5-26-2021 10:00 AM

Essays in Financial Econometrics and Machine Learning

Fred Liu, *The University of Western Ontario*

Supervisor: Stentoft, Lars, *The University of Western Ontario*
: Conley, Tim, *The University of Western Ontario*
: Saunders, Charles, *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy degree in Economics

Abstract

Financial econometrics is a highly interdisciplinary field that integrates finance, economics, probability, statistics, and applied mathematics. Machine learning is a growing area in finance that is particularly suitable for studying problems with many variables. My thesis contains three chapters that explore financial econometrics and machine learning in the fields of asset pricing and risk management.

Chapter 2 studies the implications of the new Basel 3 regulations. In 2019, the BCBS finalized the Basel 3 regulatory regime, which changes the regulatory measure of market risk and adds new complex calculations based on liquidity and risk factors. This chapter is motivated by these changes and seeks to answer the question of how regulation affects banks' choice of risk-management models, whether it incentivizes them to use correctly specified models, and if it results in more stable capital requirements.

Chapter 3 conducts, to our knowledge, the largest study ever of five-minute equity market returns using state-of-the-art machine learning models trained on the cross-section of lagged market index constituent returns, where we show that regularized linear models and nonlinear tree-based models yield significant market return predictability. Ensemble models perform the best across time and their predictability translates into economically significant Sharpe ratios of 0.98 after transaction costs. These results provide strong evidence that intraday market returns are predictable during short time horizons.

Chapter 4 studies the idiosyncratic tail risk premium and common factor. Stocks in the highest idiosyncratic tail risk decile earn 8% higher average annualized returns than in the lowest. I propose a risk-based explanation for this premium, in which shocks to intermediary funding cause idiosyncratic tail risk to follow a strong factor structure, and the factor, common idiosyncratic tail risk (CITR), comoves with intermediary funding. Consequently, firms with high idiosyncratic tail risk have high exposure to CITR shocks, and command a risk premium due to their low returns when intermediary constraints tighten. To test my explanation, I create a novel measure of idiosyncratic tail risk. Consistent with my explanation, CITR shocks

are procyclical, are correlated to intermediary factors, are priced in assets, and explain the idiosyncratic tail risk premium.

Keywords: Financial Econometrics, Machine Learning, Asset Pricing, Risk Management, Tail Risk

Summary for Lay Audience

Financial econometrics is a highly interdisciplinary field that integrates finance, economics, probability, statistics, and applied mathematics. Machine learning is a growing area in finance that is particularly suitable for studying problems with many variables. My thesis contains three chapters that explore financial econometrics and machine learning in the fields of asset pricing and risk management.

Chapter 2 is motivated by the new Basel 3 market risk regulation, which introduces new complex calculations for global banks. This chapter has three main findings. First, under Basel 3, banks are incentivized towards riskier models. Second, banks are incentivized toward inaccurate models meaning that the Basel 3 penalty for inaccuracy may be insufficient. Third, Basel 3 results in more stable capital requirements.

Chapter 3 is motivated by the idea that markets may be predictable in very short time horizons, since it takes time for traders to incorporate information into prices. To test this idea, we conduct the largest study of intraday (i.e. five-minute) market return predictability using machine learning techniques. This chapter has three main findings. First, intraday market returns are predictable, though this predictability has decreased across the years. Second, this predictability is profitable after transaction costs. Third, consistent with slow traders, predictability is higher during the middle of the day, and during volatile or illiquid days.

Chapter 4 uses a new tail risk factor to provide an economic explanation for a recent asset pricing puzzle, which uncovers a hidden risk that emerges during bad times. Specifically, this chapter is on idiosyncratic tail risk, which measures the severe losses of an individual stock that are uncorrelated to the market. I propose that idiosyncratic tail risk is caused by large investment firms impacting individual stocks due to the large size of their trades. When they're flush with cash, they conduct more trades, causing tail risk to go up. This means the aggregate level of tail risk is informative of how much cash big investment firms have. As these are key players in financial markets, then idiosyncratic tail risk matters for prices and financial stability.

Co-Authorship Statement

This thesis contains co-authored materials. Chapter 2 is co-authored with Lars Stentoft. Chapter 3 is co-authored with Dillon Huddleston and Lars Stentoft. All authors are equally responsible for the work.

A version of Chapter 2 has been published in the Journal of Financial Econometrics with the following citation: "Fred Liu, Lars Stentoft, Regulatory Capital and Incentives for Risk Model Choice under Basel 3*, Journal of Financial Econometrics, Volume 19, Issue 1, Winter 2021, Pages 53–96." A version of Chapter 3 is forthcoming in the Journal of Financial Econometrics. A version of Chapter 4 has been submitted to the Review of Asset Pricing Studies via Dual Submission.

Acknowledgments

It has been my great privilege to study in the Department of Economics of the University of Western Ontario, as it provides an excellent study and research environment built by its dedicated faculty and staff members, to whom I want to express my sincere gratitude.

First and foremost, I am extremely grateful to my supervisor Professor Lars Stentoft, for his invaluable advice, continuous encouragement, and great patience during my PhD study. He always makes himself available whenever I need guidance, and encourages me for whatever I achieve. I have been very fortunate to work with Professor Lars Stentoft on joint projects, where he has trained me on how to think and write as an economist. His financial econometrics course sparked my interest in the field; one of his projects led to a joint publication, a chapter of this thesis. I would like to take this opportunity to express my deepest gratitude to Professor Lars Stentoft.

I am very fortunate to receive the guidance from my committee members, Professor Tim Conley and Professor Charles Saunders, who spent a lot of time to review my thesis and provide me invaluable advice. At many times they responded to my emails late at night, early in the morning, at weekends and holidays, which touched me greatly. I have benefited tremendously from their guidance and support, without which the completion of this thesis would not have been possible. I owe Professor Tim Conley and Professor Charles Saunders a debt of gratitude.

My thesis has also greatly benefited from valuable suggestions kindly provided by the professors who are currently working or used to work at the University of Western Ontario, including but not limited to, Victor Aguiar, Roy Allen, Audra Bowlus, Andres Carvajal, Jorge Cruz Lopez, Juan Carlos Hatchondo, Maria Goltsman, Silvia Goncalves, Sermin Gungor, Nail Kashaev, Lance Lochner, Leigh MacDonald, Jim MacGee, Rogemar Mamon, Nirav Mehta, Salvador Navarro, Ananth Ramanarayanan, Mark Reesor, Chris Robinson, David Rivers, Al Slivinski, Todd Stinebrickner, and Charles Zheng. In addition, I would like to take a moment to pay tribute to late Professor John Knight, whose core econometrics course provided me with a solid foundation for econometrics research. I would also like to thank the faculty members

and graduate students at the University of Waterloo for their helpful suggestions.

I would certainly like to thank my fellow students, in particular my co-author Dillon Huddleston, as well as Francisco Adame, Samantha Black, Matthew Carew, Cecilia Diaz Campo, Yifan Gong, Aldo Sandoval Hernandez, Ali Kamranzadeh, Qian Liu, Zhuang Liu, Martin Luccioni, Shannon Potter, Rubina Siddika, Phuong Vu, Sha Wang, Zijian Wang, Ziyu Zhang, and MA classmates Aleksandar (Sasa) Cabric, Darryl Davies, Kathleen Healy, Shim Won (Isabel) Lee, Alena Samoday, Andres Arcila Vasquez, and Gabrielle Vasey for their help and friendship, which have made my PhD experience an enjoyable one.

I am indebted to all support from the staff members of the Department of Economics, in particular our graduate coordinator Sandra Augustine, who has tirelessly helped me on numerous tasks. The staff members of the department have made my PhD experience an enjoyable one.

I am deeply grateful to my parents for their unconditional love and support, as well as strong belief in me. I would like to thank my mom for her extreme support during difficult times. Finally, I would like to thank my partner Ashley Zarzecki for her love and support. I am extremely fortunate to have met her during the Master's program at Western.

Contents

List of Figures

List of Tables

List of Appendices

Chapter 1

Introduction

Financial econometrics is a highly interdisciplinary field that integrates finance, economics, probability, statistics, and applied mathematics, and is important for understanding complex financial problems. Machine learning is a growing area in finance that is particularly suitable for studying problems with many variables, and may help solve problems that have challenged traditional econometric models. Examples of problems include: risk managers need to understand the behavior of extreme losses of their financial assets and its implications for regulatory capital requirements, investors seek to explain the price movements of financial assets with respect to different types of risk, and traders are interested in anticipating short- and long-term financial market movements. These difficult problems can be solved by combining the theoretical foundation in economics with the quantitative methods in statistics and applied math.

My thesis contains three chapters that explore financial econometrics and machine learning in the fields of asset pricing and risk management. In Chapter 2, joint with Lars Stentoft, we study the implications of new market risk regulations for banks' choice of econometric models and whether it incentivizes banks to use accurate models. In Chapter 3, joint with Dillon Huddleston and Lars Stentoft, we examine the predictability of intraday (i.e. five-minute) market returns using machine learning models estimated on a large cross-section of stocks. In Chapter 4, I study two recent empirical puzzles: the idiosyncratic tail risk premium and the common factor driving the idiosyncratic tail risks. I introduce a new econometric measure of idiosyncratic tail risk using a high-frequency factor model, and resolve the idiosyncratic tail

risk premium puzzle using its common factor.

Chapter 2 studies the implications of the new Basel 3 regulations for banks' quantitative risk models. In response to the Subprime Mortgage crisis, the Basel Committee on Banking Supervision (BCBS) has spent the previous decade overhauling the regulatory framework that governs how banks calculate minimum capital requirements. In 2019, the BCBS finalized the Basel 3 regulatory regime, which changes the regulatory measure of market risk and adds new complex calculations based on liquidity and risk factors. This chapter is motivated by these changes and seeks to answer the question of how regulation affects banks' choice of risk-management models, whether it incentivizes them to use correctly specified models, and if it results in more stable capital requirements. Our results show that, although the models that minimize regulatory capital for a representative bank portfolio also result in the most stable requirements, these models are generally rejected as being correctly specified and tend to produce inferior forecasts of the regulatory risk measures.

Chapter 3 conducts, to our knowledge, the largest study ever of five-minute equity market returns using state-of-the-art machine learning models trained on the cross-section of lagged market index constituent returns, where we show that regularized linear models and nonlinear tree-based models yield significant market return predictability. Ensemble models perform the best across time and their predictability translates into economically significant Sharpe ratios of 0.98 after transaction costs. These results provide strong evidence that intraday market returns are predictable during short time horizons, beyond what can be explained by transaction costs. Furthermore, we show that constituent returns hold significant predictive information that is not contained in market returns or in price trend and liquidity characteristics. Consistent with the hypothesis that predictability is driven by slow-moving trader capital, predictability decreased post-decimalization, and market returns are more predictable during the middle of the day, on days with high volatility or illiquidity, and in financial crisis periods.

Chapter 4 studies the idiosyncratic tail risk premium and common factor. Stocks in the highest idiosyncratic tail risk decile earn 8% higher average annualized returns than in the

lowest. I propose a risk-based explanation for this premium, in which shocks to intermediary funding cause idiosyncratic tail risk to follow a strong factor structure, and the factor, common idiosyncratic tail risk (CITR), comoves with intermediary funding. Consequently, firms with high idiosyncratic tail risk have high exposure to CITR shocks, and command a risk premium due to their low returns when intermediary constraints tighten. To test my explanation, I create a novel measure of idiosyncratic tail risk that is estimated using high-frequency returns, and theoretically establish its time-aggregation properties. Consistent with my explanation, CITR shocks are procyclical, are correlated to intermediary factors, are priced in assets, and explain the idiosyncratic tail risk premium. Furthermore, volume tail risk also earns a premium, follows a strong factor structure, and its common factor is priced. This duality of idiosyncratic tail risk and volume tail risk provides evidence for my risk-based explanation, and further supports the hypothesis that intermediaries' large trades cause idiosyncratic tail risk and volume tail risk from Gabaix et al. (2006).

Chapter 2

Regulatory Capital and Incentives for Risk Model Choice under Basel 3

2.1 Introduction

The regulatory environment that governs how banks calculate minimum capital requirements has changed dramatically in recent years. First, in response to the 2008 Subprime Mortgage Crisis, the Basel Committee on Banking Supervision (BCBS) adopted the Stressed Value at Risk measure. Value at Risk (VaR) is the conditional quantile of the loss distribution at a given confidence level, and Stressed VaR is defined as the VaR on a one year historical dataset with significant financial stress (BCBS (2011)). This change significantly increased the capital charges banks faced. Second, in 2019 the BCBS finalized the Fundamental Review of the Trading Book (FRTB) regulatory regime as part of Basel 3. This will change the regulatory measure of market risk from VaR at the 99% confidence level to Expected Shortfall at the 97.5% confidence level (BCBS (2019)). Expected Shortfall (ES) is defined as the expected loss conditional on VaR being exceeded and is widely perceived to be a more appropriate risk measure as it captures both the size and likelihood of losses (BCBS (2013)). Additionally, in Basel 3 Stressed ES at the 97.5% confidence level, defined as the ES during the most severe

one year period of losses available, replaces Stressed VaR as the key risk measure.

The new capital requirements for market risk present an interesting trade-off for banks. Assuming bank capital is costly and hence banks minimize their capital requirements, Basel 3 incentivizes them to use models producing low Stressed ES. At the same time, Basel 3 penalizes banks by increasing their capital requirements when their model generates too many VaR exceedances, which incentivizes banks to use more conservative models. It is not obvious how different models for VaR and Stressed ES perform in terms of this trade-off and which model is therefore privately optimal in the sense that it minimizes banks' capital requirements. Moreover, it is likely even more costly, and sometimes impossible, for banks to raise capital during times of high volatility, so prudence might suggest building a buffer stock of capital during good times that can be drawn down during times of stress. Basel 3 aims at creating more stable capital requirements by focusing on stressed risk measures. It is also not obvious how different models for VaR and Stressed ES perform in terms of minimizing measures of capital requirement stability.

Our paper is motivated by these changes and seeks to answer the question of how regulation affects banks' choice of risk-management models and whether it incentivizes them to use correctly specified models. We also analyze whether the changes made to regulation incentivize banks to use models that lead to more stable (through-the-cycle) capital requirements rather than more dynamic (point-in-time), and therefore potentially more systemically risky, capital requirements. To preview our results, we find that, although the proposed regulation under Basel 3 seems to incentivize banks to use models that have more stable capital requirements through time, the models that minimize average capital requirements appear misspecified, in the sense that they are rejected using standard backtests, and produce inferior forecasts of the regulatory risk measures.

To answer these questions we construct a portfolio of diverse assets, with different risk factors and subject to different levels of liquidity risk, that a bank might realistically hold, and we consider three classes of models for the dynamics of the losses of this portfolio: 1) ad hoc

methods like Historical Simulation (HS) and RiskMetrics that involve no parameter estimation, 2) models within the classical GARCH framework of Engle (1983) and Bollerslev (1986) that are estimated by fitting model parameters to the entire dynamic distribution of losses, and 3) a new class of models developed by Patton, Ziegel, and Chen (2019) that are estimated by fitting model parameters directly to the dynamics of the risk measures instead, which we refer to as "FZ" models. For the dynamic models, and the GARCH models in particular, we consider several types of conditional distributions that can accommodate stylized facts like heavy tails and skewness toward losses as documented by Hansen (1994) among others.[1]

We first examine which of the models are correctly specified using both traditional VaR backtests and a battery of recent joint VaR and ES backtests, and we examine which models produce optimal forecasts of the risk measures.[2] As expected our results show that the ad hoc models and the HS model in particular fail the backtests and provide inferior forecasts of the risk measures. More interestingly though, our results also show that the FZ models fail some backtests and provide inferior forecasts of the VaR and ES compared to the subset of GARCH models that allow for skewed conditional distributions and are not rejected by the backtests.

Next, we carefully calculate the capital requirements for each model using the Basel 3 formulas and compare the results to those using previous regulatory regimes. Our results show that the HS model has the lowest average Basel 3 capital of 17.21% whereas the best model in the FZ class has average capital requirements of 17.42%. However both of these models, along with several other models that have low capital requirements, are rejected as being correctly specified or shown to produce inferior forecasts of the regulatory risk measures. The best performing model which is not rejected by the backtests, a skewed GARCH model, provides

[1] Using a rolling window estimation setup we also accommodate the recent observation that the conditional distribution and tails in particular of financial returns vary significantly over time and this particularly so during times of crisis (Bollerslev and Todorov (2014a) and Kelly and Jiang (2014a)).

[2] The methods we use to backtest predicted VaR are standard and can be found in, e.g., Christoffersen (2009). In terms of joint VaR and ES backtests, we consider McNeil and Frey (2000)'s residual test (ER), Bayer and Dimitriadis (2018)'s strict, auxiliary, and one-sided regression tests (ESR), Nolde and Ziegel (2017)'s conditional calibration test (CCa), and Gordy and McNeil (2020)'s Spectral Backtests. To asses the optimality of a model's forecasts, we consider the Model Confidence Set of Hansen, Lunde, and Nason (2011), the pairwise forecast performance tests from Diebold and Mariano (2002), and Ziegel et al. (2017)'s Murphy Diagrams.

superior risk measure forecasts but requires nearly 1.55 times higher capital than the HS model. The Filtered Historical Simulation (FHS) model, which is also not rejected and in the set of models that provide superior risk measure forecasts, has capital requirements that are roughly 1.78 times larger than the incorrectly specified HS model. Under Basel 3 there is therefore little incentive for a capital requirement minimizing bank to choose correctly specified models. Compared across the previous regimes our results show that correctly specified models are in fact never the models that minimize capital requirements.

Finally, we measure the variability of the minimum capital requirements under Basel 2, 2.5, and 3 to determine whether the new regulation is successful at increasing the stability of capital requirements. We consider several measures of the volatility of regulatory capital and we also measure peak-to-trough variation as the maximum difference in capital requirements. The BCBS has also focused on the procyclicality of regulation, often measured as peak-to-trough variation in minimum capital requirements (Gordy and Howells (2006), Heid (2007), and Shim (2013)). Our results show that capital requirements became significantly more stable from Basel 2 to Basel 2.5 due to the introduction of Stressed VaR. Basel 3 will further increase the stability of capital requirements by decreasing non-standardized and standardized volatility across most models and could also reduce the procyclicality of capital requirements as evidenced by the lower peak-to-trough variation across most models under this regime compared to previous regimes. However, the results also show that the models that result in the most stable capital requirements across regulatory regimes and across variability metrics are generally not the correctly specified models.

Our findings have important implications for current regulation. In particular, our results show that Basel 3 regulation strongly disincentivizes banks from using correctly specified models. In fact, banks can minimize both the mean and volatility of Basel 3 capital by using the hybrid FZ model of Patton, Ziegel, and Chen (2019) which is not only rejected by some of the backtests but also provides inferior risk measure forecasts for our portfolio. Although the same holds for previous regulatory regimes, the changes suggested in Basel 3 make the relative

differences even larger. For example, under Basel 2.5 FHS is only marginally worse than HS and would require only 1.09, instead of 1.78 under Basel 3, times the capital. We identify two possible reasons for this. First, and this is somewhat subtle, it appears that the requirement under Basel 3 to penalize low liquidity assets, i.e. assets with long liquidity horizons, additionally in fact disincentivizes banks from using the correctly specified skewed GARCH models due to their consistently high Stressed ES across liquidity horizons. Second, and this is more obvious, given the low level of the Basel 3 multiplier banks have little incentive to choose conservative and correctly specified models. Thus, if the regulator's objective is to incentivize the use of correctly specified models they would have to reconsider the effect of these changes.[3]

Our paper is related to at least three strands of existing literature. First of all, there is a large literature on empirically backtesting VaR (see, e.g., Christoffersen, Hahn, and Inoue (2001) and Gencay and Selcuk (2004)) and a growing literature on backtesting VaR and ES jointly. Our paper complements this literature in several ways. First, while most backtesting papers focus on one asset class (most commonly equities) we consider a large and diverse set of assets spanning the multiple risk factors and liquidity horizons that banks may be exposed to. Second, we use cutting-edge ES backtesting techniques recently developed by academics to determine if models are correctly specified and we compare these models' forecasts of VaR and ES. Finally, we go beyond simple backtesting by calculating Basel 3 capital requirements for the representative bank portfolio and by evaluating the trade-off between correctly specified models and models that minimize not only the level but also the variability of the capital requirements.

Next, our paper is related to studies of backtesting that use actual bank P&L or VaR data. For example, Berkowitz and O'Brien (2002) show that U.S. banks had conservative VaR estimates during the 1998 Asian crisis with few exceedances, but the exceedances were clustered indicating bank models did not adapt to dynamic volatility. Pérignon, Deng, and Wang (2008)

[3]Another possibility is that regulators could attempt to incentivize the use of correctly specified models by penalizing models with high standardized volatility, the only capital stability measure for which correctly specified models, in this case models based on Extreme Value Theory, perform the best.

find that Canadian banks also have conservative VaR forecasts and Pérignon and Smith (2008) extends the results to international banks. O'Brien and Szerszeń (2017), on the other hand, show that these early results were driven by a calm sample period and that U.S. banks had excessive exceedances and clustering during the 2008 financial crisis. Their findings suggest that the banks used misspecified models that do not adapt to time-varying volatility. Berkowitz, Christoffersen, and Pelletier (2011) compares the accuracy of VaR forecasts for trading desks in a commercial bank using a novel spectral backtest. Gordy and McNeil (2020) study model-implied probabilities associated with bank-reported P&L and reject the hypothesis of correct specification for many U.S. bank models. While these studies use actual bank data, they are limited in that they do not observe the bank's internal model or portfolio composition. They also do not calculate capital requirements under Basel 3. Our study complements this literature by using a transparent representative bank portfolio and well known models.

Finally, our study is similar in spirit to the annual BCBS monitoring exercise that provides a hypothetical portfolio for banks to calculate actual risk measures and capital requirements, see, e.g., BCBS (2014). However, in these exercises the banks' internal models are confidential and the BCBS only reports aggregate capital results. Also, formal backtests are not performed to determine if the internal models used by banks are correctly specified as a part of these exercises. Our research complements the BCBS exercises by examining model VaR and ES forecasts on a hypothetical portfolio. We are able to backtest if the models are correctly specified and compare the mean and variability of their capital requirements. Hence, a bank's regulatory supervisor could assess the trade-off between correct specification, capital requirements, and capital stability for the bank's internal model.

The paper is organized as follows: Section 2.2 describes how the regulatory environment has changed over time and outlines the requirements and objectives of the current regulation. Section 2.3 presents the data used, reviews the different classes of dynamic models considered, and explains how multiperiod risk measures can be calculated with these models. Section 2.4 contains extensive backtests for the models considered, analysing which of them are correctly

specified and which produce the best risk measure forecasts. Section 2.5 calculates the regulatory capital under Basel 3 and the previous regulatory regimes and assesses the stability through time of different model's capital requirement. Section 2.6 concludes. Appendix A contains further details on the regulatory calculations, on how to select optimal thresholds for Extreme Value Theory models, on the backtesting methods used for ES, and some additional results.

2.2 Regulatory Capital Calculations

The last 25 years have seen significant changes to the regulation faced by banks when it comes to the regulatory capital requirements. In particular, over this period the Basel Accord has had three regimes for calculating market risk capital requirements: Basel 2, Basel 2.5, and the incoming Basel 3. While banks are currently adapting to the new requirements of Basel 3, it remains important to consider how we arrived at this regulation as well as the motivation and implications of this changing regulation.

To set the scene for the rest of the paper, this section first provides a brief background on the regulation that led to Basel 3 paying special attention to how this has changed the calculations of minimum capital requirements. We then provide details on the proposed Basel 3 capital requirements and its formulas for calculating market risk capital. Finally, we discuss the intended and expected impact of this regulation on banks and the requirements it imposes on them.

2.2.1 Background

The Basel 2 market risk requirements were first introduced by the BCBS in the 1996 Amendment to the Basel Accord (BCBS (1996a)) and allowed banks to use their own "internal" models to calculate regulatory capital. The internal model-based approach for setting market risk capital involved calculating a VaR measure with a 10-day time horizon and at a 99% confi-

dence level. Denoting daily losses, or negative returns, on a single asset or portfolio by L_t then formally the 1-day VaR measure at time t is defined as the value VaR^p_{t+1} such that

$$Pr(L_{t+1} > VaR^p_{t+1}|\Im_t) = p, \tag{2.1}$$

where $\Im_t$ is the information at time t and p is the coverage probability. That is, at period $t + 1$, losses exceed VaR^p_{t+1} only with probability p, given the available information. The day t capital requirement for a bank with this VaR is then set as

$$CA^{B2}_t = \max(VaR_{t-1}, m_c \times \overline{VaR}_{t-1}), \tag{2.2}$$

where m_c is a multiplicative factor, VaR_{t-1} is the previous day's 10-day value at risk, and $\overline{VaR}_{t-1}$ is the average value at risk over the previous 60 days. The multiplicative factor m_c has a minimum value of 3 and adds a scaling factor between 0 and 1 that depends on the model's backtesting performance and penalizes models that backtest poorly, see BCBS (1996b). Basel 2 backtesting compares VaR with a 1-day time horizon at the 99% confidence level to realized exceedances (losses above VaR) over the previous 250 days.[4]

A given internal model's multiplication factor is set according to the "traffic light" system of exceedances reproduced in Table 2.1. Column 3 shows that if the number of exceedances in the previous 250 days is 4 or fewer, the model is in the Green Zone and the multiplicative factor m_c is 3. If the number of exceedances is between 5 and 9, the model is in the Yellow Zone and m_c is between 3.4 and 3.85. Note that the large discrete jump between 4 and 5 exceedances increases the penalty with more than 10%. If the number of exceedances is greater than 10, the model is in the Red Zone and m_c is set at 4. Additionally, during Red Zone periods, regulatory supervisors can disallow the use of a particular internal model, which forces the bank to use the standard model approach. This could potentially increase overall capital requirements significantly.

[4]Backtesting is not done on a 10-day horizon, since the portfolio composition may change within the ten days.

Table 2.1: Basel backtesting zone boundaries

Backtesting zone	Exceedances	Basel 2 Multiplier	Basel 3 Multiplier	Cumulative Probability
Green	4 or fewer	3	1.5	89.22%
Yellow/Amber	5	3.4	1.7	95.88%
	6	3.5	1.76	98.63%
	7	3.65	1.83	99.60%
	8	3.75	1.88	99.89%
	9	3.85	1.92	99.97%
Red	10 or more	4	2	99.99%

This table defines the Green, Yellow/Amber, and Red Zones that supervisors use to assess backtesting results in conjunction with the internal models approach to market risk capital requirements under Basel 2 and Basel 3, see BCBS (1996b) and BCBS (2019). The boundaries shown in the table are based on a sample of 250 observations.

The 2008 Subprime Mortgage Crisis revealed that the Basel 2 requirements were far too low to capture systemic risks, which resulted in banks holding insufficient capital before the crisis. Additionally, the sudden jump in capital requirements during the crisis caused banks to deleverage by sharply shedding risk exposures (Adrian and Shin (2014)), resulting in fire-sales, liquidity spirals, and disinflation (Brunnermeier and Sannikov (2016)). The reduction in bank balance sheets and credit at the height of the financial crisis caused by the procyclicality of Basel 2 capital requirements amplified the downturn (Adrian and Shin (2014)).

To address the shortcomings of Basel 2, the BCBS introduced the current regulatory regime, Revisions to the Basel 2 Framework (Basel 2.5), with an implementation date of December 31, 2011 (BCBS (2009)). In Basel 2.5, the level and stability of capital requirements is increased by introducing the Stressed VaR measure. Stressed VaR is VaR calculated during a 12-month period of significant financial stress. The stress period is identified as the 12 months in history that maximizes Value at Risk with a 10-day time horizon and a 99% confidence level (EBA (2012)). The Basel 2.5 capital charge is set at

$$CA_t^{B2.5} = \max(VaR_{t-1}, m_c \times \overline{VaR}_{t-1}) + \max(SVaR_{t-1}, m_s \times \overline{SVaR}_{t-1}), \tag{2.3}$$

where the first term is the Basel 2 capital charge, $SVaR_{t-1}$ is the previous day's Stressed VaR, $\overline{SVaR}_{t-1}$ is the average Stressed VaR over the previous 60 days, and m_s is the stressed multiplicative factor set by the regulatory supervisor. Since Stressed VaR is always at least as large as VaR and assuming $m_c = m_s$, the Basel 2.5 capital charge is at least double the Basel 2 charge. Also since the period for Stressed VaR rarely changes Basel 2.5 should result in more stable capital requirements.

2.2.2 Basel 3

While Basel 2.5 was implemented to temporarily increase the level and stability of capital requirements to address some of the serious flaws in the capital regulation framework exposed by the 2008 crisis, the BCBS has continued to work on implementing more stringent capital calculations. In 2014, the BCBS proposed Basel 3, formally named the Fundamental Review of the Trading Book (FRTB), as a new and comprehensive approach to determining minimum regulatory capital for market risk. The key changes to market risk calculations under Basel 3 include the use of ES instead of VaR, calculations based on liquidity horizons, calibration to periods of significant financial stress, and diversification restrictions. We now summarize each of the changes and explain the Basel 3 market risk formula based on the finalized FRTB documentation (BCBS (2019)).

In Basel 3, the measure used to determine capital changes from 10-day VaR at a 99% confidence level to 10-day ES at a 97.5% confidence level. Formally, the 1-day ES measure at time t is defined as

$$ES^p_{t+1} = E(L_{t+1}|L_{t+1} > VaR^p_{t+1}, \Im_t). \tag{2.4}$$

That is, if losses at time $t+1$ exceed VaR^p_{t+1}, then the expected loss is ES^p_{t+1}. The motivation for changing risk measures is that ES reflects tail risk better than VaR, since ES captures both the size and likelihood of losses in the tail (e.g. BCBS (2013)). Moreover, ES is a coherent risk measure (Artzner et al. (1999)) since it satisfies the subadditivity property while VaR does

not.[5]

Market liquidity played a large role during the 2008 Subprime Mortgage Crisis. At the height of the crisis, investors took large discounts to sell illiquid assets which further lowered asset prices and caused a liquidity spiral (Brunnermeier and Sannikov (2016)). Basel 3 proposes to account for liquidity risk by scaling the 10-day ES based on an asset's liquidity. Assets are assigned to one of five h_j day liquidity horizons, where $h_1 = 10$, $h_2 = 20$, $h_3 = 40$, $h_4 = 60$, and $h_5 = 120$.[6] Define LH h_j as the portfolio of the subset of assets with a liquidity horizon of h_j days or longer. For example, LH 10 is the portfolio of all assets, LH 20 is the portfolio of assets with a liquidity horizon of 20 days or more, and LH 120 is the portfolio of assets with the longest liquidity horizon of 120 days only. Next, define $ES(j)$ as the ES of portfolio LH h_j. The formula for liquidity-adjusted ES is

$$ES = \sqrt{ES(1)^2 + \sum_{j=2}^{5} \left(\sqrt{\frac{h_j - h_{j-1}}{10}} ES(j) \right)^2}, \tag{2.5}$$

where h_j is the liquidity horizon, $\sqrt{\frac{h_j - h_{j-1}}{10}}$ is the liquidity scaling based on a normality assumption on asset returns, and $ES(j)$ is the ES for assets with an liquidity horizon of h_j or longer. Hence, $ES(1)$ and $ES(2)$ have no additional liquidity scaling, $ES(3)$ and $ES(4)$ are scaled by $\sqrt{2}$, and $ES(5)$ is scaled by $\sqrt{6}$. The high scaling on illiquid assets such as credit derivatives is meant to reflect their additional risk of discounts during financial stress.

A key weakness of Basel 2 was that its risk measures were calibrated to current market conditions, which resulted in undercapitalization and procyclical capital requirements during the crisis. Basel 2.5 introduced Stressed VaR to ensure the capital charge includes periods of significant financial stress in addition to current market conditions, which may be unnecessarily duplicative (BCBS (2013)). In Basel 3, the capital charge is instead only based on a Stressed

[5] A risk measure is subadditive if the risk measure for the sum of two portfolios is no greater than the sum of the risk measures for those portfolios. Subadditivity of risk measures means that diversification may help reduce risks.

[6] See Table 2.2 for some examples of this classification.

ES, which is ES calibrated to a period of significant financial stress. Since historical data may be unavailable for the full set of risk factors, Basel 3 allows Stressed ES calculations to use a reduced set of risk factors, as long as the reduced set explains at least 75% of the variation in the full set. The formula for the bank's Internally Modeled Capital Requirement (IMCC) is

$$IMCC(C) = ES_{R,S} \times \frac{ES_{F,C}}{ES_{R,C}}, \tag{2.6}$$

where $ES_{R,S}$ is the liquidity-adjusted ES of the reduced set of risk factors calibrated to a period of significant financial stress, $ES_{F,C}$ is the liquidity-adjusted ES of the full set of risk factors calibrated to the current market, $ES_{R,C}$ is the liquidity-adjusted ES of the current reduced set of risk factors, and where the ratio $\frac{ES_{F,C}}{ES_{R,C}}$ is floored at 1.

The IMCC calculated above naturally benefits from portfolio diversification. However, during a financial crisis, systemic risk often causes correlations to increase and reduces this effect. To account for this risk, Basel 3 requires banks to calculate partial ES capital requirements, denoted $IMCC(C_i)$ for each regulatory risk class. The undiversified IMCC for risk class i is

$$IMCC(C_i) = ES_{R,S,i} \times \frac{ES_{F,C,i}}{ES_{R,C,i}}, \tag{2.7}$$

where i refers to interest rate risk, equity risk, foreign exchange risk, commodity risk, or credit spread risk. The stress period used to calculate $ES_{R,S,i}$ is the same as the period used to calculate the portfolio-wide $ES_{R,S}$. The bank's aggregate IMCC is then given by

$$IMCC = \rho(IMCC(C)) + (1-\rho)(\sum_{i=1}^{5} IMCC(C_i)), \tag{2.8}$$

where $\rho = 0.5$.

The capital charge for modellable risk factors under Basel 3 is

$$CA_t^{B3} = max(IMCC_{t-1}, m_c \times \overline{IMCC}_{t-1}), \tag{2.9}$$

where $IMCC_{t-1}$ is the previous day's aggregate IMCC, $\overline{IMCC}_{t-1}$ is the average IMCC over the previous 60 days, and m_c is the Basel 3 multiplicative factor. Similar to the previous regimes, Basel 3 backtesting compares VaR with a 1-day time horizon at the 99% confidence level to realized exceedances over the previous 250 days. However, Basel 3 essentially halves the Basel 2 multiplicative factor with a minimum value of 1.5 for m_c and a scaling factor between 0 and 0.5 that depends on the model's backtesting performance. The model's multiplication factor is set according to the system of exceedances in Column 4 in Table 2.1.

2.2.3 Discussion

Whereas Basel 2.5 introduced only one change to Basel 2 when it comes to calculating capital requirements for market risk, by requiring the use of Stressed VaR, Basel 3 involves several additional changes: 1) the use of ES as the key risk metric instead of VaR, 2) explicit penalties for exposure to assets with liquidity risk, and 3) reduction in the possible gains from diversification. Although these changes may appear innocuous they in fact severely complicate the calculations required to conduct proper risk management within banks. First of all, under Basel 3 banks face additional requirements on the data needed to calculate the relevant capital requirement. In particular, long samples of data are needed for all the combinations of risk factors and liquidity horizons towards which the bank is exposed. And if some of these risk factors are less liquidly traded or they do not have sufficient amounts of historical data, mimicking portfolios with equivalent characteristics are needed. Moreover, on each day under Basel 3 there are 3 liquidity-adjusted ES calculations ($ES_{R,S}$, $ES_{F,C}$, $ES_{R,C}$) and 21 possible liquidity horizons across the 5 risk factors, totaling 63 daily ES calculations.

Finally, while ES as a risk measure has several benefits compared to VaR, it is a coherent risk measure that captures both the size and likelihood of losses in the tail, a downside is that it is slightly more difficult to estimate and that historically it has been complicated to backtest this measure as it lacks a property called elicitability (Gneiting (2011)). A risk measure is elicitable if there exists a loss function such that the risk measure is the solution to minimizing

the expected loss and while VaR is elicitable ES is not so individually. However, ES is jointly elicitable with VaR as shown by Fissler and Ziegel (2016), and using this result it is possible not only to jointly backtest VaR and ES, see Section 2.4, but also to model these measures jointly, see Section 2.3.2. An added complication of using 10-day ES is that it cannot be scaled from the 1-day ES, and although Basel 3 does allow calculating 10-day ES by using 10-day overlapping periods in many cases this method is not applicable and instead simulation is needed to create these multiperiod forecasts.

So why would regulators consider changing the regulatory capital requirements and, in particular, make these much more complicated to calculate for banks? If the objective is to incentivize banks to use models that best predict such risk measures there now exist several tests that can be used to test the backtesting performance. Models that pass these tests are essentially correctly specified. Moreover, if the question is one of finding the "best" forecasting model for risk measures, this can be examined by using an appropriate and consistent loss function for the risk measures together with Diebold and Mariano (2002) type tests to assess significant difference between two sets of forecasts or the Model Confidence Set of Hansen, Lunde, and Nason (2011) to examine which models among a set of models provide superior forecasts. We conduct several such tests and examine which models, among a large class of statistical models, are indeed correctly specified and which provide superior risk measure forecasts.

However, for regulators it is clearly not sufficient that banks hold enough capital to cover their losses 99% of the time and instead regulators require banks to keep a buffer of capital that is generally larger than the predicted risk measures by maximizing over current risk and historical averages. Moreover, by using backtesting multipliers banks are incentivized to use more conservative models. We show that though capital requirements have increased, historically Basel regulation has failed to incentivize banks to use correctly specified models and Basel 3 is no exception. Moreover, an additional goal of Basel 3 is the ensure the stability of these capital requirements by calibrating the risk measures to periods of significant financial stress. We show that Basel 3 further dampens the cyclicality in capital requirements by focusing on

Stressed ES and by removing most point-in-time calculations. However, the models that minimize capital requirements, though misspecified, are also generally the models that generate the most stable capital requirements for banks seeking to minimize Basel capital variability.

2.3 Data, models and multiperiod risk measures

Basel 3 puts additional requirements on the data needed to calculate regulatory capital, requiring long samples of data for all the combinations of risk factors and liquidity horizons towards which the bank is exposed, on the risk measures that an internal model should be able to produce estimates of, which includes Stressed ES measures, and explicitly disallows simple scaling techniques for generating multiperiod forecasts of risk measures, which in many cases are available only using simulation techniques.

In this section we first provide an overview of the data used in this paper. Next, we introduce the various dynamic models that are used to estimate VaR and ES. Finally we explain how multiperiod VaR and ES forecasts can be generated using, in most cases, simulation techniques. Readers, who are familiar with all these issues, can skip this section and go straight to our backtesting results in Section 2.4.

2.3.1 Data

A key feature distinguishing our paper from most of the existing literature is that we consider a realistic portfolio of multiple asset classes that banks likely trade in and hold on their trading book instead of a single asset class like, e.g., equities. We use the daily indexes in Table 2.2 to proxy for the various risk exposures and liquidity horizons that enter into the calculation of capital requirements under Basel 3. These indexes are highly liquid, widely traded, and constitute a diverse sample that span all risk factors and relevant liquidity horizons. Our sample is from January 1989 to February 2020, and indexes are included when they become available. All data are obtained from Bloomberg.

Table 2.2: Sample of indexes used

Risk Factor	Index	Symbol	Start Date	Liquidity Horizon
Interest Rate	Bloomberg US Treasury	LUATTRUU	03/1994	10
Interest Rate	Bloomberg US Treasury Inflation-Linked	LBUTTRUU	04/1998	10
Equity	S&P 500	SPX	01/1989	10
Equity	Russel 2000	RTY	01/1989	20
Equity	CBOE Putwrite	PUT	01/1989	60
Commodity	Bloomberg Commodities	BCOM	01/1990	20
Commodity	Bloomberg Commodities Volatility	GSVL1027	02/2001	120
Foreign Exch.	JP Morgan USD Trade Weighted	JPMQUSD	01/1990	20
Foreign Exch.	Bloomberg Dollar	DXY	01/2005	20
Credit	Bloomberg US Aggregate	LBUSTRUU	01/1989	20
Credit	Bloomberg Mortgage Backed Securities	LUMSTRUU	01/1991	40
Credit	Bloomberg Corporate	LUACTRUU	08/1998	40
Credit	Bloomberg High Yield	LF98TRUU	08/1998	60
Credit	Bloomberg Municipal	LMBITR	12/2000	60
Credit	Credit Default Swap Investment Grade	CDXIG	04/2007	120
Credit	Credit Default Swap High Yield	CDXHY	04/2007	120

This table shows the list of indexes used to form the representative banking portfolio. The sample is from January 1989 to February 2020. Indexes are sorted by risk factor first and then Basel 3 liquidity horizon category. We also provide the symbol and starting month for each index.

For interest rate risk, we use the Bloomberg US Treasury and US Treasury Inflation-Linked indexes for exposure to U.S. government bonds. For equity risk, we include the S&P 500 Index for exposure to large-cap firms and the Russell 2000 Index for exposure to small-cap firms. For equity derivatives risk, we include the CBOE Putwrite Index for exposure to a trading strategy that sells one-month at the money S&P 500 put options and invests the proceeds in one- and three-month Treasury bills. For commodity risk, we include the Bloomberg Commodities Index for exposure to energy, grains, metals, softs, and livestock. For commodity derivatives risk, we include the Bloomberg Commodities Volatility Index. For foreign exchange risk, we include the JP Morgan USD Trade-Weighted Index for exposure to the Australian dollar, British pound, Canadian dollar, Euro, Japanese yen, Swedish krona, and Swiss franc. We also include the Bloomberg Dollar Index for exposure to the Australian dollar, British pound, Canadian

dollar, Euro, Japanese yen, Swiss franc, Mexican peso, Chinese renminbi, Korean won, and Indian rupee. For credit risk, we include the Bloomberg US Aggregate, Mortgage Backed Securities, High Yield, Corporate, and Municipal Indexes for exposures to a large variety of U.S. fixed income securities. For credit derivatives risk, we include the Credit Default Swap Investment Grade and High Yield Indexes for exposure to credit default swaps, which played an important role in the 2008 crisis.

The risk factors we use are similar to those considered in Falato, Iercosan, and Zikes (2019) who use proprietary P&L data reported to the Federal Reserve to show that bank trading desks have exposures to these risk factors. We evaluate a representative bank portfolio that takes an equal-weighted long position in each available index, adjusting the weights as new indexes become available. We choose a simple equal-weighted portfolio, since the start date of the indexes roughly coincides with their importance in the market. Hence, we expect the portfolio to be a good representation of a typical U.S. bank's trading portfolio. Since the index returns are given as simple returns and our models use log returns, we first form the representative portfolio by taking an equal-weighted mean, then transforming the portfolio to log returns to form the representative bank portfolio used to calculate VaR and ES. We perform the same log transformation for portfolios grouped by risk factor (see Section A.1 of Appendix A for further details).

Table 2.3 provides summary statistics for the representative portfolio and individual risk factors. The representative portfolio had a mean daily return of 0.025% mainly driven by the equity, credit, and interest risk factors. Equity had the highest mean return, but also the highest daily volatility of 0.968% corresponding to an annualized volatility of 15%. The representative portfolio is highly skewed towards losses and heavy-tailed, demonstrating that models must accommodate these empirical features to accurately measure tail risk. The 0.025 quantile and 0.01 quantile and minimum return for the representative portfolio are roughly 2, 3, and 13 standard deviations from the mean, highlighting the non-normality of the distribution's loss tail.

Table 2.3: Summary statistics for portfolios

Portfolios	Representative	Interest	Equity	Commodity	Foreign Exchange	Credit
Mean	0.025	0.020	0.037	0.003	0.005	0.025
Standard Deviation	0.276	0.290	0.968	0.616	0.332	0.190
Skewness	-1.016	-0.207	-0.677	-0.661	0.215	-0.815
Kurtosis	9.692	3.790	9.603	10.168	6.377	9.674
0.025	-0.567	-0.583	-2.033	-1.291	-0.648	-0.366
0.01	-0.777	-0.759	-2.865	-1.697	-0.840	-0.516
Min	-3.728	-2.170	-9.747	-9.171	-2.476	-2.118

This table shows summary statistics for the representative banking portfolio and by risk factor. The sample is from January 1989 to February 2020. The representative portfolio is formed by taking the equal-weighted mean return of all available indexes (denoted r_{Bank}) in Table 2.2, then taking the log transformation $x_{Bank} = \log(1 + r_{Bank})$. Risk factor portfolios are analogously formed by taking the equal-weighted mean return of indexes in the risk factor, then taking the log transformation. Summary statistics are calculated based on each portfolio's daily log return and include the sample mean, standard deviation, skewness, kurtosis, 0.025 quantile, 0.01 quantile, and min in percentage terms.

2.3.2 Dynamic Models for returns

We assume throughout that losses are governed by a dynamic model given by

$$L_t = \mu_t + \sigma_t \epsilon_t, \qquad t = 1, ..., T, \tag{2.10}$$

where μ_t is the conditional mean, σ_t is the conditional volatility, and ϵ_t are independent and identically distributed (i.i.d.) innovations with distribution $G(0, 1)$.

Given the dynamic model in Equation (2.10), VaR can be expressed as

$$VaR^p_{T+1} = \mu_{T+1} + \sigma_{T+1} G^{-1}_{1-p} \equiv \mu_{T+1} + \sigma_{T+1} c_{1,p}, \tag{2.11}$$

where G^{-1}_{1-p} denotes the $(1 - p)$'th quantile of G. For example, if G is the standard normal distribution and $p = 0.01$, then $G^{-1}_{0.99} = \Phi^{-1}_{0.99} = 2.33$, where Φ denotes the standard normal distribution function, and hence $VaR^p_{T+1} = \mu_{T+1} + 2.33\sigma_{T+1}$. Similarly, ES can be expressed as

$$ES^p_{T+1} = \mu_{T+1} + \sigma_{T+1} E(\epsilon_{T+1} | \epsilon_{T+1} > G^{-1}_{1-p}) \equiv \mu_{T+1} + \sigma_{T+1} c_{2,p}. \tag{2.12}$$

For example, if $\epsilon_t \sim N(0,1)$ and $p = 0.01$, it can be shown that $E(\epsilon_{T+1}|\epsilon_{T+1} > \Phi^{-1}_{0.99}) = \phi(\Phi^{-1}_{0.99})/0.01 = 2.67$, where ϕ denotes the standard normal density function, and hence $ES^p_{T+1} = \mu_{T+1} + 2.67\sigma_{T+1}$. When the innovation distribution is non-normal we can still express VaR and ES as in Equations (2.11) and (2.12), though the values of $c_{1,p}$ and $c_{2,p}$ will depend on the distribution G.

In this paper, we consider several techniques for estimating the upper tail of the innovation distribution to find the tail risk measures. We first introduce two ad hoc models that involve no estimation including the most popular model used in banks called Historical Simulation. Next we consider a popular class of models in which the dynamics are parameterized using GARCH processes with parameters fitted to the entire conditional distribution. Finally, we describe a new class of models with conditional dynamics based on Generalized Autoregressive Score type models where parameters are instead fitted directly to a relevant loss metric for the risk measures considered.

Ad hoc models

The simplest and most popular model for estimating VaR and ES is undoubtedly Historical Simulation (HS) due to ease of implementation.[7] This nonparametric and distribution-free model calculates VaR and ES using the empirical distribution of past losses. The HS estimate for VaR^p_{T+1} is

$$\text{HS} - VaR^p_{T+1} = Q_{1-p}(\{L_t\}), \tag{2.13}$$

where $Q_{1-p}(\{L_t\})$ denotes the $(1-p)$'th empirical quantile of losses $\{L_t\}^T_{t=1}$. The HS estimate for ES^p_{T+1} is

$$\text{HS} - ES^p_{T+1} = \frac{1}{\#(L_t > HS - VaR^p_{T+1})}\Big(\sum_{L_t > HS - VaR^p_{T+1}} L_t\Big), \tag{2.14}$$

where $\#(L_t > HS - VaR^p_{T+1})$ denotes the number of losses $\{L_t\}^T_{t=1}$ exceeding $HS - VaR^p_{T+1}$. While HS can capture the nonnormality commonly observed in financial returns it cannot ac-

[7] Pérignon and Smith (2010) report that 73% of international banks use HS and Mehta et al. (2012) reports that 75% of large banks use only HS.

count for the conditional dynamics in Equation (2.10).

The HS model is a "fully" nonparametric model since no assumptions are made about neither the dynamics nor the conditional distribution. Parametric models on the other hand use explicit formulas for the dynamics together with a parameterized distribution to calculate the values of $\hat{c}_{1,p}$ and $\hat{c}_{2,p}$. The simplest parametric model is the RiskMetrics (RM) model developed by JP Morgan (JPMorgan (1996)). RiskMetrics assumes losses are normally distributed, that $\mu_t = 0$, and that the conditional variance follows

$$\sigma_t^2 = 0.06(L_{t-1})^2 + 0.94\sigma_{t-1}^2, \tag{2.15}$$

and thus no estimation is required for this model either. The RM estimate for VaR_{T+1}^p is

$$\text{RM} - VaR_{T+1}^p = \sigma_{T+1} c_{1,p}^{Norm}, \tag{2.16}$$

where $c_{1,p}^{Norm} = \Phi_{1-p}^{-1}$. The RM estimate for ES_{T+1}^p is

$$\text{RM} - ES_{T+1}^p = \sigma_{T+1} c_{2,p}^{Norm}, \tag{2.17}$$

where $c_{2,p}^{Norm} = \phi(c_{1,p}^{Norm})/p$.

Dynamic location-scale models

The most popular approach for specifying the dynamic model in Equation (2.10) in a flexible manner is to let μ_t follow some ARMA process and to let σ_t^2 follow a GARCH process. In this paper we will assume that the conditional mean is constant ($\mu_t = \mu$) and that the conditional variance follows a GARCH(1,1) model given by

$$\sigma_t^2 = \omega + \alpha(L_{t-1} - \mu_{t-1})^2 + \beta\sigma_{t-1}^2, \tag{2.18}$$

where $\alpha + \beta < 1$ to ensure stationarity.[8] The RM model is a special case of this framework which sets $\mu = 0$, $\omega = 0$, $\alpha = 0.06$, and $\beta = 0.94$ in Equation (2.18).[9]

A first model that corrects the shortcomings of the HS model above is the Filtered Historical Simulation (FHS) model which computes $\hat{c}_{1,p}$ and $\hat{c}_{2,p}$ from the empirical distribution of centered innovations $\hat{\epsilon}_t - \bar{\hat{\epsilon}}$. Thus, this model uses the conditional dynamics without the need for distributional assumptions on the empirical innovations.[10] FHS was first proposed by Barone-Adesi, Bourgoin, and Giannopoulos (1998), Diebold, Schuermann, and Stroughair (2000), and Hull and White (1998). The FHS estimate of $c_{1,p}$ is

$$\hat{c}_{1,p}^{FHS} = Q_{1-p}(\{\hat{\epsilon}_t - \bar{\hat{\epsilon}}\}), \tag{2.19}$$

and the FHS estimate of $c_{2,p}$ is

$$\hat{c}_{2,p}^{FHS} = \frac{1}{\#(\hat{\epsilon}_t - \bar{\hat{\epsilon}} > c_{1,p}^{FHS})} \Big(\sum_{\hat{\epsilon}_t - \bar{\hat{\epsilon}} > c_{1,p}^{FHS}} (\hat{\epsilon}_t - \bar{\hat{\epsilon}}) \Big), \tag{2.20}$$

The FHS estimates for VaR_{T+1}^{p} and ES_{T+1}^{p} are then obtained by substituting these estimates into Equations (2.11) and (2.12), respectively.

Parametric models Whereas FHS makes no assumptions about the conditional distribution, the RM model assumes losses are Normally distributed. Other typical choices in the literature include the Student's t-distribution (STD), the Hansen (1994) Skewed Student's t-distribution (SSTD), or the Generalized Error distribution (GED) which are popular because they can capture heavy tails exhibited by financial returns. We consider the flexible Skewed Generalized t-distribution (SGT) of Theodossiou (1998), which nests many of the popular parametric dis-

[8] Naturally, our approach generalizes to more complex specifications of the conditional mean and variance.

[9] Since $\alpha + \beta = 1$, the RM model follows a IGARCH random walk process and is therefore not stationary.

[10] The parameters of ARMA-GARCH type models can be estimated consistently using Quasi-Maximum Likelihood estimation with a Gaussian likelihood even if the underlying distribution is non-Gaussian assuming the correct order of the dynamic processes is specified, see, e.g. Bollerslev and Wooldridge (1992) and Gourieroux (1997).

tributional assumptions for modeling financial returns. The probability density function of the SGT distribution is given by

$$f(x|\lambda,n,k) = \frac{k}{2\phi}n^{-\frac{1}{k}}B\left(\frac{1}{k},\frac{n}{k}\right)^{-1}\left(1+\frac{1}{n}\frac{|x-m|^k}{(1+sgn(x-m)\lambda)^k\phi^k}\right)^{-\frac{n+1}{k}}, \tag{2.21}$$

where m is the mode, ϕ is a scaling constant, $-1 < \lambda < 1$ is a skewness parameter, k and n are positive tail parameters, sgn is the sign function, and B is the Beta function. For a standardized SGT random variable with mean zero and unit variance, the mode is $m = -2\lambda G_1\phi$ and the scaling constant is $\phi = \left((1+3\lambda^2)G_2 - 4\lambda^2G_1^2\right)^{-\frac{1}{2}}$, where G_1 and G_2 are given by $G_j = n^{\frac{j}{k}}B\left(\frac{j+1}{k},\frac{n-j}{k}\right)B\left(\frac{1}{k},\frac{n}{k}\right)^{-1}$ for $j = 1, 2$.

Theodossiou (2018) shows that the closed form expression for the $(1-p)$'th quantile of the SGT distribution is

$$c_{1,p}^{SGT} = m + (1+\lambda)\phi n^{\frac{1}{k}}t_p^{\frac{1}{k}}(1-t_p)^{-\frac{1}{k}}, \tag{2.22}$$

where $t_p = IB^{-1}\left(\frac{2|(1-p)-(1-\lambda)/2|}{(1+\lambda)};\frac{1}{k},\frac{n}{k}\right)$ and IB^{-1} is the inverse incomplete Beta function ratio, and that

$$c_{2,p}^{SGT} = m + \frac{(1+\lambda)^2}{2p}\left[1 - IB\left(t_p;\frac{2}{k},\frac{n-1}{k}\right)\right]G_1\phi, \tag{2.23}$$

where IB is the incomplete Beta function ratio. We calculate the values for VaR_{T+1}^p and ES_{T+1}^p in Equations (2.11) and (2.12) under a SGT distributional assumption on the innovations by using estimated values of λ, n, and k to obtain $\hat{c}_{1,p}^{SGT}$ and $\hat{c}_{2,p}^{SGT}$ in Equations (2.22) and (2.23), respectively.

The SGT distribution is extremely flexible and nests all the parametric distributions used in this paper. For example, the SGT distribution is equivalent to the normal distribution when $\lambda = 0$, $k = 2$, and $n \to \infty$ and a SGT distribution with $\lambda = 0$, $k = 2$, and $n = d$ is equivalent to a Student's t-distribution with d degrees of freedom. When the skewness parameter $-1 < \lambda < 1$, the SGT distribution with $k = 2$, and $n = d$ is equivalent to Hansen's skewed t-distribution with d degrees of freedom and the same skewness parameter λ. Finally, the SGT distribution with

$n \to \infty$ is equivalent to the (skewed when $\lambda \neq 0$) GED distribution with shape parameter $v = k$.

Methods based on Extreme Value Theory Whereas the parametric model above imposes assumptions on the entire distribution of the innovations, Extreme Value Theory (EVT) models the behaviour of the distribution's tail. Tail values are described by the conditional excess distribution function defined by

$$F_\eta(y) = Pr\{X - \eta \leq y | X > \eta\} = \frac{F(y+\eta) - F(\eta)}{1 - F(\eta)}, \quad y > 0, \tag{2.24}$$

which is the probability that x exceeds threshold η by at most y given x exceeds the threshold.[11]

Generalized Pareto Distribution estimator Balkema and Haan (1974) show that, for a sufficiently high threshold η, the cumulative distribution function in Equation (2.24) can be approximated by the Generalized Pareto Distribution (GPD) given by

$$G(y) = \begin{cases} 1 - \left(1 + \xi \frac{y}{\sigma}\right)^{-1/\xi}, & \text{if } \xi \neq 0 \\ 1 - \exp(-y/\sigma), & \text{if } \xi = 0, \end{cases} \tag{2.25}$$

where ξ is a shape parameter and $\sigma > 0$ is a scale parameter defined for $y \geq 0$ when $\xi \geq 0$ and $0 \leq y \leq -\sigma/\xi$ when $\xi < 0$. When $\xi > 0$, the distribution becomes the heavy-tailed Pareto distribution. The probability density function of the GPD is given by

$$g(z_t; \xi, \sigma) = \begin{cases} \frac{1}{\sigma}\left[1 + \frac{\xi z_t}{\sigma}\right]^{-(1+1/\xi)}, & \text{if } \xi \neq 0 \\ \frac{1}{\sigma}\exp[-\frac{z_t}{\sigma}], & \text{if } \xi = 0, \end{cases} \tag{2.26}$$

[11] See Christoffersen (2011) for a highly accessible introduction to the use of EVT in risk management.

where the exceedances $\{z_1, ..., z_{N_\eta}\}$ are defined as $z_t = \epsilon_t - \eta$ for $1 \leq t \leq N_\eta$, and N_η is the number of exceedances above the threshold η. Maximizing the likelihood function given by

$$L(\xi, \sigma) = \prod_{i=1}^{N_\eta} \frac{1}{N_\eta} g(z_t; \xi, \sigma), \tag{2.27}$$

yields estimates of ξ and σ.

Using the estimated parameters $\hat{\xi}$ and $\hat{\sigma}$, McNeil and Frey (2000) show that the closed form expression for the $(1 - p)$'th quantile of the GPD distribution is

$$\hat{c}_{1,p}^{GPD} = \eta + \frac{\hat{\sigma}}{\hat{\xi}} \left(\left(\frac{Tp}{N_\eta} \right)^{-\hat{\xi}} - 1 \right), \tag{2.28}$$

where T is the sample size, and that

$$\hat{c}_{2,p}^{GPD} = \hat{c}_{1,p}^{GPD} \left(\frac{1}{1 - \hat{\xi}} + \frac{\hat{\sigma} - \hat{\xi}\eta}{(1 - \hat{\xi})\hat{c}_{1,p}^{GPD}} \right). \tag{2.29}$$

Hence, given a GPD assumption on exceedances, the values for VaR_{T+1}^p and ES_{T+1}^p are obtained by substituting these estimates into Equations (2.11) and (2.12), respectively.

Hill estimator Hill (1975a) provides an alternative estimation method used in EVT. The Hill estimator assumes that $\xi > 0$ and the distribution has heavy tails. Suppose the tail of the conditional distribution of innovations is approximated by the distribution function

$$F(z) = 1 - L(z)z^{-1/\xi} \approx 1 - cz^{-1/\xi}, \tag{2.30}$$

whenever $\epsilon_t > u$, where u is the threshold, and $L(z)$ is a slowly varying function, which we approximate with a constant c. Let k be the number of observations that exceed u. The Hill estimator $\hat{\xi}$ is the maximum likelihood estimator of ξ assuming innovations are i.i.d. from an

unknown distribution given in closed form by

$$\hat{\xi} = \frac{1}{k}\sum_{t=1}^{k} \ln(\hat{\epsilon}_{(T-t+1)}) - \ln(u), \tag{2.31}$$

where $\hat{\epsilon}_{(t)}$ denotes the t'th order statistic of $\hat{\epsilon}_t$ such that $\hat{\epsilon}_{(t)} \geq \hat{\epsilon}_{(t-1)}$ for $t = 2, ..., T$.

Huisman et al. (2001) provide an alternative estimator of ξ that does not require choosing a threshold u. They show that the bias in the Hill estimator is a linear and increasing function of k. Hence, a threshold-free estimate of ξ is the intercept β_0 in the regression

$$\hat{\xi}_k = \beta_0 + \beta_1 k + \nu(k), k = 1, .., K, \tag{2.32}$$

where $\hat{\xi}_k$ is the Hill estimator in Equation (2.31) with threshold $u = \hat{\epsilon}_{(k)}$. Since the variance of $\hat{\xi}_k$ depends on k, $\nu(k)$ is heteroskedastic. To correct for this, they estimate Equation (2.32) using Weighted Least Squares (WLS) with a $(K \times K)$ weighting matrix W, which has $\{\sqrt{1}, .., \sqrt{K}\}$ as diagonal elements and zeros elsewhere. We set $K = T/4$.

Given $\hat{\xi}$ we approximate the tail distribution F by setting $c = \frac{k}{T}u^{1/\hat{\xi}}$, which is derived from the condition $1 - F(u) = \frac{k}{T}$. The estimate of F is

$$\hat{F}(z) = 1 - \frac{k}{T}\left(\frac{z}{u}\right)^{-1/\hat{\xi}}. \tag{2.33}$$

Christoffersen and Gonçalves (2005) show that the $(1-p)$'th quantile of $\hat{F}(z)$ is

$$\hat{c}_{1,p}^{Hill} = u\left(\frac{pT}{k}\right)^{-\hat{\xi}}, \tag{2.34}$$

and that

$$\hat{c}_{2,p}^{Hill} = \frac{\hat{c}_{1,p}^{Hill}}{1-\hat{\xi}}. \tag{2.35}$$

We substitute these estimates into Equations (2.11) and (2.12) to obtain the corresponding estimates of VaR_{T+1}^{p} and ES_{T+1}^{p}.

In Section A.2 of the Appendix A, we conduct a simulation study to estimate the optimal threshold, $\hat{\eta}$, for GPD estimation and, $\hat{u}$, for Hill estimation at the 99%, 97.5% and 95% confidence levels. We find that GPD estimates are optimized by setting $\hat{\eta}$ equal to the 0.85 quantile of innovations for all three confidence levels and that Hill estimates are optimized by setting $\hat{u}$ equal to the same quantile as the confidence level, and we use these thresholds in this paper.

Dynamic VaR and ES models

When estimating the parameters of the location-scale models above this is typically done by fitting to the whole conditional distribution of asset returns using Maximum Likelihood type methods. Such methods are efficient if the dynamics and the distributional specification are correct and in this case the model will also yield optimal forecasts of the risk measures. However, if we think of the dynamic models only as approximations it is not obvious that the estimated model is optimal when the application is to forecast VaR and ES. Indeed, in this situation it may be possible to improve the forecasted risk measures by estimating model parameters using an alternative metric, one that is consistent for VaR and ES, rather than using a (Quasi) Maximum Likelihood approach which focuses on the conditional mean and variance.

An appropriate metric for this problem was provided by Fissler and Ziegel (2016), who showed that the class of FZ loss functions given by

$$L_{FZ}(Y, v, e; p, G_1, G_2) = (\mathbb{1}_{Y \leq v} - p)(G_1(v) - G_1(Y) + \frac{1}{p} G_2(e)v) - G_2(e)(\frac{1}{p}\mathbb{1}_{Y \leq v} Y - e) - \mathcal{G}_2(e), \quad (2.36)$$

where Y denotes the return, $-v$ is the VaR, $-e$ is the ES, G_1 is weakly increasing, G_2 is strictly increasing and strictly positive, and $\mathcal{G}_2' = G_2$, is consistent for VaR and ES. In other words, minimizing the expected FZ loss returns the true VaR and ES and

$$(-VaR_t, -ES_t) = \underset{(v,e)}{\operatorname{argmin}} E_{t-1}\left[L_{FZ}(Y_t, v, e; p, G_1, G_2)\right]. \quad (2.37)$$

To implement this approach for estimation one needs to choose the functions G_1 and G_2. Patton,

Ziegel, and Chen (2019) suggest to set $G_1(x) = 0$ and $G_2(x) = -1/x$ to obtain

$$L_{FZ0}(Y, v, e; p) = -\frac{1}{pe}\mathbb{1}_{Y \leq v}(v - Y) + \frac{v}{e} + \log(-e) - 1, \tag{2.38}$$

which they refer to as the FZ0 loss function, and they provide asymptotic theory for estimating VaR and ES models by minimizing this loss. Thus, one can now use this criterion to estimate parameters of any dynamic specifications like, e.g., an ARMA-GARCH type model.

While GARCH type dynamics could be considered, Patton, Ziegel, and Chen (2019) specify instead a dynamic model for VaR and ES using the Generalized Autoregressive Score (GAS) model of Creal, Koopman, and Lucas (2013) and Harvey (2013) where the forcing variable is a function of the derivative and the Hessian of the FZ0 loss function instead of a log-likelihood. We first consider their one-factor GAS model for VaR and ES, where the risk measures, $v_t = a\exp(\kappa_t)$ and $e_t = b\exp(\kappa_t)$ with $b < a < 0$, are driven by a factor $\kappa_t = \log(\sigma_t)$, interpreted as the log volatility, with dynamics given by

$$\kappa_t = \omega + \beta\kappa_{t-1} + \gamma\frac{1}{b\exp(\kappa_{t-1})}\left(\frac{1}{p}\mathbb{1}_{Y_{t-1} \leq a\exp(\kappa_{t-1})}Y_{t-1} - b\exp(\kappa_{t-1})\right). \tag{2.39}$$

The parameters (a, b, β, γ) can be estimated using the FZ0 loss function while setting $\omega = 0$ ensures identification, see Patton, Ziegel, and Chen (2019) for details. We also consider their Hybrid GAS/GARCH model with

$$\kappa_t = \omega + \beta\kappa_{t-1} + \gamma\frac{1}{b\exp(\kappa_{t-1})}\left(\frac{1}{p}\mathbb{1}_{Y_{t-1} \leq a\exp(\kappa_{t-1})}Y_{t-1} - b\exp(\kappa_{t-1})\right) + \delta\log|Y_{t-1}|, \tag{2.40}$$

where parameters $(a, b, \beta, \gamma, \delta)$ are estimated using the FZ0 loss function and setting $\omega = 0$ again ensures identification.[12] We refer to these two models as the FZ1 and FZH models,

[12]Patton, Ziegel, and Chen (2019) also consider a two-factor model for VaR and ES. However, they found the one-factor model performed better than the two-factor model, so we exclude it from our analysis.

respectively. The 1-day VaR and ES forecasts are given by

$$\text{FZ} - VaR^p_{T+1} = -v_{T+1} = -a\exp(\kappa_{T+1}), \tag{2.41}$$

and

$$\text{FZ} - ES^p_{T+1} = -e_{T+1} = -b\exp(\kappa_{T+1}), \tag{2.42}$$

where κ_{T+1} follows Equation (2.39) in the FZ1 model and Equation (2.40) in the FZH.

2.3.3 Multiperiod VaR and ES forecasts

Basel 2 and 2.5 capital charges are based on VaR with a 10 day time horizon whereas Basel 3 capital charges are based on ES with a 10 day time horizon. During Basel 2, banks would often approximate 10 day VaR by multiplying 1-day VaR by $\sqrt{10}$, which is correct when returns are normally distributed. Due to the non-normality of returns this scaling method is explicitly prohibited in Basel 3 (BCBS (2019)). Basel 3 regulation though, does allow 10-day forecasts to be calculated using overlapping observations, which is how we calculate HS forecasts. Define the sum of 10-day losses conditional on time t as

$$L_t[10] = \sum_{k=t+1}^{t+10} L_k. \tag{2.43}$$

The goal of multiperiod forecasts is to calculate the VaR and ES of $L_T[10]$ given the information at time T. The HS estimate for VaR^p_{T+10} is

$$\text{HS} - VaR^p_{T+10} = Q_{1-p}(\{L_t[10]\}), \tag{2.44}$$

where $Q_{1-p}(\{L_t[10]\})$ denotes the $(1-p)$'th empirical quantile of the 10-day losses $\{L_t[10]\}_{t=1}^{T-10}$. The HS estimate for ES^p_{T+10} is

$$\text{HS} - ES^p_{T+10} = \frac{1}{\#(L_t[10] > HS - VaR^p_{T+10})}\Big(\sum_{L_t[10]>HS-VaR^p_{T+10}} L_t[10]\Big), \tag{2.45}$$

where $\#(L_t[10] > HS - VaR^p_{T+10})$ denotes the number of 10-day losses exceeding $HS - VaR^p_{T+10}$.

For GARCH models, the volatility forecast and conditional distribution of 10-day losses is only available in closed form for normally distributed innovations. The 10-day GARCH variance forecast for normal innovations is

$$\sigma^2_{T+10} = \frac{\omega}{1-\gamma}\left(10 - \frac{1-\gamma^{10}}{1-\gamma}\right) + \frac{1-\gamma^{10}}{1-\gamma}\sigma^2_{T+1}, \tag{2.46}$$

where $\gamma = \alpha + \beta < 1$ (Tsay (2010)). The normal estimate for VaR^p_{T+10} is

$$\text{Norm} - VaR^p_{T+10} = 10\mu + \sigma_{T+10}c^{Norm}_{1,p}, \tag{2.47}$$

and the estimate for ES^p_{T+10} is

$$\text{Norm} - ES^p_{T+10} = 10\mu + \sigma_{T+10}c^{Norm}_{2,p}. \tag{2.48}$$

Taking the limit as $\gamma \to 1$ the 10-day RiskMetrics variance forecast is seen to be $\sigma^2_{T+10} = 10\sigma^2_{T+1}$. Hence, the RM estimate for VaR^p_{T+10} and ES^p_{T+10} is $\sqrt{10}\text{RM} - VaR^p_{T+1}$ and $\sqrt{10}\text{RM} - ES^p_{T+1}$ respectively. This relationship is referred to as the square root of time rule under RiskMetrics.

In all other cases multiperiod risk measures are obtained through simulation. We approximate the conditional distribution of $L_T[10] = L_{T+1} + .. + L_{T+10}$ by simulating future paths of losses

$$L_{T+k} = \mu + \sigma_{T+k}\epsilon_{T+k}, \tag{2.49}$$

where the conditional variance in GARCH models follows

$$\sigma^2_{T+k} = \omega + \alpha(L_{T+k-1} - \mu)^2 + \beta\sigma^2_{T+k-1}, \tag{2.50}$$

for $\sigma^2_{T+1}, .., \sigma^2_{T+10}$. We perform B = 2,000 simulations, resulting in a sample $\{L_{b,T}[10]\}_{b=1}^{B}$ of 10 day losses indexed by b. The GARCH estimate for VaR^p_{T+10} is then

$$VaR^p_{T+10} = Q_{1-p}(\{L_{b,T}[10]\}), \tag{2.51}$$

$Q_{1-p}(\{L_{b,T}[10]\})$ denotes the $(1-p)$'th empirical quantile of the simulated 10-day losses $\{L_{b,T}[10]\}_{b=1}^{B}$. The GARCH estimate for $ES^{\,p}_{T+10}$ is

$$ES^{\,p}_{T+10} = \frac{1}{\#(L_{b,T}[10] > VaR^p_{T+10})}\left(\sum_{L_{b,T}[10]>VaR^p_{T+10}} L_{b,T}[10]\right), \tag{2.52}$$

where $\#(L_{b,T}[10] > HS - VaR^p_{T+10})$ denotes the number of simulated 10-day losses $\{L_{b,T}[10]\}_{b=1}^{B}$ exceeding VaR^p_{T+10}. The simulated innovations $\{\epsilon_{T+1}, .., \epsilon_{T+10}\}$ in Equation(2.49) are drawn directly from the estimated distribution, or in the case of the Filtered Historical Simulation model drawn with replacement from the empirical distribution of centered innovations $\hat{\epsilon}_t - \bar{\hat{\epsilon}}$. A similar approach is used for the FZ models with the only change that here the innovations are given by $\eta_t = L_t/\exp(\kappa_t)$.

For the EVT models, we follow the procedure in McNeil and Frey (2000). For the GPD model, we estimate the positive and negative threshold $\eta^{\pm}$ as the 0.15 and 0.85 quantiles of $\hat{\epsilon}_t$ respectively. Using the thresholds, we estimate the positive and negative shape $\xi^{\pm}$ and scale $\sigma^{\pm}$ parameters. We draw innovations from the empirical distribution of innovations $\hat{\epsilon}_t$. If the drawn innovation is greater than η^+, we replace the innovation with $\eta^+ + y^+$, where y^+ is drawn from a GPD distribution with shape ξ^+ and scale σ^+. Similarly, if the drawn innovation is less than η^-, we replace the innovation with $\eta^- - y^-$, where y^- is GPD distributed with shape ξ^- and scale σ^-. If the drawn innovation is between η^- and η^+, we use the innovation itself. We take

a similar approach for the Hill model, estimating the positive and negative threshold $u^{\pm}$ as the p and $(1-p)$'th quantiles of the innovations $\hat{\epsilon}_t$ respectively. Using the thresholds, we estimate the positive and negative shape $\xi^{\pm}$ parameter. If the drawn innovation is greater than u^+, we replace the innovation with $u^+ + z^+$, where z^+ is drawn from a Pareto distribution with shape parameter ξ^+. If the drawn innovation is less than u^-, we replace the innovation with $u^- + z^-$, where z^- is drawn from a Pareto distribution with shape parameter ξ^-. If the drawn innovation is between u^- and u^+, we use the innovation itself.

2.4 Model Backtesting

Basel 3 changes the regulatory market risk measure from VaR at the 99% confidence level to ES at the 97.5% confidence level. With this change, banks are strongly motivated to identify which models are correctly specified under the new risk measure and regulatory supervisors are particularly interested in identifying the set of models that underestimate risk, since banks using these models may be undercapitalized prior to a financial shock. When Basel 3 was introduced, though, a major criticism against using ES for regulation was the lack of available backtesting due to the risk measure not being elicitable (Gneiting (2011)).[13] However, Fissler, Ziegel, and Gneiting (2015) showed that ES is in fact jointly elicitable with VaR, and a recent literature on joint VaR and ES backtesting has been developed by academics. This section uses these recently developed and state-of-the-art methods, which we simply refer to as ES backtests, to examine which of the models are correctly specified and provide the best risk forecasts.

Since VaR is still used for regulatory backtesting and for setting the capital multiplier, we first consider individual backtests for VaR. The methods we use to backtest predicted VaR are standard and can be found in, e.g., Christoffersen (2009). Specifically, we report the actual exceedances (Actual) and p-values from Kupiec (1995)'s Unconditional Coverage (UC) test,

[13] A variable is elicitable if it can be defined as the minimizer of a mean scoring function. Gneiting (2011) shows that ES lacks elicitability while VaR is elicitable allowing backtesting VaR but not ES individually.

Christoffersen (1998)'s Conditional Coverage (CC) test, Christoffersen and Pelletier (2004)'s Duration (Dur) test, and Engle and Manganelli (2004)'s Dynamic Quantile (DQ) test. In terms of ES backtests, Nolde and Ziegel (2017) suggest that these can be separated into two categories: 1) traditional backtests which can be used to determine the correctly specified models, and 2) comparative backtests which can be used to select the models that provide superior forecasts. For the traditional backtests, we next report p-values for McNeil and Frey (2000)'s residual test (ER), Bayer and Dimitriadis (2018)'s strict, auxiliary, and one-sided regression tests (ESR), Nolde and Ziegel (2017)'s conditional calibration test (CCa), and Gordy and McNeil (2020)'s Spectral Backtests. Finally, for the comparative backtests, we report average FZ0 losses with Hansen, Lunde, and Nason (2011)'s Model Confidence Set, Diebold and Mariano (2002) t-statistics from Patton (2019)'s FZ0 backtest, and Ziegel et al. (2017)'s Murphy Diagrams.[14]

We empirically evaluate the various models' performance in terms of estimating 1-day VaR and ES for our representative portfolio from January, 1997, to February, 2020.[15] For HS and RM, we use a 250-day rolling estimation window. While most banks use short windows for their HS estimation, dynamic models generally require a larger estimation window to reduce estimation error. We therefore choose to report results using a rolling estimation window of $T = 2,000$ days and include data from 1989 for estimation purposes. In addition to results at the 99% and 97.5% confidence levels, we also consider a 95% confidence level. The 95% confidence level is not currently used for regulatory purposes. However, due to the larger tail sample, estimation error could be reduced across models at this level and hence this confidence level could be used for future regulation. For visual clarity, hypothesis tests that are rejected with 95% confidence are bold. Previewing our results, we reject that the HS, RM, Normal, STD, and GED models are correctly specified for VaR and ES backtests at traditional levels and find that these models provide poor forecasts. We also reject the FZ models in the conditional

[14] Further details on all these ES backtests can be found in Section A.3 of Appendix A.

[15] Although Basel sets capital charges based on 10-day risk estimates, multi-horizon backtests are challenging to conduct due to overlapping observations and changing portfolio compositions. We follow the existing literature and conduct backtests at a 1-day horizon which is also the horizon used for Basel backtesting.

spectral backtests and find that these models provide inferior forecasts. We cannot reject the hypothesis of correct specification for most of the skewed GARCH models and find that these models also provide superior VaR and ES forecasts.

2.4.1 VaR Backtests

Panel A of Table 2.4 reports 1-day VaR backtesting results at the 99% confidence level for which 57 exceedances are expected. The UC and CC p-values indicate we can reject the hypothesis that the HS, RM, Normal, STD, and GED models are correctly specified. These models have too many exceedances from underestimating VaR and are likely to have serially correlated exceedances. The Dur p-values indicate we can reject the hypothesis that HS has the correct duration between exceedances, likely because the model is misspecified against volatility clustering. Additionally, the DQ p-values indicate that HS, RM, Normal, STD, and FZ1 models are misspecified. Panel B of Table 2.4 reports results at the 97.5% confidence level for which 144 exceedances are expected. The UC and CC p-values indicate we can reject the same set of models as with the 99% confidence level. The Dur p-values indicate we can reject the hypothesis that the HS, RM, STD, GED, GPD, and FZ1 models have independent durations between exceedances. Additionally, the DQ p-values indicate that the HS, RM, Normal, STD, FZ1, and now also the FZH models are misspecified. We cannot reject the hypothesis of correct specification for FHS, SSTD, SGED, SGT, Hill, and HillH for either confidence level.

Next, Panel C of Table 2.4 reports the results at the 95% confidence level for which 288 exceedances are expected. The UC, CC, and Dur test results are similar to the previous confidence levels. Interestingly, the DQ p-values indicate we can reject the hypothesis of correct specification for every model except the two semi-parametric FZ models. This finding is consistent with Manganelli and Engle (2001), who find that their CAViaR model outperforms GARCH models at the 95% confidence level. The two semi-parametric FZ models are an extension of the CAViaR model and as such are expected to outperform GARCH models at this lower confidence level.

Table 2.4: Value at Risk Backtests

	HS	RM	FHS	Norm	STD	SSTD	GED	SGED	SGT	GPD	Hill	HillH	FZ1	FZH
Panel A: Results at a 99% confidence level (Expected exceedances = 57)														
Actual	80	116	59	105	80	64	78	65	65	52	59	59	53	54
UC	**0.01**	**0.00**	0.87	**0.00**	**0.01**	0.42	**0.01**	0.35	0.35	0.44	0.87	0.87	0.52	0.61
CC	**0.00**	**0.00**	0.88	**0.00**	**0.00**	0.68	**0.01**	0.62	0.62	0.59	0.88	0.88	0.66	0.73
Dur	**0.00**	0.17	0.51	0.09	0.08	0.49	0.23	0.39	0.33	0.61	0.51	0.51	0.08	0.45
DQ	**0.00**	**0.00**	0.33	**0.00**	**0.00**	0.45	0.07	0.20	0.20	0.32	0.33	0.33	**0.00**	0.49
Panel B: Results at a 97.5% confidence level (Expected exceedances = 144)														
Actual	178	200	141	204	197	149	180	141	151	131	141	141	165	147
UC	**0.01**	**0.00**	0.77	**0.00**	**0.00**	0.70	**0.00**	0.77	0.59	0.25	0.77	0.77	0.09	0.83
CC	**0.00**	**0.00**	0.21	**0.00**	**0.00**	0.31	**0.00**	0.21	0.31	0.06	0.21	0.21	0.05	0.06
Dur	**0.00**	**0.03**	0.05	0.07	**0.04**	0.05	**0.03**	0.07	0.08	**0.02**	0.05	0.05	**0.03**	0.07
DQ	**0.00**	**0.00**	0.24	**0.00**	**0.00**	0.12	**0.00**	0.12	0.16	0.11	0.24	0.24	**0.01**	**0.02**
Panel C: Results at a 95% confidence level (Expected exceedances = 288)														
Actual	314	312	297	333	357	310	331	292	314	299	297	297	329	309
UC	0.14	0.17	0.63	**0.01**	**0.00**	0.21	**0.01**	0.85	0.14	0.55	0.63	0.63	**0.02**	0.23
CC	**0.00**	0.12	0.30	**0.02**	**0.00**	0.38	**0.04**	0.53	0.25	0.31	0.30	0.30	0.06	0.33
Dur	**0.00**	**0.02**	0.28	0.42	0.27	0.58	0.54	0.32	0.31	0.22	0.28	0.28	0.89	0.78
DQ	**0.00**	**0.00**	**0.02**	**0.00**	**0.00**	**0.04**	**0.00**	**0.02**	**0.01**	**0.01**	**0.02**	**0.02**	0.13	0.58

This table shows the VaR backtesting results of the representative portfolio from 01/1997 to 02/2020. Each panel reports results for a particular confidence level. In each panel, Row 1 shows the actual number of exceedances. Rows 2 to 5 display two-sided p-values for the Unconditional Coverage, Conditional Coverage, Duration, and Dynamic Quantile backtests. Models with p-values below 0.05 are in bold.

In summary, Table 2.4 shows that irrespective of the confidence level we can reject that the HS, RM, Normal, STD, and GED models are correctly specified. These models underestimate VaR and often have clustered exceedances. Also, since the symmetric GARCH and FZ parametric models are often rejected, we conclude that modeling skewness is crucial for estimating VaR in financial returns. At the extreme 95% confidence level, the FZH model is the only model to not be rejected by any test. However, the FHS, SSTD, SGED, SGT, Hill, and HillH models have the most accurate VaR estimates for the 99% and 97.5% confidence levels, the relevant levels for Basel regulation.

2.4.2 Traditional Backtests

Panel A of Table 2.5 reports traditional ES backtest results at the 99% confidence level. The ER p-values indicate that we can reject the hypothesis that the HS, RM, Normal, GED, and SGED models have the correct ES estimates on average. Since we evaluate the one-sided hypothesis of the unconditional ER test, we can conclude that these models systemically underestimate ES. The ESR Strict and Aux p-values indicate we can reject the hypothesis that the HS, RM,

Table 2.5: Traditional Expected Shortfall Backtests

	HS	RM	FHS	Norm	STD	SSTD	GED	SGED	SGT	GPD	Hill	HillH	FZ1	FZH
Panel A: Results at a 99% confidence level														
ER	**0.02**	**0.00**	0.53	**0.00**	0.10	0.41	**0.00**	**0.01**	0.41	0.20	0.57	0.14	0.31	0.07
ESR Strict	**0.00**	**0.00**	0.98	**0.00**	0.09	0.79	**0.01**	0.26	0.73	0.96	0.99	0.62	0.28	0.87
ESR Aux	**0.00**	**0.00**	0.93	**0.00**	0.09	0.84	**0.01**	0.23	0.80	0.94	0.90	0.69	0.19	0.92
ESR Int	**0.00**	**0.00**	0.27	**0.00**	**0.01**	0.13	**0.00**	**0.02**	0.11	0.32	0.29	0.09	0.32	0.15
CCa	**0.01**	**0.00**	0.97	**0.00**	0.07	0.87	**0.01**	0.17	0.85	0.63	1.00	0.84	0.76	0.62
Panel B: Results at a 97.5% confidence level														
ER	**0.04**	**0.00**	0.74	**0.00**	0.25	0.39	**0.00**	**0.02**	0.48	0.62	0.84	**0.00**	0.93	0.86
ESR Strict	**0.01**	**0.00**	0.90	**0.00**	**0.02**	0.79	**0.00**	0.46	0.77	0.76	0.71	0.08	0.99	0.90
ESR Aux	**0.00**	**0.00**	0.85	**0.00**	**0.02**	0.84	**0.00**	0.49	0.81	0.70	0.65	0.11	0.95	0.85
ESR Int	**0.00**	**0.00**	0.43	**0.00**	**0.00**	0.14	**0.00**	0.06	0.13	0.53	0.56	**0.01**	0.44	0.55
CCa	**0.01**	**0.00**	1.00	**0.00**	**0.01**	0.83	**0.00**	0.36	0.79	0.35	1.00	0.10	1.00	0.99
Panel C: Results at a 95% confidence level														
ER	**0.04**	**0.00**	0.84	**0.00**	**0.02**	0.42	**0.00**	0.12	0.54	0.91	0.94	**0.00**	0.96	0.93
ESR Strict	**0.00**	**0.00**	0.81	**0.00**	**0.00**	0.54	**0.00**	0.54	0.53	0.76	0.45	**0.00**	1.00	0.81
ESR Aux	**0.00**	**0.00**	0.90	**0.00**	**0.00**	0.58	**0.00**	0.57	0.55	0.80	0.54	**0.00**	0.98	0.85
ESR Int	**0.00**	**0.00**	0.39	**0.00**	**0.00**	0.07	**0.00**	0.06	0.06	0.46	0.63	**0.00**	0.26	0.37
CCa	0.05	**0.00**	1.00	**0.00**	**0.00**	0.42	**0.00**	0.42	0.38	1.00	1.00	**0.00**	1.00	1.00

This table shows the traditional ES backtesting results of the representative portfolio from 01/1997 to 02/2020. Each panel reports results for a particular confidence level. In each panel, Row 1 shows the one-sided p-values for the Exceedance Residual test. Rows 2 and 3 show the two-sided p-values for the Strict and Auxiliary ES regression backtests. Rows 4 and 5 show the one-sided p-values for the Intercept ES regression and Conditional Calibration backtests. Models with p-values below 0.05 are in bold.

Normal, and GED models have accurate ES estimates. Specifically, we can reject the hypothesis of zero intercept and unit intercept when regressing exceedances from these models on ES. The ESR Int p-values for the one-sided intercept backtest show that these models along with STD and SGED have intercepts that are too low, confirming that these models underestimate ES on average. The CCa p-values indicate we can reject the hypothesis that the HS, RM, Normal, and GED models have accurate VaR and ES estimates. Since we conduct the one-sided CCa test where the null hypothesis is that the VaR and ES estimates are weakly greater than their true values on average, we conclude that the rejected models underestimate risk. We cannot reject the hypothesis that VaR and ES are correctly specified for FHS, SSTD, SGT, GPD, Hill, HillH, FZ1, and FZH in any traditional ES backtest at this level. The traditional ES backtest results are nearly identical for the 97.5% and 95% confidence levels in Panels B and C, except that HillH is rejected at these confidence levels.

Panel A of Table 2.6 reports unconditional spectral backtest results for the wide interval from [0.95,0.995] using the Uniform (Uni), Arcsin (Arc), and Epanechnikov (Epa) continu-

ous kernel density functions. The continuous kernel p-values indicate that we can reject the hypothesis that the RM, Normal, STD, GED, and HillH models have uniformly distributed probability integral transform (PIT) values. Given the results from the traditional backtests in Table 2.5, we conclude these models have thinner tails than actual losses have, which likely results in underrepresented tail PIT values. Panel B of Table 2.6 reports unconditional spectral backtest results for the narrow interval from [0.97,0.98], which is the neighborhood around the 97.5% confidence level used for Basel 3. The p-values for all three kernels indicate we reject the same set of models as for the wide interval. Panel C of Table 2.6 reports unconditional spectral backtest results for the uniform 3-level points (0.95,0.975,0.99), the main confidence levels of interest in our paper. The p-values indicate that we reject the hypothesis that the RM, Normal, STD, and GED models have uniformly distributed PIT values. The set of models rejected is consistent across all the unconditional backtests, indicating these models have non-uniform PIT-values and are unlikely to be correctly specified. We cannot reject that HS, FHS, SSTD, SGED, SGT, GPD, Hill, FZ1, and FZH have uniformly distributed PIT-values in any unconditional spectral backtest.

Panel D of Table 2.6 reports conditional spectral backtest results for the wide interval. The continuous kernel p-values indicate that we can reject the hypothesis that the HS, RM, Normal, STD, GED, HillH, FZ1, and FZH models have uniformly distributed PIT-values and are serially independent. Interestingly, the HS, FZ1, and FZH models pass the unconditional backtests, but are rejected in the conditional backtest. This indicates that these models exhibit correlated spectrally transformed PIT-values. These models fail to use all available information when forecasting, resulting in temporal dependence between PIT-values.[16] Panel E of Table 2.6 reports conditional spectral backtest results for the narrow interval. The set of rejected models are the same as for the wide interval, reaffirming that these models have non-uniform or dependent PIT-values and are unlikely to be correctly specified. Panel F of Table 2.6 reports conditional spectral backtest results for the uniform 3-level points. The p-values indicate that

[16]This is confirmed by these models inferior performance in comparative backtests below.

Table 2.6: Spectral Backtests

	HS	RM	FHS	Norm	STD	SSTD	GED	SGED	SGT	GPD	Hill	HillH	FZ1	FZH
Panel A: Unconditional Coverage on a Wide Interval [0.95,0.995]														
Uni	0.19	**0.00**	0.87	**0.00**	**0.00**	0.17	**0.00**	0.85	0.17	0.80	0.72	**0.00**	0.90	0.93
Arc	0.23	**0.00**	0.94	**0.00**	**0.00**	0.16	**0.00**	0.74	0.15	0.93	0.68	**0.01**	0.91	1.00
Epa	0.17	**0.00**	0.82	**0.00**	**0.00**	0.20	**0.00**	0.96	0.21	0.66	0.75	**0.00**	0.90	0.88
Panel B: Unconditional Coverage on a Narrow Interval [0.97,0.98]														
Uni	0.11	**0.00**	0.62	**0.00**	**0.00**	0.43	**0.00**	0.69	0.46	0.28	0.78	**0.00**	0.99	0.93
Arc	0.13	**0.00**	0.61	**0.00**	**0.00**	0.40	**0.00**	0.68	0.43	0.29	0.78	**0.00**	0.99	1.00
Epa	0.09	**0.00**	0.63	**0.00**	**0.00**	0.48	**0.00**	0.69	0.49	0.27	0.78	**0.00**	0.97	0.86
Panel C: Unconditional Coverage on a Discrete 3-level points (0.95,0.975,0.99)														
Uni	0.21	**0.00**	0.92	**0.00**	**0.00**	0.31	**0.00**	0.83	0.21	0.77	0.73	0.11	0.88	0.92
Panel D: Conditional Coverage on a Wide Interval [0.95,0.995]														
Uni	**0.00**	**0.00**	0.32	**0.00**	**0.00**	0.09	**0.00**	0.15	0.08	0.32	0.82	0.06	**0.00**	**0.01**
Arc	**0.00**	**0.00**	0.37	**0.00**	**0.00**	0.11	**0.00**	0.19	0.10	0.38	0.86	0.18	**0.00**	**0.02**
Epa	**0.00**	**0.00**	0.29	**0.00**	**0.00**	0.08	**0.00**	0.14	0.08	0.27	0.77	**0.02**	**0.00**	**0.00**
Panel E: Conditional Coverage on a Narrow Interval [0.97,0.98]														
Uni	**0.00**	**0.00**	0.23	**0.00**	**0.00**	0.12	**0.00**	0.12	0.13	0.18	0.85	**0.01**	**0.00**	**0.00**
Arc	**0.00**	**0.00**	0.24	**0.00**	**0.00**	0.13	**0.00**	0.12	0.13	0.17	0.82	**0.01**	**0.00**	**0.00**
Epa	**0.00**	**0.00**	0.22	**0.00**	**0.00**	0.11	**0.00**	0.12	0.13	0.19	0.87	**0.01**	**0.00**	**0.00**
Panel F: Conditional Coverage on a Discrete 3-level points (0.95,0.975,0.99)														
Uni	**0.00**	**0.00**	0.36	**0.00**	**0.00**	0.12	**0.00**	0.15	0.08	0.52	0.93	0.52	**0.00**	**0.02**

This table shows the Spectral Backtest results of the representative portfolio from 01/1997 to 02/2020. Each panel reports results for a particular test type and interval. Panels A and B show p-values for the unconditional Z-test using a uniform, arcsin, and epanechnikov kernel. Panel C shows p-values for the unconditional Z-test using a discrete uniform kernel. Panels D and E show p-values for the conditional Martingale Difference test using a uniform, arcsin, and epanechnikov kernel and the conditioning variable transformation $h(p) = |2p - 1|^4$ with 4 lags. Panel F shows p-values for the conditional Z-test using a discrete uniform kernel. HS and GARCH models have the same PIT values across confidence levels. FZ PIT values are estimated at the 95% confidence level. Models with p-values below 0.05 are in bold.

we reject the hypothesis that the HS, RM, Normal, STD, GED, FZ1, and FZH models have uniformly distributed and independent PIT-values. We cannot reject that FHS, SSTD, SGED, SGT, GPD, and Hill have uniformly distributed and independent PIT-values in any conditional spectral backtest.

In summary, Tables 2.5 and 2.6 show that the HS, RM, Normal, STD, GED, SGED, HillH, FZ1, and FZH models are likely not correctly specified for VaR and ES at the 99%, 97.5%, and 95% confidence levels. These models either underestimate tail risk or have dependent forecasts. The FHS, SSTD, SGT, GPD, and Hill models are, on the other hand, never rejected by the traditional ES backtests. In particular, we cannot reject that these models have independent forecasts or PIT-values, suggesting that they are using the full information set when making forecasts. These models can accommodate the heavy tails and skewness of financial returns and appear to be correctly specified for VaR and ES. While the traditional backtests in this

section determines which models are incorrectly specified for VaR and ES, the next section compares across models to evaluate the set of best forecasting models.

2.4.3 Comparative Backtests

Table 2.7 shows results for comparative ES backtests using the FZ0 loss function. Patton, Ziegel, and Chen (2019) show that FZ0 is the only FZ loss function that generates loss differences that are homogeneous of degree zero, a property that has been shown in volatility forecasting applications to lead to higher power in Diebold and Mariano (2002) tests. We consider results at the 97.5% confidence level here and, since the results are qualitatively similar, report 99% and 95% confidence levels in Section A.4 of Appendix A. Panel A in Table 2.7 reports the average FZ0 loss for each model with a star besides models in the 75% Model Confidence Set (MCS). The FHS, SSTD, SGED, SGT, GPD, and Hill models have the lowest average loss of -0.47 and are the only models that belong to the MCS. This is consistent with the traditional backtesting results, where the same set of models fail to be rejected as correctly specified in every backtest. The Normal, STD, GED, HillH, FZ1, and FZH models, on the other hand, have average losses between -0.46 to -0.42 and do not belong to the MCS. This is consistent with the traditional backtesting results, where these models are sometimes rejected as correctly specified. HS and RM have relatively higher average losses of -0.27 and -0.37, and are not in the MCS which is consistent with their rejection for nearly every traditional backtest.

Next, Panel B in Table 2.7 reports Diebold-Mariano test statistics on average FZ0 losses. Along each row, a negative number indicates the row model outperforms the column model. A t-stat greater than 1.96 in absolute value indicates the loss difference is statistically different from zero at the 95% confidence level. The HS row is positive and statistically significant for each column model, indicating that HS performs worse than every other model. The RM row shows the model underperforms every column model except for HS. The SSTD row on the other hand is negative for every column model indicating that SSTD has the lowest average loss. However, the loss difference is not statistically significant for models in the MCS. In-

Table 2.7: Comparative Expected Shortfall Backtests at the 97.5% Confidence Level

	HS	RM	FHS	Norm	STD	SSTD	GED	SGED	SGT	GPD	Hill	HillH	FZ1	FZH
Panel A: Average FZ0 Loss														
Loss	-0.27	-0.37	-0.47*	-0.43	-0.45	-0.47*	-0.46	-0.47*	-0.47*	-0.47*	-0.47*	-0.46	-0.42	-0.44
Panel B: Diebold-Mariano Test Statistics for FZ0 Loss														
HS	NA	2.46	4.09	3.51	4.03	4.20	4.10	4.17	4.19	4.04	4.08	4.13	3.50	3.54
RM	**-2.46**	NA	3.55	2.97	3.89	3.83	4.00	3.73	3.81	3.40	3.51	3.79	1.88	2.23
FHS	**-4.09**	**-3.55**	NA	**-2.45**	-1.39	0.50	-1.11	0.32	0.41	-0.43	0.03	-0.54	**-2.74**	**-2.26**
Norm	**-3.51**	**-2.97**	2.45	NA	3.99	3.00	3.92	2.78	2.98	2.25	2.39	2.74	-0.10	0.42
STD	**-4.03**	**-3.89**	1.39	**-3.99**	NA	2.05	2.24	1.74	2.01	1.18	1.34	1.57	-1.73	-1.06
SSTD	**-4.20**	**-3.83**	-0.50	**-3.00**	**-2.05**	NA	-1.80	-0.40	-0.42	-0.58	-0.42	-1.21	**-2.97**	**-2.39**
GED	**-4.10**	**-4.00**	1.11	**-3.92**	**-2.24**	1.80	NA	1.50	1.75	0.92	1.06	1.25	**-2.03**	-1.34
SGED	**-4.17**	**-3.73**	-0.32	**-2.78**	-1.74	0.40	-1.50	NA	0.18	-0.51	-0.27	-0.87	**-2.93**	**-2.35**
SGT	**-4.19**	**-3.81**	-0.41	**-2.98**	**-2.01**	0.42	-1.75	-0.18	NA	-0.50	-0.35	-1.20	**-2.93**	**-2.35**
GPD	**-4.04**	**-3.40**	0.43	**-2.25**	-1.18	0.58	-0.92	0.51	0.50	NA	0.47	-0.18	**-2.71**	**-2.21**
Hill	**-4.08**	**-3.51**	-0.03	**-2.39**	-1.34	0.42	-1.06	0.27	0.35	-0.47	NA	-0.47	**-2.72**	**-2.25**
HillH	**-4.13**	**-3.79**	0.54	**-2.74**	-1.57	1.21	-1.25	0.87	1.20	0.18	0.47	NA	**-2.63**	**-2.09**
FZ1	**-3.50**	**-1.88**	2.74	0.10	1.73	2.97	2.03	2.93	2.93	2.71	2.72	2.63	NA	0.88
FZH	**-3.54**	**-2.23**	2.26	-0.42	1.06	2.39	1.34	2.35	2.35	2.21	2.25	2.09	-0.88	NA

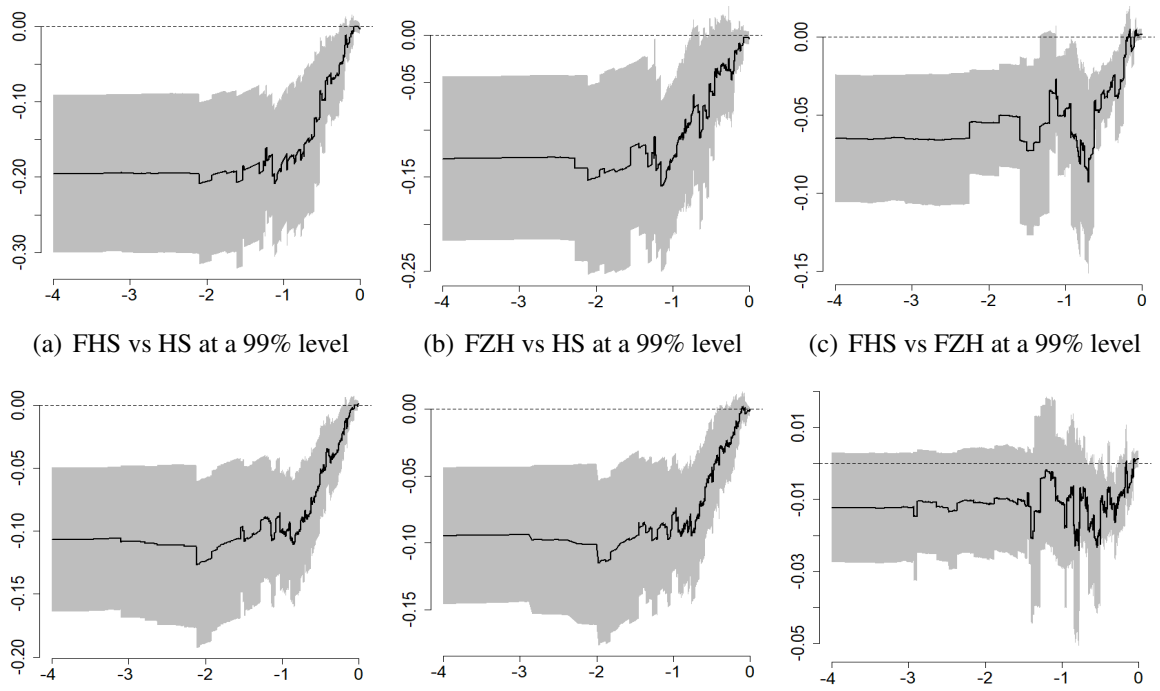
This table shows the comparative ES backtesting results of the representative portfolio from 01/1997 to 02/2020. Results are based on the FZ0 loss function at the 97.5% confidence level. Panel A shows the average FZ0 loss with a star beside models in the 75% model confidence set. Panel B shows t-statistics from Diebold-Mariano tests comparing FZ0 average losses. A negative value indicates that the row model has lower average loss than the column model. t-statistics greater than 1.96 in absolute value indicate the loss difference is significantly different from zero at the 95% confidence level. Models with t-stat below -1.96 are in bold.

terestingly, the FZ1 and FZH columns are positive and statistically significant for the models in the MCS, indicating they perform worse than this group of skewed GARCH models. This finding contradicts the results in Patton, Ziegel, and Chen (2019) who find that semi-parametric FZ models outperform GARCH models. In their analysis however, models are estimated once whereas we use a rolling window estimation procedure. Our results show that after accounting for time-varying parameters and changes to the distribution of innovations, GARCH models have superior forecasts compared to semi-parametric FZ models.[17]

Figure 2.1 plots selected Murphy diagrams at the 99% and 97.5% confidence levels, comparing pairwise FHS to HS, FZH to HS, and FHS to FZH across the entire class of consistent loss functions. Complementary to the Murphy diagrams, Table 2.8 reports minimal Westfall-Young (WY) p-values, formally testing for forecasting dominance across a grid of thresholds. Panels (a) and (d) in the figure compare FHS to HS at the 99% and 97.5% confidence levels and

[17]Section A.4 of Appendix A shows that skewed GARCH models also outperform FZ models at the 99% and 95% confidence levels, though the loss differences are not statistically significant at the 95% confidence level. Section A.5 of Appendix A confirms that skewed GARCH models have lower losses than FZ models for the risk factor portfolios and liquidity horizon portfolios.

Figure 2.1: Murphy Diagrams at the 97.5% and 99% confidence levels



(a) FHS vs HS at a 99% level (b) FZH vs HS at a 99% level (c) FHS vs FZH at a 99% level

(d) FHS vs HS at a 97.5% level (e) FZH vs HS at a 97.5% level (f) FHS vs FZH at a 97.5% level

This figure shows Murphy diagrams of the representative portfolio from 01/1997 to 02/2020 for the 99% and 97.5% confidence levels. In each panel, the vertical axis is the difference in average loss between models, where a negative value indicates the first model outperforms the second model. The horizontal axis is the threshold value, where each threshold corresponds to a different consistent loss function. Confidence intervals at the 95% level are plotted in gray.

show that the loss difference is negative at nearly every threshold, indicating FHS outperforms HS. The 95% confidence interval shows that the difference is nearly always statistically different from zero. The WY p-values in Row 1 of Table 2.8 show that we can reject the hypothesis of HS dominating FHS at the 99%, 97.5%, and 95% confidence levels. The WY p-values in Row 2 show that we cannot reject FHS dominating HS in any confidence level. Next, Panels (b) and (e) in Figure 2.1 show that FZH outperforms HS. The WY p-values in Row 3 of Table 2.8 show that we can reject the hypothesis that HS weakly dominates FZH, while Row 4 shows that we cannot reject that FZH weakly dominates HS.

Finally, Panels (c) and (f) in Figure 2.1 plot the Murphy diagram comparing FHS to FZH at the 99% and 97.5% confidence levels. The negative loss difference shows that FHS outper-

Table 2.8: Tests of Forecast Dominance

Hypothesis	99% Confidence Level	97.5% Confidence Level	95% Confidence Level
HS weakly dominates FHS	**0.002**	**0.000**	**0.000**
FHS weakly dominates HS	0.940	0.702	0.808
HS weakly dominates FZH	**0.020**	**0.004**	**0.004**
FZH weakly dominates HS	0.982	0.912	0.406
FZH weakly dominates FHS	0.050	**0.048**	**0.002**
FHS weakly dominates FZH	0.688	0.904	0.800

This table shows the tests of forecast dominance of the representative portfolio from 01/1997 to 02/2020 for the 99%, 97.5%, and 95% confidence levels. Each row shows the minimal Westfall-Young p-values for the hypothesis of weak dominance. Models with p-values below 0.05 are in bold.

forms FZH across different consistent loss functions. Rows 5 and 6 of Table 2.8 report WY p-values for the hypothesis of weak dominance between FHS and FZH. Row 6 shows that we cannot reject the hypothesis that FHS dominates FZH at the 99%, 97.5%, and 95% confidence levels. However, Row 5 shows we can reject the hypothesis that FZH dominates FHS at the 97.5% and 95% confidence levels only. The comparative backtest found FHS outperforms FZH for the FZ0 loss function, while the Murphy diagram and WY p-values confirm that FHS outperforms FZH across a large family of loss functions.

In summary, the comparative backtests show that FHS, SSTD, SGED, SGT, GPD, and Hill consistently have the best performance across the 99%, 97.5%, and 95% confidence levels. The HS, RM, Normal, STD, GED, HillH, FZ1, and FZH models have statistically higher losses compared to the best performing models. We also find that the best performing GARCH models had lower average losses than the FZ models, and we fail to reject the hypothesis that FHS weakly dominates FZH. The set of best performing models is broadly consistent with the set of correctly specified models in the traditional backtests. Considering all the results reported in this section, we conclude that skewed GARCH models are likely correctly specified and perform the best in terms of forecasting risk measures. The semi-parametric FZ1 and FZH models are generally outperformed by the skewed GARCH models. The symmetric Normal,

STD, and GED models perform worse than the previous models and are likely not correctly specified. The HS and RM models, which are frequently used by banks, are nearly always rejected as being correctly specified and are the worst performing models in terms of forecasting the relevant risk measures.

2.5 Regulatory Capital Requirements

Assuming that capital is costly and hence banks are interested in minimizing their capital requirements, Basel 3 incentivizes banks to select models that produce low Stressed Expected Shortfall.[18] However, at the same time Basel 3 also penalizes models with too many VaR exceedances through the backtesting multiplier which increases capital requirements and incentivizes banks to use more conservative models. Given this trade-off, it's not clear which model minimizes capital for banks. Moreover, from a regulatory supervisor's perspective, banks using incorrectly specified models that underestimate risk is problematic during a financial crisis, since the banks' risk management models will fail at the same time which increases systemic risk. Considering the backtesting results from Section 2.4, this section determines whether Basel regulation is in fact incentivizing banks to choose models that are correctly specified.

We address the question about correctly incentivizing banks by carefully calculating the capital requirements for a representative bank with the diverse portfolio of assets described in Section 2.3.1. We do this for each model under the Basel 3 regulatory framework to determine which models minimize capital requirements. We then use the same data to calculate requirements under Basel 2 and Basel 2.5 to determine whether the set of models that minimize capital change across the three regimes. Finally, we measure the volatility and the peak-to-trough variation as the maximum difference of capital requirements in Basel 2, 2.5, and 3 to determine whether this regulation is successful at increasing the stability of the capital requirements.

[18] While lower capital requirements may be privately optimal for banks, higher capital requirements may be socially beneficial in reducing the likelihood of systemic crisis. See Birn et al. (2020) for a review on the costs and benefits of bank capital.

We empirically determine the capital requirements of each model for the representative bank portfolio from January, 1997, to February, 2020. We calculate capital requirements for Basel 2, 2.5, and 3 by fixing the regime and closely following the regime's capital calculations over the entire period. Consistent with backtesting, we use a rolling estimation window of 250 days for HS and RM and 2,000 days for the dynamic GARCH and FZ models. In each table, columns are bold if the models are rejected in the traditional ES or spectral backtesting of Section 2.4. Previewing our results, we find that HS and FZH have the lowest capital requirement for Basel 3. In fact, the results show that incorrectly specified models generally have lower capital requirements than correctly specified ones and Basel 3 therefore does not incentivize banks to use correctly specified models. Moreover, we find that while average capital requirements nearly quadruple from Basel 2 to Basel 2.5, they decreased significantly from Basel 2.5 to Basel 3 only for incorrectly specified models. Finally, we show that capital stability generally increases with the regulatory changes. However, the models that have the lowest volatility, the FZ1 and FZH models, are not among the correctly specified models.

2.5.1 Basel 3 Capital Requirements

Table 2.9 shows the average capital requirements of each model calculated under the Basel 3 framework. We also provide a detailed breakdown of the intermediary calculations by risk factor and liquidity horizon to verify that our findings are robust to various portfolio specifications. Panel A of Table 2.9 reports the average Basel 3 capital requirement, backtest multiplier, and IMCC. Row 1 demonstrates that correctly specified models have significantly higher average Basel 3 capital than rejected models. HS has the lowest average Basel 3 capital of 17.21%, despite having the worst backtesting performance in Section 2.4. The light-tailed RM and GARCH Normal models also have low average capital of 19.47% and 21.69%, despite having poor backtesting performance. The FZ1 and FZH models outperform these models during backtesting and also have low average capital of 17.42% and 19.96%. The FHS, SSTD, SGT, GPD, and Hill models generally have the highest average capital requirements of 26.61% to

Table 2.9: Basel 3 Capital Requirements

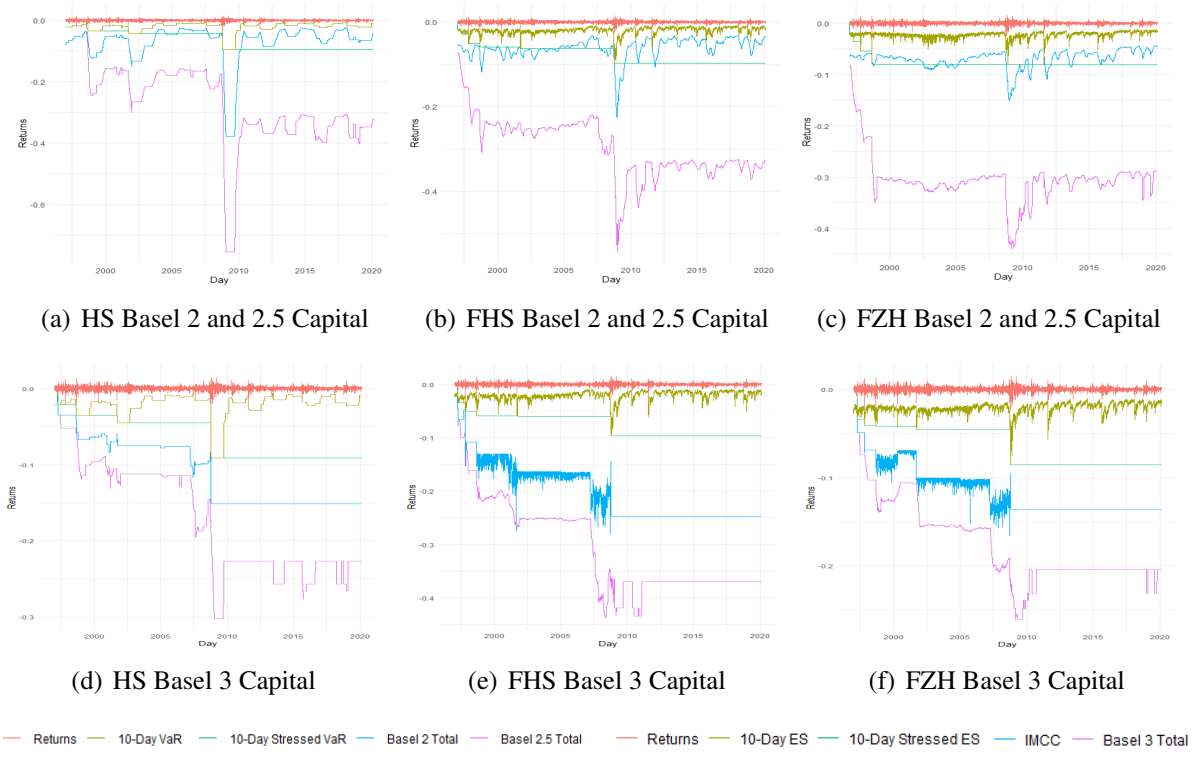
	HS	**RM**	FHS	**Norm**	**STD**	SSTD	**GED**	**SGED**	SGT	GPD	Hill	**HillH**	**FZ1**	**FZH**
Panel A: Basel 3 Average Capital Requirement														
Basel 3 Capital	**17.21**	**19.47**	30.64	**21.69**	**25.77**	26.61	**26.33**	**27.48**	27.38	32.38	29.85	**32.46**	**19.96**	**17.42**
Basel 3 Multiplier	**1.57**	**1.65**	1.53	**1.63**	**1.58**	1.55	**1.58**	**1.55**	1.56	1.52	1.53	**1.53**	**1.53**	**1.53**
IMCC	**10.96**	**11.82**	20.04	**13.27**	**16.29**	17.20	**16.67**	**17.76**	17.66	21.37	19.48	**21.22**	**13.14**	**11.43**
Panel B: Liquidity-Adjusted Stressed Expected Shortfall by Risk Factor														
All	**10.18**	**9.92**	16.94	**11.11**	**13.70**	15.38	**14.45**	**15.78**	14.96	18.98	17.23	**19.27**	**11.45**	**9.59**
Interest	**0.25**	**0.39**	0.47	**0.38**	**0.45**	0.43	**0.42**	**0.42**	0.45	0.47	0.44	**0.43**	**0.45**	**0.50**
Equity	**5.82**	**6.71**	12.27	**7.72**	**9.81**	9.67	**9.46**	**10.22**	10.91	13.57	12.43	**13.06**	**8.15**	**6.68**
Commodities	**1.46**	**1.55**	1.60	**1.28**	**1.41**	1.47	**1.44**	**1.46**	1.44	1.46	1.55	**1.47**	**1.48**	**1.40**
Foreign Exch.	**0.37**	**0.79**	0.67	**0.65**	**0.67**	0.74	**0.70**	**0.72**	0.72	0.64	0.66	**0.67**	**0.47**	**0.51**
Credit	**3.85**	**4.29**	8.13	**5.38**	**6.54**	6.71	**6.86**	**6.92**	6.84	7.62	6.64	**7.54**	**4.28**	**4.17**
Panel C: Stressed Expected Shortfall of Representative Portfolio by Liquidity Horizon														
LH 10	**6.43**	**6.21**	7.61	**5.73**	**6.70**	7.74	**6.83**	**7.20**	7.32	7.78	8.15	**7.32**	**6.71**	**6.35**
LH 20	**4.98**	**4.57**	5.86	**4.31**	**5.10**	5.69	**5.14**	**5.69**	5.59	6.14	5.45	**5.85**	**4.56**	**3.64**
LH 40	**3.03**	**2.82**	6.56	**4.12**	**5.26**	5.94	**5.45**	**5.85**	5.84	7.48	5.87	**7.02**	**3.52**	**2.99**
LH 60	**2.42**	**2.36**	6.43	**3.76**	**4.73**	5.16	**5.43**	**6.17**	5.00	7.60	6.83	**7.15**	**3.29**	**2.46**
LH 120	**0.33**	**0.44**	0.50	**0.40**	**0.42**	0.45	**0.43**	**0.45**	0.47	0.42	0.47	**0.40**	**0.35**	**0.37**

This table shows the Basel 3 capital requirement results of the representative bank portfolio from 01/1997 to 02/2020. Panel A shows the average daily Basel 3 capital requirement in in percentage terms, backtesting multiplier, and IMCC in in percentage terms. Panel B shows in percentage terms the average liquidity-adjusted Stressed ES of the representative portfolio and separated by risk factor. Panel C shows in percentage terms the Stressed ES of the representative portfolio for each individual liquidity horizon. Models rejected by the traditional ES or spectral backtests are bold.

32.38%, despite failing to be rejected in any ES backtest and having the lowest average FZ0 losses in Table 2.7. While FHS weakly dominates HS and FZH across the set of consistent loss functions, using this model requires 1.78 times the average daily capital. SSTD has the lowest average capital among correctly specified models, but using this model still requires 1.55 times the capital relative to HS and FZH.

The multiplier and IMCC in Panel A of Table 2.9 explain why correctly specified models have higher average capital than rejected models. Despite poor backtesting performance, HS has an average multiplier of 1.57. The average multiplier of correctly specified models ranges from 1.52 to 1.56. Since Basel 3 capital is equal to IMCC times the multiplier, we see that differences in average Basel 3 capital across models are almost entirely attributed to differences in IMCC. HS has the lowest IMCC of 10.96%, while FHS has nearly double the IMCC at 20.04% due to its dynamic volatility and heavy tails. The difference between HS and FHS is illustrated in Panels (d) and (e) of Figure 2.2, which show that FHS has a higher average capital and IMCC on every day after the 1998 Asian crisis. SSTD has the lowest IMCC among correctly specified models at 17.20%, which is still substantially higher than HS. The RM and Normal models have the highest multipliers at 1.65 and 1.63, but much lower average

Figure 2.2: Basel Capital Requirements



(a) HS Basel 2 and 2.5 Capital (b) FHS Basel 2 and 2.5 Capital (c) FZH Basel 2 and 2.5 Capital

(d) HS Basel 3 Capital (e) FHS Basel 3 Capital (f) FZH Basel 3 Capital

This figure shows the returns, negative risk measures, and capital requirements of the representative portfolio from 01/1997 to 02/2020 for HS, FHS, and FZH. The top panels plot the daily returns, negative 10-day VaR, Basel 2 capital requirement, 10-day Stressed VaR, and Basel 2.5 capital requirement. The bottom panels plot the daily returns, negative 10-day ES, 10-day Stressed ES, IMCC, and Basel 3 capital requirement.

capital requirements than SSTD due to their low IMCC. The FZ models have both low average multipliers of 1.53 and low IMCC, resulting in low Basel 3 capital. Generally, the least costly model for banks is the one that minimizes IMCC, or equivalently minimizes Stressed ES at the 97.5% confidence level. The HS, RM, Normal, FZ1, and FZH models minimize Stressed ES, resulting in significantly lower Basel 3 average capital requirements. However, these models are either rejected using traditional backtests or provide inferior forecasts of risk measures.

Panel B of Table 2.9 shows the liquidity-adjusted Stressed ES of the representative bank portfolio as well as by individual risk factors. From Equation (2.8), IMCC is calculated by taking a weighted sum of the representative portfolio and risk factor portfolios' liquidity-adjusted Stressed ES. Row 1 of Panel B shows that for the representative portfolio, correctly specified

models have high liquidity-adjusted Stressed ES ranging from 14.96% to 18.98%. The rejected HS, RM, Normal, FZ1, and FZH models have the lowest liquidity-adjusted Stressed ES, the same set of models that minimize IMCC and Basel 3 capital. Rows 2-6 of Panel B show that HS has the lowest ES for the interest, equity, foreign exchange, and credit risk factors, while the Normal model has the lowest ES for commodities. The FZH model has the lowest ES for the representative portfolio of 9.59%, but it has moderate levels of ES for the individual risk factors explaining why its IMCC is higher than HS. The correctly specified models have high ES for every risk factor, verifying that our results are not sensitive to risk factor weights. FHS has more than twice the ES of HS for equity and credit, the two largest risk factors. Across risk factors, the models that minimize liquidity-adjusted Stressed ES are rejected by backtests, while correctly specified models have the highest ES.

Panel C of Table 2.9 shows the Stressed ES of the representative bank portfolio for each liquidity horizon. The liquidity-adjusted Stressed ES of the representative portfolio is calculated as a function of these liquidity portfolios. The panel shows that the HS, RM, Normal, FZ1, and FZH models minimize Stressed ES across liquidity horizons. Since differences in Basel 3 capital are driven by Stressed ES, this explains why these models have the lowest average Basel 3 capital requirements. The ES backtests in Section 2.4 are only for the representative portfolio (LH 10). For this portfolio, correctly specified models have the highest Stressed ES of 7.32% to 8.15%. We show in Section A.5 of Appendix A that for the LH 20-60 portfolios, skewed GARCH and FZ models are likely to be correctly specified. The HS, RM, Normal, STD, GED, SGED models are rejected for the LH 20-60 portfolios, explaining their lower Stressed ES. For the LH 20-60 portfolios, FZ1 and FZH are never rejected and also have a low average Stressed ES. These results also show that every model except STD is rejected for the LH 120 portfolio. However, the HS, FZ1, and FZH models have the lowest Stressed ES for LH 120.

In Basel 3, capital requirements are minimized for models that consistently minimize Stressed ES at the 97.5% confidence level across risk factors and liquidity horizons. HS, RM, Normal, FZ1, and FZH consistently minimize Stressed ES across portfolios, resulting in low

Basel 3 capital requirements. The HS, RM, Normal, FZ1 and FZH models are misspecified since they are rejected in nearly every ES backtest, have the largest FZ0 losses, or provide inferior forecasts of risk measures. The FHS, SSTD, SGT, GPD, and Hill models are never rejected in the traditional backtests and minimize FZ0 losses. However, they have high Stressed ES estimates, resulting in much higher average Basel 3 capital requirements. Thus, a first result is that under Basel 3 there is little incentive for a capital requirement minimizing bank to choose correctly specified models. In particular, given the level of the Basel 3 multiplier, this tool that regulators can use to punish misspecified models has only a marginal effect on average capital requirements and banks have little incentive to choose conservative and correctly specified models.

2.5.2 Evolution of Basel capital requirements

Table 2.10 summarizes the average Basel 2, 2.5, and 3 capital requirements, risk measures, and Basel backtesting results. Panel A of Table 2.10 shows the average 10-day VaR at the 99% confidence level and average capital requirements of the entire sample calculated under the Basel 2 framework. Row 1 shows that the HS average 10-day VaR is 2.17%, which is roughly 2.5 times the 10-day volatility ($0.276\sqrt{10} = 0.87$) of the representative portfolio in Table 2.3. All dynamic models have a lower 10-day VaR compared to HS, since dynamic volatility decreases risk levels faster after a shock. This phenomenon is illustrated in panels (a) and (b) of Figure 2.2, which plots HS and FHS 10-day VaR. After a shock like that following the 2008 Subprime Mortgage crisis, the FHS model quickly decreases VaR levels, while HS maintains high levels of VaR for the entire estimation window. The model with the lowest VaR though is the Normal model. Row 2 shows that the Normal model also has the lowest average Basel 2 capital requirement of 4.98%, despite the Basel backtest results in Panel D showing that this model has the second highest average exceedances and multiplier of 4.48 and 3.26, respectively. In contrast to Basel 3 results, HS has the highest average Basel 2 requirement of 7.12% due to its high 10-day VaR and capital multiplier of 3.15. The correctly specified

Table 2.10: Basel 2, 2.5, and 3 Average Capital Requirements

	HS	**RM**	FHS	**Norm**	**STD**	SSTD	**GED**	**SGED**	SGT	GPD	Hill	**HillH**	**FZ1**	**FZH**
Panel A: Basel 2 Average Capital Requirements														
VaR	**2.17**	**1.80**	2.06	**1.23**	**1.73**	1.86	**1.71**	**1.83**	1.86	2.04	2.14	**1.89**	**1.92**	**2.08**
Basel 2 Capital	**7.12**	**5.98**	6.09	**4.98**	**5.38**	5.61	**5.35**	**5.56**	5.60	5.98	6.12	**5.91**	**5.93**	**6.88**
Panel B: Basel 2.5 Average Capital Requirements														
Stressed VaR	**6.45**	**6.18**	7.76	**6.70**	**6.60**	7.03	**6.49**	**6.85**	7.04	7.35	7.89	**7.48**	**5.78**	**7.75**
Basel 2.5 Capital	**27.34**	**26.27**	29.73	**26.83**	**26.17**	27.26	**25.78**	**26.67**	27.31	28.19	30.15	**28.71**	**23.44**	**30.48**
Panel C: Basel 3 Average Capital Requirements														
Stressed ES	**6.43**	**6.21**	7.61	**5.73**	**6.70**	7.74	**6.83**	**7.20**	7.32	7.78	8.15	**7.32**	**6.71**	**6.35**
Basel 3 Capital	**17.21**	**19.47**	30.64	**21.69**	**25.77**	26.61	**26.33**	**27.48**	27.38	32.38	29.85	**32.46**	**19.96**	**17.42**
Panel D: Basel Backtests on 1-day VaR														
Exceedances	**3.49**	**4.93**	2.48	**4.48**	**3.42**	2.82	**3.42**	**2.82**	2.91	2.22	2.48	**2.48**	**2.26**	**2.04**
Basel 2 Multiplier	**3.15**	**3.30**	3.07	**3.26**	**3.16**	3.10	**3.16**	**3.10**	3.11	3.04	3.07	**3.07**	**3.05**	**3.06**

This table shows the Basel 2, 2.5, and 3 risk measures and capital requirements of the representative bank portfolio from 01/1997 to 02/2020. Panel A shows the average daily VaR estimate and Basel 2 capital requirement in percentage terms. Panel B shows the average daily Stressed VaR estimate and Basel 2.5 capital requirement in percentage terms. Panel C shows the average Stressed ES (LH 10) and Basel 3 capital requirements in percentage terms. Panel D shows the average number of exceedances and the Basel 2.5 multiplier. Models rejected by the traditional ES or spectral backtests are bold.

models have high average Basel 2 capital of 5.60%-6.12%, despite having low multipliers of 3.04-3.11. This shows that Basel 2 failed to incentivize banks to choose correctly specified models.

Next, Panel B of Table 2.10 reports the average 10-day Stressed VaR and Basel 2.5 capital requirement. Stressed VaR is defined as the maximum historical 10-day VaR at the 99% confidence level and the panel shows that the average Stressed VaR is significantly higher than VaR. In contrast to non-stressed VaR, all of the GARCH models have higher Stressed VaR than HS does, since the stressed measure never decreases. Put differently, the VaR-minimizing benefit of dynamic volatility disappears for Stressed VaR. The top panels of Figure 2.2 show that the stressed period for HS, FHS, and FZH is set during the Asian crisis, the Dot-Com crisis, and the 2008 recession resulting in high values of Stressed VaR. FZ1 has the lowest Stressed VaR across all models at 5.78%. Row 2 shows the FZ1 model also has the lowest average Basel 2.5 capital, partially due to its low multiplier. HS has a low Stressed VaR but a high multiplier resulting in moderate Basel 2.5 capital. The correctly specified FHS, SSTD, SGT, GPD, and Hill models have a high Stressed VaR but a low multiplier and as a result reasonable capital requirements. Consequently, the correctly specified SSTD and SGT models require less average capital than HS. These results show that Basel 2.5 has better incentives to use correctly specified models relative to Basel 2.

Finally, Panels A-C of Table 2.10 together allow us to compare capital calculated under Basel 2, 2.5, or 3. We first note that models generally require more than four times the average capital under Basel 2.5 than under Basel 2. As illustrated in the top panels of Figure 2.2, this increase in capital is driven by shocks during the 1998, 2001, and 2008 crises leading to a large Stressed VaR. Since Stressed VaR never decreases, capital levels stay elevated after the 2008 crisis across all models. The comparison of Basel 2.5 to Basel 3 is more nuanced. For HS, there is a significant decrease in required capital from 27.34% under Basel 2.5 to 17.21% under Basel 3, and Panels (a) and (d) of Figure 2.2 illustrate that HS requires more capital under Basel 2.5 on each day. The capital requirements under Basel 3 are also much lower than under Basel 2.5 for the FZ and light-tailed RM and Normal models but remain mostly unchanged from Basel 2.5 to Basel 3 for the skewed GARCH models.

So what drives these differences from Basel 2.5 to Basel 3? First of all, comparing HS Stressed 10-day VaR at the 99% level in Panel B with Stressed 10-day ES at the 97.5% level in Panel C shows that the two stressed risk measures are nearly identical. However, Table 2.9 shows that HS has substantially lower Stressed ES for the 40, 60, and 120 day liquidity-horizon portfolios, also causing low liquidity-adjusted Stressed ES across several risk factors. The capital requirements under Basel 3 are also much lower than under Basel 2.5 for the FZ and light-tailed RM and Normal models due to their relatively low Stressed ES estimates for longer liquidity-horizon portfolios and low liquidity-adjusted Stressed ES. Interestingly, it appears that the requirement under Basel 3 to penalize low liquidity assets, (i.e. assets with long liquidity horizons), additionally in fact disincentivizes banks from using the correctly specified skewed GARCH models due to their consistently high Stressed ES across liquidity horizons. Instead, under Basel 3 banks are incentivized to use misspecified models that consistently minimize Stressed ES across liquidity horizons.

Table 2.11: Basel 2, 2.5, and 3 Capital Variability

	HS	**RM**	FHS	**Norm**	STD	SSTD	**GED**	**SGED**	SGT	GPD	Hill	**HillH**	FZ1	FZH
Panel A: Volatility in Percentage Terms														
Basel 2 Vol	**6.57**	**2.91**	2.38	**2.73**	**2.33**	2.27	**2.32**	**2.25**	2.27	2.16	2.36	**2.31**	**1.85**	**1.64**
Basel 2.5 Vol	**13.43**	**6.15**	6.85	**8.39**	**6.71**	6.14	**5.87**	**5.89**	5.89	5.71	6.65	**6.13**	**4.06**	**4.46**
Basel 3 Vol	**7.17**	**5.72**	8.85	**7.51**	**7.54**	7.63	**7.92**	**7.66**	7.02	7.36	6.12	**7.26**	**5.02**	**4.66**
Panel B: Volatility standardized by Average Capital														
Basel 2 Vol/CA	**0.92**	**0.49**	0.39	**0.55**	**0.43**	0.40	**0.43**	**0.40**	0.40	0.36	0.39	**0.39**	**0.31**	**0.24**
Basel 2.5 Vol/CA	**0.49**	**0.23**	0.23	**0.31**	**0.26**	0.23	**0.23**	**0.22**	0.22	0.20	0.22	**0.21**	**0.17**	**0.15**
Basel 3 Vol/CA	**0.42**	**0.29**	0.29	**0.35**	**0.29**	0.29	**0.30**	**0.28**	0.26	0.23	0.21	**0.22**	**0.25**	**0.27**
Panel C: Maximum Annual Difference in Percentage Terms														
Basel 2 Max Diff	**33.58**	**17.38**	15.63	**18.60**	**15.22**	14.19	**15.16**	**14.43**	14.45	13.57	15.75	**15.13**	**10.20**	**9.18**
Basel 2.5 Max Diff	**58.05**	**24.96**	24.07	**33.11**	**25.45**	20.76	**21.37**	**21.12**	21.14	18.76	23.74	**21.42**	**17.31**	**16.23**
Basel 3 Max Diff	**15.71**	**13.72**	16.94	**18.85**	**20.87**	17.52	**21.07**	**19.74**	22.51	23.91	29.06	**21.61**	**52.79**	**10.26**

This table shows the Basel 2, 2.5, and 3 variability measures of the representative bank portfolio from 01/1997 to 02/2020. Panel A shows the daily volatility of capital in percentage terms. Panel B shows the daily volatility divided by the average capital requirement in percentage terms. Panel C shows the maximum annual difference (250-day difference) of capital requirements in percentage terms. Models rejected by the traditional ES or spectral backtests are bold.

2.5.3 Basel Capital Variability

Table 2.11 reports several measures of capital variability across the Basel 2, 2.5, and 3 regimes. We measure the daily capital volatility to capture day-to-day fluctuations in capital and report volatility standardized by capital to compare across regimes as well. We also measure the maximum of annual differences in capital to capture peak-to-trough variation in capital requirements. Panel A of Table 2.11 shows the capital volatility, which is suitable for comparing across models holding the regime fixed. Row 1 shows that FZH has the lowest Basel 2 volatility of 1.64% followed by FZ1 with 1.85%. The non-normal GARCH models have Basel 2 volatility ranging from 2.16% for GPD to 2.38% for FHS. The RM and Normal models have significantly higher volatility at 2.73% and 2.91%, but HS has the absolute highest Basel 2 volatility with 6.57%. Row 2 shows that FZ1 and FZH have the lowest Basel 2.5 volatility at 4.06% and 4.46%. Compared to this, HS has more than three times the volatility of FZ models and double the volatility of most GARCH models. Panel (a) of Figure 2.2 shows that the high HS volatility is caused by its short estimation window and the Basel multiplier. Panel (c) of Figure 2.2, on the other hand, shows that FZH has stable estimates of 10-day VaR and is generally unaffected by the Basel multiplier. Row 3 shows that HS has a moderate Basel 3 volatility of 7.17% whereas FHS has the highest volatility at 8.85% under this regulation. FZ1 and FZH continues to have the lowest Basel 3 volatility at 5.02% and 4.66%. These numbers are close

Figure 2.3: Basel Models Mean-Volatility Plots

(a) Basel 2 Capital Requirements

(b) Basel 2.5 Capital Requirements

(c) Basel 3 Capital Requirements

This figure shows mean and volatility plots of the representative portfolio from 01/1997 to 02/2020 for Basel 2, 2.5, and 3. In each panel, the vertical axis is the mean capital requirement and the horizontal axis is the volatility of capital requirements. Models in the bottom left minimize mean and volatility of capital requirements, while models at the top right maximize mean and volatility.

to the Basel 2.5 volatility levels for the FZ models whereas Basel 3 volatility is significantly higher than Basel 2.5 volatility for most of the GARCH models. Seen across the regulatory regimes, the table shows that the FZ models have the lowest capital volatility for Basel 2, 2.5, and 3, and banks seeking to minimize capital volatility in Basel 3, in particular, since it may be costly to raise capital during times of high volatility, should consider using the FZH model.

Figure 2.3 visualizes the trade-off between capital minimization and stability of capital requirements by plotting the mean against the volatility of model capital. We color code the rejected models in red and the correctly specified models in blue. For each regime, the bottom

models minimize mean capital and the left models minimize capital variability. Panel (a) shows that HS in the top-right maximizes both mean and volatility of Basel 2 capital. The GARCH models are clustered with the Normal model minimizing mean and the correctly specified models having high means. The FZ models minimize volatility with a moderate mean capital. Panel (b) shows that HS maximizes Basel 2.5 volatility with a moderate mean capital. The GARCH models are clustered with moderate mean and volatility of capital. FZ models minimize Basel 2.5 volatility and FZ1 has the lowest mean capital as evidenced by its position in the bottom left, and banks can minimize both the mean and volatility of Basel 2.5 capital requirements by using the FZ1 model. Panel (c) shows that the FZ models minimize Basel 3 volatility with low means as evidenced by their positions in the bottom left. HS has the lowest mean capital, but has higher volatility than the FZ models. GARCH models have both high mean and volatility as evidenced by their positions in the top right. Basel 3 disincentivizes banks to use these correctly specified models, since they have the highest means and volatilities, despite never being rejected in any ES backtest and having the lowest FZ losses. Banks are instead incentivized to use FZ models, which minimize mean and variability of capital, but perform worse than GARCH models in the comparative backtests. Many banks may remain with HS, since the model minimizes mean capital and is simplest to implement. This may be problematic to regulators, since HS has the worst ES backtest performance across all models and systematically underestimates ES. However, since there is negligible punishment for misspecification and the Basel backtests only VaR, banks have little incentive to move away from HS. Regulators may be able to incentivize banks to move towards correctly specified models if they penalize misspecification in VaR and ES.

Next, we compare stability across Basel regimes using volatility standardized by capital. We use standardized volatility to capture models with either a high volatility or low capital. Panel B of Table 2.11 shows that standardized volatility significantly declines between Basel 2 and Basel 2.5 as a result of including Stressed VaR. This decline is most apparent for HS, where standardized volatility decreases from 0.92 to 0.49. Rows 2 and 3 show that HS capital

volatility further declines to 0.42 from Basel 2.5 to Basel 3, demonstrating that Basel 3 indeed induces more stability when using HS. However, nearly every other model experiences an increase in capital volatility from Basel 2.5 to Basel 3. Panel (d) of Figure 2.2 illustrates the stability of HS under Basel 3 relative to Basel 2.5. Panels (e) and (f) of Figure 2.2 show that for FHS and FZH, the increase in Basel 3 volatility is largely driven by variation in $\frac{ES_{F,C}}{ES_{R,C}}$, the ratio that Basel 3 uses to adjust the reduced portfolio to the current. This point-in-time adjustment ratio causes day-to-day capital changes in dynamic models as evidenced by the erratic IMCC changes between 1999 to 2008, but is set to 1 after the reduced set of portfolios becomes the full set in 2008. The increase in standardized volatility for FZ models also occurs due to their significant decrease in capital requirements from Basel 2.5 to Basel 3. In general, most of the models are more stable in Basel 3 than the previous regimes.

Finally, Panel C of Table 2.11 shows the maximum annual difference in percentage terms. The max difference measures the largest annual peak-to-trough variation in capital requirements, which occurs during the 2008 financial crisis for most models. Peak-to-trough variation in minimum capital requirements is a common measure of regulatory procyclicality (Gordy and Howells (2006)). Rows 1 and 2 show that the max difference increases from Basel 2 to Basel 2.5 due to the introduction of Stressed VaR. The top panels in Figure 2.2 illustrate that the new stressed period in 2008 combined with the multiplier causes a large jump in capital requirements. This is particularly apparent for HS in Panel (a), where the peak-to-trough variation in Basel 2.5 capital requirements is 58%. While Basel 2.5 was implemented after 2008, the results show that staying with Basel 2.5 can cause a large peak-to-trough variation in capital requirements if a future shock moves the stressed period, although this risk is somewhat mitigated by the higher Basel 2.5 capital requirements. Rows 2 and 3 show that the maximum difference substantially decreases from Basel 2.5 to Basel 3 for most models. This decrease is largest for HS, where the peak-to-trough variation in Basel 3 is only 15.71% as shown in Panel (d) of Figure 2.2. FZH has the lowest maximum difference across all models for every Basel regime, making the model suitable for banks seeking to minimize large jumps in capital

requirements.[19]

In summary, our results show that capital requirements became significantly more stable from Basel 2 to Basel 2.5 due to the introduction of Stressed VaR. Basel 3 will further increase the stability of capital requirements by decreasing non-standardized and standardized volatility across most models and could also reduce the procyclicality of capital requirements as evidenced by the lower peak-to-trough variation across most models under this regime compared to previous regimes. Procyclicality is particularly concerning to regulators, since it may amplify severe shocks and increase systemic risk. However, the results also show that the models with the most stable capital requirements across regulatory regimes and across variability metrics are generally not the correctly specified models. In particular, FZ models most often result in the least variability although they provide inferior forecasts of the risk measures. The exception to this is when considering the standardized volatility of the Basel 3 capital which is the lowest for the GPD and Hill models. Thus, while banks can minimize both the mean and volatility of Basel 3 capital by using the FZH model, regulators could attempt to incentivize the use of correctly specified models under this regime by penalizing models with high standardized volatility.

2.6 Conclusion

In 2019, the Basel Committee on Banking Supervision (BCBS) finalized the Basel 3 regulatory regime, which changes the regulatory measure of market risk from Value at Risk (VaR) at the 99% confidence level to Expected Shortfall (ES) at the 97.5% confidence level and adds new complex calculations based on liquidity and risk factors resulting in the most complex market risk framework to date. As a result, banks are eager to know which models minimize these new capital requirements. Additionally, regulatory supervisors are eager to know if the models incentivized by the new regulations are correctly specified. Finally, regulators are also

[19]The FZ1 model had a large shock during the dot-com crisis resulting in high one-day Basel 3 capital requirements.

interested in whether Basel 3 increases the stability and reduces the procyclicality of capital requirements. This paper answers these questions using cutting-edge ES backtests by carefully calculating regulatory capital under the new Basel 3 framework for realistic portfolios banks may hold.

Our backtesting results show that Historical Simulation (HS), RiskMetrics, and symmetric GARCH models are misspecified and systematically underestimate VaR and ES. In particular, HS is rejected by every traditional backtest and has the worst comparative backtest results. A new class of FZ models due to Patton, Ziegel, and Chen (2019) perform much better, but are rejected by conditional spectral backtests and underperform the class of skewed GARCH models traditionally used in empirical finance in comparative backtests. GARCH Filtered Historical Simulation, Skewed Student's t, Skewed Generalized t, Generalized Pareto Distribution, and Extreme Value Theory based Hill models are not rejected in any traditional ES backtest. These skewed GARCH models also have the best performance for comparative backtests, often belonging to the model confidence set and weakly dominating the class FZ models.

Our regulatory capital results demonstrate that Basel 3 incentivizes banks to choose models that minimize Stressed ES at the 97.5% level, since there is nearly no penalty for using misspecified models. As a result, HS is the model that minimizes capital requirements for Basel 3 despite failing every backtest. The Hybrid FZ model requires the second lowest regulatory capital and leads to the most stable capital requirements whereas correctly specified GARCH models require the highest regulatory capital and have the highest capital requirement variability using most metrics. Our results thus show that, although the proposed regulation under Basel 3 seem to incentivize banks to use models that have more stable capital requirements through time, the models that minimize average capital requirements appear misspecified and produce inferior forecasts of the regulatory risk measures.

Our findings have important implications for the current regulation. In particular, we show that Basel 3 regulation strongly disincentivizes banks from using correctly specified models. Although the same holds for previous regulatory regimes, the changes suggested in Basel 3

makes the relative differences even larger. We identify two possible reasons for this. First of all, it appears that the requirement under Basel 3 to penalize low liquidity assets, i.e. assets with long liquidity horizons, additionally in fact disincentivizes banks from using the correctly specified skewed GARCH models due to their consistently high Stressed ES across liquidity horizons. Secondly, given the low level of the Basel 3 multiplier banks have little incentive to choose conservative and correctly specified models. Thus, if regulator's objective is to also incentivize the use of correctly specified models they would have to reconsider the effect of these changes.

Chapter 3

Intraday Market Predictability: A Machine Learning Approach

3.1 Introduction

The previous chapter focuses on the predictability of tail risk. In this chapter, we study the mean predictability of the aggregate market. We focus on intraday market predictability using machine learning models.

The predictability of the aggregate market is a central topic in financial economics. While long-horizon (i.e. monthly or quarterly) market predictability has been extensively studied, intraday (i.e. within a trading day) market predictability has received relatively less attention. Traders require time to incorporate new information about cash flows and discount rates, and over short time horizons equity prices can differ from their adjusted fundamental values, particularly when market frictions are high. This process may introduce short-horizon predictability in equity returns and raises several interesting questions. First of all, if markets are predictable at short horizons what is the magnitude of this predictability? Secondly, is intraday return predictability economically profitable, and if so, does this profitability survive transaction costs? Finally, if markets are predictable it is obviously interesting to know which characteristics, like,

e.g., lagged liquidity or price trends, are in fact important for predicting intraday returns.

Motivated by these questions, we study intraday market predictability using a cross-section of lagged returns of the market and its constituent stocks as predictor variables.[1] We believe our paper is the first to conduct such a study and speculate that the lack of previous studies on this topic may be due to statistical challenges associated with the high-dimensional inputs and the computational difficulties of estimating models using large panels of high-frequency data. We overcome the first issue by using a variety of cutting-edge machine learning models necessary to accommodate the long list of predictors and rich functional forms. We consider candidate methods from Gu, Kelly, and Xiu (2020c) and Hastie, Tibshirani, and Friedman (2009), including linear models with regularization and dimension reduction using lasso (LAS), elastic net (EN), and principal component regression (PCR), and nonlinear tree based models like random forests (RF) and gradient-boosted regression trees (GBRT) along with artificial neural networks (ANN). We also consider the ensemble mean and median of these models. Our baseline model uses all the five-minute returns in an expanding estimation window. Training machine learning models on such a large data panel is computationally challenging and this second issue is overcome using the Apache Sparkling Water and H2O.ai computing framework, which allows us to efficiently estimate machine learning models on large datasets. For example, in October, 2016, the estimation window covers 285 months and contains roughly 450,000 five-minute returns for each S&P constituent, requiring approximately 48 hours to estimate all models.

Our null hypothesis throughout this paper is that markets are not predictable, since any predictability should be removed by active traders. Our alternative hypothesis is that the information in lagged returns is *not* instantly reflected in market prices, and as a result, lagged returns *are* predictive of short-term market returns. These hypotheses are tested by examining the statistical predictability and the economic significance thereof for each of the machine learning models considered. If models can forecast returns and consistently profit in so doing,

[1] While individual stocks are likely more predictable than the market, we begin with market predictability to avoid issues related to data mining and concerns about lack of liquidity.

then markets are likely to be predictable on short-horizons. Our second alternative hypothesis is that there is predictability beyond that explained by transaction costs. This is tested by evaluating the economic significance of model predictions after accounting for realistic transaction costs. If model predictions remain profitable after such costs, then there is likely incorrect pricing beyond that explained by transaction costs.

To examine if there is statistical predictability of intraday market returns over five-minute time intervals, our estimated models are trained on lagged returns with an expanding estimation window from 1993 to 2016. For non-ensemble models, linear as well as nonlinear, out-of-sample R^2 (R^2_{OOS}) values range up to 2.00% for LAS, followed by 1.95% for EN, and 1.71% for the nonlinear RF model. The ensemble mean (median) model yields R^2_{OOS} of 2% (2.01%), illustrating the strength of combined forecasts. However, most of this predictability is concentrated in the pre-decimalization period from 1993 to 2000. In the early post-decimalization period from 2001 to 2004, the LAS, EN, and RF models still have positive R^2_{OOS} values of 0.91%, 0.81%, and 0.85%, respectively. The ensemble mean and median models have respective post-decimalization R^2_{OOS} values of 1.04% and 1.01%. However, during the late decimalization period from 2005 to 2016, we find model predictability significantly decreases due to decreased transaction costs, but remains positive for the LAS, PCR, and ensemble models.

Next, to examine the economic significance of model predictions, a market-timing strategy that buys (sells) the market on positive (negative) predictions is considered. Our results demonstrate that a small intraday R^2_{OOS} value can yield large economic profits, especially given the numerous trading opportunities available at a five-minute interval. Using the baseline model from 1993 to 2016, all models are found to have positive returns. The LAS, EN, RF, and GBRT models have the highest non-ensemble statistical predictability and so too high annualized returns (Sharpe ratios) of 191%, 188%, 198%, and 192% (2.71, 2.68, 2.90, and 2.84). The ensemble mean and median have annualized returns (Sharpe ratios) of 205% and 204% (2.90 and 2.82). Again, most returns are concentrated in the pre-decimalization period. However, in the late post-decimalization period from 2005 to 2016, all models still earn economically

significant profits, with ANN earning the lowest returns (Sharpe ratios) of 16% (0.83). The consistently positive returns and high Sharpe ratios provide further evidence that intraday markets are predictable.

Lastly, economic significance is analyzed after transaction costs. To do this the market-timing strategy is modified to only trade when the signal is strong, i.e., when the model prediction exceeds the transaction cost. Using the baseline model from 1993 to 2016, all models, with the exception of ANN, are found to have positive returns even after accounting for transaction costs. The PCR, RF, and ensemble mean and median models have Sharpe ratios of 0.68, 0.77, 0.67, and 0.98, respectively, vastly exceeding the Sharpe ratio of 0.48 for the benchmark buy-and-hold SPDR S&P 500 (SPY) portfolio. Again, most returns are concentrated in the pre-decimalization period. In the late post-decimalization period from 2005 to 2016, the PCR and RF models have annual returns (Sharpe ratios) of 9% and 8% (0.88 and 0.81) after transaction costs. The ensemble mean and median have respective annual returns (Sharpe ratios) of 4% and 5% (0.35 and 0.68) after transaction costs. As a comparison, from 2005 to 2016 the SPY returned 7% with a Sharpe ratio of 0.46. Thus, in this recent sample PCR, RF, and the ensemble median models continue to earn significantly higher Sharpe ratios than the market does even after transaction costs, providing strong evidence that markets are predictable even after accounting for transaction costs.

We hypothesize that the demonstrated significant predictability of intraday market returns through time is driven by slow-moving trader capital, i.e. by infrequent portfolio rebalancing (Bogousslavsky (2016) and Duffie (2010)), at a high frequency. If some traders rebalance their portfolios infrequently, they may be slow to incorporate shocks to individual stock returns into the aggregate market, particularly when traders face severe volatility or illiquidity. To analyse this further, we examine if our results differ within the trading day, in periods of high versus low volatility or illiquidity, and during periods of financial crisis. First, since traders are most active at the beginning and end of each day, we expect predictability to be low during those times. Our results show that predictability is indeed stronger when traders are less active and

exhibits an inverse-U shape. Second, during periods of high volatility and high illiquidity, traders encounter significant market frictions. Consistent with our hypothesis we find that predictability and its economic significance increases when market volatility and illiquidity are high. Finally, since crisis periods are associated with significant market frictions we expect to find that predictability is also relative high in these periods. Our results demonstrate that predictability is indeed stronger during the Subprime Mortgage crisis and the EU debt crisis.

The baseline predictor variables used here are the cross-section of lagged intraday returns for the market constituents. A natural comparison to these cross-sectional models are autoregressive models for the market return itself, since our predictability results could simply be capturing intraday momentum. For example, Heston, Korajczyk, and Sadka (2010) find significant auto-correlation of half-hour returns at daily intervals and Gao et al. (2018) find that the first half-hour return of the SPY predicts the last. Thus, as an additional analysis we validate our cross-sectional model's results by contrasting them against the results obtained with AR(1), AR(p) where the lag-order p is chosen to minimize the validation error, and AR(500) models estimated using OLS, LAS, EN, and PCR. We confirm that our baseline cross-sectional models significantly outperform the autoregressive models, indicating that the cross-section of lagged constituent returns has significant predictive information that is not contained in lagged market returns.

As a final additional analysis we evaluate intraday market predictability using additional lagged stock characteristics as predictors. In this analysis we include market beta, momentum, illiquidity, extreme returns, trading volume, volatility, skewness, and kurtosis, all estimated over the previous day.[2] We also consider the lagged bid-ask spread. In most cases adding additional variables is found to *decrease* model predictability, indicating that these characteristics do not help lagged returns predict the market portfolio returns. These findings have important implications for the possible economic mechanisms that drive such predictability. For example, could the predictability be caused by intraday momentum, as argued by Heston, Korajczyk, and

[2]Gu, Kelly, and Xiu (2020c) demonstrate that price trend and liquidity have the strongest predictive ability.

Sadka (2010) and Gao et al. (2018)? This explanation seems unlikely given that price trend variables fail to improve model predictability. A big picture implication of our findings is that a careful exploration of the economic mechanisms driving predictability is warranted.

Our paper is related to at least three existing strands of literature. First, our results are naturally related to the recent literature examining intraday return predictability using lagged returns and trading volume. Chordia, Roll, and Subrahmanyam (2005) and Chordia, Roll, and Subrahmanyam (2008) study the predictability of short-run stock returns, finding that intraday returns cannot be predicted by past prices, but that order imbalances do forecast short-horizon returns. Heston, Korajczyk, and Sadka (2010) find significant autocorrelation of returns at daily intervals, for up to 20 days. Gao et al. (2018) demonstrate that the first half-hour return of the SPY market exchange-traded fund (ETF) predicts the last half-hour return. Bogousslavsky (2016) theoretically establishes that seasonality in intraday returns may be caused by traders' infrequent rebalancing. Chinco, Clark-Joseph, and Ye (2019) use a LAS model on the cross-section of NYSE lagged returns to show that one-minute returns are predictable. These studies use *linear* models to forecast returns, whereas our models use information from the entire *cross-section* of lagged returns as well as other characteristics, and permit more flexible functional forms accommodating variable interactions and other *non*linear effects. Furthermore, Ke, Kelly, and Xiu (2019) and Renault (2017) show that text data forecasts intraday returns.

Our paper also relates to the rapidly expanding literature applying machine learning techniques in financial economics.[3] Our paper is most closely related to Gu, Kelly, and Xiu (2020c), which applies an extensive array of machine learning techniques to the problem of predicting equity risk premiums (see also Fischer and Krauss (2018), Long, Lu, and Cui (2019) and Marković et al. (2017)).[4] Bianchi, Büchner, and Tamoni (2021) consider the prediction of

[3]See Weigand (2019) for a recent concise summary of asset pricing via machine learning and Heston and Sinha (2017) and Hajek and Barushka (2018) for reviews of the history of financial applications of machine learning.

[4]Other papers similar to that of Gu, Kelly, and Xiu (2020c), which are also concerned with the use of machine learning for predicting returns, include those of Chong, Han, and Park (2017), Feng, He, and Polson (2018), Kelly, Pruitt, and Su (2019b), Sutherland, Jung, and Lee (2018), and Xue et al. (2018).

bond, rather than stock, risk premiums, and other financial applications of machine learning include their use for derivatives pricing (Ye and Zhang (2019)), hedge fund selection and return prediction (Chen, Wu, and Tindall (2016)), credit risk management (Barboza, Kimura, and Altman (2017)), portfolio management and optimization (Yun et al. (2020) and Day and Lin (2019)), cryptocurrency (Dutta, Kumar, and Basu (2020) and Alessandretti et al. (2018)), stochastic discount factors (Korsaye, Quaini, and Trojani (2019)), and factor models (Bryzgalova, Pelger, and Zhu (2019), Chen, Pelger, and Zhu (2019), Feng, Polson, and Xu (2019), Gu, Kelly, and Xiu (2020a), and Kelly, Pruitt, and Su (2019a)).

Finally, our study is similar in spirit to the literature on aggregate market return predictability, which focuses on long horizons over months, quarters and years.[5] Fama (1991) established the general view that return predictability at long horizons is driven by time-varying expected returns, consistent with market efficiency. Our research differs fundamentally, since the established predictability and its economic value is driven by market frictions at high frequencies.

The paper is organized as follows: Section 3.2 introduces the machine learning methods we consider and explains how we evaluate intraday predictability. Section 3.3 presents results for statistical predictability, economic significance before and after transaction costs, and a robustness check using different training window sizes. Section 3.4 examines if the results differ across the time of day, across levels of volatility and illiquidity, or during crisis periods, whether autoregressive models can also forecast returns, and whether predictability can be improved by including additional input variables. Section 3.5 concludes. The appendices contain further details on the data used, variable cleaning, the stock characteristics used as inputs, and on the implementation of the machine learning methods, in general, and hyperparameter optimization, in particular.

[5] See, among others, Ang and Bekaert (2007), Campbell and Thompson (2008), Chen, Da, and Zhao (2013), Kelly and Pruitt (2015), Neely et al. (2014), and Welch and Goyal (2008). See Koijen and Van Nieuwerburgh (2011) and Rapach and Zhou (2013) for recent surveys.

3.2 Methodology

Given T high-frequency price observations, indexed by $1 \leq t \leq T$, and denoting by p_t the natural logarithm of the t^{th} observed price, the corresponding logarithmic return is given by

$$r_t \equiv p_t - p_{t-1}. \tag{3.1}$$

Of particular interest in this paper is the market return which we denote by r_t^M and whether or not this can be predicted. To assess this, we conduct one of the largest empirical studies of market prediction in the high-frequency literature. The primary dataset is the trade and quotes (TAQ) database, containing intraday transaction data for all stocks on the New York stock exchange (NYSE), American stock exchange (AMEX), NASDAQ, and other American regional exchanges, from February, 1993, to October, 2016. Our goal is to predict five-minute changes in the aggregate market, which we proxy by the SPDR S&P 500 (SPY) ETF. The baseline predictor variables are lagged returns of the SPY and S&P 500 constituents. We obtain the list of S&P 500 constituents from the center for research in security prices (CRSP), and update the 500 constituents monthly.[6]

Given a vector of lagged characteristics of the SPY and S&P 500 constituents, $\mathbf{X}_t^\ell$, indexed by integers, $0 \leq \ell \leq 500$, the objective of our paper is to use state of the art machine learning methods to approximate the empirical model given by

$$r_{t+1}^M = f\left(\mathbf{X}_t^\ell\right). \tag{3.2}$$

In the baseline model, the covariates are lagged returns, such that $\mathbf{X}_t^\ell = r_t^\ell$, resulting in 501 covariates. We consider in addition models that include other firm-level characteristics, such that $\mathbf{X}_t^\ell = [\ r_t^\ell \quad z_t^\ell\ ]$, where [.] denotes matrix concatenation, and z_t^0 (z_t^ℓ with $\ell \geq 1$) is the market beta (of firm $\ell \geq 1$), illiquidity, kurtosis, maximum, minimum, momentum, skewness,

[6] See Appendix B.1.1 for more details on the databases used.

volatility, trading volume, or bid-ask spread. These expanded models have 1002 covariates (501 lagged returns and 501 characteristics).[7]

The rest of this section introduces several methods for estimating the function f in the empirical model in Equation (3.2) above. In this paper we consider specifically 1) linear models, 2) nonlinear models, and 3) so-called ensemble models that weigh together multiple individual models. Finally, we explain how, given an estimator of f, possible predictability can be assessed both statistically and economically.

3.2.1 Linear models

When restricting the focus to linear models the optimization problem is given by

$$\inf_{\beta \in \mathbb{R}^K} m[\mathbf{y} - \mathbf{X}\beta], \tag{3.3}$$

where we use notation from the machine learning literature and refer to the market returns as the targets denoted by $\mathbf{y}$, an $n \times 1$ column vector, and where the predictors, e.g. lagged returns or other characteristics, are denoted by $\mathbf{X}$, an $n \times K$ matrix in the case of K predictors. In Equation (3.3), $m[\cdot]$ denotes a metric or loss for the fit of the model and β denotes the relevant parameters in the model. For example, in the case of OLS regression the metric is taken to be the Euclidean/ℓ^2 norm, i.e. $m \equiv \|\cdot\|_2$, and the solution to the optimization problem in Equation (3.3) is given by the classical OLS estimator

$$\hat{\beta} \equiv (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

In the presence of many predictors, the simple linear regression model easily becomes inefficient leading to in-sample over-fitting which is detrimental to our objective of out-of-sample prediction. We present here two approaches to deal with this: 1) regularization and 2) principle

[7] The only exception to this is the expanded model with z_t^ℓ equal to market beta which only has 1001 covariates, since the SPY market beta $z_t^0 = 1$. See Appendix B.1.2 for details on how these characteristics are calculated.

component regression.

Regularization

One primary objective of *regularization* techniques (see, e.g., Friedman, Hastie, and Tibshirani (2010)) is to avoid over-fitting in statistical models. This is often accomplished by adding a penalty term to the optimization problem in Equation (3.3) as follows

$$\inf_{\beta \in \mathbb{R}^K} \{m[\mathbf{y} - \mathbf{X}\beta] + \lambda n[\beta]\}. \tag{3.4}$$

Here, the functional $n[\cdot]$, often a norm, penalizes non-zero estimators and the *regularization parameter* λ regulates the penalty's impact as a multiplicative scale. A classical example is ridge regression, in which the optimization problem in Equation (3.4) is modified to

$$\inf_{\beta \in \mathbb{R}^K} \left\{ ||\mathbf{y} - \mathbf{X}\beta||_2^2 + \lambda ||\beta||_2^2 \right\}.$$

The smoothness of ridge regression (see, e.g., Marquaridt (1970)) resulting from using the ℓ^2 norm is computationally advantageous, but may result in many 'near-but-non-zero' coefficients and thus may not reduce the dimensionality of the optimization problem in a sufficient manner. In this paper we instead consider lasso regression (see, e.g., Tibshirani (1996)) and elastic nets (see, e.g., Zou and Hastie (2005)).

In the case of a lasso regression (LAS) the optimization problem in Equation (3.4) is modified to

$$\inf_{\beta \in \mathbb{R}^K} \left\{ \frac{1}{2} ||\mathbf{y} - \mathbf{X}\beta||_2^2 + \lambda ||\beta||_1 \right\}.$$

Hence LAS employs the computationally difficult (i.e., non-smooth) ℓ^1 norm, but has the resulting advantage that many coefficients are driven to zero exactly, leaving out only those of sufficient predictive importance. The resulting β, is non-zero only for those predictors which

most significantly determine the target and may therefore be of much lower dimension than the original problem which is indeed one of the primary objectives of our use of regularization techniques.

In the case of an elastic net regression (EN) the optimization problem in Equation (3.4) is modified to

$$\inf_{\beta\in\mathbb{R}^K}\left\{\frac{1}{2}\|\mathbf{y}-\mathbf{X}\beta\|_2^2+\alpha\lambda_1\|\beta\|_1+\frac{1-\alpha}{2}\lambda_2\|\beta\|_2^2\right\}.$$

Hence EN convexly combines ridge and lasso penalties to balance these two competing properties; smoothness and perfect elimination of unimportant predictors. Here, $\alpha \in [0,1]$ is the coefficient of the convex combination of ℓ^1 and ℓ^2 norms of the regression coefficients, β, and in general, each may have its own regularization parameter, respectively, λ_1 and λ_2. In the particular cases of $\alpha \in \{0,1\}$, elastic nets respectively reduce to ridge and lasso regressions.[8]

Principal component regression

Principal component regression (PCR) is a dimension-reduction technique used to summarize variation within a data set using a small number of linear combinations thereof (see, e.g., Jolliffe (2002)). Given a data set $\mathbf{X}$, consisting of n observations of K predictors, PCR solves the following problem

$$\sup_{\mathbf{w}\in\mathbb{R}^K}\frac{\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}}{\mathbf{w}^T\mathbf{w}},$$

where $\mathbf{w} \in \mathbb{R}^K$ are the predictor weight vectors and Σ is the covariance of the predictors. The motivation for PCR is clear from this formulation: since an eigenvector of $\boldsymbol{\Sigma}$, $\mathbf{w}$, solves the eigenvalue problem, $\boldsymbol{\Sigma}\mathbf{w} = \lambda\mathbf{w}$, for a corresponding eigenvalue of $\boldsymbol{\Sigma}$, λ, the best *eigenvector* solution of this problem is obviously that for which the corresponding eigenvalue, λ, is largest. This follows since the Raleigh quotient to be optimized, $\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}/\mathbf{w}^T\mathbf{w}$, which measures nor-

[8] Appendix B.2 explains how the hyper-parameters α, λ_1, and λ_2, or in the case of LAS only λ, are selected.

malized variation/variance of the data set along the weight vector, $\mathbf{w}$, simplifies in this case to λ. So in applying PCR, it is necessary only to compute an eigenvalue decomposition of $\mathbf{\Sigma}$, sort its eigenvalues, and take as many corresponding eigenvectors as are desired principal components, to yield the principal directions. Normalizing each by its corresponding square-root eigenvalue, the desired principal components are obtained.

In the context of the optimization problem in Equation (3.3), PCR is first applied to the predictors, $\mathbf{X}$. Supposing $\kappa \ll K$ components are desired, the projected predictors, $\mathbf{Z}$, are then used in a linear regression yielding the classical OLS estimator given by

$$\hat{\beta} \equiv (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{y},$$

in which

$$\mathbf{Z} \equiv (\mathbf{X} - \mu)\mathbf{W},$$

where μ is the mean of the predictors and $\mathbf{W} \equiv [\ \mathbf{w}_{(1)} \quad \mathbf{w}_{(2)} \quad \cdots \quad \mathbf{w}_{(\kappa)}\ ]$ is the matrix of principal components to be used. The resulting (demeaned) projected predictors, $\mathbf{Z}$, are thus only of dimension $\kappa \ll K$, yielding a potentially significant reduction in the number of predictors and resulting regression model complexity, while preserving, to the extent possible, the richness of variation in the original data.[9]

3.2.2 Nonlinear models

More generally, the fundamental problem to be addressed in this paper is to solve the following optimization problem

$$\inf_{\mathbf{f}\in\mathcal{F}} m[\mathbf{y} - \mathbf{f}(\mathbf{X})], \tag{3.5}$$

[9]Selecting the value of κ can be a complex problem. Appendix B.2 explains how we set this hyper-parameter.

where $\mathcal{F}$ denotes the set of functions from which candidate predictors are drawn and m denotes a metric or loss for the fit of the model. Above we considered several linear models and we now consider two general examples of nonlinear models: 1) models based on decision trees and 2) models based on artificial neural networks.

Tree-based models

Decision trees are nonparametric, hierarchical sequences of decisions, which optimally construct, based on the training and validation data, sequences of decisions to classify or regress arbitrary input predictors. Individual decision trees train in logarithmic time with the number of training points, but over-fitting is common, as more decisions (greater 'depth') are needed to better model training data, which may not generalize well. Individual trees often behave chaotically, too, in that the optimal structure may change drastically in response to the addition or removal of a handful of training data. Ensembles of trees mitigate the resulting large variance of individual decision trees, and avoid over-fitting by restriction to simple individuals across the ensemble, but inherit from such individuals some degree of the advantages of interpretation and efficient training. The fundamentals and many refinements of decision tree training are provided by, e.g., Breiman et al. (1984).

Random forest (RF) models (see, e.g., Breiman (2001)) independently and pseudorandomly generate decision trees, which are separately trained and whose predictions are then averaged to yield the ensemble prediction. In RF prediction robustness improvements are achieved via variance reductions implicit in the law of large numbers. Given predictors $\mathbf{X}$, denote an individual decision tree's estimator, say that of the i^{th} tree generated in an ensemble, by $\mathbf{g}_i(\mathbf{X})$. Supposing there are N decision trees in the ensemble, the prediction of RF is simply

$$\mathbf{f}_{\text{RF}}(\mathbf{X}) \equiv \frac{1}{N} \sum_{i=1}^{N} \mathbf{g}_i(\mathbf{X}).$$

That is, RF amounts to considering the (arithmetic) average of the predictions of individual

trees in the forest.[10]

Each individual decision tree in the ensemble is trained on data drawn pseudorandomly with replacement from the training data. I.e., each individual tree is trained on a bootstrapped sample of the training data, and aggregated in making predictions via the average of those from each individual, and so random forests constitute one instance of bootstrap aggregated (*bagged*) predictors (see, e.g., Breiman (1996)). This *bagging* yields many individuals with uncorrelated prediction errors and less variance on average, whereas individuals tend to have high variance and over-fit. The construction may however introduce estimator bias, which is the problem addressed below by gradient-boosted regression trees. The bagging implicit in random forests also addresses the NP-completeness of the problem of training a *globally* optimal individual decision tree, which necessitates the use of heuristics, including typical greedy implementations, in training such individuals and bagging mitigates much of the bias that such heuristics introduce during individuals' training.

As their name suggests, gradient-boosted regression trees (GBRT) implement *boosting*, a second case of ensemble methods. As opposed to averaging methods, the simple estimators are sequentially (and so not independently) generated in a manner which progressively eliminates bias from the ensemble (a standard reference is Schapire and Freund (2013)). Boosting refers to the notion of developing a strong learner (a predictor with metric, m, approaching zero on arbitrary data) from a weak one (a predictor which performs marginally better than random guessing). Schapire (1990) first affirmatively answered this *hypothesis boosting* question, and the adaptive resampling and combining, or *arcing*, algorithm of Freund and Schapire (1997) is regarded as the canonical method for achieving such boosting in machine learning. *Gradient* boosting refers to the generalization of this and other boosting algorithms to the case of arbitrary differentiable loss functions, first achieved explicitly by Friedman (2001) and Friedman (2002).

[10]The crucial hyper-parameters to optimize for RF are the number of trees in the forest, the maximum number of predictors or *features* considered at each node of each tree, and the maximum number of nodes between the leaves and root of any tree, known as the *depth*. Appendix B.2 explains how these are selected.

In the case of GBRT, the ensemble prediction is a weighted combination of individual predictions, with weights γ_i determined as part of training leading to the following representation

$$\mathbf{f}_{\text{GBRT}}(\mathbf{X}) \equiv \sum_{i=1}^{N} \gamma_i \mathbf{g}_i(\mathbf{X}).$$

Specifying $\mathbf{f}_0 \equiv 0$ and denoting $\mathbf{f}_{\text{GBRT}} \equiv \mathbf{f}_N$, the following holds for $1 \leq i \leq N$

$$\mathbf{f}_i(\mathbf{X}) = \mathbf{f}_{i-1}(\mathbf{X}) + \gamma_i \mathbf{g}_i(\mathbf{X}).$$

At each stage, i, of training, the decision tree predictor, $\mathbf{g}_i$, is greedily chosen to solve

$$\inf_{\mathbf{g} \in \mathcal{G}} \mathcal{L}[\mathbf{y}, \mathbf{f}_{i-1}(\mathbf{X}) + \gamma_i \mathbf{g}_i(\mathbf{X})]. \tag{3.6}$$

Here, $\mathcal{L}$ is some loss function, typically expressed as a sum of losses, say L, between corresponding targets, $y_j \in \mathbf{y} \equiv \{y_j\}$, and predictors, $\mathbf{x}_{j\cdot} \in \mathbf{X} \equiv \{\mathbf{x}_{j\cdot}\}$, as

$$\mathcal{L}[\mathbf{y}, \mathbf{f}_{i-1}(\mathbf{X}) + \gamma_i \mathbf{g}_i(\mathbf{X})] \equiv \sum_{j=1}^{n} L[y_j, \mathbf{f}_{i-1}(\mathbf{x}_{j\cdot}) + \gamma_i \mathbf{g}_i(\mathbf{x}_{j\cdot})]. \tag{3.7}$$

Note that the solution to the optimization problem in Equation (3.6) is conditional on the current ensemble, $\mathbf{f}_{i-1}$, as opposed to the independent generation of individual decision tree predictors, $\mathbf{g}_i$, in a RF.

In the case where the loss, L, is differentiable, gradient boosting solves the optimization problem in Equation (3.6) via gradient descent as

$$\mathbf{f}_i(\mathbf{X}) = \mathbf{f}_{i-1}(\mathbf{X}) - \gamma_i \sum_{j=1}^{n} \nabla_{\mathbf{f}} L[y_j, \mathbf{f}_{i-1}(\mathbf{x}_{j\cdot})],$$

where the the weights, γ_i, are chosen to optimize a loss similar to that of Equation (3.7) given

by

$$\inf_{\gamma>0} \sum_{j=1}^{n} L\left(y_j, \mathbf{f}_{i-1}(\mathbf{x}_{j\cdot}) - \gamma \partial_{\mathbf{f}} L[y_j, \mathbf{f}_{i-1}(\mathbf{x}_{j\cdot})]\right).$$

Thus, the weights can be interpreted as step sizes/learning rates in this gradient descent procedure.[11]

Artificial neural networks

As their name suggests, artificial neural networks (ANNs) are motivated by the neural networks found in animal brains, and theoretical neuroscientific models thereof (see Bishop (1995) and Bishop (2006)). The most general-purpose and well-known neural network architecture is feedforward in which an *input layer* consisting of the inputs, $\mathbf{y}^0 \equiv \mathbf{x}_{i\cdot}$, is followed by a sequence of *hidden layers*, each consisting of a number of neurons. Initially, the inputs are weighted by a set of learned parameters and added to another, the so-called *bias*, to yield an input for each neuron in the succeeding layer. Denoting by k_1 the number of neurons in this layer and $k_0 \equiv K$ the number of inputs, there are thus $k_1(k_0 + 1)$ parameters which yield this layer's inputs as

$$(\forall 1 \leq j \leq k_1)\ x_j^1 \equiv \mathbf{y}^0 \mathbf{w}_j^1 + b_j^1.$$

Here, $\mathbf{w}_j^1$ is the learned column vector of weights for the row vector of inputs, $\mathbf{y}^0$, and b_j^1 the corresponding learned additive bias. To introduce non-linearity into the output of a neuron $y_j^1 \equiv \phi_j^1(x_j^1)$ where ϕ_j^1 is the so-called activation function of neuron j.

More concisely we can write the mapping as

$$\begin{aligned} \mathbf{x}^1 &\equiv \mathbf{y}^0 \mathbf{W}^1 + \mathbf{b}^1 \\ \mathbf{y}^1 &= \mathbf{\Phi}^1(\mathbf{x}^1). \end{aligned} \tag{3.8}$$

[11] In addition to the hyper-parameters discussed for RF, in the case of GBRT the loss function, L, used for training and the additional uniform multiplier/scale, ν, which may be factored out of the step sizes, γ_i, as an explicit learning rate, can also be important. See Appendix B.2 for further details.

Here, $\mathbf{W}^1$ is the matrix with columns $\mathbf{w}_j^1$, $\mathbf{b}^1$ is the row vector with entries b_j^1, and $\mathbf{\Phi}^1$ is the mapping from the row vector $\mathbf{x}^1$, with entries x_j^1, to the row vector $\mathbf{y}^1$, with entries, y_j^1. In succeeding hidden layers, if applicable, the process is iterated as follows

$$\begin{aligned}(\forall 2 \leq \nu \leq N)\, \mathbf{x}^\nu &\equiv \mathbf{y}^{\nu-1}\mathbf{W}^\nu + \mathbf{b}^\nu \\ \mathbf{y}^\nu &= \mathbf{\Phi}^\nu(\mathbf{x}^\nu),\end{aligned} \tag{3.9}$$

where N is the total number of hidden layers. Finally, the *output layer* yields predictions given by

$$\begin{aligned}\mathbf{x}^{N+1} &\equiv \mathbf{y}^N\mathbf{W}^{N+1} + \mathbf{b}^{N+1} \\ \mathbf{y}^{N+1} &= \mathbf{\Phi}^{N+1}(\mathbf{x}^{N+1}).\end{aligned} \tag{3.10}$$

In our application, the output layer is linear and the number of neurons, k_{N+1}, is naturally set to one, resulting in

$$\mathbf{f}_{\text{ANN}}(\mathbf{X}) \equiv \mathbf{y}^N\mathbf{W}^{N+1} + \mathbf{b}^{N+1}. \tag{3.11}$$

Hence, the neurons are simply linearly aggregated into the forecast.

Collecting all weights and biases across layers, the feedforward network has a total parameter count of $\sum_{\nu=0}^{N} k_{\nu+1}(k_\nu + 1)$ as there are $\sum_{\nu=0}^{N} k_{\nu+1}$ biases and $\sum_{\nu=0}^{N} k_{\nu+1}k_\nu$ weights and such networks may indeed be highly parametric. Classically, these are optimized via stochastic gradient descent, but several adaptive/'momentum'-based generalizations have been proposed.[12]

[12]In ANN, the optimization algorithm employed for training, the loss function it uses, the ℓ^1 and ℓ^2 regularization parameters, the number of epochs and batches, and the level of dropout all constitute *hyperparameters* which impact the quality of predictions from a model with trained *parameters*. However, the actual network architecture, characterized by the number of hidden layers, the number of neurons in each layer, and the activation function associated with each neuron, often have an even greater impact. We fix the number of hidden layers at 3 and use hyperparameter optimization to determine the number of nodes in each layer. Appendix B.2 provides further details on these hyperparameters.

3.2.3 Ensemble methods

Ensemble methods seek to combine multiple predictors' results, for both variance reduction and 'crowd wisdom' purposes. For example, various measures of central tendency or location parameters, including the median and alternate (weighted) means, such as the harmonic, geometric and arithmetic mean, may act as ensemble aggregation methods and represent an 'average' prediction based on all the predictors' outputs.[13] For regression problems, the two most prominent machine learning ensemble aggregation methods are boosting and bagging, as respectively outlined for GBRT and RF models. Abstractly, given any of these ensemble aggregation methods, say, AM $\in$ {Bag, boost, (weighted) arithmetic/geometric/harmonic mean, median}, etc., and a set of predictors' outputs, e.g., $\{\mathbf{f}_{\text{LAS}}(\mathbf{X}), \mathbf{f}_{\text{EN}}(\mathbf{X}), \mathbf{f}_{\text{PCR}}(\mathbf{X}), \mathbf{f}_{\text{RF}}(\mathbf{X}), \mathbf{f}_{\text{GBRT}}(\mathbf{X}), \mathbf{f}_{\text{ANN}}(\mathbf{X})\}$, the ensemble prediction is

$$\mathbf{f}_{\text{AM}}(\mathbf{X}) \equiv \text{AM}\left(\{\mathbf{f}_{\text{LAS}}(\mathbf{X}), \mathbf{f}_{\text{EN}}(\mathbf{X}), \mathbf{f}_{\text{PCR}}(\mathbf{X}), \mathbf{f}_{\text{RF}}(\mathbf{X}), \mathbf{f}_{\text{GBRT}}(\mathbf{X}), \mathbf{f}_{\text{ANN}}(\mathbf{X})\}\right).$$

In our application we will consider the average and the median as aggregation methods and as such this method does not introduce any additional hyperparameters.

It should be noted that regularization may also be applied to the ensemble aggregation methods just as it may be in the particular cases of GBRT and RF models.[14] The machine learning ensemble aggregation methods are more flexible than straightforward computation of some measure of central tendency or location parameter of the various models' predictions, as the former permit regularization and hyperparameter tuning. Which ensemble aggregation method to use is itself a hyperparameter optimization problem, albeit a small one which may be largely mitigated by simply computing many ensembles, e.g., boosting, bagging, and a variety of 'popular' measures of central tendency or location parameters. The justification for any particular aggregation method may come from either the statistical (variance-reduction and

[13] Timmermann (2006) provides an overview of forecast combinations and Genre et al. (2013) show that forecast combinations using a simple average often outperform methods that rely on estimated combination weights.

[14] In fact, Koren (2009) aggregated model predictions using GBRT, in the winning solution of the Netflix Prize.

bias) properties of bagging or boosting, or laws of large numbers, central limit theorems, or other asymptotic results applicable to the computed measures of central tendency or location parameters.

3.2.4 Evaluation criteria

Our objective is to assess the predictability of market returns. We do so using two types of criteria: purely statistical criteria based on a relevant metric for out of sample model fit and economic criteria based on the obtained returns from trading on a given model's predictions.

Statistical significance

To evaluate the predictive performance of the high-frequency market return forecasts, we calculate the out-of-sample R^2 metric proposed by Gu, Kelly, and Xiu (2020c). Given the market return, r^M_{t+1}, and a corresponding model prediction given the history up to time t, $\hat{r}^M_{t+1}$, the out-of-sample R^2 is calculated as

$$R^2_{\text{OOS}} \equiv 1 - \frac{\sum_{t\in\text{Test}}\left(r^M_{t+1} - \hat{r}^M_{t+1}\right)^2}{\sum_{t\in\text{Test}}\left(r^M_{t+1}\right)^2}. \tag{3.12}$$

Note that the R^2_{OOS} metric is only calculated over the test samples, indexed by times t in the set Test. The denominator of R^2_{OOS} is the squared sum of market returns without demeaning. As discussed by Gu, Kelly, and Xiu (2020c), the historical mean underperforms a zero forecast. The historical mean return is noisy, resulting in artificially high estimates of R^2. Hence, we benchmark against zero rather than the historical mean.[15]

Following Gao et al. (2018) and Chinco, Clark-Joseph, and Ye (2019), we also consider the estimated coefficients from running a simple predictive regression given by

$$r^M_{t+1} = \alpha + \beta\hat{r}^M_{t+1} + \epsilon_{t+1}, \tag{3.13}$$

[15]When we benchmark against the historical mean, the R^2 increases by approximately 0.01% across methods.

where r^M_{t+1} is the market return and $\hat{r}^M_{t+1}$ is the model prediction for five-minute time interval $t+1$.

Economic significance

In addition to evaluating predictive performance, we assess the ability of each model to time the market. We implement a trading strategy that takes a long (short) position in the market if the model predicts a positive (negative) return. The profitability can be expressed as $\pi \equiv \sum_{t\in\text{Test}} \pi_t$ where the individual daily profits, π_t, are defined as

$$\pi_t\left(\hat{r}^M_{t+1}\right) \equiv \begin{cases} -r^M_{t+1} & \text{if } \hat{r}^M_{t+1} < 0, \\ r^M_{t+1} & \text{if } \hat{r}^M_{t+1} > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{3.14}$$

Reported are annualized excess arithmetic returns and Sharpe ratios of this trading strategy across the test sample. The Sharpe ratio is calculated as the monthly excess return divided by the corresponding standard deviation and scaled by $\sqrt{12}$.

The economic significance of the trading strategy could be completely driven by small return fluctuations such as the bid-ask bounce discussed by Roll (1984), which is not useful to traders. To assess the models' ability to predict larger returns, we consider a trading strategy that accounts for transaction costs. Given national best bid (ask) price, Bid_t (Ask_t), and midquote $\text{M}_t = \frac{\text{Bid}_t+\text{Ask}_t}{2}$, we estimate the transaction cost by the relative national-best bid-offer (NBBO) spread, given by

$$\text{Spread}_t = \frac{\text{Ask}_t - \text{Bid}_t}{\text{M}_t}. \tag{3.15}$$

Following Chinco, Clark-Joseph, and Ye (2019), we evaluate the economic significance of returns after transaction costs by implementing a trading strategy that only trades when model

predictions exceed the transaction costs. The returns of this trading strategy are

$$\phi_t\left(\hat{r}^M_{t+1}\right) \equiv \begin{cases} -r^M_{t+1} - \text{Spread}_t & \text{if } \left|\hat{r}^M_{t+1}\right| > \text{Spread}_t \text{ and } \hat{r}^M_{t+1} < 0, \\ r^M_{t+1} - \text{Spread}_t & \text{if } \left|\hat{r}^M_{t+1}\right| > \text{Spread}_t \text{ and } \hat{r}^M_{t+1} > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{3.16}$$

Again, we report annualized excess returns and Sharpe ratios for this trading strategy. For models to be profitable, their predicted returns must be properly directed and exceed transaction costs, that is $\hat{r}^M_{t+1} > 0$ and $|\hat{r}^M_{t+1}| > \text{Spread}_t$ for a long position to be implemented. This simple trading strategy presents a significant hurdle for model validation and provides a benchmark for returns available to traders. We use a simple strategy to avoid data mining concerns, but more sophisticated strategies can yield higher returns.

3.3 Empirical results

In this section we present the results for the eight models considered: lasso (LAS), elastic net (EN), principal component regression (PCR), random forest (RF), gradient-boosted regression trees (GBRT), artificial neural networks (ANN), and the ensemble (arithmetic) average (MEAN) and median (MED). We omit results for OLS from our analysis, since we find that the simple linear model is highly inaccurate, resulting in a negative out-of-sample R^2 in most periods. All models are estimated using Sparkling Water from H2O.ai, which combines the machine learning algorithms of H2O with the big data capabilities of Apache Spark, and permits efficient estimation of machine learning models on much larger datasets than the existing literature. We train the models using all intraday and overnight observations, however we report results from 9:35 to 3:55 only excluding the opening and closing returns to avoid concerns regarding the accuracy of these auction-based prices and the effect of overnight/weekend effects. For each model, we tune hyperparameters in the validation set using random search on mean squared error.

We evaluate the predictive performance of each model by out-of-sample R^2, R^2_{OOS}, and the predictive regression coefficients and t-statistics. We evaluate the economic significance by the excess returns and Sharpe ratios of the market timing trading strategy with and without transaction costs. In each case, we examine model performance over the entire testing sample from April, 1993, to October, 2016. The results are also presented for significant sub-periods, including the 1/8 tick size sample from 1993 to 1996, the 1/16 tick size sample from 1997 to 2000, the early post-decimalization sample from 2001 to 2004, and the late post-decimalization sample from 2005 to 2016. Finally, we consider the robustness of our results to using models trained on 1, 4, 7, 10, 16, 22, 34, and 58 months of returns instead of our baseline expanding window.

3.3.1 Market predictability

Table 3.1 reports coefficients from the predictive regression in Equation (3.13) along with Newey-West t-statistics with 79 lags, R^2_{OOS} percentages, and the ratio of R^2_{OOS} to mean transaction cost (R^2_{OOS}/(Bid-Ask Spread scaled by 1,000)) for the eight machine learning techniques considered. Panel A presents results for the entire testing sample from April 1993 to October 2016. Across models, slope coefficents are roughly 1 with most t-statistics exceeding 40. The magnitude of the intercept and slope coefficients are not exactly 0 and 1 respectively, since the relationship between realized and predicted returns is non-linear (for example, due to the prevalence of zero return observations in five-minute returns). The table shows that all the models predict the market better than a naive forecast of zero, since each of them have a positive R^2_{OOS}. The LAS and EN models have R^2_{OOS} of 2% and 1.95%, respectively. These regularized linear models have the highest R^2_{OOS} among non-ensemble models. Among non-linear models, RF performs best with an R^2_{OOS} of 1.71%. Surprisingly, ANN performs poorly despite performing well in Gu, Kelly, and Xiu (2020c). This performance difference is likely because they forecast using an ensemble of neural networks, while we are limited by computational time to a single neural network prediction each period, which typically has higher variance forecasts. Across

all models, the ensemble mean and median perform best with respective R^2_{OOS} values of 2% and 2.01%. Note that most models have an R^2_{OOS} exceeding 1.6%, which is the R^2_{OOS} documented in Gao et al. (2018) for 30-minute returns and roughly twice the monthly R^2_{OOS} reported by Gu, Kelly, and Xiu (2020c). Having a similar level of R^2_{OOS} at a higher-frequency (like five-minutes in this paper) than at a lower frequency is interesting since more trades can be carried out based on the predictability. Our findings are different than Chordia, Roll, and Subrahmanyam (2005), which uses linear models and find that lagged returns cannot forecast five-minute equity returns. Our results reveal that by including lagged returns of market constituents using high-dimensional models, market returns are predictable using only lagged returns.

Next, we break down the predictability by time period based on observed structural breaks caused by changes to tick size and transaction costs. Panel B presents results for the early sample from 1993 to 1996, coinciding with a 1/8 tick size. This sample has the strongest predictability, with R^2_{OOS} exceeding 6% across all models and slope t-statistics exceeding 30. This finding makes sense intuitively, since predictability should be high when a large tick size prevents traders from bringing prices to their fundamental values. Also, trading volumes were relatively low and transaction costs were high during this time period, which further increased market frictions. Panel C presents results for the period with 1/16 tick size from 1997 to 2000. This period has the second highest predictability with R^2_{OOS} exceeding 3% for all models except PCR. The R^2_{OOS}/Cost ratio decreased by roughly 90% relative to the previous period, demonstrating that most of the decreases in predictability were due to factors other than the decreases in transaction costs, and possibly due to technological improvements for traders (i.e. faster computer driven trading).[16]

Beginning in 2001, exchanges adopted decimal ($0.01) trading ticks, which is the tick size in the remainder of our sample. For this reason, we refer to 2001 to 2016 as the post-decimalization period. Panel D reports results for the early post-decimalization period from 2001 to 2004. Coefficients are statistically significant at the 1% level across all models, and

[16]The simple intuition for this is the following: if the decrease in R^2_{OOS} could be fully explained by a decrease in transaction costs, then the R^2_{OOS}/Cost ratio should remain the same across periods.

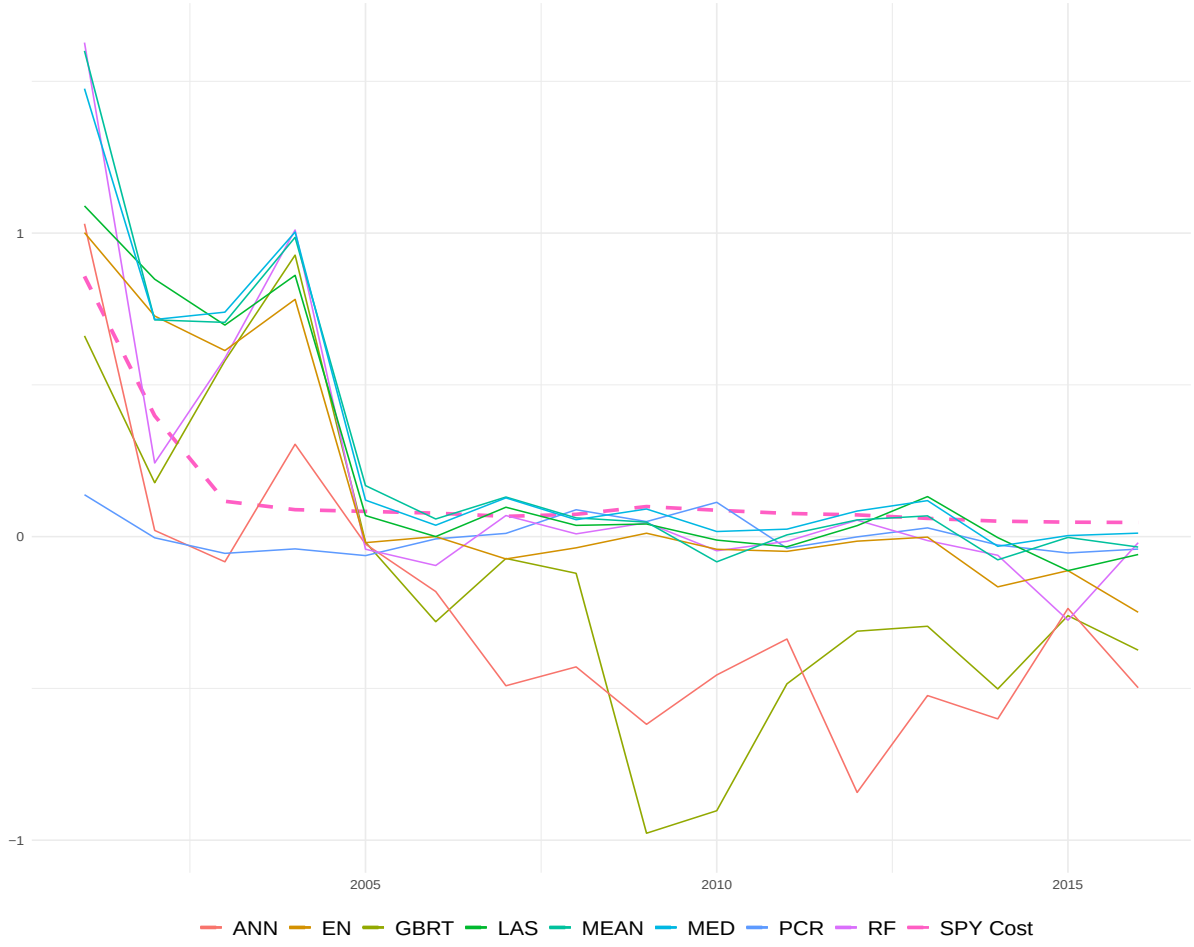
while the R^2_{OOS} are lower than pre-decimalization, they remain relatively high. Furthermore, the R^2_{OOS}/Cost ratio decreased by roughly 50% relative to the previous period, indicating that much of the decrease in predictability can be explained by decreases in transaction costs. The highest non-ensemble R^2_{OOS} are for the LAS, EN, and RF models at 0.91%, 0.81%, and 0.85%, respectively. The ensemble mean and median reach an R^2_{OOS} of 1.04% and 1.01%, respectively. These R^2_{OOS} are still in line with the market predictability results from Gao et al. (2018) and Gu, Kelly, and Xiu (2020c) for the 30-minute and monthly time horizons, respectively. It should be noted that from 2001 to 2004 there was a significant decrease in transaction costs with no change to tick size. Figure 3.1 plots the median transaction cost (Bid-Ask Spread scaled by 1,000) against the R^2_{OOS} for each model during the post-decimalization period. Consistent with our hypothesis, the decrease in transaction costs in the early post-decimalization period significantly decreased the R^2_{OOS} across all models.

Finally, Panel E reports results for the late post-decimalization period from 2005 to 2016. The LAS and PCR models have positive R^2_{OOS} of 0.02% and 0.04%, respectively, during this period. Interestingly, the PCR model has a positive R^2_{OOS} during this period despite having the lowest R^2_{OOS} in all previous periods. This is likely because the Subprime crisis and European debt crisis occurred in this period, and as a result of asset correlation increasing to near one in these crisis periods, returns had a strong factor structure.[17] The EN, RF, GBRT, and ANN models have negative R^2_{OOS} of -0.04, -0.01, -0.39, and -0.45, respectively, indicating that they perform worse than a naive constant prediction of 0. The poor performance of the non-linear models suggests that non-linear models may be over-fitting a simpler predictive relationship during this recent period. The ensemble mean and median, however, have the highest R^2_{OOS} of 0.04% and 0.06% respectively, benefiting from PCR's strong performance during crisis periods. This result shows that models that perform poorly on average can still improve ensemble models.

At a first glance, it may appear that after 2005 all of the predictability is gone. However,

[17] In Section 3.4.3 we confirm that the PCR model indeed performs well during both these crisis periods.

Figure 3.1: R^2_{OOS} and trading costs



Plot of the post-decimalization R^2_{OOS} for the SPY using the lasso (LAS), elastic net (EN), principal component regression (PCR), random forest (RF), gradient-boosted regression trees (GBRT), neural network (ANN), mean ensemble (MEAN), and median ensemble (MED). Also plotted is the median transaction cost scaled by 1,000 (SPY Cost).

we argue that this is not the case. In particular, Figure 3.1 shows that the R^2_{OOS} for the two ensemble methods stay consistently positive, although it is small in magnitude, and the ensemble median R^2_{OOS} is positive in every year except for 2015. Relative to the previous period, the R^2_{OOS} decreased by roughly 95%, but the R^2_{OOS}/Cost ratio only decreased by 33-66%, indicating that some of the decrease in predictability can be explained by decreases in transaction costs. Given the substantial decrease in transaction costs for five-minute returns, a high R^2_{OOS} would be unreasonable, and a small but positive R^2_{OOS} should be expected. As argued by Campbell and

Thompson (2008) and Rapach and Zhou (2013), even a small R^2 can generate economically large portfolio returns. This can be especially true given the number of trading opportunities at a five-minute interval.

In summary, our first set of results demonstrate that the market is remarkably predictable at the five-minute frequency. This predictability is highest prior to the decimalization of exchanges in 2001. Post-decimalization, markets became faster in integrating lagged intraday information. However, the R^2_{OOS} remains positive for the LAS, PCR, and ensemble models, demonstrating that some predictability persists even after the decimalization of exchanges and indicating that decreases in transaction costs also decreased market frictions.

3.3.2 Economic significance

As we mentioned above, even a small predictability at five-minute intervals can result in large and economically significant returns. In this section we therefore evaluate the economic significance of our models' predictions by implementing the simple trading strategy specified in Equation (3.14). Since the models are optimized to minimize forecasting error, the economic significance of forecasts provides an indirect evaluation of model performance. Table 3.2 presents the annualized excess returns, Newey-West t-statistics with 1 lag and Sharpe ratios, denoted SR in the table, of the market timing strategy. Columns (1) - (8) presents results for the eight machine learning models and columns (9) and (10) report the intraday SPY returns as well as the benchmark buy-and-hold SPY returns, respectively, for comparison.

Panel A of Table 3.2 reports the results for the entire sample from 1993 to 2016. Rankings across methods are mostly consistent with their R^2_{OOS} percentages, and all models have positive returns and Sharpe ratios. Among the non-ensemble models, LAS, EN, RF, and GBRT have the highest R^2_{OOS} and also have high returns (Sharpe ratios) of, respectively, 191%, 188%, 198%, and 192% (2.71, 2.68, 2.90, and 2.84), indicating that dimension reduction is important for predicting returns. In particular, the tree-based RF and GBRT have the highest non-ensemble returns (Sharpe ratios), showing that modeling non-linearities and interaction effects are impor-

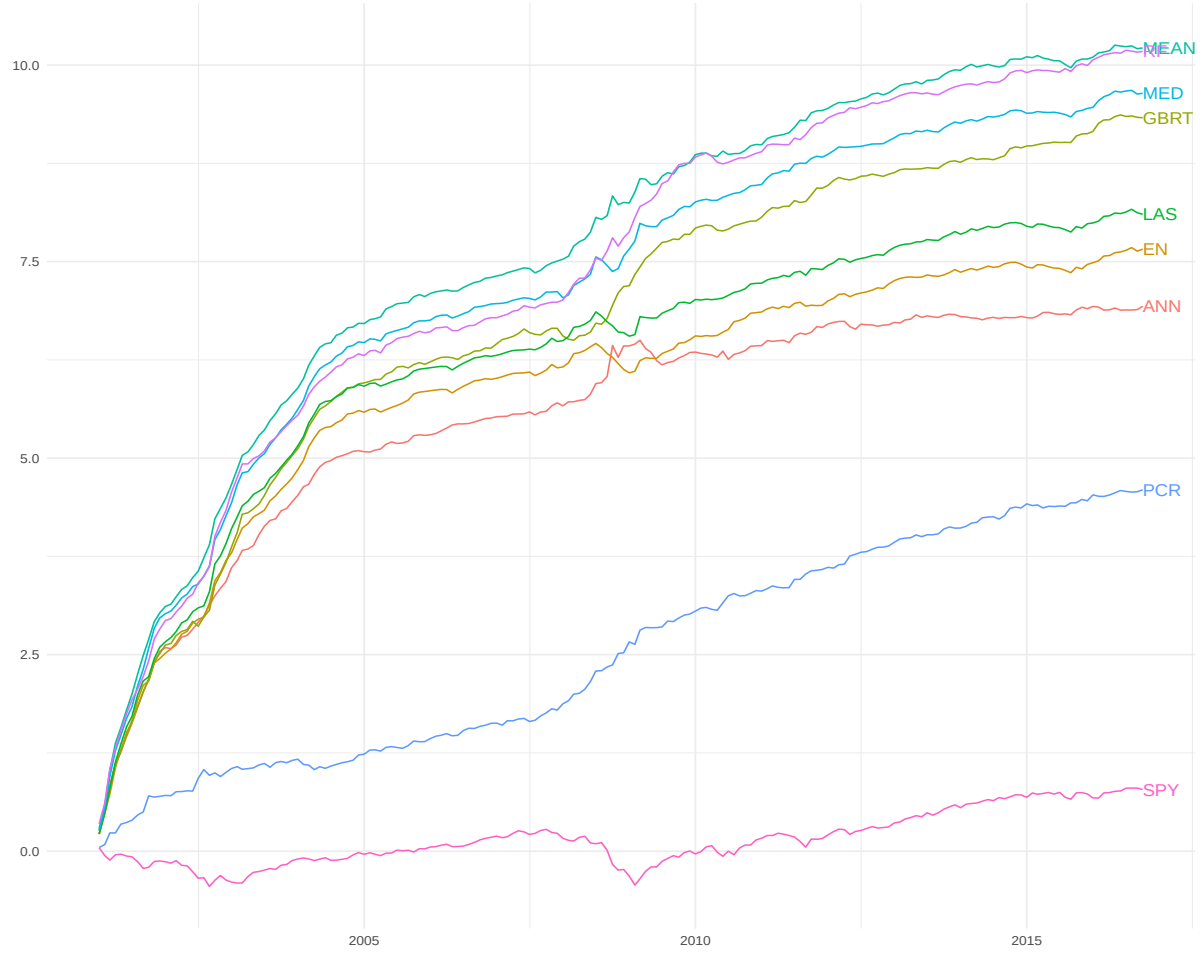
tant for return prediction. The ANN model has lower returns (Sharpe ratios) of 172% (2.66) due to its high variance predictions. PCR has positive but relatively low excess returns of 75% matching its low R^2_{OOS} over the entire sample. Consistent with their R^2_{OOS} percentages, the ensemble mean and median have the highest returns (Sharpe ratios) of, respectively, 205% and 204% (2.90 and 2.82), illustrating that combining forecasts yields higher economic predictability. It is noteworthy that every model significantly out-performs the benchmark buy-and-hold SPY strategy, which yields excess returns (Sharpe ratios) of 7% (0.48).

Next, Panel B of Table 3.2 reports results for the 1/8 tick period from 1993 to 1996. The excess returns (Sharpe ratios) of all models exceed 300% (5). The large economic returns are consistent with the large R^2_{OOS} observed for this period. Panel C reports results for the 1/16 tick period from 1997 to 2000. While Table 3.1 showed that the R^2_{OOS} decreased relative to the previous period, surprisingly returns and Sharpe ratios increased during this period for nearly all models. This volatile period contains both the Asian crisis and the dot-com bust, and in Section 3.4.3 we show that the economic significance of model predictions increase during financial crises. PCR has a highly negative R^2_{OOS} during this period resulting in a low predictive return, suggesting that there is not a strong factor structure during this period.

Finally, we consider the results in the post-decimalization period. Panel D reports results for the early post-decimalization period from 2001 to 2004 and shows that, consistent with the finding that the R^2_{OOS} decreased post-decimalization, excess returns and Sharpe ratios, although significant and larger than the benchmark buy-and-hold SPY returns, are lower than in previous periods. Panel E reports results for the late post-decimalization period from 2005 to 2016. Table 3.1 showed that this period had a substantial decrease in R^2_{OOS}, and consequently returns and Sharpe ratios are significantly lower than in previous periods. However, returns and Sharpe ratios remain high relative to the benchmark buy-and-hold SPY. In particular, all models have returns exceeding 16% and most models have Sharpe ratios exceeding 1. In comparison, the buy-and-hold SPY returns (Sharpe ratios) were 7% (0.46).

Figure 3.2 plots the cumulative log returns for each model and of the SPY from 2001 to

Figure 3.2: Cumulative returns by model



Plot of the post-decimalization cumulative log returns of the market-timing strategy for the SPY using the lasso (LAS), elastic net (EN), principal component regression (PCR), random forest (RF), gradient-boosted regression trees (GBRT), neural network (ANN), mean ensemble (MEAN), and median ensemble (MED).

2016. Every model has higher cumulative returns than the market portfolio with the RF and ensemble models having significantly higher cumulative returns. In summary, the consistency of the results across models supports the hypothesis that the intraday market is predictable and demonstrates that the predictability is economically significant. Next we show that profitability remains even with (large) transaction costs.

3.3.3 Economic significance after transaction costs

The economic significance established above does not account for transaction costs, which are substantial when trading at a five-minute frequency. If models are only accurate for small return fluctuations but cannot forecast large returns, then they are not useful to traders. This section shows that the predictability of the market portfolios is economically significant even after accounting for transaction costs. Additionally, we documented above that after 2005, R^2_{OOS}, excess returns, and transaction costs significantly decreased at the same time, so it is not obvious if model predictions remain profitable after accounting for transaction costs in this recent sample.

Table 3.3 presents the annualized excess returns and Sharpe ratios of the market timing strategy with transaction costs along with the average percentage of executed trades. Panel A of Table 3.3 reports the results for the entire sample from 1993 to 2016. Columns (1) - (8) presents results for the eight machine learning models and columns (9) and (10) report the intraday SPY and the benchmark buy-and-hold SPY returns respectively. As expected, all returns are significantly lower due to transaction costs and infrequent trading. However, even after accounting for transaction costs, every model (with the exception of ANN) has positive returns. Among non-ensemble models, PCR and RF have the highest returns (Sharpe ratios), respectively yielding 5% and 6% (0.68 and 0.77). The two ensemble models have high returns (Sharpe ratios), both yielding 6% (0.67 and 0.98). As a benchmark, a buy-and-hold SPY strategy has 7% return and a Sharpe ratio of 0.48. Thus, even after transaction costs, the PCR, RF, and ensemble models outperform holding the market.[18] These findings demonstrate that such models can predict large returns that exceed the transaction costs very well. Our results are similar in magnitude to the strategies in Gao et al. (2018) and Chinco, Clark-Joseph, and Ye (2019) that have after-transaction cost annualized returns of 4.46% (Sharpe ratio of 0.98) and

[18]When regressing the returns after transaction costs of our trading strategy onto the benchmark buy-and-hold SPY we find that the alphas are positive and statistically significant for the PCR, RF, and ensemble models. The betas, on the other hand, are mostly small, and if they are large and significant, they tend to be negative indicating that our trading strategy if anything hedges systematic risk. Thus, there is strong evidence that our trading strategy out-performs the benchmark portfolio and little evidence that our models are simply buying systematic risk.

4.92%, respectively. Consistent with the previous sections, ANN predictions have a high variance and hence fail to forecast large returns. Similarly, the linear LAS and EN models had high R^2_{OOS}, but the statistical predictability did not translate well into economic profitability, earning returns (Sharpe ratios) after transaction costs of 3% and 1% (0.42 and 0.20), respectively. We show in Section 3.4.3 that LAS and EN had highly negative returns during the Subprime Mortgage crisis, since these models generally perform worse when predictors are highly correlated (Wang et al. (2011)).[19]

Next, Panel B of Table 3.3 reports results for the 1/8 tick period from 1993 to 1996. The models returns are low since they only trade roughly 1% of the time. However, the large and positive Sharpe ratios of at least 0.8 indicate model predictions are economically significant even after paying transaction costs. The LAS, EN, GBRT, and ensemble models have Sharpe ratios exceeding 2 and exceeding the benchmark buy-and-hold SPY Sharpe ratio of 1.6. Panel C of Table 3.3 reports results for the 1/16 tick period from 1997 to 2000, where models trade less than 0.25% of the time due to large transaction costs. Whereas Table 3.2 showed that pre-transaction cost returns are higher than in the previous period, Table 3.3 shows that after accounting for transaction costs returns (Sharpe ratios) are now lower than in the previous period. LAS, EN, PCR, RF, ANN, and the ensemble models, though, continue to have higher Sharpe ratios than the buy-and-hold benchmark Sharpe ratio of 0.71.

Finally, Panel D and E reports results for the post-decimalization period from 2001 to 2016, where models trade much more frequently due to lower transaction costs. Panel D reports results for the early post-decimalization period from 2001 to 2004 and shows that excess returns (Sharpe ratios) increase for the LAS, EN, RF, GBRT, and ensemble models relative to the period before 2001 due to the significantly lower transaction costs. All models, except ANN,

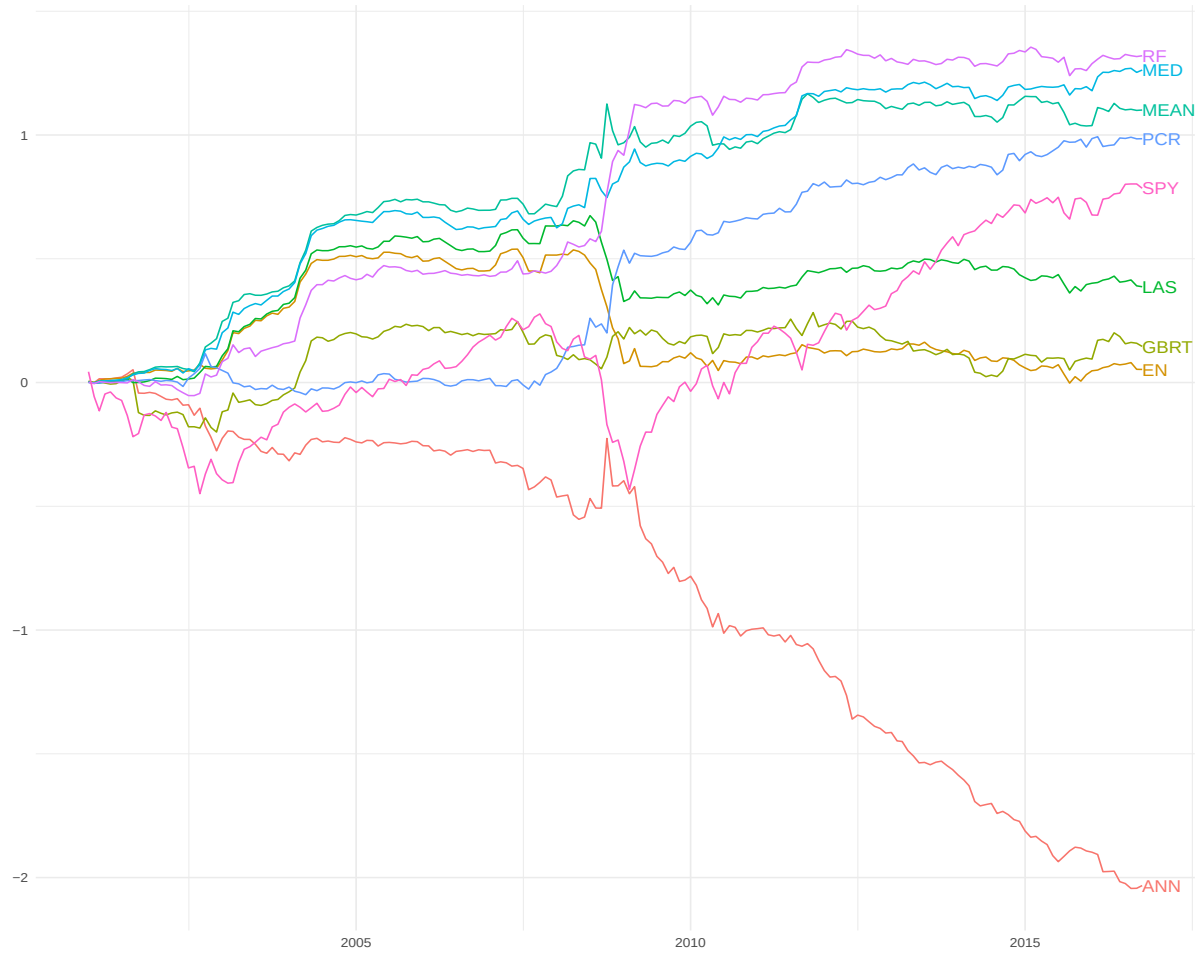
[19] In unreported results we confirm that when correlations among constituent returns increase LAS and EN perform worse while PCR performs better. Intuitively, LAS and EN may encounter difficulties when stocks are highly correlated. For example, if several highly correlated stocks are relevant for prediction, then LAS may only select one from the group and shrinks the rest to zero (Zou and Hastie (2005)). EN mitigates this issue by using the ridge penalty. However, the ridge penalty forces the estimated coefficients of highly correlated predictors to be close together, which is problematic since the coefficients on our predictor stocks likely have different magnitudes or different signs (Wang et al. (2011)). Conversely, PCR has stronger predictability when correlations increase, since the model creates new predictors that summarize the variation of the constituent stocks.

earn much higher returns (Sharpe ratios) than the benchmark -0.01% (-0.07) for the buy-and-hold SPY. Panel E of Table 3.3 reports results for the late post-decimalization period from 2005 to 2016. In the previous sections, we documented that this period had substantially lower R^2_{OOS}, returns, and Sharpe ratios relative to previous periods. However, due to the decrease in transaction costs, the after-transaction cost returns and Sharpe ratios are still high. During this period, the benchmark buy-and-hold SPY earned 7% returns with a 0.46 Sharpe ratio. The PCR and RF models out-perform the benchmark with returns (Sharpe ratios) of 9% and 8% (0.88 and 0.81), respectively, which we in Section 3.4.3 show is partially due to their strong performance during the Subprime mortgage crisis. The ensemble median also beat the buy-and-hold SPY with returns (Sharpe ratios) of 5% (0.68).[20]

Figure 3.3 plots the cumulative log returns after transaction costs for each model and of the SPY from 2001 to 2016. After accounting for transaction costs, the PCR, RF and ensemble models have higher cumulative returns relative to the market, despite only trading infrequently. Even accounting for the drop in R^2_{OOS} after 2005, the after-transaction cost returns remain consistently large. That is, even after accounting for transaction costs, the considered models earn economically significant returns with low variance. We demonstrate economic gains available to traders using such model forecasts supporting the hypothesis that markets are predictable at the five-minute frequency. However, the market is notably less predictable post-decimalization, particularly after 2005, as expected, evidenced by the lower returns and Sharpe ratios. Furthermore, the LAS and EN models outperformed PCR and RF prior to the 2008 crisis, but performed poorly during and after the crisis due to the higher level of correlation among stocks.

[20]These results are robust to assuming fixed transaction costs of, say, 0.01% or 0.1% instead of the bid-ask spread (the median post-decimalization spread was 0.008%). In particular, with 0.01% fixed transaction costs economic profitability generally increases across models and though the models nearly never trade with a 0.1% fixed transaction cost Sharpe ratios remain positive for the LAS, EN, and ensemble models. This verifies that our results are not driven by models that only trade when the spread is low.

Figure 3.3: Cumulative returns after transaction costs by model



Plot of the post-decimalization cumulative log returns of the market-timing strategy with transaction costs for the SPY using the lasso (LAS), elastic net (EN), principal component regression (PCR), random forest (RF), gradient-boosted regression trees (GBRT), neural network (ANN), mean ensemble (MEAN), and median ensemble (MED).

3.3.4 Robustness to training sample size

The baseline model uses an expanding window for training and one month each for validation and for out-of-sample testing. If the predictive relationship that we document is stable, then forecasting accuracy should be increasing in training sample size, since more observations yield lower variance forecasts. However, financial time series are notorious for containing structural breaks, time-varying volatility, and other nonstationarities (Timmermann (2008)). A shorter training sample may therefore outperform a longer one if the empirical model in

Equation (3.2) changes due to this nonstationarity, in which case using a longer training sample may yield biased forecasts. We test the importance of various training periods by evaluating R^2_{OOS} and after-transaction cost Sharpe ratios using 58-, 34-, 22-, 16-, 10-, 7-, 4-, and 1-month samples for training and compare the results with the baseline training duration.

Panel A of Table 3.4 reports the post-decimalization R^2_{OOS} of the eight machine learning models using different training window lengths. This sample includes several possible structural breaks, including the Subprime mortgage crisis and EU debt crisis. Across nearly all models, R^2_{OOS} is increasing in training size, indicating that the predictive relationship is mostly stable. However, the R^2_{OOS} do not increase monotonically, which demonstrates that non-stationarities do have some effect on forecasts. This is particularly apparent comparing the 58-month estimation window to expanding, suggesting that results could be improved by starting the expanding window later in the sample. Furthermore, the R^2_{OOS} are positive for the ensemble models across nearly all training periods and positive for most individual models. Panel B of Table 3.4 reports the after-transaction cost Sharpe ratios using different training window lengths for the post-decimalization period. Consistent with the results for the R^2_{OOS}, model Sharpe ratios are mostly increasing in training size across models.

The results in Table 3.4 first of all show that our predictability findings for intraday market returns are largely robust to using different training window specifications. Secondly, they importantly show that predictability increases with the training window size. Chinco, Clark-Joseph, and Ye (2019) theorize that market predictability could be driven by very short-term sparse signals. Our results indicate instead that predictability may be consistently exploiting inefficiencies across time and is not necessarily driven by infrequent signals.

3.4 Additional analysis

Our hypothesis for intraday predictability is based on slow-moving trader capital. In this section we examine if our results differ within the trading day, in periods of high versus low

volatility or illiquidity, and during periods of financial crisis. Finally, we compare our baseline model performance to that of autoregressive models for the market return and we consider whether forecasting accuracy may be improved by including additional lagged variables.

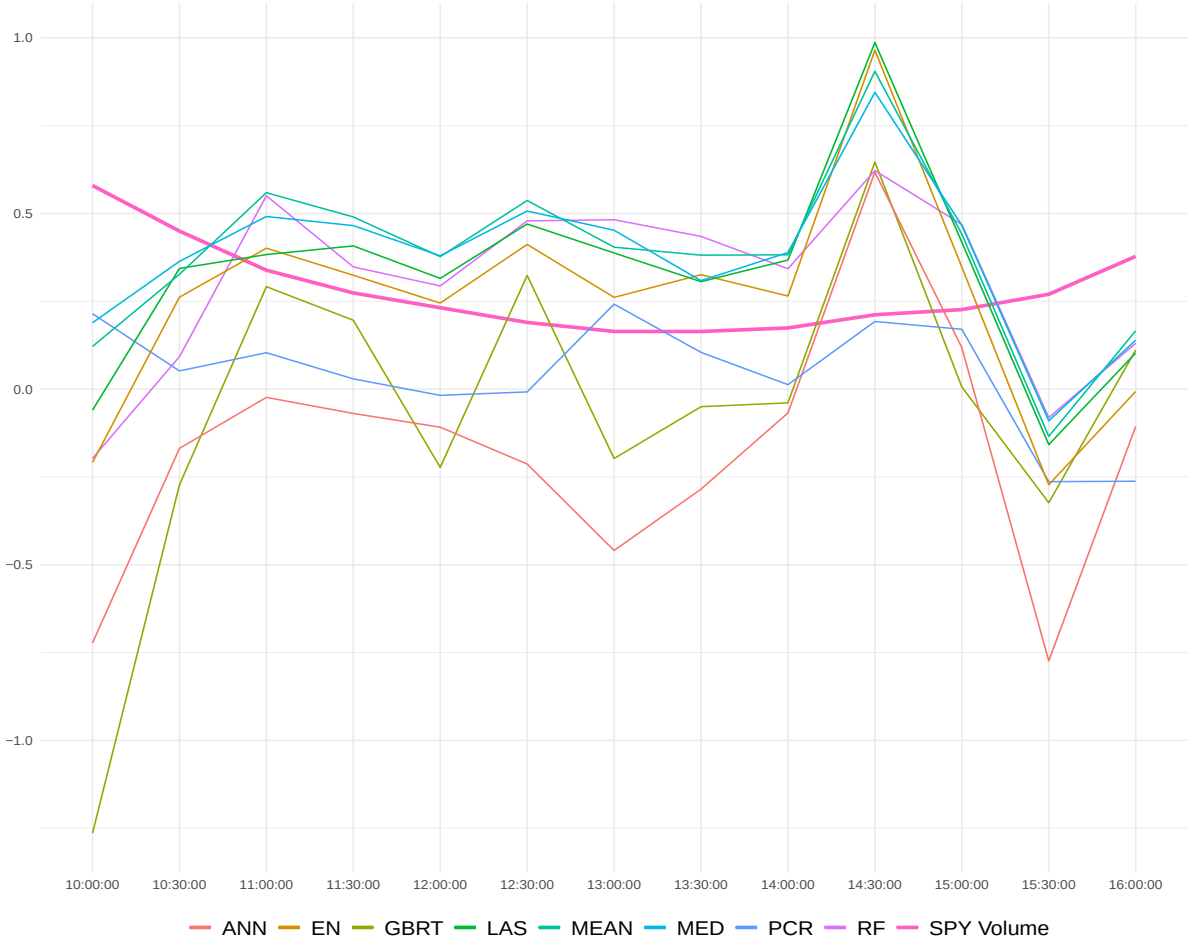
3.4.1 Time-of-day patterns

This section tests the intraday implications of slow traders on intraday predictability. Since traders are most active at the beginning and end of each day, we expect predictability to be low during those times. Also, since intraday trading volume exhibits a "U" shape, we expect predictability to exhibit an inverse-U shape.

Panel A of Table 3.5 reports the R^2_{OOS} of each half-hour interval in the post-decimalization period. For every model, the R^2_{OOS} is low in the first half hour and last hour of the day, when traders are most active. On the other hand, these models have high R^2_{OOS} between 10:00 - 15:00, demonstrating that nearly all of the predictability occurs during the middle of the day when traders are less active. Panel B of Table 3.5 reports after-transaction cost Sharpe ratios of each half-hour interval in the post-decimalization period. For most models, the Sharpe ratios are highest between 10:00 - 15:00, demonstrating that the increased R^2_{OOS} translates into economically meaningful returns after transaction costs. The Sharpe ratio for the intraday SPY in column (9) does not display the same pattern, showing that our models are not simply buying the SPY portfolio.

Figure 3.4 plots the R^2_{OOS} and median trading volume in each half-hour interval in the post-decimalization period. Across all models (except ANN), there is a remarkably similar inverse-U pattern, suggesting that these models are approximating the same function. Every model's R^2_{OOS} jumps up between 14:00 - 14:30, suggesting that this is the least active time for traders. Every model's R^2_{OOS} jumps down between 15:00 - 15:30, suggesting that traders are active at this time, which is consistent with Lou, Polk, and Skouras (2019)'s finding that institutional investors tend to initiate trades near the close. Our discovery that predictability is stronger when traders are less active is consistent with our hypothesis that the predictability is

Figure 3.4: R^2_{OOS} and trading volume



Plot by time of day of the post-decimalization R^2_{OOS} for the SPY using the lasso (LAS), elastic net (EN), principal component regression (PCR), random forest (RF), gradient-boosted regression trees (GBRT), neural network (ANN), mean ensemble (MEAN), and median ensemble (MED). Also plotted is the median SPY trading volume in millions.

driven by slow traders.

Gao et al. (2018) find that returns in the first half hour forecast returns in the last half hour. The results in Table 3.5 show that the predictability of our models are orthogonal to their findings and complement their paper meaningfully. Economically, there are likely several sources of risk and trading behaviors driving intraday patterns in returns.

3.4.2 Volatility and illiquidity effects

During periods of high volatility or illiquidity (i.e. periods of low liquidity), traders encounter higher market frictions. According to our slow trader hypothesis, predictability should be higher when volatility or illiquidity is higher. We test this hypothesis by first sorting days into 3 equal groups (tertiles) based on their daily realized volatility and Amihud (2002) measure of illiquidity, respectively. We then study the predictability individually for each tertile.

Panel A of Table 3.6 reports the R^2_{OOS} of each volatility group during the post-decimalization period. Consistent with our hypothesis, R^2_{OOS} is strictly increasing in volatility across every model (except PCR). Likewise, Panel B of Table 3.6 shows that after-transaction cost Sharpe ratios are increasing in volatility for most models. Note that column (9) shows that the intraday SPY Sharpe ratio decreases in volatility, so model Sharpe ratios relative to the benchmark are strongly increasing across the volatility tertiles.[21]

Panel A of Table 3.7 reports the R^2_{OOS} of each illiquidity group during the post-decimalization period. Interestingly, R^2_{OOS} is small for low and mid illiquidity days, but extremely large only for high illiquidity days. This suggests that most of the predictability occurs during days when the market is illiquid and when it's more expensive for traders to rebalance their portfolios. Panel B of Table 3.7 reports after-transaction cost Sharpe ratios of each illiquidity group during the post-decimalization period. For most models, Sharpe ratios are increasing in illiquidity.

In summary, predictability and economic significance increase when market volatility and illiquidity are high. These findings are similar to Gao et al. (2018) who show that 30-minute predictability is higher on high volatility and illiquidity days. Our results are consistent with their findings and with our hypothesis that market predictability is driven by slow-moving capital from traders that face market frictions. These findings also support the argument in Chordia, Roll, and Subrahmanyam (2005) and Chordia, Roll, and Subrahmanyam (2008) that prices can

[21] To examine this further, we considered a volatility timing strategy that trades only when the previous day's volatility is in a given tertile group using only past information to create the groups, i.e. with no look-ahead bias. The results show that Sharpe ratios are increasing in volatility for all models except LAS and EN, demonstrating that our ex-post volatility analysis could be converted into a tradeable volatility timing strategy.

be separated from fundamentals and are predictable in short horizons due to insufficient liquidity.

3.4.3 Impact of financial crisis

The recent 2005 to 2016 period contains the Subprime mortgage and European debt crises. These crisis periods are associated with significant market frictions. According to our hypothesis, predictability should be higher during these crisis periods. This is tested in this section by studying several interesting sub-periods of the late post-decimalization period illustrating crisis and non-crisis periods.

Panel A of Table 3.8 reports the R^2_{OOS} of crisis and non-crisis periods. Across most models, the R^2_{OOS} during the Subprime and EU crises are higher than during the 2010 - 2011 and 2014 - 2016 periods. However, the R^2_{OOS} was high during the earliest 2005-2007 period, possibly due to technological factors (i.e. less sophisticated trading). PCR had the highest R^2_{OOS} during the Subprime mortgage crisis and a positive R^2_{OOS} during the EU debt crisis. This finding confirms our assertion in Section 3.3.1 that PCR has stronger predictability when equities are more correlated and follows a strong factor structure, as is the case during a financial crises. From 2014 - 2016, model predictability is negative across models. This may be because the period was relatively stable, but could also be caused by reduced trader frictions.[22] These results demonstrate that the predictability is higher during a financial crisis. In the slow moving theory of capital, traders may not invest in arbitrage opportunities if they have insufficient capital or can invest in assets with higher expected returns (Duffie (2010)). These results suggest that during the financial crisis, the market portfolio may have been particularly predictable due to slow moving capital. These findings are similar to Gao et al. (2018), who show 30-minute predictability is higher during the Subprime mortgage crisis.

The economic significance results of crisis and non-crisis periods are somewhat mixed.

[22] It will be interesting to study if predictability increases during the ongoing Coronavirus crisis once high-frequency data becomes available for this period.

Panel B of Table 3.8 reports the pre-transaction cost Sharpe ratios of crisis and non-crisis periods. The pre-transaction cost Sharpe ratios are mostly consistent with the R^2_{OOS} results, with the exception of the 2010 - 2011 period which earned higher Sharpe ratios than expected given the low R^2_{OOS}. Panel C of Table 3.8 reports the after-transaction cost Sharpe ratios. Surprisingly, the results differ from the findings in Panel A and B for the 2005 - 2007 period and the EU debt crisis period. It appears that high transaction costs during these periods removed most of the predictive profits. Similarly, the 2010 - 2011 period had the highest after-transaction cost Sharpe ratios, potentially due to reduced transaction costs after the Subprime crisis. Interestingly, the LAS and EN models had extremely low returns during the financial crisis, likely because these models often perform poorly when predictors are highly correlated as explained in Section 3.3.3.

3.4.4 Comparison to autoregressive models

A natural comparison to our cross-sectional models with lagged intraday returns for the market constituents are autoregressive models for the market return itself, since our predictability results could simply be capturing intraday momentum. For example, Heston, Korajczyk, and Sadka (2010) find significant auto-correlation of half-hour returns at daily intervals and Gao et al. (2018) find that the first half-hour return of the SPY predicts the last. This section evaluates the predictability of linear models estimated on up to 500 lagged SPY returns using the same specifications as our baseline machine learning models. We consider a simple AR(1) that uses the SPY return with 1 lag (i.e. our baseline model without constituent returns) and an AR(p) model where the number of lagged returns (up to 500) are chosen to minimize the validation mean squared error. We also consider the linear LAS, EN, and PCR models to perform dimension reduction on the 500 lagged SPY returns.

Panel A of Table 3.9 reports the R^2_{OOS} of the linear autoregressive models together with the linear baseline models for comparison. During the overall period from 1993 to 2016, LAS and EN generally have the highest R^2_{OOS} among AR models of 0.6% and 0.58%, suggesting that

certain market return lags contain predictive information. However, these regularized models only slightly improve on the simplest AR(1) model's R^2_{OOS} of 0.55%, indicating that the most recent lag is the most important for prediction. This is reinforced by the AR(p), which uses a median of 29 lags and performs worse than the AR(1) model with an R^2_{OOS} of 0.53%. The AR(500) model without dimension reduction, i.e. estimated with OLS, performs poorly due to the high-dimensional inputs as expected. In comparison, our baseline LAS and EN models significantly outperform every AR model with an R^2_{OOS} of 2% and 1.95% respectively. Additionally, the baseline PCR's R^2_{OOS} of 0.31% is higher than the AR(500) PCR model's R^2_{OOS} of -0.07%. These results demonstrate that there is significant predictive information embedded in the lagged returns of the S&P 500 constituents.

Panel B of Table 3.9 reports the after-transaction cost Sharpe ratios of the linear autoregressive models together with the linear baseline models for comparison. During the overall period from 1993 to 2016, every AR model has a negative Sharpe ratio except for the AR(1), which has a Sharpe ratio of 0. In contrast, the baseline linear models all have positive Sharpe ratios, with the PCR's Sharpe ratio of 0.68 even outperforming the buy-and-hold S&P 500's Sharpe ratio of 0.48. These results demonstrate that using the lagged returns of the S&P 500 constituents is necessary for improving the economic significance of our predictions. The panel also shows that this holds true for all the sub-periods considered in this paper.

3.4.5 Effect of additional variables

Finally, we consider whether including additional lagged characteristics of the S&P 500 constituents may improve forecasting accuracy. Due to computational constraints (both in terms of memory requirements and training time), we only consider results using the four-month training sample during the post-decimalization period. We consider characteristics that proxy short-term changes in liquidity and trading trends. The characteristics include firm-level market beta, momentum, maximum, minimum, volatility, illiquidity, trading volume, kurtosis, and skewness calculated over preceding days as well as the lagged observed bid-ask spread.

Table 3.10 reports R^2_{OOS} when including different characteristics during the post-decimalization period. The first row reports R^2_{OOS} for the baseline model using lagged returns only. The panel shows that the LAS model is improved by including illiquidity, momentum, or skewness, indicating that there may be some index constituents with useful characteristics for predicting the market return. PCR is also slightly improved by adding momentum and volume. However, for most models including any single characteristic reduces the R^2_{OOS}. One possibility for the poor performance of the price trend and liquidity characteristics is that they are estimated over the preceding day and potentially noisy. However when we include the lagged bid-ask spread, which is not estimated, the R^2_{OOS} also decreases across most models. These result may suggest that including additional characteristics simply adds noise to the model without providing additional predictive information. While adding these characteristics generally does not improve market forecasts beyond the baseline model, it remains possible that such characteristics may help predict individual stock returns.

Panel B of Table 3.10 reports the after-transaction cost Sharpe ratios when including different characteristics during the post-decimalization period. This panel shows, similarly to Panel A, that including some characteristics may slightly improve the Sharpe ratios of certain models. For example, LAS, EN, PCR, RF, ANN, and the ensemble median can be slightly improved by including some characteristics. In particular, the median ensemble after including the bid-ask spread achieves a Sharpe ratio of 0.44. However, these results are not robust and could just be due to random chance.

One concern may be that including other characteristics may increase the dimensionality of the data and increase estimation error. However, we also analyzed the predictability of only the characteristics, removing lagged returns from the model, and again found no evidence of predictability. In summary, we find little evidence that including other characteristics can improve predictions. Among our tested characteristics, only lagged returns consistently predict market returns.

3.5 Conclusion

This paper conducts, to our knowledge, the largest study ever of intraday market return predictability using state-of-the-art machine learning models trained on the cross-section of lagged market index constituent returns and other characteristics to forecast five-minute market returns over the longest possible time period. The paper demonstrates that there is significant statistical predictability of intraday market returns and establishes that this return predictability translates into economically significant profits even after accounting for transaction costs. Furthermore, we show that the lagged constituent returns holds significant predictive information that is not contained in lagged market returns or in lagged price trend and liquidity characteristics.

Specifically, the paper shows that regularized linear models such as lasso and elastic nets and nonlinear tree-based models such as random forests yield the largest positive out-of-sample R^2s. Linear models such as principal component analysis had a high out-of-sample R^2s during the Subprime crisis and EU debt crisis, providing returns that hedge these crisis states. Ensemble models that combine individual model predictions perform the best across time and the return predictability from these models translates into economically significant profits with Sharpe ratios after transaction costs of 0.98. This Sharpe ratio is much higher than what is obtained from holding the index intraday and significantly exceeds the Sharpe ratio of the benchmark buy-and-hold strategy.

Across time, we show that the statistical predictability has suffered somewhat as transaction costs were reduced post-decimalization. We argue that this strongly suggests that predictability could be a result of slow-moving trader capital. Consistent with the hypothesis of slow traders, market returns are also shown to be more predictable during the middle of the day when trading activity is lower, on days with high volatility or high illiquidity where prices can be driven further away from their fundamental values, and during years of financial crisis which are plagued by market frictions. Nevertheless, the best ensemble models retain some predictability and trading based on the model's signals remain profitable throughout the sample, even after adjusting for transaction costs.

Our results provide strong evidence that market returns are predictable over short-horizons. Our empirical findings suggest that further investigation into the economic mechanisms driving such short-horizon predictability is warranted. The late-informed investor explanation in Gao et al. (2018) is supported by our evidence. Another explanation in Chinco and Fos (2019) theorizes that computational complexity of traders' rebalancing introduces noise. We believe this can explain some of our predictability results, wherein models appear able to capture the systematic behaviour of traders' rebalancing. In particular, this could explain the persistent profitability of our models in recent years. However, verifying these economic mechanism requires further investigation which we leave for future research.

Table 3.1: Market predictability

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	LAS	EN	PCR	RF	GBRT	ANN	MEAN	MED
Panel A: Overall Period (1993 - 2016)								
Intercept	-0.64	-0.59	-0.31	-0.45	-0.36	-0.32	-0.64	-0.67
t-stat	-4.02	-3.72	-1.98	-2.84	-2.28	-1.98	-4.04	-4.24
Slope	1.38	1.29	0.65	1.22	0.90	0.83	1.53	1.54
t-stat	47.26	46.06	24.14	54.11	43.23	42.49	58.28	58.97
R^2_{OOS}	2.00	1.95	0.31	1.71	1.64	1.40	2.00	2.01
R^2_{OOS}/Cost	1.16	1.13	0.18	0.99	0.95	0.81	1.15	1.16
Panel B: 1/8 Tick Size Period (1993 - 1996)								
Intercept	-0.22	-0.17	-0.16	-0.14	-0.32	0.16	-0.21	-0.26
t-stat	-1.16	-0.90	-0.95	-0.84	-1.68	0.79	-1.29	-1.59
Slope	1.55	1.45	0.92	1.70	1.18	1.17	1.59	1.65
t-stat	36.53	35.80	44.92	41.48	38.60	41.24	42.96	41.07
R^2_{OOS}	9.27	9.32	6.52	6.78	8.95	7.41	9.16	9.14
R^2_{OOS}/Cost	10.82	10.88	7.61	7.92	10.46	8.66	10.70	10.67
Panel C: 1/16 Tick Size Period (1997 - 2000)								
Intercept	-1.71	-1.67	-0.61	-1.28	-1.22	-1.04	-1.67	-1.73
t-stat	-3.91	-3.82	-1.40	-2.88	-2.71	-2.23	-3.80	-3.93
Slope	1.42	1.38	0.18	1.27	1.05	0.92	1.62	1.56
t-stat	35.52	35.31	2.77	46.73	41.86	38.57	43.56	45.09
R^2_{OOS}	3.72	3.69	-0.76	3.48	3.45	3.21	3.60	3.66
R^2_{OOS}/Cost	0.68	0.67	-0.14	0.64	0.63	0.59	0.66	0.67
Panel D: Early Post-Decimalization Period (2001 - 2004)								
Intercept	-0.65	-0.57	-0.42	-0.25	-0.30	-0.14	-0.55	-0.57
t-stat	-1.40	-1.24	-0.91	-0.55	-0.68	-0.32	-1.19	-1.24
Slope	1.14	0.99	0.59	0.87	0.63	0.61	1.36	1.38
t-stat	10.52	9.67	4.67	16.09	13.46	10.87	15.12	14.65
R^2_{OOS}	0.91	0.81	0.03	0.85	0.48	0.38	1.04	1.01
R^2_{OOS}/Cost	0.30	0.27	0.01	0.28	0.16	0.13	0.35	0.34
Panel E: Late Post-Decimalization Period (2005 - 2016)								
Intercept	-0.10	-0.03	-0.12	-0.07	0.08	0.07	-0.14	-0.20
t-stat	-0.43	-0.11	-0.52	-0.30	0.35	0.29	-0.60	-0.84
Slope	0.58	0.38	0.64	0.48	0.13	0.10	0.68	0.82
t-stat	3.70	3.17	4.80	4.85	1.53	1.25	4.02	4.59
R^2_{OOS}	0.02	-0.04	0.04	-0.01	-0.39	-0.45	0.04	0.06
R^2_{OOS}/Cost	0.07	-0.14	0.13	-0.03	-1.22	-1.42	0.12	0.18

This table reports coefficients from the predictive regression in Equation (3.13) along with Newey-West t-statistics with 79 lags, R^2_{OOS} percentages, and R^2_{OOS} scaled by transaction cost for the SPY using the lasso (LAS), elastic net (EN), principal component regression (PCR), random forest (RF), gradient-boosted regression trees (GBRT), neural network (ANN), and ensemble mean (MEAN) and median (MED). The intercept is reported as a percentage scaled by 1,000. The predictors used are lagged returns for the market and all the S&P 500 constituents. Results are reported for the full, 1/16 tick size, 1/8 tick size, early post-decimalization, and late post-decimalization samples.

Table 3.2: Excess Returns and Sharpe ratios

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	LAS	EN	PCR	RF	GBRT	ANN	MEAN	MED	Intraday SPY	Hold SPY
Panel A: Overall Period (1993-2016)										
Return	1.91	1.88	0.75	1.98	1.92	1.72	2.05	2.04	-0.03	0.07
t-stat	4.81	4.92	4.62	4.50	4.26	5.11	4.60	4.91	-1.11	2.19
SR	2.71	2.68	1.97	2.90	2.84	2.66	2.90	2.82	-0.24	0.48
Panel B: 1/8 Tick Size Period (1993 - 1996)										
Return	3.74	3.73	3.15	3.47	3.47	3.39	3.83	3.88	-0.01	0.11
t-stat	5.73	6.00	10.73	6.43	6.35	8.13	6.77	6.72	-0.37	2.62
SR	5.63	5.60	8.49	6.81	5.73	6.90	6.63	6.19	-0.17	1.19
Panel C: 1/16 Tick Size Period (1997 - 2000)										
Return	5.61	5.57	0.33	5.72	5.61	5.11	5.79	5.83	-0.14	0.12
t-stat	25.49	26.00	1.35	16.05	18.06	18.92	17.92	20.74	-1.78	1.62
SR	12.35	12.20	1.01	11.30	12.73	10.35	11.73	11.28	-0.89	0.71
Panel D: Early Post-Decimalization Period (2001-2004)										
Return	1.61	1.51	0.31	1.74	1.61	1.37	1.85	1.77	-0.04	-0.01
t-stat	5.12	5.74	3.32	5.37	6.31	4.06	4.46	4.60	-0.63	-0.12
SR	4.55	4.62	1.56	4.29	4.84	4.42	4.67	4.64	-0.35	-0.07
Panel E: Late Post-Decimalization Period (2005 - 2016)										
Return	0.18	0.17	0.29	0.33	0.29	0.16	0.30	0.27	0.01	0.07
t-stat	4.48	3.89	7.75	6.02	4.84	3.27	6.22	4.96	0.32	1.37
SR	1.36	1.41	2.10	2.06	2.03	0.83	1.80	1.73	0.09	0.46

Reported are the annualized excess returns, t-statistics, and Sharpe ratios for the SPY market-timing strategy. Models include the lasso (LAS), elastic net (EN), principal component regression (PCR), random forest (RF), gradient-boosted regression trees (GBRT), neural network (ANN), mean ensemble (MEAN), and median ensemble (MED). The independent variables are lagged returns of S&P 500 constituents. Results are reported for the full, 1/16 tick size, 1/8 tick size, early post-decimalization, and late post-decimalization samples.

Table 3.3: Excess Returns and Sharpe ratios with transaction costs

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	LAS	EN	PCR	RF	GBRT	ANN	MEAN	MED	Intraday SPY	Hold SPY
Panel A: Overall Period (1993 - 2016)										
Return	0.03	0.01	0.05	0.06	0.02	-0.07	0.06	0.06	-0.03	0.07
t-stat	1.71	0.77	3.29	2.70	1.10	-3.87	3.38	4.27	-1.11	2.19
SR	0.42	0.20	0.68	0.77	0.24	-0.66	0.67	0.98	-0.24	0.48
% Trades	11.63	12.73	11.45	12.39	14.86	19.26	11.21	11.00		
Panel B: 1/8 Tick Size Period (1993 - 1996)										
Return	0.04	0.05	0.04	0.01	0.07	0.03	0.05	0.04	-0.01	0.11
t-stat	4.79	4.86	2.68	1.22	2.82	2.77	4.98	4.28	-0.37	2.62
SR	2.50	2.60	1.48	0.80	2.27	1.20	2.70	2.34	-0.17	1.19
% Trades	0.47	0.60	1.53	0.27	1.25	0.75	0.52	0.48		
Panel C: 1/16 Tick Size Period (1997 - 2000)										
Return	0.01	0.01	0.01	0.01	0.00	0.01	0.01	0.01	-0.14	0.12
t-stat	1.83	1.67	1.48	1.88	0.46	1.38	2.15	2.42	-1.78	1.62
SR	1.01	0.95	0.75	0.97	0.21	1.14	1.04	0.98	-0.89	0.71
% Trades	0.11	0.12	0.07	0.13	0.20	0.25	0.08	0.10		
Panel D: Early Post-Decimalization Period (2001 - 2004)										
Return	0.14	0.13	0.00	0.11	0.06	-0.06	0.17	0.17	-0.04	-0.01
t-stat	3.17	3.05	0.05	2.09	0.85	-1.29	3.17	3.70	-0.63	-0.12
SR	2.00	1.97	0.02	1.20	0.49	-0.61	2.20	2.19	-0.35	-0.07
% Trades	10.57	11.31	7.23	12.76	15.01	17.54	10.05	10.05		
Panel E: Late Post-Decimalization Period (2005 - 2016)										
Return	-0.01	-0.04	0.09	0.08	0.00	-0.14	0.04	0.05	0.01	0.07
t-stat	-0.58	-1.41	3.07	1.95	-0.18	-4.44	1.41	2.32	0.32	1.37
SR	-0.18	-0.49	0.88	0.81	-0.05	-0.98	0.35	0.68	0.09	0.46
% Trades	19.45	21.35	19.88	20.29	24.12	32.18	18.78	18.37		

Reported are the annualized excess returns, t-statistics, and Sharpe ratios for the SPY market-timing strategy with transaction costs. Models include the lasso (LAS), elastic net (EN), principal component regression (PCR), random forest (RF), gradient-boosted regression trees (GBRT), neural network (ANN), mean ensemble (MEAN), and median ensemble (MED). The independent variables are lagged returns of S&P 500 constituents. Results are reported for the full, 1/16 tick size, 1/8 tick size, early post-decimalization, and late post-decimalization samples.

Table 3.4: Market predictability (percentage R^2_{OOS}) and profitability (Sharpe ratio) by training duration post-decimalization

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Post-Decimalization Out-of-Sample R^2								
Training months	LAS	EN	PCR	RF	GBRT	ANN	MEAN	MED
Expanding	0.30	0.22	0.04	0.26	-0.12	-0.19	0.35	0.35
58 month	0.33	0.30	-0.06	0.25	-0.13	-0.15	0.38	0.39
34 month	0.23	0.26	0.09	0.23	-0.29	-0.20	0.36	0.38
22 month	0.29	0.27	0.07	0.22	-0.42	-0.30	0.33	0.34
16 month	0.19	0.18	0.08	-0.01	-0.39	-0.36	0.28	0.30
10 month	0.07	0.05	0.07	-0.06	-0.36	-0.29	0.22	0.24
7 month	0.05	0.05	0.00	-0.03	-0.36	-0.42	0.20	0.21
4 month	0.05	0.05	-0.08	-0.22	-0.63	-0.65	0.12	0.14
1 month	-0.09	-0.09	-0.33	-0.68	-1.35	-0.40	-0.08	-0.04
Panel B: Post-Decimalization Sharpe Ratio after Transaction Costs								
	LAS	EN	PCR	RF	GBRT	ANN	MEAN	MED
Expanding	0.34	0.06	0.72	0.90	0.12	-0.90	0.69	1.04
58 month	0.26	0.21	-0.16	0.85	0.23	-0.73	0.80	0.81
34 month	-0.07	-0.21	0.19	0.69	-0.30	-0.72	0.45	0.56
22 month	-0.13	-0.26	0.43	0.19	-0.30	-1.27	0.50	0.37
16 month	-0.10	-0.28	0.41	0.06	-0.24	-1.04	0.41	0.57
10 month	-0.11	-0.14	0.27	-0.43	-0.64	-0.54	0.50	0.55
7 month	-0.04	-0.06	0.28	-0.03	-0.22	-0.63	0.59	0.61
4 month	-0.41	-0.40	0.02	-0.30	-0.66	-0.89	0.34	0.15
1 month	-0.16	-0.15	-0.40	-1.50	-2.13	-1.29	-0.29	0.09

Reported are the out-of-sample predictive R^2 percentages and annualized Sharpe ratios for the SPY market-timing strategy with transaction costs. Models include the lasso (LAS), elastic net (EN), principal component regression (PCR), random forest (RF), gradient-boosted regression trees (GBRT), neural network (ANN), mean ensemble (MEAN), and median ensemble (MED). We compare R^2 values across 1-, 4-, 7-, 10-, 16-, 22-, 34-, and 58-month training windows. The independent variables are lagged returns of S&P 500 constituents. Reported are results for the *post-decimalization* subsamples.

Table 3.5: Market predictability (percentage R^2_{OOS}) and profitability (Sharpe ratio) by time post-decimalization

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Post-Decimalization Out-of-Sample R^2									
Time	LAS	EN	PCR	RF	GBRT	ANN	MEAN	MED	
9:35 - 10:00	-0.06	-0.21	0.21	-0.20	-1.26	-0.72	0.12	0.19	
10:00 - 10:30	0.34	0.26	0.05	0.09	-0.27	-0.17	0.33	0.36	
10:30 - 11:00	0.38	0.40	0.10	0.55	0.29	-0.02	0.56	0.49	
11:00 - 11:30	0.41	0.32	0.03	0.35	0.20	-0.07	0.49	0.47	
11:30 - 12:00	0.32	0.24	-0.02	0.29	-0.22	-0.11	0.38	0.38	
12:00 - 12:30	0.47	0.41	-0.01	0.48	0.32	-0.21	0.54	0.51	
12:30 - 13:00	0.39	0.26	0.24	0.48	-0.20	-0.46	0.40	0.45	
13:00 - 13:30	0.31	0.33	0.10	0.43	-0.05	-0.28	0.38	0.31	
13:30 - 14:00	0.37	0.26	0.01	0.34	-0.04	-0.07	0.38	0.39	
14:00 - 14:30	0.99	0.96	0.19	0.62	0.65	0.62	0.90	0.84	
14:30 - 15:00	0.42	0.35	0.17	0.47	0.01	0.12	0.44	0.47	
15:00 - 15:30	-0.16	-0.27	-0.26	-0.08	-0.32	-0.77	-0.13	-0.09	
15:30 - 15:55	0.10	-0.01	-0.26	0.13	0.11	-0.11	0.17	0.14	
Panel B: Post-Decimalization Sharpe Ratio after Transaction Costs									
Time	LAS	EN	PCR	RF	GBRT	ANN	MEAN	MED	Intraday SPY
9:35 - 10:00	-0.20	-0.35	0.23	0.26	-0.64	-0.63	0.55	0.21	-0.25
10:00 - 10:30	0.57	0.47	0.60	0.08	0.19	-0.19	0.50	0.69	-0.02
10:30 - 11:00	0.18	0.23	0.35	0.07	0.17	0.03	0.62	0.31	-0.21
11:00 - 11:30	-0.11	-0.30	-0.08	0.00	0.18	-0.17	0.08	-0.08	-0.29
11:30 - 12:00	0.13	0.15	0.21	0.52	0.08	-0.23	0.40	0.32	0.12
12:00 - 12:30	0.00	-0.09	0.44	0.38	0.20	-0.50	0.28	0.24	-0.04
12:30 - 13:00	0.22	0.32	0.40	0.39	-0.52	-0.89	0.18	0.62	0.47
13:00 - 13:30	0.13	0.11	0.55	0.74	0.23	-0.29	0.18	0.57	0.16
13:30 - 14:00	0.17	0.00	0.09	0.39	-0.26	-0.59	0.34	0.34	-0.31
14:00 - 14:30	0.01	0.19	0.51	0.44	0.64	-0.17	0.27	0.34	-0.23
14:30 - 15:00	0.31	0.17	0.49	0.52	0.24	-0.30	0.35	0.59	0.43
15:00 - 15:30	-0.13	-0.30	-0.30	-0.34	-0.51	-0.99	-0.59	-0.39	0.27
15:30 - 15:55	-0.11	-0.28	-0.44	0.43	0.40	0.27	-0.05	-0.11	-0.09

Reported are the out-of-sample predictive R^2 percentages and annualized Sharpe ratios for the SPY market-timing strategy with transaction costs. Models include the lasso (LAS), elastic net (EN), principal component regression (PCR), random forest (RF), gradient-boosted regression trees (GBRT), neural network (ANN), mean ensemble (MEAN), and median ensemble (MED). We compare results across 30-minute windows throughout the trading day. The independent variables are lagged returns of S&P 500 constituents. Reported are results for the *post-decimalization* subsamples.

Table 3.6: Market predictability (percentage R^2_{OOS}) and profitability (Sharpe ratio) by volatility in the post-decimalization period

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Post-Decimalization Out-of-Sample R^2									
Volatility	LAS	EN	PCR	RF	GBRT	ANN	MEAN	MED	
Low	0.09	-0.02	-0.03	0.02	-0.41	-0.76	0.12	0.17	
Mid	0.19	0.11	-0.09	0.11	-0.11	-0.37	0.24	0.25	
High	0.34	0.26	0.07	0.31	-0.09	-0.11	0.40	0.39	
Panel B: Post-Decimalization Sharpe Ratio after Transaction Costs									
Volatility	LAS	EN	PCR	RF	GBRT	ANN	MEAN	MED	Intraday SPY
Low	0.47	0.09	0.20	0.23	-0.81	-1.79	-0.03	0.44	2.88
Mid	0.40	0.39	0.43	0.28	0.26	-1.08	0.81	0.89	-0.20
High	0.14	-0.15	0.73	1.04	0.23	-0.33	0.60	0.87	-0.93

Reported are the out-of-sample predictive R^2 percentages and annualized Sharpe ratios for the SPY market-timing strategy with transaction costs. Models include the lasso (LAS), elastic net (EN), principal component regression (PCR), random forest (RF), gradient-boosted regression trees (GBRT), neural network (ANN), mean ensemble (MEAN), and median ensemble (MED). We compare results across 3 groups sorted on realized volatility. The independent variables are lagged returns of S&P 500 constituents. Reported are results for the *post-decimalization* subsamples.

Table 3.7: Market predictability (percentage R^2_{OOS}) and profitability (Sharpe ratio) by illiquidity in the post-decimalization period

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Post-Decimalization Out-of-Sample R^2									
Illiquidity	LAS	EN	PCR	RF	GBRT	ANN	MEAN	MED	
Low	0.00	-0.08	0.06	-0.06	-0.41	-0.46	0.04	0.04	
Mid	0.01	-0.04	0.02	0.02	-0.43	-0.34	0.05	0.07	
High	0.76	0.66	0.04	0.69	0.37	0.11	0.85	0.83	
Panel B: Post-Decimalization Sharpe Ratio after Transaction Costs									
Illiquidity	LAS	EN	PCR	RF	GBRT	ANN	MEAN	MED	Intraday SPY
Low	-0.25	-0.43	0.42	0.11	-0.34	-1.08	0.24	0.26	0.33
Mid	-0.02	-0.21	0.93	0.83	0.11	-0.47	0.23	0.59	-0.22
High	1.03	0.85	0.18	1.03	0.47	-0.56	1.50	1.22	-0.07

Reported are the out-of-sample predictive R^2 percentages and annualized Sharpe ratios for the market-timing strategy with transaction costs. Models include the lasso (LAS), elastic net (EN), principal component regression (PCR), random forest (RF), gradient-boosted regression trees (GBRT), neural network (ANN), mean ensemble (MEAN), and median ensemble (MED). We compare results across 3 groups sorted on Amihud (2002) illiquidity. The independent variables are lagged returns of S&P 500 constituents. Reported are results for the *post-decimalization* subsamples.

Table 3.8: Market predictability (percentage R^2_{OOS}) and profitability (Sharpe ratio) by crisis period

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
					Panel A: Post-Decimalization Out-of-Sample R^2					
	LAS	EN	PCR	RF	GBRT	ANN	MEAN	MED		
2005-2007	0.07	-0.04	-0.01	0.00	-0.11	-0.30	0.12	0.10		
Subprime	0.04	-0.02	0.08	0.02	-0.37	-0.48	0.06	0.07		
2010-2011	-0.02	-0.05	0.03	-0.03	-0.67	-0.39	-0.03	0.02		
EU Debt	0.08	-0.01	0.01	0.03	-0.30	-0.70	0.06	0.10		
2014-2016	-0.07	-0.17	-0.04	-0.14	-0.36	-0.41	-0.03	0.00		
					Panel B: Post-Decimalization Sharpe Ratio before Transaction Costs					
	LAS	EN	PCR	RF	GBRT	ANN	MEAN	MED	Intraday SPY	Hold SPY
2005-2007	1.64	1.48	1.65	1.82	2.08	2.06	2.38	1.97	-1.09	0.56
Subprime	1.06	0.96	3.11	3.78	2.67	0.94	2.22	1.91	0.07	-0.40
2010-2011	2.18	2.07	2.22	1.87	2.32	1.14	2.42	3.22	0.20	0.54
EU crisis	3.07	2.88	2.82	3.08	2.30	0.84	3.52	3.08	1.34	2.29
2014-2016	0.79	0.94	1.70	1.68	1.88	0.51	0.91	1.35	0.61	0.68
					Panel C: Post-Decimalization Sharpe Ratio after Transaction Costs					
	LAS	EN	PCR	RF	GBRT	ANN	MEAN	MED	Intraday SPY	Hold SPY
2005-2007	0.35	-0.06	0.17	0.08	-0.21	-0.92	0.16	-0.04	-1.09	0.56
Subprime	-1.09	-1.55	1.38	2.00	-0.12	-0.52	0.73	0.87	0.07	-0.40
2010-2011	0.77	0.30	1.97	1.07	0.38	-1.35	0.69	1.82	0.20	0.54
EU crisis	0.63	-0.22	0.68	0.14	-1.05	-2.68	-0.09	0.57	1.34	2.29
2014-2016	-0.58	-0.38	0.64	0.12	0.17	-2.18	-0.06	0.40	0.61	0.68

Reported are the out-of-sample predictive R^2 percentages and annualized Sharpe ratios for the SPY market-timing strategy without and with transaction costs. Models include the lasso (LAS), elastic net (EN), principal component regression (PCR), random forest (RF), gradient-boosted regression trees (GBRT), neural network (ANN), mean ensemble (MEAN), and median ensemble (MED). The independent variables are lagged returns of S&P 500 constituents. Reported are results for the late post-decimalization subsample which include the Subprime Mortgage crisis, during 2008-2009, and the European Debt crisis, during 2012-2013.

Table 3.9: Market predictability (percentage R^2_{OOS}) and profitability (Sharpe ratio) for autoregressive models

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Out-of-Sample R^2									
	Autoregressive Models						Baseline Models		
	AR(1)	AR(p)	OLS	LAS	EN	PCR	LAS	EN	PCR
1993-2016	0.55	0.53	-0.49	0.58	0.60	-0.07	2.00	1.95	0.31
1993-1996	0.22	0.37	-4.51	0.12	0.40	-0.04	9.27	9.32	6.52
1997-2000	1.87	1.95	1.07	1.78	1.79	-0.01	3.72	3.69	-0.76
2001-2004	0.17	0.08	-0.81	0.30	0.27	-0.13	0.91	0.81	0.03
2005-2016	-0.10	-0.20	-0.67	-0.02	-0.03	-0.08	0.02	-0.04	0.04
Panel B: Sharpe Ratio after Transaction Costs									
	Autoregressive Models						Baseline Models		
	AR(1)	AR(p)	OLS	LAS	EN	PCR	LAS	EN	PCR
1993-2016	0.00	-0.63	-1.25	-0.27	-0.38	-0.03	0.42	0.20	0.68
1993-1996	-0.49	-0.71	-0.56	-0.68	-0.68	0.14	2.50	2.60	1.48
1997-2000	0.72	0.73	0.41	0.49	0.65	0.33	1.01	0.95	0.75
2001-2004	0.10	-0.39	-2.08	0.63	0.27	0.38	2.00	1.97	0.02
2005-2016	-0.04	-0.91	-1.70	-0.50	-0.62	-0.10	-0.18	-0.49	0.88

Reported are the out-of-sample predictive R^2 percentages and annualized Sharpe ratios for the SPY market-timing strategy with transaction costs. Models include the AR(1), AR(p), lasso (LAS), elastic net (EN), and principal component regression (PCR). The independent variables are 500 lagged returns of S&P 500. Results are reported for the full, 1/16 tick size, 1/8 tick size, early post-decimalization, and late post-decimalization samples.

Table 3.10: Market predictability (percentage R^2_{OOS}) and profitability (Sharpe ratio) with additional characteristics in the post-decimalization period

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Panel A: Post-Decimalization Out-of-Sample R^2							
	LAS	EN	PCR	RF	GBRT	ANN	MEAN	MED
Base	0.05	0.05	-0.08	-0.22	-0.63	-0.65	0.12	0.14
BETA	-0.98	-1.41	-2.44	-1.92	-1.54	-4.10	-0.39	-0.03
ILLIQ	0.09	-0.44	-8.73	-2.70	-7.74	-3.22	-0.62	0.05
KURT	-0.64	-2.27	-1.1E+18	-2.12	-1.62	-4.25E+23	-1.18E+22	-0.01
MAX	0.03	-0.50	-0.10	-0.56	-1.14	-1.27	-0.02	0.09
MIN	-0.37	-0.36	-0.10	-5.53	-0.68	-1.38	-0.25	0.01
MOM	0.07	-1.24	-0.07	-0.42	-2.59	-1.05	-0.08	0.08
SKEW	0.06	-1.13	-4.8E+12	-0.72	-1.59	-1.8E+15	-5.5E+13	-0.01
VOL	-0.02	-0.35	-0.16	-1.69	-2.26	-1.09	-0.07	0.07
VOLUME	-1.95	-2.24	-0.06	-3.05	-7.49	-1.35	-0.47	-0.24
SPREAD	-1.51	-1.20	-3.67	-0.12	-0.42	-1.40	-0.08	0.13
	Panel B: Post-Decimalization Sharpe Ratio after Transaction Costs							
	LAS	EN	PCR	RF	GBRT	ANN	MEAN	MED
Base	-0.41	-0.40	0.02	-0.30	-0.66	-0.89	0.34	0.15
BETA	-0.27	-0.30	0.14	-0.80	-2.18	-1.38	-0.17	-0.08
ILLIQ	-0.29	-0.36	-0.48	-1.12	-2.77	-1.29	-0.95	-0.18
KURT	-0.47	-0.46	0.22	-0.90	-1.74	-1.19	-0.07	0.11
MAX	-0.21	-0.31	-0.27	-0.71	-1.76	-1.00	-0.53	0.02
MIN	-0.18	-0.18	-0.06	-0.47	-1.28	-0.99	-0.20	0.06
MOM	-0.29	-0.32	0.12	-0.80	-1.72	-0.92	-0.18	0.07
SKEW	-0.26	-0.37	0.27	-0.62	-1.58	-1.00	-0.10	0.32
VOL	-0.53	-0.61	-0.21	-0.94	-2.33	-1.14	-0.17	-0.12
VOLUME	-0.24	-0.27	0.28	-1.02	-2.75	-1.11	-0.68	0.08
SPREAD	-0.21	-0.25	-0.03	0.16	-0.56	-1.00	0.17	0.44

Reported are the out-of-sample predictive R^2 percentages and annualized Sharpe ratios for the SPY using the lasso (LAS), elastic net (EN), principal component regression (PCR), random forest (RF), gradient-boosted regression trees (GBRT), neural network (ANN), mean ensemble (MEAN), and median ensemble (MED). In addition to lagged returns (LAG), the independent variables include market beta (BETA), illiquidity (ILLIQ), kurtosis (KURT), maximum (MAX), minimum (MIN), momentum (MOM), skewness (SKEW), volatility (VOL), or volume (VOLUME) of S&P 500 constituents calculated over the previous day as well as the percent bid-ask spread (SPREAD). See Appendix B.1.2 for details on how these characteristics are calculated.

Chapter 4

Can the Premium for Idiosyncratic Tail Risk be Explained by Exposures to its Common Factor?

Introduction

Chapters 2 and 3 study the predictability of tail risk and the expected returns of the market portfolio, respectively. This chapter is motivated by the previous research and studies the relationship between the tail risk and expected returns of individual stocks. To analyze the relationship, this chapter introduces a new measure of firm-level tail risk estimated using high-frequency data.

A stock's idiosyncratic tail risk (ITR), denoted as ξ in this paper, measures the density of the left tail of idiosyncratic returns, the residual returns after removing the effect of systematic factors. Empirically, stocks with higher idiosyncratic tail risk have higher average excess returns. This idiosyncratic tail risk premium is recently documented in Savor (2012), Jiang and Zhu (2017), Begin, Dorion, and Gauthier (2019) and Kapadia and Zekhnini (2019), and it contradicts classical asset pricing theory where only systematic risk earns a risk premium. These

papers speculate that the premium is caused by either the inability for investors to diversify due to frictions or market under-reaction to firm-specific news.[1] However, none of them find that the premium can be explained in the conventional manner by exposures to systematic risk.

In this paper, I propose a risk-based explanation for the idiosyncratic tail risk premium that combines two ideas. First, I hypothesize that idiosyncratic tail risk is caused by large intermediaries trading on news or propriety analysis, which is motivated by Gabaix et al. (2006).[2] Since intermediaries require funding to trade, shocks to intermediary funding cause idiosyncratic tail risks of different firms to comove over time and share a common factor, which is denoted as the common idiosyncratic tail risk (CITR). This explanation is consistent with the recent finding of commonality in idiosyncratic tail risk (Qin and Todorov (2019) and Begin, Dorion, and Gauthier (2019)). Since CITR decreases when intermediary constraints tighten, it is correlated to the intermediary marginal value of wealth and is *procyclical*.

Second, the recent intermediary asset pricing literature shows that intermediaries are the marginal investor in many financial assets. Subsequently, shocks to their funding, correlated with shocks to CITR, are an important source of undiversifiable risk and may explain cross-sectional differences in average returns (see e.g. Brunnermeier and Pedersen (2009)). Since intermediaries play a central role in explaining asset prices, they may explain the idiosyncratic tail risk premium. I hypothesize that firms with higher idiosyncratic tail risk have higher exposure to shocks to CITR, thus commanding a risk premium. Differences in exposure to CITR shocks provide a risk-based explanation for the idiosyncratic tail risk premium.

To test my explanation, I first introduce a new measure of *idiosyncratic* tail risk, that is the shape parameter of the power-law distribution of idiosyncratic returns, denoted as ξ. A large ξ corresponds to high density in the left tail of idiosyncratic returns. Therefore ξ, which is inversely proportional the slope of the tails, is an intuitive measure of idiosyncratic tail risk. To measure ξ, I use a high-frequency factor model, then use Extreme Value Theory to prove

[1] Examples of frictions include short-sale constraints, narrow framing, or market incompleteness due to non-traded assets (e.g. human capital or private businesses).

[2] Examples of intermediaries include hedge funds, mutual funds, or banks' propriety trading desks.

that my idiosyncratic tail risk measure is inherited under time-aggregation, meaning that within each month, the idiosyncratic tail risk calculated on high-frequency returns can be extrapolated to longer time horizons (i.e. daily, weekly, or monthly), even with a finite number of return observations. Moreover, the measure is robust to certain microstructure noise processes.

Testing my explanation requires a large sample of medium and small stocks, since their idiosyncratic tail risks are most affected by large intermediary trades. Existing options-based measures of idiosyncratic tail risk require a large cross-sections of options at different strike prices (i.e. Begin, Dorion, and Gauthier (2019), Kapadia and Zekhnini (2019), and Qin and Todorov (2019)), which are available only for a limited number of large stocks. Instead, I use the richness of high-frequency data to study a much larger sample of stocks. Additionally, existing options-based measures only extract the market factor, while my high-frequency model can remove multiple factors from returns.

I begin my empirical analysis by establishing that the idiosyncratic tail risk premium is a significant and prevalent phenomenon. I estimate the factor model with respect to the Fama and French (2015) five factor model and estimate the idiosyncratic tail risk as the Hill estimate of idiosyncratic returns. In each month, I sort stocks into deciles based on their idiosyncratic tail risk and hold the portfolios for a month. Stocks in the decile with the highest idiosyncratic tail risk earn 0.66% (t-stat of 3.16), approximately 8% annually, higher value-weighted returns than stocks in the lowest decile. This idiosyncratic tail risk premium cannot be explained by the market, size, value, profitability, investment, or momentum factors. The premium exists even when stocks are conditionally sorted on other firm characteristics and is robust to different factor model specifications and tail risk estimation methods. Additionally, Gabaix et al. (2006) theorize that large intermediaries cause both idiosyncratic tail risk and trading volume tail risk, predicting that their tail distributions are proportional. Under my explanation, volume tail risk (VTR) as a proxy for idiosyncratic tail risk should also earn a premium. To test my explanation, I estimate volume tail risk as the Hill estimate of the right tail of changes in trading volume. Stocks in the decile with the highest volume tail risk earn 0.67% (t-stat of 3.70) higher value-

weighted returns than stocks in the lowest decile. This volume tail risk premium is highly correlated to the idiosyncratic tail risk premium, indicating that these premia are driven by the same factor.

The idiosyncratic tail risk and volume tail risk premia are highly persistent. In each month, I sort stocks into deciles based on their idiosyncratic tail risk and hold the portfolios for up to two years. For the two-year holding period, stocks in the decile with the highest idiosyncratic tail risk earn 16.60% (t-stat of 3.44) higher value-weighted returns than stocks in the lowest decile. Likewise, stocks in the decile with the highest volume tail risk earn 22.46% (t-stat of 3.71) higher value-weighted returns than stocks in the lowest decile for the two-year holding period. This finding is consistent with the intuition that firm exposure to CITR shocks does not frequently change. If the idiosyncratic tail risk premium is compensation for exposure to CITR shocks, and stocks' risk exposures do not frequently change, then the risk premium should be persistent, which is exactly what I document.

Next, I test the large intermediary hypothesis of idiosyncratic tail risk by conducting firm-level regressions. A cross-sectional regression shows that idiosyncratic tail risk is correlated to intermediary trading volume as a percentage of total trading volume, where a one percent increase in intermediary volume increases idiosyncratic tail risk by 0.06. Furthermore, idiosyncratic tail risk is highly correlated to volume tail risk, providing empirical evidence that intermediary trades cause both idiosyncratic tail risk and volume tail risk (Gabaix et al. (2006)). Finally, idiosyncratic tail risk is highly autocorrelated, which provides evidence against the explanation that the measure is only driven by fundamental news-shocks. This finding is consistent with previous studies that show large intermediaries frequently trade the same stocks (Sias (2004)).

Comovement in idiosyncratic tail risks is pervasive. Commonality of idiosyncratic tail risk exists for firms with different sizes, values, and industries. Their common factor, CITR estimated as the average cross-sectional idiosyncratic tail risk, explains 42.6% of the time variation in firm-level tail risk. This synchronization of idiosyncratic tail risks is robust across

various specifications and does not arise from omitted factors, since the factor model residuals are virtually uncorrelated. Consistent with my explanation, the CITR factor, defined as shocks to CITR, is procyclical as evidenced by its positive correlation to log changes in the price-to-earnings ratio, market return, gross domestic product, investment, and consumption. In addition, the CITR factor is highly correlated to existing intermediary factors, demonstrated by its strong positive correlation to the intermediary capital factor (He, Kelly, and Manela (2017)) and the broker-dealer leverage factor (Adrian, Etula, and Muir (2014)), and its negative correlation to the Leverage Constraint Tightness of mutual funds (Boguth and Simutin (2018)). Furthermore, the CITR factor is uncorrelated with shocks to market volatility, shocks to the VIX, and the common idiosyncratic volatility factor of Herskovic et al. (2016), showing that it's not driven by volatility.

The CITR factor is a systematic risk factor, since CITR is correlated to the intermediary marginal value of wealth. Consequently, exposure to this risk should be priced in equilibrium. To test that, I examine whether the CITR factor is a priced risk factor and find support for this hypothesis in stock returns. First, I sort stocks into portfolios based on their CITR-betas, which are estimated by regressing individual stock excess returns on the CITR factor, then hold the portfolios for a month. Excess returns and alphas monotonically increase in CITR-betas. Stocks in the highest CITR-beta quintile earns a 0.62% (t-stat of 2.01) higher value-weighted returns than the lowest quintile, reflecting their compensation for higher exposures to shocks to CITR.

The main question of this paper is whether high idiosyncratic tail risk firms have high exposure to the CITR factor and if that can explain their cross-sectional differences in average returns, which are confirmed by my following findings. My analysis shows that betas to the CITR factor are increasing in the idiosyncratic tail risk deciles and the relationship is nearly monotonic. Furthermore, the CITR factor alone explains 73%, up to 86% along with the market factor, of the cross-sectional variation of average excess returns for the portfolios sorted by idiosyncratic tail risk. These findings confirm my hypothesis that exposure to the CITR factor

explain most of the idiosyncratic tail risk premium. The CITR factor alone also explains nearly 32% of its cross-sectional variation in the volume tail risk deciles and adding the market factor increases the cross-sectional R^2 to 68%.

My explanation for the idiosyncratic tail risk premium is supported by a battery of robustness checks. First, the CITR factor also helps to explain cross-sectional differences in average returns of portfolios conditionally double-sorted on size and then idiosyncratic tail risk or volume tail risk. The CITR factor is also priced in anomaly portfolios independently double-sorted on size and the following characteristics: operating profitability, investment, momentum, or idiosyncratic volatility. Second, the CITR factor risk price is positive for all asset classes used in He, Kelly, and Manela (2017) and is statistically significant for the sophisticated options, CDS, commodities, and foreign exchange portfolios. Third, some intermediary models predict that the intermediary factor negatively forecasts market returns (He, Kelly, and Manela (2017)). I test this hypothesis by regressing equity market returns on CITR. A one-standard-deviation increase in CITR forecasts a decrease in annualized excess market returns of -9.56%, -7.09%, -5.31%, and -3.65% at the one-month, six-month, one-year, and two-year horizons, respectively. The results are statistic significant with Hodrick t-statistics of -2.02, -2.31, -2.36, and -2.05 and produce an R^2 of 9.83% at the annual frequency. Furthermore, I use the idiosyncratic tail risk long-short portfolio as a factor-mimicking portfolio for the non-traded CITR factor, and find that this traded CITR factor is also priced in equities and sophisticated asset classes.

Finally, in my explanation volume tail risk is a substitute for idiosyncratic tail risk, since they are both caused by large intermediary trades (Gabaix et al. (2006)). Hence, a litmus test is to evaluate whether volume tail risk exhibits commonality that is correlated to the intermediary marginal value of wealth and helps explain average returns across multiple asset classes. Consistent with my explanation, volume tail risk also exhibits a strong factor structure and common volume tail risk (CVTR), estimated as the average cross-sectional volume tail risk, explains 46.3% of the time variation in volume tail risks. The common volume tail risk factor, defined as 3-month shocks to CVTR is procyclical, correlated to intermediary factors, and

helps explain cross-sectional differences in average returns for anomalies and sophisticated assets. The CITR and CVTR factors are highly correlated and have similar prices of risk, supporting the theory that they are driven by shocks to intermediary funding. This paper is the first to show duality in the *tail distributions* of idiosyncratic returns and trading volume in empirical asset pricing, providing strong support for the large intermediary hypothesis of tail risk and my explanation of their risk premia.

My paper is related to several strands of literature. My power-law measure of idiosyncratic tail risk is closest to the seminal work of Bollerslev and Todorov (2011a), which measures the total tail risk (tail risk of returns) from high-frequency returns. My measure complements their work by measuring idiosyncratic tail risk after removing common factors. Danielsson and De Vries (1997) use power-law to estimate the tail distribution of high-frequency foreign exchange data. Kelly and Jiang (2014b) use power-law to measure market tail risk. Bollerslev and Todorov (2011b), Bollerslev, Todorov, and Li (2013), and Bollerslev and Todorov (2014b) use power-law to measure tail risk implied from options prices. Qin and Todorov (2019) study the asymmetry of power-law measures of idiosyncratic tail risk using high-frequency returns and options, but focusing on the latter. Van Oordt and Zhou (2016) uses power-law theory to estimate exposures to systematic tail risk. My paper complements these studies by measuring the realized idiosyncratic tail risk estimated using high-frequency equity returns. My model allows removal of multiple factors and provides a much larger sample of stocks. Additionally, I theoretically demonstrate the time-aggregation properties of my idiosyncratic tail risk measure, which justifies extrapolating the high-frequency idiosyncratic tail risk measure to longer time horizons.

Next, my research on the idiosyncratic tail risk premium is motivated by the recent literature on *idiosyncratic* tail risk and returns. Begin, Dorion, and Gauthier (2019) shows that idiosyncratic jump risk explains 28% of the variation in risk premium on a stock. Kapadia and Zekhnini (2019) show ex-ante jump probabilities predict cross-sectional average returns. Pederzoli (2018) shows that idiosyncratic skewness is priced in individual stocks. Kelly, Lustig,

and Van Nieuwerburgh (2016) show that for firms in the financial sector, idiosyncratic risk had a higher price than sector risk during the financial crisis. Long, Jiang, and Zhu (2018) study idiosyncratic tail risk in Chinese stock markets and finds that it negatively predicts stock returns. My paper complement this literature by demonstrating that idiosyncratic tail risk estimated using high-frequency returns predicts cross-sectional average returns. Furthermore, I demonstrate that volume tail risk measured from high-frequency trading volume also predicts cross-sectional average returns. Finally, I show both the idiosyncratic tail risk and volume tail risk premia are highly persistent and can be explained by exposures to a common factor.

In addition, my paper is related to the literature on *total* tail risk and returns. Savor (2012) finds momentum after large absolute returns with information and reversals in the absence of information. Jiang and Zhu (2017) find that markets under-react to jumps interpreted as information shocks and Jiang and Yao (2013) argue that jumps are due to new information and not systematic shocks. My paper complements these studies by measuring *idiosyncratic* tail risk with systematic factors removed. Additionally, I demonstrate that idiosyncratic tail risk has a strong factor structure, which provides evidence against their interpretation that tail risk is only caused by firm-specific news and that the premium is due to market under-reaction.

Furthermore, my research on the commonality of idiosyncratic tail risk is connected to the literature on cross-sectional studies of firm-level risk. Campbell et al. (2001) show between 1962 to 1997 that firm-level volatility increased relative to market volatility. Herskovic et al. (2016) find that idiosyncratic volatility follows a strong factor structure, where their common idiosyncratic volatility factor is priced and related to labor income risk. Dew-Becker and Giglio (2020) study the cross-section of implied volatility and find a strong factor structure. Begin, Dorion, and Gauthier (2019) find a high degree of comovement in idiosyncratic tail risk measured from options and daily returns. Lin and Todorov (2019) study the asymmetry between positive and negative idiosyncratic tail risk from options and find their asymmetry measure predicts the market risk premium. My paper compliments this literature in several ways. First, I study the *realized* idiosyncratic tail risk measured from high-frequency returns and show that

idiosyncratic tail risk follows a strong factor structure. Second, I show that CITR is correlated to the intermediary marginal utility wealth and is procyclical. Finally, my paper studies the cross-sectional asset pricing implications of the CITR factor, showing exposures to this factor earn higher average returns and that the factor explains the idiosyncratic tail risk premium.

Finally, my study is similar in spirit to the recent work on intermediary asset pricing including Adrian and Shin (2010), Adrian, Etula, and Muir (2014), Brunnermeier and Pedersen (2009), Brunnermeier and Sannikov (2014), Boguth and Simutin (2018), Fontaine and Garcia (2012), Gromb and Vayanos (2002), He and Krishnamurthy (2013), He, Kelly, and Manela (2017). My paper complements this literature by showing that intermediaries also drive the tail distribution of returns and trading volume and that common idiosyncratic tail risk and common volume tail risk are both correlated to the intermediary marginal value of wealth. I also demonstrate that the common idiosyncratic tail risk and common volume tail risk factors are priced in many sophisticated asset classes, which further supports the hypothesis that intermediaries are the marginal investor in these assets.

The paper is organized as follows: Section 4.1 provides my explanation for the idiosyncratic tail risk premium. Section 4.2 presents the econometric model, idiosyncratic tail risk measure, time-aggregation properties, robustness to microstructure noise, and estimation of idiosyncratic tail risk. Section 4.3 describes the data and factors, then documents the idiosyncratic tail risk and volume tail risk premia, persistence of the tail risk premia, and firm-level evidence that idiosyncratic tail risk is driven by large intermediary trades. Section 4.4 explores the factor structure in idiosyncratic tail risk, defines the CITR factor and shows that it's priced and explains the idiosyncratic tail risk premium. Section 4.5 demonstrates robustness of the results by showing that the CITR factor is priced in anomaly portfolios and in sophisticated assets, the common volume tail risk factor has similar pricing abilities, the traded idiosyncratic tail risk factor has similar pricing abilities, and CITR predicts market returns. Section 4.6 concludes.

4.1 Explanation for the Idiosyncratic Tail Risk Premium

In this section, I describe my explanation for the idiosyncratic tail risk premium using a highly stylized model, where intermediary funding drives the comovement in idiosyncratic tail risk. Motivated by Gabaix et al. (2006), I hypothesize that intermediaries' large trades, either due to fundamental news or proprietary analysis (absent news), causes idiosyncratic tail risk. The assumptions of this model are, in turn, motivated by existing empirical evidence on intermediaries and the price impact due to their large trades.

Empirically, intermediaries hold the majority of equities, approximately 78.5% in 2007 (French (2008)). Most equities in those institutions are actively managed; the fraction of actively managed mutual funds is 83% in 2012 and the fraction of actively managed institutional owned equities is 59% (Stambaugh (2014)). Intermediaries conduct large trades based on investment opportunities from news or proprietary analysis, executing 86% of large block trades with 10,000 shares or more (Griffin, Harris, and Topaloglu (2003)). While they moderate trading volume to mitigate the price impact, the desired trading volume will still cause a large temporary price impact, particularly in small and medium sized stocks. The average price pressure (i.e. temporary price impact) is 0.49% for a one standard deviation inventory change and the average price pressure is 1.18% for the smallest quintile of stocks (Hendershott and Menkveld (2014)). These large price impacts cause the tail distribution of stock returns to become heavier (i.e. more extreme returns).

Intermediaries require funding to trade on investment opportunities (Brunnermeier and Pedersen (2009)). Intuitively, as funding constraints tighten, intermediaries forgo trading on some investment opportunities (especially in riskier investments in small and medium sized stocks (Naes, Skjeltorp, and Odegaard (2011)), and subsequently idiosyncratic tail risk *decreases* on average. Hence, the common component of each stock's idiosyncratic tail risk is a decreasing function of the shadow cost of capital in Brunnermeier and Pedersen (2009), which is a measure of the tightness of intermediary funding constraints. A higher shadow cost of capital means low intermediary funding, indicating that investment opportunities are higher and

therefore idiosyncratic tail risks are lower across stocks due to fewer large intermediary trades. My explanation is consistent with intermediaries deleveraging when their funding constraints tighten as demonstrated by Adrian, Etula, and Muir (2014), impairing their ability to conduct large trades. Consequently, the common factor of the idiosyncratic tail risks is correlated to their leverage factor.

4.1.1 Stylized Model of Idiosyncratic Tail Risk

Denote the idiosyncratic tail risk of firm $l = 1...L$, as $\xi_{l,1}$. Each firm's time-one idiosyncratic tail risk is driven by intermediary funding and firm-specific news $N_{l,1}$, and equals

$$\xi_{l,1} = b - c\phi_1 + N_{l,1}, \tag{4.1}$$

where $N_{l,1}$ is i.i.d., with mean μ_N, variance σ_N^2, independent of ϕ_1, $b, c > 0$ are constants, and ϕ_1 is the time-one intermediary shadow cost of capital from Brunnermeier and Pedersen (2009). Taking an average over a large number of firms yields approximation

$$CITR_1 \equiv \frac{1}{L}\sum_{l=1}^{L} \xi_{l,1} \approx b + \mu_N - c\phi_1, \tag{4.2}$$

where $CITR_1$ is the common idiosyncratic tail risk. $CITR_1$ is a negative function of the intermediary shadow cost of capital, since

$$\phi_1 \approx \frac{b + \mu_N - CITR_1}{c}, \tag{4.3}$$

such that a lower $CITR_1$ corresponds to tighter funding constraints. Also, each firm's idiosyncratic tail risk comoves with $CITR_1$, since substituting (4.3) into (4.1) yields approximate factor structure

$$\xi_{l,1} \approx -\mu_N + CITR_1 + N_{l,1}, \tag{4.4}$$

implying that $Cov_0(CITR_1, \xi_{l,1}) > 0$ for all firms. Brunnermeier and Pedersen (2009) demonstrate how funding liquidity enters the pricing kernel even when investors are risk-neutral. Let W_1 be time-one wealth such that the risk-neutral investor maximizes $E_0[\phi_1 W_1]$. Then, the stochastic discount factor is $\phi_1/E_0[\phi_1]$ and asset k's time-zero expected excess return $R^e_{k,1}$ is

$$E_0[R^e_{k,1}] = -\frac{Cov_0(\phi_1, R^e_{k,1})}{E_0[\phi_1]}. \tag{4.5}$$

Equation (4.5) states that stocks with a negative covariance term have higher expected excess returns, since the stock has a low payoff during future funding liquidity crises when ϕ_1 is high. Substituting Equation (4.3) into Equation (4.5) yields approximation

$$E_0[R^e_{k,1}] \approx \lambda_{CITR} Cov_0(CITR_1, R^e_{k,1}), \tag{4.6}$$

where price of risk $\lambda_{CITR} > 0$. Thus, CITR is a priced risk factor and assets that covary with CITR are risky and earn a higher risk premium. A risk-based explanation of the idiosyncratic tail risk premium requires a further assumption that exposure to CITR is increasing in idiosyncratic tail risk, that is if $\xi_{n,1} > \xi_{m,1}$, then

$$Cov_0(CITR_1, R^e_{n,1}) > Cov_0(CITR_1, R^e_{m,1}), \tag{4.7}$$

and consequently $E_0[R^e_{n,1}] > E_0[R^e_{m,1}]$ by (4.6). One possible economic mechanism for Assumption (4.7) is flight-to-quality. When funding is tight, intermediaries decrease ownership in small and medium stocks (Naes, Skjeltorp, and Odegaard (2011), Papaioannou et al. (2013)) due to their higher risk and increased margins on leverage. Since small and medium stocks have high idiosyncratic tail risks, their exposure to intermediary funding is high, increasing their riskiness.

My explanation relies on two central assumptions. First, the common component of idiosyncratic tail risk is correlated to the intermediary marginal value of wealth, which implies CITR is a priced risk factor. Second, exposure to CITR is increasing in idiosyncratic tail risk.

Large exposures to CITR for firms with high idiosyncratic tail risk can explain their premium. These simple assumptions have several testable implications, listed below, that allow me to distinguish my theory from existing theories of what causes idiosyncratic tail risk and its premium.

a) CITR is correlated to the intermediary marginal value of wealth and can be estimated as the cross-sectional mean of idiosyncratic tail risks. Since intermediary funding is pro-cyclical, then CITR should be as well. This is the opposite prediction of several existing explanations of what drives the commonality of idiosyncratic tail risk like fire-sales and labour tail risk, which predict counter-cyclical comovement. Also, if idiosyncratic tail risk were only caused by firm-specific news, there should be no commonality.

b) CITR is a risk factor with a positive price. Stocks with higher exposures to CITR should earn higher average returns. The prediction that CITR has a positive price of risk distinguishes the model from existing possibilities like fire-sales and labour tail risk, which predict a negative price of risk, and firm-specific news which should have no price of risk. Additionally, CITR's correlation to the intermediary marginal value of wealth means it should negatively forecast market returns as predicted by some intermediary models (i.e. He, Kelly, and Manela (2017)).

c) The idiosyncratic tail risk premium is explained by exposures to CITR. Firms with high idiosyncratic tail risk also have high betas to CITR and are compensated with high returns. This is the main test that distinguishes my explanation from the inability of investors to diversify or under-reaction to news, which should be firm-specific.

d) Firms with higher idiosyncratic tail risk should continue to enjoy higher future expected returns. If the idiosyncratic tail risk is driven by intermediary trading, then it should be persistent, since intermediaries can frequently trade these stocks according to their proprietary models. If the idiosyncratic tail risk premium is a systematic risk premium, then firms with higher idiosyncratic tail risks are riskier due to their higher exposures

to the CITR factor. If the relationship between idiosyncratic tail risk and CITR do not change rapidly, then this risk premium should be persistent. This test allows me to distinguish my explanation from firm-specific explanations like under-reaction to fundamental news-shocks, which should be short-lived.

A litmus test for my explanation uses the relationship between the tail distribution of returns and trading volume predicted by Gabaix et al. (2006). Under their model, the tail distribution of returns and volume are proportional, meaning volume tail risk is a proxy for idiosyncratic tail risk. This implies that trading volume should follow the same predictions as above, where volume tail risk earns a premium, exhibits commonality, its common factor is driven by intermediary funding, is a priced risk factor, and explains the volume tail risk premium.

This paper uses a power-law measure of idiosyncratic tail risk, which is motivated by the power-law economic model in Gabaix et al. (2006). Power-law conveniently summarizes the tail distribution using a single parameter corresponding to the idiosyncratic tail risk ξ_l. Additionally, tail risk under power-law is preserved under time aggregation, which justifies measuring tail risk using high-frequency returns resulting in a time-varying measure of idiosyncratic tail risk. Section C.1 presents empirical evidence that idiosyncratic returns and trading volume have power-law distributed tails.

4.2 Econometric Framework

4.2.1 Idiosyncratic Tail Risk Measure

In each month t, I observe a large series of high-frequency observations $i = 1, .., N$, where N is the number of observations. For example, in a month with 21 trading days and 5-minute time intervals, $N = 1659$.[3] The high-frequency log return during month t for firm $l = 1, .., L$ is

$$r_{l,t,i} = p_{l,t-1+\frac{i}{N}} - p_{l,t-1+\frac{i-1}{N}}, \tag{4.8}$$

[3]In a trading day, there are 78 intraday returns when sampled in 5-minute intervals and 1 overnight return.

where p is the natural log of price. I assume return dynamics follow a standard factor model

$$r_{l,t,i} = \boldsymbol{\beta}_{l,t}^{\intercal} \mathbf{f}_{t,i} + x_{l,t,i}, \tag{4.9}$$

where the $\boldsymbol{\beta}_{l,t}$ are the factor loadings of firm l on systematic factors $\mathbf{f}_{t,i}$, and $x_{l,t,i}$ are unobservable idiosyncratic returns. The factors are assumed to be observable and log returns of equity portfolios, for example in the CAPM $\mathbf{f}_{t,i}$ is the high-frequency return of the market portfolio. $\boldsymbol{\beta}_{l,t}$ is assumed to be constant throughout each month t, but can vary month-to-month. The constant $\boldsymbol{\beta}_{l,t}$ assumption is common in the literature (e.g., Todorov and Bollerslev (2010), Dai, Lu, and Xiu (2019)). Furthermore, I assume that the tails of idiosyncratic returns are power-law distributed.

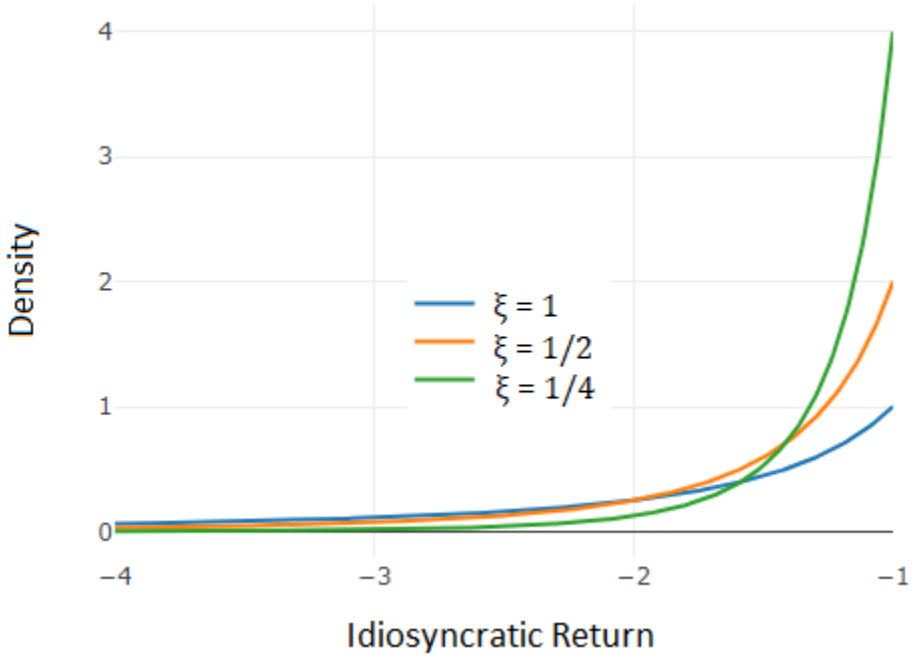
Assumption 1 *(a) For each firm l, the idiosyncratic returns $x_{l,t,i}$, $i = 1, ..., N$, are independent and identically distributed, and are regularly varying with tail risk parameter $\xi_{l,t} > 0$. That is, $P(|x_{l,t,i}| > y) = y^{-\frac{1}{\xi_{l,t}}} L(y)$, where L is a slowly varying function.*[4]

(b) $\lim_{y \to \infty} \frac{P(x_{l,t,i} \leq -y)}{P(|x_{l,t,i}| > y)} = \theta \in (0, 1]$

Assumption 1(a) postulates that the tails of idiosyncratic returns are power-law distributed with tail risk parameter $\xi_{l,t}$, which is supported by empirical evidence presented in Figure C.2 in Section C.1. Tail risk parameter $\xi_{l,t}$ is firm-specific, dynamic, and able to capture monthly changes in idiosyncratic tail risk. The tail risk parameter $\xi_{l,t}$ is inversely proportional to the slope of the tails. Figure 4.1 illustrates the relationship between tail risk parameter $\xi_{l,t}$ and the slope of the left tail. A large tail risk parameter corresponds to a small slope and a high probability of extreme observations. Hence, $\xi_{l,t}$ is an intuitive measure of idiosyncratic tail risk. Assumption 1(b) is a standard regularity condition that ensures the right tail of idiosyncratic returns does not dominate the left tail. Assumption 1(b) can be economically interpreted as the probability of bankruptcy is non-zero for any firm. For simplicity, I assume that idiosyncratic

[4]Function L is slowly varying if L is strictly positive and $\lim_{x \to \infty} L(tx)/L(x) = 1$ for all $t > 0$. Prototypical examples include $L(x) = log(x)$ and $L(x) = c$ for a $c > 0$.

Figure 4.1: Probability Density of Power-Law Distributed Tails



The figure plots the probability density of Pareto distributed tails with various tail risk parameters ξ and L(y)=1. Pareto tails become heavier as the tail risk parameter ξ increases. A larger ξ has a smaller slope, which increases the severity of extreme negative returns.

returns are independent and identically distributed within each month. The i.i.d. assumption is stronger than necessary for the theoretical results and estimation methods in my paper, which can still hold for idiosyncratic returns that are dependent and heterogeneous.[5]

The strict positivity of $\xi_{l,t}$ rules out light tails that understate the probability of rare events, such as normally distributed tails. Most parametric models of financial returns are power-law distributed. For example, a Student's t distribution with d degrees of freedom satisfies Assumption 1 with tail risk $\xi_{l,t} = 1/d$. Other prototypical examples of power-law distributions are the Pareto, Levy-Stable, and Cauchy distributions (see Embrechts, Kluppelberg, and Mikosch (1997) for more examples). Assumption 1 is semi-parametric and agnostic on the body (region of the distribution not belonging to the tail) of idiosyncratic returns, allowing for skewness and

[5]See Hill (2010) for tail risk estimation using dependent and heterogeneous data.

excess kurtosis commonly exhibited by returns.[6]

Factor models of equity returns are commonly used in financial economics due to their connection to investor preferences. In the absence of arbitrage, expected stock returns can be expressed as a function of the stochastic discount factor (SDF), where in equilibrium the stochastic discount factor is the investor's marginal utility. As a result, systematic factors correlated with the stochastic discount factor determine expected returns.[7] For example, in the Capital Asset Pricing model (CAPM) by Sharpe (1964) and Lintner (1965), the market portfolio is the single factor determining expected returns. Models with multiple factors include the Fama and French (1993) and Carhart (1997) four factor model, the Fama and French (2015) five factor model, and statistical factor models. However, most factor models of returns assume the idiosyncratic returns are light-tailed. Light-tailed idiosyncratic returns understate the probability of rare events and fitting a light-tailed model on power-law distributed data causes severe econometric issues (Balkema and Embrechts (2018)).

The model assumes *high-frequency* idiosyncratic returns have power-law distributed tails. High-frequency returns have more observations than daily, which should improve the precision of tail risk estimates. However, this gain in precision assumes that the idiosyncratic tail risk is the same at longer time-intervals (e.g. daily, monthly). Assumption 1 describes the tail behaviour of high frequency idiosyncratic returns. This is different from many factor models and risk measures that are measured at longer time horizons. The next section demonstrates how the high-frequency idiosyncratic tail risk $\xi_{l,t}$ maps to longer time horizons.

4.2.2 Temporal Aggregation of Idiosyncratic Tail Risk

In this section, I theoretically demonstrate that within each month, longer-horizon (i.e. daily, weekly, or monthly) idiosyncratic returns inherit the tail risk of the underlying high-frequency returns, which I refer to as the inheritance property under time-aggregation. Heuristically, a

[6] While I assume the left and right tails have the same tail risk parameter ξ, the body of the distribution can still exhibit skewness. For example, a Levy-Stable distribution has a tail parameter and a separate skewness parameter.

[7] For a further explanation, see Chapter 6 of Cochrane (2009).

property of power-law distributed random variables is that their sum inherits the tail risk of the underlying random variables. For example, the sum of two t-distributed random variables with tail risk ξ will also have tail risk ξ, despite not being t-distributed. This section uses extreme value theory to formalize these heuristic arguments. I show that the inheritance property under time-aggregation is theoretically guaranteed even with a finite number of observed returns. Additionally, the measure is robust to certain microstructure noise. To simplify the notation in this section, I drop the firm l superscript with the understanding that all idiosyncratic returns $x_{t,i}$ and idiosyncratic tail risk parameters ξ_t are firm-specific.

I prove that idiosyncratic tail risk is inherited under time-aggregation using progressively weaker assumptions. First, I show that the monthly idiosyncratic returns weakly converge to a Levy distribution with the same tail risk as the high-frequency returns, when returns are continuously observed in a month. Second, I show that when the number of returns is finite and the distribution of the monthly idiosyncratic return is unknown, the inheritance property under time-aggregation still holds. Finally, I show the high-frequency idiosyncratic tail risk can also be measured when certain microstructure noise exists, which justifies using my measure of idiosyncratic tail risk even when returns are contaminated with light-tailed noise such as the bid-ask bounce.

Monthly Idiosyncratic Tail Risk

The literature overwhelmingly specifies factor models at the monthly time horizon, so this section uses only the definition of log returns to show the connection between high-frequency and monthly idiosyncratic returns. To do so, I first construct a monthly factor model defined analogously to the high-frequency model. By definition, monthly log returns are $R_t = p_t - p_{t-1}$, where p_t is the log price at month t. I then assume the monthly factor model of each firm follows

$$R_t = \boldsymbol{\beta}_t^\intercal \mathbf{F}_t + X_t, \tag{4.10}$$

where returns R_t, factors $\mathbf{F}_t$, and idiosyncratic returns X_t are *monthly* log returns. By definition, the observable returns and factors can be written as the sum of their underlying high-frequency returns by using the aggregation properties of log returns, such that $R_t = \sum_{i=1}^{N} r_{t,i}$ and $F_t = \sum_{i=1}^{N} f_{t,i}$. Since $\boldsymbol{\beta}_t$ are constant for each month t, $\boldsymbol{\beta}_t$ is assumed to be the same in the monthly and high-frequency models. Furthermore, I impose no direct assumptions on the distribution of the unobservable monthly idiosyncratic returns X_t. Using only the definition of the models and log returns, the following Lemma shows the monthly idiosyncratic return can be expressed as the sum of the underlying high-frequency idiosyncratic returns.

Lemma 4.2.1 *The monthly idiosyncratic return is equal to the sum of high-frequency idiosyncratic returns, that's* $X_t = \sum_{i=1}^{N} x_{t,i}$.

Proof See Section C.2.1 in the appendix.

Lemma 4.2.1 shows the identity that the monthly idiosyncratic return is equal to the sum of high-frequency idiosyncratic returns. However, the distribution of the monthly idiosyncratic return is unknown. In general, the distribution of $x_{t,i}$ will be different from the distribution of $X_t = \sum_{i=1}^{N} x_{t,i}$. For example, the sum of two t-distributed random variables is generally not t-distributed. The monthly idiosyncratic return $X_t = \sum_{i=1}^{N} x_{t,i}$ has the same distribution as $x_{t,i}$ if and only if $x_{t,i}$ is a stable distribution.[8] However, the monthly idiosyncratic tail risk inherits the high-frequency idiosyncratic tail risk ξ_t, which is proved in the next section.

Idiosyncratic Tail Risk with Continuously Observed Returns

I first assume that high-frequency returns are continuously observed, such that $N \to \infty$. As the number of returns approach infinity, the limiting distribution of the sum of returns can be derived using Levy's Theorem. Levy's Theorem in Section C.2.2 is the generalization of the Central Limit Theorem for heavy-tailed random variables. The theorem states that the sum of

[8] A distribution is stable if a sum two i.i.d. random variables with this distribution has the same distribution. The normal distribution is the only stable distribution to be light-tailed, while all other stable distributions are heavy-tailed.

i.i.d. power-law distributed random variables converge to a Levy distribution with the same tail risk. I use Levy's Theorem to derive the limiting distribution and tail risk of the monthly idiosyncratic returns in the following theorem.

Theorem 4.2.2 *Suppose the high-frequency idiosyncratic returns $x_{t,i}$, $i = 1, .., N$, satisfy Assumption 1 and $\xi_t > \frac{1}{2}$. Then there exist a_N and b_N such that the monthly idiosyncratic return X_t satisfies*

$$(\sum_{i=1}^{N} x_{t,i} - b_N)/a_N \xrightarrow{d} u_t \text{ as } N \to \infty, \tag{4.11}$$

where u_t is a Levy-distributed random variable with tail risk ξ_t, and $a_N = inf\{y : P(|x_{t,1}| > y) \leq N^{-1}\}$, and $b_N = NE(x_{t,1}\mathbb{1}_{(x_{t,1} \leq a_N)})$.

Proof See Section C.2.3 in the appendix.

If the high-frequency returns are continuously observed, then the distribution and tail risk of the monthly idiosyncratic returns is derived in Theorem 4.2.2. In practice, microstructure effects prevent returns from being sampled continuously, since microstructure noise dominates the returns process at ultra high frequencies. Therefore, returns must be sampled discretely, where the number of returns N is finite. When N is finite, monthly idiosyncratic returns are no longer Levy distributed and in general the distribution cannot be found. However, the next section shows that the monthly idiosyncratic *tail risk* can still be measured.

Idiosyncratic Tail Risk with a Finite Number of Returns

The following theorem is the main theoretical result for the tail risk of the monthly idiosyncratic return when the number of returns observed in a month is finite.

Theorem 4.2.3 *Suppose the high-frequency idiosyncratic returns $x_{t,i}$, $i = 1, .., N$, satisfy Assumption 1 and $N \geq 2$. Then,*

$$P(X_t \leq -y) \sim NP(x_{t,i} \leq -y), \tag{4.12}$$

where $f(y) \sim g(y)$ *means* $\lim_{y\to\infty} f(y)/g(y) \to 1$. *This implies the left tail of the monthly idiosyncratic return* X_t *is power-law distributed with tail risk parameter* ξ_t.

Proof See Section C.2.5 in the appendix.

The above implies that when the number of returns N is finite and the distribution of the monthly idiosyncratic return X_t is unknown, the left tail of X_t still inherits the tail risk of the high-frequency idiosyncratic returns ξ_t under very general power-law assumptions.

In practice, microstructure effects introduce an unobservable noise process to returns. Separating the noise from the idiosyncratic returns is difficult, since both processes are unobservable. However, the next section shows that the high-frequency idiosyncratic tail risk can still be measured under certain microstructure noise processes.

Idiosyncratic Tail Risk with Microstructure Noise

Suppose high-frequency returns are observed with microstructure noise process $\eta_{t,i}$. While papers often assume microstructure noise follows a normal or light-tailed distribution, I allow the microstructure noise to be power-law distributed in the following assumption.

Assumption 2 *(a) The microstructure noise* $\eta_{t,i}$ *are independent and identically distributed, and are regularly varying with tail risk* $\xi_t > \gamma_t > 0$. *That's* $P(|\eta_{t,i}| > y) = y^{-\frac{1}{\gamma_t}} L_\eta(y)$, *where* L_η *is a slowly varying function.*

(b) $\lim_{y\to\infty} \frac{P(\eta_{t,i} \leq -y)}{P(|\eta_{t,i}|>y)} = p \in [0, 1]$

(c) $\eta_{t,i}$ *is independent of idiosyncratic returns* $x_{t,i}$ *and systematic factors* $\boldsymbol{f}_{t,i}$.

The key condition is that $\xi_t > \gamma_t$, meaning the tails of the idiosyncratic returns are heavier than the noise. If the condition is violated, the idiosyncratic tail risk cannot be measured. Next, I

assume high-frequency returns are contaminated with additive microstructure noise, such that the returns contaminated with noise follow

$$\begin{aligned} r^*_{t,i} &= \boldsymbol{\beta}^\intercal_{l,t}\mathbf{f}_{t,i} + x^*_{t,i}, \\ x^*_{t,i} &= x_{t,i} + \eta_{t,i}, \end{aligned} \tag{4.13}$$

where the $x^*_{t,i}$ are unobservable noisy idiosyncratic returns. Additive noise is by far the most common assumption in the microstructure literature.[9] In general, the distribution of the idiosyncratic returns contaminated with noise depend on the distribution of the microstructure noise. For example the volatility of $x^*_{t,i}$ is a function of the volatility of $x_{t,i}$ and $\eta_{t,i}$. However, if the tails of the microstructure noise $\eta_{t,i}$ are lighter than the tails of the idiosyncratic returns $x_{t,i}$, i.e. when $\xi_t > \gamma_t$, then the idiosyncratic returns contaminated with noise $x^*_{t,i}$ only inherit the tail risk of the idiosyncratic returns $x_{t,i}$ without the noise, which is demonstrated in the following theorem.

Theorem 4.2.4 *Under Assumption 1 for idiosyncratic returns $x_{t,i}$, Assumption 2 for noise $\eta_{t,i}$, and additive microstructure noise* (4.13)*, then $\forall i$,*

$$P(x^*_{t,i} \leq -y) \sim P(x_{t,i} \leq -y), \tag{4.14}$$

where the $x^*_{t,i}$ are idiosyncratic returns contaminated with noise defined in Equation (4.13)*.*

Proof See Section C.2.6 in the appendix.

Since microstructure noise such as the bid-ask bounce (Roll (1984)) have lighter tails than returns, Theorem (4.2.4) indicates that the bid-ask bounce will not affect idiosyncratic tail risk measurement. However, γ_t is larger than ξ_t during rare microstructure events such as the 2013 Flash Crash, meaning the idiosyncratic tail risk cannot be measured during those events.

[9]The theory also holds for other noise processes, such as multiplicative microstructure noise.

4.2.3 Volume Tail Risk Measure

In Gabaix et al. (2006), tail returns are caused by the price impact of large intermediaries trading an abnormally large volume of shares, which implies the tails of idiosyncratic returns and trading volume are power-law distributed and proportional. In this section, I define a power-law measure of the right tail of trading volume, which is supported by the empirical evidence in Figure C.3 of Section C.1. Let $s_{t,i}$ denote the total trading volume between time interval i and $i-1$. Since trading volume has intraday seasonality, I use daily differences and measuring changes in trading volume $v_{t,i} = s_{t,i} - s_{t,i-d}$, where d denotes the number of intraday observations in a day. Next, I assume that changes in trading volume are power-law distributed with tail risk parameter ν_t.

Assumption 3 *(a) Changes in trading volume $v_{t,i}$, $i = 1, ..., N$, are independent and identically distributed, and regularly varying with volume tail risk parameter $\nu_t > 0$. That is, $P(v_{t,i} > y) = y^{-\frac{1}{\nu_t}} L_{\nu_t}(y)$, where L_ν is a slowly varying function.*

(b) $\lim_{y\to\infty} \frac{P(v_{t,i}\leq -y)}{P(|v_{t,i}|>y)} = p \in [0,1]$

I define volume tail risk (VTR) as the tail risk parameter ν_t of changes in trading volume $v_{t,i}$. Gabaix et al. (2006) predict that large intermediary trades cause idiosyncratic tail risk and volume tail risk to be related according to $\xi_t \sim \rho\nu_t$ for some price impact measure $0 \leq \rho \leq 1$. This theoretical relationship can be tested by substituting volume tail risk in the place of idiosyncratic tail risk in the empirical asset pricing tests. If ITR and VTR are proportional, they should produce similar asset pricing results.

4.2.4 Estimation of Idiosyncratic Tail Risk and Volume Tail Risk

The monthly idiosyncratic tail risk ξ_t for each firm is the main parameter of interest. Section 4.2.2 demonstrates that the monthly idiosyncratic tail risk inherits the tail risk of high-frequency idiosyncratic returns. However, idiosyncratic returns $x_{t,i}$ are unobservable and must

be estimated. The factor model's filtered residuals are frequently used as the estimator of idiosyncratic returns.[10] Filtered residuals depend on the estimates of the factor betas, hence this section also discusses estimation of betas in heavy-tailed regressions. The asset pricing literature overwhelmingly uses ordinary least squares (OLS) to estimate factor model betas. When $\xi_t \leq 1/2$, the idiosyncratic returns have finite variance, and the Gauss-Markov Theorem proves the OLS estimate has the minimum variance of all linear unbiased estimators. However, when $\xi_t > 1/2$, the idiosyncratic returns have infinite variance and OLS is no longer efficient.[11]

Least Absolute Deviations (LAD) is a robust alternative to OLS when the errors have heavy tails, in the sense that LAD is generally more efficient than OLS when the regression errors have infinite variance.[12] Blattberg and Sargent (1971) show that the LAD estimator is the maximum likelihood estimate when the i.i.d. regression errors have infinite variance with a Laplace distribution (two-tailed exponential distribution). It is well known that the LAD estimator is consistent and asymptomatically normal when the errors have a distribution function that is differentiable at 0 with the derivative positive (See Koenker and Bassett (1978)). Knight (1998) shows that the LAD estimate is asymptotically normal under more general regularity conditions. To find the limiting distribution of the LAD-estimated betas $\hat{\boldsymbol{\beta}}_t$, requires the following assumption.

Assumption 4 *(a) The returns and factors* $(r_{t,i}, \boldsymbol{f}_{t,i}^{\intercal})^{\intercal}$ *are independently and identically distributed across i.*

(b) The factors have bounded second moment, i.e., $E[\|\boldsymbol{f}_{t,i}\|^2] < \infty$.

(c) Idiosyncratic returns $x_{t,i}$ *are continuously distributed given systematic factors* $\boldsymbol{f}_{t,i}$*, with*

[10] For example, see Ang et al. (2006), Kapadia and Zekhnini (2019).

[11] Davis and Wu (1997) show that when $\xi_t > 1/2$, the least square estimate converges weakly to a ratio of stable distributions with heavy tails. Mikosch and Vries (2013) derive explicit finite sample expressions for the tails of the OLS estimate. They show the OLS estimate is heavy-tailed in finite samples. OLS estimates that are heavy-tailed can have large errors. Balkema and Embrechts (2018) perform a simulation study and demonstrate the shortcomings of OLS and other estimators when the errors are heavy-tailed.

[12] This may not be true if the regressors have infinite variance, as discussed in Balkema and Embrechts (2018). While I find the factors are heavy-tailed, there is no evidence to suggest they have infinite variance. Additionally, the results are quantitatively similar when the factors are estimated using the weighted-LAD from Ling (2005), which is robust to infinite variance regressors.

conditional density $g(x_{t,i}|\boldsymbol{f}_{t,i})$, and with median zero conditional on the systematic factors, i.e., $\int_{-\infty}^{0} g(\lambda|\boldsymbol{f}_{t,i})d\lambda = \frac{1}{2}$.

(d) The systematic factors and idiosyncratic return density satisfy a "local identification" condition, meaning the matrix $C = E[g(0|\boldsymbol{f}_{t,i})\boldsymbol{f}_{t,i}\boldsymbol{f}_{t,i}^{\intercal}]$ is positive definite.

From Assumption 4, the LAD estimates of the betas $\hat{\boldsymbol{\beta}}_t$ converge to in distribution to a multivariate normal distribution with covariance matrix $\frac{1}{4}C^{-1}E[\mathbf{f}_{t,i}\mathbf{f}_{t,i}^{\intercal}]C^{-1}$ (Powell (1991)).[13] Estimates of idiosyncratic returns are then the LAD filtered residuals

$$\hat{x}_{t,i} = r_{t,i} - \hat{\boldsymbol{\beta}}_t \mathbf{f}_{t,i}. \tag{4.15}$$

Following power-law literature, I use the Hill (1975b) method to estimate the idiosyncratic tail risk ξ_t. When a distribution's tail is exactly Pareto (i.e. when $L(x)$ is a constant) the Hill estimate is the maximum likelihood estimate of ξ_t. If the idiosyncratic returns are observable, the Hill estimate is

$$\hat{\xi}_t^{Hill} = \frac{1}{K_t}\sum_{k=1}^{K_t} log\frac{x_{t,(k)}}{x_{t,(K_t+1)}}, \tag{4.16}$$

where $x_{t,(k)}$ is the k-th order statistic of high-frequency idiosyncratic returns ($x_{t,(1)} \leq x_{t,(2)} \leq ... \leq x_{t,(N)}$) and K_t is the total number of returns below threshold $x_{t,(K_t+1)}$ in month t. The Hill estimator only uses idiosyncratic returns less than negative threshold $x_{t,(K_t+1)}$, disregarding returns that do not belong to the left tail. The threshold is chosen to be sufficiently negative so that the estimator only uses returns belonging to the left tail. Panel (a) of Figure C.2 shows idiosyncratic returns greater than 2 standard deviations (approximately the 0.95 quantile of absolute idiosyncratic returns) are power-law distributed. Hence, a natural value for threshold $x_{t,(K_t+1)}$ is the 0.05 quantile of the stock's idiosyncratic returns in month t.[14] When $K_t \to \infty$ and

[13] LAD estimation is only used in the creation of the idiosyncratic tail risk estimates and not in the asset pricing sections. All the regressions in the asset pricing sections follow the standard procedure by estimating betas with OLS.

[14] Kelly and Jiang (2014b) also use the 0.05 quantile for their Hill estimator. As a robustness check, I show

$K_t/N \to 0$, the Hill estimate is consistent and asymptotically normal with variance ξ_t^2.

The Hill estimate in Equation (4.16) is infeasible, because the idiosyncratic returns are unobservable. Hill (2015) extends the Hill estimator to residuals filtered from a regression. According to Lemma 2 in Hill (2015), the Hill estimator applied to LAD filtered residuals is asymptotically normal with variance ξ_t^2. Hence, the firm's idiosyncratic tail risk estimate is

$$\hat{\xi}_t^{Hill-Res} = \frac{1}{K_t} \sum_{k=1}^{K_t} log \frac{\hat{x}_{t,(k)}(\hat{\boldsymbol{\beta}}_t)}{\hat{x}_{t,(K_t+1)}(\hat{\boldsymbol{\beta}}_t)}, \tag{4.17}$$

where $\hat{x}_{t,(K_t+1)}(\hat{\boldsymbol{\beta}}_t)$ is the k-th order statistic of the estimated idiosyncratic returns in Equation (4.15).

To estimate the volume tail risk ν_t, I use the Hill (1975b) estimator

$$\hat{\nu}_t^{Hill} = \frac{1}{M_t} \sum_{m=M_t+1}^{N} log \frac{v_{t,(m)}}{v_{t,(M_t+1)}}, \tag{4.18}$$

where $v_{t,(k)}$ is the m-th order statistic of changes in trading volume ($v_{t,(1)} \leq v_{t,(2)} \leq ... \leq v_{t,(N)}$) and M_t is the total number of $v_{t,i}$ above threshold $v_{t,(M_t+1)}$ in month t. Figure C.3 shows trading volume greater than 3 absolute deviations (approximately the 0.9 quantile of trading volume) are power-law distributed. Hence, a natural value for threshold $v_{t,(K_t+1)}$ is the 0.9 quantile of the stock's trading volume in month t. When $M_t \to \infty$ and $M_t/N \to 0$, the Hill estimate is consistent and asymptotically normal with variance ν_t^2. This section's procedures result in a monthly idiosyncratic tail risk and volume tail risk estimate for each firm.

that my empirical results are similar for the 0.025 and 0.1 quantiles. Liu and Stentoft (2020a) conducts a large simulation study to show that the 0.05 quantile is accurate for market tail risk estimation.

4.3 Empirical Firm-Level Tail Risk

4.3.1 Data

I conduct a large empirical study of high-frequency equity returns, merging the Trade and Quotes (TAQ), the Center for Research in Security Prices (CRSP), and Compustat databases. The TAQ is the primary dataset, containing intraday transactions data for all stocks on the New York Stock Exchange (NYSE), American Stock Exchange (AMEX), NASDAQ, and other U.S. regional exchanges. I use the Monthly TAQ database up to 2003 then use the Daily TAQ database from 2004 to 2016.[15] I analyze all common stocks (Share Code 10 or 11) and price above $5. I clean the TAQ according to the procedures in Barndorff-Nielsen et al. (2008), and extract second-by-second price, trade size, bid, and ask data between 9:30am - 4:00pm. The data is then aggregated into 5-minute intervals and merged with the monthly CRSP by the TAQ CUSIP key to obtain overnight returns. During each month, I keep stocks with more than 150 negative returns (roughly 10% of 5-minute returns in a month) to ensure enough left tail observations and liquidity. Section C.3.1 in Appendix C.3 provides further details on my TAQ data cleaning procedures. The empirical analysis includes 6,213 unique securities and 414,336 firm-month observations during the post decimalization sample period from January 2001 to December 2016,[16] and the least number of firms in a month is 2,616.[17]

The CRSP monthly database is used to obtain each stock's price, shares outstanding, and returns. Delisted stocks are adjusted using the delisting return from CRSP to avoid survivorship bias. Daily and monthly portfolio returns for the Fama and French and momentum factors are downloaded from Kenneth French's website. Anomaly portfolios and industry classifications are also downloaded from French's website.

[15]The monthly database reports data in second time-intervals, while the daily database reports data in milliseconds from January 2004 to July 27, 2015, microseconds to October 2016, and from November 2016 onwards.

[16]While the TAQ data is available from January 1993, I find a structural break in tail risk measures in 1997 across all stocks, caused by microstructure changes (tick size changes from 1/8 to 1/16). To avoid influence from tick size changes, the asset pricing analysis focuses on the post decimalization period.

[17]Factor betas are estimated using a 36 month rolling window, so idiosyncratic tail risk data is used from January 1998 to December 2001 to estimate the factor betas.

The Compustat annual database is used to create the Fama and French characteristics from firm fundamentals. Compustat is merged with CRSP to replicate the Fama and French portfolio sorts and determine daily constituents for each portfolio. The daily constituents are merged with the TAQ database to create the high-frequency Fama and French factors. My exact procedures to create the high-frequency factors is described in Section C.3.2 of Appendix C.3.

I replicate an extensive set of control variables that may be correlated with idiosyncratic tail risk, and have demonstrated predictive power for cross-sectional stock returns. I include characteristics from Fama and French (2015), which are market beta, size, book-to-market, operating profitability, and investment. Momentum, short-term reversal, and illiquidity are included, since they can lead to tail risk. Finally, idiosyncratic volatility, coskewness, downside beta, and extreme positive returns are included, because they are related to the distribution of returns. The control variables are created using the CRSP and Compustat databases. Section C.3.3 in Appendix C.3 describes the variables and procedures to create them.

4.3.2 High-Frequency Systematic Factors

The high-frequency factor literature generally uses two types of systematic factors: observable or latent. Observable factors are often created using portfolios that are sorted on characteristics known to explain the cross-sectional variation of expected returns (i.e. Fama and French (1993), Carhart (1997), Fama and French (2015)). In this paper, I adopt the approach of Ait-Sahalia, Kalnina, and Xiu (2020) to create high-frequency Fama and French (2015) factors using portfolios sorted on characteristics (see Section C.3.2 of Appendix C.3 for further details). The baseline model in this paper is the well known Fama and French (2015) five factor model. I also use the market and Fama-French-Carhart four factor models as robustness checks and find similar results.[18] Furthermore, I create a five factor model that uses high-frequency

[18]Using high-frequency equity factors may have been controversial pre-decimalization due to high trading costs. However, in the post-decimalization sample used in this paper, trading costs have significantly decreased so that they are no longer a barrier to forming these traded factors. Note that the observable factors used in this paper are also openly traded as Exchange Traded Funds (ETFs), which accurately track the factors due to tracking-error constraints.Examples of ETFs that track factors include the SPY/VTI market, VTV value, VB size, VFQY quality

cross-sectional moments and liquidity variables. Specifically, I calculate the high-frequency cross-sectional mean absolute deviation around the median as a robust measure of dispersion, Hinkley (1975)'s robust coefficient of asymmetry skewness at the 0.95 quantile, the power-law tail risk measure of Kelly and Jiang (2014b) using the Hill estimate at the 0.05 quantile, the mean percentage NBBO spread, and the mean dollar trading volume.

As a further robustness check, I consider an alternative latent factor model. Latent factors use statistical methods to create systematic factors. Pelger (2020) shows that the high-frequency statistical factors can be proxied by the equal-weighted market, oil, finance, and electricity portfolios. Motivated by their findings, I use their methods to construct the above market and industry factors for my sample of stocks, to create the industry statistical model.

4.3.3 The Idiosyncratic Tail Risk Premium

In this section, I verify the idiosyncratic tail risk premium by documenting that stocks with higher idiosyncratic tail risk have higher average excess returns. For each month from January 2001 until December 2016, each firm's idiosyncratic tail risk estimate is the Hill estimate of LAD residuals from the Fama and French (2015) five factor model. Stocks are sorted into decile groupings based on their idiosyncratic tail risk in that month. Decile 1 (low) contains stocks with the lowest idiosyncratic tail risk and decile 10 (high) contains stocks with the highest idiosyncratic tail risk. I form an equally-weighted portfolio and a value-weighted portfolio in each decile, and hold each portfolio for 1 month.

Panel A of Table 4.1 reports the average idiosyncratic tail risk in each decile. The high values of idiosyncratic tail risk illustrates that the returns of some stocks have infinite moments. A random variable's moment is infinite if its tail risk parameter exceeds the reciprocal of the moment. The skewness is infinite for deciles 2-10, since $\xi > 1/3$ in these deciles. The variance is infinite for deciles 9 and 10, because $\xi > 1/2$ in these deciles. All deciles have $\xi > 1/4$ indicating infinite kurtosis. These findings highlight the usefulness of power-law approximations

(investment and profitability), VFMO momentum, and VFMF multi-factor ETFs).

Table 4.1: Excess Returns and Alphas of Portfolios Sorted on Idiosyncratic Tail Risk

	1 (Low)	2	3	4	5	6	7	8	9	10 (High)	(10-1)
Panel A: Average Portfolio Idiosyncratic Tail Risk											
Idiosyncratic Tail Risk	0.32	0.36	0.38	0.39	0.41	0.43	0.45	0.47	0.51	0.65	
Panel B: Univariate Sort on Idiosyncratic Tail Risk (Equal-Weighted)											
Excess Return	0.47	0.51	0.67	0.63	0.62	0.75	0.71	0.73	0.87	1.03	0.57
t-stat	(1.04)	(1.16)	(1.52)	(1.40)	(1.44)	(1.76)	(1.61)	(1.76)	(2.11)	(2.76)	(2.77)
FFC4 alpha	0.51	0.57	0.72	0.69	0.66	0.80	0.74	0.78	0.89	1.05	0.54
FFC4 t-stat	(1.23)	(1.38)	(1.75)	(1.70)	(1.67)	(1.98)	(1.83)	(2.07)	(2.30)	(2.96)	(2.69)
FF5 alpha	0.68	0.70	0.85	0.84	0.78	0.94	0.85	0.87	0.96	1.11	0.43
FF5 t-stat	(1.56)	(1.59)	(1.94)	(1.96)	(1.88)	(2.24)	(2.01)	(2.19)	(2.39)	(3.04)	(2.07)
Panel C: Univariate Sort on Idiosyncratic Tail Risk (Value-Weighted)											
Excess Return	0.23	0.39	0.50	0.52	0.41	0.42	0.53	0.48	0.77	0.89	0.66
t-stat	(0.61)	(1.15)	(1.46)	(1.41)	(1.20)	(1.23)	(1.34)	(1.26)	(2.05)	(2.23)	(3.16)
FFC4 alpha	0.27	0.45	0.55	0.56	0.45	0.46	0.55	0.56	0.80	0.88	0.61
FFC4 t-stat	(0.77)	(1.41)	(1.72)	(1.77)	(1.42)	(1.50)	(1.61)	(1.65)	(2.21)	(2.40)	(2.90)
FF5 alpha	0.43	0.58	0.59	0.69	0.50	0.55	0.67	0.68	0.84	1.00	0.57
FF5 t-stat	(1.15)	(1.82)	(1.78)	(2.19)	(1.57)	(1.80)	(1.90)	(1.93)	(2.24)	(2.57)	(2.61)

The table reports monthly average idiosyncratic tail risk, excess returns, and alphas for portfolios sorted on idiosyncratic tail risk between January 2001 to December 2016. Panel A reports the average idiosyncratic tail risk for portfolios sorted on idiosyncratic tail risk. Panel B reports equally-weighted excess returns and alphas for portfolios sorted on idiosyncratic tail risk. Panel C reports value-weighted excess returns and alphas for portfolios sorted on idiosyncratic tail risk.

that can model returns with infinite moments.

Panel B of Table 4.1 reports the average equal-weighted excess return of each decile portfolio and a portfolio that goes long the high idiosyncratic tail risk decile and shorts the low idiosyncratic tail risk decile (10-1). Average excess returns increase as the portfolio's average idiosyncratic tail risk increases. The long-short portfolio has an average monthly return of 0.57% with a t-statistic of 2.77.[19] The next 4 rows report returns and t-statistics relative to the Fama and French (1993) and Carhart (1997) four factor model (FFC4) and the Fama and French (2015) five factor model (FF5). Alphas are increasing in idiosyncratic tail risk. The long-short portfolio has a four factor alpha of 0.54% (t-stat of 2.69) and a five factor alpha of 0.43% (t-stat of 2.07).

Panel C of Table 4.1 reports the average value-weighted excess return of each decile portfolio and the long-short portfolio. Average value-weighted excess returns and alphas increase as

[19] All t-statistics are calculated using Newey and West (1987b) standard errors with one lag.

the portfolio's average idiosyncratic tail risk increases. The long-short portfolio has an average return of 0.66% with a t-statistic of 3.16. The long-short portfolio has a four factor alpha of 0.61% (2.90) and a five factor alpha of 0.57% (2.61). In summary, stocks with higher idiosyncratic tail risk have statistically and economically higher average returns than stocks with lower idiosyncratic tail risk.

Appendix C.4 verifies the robustness of these results. Appendix C.4.1 shows the cross-sectional relationship remains when estimating idiosyncratic tail risk using residuals from the CAPM, the Fama-French-Carhart four factor model, the high-frequency cross-sectional variable model, and the industry statistical model. The premium is also robust to using the 0.025 and 0.1 Extreme Value Theory thresholds $x_{t,(K_t+1)}$. In addition, results are similar for idiosyncratic tail risk estimated using midquotes instead of trades. Appendix C.4.2 reports double sorts on other firm characteristics that have been shown to predict returns. The double sorts demonstrate that conditional on other variables, average returns are still increasing in idiosyncratic tail risk.

I conclude that the idiosyncratic tail risk premium is a large economic phenomenon, highly statistically significant, and robust to different high-frequency factor models, tail risk thresholds, control variables, and midquotes. This section's cross-sectional findings are remarkably similar in magnitude to the cross-sectional results using jump risk implied by equity options. Kapadia and Zekhnini (2019) and Begin, Dorion, and Gauthier (2019) also find a positive cross-sectional relationship between implied jump risk and average returns and argue the premium is caused by the inability to diversity jump risk and not due to a systematic factor. Option implied jump risk can be considered as the price of insurance against jumps. It's not obvious that the realized idiosyncratic tail risk estimated from historical returns will also earn a premium, since it is a realized measure of jumps. This section demonstrates that the premium is a pervasive phenomenon that also exists for realized measures of idiosyncratic tail risk.

Table 4.2: Excess Returns and Alphas of Portfolios Sorted on Volume Tail Risk

	1 (Low)	2	3	4	5	6	7	8	9	10 (High)	(10-1)
Panel A: Average Portfolio Volume Tail Risk											
Volume Tail Risk	0.45	0.52	0.57	0.61	0.65	0.69	0.73	0.79	0.87	1.07	
Panel B: Univariate Sort on Volume Tail Risk (Equal-Weighted)											
Excess Return	0.30	0.58	0.61	0.67	0.65	0.63	0.79	0.73	0.84	1.17	0.88
t-stat	(0.66)	(1.38)	(1.44)	(1.57)	(1.53)	(1.43)	(1.82)	(1.75)	(1.99)	(2.83)	(3.88)
FFC4 alpha	0.34	0.62	0.66	0.73	0.70	0.69	0.85	0.80	0.88	1.16	0.83
FFC4 t-stat	(0.82)	(1.64)	(1.66)	(1.83)	(1.73)	(1.66)	(2.04)	(1.98)	(2.21)	(3.14)	(3.80)
FF5 alpha	0.48	0.77	0.78	0.86	0.80	0.82	0.96	0.89	0.97	1.25	0.77
FF5 t-stat	(1.16)	(1.95)	(1.88)	(2.02)	(1.88)	(1.89)	(2.22)	(2.10)	(2.32)	(3.21)	(3.85)
Panel C: Univariate Sort on Volume Tail Risk (Value-Weighted)											
Excess Return	0.31	0.50	0.69	0.77	0.77	0.63	0.68	0.78	0.79	0.98	0.67
t-stat	(0.89)	(1.41)	(1.93)	(2.11)	(2.12)	(1.65)	(1.64)	(2.03)	(2.01)	(2.47)	(3.70)
FFC4 alpha	0.35	0.54	0.72	0.83	0.80	0.68	0.72	0.82	0.85	0.97	0.61
FFC4 t-stat	(1.09)	(1.68)	(2.32)	(2.59)	(2.38)	(1.96)	(1.97)	(2.29)	(2.48)	(2.84)	(3.22)
FF5 alpha	0.46	0.65	0.82	0.95	0.87	0.80	0.79	0.90	0.96	1.03	0.58
FF5 t-stat	(1.42)	(1.98)	(2.55)	(2.91)	(2.49)	(2.22)	(2.16)	(2.45)	(2.63)	(2.93)	(3.17)

The table reports monthly average volume tail risk, excess returns, and alphas for portfolios sorted on Volume Tail Risk between January 2001 to December 2016. Panel A reports the average volume tail risk for portfolios sorted on volume tail risk. Panel B reports equally-weighted excess returns and alphas sorted on volume tail risk. Panel C reports value-weighted excess returns and alphas for portfolios sorted on volume tail risk.

4.3.4 The Volume Tail Risk Premium

The large intermediary hypothesis of tail risk predicts that there is a strong firm-level relationship between idiosyncratic tail risk and volume tail risk. As a litmus test, I evaluate whether stocks with higher Volume Tail Risk earn higher average returns. Volume tail risk is distinct from idiosyncratic tail risk, since it measured using trading volume data. For each month from January 2001 to December 2016, I estimate each firm's volume tail risk as the Hill estimate of the firm's changes in trading volume $v_{t,i}$. I sort stocks into decile groupings based on their VTR in that month. Decile 1 (low) holds stocks with the lowest VTR and decile 10 (high) holds with the highest VTR. I form an equally-weighted and a value-weighted portfolio in each decile, and hold the portfolio for 1 month.

Panel A of Table 4.2 reports the average volume tail risk in each decile. VTR has very heavy tails. In deciles 2-10 trading volume has an infinite variance as indicated by $\xi > 1/2$. In decile

10 trading volume has an infinite expectation as indicated by $\xi > 1$. Power-law is particularly suitable for modeling trading volume, since moment-based measures become intractable with such heavy tails.

Panel B of Table 4.2 reports the average equal-weighted excess return of each decile portfolio and a portfolio that goes long the highest VTR decile and shorts the lowest VTR decile. The results are remarkably similar to the idiosyncratic tail risk anomaly in Section 4.3.3. Average excess returns and alphas increase as portfolio VTR increases. The long-short portfolio has an average return of 0.88% with a t-stat of 3.88. The long-short portfolio has a four-factor alpha of 0.83% (t-stat of 3.80) and five-factor alpha of 0.77% (t-stat of 3.85).

Panel C of Table 4.2 reports the average value-weighted excess returns of each decile portfolio and a long-short portfolio. The value-weighted returns and alphas of the long-short portfolio are nearly identical to the average returns and alphas of the idiosyncratic tail risk long-short portfolio in Section 4.3.3. These results confirm that VTR has predictive ability for the cross-section of average stock returns. The returns of the long-short portfolio for ITR and VTR are highly correlated. The equal-weighted long-short portfolios have a correlation of 0.71 (t-stat of 13.88) while the value-weighted long-short portfolios have a correlation of 0.43 (t-stat of 6.57). These results suggest that the ITR and VTR premiums may be driven by the same risk factor.

4.3.5 Persistence in Idiosyncratic Tail Risk and Volume Tail Risk Premia

Existing theories argue that the idiosyncratic tail risk premium is caused by an inability to fully diversity or an under-reaction to firm-specific news. In both theories, the premium should be short-lived as markets will quickly adjust to the firm-specific information. My risk-based theory predicts the premium is highly persistent as long as firms' tail risks and intermediary betas are fairly stable. If risk exposures do not change rapidly over time, then lagged idiosyncratic tail risk is correlated to current idiosyncratic tail risk and the risk premium should be persistent across stocks, that is those with high (low) past idiosyncratic tail risks should continue to

Table 4.3: Persistence in Excess Returns of Portfolios Sorted on Tail Risk

	Idiosyncratic Tail Risk				Volume Tail Risk			
Hold Period (Months)	3	6	12	24	3	6	12	24
Panel A: Univariate Sort on Idiosyncratic Tail Risk and Volume Tail Risk (Equal-Weighted)								
(10-1) Return	2.97	4.39	6.16	7.98	3.84	7.23	13.42	25.01
t-stat	(5.08)	(5.56)	(5.00)	(3.13)	(4.40)	(3.95)	(3.77)	(3.51)
Panel B: Univariate Sort on Idiosyncratic Tail Risk and Volume Tail Risk (Value-Weighted)								
(10-1) Return	2.85	5.14	9.35	16.60	2.96	5.97	11.18	22.46
t-stat	(4.39)	(3.78)	(4.04)	(3.44)	(3.85)	(4.13)	(3.37)	(3.71)

The table reports long-horizon excess returns for portfolios sorted on idiosyncratic tail risk and volume tail risk between January 2001 to December 2016. Panel A reports equally-weighted excess returns for the idiosyncratic tail risk and volume tail risk long-short portfolios. Panel B reports value-weighted excess returns for the idiosyncratic tail risk and volume tail risk long-short portfolios.

experience high (low) future average returns.

To evaluate this hypothesis, in each month I sort stocks into decile groupings based on their ITR in that month, form an equally-weighted and a value-weighted portfolio in each decile, and hold the portfolio for 3, 6, 12, or 24 months. I perform an analogous portfolio sort for volume tail risk.

Table 4.3 reports the average excess returns of portfolios that buys the highest tail risk decile and shorts the lowest tail risk decile for 3, 6, 12, or 24 months. The left side of the table shows results for idiosyncratic tail risk long-short portfolios and the right side shows results for volume tail risk long-short portfolios. Panel A shows that the equal-weighted idiosyncratic tail risk and volume tail risk premia are highly persistent. Excess returns of the idiosyncratic tail risk long-short portfolio range from 2.97% (t-stat of 5.08) at the three-month horizon to 7.98% (t-stat of 3.13) at the two-year horizon. Similarly, excess returns of the volume tail risk long-short portfolio range from 3.84% (t-stat of 4.40) at the three-month horizon to 25.01% (t-stat of 3.51) at the two-year horizon.

Panel B shows that the value-weighted idiosyncratic tail risk and volume tail risk premia are also highly persistent and have higher magnitudes than the equal-weighted portfolios. Excess returns of the idiosyncratic tail risk long-short portfolio range from 2.85% (t-stat of 4.39) at

the three-month horizon to 16.60% (t-stat of 3.44) at the two-year horizon. Similarly, excess returns of the volume tail risk long-short portfolio range from 2.96% (t-stat of 3.85) at the three-month horizon to 22.46% (t-stat of 3.71) at the two-year horizon.

The persistence of the idiosyncratic tail risk and volume tail risk long-short portfolio returns is consistent with the risk-based explanation proposed in this paper. This finding also provides evidence against existing theories of the premium, which should dissipate as the holding period increases. Instead, I find the returns increase as the holding period increases. Additionally, transaction costs should not affect the premium, since turnover is likely to be low when returns are so highly persistent.

4.3.6 Idiosyncratic Tail Risk Covariates

My explanation for the idiosyncratic tail risk premium relies on Gabaix et al. (2006)'s hypothesis that large intermediaries cause tail risk through large trades. This section provides empirical evidence for their large intermediary hypothesis. In their economic model, the price impact of a trade scales with the trading volume, and the tail distribution of idiosyncratic returns is proportional to the tail distribution of trading volume. Hence, their hypothesis can be tested by analyzing the firm-level relationship between idiosyncratic tail risk and volume tail risk. Additionally, I use the Lee and Radhakrishna (2000) algorithm to identify intermediary trades and provide direct evidence on whether a higher percentage of intermediary trades is associated with a higher idiosyncratic tail risk.[20] I also classify the direction of intermediary trades using the Lee and Ready (1991) algorithm to test whether a higher percentage of intermediary selling (buying) as a percentage of total trades is associated with a higher idiosyncratic tail risk. Finally, I verify the persistence of idiosyncratic tail risk. In each month from January 2001 to December 2016, I run Fama-MacBeth cross-sectional regressions on nested versions of

$$ITR_{l,t} = \pi_{0,t} + \pi_{1,t} VTR_{l,t} + \pi_{2,t} \%IntermediaryVolume_{l,t} + \pi_{3,t} ITR_{l,t-k} + \pi_{4,t} Controls_{l,t} + \epsilon_{l,t}, \quad (4.19)$$

[20] Large trades are classified as intermediary driven if their dollar volume exceeds \$100,000 for the largest, \$50,000 for the medium, and \$20,000 for the smallest tripartite of stocks (Lee and Radhakrishna (2000)).

Table 4.4: Idiosyncratic Tail Risk Covariates

	Panel A: VTR		Panel B: Intermediary Dollar Volume						Panel C: Lagged ITR					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Intercept	0.32	0.31	0.39	0.37	0.39	0.38	0.40	0.38	0.28	0.28	0.32	0.32	0.35	0.35
t-stat	52.17	60.62	174.66	125.99	163.22	131.64	234.91	141.00	47.08	49.80	68.90	52.61	106.43	73.52
VTR	0.14	0.12												
t-stat	14.01	16.21												
%Intermediary Total			0.06	0.06										
t-stat			17.98	13.60										
%Intermediary Sell					0.12	0.10								
t-stat					15.64	13.31								
%Intermediary Buy							0.09	0.09						
t-stat							14.86	14.21						
ITR t-1									0.32	0.27				
t-stat									23.90	17.02				
ITR t-12											0.22	0.17		
t-stat											18.36	12.29		
ITR t-24													0.14	0.10
t-stat													13.68	9.05
Size		0.00		0.00		0.00		0.00		0.00		0.00		0.00
t-stat		1.59		-1.93		-1.73		-1.87		-0.36		-0.40		-0.05
ILLIQ		0.12		0.13		0.12		0.13		0.10		0.11		0.11
t-stat		3.96		4.58		4.56		4.61		4.54		5.55		5.93
IVOL		0.01		0.01		0.01		0.01		0.01		0.01		0.01
t-stat		17.71		14.53		14.18		14.99		17.99		16.73		16.88
% Adj. R^2	10.19	16.75	1.96	13.35	2.24	13.18	1.25	12.84	10.87	18.63	5.20	13.87	2.43	11.32

Monthly firm-level regression of idiosyncratic tail risk on volume tail risk, percent intermediary dollar volume, and lagged idiosyncratic tail risk, controlling for size, illiquidity, and idiosyncratic volatility. The table reports point estimates, Newey-West t-statistics with one lag, and adjusted R^2.

where for each firm l and month t, $ITR_{l,t}$ is the idiosyncratic tail risk, $VTR_{l,t}$ is the volume tail risk, $\%IntermediaryVolume_{l,t}$ is the dollar volume of intermediaries' large trades as a percentage of the stock's total dollar volume, $ITR_{l,t-k}$ is idiosyncratic tail risk with lag k, and $Controls_{l,t}$ include size, illiquidity, and idiosyncratic volatility.[21]

Table 4.4 reports point estimates, Newey-West t-statistics with one lag, and the adjusted R^2 for regression (4.19). Column (1) of Panel A regresses idiosyncratic tail risk on just volume tail risk. The intercept term of 0.32 can be interpreted as the base level of ITR. Volume tail risk contributes $0.14\times VTR$ to the firm's monthly idiosyncratic tail risk. VTR also has a high degree of explanatory power for the cross-section of ITR, with an adjusted R^2 of 10%. In column (2), VTR coefficient does not substantially change after controlling for size, idiosyncratic volatility, illiquidity. This finding supports the theoretical link between idiosyncratic tail risk and volume tail risk.

Column (3) in Panel B reports results for a regression of idiosyncratic tail risk on percent intermediary dollar volume. There is a strong relationship between ITR and percent interme-

[21]The results are robust to controlling for firm-level and month fixed effects, available upon request.

diary dollar volume. Stocks with 1% higher intermediary dollar volume have 0.06 high ITR. Since the average ITR is roughly 0.44 with a standard deviation of 0.1, this increase is highly economically significant. One potential cause of this relationship may be that stocks that are small, illiquid or volatile may be associated with higher intermediary dollar volume. However, after controlling for these characteristics in column (4), the coefficient is unchanged. Column (5) and (8) report results for a regression of ITR on percent intermediary selling and buying volume respectively. Since ITR is a measure of the left tail, it should be more affected by intermediary selling volume, which is what I show. Stocks with 1% higher intermediary selling volume have 0.12 high ITR. ITR is also associated with higher buying volume, which is consistent with large trades absent of information being quickly reversed by arbitragers. These results provide strong empirical support for the large intermediary theory of idiosyncratic tail risk.

In contrast to existing theories, my explanation predicts idiosyncratic tail risk is persistent. Columns (9), (11), and (13) in Panel C report results for a regression of idiosyncratic tail risk on itself with a 1-month, 1-year, and 2-year lag. Idiosyncratic tail risk is highly persistent at the 1-month horizon with an adjusted R^2 of nearly 11%. The persistence decreases at the 1-year and 2-year horizons, but their coefficients remain statistically significant. Since, idiosyncratic tail risk is estimated with non-overlapping data, this persistence is not mechanical. This finding is consistent with investor herding, where intermediaries frequently purchase the same stocks resulting in autocorrelated stock trades as documented in Sias (2004). The persistence of idiosyncratic tail risk is also consistent with the persistence of its premium, which requires the risk measure to be fairly stable. This finding provides evidence against the existing theory that idiosyncratic tail risk is only caused by inability to diversify or firm-specific news, which should have no persistence.

4.4 Commonality in Idiosyncratic Tail Risk and Pricing Implications

A key hypothesis in my explanation of the ITR premium is that idiosyncratic tail risk follows a strong factor structure and that common idiosyncratic tail risk is correlated to the intermediary marginal value of wealth. In this section, I verify that idiosyncratic tail risk follows a strong factor structure and that a single factor explains a high degree of the time variation in firm-specific tail risk. The common idiosyncratic tail risk factor is highly correlated to existing intermediary factors and is procyclical. Stocks with higher exposures to this common idiosyncratic tail risk factor earn higher average returns, suggesting it's a priced risk factor. Asset pricing tests show that the common idiosyncratic tail risk factor explains the idiosyncratic tail risk premium.

4.4.1 Factor Structure in Idiosyncratic Tail Risk

Firms with different characteristics share common idiosyncratic tail risk dynamics over time. Figure 4.2 plots average total tail risk (tail risk of returns) and idiosyncratic tail risk for portfolios formed on size, book-to-market, and industry. Panels (a) and (b) plot average total tail risk and idiosyncratic tail risk of size quintiles. As expected, levels of tail risk decreases as firm size increases. However, the idiosyncratic tail risk of different size quintiles are also strongly correlated through time. The largest firms are several orders of magnitude larger than the smallest firms, but their idiosyncratic tail risk has a time-series correlation of 44.6%. Panels (c) and (d) plot average total tail risk and idiosyncratic tail risk of firms sorted by book-to-market. The idiosyncratic tail risk of quintiles with the highest and lowest book-to-market have a correlation of 60.5%. Panels (e) and (f) plots average idiosyncratic tail risk for firms sorted into the five-industry SIC code categories from Kenneth French's website. The idiosyncratic tail risk of the industries have an average pairwise correlation of 68.7% with a minimum correlation of 58.4% between the health and manufacturing sectors. Firms with different size, book-to-market and

in different industries all share common idiosyncratic tail risk dynamics. Additionally, the total tail risk time-series dynamics in Panels (a), (c), and (e) are nearly identical to the idiosyncratic tail risk dynamics in Panels (b), (d), and (f), indicating that most of the time-series variation in total tail risk is coming from the idiosyncratic component, and not the systematic component.

Tail risk of returns is expected to be correlated by Equation (1), since the systematic factors are heavy-tailed.[22] Returns inherit the tail behaviour of systematic factors, and time variation in factor tail risk causes the tail risk of returns to be correlated. For example in the CAPM, returns inherits the tail risk of the market factor. However, the high correlation of idiosyncratic tail risks is unexpected, since idiosyncratic returns are uncorrelated to systematic factors. If idiosyncratic tail risk is only caused by firm-specific shocks as in Merton (1976), then there should be no correlation of firm-specific tail risks. The observed commonality suggests firm-level tail risk is driven by a common factor.

The high correlation of idiosyncratic tail risks suggests modeling the dynamics using a single factor model. Consistent with my explanation in Equation (4.2), I define common idiosyncratic tail risk (CITR) as the mean cross-sectional idiosyncratic tail risk in each month. For each firm, I run monthly and annual time-series regressions of idiosyncratic tail risk on CITR,[23]

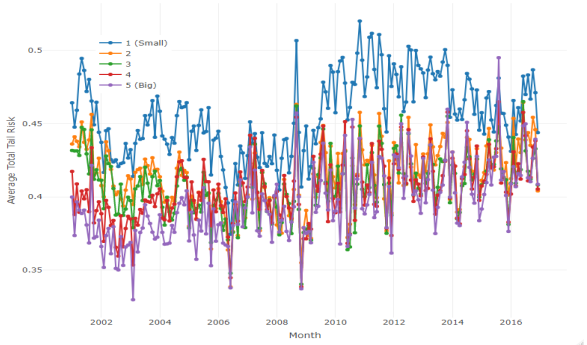
$$\xi_{l,t} = \kappa_{l,0} + \kappa_{l,1} CITR_t + e_{l,t}, \tag{4.20}$$

which is analogous to the hypothesized factor structure in Equation (4.4) with $\kappa_{l,0} = -\mu_N$ and $\kappa_{l,1} = 1$. This single factor model for idiosyncratic tail risk has an average R^2 of 13.6% for the monthly regression and an average R^2 of 42.6% for the annual regression. In comparison, Begin, Dorion, and Gauthier (2019) document an R^2 of 56.4%. However, their sample of 260 S&P stocks is much smaller and more homogeneous than the 6,213 equities sampled in this paper. Considering that my sample contains small and medium stocks, the 42.6% average R^2

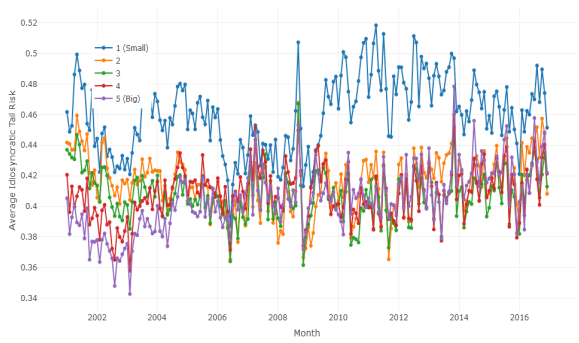
[22] Since the systematic factors are portfolios of heavy-tailed returns, they will also be heavy-tailed. For example, it is well documented that the market return is heavy-tailed (Kelly and Jiang (2014b))

[23] Annual idiosyncratic tail risk and CITR is calculated as the average of their monthly values within each year.

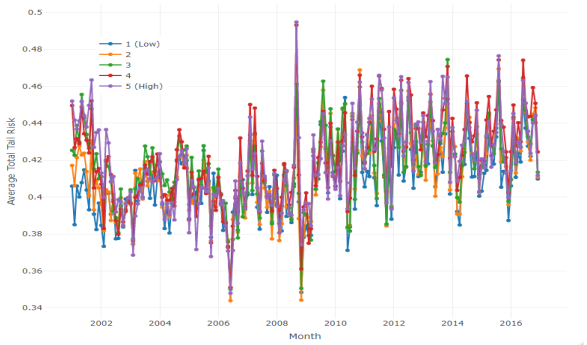
Figure 4.2: Average Idiosyncratic Tail Risk of Portfolios Formed on Firm Characteristics



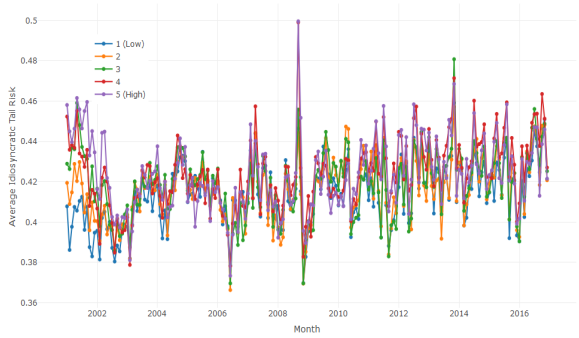
(a) Average Total Tail Risk by Size



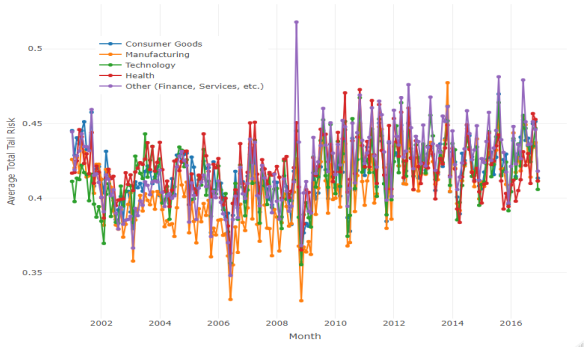
(b) Average Idiosyncratic Tail Risk by Size



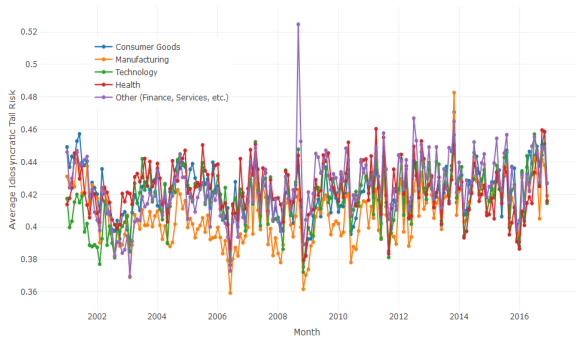
(c) Average Total Tail Risk by Book-to-Market



(d) Average Idiosyncratic Tail Risk by Book-to-Market



(e) Average Total Tail Risk by Industry



(f) Average Idiosyncratic Tail Risk by Industry

The figure plots monthly total tail risk and idiosyncratic tail risk averaged within size quintiles, book-to-market quintiles, and industries from January 2001 to December 2016. Each month, total tail risk for each stock is the Hill estimator of log returns, and idiosyncratic tail risk for each stock is calculated as the Hill estimator of residuals from the Fama and French (2015) five factor model. Panel (a) and (b) shows total tail risk and idiosyncratic tail risk averaged within market capitalization quintiles. Panel (c) and (d) shows total tail risk and idiosyncratic tail risk averaged within book-to-market quintiles. Panel (e) and (f) shows total tail risk and idiosyncratic tail risk averaged within the five-industry SIC categorizes on Kenneth French's website.

is economically significant.

Another natural comparison can be done with common idiosyncratic volatility (CIV) from Herskovic et al. (2016).[24] They document that the annual regression of idiosyncratic volatility on CIV has an average R^2 of 35%. However, they examine a larger sample at a longer time-span, so it is again difficult to make direct comparisons. However, it is evident that idiosyncratic tail risk follows a strong factor structure, where CITR explains much of the variation in idiosyncratic tail risk.

4.4.2 Robustness to Omitted Variables

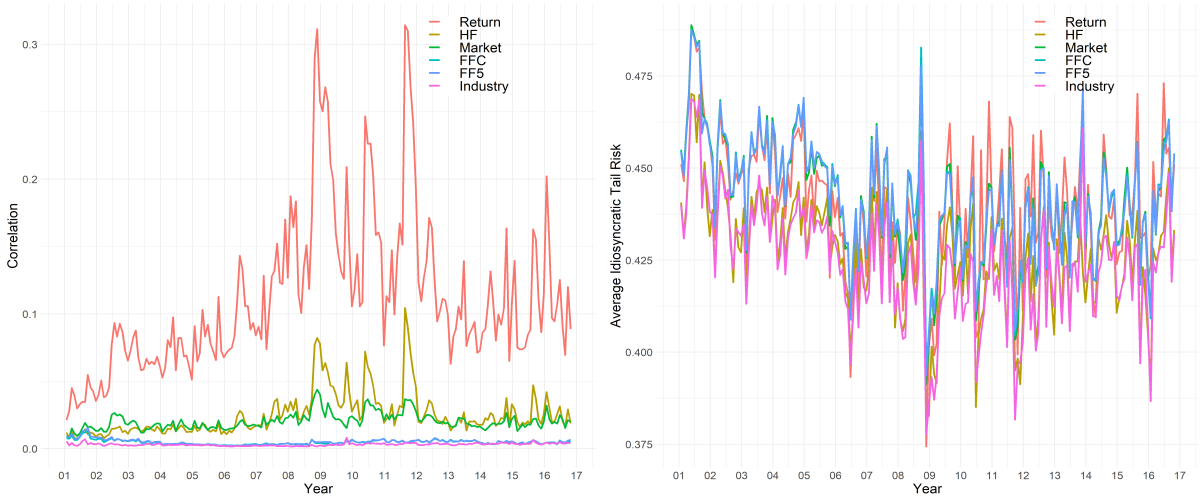
The commonality of the idiosyncratic tail risks cannot be explained by omitted common factors if any. Panel A of Figure 4.3 shows that high-frequency returns exhibit substantial common variation, with an average monthly pairwise correlation of 11% during 2001 to 2016, and a monthly pairwise correlation exceeding 30% during the Subprime Mortgage and Euro Debt crises. Even after removing the market factor, the residuals continue to exhibit some common variation, with an average and maximum pairwise correlation of 2% and 4.4% respectively. The Fama and French five factor model captures nearly all the common variation of high-frequency returns, as the average pairwise correlation among its residuals is only 0.05% and the maximum is 1.5% during the 9/11 terrorist attack (after 2001, the average pairwise correlation never exceeds 0.9%).[25] The Fama-French-Carhart four factor and industry statistical models also remove most of the covariation from returns.

Panel B of Figure 4.3 shows that the average total tail risk of returns and the idiosyncratic tail risk from the various factor models are nearly identical, despite the fact that the Fama and

[24]The CITR factor only has a pairwise correlation of 0.08 (t-stat of 0.97) with the CIV factor in Herskovic et al. (2016) and the opposite price of risk. Additionally, while the volatility factor of Herskovic et al. (2016) is countercyclical, I affirm that the idiosyncratic tail risk factor is strongly procyclical indicating they are separate factors.

[25]Conversely, the high-frequency cross-sectional variable model does not remove as much of the covariation as the Fama and French models. The failure of the high-frequency cross-sectional variable model is consistent with the finding from Huddleston, Liu, and Stentoft (2020) that price trends and liquidity variables do not help forecast high-frequency market returns.

Figure 4.3: Average Pairwise Correlations and Average Idiosyncratic Tail Risks



(a) Average pairwise correlations

(b) Average idiosyncratic tail risks

The figure plots the average correlation and idiosyncratic tail risks of returns and factor model residuals from 2001 to 2016. Idiosyncratic tail risk is the Hill estimate of residuals from the market model (Market), high-frequency cross-section variable model (HF), industry statistical model (Industry), Fama-French-Carhart four factor model (FFC), or Fama and French five factor model (FF5). Panel A shows the average pairwise Spearman correlation for returns and residuals for each month. The figures for Pearson and Kendall correlations look nearly identical. Panel B shows the average idiosyncratic tail risk across firms for each month.

French models and industry model saturate nearly all of the covariation in returns. This makes omitted factors an unlikely explanation for the strong commonality in idiosyncratic tail risks. Furthermore, Section C.4.1 shows that the idiosyncratic tail risk premium is robust to these alternative factor models, verifying that omitted factors also cannot explain the premium.

4.4.3 Link Between Common Idiosyncratic Tail Risk and Intermediary Factors

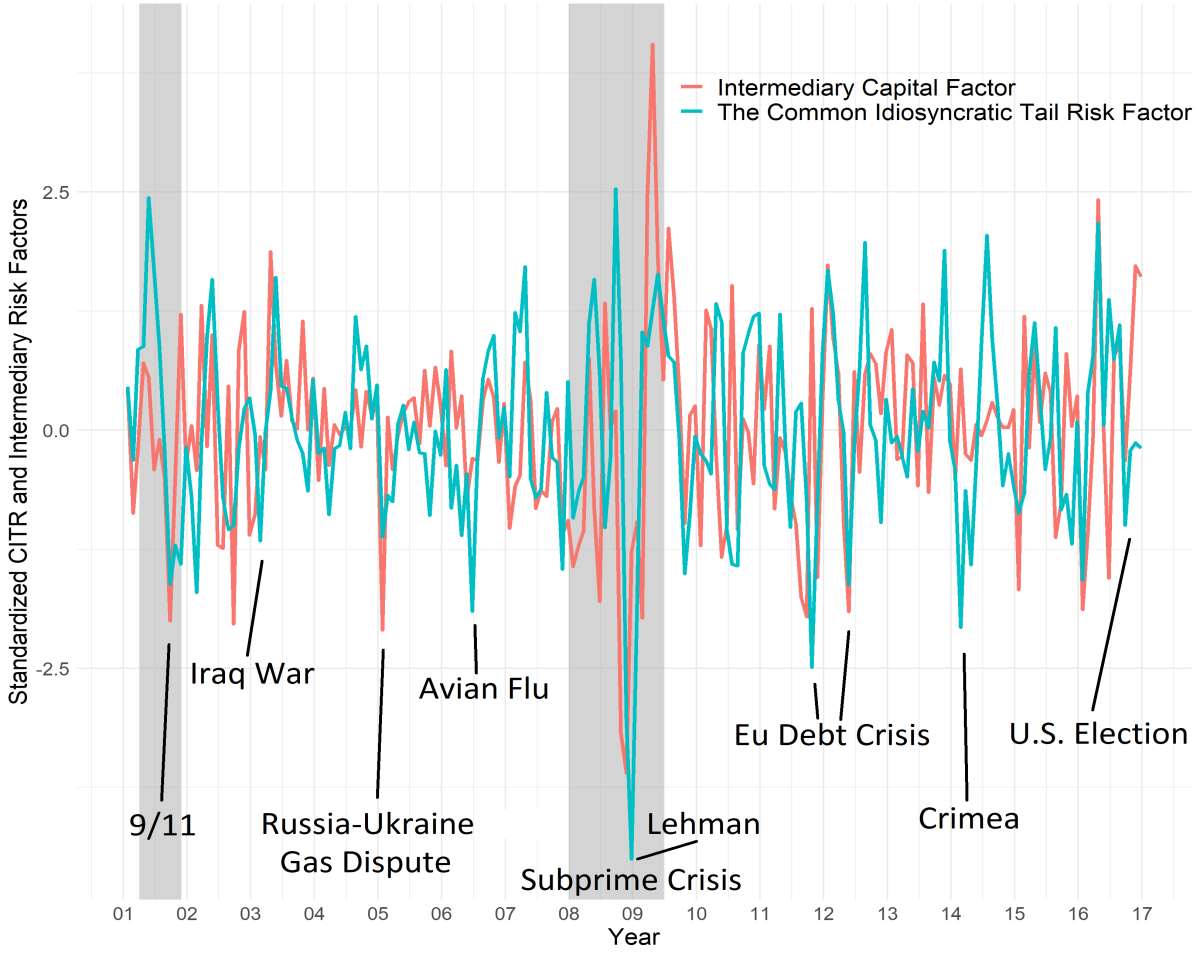
A key assumption in my explanation is that CITR, the factor driving idiosyncratic tail risks, is correlated to the intermediary marginal value of wealth. When intermediary funding is high, they trade on firm-specific signals, and the average idiosyncratic tail risk increases. When constraints tighten, investment opportunities increase as intermediaries are unable to fund many firm-specific trades, and the average idiosyncratic tail risk falls. A testable implication is that

CITR is highly correlated to intermediary factors in the literature and negative economic news should be associated with *decreases* to CITR.

As empirical evidence, I plot the CITR factor and the intermediary capital factor from He, Kelly, and Manela (2017), downloaded from Asaf Manela's website. Since their intermediary factor is defined as quarterly shocks to intermediary capital levels, I define the CITR factor $\triangle CITR$ as 3-month differences in CITR levels, that is $\triangle CITR_t = CITR_t - CITR_{t-3}$. The intermediary capital factor is the change in the equity capital ratio of large primary dealers, a prominent example of sophisticated intermediaries. Figure 4.4 shows that the CITR factor and intermediary capital factor are remarkably correlated, despite the fact that CITR is measured using high-frequency equity returns and intermediary capital is measured using quarterly accounting data. The CITR factor and the intermediary capital factor have a monthly correlation of 0.32 (t-stat of 4.58) and a quarterly correlation of 0.54 (t-stat of 4.30). Additionally, the CITR factor and the broker-dealer leverage factor from Adrian, Etula, and Muir (2014) have a quarterly correlation of 0.31 (t-stat of 2.15). The CITR factor and the (negative) Leverage Constraint Tightness factor (the average market beta of actively managed mutual funds is correlated to intermediary funding liquidity tightness) from Boguth and Simutin (2018) has a correlation of 0.20 (t-stat of 2.67). The empirical correlation between CITR and existing intermediary factors provide strong support for my hypothesis that CITR is correlated to the intermediary marginal value of wealth.

Figure 4.4 also illustrates that there is strong evidence that CITR decreases during negative news shocks. Significant negative news like the 9/11 terrorist attack, Iraq war, Russia-Ukraine gas dispute, Lehman Brothers bankruptcy, EU debt crisis, and 2016 election are associated with large decreases in both the CITR and intermediary factors. The largest decrease to CITR occurred in October 2008 after the bankruptcy of Lehman Brothers when intermediary funding was at its lowest. Interestingly, the 2004 Avian flu outbreaks and 2014 Crimea Annexation resulted in much larger decreases in CITR relative to the intermediary capital factor, illustrating that CITR contains additional information. These two events show that large intermediaries

Figure 4.4: The Common Idiosyncratic Tail Risk and Intermediary Capital Factors



The common idiosyncratic tail risk factor and the intermediary capital factor. Both time-series are standardized to zero mean and unit variance. The monthly sample is from January 2001 to December 2016. CITR is the mean monthly idiosyncratic tail risk.

decreased large trades as possible risks increased, but quickly reversed course when the risk failed to materialize.

Consistent with my hypothesis, CITR is correlated to existing intermediary factors. Negative economic news is associated with decreases to CITR, which contradicts the fire-sale and labor risk theories of idiosyncratic tail risk. The strong empirical connection between shocks to CITR and negative economic news rules out the theory that idiosyncratic tail risk is only caused by firm-specific news shocks.

Table 4.5: Correlations between the CITR factor and Economic Variables

Monthly Variables	P/E Growth	Market Volatility Growth	Market Return
Pairwise Correlation	0.22	-0.07	0.15
t-stat	(3.16)	(-0.93)	(2.14)
Quarterly Variables	**GDP Growth**	**Investment Growth**	**Consumption Growth**
Pairwise Correlation	0.27	0.25	0.18
t-stat	(2.23)	(1.98)	(1.42)

Time-series correlations between the common idiosyncratic tail risk factor and economic variables from January 2001 to December 2016. The common idiosyncratic tail risk factor each month is the mean firm-level tail risk. Monthly growth (log changes) factors are Price-to-Earnings of the S&P 500, market volatility, and market prices. Quarterly growth (log changes) factors are GDP, investment, and consumption.

4.4.4 Cyclicality of Common Idiosyncratic Tail Risk

In my explanation, idiosyncratic tail risk comoves with intermediary funding, which is a negative function of the shadow cost of capital ϕ_1. A testable implication is that the common idiosyncratic tail risk factor should be procyclical, since ϕ_1 is counter-cyclical. In this section, I examine the empirical correlation of CITR to innovations in financial and real economic variables.

Table 4.5 reports correlations of the common idiosyncratic tail risk factor and growth (log changes) of aggregate macro variables. Monthly economic variables include Robert Shiller's (Shiller (2015)) Cyclically Adjusted Price-to-Earnings Ratio of the S&P 500, market volatility, and market prices. Seasonally adjusted quarterly economic variables include real GDP, gross private domestic investment, and real personal consumption expenditures. All correlations support the procyclicality of idiosyncratic tail risk. Shocks to common idiosyncratic tail risk are positively correlated with positive economic growth, measured as increases to Price-to-Earnings, market prices, GDP, investment, and consumption. CITR and market volatility have nearly zero correlation, further illustrating that volatility does not drive tail risk.

The commonality and procyclicality of idiosyncratic tail risk is difficult to reconcile with the theories positing that idiosyncratic tail risk is caused by firm-specific shocks (Merton

(1976)), investor fire-sales, or labor risk. If idiosyncratic tail risk is only caused by firm-specific news, then there would be no commonality in idiosyncratic tail risk. Alternatively, if idiosyncratic tail risk is primarily caused by fire-selling during a liquidity crisis or labor risk during an economic crisis, then there may be common variation in idiosyncratic tail risk, but it would be countercyclical. The procyclical common idiosyncratic tail risk factor documented in this section strongly supports my hypothesis that idiosyncratic tail risk is driven by shocks to intermediary funding.

4.4.5 CITR Exposure and Average Returns

In my explanation for the idiosyncratic tail risk premium, CITR is a priced risk factor. This section documents that stocks' exposure to the common idiosyncratic tail risk factor helps to explain differences in the cross-section of expected returns. For each month from January 2001 to December 2016, I estimate each stock's CITR-beta by regressing the stock's monthly excess return on the CITR factor using a 36-month trailing window.[26] This procedure results in a CITR-beta for each stock in each month. A stock's CITR-beta is a measure of its exposure to the common idiosyncratic tail risk factor.

For the analysis, I sort stocks in each month into quintiles based on their CITR-betas. Stocks in quintile 1 have the least exposure to CITR and stocks in quintile 5 have the most CITR exposure. I form an equal-weighted portfolio and a value-weighted portfolio in each quintile and hold the portfolios for 1 month. Panel A of Table 4.6 reports the average CITR-beta in each quintile. There is a large range of CITR-betas from -0.97 in the lowest quintile to 2.28 in the highest quintile. Stocks in quintile 1 hedge states of low CITR, while stocks in quintile 5 lose value when CITR is low.

Panel B reports excess returns for the equal-weighted portfolios, and a portfolio that goes long the highest CITR-beta quintile and shorts the lowest CITR-beta quintile. Excess returns

[26] A stock is included only if it has no missing return in the 36-month estimation window. Additional returns from 1998-2001 are included in estimating betas.

Table 4.6: Excess Returns and Alphas of Portfolios Sorted on CITR-betas

	1 (Low)	2	3	4	5 (High)	(5-1)
Panel A: Average CITR-beta						
Average CITR-beta	-0.97	0.05	0.57	1.13	2.28	
Panel B: Univariate Sort on CITR-beta (Equal-Weighted)						
Excess Return	0.50	0.70	0.70	0.89	0.90	0.40
t-stat	(1.11)	(1.82)	(1.79)	(2.07)	(1.72)	(2.06)
FFC4 alpha	0.55	0.74	0.74	0.94	0.95	0.40
FFC4 t-stat	(1.30)	(2.08)	(2.00)	(2.28)	(1.93)	(2.21)
FF5 alpha	0.73	0.86	0.83	0.99	1.01	0.28
FF5 t-stat	(1.72)	(2.31)	(2.10)	(2.28)	(1.97)	(1.77)
Panel C: Univariate Sort on CITR-beta (Value-Weighted)						
Excess Return	0.14	0.37	0.54	0.63	0.76	0.62
t-stat	(0.37)	(1.20)	(1.67)	(1.71)	(1.42)	(2.01)
FFC4 alpha	0.15	0.42	0.60	0.70	0.80	0.65
FFC4 t-stat	(0.42)	(1.44)	(1.98)	(1.96)	(1.65)	(2.21)
FF5 alpha	0.29	0.50	0.70	0.78	0.84	0.55
FF5 t-stat	(0.86)	(1.72)	(2.25)	(2.08)	(1.73)	(2.08)

The table reports monthly average CITR-beta, excess returns, and alphas for portfolios sorted on CITR-betas between January 2001 to December 2016. Panel A reports the average CITR-beta for portfolios sorted on CITR-beta. Panel B reports equally-weighted excess returns and alphas sorted on CITR-beta. Panel C reports value-weighted excess returns and alphas sorted on CITR-beta.

and four factor alphas are monotonically increasing in CITR-betas. The long-short portfolio has an average return of 0.40% (t-stat of 2.06). The next 4 rows report excess returns relative to the Fama-French-Carhart four factor model and Fama and French five factor model. The long-short portfolio has a four factor alpha of 0.40% (t-stat of 2.21) and five-factor alpha of 0.28% (t-stat of 1.77). Panel C reports excess returns and alphas for value-weighted portfolios. Value-weighted excess returns and alphas are monotonically increasing in CITR-betas. The long-short portfolio has an average monthly return of 0.62% (t-stat of 2.01), four factor alpha of 0.65% (t-stat of 2.11), and five factor alpha of 0.55% (t-stat of 2.08).

In summary, stocks with high CITR-betas have economically and statistically higher returns than stocks with low CITR-betas. These results support the hypothesis that stocks with low or negative CITR exposure provide hedges for states of low CITR, and stocks with high CITR

exposure are compensated for the additional risk. These results support my hypothesis that CITR is a priced risk factor. Additionally, the positive price of risk for CITR contradicts the existing fire-sale or labour risk theories of tail risk that imply a negative price of risk, and also contradicts the news explanation that implies no price of risk.

4.4.6 Pricing Idiosyncratic Tail Risk and Volume Tail Risk Portfolios

This section tests the risk-based explanation for the idiosyncratic tail risk premium using formal asset pricing procedures. To do so, I demonstrate that the CITR factor is priced in portfolios sorted on idiosyncratic tail risk and volume tail risk. I also show that differences in CITR exposure of idiosyncratic tail risk and volume tail risk portfolios can account for most of the differences in average returns. I conduct a two-stage Fama and MacBeth (1973) estimation procedure from January 2001 to December 2016. In the first stage, I estimate factor betas for each test asset k from time-series regression,

$$R_{k,t+1} - R_{f,t} = a^k + \beta_{k,CITR} \triangle CITR_{t+1} + \beta_{k,M}(R_{M,t+1} - R_{f,t}) + e_{k,t+1}, \tag{4.21}$$

where $R_{k,t+1} - R_{f,t}$ are monthly excess returns for test asset k, $\triangle CITR_{t+1}$ is the common idiosyncratic tail risk factor in percentage terms (that is, three-month changes in CITR times 100), and $R_{M,t+1} - R_{f,t}$ is the excess return on the market portfolio. In the second stage, I use the estimated betas to run cross-sectional regression,

$$E[R_{k,t+1} - R_{f,t}] = \alpha + \beta_{k,CITR}\lambda_{CITR} + \beta_{k,M}\lambda_M + \epsilon_k, \tag{4.22}$$

where $E[R_{k,t+1} - R_{f,t}]$ is the average excess return of test asset k, λ_{CITR} is the risk price for the common idiosyncratic risk factor, and λ_M is the risk price of the market, reported in percentage terms. In addition to point estimates of the risk prices, I adjust the Fama and MacBeth standard errors for time-series correlation by reporting Newey-West t-statistics with one lag and Shanken t-statistics (Shanken (1992)). To evaluate model fit, I report the cross-sectional R^2

Table 4.7: Cross-sectional Asset Pricing Tests on ITR and VTR Deciles

	Panel A: Idiosyncratic Tail Risk Deciles		Panel B: Volume Tail Risk Deciles	
	(1)	(2)	(3)	(4)
Intercept	0.00	1.49	0.20	3.33
NW t-stat	(0.01)	(1.73)	(0.55)	(4.15)
Shanken t-stat	[0.00]	[1.28]	[0.47]	[2.99]
$R^M - R^f$		-1.04		-2.84
NW t-stat		(-1.13)		(-3.31)
Shanken t-stat		[-0.84]		[-2.43]
CITR	1.21	0.99	1.08	0.73
NW t-stat	(2.69)	(2.27)	(2.90)	(2.03)
Shanken t-stat	[2.10]	[1.83]	[2.19]	[1.42]
% Adj. R^2	73.45	86.05	31.60	68.05
% MAE	0.08	0.06	0.11	0.08
Months	192	192	192	192

In Panel A, the test assets are decile portfolios sorted on idiosyncratic tail risk. In Panel B, the test assets are decile portfolios sorted on volume tail risk. The Fama MacBeth analysis is from January 2001 to December 2016. The model in columns (1) and (3) uses innovations in CITR as the factor. The model in columns (2) and (4) uses both the market portfolio and innovations in CITR as the factors. The table reports the risk price estimates, R^2, and mean absolute pricing errors in percentage terms, Newey-West t-statistics with one lag, and Shanken t-statistics.

and the mean absolute pricing error (MAPE), both in percentage terms.

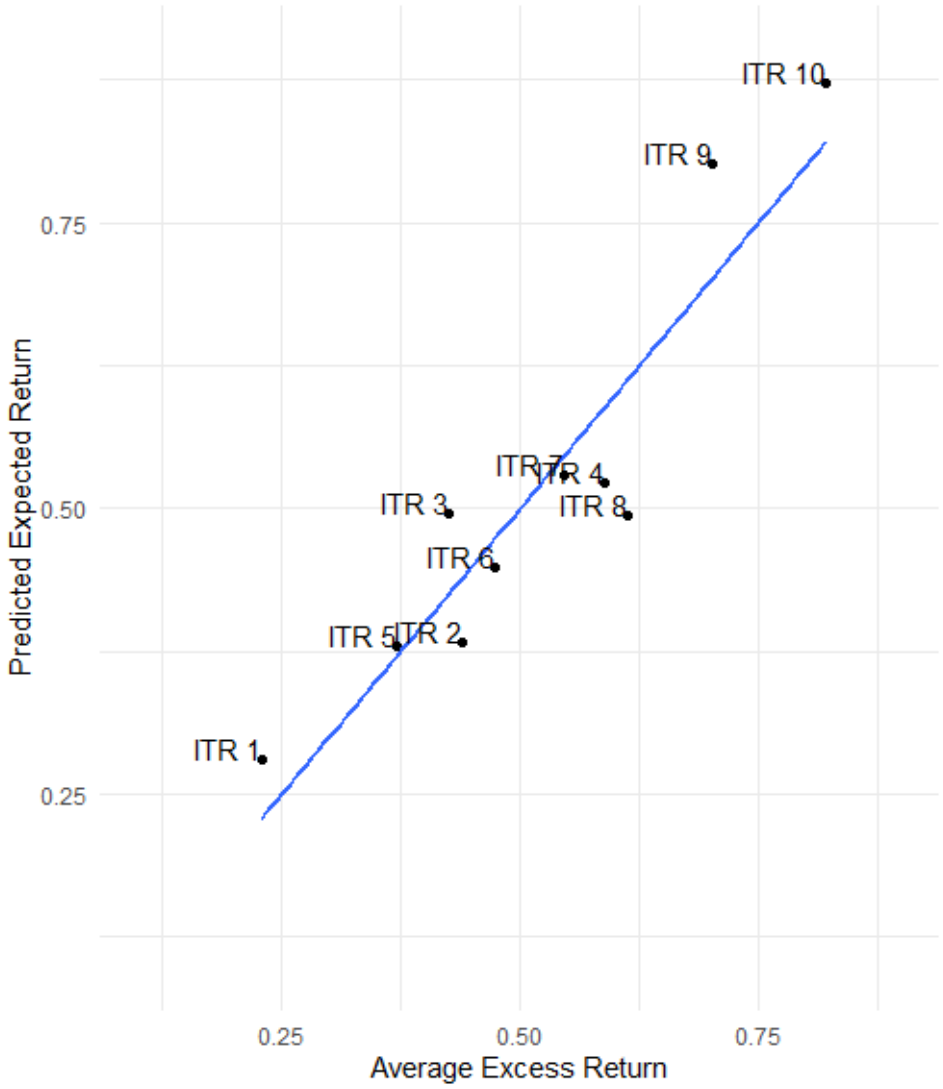
Panel A in Table 4.7 reports results for using the decile portfolios sorted on idiosyncratic tail risk from Section 4.3.3 as test assets, to determine whether the CITR factor explains the idiosyncratic tail risk premium. Column (1) reports results for the pricing model using only the CITR factor. Consistent with the risk-based explanation, CITR is priced in the idiosyncratic tail risk deciles with a statistically significant risk price λ_{CITR} of 1.21%. CITR explains nearly all the differences in average returns of the idiosyncratic tail risk decile portfolios, with an adjusted R^2 of 74% and a small pricing error of 0.08%. Column (2) shows that adding the market factor increases the adjusted R^2 to 86% and reduces the pricing error to 0.06%. In the two-factor model, CITR has positive and statistically significant risk price of 0.99%. Figure 4.5 plots the expected excess returns predicted in the two-factor model against actual returns for the idiosyncratic tail risk deciles. Test assets line up closely on the 45 degree line, indicating

that the model does a good job of pricing these assets. Panel B reports results using the decile portfolios sorted on volume tail risk from Section 4.3.4 as test assets. Column (3) shows that the CITR single factor model has some ability to explain the VTR premium with an adjusted R^2 of 32% and pricing error of 0.11%. CITR has a statistically significant risk price of 1.08%, showing the factor is priced in these portfolios. Column (4) shows that adding the market factor increases the R^2 to 68% and reduces the pricing error to 0.08%. The CITR factor remains statistically significant with a risk price of 0.73%.

To confirm that exposures to CITR explains the idiosyncratic tail risk premium, I examine the $\beta_{k,CITR}$ in each of the idiosyncratic tail risk deciles. Table 4.8 reports the $\beta_{k,M}$ and $\beta_{k,CITR}$ of each decile and the difference between the highest and lowest decile. Panel A shows that in the single factor model, $\beta_{k,CITR}$ is nearly monotonically increasing in idiosyncratic tail risk. The highest idiosyncratic tail risk decile has 0.35 higher $\beta_{k,CITR}$ than the lowest decile. Since the CITR risk price is 1.21%, the factor explains 0.42% ($0.35 \times 1.21\% = 0.42\%$) or roughly two-thirds of the 0.66% idiosyncratic tail risk premium. Hence, the idiosyncratic tail risk can be mostly explained by exposures to CITR. In Panel B, I add the market factor and find that $\beta_{k,CITR}$ is still increasing in idiosyncratic tail risk. The highest idiosyncratic tail risk decile has 0.42 higher $\beta_{k,CITR}$ than the lowest decile. Since the CITR risk price is 0.99%, the factor explains 0.41% of the premium and the two factors combined explain 0.59% (0.18 + 0.41) of the 0.66% premium.

In summary, exposures to CITR helps to explain the abnormal returns in the idiosyncratic tail risk and volume tail risk portfolios. The CITR factor provides a risk-based explanation for the idiosyncratic tail risk premium. Portfolios with high idiosyncratic tail risk earn high average returns due to their high exposure to the CITR factor. Portfolios with low idiosyncratic tail risk earn low average returns, since they have less CITR exposure and hedge against states of low CITR.

Figure 4.5: Realized Versus Predicted Mean Excess Returns of ITR Deciles



Actual average percent excess returns on all anomaly portfolios versus predicted expected returns using exposures to CITR and the market portfolio. The test assets are decile portfolios sorted on idiosyncratic tail risk. Distance from the 45-degree line represents pricing errors (alphas). The monthly sample is from January 2001 to December 2016.

4.5 Robustness

This section examines a battery of robustness tests for my explanation of the idiosyncratic tail risk premium. In my explanation, CITR is correlated to the intermediary marginal value of wealth and its shocks should be positively priced in anomaly portfolios and asset classes

Table 4.8: CITR Betas for Portfolios Sorted on Idiosyncratic Tail Risk

Panel A: One Factor Model: CITR											
	1 (Low ITR)	2	3	4	5	6	7	8	9	10 (High ITR)	(10-1) Beta
β_{CITR}	0.29	0.35	0.37	0.45	0.23	0.34	0.50	0.55	0.55	0.64	0.35
Panel B: Two Factor Model: CITR and Market											
	1 (Low ITR)	2	3	4	5	6	7	8	9	10 (High ITR)	(10-1) Beta
β_M	1.07	0.97	0.99	0.93	0.93	0.94	0.99	0.98	0.92	0.90	-0.17
β_{CITR}	-0.15	-0.05	-0.04	0.06	-0.15	-0.05	0.09	0.14	0.17	0.26	0.42

The table presents factor betas $\beta_{k,CITR}$ and $\beta_{k,M}$ for each idiosyncratic tail risk decile and the difference in factor betas between the highest and lowest decile. CITR beta $\beta_{k,CITR}$ and market beta $\beta_{k,M}$ are estimated in stage one of the Fama and MacBeth regression in (4.21) using the full sample from January 2001 to December 2016.

traded by intermediaries, and should negatively forecast market returns. Additionally, Gabaix et al. (2006)'s economic model theorizes that volume tail risk is a proxy for idiosyncratic tail risk, hence I create a common volume tail risk factor and test whether it is a risk factor that prices the test assets above. Finally, if the idiosyncratic tail risk premium is caused by different exposures to the CITR factor, then the long-short idiosyncratic tail risk portfolio is factor-mimicking portfolio for CITR and can be seen as a traded CITR factor. I test whether this traded CITR factor is priced in the test assets above. I find that the results of these tests are consistent with my explanation.

4.5.1 Pricing Anomaly Portfolios

In my explanation of the ITR premium, CITR shocks are a risk factor and should be priced in portfolios traded by intermediaries. In this section, I examine if the CITR factor can price test assets with anomalous returns. Recent work by Lewellen, Nagel, and Shanken (2010) advocates expanding the set of test assets beyond book-to-market. Motivated by their recommendations, I focus my analysis on less known anomalies in which intermediaries are likely to be the marginal investor. My test assets include portfolios independently double-sorted on size and another asset characteristic, including operating profitability, investment, momentum,

Table 4.9: Asset Pricing Tests on Characteristic Portfolios

	Panel A: Size and Tail Risk Portfolios		Panel B: Size and Characteristic Portfolios				
	25 ITR	25 VTR	25 OP	25 INV	25 MOM	25 IVOL	ALL
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Intercept	0.03	0.44	0.89	1.62	1.35	1.52	1.15
NW t-stat	(0.08)	(1.24)	(2.34)	(3.51)	(2.88)	(4.29)	(3.54)
Shanken t-stat	[0.05]	[0.86]	[2.10]	[3.31]	[2.26]	[3.63]	[2.95]
$R^M - R^f$	0.56	0.21	-0.20	-0.91	-0.61	-0.77	-0.43
NW t-stat	(1.19)	(0.47)	(-0.37)	(-1.56)	(-1.03)	(-1.54)	(-0.90)
Shanken t-stat	[0.92]	[0.36]	[-0.38]	[-1.63]	[-0.97]	[-1.56]	[-0.90]
CITR	1.05	1.05	0.70	0.97	1.01	0.89	0.84
NW t-stat	(2.66)	(2.18)	(1.76)	(2.27)	(2.53)	(2.23)	(2.34)
Shanken t-stat	[2.25]	[1.81]	[1.61]	[1.83]	[1.95]	[1.80]	[2.03]
% Adj. R^2	67.49	50.16	25.08	45.07	43.92	35.17	26.89
% MAE	0.11	0.12	0.17	0.18	0.16	0.20	0.18
Months	192	192	192	192	192	192	192

This table presents asset pricing tests on double-sorted portfolios using the CITR and market two-factor model from 2001 to 2016. In Panel A, the test assets are 25 portfolios conditionally sorted on size and idiosyncratic tail risk or volume tail risk. Stocks are first grouped into size quintiles, then within each size quintile, stocks are grouped by their ITR or VTR. In Panel B, test assets are 25 portfolios independently sorted by size and the characteristic. Stocks are grouped by the intersection of 5 quintiles sorted on size and 5 quintiles sorted on the characteristic. These anomaly portfolios are downloaded from from Kenneth French's website and include operating profitability, investment, momentum, and idiosyncratic volatility. The table reports the risk premia estimates, R^2, and mean absolute pricing errors in percentage terms, Newey-West t-statistics with one lag, and Shanken t-statistics.

and idiosyncratic volatility from January 2001 to December 2016.[27] Anomaly portfolios are sorted by size to have more granular test assets for cross-sectional tests. All anomaly portfolios are downloaded from Kenneth French's website. Additionally, Lewellen, Nagel, and Shanken (2010) advocates including portfolios sorted by exposures to the CITR factor in tests. Since Section 4.4.6 shows that average returns for ITR and VTR sorted portfolios are largely driven by their exposures to the CITR factor, I include portfolios conditionally double-sorted on size then idiosyncratic tail risk and size then volume tail risk.[28]

Table 4.9 reports results for the CITR and market two-factor model. In each column, the

[27] In each month, stocks are sorted into 25 (5x5) groups based on their size and characteristic simultaneously, as the intersection of 5 quintiles sorted on size and 5 quintiles sorted on the characteristic. Value-weighted portfolios are formed in each grouping, and held for one month. See Kenneth French's website for more details.

[28] In each month, stocks are sorted into quintile groupings based on their size. Then, within each size quintile, stocks are sorted into idiosyncratic tail risk or volume tail risk quintiles, value-weighted portfolios are formed in each grouping, and held for one month.

CITR risk price λ_{CITR} is positive and statistically significant at the 10% confidence level. The magnitude of λ_{CITR}s are remarkably similar, ranging from 0.70% to 1.05%, and in line with the tail risk decile long-short return of 0.66% in Table 4.7. The two-factor model explains a large degree of the variation in average returns as indicated by the high adjusted R^2s ranging from 25% for the portfolios sorted by size and operating profitability up to 67% for the portfolios sorted by size and ITR. Likewise, pricing errors are relatively low, ranging from 0.11% for the portfolios sorted by size and ITR to 0.2% for the portfolios sorted by size and idiosyncratic volatility. In particular, the high R^2s and low pricing errors for the tail risk portfolios in Panel A confirm that the CITR factor explains most of the idiosyncratic tail risk and volume tail risk premia. Column (7) reports the results for the all-in portfolio, which includes all 150 test assets from Columns (1) to (6). The CITR risk price remains statistically and economically significant. The cross-section standard deviation of CITR-betas across all 150 test assets is 0.19, hence a one standard deviation increase in an asset's CITR-beta corresponds to a 2% ($0.19 \times 0.84 \times 12 = 2\%$) increase in its annual risk premia. Consistent with my risk-based explanation, CITR shocks are a risk factor that prices anomaly portfolios traded by intermediaries.

4.5.2 Is CITR Just a Proxy for Other Pricing Factors?

A large number of factors have been documented to explain the cross-section of expected equity returns. This section demonstrates that the CITR factor's pricing ability is not explained by existing non-traded factors that have been documented in the literature.[29] Using idiosyncratic tail risk and volume tail risk portfolios as test assets, this section compares the CITR factor's pricing ability to existing liquidity, volatility, tail risk, and intermediary factors.

In Table 4.10, I compare the pricing power of the CITR factor relative to similar factors

[29]Because the CITR factor is non-traded and economically motivated, statistical models such as the Fama and French three factor and five factor models are not useful benchmarks. Traded factors almost always outperform economic factors due to measurement error in the latter (Cochrane (2009)). Hence, this section only compares the CITR factor against other non-traded pricing factors.

Table 4.10: Comparison with Existing Pricing Factors

	Market	CITR	PS-Liqu.	ICR	LCT	CIV	KJ-MTR
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Intercept	-0.01	0.25	0.31	0.35	0.02	0.33	-0.01
NW t-stat	(-0.04)	(0.77)	(0.98)	(1.10)	(0.04)	(1.02)	(-0.04)
Shanken t-stat	[-0.03]	[0.53]	[0.65]	[0.76]	[0.03]	[0.74]	[-0.02]
$R^M - R^f$	0.66	0.36	0.30	0.25	0.55	0.21	0.59
NW t-stat	(1.41)	(0.81)	(0.68)	(0.56)	(1.13)	(0.49)	(1.29)
Shanken t-stat	[1.22]	[0.62]	[0.51]	[0.43]	[0.85]	[0.38]	[0.93]
CITR		1.15	1.19	0.93	0.88	0.81	1.23
NW t-stat		(3.05)	(3.01)	(2.99)	(2.59)	(2.43)	(3.18)
Shanken t-stat		[2.46]	[2.40]	[2.35]	[2.04]	[2.10]	[2.35]
Existing Factor			-0.17	2.30	-4.36	-11.61	0.85
NW t-stat			(-0.23)	(2.34)	(-2.26)	(-2.45)	(2.07)
Shanken t-stat			[-0.17]	[1.93]	[-1.65]	[-2.17]	[1.36]
% Adj. R^2	25.16	58.64	59.50	62.75	62.56	69.53	63.13
% MAE	0.16	0.12	0.12	0.11	0.11	0.10	0.11
Months	192	192	192	192	168	192	192

This table presents asset pricing tests using the CITR single-factor model, the CITR and market two-factor model, and a three-factor model with CITR, the market, and an existing factor at a monthly frequency from January 2001 to December 2016. The test portfolios are the idiosyncratic tail risk deciles and volume tail risk deciles as well as the 25 portfolios conditionally sorted on size and idiosyncratic tail risk and volume tail risk from Table 4.9 for a total of 70 test portfolios. Existing factors include the liquidity factor (PS-Liqu.), intermediary capital factor (ICR), leverage constraint tightness factor (LCT), common idiosyncratic volatility factor (CIV), market tail risk factor (KJ-MTR). The table reports the risk premium estimates, R^2, and mean absolute pricing errors in percentage terms, Newey-West t-statistics with one lag, and Shanken t-statistics.

in the literature by adding the existing factor to the CITR and market two-factor model. Factors considered include the liquidity factor (PS-Liqu.) from Pastor and Stambaugh (2003), intermediary capital ratio factor (ICR) from He, Kelly, and Manela (2017), leverage constraint tightness factor (LCT) from Boguth and Simutin (2018), common idiosyncratic volatility factor (CIV) from Herskovic et al. (2016), market tail risk factor (KJ-MTR) from Kelly and Jiang (2014b) (defined as monthly changes in market tail risk). The test portfolios are the idiosyncratic tail risk deciles and volume tail risk deciles from Table 4.7, and the 25 portfolios conditionally sorted on size and idiosyncratic tail risk and volume tail risk from Table 4.9 for a total of 70 portfolios. Column (1) indicates that the one-factor market model fails to price these assets with a statistically insignificant price of risk and an R^2 of 25.16%. Column (2) shows

that adding the CITR factor increases the R^2 to 58.64%. Columns (3) - (7) show that adding an existing factor only slightly increases the R^2. Adding the CIV factor in column (6) obtains the highest R^2 of 69.53%, which is consistent with the literature that demonstrates volatility and tail risk contain separate information about expected returns (Andersen, Fusari, and Todorov (2020)). Additionally, the CITR factor risk price is essentially unchanged by the inclusion of any existing factor and stays statistically significant.

4.5.3 Pricing Sophisticated Asset Classes

In my explanation, the common idiosyncratic tail risk factor is correlated to the intermediary marginal value of wealth. A natural test of my explanation is to evaluate whether the CITR factor can price the sophisticated asset classes in which large intermediaries are the marginal investor. I use the well-known asset returns from He, Kelly, and Manela (2017) as test assets. The test assets are downloaded from Asaf Manela's website, and include quarterly returns for equities, US government and corporate bonds, sovereign bonds, options, credit default swaps, commodities, and foreign exchange up to December 2012.

To be consistent with their main analysis, I conduct pricing tests at the quarterly frequency. I construct quarterly CITR levels as the average monthly CITR within each quarter and define the quarterly CITR factor as first differences in quarterly CITR levels. I then conduct a two-stage Fama and MacBeth (1973) estimation procedure from January 2001 to December 2012 to estimate the CITR risk premium for each asset class. I analyze both the CITR single-factor model and the CITR and market two-factor model.

Table 4.11 reports the pricing ability of the CITR factor for the asset classes investigated in He, Kelly, and Manela (2017) from January 2001 to December 2012. Due to its correlation with the intermediary marginal value of wealth, the CITR factor should have a positive risk premium, which is what I find in every asset class for the single-factor model. In odd columns, λ_{CITR} is positive and economically large. λ_{CITR} is also statistically significant for the options, CDS, commodities, and foreign exchange test assets in columns (7), (9), (11), and (13), which

Table 4.11: Asset Pricing Tests on Sophisticated Assets, Quarterly

	FF25		Bond		Sov		Options		CDS		Commod		FX		All	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
Intercept	1.27	1.09	0.82	0.82	1.24	1.20	-3.16	-5.49	-0.28	-0.21	-0.09	0.34	-0.57	-1.30	0.12	0.28
NW t-stat	0.96	1.04	4.02	4.05	2.92	3.04	-4.08	-5.08	-5.09	-4.07	-0.07	0.37	-1.05	-2.25	0.30	0.97
Shanken t-stat	1.06	0.77	4.03	4.08	2.80	2.81	-1.86	-3.06	-3.67	-1.72	-0.07	0.30	-0.84	-1.47	0.35	0.78
$R^M - R^f$		0.71		1.35		1.70		5.55		10.09		0.99		9.35		0.95
NW t-stat		0.52		0.68		0.64		2.63		3.93		0.41		3.03		0.67
Shanken t-stat		0.37		0.89		0.62		2.22		2.86		0.42		2.59		0.67
CITR	0.19	0.18	0.56	0.61	0.49	0.44	2.00	-1.07	0.93	-1.16	0.75	0.79	1.24	0.86	0.51	0.75
NW t-stat	0.74	0.65	1.00	1.51	1.33	1.11	2.50	-1.96	2.21	-2.77	2.04	2.18	2.96	1.86	1.42	2.87
Shanken t-stat	0.65	0.59	1.39	1.30	1.52	1.17	1.45	-1.34	2.27	-1.74	1.57	1.63	2.77	1.75	1.35	2.48
% R^2	5.21	5.28	63.37	64.03	72.90	72.97	88.61	95.07	42.83	73.64	26.53	29.88	20.08	37.26	29.91	34.55
% MAE	0.67	0.68	0.30	0.30	0.46	0.47	0.36	0.27	0.19	0.18	1.42	1.41	0.71	0.65	0.92	0.90
Quarters	48	48	48	48	41	41	44	44	47	47	48	48	36	36	48	48

This table presents asset pricing tests using the CITR single-factor model and the CITR and market two-factor model at a quarterly frequency from January 2001 to December 2012. Quarterly CITR is defined as the mean of monthly CITRs within the quarter, and quarterly CITR shocks are defined as quarterly changes in CITR. Test assets are the portfolios in He, Kelly, and Manela (2017) downloaded from Asaf Manela's website. Assets include equities, US government and corporate bonds, sovereign bonds, options, credit default swaps, commodities, and foreign exchange. The table reports the risk premium estimates, R^2, and mean absolute pricing errors in percentage terms, Newey-West t-statistics with one lag, and Shanken t-statistics.

are highly sophisticated assets and are most likely to be traded by large intermediaries. However, for the two-factor model, λ_{CITR} is negative for options and CDS in columns (8) and (10), suggesting that some of the CITR factor's ability to price these assets may be due to its correlation with the market factor. Columns (15) and (16) report results for the one- and two-factor models using all test assets, where λ_{CITR} is positive in both columns and statistically significant in the two-factor model with a risk price of 0.75% (t-stat of 2.87). Furthermore, the two-factor model achieves an R^2 of 34.55% when pricing all assets in Column (16). λ_{CITR} is not statistically significant for equities, bonds, and sovereign bonds in columns (1) to (6), which is likely because these assets are frequently traded by less sophisticated investors, hence intermediaries may not be the marginal investor. In summary, there is some evidence that the CITR factor is priced in highly sophisticated asset classes, further supporting the factor's correlation to the intermediary marginal value of wealth.

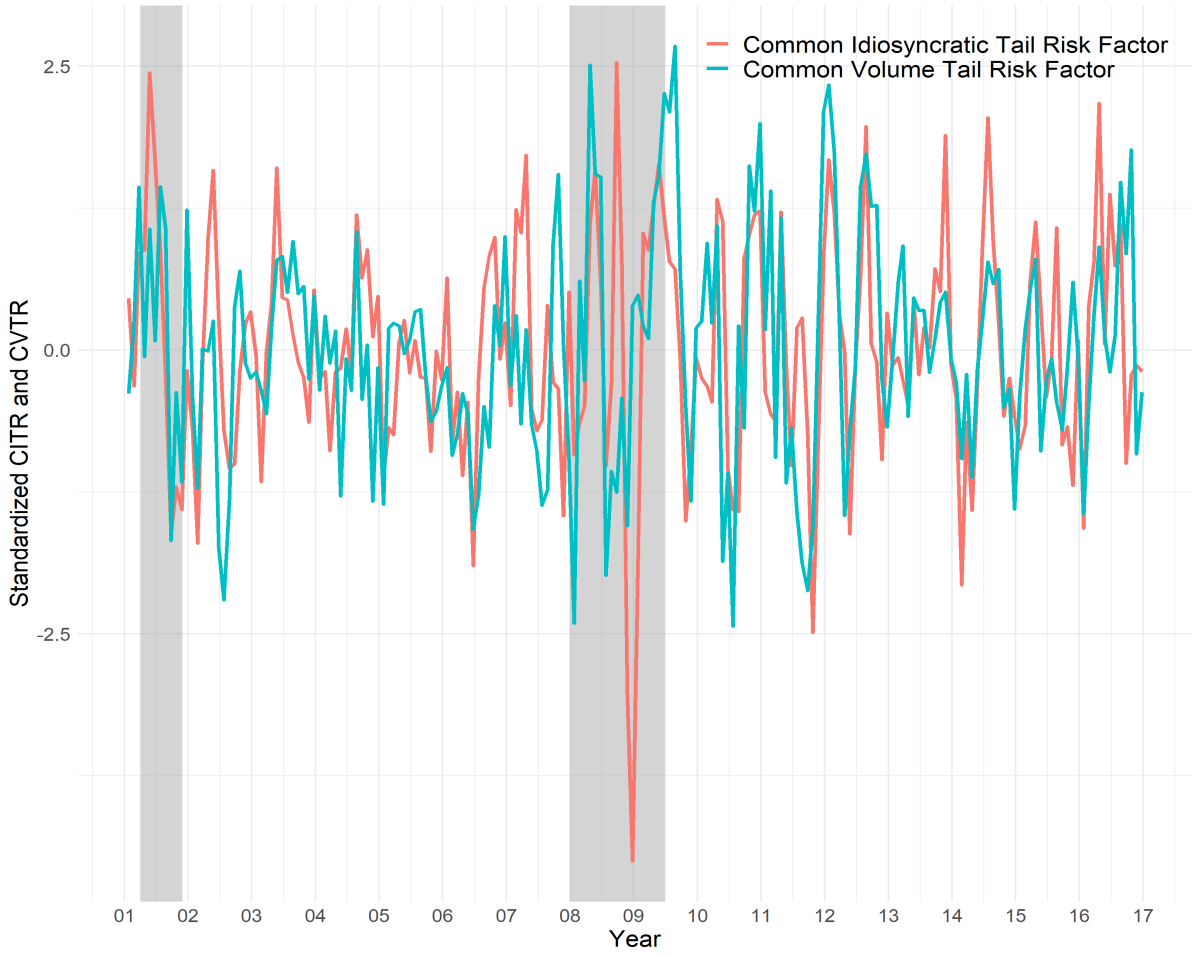
4.5.4 Duality of the Common Idiosyncratic Tail Risk and Common Volume Tail Risk Factors

In my explanation, volume tail risk is a proxy for idiosyncratic tail risk. Section 4.3 presents new evidence of duality in idiosyncratic tail risk and volume tail risk, showing that stocks with higher volume tail risks earn higher average returns, this premium is highly persistent, and that volume tail risk is cross-sectionally correlated to idiosyncratic tail risk. A further test of my explanation is if volume tail risk has a factor structure and if the common volume tail risk factor has similar empirical asset pricing results as the CITR factor, which is exactly what I show in this section, illustrating the duality of the common idiosyncratic tail risk and common volume tail risk factors.

Analogous to Section 4.4 and Equation (4.20), I test the commonality of volume tail risk by regressing each firm's VTR on the mean VTR, denoted as the common volume tail risk (CVTR). Volume tail risk has a stronger factor structure than idiosyncratic tail risk, where CVTR explains 25.18% of the monthly variation in firm-level volume tail risk and 46.27% of the annual variation in firm-level volume tail risk.

In my explanation, volume tail risk is driven by large intermediary trades, and innovations in CVTR should be driven by innovations in intermediary funding. I define the CVTR factor $\triangle CVTR$ as 3-month differences in CVTR levels, that is $\triangle CVTR_t = CVTR_t - CVTR_{t-3}$. Figure 4.6 plots the CVTR and CITR factors, showing that they are highly correlated and procyclical. The common volume tail risk factor has a monthly pair-wise correlation of 0.46 (t-stat of 7.18) with the common idiosyncratic tail risk factor. While idiosyncratic tail risk and volume tail risk may be cross-sectionally correlated due to other characteristics like size or idiosyncratic volatility, the high correlation between CITR and CVTR is not automatic and there is no mechanical reason why the two factors should be correlated over time. Furthermore, the CVTR factor has a monthly pair-wise correlation of 0.35 (t-stat of 5.19) with the intermediary capital factor of He, Kelly, and Manela (2017). These empirical findings strongly support Gabaix et al. (2006)'s hypothesis that both idiosyncratic tail risk and volume tail risk are caused by large

Figure 4.6: Common Idiosyncratic Tail Risk and Volume Tail Risk Factors



The common idiosyncratic tail risk factor and the common volume tail risk factor. Both time-series are standardized to zero mean and unit variance. The monthly sample is from January 2001 to December 2016. CITR is the mean monthly idiosyncratic tail risk and common volume tail risk is the mean monthly volume tail risk.

intermediary trades and my explanation that the CITR and CVTR factors are dually driven by shocks to intermediary funding.

Next, I test whether shocks to CVTR is a priced risk factor by analyzing portfolios with different exposures to CVTR using the same methodology as in Section 4.4.5. For each month from January 2001 to December 2016, I estimate each stock's CVTR-beta by regressing the stock's monthly excess return on the CVTR factor using a 36-month trailing window. I sort stocks in each month into quintiles based on their CVTR-betas. Stocks in quintile 1 have the

Table 4.12: Excess Returns and Alphas of Portfolios Sorted on CVTR-betas

	1 (Low)	2	3	4	5 (High)	(5-1)
Panel A: Average CVTR-beta						
Average CVTR-beta	-0.54	0.11	0.47	0.85	1.67	
Panel B: Univariate Sort on CVTR-beta (Equal-Weighted)						
Excess Return	0.52	0.71	0.78	0.84	0.84	0.32
t-stat	(1.24)	(1.82)	(2.04)	(1.93)	(1.46)	(1.07)
FFC4 alpha	0.57	0.75	0.83	0.89	0.88	0.30
FFC4 t-stat	(1.40)	(2.05)	(2.26)	(2.18)	(1.71)	(1.10)
FF5 alpha	0.78	0.86	0.92	0.95	0.92	0.14
FF5 t-stat	(1.94)	(2.22)	(2.35)	(2.23)	(1.75)	(0.56)
Panel C: Univariate Sort on CVTR-beta (Value-Weighted)						
Excess Return	0.07	0.38	0.60	0.52	0.85	0.77
t-stat	(0.17)	(1.21)	(1.83)	(1.43)	(1.66)	(2.25)
FFC4 alpha	0.10	0.41	0.66	0.56	0.91	0.80
FFC4 t-stat	(0.27)	(1.40)	(2.10)	(1.65)	(1.95)	(2.21)
FF5 alpha	0.27	0.51	0.76	0.65	0.98	0.72
FF5 t-stat	(0.74)	(1.71)	(2.34)	(1.92)	(2.09)	(2.09)

The table reports monthly average CVTR-beta, excess returns, and alphas for portfolios sorted on CVTR-betas between January 2001 to December 2016. Panel A reports the average CVTR-beta for portfolios sorted on CVTR-beta. Panel B reports equally-weighted excess returns and alphas sorted on CVTR-beta. Panel C reports value-weighted excess returns and alphas sorted on CVTR-beta.

least exposure to CVTR and stocks in quintile 5 have the most CVTR exposure.

Table 4.12 reports results for the univariate sort. Panel A shows that CVTR-betas range from -0.54 to 1.67, which is slightly less disperse than CITR-betas. Panel B shows that for the equal-weighted results, the long-short portfolio return is not statistically significant, but still has a positive average return of 0.32%. In comparison, the CITR-beta long-short portfolio has a similar equal-weighted return of 0.40%. Panel C shows that for the value-weighted results, the long-short portfolio return is 0.77% (t-stat of 2.25), which is higher than the 0.62% for CITR-beta. Both equal-weighted and value-weighted returns and alphas are generally increasing in CVTR-betas. These results are consistent with my explanation that shocks to CVTR are a priced risk factor, are driven by shocks in intermediary funding, and has a pricing duality with CITR.

Table 4.13: Asset Pricing Tests on Anomaly Portfolios using the CVTR factor

	Panel A: Tail Risk Deciles		Panel B: Size & Tail Risk		Panel C: Size & Characteristic				
	ITR Dec.	VTR Dec.	25 ITR	25 VTR	25 OP	25 INV	25 MOM	25 IVOL	ALL-DS
Intercept	1.05	1.58	-0.47	0.03	0.79	1.61	0.94	1.38	0.86
NW t-stat	(1.02)	(1.90)	(-1.16)	(0.08)	(2.08)	(3.52)	(2.19)	(4.18)	(2.92)
Shanken t-stat	[0.71]	[1.23]	[-0.86]	[0.06]	[1.70]	[3.23]	[1.72]	[2.84]	[2.41]
$R^M - R^f$	-0.59	-1.08	1.05	0.64	-0.07	-0.88	-0.21	-0.72	-0.13
NW t-stat	(-0.53)	(-1.15)	(1.97)	(1.33)	(-0.15)	(-1.55)	(-0.35)	(-1.44)	(-0.27)
Shanken t-stat	[-0.38]	[-0.79]	[1.62]	[1.09]	[-0.13]	[-1.56]	[-0.34]	[-1.29]	[-0.27]
CVTR	1.39	2.02	1.15	1.08	0.88	1.50	1.48	2.26	0.86
NW t-stat	(1.45)	(2.51)	(2.19)	(2.26)	(1.20)	(2.21)	(2.07)	(2.97)	(1.75)
Shanken t-stat	[1.11]	[1.64]	[2.14]	[1.97]	[1.30]	[2.02]	[1.67]	[2.36]	[1.83]
% R^2	47.75	80.80	50.77	29.82	7.04	30.00	28.62	40.85	8.47
% MAE	0.11	0.06	0.14	0.14	0.19	0.20	0.18	0.20	0.21
Months	192	192	192	192	192	192	192	192	192

This table presents asset pricing tests on double-sorted portfolios using the CVTR and market two-factor model from 2001 to 2016. In Panel A, the test assets are the decile portfolios sorted on idiosyncratic tail risk or volume tail risk examined in Table 4.7. In Panel B, the test assets are 25 portfolios conditionally sorted on size and idiosyncratic tail risk or volume tail risk. Stocks are first grouped into size quintiles, then within each size quintile, stocks are grouped by their ITR or VTR. In Panel C, test assets are 25 portfolios independently sorted by size and the characteristic. Stocks are grouped by the intersection of 5 quintiles sorted on size and 5 quintiles sorted on the characteristic. These anomaly portfolios are downloaded from from Kenneth French's website and include operating profitability, investment, momentum, and idiosyncratic volatility. The table reports the risk premia estimates, R^2, and mean absolute pricing errors in percentage terms, Newey-West t-statistics with one lag, and Shanken t-statistics.

Next, Table 4.13 reports the ability of the CVTR and market two-factor model to price the test assets in Sections 4.4.6 and 4.5.1. CVTR explains most of the VTR deciles with an R^2 of 81%, and has some ability to price the idiosyncratic tail risk deciles with an R^2 of 48%. The VTR risk price λ_{CVTR} is positive for all test assets and statistically significant for many. Column (9) reports results using all double-sorted portfolios as test assets, where the CVTR factor has a risk price of 0.86% (t-stat of 1.75). The magnitude of the CVTR risk prices λ_{CVTR} are in line with the magnitude of the CITR risk prices λ_{CITR}, further supporting the duality between risk factors.

Finally, Table 4.14 reports the pricing ability of the CVTR single-factor and the CVTR and market two-factor models on the sophisticated asset classes investigated in He, Kelly, and Manela (2017) from January 2001 to December 2012 at the quarterly frequency. In each asset class, the CVTR risk price is positive for both the single-factor and two-factor models.

Table 4.14: Asset Pricing Tests on Sophisticated Assets using the CVTR factor, Quarterly

	FF25		Bond		Sov		Options		CDS		Commod		FX		All	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
Intercept	0.60	1.11	0.77	0.83	0.94	0.91	-2.71	8.06	-0.32	-0.20	-0.52	-0.01	-1.78	-1.77	0.03	0.00
NW t-stat	0.50	0.94	4.11	4.88	2.72	2.46	-3.58	2.08	-5.31	-5.63	-0.50	-0.01	-2.89	-2.91	0.12	0.44
Shanken t-stat	0.57	0.67	3.99	4.83	2.14	1.81	-2.57	0.54	-3.89	-2.22	-0.55	-0.01	-1.34	-1.44	0.15	0.41
$R^M - R^f$		0.56		1.55		1.50		-9.38		-0.75		-0.17		7.92		0.86
NW t-stat		0.39		0.81		0.60		-1.66		-0.28		-0.07		2.53		0.62
Shanken t-stat		0.26		0.72		0.52		-0.55		-0.24		-0.08		1.52		0.60
VITR	0.48	0.61	0.72	1.06	0.99	1.22	1.41	6.35	1.56	3.04	0.77	0.86	3.52	3.27	0.62	1.05
NW t-stat	1.00	1.30	0.99	1.47	1.32	1.00	2.57	3.20	2.64	2.97	1.85	2.25	4.00	4.00	1.23	2.62
Shanken t-stat	0.96	1.22	1.47	1.27	1.44	0.95	2.45	0.79	2.84	2.33	1.67	1.86	2.29	1.82	1.23	2.07
% R^2	9.16	10.23	65.14	61.98	72.31	72.42	95.21	98.34	67.11	71.81	23.27	34.05	45.56	45.74	27.39	37.49
% MAE	0.66	0.65	0.29	0.30	0.50	0.49	0.25	0.15	0.19	0.16	1.48	1.38	0.66	0.65	0.93	0.85
Months	48	48	48	44	41	41	44	44	47	47	48	48	36	36	48	48

This table presents asset pricing tests using the CVTR single-factor model and the CVTR and market two-factor model at a quarterly frequency from January 2001 to December 2012. Quarterly CVTR is defined as the mean of monthly CVTRs within the quarter, and quarterly CVTR shocks are defined as quarterly changes in CVTR. Test assets are the portfolios in He, Kelly, and Manela (2017) downloaded from Asaf Manela's website. Assets include equities, US government and corporate bonds, sovereign bonds, options, credit default swaps, commodities, and foreign exchange. The table reports the risk premium estimates, R^2, and mean absolute pricing errors in percentage terms, Newey-West t-statistics with one lag, and Shanken t-statistics.

Analogous to the CITR results, the CVTR risk price is statistically significant for options, CDS, commodities, and foreign exchange while not being statistically significant for equities, bonds, and sovereign bonds. Using all assets in column (16), CVTR has a risk price of 1.05 (t-stat of 2.62). Furthermore, for each asset class the R^2 and pricing errors are similar in magnitude for the CITR and CVTR results. The duality of CITR and CVTR in pricing sophisticated assets further links these factors to intermediaries.

This section presents new results on the duality of the tail distributions of idiosyncratic returns and trading volume, estimated on two distinct sources of data. Like idiosyncratic tail risk, volume tail risk follows a factor structure and shocks to common volume tail risk is a priced risk factor. These results are not automatic and strongly support the hypothesis in this paper that shocks to intermediary funding drive the time-series variation in common idiosyncratic tail risk and common volume tail risk.

4.5.5 Pricing Assets using a Traded CITR Factor

Section 4.4.6 demonstrates that the idiosyncratic tail risk premium is explained by exposures to the CITR factor. If the idiosyncratic tail risk deciles are primarily driven by the CITR factor,

then the ITR long-short portfolio itself can be used as a traded factor that mimics the non-traded CITR factor. I define the traded CITR factor as the value-weighted return on a portfolio that goes long the highest idiosyncratic tail risk decile and shorts the lowest idiosyncratic tail risk decile.[30] The traded CITR factor should price assets better than the non-traded CITR factor, due to the measurement error in the latter.[31]

This section investigates the ability of the traded CITR factor to price assets. According to my risk-based explanation, the traded CITR factor is correlated to the intermediary marginal value of wealth, and should price assets in which intermediaries are the marginal investor. I test this hypothesis by evaluating the pricing of the traded CITR factor on the previous anomalies and the sophisticated assets classes in He, Kelly, and Manela (2017).

Table 4.15 reports the ability of the traded CITR and the market two-factor model to price the same portfolios in Sections 4.4.6 and 4.5.1. The traded CITR factor prices the test assets similarly to the non-traded CITR factor, validating that it is an appropriate factor-mimicking portfolio. In each column, the traded CITR factor risk price is statistically significant and economically large. Since the traded CITR factor is a traded portfolio, its risk price can be interpreted as a monthly excess return. Estimates of the traded CITR factor risk premium is economically large, ranging from 0.63% for the idiosyncratic tail risk deciles to 1.80% for the portfolios sorted on idiosyncratic volatility. Column (9) reports results using all double-sorted portfolios as test assets, where traded CITR factor has a risk price of 0.76% (t-stat of 2.5), R^2 of 71.74%, and a pricing error of 0.09. This traded CITR factor risk price is in line with the long-short return of 0.62% for quintile portfolios sorted on CITR-betas in Table 4.6.

Table 4.16 reports the pricing ability of the traded CITR factor for the asset classes inves-

[30] Motivated by to Fama and French (1993), I use a characteristic-managed portfolio to create the traded CITR factor. The finance literature on characteristic-managed portfolios include Feng, Giglio, and Xiu (2019), Giglio and Xiu (2018), Gu, Kelly, and Xiu (2020b), Kelly, Pruitt, and Su (2019b), and Kozak, Nagel, and Santosh (2020).

[31] Economically motivated macro variables, such as CITR, will always have measurement error. Due to measurement error, the factor-mimicking portfolio will always price assets better than an estimate of the underlying factor that uses measured macroeconomic variables Cochrane (2009). Additionally, the traded CITR factor can be a factor-mimicking portfolio for CITR, since they both are both driven by intermediary funding. Using the idiosyncratic tail risk premium may be statistically preferred to the projection of non-traded CITR on stocks, since it avoids the errors-in-variables bias and variance issues associated with estimating factor loadings plagued by many other factor-mimicking approaches.

Table 4.15: Asset Pricing Tests on Anomaly Portfolios using the Traded CITR factor

	Panel A: Tail Risk Deciles		Panel B: Size & Tail Risk		Panel C: Size & Characteristic				
	ITR Dec.	VTR Dec.	25 ITR	25 VTR	25 OP	25 INV	25 MOM	25 IVOL	ALL-DS
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Intercept	0.35	2.81	-0.03	0.54	0.46	0.73	0.44	0.58	-0.28
NW t-stat	(0.35)	(3.71)	(-0.10)	(1.80)	(1.39)	(2.07)	(1.02)	(2.11)	(-0.72)
Shanken t-stat	[0.30]	[2.92]	[-0.55]	[-0.07]	[1.06]	[1.70]	[0.82]	[1.62]	[1.27]
$R^M - R^f$	0.12	-2.39	0.55	-0.03	0.06	-0.21	0.02	-0.13	0.78
NW t-stat	(0.12)	(-2.81)	(1.19)	(-0.05)	(0.11)	(-0.40)	(0.03)	(-0.25)	(1.57)
Shanken t-stat	[0.10]	[-2.31]	[1.32]	[0.98]	[0.11]	[-0.39]	[0.03]	[-0.26]	[0.08]
Traded CITR	0.63	1.44	1.05	1.50	1.37	1.48	1.79	1.80	0.76
NW t-stat	(2.74)	(2.90)	(2.38)	(3.46)	(2.65)	(3.09)	(3.16)	(3.82)	(2.50)
Shanken t-stat	[2.50]	[2.46]	[2.17]	[2.15]	[2.67]	[2.96]	[2.46]	[3.31]	[3.06]
% Adj. R^2	80.50	77.03	73.43	71.94	75.54	82.16	66.87	88.75	71.74
% MAE	0.06	0.08	0.09	0.11	0.09	0.10	0.12	0.08	0.09
Months	192	192	192	192	192	192	192	192	192

This table presents asset pricing tests on double-sorted portfolios using the traded CITR and market two-factor model from 2001 to 2016. In Panel A, the test assets are the decile portfolios sorted on idiosyncratic tail risk or volume tail risk examined in Table 4.7. In Panel B, the test assets are 25 portfolios conditionally sorted on size and idiosyncratic tail risk or volume tail risk. Stocks are first grouped into size quintiles, then within each size quintile, stocks are grouped by their ITR or VTR. In Panel C, test assets are 25 portfolios independently sorted by size and the characteristic. Stocks are grouped by the intersection of 5 quintiles sorted on size and 5 quintiles sorted on the characteristic. These anomaly portfolios are downloaded from from Kenneth French's website and include operating profitability, investment, momentum, and idiosyncratic volatility. The table reports the risk premia estimates, R^2, and mean absolute pricing errors in percentage terms, Newey-West t-statistics with one lag, and Shanken t-statistics.

tigated in He, Kelly, and Manela (2017) from January 2001 to December 2012 at the monthly frequency. In each asset class except for sovereign bonds, the traded CITR factor risk price is positive, economically large, and statistically significant at the 10% level. This is strong evidence that supports the link between the traded CITR factor and the intermediary marginal value of wealth. Column (16) reports the two-factor model results using all test assets. The traded CITR factor risk premium across all the asset classes is 1.80% and statistically significant at the 1% level. The two-factor model provides the best fit for options with an R^2 of 95% and an R^2 of 29.74% in Column (16) using all asset classes.

The results in this section show that the traded CITR factor, defined as the long-short ITR portfolio, has similar pricing results as the non-traded CITR factor, supporting my explanation that the idiosyncratic tail risk premium is driven by the CITR factor. Furthermore, if the ITR

Table 4.16: Tests on HKM Portfolios using CITR Factor-Mimicking Portfolio, Monthly

	FF25		Bond		Sov		Options		CDS		Commod		FX		All	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
Intercept	0.61	0.16	0.24	0.22	0.50	0.29	5.93	0.67	0.00	-0.10	0.38	0.27	0.06	-0.66	0.35	0.15
NW t-stat	1.09	0.35	3.63	3.45	2.00	1.63	1.82	1.09	-0.02	-5.08	0.90	0.98	0.27	-2.97	1.43	1.48
Shanken t-stat	1.26	0.28	2.88	3.17	1.65	1.40	0.36	0.42	-0.01	-2.53	0.95	0.87	0.31	-2.11	1.68	1.46
$R^M - R^f$		0.14		0.25		1.53		-0.46		-0.03		-0.09		2.73		0.03
NW t-stat		0.21		0.33		1.64		-0.56		-0.03		-0.11		2.28		0.06
Shanken t-stat		0.20		0.57		1.65		-0.28		-0.02		-0.12		1.88		0.07
Tr. CITR	1.19	1.15	4.11	2.19	-5.11	-2.26	27.41	9.44	12.94	7.54	1.60	1.66	2.61	3.01	1.53	1.80
NW t-stat	1.72	1.64	2.43	1.77	-2.06	-1.73	2.03	2.93	4.73	4.65	1.87	1.87	1.95	2.34	2.69	3.44
Shanken t-stat	2.09	2.00	2.33	2.19	-1.32	-1.21	0.41	1.30	1.46	2.36	1.47	1.49	1.89	1.90	2.36	3.01
% R^2	45.01	48.49	26.05	69.79	64.00	74.69	88.67	95.63	59.82	81.71	12.25	13.12	10.51	34.93	16.36	29.74
% MAE	0.17	0.16	0.14	0.08	0.15	0.14	0.14	0.09	0.09	0.06	0.56	0.55	0.25	0.23	0.35	0.29
Months	144	144	144	144	124	124	133	133	143	143	144	144	109	109	144	144

This table presents asset pricing tests using the traded CITR single-factor model and the traded CITR and market two-factor model at a monthly frequency from January 2001 to December 2012. Traded CITR is defined as the value-weighted return on a portfolio that goes long the highest idiosyncratic tail risk decile and shorts the lowest idiosyncratic tail risk decile. Test assets are the portfolios in He, Kelly, and Manela (2017) downloaded from Asaf Manela's website. Assets include equities, US government and corporate bonds, sovereign bonds, options, credit default swaps, commodities, and foreign exchange. The table reports the risk premium estimates, R^2, and mean absolute pricing errors in percentage terms, Newey-West t-statistics with one lag, and Shanken t-statistics.

and VTR premia are both driven by shocks to the intermediary marginal value of wealth, then the traded VTR factor, defined as the VTR long-short portfolio, should have similar pricing abilities as the traded CITR factor, which is what I find in the data. Section C.4.4 investigates the pricing ability of traded VTR to price anomalies and sophisticated assets. Tables C.10 and C.11 show that traded VTR has nearly the same pricing ability as the traded CITR factor, providing further support for the empirical duality between ITR and VTR and their risk premia.

4.5.6 Forecasting the Equity Market Premium

A common prediction of intermediary asset pricing models is that expected returns are a function of lagged state variables that capture financial sector distress. In my explanation, CITR is negatively related to the shadow cost of capital for intermediaries, implying it should negatively forecast the equity market risk premium. I test this hypothesis using a monthly time-series regression of market returns on CITR. Denote $r_{[t,t+h]}$ as the CRSP value-weighted return from time t to $t + h$, where time t is measured in months. The return regression is

$$r_{[t,t+h]} = a_h + b_h CITR_t + c_h X_t + \epsilon_{t,t+h}, \quad t = 1, 2, ..., T - h, \tag{4.23}$$

Table 4.17: Equity Market Premium Time-Series Predictability

Horizon	1	4	6	8	12	16	20	24
				Panel A: Univariate Regression				
CITR	-9.56	-7.24	-7.09	-5.79	-5.31	-4.81	-4.09	-3.65
Hodrick t-stat	(-2.02)	(-2.05)	(-2.31)	(-2.20)	(-2.36)	(-2.01)	(-1.85)	(-2.05)
Intercept	4.28	4.87	4.99	5.12	5.40	5.68	6.05	6.47
Hodrick t-stat	(1.17)	(1.35)	(1.37)	(1.40)	(1.49)	(1.57)	(1.69)	(1.87)
Adj. R^2	3.03	6.78	9.02	7.75	9.83	10.86	9.57	8.72
				Panel B: Fama French Carhart Regression				
CITR	-8.48	-6.57	-5.97	-5.17	-5.12	-4.79	-4.01	-3.39
Hodrick t-stat	(-1.85)	(-1.93)	(-2.03)	(-1.99)	(-2.26)	(-2.03)	(-1.84)	(-1.96)
Intercept	4.32	5.14	5.56	5.47	5.62	5.69	6.08	6.55
Hodrick t-stat	(1.09)	(1.37)	(1.49)	(1.49)	(1.53)	(1.55)	(1.67)	(1.86)
Market	1.88	0.62	0.32	0.25	0.03	0.17	0.13	0.05
Hodrick t-stat	(1.29)	(0.72)	(0.44)	(0.40)	(0.06)	(0.40)	(0.34)	(0.15)
SMB	-2.08	-1.18	-1.30	-0.82	-0.32	-0.16	-0.23	-0.09
Hodrick t-stat	(-1.27)	(-1.41)	(-1.95)	(-1.37)	(-0.73)	(-0.47)	(-0.76)	(-0.34)
HML	-0.38	-0.25	-0.81	-0.40	-0.06	0.08	0.07	-0.22
Hodrick t-stat	(-0.19)	(-0.28)	(-1.26)	(-0.88)	(-0.17)	(0.28)	(0.27)	(-0.90)
UMD	-0.15	-0.67	-0.74	-0.70	-0.56	-0.26	-0.27	-0.24
Hodrick t-stat	(-0.14)	(-1.54)	(-2.10)	(-2.28)	(-2.32)	(-1.21)	(-1.29)	(-1.34)
% Adj. R^2	3.97	9.08	13.42	11.10	11.27	10.49	9.31	8.18

This table shows the results of regression (14) for a horizon of 1 - 24 months. The top panel is a regression of returns on only CITR and the bottom panel is a regression of returns on CITR and the Fama French Carhart factors. The first row in each panel shows the regression coefficient for a one standard deviation increase in CITR. The second row shows the t-statistic calculated using Hodrick standard errors. The third row shows the adjusted-R^2.

where $CITR_t$ is standardized to have zero mean and unit variance, X_t refers to other explanatory variables, and the horizon ranges from $h = 1$ (one month) to $h = 24$ (two years). I report t-statistics calculated using Hodrick (1992) standard errors to account for the overlapping returns when $h > 1$.[32]

Panel A in Table 4.17 presents the regression results with no control variables. A one standard deviation increase in CITR results in a decrease in future returns of 9.6%, 5.3%, 3.7% at the 1 month, 1 year, and 2 year horizons. The respective t-statistics are all significant at the

[32]Ang and Bekaert (2007) demonstrate through simulation that Hodrick (1992)'s standard error correction provide the most conservative test statistics relative to other commonly used procedures. I also find that Hodrick's correction produces more conservative t-statistics than those calculated using Newey and West (1987b) standard errors lag length equal to $2 \times h$.

95% confidence level. Panel B in Table 4.17 presents the results for the regression controlling for Fama-French-Carhart factors. The coefficients remain economically and statistically significant at the 10% confidence level. The regression R^2s are also relatively high, increasing from 3% at the 1-month horizon up to 11% at the 16-month horizon. These results further support the hypothesis that CITR is a priced risk factor and is driven by intermediary funding. The results are inconsistent with the fire-sale or labour income risk explanation of idiosyncratic tail risk, in which a high average ITR should predict positive average returns.

4.6 Conclusion

The idiosyncratic tail risk premium is a significant recent discovery in the asset pricing literature. Savor (2012) and Jiang and Zhu (2017) interpret the idiosyncratic tail risk as news shocks, and argues its premium is driven by under-reaction to firm-specific news. Begin, Dorion, and Gauthier (2019) and Kapadia and Zekhnini (2019) argue that the premium is caused by the inability to diversify and persists due to limits to arbitrage.

My paper offers a risk-based explanation for the idiosyncratic tail risk premium. I show that idiosyncratic tail risk is driven by intermediary funding and the common idiosyncratic tail risk factor is correlated to the intermediary marginal value of wealth. Stocks with high idiosyncratic tail risk also have high exposures to the common idiosyncratic tail risk (CITR) factor, earning a risk premium due to their low returns when intermediary constraints tighten.

I test my explanation using a new measure of idiosyncratic tail risk. First, I show that stocks with high idiosyncratic tail risk earn higher average returns that persists over years. Second, idiosyncratic tail risk has a strong firm-level correlation to volume tail risk and intermediary trading volume as a percentage of total trading volume. Third, idiosyncratic tail risk follows a strong factor structure, and the common idiosyncratic tail risk factor is procyclical and correlated to existing intermediary factors. Fourth, the common idiosyncratic tail risk factor explains cross-sectional differences in average returns, including the idiosyncratic tail

risk premium. I confirm that high (low) idiosyncratic tail risk portfolios have high (low) exposures to the common idiosyncratic tail risk factor, and that the difference can explain most of the premium. Fifth, the common idiosyncratic tail risk factor is priced in other anomalies and sophisticated assets, and forecasts the equity market premium. Sixth, the idiosyncratic tail risk long-short portfolio is a factor-mimicking portfolio for the CITR factor and is priced in anomalies and sophisticated assets. Finally, I document that these asset pricing results are similar when using volume tail risk in the place of idiosyncratic tail risk. Volume tail risk earns a persistent premium and exhibits commonality that's correlated to intermediary factors. The common volume tail risk factor is also priced in anomaly portfolios and sophisticated assets. This duality between idiosyncratic tail risk and volume tail risk in asset pricing provides new empirical evidence for the intermediary hypothesis of idiosyncratic tail risk from Gabaix et al. (2006).

Some of these results allow me to distinguish my hypothesis from leading alternative explanations of the idiosyncratic tail risk premium. Savor (2012) and Jiang and Zhu (2017) propose behavioral explanations for the premium based on short-term under-reaction to news shocks caused by limited investor inattention. Their short-term explanation is inconsistent with my finding that the idiosyncratic tail risk premium is persistent. Additionally, if idiosyncratic tail risk were only caused by news shocks, then it would not exhibit such a strong factor structure. Begin, Dorion, and Gauthier (2019) and Kapadia and Zekhnini (2019) propose that the premium is caused by the inability to hedge idiosyncratic tail risk in a diversified portfolio, and that the premium persists due to limits to arbitrage. This explanation is inconsistent with my finding that the premium can be explained by differences in exposure to the common idiosyncratic tail risk factor.

Finally, my explanation and empirical findings support the large intermediary hypothesis of idiosyncratic tail risk and volume tail risk by Gabaix et al. (2006). Their economic model assumes intermediary funding is constant, while my explanation allows intermediary funding to change, which provides a stylized framework linking intermediaries, asset prices, and the

tail distributions of returns and trading volume.

Bibliography

[1] Tobias Adrian, Erkko Etula, and Tyler Muir. "Financial intermediaries and the cross-section of asset returns". In: *The Journal of Finance* 69.6 (2014), pp. 2557–2596.

[2] Tobias Adrian and Hyun Song Shin. "Liquidity and leverage". In: *Journal of financial intermediation* 19.3 (2010), pp. 418–437.

[3] Tobias Adrian and Hyun Song Shin. "Procyclical Leverage and Value-at-Risk". In: *The Review of Financial Studies* 27.2 (2014), pp. 373–403.

[4] Yacine Ait-Sahalia, Ilze Kalnina, and Dacheng Xiu. "High-frequency factor models and regressions". In: *Journal of Econometrics* (2020).

[5] Laura Alessandretti et al. "Anticipating Cryptocurrency Prices Using Machine Learning". In: *Complexity* 2018 (2018).

[6] Diego Amaya et al. "Does Realized Skewness Predict the Cross-Section of Equity Returns?" In: *Journal of Financial Economics* 118.1 (2015), pp. 135–167.

[7] Yakov Amihud. "Illiquidity and Stock Returns: Cross-Section and Time-Series Effects". In: *Journal of Financial Markets* 5.1 (2002), pp. 31–56.

[8] Torben G Andersen, Nicola Fusari, and Viktor Todorov. "The pricing of tail risk and the equity premium: Evidence from international option markets". In: *Journal of Business and Economic Statistics* 38.3 (2020), pp. 662–678.

[9] Andrew Ang and Geert Bekaert. "Stock Return Predictability: Is It There?" In: *The Review of Financial Studies* 20.3 (2007), pp. 651–707.

[10] Andrew Ang, Joseph Chen, and Yuhang Xing. “Downside Risk”. In: *The Review of Financial Studies* 19.4 (2006), pp. 1191–1239.

[11] Andrew Ang et al. “The Cross-Section of Volatility and Expected Returns”. In: *The Journal of Finance* 61.1 (2006), pp. 259–299.

[12] Philippe Artzner et al. “Coherent Measures of Risk”. In: *Mathematical Finance* 9.3 (1999), pp. 203–228.

[13] Turan G Bali, Nusret Cakici, and Robert F Whitelaw. “Maxing Out: Stocks as Lotteries and the Cross-Section of Expected Returns”. In: *Journal of Financial Economics* 99.2 (2011), pp. 427–446.

[14] August Balkema and Laurens de Haan. “Residual Life Time at Great Age”. In: *The Annals of Probability* 2.5 (1974), pp. 792–804.

[15] Guus Balkema and Paul Embrechts. “Linear regression for heavy tails”. In: *Risks* 6.3 (2018), p. 93.

[16] Flavio Barboza, Herbert Kimura, and Edward Altman. “Machine Learning Models and Bankruptcy Prediction”. In: *Expert Systems with Applications* 83 (2017), pp. 405–417.

[17] Ole E Barndorff-Nielsen et al. “Designing Realized Kernels to Measure the Ex Post Variation of Equity Prices in the Presence of Noise”. In: *Econometrica* 76.6 (2008), pp. 1481–1536.

[18] Giovanni Barone-Adesi, Frederick Bourgoin, and Kostas Giannopoulos. “Don’t Look Back”. In: *Working Paper* (1998).

[19] Sebastian Bayer and Timo Dimitriadis. “Regression Based Expected Shortfall Backtesting”. In: *arXiv preprint arXiv:1801.04112* (2018).

[20] BCBS. “Amendment to the Capital Accord to Incorporate Market Risks”. In: *Bank for International Settlements* (1996). https://www.bis.org/publ/bcbs24.pdf.

[21] BCBS. “Analysis of the Trading Book Hypothetical Portfolio Exercise”. In: *Bank for International Settlements* (2014). https://www.bis.org/publ/bcbs288.pdf.

[22] BCBS. “Fundamental Review of the Trading Book: A Revised Market Risk Framework”. In: *Bank for International Settlements* (2013). https://www.bis.org/publ/bcbs265.pdf.

[23] BCBS. “Interpretive Issues with respect to the Revisions to the Market Risk Framework”. In: *Bank for International Settlements* (2011). http://www.bis.org/publ/bcbs193a.pdf.

[24] BCBS. “Minimum Capital Requirements for Market Risk”. In: *Bank for International Settlements* (2019). https://www.bis.org/bcbs/publ/d457.pdf.

[25] BCBS. “Revisions to the Basel II Market Risk Framework”. In: *Bank for International Settlements* (2009). https://www.bis.org/publ/bcbs193.pdf.

[26] BCBS. “Supervisory Framework for the Use of “Backtesting” in Conjunction with the Internal Models Approach to Market Risk Capital Requirements”. In: *Bank for International Settlements* (1996). http://www.bis.org/publ/bcbs22.pdf.

[27] Jean-Francois Begin, Christian Dorion, and Genevieve Gauthier. “Idiosyncratic jump risk matters: Evidence from equity returns and options”. In: *The Review of Financial Studies* 33.1 (2019), pp. 155–211.

[28] James Bergstra and Yoshua Bengio. “Random Search for Hyper-Parameter Optimization”. In: *The Journal of Machine Learning Research* 13.1 (2012), pp. 281–305.

[29] Jeremy Berkowitz, Peter Christoffersen, and Denis Pelletier. “Evaluating Value-at-Risk Models with Desk-Level Data”. In: *Management Science* 57.12 (2011), pp. 2213–2227.

[30] Jeremy Berkowitz and James O’Brien. “How Accurate are Value-at-Risk Models at Commercial Banks?” In: *The Journal of Finance* 57.3 (2002), pp. 1093–1111.

[31] Daniele Bianchi, Matthias Büchner, and Andrea Tamoni. “Bond Risk Premia with Machine Learning”. In: *The Review of Financial Studies* 34 (2 2021), pp. 1046–1089.

[32] Martin Birn et al. “The Costs and Benefits of Bank Capital - A Review of the Literature”. In: *Journal of Risk and Financial Management* 13.4 (2020), p. 74.

[33] Christopher M Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[34] Christopher M Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[35] Robert Blattberg and Thomas Sargent. “Regression with non-Gaussian stable disturbances: Some sampling results”. In: *Econometrica: Journal of the Econometric Society* (1971), pp. 501–510.

[36] Vincent Bogousslavsky. “Infrequent Rebalancing, Return Autocorrelation, and Seasonality”. In: *The Journal of Finance* 71.6 (2016), pp. 2967–3006.

[37] Oliver Boguth and Mikhail Simutin. “Leverage constraints and asset prices: Insights from mutual fund risk taking”. In: *Journal of Financial Economics* 127.2 (2018), pp. 325–341.

[38] Tim Bollerslev. “Generalized Autoregressive Conditional Heteroskedasticity”. In: *Journal of Econometrics* 31 (3 1986), pp. 307–327.

[39] Tim Bollerslev and Viktor Todorov. “Estimation of jump tails”. In: *Econometrica* 79.6 (2011), pp. 1727–1783.

[40] Tim Bollerslev and Viktor Todorov. “Tails, fears, and risk premia”. In: *The Journal of Finance* 66.6 (2011), pp. 2165–2211.

[41] Tim Bollerslev and Viktor Todorov. “Time-Varying Jump Tails”. In: *Journal of Econometrics* 183.2 (2014), pp. 168–180.

[42] Tim Bollerslev and Viktor Todorov. “Time-varying jump tails”. In: *Journal of Econometrics* 183.2 (2014), pp. 168–180.

[43] Tim Bollerslev, Viktor Todorov, and Sophia Zhengzi Li. "Jump tails, extreme dependencies, and the distribution of stock returns". In: *Journal of Econometrics* 172.2 (2013), pp. 307–324.

[44] Tim Bollerslev and Jeffrey M Wooldridge. "Quasi-Maximum Likelihood Estimation and Inference in Dynamic Models with Time-Varying Covariances". In: *Econometric Reviews* 11.2 (1992), pp. 143–172.

[45] Leo Breiman. "Bagging Predictors". In: *Machine Learning* 24.2 (1996), pp. 123–140.

[46] Leo Breiman. "Random Forests". In: *Machine Learning* 45.1 (2001), pp. 5–32.

[47] Leo Breiman et al. "Classification and Regression Trees". In: *International Group* 432 (1984), pp. 151–166.

[48] Markus K Brunnermeier and Lasse Heje Pedersen. "Market Liquidity and Funding Liquidity". In: *The Review of Financial Studies* 22.6 (2009), pp. 2201–2238.

[49] Markus K Brunnermeier and Yuliy Sannikov. "A macroeconomic model with a financial sector". In: *American Economic Review* 104.2 (2014), pp. 379–421.

[50] Markus K Brunnermeier and Yuliy Sannikov. "The I Theory of Money". In: *NBER Working Paper* (2016).

[51] Svetlana Bryzgalova, Markus Pelger, and Jason Zhu. "Forest Through the Trees: Building Cross-Sections of Stock Returns". In: *Available at SSRN 3493458* (2019).

[52] John Y Campbell and Samuel B Thompson. "Predicting Excess Stock Returns Out of Sample: Can Anything Beat the Historical Average?" In: *The Review of Financial Studies* 21.4 (2008), pp. 1509–1531.

[53] John Y Campbell et al. "Have individual stocks become more volatile? An empirical exploration of idiosyncratic risk". In: *The journal of finance* 56.1 (2001), pp. 1–43.

[54] Mark M Carhart. "On persistence in mutual fund performance". In: *The Journal of finance* 52.1 (1997), pp. 57–82.

[55] Jiaqi Chen, Wenbo Wu, and Michael L Tindall. “Hedge Fund Return Prediction and Fund Selection: A Machine-Learning Approach”. In: *Ocasional Papers - Dallas Federal Reserve* 16 (2016), p. 04.

[56] Long Chen, Zhi Da, and Xinlei Zhao. “What Drives Stock Price Movements?” In: *The Review of Financial Studies* 26.4 (2013), pp. 841–876.

[57] Luyang Chen, Markus Pelger, and Jason Zhu. “Deep Learning in Asset Pricing”. In: *Available at SSRN 3350138* (2019).

[58] Alex Chinco, Adam D Clark-Joseph, and Mao Ye. “Sparse Signals in the Cross-Section of Returns”. In: *The Journal of Finance* 74.1 (2019), pp. 449–492.

[59] Alex Chinco and Vyacheslav Fos. “The Sound of Many Funds Rebalancing”. In: *Available at SSRN 3346352* (2019).

[60] Eunsuk Chong, Chulwoo Han, and Frank C Park. “Deep Learning Networks for Stock Market Analysis and Prediction: Methodology, Data Representations, and Case Studies”. In: *Expert Systems with Applications* 83 (2017), pp. 187–205.

[61] Tarun Chordia, Richard Roll, and Avanidhar Subrahmanyam. “Evidence on the Speed of Convergence to Market Efficiency”. In: *Journal of Financial Economics* 76.2 (2005), pp. 271–292.

[62] Tarun Chordia, Richard Roll, and Avanidhar Subrahmanyam. “Liquidity and Market Efficiency”. In: *Journal of Financial Economics* 87.2 (2008), pp. 249–268.

[63] Tarun Chordia, Avanidhar Subrahmanyam, and V Ravi Anshuman. “Trading Activity and Expected Stock Returns”. In: *Journal of Financial Economics* 59.1 (2001), pp. 3–32.

[64] Peter Christoffersen. “Evaluating Interval Forecasts”. In: *International Economic Review* 39.4 (1998), pp. 841–862.

[65] Peter Christoffersen. “Value-at-Risk Models”. In: *Handbook of Financial Time Series*. Ed. by T.G. Andersen et al. Springer, 2009.

[66] Peter Christoffersen and Silvia Gonçalves. “Estimation Risk in Financial Risk Management”. In: *The Journal of Risk* 7.3 (2005), pp. 1–28.

[67] Peter Christoffersen, Jinyong Hahn, and Atsushi Inoue. “Testing and Comparing Value-at-Risk Measures”. In: *Journal of Empirical Finance* 8.3 (2001), pp. 325–342.

[68] Peter Christoffersen and Denis Pelletier. “Backtesting Value-at-Risk: A Duration-Based Approach”. In: *Journal of Financial Econometrics* 2.1 (2004), pp. 84–108.

[69] Peter F. Christoffersen. *Elements of Financial Risk Managemnet*. Elsevier, 2011.

[70] John H Cochrane. *Asset pricing: Revised edition*. Princeton university press, 2009.

[71] Drew Creal, Siem Jan Koopman, and André Lucas. “Generalized Autoregressive Score Models with Applications”. In: *Journal of Applied Econometrics* 28.5 (2013), pp. 777–795.

[72] Chaoxing Dai, Kun Lu, and Dacheng Xiu. “Knowing factors or factor loadings, or neither? Evaluating estimators of large covariance matrices with noisy and asynchronous data”. In: *Journal of Econometrics* 208.1 (2019), pp. 43–79.

[73] Jon Danielsson and Casper G De Vries. “Tail index and quantile estimation with very high frequency data”. In: *Journal of empirical Finance* 4.2-3 (1997), pp. 241–257.

[74] Richard A Davis and Wei Wu. “Bootstrapping M-estimates in regression and autoregression with infinite variance”. In: *Statistica Sinica* (1997), pp. 1135–1154.

[75] Min-Yuh Day and Jian-Ting Lin. “Artificial Intelligence for ETF Market Prediction and Portfolio Optimization”. In: *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 2019, pp. 1026–1033.

[76] Ian Dew-Becker and Stefano Giglio. “Cross-sectional uncertainty and the business cycle: evidence from 40 years of options data”. In: (2020).

[77] Francis X Diebold and Robert S Mariano. “Comparing Predictive Accuracy”. In: *Journal of Business & Economic Statistics* 20.1 (2002), pp. 134–144.

[78] Francis X. Diebold, Til Schuermann, and John D. Stroughair. “Pitfalls and Opportunities in the use of Extreme Value Theory in Risk Management”. In: *The Journal of Risk Finance* 1.2 (2000), pp. 30–35.

[79] Darrell Duffie. “Presidential Address: Asset Price Dynamics with Slow-Moving Capital”. In: *The Journal of Finance* 65.4 (2010), pp. 1237–1267.

[80] Rick Durrett. *Probability: theory and examples*. Vol. 49. Cambridge university press, 2019.

[81] Aniruddha Dutta, Saket Kumar, and Meheli Basu. “A Gated Recurrent Unit Approach to Bitcoin Price Prediction”. In: *Journal of Risk and Financial Management* 13.2 (2020), p. 23.

[82] EBA. “EBA Guidlines on Stressed Value at Risk”. In: *European Banking Authority* (2012).

[83] Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. CRC Press, 1994.

[84] Werner Ehm et al. “Of Quantiles and Expectiles: Consistent Scoring Functions, Choquet Representations and Forecast Rankings”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 78.3 (2016), pp. 505–562.

[85] Paul Embrechts, Claudia Kluppelberg, and Thomas Mikosch. *Modelling extremal events, volume 33 of Applications of Mathematics*. 1997.

[86] Robert F Engle and Simone Manganelli. “CAViaR: Conditional Autoregressive Value at Risk by Regression Quantiles”. In: *Journal of Business & Economic Statistics* 22.4 (2004), pp. 367–381.

[87] Robert F. Engle. “Estimates of the Variance of U.S. Inflation Based on the ARCH Model”. In: *Journal of Money, Credit and Banking* 15 (1983), pp. 286–301.

[88] Antonio Falato, Diana Iercosan, and Filip Zikes. “Banks as Regulated Traders”. In: *Working Paper* (2019).

[89] Eugene F Fama. “Efficient Capital Markets: II”. In: *The Journal of Finance* 46.5 (1991), pp. 1575–1617.

[90] Eugene F Fama. “Mandelbrot and the stable Paretian hypothesis”. In: *The journal of business* 36.4 (1963), pp. 420–429.

[91] Eugene F Fama and Kenneth R French. “A five-factor asset pricing model”. In: *Journal of financial economics* 116.1 (2015), pp. 1–22.

[92] Eugene F Fama and Kenneth R French. “Common risk factors in the returns on stocks and bonds”. In: *Journal of financial economics* 33.1 (1993), pp. 3–56.

[93] Eugene F Fama and James D MacBeth. “Risk, return, and equilibrium: Empirical tests”. In: *Journal of political economy* 81.3 (1973), pp. 607–636.

[94] Guanhao Feng, Stefano Giglio, and Dacheng Xiu. *Taming the Factor Zoo: A Test of New Factors*. Tech. rep. National Bureau of Economic Research, 2019.

[95] Guanhao Feng, Jingyu He, and Nicholas G Polson. “Deep Learning for Predicting Asset Returns”. In: *arXiv preprint arXiv:1804.09314* (2018).

[96] Guanhao Feng, Nick Polson, and Jianeng Xu. “Deep Learning in Characteristics-Sorted Factor Models”. In: *Available at SSRN 3243683* (2019).

[97] Thomas Fischer and Christopher Krauss. “Deep Learning with Long Short-Term Memory Networks for Financial Market Predictions”. In: *European Journal of Operational Research* 270.2 (2018), pp. 654–669.

[98] Tobias Fissler and Johanna F Ziegel. “Higher Order Elicitability and Osband’s Principle”. In: *The Annals of Statistics* 44.4 (2016), pp. 1680–1707.

[99] Tobias Fissler, Johanna F Ziegel, and Tilmann Gneiting. “Expected Shortfall is Jointly Elicitable with Value at Risk - Implications for Backtesting”. In: *arXiv preprint arXiv:1507.00244* (2015).

[100] Jean-Sebastien Fontaine and Rene Garcia. “Bond liquidity premia”. In: *The Review of Financial Studies* 25.4 (2012), pp. 1207–1254.

[101] Alain-Philippe Fortin, Jean-Guy Simonato, and Georges Dionne. “Forecasting Expected Shortfall: Should we use a Multivariate Model for Stock Market Factors?” In: *Available at SSRN 3203049* (2018).

[102] Kenneth R French. “Presidential address: The cost of active investing”. In: *The Journal of Finance* 63.4 (2008), pp. 1537–1573.

[103] Yoav Freund and Robert E Schapire. “A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting”. In: *Journal of Computer and System Sciences* 55.1 (1997), pp. 119–139.

[104] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. “Regularization Paths for Generalized Linear Models via Coordinate Descent”. In: *Journal of Statistical Software* 33.1 (2010), p. 1.

[105] Jerome H Friedman. “Greedy Function Approximation: A Gradient Boosting Machine”. In: *Annals of Statistics* (2001), pp. 1189–1232.

[106] Jerome H Friedman. “Stochastic Gradient Boosting”. In: *Computational Statistics & Data Analysis* 38.4 (2002), pp. 367–378.

[107] Xavier Gabaix. “Power laws in economics and finance”. In: *Annu. Rev. Econ.* 1.1 (2009), pp. 255–294.

[108] Xavier Gabaix et al. “Institutional investors and stock market volatility”. In: *The Quarterly Journal of Economics* 121.2 (2006), pp. 461–504.

[109] Lei Gao et al. “Market Intraday Momentum”. In: *Journal of Financial Economics* 129.2 (2018), pp. 394–414.

[110] R. Gencay and F. Selcuk. “Extreme Value Theory and Value-at-Risk: Relative Performance in Emerging Markets”. In: *International Journal of Forecasting* 20 (2004), pp. 287–303.

[111] Véronique Genre et al. “Combining Expert Forecasts: Can Anything Beat the Simple Average?” In: *International Journal of Forecasting* 29.1 (2013), pp. 108–121.

[112] Raffaella Giacomini and Halbert White. “Tests of Conditional Predictive Ability”. In: *Econometrica* 74.6 (2006), pp. 1545–1578.

[113] Stefano Giglio and Dacheng Xiu. “Asset Pricing with Omitted Factors”. In: *Chicago Booth Research Paper* 16-21 (2018).

[114] Tilmann Gneiting. “Making and Evaluating Point Forecasts”. In: *Journal of the American Statistical Association* 106.494 (2011), pp. 746–762.

[115] Ian J Goodfellow et al. “Maxout Networks”. In: *arXiv preprint arXiv:1302.4389* (2013).

[116] Michael B Gordy and Bradley Howells. “Procyclicality in Basel II: Can we Treat the Disease Without Killing the Patient?” In: *Journal of Financial Intermediation* 15.3 (2006), pp. 395–417.

[117] Michael B Gordy and Alexander J McNeil. “Spectral Backtests of Forecast Distributions with Application to Risk Management”. In: *Journal of Banking & Finance* 116 (2020), p. 105817.

[118] Christian Gourieroux. *ARCH Models and Financial Applications*. Springer Series in Statistics, 1997.

[119] John M Griffin, Jeffrey H Harris, and Selim Topaloglu. “The dynamics of institutional and individual trading”. In: *The Journal of Finance* 58.6 (2003), pp. 2285–2320.

[120] Denis Gromb and Dimitri Vayanos. “Equilibrium and welfare in markets with financially constrained arbitrageurs”. In: *Journal of financial Economics* 66.2-3 (2002), pp. 361–407.

[121] Shihao Gu, Bryan Kelly, and Dacheng Xiu. “Autoencoder Asset Pricing Models”. In: *Journal of Econometrics* (2020).

[122] Shihao Gu, Bryan Kelly, and Dacheng Xiu. “Autoencoder Asset Pricing Models”. In: *forthcoming in Journal of Econometrics* (2020).

[123] Shihao Gu, Bryan Kelly, and Dacheng Xiu. “Empirical Asset Pricing via Machine Learning”. In: *The Review of Financial Studies* 33.5 (2020), pp. 2223–2273.

[124] Petr Hajek and Aliaksandr Barushka. “Integrating Sentiment Analysis and Topic Detection in Financial News for Stock Movement Prediction”. In: *Proceedings of the 2nd International Conference on Business and Information Management*. 2018, pp. 158–162.

[125] Bruce Hansen. “Autoregressive Conditional Density Estimation”. In: *International Economic Review* 35.3 (1994), pp. 705–730.

[126] Peter R Hansen, Asger Lunde, and James M Nason. “The Model Confidence Set”. In: *Econometrica* 79.2 (2011), pp. 453–497.

[127] Andrew C Harvey. *Dynamic Models for Volatility and Heavy Tails: With Applications to Financial and Economic Time Series*. Vol. 52. Cambridge University Press, 2013.

[128] Campbell R Harvey and Akhtar Siddique. “Conditional Skewness in Asset Pricing Tests”. In: *The Journal of Finance* 55.3 (2000), pp. 1263–1295.

[129] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media, 2009.

[130] Zhiguo He, Bryan Kelly, and Asaf Manela. “Intermediary asset pricing: New evidence from many asset classes”. In: *Journal of Financial Economics* 126.1 (2017), pp. 1–35.

[131] Zhiguo He and Arvind Krishnamurthy. “Intermediary asset pricing”. In: *American Economic Review* 103.2 (2013), pp. 732–70.

[132] Frank Heid. “The Cyclical Effects of the Basel II Capital Requirements”. In: *Journal of Banking & Finance* 31.12 (2007), pp. 3885–3900.

[133] Terrence Hendershott and Albert J Menkveld. “Price pressures”. In: *Journal of Financial Economics* 114.3 (2014), pp. 405–423.

[134] Bernard Herskovic et al. “The common factor in idiosyncratic volatility: Quantitative asset pricing implications”. In: *Journal of Financial Economics* 119.2 (2016), pp. 249–283.

[135] Steven L Heston, Robert A Korajczyk, and Ronnie Sadka. “Intraday Patterns in the Cross-Section of Stock Returns”. In: *The Journal of Finance* 65.4 (2010), pp. 1369–1407.

[136] Steven L Heston and Nitish Ranjan Sinha. “News vs. Sentiment: Predicting Stock Returns From News Stories”. In: *Financial Analysts Journal* 73.3 (2017), pp. 67–83.

[137] Bruce Hill. “A Simple General Approach to Inference About the Tail of a Distribution”. In: *The Annals of Statistics* 3.5 (1975), pp. 1163–1174.

[138] Bruce M Hill. “A simple general approach to inference about the tail of a distribution”. In: *The annals of statistics* (1975), pp. 1163–1174.

[139] Jonathan B Hill. “On tail index estimation for dependent, heterogeneous data”. In: *Econometric Theory* 26.5 (2010), pp. 1398–1436.

[140] Jonathan B Hill. “Tail index estimation for a filtered dependent time series”. In: *Statistica Sinica* (2015), pp. 609–629.

[141] David V Hinkley. “On power transformations to symmetry”. In: *Biometrika* 62.1 (1975), pp. 101–111.

[142] Robert J Hodrick. “Dividend yields and expected stock returns: Alternative procedures for inference and measurement”. In: *The Review of Financial Studies* 5.3 (1992), pp. 357–386.

[143] Dillon Huddleston, Fred Liu, and Lars Stentoft. “Intraday market predictability: A machine learning approach”. In: (2020).

[144] Ronald Huisman et al. “Tail-index Estimates in Small Samples”. In: *Journal of Business & Economic Statistics* 19.2 (2001), pp. 208–216.

[145] John Hull and Alan White. “Incorporating Volatility Updating into the Historical Simulation Method for Value at Risk”. In: *Journal of Risk* 1.1 (1998), pp. 5–19.

[146] Dennis W Jansen and Casper G De Vries. “On the frequency of large stock returns: Putting booms and busts into perspective”. In: *The review of economics and statistics* (1991), pp. 18–24.

[147] Narasimhan Jegadeesh. “Evidence of Predictable Behavior of Security Returns”. In: *The Journal of Finance* 45.3 (1990), pp. 881–898.

[148] Narasimhan Jegadeesh and Sheridan Titman. “Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency”. In: *The Journal of Finance* 48.1 (1993), pp. 65–91.

[149] Anders Hedegaard Jessen and Thomas Mikosch. “REGULARLY VARYING FUNCTIONS.” In: *Publications de L’institut Mathematique* 94 (2006).

[150] George J Jiang and Tong Yao. “Stock price jumps and cross-sectional return predictability”. In: *Journal of Financial and Quantitative Analysis* 48.5 (2013), pp. 1519–1544.

[151] George J Jiang and Kevin X Zhu. "Information shocks and short-term market underreaction". In: *Journal of Financial Economics* 124.1 (2017), pp. 43–64.

[152] I.T. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics. Springer, 2002. ISBN: 9780387954424. URL: https://books.google.ca/books?id=%5C_olByCrhjwIC.

[153] JPMorgan. "Riskmetrics - Technical Document". In: *Technical Document, 4th ed., JP Morgan, New York* (1996).

[154] Nishad Kapadia and Morad Zekhnini. "Do idiosyncratic jumps matter?" In: *Journal of Financial Economics* 131.3 (2019), pp. 666–692.

[155] Jovan Karamata. *Some theorems concerning slowly varying functions*. Tech. rep. WISCONSIN UNIV-MADISON MATHEMATICS RESEARCH CENTER, 1962.

[156] Zheng Tracy Ke, Bryan T Kelly, and Dacheng Xiu. *Predicting Returns with Text Data*. Tech. rep. 2019.

[157] Phillip Kearns and Adrian Pagan. "Estimating the density tail index for financial time series". In: *Review of Economics and Statistics* 79.2 (1997), pp. 171–175.

[158] Bryan Kelly and Hao Jiang. "Tail Risk and Asset Prices". In: *The Review of Financial Studies* 27.10 (2014), pp. 2841–2871.

[159] Bryan Kelly and Hao Jiang. "Tail risk and asset prices". In: *The Review of Financial Studies* 27.10 (2014), pp. 2841–2871.

[160] Bryan Kelly, Hanno Lustig, and Stijn Van Nieuwerburgh. "Too-systemic-to-fail: What option markets imply about sector-wide government guarantees". In: *American Economic Review* 106.6 (2016), pp. 1278–1319.

[161] Bryan Kelly and Seth Pruitt. "The Three-Pass Regression Filter: A New Approach to Forecasting Using Many Predictors". In: *Journal of Econometrics* 186.2 (2015), pp. 294–316.

[162] Bryan Kelly, Seth Pruitt, and Yinan Su. “Instrumented Principal Component Analysis”. In: *Available at SSRN 2983919* (2019).

[163] Bryan T Kelly, Seth Pruitt, and Yinan Su. “Characteristics are Covariances: A Unified Model of Risk and Return”. In: *Journal of Financial Economics* 134.3 (2019), pp. 501–524.

[164] Keith Knight. “Limiting distributions for L1 regression estimators under general conditions”. In: *Annals of statistics* (1998), pp. 755–770.

[165] Roger Koenker and Gilbert Bassett. “Regression quantiles”. In: *Econometrica: journal of the Econometric Society* (1978), pp. 33–50.

[166] Ralph SJ Koijen and Stijn Van Nieuwerburgh. “Predictability of Returns and Cash Flows”. In: *Annual Review of Financial Economics* 3.1 (2011), pp. 467–491.

[167] Yehuda Koren. “The Bellkor Solution to the Netflix Grand Prize”. In: *Netflix Prize Documentation* 81.2009 (2009), pp. 1–10.

[168] Sofonias A Korsaye, Alberto Quaini, and Fabio Trojani. “Smart SDFs”. In: *Available at SSRN 3475451* (2019).

[169] Serhiy Kozak, Stefan Nagel, and Shrihari Santosh. “Shrinking the Cross-Section”. In: *Journal of Financial Economics* 135.2 (2020), pp. 271–292.

[170] Paul Kupiec. “Techniques for Verifying the Accuracy of Risk Measurement Models”. In: *The Journal of Derivatives* 3.2 (1995), pp. 73–84.

[171] Charles MC Lee and Balkrishna Radhakrishna. “Inferring investor behavior: Evidence from TORQ data”. In: *Journal of Financial Markets* 3.2 (2000), pp. 83–111.

[172] Charles MC Lee and Mark J Ready. “Inferring trade direction from intraday data”. In: *The Journal of Finance* 46.2 (1991), pp. 733–746.

[173] Bruce N Lehmann. “Fads, Martingales, and Market Efficiency”. In: *The Quarterly Journal of Economics* 105.1 (1990), pp. 1–28.

[174] Jonathan Lewellen, Stefan Nagel, and Jay Shanken. “A skeptical appraisal of asset pricing tests”. In: *Journal of Financial economics* 96.2 (2010), pp. 175–194.

[175] Huidi Lin and Viktor Todorov. “Aggregate Asymmetry in Idiosyncratic Jump Risk”. In: (2019).

[176] Shiqing Ling. “Self-weighted least absolute deviation estimation for infinite variance autoregressive models”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.3 (2005), pp. 381–393.

[177] John Lintner. “Security Prices, Risk, and Maximal Gains from Diversification”. In: *The Journal of Finance* 20.4 (1965), pp. 587–615.

[178] Fred Liu and Lars Stentoft. “Regulatory Capital and Incentives for Risk Model Choice under Basel 3”. In: *Journal of Financial Econometrics* (2020).

[179] Fred Liu and Lars Stentoft. “Regulatory Capital and Incentives for Risk Model Choice under Basel 3”. In: *Working Paper* (2020).

[180] Huaigang Long, Yuexiang Jiang, and Yanjian Zhu. “Idiosyncratic tail risk and expected stock returns: Evidence from the Chinese stock markets”. In: *Finance Research Letters* 24 (2018), pp. 129–136.

[181] Wen Long, Zhichen Lu, and Lingxiao Cui. “Deep Learning-Based Feature Engineering for Stock Price Movement Prediction”. In: *Knowledge-Based Systems* 164 (2019), pp. 163–173.

[182] Dong Lou, Christopher Polk, and Spyros Skouras. “A Tug of War: Overnight Versus Intraday Expected Returns”. In: *Journal of Financial Economics* 134.1 (2019), pp. 192–213.

[183] Benoit Mandelbrot. “New methods in statistical economics”. In: *Journal of political economy* 71.5 (1963), pp. 421–440.

[184] Simone Manganelli and Robert F Engle. “Value at Risk Models in Finance”. In: *Working Paper* (2001).

[185] Ivana Marković et al. “Stock Market Trend Prediction Using AHP and Weighted Kernel LS-SVM”. In: *Soft Computing* 21.18 (2017), pp. 5387–5398.

[186] Donald W Marquaridt. “Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation”. In: *Technometrics* 12.3 (1970), pp. 591–612.

[187] Alexander McNeil and Rudiger Frey. “Estimation of Tail-Related Risk Measures for Heteroscedastic Financial Time Series: An Extreme Value Approach”. In: *Journal of Empirical Finance* 7 (3-4 2000), pp. 271–300.

[188] Amit Mehta et al. “Managing Market Risk: Today and Tomorrow”. In: *McKinsey & Company Working Papers on Risk 32* (2012).

[189] Robert C Merton. “Option pricing when underlying stock returns are discontinuous”. In: *Journal of financial economics* 3.1-2 (1976), pp. 125–144.

[190] Thomas Mikosch and Casper G de Vries. “Heavy tails of OLS”. In: *Journal of Econometrics* 172.2 (2013), pp. 205–221.

[191] Randi Naes, Johannes A Skjeltorp, and Bernt Arne Odegaard. “Stock market liquidity and the business cycle”. In: *The Journal of Finance* 66.1 (2011), pp. 139–176.

[192] Vinod Nair and Geoffrey E Hinton. “Rectified Linear Units Improve Restricted Boltzmann Machines”. In: *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*. 2010, pp. 807–814.

[193] Christopher J Neely et al. “Forecasting the Equity Risk Premium: The Role of Technical Indicators”. In: *Management Science* 60.7 (2014), pp. 1772–1791.

[194] Whitney K Newey and Kenneth D West. “A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix”. In: *Econometrica* 55.3 (1987), pp. 703–708.

[195] Whitney K Newey and Kenneth D West. “Hypothesis Testing with Efficient Method of Moments Estimation”. In: *International Economic Review* (1987), pp. 777–787.

[196] Natalia Nolde and Johanna F Ziegel. “Elicitability and Backtesting: Perspectives for Banking Regulation”. In: *The Annals of Applied Statistics* 11.4 (2017), pp. 1833–1874.

[197] James O’Brien and Paweł J Szerszeń. “An Evaluation of Bank Measures for Market Risk Before, During and After the Financial Crisis”. In: *Journal of Banking & Finance* 80 (2017), pp. 215–234.

[198] Mr Michael G Papaioannou et al. *Procyclical behavior of institutional investors during the recent financial crisis: Causes, impacts, and challenges*. 13-193. International Monetary Fund, 2013.

[199] Lubos Pastor and Robert F Stambaugh. “Liquidity risk and expected stock returns”. In: *Journal of Political economy* 111.3 (2003), pp. 642–685.

[200] Andrew J Patton. “Comparing Possibly Misspecified Forecasts”. In: *Journal of Business & Economic Statistics* (2019), pp. 1–23.

[201] Andrew J Patton, Johanna F Ziegel, and Rui Chen. “Dynamic Semiparametric Models for Expected Shortfall (and Value-at-Risk)”. In: *Journal of Econometrics* 211.2 (2019), pp. 388–413.

[202] Paola Pederzoli. “Crash Risk in Individual Stocks”. In: *Available at SSRN 2862372* (2018).

[203] Markus Pelger. “Understanding systematic risk: A high-frequency approach”. In: *The Journal of Finance* 75.4 (2020), pp. 2179–2220.

[204] Christophe Pérignon, Zi Yin Deng, and Zhi Jun Wang. “Do Banks Overstate their Value-at-Risk?” In: *Journal of Banking & Finance* 32.5 (2008), pp. 783–794.

[205] Christophe Pérignon and Daniel R Smith. “A new Approach to Comparing VaR Estimation Methods”. In: *The Journal of Derivatives* 16.2 (2008), pp. 54–66.

[206] Christophe Pérignon and Daniel R Smith. “The Level and Quality of Value-at-Risk Disclosure by Commercial Banks”. In: *Journal of Banking & Finance* 34.2 (2010), pp. 362–377.

[207] James L Powell. “Estimation of monotonic regression models under quantile restrictions”. In: (1991).

[208] Likuan Qin and Viktor Todorov. “Nonparametric implied Levy densities”. In: *The Annals of Statistics* 47.2 (2019), pp. 1025–1060.

[209] David Rapach and Guofu Zhou. “Forecasting Stock Returns”. In: *Handbook of Economic Forecasting*. Vol. 2. Elsevier, 2013, pp. 328–383.

[210] Thomas Renault. “Intraday Online Investor Sentiment and Return Patterns in the US Stock Market”. In: *Journal of Banking & Finance* 84 (2017), pp. 25–40.

[211] Richard Roll. “A Simple Implicit Measure of the Effective Bid-Ask Spread in an Efficient Market”. In: *The Journal of Finance* 39.4 (1984), pp. 1127–1139.

[212] Ritirupa Samanta and Blake LeBaron. *Extreme value theory and fat tails in equity markets*. Tech. rep. Society for Computational Economics, 2005.

[213] Pavel G Savor. “Stock returns after major price shocks: The impact of information”. In: *Journal of financial Economics* 106.3 (2012), pp. 635–659.

[214] Robert E Schapire. “The Strength of Weak Learnability”. In: *Machine Learning* 5.2 (1990), pp. 197–227.

[215] Robert E Schapire and Yoav Freund. “Boosting: Foundations and Algorithms”. In: *Kybernetes* (2013).

[216] Jay Shanken. “On the estimation of beta-pricing models”. In: *The review of financial studies* 5.1 (1992), pp. 1–33.

[217] William F Sharpe. “Capital Asset Prices: A Theory of Market Equilibrium Under Conditions of Risk”. In: *The Journal of Finance* 19.3 (1964), pp. 425–442.

[218] Robert J Shiller. *Irrational exuberance: Revised and expanded third edition*. Princeton university press, 2015.

[219] Jeungbo Shim. "Bank Capital Buffer and Portfolio Risk: The Influence of Business Cycle and Revenue Eiversification". In: *Journal of Banking & Finance* 37.3 (2013), pp. 761–772.

[220] Richard W Sias. "Institutional herding". In: *The Review of Financial Studies* 17.1 (2004), pp. 165–206.

[221] Nitish Srivastava et al. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". In: *The Journal of Machine Learning Research* 15.1 (2014), pp. 1929–1958.

[222] Robert F Stambaugh. "Presidential address: Investment noise and trends". In: *The Journal of Finance* 69.4 (2014), pp. 1415–1453.

[223] H Eugene Stanley, Vasiliki Plerou, and Xavier Gabaix. "A statistical physics view of financial fluctuations: Evidence for scaling and universality". In: *Physica A: Statistical Mechanics and its Applications* 387.15 (2008), pp. 3967–3981.

[224] Ian Sutherland, Yesuk Jung, and Gunhee Lee. "Statistical Arbitrage on the KOSPI 200: An Exploratory Analysis of Classification and Prediction Machine Learning Algorithms for Day Trading". In: *Journal of Economics and International Business Management* 6.1 (2018), pp. 10–19.

[225] Panayiotis Theodossiou. "Financial Data and the Skewed Generalized t-distribution". In: *Management Science* 44.12 (1998), pp. 1650–1661.

[226] Panayiotis Theodossiou. "Risk Measures for Investment Values and Returns Based on Skewed-Heavy Tailed Distributions: Analytical Derivations and Comparison". In: *SSRN Working Paper* (2018).

[227] Robert Tibshirani. "Regression Shrinkage and Selection via the Lasso". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996), pp. 267–288.

[228] Allan Timmermann. “Elusive Return Predictability”. In: *International Journal of Forecasting* 24.1 (2008), pp. 1–18.

[229] Allan Timmermann. “Forecast Combinations”. In: *Handbook of Economic Forecasting* 1 (2006), pp. 135–196.

[230] Viktor Todorov and Tim Bollerslev. “Jumps and betas: A new framework for disentangling and estimating systematic risks”. In: *Journal of Econometrics* 157.2 (2010), pp. 220–235.

[231] Ruey S Tsay. *Analysis of Financial Time Series*. Wiley, Third Edition, 2010.

[232] Maarten RC Van Oordt and Chen Zhou. “Systematic tail risk”. In: *Journal of Financial and Quantitative Analysis* (2016), pp. 685–705.

[233] Sijian Wang et al. “Random Lasso”. In: *The Annals of Applied Statistics* 5.1 (2011), p. 468.

[234] Alois Weigand. “Machine Learning in Empirical Asset Pricing”. In: *Financial Markets and Portfolio Management* 33.1 (2019), pp. 93–104.

[235] Ivo Welch and Amit Goyal. “A Comprehensive Look at the Empirical Performance of Equity Premium Prediction”. In: *The Review of Financial Studies* 21.4 (2008), pp. 1455–1508.

[236] Peter H Westfall and S Stanley Young. *Resampling-based Multiple Testing: Examples and Methods for p-value Adjustment*. Vol. 279. John Wiley & Sons, 1993.

[237] Jingming Xue et al. “Financial Time Series Prediction Using L2, 1RF-ELM”. In: *Neurocomputing* 277 (2018), pp. 176–186.

[238] Tingting Ye and Liangliang Zhang. “Derivatives Pricing via Machine Learning”. In: *Available at SSRN 3352688* (2019).

[239] Hyungbin Yun et al. “Portfolio Management via Two-Stage Deep Learning with a Joint Cost”. In: *Expert Systems with Applications* 143 (2020), p. 113041.

[240] Matthew D Zeiler. "Adadelta: An Adaptive Learning Rate Method". In: *arXiv preprint arXiv:1212.5701* (2012).

[241] Johanna F Ziegel et al. "Murphy Diagrams: Forecast Evaluation of Expected Shortfall". In: *arXiv preprint arXiv:1705.04537* (2017).

[242] Hui Zou and Trevor Hastie. "Regularization and Variable Selection via the Elastic Net". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.2 (2005), pp. 301–320.

Appendix A

Appendices for Chapter 2

A.1 Regulatory Calculation Details

This section provides additional technical details on our regulatory calculations for Basel 3. We discuss portfolio formation, dynamic model estimation, stressed periods, and calculation of the adjustment ratio $ES_{F,C}/ES_{R,C}$.

A.1.1 Portfolio Formation

Under Basel 3, each trading desk is required to calculate 3 liquidity-adjusted ES measures ($ES_{R,S}$, $ES_{F,C}$, $ES_{R,C}$) for 21 possible liquidity horizons across the 5 risk factors, totaling 63 daily ES calculations. To implement this efficiently, we calculate the ES of each portfolio using univariate models rather than multivariate models, since they have been shown to perform equally well (Fortin, Simonato, and Dionne (2018)). Each portfolio is formed by taking the equal-weighted mean return of assets within the portfolio (denoted r_{Port}), then taking the log transformation $x_{Port} = \log(1 + r_{Port})$. We use the log returns from the univariate portfolio to estimate models and conduct forecasts. Finally, we weigh the forecasts by the portfolio weight relative to the number of assets in the representative portfolio. For example, for the 20-day liquidity horizon portfolio for equities in 2010, the weight used for the data used in Liu and

Stentoft (2020b) would be 2/16 (the Russel 2000 and the CBOE Putwrite index divided by 16 total indexes).

A.1.2 Dynamic Model Estimation

We use a 2,000 day rolling window to estimate the GARCH and FZ models. However, some portfolios begin after 1989 and we want to estimate their risk measures without look-ahead bias. To calculate risk measures for those portfolios during the first 2,000 days, we use parameters estimated from the first 2,000 days of the representative portfolio, then filter using returns from the actual portfolio. For example, suppose we are estimating FHS for the LH 120 portfolio, which begins in 02/2001 in Liu and Stentoft (2020b). We first estimate GARCH parameters and $\hat{c}_{1,p}^{FHS}$ and $\hat{c}_{2,p}^{FHS}$ using returns from the representative portfolio from 01/89 to 12/96. Then, from 02/2001 to 01/2009, we calculate risk measures using the representative portfolio parameters, but update σ_t using returns from LH 120. After 02/2009, we use parameters estimated from LH 120 returns using a rolling-window as usual. We find this method provides good risk estimates without any look-ahead bias.

A.1.3 Stressed Period

The BCBS sets the estimation period for Stressed VaR and ES as 12-months to ensure nonparametric models like HS are sufficiently stressed. Setting a fixed period for HS prevents banks from "gaming" the regulation by stretching the estimation window and lowering Stressed VaR or ES. However, this choice of estimation period severely limits the class of available models, since dynamic models require a longer estimation period. To allow for dynamic models, we reinterpret Basel's stress period to be the estimation period that produces the maximum risk measure. This is justified, since even with the longer estimation window most dynamic models have a higher Stressed VaR and ES relative to 12-month HS.

While the stressed period is identified as "the 12-month period of stress over the observation horizon in which the portfolio experiences the largest loss", the exact choice of stressed period

is subject to the regulatory supervisor's judgment (BCBS (2019), 33.7-33.8). For Basel 2.5, we find the stressed period each day by searching for the maximum historical (up to that day) representative portfolio 10-day VaR at the 99% confidence level for each model. Analogously for Basel 3, we find the stressed period each day by searching for the maximum historical (up to that day) representative portfolio 10-day ES at the 97.5% confidence level for each model. Then, we use that same stressed period to calculate the liquidity horizons portfolios across the 5 risk factors. The stressed period is stable and only changes during periods of high volatility. Across all models, the Basel 3 stressed periods are the 1998 Asian crisis, the 2001 dot-com bust, and the 2008 financial crisis. As a robustness check, we also calculate the stressed period based on the liquidity-adjusted ES and find that the stressed periods are the same and the capital requirements are nearly identical.

A.1.4 Reduced Set Adjustment Ratio

Due to data limitations, BCBS allows banks to use the ratio $ES_{F,C}/ES_{R,C}$ with a floor of 1, to adjust the reduced set Stressed ES to the full set. The main stressed periods are the 1998 Asian crisis, the 2001 dot-com bust, and the 2008 financial crisis. Between 1998 to 2001, we calculate the reduced set $ES_{R,C}$ using only indexes that begin before 09/1998, while the full set $ES_{F,C}$ contains all available indexes. We follow the same procedure between 2001 and 2008 by calculating the reduced set $ES_{R,C}$ using only indexes that begin before 9/2001 and the full set $ES_{F,C}$ containing all available indexes. Finally, after the 2008 financial crisis, the reduced set is equal to the full set and $ES_{F,C}/ES_{R,C} = 1$.

A.2 Threshold values for EVT models

In this section we conduct an extensive Monte Carlo study with the objective of determining appropriate threshold values for the GPD and Hill models. When choosing the threshold, the problem faced is analogous to the classic Bias-Variance trade-off: a threshold that is too low

violates the asymptotic basis of the model and increases bias whereas a threshold that is too high will have few exceedances leading to high variance. The standard practice is to have as low a threshold as possible, subject to the EVT model providing a reasonable approximation.

Our simulation study considers five versions of the SGT distribution for the innovations in the dynamic loss model in Equation (2.10) from Liu and Stentoft (2020b).[1] The five cases are:

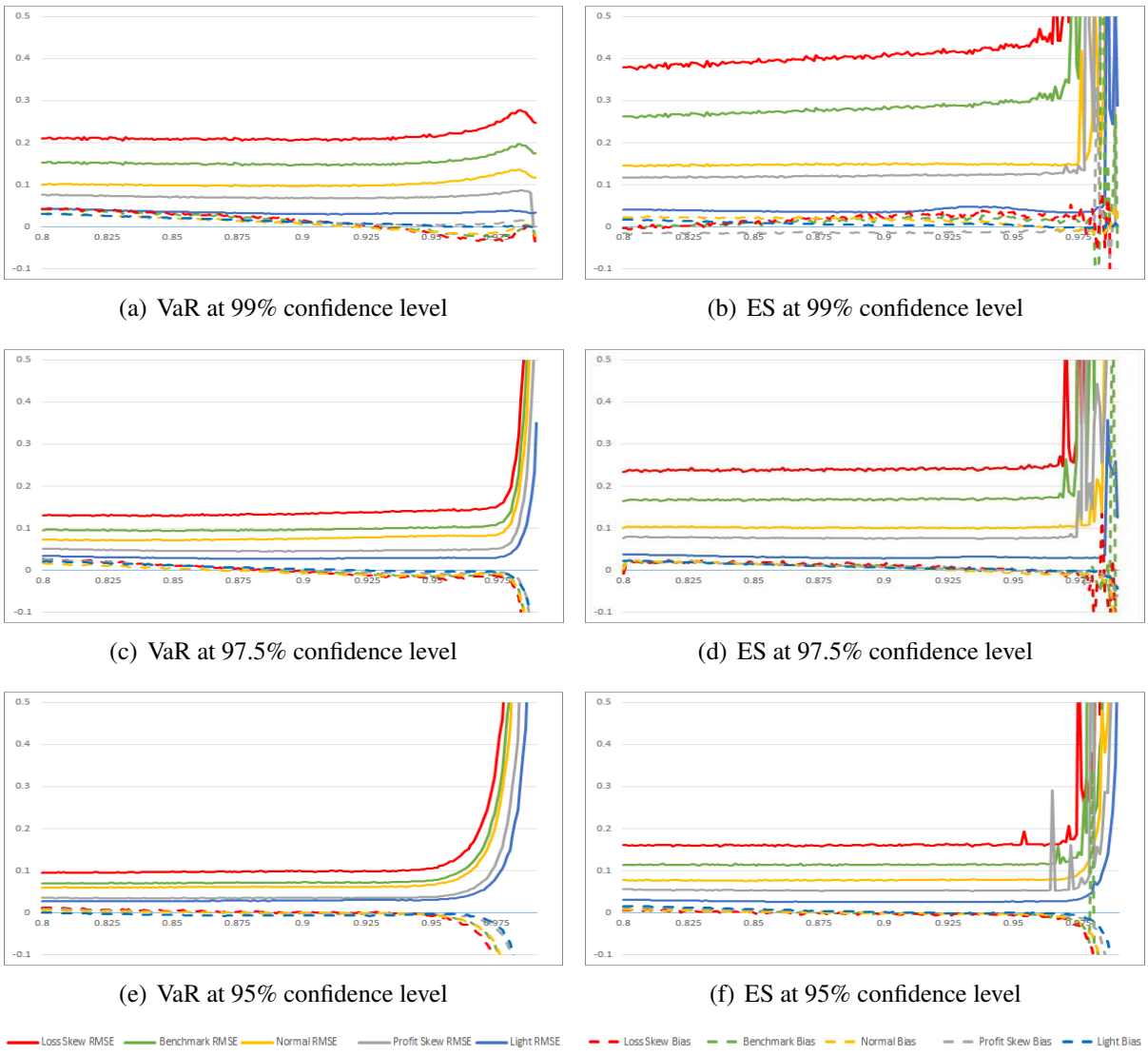
1. Benchmark: $k = 2, n = 8, \lambda = 0$
2. Loss skewed: $k = 2, n = 8, \lambda = 0.5$
3. Profit skewed: $k = 2, n = 8, \lambda = -0.5$
4. Normal: $k = 2, n = 500, \lambda = 0$
5. Light-tailed: $k = 10, n = 500, \lambda = 0$

The five data generating processes (DGPs) capture the heavy tails and skewness exhibited by financial returns and reported in, e.g., Section 2.3.1 of Liu and Stentoft (2020b). We set the estimation window $T = 1,000$.

McNeil and Frey (2000) describe a method of threshold choice using a threshold at the (m)'th order statistic of innovations where $N_\eta = T - m$ is the number of tail observations. Let $\epsilon_{(T)} \geq \epsilon_{(T-1)} \geq ... \geq \epsilon_{(1)}$ denote the ordered innovations. Then the GPD with parameters ξ and σ are fit to the excesses of innovations over estimated threshold $\epsilon_{(m)}$, i.e., $(\epsilon_{(T)} - \epsilon_{(m)}, ..., \epsilon_{(m+1)} - \epsilon_{(m)})$. McNeil and Frey (2000) use Monte Carlo simulation at the 99% confidence level and estimate the optimal threshold for VaR as $\hat{\eta} = \hat{\epsilon}_{(m)} = Q_{0.9}(\{\hat{\epsilon}_t\})$, where $Q_{0.9}(\{\hat{\epsilon}_t\})$ denotes the 0.9 quantile of the empirical innovations. To estimate the optimal threshold at the 97.5% and 95% confidence levels we perform the simulation exercise described above, which is analogous to the threshold choice Monte Carlo experiment in Christoffersen and Gonçalves (2005) to find the optimal threshold of the Hill model. Moreover, while McNeil and Frey (2000) and Christoffersen and Gonçalves (2005) only study threshold selection with simulations from the t-distribution we simulate skewed and light-tailed DGPs as well.

[1] Since the paper's focus is primarily on the distribution of innovations the GARCH parameters are kept constant at $\alpha = 0.1$, $\beta = 0.8$, and $\omega = (20^2/252) * (1 - \alpha - \beta)$, resulting in an annual unconditional volatility of 20%.

Figure A.1: GPD estimator errors for the 99%, 97.5%, and 95% confidence levels



(a) VaR at 99% confidence level

(b) ES at 99% confidence level

(c) VaR at 97.5% confidence level

(d) ES at 97.5% confidence level

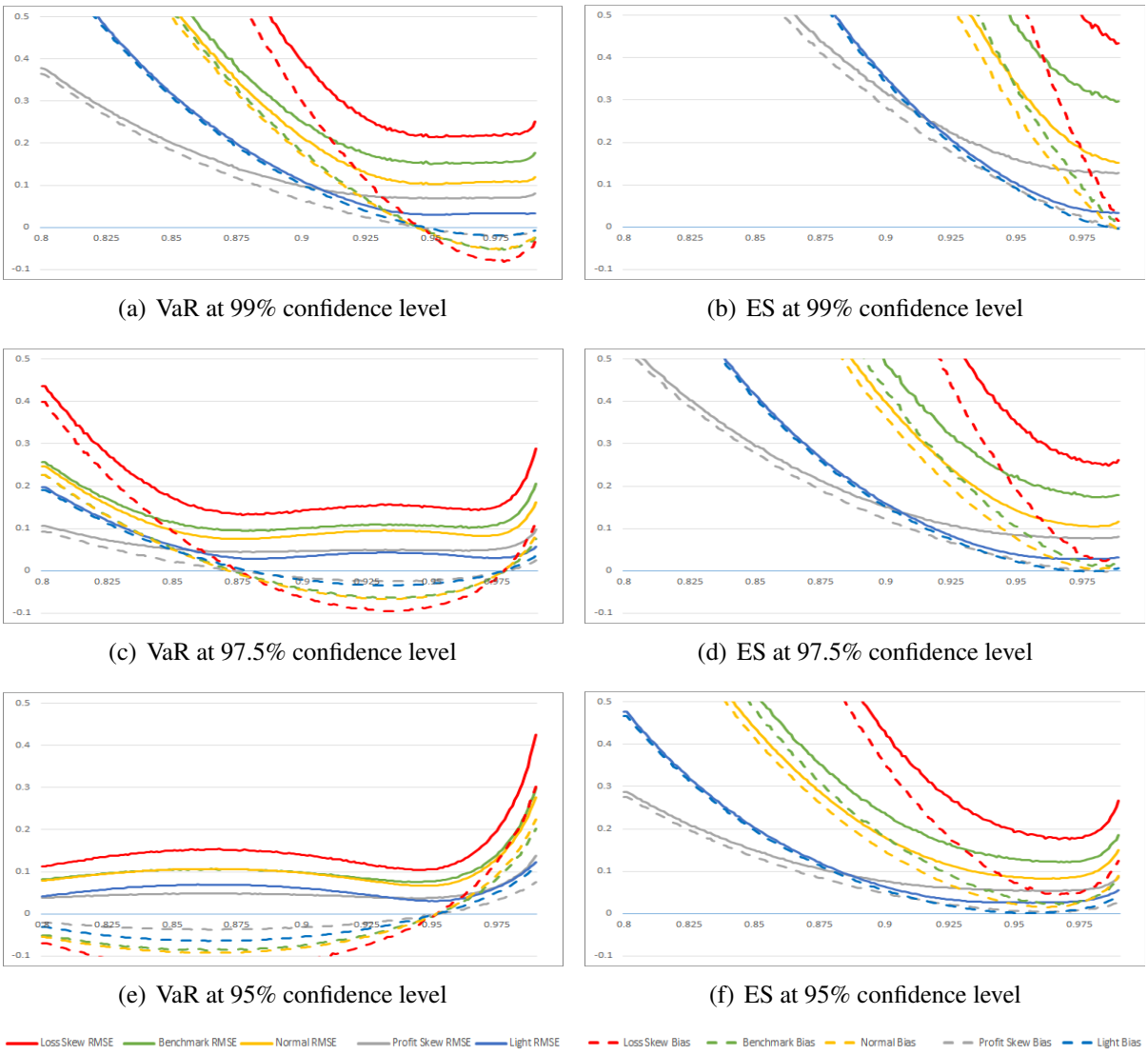
(e) VaR at 95% confidence level

(f) ES at 95% confidence level

This figure shows the RMSE and Bias for the GPD estimator as a function of the thresholds. Rows report results for different confidence levels. Left hand plots reports results for estimated VaR and right hand plots results for estimated ES.

To select the optimal threshold, $\hat{\eta}$, for GPD estimation, we compare the Bias and the root mean squared error (RMSE) for threshold quantiles between 0.8 and 0.99 using a 0.001 grid. Figure A.1 plots these as a function of threshold quantile choice for both VaR and ES at the 99%, 97.5% and 95% confidence levels. The figure shows that both VaR and ES have stable threshold choices up to the 0.94 quantile. The instability after the 0.94 quantile is caused by too

Figure A.2: Hill estimator errors for the 99%, 97.5%, and 95% confidence levels



(a) VaR at 99% confidence level

(b) ES at 99% confidence level

(c) VaR at 97.5% confidence level

(d) ES at 97.5% confidence level

(e) VaR at 95% confidence level

(f) ES at 95% confidence level

This figure shows the RMSE and Bias for the Hill estimator as a function of the thresholds. Rows report results for different confidence levels. Left hand plots reports results for estimated VaR and right hand plots results for estimated ES.

few excesses resulting in high variance. We find the 0.85 quantile of innovations a reasonable threshold estimate for both the 97.5% and 95% confidence levels, since it minimizes the absolute Bias and RMSE in all five DGPs. For the 99% confidence level, Panel (a) confirms that the 0.9 quantile is reasonable for VaR as in McNeil and Frey (2000). However, Panel (b) shows that the ES Bias and RMSE increases with the threshold, especially in the Benchmark and Loss

Skew cases. Hence, for VaR and ES the 0.85 quantile of innovations is more reasonable for the 99% confidence level.

We employ the same strategy to estimate the optimal threshold, $\hat{u}$, for Hill estimation at the 99%, 97.5% and 95% confidence levels. Figure A.2 plots the Bias and RMSE as a function of threshold choice for both VaR and ES at the 99%, 97.5% and 95% confidence levels. The figure shows that for the 97.5% confidence level, the threshold estimate $\hat{u}$ is optimized at the 0.975 quantile of innovations for evaluating both VaR and ES. For the 95% confidence level, the threshold estimate $\hat{u}$ is optimized at the 0.95 quantile of innovations for evaluating both VaR and ES. Our choice of Hill estimation method threshold at the 99% level confirms the results from Christoffersen and Gonçalves (2005) where $\hat{u}$ is the 0.97 quantile of innovations for VaR calculations and the 0.99 quantile of innovations for evaluating ES.

At each confidence level, the optimal threshold for both EVT models are remarkably robust across the different DGPs. This justifies holding the threshold fixed at the optimal quantile for empirical returns, where the true DGP is unknown. Thus, we recommend setting the threshold at the 0.85 quantile for the GPD method and equal to the confidence level for the Hill method.

A.3 Backtesting methods for ES

This section summarizes the ES backtesting procedures used in Section 2.4 of Liu and Stentoft (2020b).

A.3.1 Residual Test

The Residual Test by McNeil and Frey (2000) uses simulation methods to test residuals defined as

$$R_{t+1} = \{L_{t+1} - ES^{p}_{t+1} | L_{t+1} > VaR^{p}_{t+1}\}, \tag{A.1}$$

where under the true model these residuals are i.i.d. and have expected value 0 conditional on the losses exceeding VaR^p_{t+1}.[2] They test the one-sided hypothesis of mean zero residuals

$$H_0 : E[R_{t+1}|L_{t+1} > VaR^p_{t+1}] \leq 0 \text{ against } H_1 : E[R_{t+1}|L_{t+1} > VaR^p_{t+1}] > 0, \tag{A.2}$$

using a bootstrap test that makes no assumptions about the underlying distribution of residuals (Efron and Tibshirani (1994)). A rejection of the hypothesis is evidence that the model systematically underestimates ES. We report the p-value P_{ER} of their one-sided test.

A.3.2 Conditional Calibration Backtest

Nolde and Ziegel (2017) introduce conditional calibration backtests using identification functions (i.e. moment conditions). The strict identification function of VaR and ES for probability coverage p is

$$V(L, VaR, ES) = \begin{pmatrix} p - \mathbb{1}_{L>VaR} \\ VaR - ES - \mathbb{1}_{L>VaR}(VaR - L)/p \end{pmatrix}, \tag{A.3}$$

which has expectation zero if and only if we input the true VaR and ES of losses L. They test the forecasts $\widehat{VaR}$ and $\widehat{ES}$ based on the hypothesis

$$H_0 : E(V(L_t, \widehat{VaR}_t, \widehat{ES}_t)|\mathfrak{I}_{t-1}) \geq 0 \text{ against } H_1 : E(V(L_t, \widehat{VaR}_t, \widehat{ES}_t)|\mathfrak{I}_{t-1}) < 0, \tag{A.4}$$

component-wise for all $t = 1, .., T$, where $\mathfrak{I}_t$ is the information at time t. The requirement $E(V(L_t, \widehat{VaR}_t, \widehat{ES}_t)|\mathfrak{I}_{t-1}) \geq 0$ is equivalent to stating $E(h_t' V(L_t, \widehat{VaR}_t, \widehat{ES}_t)) \geq 0$ for all $\mathfrak{I}_{t-1}$-measurable R^2-valued functions $\mathbf{h}_t$. Since this is infeasible, they consider a $\mathfrak{I}_{t-1}$-predictable sequence of qx2-matrices $\mathbf{h}_t$ called test functions to construct Wald-type test statistic

$$T_{CC} = T\left(\frac{1}{T}\sum_{t=1}^{T} \mathbf{h}_t V(L_t, \widehat{VaR}_t, \widehat{ES}_t)\right)' \hat{\Omega}_T^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} \mathbf{h}_t V(L_t, \widehat{VaR}_t, \widehat{ES}_t)\right) \tag{A.5}$$

[2] McNeil and Frey (2000) also define a second version of residuals that are standardized by volatility. These residuals are conditional on volatility estimates and cannot be used to backtest ES for HS.

where $\hat{\Omega}_T^{-1} = \frac{1}{T}\sum_{t=1}^{T}(\mathbf{h}_t V(L_t, \widehat{VaR}_t, \widehat{ES}_t))(\mathbf{h}_t V(L_t, \widehat{VaR}_t, \widehat{ES}_t))'$. Under H_0, the test statistic is asymptotically χ^2_q distributed. We report the p-value P_{CCa} of their simple conditional calibration test, where the test function is the identity matrix, $\mathbf{h}_t = I_2$.

A.3.3 Regression Based Backtest

Bayer and Dimitriadis (2018) test whether a series of ES forecasts is correctly specified relative to returns Y_t. They regress returns on expected shortfall forecasts and an intercept term

$$Y_t = \gamma_1 + \gamma_2 \hat{e}_t + u_t^e, \tag{A.6}$$

where $-\hat{e}_t$ is the ES forecast and $E(u_t^e | Y_t < -VaR_t, \mathfrak{I}_{t-1}) = 0$. Given (A.6), the condition on the error term is equivalent to

$$E(Y_t | Y_t < -VaR_t, \mathfrak{I}_t) = \gamma_1 + \gamma_2 \hat{e}_t. \tag{A.7}$$

They test hypothesis

$$H_0 : (\gamma_1, \gamma_2) = (0, 1) \text{ against } H_1 : (\gamma_1, \gamma_2) \neq (0, 1), \tag{A.8}$$

where under H_0 the ES forecasts are correctly specified and $e_t = E(Y_t | Y_t < -VaR_t, \mathfrak{I}_{t-1})$. Using the FZ0 loss function to estimate model parameters, they propose the Auxiliary, Strict, and Intercept backtests. The Auxiliary backtest jointly tests VaR and ES by specifying regression (A.6) and VaR regression

$$Y_t = \beta_1 + \beta_2 \hat{v}_t + u_t^q, \tag{A.9}$$

where $-\hat{v}_t$ is the VaR forecast and $P(Y_t < u_t^q | \mathfrak{I}_{t-1}) = p$. They test hypothesis (A.8) using Wald-type test statistic

$$T_{ESR-Aux} = T(\hat{\gamma}_T - (0, 1))\hat{\Omega}_\gamma^{-1}(\hat{\gamma}_T - (0, 1))', \tag{A.10}$$

based on covariance estimator $\hat{\Omega}_{\gamma}^{-1}$ for the covariance of $\gamma = (\gamma_1, \gamma_2)$. They further consider a Strict Backtest using only ES forecasts as inputs. They approximate $\hat{e} = c\hat{v}_t$ for some constant c and replace regression (A.9) with

$$Y_t = \beta_1 + \beta_2 \hat{e}_t + u_t^q, \tag{A.11}$$

and test hypothesis (A.8) using test statistic (A.10). Finally, they propose a one-sided backtest to evaluate if ES forecasts are systematically low. Their Intercept backtest regresses forecasting errors $Y_t - \hat{e}_t$ on an intercept term

$$Y_t - \hat{e}_t = \beta_1 + u_t^q \text{ and } Y_t - \hat{e}_t = \gamma_1 + u_t^e, \tag{A.12}$$

where $P(Y_t < u_t^q | \mathfrak{I}_{t-1}) = p$ and $E(u_t^e | Y_t < -VaR_t, \mathfrak{I}_{t-1}) = 0$. Using this restricted regression, they test hypothesis

$$H_0 : \gamma_1 \geq 0 \text{ against } H_0 : \gamma_1 < 0, \tag{A.13}$$

using test statistic (A.10). We report two-sided p-values for their Strict $P_{ESR-Strict}$ and Auxiliary $P_{ESR-Aux}$ tests. We also report one-sided p-values for their Intercept $P_{ESR-Int}$ test.

A.3.4 Spectral Backtests

Gordy and McNeil (2020) propose backtests using new data reported to regulatory supervisors. Since 2013, banks are required to report the probability associated with the realized return based on the previous day's forecast distribution. This is mathematically expressed as the probability integral transformation (PIT) process $P_t = \hat{F}_t(L_t)$. If distribution function $\hat{F}_t$ is continuous and correctly specified, then P_t should be i.i.d standard uniform. Also, for any $\alpha \in (0, 1)$, let $\widehat{VaR}_{\alpha,t} = \hat{F}_t^{-1}(\alpha)$ be the estimated α-VaR constructed at time $t-1$, then a VaR exceedance event $\{L_T \geq \widehat{VaR}_{\alpha,t}\}$ is equivalent to the event $\{P_t \geq \alpha\}$. Hence, PIT-values contain information about VaR exceedances at every level α, allowing powerful tests of the deficiencies

in model $\hat{F}$ to be constructed.

Gordy and McNeil (2020) create tests based on spectral transformations of PIT exceedances

$$W_t = \int_{[0,1]} \mathbb{1}_{P_t \geq u} d\nu(u), \tag{A.14}$$

where kernel measure ν is a Lebesgue-Stieltjes measure on [0,1]. The choice of measure ν allow regulators to customize weights for quantiles of interest (for example, regulators are mainly concerned about large losses corresponding to α close to 1). Let $\Im^*_{t-1}$ denote the regulator's filtration generated by the PIT values, the null hypothesis is

$$W_t \sim F^0_W \text{ and } W_t \perp \Im^*_{t-1}, \forall t, \tag{A.15}$$

where F^0_W denotes the distribution of W_t when P_t is uniform, which changes according the kernel measure. The null hypothesis implies that the W_t are i.i.d., but is weaker than a null hypothesis that the P_t are i.i.d. uniform. Under the null hypothesis, $\overline{W}_n = n^{-1} \sum_{t=1}^n W_t$ is asymptotically normal and their spectral Z-test of unconditional coverage is

$$Z_n = \frac{\sqrt{n}\overline{W}_n - \mu_W}{\sigma_W} \xrightarrow[n\to\infty]{d} N(0,1), \tag{A.16}$$

where $\mu_W = E(W_t)$ and $\sigma^2_W = var(W_t)$ are the moments in null model F^0_W. We report p-values for the unconditional test using the Uniform (Uni), Arcsin (Arc), and Epanechnikov (Epa) continuous kernel density functions on a wide interval [0.95,0.995] and narrow interval [0.97,0.98]. We also use the discrete 3-level uniform kernel on points (0.95,0.975,0.99).

Gordy and McNeil (2020) also test for conditional coverage, which detects dependence in exceedances (i.e. Christoffersen (1998), Christoffersen and Pelletier (2004), and Nolde and Ziegel (2017)) based on the martingale difference (MD) property $E(W_t - \mu_W | \Im^*_{t-1}) = 0$. When the MD property holds, then $E(h_{t-1}(W_t - \mu_W)) = 0$ for any $\Im^*_{t-1}$-measurable random variable h_{t-1}. Using function $h(p)$, called the conditioning variable transformation (CVT), they form

a lagged vector $\mathbf{h}_{t-1} = (1, h(P_{t-1}), ..., h(P_{t-k}))$. In our empirical analysis, we set $k = 4$ lags and $h(p) = |2p - 1|^4$, which places heavier weight on tail PIT values in the recent past. Let $\mathbf{Y}_t = \mathbf{h}_{t-1}(W_t - \mu_W)$, where under the null hypotheses $E(\mathbf{Y}_t|\Im^*_{t-1}) = 0$. Similar to Nolde and Ziegel (2017), they construct Wald-type test statistic

$$T_{MD} = (n-k)\overline{\boldsymbol{Y}}'_{n,k}\hat{\Sigma}_Y^{-1}\overline{\boldsymbol{Y}}_{n,k} \tag{A.17}$$

where $\overline{\boldsymbol{Y}}_{n,k} = (n-k)^{-1}\sum_{t=k+1}^{n}\mathbf{Y}_t$ and $\hat{\Sigma}_Y^{-1}$ a consistent estimator of $\Sigma_Y = Cov(\mathbf{Y}_t)$. Under the null hypothesis, they use asymptotic theory from Giacomini and White (2006) to show that T_{MD} is χ^2_{k+1} distributed. We report p-values for the conditional tests using the same kernels and intervals as the unconditional tests.

A.3.5 FZ Loss Comparative Backtest

Fissler, Ziegel, and Gneiting (2015) propose a joint VaR and ES backtest using their FZ loss function from Equation (2.36) in Liu and Stentoft (2020b). They propose pairwise comparative backtests by taking differences in average losses between two models and defining the test statistic

$$T_{1,2} = \frac{\bar{L}_{FZ}(Y_t, v_{1t}, e_{1t}; \alpha, G_1, G_2) - \bar{L}_{FZ}(Y_t, v_{2t}, e_{2t}; \alpha, G_1, G_2)}{\sigma_T} = \frac{\Delta\bar{L}_{FZ}(Y_t, v_{1t}, e_{1t}, v_{2t}, e_{2t}))}{\sigma_T}, \tag{A.18}$$

where $\bar{L}_{FZ}(Y, v_i, e_i; \alpha, G_1, G_2) = \frac{1}{T}\sum_{t=1}^{T} L_{FZ}(Y, v_i, e_i; \alpha, G_1, G_2)$ for models $i = 1, 2$ and σ_T is an estimate of the respective standard deviation. Using the test statistic requires choosing functions G_1 and G_2. We use the FZ0 loss function from Patton, Ziegel, and Chen (2019) and Diebold and Mariano (2002) test statistics for the loss differences between models, to compare out-of-sample predictions of VaR and ES. Patton, Ziegel, and Chen (2019) show that FZ0 is the only FZ loss function that generates loss differences that are homogeneous of degree zero, a property that has been shown in volatility forecasting applications to lead to higher power in Diebold and Mariano (2002) tests. We also follow Patton, Ziegel, and Chen (2019) by

calculating the t-statistic using Newey-West standard errors with 20 lags (Newey and West (1987a)). Finally, we use FZ0 losses to find the Hansen, Lunde, and Nason (2011) Model Confidence Set, which contains the best model with a given level of confidence.

A.3.6 Murphy Diagrams

FZ loss comparisons require specifying the loss function. Fissler and Ziegel (2016) show that a large family of FZ loss functions is consistent and can be used to evaluate forecast performance. Ziegel et al. (2017) define the forecast dominance of model 1 over model 2 as

$$E(L_{FZ}(Y, v_1, e_1; p)) \leq E(L_{FZ}(Y, v_2, e_2; p)), \text{for all } L_{FZ} \in \mathcal{S}, \tag{A.19}$$

where $\mathcal{S}$ is the class of consistent LZ loss functions. In the previous section, we used the FZ0 scoring function for model comparison. However, Patton (2019) shows that in the presence of estimation error or model misspecification, forecast rankings can be sensitive to the choice of loss function.

In a recent paper, Ehm et al. (2016) show that any loss function that is consistent for a quantile or expectile has a mixture representation, providing a method to compare models across all consistent loss functions. Ziegel et al. (2017) derive a mixture representation for their joint VaR and ES FZ loss function. Defining elementary loss functions $L_{v_1}(v, Y)$ and $L_{v_2}(v, e, Y)$ for coverage probability p as

$$L_{\theta_1}(v, Y) = (\mathbb{1}_{Y \leq v} - p)(\mathbb{1}_{\theta_1 \leq v} - \mathbb{1}_{\theta_1 \leq Y}), \tag{A.20}$$

$$L_{\theta_2}(v, e, Y) = \mathbb{1}_{\theta_2 \leq e}(\frac{1}{p}\mathbb{1}_{Y \leq v}(v - Y) - (v - \theta_2)) + \mathbb{1}_{\theta_2 \leq v}(Y - \theta_2), \tag{A.21}$$

where θ_1 and θ_2 are real-valued thresholds, they show that all FZ loss functions can be written

as a mixture representation

$$L_{FZ}(Y, v, e; p) = \int L_{\theta_1}(v, Y)dH_1(\theta_1) + \int L_{\theta_2}(v, e, Y)dH_2(\theta_2), \tag{A.22}$$

where H_1 is a locally finite measure and H_2 is a measure that is finite on all intervals of the form $(-\infty, x], x \in \mathbb{R}$. Since they focus on comparing ES forecasts, they define $\mathcal{S}_2 \subset \mathcal{S}$ as the class of all consistent scoring functions with $H_1 = 0$. The mixture representation in Equation (A.22) can be graphically displayed using Murphy diagrams (Ehm et al. (2016)) by plotting the average score $\frac{1}{T}\sum_{t=1}^{T} L_{\theta_2}(v_t, e_t, Y_t)$ on θ_2 for loss functions in $\mathcal{S}_2$. They use the representation to formally test forecast dominance under hypothesis

$$H_{0,\theta_2} : E(L_{\theta_2}(Y, v_1, e_1; p)) = E(L_{\theta_2}(Y, v_2, e_2; p)), \tag{A.23}$$

$$H_{1,\theta_2} : E(L_{\theta_2}(Y, v_1, e_1; p)) > E(L_{\theta_2}(Y, v_2, e_2; p)), \tag{A.24}$$

where a rejection of H_{0,θ_2} suggests that model 2 performs strictly better. The test is repeated for threshold values θ_2 on a grid, yielding a sequence of point-wise p-values. They then compute WestFall-Young corrected p-values, P_{WY}, to control the family-wise error rate and report the minimal p-value (Westfall and Young (1993)).

A.4 Comparative Backtests for the 99% and 95% Confidence Levels

Table A.1: Comparative Expected Shortfall Backtests at the 99% Confidence Level

	HS	RM	FHS	Norm	STD	SSTD	GED	SGED	SGT	GPD	Hill	HillH	FZ1	FZH
Panel A: Average FZ0 Loss														
Loss	0.05	-0.04	-0.24	-0.14	-0.22	-0.24*	-0.22	-0.24	-0.24	-0.24	-0.24	-0.24	-0.13	-0.14
Panel B: Diebold-Mariano Test Statistics for FZ0 Loss														
HS	NA	1.34	3.69	2.70	3.79	3.83	3.71	3.84	3.83	3.64	3.68	3.73	2.34	2.65
RM	-1.34	NA	3.26	2.74	3.82	3.55	3.80	3.62	3.55	3.21	3.25	3.36	1.43	1.67
FHS	**-3.69**	**-3.26**	NA	**-2.57**	-0.85	0.99	-1.16	0.24	0.85	0.14	-0.62	0.97	**-4.14**	**-3.94**
Norm	**-2.70**	**-2.74**	2.57	NA	3.73	3.05	3.83	3.19	3.04	2.47	2.55	2.67	-0.24	0.11
STD	**-3.79**	**-3.82**	0.85	**-3.73**	NA	1.62	**-2.48**	1.65	1.59	0.79	0.83	1.03	**-3.54**	**-3.16**
SSTD	**-3.83**	**-3.55**	-0.99	**-3.05**	-1.62	NA	-1.88	-1.13	-1.30	-0.78	-1.01	-0.45	**-4.63**	**-4.38**
GED	**-3.71**	**-3.80**	1.16	**-3.83**	2.48	1.88	NA	2.00	1.85	1.09	1.14	1.32	**-3.10**	**-2.71**
SGED	**-3.84**	**-3.62**	-0.24	**-3.19**	-1.65	1.13	**-2.00**	NA	0.92	-0.15	-0.26	0.07	**-4.52**	**-4.27**
SGT	**-3.83**	**-3.55**	-0.85	**-3.04**	-1.59	1.30	-1.85	-0.92	NA	-0.64	-0.87	-0.33	**-4.57**	**-4.33**
GPD	**-3.64**	**-3.21**	-0.14	**-2.47**	-0.79	0.78	-1.09	0.15	0.64	NA	-0.19	0.31	**-4.16**	**-3.96**
Hill	**-3.68**	**-3.25**	0.62	**-2.55**	-0.83	1.01	-1.14	0.26	0.87	0.19	NA	1.00	**-4.13**	**-3.92**
HillH	**-3.73**	**-3.36**	-0.97	**-2.67**	-1.03	0.45	-1.32	-0.07	0.33	-0.31	-1.00	NA	**-4.10**	**-3.87**
FZ1	**-2.34**	-1.43	4.14	0.24	3.54	4.63	3.10	4.52	4.57	4.16	4.13	4.10	NA	0.86
FZH	**-2.65**	-1.67	3.94	-0.11	3.16	4.38	2.71	4.27	4.33	3.96	3.92	3.87	-0.86	NA

This table shows the comparative ES backtesting results of the representative portfolio from 01/1997 to 02/2020. Results are based on the FZ0 loss function at the 99% confidence level. Panel A shows the average FZ0 loss with a star beside models in the 75% model confidence set. Panel B shows t-statistics from Diebold-Mariano tests comparing FZ0 average losses. A negative value indicates that the row model has lower average loss than the column model. t-statistics greater than 1.96 in absolute value indicate the loss difference is significantly different from zero at the 95% confidence level. Models with t-stat below -1.96 indicate the row model is superior and are in bold.

Table A.2: Comparative Expected Shortfall Backtests at the 95% Confidence Level

	HS	RM	FHS	Norm	STD	SSTD	GED	SGED	SGT	GPD	Hill	HillH	FZ1	FZH
Panel A: Average FZ0 Loss														
Loss	-0.50	-0.60	-0.65*	-0.63	-0.64	-0.65*	-0.64	-0.65*	-0.65	-0.65	-0.65*	-0.64	-0.64	-0.64
Panel B: Diebold-Mariano Test Statistics for FZ0 Loss														
HS	NA	3.19	4.28	4.10	4.20	4.42	4.35	4.41	4.40	4.33	4.26	4.22	3.92	3.92
RM	**-3.19**	NA	3.59	3.10	3.34	3.95	3.81	3.91	3.88	3.64	3.51	3.64	2.40	2.40
FHS	**-4.28**	**-3.59**	NA	-1.87	-1.51	-0.02	-0.99	0.71	-0.22	0.14	-0.06	-1.88	-1.13	-1.08
Norm	**-4.10**	**-3.10**	1.87	NA	2.60	2.81	3.92	2.44	2.68	2.02	1.76	1.68	0.26	0.40
STD	**-4.20**	**-3.34**	1.51	**-2.60**	NA	2.40	2.91	2.03	2.25	1.65	1.41	0.95	-0.09	0.07
SSTD	**-4.42**	**-3.95**	0.02	**-2.81**	**-2.40**	NA	-1.88	0.87	-0.93	0.08	0.01	**-2.97**	-1.18	-1.06
GED	**-4.35**	**-3.81**	0.99	**-3.92**	**-2.91**	1.88	NA	1.60	1.65	1.11	0.91	-0.27	-0.59	-0.42
SGED	**-4.41**	**-3.91**	-0.71	**-2.44**	**-2.03**	-0.87	-1.60	NA	-1.11	-0.84	-0.61	**-2.88**	-1.31	-1.22
SGT	**-4.40**	**-3.88**	0.22	**-2.68**	**-2.25**	0.93	-1.65	1.11	NA	0.33	0.18	**-2.82**	-1.12	-1.00
GPD	**-4.33**	**-3.64**	-0.14	**-2.02**	-1.65	-0.08	-1.11	0.84	-0.33	NA	-0.14	**-2.04**	-1.17	-1.10
Hill	**-4.26**	**-3.51**	0.06	-1.76	-1.41	-0.01	-0.91	0.61	-0.18	0.14	NA	-1.62	-1.12	-1.07
HillH	**-4.22**	**-3.64**	1.88	-1.68	-0.95	2.97	0.27	2.88	2.82	2.04	1.62	NA	-0.52	-0.36
FZ1	**-3.92**	**-2.40**	1.13	-0.26	0.09	1.18	0.59	1.31	1.12	1.17	1.12	0.52	NA	0.27
FZH	**-3.92**	**-2.40**	1.08	-0.40	-0.07	1.06	0.42	1.22	1.00	1.10	1.07	0.36	-0.27	NA

This table shows the comparative ES backtesting results of the representative portfolio from 01/1997 to 02/2020. Results are based on the FZ0 loss function at the 95% confidence level. Panel A shows the average FZ0 loss with a star beside models in the 75% model confidence set. Panel B shows t-statistics from Diebold-Mariano tests comparing FZ0 average losses. A negative value indicates that the row model has lower average loss than the column model. t-statistics greater than 1.96 in absolute value indicate the loss difference is significantly different from zero at the 95% confidence level. Models with t-stat below -1.96 indicate the row model is superior and are in bold.

A.5 Backtests for the Risk Factor and Liquidity Horizon Portfolios

Table A.3: Traditional ES Backtests for Risk Factors at the 97.5% Confidence Level

	HS	RM	FHS	Norm	STD	SSTD	GED	SGED	SGT	GPD	Hill	HillH	FZ1	FZH
Panel A: Results for Interest Rate Risk at a 97.5% confidence level														
ER	**0.00**	**0.00**	0.87	**0.00**	0.71	0.73	0.05	0.15	0.85	0.36	0.96	**0.00**	0.93	0.75
ESR Strict	**0.00**	**0.02**	0.96	**0.01**	0.89	0.71	0.63	0.99	0.64	0.93	0.68	0.07	0.82	0.82
ESR Aux	**0.00**	**0.02**	0.96	**0.01**	0.86	0.75	0.60	0.99	0.71	0.97	0.75	0.07	0.86	0.93
ESR Int	**0.00**	**0.00**	0.42	**0.00**	0.20	0.66	0.09	0.34	0.70	0.44	0.65	**0.00**	0.59	0.53
CCa	0.05	**0.01**	1.00	**0.00**	0.86	1.00	0.45	0.69	1.00	0.81	1.00	**0.03**	1.00	0.89
Loss	-0.29	-0.35	-0.36	-0.35	-0.36	-0.35	-0.36*	-0.35	-0.35	-0.35	-0.36	-0.35	-0.32	-0.33
Panel B: Results for Equity Risk at a 97.5% confidence level														
ER	0.06	**0.00**	0.57	**0.00**	0.28	0.71	**0.01**	0.16	0.54	0.72	0.72	**0.00**	0.89	0.91
ESR Strict	**0.00**	**0.00**	0.44	**0.00**	**0.00**	0.12	**0.00**	**0.04**	0.09	0.50	0.52	**0.00**	0.12	0.25
ESR Aux	**0.00**	**0.00**	0.29	**0.00**	**0.00**	0.09	**0.00**	**0.04**	0.05	0.36	0.33	**0.00**	0.09	0.17
ESR Int	**0.00**	**0.00**	0.15	**0.00**	**0.00**	**0.02**	**0.00**	**0.01**	**0.01**	0.34	0.21	**0.00**	**0.04**	0.09
CCa	**0.00**	**0.00**	1.00	**0.00**	**0.00**	0.34	**0.00**	0.15	0.28	1.00	1.00	**0.00**	0.38	0.42
Loss	1.11	1.06	0.94	1.02	0.98	0.94	0.96	0.94	0.94	0.93*	0.94	0.94	0.99	0.98
Panel C: Results for Commodity Risk at a 97.5% confidence level														
	HS	RM	FHS	Norm	STD	SSTD	GED	SGED	SGT	GPD	Hill	HillH	FZ1	FZH
ER	**0.00**	**0.00**	0.18	**0.00**	**0.01**	0.26	**0.00**	**0.01**	0.23	0.29	0.35	**0.00**	0.65	0.72
ESR Strict	**0.00**	**0.00**	0.16	**0.00**	**0.00**	0.11	**0.00**	0.06	0.10	0.20	0.22	**0.00**	0.19	0.78
ESR Aux	**0.00**	**0.00**	0.18	**0.00**	**0.00**	0.11	**0.00**	0.06	0.12	0.22	0.23	**0.00**	0.19	0.70
ESR Int	**0.00**	**0.00**	0.16	**0.00**	**0.00**	0.12	**0.00**	**0.04**	0.11	0.25	0.23	**0.00**	0.25	0.44
CCa	**0.02**	**0.00**	0.83	**0.00**	**0.01**	0.64	**0.00**	0.24	0.60	0.77	0.97	**0.01**	1.00	0.97
Loss	0.48	0.49	0.43*	0.46	0.44	0.43*	0.44	0.43*	0.43*	0.43*	0.43*	0.43	0.46	0.44
Panel D: Results for Foreign Exchange Risk at a 97.5% confidence level														
	HS	RM	FHS	Norm	STD	SSTD	GED	SGED	SGT	GPD	Hill	HillH	FZ1	FZH
ER	**0.01**	**0.00**	0.40	**0.00**	0.79	0.85	0.38	0.34	0.76	0.34	0.62	**0.00**	0.26	0.22
ESR Strict	**0.01**	**0.00**	0.88	**0.01**	0.92	0.69	0.99	0.84	0.70	0.94	0.96	**0.01**	0.61	0.82
ESR Aux	**0.01**	**0.00**	0.83	**0.01**	0.91	0.69	0.99	0.85	0.69	0.92	0.93	**0.01**	0.62	0.83
ESR Int	**0.00**	**0.00**	0.18	**0.00**	0.46	0.62	0.33	0.52	0.61	0.22	0.28	**0.00**	0.07	0.15
CCa	**0.03**	**0.00**	1.00	**0.01**	1.00	1.00	0.77	0.41	1.00	0.94	1.00	**0.01**	0.60	0.81
Loss	-0.13	-0.20	-0.22*	-0.22	-0.22*	-0.22*	-0.22*	-0.22*	-0.22*	-0.22*	-0.22*	-0.21	-0.17	-0.18
Panel E: Results for Credit Risk at a 97.5% confidence level														
	HS	RM	FHS	Norm	STD	SSTD	GED	SGED	SGT	GPD	Hill	HillH	FZ1	FZH
ER	**0.03**	**0.00**	0.60	**0.00**	0.37	0.44	**0.00**	**0.00**	0.53	0.53	0.71	**0.00**	0.93	0.91
ESR Strict	**0.00**	**0.00**	0.69	**0.00**	0.60	0.76	0.33	0.72	0.76	0.60	0.53	0.44	0.49	0.15
ESR Aux	**0.00**	**0.00**	0.61	**0.00**	0.62	0.73	0.35	0.72	0.70	0.53	0.45	0.58	0.58	0.20
ESR Int	**0.00**	**0.00**	0.31	**0.00**	0.05	0.23	**0.02**	0.09	0.23	0.36	0.40	**0.02**	0.63	0.87
CCa	0.06	**0.00**	0.41	**0.00**	0.51	0.72	0.19	0.05	0.69	0.12	0.41	0.21	1.00	0.78
Loss	-0.70	-0.85	-0.96	-0.95	-0.97*	-0.97*	-0.97*	-0.96	-0.97*	-0.97	-0.96	-0.96	-0.90	-0.91

This table shows the traditional ES backtesting results of the five risk factor portfolios from 01/1997 to 02/2020. Each panel reports results for a particular confidence level. In each panel, Row 1 shows the one-sided p-values for the Exceedance Residual test. Rows 2 and 3 show the two-sided p-values for the Strict and Auxiliary ES regression backtests. Rows 4 and 5 show the one-sided p-values for the Intercept ES regression and Conditional Calibration backtests. Row 6 shows the average FZ0 loss with a star beside models in the 75% model confidence set. Models with p-values below 0.05 are in bold.

Table A.4: Traditional ES Backtests for Liquidity Horizons at the 97.5% Confidence Level

	HS	RM	FHS	Norm	STD	SSTD	GED	SGED	SGT	GPD	Hill	HillH	FZ1	FZH
Panel A: Results for LH 20 at a 97.5% confidence level														
ER	**0.04**	**0.00**	0.76	**0.00**	**0.04**	0.63	**0.00**	**0.04**	0.74	0.82	0.84	**0.03**	0.93	0.91
ESR Strict	**0.00**	**0.00**	0.92	**0.00**	**0.00**	0.51	**0.00**	0.19	0.53	0.92	0.83	0.11	0.90	0.75
ESR Aux	**0.00**	**0.00**	0.92	**0.00**	**0.00**	0.54	**0.00**	0.20	0.57	0.88	0.86	0.15	0.96	0.77
ESR Int	**0.00**	**0.00**	0.37	**0.00**	**0.00**	0.09	**0.00**	**0.02**	0.10	0.45	0.48	**0.01**	0.38	0.37
CCa	**0.01**	**0.00**	1.00	**0.00**	**0.00**	0.35	**0.00**	0.11	0.35	1.00	1.00	0.09	1.00	1.00
Loss	-0.31	-0.42	-0.52	-0.46	-0.49	-0.52*	-0.50	-0.52*	-0.52	-0.52*	-0.52	-0.51	-0.48	-0.48
Panel B: Results for LH 40 at a 97.5% confidence level														
ER	**0.02**	**0.00**	0.79	**0.00**	0.48	0.76	**0.00**	**0.01**	0.68	0.69	0.83	0.30	0.77	0.80
ESR Strict	**0.00**	**0.00**	0.54	**0.00**	**0.04**	0.81	**0.01**	0.37	0.76	0.54	0.32	0.88	0.84	0.91
ESR Aux	**0.00**	**0.00**	0.38	**0.00**	0.06	0.86	**0.02**	0.40	0.83	0.39	0.20	0.95	0.97	0.89
ESR Int	**0.00**	**0.00**	0.51	**0.00**	**0.00**	0.11	**0.00**	**0.03**	0.09	0.50	0.62	0.10	0.08	0.14
CCa	**0.03**	**0.00**	0.54	**0.00**	**0.03**	0.80	**0.01**	0.29	0.76	0.09	0.54	0.54	1.00	1.00
Loss	-0.31	-0.42	-0.66	-0.59	-0.64	-0.66	-0.65	-0.67*	-0.67*	-0.66	-0.66	-0.66*	-0.52	-0.54
Panel C: Results for LH 60 at a 97.5% confidence level														
ER	**0.02**	**0.00**	0.75	**0.00**	0.76	0.81	**0.00**	**0.01**	0.63	0.75	0.82	0.40	0.86	0.89
ESR Strict	**0.00**	**0.00**	0.09	**0.00**	**0.01**	0.43	**0.01**	0.13	0.37	0.06	**0.02**	0.71	0.95	0.43
ESR Aux	**0.00**	**0.00**	0.08	**0.00**	**0.02**	0.37	**0.01**	0.15	0.40	0.07	**0.01**	0.63	0.85	0.37
ESR Int	**0.00**	**0.00**	0.51	**0.00**	**0.00**	**0.03**	**0.00**	**0.01**	**0.02**	0.57	0.69	0.19	0.33	0.38
CCa	**0.09**	**0.00**	0.06	**0.00**	**0.01**	0.25	**0.04**	0.15	0.32	**0.03**	0.06	0.06	1.00	1.00
Loss	-0.05	-0.11	-0.36	-0.28	-0.32	-0.34	-0.36	-0.36	-0.36	-0.36	-0.36	-0.37*	-0.24	-0.26
Panel D: Results for LH 120 at a 97.5% confidence level														
ER	0.17	**0.00**	0.63	**0.00**	0.60	0.49	**0.00**	**0.04**	0.49	0.46	0.72	0.10	0.81	0.87
ESR Strict	**0.00**	**0.00**	**0.00**	**0.02**	0.19	**0.01**	0.17	**0.02**	**0.01**	**0.00**	**0.00**	0.07	**0.00**	**0.00**
ESR Aux	**0.00**	**0.00**	**0.00**	**0.02**	0.15	**0.00**	0.16	**0.02**	**0.00**	**0.00**	**0.00**	0.07	**0.03**	**0.00**
ESR Int	**0.03**	**0.00**	0.84	**0.00**	0.23	0.68	0.15	0.51	0.68	0.87	0.88	0.57	0.97	0.98
CCa	0.60	**0.00**	**0.01**	**0.04**	1.00	**0.04**	0.18	**0.00**	**0.04**	**0.00**	**0.01**	**0.01**	**0.03**	**0.04**
Loss	-0.37	-0.49	-0.51	-0.53	-0.55*	-0.52	-0.54	-0.51	-0.52	-0.50	-0.51	-0.53	-0.48	-0.49

This table shows the traditional ES backtesting results of the 20-120 day Liquidity Horizon portfolios from 01/1997 to 02/2020. Each panel reports results for a particular confidence level. In each panel, Row 1 shows the one-sided p-values for the Exceedance Residual test. Rows 2 and 3 show the two-sided p-values for the Strict and Auxiliary ES regression backtests. Rows 4 and 5 show the one-sided p-values for the Intercept ES regression and Conditional Calibration backtests. Row 6 shows the average FZ0 loss with a star beside models in the 75% model confidence set. Models with p-values below 0.05 are in bold.

Appendix B

Appendices for Chapter 3

B.1 Data

This appendix provides further details on our trade and quote (TAQ) data cleaning procedures and describes how the additional characteristics used in the empirical analysis are calculated.

B.1.1 Trade and quotes cleaning

The trade and quotes (TAQ) data require substantial cleaning due to contamination from market microstructure noise. We use the Monthly TAQ Second database up to 2003 and the Daily TAQ Millisecond/Microsecond database from 2004-2016 and filter noisy observations following the procedures similar to Barndorff-Nielsen et al. (2008). Returns are primarily calculated using prices from the trades database. Bid and ask observations from the quotes database are also used for cleaning trades and calculating transaction costs. For both trades and quotes, entries with time stamps outside of the trading day (9:30 AM to 4:00 PM) are removed. Entries with a zero or negative bid, ask, or price are also removed. For each stock, we keep only entries from the exchange with the highest volume in the month, and delete entries from other exchanges.

For only the quotes database, we construct the national-best bid-ask (NBBO) following procedures from (https://support.sas.com/resources/papers/proceedings16/11201-2016.pdf) using

quotes from all exchanges at a second interval. We remove all entries with a negative spread (ask minus bid) and quotes larger than 50 times the median spread on each day.

For only the trades data, entries with corrected trades (CORR $\neq$ 0) are deleted. Entries with an abnormal sale condition are removed, only keeping entries with COND equal to E, F, @, *, @E, @F, *E and *F. Multiple trades with the same second timestamp are replaced with an entry with the median price. Finally, quotes are used to discipline the trade prices: Entries are removed if their price is above the ask plus spread or below the bid minus spread.

The cleaned second-by-second price, bid and ask data is then aggregated into five-minute intervals and merged with the daily center for research in security prices (CRSP) data file by the CUSIP key. Merging the TAQ and CRSP data is challenging, because their tickers often differ. Additionally, tickers change in time due to mergers, acquisitions, and other corporate events. We instead merge the TAQ and CRSP databases using CUSIPs. The CUSIP identifier of each stock is obtained by merging trades with the TAQ master files. Finally, each stock is indexed by CRSP PERMINOs, which are unique and do not change in time.

B.1.2 Additional firm characteristics

In addition to returns with one lag, we consider other lagged characteristics of the SPY and S&P 500 constituents. These characteristics proxy short-term changes in liquidity and price trends. We create high-frequency analogs of characteristics from the asset pricing literature that have been shown to predict returns, including firm-level market beta, momentum, illiquidity, maximum, minimum, and trading volume. We also consider higher order moments, including volatility, skewness, and kurtosis. These characteristics are all calculated over preceding days. The particular variables used are the following:

Market beta: The market beta is motivated by the capital asset pricing model of Sharpe (1964) and Lintner (1965). At time $1 \leq t \leq T$, given (SPY market) return of stock $1 \leq \ell \leq 500$,

r_t^ℓ (r_t^m), it is based on a single factor model given by

$$r_t^\ell = \alpha^\ell + \beta^\ell(r_t^m) + \epsilon_t^\ell,$$

which is computed for each five-minute interval using a rolling window of preceding days, yielding one overnight and 78 intraday observations. BETA of stock ℓ during each five-minute interval, t, is the least-squares estimate, $\hat{\beta}^\ell$.

Momentum: Motivated by Jegadeesh and Titman (1993), the MOM of stock ℓ during each five-minute interval, t, is the cumulative return for the preceding day, yielding one overnight and 78 intraday observations; that is, $\sum_{t=1}^{T} r_t^\ell$ with $T = 79$.

Illiquidity: Motivated by Amihud (2002), the ILLIQ of stock ℓ, during each five-minute interval, t, is the ratio of the absolute stock return, $|r_t^\ell|$, to the dollar trading volume, averaged in a day excluding the overnight volume, i.e., over the $T = 78$ previous trading intervals in the day. Given the corresponding five-minute trading volume (price times number of shares traded) in dollars at time t, VOLD_t^ℓ, this yields

$$\text{ILLIQ}^\ell \equiv \frac{1}{T}\sum_{t=1}^{T}\frac{|r_t^\ell|}{\text{VOLD}_t^\ell}.$$

Maximum (minimum) return: Motivated by Bali, Cakici, and Whitelaw (2011), the MAX (MIN) of stock ℓ during each five-minute interval, t, is defined as the maximum (minimum) five-minute return within the preceding day (i.e., the previous 79 observations), that is, $\max_{1\le t\le T} r_t^\ell$ ($\min_{1\le t\le T} r_t^\ell$).

Trading volume: Motivated by Chordia, Subrahmanyam, and Anshuman (2001), the VOLUME of stock ℓ during each five-minute interval, t, is defined as the average number of shares traded within the preceding day excluding the overnight volume (i.e., the previous 78 observa-

tions):

$$\text{VOLUME}^{\ell} \equiv \frac{1}{T}\sum_{t=1}^{T} \text{SharesTraded}_t^{\ell}.$$

Volatility: Motivated by Ang et al. (2006), the VOL of stock ℓ during each five-minute interval, t, is the standard deviation of five-minute returns within the preceding day (i.e., the previous 79 observations):

$$\text{VOL}^{\ell} \equiv \sqrt{\frac{1}{T}\sum_{t=1}^{T}(r_t^{\ell} - \bar{r})^2} \equiv \text{sd}(r_t^{\ell}).$$

Here, $\bar{r} \equiv \frac{1}{T}\sum_{t=1}^{T} r_t^{\ell}$ denotes the corresponding arithmetic mean.

Skewness (kurtosis): Motivated by Amaya et al. (2015), the SKEW (KURT) of stock ℓ during each five-minute interval, t, is the skewness (kurtosis) of five-minute returns within the preceding day (i.e., the previous 79 observations):

$$\text{SKEW}^{\ell} \equiv \frac{1}{T}\sum_{t=1}^{T}\frac{(r_t^{\ell} - \bar{r})^3}{\text{sd}(r_t^{\ell})^3} \qquad \left(\text{KURT}^{\ell} \equiv \frac{1}{T}\sum_{t=1}^{T}\frac{(r_t^{\ell} - \bar{r})^4}{\text{sd}(r_t^{\ell})^4}\right).$$

B.2 Hyperparameters

This appendix discusses in more detail the most relevant hyperparameters that are tuned for each of the models used in the current paper. First, and applicable to all models, inherent in regularization is the particular binary problem of *whether* to regularize or not. Generally, the hyperparameter optimization problem is that of tuning the regularization parameters, λ in the case of lasso (LAS) regularization, with the binary problem of whether this parameters is zero or not. In elastic net (EN) regularization, the convex combination coefficient $\alpha \in [0, 1]$ is another hyperparameter to be optimized, with the endpoints $\{0, 1\}$ respectively corresponding to the discrete choices of only ridge or only lasso regularization, respectively. In principal component regression (PCR) the number of principal components, $1 \leq \kappa \leq K$, can be considered a

hyperparameter to be optimized.[1]

There are two crucial hyperparameters for random forest (RF) models: (1) the number of decision trees (or base learners) in the forest, and (2) the maximum depth of each of these trees, i.e., the maximum path length from any tree root to any of its leaves. More trees in the forest yield greater prediction variance reduction at increased computational cost. Restricting maximum tree depth limits the complexity, both saving computational time and avoiding overfitting. Similar goals may be achieved less directly, e.g., by limiting the number of training samples per leaf, which generally decreases with increased tree size. The same hyperparameters are important for gradient-boosted regression tree (GBRT) models, in addition to the explicit *learning rate*, ν, which may be factored out from the step sizes or weights γ_i. A smaller learning rate typically leads to better testing performance, but requires more training steps/additional decision trees in the ensemble, to maintain a given training error. Though these are the principal hyperparameters to consider for RF and GBRT models, several other, generally less important ones do exist. For example, bounding above the number of predictors determining node splits in base learners is relatively important for RF models in particular. Greater bounds permit more complexity, resulting in bias reduction at the expense of increased prediction variance and computational cost. Note that base learner complexity may be directly limited by bounding above the number of node splits or the total number of tree nodes, i.e., the tree size, itself.

For artificial neural networks (ANN), in particular, the loss function used for training is important. Standard choices include mean squared error, used in the current paper, as well as mean absolute error, logarithmic hyperbolic cosine, and a variety of others adapted to more specific scenarios.[2] We guard against model overfitting by adding ℓ^1 and ℓ^2 regularization to the loss function. Other important factors involved in training include the number of iterations,

[1] Other heuristics for determining κ is to accept the smallest value which achieves some level of total explained variance or to increase κ so long as the increase in variance explained exceeds a given threshold.

[2] Of particular interest is the Huber loss function, a combination of the ℓ^1 and ℓ^2 norms that permits control of the sensitivity to outliers, used in Gu, Kelly, and Xiu (2020c). Preliminary results show that this metric may outperform the MSE metric used for both the ANN and the GBRT models.

or epochs, and the level of *dropout*, the percentage of training data discarded in each epoch to avoid over-fitting and regularize the optimization problem (see, e.g., Srivastava et al. (2014) for a discussion).

While the optimization algorithm employed for training, the loss function it uses, the number of epochs, and the level of dropout, all constitute important *hyperparameters* impacting the quality of predictions from ANN models with trained *parameters*, the actual network architecture, i.e. the number of hidden layers, the number of neurons in each hidden layer, and the particular activation function associated with each neuron, may often have an even greater impact on the results. The activation functions used in this paper are the Rectified linear unit (ReLU) given by $\phi(x) = \max\{0, x\}$ (see Nair and Hinton (2010)), the Maxout given by $\phi(x_1, x_2) = \max\{x_1, x_2\}$ (see Goodfellow et al. (2013)), and the Hyperbolic tangent given by $\phi(x) = \tanh x$. We use a hyperparameter grid holding the number of hidden layers fixed at 3. Models are estimated using stochastic gradient descent optimization with the ADADELTA (Zeiler (2012)) adaptive learning rate method for faster estimation. We use dropout and early stopping as regularization techniques to prevent overfitting.

Table B.1 summarizes the set of hyperparameters that are tuned for each of the models considered in the current paper. To tune these at a particular time, the sample of data is partitioned into training, testing, and validation sets and used specifically in the following way: the optimization problem in Equation (3.5) is repeatedly solved using training predictors and targets, validating at each training iteration to determine whether to terminate training and minimize over-fitting using validation predictors and targets. When it comes to training, all decisions are made according to the goal of minimizing the metric, $m[\cdot]$, modifying internal parameters, like β in the case of the linear models, according to a specific optimization algorithm, e.g., classical stochastic gradient descent in the case of ANN models. When it comes to validation of the hyperparameters the mean squared error from the validation predictors and targets is used. For LAS and EN, λ is selected using coordinate descent (see Friedman, Hastie, and Tibshirani (2010)). For PCR grid search is used. For all other algorithms, random search is

Table B.1: Model Hyperparameters

LAS	λ
EN	λ, α = {0, 0.25, 0.5, 0.75, 1}
PCR	κ = 1-99 by 2
RF	#trees = 100-700 by 100, tree depth = 1-40 by 5, #minimum rows = {1, 25, 50}, columns randomly selected = floor(#number of features * {.05, .15, .25, .333, .4})
GBRT	#trees = 100-700 by 100, tree depth = 2-40 by 3, learning rate = 0.01 - 0.1 by 0.03, sample rate = 0.5-1 by 0.2, column sample rate = 0.1 - 1 by 0.2
ANN	Activation Functions = Relu, Maxout, Tanh; with and without dropout, Hidden nodes = {1000,500,10}, {100,50,10}, {16,14,12},{20,15,10},{25,17,10},{15,10,5}, epochs = {50,100}, dropout ratio = {0, 0.1, 0.2}, max w^2 = {10, 100, 1000, 3.4028235e+38}, ℓ^1 = {0, 0.00001, 0.0001}, ℓ^2 = {0, 0.00001, 0.0001}, rho (rate time decay) = {0.9, 0.95, 0.99, 0.999}, epsilon (rate time smoothing) = {1e-10, 1e-8, 1e-6, 1e-4}

This table describes the set of hyperparameters that are tuned for each individual model considered in the paper.

used with 50 models. Random search has been shown to be more efficient than grid search, see, e.g., Bergstra and Bengio (2012). In contrast to Gu, Kelly, and Xiu (2020c), we do not use ensembles of neural networks due to computational limitations.

Appendix C

Appendices for Chapter 4

C.1 Empirical Evidence of Power-Law Distributed Tails

In this section, I provide empirical evidence for the power-law definition of idiosyncratic tail risk ξ used in this paper. Since Mandelbrot (1963), numerous studies have documented that the tails of *daily* and *monthly* equity returns are power-law distributed.[1] Recently, researchers have been using high-frequency equity data to show the tails of *high-frequency* returns are power-law distributed.[2] Gabaix et al. (2006) shows the tail distribution of 15-minute US equity returns for the 1000 largest stocks from 1994-1995 has power-law distributed tails according to

$$P(|r| > x) \sim x^{-1/\xi} L(x) \tag{C.1}$$

where r is the log return (log denotes natural logarithm), $L(x)$ is a slowly varying function, and $\sim$ denotes asymptotic equivalence.[3]

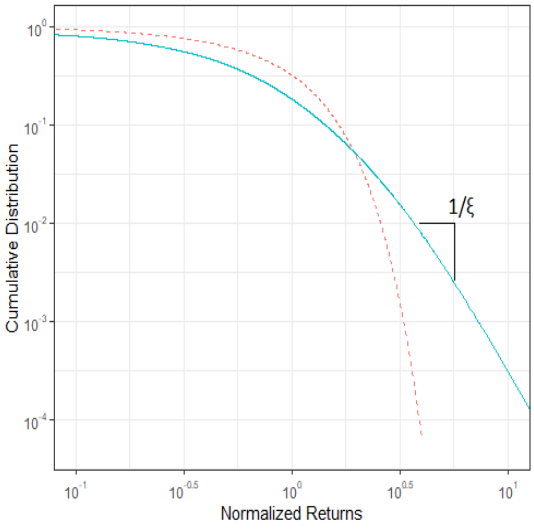
Following similar procedures as in Gabaix et al. (2006), Figure C.1 illustrates the empirical

[1] Power law tail behaviour of equity returns has been thoroughly documented in, for example, Mandelbrot (1963), Fama (1963), Jansen and De Vries (1991), Kearns and Pagan (1997), Samanta and LeBaron (2005), and Kelly and Jiang (2014b). For an overview, refer to Gabaix (2009).

[2] For examples, see Danielsson and De Vries (1997), Gabaix et al. (2006), Stanley, Plerou, and Gabaix (2008), and Bollerslev and Todorov (2011a)

[3] Function L is slowly varying if L is strictly positive and $\lim_{x\to\infty} L(tx)/L(x) = 1$ for all $t > 0$. Prototypical examples include $L(x) = log(x)$ and $L(x) = c$ for a $c > 0$. $f(x) \sim g(x)$ means $\lim_{x\to\infty} f(x)/g(x) \to 1$.

Figure C.1: Complementary Cumulative Distribution of High-Frequency Returns



This figure plots the empirical complementary cumulative distribution function $P(|r| > x)$ of absolute values of a sample of 5-minute log returns for the 1000 largest U.S stocks from 2001 to 2016. 10% of the returns are sampled resulting in approximately 30 million observations. Returns are normalized by stock to have mean 0 and standard deviation 1. logx is plotted on the horizontal axis and $logP(|r| > x)$ on the vertical axis.

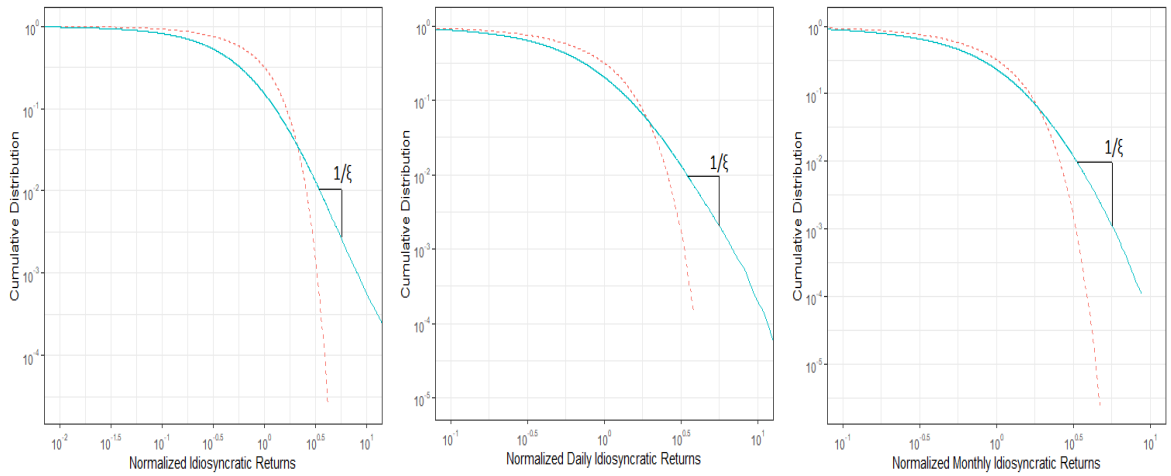
complementary cumulative distribution of a sample of 5-minute normalized absolute returns for the 1000 largest stocks from 2001 to 2016 using a natural log scale for the horizontal and vertical axis. The standard normal distribution is plotted in dashed lines for comparison. According to Figure C.1, the natural log of probability that returns are greater than 2 standard deviations is approximately linear in log of absolute returns following,

$$logP(|r| > x) \approx -\frac{1}{\xi} logx + c, \tag{C.2}$$

and confirming return tails are power-law distributed. Figure C.1 shows that return tails are much heavier than normal tails. Furthermore, in the data empirical returns greater than 3 standard deviations constitute 1.76% of the sample – approximately 6 times the probability implied by a normal distribution.

While empirical evidence shows that return tails obey a power-law, this does not guarantee that idiosyncratic return tails are power-law distributed. In factor models, returns are a function

Figure C.2: Complementary Cumulative Distribution of High-Frequency, Daily and Monthly Idiosyncratic Returns



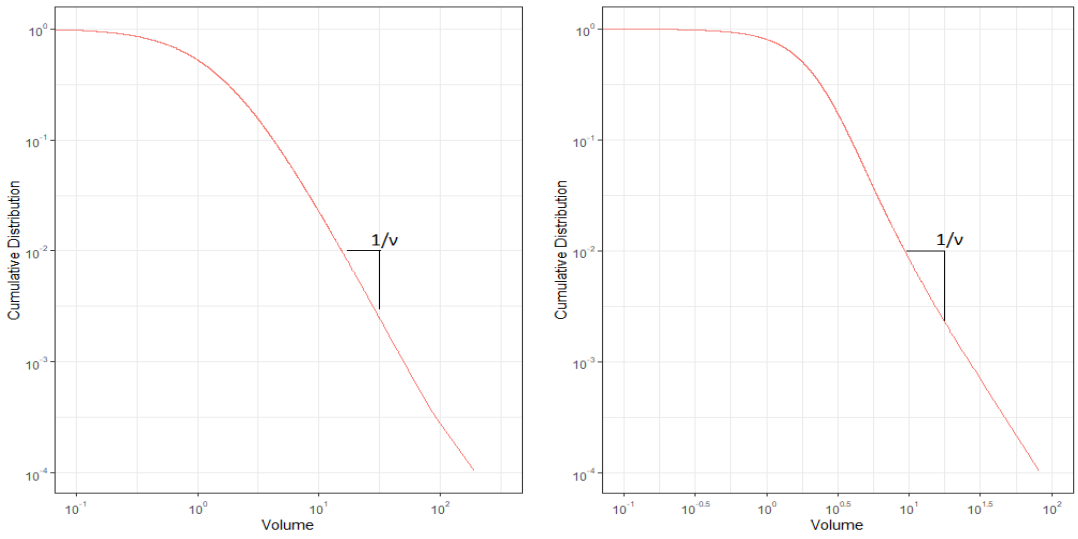
(a) High-Frequency Complementary Cumulative Distribution (b) Daily Complementary Cumulative Distribution (c) Monthly Complementary Cumulative Distribution

Panel (a) plots the empirical complementary cumulative distribution function of absolute values of a sample of normalized 5-minute residuals for the 1000 largest U.S stocks from 2001 to 2016. Residuals are estimated every month from regressing log returns on the high-frequency Fama and French (2015) five factors. 10% of the residuals are sampled resulting in approximately 30 million observations. Panel (b) plots the empirical cumulative distribution function of absolute values of daily residuals for the 1000 largest U.S stocks from 2001 to 2016. Panel (c) plots the empirical cumulative distribution function of absolute values of normalized monthly residuals for all U.S common stocks from 1963 to 2016. Daily and Monthly residuals are estimated for each firm by regressing log excess returns on the log of Fama and French (2015) five factors over the entire period. For further details on estimation see Section 4.2.4. High-frequency, daily, and monthly residuals are normalized by stock to have mean 0 and standard deviation 1. logx is plotted on the horizontal axis and $logP(|r| > x)$ on the vertical axis.

of systematic factors and idiosyncratic returns. Returns could be inheriting the power-law behaviour only from the tails of systematic factors and not the idiosyncratic returns. I present new evidence that the tails of idiosyncratic returns are also power-law distributed. Figure C.2 plots the complementary cumulative distribution of residuals from Fama and French (2015) five factor model.[4] Panel (a) plots high-frequency residuals, panel (b) plots daily residuals, and panel (c) plots monthly residuals. The distribution is linear in logs, confirming that idiosyncratic returns are power-law distributed. Returns and residuals have similar slopes $1/\xi$, suggesting

[4]For each firm, high-frequency, daily, or monthly residuals are created by regressing returns on high-frequency, daily, or monthly Fama and French (2015) five factors, where the factor betas are estimated using Least Absolute Deviation. See Section 4.2.4 for more details on estimation.

Figure C.3: Cumulative Distribution of High-Frequency and Daily Trading Volume



(a) High-Frequency Volume Cumulative Distribution

(b) Daily Volume Cumulative Distribution

Panel (a) plots the empirical cumulative distribution function $P(|r| > x)$ of high-frequency trading volume for the 1000 largest U.S stocks from 2001 to 2016. 10% of the volume observations are sampled resulting in approximately 30 million observations. Panel (b) plots the empirical cumulative distribution function $P(|v| > x)$ of daily trading volumes for all U.S common stocks from 2001 to 2016. For each stock, high-frequency and daily trading volumes are normalized by the stock's mean absolute deviation $\frac{v}{\frac{1}{N}\sum_{i=1}^{N}|v-\bar{v}|}$. logx is plotted on the horizontal axis and $logP(|v| > x)$ on the vertical axis.

that returns are inheriting their power-law behaviour mostly from idiosyncratic returns.[5] Additionally, slopes $1/\xi$ appear related for high-frequency, daily, and monthly idiosyncratic returns. Section 4.2.2 proves that that idiosyncratic tail risk ξ is preserved under time aggregation and that ξ is theoretically equivalent for high-frequency, daily, and monthly idiosyncratic returns.

Gabaix et al. (2006) shows that high-frequency and daily trading volume is power-law distributed. Following the procedures in their paper, Figure C.3 plots the cumulative distribution function of 5-minute and daily trading volume for the 1000 largest stocks from January 2001 to December 2016. Each stock's volume v is normalized by its mean absolute deviation.

Panel (a) of Figure C.3 shows that high-frequency volume greater than 1 absolute deviation follows a power law with volume tail risk parameter ν. Trading volume has much heavier tails

[5] ξ here is pooled across firms and time, while $\xi_{l,1}$ in Equation (4.1) is for a single firm-month.

than returns. The 10% largest 5-minute volumes represents 65.82% of the total volume traded and the 1% largest volumes represents 27.81% of the total volume traded.[6] Panel (b) of Figure C.3 shows that the tail of daily volume is also power-law distributed. This is unsurprising, since daily volume is the sum of high-frequency volumes with power-law distributed tails and power-law random variables are preserved under time aggregation. Additionally, high-frequency and daily trading volume have similar slopes $1/\nu$, suggesting that daily trading volume also inherits the tail distribution of high-frequency trading volume.

In summary, high-frequency returns are power-law distributed. Their availability provides a new means to effectively capture time-varying idiosyncratic tail risk. In addition, the empirical evidence shows that idiosyncratic returns are also power-law distributed, as such this is used as one of the key assumptions in this paper. Furthermore, daily and monthly idiosyncratic returns inherit the power-law distribution and tail risk parameter ξ from high-frequency idiosyncratic returns. Trading volume is also power-law distributed and its tail risk can be summarized by slope parameter ν. The next section introduces a new power-law factor model and idiosyncratic tail risk measure.

C.2 Proofs

C.2.1 Proof for Lemma 4.2.1

Proof Since $\boldsymbol{\beta}_t$ are constant for each month t, $\boldsymbol{\beta}_t$ is assumed to be the same in the monthly and high-frequency models. Take the sum of both sides of Equation (4.9) for periods $i = 1, .., N$, to obtain

$$\sum_{i=1}^{N} r_{t,i} = \boldsymbol{\beta}_t \sum_{i=1}^{N} \mathbf{f}_{t,i} + \sum_{i=1}^{N} x_{t,i}. \tag{C.3}$$

[6]The 0.1% largest volumes represents 9.52% of the total volume traded.

Then by the aggregation property of log returns, (C.3) can be written as

$$R_t = \boldsymbol{\beta}_t \mathbf{F}_t + \sum_{i=1}^{N} x_{t,i}, \tag{C.4}$$

which combined with (4.10) implies $X_t = \sum_{i=1}^{N} x_{t,i}$.

C.2.2 Levy's theorem (Auxiliary)

Theorem C.2.1 *Durrett (2019), Page 186. Suppose $Z_1, Z_2, ...$ are i.i.d. with a distribution that satisfies*

(i) $\lim_{z\to\infty} \frac{P(Z_1>z)}{P(|Z_1|>z)} = \theta \in [0,1]$

(ii) $P(|Z_1| > z) = z^{-\alpha} L(z)$,

where $\alpha < 2$ and L is a slowly varying function. Let $S_n = Z_1 + ... + Z_n$,

$$a_n = inf\{z : P(|Z_1| > z) \leq n^{-1}\} \text{ and } b_n = nE(Z_1 \mathbb{1}_{(Z_1 \leq a_n)})$$

As $n \to \infty$, $(S_n - b_n)/a_n \Rightarrow Y$ where Y has a nondegenerate distribution. Y follows a Levy distribution with tail shape α.

C.2.3 Proof for Theorem 4.2.2

Proof Let $\alpha_t = \frac{1}{\xi_t}$ in Levy's Theorem C.2.1 in Section C.2.2. Parameter α is often called the tail exponent. Assumption 1 (a) and (b) satisfy conditions (i) and (ii) of Levy's Theorem. Hence, by Levy's Theorem,

$$(\sum_{i=1}^{N} x_{t,i} - b_N)/a_N \xrightarrow{d} u_t \text{ as } N \to \infty, \tag{C.5}$$

where u is a Levy-distributed random variable with tail risk ξ_t, $a_N = inf\{y : P(|x_{t,1}| > y) \leq N^{-1}\}$, and $b_N = NE(x_{t,1} \mathbb{1}_{(x_{t,1} \leq a_N)})$. By Lemma 4.2.1, the monthly idiosyncratic return is equal to the

sum of high-frequency returns, that is $X_t = \sum_{i=1}^{N} x_{t,i}$. Substituting $X_t = \sum_{i=1}^{N} x_{t,i}$ into (C.5) gives the result in (4.11).

C.2.4 Jessen and Mikosch Lemma (Auxiliary)

Lemma C.2.2 (Jessen and Mikosch (2006), Lemma 3.1) *Assume $|Z_1|$ is regularly varying with shape $\alpha \geq 0$. Assume $Z_1, ..., Z_n$ are random variables satisfying*

$$\lim_{z\to\infty} \frac{P(Z_i > z)}{P(|Z_1| > z)} = c_i^+ \text{ and } \lim_{z\to\infty} \frac{P(Z_i \leq -z)}{P(|Z_1| > z)} = c_i^-, \ i = 1,..,n, \tag{C.6}$$

for some non-negative numbers $c_i^{\pm}$ and

$$\lim_{z\to\infty} \frac{P(Z_i > z, Z_j > z)}{P(|Z_1| > z)} = \lim_{z\to\infty} \frac{P(Z_i \leq -z, Z_j > z)}{P(|Z_1| > z)} = \lim_{z\to\infty} \frac{P(Z_i \leq -z, Z_j \leq -z)}{P(|Z_1| > z)} = 0, \ i \neq j, \tag{C.7}$$

then

$$\lim_{z\to\infty} \frac{P(Z_1 + ... + Z_n > z)}{P(|Z_1| > z)} = c_1^+ + ... + c_n^+ \text{ and } \lim_{z\to\infty} \frac{P(Z_1 + ... + Z_n \leq -z)}{P(|Z_1| > z)} = c_1^- + ... + c_n^-. \tag{C.8}$$

In particular if the Z_i's are independent non-negative regularly varying random variables then

$$P(Z_1 + ... + Z_n > z) \sim P(Z_1 > z) + ... + P(Z_n > z). \tag{C.9}$$

C.2.5 Proof for Theorem 4.2.3

Proof The result is an application of Lemma 3.2 (a) in Mikosch and Vries (2013), and I follow their proof closely with additional details. The proof uses the result from auxiliary Lemma C.2.2 (Jessen and Mikosch (2006)) in Section C.2.4 for the sum of regularly varying random variables. I show the high-frequency idiosyncratic returns, $x_{t,1}, .., x_{t,N}$, satisfy the conditions of

Lemma C.2.2.[7] Condition (C.6) is satisfied by the tail balance condition in Assumption 1(b),

$$\lim_{y\to\infty}\frac{P(x_{t,1} > y)}{P(|x_{t,1}| > y)} = (1-\theta) \text{ and } \lim_{y\to\infty}\frac{P(x_{t,1} \leq -y)}{P(|x_{t,1}| > y)} = \theta, \tag{C.10}$$

where $\theta \in (0, 1]$. Since the idiosyncratic returns are i.i.d., then

$$\lim_{y\to\infty}\frac{P(x_{t,i} > y)}{P(|x_{t,1}| > y)} = (1-\theta) \text{ and } \lim_{y\to\infty}\frac{P(x_{t,i} \leq -y)}{P(|x_{t,1}| > y)} = \theta, \text{ i = 1,..,N.} \tag{C.11}$$

To show condition (C.7), note that

$$\begin{aligned}\lim_{y\to\infty}\frac{P(x_{t,i} \leq -y, x_{t,j} \leq -y)}{P(|x_{t,1}| > y)} &= \lim_{y\to\infty}\frac{P(x_{t,i} \leq -y)P(x_{t,j} \leq -y)}{P(|x_{t,1}| > y)} \\ &= \lim_{y\to\infty}\frac{P(x_{t,i} \leq -y)}{P(|x_{t,1}| > y)}\frac{P(x_{t,j} \leq -y)}{P(|x_{t,1}| > y)}P(|x_{t,1}| > y),\end{aligned} \tag{C.12}$$

where the first equality uses the independence of $x_{t,i}$ and the second equality is from multiplying the denominator and numerator by $P(|x_{t,1}| > y)$. It has been proved that for any $\rho > 0$ and slowly varying function $L_\rho(y)$, $\lim_{y\to\infty} y^{-\frac{1}{\rho}}L_\rho(y) = 0$ (Karamata (1962)), hence $\lim_{y\to\infty} P(|x_{t,1}| > y) = \lim_{y\to\infty} y^{-\frac{1}{\xi_t}}L(y) = 0$. Applying the product rule of limits to the last equality in Equation (C.12) and using Assumption 1(b) gives,

$$\lim_{y\to\infty}\frac{P(x_{t,i} \leq -y, x_{t,j} \leq -y)}{P(|x_{t,1}| > y)} = \lim_{y\to\infty}\frac{P(x_{t,i} \leq -y)}{P(|x_{t,1}| > y)}\lim_{y\to\infty}\frac{P(x_{t,j} \leq -y)}{P(|x_{t,1}| > y)}\lim_{y\to\infty}P(|x_{t,1}| > y) = \theta^2\lim_{y\to\infty}y^{-\frac{1}{\xi_t}}L(y) = 0. \tag{C.13}$$

The other 2 conditions in (C.7) are proved analogously. Since the conditions in Lemma C.2.2 are satisfied, then

$$\lim_{y\to\infty}\frac{P(x_{t,1} + ... + x_{t,N} \leq -y)}{P(|x_{t,1}| > y)} = N\theta = N\lim_{y\to\infty}\frac{P(x_{t,1} \leq -y)}{P(|x_{t,1}| > y)}, \tag{C.14}$$

[7] In the notation of Lemma C.2.2, $Z_1 = x_{t,1}, Z_2 = x_{t,2}, ..., Z_N = x_{t,N}$.

where the first equality is by Lemma C.2.2 and the second equality is by Assumption 1(b). Hence,

$$P(x_{t,1} + ... + x_{t,N} \leq -y) \sim NP(x_{t,1} \leq -y), \tag{C.15}$$

and with the identity in Lemma 4.2.1 implies (4.12).

C.2.6 Proof for Theorem 4.2.4

Proof The result an application of Lemma 3.2 (c) in Mikosch and Vries (2013), and I follow their proof closely with additional details. Similar to the previous theorem, this proof uses the result from auxiliary Lemma C.2.2 (Jessen and Mikosch (2006)) in Appendix C.2.4 for the sum of regularly varying random variables. However, this theorem requires showing the conditions of Lemma C.2.2 are satisfied for any i by the pair $x_{t,i}$ and $\eta_{t,i}$.[8]

Fix any $i \in [1, N]$, then $|x_{t,i}|$ is regularly varying with shape $\frac{1}{\xi_t} > 0$. As in Theorem (4.2.3), $x_{t,i}$ satisfies condition (C.6) by Assumption 1(b). To show $\eta_{t,i}$ satisfies condition (C.6), note that

$$\lim_{y\to\infty} \frac{P(\eta_{t,i} \leq -y)}{P(|x_{t,i}| > y)} = \lim_{y\to\infty} \frac{P(\eta_{t,i} \leq -y)}{P(|\eta_{t,i}| > y)} \frac{P(|\eta_{t,i}| > y)}{P(|x_{t,i}| > y)}. \tag{C.16}$$

By Assumption 2(b), $\lim_{y\to\infty} \frac{P(\eta_{t,i}\leq -y)}{P(|\eta_{t,i}|>y)} = p$. Also,

$$\lim_{y\to\infty} \frac{P(|\eta_{t,i}| > y)}{P(|x_{t,i}| > y)} = \lim_{y\to\infty} \frac{y^{-\frac{1}{\gamma_t}} L_\eta(y)}{y^{-\frac{1}{\xi_t}} L(y)} = \lim_{y\to\infty} y^{-(\frac{1}{\gamma_t} - \frac{1}{\xi_t})} \frac{L_\eta(y)}{L(y)} = 0, \tag{C.17}$$

where the last equality is because $\frac{1}{\gamma_t} - \frac{1}{\xi_t} > 0$, $\frac{L_\eta(y)}{L(y)}$ is a slowly varying function, and for any $\rho > 0$ and slowly varying function $L_\rho(y)$, $\lim_{y\to\infty} y^{-\frac{1}{\rho}} L_\rho(y) = 0$ (see Karamata (1962)). Furthermore, by the product rule of limits,

[8] In the notation of Lemma C.2.2, $Z_1 = x_{t,i}$ and $Z_2 = \eta_{t,i}$

$$\lim_{y\to\infty}\frac{P(\eta_{t,i}\leq -y)}{P(|x_{t,i}|>y)}=\lim_{y\to\infty}\frac{P(\eta_{t,i}\leq -y)}{P(|\eta_{t,i}|>y)}\lim_{y\to\infty}\frac{P(|\eta_{t,i}|>y)}{P(|x_{t,i}|>y)}=p0=0. \tag{C.18}$$

The other case in condition (C.6) is proved analogously. To show condition (C.7), note that

$$\begin{aligned}\lim_{y\to\infty}\frac{P(x_{t,i}\leq -y,\eta_{t,i}\leq -y)}{P(|x_{t,1}|>y)}&=\lim_{y\to\infty}\frac{P(x_{t,i}\leq -y)P(\eta_{t,i}\leq -y)}{P(|x_{t,1}|>y)}\\&=\lim_{y\to\infty}\frac{P(x_{t,i}\leq -y)}{P(|x_{t,1}|>y)}\frac{P(\eta_{t,i}\leq -y)}{P(|\eta_{t,1}|>y)}P(|\eta_{t,1}|>y),\end{aligned} \tag{C.19}$$

where the first equality uses the independence of $x_{t,i}$ and $\eta_{t,i}$ and the second equality is from multiplying the denominator and numerator by $P(|\eta_{t,1}|>y)$. Applying the product rule of limits to the last equality in Equation (C.19) gives,

$$\lim_{y\to\infty}\frac{P(x_{t,i}\leq -y,x_{t,j}\leq -y)}{P(|x_{t,1}|>y)}=\lim_{y\to\infty}\frac{P(x_{t,i}\leq -y)}{P(|x_{t,1}|>y)}\lim_{y\to\infty}\frac{P(\eta_{t,i}\leq -y)}{P(|\eta_{t,1}|>y)}\lim_{y\to\infty}P(|\eta_{t,1}|>y)=\theta p\lim_{y\to\infty}y^{-\frac{1}{\eta_t}}L_\eta(y)=0. \tag{C.20}$$

The other 2 conditions in (C.7) are proved analogously. Since the assumptions of Lemma C.2.2 are satisfied for $x_{t,i}$ and $\eta_{t,i}$, then

$$\begin{aligned}\lim_{y\to\infty}\frac{P(x^*_{t,i}\leq -y)}{P(|x_{t,i}|>y)}=\lim_{y\to\infty}\frac{P(x_{t,i}+\eta_{t,i}\leq -y)}{P(|x_{t,i}|>y)}=\lim_{y\to\infty}\frac{P(x_{t,i}\leq -y)+P(\eta_{t,i}\leq -y)}{P(|x_{t,i}|>y)}=\\\lim_{y\to\infty}\frac{P(x_{t,i}\leq -y)}{P(|x_{t,i}|>y)}+\lim_{y\to\infty}\frac{P(\eta_{t,i}\leq -y)}{P(|x_{t,i}|>y)}=\lim_{y\to\infty}\frac{P(x_{t,i}\leq -y)}{P(|x_{t,i}|>y)},\end{aligned} \tag{C.21}$$

which implies that

$$P(x^*_{t,i}\leq -y)\sim P(x_{t,i}\leq -y). \tag{C.22}$$

C.3 Data

C.3.1 TAQ Cleaning

The TAQ requires substantial cleaning due to contamination from market microstructure noise. As demonstrated in Section 4.2.2, large outliers from microstructure effects must be filtered for tail risk measurement. I filter noisy observations following the procedures in Barndorff-Nielsen et al. (2008). My econometric analysis primarily uses prices and trade size from the trades database. Bid and ask observations from the quotes database are also used for cleaning trades and robustness analysis. For both trades and quotes, entries with time stamps outside of the trading day (9:30am to 4:00pm) are removed. Entries with a zero or negative bid, ask, or price are also removed. For each stock, I keep only entries from the exchange with the highest volume in the month, and delete entries from other exchanges.

For only the quotes database, entries with a negative spread (ask minus bid) are deleted. Multiple quotes with the same second timestamp are replaced with a single entry with the median bid and median ask. Entries with a spread more than 50 times the median spread on that day are removed. Finally, entries are deleted when the midquote (midpoint of bid and ask) deviates by more than 10 mean absolute deviations (excluding the considered observation) from a rolling centered mean of 50 observations (25 observations before and 25 after).

For only the trades data, entries with corrected trades (CORR $\neq$ 0) are deleted. Entries with an abnormal sale condition are removed, only keeping entries with COND equal to E, F, @, *, @E, @F, *E, and *F. Multiple trades with the same second timestamp are replaced with an entry with the median price. Finally, quotes are used to discipline the trade prices. Entries are removed if their price is above the ask plus spread or below the bid minus spread.

Merging the TAQ data with the CRSP data is challenging, because CRSP tickers often differ from TAQ tickers. Additionally, tickers change over time due to mergers, acquisitions, and other corporate events. I instead merge the TAQ and CRSP databases using CUSIPs. Each stock's CUSIP identifier is obtained by merging trades with the TAQ master files. Finally, each

stock is indexed by CRSP PERMINOs, which are unique and do not change over time.

C.3.2 Constructing High-Frequency Factors

I construct the high-frequency equity factors in the heavy-tailed factor model in Equation (4.9) following the procedures in Ait-Sahalia, Kalnina, and Xiu (2020). The models considered are the one-factor CAPM model (Sharpe (1964);Lintner (1965)), the Fama and French (1993) and Carhart (1997) four-factor model, and the Fama and French (2015) five-factor model. Factors in the models are reconstructed at 5-minute time intervals. Following Ait-Sahalia, Kalnina, and Xiu (2020), the construction combines the TAQ, CRSP, and COMPUSTAT databases in a multiple step procedure.

Since the constituents for the Fama and French factors are not publicly available, the first step is to replicate their factor construction. Compustat is used to create annual book equity (BE), operating profit (OP), and investment (INV) variables for each firm in North America. Only firms on Compustat for more than two years are included, to avoid survival bias. The firm-year variables are then merged with the CRSP monthly dataset. CRSP data is filtered according to Fama and French (1993), keeping only ordinary common equity (Share Code 10 or 11) from the NYSE, AMEX, and NASDAQ exchanges. Market equity (ME) for each month is the product of a firm's price and shares outstanding. Book-to-market (BE/ME) for June of year y is the ratio of book equity for y-1 fiscal year end divided by market equity for December of year y-1. Operating profitability and investment are also calculated during June of year y according to the formulas in Section C.3.3.

NYSE breakpoints are calculated during June of year y for ME, BE/ME, OP, and INV. Breakpoints determine portfolio groupings for each variable. Size groupings divide stocks with ME smaller (S) and bigger (B) than the median NYSE ME. Book-to-market groupings are denoted growth (L), neutral (N), and value (H), based on breakpoints for the lowest 30%, neutral 40%, and highest 30% of ranked BE/ME values for NYSE stocks. Six value-weighted portfolios are formed from the intersection of the two ME and three BE/ME groupings. The

value factor (HML) is the return of the high value portfolios minus low value (growth) portfolios, given by

$$HML = \frac{1}{2}(Small\ Value + Big\ Value) - \frac{1}{2}(Small\ Growth + Big\ Growth). \tag{C.23}$$

The size factor ($SML_{B/M}$) is the return of the small portfolios minus big portfolios, given by

$$SMB_{B/M} = \frac{1}{3}(Small\ Value + Small\ Neutral + Small\ Growth) - \frac{1}{3}(Big\ Value + Big\ Neutral + Big\ Growth). \tag{C.24}$$

In Fama and French (2015), the operating profitability groupings are denoted robust (R), neutral (N), and weak (W). The investment groupings are conservative (C), neutral (N), and aggressive (A). The operating profitability and investment breakpoints are the 30th and 70th NYSE percentiles of OP and INV. The profitability factor (RMW) is

$$RMW = \frac{1}{2}(Small\ Robust + Big\ Robust) - \frac{1}{2}(Small\ Weak + Big\ Weak),$$

where the value-weighted portfolios are formed by the intersection of OP and ME groupings. The investment factor (CMA) is

$$CMA = \frac{1}{2}(Small\ Conservative + Big\ Conservative) - \frac{1}{2}(Small\ Aggressive + Big\ Aggressive),$$

where the value-weighted portfolios are formed by the intersection of INV and ME groupings. Additionally, the size factor (SMB) in Fama and French (2015) is adjusted to

$$SMB = \frac{1}{3}(SMB_{B/M} + SMB_{OP} + SMB_{INV}), \tag{C.25}$$

where

$$SMB_{OP} = \frac{1}{3}(Small\ Robust + Small\ Neutral + Small\ Weak) - \frac{1}{3}(Big\ Robust + Big\ Neutral + Big\ Weak), \tag{C.26}$$

$$SMB_{INV} = \frac{1}{3}(Small\ Conservative + Small\ Neutral + Small\ Aggressive) - \frac{1}{3}(Big\ Conservative + Big\ Neutral + Big\ Aggressive), \tag{C.27}$$

and $SMB_{B/M}$ is given in Equation (C.24).

A stock's momentum is its return in the previous 12 months excluding the most recent month. Momentum portfolios are denoted up (U), neutral (N), and down (D) based on breakpoints at the 30th and 70th NYSE percentiles. Following Ken French's website, six value-weighted portfolios are reformed every month by intersecting MOM and ME groupings. The momentum factor (UMD) is

$$UMD = \frac{1}{2}(Small\ Up + Big\ Up) - \frac{1}{2}(Small\ Down + Big\ Down).$$

In each model, the market factor (MKT) is the value-weighted portfolio of all stocks considered in Fama and French (1993).

The replication procedure provides daily portfolio constituents and weights for each equity factor. The daily constituents are merged with the TAQ to create intraday factors at 5-minute intervals. The constituents are also merged with the CRSP daily database to create overnight factors. Portfolios are weighted using the previous day's market capitalization. The high-

frequency factors are transformed from simple returns to log returns using

$$f_{t,i} = log(1 + f_{t,i}^{Simple}),$$

where $f_{t,i}^{Simple}$ is the simple return of the high-frequency factor created from the procedures in this section.

In addition to intraday returns, I also include the overnight return of each stock in estimation procedures, which is consistent with the daily estimation procedures in Savor (2012), Jiang and Yao (2013), and Jiang and Zhu (2017). Overnight returns are calculated using daily CRSP data according to the formula in Lou, Polk, and Skouras (2019), which accounts for dividend adjustments, share splits, and other corporate events. I take the natural logarithm of their simple return to obtain log overnight return,

$$R_{overnight} = log(\frac{1 + R_{close-to-close}}{1 + R_{intraday}}), \tag{C.28}$$

where $R_{close-to-close}$ is the standard return from the daily CRSP, and $R_{intraday} = \frac{P_{close}}{P_{open}} - 1$. Overnight returns are merged with intraday returns for idiosyncratic tail risk estimation.

C.3.3 Additional Firm Characteristics

Market Beta: Market beta is based on single factor model

$$R_{l,d} - r_d^f = \alpha_l + \beta_l(R_d^m - r_d^f) + \epsilon_{l,d}, \tag{C.29}$$

where $R_{l,d}$ is the return of stock l on day d, $R_{m,d}$ is the market return, and r_d^f is the risk-free rate. Equation (C.29) is estimated monthly using a rolling window of the previous 252 days. BETA for stock l is the least squares estimate $\hat{\beta}_l$.

Size: Following Fama and French (1993), a firm's SIZE is the natural logarithm of the market value of equity. Market equity (ME) is the product of price and number of shares outstanding

in millions of dollars. The size factor (SMB) includes all common stocks in the NYSE, AMEX, and NASDAQ exchanges with non-missing market equity for June of year y.

Book-to-Market: Following Fama and French (1993), the BM ratio in June of year y is the firm's ratio of book value of common equity in fiscal year y-1 and market value of equity in December of year y-1. Book value of common equity is defined as book value of shareholders' equity, plus balance sheet deferred taxes and investment tax credit (if available), minus book value of preferred stock. The book value of common equity for firm l in fiscal year y-1 is

$$BE_{l,y-1} = SEQ_{l,y-1} + TXDB_{l,y-1} + ITCB_{l,y-1} - BVPS_{l,y-1},$$

where $SEQ_{l,y-1}$ is book value of shareholder's equity, $TXDB_{l,y-1}$ is deferred taxes, $ITCB_{l,y-1}$ is investment tax credit, and $BVPS_{l,y-1}$ is book value of preferred stock. Depending on availability, book value of preferred stock is estimated using the redemption ($PSTKRV_{l,y-1}$), liquidation ($PSTKL_{l,y-1}$), or par ($PSTK_{l,y-1}$) value in that order. The book-to-market factor (HML) includes all common stocks in the NYSE, AMEX, and NASDAQ exchanges with non-missing market equity for December of year y-1 and June of year y, and non-missing book equity for fiscal year y-1.

Operating Profitability: Following Fama and French (2015), the OP of firm l for June of year y is annual revenues minus cost of goods sold, interest expense, and selling, general, and administrative expenses divided by book equity for fiscal year y-1, that is,

$$OP_{l,y} = \frac{REVT_{l,y-1} - COGS_{l,y-1} - XINT_{l,y-1} - XSGA_{l,y-1}}{BE_{l,y-1}}$$

where $REVT_{l,y-1}$ is annual revenues, $COGS_{l,y-1}$ is cost of goods sold, $XINT_{l,y-1}$ is interest and related expenses, $XSGA_{l,y-1}$ is selling, general, and administrative expenses, and $BE_{l,y-1}$ is book equity for fiscal year y-1. The profitability factor (RMW) includes all common stocks on the NYSE, AMEX, and NASDAQ with market equity data for June of year y, positive $BE_{l,y-1}$, non-missing $REVT_{l,y-1}$, and non-missing data for at least one of $COGS_{l,y-1}$, $XINT_{l,y-1}$,

or $XSGA_{l,y-1}$.

Investment: Following FFama and French (2015), the INV of firm l for June of year y is the change of total assets from fiscal year y-2 to fiscal year y-1, divided by the total assets in fiscal year y-2, that is,

$$INV_{l,y} = \frac{AT_{l,y-1} - AT_{l,y-2}}{AT_{l,y-2}},$$

where $AT_{l,y-1}$ is the firm's total assets in fiscal year y-1. The investment factor (CMA) includes all common stocks on the NYSE, AMEX, and NASDAQ exchanges with market equity data for June of year y and total assets data for fiscal years y-1 and y-2.

Momentum: Following Jegadeesh and Titman (1993), the MOM of stock l in month t is the cumulative return for the 11 months from month t-12 to t-2, skipping the most recent month.

Short-term reversals: Following Jegadeesh (1990) and Lehmann (1990), the REV of stock l in month t is the return of the stock in the previous month, that is the return in month t-1.

Illiquidity: Following Amihud (2002), the ILLIQ of stock l in month t is the average daily ratio of the absolute stock return to the dollar volume, averaged in month t, that is,

$$ILLIQ_l = \frac{1}{D}\sum_{d=1}^{D}\frac{|R_{l,d}|}{VOLD_{l,d}},$$

where D is the number of trading days in the month, R_d is the return on day d, $VOLD_d$ is the respective daily volume in millions of dollars.

Maximum Daily Return: Following Bali, Cakici, and Whitelaw (2011), the MAX is defined as the maximum daily return within month t.

Idiosyncratic Volatility: Following Ang et al. (2006), idiosyncratic volatility is based on the Fama and French (1993) three factor model

$$R_{l,d} - r_d^f = \alpha_l + \beta_l(R_d^m - r_d^f) + \gamma_l SMB_d + \phi_l HML_d + \epsilon_{l,d},$$

where SMB_d and HML_d are daily returns on the size and book-to-market factors defined

in Section C.3.2. The IVOL of stock l in month t is the standard deviation of daily residuals in month t, that is, $IVOL_l = \sqrt{var(\epsilon_{l,d})}$.

Coskewness: Following Harvey and Siddique (2000), the COSKEW for stock l in month t is

$$COSKEW_l = \frac{E[(R_{l,d} - \bar{r}_{l,t})(R_d^m - \bar{r}_t^m)^2]}{\sqrt{var(R_{l,d})}var(R_d^m)},$$

where $R_{l,d}$ is the return of stock l on day d and R_d^m is the market return on day d over the past year. Variables $\bar{r}_{l,t}$ and $\bar{r}_t^m$ are respectively the average daily returns of stock l and the market portfolio over the past year.

Downside Risk: Following Ang, Chen, and Xing (2006), the $\beta_{l,down}$ for stock l in month t is calculated over the past year using excess daily stock returns and market returns, conditional on the excess market return moving below its average value, that is,

$$\beta_{l,down} = \frac{cov(R_{l,d} - r_d^f, R_d^m - r_d^f | R_d^m - r_d^f < \mu^m)}{var(R_d^m - r_d^f | R_d^m - r_d^f < \mu^m)},$$

where $R_{l,d}$ is the return of stock l on day d, R_d^m is the market return on day d, and μ^m is the average daily excess market return over the past year.

C.4 Robustness for Idiosyncratic Tail Risk and Expected Returns

C.4.1 Additional Univariate Sorts

Table C.1: Portfolios Sorted on Idiosyncratic Tail Risk estimated using the Market model residuals

	1 (Low)	2	3	4	5	6	7	8	9	10 (High)	(10-1)
Panel A: Univariate Sort on Idiosyncratic Tail Risk (Equal-Weighted)											
Excess Return	0.50	0.51	0.67	0.62	0.66	0.66	0.76	0.67	0.90	1.02	0.52
t-stat	(1.12)	(1.15)	(1.50)	(1.40)	(1.56)	(1.57)	(1.73)	(1.55)	(2.25)	(2.72)	(2.61)
FFC4 alpha	0.55	0.58	0.71	0.67	0.73	0.70	0.81	0.70	0.93	1.03	0.48
FFC4 t-stat	(1.32)	(1.39)	(1.73)	(1.66)	(1.84)	(1.79)	(1.99)	(1.78)	(2.45)	(2.91)	(2.40)
FF5 alpha	0.73	0.70	0.86	0.79	0.86	0.80	0.93	0.81	1.00	1.09	0.36
FF5 t-stat	(1.65)	(1.61)	(1.97)	(1.85)	(2.11)	(1.97)	(2.20)	(1.92)	(2.51)	(3.00)	(1.67)
Panel B: Univariate Sort on Idiosyncratic Tail Risk (Value-Weighted)											
Excess Return	0.28	0.39	0.47	0.41	0.40	0.49	0.45	0.48	0.73	1.02	0.74
t-stat	(0.73)	(1.12)	(1.27)	(1.20)	(1.21)	(1.45)	(1.22)	(1.14)	(1.97)	(2.84)	(3.75)
FFC4 alpha	0.33	0.45	0.50	0.43	0.44	0.52	0.53	0.49	0.78	1.05	0.71
FFC4 t-stat	(0.89)	(1.37)	(1.53)	(1.47)	(1.40)	(1.70)	(1.67)	(1.30)	(2.29)	(3.17)	(3.48)
FF5 alpha	0.49	0.55	0.57	0.51	0.54	0.60	0.63	0.63	0.80	1.18	0.68
FF5 t-stat	(1.28)	(1.66)	(1.70)	(1.72)	(1.81)	(1.88)	(1.95)	(1.61)	(2.23)	(3.55)	(3.18)

The table reports monthly average idiosyncratic tail risk, excess returns, and alphas for portfolios sorted on idiosyncratic tail risk between January 2001 to December 2016. Idiosyncratic tail risk is estimated from the high-frequency the market model residuals. Panel A reports equally-weighted excess returns and alphas sorted on idiosyncratic tail risk. Panel B reports value-weighted excess returns and alphas sorted on idiosyncratic tail risk.

Table C.2: Portfolios Sorted on Idiosyncratic Tail Risk estimated using the FFC residuals

	1 (Low)	2	3	4	5	6	7	8	9	10 (High)	(10-1)
Panel A: Univariate Sort on Idiosyncratic Tail Risk (Equal-Weighted)											
Excess Return	0.51	0.56	0.68	0.66	0.56	0.76	0.72	0.62	0.88	1.01	0.50
t-stat	(1.16)	(1.26)	(1.61)	(1.51)	(1.27)	(1.77)	(1.66)	(1.45)	(2.13)	(2.70)	(2.45)
FFC4 alpha	0.57	0.60	0.76	0.73	0.60	0.81	0.76	0.65	0.90	1.03	0.46
FFC4 t-stat	(1.34)	(1.47)	(1.88)	(1.83)	(1.47)	(2.02)	(1.92)	(1.69)	(2.32)	(2.90)	(2.25)
FF5 alpha	0.74	0.72	0.87	0.88	0.74	0.93	0.88	0.76	0.97	1.09	0.35
FF5 t-stat	(1.63)	(1.68)	(2.04)	(2.10)	(1.72)	(2.19)	(2.16)	(1.89)	(2.39)	(2.99)	(1.60)
Panel B: Univariate Sort on Idiosyncratic Tail Risk (Value-Weighted)											
Excess Return	0.34	0.43	0.36	0.61	0.29	0.57	0.42	0.52	0.60	0.91	0.57
t-stat	(0.93)	(1.25)	(1.07)	(1.77)	(0.82)	(1.69)	(1.10)	(1.31)	(1.40)	(2.37)	(2.65)
FFC4 alpha	0.40	0.46	0.42	0.66	0.30	0.63	0.45	0.57	0.62	0.91	0.51
FFC4 t-stat	(1.09)	(1.46)	(1.33)	(2.10)	(0.96)	(2.13)	(1.37)	(1.62)	(1.55)	(2.64)	(2.39)
FF5 alpha	0.54	0.55	0.52	0.74	0.37	0.75	0.58	0.65	0.66	1.02	0.48
FF5 t-stat	(1.42)	(1.75)	(1.64)	(2.33)	(1.11)	(2.51)	(1.75)	(1.80)	(1.58)	(2.95)	(2.24)

The table reports monthly average idiosyncratic tail risk, excess returns, and alphas for portfolios sorted on idiosyncratic tail risk between January 2001 to December 2016. Idiosyncratic tail risk is estimated from the high-frequency Fama-French-Carhart model residuals. Panel A reports equally-weighted excess returns and alphas sorted on idiosyncratic tail risk. Panel B reports value-weighted excess returns and alphas sorted on idiosyncratic tail risk.

Table C.3: Portfolios Sorted on Idiosyncratic Tail Risk estimated using the Industry statistical model residuals

	1 (Low)	2	3	4	5	6	7	8	9	10 (High)	(10-1)
Panel A: Univariate Sort on Idiosyncratic Tail Risk (Equal-Weighted)											
Excess Return	0.45	0.57	0.64	0.61	0.62	0.73	0.73	0.71	0.81	0.98	0.53
t-stat	(1.00)	(1.20)	(1.36)	(1.28)	(1.31)	(1.51)	(1.53)	(1.50)	(1.77)	(2.24)	(2.95)
FFC4 alpha	0.51	0.62	0.70	0.67	0.66	0.78	0.78	0.75	0.84	0.98	0.47
FFC4 t-stat	(1.21)	(1.43)	(1.64)	(1.52)	(1.52)	(1.74)	(1.76)	(1.73)	(1.96)	(2.46)	(2.69)
FF5 alpha	0.68	0.75	0.85	0.83	0.77	0.90	0.90	0.85	0.94	1.07	0.39
FF5 t-stat	(1.49)	(1.66)	(1.89)	(1.79)	(1.68)	(1.91)	(1.96)	(1.87)	(2.07)	(2.57)	(2.25)
Panel B: Univariate Sort on Idiosyncratic Tail Risk (Value-Weighted)											
Excess Return	0.21	0.42	0.48	0.59	0.50	0.57	0.72	0.60	0.67	0.91	0.70
t-stat	(0.63)	(1.11)	(1.35)	(1.55)	(1.25)	(1.49)	(1.83)	(1.45)	(1.53)	(2.29)	(3.53)
FFC4 alpha	0.27	0.45	0.54	0.63	0.55	0.60	0.76	0.64	0.68	0.94	0.67
FFC4 t-stat	(0.81)	(1.34)	(1.67)	(1.95)	(1.57)	(1.78)	(2.19)	(1.70)	(1.74)	(2.67)	(3.39)
FF5 alpha	0.37	0.53	0.64	0.74	0.65	0.72	0.89	0.72	0.84	1.02	0.65
FF5 t-stat	(1.06)	(1.56)	(1.88)	(2.23)	(1.82)	(2.07)	(2.61)	(1.94)	(2.11)	(2.92)	(3.14)

The table reports monthly average idiosyncratic tail risk, excess returns, and alphas for portfolios sorted on idiosyncratic tail risk between January 2001 to December 2016. Idiosyncratic tail risk is estimated from the high-frequency industry statistical model residuals. Panel A reports equally-weighted excess returns and alphas sorted on idiosyncratic tail risk. Panel B reports value-weighted excess returns and alphas sorted on idiosyncratic tail risk.

Table C.4: Portfolios Sorted on Idiosyncratic Tail Risk estimated using the High-Frequency Cross-Sectional Variable model residuals

	1 (Low)	2	3	4	5	6	7	8	9	10 (High)	(10-1)
Panel A: Univariate Sort on Idiosyncratic Tail Risk (Equal-Weighted)											
Excess Return	0.46	0.53	0.61	0.62	0.58	0.69	0.76	0.72	0.86	0.96	0.50
t-stat	(1.00)	(1.12)	(1.32)	(1.27)	(1.21)	(1.41)	(1.57)	(1.50)	(1.87)	(2.15)	(2.80)
FFC4 alpha	0.51	0.59	0.68	0.68	0.62	0.73	0.80	0.75	0.89	0.96	0.45
FFC4 t-stat	(1.20)	(1.35)	(1.59)	(1.52)	(1.43)	(1.66)	(1.79)	(1.70)	(2.06)	(2.37)	(2.50)
FF5 alpha	0.65	0.74	0.85	0.80	0.73	0.87	0.95	0.85	0.98	1.05	0.40
FF5 t-stat	(1.43)	(1.64)	(1.89)	(1.70)	(1.59)	(1.92)	(2.03)	(1.80)	(2.13)	(2.50)	(2.15)
Panel B: Univariate Sort on Idiosyncratic Tail Risk (Value-Weighted)											
Excess Return	0.30	0.31	0.50	0.55	0.32	0.69	0.66	0.76	0.62	0.72	0.42
t-stat	(0.85)	(0.86)	(1.23)	(1.57)	(0.80)	(1.85)	(1.70)	(1.74)	(1.34)	(1.66)	(1.88)
FFC4 alpha	0.33	0.37	0.50	0.58	0.38	0.75	0.73	0.78	0.62	0.72	0.40
FFC4 t-stat	(1.02)	(1.13)	(1.45)	(1.79)	(1.06)	(2.21)	(2.05)	(2.20)	(1.49)	(1.88)	(1.74)
FF5 alpha	0.40	0.46	0.65	0.62	0.49	0.88	0.88	0.89	0.73	0.87	0.47
FF5 t-stat	(1.18)	(1.42)	(1.81)	(1.90)	(1.32)	(2.62)	(2.56)	(2.49)	(1.73)	(2.27)	(1.97)

The table reports monthly average idiosyncratic tail risk, excess returns, and alphas for portfolios sorted on idiosyncratic tail risk between January 2001 to October 2016. Idiosyncratic tail risk is estimated from the high-frequency cross-Sectional variable model residuals. Panel A reports equally-weighted excess returns and alphas sorted on idiosyncratic tail risk. Panel B reports value-weighted excess returns and alphas sorted on idiosyncratic tail risk.

Table C.5: Portfolios Sorted on Idiosyncratic Tail Risk estimated using Midquotes

	1 (Low)	2	3	4	5	6	7	8	9	10 (High)	(10-1)
Panel A: Univariate Sort on Idiosyncratic Tail Risk (Equal-Weighted)											
Excess Return	0.52	0.48	0.59	0.37	0.58	0.56	0.61	0.79	0.83	0.96	0.44
t-stat	(1.13)	(1.09)	(1.37)	(0.80)	(1.31)	(1.29)	(1.38)	(1.82)	(1.88)	(2.35)	(3.05)
FFC4 alpha	0.53	0.54	0.67	0.39	0.63	0.61	0.64	0.83	0.83	0.97	0.43
FFC4 t-stat	(1.32)	(1.26)	(1.71)	(0.93)	(1.51)	(1.49)	(1.60)	(2.07)	(2.06)	(2.57)	(3.13)
FF5 alpha	0.67	0.71	0.84	0.52	0.77	0.77	0.74	0.91	0.88	1.03	0.36
FF5 t-stat	(1.60)	(1.54)	(2.11)	(1.17)	(1.75)	(1.85)	(1.72)	(2.18)	(2.10)	(2.62)	(2.67)
Panel B: Univariate Sort on Idiosyncratic Tail Risk (Value-Weighted)											
Excess Return	0.25	0.13	0.50	0.15	0.28	0.54	0.32	0.61	0.59	0.97	0.72
t-stat	(0.61)	(0.34)	(1.49)	(0.38)	(0.75)	(1.41)	(0.76)	(1.52)	(1.46)	(2.65)	(3.10)
FFC4 alpha	0.31	0.14	0.58	0.18	0.32	0.57	0.30	0.63	0.60	1.00	0.68
FFC4 t-stat	(0.79)	(0.39)	(1.81)	(0.48)	(0.96)	(1.57)	(0.89)	(1.76)	(1.70)	(2.86)	(2.81)
FF5 alpha	0.49	0.27	0.66	0.23	0.42	0.72	0.47	0.67	0.71	1.11	0.62
FF5 t-stat	(1.21)	(0.74)	(1.95)	(0.62)	(1.18)	(1.90)	(1.34)	(1.85)	(1.96)	(3.13)	(2.75)

The table reports monthly average idiosyncratic tail risk, excess returns, and alphas for portfolios sorted on idiosyncratic tail risk between January 2001 to October 2016. Idiosyncratic tail risk is estimated using midquote data and the high-frequency Fama-French five factor model residuals. Panel A reports equally-weighted excess returns and alphas sorted on idiosyncratic tail risk. Panel B reports value-weighted excess returns and alphas sorted on idiosyncratic tail risk.

Table C.6: Portfolios Sorted on Idiosyncratic Tail Risk estimated using 0.025 quantile

	1 (Low)	2	3	4	5	6	7	8	9	10 (High)	(10-1)
Panel A: Univariate Sort on Idiosyncratic Tail Risk (Equal-Weighted)											
Excess Return	0.46	0.62	0.58	0.52	0.68	0.75	0.80	0.80	0.85	0.90	0.45
t-stat	(1.02)	(1.40)	(1.31)	(1.15)	(1.59)	(1.75)	(1.94)	(1.94)	(2.08)	(2.31)	(2.77)
FFC4 alpha	0.50	0.66	0.63	0.56	0.72	0.80	0.84	0.85	0.90	0.93	0.43
FFC4 t-stat	(1.23)	(1.61)	(1.53)	(1.37)	(1.82)	(1.98)	(2.17)	(2.22)	(2.39)	(2.52)	(2.66)
FF5 alpha	0.66	0.79	0.76	0.69	0.85	0.91	0.97	0.96	1.00	0.99	0.33
FF5 t-stat	(1.52)	(1.81)	(1.75)	(1.61)	(2.04)	(2.12)	(2.38)	(2.42)	(2.58)	(2.57)	(1.99)
Panel B: Univariate Sort on Idiosyncratic Tail Risk (Value-Weighted)											
Excess Return	0.23	0.32	0.49	0.54	0.39	0.49	0.54	0.56	0.56	0.72	0.49
t-stat	(0.62)	(0.94)	(1.31)	(1.55)	(1.04)	(1.33)	(1.65)	(1.54)	(1.38)	(1.92)	(2.71)
FFC4 alpha	0.26	0.36	0.54	0.57	0.40	0.53	0.60	0.60	0.60	0.80	0.54
FFC4 t-stat	(0.77)	(1.10)	(1.55)	(1.84)	(1.25)	(1.60)	(2.00)	(1.83)	(1.67)	(2.37)	(2.72)
FF5 alpha	0.36	0.44	0.65	0.68	0.46	0.61	0.68	0.75	0.73	0.83	0.46
FF5 t-stat	(1.00)	(1.29)	(1.80)	(2.12)	(1.45)	(1.80)	(2.23)	(2.31)	(2.03)	(2.36)	(2.32)

The table reports monthly average idiosyncratic tail risk, excess returns, and alphas for portfolios sorted on idiosyncratic tail risk between January 2001 to December 2016. Idiosyncratic tail risk is estimated using a Hill threshold of 0.025 and the high-frequency market model residuals. Panel A reports equally-weighted excess returns and alphas sorted on idiosyncratic tail risk. Panel B reports value-weighted excess returns and alphas sorted on idiosyncratic tail risk.

Table C.7: Portfolios Sorted on Idiosyncratic Tail Risk using 0.01 quantile

	1 (Low)	2	3	4	5	6	7	8	9	10 (High)	(10-1)
Panel A: Univariate Sort on Idiosyncratic Tail Risk (Equal-Weighted)											
Excess Return	0.45	0.55	0.63	0.63	0.70	0.62	0.68	0.79	0.89	1.01	0.56
t-stat	(1.05)	(1.26)	(1.43)	(1.42)	(1.58)	(1.45)	(1.54)	(1.85)	(2.08)	(2.73)	(2.78)
FFC4 alpha	0.51	0.60	0.67	0.68	0.76	0.67	0.73	0.84	0.91	1.02	0.51
FFC4 t-stat	(1.29)	(1.50)	(1.62)	(1.70)	(1.87)	(1.68)	(1.78)	(2.05)	(2.28)	(2.89)	(2.58)
FF5 alpha	0.68	0.74	0.81	0.83	0.87	0.81	0.83	0.93	1.00	1.08	0.40
FF5 t-stat	(1.61)	(1.77)	(1.85)	(1.94)	(2.05)	(1.95)	(1.93)	(2.15)	(2.38)	(2.97)	(1.95)
Panel B: Univariate Sort on Idiosyncratic Tail Risk (Value-Weighted)											
Excess Return	0.18	0.59	0.47	0.60	0.49	0.71	0.70	0.74	0.72	0.86	0.68
t-stat	(0.55)	(1.74)	(1.27)	(1.62)	(1.31)	(1.84)	(1.86)	(1.83)	(1.66)	(2.54)	(3.22)
FFC4 alpha	0.23	0.64	0.47	0.65	0.53	0.75	0.74	0.79	0.72	0.86	0.63
FFC4 t-stat	(0.76)	(2.01)	(1.43)	(1.99)	(1.67)	(2.26)	(2.16)	(2.09)	(1.82)	(2.72)	(2.88)
FF5 alpha	0.33	0.74	0.58	0.76	0.63	0.88	0.87	0.89	0.80	0.93	0.61
FF5 t-stat	(1.06)	(2.31)	(1.69)	(2.27)	(1.92)	(2.64)	(2.48)	(2.24)	(2.05)	(2.88)	(2.79)

The table reports monthly average idiosyncratic tail risk, excess returns, and alphas for portfolios sorted on idiosyncratic tail risk between January 2001 to December 2016. Idiosyncratic tail risk is estimated using a Hill threshold of 0.1 and the high-frequency market model residuals. Panel A reports equally-weighted excess returns and alphas sorted on idiosyncratic tail risk. Panel B reports value-weighted excess returns and alphas sorted on idiosyncratic tail risk.

C.4.2 Conditional Portfolio Sorts on Additional Firm Characteristics

Table C.8: Double-Sorted Portfolios on Firm Characteristic then Idiosyncratic Tail Risk

Firm Characteristic	BETA	SIZE	BM	OP	INV	MOM	COSKEW	IVOL	ILLIQ	MAX	$\beta^{Downside}$	REV
Panel A: Double-Sort on Firm Characteristic then Idiosyncratic Tail Risk (Equal-Weighted)												
1 (Low ITR)	0.53	0.54	0.53	0.63	0.54	0.53	0.52	0.49	0.60	0.50	0.52	0.50
2	0.70	0.71	0.71	0.70	0.68	0.68	0.67	0.65	0.68	0.67	0.68	0.68
3	0.72	0.65	0.72	0.76	0.71	0.66	0.70	0.72	0.66	0.67	0.70	0.67
4	0.79	0.83	0.79	0.81	0.80	0.76	0.79	0.76	0.81	0.79	0.81	0.78
5 (High ITR)	0.85	0.85	0.87	1.06	0.91	0.94	0.92	0.97	0.84	0.96	0.89	0.94
Excess Return	0.32	0.31	0.34	0.44	0.38	0.41	0.40	0.48	0.24	0.45	0.37	0.44
t-stat	(2.68)	(2.25)	(2.38)	(2.92)	(2.82)	(2.77)	(2.75)	(2.73)	(2.16)	(2.91)	(3.03)	(2.97)
Panel B: Double-Sort on Firm Characteristic then Idiosyncratic Tail Risk (Value-Weighted)												
	BETA	SIZE	BM	OP	INV	MOM	COSKEW	IVOL	ILLIQ	MAX	$\beta^{Downside}$	REV
1 (Low ITR)	0.36	0.54	0.45	0.38	0.33	0.35	0.40	0.29	0.56	0.29	0.39	0.36
2	0.44	0.65	0.58	0.48	0.50	0.50	0.44	0.36	0.64	0.42	0.45	0.45
3	0.45	0.65	0.57	0.50	0.47	0.42	0.50	0.38	0.63	0.32	0.50	0.48
4	0.58	0.77	0.68	0.61	0.57	0.48	0.56	0.57	0.72	0.51	0.59	0.46
5 (High ITR)	0.67	0.82	0.74	0.92	0.75	0.69	0.79	0.74	0.75	0.69	0.71	0.76
Excess Return	0.30	0.28	0.29	0.53	0.42	0.34	0.39	0.45	0.18	0.40	0.32	0.39
t-stat	(2.59)	(2.03)	(2.06)	(3.24)	(3.13)	(2.33)	(2.68)	(2.25)	(1.86)	(2.51)	(2.40)	(2.42)

Double-Sorted equal-weighted and value-weighted portfolios are formed every month from January 2001 to December 2016. Portfolios are created by sorting by idiosyncratic tail risk after first controlling for beta, size, book-to-market, operating profitability, investment, momentum, coskewness, idiosyncratic volatility, max, downside beta, or reversal. In each case, I first sort stocks in quintiles using the control variable, then in each quintile, I sort the stocks into quintiles based on idiosyncratic tail risk. The table presents average returns in percentage terms across the 5 control quintiles to produce quintile portfolios with different idiosyncratic tail risks but similar levels of the control variable. The table also reports the average return and Newey-West t-statistic with 1 lag of a portfolio that goes long the highest idiosyncratic tail risk and shorts the lowest idiosyncratic tail risk for the returns averaged across the 5 control quintiles.

C.4.3 Descriptive Statistics

Table C.9: Descriptive Statistics of Portfolios Sorted by Idiosyncratic Tail Risk and Unconditional Correlations

	ITR	BETA	SIZE	BM	OP	INV	MOM	IVOL	COSKEW	ILLIQ	MAX	$\beta^{Downside}$	REV
Panel A: Unconditional Correlations													
ITR	1	-0.18	-0.09	0.08	0.00	0.01	0.02	0.14	0.01	0.25	0.07	-0.16	0.04
BETA		1	-0.07	-0.03	-0.03	-0.01	0.01	0.19	0.17	-0.04	0.21	0.80	0.00
SIZE			1	-0.06	0.01	0.00	-0.03	-0.14	0.02	-0.05	-0.10	-0.05	-0.01
BM				1	-0.06	0.00	0.15	-0.05	-0.03	0.04	-0.02	-0.01	0.05
OP					1	-0.01	0.00	-0.02	0.00	-0.01	-0.02	-0.02	0.00
INV						1	0.00	0.00	0.00	0.00	0.00	-0.01	0.00
MOM							1	0.08	-0.05	0.00	0.03	0.06	0.00
IVOL								1	-0.04	0.09	0.83	0.18	-0.01
COSKEW									1	0.01	0.01	-0.23	-0.02
ILLIQ										1	0.07	-0.05	0.00
MAX											1	0.18	-0.04
$\beta^{Downside}$												1	0.01
REV													1
Panel B: Descriptive Statistics for Stocks Sorted by Idiosyncratic Tail Risk													
1 (Low)	0.32	1.22	9.14	0.58	0.32	0.20	0.18	1.44	0.00	0.06	0.05	1.18	0.01
2	0.36	1.19	9.07	0.57	0.33	0.18	0.19	1.50	0.00	0.06	0.05	1.17	0.02
3	0.38	1.18	9.06	0.57	0.37	0.18	0.19	1.55	0.00	0.07	0.05	1.15	0.02
4	0.39	1.16	9.01	0.57	0.38	0.17	0.20	1.59	0.00	0.07	0.05	1.14	0.02
5	0.41	1.14	8.96	0.56	0.35	0.17	0.20	1.64	0.00	0.08	0.05	1.13	0.02
6	0.43	1.13	8.95	0.56	0.33	0.17	0.21	1.68	0.00	0.09	0.05	1.12	0.02
7	0.44	1.11	8.90	0.57	0.35	0.16	0.22	1.74	0.00	0.10	0.05	1.10	0.02
8	0.47	1.09	8.77	0.57	0.35	0.16	0.22	1.83	0.00	0.14	0.05	1.07	0.02
9	0.51	1.05	8.48	0.60	0.37	0.98	0.22	1.95	0.00	0.23	0.05	1.03	0.02
10 (High)	0.65	0.95	7.35	0.70	0.30	1.79	0.21	1.98	0.01	0.65	0.06	0.91	0.02

Panel A reports the unconditional correlations between all the variables from January 2001 to December 2016. Panel B reports the average value of each control variable for decile portfolios sorted on idiosyncratic tail risk. Control variables include beta, size, book-to-market, operating profitability, investment, momentum, coskewness, idiosyncratic volatility, max, downside beta, and reversal.

Table C.10: Asset Pricing Tests on Anomaly Portfolios using Traded VTR

	Panel A: Tail Risk Deciles		Panel B: Tail Risk ME Portfolios		Panel C: Characteristic ME Portfolios				
	ITR Dec.	VTR Dec.	25 ITR	25 VTR	25 OP	25 INV	25 MOM	25 IVOL	ALL
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Intercept	1.28	1.81	0.58	0.25	1.20	1.55	0.64	0.97	0.79
NW t-stat	(1.50)	(2.10)	(1.73)	(0.73)	(2.94)	(3.52)	(1.64)	(3.38)	(2.71)
Shanken t-stat	[1.25]	[1.84]	[1.24]	[0.57]	[2.85]	[3.62]	[1.44]	[2.85]	[2.26]
$R^M - R^f$	-0.87	-1.38	-0.12	0.24	-0.68	-1.05	-0.08	-0.49	-0.28
NW t-stat	(-0.96)	(-1.42)	(-0.25)	(0.54)	(-1.14)	(-1.75)	(-0.14)	(-0.97)	(-0.56)
Shanken t-stat	[-0.80]	[-1.30]	[-0.21]	[0.45]	[-1.29]	[-1.99]	[-0.15]	[-1.06]	[-0.59]
Traded VTR	0.89	0.69	0.89	0.62	0.86	1.00	0.70	0.96	0.79
NW t-stat	(2.56)	(3.82)	(3.30)	(2.44)	(2.68)	(3.27)	(2.55)	(3.60)	(3.17)
Shanken t-stat	[2.31]	[3.55]	[2.79]	[2.30]	[2.60]	[3.18]	[2.32]	[3.34]	[2.95]
% R^2	76.53	93.05	84.71	79.85	64.48	89.33	71.19	80.91	67.85
% MAE	0.07	0.04	0.07	0.07	0.11	0.08	0.11	0.09	0.11
Months	192	192	192	192	192	192	192	192	192

This table presents asset pricing tests on double-sorted portfolios using the traded VTR and market two-factor model from 2001 to 2016. In Panel A, the test assets are the decile portfolios sorted on idiosyncratic tail risk or volume tail risk examined in Table 4.7. In Panel B, the test assets are 25 portfolios conditionally sorted on size and idiosyncratic tail risk or volume tail risk. Stocks are first grouped into size quintiles, then within each size quintile, stocks are grouped by their ITR or VTR. In Panel C, test assets are 25 portfolios independently sorted by size and the characteristic. Stocks are grouped by the intersection of 5 quintiles sorted on size and 5 quintiles sorted on the characteristic. These anomaly portfolios are downloaded from from Kenneth French's website and include operating profitability, investment, momentum, reversal, and idiosyncratic volatility. The table reports the risk premia estimates, R^2, and mean absolute pricing errors in percentage terms, Newey-West t-statistics with one lag, and Shanken t-statistics.

Table C.11: Tests on HKM Portfolios using Traded VTR

	FF		Bond		Sov		Options		CDS		Conmod		FX		All	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
Intercept	0.63	1.25	0.26	0.28	0.84	0.34	0.90	0.78	0.02	-0.11	0.24	0.10	0.22	-0.34	0.34	0.14
NW t-stat	1.12	1.92	4.16	4.91	2.46	1.76	1.07	0.94	0.41	-4.99	0.59	0.34	0.95	-1.29	1.34	1.36
Shanken t-stat	1.27	1.83	2.72	5.20	1.94	1.54	1.01	0.74	0.27	-4.79	0.62	0.30	0.73	-0.78	1.59	1.37
$R^M - R^f$		-1.01		1.17		2.04		-0.89		2.51		-0.07		0.98		-0.08
NW t-stat		-1.20		0.99		2.07		-0.86		3.45		-0.08		0.72		-0.17
Shanken t-stat		-1.30		1.16		1.85		-0.76		3.57		-0.09		0.49		-0.18
CITR	1.02	1.27	3.75	-0.17	-3.21	-1.57	2.70	2.59	5.10	-0.77	0.94	0.95	4.51	4.26	1.12	1.21
NW t-stat	2.49	2.51	1.22	-0.10	-2.06	-1.29	2.00	2.18	2.57	-0.99	1.63	1.62	3.64	3.35	3.05	3.25
Shanken t-stat	2.28	2.65	1.13	-0.12	-1.37	-1.11	2.15	1.99	1.48	-0.78	1.34	1.34	2.82	2.53	2.64	2.92
% Adj. R^2	58.06	63.23	34.96	58.98	30.18	75.30	97.19	97.19	14.91	66.95	19.70	20.96	14.91	33.22	28.93	43.37
% MAE	0.12	0.12	0.11	0.10	0.26	0.14	0.07	0.07	0.09	0.07	0.50	0.49	0.26	0.23	0.31	0.24
Months	144	144	132	132	124	124	133	133	143	143	144	144	109	109	144	144

Test assets are the portfolios in He, Kelly, and Manela (2017) downloaded from Asaf Manela's website. Assets include equities, US government and corporate bonds, sovereign bonds, options, credit default swaps, commodities, and foreign exchange. The Fama MacBeth analysis is from January 2001 to December 2012. The model uses the market portfolio and traded CITR as the factors. Traded ITR is the value-weighted return on a portfolio that goes long the highest ITR decile and shorts the lowest ITR decile. The table reports the risk premia estimates, R^2, and mean absolute pricing errors in percentage terms, Newey-West t-statistics with one lag, and Shanken t-statistics.

C.4.4 Traded Volume Tail Risk Pricing

Curriculum Vitae

Education

- Ph.D. Economics, University of Western Ontario 2021
- M.A. Economics, University of Western Ontario 2015
- B.A. Economics, University of Waterloo 2014
- B.A. Business Administration, Wilfrid Laurier University 2012

Publications:

- Regulatory Capital and Incentives for Risk Model Choice under Basel 3 (with Lars Stentoft), *Journal of Financial Econometrics*, 2021.
- Intraday Market Predictability: A Machine Learning Approach (with Dillon Huddleston and Lars Stentoft), *Journal of Financial Econometrics*, forthcoming.

Honors, Awards, and Fellowships

- Social Sciences & Humanities Research Council Fellowship, Canada 2018-2020
- Ontario Graduate Scholarship, Ontario, Canada 2016, 2017
- Economics Graduate Scholarship, University of Western Ontario 2014-2019
- Tutorial Teaching Assistant of the Year, University of Western Ontario 2015

- Award for Distinguished Academic Achievement, University of Waterloo 2014
- Economics Achievement Award, University of Waterloo 2014

Employment

- Research Assistant for Lars Stentoft, Department of Economics, UWO 2017 - 2021
- Research Assistant for Davin Raiha, Ivey School of Business 2015 - 2017
- Tutorial Instructor, Department of Economics, UWO 2014 - 2016
- Teaching Assistant, Department of Economics, UWO 2016 - 2017