Particularly, "$R^2_{os^*}$" denotes the R^2_{os} obtained from original sample. "Mean($R^2_{os,B}$)" and "Std($R^2_{os,B}$)" represent the mean and standard deviation, respectively, of the R^2_{os} that are estimated on the 1000 bootstrapped samples. "Prob($R^2_{os,B} < 0$)" represents the probability that bootstrapped R^2_{os} falls below zero. Panel B presents the comparisons of R^2_{os} between different characteristic-based factor models. Particularly, "Prob($R^2_{os,B}(X) > R^2_{os,B}(Y)$)" represents the probability that the bootstrapped R^2_{os} of a model X (listed in the first column) is greater than that of a model Y (listed in the first row).The sample is from 01/1963 to 12/2015.

Panel A: R_{os}					
	FF3	FF5	Q4	M4	BF3
$R^2_{os^*}$, %	19.95	13.59	4.84	5.90	3.33
Mean($R^2_{os,B}$), %	22.52	16.98	6.09	11.54	7.70
Std($R^2_{os,B}$), %	4.38	3.56	2.59	4.04	3.60
Prob($R^2_{os,B} < 0$), %	0.00	0.00	0.50	0.00	0.90
Panel B: Prob($R^2_{os,B}(X) > R^2_{os,B}(Y)$) %					
X \ Y	FF3	FF5	Q4	M4	BF3
FF3	-				
FF5	6.50	-			
Q4	0.00	0.00	-		
M4	1.90	6.40	94.50	-	
BF3	0.40	1.10	66.70	15.20	-

Conclusion

Characteristic-based factor models are very popular in empirical asset pricing. They perform well in describing the cross-sectional variation in stock returns, but they have long been criticized for the lack of economic interpretations. The main research question of this chapter is do characteristic-based factors represent macroeconomic shocks? I answer this question by regressing the SDFs of those models on a comprehensive set of 131 macroeconomic shocks and applying machine learning techniques to prevent overfitting. R^2_{os} is the metric which measures to what extent the variation of an SDF is explained by macroeco-

nomic shocks out-of-sample.

This chapter simultaneously studies five prominent characteristic-based factor models, that is FF3, FF5, Q4, M4 and BF3, and convey three main findings. First, characteristic-based factors can be better understand under the behavioral framework rather than risk-based framework. I find that 131 macroeconomic shocks can only explain 4% to 20% variation of the SDFs of those model. I attribute the remaining variation of the SDFs that is unexplained by such a comprehensive set of macroeconomic shocks, to market inefficiency or behavioral-based explanations. Second, the relation between macroeconomic shocks and SDF is sparse. Three prominent shocks can have explanatory power comparable to 131 macroeconomic shocks. Third, models vary in terms of their macroeconomic content. Models that are arguably risk-based, such as FF3 and FF5, can be better explained by macroeconomic shocks than models that are motivated by market mispricing, such as M4 and BF3.

In addition, I also illustrate the application of my approach in the context of investigating whether intermediary asset pricing factors contain unique information for SDFs compared to macroeconomic shocks. I find that intermediary asset pricing factors do not fully subsume the explanatory power of 131 macroeconomic shocks, and can only provide additional information for SDF of the FF3.

My study contributes to the literature in three ways. First, it provides a unique application of machine learning to the analysis of economic interpretability of asset pricing factors. Second, it proposes a new approach to investigate characteristic-based factor models that focuses on models' SDFs rather than individual factors. Last, it simultaneously considers five characteristic-based factor models, and is thus the most comprehensive study of such

type of models up to date.

This chapter suggests several directions for future research. For example, it could be interesting to further pin down the specific macroeconomic shocks that are represented by characteristic-based factors. Further, it is possible that one can use the SDFs of characteristic-based factor models to guide us construct better asset pricing models from macroeconomic shocks. Third, it could be exciting to investigate the unexplained variation in SDFs, which do not admit macroeconomic interpretations, under the behavioral frame-work.

CHAPTER 3

DO INVESTORS EFFICIENTLY AGGREGATE ANALYSTS' FORECASTS?

Introduction

Prior literature has documented that there is a cross-sectional variation in the accuracy of analysts' earnings forecasts (Stickel 1992) as well as persistence in such accuracy (Sinha, Brown, and Das 1997). Moreover, the market reaction to analysts' revisions is stronger for the analysts with higher accuracy (Loh and Mian 2006), which implies that investors learn about the analysts' reliability (Chen, Francis, and Jiang 2005).

In this study, we examine whether investors aggregate analysts' forecasts efficiently. If investors efficiently aggregate analysts' earnings forecasts, a reasonable proxy for the market expectations of a firm's earnings should be the accuracy-weighted average of all analysts' earning forecasts. If investors do not take analysts' accuracy into account, market expectations should be equal to the simple average of all analysts' earnings forecasts. Our contribution is to document that investors do not efficiently aggregate analysts' forecasts by factoring in the variation in analysts' accuracy. The misweighting on analysts' forecasts is especially concentrated in stocks with high participation by retail investors, who have greater difficulty processing data relative to institutional investors.

We start by defining the following: (1) Naive Consensus, which is the simple average of all outstanding forecasts three days prior to an earnings announcement; (2) Smart Consen-

sus, which is the accuracy-weighted average of all outstanding forecasts three days prior to an earnings announcement. Specifically, for any firm, the accuracy-based weight on an analyst is the ratio of his/her accuracy, which is measured as the inversed absolute forecast error in the preceding quarter, against the sum of accuracy over all analysts following the firm. By construction, higher weights are assigned to forecasts provided by more accurate analysts and lower weights are assigned to forecasts provided by less accurate analysts. We also allow analysts' accuracy to vary depending on the firms they cover (Brown, Richardson, and Schwager 1987, Lang and Lundholm 1996, Behn, Choi, and Kang 2008), and to vary over time (Mikhail, Walther, and Willis 1997, Clement 1999), so that the impact of change in analyst experience is incorporated.

We find that, on average, the absolute earnings forecast error of Smart Consensus is 6.7 cents. The ability of Smart Consensus to predict actual EPS, by itself, does not indicate whether the forecast aggregation is efficient or not. We thus examine whether Smart Consensus is an improved earning predictor over the traditional Naive Consensus. The absolute earnings forecast error of Naive Consensus is 0.4 cent higher than that of Smart Consensus, which is also statistically significant according to a pair-wise t-test. Based on the above evidence, we conclude that Smart Consensus is more accurate than Naive Consensus. In addition, the comparative advantage of Smart Consensus over Naive Consensus is mainly driven by firms whose analyst coverage is lower or analyst forecast dispersion is higher. This is consistent with the fact that the choice of weighting schemes makes a greater impact on the value of aggregated forecasts for these firms.

As another test for the superiority of Smart Consensus over Naive Consensus, we investigate whether Smart Consensus predicts Naive Surprise, which is the difference between

reported EPS and Naive Consensus.[1] The logic here is simple: if Smart Consensus is a more efficient aggregate of analysts' earnings forecasts than Naive Consensus, it should predict the forecast error of Naive Consensus. We introduce a new variable, termed "Expectation Gap", which is the difference between Smart Consensus and Naive Consensus. We first examine if the sign of Expectation Gap predicts the sign of Naive Surprise. Investors who trade before earnings announcements are particularly concerned about whether the consensus forecast is overly optimistic or pessimistic about a firm's earnings. Therefore, predicting the sign of Naive Surprise deserves a separate inspection along with predicting the value of Naive Surprise. In an univariate test, conditional on Expectation Gap being positive, 70% of the firms will have a positive Naive Surprise (i.e., reported earnings exceed naive expectations). On the other hand, conditional on Expectation Gap being negative, 51% of the firms will have a negative Naive Surprise (i.e., reported earnings underperform naive expectations). It is a substantial improvement over a unconditional prediction for the sign of Naive Surprise, which is 58% instead of 70% for a positive Naive Surprise, and is 42% instead of 51% for a negative Naive Surprise. More formally, we regress Naive Surprise on Expectation Gap and control for several variables documented in prior literature to predict Naive Surprise. We find the coefficient on Expectation Gap to be 0.63, which is significantly positive but also significantly different from one. In terms of economic significance, for every $1 of Expectation Gap, the firm will report a Naive Surprise of 63 cents.

Having established the ability of Smart Consensus to predict a firm's earnings, we test whether investors efficiently aggregate analysts' earnings forecasts. We introduce a new

[1] In some papers, this forecast error is also termed "earnings surprise."

variable, "Smart Surprise", which is the difference between actual EPS and Smart Consensus. We then examine market reaction to Naive Surprise and Smart Surprise to deduce investors' expectations because the market reaction reflects surprises. If investors efficiently weight each analyst's forecast, the market reaction to earnings should be related to Smart Surprise but unrelated to Naive Surprise. Our question can only be answered when firms release their earnings, and not by examining the market reaction when analysts issue their revisions or by examining forecasts made months in advance of earnings announcements.

To test whether investors efficiently weight analyst's earnings forecast, we estimate regressions of three-day cumulative returns (CR) around an earnings announcement on Smart Surprise and Naive Surprise simultaneously. The key observation is that earnings announcement returns are positively related to Smart Surprise, as well as Naive Surprise, the latter implying that investors do not efficiently weight analysts' forecasts. We further examine whether there is cross-sectional differences (across stocks) in the way investors weight analysts' forecasts. Specifically, individual investors are unlikely to weight analysts' forecasts efficiently given that they have fewer resources (relative to institutional investors) to identify accurate analysts. Consistent with this notion, we find that such misweighting on analysts' forecasts is more pronounced in firms with low institutional (and, therefore, correspondingly high individual) ownership. For robustness, we also consider three alternative cumulative abnormal returns (CAR) to exclude the impact of sysmatic risk. The three CARs are constructed from market-adjusted, Carhart four-factor model adjusted, and DGTW-benchmarked (Daniel, Grinblatt, Titman, and Wermers 1997) daily returns. Our empirical results are insensitive to the choice of risk adjustment for returns.

Finally, we examine whether we can implement a trading strategy around earnings an-

nouncement days that exploits the predictive power of Expectation Gap on Naive Surprise induced by investors' misweighting. For each fiscal calendar quarter, we first construct Expectation Gap for all the firms. Next, we form a long-short portfolio at the end of trading day 't–2' prior to an earnings announcement, which buys stocks in the top quintile sorted by Expectation Gap and short stocks in the bottom quintile sorted by Expectation Gap. We hold this long-short portfolio for three days and reverse the trade at the end of day 't+1' after an earnings announcement. We find the annualized cumulative abnormal returns to be 1.3%. Given that misweighting on analyst earnings forecasts is likely to occur in firms with more retail investors, we construct trading strategies separately for the subsamples with high and low institutional ownership. We find that the same strategy implemented in the firms with low institutional ownership earns a statistically significant annualized cumulative abnormal return of 1.9%, which is substantially higher than the strategy implemented in all firms. In contrast, the same strategy implemented in the firms with high institutional ownership does not produce a significant cumulative abnormal return.

This chapter is closely related to the literature on measuring unobservable earnings expectations of capital market participants. Many empirical studies in finance and accounting rely on the identification of investors' earnings expectations. The first generation of such measures of investors' earnings expectations are produced by time-series models of EPS (Watts and Leftwich 1977, Griffin 1997, Foster 1977, Brown and Rozeff 1979). However, researchers in the 1980s discovered that analysts' earnings forecasts are better proxies for firms' future earnings because analysts incorporate broader and more timely information than used by the time-series models (Brown, Richardson, and Schwager 1987, Brown, Hagerman, Griffin, and Zmijewski 1987a). Ever since, analyst earnings forecast consensus

has been introduced to the literature as a proxy for market earnings expectations and, traditionally, the consensus is constructed as the average or median of outstanding analysts' forecasts (Elton, Gruber, and Gultekin 1981, Fried and Givoly 1982, Brown, Foster, and Noreen 1985, O'Brien 1988). More recent studies cast doubt on the ability of such traditional consensus measures to predict earnings. Barron, Kim, Lim, and Stevens (1998) and Kim, Lim, and Shaw (2001) demonstrate that a naive mean consensus of analysts' forecasts overweights public information; Brown and Kim (1991) remove stale forecasts and create a consensus based on recent forecasts; Butler, Markov, and Kraft (2007) create a consensus that weights analysts with respect to characteristics (e.g., forecast age, broker size, and analyst experience) that have been shown to be associated with their accuracy. Our study contributes to the literature by proposing an accuracy-based consensus measure, Smart Consensus, to proxy market expectations. Distinct from Butler, Markov, and Kraft (2007), our construction of accuracy-based weights are solely determined by analysts' past forecast accuracy rather than by various analysts' characteristics. Using a model evaluation methodology (Beaver and Dukes 1972, Collins 1975, Patell 1976, Ball and Brown 1968, Fried and Givoly 1982, O'Brien 1988), we find that the forecast error of Smart Consensus (i.e., the earnings surprise if Smart Consensus is specified as market expectations) exhibits a higher association with stock returns than that of Naive Consensus. Therefore, Smart Consensus is, on average, a better surrogate for market earnings expectations than Naive Consensus.

As we construct a consensus measure based on expected analysts' forecast accuracy, our study is also closely related to the literature about analysts' forecast accuracy. Forecast accuracy is one of the most important determinants for analysts' job security (Mikhail,

Walther, and Willis 1999) and reputation (Jackson 2005). Prior literature has documented a large cross-sectional variation in analysts' forecast accuracy that can be well explained by the variation in the characteristics of analysts and brokerages. For instance, Mikhail, Walther, and Willis (1997) document that analysts improve forecast accuracy as they gain firm-specific experience; Clement (1999) finds that analysts' accuracy is positively related to analysts' overall and firm-specific forecasting experience (ability and skill) and brokerage size (resources available), and it is negatively associated with the number of firms and industries followed by the analyst (task complexity). Jacob, Lys, and Neal (1999) document how analysts' accuracy is positively associated with their forecast frequency, broker-industry specialization, and broker size, but is negatively associated with broker-analyst turnover. In addition to analyst-specific characteristics, analysts' forecast accuracy also depends on firms' information environment. For instance, Brown, Richardson, and Schwager (1987), Lang and Lundholm (1996), Abarbanell and Bushee (1997), Hope (2003), Behn, Choi, and Kang (2008) document a positive association between firms' financial disclosure quality and analysts' accuracy. Such a positive association with analysts' accuracy also holds for firms' non-financial (CSR) disclosure quality (Dhaliwal, Radhakrishnan, Tsang, and Yang 2012) and corporate governance transparency (Bhat, Hope, and Kang 2006). In contrast to that literature, our accuracy measure of analysts' accuracy is based on the past accuracy instead of identified analyst and firm characteristics. The identified superiority of Smart Consensus over Naive Consensus in predicting future earnings adds to evidence that aside from numerous analyst and firm characteristics, past forecast accuracy is a comprehensive and parsimonious proxy for analysts' expected accuracy.

Our study is also related to the literature about investors' ability to assess the reliability

of information provided by different analysts. Studies find evidence in support of investors' sophistication in differentiating analysts. The theoretical model in Abarbanell, Lanen, and Verrecchia (1995) predict that investors' respond to forecast revisions increases with the expected accuracy of the forecast, and has been empirically proved by Stickel (1992), Park and Stice (2000). Hilary and Hsu (2013) also document that investors can recover systematic bias in analysts' forecasts and thus react more to analysts with consistent forecast error. However, evidence also exists in support of investors inability of assessing the reliability of information provided by different analysts. Clement and Tse (2003) find that only some analyst characteristics, that determine analysts' forecast accuracy, can also predict market reaction to analysts' revisions. Gleason and Lee (2003) further document a slower and less complete return adjustment for obscure yet highly accurate analysts than for "celebrity" analysts (Institutional Investor All-Stars). Consistent with prior literature, we find evidence that not all investors can efficiently weight analysts' forecasts. In addition, we document that investors have heterogeneous ability in processing information. Particularly, institutional investors are sophisticated enough to assign efficient weights on earnings forecasts with respect to analysts' accuracy, in contrast, retail investors are less capable of extracting and exploiting the implicit information regarding analysts' accuracy when they form their earnings expectations.

Two recent papers that are closely related to our study are Chen and Jiang (2006) and So (2013). While we examine whether investors weight each analyst's forecast efficiently given the cross-sectional variation in analysts' forecast accuracy, Chen and Jiang (2006) examine whether analysts weight their private and public signals efficiently to arrive at their forecast and So (2013) examines whether investors efficiently weight forecasts of analysts

and forecast of earnings obtained purely from firm characteristics. Chen and Jiang (2006) find that analysts overweight the more optimistic of the two signals, while So (2013) finds that investors overweight analyst consensus. In our study, we find that investors do not efficiently weight individual analyst's forecasts.

Data and Variables

We use the I/B/E/S Unadjusted Detail and Unadjusted Actual files to obtain data on analysts' earnings forecasts and firms' earnings announcements. We use Compustat for data on firms' fundamentals, and CRSP for the data of stock returns.

Sample Selection

We start with all available quarterly earnings announcements in the I/B/E/S Unadjusted Actual file from the first quarter of 1993 to the last quarter of 2015. We choose 1993 as the starting year of our sample because the earnings announcement dates on I/B/E/S are suspect in prior years (Ma and Markov 2017). We apply several filters to the firm-quarter observations in our sample that are similar to Ma and Markov (2017): (1) We require that a firm's fiscal quarter end is aligned with calendar quarters, that is March, June, September, or December; (2) We require that the announcement date of earnings are within 45 days of the quarter end for the first three fiscal quarters and within 90 days of the quarter end for the fourth fiscal quarter; (3) As in Barth and So (2014) and Dellavigna and Pollet (2009) we compare I/B/E/S and Compustat announcement dates and assume the earlier date is correct; (4) We use the I/B/E/S timestamp to adjust the announcement date one trading day forward

if the announcement occurred after markets close; (5) We require earnings announcement dates reported in the I/B/E/S Actual Unadjusted file and Compustat to differ by no more than three days (in contrast to one day in Ma and Markov (2017)).

We merge the firm-quarter earnings announcements data with the corresponding analysts' quarterly EPS forecasts from the I/B/E/S Unadjusted Detail file. As with prior literature (So 2013, Chen, De, Hu, and Hwang 2014, Ma and Markov 2017), we use the I/B/E/S Unadjusted Details and Unadjusted Actual files to avoid rounding errors induced by retroactive split adjustment. As in Livnat and Mendenhall (2006), we use the CRSP adjustment factor to adjust analysts' EPS forecasts for stock splits that occurred between the analysts' forecasts reporting date and the firm's earnings announcement date. For an analyst, we keep their latest EPS forecast of a firm-quarter that has been made at least two days before an earnings announcement. Since we construct trading strategies that buy-and-hold stocks for a three-day window around the earnings announcement days, the analysts' forecasts made within two days before the earnings announcements are ignored to avoid look-ahead bias. We also eliminate firm-quarters where less than two analysts follow the firm because there are no adjustment in weights to be made in terms of forecast accuracy and, by construction, Smart Consensus will be the same as Naive Consensus.

Lastly, we obtain the data of firms' fundamentals and stock returns from Compustat and CRSP. Following the convention of prior literature, we drop firms in financial or utility sector. We require the market value of a firm to be higher than $5 million at the end of previous quarter. We drop firms with closing stock price lower than $5 (Chen and Jiang 2006, So 2013) on day 't–2' prior to the earnings announcement date (initial trade date) and day 't+1' after the earnings announcement (final trade date) because our tests are based

on three-day announcement returns that involves buying and selling at the closing prices on these two days. This avoids contamination of returns by bid-ask bounce.

Naive Consensus and Smart Consensus

As in prior literature, we estimate the Naive Consensus for a given firm-quarter as the simple average of the latest outstanding forecasts of all analysts covering the firm:

$$Naive\ Consensus_{j,t} = \frac{1}{N}\sum_{i=1}^{N} F_{j,t}^{i}, \tag{3.1}$$

where F denotes an analyst's forecast of EPS for a firm, while i, j, and t are indexes for the analyst, firm, and quarter, respectively. We estimate consensus as of the day t-2 relative to an earnings announcement date. We do not use the numbers given in the I/B/E/S Summary file because it is available only once a month (as of the third Thursday of the month). A large fraction of the earnings announcements have at least one analyst updating their forecast after the latest date for which I/B/E/S Summary is available; thus, the use of the Summary file will result in measurement errors.

Ideally, Smart Consensus should weight each forecast by the analyst's current and actual accuracy, which is unobservable until the earnings announcement. We, instead, use the past forecast accuracy to proxy the current accuracy of an analyst. Specifically, inversed forecast accuracy is the absolute difference between an analyst's outstanding forecast before earnings announcement and the firm's reported earnings. In the main analysis, we determine the weight of an analyst as the ratio of his/her accuracy in the immediate preceding quarter against the sum of accuracy over all analysts following the same firm in the same

quarter. Smart Consensus is therefore constructed as:

$$Smart\ Consensus_{j,t} = \sum_{i=1}^{N} AW_{j,t}^{i} F_{j,t}^{i} = \sum_{i=1}^{N} \frac{1/|E_{j,t-1} - F_{j,t-1}^{i}|}{\sum_{i=1}^{N} \left(1/|E_{j,t-1} - F_{j,t-1}^{i}|\right)} F_{j,t}^{i}, \tag{3.2}$$

where E and F represent the reported and analyst's forecast EPS of a firm, i, j, and t are indexes for the analyst, firm, and quarter, respectively. We allow analyst-level accuracy to vary with time and firm. If an analyst has zero earnings forecast error in the previous quarter, we assign the error to be 0.0001, which is lowest possible error, to avoid unidentified accuracy or weight. If an analyst did not make earnings forecast during the past quarter, we drop his/her forecast from an Accuracy Consensus. According to the unreported tests, our findings are insensitive to constructing the weight of an analyst based on his/her relative accuracy in the preceding four quarters.

We winsorize the actual EPS (E), the constructed Naive and Smart Consensuses at the 1st and 99th percentile values. Panel A of Table 1 reports the descriptive summary statistics. We end up with 179174 firm-quarter observations in the sample. The sample average of actual EPS and the two consensuses are similar and close to 0.29, which means that analysts forecasts are overall unbiased. However the two consensuses have smaller standard deviation (0.44) and smaller magnitude of extreme values (range over -0.80 to 1.95) as compared to actual EPS (with standard deviation of 0.48, and range over -1.07 to 2.08). This indicates that analysts, are more conservative in issuing extremely high or low earnings forecasts.

Naive Surprise, Smart Surprise, and Expectation Gap

As in Chen, De, Hu, and Hwang (2014), Naive Surprise is the difference between the reported quarterly EPS and Naive Consensus,

$$Naive\ Surprise_{j,t} = E_{j,t} - Naive\ Consensus_{j,t}.$$

As Naive Consensus is the traditional proxy for market expectations of EPS as of the day prior to an earnings announcement, Naive Surprise is thus a conventional measure of earnings surprise in the market. In the case of a positive Naive Surprise, firms' reported EPS outperforms market expectations, and thus generates a positive earnings surprise to the market. On the contrary, in the case of a negative Naive Surprise, the reported EPS underperforms market expectations and is thus a negative surprise to the market. In prior literature, a positive (negative) Naive Surprise is followed by an upward (downward) correction of stock price, which results in a positive (negative) abnormal stock return and long-run return drift.

Similar to the definition of Naive Surprise, Smart Surprise is constructed as the difference between the reported quarterly EPS and Smart Consensus,

$$Smart\ Surprise_{j,t} = E_{j,t} - Smart\ Consensus_{j,t}.$$

If Smart Consensus is a better predictor for earnings, we should observe a smaller magnitude of Smart Surprise compared to the magnitude of Naive Surprise. Separately, if investors efficiently weight analysts according to their accuracy, Smart Consensus is likely to

be a better proxy for the unobservable market expectations of earnings than Naive Consensus would be. Therefore, Smart Surprise should turn out to be a better proxy for earnings surprise to market, and is associated with stronger market reaction as compared to Naive Surprise.

We also construct Expectation Gap as the difference between Smart Consensus and Naive Consensus,

$$Expecation\ Gap_{j,t} = Smart\ Consensus_{j,t} - Naive\ Consensus_{j,t}.$$

Expectation Gap captures the discrepancy between two consensuses. Knowing that Smart Consensus predicts earnings better than Naive Consensus, Expectation Gap is helpful for predicting Naive Surprise, and thus can be exploited to build trading strategy around earnings announcements if investors do not weight analysts' forecasts efficiently. In our main results, we use the unscaled Surprises and Expectation Gap. We also scaled them by the absolute value of respective EPS in unreported robustness tests, where the results and implications hold.

Earnings Announcement Returns

As in Cohen, Dey, Lys, and Sunder (2007), we estimate the accumulative abnormal returns over the three-day announcement window to proxy the market reaction to unexpected news in earnings announcement. This circumvents the need to be precise about the timing of announcement release, though we do our best to identify the correct date and time. As described in the section of sample selection, we compare the earnings announcement dates

reported from I/B/E/S and Compustat. To summarize, we requires the dates reported by the two different database to differ by no more than three days and use the earlier date as the correct earnings announcement date. We use the actual announcement date rather than expected or predicted announcement date because our focus is not in estimating the magnitude of earnings announcement premia as compensation for risk (Cohen, Dey, Lys, and Sunder 2007, Savor and Wilson 2014). Therefore, we do not attempt to forecast the expected announcement date. Anecdotally, even when firms announce early, they typically follow it with a formal earnings announcement and we use this formal date as the earnings announcement date. For early announcers, we still capture the announcement return on the formal announcement date, but we do not expect much of a reaction given the early release of news.

In the baseline setting, we consider the cumulative returns (CR) over the three-day window (t-1, t+1) around earnings announcement days as a measurement for earnings announcement returns. Although risk adjustment might only have negligible impact on the daily returns of individual stocks, for robustness, we also consider three alternative three-day cumulative abnormal returns (CAR) in place of raw returns.

One way to adjust individual stock returns for systematic risks is to estimate the abnormal returns of the stock in a factor pricing model. We consider both CAPM and Carhart four-factor model as the benchmark pricing model.[2] The daily data on market (MKT), size (SMB), value (HML), and momentum (MOM) factors are taken from Ken French's data library. Over an estimation window of (t-255, t-3) before earnings announcement, we es-

[2]CAPM contains a single market factor. Carhart four-factor model is obtained by augmenting Fama-French three-factor model (Fama and French 1993) by a momentum factor (Carhart 1997)

timate the factor loadings (β^i_{MKT} for CAPM, as well as β^i_{MKT}, β^i_{SMB}, β^i_{HML}, and β^i_{MOM} for Carhart model) for each stock. Next, the daily abnormal returns over the event window of (t-1, t+1) around earnings announcement days are obtained for each stock as:

$$AR_t^{i,Mkt} = r_t^i - \hat{\beta}^i_{MKT} MKT_t,$$

$$AR_t^{i,Carhart} = r_t^i - \hat{\beta}^i_{MKT} MKT_t - \hat{\beta}^i_{SMB} SMB_t - \hat{\beta}^i_{HML} HML_t - \hat{\beta}^i_{MOM} MOM_t.$$

The three-day cumulative abnormal returns, $\text{CAR}^{\text{i,Mkt}}$ and $\text{CAR}^{\text{i,Carhart}}$, are constructed by compounding $\text{AR}^{\text{i,Mkt}}_{\text{t}}$ and $\text{AR}^{\text{i,Carhart}}_{\text{t}}$, respectively, over the three-day (t-1, t+1) window around earnings announcement days

The DGTW benchmark proposed by Daniel, Grinblatt, Titman, and Wermers (1997) is another risk-adjustment for returns and is widely applied in the literature of mutual fund. Based on the assumption that stocks with similar firm characteristics — size, book-to-market ratio, and momentum — are likely to have comparable exposure to risks, DGTW method obtains abnormal returns on a stock $AR_t^{i,DGTW}$ by benchmarking its raw returns against the value-weighted average returns on a portfolio of stocks which have the similar characteristics. Data on the assignments of benchmark portfolios for all common stocks comes from Russ Wermers' data library. The three-day cumulative abnormal returns, $\text{CAR}^{\text{i,DGTW}}$, is again constructed by compounding $\text{AR}^{\text{i,DGTW}}_{\text{t}}$ over the three-day (t-1, t+1) window around earnings announcement days

The summary statistics for the cumulative returns (CR) and three alternative cumulative abnormal returns (CAR^{Mkt}, $\text{CAR}^{\text{Carhart}}$, and CAR^{DGTW}) are reported in Panel B of Table 3.1, and are presented in percentage values. All CR and CARs have similar variance and

percentile values. Specifically, they all have mean and median values that are close to zero in our sample. As the data for DGTW benchmark assignments was lastly updated on 2012, the sample size of CAR^{DGTW} is substantially smaller than the other CARs.

Control Variables

In addition to the main variables described above, we also construct a large set of control variables that have been used in prior literature for either reported earnings or earnings announcement returns. This subsection describes the control variables. We include only the citation for the first published paper that, to the best of our knowledge, proposed the control variable in regressions explaining the variation in earnings surprise or earnings announcement returns.

Analyst Coverage is the number of analysts following a firm (Barton and Simko 2002) as of the beginning of the current quarter.

Dispersion represents the analysts' forecast dispersion (Payne and Robb 2000) which is constructed as the standard deviation of the analysts' EPS forecasts for a given firm-year.

LTG represents the expected long-term growth of the firm (LaPorta 1996) which measured as the long-term earnings growth forecast from I/B/E/S. If missing, we replace it with the growth rate implied by the one-year ahead (FY1) and two-year ahead (FY2) earnings forecast that are also from I/B/E/S (Frankel and Lee 1998).

SG represents the sales growth (Lakonishok, Shleifer, and Vishny 1994). It is defined as the growth rate in annual sales over the past three years. Because our study is based on

Table 3.1. Summary Statistics of Main Variables

Summary statistics are presented for all variables. "Obs.", "Mean", "SD", "Min", and "Max" denote, respectively, the number of observations, mean, standard deviation, minimum and maximum values for each variable. "P25", "P50", and "P75" denote, respectively, the 25th, 50th, and 75th percentile for each variable. Panel A presents summary statistics for actual EPS (E) and the two consensuses on EPS. Panel B presents summary statistics for four alternative measurements of three-day cumulative (abnormal) returns around earnings announcement. All C(A)Rs are reported as percent values. The sample starts from 1993:Q1 to 2015:Q4.

Panel A: Analysts' EPS forecasts								
	Obs.	Mean	SD	Min	P25	P50	P75	Max
E	179174	0.29	0.48	-1.07	0.03	0.23	0.49	2.08
Naive Consensus	179174	0.29	0.44	-0.80	0.04	0.22	0.47	1.92
Smart Consensus	179174	0.28	0.44	-0.83	0.03	0.22	0.47	1.95
Panel B: three-day cumulative (abnormal) returns around earnings announcements								
	Obs.	Mean	SD	Min	P25	P50	P75	Max
CR,%	179141	0.00	8.50	-27.82	-4.00	0.10	4.30	25.80
CAR^{Mkt},%	179165	0.20	8.90	-27.40	-4.20	0.00	4.50	28.06
$CAR^{Carhart}$,%	179165	0.00	8.60	-26.99	-4.10	0.00	4.20	26.80
CAR^{DGTW},%	122297	0.00	8.00	-25.10	-3.90	0.00	4.00	24.68
Panel C: Control variables								
	Obs.	Mean	SD	Min	P25	P50	P75	Max
Analyst Coverage	179174	9.71	6.93.00	3.00	5.00	8.00	13.00	58.00
Dispersion	178846	0.45	1.12	0.00	0.05	0.12	0.32	8.33
LTG	179076	0.21	0.25	-0.28	0.11	0.16	0.24	1.89
SG	148161	1.23	3.54	-0.76	0.07	0.37	0.97	27.96
ME/BE	176401	1.09	0.40	0.54	0.81	0.99	1.28	2.49
ME/BE^{+}	179174	2.24	1.75	0.00	1.22	1.66	2.56	10.91
ME/BE^{-}	179174	0.02	0.12	0.00	0.00	0.00	0.00	0.00
MOM	179174	0.01	0.14	-0.41	-0.06	0.01	0.07	0.49
Accrual	173559	-177.00	567.00	-4078.00	-97.00	-17.00	1.00	272.00
NOA	170176	2.23	4.31	-18.10	0.85	1.69	2.97	24.69
Shares	179026	10.81	1.18	8.56	9.99	10.66	11.49	14.30
Size	179026	13.71	1.67	10.25	12.52	13.59	14.80	18.09
AD	176928	0.84	0.37	0.00	1.00	1.00	1.00	1.00
IO	174730	0.61	0.26	0.03	0.40	0.64	0.82	1.00
Durable	179074	0.39	0.49	0.00	0.00	0.00	1.00	1.00
LTR	179074	0.27	0.45	0.00	0.00	0.00	1.00	1.00
R^{1Y},%	179166	14.60	60.80	-86.13	-21.60	6.80	36.60	291.43
R^{1M},%	179164	0.60	14.10	-40.48	-6.70	0.30	7.40	50.00
CAR^{1Q},%	178168	0.10	5.20	-16.22	-2.50	0.10	2.80	15.01
CAR^{4Q},%	162839	0.10	8.30	-26.16	-3.90	0.10	4.30	24.94
BidAsk	179174	0.04	0.03	0.01	0.02	0.03	0.05	0.15
Volume	179174	8.69	0.91	6.31	8.11	8.71	9.31	10.85
Quarter End	179174	0.25	0.43	0.00	0.00	0.00	0.00	1.00

quarterly data, we use the average growth rate of four quarters of sales (t–5Q to t–2Q) and compare it to the four quarters of sales three years earlier (t–17Q to t–14Q).

ME/BE represents the market-to-book ratio (Fama and French 1993) is estimated using quarterly book value (as of t–2Q) and market price (as of t–1Q).

MOM represents the momentum of stock returns. As in (Chen, De, Hu, and Hwang 2014), MOM is the cumulative market-adjusted daily returns over the period [t–3D, t–30D]

Accrual is defined as the Total Operating Net Accruals (Sloan 1996, Bradshaw, Richardson, and Sloan 2001), which equals to the difference between Income before Extraordinary Items and Net Cash from Operating Activities.

NOA is the abbreviation for net operating assets. As in Barton and Simko (2002) it is constructed as the shareholders equity minus cash and marketable securities plus total debt. We scale NOA by Sales. Considering that we quarterly data is in this chapter, to avoid cross-sectional differences induced by seasonality in quarterly sales, we scale NOA by cumulative sales of prior four quarters rather than the latest quarter's sales.

Shares represents shares outstanding (Barton and Simko 2002), which is also the number of shares used to calculate the I/B/E/S EPS consensus of quarter end t–2. **Size** is the natural logarithm of market capitalization as of the end of t–1 (Das, Levine, and Sivaramakrishnan 1998).

AD represents auditor quality (Barton and Simko 2002). It is an indicator variable that equals 1 if the auditor is part of Big 8 before 1989, part of Big 5 between 1998–2002, and part of Big 4 after 2002.

IO represents institutional ownership (Mohanram 2005). It is the shares held by financial institutions over the total number of shares outstanding. It is averaged over the prior four quarters, specifically, at the end of t–1 to t–4.

Durable represents durable goods (Matsumoto 2002), which is used to capture a firm's reliance on implicit claims. It constructed as a dummy variable which equals 1 if the firm belongs to one of the SIC codes: 150-179, 245, 250-259, 283, 301, and 324-399.

LTR represents litigation risk (Soffer, Thiagarajan, and Walther 2000, Skinner 1994, Matsumoto 2002). It is an indicator variable that equals 1 if a firm is in the industries with high litigation risk, and equals 0 otherwise. Specifically, firms in four industries are considered to have high litigation risk: biotechnology (SIC codes 2833-2836 and 8731-8734), computers (SIC codes 3570-3577 and 7370-7374), electronics (SIC codes 3600-3674), and retailing (SIC codes 5200-5961).

$\mathbf{R^{1Y}\%}$ and $\mathbf{R^{1M}\%}$ are two measurements of past returns. R^{1Y},% is the average monthly return of the stock during the prior year. R^{1M},% is the monthly return of the stock in the prior month.

$\mathbf{CAR^{1Q}\%}$ and $\mathbf{CAR^{4Q}\%}$ are two measurements of past earnings announcement returns. CAR^{1Q},% is three-day cumulative abnormal returns of the stock around the earnings announcement date in the previous quarter.CAR^{4Q},% is average three-day cumulative abnormal returns of the stock around the earnings announcement date in the previous year (four quarters).

BidAsk and **Volume** are two liquidity measures that are included. BidAsk is the average

bid-ask spread over the previous month. Volume is the trading volume of the stock over the previous month.

QuarterEnd is the fourth quarter indicator (Barton and Simko 2002) is a dummy variable that equals 1 if the reporting quarter corresponds to the firm's fourth quarter of the fiscal year, and equals 0 otherwise.

The summary statistics for the control variables are provided in Table 3.1 Panel C. Overall, their magnitudes and variance are consistent with the literature.

Empirical Tests

In order to examine whether investors efficiently aggregate analysts' earnings forecasts, we consecutively test the following empirical questions. (1) Are accuracy-based weights efficient for aggregating analysts' forecasts? (2) Do investors efficiently weight analysts' forecasts with respect to their accuracy? (3) If investors misweight analysts' forecasts, can we formulate a profitable trading strategy around earnings announcements?

Validity of Smart Consensus

There exists a cross-sectional variation in analysts' forecast accuracy. It is intuitive to postulate that forecasts provided by different analysts are not equally reliable to investors. Therefore, when we aggregate all analysts' forecasts to form a consensus measure, information on analysts' forecast accuracy should be incorporated to determine weights we put on each forecast. We construct Smart Consensus such that higher weights are assigned to forecasts provided by more accurate analysts and lower weights are assigned to forecasts

provided by less accurate analysts. In this section, we test if our accuracy-based weights are efficient or not. Particularly, we use the traditional consensus forecast, Naive Consensus, as a benchmark and examine whether the more sophisticated Smart Consensus is a better predictor for the actual EPS of a firm.

Is Smart Consensus More Accurate Than Naive Consensus?

An analyst's relative earnings forecast accuracy is likely to persist over time (Sinha, Brown, and Das 1997). Based on that, we propose to weight analysts' earnings forecasts for a given firm-quarter with respect to how accurately they predicted the firm's actual EPS in the preceding quarter. Construction of accuracy-based weights is described in the section of data and variables. A consensus measure is preferred over a single forecast that is considered the most accurate because an analyst's past accuracy will not perfectly predict future accuracy, furthermore, combining individual forecasts diversifies away idiosyncratic forecast error. If our accuracy-based weights are efficient for aggregating analysts' forecasts, Smart Consensus will outperform the traditional Naive Consensus in predicting the actual EPS of the firm. Thus, our first hypothesis is

Hypothesis 1: Smart Consensus predicts actual EPS more accurately than Naive Consensus.

We construct Naive $FE_t = |\text{Naive Surprise}_t|$ and Smart $FE_t = |\text{Smart Surprise}_t|$ to measure the forecast error of Naive and Smart Consensuses, respectively. By construction, the FE measures are inversely related to the accuracy of a consensus.

We compare the forecast error of Naive and Smart Consensuses using a pair-wise t-test under the null of Naive FE = Smart FE over all firm-quarters in our sample. The

test results are presented in column (1) of Table 3.2. Panel A and B report the summary statistics for the forecast error of Naive Consensus (Naive FE) and Smart Consensus (Smart FE), respectively. In the two panels, statistics included are: the number of observations (Obs.), sample mean (Mean), standard deviation (SD), minimum value (Min), maximum value (Max), as well as the 25%, 50%, 75% percentiles (P25, P50, P75, respectively). Panel C presents the average difference between Naive FE and Smart FE, together with the corresponding t-values.

The average absolute value of forecasting errors for Naive Consensus (7.1 cents) is slightly higher than that of Smart Consensus (6.7 cents). Besides the mean, Naive FE is above Smart FE at each percentile considered in the table. A simple comparison of Panel A and B provides a straightforward observation that the accuracy of the consensus improves (forecasting error of the consensus declines) when we switch from the traditional equal-weighting method to a more sophisticated accuracy-based weighting scheme. Panel C confirms this observation by showing that, on average, Naive FE $-$ Smart FE is equal to 0.4 cent, which is statistically significant at 1% level.

The difference between accuracy-based weights and equal weights are expected to be greater when the number of analysts following the firm is smaller or the dispersion in analysts' earnings forecasts for the firm is higher. More specifically, when fewer analysts follow the firm, each individual analyst pulls a significant weight when they are equally-weighted, and therefore switching to accuracy-based weighing scheme makes a greater difference to the value of consensus. Also, the choice weighting scheme matters only if there exists substantial dispersion in analysts' earnings forecasts. For instance, in an extreme scenario where the analysts following a firm produce homogeneous forecast numbers, the

value of consensus becomes invariant to alternative weighting schemes. According to the rationale above, we further test the following hypothesis:

Hypothesis 1a: The advantage of Smart Consensus over Naive Consensus in predicting actual EPS is greater with lower level of analysts coverage.

Hypothesis 1b: The advantage of Smart Consensus over Naive Consensus in predicting actual EPS is greater when the dispersion of analysts' forecasts is higher.

To test *Hypothesis 1a*, we compare the forecast error of Naive and Smart Consensuses in two subsamples with high and low level of analyst coverage. As the median Analyst Coverage in our sample is eight (Table 3.1), we place a firm-quarter into the high analyst coverage group if Analyst Coverage $\geq$ 8, and place it into the low analyst coverage group otherwise. Next, we conduct pair-wise comparison on Naive FE and Smart FE in the two subsamples separately. The results are reported in column (2) and (3) of Table 3.2. Both consensuses have slightly higher forecasting errors in Low(Analyst Coverage) group than in the High(Analyst Coverage) group, which is consistent with findings in prior literature that analysts' forecast accuracy is generally higher for firms with higher level analyst coverage. Naive FE is 0.4 cent larger than Smart FE in both subsamples, which are both statistically significant at 1%. Moreover, consistent with H_{1b}, we observe that the difference between Naive FE and Smart FE in Low(Analyst Coverage) group is 0.01 cent higher than that in High(Analyst Coverage) group. This difference is significant at 1% level.

Table 3.2. Comparisons for Forecasting Errors of Two Consensus

The comparisons of Naive and Smart Consensuses are conducted separately on five sets of samples: the full sample (Column "Full Sample"), the two subsamples with low and high analyst coverage (Columns "Low(Analyst Coverage)" and "High(Analyst Coverage)"), and the two subsamples with low and high analyst forecast dispersions (Columns "Low(Dispersion)" and "High(Dispersion)"). Panel A and B present the summary statistics for Naive FE and Smart FE. Panel C presents the t-test on H_0: Naive FE - Smart FE = 0.

Panel A: Summary statistics for Naive FE,%					
	Full sample	Low(Analyst Coverage)	High(Analyst Coverage)	Low(Dispersion)	High(Dispersion)
Obs.	179174	88915	90259	89587	89587
Mean	7.10	7.20	7.00	3.40	10.80
SD	10.30	10.50	10.10	5.00	12.60
Min	0.00	0.00	0.00	0.00	0.00
P25	1.30	1.30	1.30	0.80	2.70
P50	3.30	3.30	3.40	1.90	6.30
P75	8.10	8.30	8.00	4.00	13.50
Max	62.00	62.00	62.00	62.00	62.00
Panel B: Summary statistics for Smart FE					
	Full sample	Low(Analyst Coverage)	High(Analyst Coverage)	Low(Dispersion)	High(Dispersion)
Obs.	179174	88915	90259	89587	89587
Mean	6.70	0.80	6.60	3.30	10.00
SD	9.70	9.90	9.50	4.90	12.00
Min	0.00	0.00	0.00	0.00	0.00
P25	1.10	1.10	1.10	0.90	2.20
P50	3.10	3.10	3.00	2.00	5.70
P75	7.80	7.90	7.60	4.00	12.80
Max	56.70	56.70	56.70	56.70	56.70
Panel C: T-test of H_0: Naive FE = Smart FE					
	Full sample	Low(Analyst Coverage)	High(Analyst Coverage)	Low(Dispersion)	High(Dispersion)
Naive FE - Smart FE	0.40***	0.40***	0.40***	0.01***	0.70***
t-stats	39.74	26.10	30.03	4.60	39.67
		High(Analyst Coverage) − Low(Analyst Coverage)		High(Dispersion) − Low(Dispersion)	
Naive FE - Smart FE		0.01***		0.69***	
t-stats		(3.27)		(38.47)	

Lastly, we compare the forecast error of Naive and Smart Consensuses in subsamples with high and low level of analyst forecast dispersion. Similar to the previous test, we split the sample into high and low dispersion groups using the median value of dispersion (0.32) as the cutting point. We present the results for the two groups in column (4) and (5) of 3.2. As shown in Panel A and B, both Naive and Smart Consensuses have much higher average forecast errors in the High(Dispersion) group (10.8 cents and 10 cents, respectively) than in the Low(Dispersion) group (3.4 and 3.3, respectively). This is consistent with findings in prior literature that an analyst's accuracy is generally lower for firms with higher information uncertainty, which is measured by analysts' forecast dispersion. Most importantly, consistent with *Hypothesis 1b*, Naive FE $-$ Smart FE is 0.69 cent higher in High(Dispersion) group than in Low(Dispersion) group, and such a difference significant at 1% according to Panel C.

Does Expectation Gap Predict Naive Surprise?

In our sample, 54% of the time a reported actual EPS is higher than both Naive and Smart Consensuses (Naive Surprise ≥ 0 and Smart Surprise ≥ 0) and 35% of the time actual EPS is lower than both Naive and Smart Consensuses (Naive Surprise < 0 and Smart Surprise < 0). This means that 89% of the time Naive and Smart Consensuses are on the same side of actual EPS. Combined with the fact that average Smart FE is smaller than average Naive FE, this observation indicates that when Naive Consensus is too pessimistic (optimistic), Smart Consensus is likely to be less pessimistic (optimistic).

Based on the above observation, the difference between Smart and Naive Consensus helps predict whether Naive Consensus is too pessimistic or optimistic about future earn-

ings. Also, the magnitude by which a firm's actual EPS will exceed Naive Consensus can be predicted by Expectation Gap before earnings announcements. As Naive Surprise is the traditional measurement of earning surprise, it is meaningful to test if Expectation Gap helps correct market expectations towards the actual EPS.

Hypothesis 2: Expectation Gap predicts the sign and value of Naive Surprise.

Before a formal test of *Hypothesis 1*, we visualize the relationship between Expectation Gap and Naive Surprise with a set of 2-by-2 matrices in Table 3.3. To form a 2-by-2 matrix, we first sort the sample firm-quarters into two groups based on whether the Expectation Gap is greater than or equal to zero (Smart Consensus $\geq$ Naive Consensus) or less than zero (Smart Consensus $<$ Naive Consensus). Next, we further sort firm-quarters from each group into two more groups based on whether Naive Surprise is greater than or equal to zero (Naive Consensus ≥ 0) or less than zero (Naive Surprise ≥ 0). A 2-by-2 matrix reports the number and proportional percentage of observations for each of the four groups.

In the full sample, Expectation Gap is greater than or equal to zero with a probability of 47%. Consistent with prior literature about firms managing earnings to just meet or beat the market expectations formed as Naive Consensus, we find that 58% of the time a firm will have Naive Surprise ≥ 0. Conditional on a non-negative Expectation Gap, the probability for Naive Surprise ≥ 0 is 70% (which is 58% unconditionally). In contrast, conditioned on a negative Expectation Gap, the probability for negative Naive Surprise is 51% (which is only 42% unconditionally). Knowing the sign of Expectation Gap helps improving the predictability for the sign of Naive Surprise by 20%. To compare the economic significance of 20%, consider a person with a crystal ball who is able to make accurate prediction

Table 3.3. The Signs of Naive Surprise Conditioning on Expectation Gap

Each 2-by-2 matrix in the table is constructed to summarize the number (fraction) of observations in four categories: (Naive Surprise ≥ 0 & Naive Surprise ≥ 0), (Naive Surprise ≥ 0 & Naive Surprise < 0), (Naive Surprise < 0 & Naive Surprise ≥ 0), (Naive Surprise < 0 & Naive Surprise < 0). Such a matrix is constructed separately for five sets of samples: the full sample (Panel "Full Sample"), the two subsamples with low and high analyst coverage (Panel "Low(Analyst Coverage)" and "High(Analyst Coverage)"), and the two subsamples with low and high analyst forecast dispersions (Panel "Low(Dispersion)" and "High(Dispersion)")

	Full Sample		
	Naive Surprise ≥ 0	Naive Surprise < 0	Total
Expectation Gap ≥ 0	58314 (33%)	26166 (15%)	84480 (47%)
Expectation Gap < 0	45632 (25%)	49062 (27%)	94694 (53%)
Total	103946 (58%)	75228 (42%)	179174 (100%)
	Low(Analyst Coverage)		
	Naive Surprise ≥ 0	Naive Surprise < 0	Total
Expectation Gap ≥ 0	34937 (34%)	13237 (13%)	48174 (46%)
Expectation Gap < 0	27164 (26%)	28307 (27%)	55471 (54%)
Total	62101 (60%)	41544 (40%)	103645 (100%)
	High(Analyst Coverage)		
	Naive Surprise ≥ 0	Naive Surprise < 0	Total
Expectation Gap ≥ 0	23377 (31%)	12929 (17%)	36306 (48%)
Expectation Gap < 0	18468 (24%)	20755 (27%)	39223 (52%)
Total	41845 (55%)	33684 (45%)	75529 (100%)
	Low(Dispersion)		
	Naive Surprise ≥ 0	Naive Surprise < 0	Total
Expectation Gap ≥ 0	23379 (26%)	13579 (15%)	36958 (41%)
Expectation Gap < 0	20038 (22%)	32706 (36%)	52744 (59%)
Total	43417 (48%)	46285 (52%)	89702 (100%)
	High(Dispersion)		
	Naive Surprise ≥ 0	Naive Surprise < 0	Total
Expectation Gap ≥ 0	34935 (39%)	12587 (14%)	47522 (53%)
Expectation Gap < 0	25594 (29%)	16356 (18%)	41950 (47%)
Total	60529 (68%)	28943 (32%)	89472 (100%)

100% of the time. Thus, the difference in conditional probability will be 100% for such a perfect forecaster. On the contrary, a naive forecaster will predict 50% of the firms will meet or beat earnings. But only 58% will actual meet or beat earnings. Even among the firms that the naive forecaster predicts will miss earnings, 58% of the firms will meet or beat earnings. Thus the difference in conditional probability is 0% for a native forecaster. The difference in conditional probability of 20% with accuracy-weighted consensus seems reasonably large in relation to the native forecaster. Going forward, we will focus on the difference in conditional probability.

As the change of weighting scheme has different impact on the value of consensus with respect to the level of analyst coverage and analyst forecast dispersion, we expand *Hypothesis 1* as follows:

Hypothesis 2a: The ability of Expectation Gap to predict Naive Surprise is stronger with lower level of analysts coverage.

Hypothesis 2b: The ability of Expectation Gap to predict Naive Surprise is stronger with higher level of analyst forecast dispersion.

We create a 2-by-2 matrix for four subsamples, i.e., Low(Analyst Coverage), High(Analyst Coverage), Low(Dispersion) and High(Dispersion) separately. The partitioning of the four subsamples is described in the previous section. We observe that the unconditional probability of non-negative Naive Surprise to be 60% and 68%, respectively, for subsamples with low analyst coverage and high analyst forecast dispersion. This observation is consistent with prior literature claiming that corporate managers are more engaged in earnings management to beat or meet Naive Consensus when the firm has fewer analysts following and

higher uncertainty. Consistent with *Hypothesis 2a*, Expectation Gap makes a greater improvement in predicting Naive Surprise in the subsample of Low(Analyst Coverage) firms (24%) than the subsample of High(Analyst Coverage) firms (17%). However, inconsistent with *Hypothesis 2b*, Expectation Gap improves the predictability of Naive Surprise by only 12% for High(Dispersion) firms, and by 24% for Low(Dispersion) firms.

We formally test *Hypothesis 2* with three sets of regressions. Firstly, we examine whether the sign of Expectation Gap significantly predicts the sign of Naive Surprise. Specifically, we create I(Expectation Gap ≥ 0) and I(Naive Surprise ≥ 0) as indicators that equal to one for non-negative Expectation Gap and non-negative Naive Surprise, respectively. Then we run a Logistic regression of I(Naive Surprise ≥ 0) on I(Expectation Gap ≥ 0). The results for the described regressions without and with control variables are reported in Columns (1) and (2), respectively, of Table 3.4. The slope coefficients on I(Expectation Gap ≥ 0) are positive and significant at 1%, implying that a firm-quarter with positive Expectation Gap will be more likely to have positive Naive Surprise than a firm-quarter with negative Expectation Gap. The marginal effect of I(Expectation Gap ≥ 0) is reported below the coefficients. Without control variables, a firm-quarter with non-negative Expectation Gap has 20.8% higher chance of having non-negative Naive Surprise, which is consistent with the improvement of predictability in the counterpart univariate analysis. After including control variables, the marginal effect of I(Expectation Gap ≥ 0) shrinks to 17.7% (at the sample average of all control variables), but is still statistically significant at 1% level.

Secondly, we examine whether the magnitude of Expectation Gap significantly predicts the sign of Naive Surprise. We run a Logistic regression of I(Naive Surprise ≥ 0) on

Table 3.4. Predict Naive Surprise with Expectation Gap

Columns (1) - (4) present the results of logistic regressions where indicator I(Naive Surprise ≥ 0) is the dependent variable. The marginal effects of Expectation Gap and I(Expectation Gap ≥ 0) are presented underneath the coefficients. Column (5) and (6) show the results of OLS regressions where Naive Surprise is the dependent variable. Firm and year-quarter fixed effects are included. Presented parentheses are the t-values based on industry and year-quarter clustered standard errors. All control variables are described in the Data and Variables section.

	(1)	(2)	(3)	(4)	(5)	(6)
	Logit:I(Naive Surprise ≥ 0)				OLS: Naive Surprise	
Intercept	0.07**	1.71***	-0.40***	1.23***	0.004***	-0.07***
	(2.33)	(10.02)	(-12.73)	(7.29)	(4.24)	(-3.48)
I(Expectation Gap ≥ 0)	0.87***	0.74***				
Marginal effect	0.21	0.18				
	(20.97)	(21.29)				
Expectation Gap			13.04***	10.69***	0.86***	0.69***
Marginal effect			3.173	2.590		
			(14.50)	(15.54)	(19.80)	(19.59)
LTG		-0.01		-0.01		0.002
		(-0.20)		(-0.16)		(0.77)
SG		-0.01		-0.004		0.000
		(-1.47)		(-1.04)		(0.27)
ME/BE		0.17***		0.14***		-0.01***
		(4.32)		(3.73)		(-3.12)
MOM		1.21***		1.26***		0.06***
		(10.14)		(10.09)		(9.65)
Accrual		0.00		0.00		-0.00
		(0.70)		(0.18)		(-0.65)
NOA		-0.01**		-0.01***		-0.001***
		(-2.57)		(-2.64)		(-4.96)
Shares		-0.15***		-0.18***		-0.01***
		(-6.14)		(-7.29)		(-5.34)
Size		0.22***		0.27***		0.02***
		(9.86)		(10.49)		(6.94)
AD		0.11***		0.11***		0.002
		(4.25)		(4.21)		(0.71)
IO		0.28***		0.29***		-0.003
		(6.31)		(6.07)		(-0.62)
Durable		0.05		0.06		-0.01
		(1.27)		(1.47)		(-0.74)
QuarterEnd		-0.07		-0.06		-0.01**
		(-1.52)		(-1.26)		(-2.35)
Dispersion		-0.17***		-0.16***		-0.01***
		(-14.28)		(-12.96)		(-10.94)
Analyst Coverage		-0.01**		-0.01*		-0.001***
		(-2.08)		(-1.90)		(-7.30)
LTR		0.19***		0.19***		0.01
		(3.53)		(3.51)		(0.22)
Naive Surprise^{-1Q}		2.83***		2.63***		0.14***
		(11.10)		(10.42)		(8.98)
N	179174	134309	179174	134309	179174	134309
R^2	0.03	0.09	0.04	0.09	0.16	0.20

Expectation Gap with and without controls, and report the results in Columns (3) and (4) of Table 3.4, respectively. The coefficients on Expectation Gap is significantly positive, indicating that a firm-quarter with greater Expectation Gap is more likely to have non-negative Naive Surprise. The marginal effect is also reported underneath the coefficient. At the average levels of all variables, 1 cent of increase in Expectation Gap increases the chance of occurrence of non-negative Naive Surprise by 3.17%. Again the marginal impact of Expectation Gap on the probability of non-negative Naive Surprise is declined (to 2.59%) when control variables are added, however it remains statistically significant at 1% level.

Thirdly, we examine whether the value of Expectation Gap predicts the value of Naive Surprise. A simple OLS is applied with Naive Surprise as the dependent variable and Expectation Gap as the independent variable. The regression results are reported in Column (5) of Table 3.4. The coefficients on Expectation Gap is 0.86 and significant at 1% level. This observation indicates that when Expectation Gap increase by 1 cent, Naive Surprise is going to increase by 0.86 cent. As in Column (6) of Table 3.4, after controlling for a set of other variables that also predict earnings surprises, the coefficient of Expectation Gap becomes 0.70 and remains statistically significant at 1% level.

In all three sets of regressions, we include firm- and year-quarter fixed effects. We also use two-way clustered standard errors with respect to industry and year-quarter to account for serial correlation, cross-correlation, and heteroscedasticity (Petersen 2009, Gow, Ormazabal, and Taylor 2010). The construction of control variables are described in section 3. We favor a "kitchen-sink" approach of including as many control variables that have been documented in the literature because we want to (i) isolate the impact of Expectation Gap, if any, on Naive Surprise and (ii) quantify the economic significance of Expectation

Gap in explaining Naive Surprise relative to all prior variables that explain the variation in Naive Surprise.

The signs of the coefficients of control variables are mostly consistent with prior literature. LTG and SG negatively predict Naive Surprise (Frankel and Lee 1998, Lakonishok, Shleifer, and Vishny 1994), because they are positively related to analyst optimism. ME/BE is negatively related to Naive Surprise for the opposite reason. Momentum affects future returns, and to the extent that stock returns are a leading indicator of future performance, MOM is positively related to Naive Surprise (Chen, De, Hu, and Hwang 2014). A trading strategy of shorting (buying) stocks with high (low) accruals was found profitable (Sloan 1996, Bradshaw, Richardson, and Sloan 2001), implying that investors overestimate the persistence of accruals. Therefore firms with high accruals are more likely to report earnings that fall short of analysts' forecasts (negative Naive Surprise). The sign on Accruals is ambiguous in our regression. The accumulation of prior earnings management results in overstatement of the balance sheet (Barton and Simko 2002), and consistent with this logic, we find that Naive Surprise is negatively related to the net operating assets. Additionally, shares outstanding is negatively related to Naive Surprise, as is serves as a constraint on earnings management in a way that the dollar amount of earnings management required to inflate earnings by a penny increases in the number of shares outstanding (Barton and Simko 2002). A firm with a top auditor is subject to a higher level constraint on earnings management, and has a lower level of discretionary accruals (Barton and Simko 2002). Therefore, AD should negatively predict Naive Surprise. However, we find the opposite results in our test. Firms with higher institutional ownership have greater incentives to report a positive earnings surprise because they want to avoid the big market reaction as-

sociated with institutional investors exiting their position (Mohanram 2005). Consistent with this logic, we find a significantly positive relation between Naive Surprise and IO. Since stakeholders assess the financial strength of firms using firms' reported earnings, firms that greatly rely on implicit claims by stakeholders (measured by Durable goods) are more likely are more likely to manage earnings upward. We find ambiguous sign on the coefficient of Durable in our test. Firms in industries that face high litigation risk are more likely to report Naive Surprise in order to mitigate litigation costs with bad newsMatsumoto (2002). Consistently, we find statistically significant coefficient on LTR.

In sum, this section documents that Smart Consensus is a superior predictor for actual EPS than the traditional Naive Consensus. Since Smart Consensus is often less optimistic (pessimistic) when Naive Consensus is too optimistic (pessimistic), Expectation Gap helps predict and thus correct for the forecast error of Naive Consensus. In addition, we observe a greater superiority of Smart Consensus over Naive Consensus in forecasting earnings for firms with lower analyst coverage or higher analyst forecast dispersion, since the choice of weighting schemes matters more for such firms.

Do Investors Efficiently Weight Analysts Forecasts?

Having concluded that Smart Consensus is a better predictor for a firm's future earnings than Naive Consensus, this section studies whether investors are sophisticated enough to weight analysts' forecasts efficiently with respect to analysts' accuracy.

Earnings Announcement Returns in Reaction to Naive and Smart Surprises

Ideally, if investors' earnings expectations are observable, our empirical question can be answered by simply comparing the closeness of investors' expectations with the two competing consensuses. For instance, if investors use accuracy-based weights instead of equal weights when they aggregate analysts' forecasts, their expected EPS should be observed to coincide with Smart Consensus and deviate from Naive Consensus. Without directly observing the earnings expectations of investors, we examine their closeness to a particular consensus by measuring the market reaction to that consensus's forecast error upon the announcement of actual earnings. Particularly, the consensus, whose forecast error is associated with stronger market reaction, is likely to be a better proxy for the market earnings expectations. The intuition is simple: if investors consider Smart Consensus rather than Naive Consensus as their earnings expectations, the market reaction to earnings announcements should be solely driven by Smart Surprise (to what extent the reported earnings surprise the investors) and should be irrelevant to Naive Surprise. Similar methodologies are also used to compare surrogates for market expectations by Beaver and Dukes (1972), Collins (1975), Patell (1976), Ball and Brown (1968), Fried and Givoly (1982), O'Brien (1988).

Hypothesis 3: Investors aggregate analysts' forecasts efficiently and their reaction to earnings announcements are completely explained by Smart Surprise and are irrelevant to Naive Surprise.

We measure the market reaction to earnings announcements as the three-day cumulative returns (CR) around earnings announcement days. The main test involves regressing CR

on Smart Surprise and Naive Surprise simultaneously as the key independent variables. Similar to the panel regressions in the previous section, we include firm and year-quarter fixed effect, and report industry and year-quarter clustered standard errors. The results are reported in Table 3.5.

Column (1) of Table 3.5 presents the regression of CR on Naive and Smart Surprises without any control variables. The coefficients on both Surprises are significantly positive, indicating that not all investors weight analysts' forecasts efficiently with respect analysts' accuracy. The market reaction to Smart Surprise is stronger (a 11.44% increase in CR corresponding to $1 increase in Smart Surprise) compared to the market reaction to Naive Surprise (a 3.99% increase in CR corresponding to a $1 increase Naive Surprise.)

We lean on recent literature, e.g., So (2013), Chen, De, Hu, and Hwang (2014), and Savor and Wilson (2016), to guide our choice of control variables that explain market reaction to earnings announcements. The construction of control variables is described in the section of data and variables. The regression results with control variables are presented in Column (2) of Table 3.5. The coefficients on both surprises remain statistically significant. The market reaction to Smart Surprise shrinks to 10.02% but is still higher than that of Naive Surprise, 6.53%. Overall, Smart Consensus is likely more relevant for investors' expectations.

Investors' ability to aggregate analysts' forecasts is heterogeneous. Institutional investors are in general more sophisticated than retail investors. We expect institutional investors to be able to weight analysts' earnings forecasts more efficiently than retail investors, thus their trading activity around earnings announcements will be more sensitive to Smart Surprise and less sensitive to Naive Surprise. Retail investors are less sophisticated

Table 3.5. Explain Earnings Announcement Returns (CR) with Naive and Smart Surprises

The dependent variable CR is a three-day cumulative returns around earnings announcements. Naive and Smart Surprises and their interactions with high institutional ownership indicator IO^H are the main independent variables. Firm and year-quarter fixed effects are included. Reported in parentheses are the t-values based on industry and year-quarter clustered standard errors. All control variables are described in the Data and Variable section and are omitted from the table for brevity.

	CR		
	(1)	(2)	(3)
Intercept	0.11	12.42***	12.42***
	(0.93)	(7.70)	(7.69)
Naive Surprise	3.99***	6.53***	7.77***
	(6.01)	(9.93)	(7.21)
Smart Surprise	11.44***	10.02***	8.01***
	(9.18)	(9.47)	(6.91)
Naive Surprise $\times$ IO^H			-2.50*
			(-1.78)
Smart Surprise $\times$ IO^H			4.02***
			(2.99)
IO^H			0.004
			(0.03)
ME/BE^+		-0.01	-0.01
		(-0.18)	(-0.20)
ME/BE^-		-0.16	-0.17
		(-0.42)	(-0.46)
$R^{1Y}, \%$		-0.001	-0.001
		(-0.72)	(-0.73)
$R^{1M}, \%$		0.004	0.004
		(0.45)	(0.46)
$CAR^{1Q}, \%$		-0.09***	-0.09***
		(-9.87)	(-9.87)
$CAR^{4Q}, \%$		-0.02***	-0.02***
		(-3.97)	(-3.97)
BidAsk		8.84	8.78
		(1.10)	(1.09)
Volume		-0.17*	-0.17*
		(-1.86)	(-1.81)
Size		-0.80***	-0.80***
		(-6.85)	(-6.84)
IO		-0.12	-0.23
		(-0.24)	(-0.36)
QuarterEnd		0.19	0.18
		(0.80)	(0.80)
N	179165	159052	159052
R^2_{adj}	0.06	0.07	0.07

and are likely to simply average analysts' earnings forecasts to form earnings expectations. Therefore, in contrast to institutional investors, we expect that the trading activity of retail investors on earnings announcement day will be more related to Naive Surprise and less related to Smart Surprise.

Hypothesis 3a: Institutional investors are more likely to adopt accuracy-based weights than retail investors, thus, market reaction to earnings announcements has stronger association with Smart Surprise (Naive Surprise) in firms with high (low) institutional ownership.

To test this hypothesis, we split sample into two groups with respect to firms' proportional percentage of institutional ownership. The median value of institutional ownership in our sample is 64%. We create a high institutional ownership indicator (denoted by IO^H), which is equal to 1 if the institutional ownership for a firm-quarter is greater than or equal to 64%, and is equal to 0 otherwise. We add IO^H and its interaction with the two surprises ($IO^H \times$ Smart Surprise and $IO^H \times$ Naive Surprise) into the regression.

As in Column (3) of Table 3.5, the coefficients on Naive and Smart Surprises are 7.77 and 8.01 respectively, indicating that investors are equally responsive to the two surprises in firms with low institutional ownership. Most interestingly, coefficients on the interactions of IO^H and the two surprises have the opposite signs. The coefficient on $IO^H \times$ Smart Surprise is 4.02 and significant at 1% level. This implies that \$1 increase of Smart Surprise leads to 4.2% higher cumulative returns for firms with high institutional ownership than in firms with low institutional ownership. In contrast, the coefficient on $IO^H \times$ Naive Surprise is -2.5 and significant at 10% level. This indicates that \$1 increase of Naive Surprise leads to 2.5% lower cumulative returns in firms with high institutional ownership than in

firms with low institutional ownership. This observation is consistent with our expectation that institutional shareholders are more likely to weight analysts forecasts with respect to analysts' accuracy and thus are more sensitive to Smart Surprise and less sensitive to Naive Surprise.

For robustness, we also consider three alternative risk-adjusted cumulative abnormal returns (CAR), which are constructed based on different benchmark asset pricing models. Specifically, we estimated the abnormal returns based on CAPM to obtain CAR^{Mkt}, based on Carhart four-factor model to obtain $CAR^{Carhart}$, and based on DGTW-benchmark portfolio assignments to obtain CAR^{DGTW}. The construction of CARs is described in detail in the section of data and variables. We conduct robustness tests as regressing alternative CARs, in place of CR, on Naive and Smart Surprises simultaneously. The regression results are reported in Table 3.6. Since the coefficients of control variables have similar signs and magnitudes with Table 3.5, we omit them in Table 3.6 for brevity.

Table 3.6. Explain Alternative CARs with Naive and Smart Surprises

The three dependent variables are cumulative abnormal returns (CAR) around earnings announcements. Naive and Smart Beats and their interactions with IO^H are the main explanatory variables. Columns (1) - (3), (4) - (6), and (7) - (9) present the regression results for CAR^{Mkt}, $CAR^{Carhart}$, and CAR^{DGTW}, respectively. Firm and year-quarter fixed effects are included. Reported in the parentheses are the t-values based on industry and year-quarter clustered standard errors. All the control variables are described in the Data and Variable section and their coefficients are omitted from the table for brevity.

	CAR^{Mkt}			$CAR^{Carhart}$			CAR^{DGTW}		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Intercept	-0.01	11.99***	11.99***	-0.06	13.09***	13.10***	-0.03	11.99***	12.02***
	(-0.17)	(8.19)	(8.19)	(-1.44)	(8.68)	(8.70)	(-0.78)	(8.51)	(8.52)
Naive Surprise	4.03***	6.17***	7.53***	2.80***	5.86***	7.10***	3.52***	5.72***	7.78***
	(7.00)	(10.26)	(7.64)	(4.95)	(10.36)	(7.84)	(5.31)	(7.33)	(6.15)
Smart Surprise	12.26***	11.37***	10.26***	10.51***	8.426***	8.10***	12.39***	11.27***	9.77***
	(9.65)	(9.86)	(7.27)	(10.03)	(9.80)	(7.67)	(10.47)	(10.52)	(7.00)
Naive Surprise × IO^H			-2.70*			-2.49*			-3.43**
			(-1.85)			(-1.69)			(-2.09)
Smart Surprise × IO^H			4.31***			4.15***			2.51
			(3.13)			(3.13)			(1.48)
IO^H			-0.04			-0.09			-0.06
			(-0.43)			(-1.06)			(-0.60)
N	179165	159052	159052	179165	159052	159052	122297	117013	117013
R^2_{adj}	0.06	0.08	0.08	0.06	0.07	0.07	0.07	0.08	0.08
Controls	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes

In Table 3.6, Column (1) - (3) report the regression results based on CAR^{Mkt}, Column (4) - (6) report the regression results based on $CAR^{Carhart}$, and the rest three columns report the regression results based on CAR^{DGTW}. Overall, the findings and implications with alternative CARs are consistent with those of CR in Table 3.5. We observe the largest difference in coefficients between Naive Surprise and Smart Surprise when no control variables are included, i.e., they are 4.03 versus 12.26 for CAR^{Mkt}, 2.80 versus 10.51 for and $CAR^{Carhart}$, and 3.52 versus 12.39 for CAR^{DGTW}. This difference in coefficients shrinks when controls are added, and becomes negligible for firms with low institutional ownership. The impact of institutional ownership on the association of CAR and the two surprises also holds when CAR^{Mkt} and $CAR^{Carhart}$ are used. Results are slightly different when CAR^{DGTW} is considered to be the dependent variable. The coefficient on $IO^{H} \times$ Smart Surprise is 2.51, which is smaller than what we observe from regressions with other CARs, and are insignificant. However $IO^{H} \times$ Naive Surprise is -3.43 is for CAR^{DGTW}, which is greater in magnitudes than what we observe from regressions with other CARs and significant at 1%.

Subsample Analysis

Since the correlation between Naive Surprise and Smart Surprise is 0.877, there could be potential multicolinearity issue in the regression of CAR on both surprises. We therefore refine our inspection to a much smaller subsample where Naive Surprise contradicts Smart Surprise, i.e., where Naive and Smart Consensuses fall on the opposite sides of the reported actual EPS. Specifically, we utilize two scenarios: (1) Smart Surprise ≥ 0 and Naive Surprise < 0, or (2) Smart Surprise < 0 and Naive Surprise ≥ 0. In scenario (1), investors using Smart Consensus as their earnings expectations will find the earnings announcement

a positive surprise, which results in a positive CAR around earnings announcement day. In contrast, for investors who use Naive Consensus as their expectations will find the earnings announcement a negative surprise, which results in a negative CAR around earnings announcement day. Scenario (2) is the opposite to scenario (1). We create an indicator variable I(Smart Surprise ≥ 0 & Naive Surprise < 0), which equals 1 in scenario (1), equals 0 in scenario (2). In Table 3.7, we regress CAR on the I(Smart Surprise ≥ 0 & Naive Surprise < 0) with or without its interaction with the indicator of high institutional ownership IO^{H}.

Table 3.7. Explain CR or CAR with Naive and Smart Surprises in Subsamples

The four dependent variables are three-day cumulative (abnormal) returns around earnings announcement days. Naive and Smart Beats and their interactions with IO^H are the main explanatory variables. Columns (1) - (2), (3) - (4), (5) - (6), and (7) - (8) present the regression results for CR, CAR^{Mkt}, $CAR^{Carhart}$, and CAR^{DGTW}, respectively. Firm and year-quarter fixed effects are included. Reported in the parentheses are the t-values based on industry and year-quarter clustered standard errors. All the control variables are described in the Data and Variable section, and their coefficients are omitted from the table for brevity.

	CR		CAR^{Mkt}		$CAR^{Carhart}$		CAR^{DGTW}	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Intercept	8.82***	9.16***	8.27***	8.59***	10.12***	10.44***	9.28***	9.59***
	(3.24)	(3.37)	(3.17)	(3.28)	(3.66)	(3.79)	(3.68)	(3.80)
I(Smart Surprise $\geq$	0.39**	-0.19	0.43**	-0.11	0.48***	-0.03	0.28	-0.48*
& Naive Surprise < 0)	(2.14)	(-0.80)	(2.50)	(-0.48)	(2.78)	(-0.16)	(1.44)	(-1.79)
I(Smart Surprise $\geq$		1.03***		0.96***		0.92***		1.21***
& Naive Surprise < 0) $\times$ IO^H		(2.97)		(2.81)		(2.74)		(2.74)
IO^H		-0.33		-0.30		-0.47		-0.26
		(-0.75)		(-0.75)		(-1.17)		(-0.51)
N	18436	18436	18436	18436	18437	18437	13595	13595
R^2_{adj}	0.06	0.06	0.07	0.07	0.08	0.08	0.07	0.07
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Columns (1) and (2) of Table 3.7 report the results with CR (cumulative raw returns) as the dependent variable. Without distinguishing firms with high or low institutional ownership, an intercept of 8.82 represents the average CAR (in percentage values) in the benchmark group where Smart Surprise < 0 and Naive Surprise ≥ 0. When Smart Surprise ≥ 0 and Naive Surprise < 0, the CR is 9.21%, which is significantly higher than the benchmark group by 0.39%. This indicates that investors use both Smart and Naive Consensus as expectations and their reaction to good earnings surprise always dominate their reaction to bad earnings surprise. After adding the interaction term, I(Smart Surprise ≥ 0 & Naive Surprise < 0) $\times$ IO^H, we observe that coefficient on it to be significantly positive indicating that investors in firms with high institutional ownerships are more likely to weight analysts' forecasts with respect to the analyst's accuracy. The results are very similar when the alternative three CARs are used as the dependent variable.

Trading Strategies

Finally, we determine whether a trading strategy can be formulated to exploit the investors' misweighting of analysts earnings forecasts and earn positive abnormal returns. The key is to implement a strategy that only uses information available at the point of trading.

Knowing that Expectation Gap can predict Naive Surprise, we sort firms into five portfolios with respect to Expectation Gap in each year-quarter. The trading strategy is formulated as long (short) stocks in the portfolio with highest (lowest) Expectation Gap at the end of day 't-2' before the earnings announcement, and hold the portfolio for three days till the end of day 't+1' after the earnings announcement. The CR and CARs obtained from such a long-short trading strategy are reported in Table 3.8.

Table 3.8. Abnormal Returns on A Trading Strategy Based on Expectation Gap

The four trading strategy returns: CR, CAR^{Mkt}, $CAR^{Carhart}$, and CAR^{DGTW}, are alternative measurements of three-day cumulative (abnormal) returns around earnings announcements. The performance of the trading strategy is estimated in three sets of samples: full sample (Panel A), the subsample with low institutional ownership (Panel B), and high institutional ownership (Panel C). Portfolio returns equally weighted and presented in percent values.

	Panel A: Full Sample					
	1 (Lowest)	2	3	4	5 (Highest)	5 - 1
CR, %	-0.12	0.19	0.32	0.33	0.20	0.32
	(-0.91)	(1.39)	(2.48)	(2.62)	(1.67)	(1.80)
CAR^{Mkt}, %	-0.24	0.05	0.18	0.19	0.08	0.32
	(-3.01)	(0.66)	(2.32)	(2.68)	(1.16)	(3.02)
$CAR^{Carhart}$, %	-0.12	0.05	0.08	0.06	-0.06	0.07
	(-2.13)	(0.71)	(1.29)	(1.15)	(-0.94)	(0.78)
CAR^{DGTW}, %	-0.25	0.09	0.14	0.18	0.05	0.30
	(-3.62)	(1.43)	(2.27)	(3.47)	(0.80)	(3.21)
	Panel B: Low Institutional Ownership					
	1 (Lowest)	2	3	4	5 (Highest)	5 - 1
CR, %	-0.45	-0.02	0.07	0.03	-0.01	0.45
	(-3.26)	(-0.11)	(0.50)	(0.24)	(-0.05)	(2.38)
CAR^{Mkt}, %	-0.55	-0.16	-0.06	-0.09	-0.12	0.42
	(-5.85)	(-1.40)	(-0.62)	(-1.06)	(-1.44)	(3.35)
$CAR^{Carhart}$, %	-0.37	-0.13	-0.12	-0.17	-0.19	0.18
	(-4.79)	(-1.37)	(-1.42)	(-2.46)	(-2.67)	(1.75)
CAR^{DGTW}, %	-0.61	-0.12	-0.18	-0.11	-0.14	0.47
	(-6.01)	(-1.29)	(-1.95)	(-1.30)	(-1.49)	(3.32)
	Panel C: High Institutional Ownership					
	1 (Lowest)	2	3	4	5 (Highest)	5 - 1
CR, %	0.27	0.38	0.63	0.55	0.41	0.14
	(1.94)	(2.86)	(4.91)	(3.97)	(3.03)	(-0.72)
CAR^{Mkt}, %	0.123	0.24	0.47	0.40	0.28	0.15
	(1.41)	(2.98)	(5.75)	(4.62)	(3.39)	(-1.29)
$CAR^{Carhart}$, %	0.18	0.21	0.34	0.24	0.10	-0.08
	(2.79)	(2.67)	(4.66)	(3.20)	(1.32)	(-0.85)
CAR^{DGTW}, %	0.06	0.23	0.43	0.35	0.19	0.13
	(0.89)	(2.92)	(6.36)	(5.51)	(2.92)	(-1.36)

Such a trading strategy generates 0.3% (significant at 1%) of CAR^{Mkt} and CAR^{DGTW} when all firms are considered. These profit increases to roughly 0.45% for firms with low institutional ownership. This observation is consistent with the finding that retail investors are more likely to misweight analysts' forecasts. None of the cumulative (abnormal) returns are significantly positive for the long-short trading strategy in firms with high institutional ownership. This is consistent with the findings that institutional investors are more sophisticated and have already incorporated the accuracy-based weights when they form earnings expectations. Therefore, a trading strategy based on the accuracy-based weights is no longer profitable for stocks mostly held by institutional investors.

Conclusion

In this chapter, we propose a new consensus based on analysts' prior earnings forecast accuracy. The traditional consensus forecast is the simple average of all outstanding analysts' forecasts, which is unsophisticated. Motivated by the existence of a cross-sectional variation in analysts' forecast accuracy and persistence in such accuracy, we propose to construct a consensus that is an accuracy-weighted average of all analysts' forecasts (Smart Consensus). We confirm that our Smart Consensus is a more accurate predictor of a firm's reported EPS than Naive Consensus. Further, Expectation Gap, which is the difference between Smart and Naive Consensuses, can predict Naive Surprise. Whether investors aggregate analysts' forecasts efficiently by assigning accuracy-based weights to analysts is a separate empirical question. We document that investors' reaction to earnings announcements, measured by three-day (abnormal) cumulative returns around earnings announcements, is

positively related to both Smart Surprise and Naive Surprise. It appears that while investors take analysts' accuracy into account when forming expectations, they do not weight analysts efficiently. We also find that the earnings announcement returns are related more to Naive Surprise in stocks with greater ownership by retail investors (who have greater difficulty processing data relative to institutional investors). Lastly, based on the ability of Expectation Gap to predict Naive Surprise, we implement a trading strategy in a three-day window around earnings announcements. It generates positive risk-adjusted returns, which is mainly driven by stocks with low institutional ownership.

In this chapter, the firm's accounting metric we consider is EPS, which is also the major focus of prior literature. Future studies should consider two alternative key metrics: sales and EBITDA. There are several reasons for including these two additional metrics. First, analysts track these metrics closely because the two primary drivers of EPS growth are sales growth and margin expansion. Second, Wall Street considers a firm that reports an EPS surprise but not a sales surprise as having lower earnings quality. Third, news articles on earnings announcement is likely to compare both reported sales and EPS to consensus numbers, with a comparison of reported margins to consensus coming in a close third. Fourth and finally, the market reaction to earnings announcement depends not only on whether the firms' reported EPS surprise consensus, but also on whether the firms' reported sales and margins surprise their respective consensus. It would be fruitful to study whether investors also efficiently aggregate each analyst's forecast on sales and EBITDA.

BIBLIOGRAPHY

Abarbanell, Jeffrey S., and Brian J. Bushee, 1997, Fundamental analysis, future earnings, and stock prices, *Journal of Accounting Research* 35, 1–24.

Abarbanell, Jeffery S., and Brian J. Bushee, 1998, Abnormal returns to a fundamental analysis strategy, *The Accounting Review* 73, 19–45.

Abarbanell, Jeffery S., William N. Lanen, and Robert E. Verrecchia, 1995, Analysts' forecasts and proxies for investor beliefs, *Journal of Accounting and Economics* 20, 31–60.

Adrian, Tobias, Erkko Etula, and Tyler Muir, 2014, Financial intermediaries and the cross-section of asset returns, *Journal of Finance* 69, 2557–2596.

Aretz, Kevin, Sohnke M. Bartram, and Peter F. Pope, 2010, Macroeconomic risks and characteristic-based factor models, *Journal of Banking and Finance* 34, 1383–1399.

Bai, Jushan, and Serena Ng, 2008, Forecasting economic time series using targeted predictors, *Journal of Econometrics* 146, 304–317.

Bailey, David H., and Marcos Lopez de Prado, 2014, The deflated sharpe ratio: Correcting for selection bias, backtest overfitting and non-normality, *Journal of Portfolio Management* 40, 94–107.

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

Baker, Malcolm, and Jeffrey Wurgler, 2000, The equity share in new issues and aggregate stock returns, *Journal of Finance* 55, 2219–2257.

Ball, Ray, and Philip Brown, 1968, An empirical evaluation of accounting index numbers, *Journal of Accounting Research* 6, 159–178.

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

Barron, Orie E., Oliver Kim, Steve C. Lim, and Douglas E. Stevens, 1998, Using analysts' forecasts to measure properties of analysts' information environment, *The Accounting Review* 73, 421–433.

Barth, Mary E., and Eric C. So, 2014, Non-diversifiable volatility rsk and risk premiums at earnings announcements, *The Accounting Review* 89, 1579–1607.

Barton, Jan, and Paul J. Simko, 2002, The balance sheet as an earnings management constraint., *The Accounting Review* 77, 1–27.

Bartsch, Viktoria-Sophie, Hubert Dichtl, Wolfgang Drobetz, and Andreas Neuhierl, 2018, Data snooping in equity premium prediction, Working Paper.

Basu, Sudipta, Burt Holland, and Fang Sun, 2012, Neglect of multiplicity when testing families of related hypotheses, Working Paper.

Basu, Sudipta, and Han-Up Park, 2014, Publication bias in recent empirical accounting research, Working Paper.

Bates, John M., and Clive W. J. Granger, 1969, The combination of forecasts, *Operational Research Quarterly* 20, 451–468.

Beaver, William, and Roland Dukes, 1972, Interperiod tax allocation, earnings expectations, and the behavior of security prices, *The Accounting Review* 42, 320 – 333.

Behn, Bruce K., Jong-Hag Choi, and Tony Kang, 2008, Audit quality and properties of analyst earnings forecasts, *The Accounting Review* 83, 327–349.

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

Bhat, Gauri, Ole-Kristian Hope, and Tony Kang, 2006, Does corporate governance transparency affect the accuracy of analyst forecast?, *Accounting and Finance* 46, 715–732.

Boivin, Jean, and Serena Ng, 2006, Are more data always better for factor analysis?, *Journal of Econometrics* 132, 169–194.

Boons, Martijn, 2016, State variables, macroeconomic activity, and the cross section of individual stocks, *Journal of Financial Economics* 119, 489–511.

Bradshaw, Mark T., Scott A. Richardson, and Richard G. Sloan, 2001, Do analysts and auditors use information in accruals?, *Journal of Accounting Research* 39, 45–74.

Breeden, Douglas T., Michael R. Gibbons, and Robert H. Litzenberger, 1989, Empirical test of the consumption-oriented CAPM, *Journal of Finance* 44, 231–262.

Brown, Lawrence D., Robert L. Hagerman, Paul A. Griffin, and Mark E. Zmijewski, 1987a, Security analyst superiority relative to univariate time-series models in forecasting quarterly earnings, *Journal of Accounting and Economics* 9, 61–87.

Brown, Lawrence D., and Kwon-Jung Kim, 1991, Timely aggregate analyst forecasts as better proxies for market earnings expectations, *Journal of Accounting Research* 29, 382–385.

Brown, Lawrence D., Gordon D. Richardson, and Steven J. Schwager, 1987, An information interpretation of financial analyst superiority in forecasting earnings, *Journal of Accounting Research* 25, 49–67.

Brown, Lawrence D., and Michael S. Rozeff, 1979, Univariate time-series models of quarterly accounting earnings per share: A proposed model, *Journal of Accounting Research* 17, 179–189.

Brown, Philip, George Foster, and Eric Noreen, 1985, Securify analyst multi-year earnings forecajts revisions and securify price changes, *American Accounting Association* 61.

Bryzgalova, Svetlana, 2016, Spurious factors in linear asset pricing models, *Working paper*.

Butler, Marty, Stanimir Markov, and Arthur G. Kraft, 2007, On the weighting of individual analyst forecasts in the consensus, *Working paper*.

Campbell, John Y., 1987, Stock returns and the term structure, *Journal of Financial Economics* 18, 373–399.

Cao, Jie, and Bing Han, 2016, Idiosyncratic risk, costly arbitrage, and the cross-section of stock returns, *Journal of Banking and Finance* 73, 1–15.

Carhart, Mark M., 1997, On persistence in mutual fund performance, *Journal of Finance* 52, 57–82.

Chan, Louis K. C., Stephen G. Dimmock, and Josef Lakonishok, 2009, Benchmarking money manager performance: Issues and evidence, *Review of Financial Studies* 22, 4553–4599.

Chen, Hailiang, Prabuddha De, Yu Hu, and Byoung-Hyoun Hwang, 2014, Wisdom of crowds: The value of stock opinions transmitted through social media, *Review of Financial Studies* 27, 1367–1403.

Chen, Long, 2009, On the reversal of return and dividend growth predictability: A tale of two periods, *Journal of Financial Economics* 92, 128–151.

Chen, Nai-Fu, Richard Roll, and Stephen A. Ross, 1986, Economic forces and the stock market, *Journal of Business* 59, 383–403.

Chen, Qi, Jennifer Francis, and Wei Jiang, 2005, Investor learning about analyst predictive ability, *Journal of Accounting and Economics* 39, 3–24.

Chen, Qi, and Wei Jiang, 2006, Analysts' weighting of private and public information, *The Review of Financial Studies* 19, 319–355.

Chinco, Alex, Adam D. Clark-Joseph, and Mao Ye, 2017, Sparse signals in the cross-section of returns, *Journal of Finance* (forthcoming).

Clark, Todd E., and Michael W. McCracken, 2001, Tests of equal forecast accuracy and encompassing for nested models, *Journal of Econometrics* 105, 85–110.

Clark, Todd E., and Kenneth D. West, 2007, Approximately normal tests for equal predictive accuracy in nested models, *Journal of Econometrics* 138, 291–311.

Clement, Michael B., 1999, Analyst forecast accuracy: Do ability, resources, and portfolio complexity matter?, *Journal of Accounting and Economics* 27, 285–303.

——— , and Senyo Y. Tse, 2003, Do investors respond to analysts' forecast revisions as if forecast accuracy is all that matters?, *The Accounting Review* 78, 227–249.

Cochrane, John H., 2005, *Asset Pricing* (Princeton University Press: Princeton, NJ).

——— , 2011, Presidential address: Discount rates, *Journal of Finance* 66, 1047–1108.

Cohen, Daniel A., Aiyesha Dey, Thomas Z. Lys, and Shyam V. Sunder, 2007, Earnings announcement premia and the limits to arbitrage, *Journal of Accounting and Economics* 43, 153–180.

Collins, Daniel W., 1975, Sec product line reporting and market efficiency, *Journal of Financial Economics* 2, 125–164.

Daniel, Kent, Mark Grinblatt, Sheridan Titman, and Russ Wermers, 1997, Measuring mutual fund performance with characteristic-based benchmarks, *Measuring Mutual Fund Performance with Characteristic-Based Benchmarks* 52, 1035–1058.

Daniel, Kent, David Hirshleifer, and Lin Sun, 2018, Short- and long-horizon behavioral factors, *Working paper*.

Das, Somnath, Carolyn B. Levine, and K. Sivaramakrishnan, 1998, Earnings predictability and bias in analysts' earnings forecasts, *The Accounting Review* 73, 277–294.

Dellavigna, Stefano, and Joshua M. Pollet, 2009, Investor inattention and friday earnings announcements, *The Journal of Finance* 64, 709–749.

DeMiguel, Victor, Alberto Martin-Utrera, Francisco J. Nogales, and Raman Uppal, 2018, A portfolio perspective on the multitude of firm characteristics, *Working paper*.

Dhaliwal, Dan S., Suresh Radhakrishnan, Albert Tsang, and Yong George Yang, 2012, Nonfinancial disclosure and analyst forecast accuracy: International evidence on corporate social responsibility disclosure, *The Accounting Revew* 87, 723–759.

Elton, Edwin J., Martin J. Gruber, and Mustafa Gultekin, 1981, Expectations and share prices, *Management Science* 27, 975–987.

Fama, Eugene F., and Kenneth R. French, 1989, Business conditions and expected returns on stocks and bonds, *Journal of Financial Economics* 25, 23–49.

——— , 1993, Common risk factors in the returns on stocks and bonds, *Journal of Financial Economics* 33, 3–56.

——— , 2008, Dissecting anomalies, *Journal of Finance* 63, 1653–1678.

——— , 2015, Incremental variables and the investment opportunity set, *Journal of Financial Economics* 117, 470–488.

Feng, Guanhao, Stefano Giglio, and Dacheng Xiu, 2017, Taming the factor zoo, *Working paper*.

Ferson, Wayne E., Sergei Sarkissian, and Timothy Simin, 1999, The alpha factor asset pricing model: A parable, *Journal of Financial Markets* 2, 49–68.

Ferson, Wayne E., Sergei Sarkissian, and Timothy T. Simin, 2003, Spurious regressions in financial economics?, *Journal of Finance* 58, 1393–1413.

Forni, Mario, Marc Hallin, Marco Lippi, and Lucrezia Reichlin, 2003, Do financial variables help forecasting inflation and real activity in the euro area?, *Journal of Monetary Economics* 50, 1243–1255.

Foster, F. Douglas, Tom Smith, and Robert E. Whaley, 1997, Assessing goodness-of-fit of asset pricing models: The distribution of the maximal R^2, *Journal of Finance* 52, 591–607.

Foster, George, 1977, Quarterly accounting data: Time-series properties and predictiveability results, *The Accounting Review* 52, 1–21.

Frankel, Richard, and Charles M. C. Lee, 1998, Accounting valuation, market expectation, and cross-sectional stock returns, *Journal of Accounting and Economics* 25, 283–319.

Freyberger, Joachim, Andreas Neuhierl, and Michael Weber, 2018, Dissecting characteristics nonparametrically, *Working paper*.

Fried, Dov, and Dan Givoly, 1982, Financial analysts' forecasts of earnings: A better surrogate for market expectations, *Journal of Accounting and Economics* 4, 85–107.

Gleason, Cristi A., and Charles M. C. Lee, 2003, Analyst forecast revisions and market price discovery, *The Accounting Review* 78, 193–225.

Gow, Ian D., Gaizka Ormazabal, and Daniel J. Taylor, 2010, Correcting for cross-sectional and time-series dependence in accounting research, *The Accounting Review* 85, 483–512.

Green, Jeremiah, John R. M. Hand, and X. Frank Zhang, 2013, The supraview of return predictive signals, *Review of Accounting Studies* 18, 692–730.

——— , 2017, The characteristics that provide independent information about average U.S. monthly stock returns, *Review of Financial Studies* 30, 4389–4436.

Griffin, John M., Xiuqing Ji, and J. Spencer Martin, 2003, Momentum investing and business cycle risk: Evidence from pole to pole, *Journal of Finance* 58, 2515–2547.

Griffin, Pa, 1997, Time-series of behavior of quarterly earnings - preliminary evidence, *Journal of Accounting Research* 15, 71–83.

Grinold, Richard, 2010, Signal weighting, *Journal of Portfolio Management* 36, 24–34.

Grover, Sean, and Michael W. McCracken, 2014, Factor-based prediction of industry-wide bank stress, *Federal Reserve Bank of St. Louis Review* 96, 173–193.

Gu, Shihao, Bryan Kelly, and Dacheng Xiu, 2018, Empirical asset pricing via machine learning, *Working paper*.

Haddad, Valentin, Serhiy Kozak, and Shrihari Santosh, 2017, Predicting relative returns, Working Paper.

Han, Yufeng, Ai He, David E. Rapach, and Guofu Zhou, 2018, What firm characteristics drive US stock returns?, *Working paper*.

Han, Yufeng, Guofu Zhou, and Yingzi Zhu, 2016, A trend factor: Any economic gains from using information over investment horizons?, *Journal of Financial Economics* 122, 352–375.

Hanna, J. Douglas, and Mark J. Ready, 2005, Profitable predictability in the cross section of stock returns, *Journal of Financial Economics* 78, 463–505.

Hansen, Peter Reinhard, 2005, A test for superior predictive ability, *Journal of Business & Economic Statistics* 23, 365–380.

Harvey, Campbell R., 2017, Presidential address: The scientific outlook in financial economics, *Journal of Finance* 72, 1399–1440.

——— , and Yan Liu, 2014, Evaluating trading strategies, *Journal of Portfolio Management* 40, 108–118.

——— , 2015, Backtesting, *Journal of Portfolio Management* 42, 13–28.

——— , and Heqing Zhu, 2016, ... and the cross-section of expected returns, *Review of Financial Studies* 29, 5–68.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman, 2016, *The elements of statistical learning: Data mining, inference, and prediction* (Springer: New York, NY) 2nd edition edn.

Haugen, Robert A., and Nardin L. Baker, 1996, Commonality in the determinants of expected stock returns, *Journal of Financial Economics* 41, 401–439.

He, Zhiguo, Bryan Kelly, and Asaf Manela, 2017, Intermediary asset pricing: New evidence from many asset classes, *Journal of Financial Economics* 126, 1–35.

Head, Megan L., Luke Holman, Rob Lanfear, Andrew T. Kahn, and Michael D. Jennions, 2015, The extent and consequences of p-hacking in science, *PLoS Biology* 13, e1002106.

Hilary, Gilles, and Charles Hsu, 2013, Analyst forecast consistency, *The Journal of Finance* 68, 271–297.

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

Hope, Ole-Kristian, 2003, Disclosure practices, enforcement of accounting standards, and analysts' forecast accuracy: An international study, *Journal of Accounting Research* 41, 235–272.

Hou, Kewei, Chen Xue, and Lu Zhang, 2015, Digesting anomalies: An investment approach, *Review of Financial Studies* 28, 650–705.

——— , 2017, A comparison of new factor models, Working Paper.

Huang, Dashan, Fuwei Jiang, Jun Tu, and Guofu Zhou, 2015, Investor sentiment aligned: A powerful predictor of stock returns, *Review of Financial Studies* 28, 791–837.

Jackson, Andrew R., 2005, Trade generation, reputation, and sell-side analysts, *The Journal of Finance* 60, 673–717.

Jacob, John, Thomas Z. Lys, and Margaret A. Neal, 1999, Expertise in forecasting performance of security analysts, *Journal of Accounting and Economics* 28, 51–82.

Jacobs, Bruce I., and Kenneth N. Levy, 2014, Investing in a multidimensional market, *Financial Analysts Journal* 70, 6–12.

Jagannathan, Ravi, and Zhenyu Wang, 2002, Empirical evaluation of asset-pricing models: A comparison of the SDF and beta methods, *Journal of Finance* 57, 2337–2367.

Jiang, Fuwei, Guohao Tang, and Guofu Zhou, 2018, Firm characteristics and Chinese stocks, Working Paper.

Kelly, Bryan, and Seth Pruitt, 2013, Market expectations in the cross section of present values, *Journal of Finance* 68, 1721–1756.

——— , 2015, The three-pass regression filter: A new approach to forecasting using many predictors, *Journal of Econometrics* 186, 294–316.

——— , and Yinan Su, 2018, Characteristics are covariances: A unified model of risk and return, *Journal of Financial Economics* (forthcoming).

Kim, Oliver, Steve C. Lim, and Kenneth W. Shaw, 2001, The inefficiency of the mean analyst forecast as a summary forecast of earnings, *Journal of Accounting Research* 39, 329–336.

Kogan, Leonid, and Mary Tian, 2015, Firm characteristics and empirical factor models: A model-mining experiment, Working Paper.

Kothari, S.P., and Jay Shanken, 1997, Book-to-market, dividend yield, and expected market returns: A time-series analysis, *Journal of Financial Economics* 44, 169–203.

Kozak, Serhiy, Stefan Nagel, and Shrihari Santosh, 2017a, Shrinking the cross section, *Working paper*.

——— , 2018, Interpreting factor models, *Journal of Finance* 73, 1183–1223.

Lakonishok, Josef, Andrei Shleifer, and Robert W. Vishny, 1994, Contrarian investment, extrapolation, and risk, *The Journal of Finance* 49, 1541–1578.

Lang, Mark H., and Russell J. Lundholm, 1996, Corporate disclosure policy and analyst behavior, *The Accounting Review* 71, 467–492.

LaPorta, Rafael, 1996, Expectations and the cross-section of stock returns, *The Journal of Finance* 51, 1715–1742.

Lev, Baruch, and S. Ramu Thiagarajan, 1993, Fundamental information analysis, *Journal of Accounting Research* 31, 190–215.

Lewellen, Jonathan, 2015, The cross section of expected stock returns, *Critical Finance Review* 4, 1–44.

Liew, Jimmy, and Maria Vassalou, 2000, Can book-to-market, size and momentum be risk factors that predict economic growth?, *Journal of Financial Economics* 57, 221–245.

Light, Nathaniel, Denys Maslov, and Oleg Rytchkov, 2017, Aggregation of information about the cross section of stock returns: A latent variable approach, *Review of Financial Studies* 30, 1339–1381.

Lin, Hai, Chunchi Wu, and Guofu Zhou, 2016, Forecasting corporate bond returns: An iterated combination approach, *Management Science* (forthcoming).

Linnainmaa, Juhani T., and Michael R. Roberts, 2018, The history of the cross section of stock returns, *Review of Financial Studies* 31, 2606–2649.

Liu, Laura Xiaolei, and Lu Zhang, 2008, Momentum profits, factor pricing, and macroeconomic risk, *Review of Financial Studies* 21, 2417–2448.

Livnat, Joshua, and Richard R. Mendenhall, 2006, Comparing the post–earnings announcement drift for surprises calculated from analyst and time series forecasts, *Journal of Accounting*

Research 44, 177–205.

Lo, Andrew W., and A. Craig MacKinlay, 1990, Data-snooping biases in tests of financial asset pricing models, *Review of Financial Studies* 3, 431–467.

Loh, Roger K., and G. Mujtaba Mian, 2006, Do accurate earnings forecasts facilitate superior investment recommendations?, *Journal of Financial Economics* 80, 455–483.

Ludvigson, Sydney C, and Serena Ng, 2007, The empirical risk–return relation: A factor analysis approach, *Journal of Financial Economics* 83, 171–222.

Ma, Guang, and Stanimir Markov, 2017, The market's assessment of the probability of meeting or beating the consensus, *Contemporary Accounting Research* 34, 314–342.

Maio, Paulo, and Pedro Santa-Clara, 2012, Multifactor models and their consistency with the ICAPM, *Journal of Financial Economics* 106, 586–613.

Masicampo, E. J., and Daniel R. Lalande, 2012, A peculiar prevalence of p values just below .05, *Quarterly Journal of Experimental Psychology* 65, 2271–2279.

Matsumoto, Dawn A., 2002, Management's incentives to avoid negative earnings surprises, *The Accounting Review* 77, 483–514.

McCracken, Michael W., and Serena Ng, 2015, FRED-MD: A monthly database for macroeconomic research, *Working paper*.

Merton, Robert C., 1973, An intertemporal capital asset pricing model, *Econometrica* 41, 867–887.

Mikhail, Michael B., Beverly R. Walther, and Richard H. Willis, 1997, Do security analysts improve their performance with experience?, *Journal of Accounting Research* 35, 131–157.

Mikhail, Michael B, Beverly R Walther, and Richard H Willis, 1999, Does forecast accuracy matter to security analysts?, *The Accounting Review* 74, 185–200.

Miller, Rupert G., 1981, *Simultaneous statistical inference* (Springer-Verlag: New York) 2nd edn.

Mohanram, Partha S., 2005, Separating winners from losers among low book-to-market stocks using financial statement analysis, *Review of Accounting Studies* 10, 133–170.

Nagel, Stefan, 2013, Empirical cross-sectional asset pricing, *Annual Review of Financial Economics* 5, 167–199.

Neely, Christopher J., David E. Rapach, Jun Tu, and Guofu Zhou, 2014, Forecasting the equity risk premium: The role of technical indicators, *Management Science* 60, 1772–1791.

Neuhierl, Andreas, and Bernd Schlusche, 2011, Data snooping and market-timing rule performance, *Journal of Financial Econometrics* 9, 550–587.

Novy-Marx, Robert, 2016, Backtesting strategies based on multiple signals, Working Paper.

O'Brien, Patricia C., 1988, Analysts' forecasts as earnings expectations, *Journal of Accounting and Economics* 10, 53–83.

Ou, Jane A., and Stephen H. Penman, 1989, Financial statement analysis and the prediction of stock returns, *Journal of Accounting and Economics* 11, 295–329.

Park, Chul W., and Earl K. Stice, 2000, Analyst forecasting ability and the stock price reaction to forecast revisions, *Review of Accounting Studies* 5, 259–272.

Patell, James M., 1976, Corporate forecasts of earnings per share and stock price behavior: Empirical tests, *Journal of Accounting Research* 14, 246–276.

Payne, Jeff L, and Sean W. G Robb, 2000, Earnings management: The effect of ex ante earnings expectations, *Journal of Accounting* 15, 371–392.

Petersen, Mitchell A., 2009, Estimating standard errors in finance panel data sets: Comparing approaches, *Review of Financial Studies* 22, 435–480.

Petkova, Ralitsa, 2006, Do the Fama-French factors proxy for innovations in predictive variables, *Journal of Finance* 61, 581–612.

Piotroski, Joseph D., 2000, Value investing: The use of historical financial statement information to separate winners from losers, *Journal of Accounting Research* 38, 1–41.

Pontiff, Jeffrey, and Lawrence D. Schall, 1998, Book-to-market ratios as predictors of market returns, *Journal of Financial Economics* 49, 141–160.

Rapach, David, and Guofu Zhou, 2019, Sparse macro factors, *Working paper*.

Rapach, David E., Jack K. Strauss, and Guofu Zhou, 2010, Out-of-sample equity premium prediction: Combination forecasts and links to the real economy, *Review of Financial Studies* 23, 821–862.

Rapach, David E., and Guofu Zhou, 2013, Forecasting stock returns, in Graham Elliott, and Allan Timmermann, ed.: *Handbook of Economic Forecasting* vol. 2A . pp. 328–383 (Elsevier: Amsterdam).

Ready, Mark J., 2002, Profits from technical trading rules, *Financial Management* 31, 43–61.

Romano, Joseph P., and Michael Wolf, 2005, Stepwise multiple testing as formalized data snooping, *Econometrica* 73, 1237–1282.

Rosenthal, Robert, 1979, The "file drawer problem" and tolerance for null results, *Psychological Bulletin* 86, 638–641.

Sarkar, Sanat K., 2006, False discovery and false nondiscovery rates in single-step multiple testing procedures, *Annals of Statistics* 34, 394–415.

Savor, Pavel, and Mungo Wilson, 2014, Asset pricing: A tale of two days, *Journal of Financial Economics* 113, 171–201.

——— , 2016, Earnings announcements and systematic risk, *The Journal of Finance* 71, 83–138.

Schmidt, Lawrence D. W., 2016, Climbing and falling off the ladder: Asset pricing implications of labor market event risk, Working Paper.

Simonsohn, Uri, Leif D. Nelson, and Joseph P. Simmons, 2014, P-curve: A key to the file-drawer, *Journal of Experimental Psychology* 143, 534–547.

Sinha, Praveen, Lawrence D. Brown, and Somnath Das, 1997, A re-examination of financial analysts' differential earnings forecast accuracy, *Contemporary Accounting Research* 14, 1–42.

Skinner, Douglas J., 1994, Why firms voluntarily disclose bad news, *Journal of Accounting Research* 32, 1994).

Sloan, Richard G., 1996, Do stock prices fully reflect information in accruals and cash flows about future earnings?, *The Accounting Review* 71, 289–315.

So, Eric C., 2013, A new approach to predicting analyst forecast errors: Do investors overweight analyst forecasts?, *Journal of Financial Economics* 108, 615–640.

Soffer, Leonard, S. Ramu Thiagarajan, and Beverly Walther, 2000, Earnings preannouncement strategies, *Review of Accounting Studies* 5, 5–26.

Sorensen, Eric H., Edward Qian, Robert Schoen, and Ronald Hua, 2004, Multiple alpha sources and active management, *Journal of Portfolio Management* 30, 39–45.

Stambaugh, Robert F., Jianfeng Yu, and Yu Yuan, 2015, Arbitrage asymmetry and the idiosyncratic volatility puzzle, *Journal of Finance* 70, 1903–1948.

Stambaugh, Robert F., and Yu Yuan, 2017, Mispricing factors, *Review of Financial Studies* 30, 1270–1315.

Sterling, Theodor D., Wilfred L. Rosenbaum, and James J. Weinkam, 1995, Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa, *American Statistician* 49, 108–112.

Stickel, Scott E., 1992, Reputation and performance among security analysts, *The Journal of Finance* 47, 1811–1836.

Stock, James H., and Mark W. Watson, 1999, Forecasting inflation, *Journal of Monetary Economics* 44, 293–335.

——— , 2002a, Forecasting using principal components from a large number of predictors, *Journal of the American Statistical Association* 97, 1167–1179.

——— , 2002b, Macroeconomic forecasting using diffusion indexes, *Journal of Business & Economic Statistics* 20, 147–162.

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

——— , 2001, Dangers of data mining: The case of calendar effects in stock returns, *Journal of Econometrics* 105, 249–286.

Tibshirani, Robert, 1996, Regression shrinkage and selection via the Lasso, *Journal of the Royal Statistical Society* 58, 267–288.

Timmermann, Allan, 2006, Forecast combinations, in Clive W. J Granger, Graham Elliott, and Allan Timmermann, ed.: *Handbook of Economic Forecasting*vol. 1 . pp. 135–196 (Elsevier: Amsterdam).

Vassalou, Maria, 2003, News related to future GDP growth as a risk factor in equity returns, *Journal of Financial Economics* 68, 47–73.

Watts, Ross L., and Richard W. Leftwich, 1977, The time series of annual accounting earnings, *Journal of Accounting Research* 15, 253–271.

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

West, Kenneth D., 2006, Forecast evaluation, in Clive W.J. Granger, Graham Elliott, and Allan Timmermann, ed.: *Handbook of Economic Forecasting*vol. 1 . pp. 99–134 (Elsevier B.V.).

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

Yan, Xuemin (Sterling), and Lingling Zheng, 2017, Fundamental analysis and the cross-section of stock returns: A data-mining approach, *Review of Financial Studies* 30, 1382–1423.

Zou, Hui, and Trevor Hastie, 2005, Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society* 67, 301–320.

APPENDIX

Table A.1. Description of 131 Macroeconomic Shocks

This table lists the series numbers, names, data transformations, and short descriptions of the 131 macroeconomic shocks. The names and descriptions directly follow McCracken and Ng (2015). "tcode" represents the data transformations: (1)=first order difference, (2)=second order difference, (3)=first order difference on log value, (4)=second order difference on log value, (5)= first order difference in rate of change.

id	variable	tcode	description
		Category 1: Output and income	
1	RPI	3	Real Personal Income
2	W875RX1	3	Real Personal Income Ex Transfer Receipts
3	INDPRO	3	IP Index
4	IPFPNSS	3	IP: Final Products and Nonindustrial Supplies
5	IPFINAL	3	IP: Final Products (Market Group)
6	IPCONGD	3	IP: Consumer Goods
7	IPDCONGD	3	IP: Durable Consumer Goods
8	IPNCONGD	3	IP: Nondurable Consumer Goods
9	IPBUSEQ	3	IP: Business Equipment
10	IPMAT	3	IP: Materials
11	IPDMAT	3	IP: Durable Materials
12	IPNMAT	3	IP: Nondurable Materials
13	IPMANSICS	3	IP: Manufacturing (SIC)
14	IPB51222S	3	IP: Residential Utilities
15	IPFUELS	3	IP: Fuels
16	NAPMPI	1	ISM Manufacturing: Production Index
17	CUMFNS	1	Capacity Utilization: Manufacturing
		Category 2: Labor market	
18	HWI	1	Help-Wanted Index for United States
19	HWIURATIO	1	Ratio of Help Wanted/No. Unemployed
20	CLF16OV	3	Civilian Labor Force
21	CE16OV	3	Civilian Employment
22	UNRATE	1	Civilian Unemployment Rate
23	UEMPMEAN	1	Average Duration of Unemployment (Weeks)
24	UEMPLT5	3	Civilians Unemployed - Less Than 5 Weeks
25	UEMP5TO14	3	Civilians Unemployed for 5-14 Weeks
26	UEMP15OV	3	Civilians Unemployed - 15 Weeks & Over
27	UEMP15T26	3	Civilians Unemployed for 15-26 Weeks
28	UEMP27OV	3	Civilians Unemployed for 27 Weeks and Over
29	CLAIMSx	3	Initial Claims
30	PAYEMS	3	All Employees: Total nonfarm
31	USGOOD	3	All Employees: Goods-Producing Industries
32	CES1021000001	3	All Employees: Mining and Logging: Mining
33	USCONS	3	All Employees: Construction
34	MANEMP	3	All Employees: Manufacturing
35	DMANEMP	3	All Employees: Durable goods

Table A.1 – *Continued from previous page*

id	variable	tcode	description
36	NDMANEMP	3	All Employees: Nondurable goods
37	SRVPRD	3	All Employees: Service-Providing Industries
38	USTPU	3	All Employees: Trade, Transportation & Utilities
39	USWTRADE	3	All Employees: Wholesale Trade
40	USTRADE	3	All Employees: Retail Trade
41	USFIRE	3	All Employees: Financial Activities
42	USGOVT	3	All Employees: Government
43	CES0600000007	1	Avg Weekly Hours : Goods-Producing
44	AWOTMAN	1	Avg Weekly Overtime Hours : Manufacturing
45	AWHMAN	1	Avg Weekly Hours : Manufacturing
46	NAPMEI	1	ISM Manufacturing: Employment Index
47	CES0600000008	4	Avg Hourly Earnings : Goods-Producing
48	CES2000000008	4	Avg Hourly Earnings : Construction
49	CES3000000008	4	Avg Hourly Earnings : Manufacturing
			Category 3: Housing
50	HOUST	3	Housing Starts: Total New Privately Owned
51	HOUSTNE	3	Housing Starts, Northeast
52	HOUSTMW	3	Housing Starts, Midwest
53	HOUSTS	3	Housing Starts, South
54	HOUSTW	3	Housing Starts, West
55	PERMIT	3	New Private Housing Permits (SAAR)
56	PERMITNE	3	New Private Housing Permits, Northeast (SAAR)
57	PERMITMW	3	New Private Housing Permits, Midwest (SAAR)
58	PERMITS	3	New Private Housing Permits, South (SAAR)
59	PERMITW	3	New Private Housing Permits, West (SAAR)
			Category 4: Consumption, orders, and inventories
60	DPCERA3M086SBEA	3	Real personal consumption expenditures
61	CMRMTSPLx	3	Real Manu. and Trade Industries Sales
62	RETAILx	3	Retail and Food Services Sales
63	NAPM	1	ISM : PMI Composite Index
64	NAPMNOI	1	ISM : New Orders Index
65	NAPMSDI	1	ISM : Supplier Deliveries Index
66	NAPMII	1	ISM : Inventories Index
67	AMDMNOx	3	New Orders for Durable Goods
68	AMDMUOx	3	Unfilled Orders for Durable Goods
69	BUSINVx	3	Total Business Inventories
70	ISRATIOx	1	Total Business: Inventories to Sales Ratio
			Category 5: Money and credit
71	M1SL	4	M1 Money Stock
72	M2SL	4	M2 Money Stock
73	M2REAL	3	Real M2 Money Stock
74	AMBSL	4	St. Louis Adjusted Monetary Base
75	TOTRESNS	4	Total Reserves of Depository Institutions
76	NONBORRES	5	Reserves Of Depository Institutions
77	BUSLOANS	4	Commercial and Industrial Loans
78	REALLN	4	Real Estate Loans at All Commercial Banks
79	NONREVSL	4	Total Nonrevolving Credit
80	CONSPI	1	Nonrevolving consumer credit to Personal Income
81	MZMSL	4	MZM Money Stock

Table A.1 – *Continued from previous page*

id	variable	tcode	description
82	DTCOLNVHFNM	4	Consumer Motor Vehicle Loans Outstanding
83	DTCTHFNM	4	Total Consumer Loans and Leases Outstanding
84	INVEST	4	Securities in Bank Credit at All Commercial Banks N.A.
			Category 6: Interest and exchange rates
85	FEDFUNDS	1	Effective Federal Funds Rate
86	CP3Mx	1	3-Month AA Financial Commercial Paper Rate
87	TB3MS	1	3-Month Treasury Bill
88	TB6MS	1	6-Month Treasury Bill
89	GS1	1	1-Year Treasury Rate
90	GS5	1	5-Year Treasury Rate
91	GS10	1	10-Year Treasury Rate
92	AAA	1	Moody's Seasoned Aaa Corporate Bond Yield
93	BAA	1	Moody's Seasoned Baa Corporate Bond Yield
94	COMPAPFFx	1	3-Month Commercial Paper Minus FEDFUNDS
95	TB3SMFFM	1	3-Month Treasury C Minus FEDFUNDS
96	TB6SMFFM	1	6-Month Treasury C Minus FEDFUNDS
97	T1YFFM	1	1-Year Treasury C Minus FEDFUNDS
98	T5YFFM	1	5-Year Treasury C Minus FEDFUNDS
99	T10YFFM	1	10-Year Treasury C Minus FEDFUNDS
100	AAAFFM	1	Moody's Aaa Corporate Bond Minus FEDFUNDS
101	BAAFFM	1	Moody's Baa Corporate Bond Minus FEDFUNDS
102	EXSZUSx	3	Switzerland / U.S. Foreign Exchange Rate
103	EXJPUSx	3	Japan / U.S. Foreign Exchange Rate
104	EXUSUKx	3	U.S. / U.K. Foreign Exchange Rate
105	EXCAUSx	3	Canada / U.S. Foreign Exchange Rate
			Category 7: Prices
106	WPSFD49207	4	PPI: Finished Goods
107	WPSFD49502	4	PPI: Finished Consumer Goods
108	WPSID61	4	PPI: Intermediate Materials
109	WPSID62	4	PPI: Crude Materials
110	OILPRICEx	4	Crude Oil, spliced WTI and Cushing
111	PPICMM	4	PPI: Metals and metal products
112	NAPMPRI	1	ISM Manufacturing: Prices Index
113	CPIAUCSL	4	CPI : All Items
114	CPIAPPSL	4	CPI : Apparel
115	CPITRNSL	4	CPI : Transportation
116	CPIMEDSL	4	CPI : Medical Care
117	CUSR0000SAC	4	CPI : Commodities
118	CUUR0000SAD	4	CPI : Durables
119	CUSR0000SAS	4	CPI : Services
120	CPIULFSL	4	CPI : All Items Less Food
121	CUUR0000SA0L2	4	CPI : All items less shelter
122	CUSR0000SA0L5	4	CPI : All items less medical care
123	PCEPI	4	Personal Cons. Expend.: Chain Index
124	DDURRG3M086SBEA	4	Personal Cons. Exp: Durable goods
125	DNDGRG3M086SBEA	4	Personal Cons. Exp: Nondurable goods
126	DSERRG3M086SBEA	4	Personal Cons. Exp: Services
			Category 8: Stock market
127	S&P 500	3	S&P's Common Stock Price Index: Composite

Table A.1 – *Continued from previous page*

id	variable	tcode	description
128	S&P Indust	3	S&P's Common Stock Price Index: Industrials
129	S&P DivYield	1	S&P's Composite Common Stock: Dividend Yield
130	S&P PERatio	3	S&P's Composite Common Stock: Price-Earnings Ratio
131	VXOCLSx	1	VXO

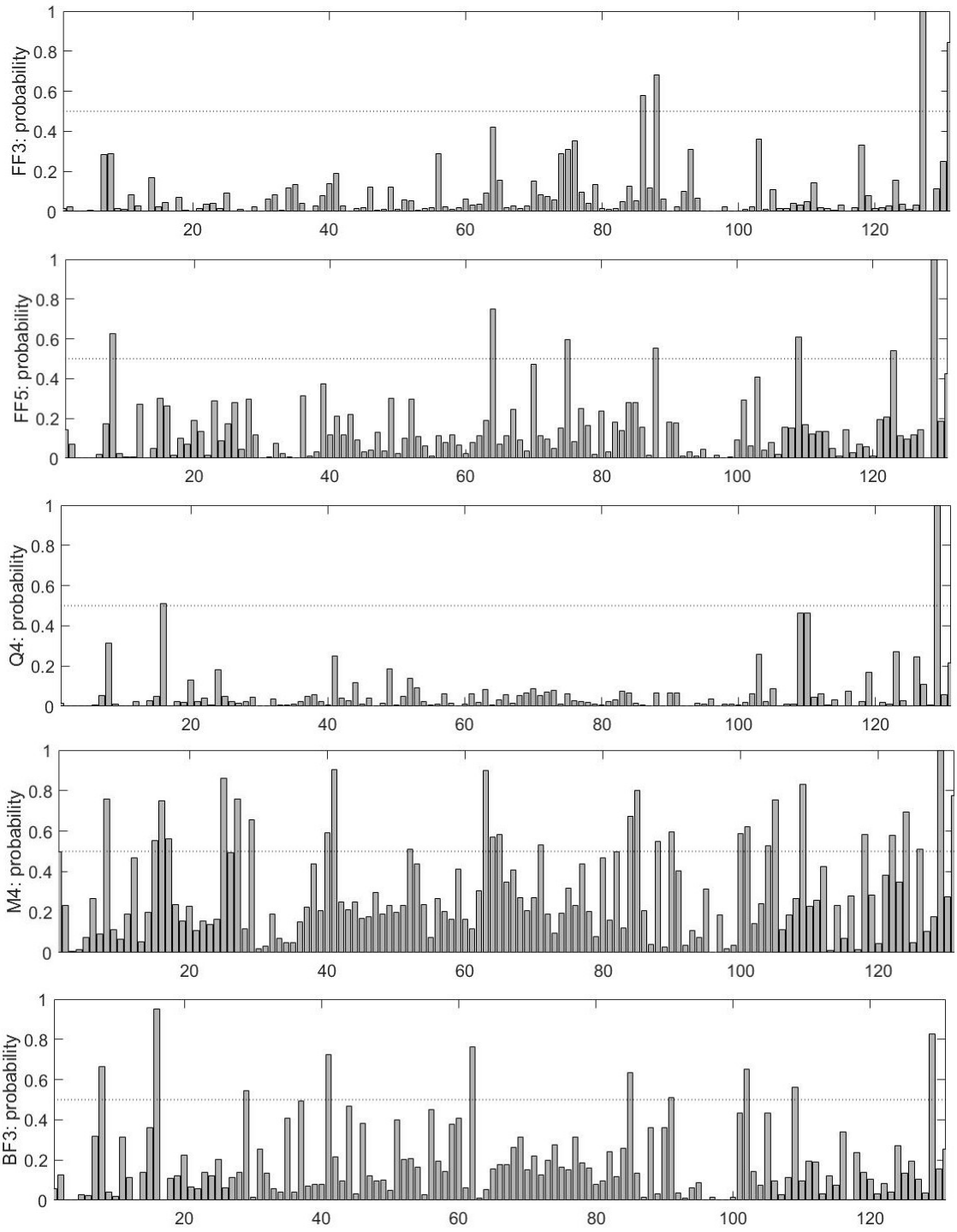

Figure A1. Variable Selection by Elastic Net

Y-axis of the figure shows the probability that a particular macroeconomic shock is selected in an elastic net regression of SDF on 131 macroeconomic shocks. SDFs of five characteristic-based factor models are considered, that is, FF3, FF5, Q4, M4, and BF3. The macroeconomic shocks are listed on the horizontal axis following the ascending order of their series numbers. The series numbers, names, short descriptions, and variable transformations are presented in Table A1.