CHARLES UNIVERSITY
FACULTY OF SOCIAL SCIENCES
Institute of Economic Studies

Asset Pricing and Portfolio Selection via Machine Learning

Dissertation thesis

Author: Mgr. Martin Hronec

Study program: Economics and Finance

Supervisor: doc. PhDr. Jozef Baruník, Ph.D.

Year of defense: 2025

Declaration of Authorship

The author hereby declares that he compiled this thesis independently, using only the listed resources and literature, and the thesis has not been used to obtain any other academic title.

Prague, Saturday 8th March, 2025

Mgr. Martin Hronec

Abstract

This thesis consists of three papers investigating stock returns forecasting, empirical asset pricing, and portfolio selection through the application of machine learning and quantitative methods. The first paper examines out-of-sample returns on 153 documented anomalies in equities, demonstrating that machine learning techniques, which aggregate these anomalies into a single mispricing signal, are profitable globally and effective even within a liquid universe of stocks. Notably, while past performance of quantitative strategies outside the U.S. does not reliably predict out-of-sample success within the U.S., U.S.-based evidence is broadly predictive of returns in international markets. The second paper introduces a novel method for predicting the full distribution of stock returns using a comprehensive set of 194 stock characteristics and market variables. This approach, free from restrictive model assumptions, leverages a two-stage quantile neural network with spline interpolation to explore non-Gaussian, heavy-tailed data and their non-linear interactions. The results indicate that this method outperforms traditional models in terms of out-of-sample losses and provides valuable alternative empirical estimates for mean estimation and forecasting across a wide range of U.S. and international data. The third paper investigates the out-of-sample performance of portfolios constructed by investors with quantile preferences, which focus on maximizing the τ-quantile of future portfolio returns rather than relying on expected utility. The study reveals significant differences in portfolio diversification and turnover based on varying τ-quantiles and shows that for risk-averse investors, a global minimum variance portfolio often yields a higher out-of-sample τ-quantile of returns compared to direct quantile maximization approaches. Through Monte Carlo simulations, the research attributes the underperformance of τ-quantile maximizers to small sample bias, which leads to deviations from optimal portfolio weights.

JEL Classification	C45, C53, C55, G11, G12, G15, G17, F30
Keywords	Cross-section of stock returns, Anomalies, Quantile forecasting, Machine learning, Neural networks, International markets, Portfolio selection, Distribution forecasting, Quantile preferences
Title	Asset Pricing and Portfolio Selection via Machine Learning

Abstrakt

Tato disertační práce se skládá ze tří článků, které zkoumají předpovídání výnosů akcií, empirické oceňování aktiv a výběr portfolia prostřednictvím aplikace strojového učení a kvantitativních metod. První článek zkoumá výnosy mimo vzorek na 153 dokumentovaných anomáliích v akciích a ukazuje, že techniky strojového učení, které agregují tyto anomálie do jednoho signálu nesprávného ocenění, jsou ziskové globálně a účinné i v rámci likvidního vesmíru akcií. Zejména zatímco minulá výkonnost kvantitativních strategií mimo USA nespolehlivě předpovídá úspěšnost mimo vzorek v USA, důkazy z USA jsou obecně prediktivní pro výnosy na mezinárodních trzích.

Druhý článek představuje nový způsob předpovídání celé distribuce výnosů akcií za použití komplexního souboru 194 charakteristik akcií a tržních proměnných. Tento přístup, osvobozený od omezujících modelových předpokladů, využívá dvoustupňovou kvantilovou neuronovou síť s interpolací pomocí spline k prozkoumání negaussovských, těžkých ocasů dat a jejich nelineárních interakcí. Výsledky naznačují, že tato metoda překonává tradiční modely z hlediska ztrát mimo vzorek a poskytuje cenné alternativní empirické odhady pro odhad a předpovídání průměru ve velkém spektru dat z USA a mezinárodních trhů.

Třetí článek zkoumá výkonnost portfolií mimo vzorek vytvořených investory s kvantilovými preferencemi, které se zaměřují na maximalizaci τ-kvantilu budoucích výnosů portfolia namísto spoléhání se na očekávaný užitek. Studie odhaluje významné rozdíly v diverzifikaci portfolií a obrátkovosti na základě různých τ-kvantilů a ukazuje, že pro rizikově averzní investory často konstrukce globálního portfolia s minimální variancí přináší vyšší mimo vzorek τ-kvantil výnosů než přímé kvantilové maximalizační přístupy. Prostřednictvím simulací Monte Carlo výzkum přisuzuje nedostatečnou výkonnost τ-kvantilových maximalizátorů zkreslení malých vzorků, které vede k odchylkám od optimálních vah portfolia.

Klasifikace JEL	C45, C53, C55, G11, G12, G15, G17, F30
Klíčova slova	Průřezová data výnosů akcií, Anomálie, Kvantilové předpovídání, Strojové učení, Neuronové sítě, Mezinárodní trhy, Výběr portfolia, Předpovídání distribuce, Kvantilové preference
Název práce	Oceňování aktiv a výběr portfolia pomocí strojového učení

Acknowledgments

I would like to express my sincere gratitude to my supervisor, doc. PhDr. Jozef Baruník, Ph.D., for his guidance, support, patience and drive for high-quality research. I've learned a lot from him, and I appreciate the opportunity to work with him.

I am very grateful to my coauthor and dear friend Ondřej Tobek PhD, who initiated our collaboration, was a pleasure to work with and offered help whenever I needed it.

I am also grateful to my partner, Lenka, for all her patience, support, and love that allowed me to deliver this thesis without going insane.

I also wish to thank my family, who always put my education first and supported me in all my endeavours.

Throughout my studies, I participated in several conferences and seminars. I would like to thank all my fellow researchers who discussed my work and provided valuable feedback. I also deeply appreciate my close colleagues from our research group for all the discussions and suggestions. Let me also acknowledge the support from the Charles University Grant Agency (GAUK) projects No. 910680 and 316521, the Czech Science Foundation under EXPRO GX19-28231X and GA16-14151S projects, and the UNCE Fellowship. All errors are solely the author's responsibility.

Typeset in LaTeXusing the IES Thesis Template.

Bibliographic Record

Hronec, Martin: *Asset Pricing and Portfolio Selection via Machine Learning.* Dissertation thesis. Charles University, Faculty of Social Sciences, Institute of Economic Studies, Prague. 2025, pages 226. Advisor: doc. PhDr. Jozef Baruník, Ph.D.

Contents

List of Tables **ix**

List of Figures **xi**

Introduction **1**

References 5

1 Does it pay to follow anomalies research? **6**

1.1 Data and methodology 11

1.1.1 Data 11

1.1.2 Anomalies 13

1.1.3 Mispricing strategy 14

1.1.4 Portfolio construction 20

1.1.5 Liquidity measures 21

1.2 Profitability of the mispricing strategy estimated in the U.S. . . 21

1.2.1 Long-only and short-only components of the strategy . . 23

1.2.2 Risk-adjusted performance of the strategy 24

1.3 The role of international evidence 24

1.3.1 Locally-trained mispricing strategy 25

1.3.2 Globally-trained mispricing strategy 26

1.3.3 Variable importance 27

1.3.4 Marginal predictive contributions of signals 28

1.4 Transaction costs 29

1.4.1 Transaction costs on the strategy 29

1.4.2 Performance of the strategy after transaction costs . . . 30

1.5 Conclusion 31

References 54

Appendix

1.A Adjustments of returns in datastream 55

1.B List of the anomalies . . . 56
1.C Sample splitting and pool of methods . . . 60
1.D Optimal hyperparameters . . . 62
1.E Definition of liquidity proxies . . . 67

2 Predicting the distributions of stock returns around the globe 69
2.1 Introduction . . . 70
2.1.1 Relation of our work to the literature . . . 74
2.2 Two-stage quantile neural networks . . . 78
2.2.1 Network architecture . . . 79
2.2.2 Training with quantile loss function . . . 82
2.2.3 Approximating probability density and its moments from quantiles using B-Splines . . . 86
2.3 Empirical results . . . 88
2.3.1 Data and features . . . 88
2.3.2 Description of the estimated moments . . . 91
2.3.3 Predictability . . . 93
2.3.4 Distribution and the cross-section of stock returns . . . 96
2.4 Conclusion . . . 102
References . . . 111
Appendix
2.A List of the anomalies . . . 112
2.B Alternative models . . . 116
2.B.1 Feed-forward neural networks . . . 116
2.B.2 GARCH benchmark . . . 118
2.C Additional empirical results . . . 121
2.D Hyperparameter search and validation . . . 125
2.D.1 Hyperparameter search . . . 125
2.D.2 Validation . . . 129
2.E Algorithms for approximating probability distribution function and moments from quantiles . . . 130
2.E.1 Fallback to linear B-Splines . . . 130
2.E.2 Central moments . . . 131
2.E.3 Adjustment of moments . . . 133
2.E.4 Two-stage quantile neural network pseudo-algorithm . . . 135
2.E.5 Addressing quantile crossing problem via penalization . . . 137
2.E.6 2019-2023 out-of-sample results in the U.S. . . . 138

3 Portfolio selection under quantile preferences 143
3.1 Introduction . . . 144
3.2 Quantile preferences . . . 148
3.3 Portfolio optimization via quantile maximization . . . 150
3.3.1 Quantile maximization via mixed integer linear programming . . . 155
3.4 Simulation . . . 156
3.5 Empirical results . . . 162
3.5.1 Data . . . 162
3.5.2 Weights . . . 164
3.5.3 Out-of-sample evaluation . . . 165
3.6 Conclusion . . . 169
References . . . 172
Appendix
3.A Evaluation . . . 173
3.B Global minimum variance portfolio . . . 174
3.C Additional simulation results . . . 175
3.D Additional empirical results . . . 177
3.E Binomial tests for number of outperformance cases . . . 178

4 Conclusion 187

5 Responses to opponents' reports 189
5.1 Dr. Maria Grith PhD . . . 189
5.1.1 Chapter 2 . . . 189
5.1.2 Chapter 3 . . . 194
5.2 Prof. Jan Hanousek Ph.D. . . . 195
5.2.1 Chapter 1 . . . 195
5.2.2 Chapter 2 . . . 198
5.2.3 Chapter 3 . . . 201
5.3 Dr Rodrigo Hizmeri PhD . . . 202
5.3.1 Introduction . . . 202
5.3.2 Chapter 1 . . . 203
5.3.3 Chapter 2 . . . 203
5.3.4 Chapter 3 . . . 210
References . . . 214

List of Tables

1.1 Number of stocks in cross-section 32
1.2 Performance of the mispricing strategy estimated in the U.S. . . 33
1.3 Performance of the mispricing strategy estimated on the international data 37
1.4 Spearman's correlation matrices for region-specific variable importance 39
1.5 Performance of the strategies after transaction costs 40
1.6 List of anomalies 56

2.1 Number of stocks in the cross-section across the regions 89
2.2 Summary statistics for the moments in the U.S. 91
2.3 Average cross-sectional correlation between variables in the U.S. 92
2.4 Out-of-sample quantile forecasts evaluation 94
2.5 Mean forecast evaluation 95
2.6 Volatility forecast evaluation 96
2.7 Quantiles and the cross-section of stock returns in the U.S. . . . 97
2.8 Mean and median portfolios 98
2.9 Moments decile portfolios 100
2.10 List of anomalies 112
2.11 Performance of various GARCH specifications 119
2.12 Average cross-sectional correlation between moments and median internationally 122
2.13 Quantiles and long-short portfolios internationally 123
2.14 Moments value-weighted decile portfolios 124
2.15 Fixed hyperparameters 126
2.16 Hyperparameter search for learning rate and dropout 127
2.17 Hyperparameter search for architecture 128
2.18 Adjustment fit of density-to-moments algorithm 134

2.19 Mean forecast evaluation in the U.S. 2019-2013 139
2.20 Volatility forecast evaluation in the U.S. 2019-2023 140
2.21 Out-of-sample quantile forecasts evaluation in the U.S. 2019-2023 141
2.22 Mean and median portfolios in the U.S. 2019-2023 142

3.1 Herfindahl-Hirschman index and turnover 157
3.2 Out-of-sample performance of the simulated portfolios 158
3.3 Datasets used in the empirical results 163
3.4 Herfindahl-Hirschman index and turnover 167
3.5 Out-of-sample 10%, 50% and 90% quantiles 168
3.6 Herfindahl-Hirschman index and turnover 176
3.7 Out-of-sample performance of the simulated portfolios 176
3.8 Out-of-sample portfolio risk metrics for 60 months long estimation window 179
3.9 Herfindahl-Hirschman index and turnover of portfolio weights for 36 months long estimation window 180
3.10 Out-of-sample 10%, 50% and 90% quantiles of portfolio returns for 36 months long estimation window 181
3.11 Out-of-sample portfolio risk metrics for 36 months long estimation window 182
3.12 Herfindahl-Hirschman index and turnover of portfolio weights for 63 days long estimation window 183
3.13 Out-of-sample 10%, 50% and 90% quantiles of portfolio returns for 63 days long estimation window 184
3.14 Out-of-sample portfolio risk metrics for 63 days long estimation window 185
3.15 Binomial tests of outperformance for 60 months long estimation window 186
3.16 Binomial tests of outperformance for 36 months long estimation window 186
3.17 Binomial tests of outperformance for 63 days long estimation window 186

5.1 Comparison of single-output vs. multi-output 2sNN models . . . 193
5.2 Double sorting on MSE mean and quantile mean 209

List of Figures

1.1 Number of the published anomalies over time 41
1.2 Decision tree 41
1.3 Neural network 41
1.4 Cumulative returns on the mispricing strategy using neural networks 42
1.5 Variable importance 43
1.6 Marginal predictive contributions of Size, BM, and Mom 44
1.7 Marginal predictive contributions of AGr, OPoE, and STR . . . 45
1.8 Monthly transaction costs 46
1.9 Validation logic - training, validation, and test split 60
1.10 L1 and L2 mixing parameter for penalized weighted least square 63
1.11 Number of trees for gradient boosting trees 64
1.12 Maximum depth of the tree of random forests 65
1.13 Dropout and the architecture for neural networks 66

2.1 Example of forecasted probability density function over time . . 72
2.2 Two-stage neural network (2sNN) architecture 81
2.3 Interpolated cumulative distribution function and probability density function 87
2.4 Histogram of central distribution moments 92
2.5 Architectures of benchmark multi-head feed-forward neural networks 117
2.6 The cross-sectional average of volatilities as a scaling factor . . . 121
2.7 Validation sample-splitting schema 129
2.8 Example of density with a large mass around zero 131

3.1 τ-quantile maximization illustration 152
3.2 Selected portfolio weights across τs 160
3.3 Portfolio weights & in-sample objective values 161
3.4 Evolution of portfolio weights over time 177

Introduction

The study of asset pricing and portfolio selection has long been central to financial economics, seeking to understand how assets are valued and how investors should allocate their wealth amidst uncertainty. Both topics share the common need to understand and predict asset returns. In asset pricing, measuring an asset's risk premium is inherently a prediction problem, as the risk premium represents the conditional expectation of future returns. In portfolio selection, we must balance the trade-off between risk and expected returns. However, forecasting asset returns, particularly stock returns, is extremely challenging due to the noisy nature of financial markets, driven by the dynamic forces within them. The presence of heavy-tailed distributions, non-linearities, a large number of potential predictor variables, and the lack of clear guidance on the functional form of the relationship between predictors and returns all contribute to the difficulty of this forecasting problem and challenge the assumptions underlying many traditional decision-making frameworks.

In this thesis, we leverage the power of big data, machine learning, and quantitative methods applied to large financial datasets, including international stock returns, to address these challenges. The first two chapters utilize machine learning methods to forecast stock returns globally, drawing on a broad set of predictors. In the first chapter, we demonstrate significant profitability gains in long-short portfolios formed using machine learning return predictions and examine the added value of incorporating data from outside the U.S. In the second chapter, we apply machine learning methods to accurately forecast the full distribution of out-of-sample stock returns, both in the U.S. and internationally. The third chapter shifts focus to portfolio selection, where we explore how an investor with quantile preferences can form portfolios, providing novel insights into their out-of-sample performance.

Throughout this thesis, I use the first-person plural pronoun "we" since all papers are the result of collaboration. The first chapter was co-authored with Ondřej Tobek, the second chapter with Ondřej Tobek and my supervisor Jozef

Baruník, and the third chapter was co-authored solely with Jozef Baruník. My contribution to all chapters was substantial across all stages of the research.

In Chapter 1 - *Does it Pay to Follow Anomalies Research? A Machine Learning Approach with International Evidence* we investigate the profitability of stock return anomalies and the benefits of incorporating international data into predictive models. While much of the literature focuses on U.S. markets, this chapter highlights the importance of international evidence in understanding the true robustness of asset pricing anomalies. We focus on a comprehensive set of anomalies to predict stock returns using a range of machine-learning techniques, including random forests, gradient boosting, and neural networks. We demonstrate that the out-of-sample profitability of a long-short strategy based on machine learning return predictions is significantly higher than that of traditional linear models. This outperformance is consistent across global markets, reinforcing the findings of Gu et al. (2020) that machine learning methods enhance return predictability. Our analysis is based on a liquid universe of stocks, ensuring that our results are relevant to investors by minimizing the impact of microstructure noise. Moreover, we consider the economic implications of our findings by analyzing the effects of transaction costs and short-selling constraints. We also uncover that expanding the estimation sample to include data outside the U.S. does not improve predictability for U.S. investors. The regional differences in variable importance suggest that the benefits of a larger sample are offset by local market idiosyncrasies. We demonstrate that while U.S.-based models capture much of the predictability in international markets, incorporating global data provides additional insights into the nature of stock return drivers, both local and global.

In Chapter 2 - *Predicting the Distributions of Stock Returns: A Machine Learning Approach*, we address the growing challenge of accurately forecasting the full distribution of stock returns, which is crucial for various financial applications, including asset pricing, risk management, and portfolio selection. Traditional methods in finance have predominantly focused on predicting the first two moments of return distributions, mean and variance, using economic variables (Ang and Bekaert 2007; Ang, Hodrick, et al. 2006). However, these methods often rely on restrictive assumptions such as normality and linearity, which fail to capture the complexities of financial data, including non-linearities, fat tails, and volatility clustering (Cont 2001). To overcome these limitations, we leverage the power of neural networks, which are increasingly recognized

for their ability to handle high-dimensional data and capture complex, non-linear relationships (Athey and Imbens 2019; Mullainathan and Spiess 2017). Specifically, we propose a novel distribution forecasting method that combines a two-stage quantile neural network with cubic B-splines interpolation. This approach allows us to predict a comprehensive set of τ-quantiles of returns, which are then interpolated to generate the full distribution of future stock returns. Our methodology not only improves the accuracy of mean and variance forecasts but also provides deeper insights into other distributional characteristics, such as skewness and kurtosis. By applying this method to both U.S. and international stock data, we demonstrate its superior out-of-sample performance compared to traditional parametric models and other neural network approaches. Moreover, we explore how these predicted distributional characteristics are reflected in the cross-section of stock returns, revealing that central quantiles are consistently priced, while the importance of tails diminishes.

In Chapter 3 - *Portfolio Selection under Quantile Preferences: An Out-of-Sample Perspective*, we keep our focus beyond the first two moments of the distribution but move from forecasting to portfolio selection. While modern portfolio theory, pioneered by Markowitz (1952) and Roy (1952), has been instrumental in financial decision-making, it often fails to account for real-world investor behaviours such as loss aversion, preference for positive skewness, and ambiguity aversion (Barberis 2013; Gilboa and Schmeidler 1989; Kahneman and Tversky 1979; Rabin 2013). Moreover, assumptions of normality and linearity in returns are inadequate given the presence of fat-tailed distributions and volatility clustering (Cont 2001). This chapter introduces quantile preferences as a promising alternative decision-making framework, shifting the focus from expected utility to the maximization of specific quantiles of future portfolio returns (de Castro and Galvao 2019; de Castro and Galvao 2022; Chambers 2009; Manski 1988; Rostek 2010). Despite quantile preferences' theoretical promise, using quantile maximization as a necessary tool presents significant practical challenges, including computational complexity and small sample biases that can distort portfolio weight selection. Our study is the first to thoroughly examine the out-of-sample performance of portfolios formed under quantile preferences, using both empirical and simulated datasets. We find that, for risk-averse investors, a global minimum variance portfolio often outperforms a τ-quantile maximization strategy in out-of-sample settings. Furthermore, we document significant differences in portfolio compositions between different τ-quantile maximization portfolios and the global minimum variance portfolio,

as well as equal-weighted portfolios.

References

Ang, A., & Bekaert, G. (2007). Stock return predictability: Is it there? *The Review of Financial Studies*, *20*(3), 651–707.

Ang, A., Hodrick, R. J., Xing, Y., & Zhang, X. (2006). The cross-section of volatility and expected returns. *The Journal of Finance*, *61*(1), 259–299.

Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, *11*, 685–725.

Barberis, N. (2013). Thirty years of prospect theory in economics: A review and assessment. *Journal of Economic Perspectives*, *27*(1), 173–196.

Cont, R. (2001). Empirical properties of asset returns: Stylized facts and statistical issues. *Quantitative Finance*, *1*(2), 223.

de Castro, L., & Galvao, A. F. (2019). Dynamic quantile models of rational behavior. *Econometrica*, *87*(6), 1893–1939.

de Castro, L., & Galvao, A. F. (2022). Static and dynamic quantile preferences. *Economic Theory*, *73*(2), 747–779.

Gilboa, I., & Schmeidler, D. (1989). Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics*, *18*(2), 141–153.

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical Asset Pricing via Machine Learning. *The Review of Financial Studies*, *33*(5), 2223–2273.

Chambers, C. P. (2009). An axiomatization of quantiles on the domain of distribution functions. *Mathematical Finance: An International Journal of Mathematics, Statistics and Financial Economics*, *19*(2), 335–342.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 363–391.

Manski, C. F. (1988). Ordinal utility models of decision making under uncertainty. *Theory and Decision*, *25*, 79–104.

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, *7*(1), 77–91.

Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87–106.

Rabin, M. (2013). Risk aversion and expected-utility theory: A calibration theorem. In *Handbook of the fundamentals of financial decision making: Part i* (pp. 241–252). World Scientific.

Rostek, M. (2010). Quantile maximization in decision theory. *The Review of Economic Studies*, *77*(1), 339–371.

Roy, A. D. (1952). Safety first and the holding of assets. *Econometrica: Journal of the Econometric Society*, 431–449.

Chapter 1

Does it pay to follow anomalies research? Machine learning approach with international evidence

Abstract

We study out-of-sample returns on 153 anomalies in equities documented in the academic literature. We show that machine learning techniques that aggregate all the anomalies into one mispricing signal are profitable around the globe and survive on a liquid universe of stocks. We investigate the value of international evidence for the selection of quantitative strategies that outperform out-of-sample. Past performance of quantitative strategies in regions other than the United States does not help to pick out-of-sample winning strategies in the U.S. Past evidence from the U.S., however, captures most of the return predictability outside the U.S.

This chapter was co-authored with Ondřej Tobek and published in the *Journal of Financial Markets*. I gratefully acknowledge support from the Czech Science Foundation under the 19-28231X (EXPRO) project. This research was also supported by the Grant Agency of Charles University (grant no. 910680) and Charles University Research Centre program No. UNCE/HUM/035. We have benefited greatly from comments from Andrew Harvey, Mark Salmon, Söhnke Bartram, Jozef Baruník and the participants at the FMND 2019 and ECOSTA 2019 conferences.

The empirical asset pricing literature is predominantly built on evidence from the financial markets in the United States. This phenomenon is termed the "academic home bias puzzle" by Andrew Karolyi (2016), who documents that only 16% of the empirical papers published in the four leading finance journals investigate markets outside the United States. Yet essential research questions can be answered very differently when looking at the international evidence. One such question is the post-publication decline in the returns of anomalies. McLean and Pontiff (2016) document a 58% decrease in post-publication profitability relative to the in-sample profitability of portfolios based on the underlying anomalies. Jacobs and Müller (2020), however, show that the U.S. is the only country with a reliable post-publication decline in the returns of anomalies, emphasizing the importance of international evidence in asset pricing. On the other hand, in the case of predictive regression for stock returns, evidence from the U.S. and around the globe is more similar. Green et al. (2017) in the U.S., and Jacobs and Müller (2018) internationally, find that combining anomalies into one mispricing signal using the least squares approach leads to superior out-of-sample risk-adjusted returns relative to focusing on individual anomalies. The benefit of combining individual anomalies through predictive regressions is further amplified by Gu et al. (2020), who conclude that more sophisticated machine learning methods offer higher out-of-sample predictability in the U.S. compared to the traditional methods in Jacobs and Müller (2018).

In this study, we examine international evidence in machine learning-based predictive regressions for stock returns using anomalies as predictors (mispricing strategy hereafter). We offer two main contributions. Firstly, using machine learning models instead of traditional linear models leads to substantial and superior out-of-sample long/short returns on the mispricing strategy both globally as well as in all the individual regions. This profitability is documented among large-cap stocks, after accounting for transaction costs, as well as short-selling constraints, and provides strong support to the findings of Gu et al. (2020). Secondly, unexpectedly, extending the estimation sample with observations from outside the U.S. does *not* benefit the U.S. investor leveraging predictive-regression-based strategies using firm characteristics. The benefits of a larger estimation sample are offset by the region-specific differences. The differences across regions are visible in the heterogeneity of marginal importance of variables, which, however, does not translate into profitability differences in the mispricing strategy.

The mispricing strategy is based on the estimated historical relation between

the past characteristics and future returns of individual stocks in the U.S., Japan, Europe, and Asia Pacific. We use 153 anomalies documented in the literature. An anomaly describes an individual stock characteristic that was shown to predict future returns in a cross-section. These anomalies are, for example, earnings over price of Basu (1977), accruals of Sloan (1996), R&D over market equity of L. K. Chan et al. (2001), and composite equity issuance of Daniel and Titman (2006).

The historical relationships are typically linearly approximated using Fama and MacBeth (1973) least squares regressions in the literature, as in Lewellen et al. (2015), Green et al. (2017), or Jacobs and Müller (2018). The topic of this study is the closest to Jacobs and Müller (2018) and Jacobs and Müller (2020), who analyze returns on anomalies outside the U.S. Our study is, however, different in many aspects. Firstly, it focuses on a liquid universe of stocks that contains stocks with capitalization in the top 95% of the overall market's capitalization and dollar trading volume over the previous year in the top 95% of the overall market's volume in the individual regions. Only about 1000 of the most liquid stocks pass the criteria in the 2010s in a given month in the U.S. Excluding small-cap stocks leads to results more relevant to investors and limits the effect of microstructure noise.[1]

Secondly, we investigate the role of international evidence in the strategies. Jacobs and Müller (2018) and Jacobs and Müller (2020) focus solely on strategies that use data in the respective regions without evaluating the possible benefits of using global data to predict future returns.

Thirdly, the prediction methods differ among the studies. Our study is the closest in methodology and application of machine learning techniques to Gu et al. (2020), who show that machine learning methods significantly outperform the linear approximation in the U.S. We extend the use of machine learning methods from the U.S. to international markets. Specifically, we compare the baseline least squares regressions to more complex gradient-boosting regression trees, random forests, and neural networks. We find that the machine learning methods lead to significant gains in the performance of the mispricing strategy in all the regions, which provide support for the conclusions of Gu et al. (2020) in true out-of-sample fashion.

A large difference with respect to Gu et al. (2020) is that we allow only anomalies documented in the previously published studies to enter predictions in each year. That is, the information set of anomalies available to investors

[1]See Asparouhova et al. (2010) for a description of the effect of microstructure noise.

at the time they make an investment decision. Ignoring this assumption can lead to illusory profits that cannot be obtained in practice. Another difference is their focus on the full universe of stocks, which has profound effects on their conclusions. The most important anomalies in their estimations are liquidity, size, and return over the past month (short-term return reversal). Asparouhova et al. (2010) argue that these variables are connected to future returns mainly through microstructure biases and have nothing to do with the true predictability of stock returns that is of interest to investors.[2]

An alternative potential explanation for the profitability of the mispricing strategy could be the well-documented limits to arbitrage from the anomalies literature, e.g. (Avramov, Chordia, et al. 2013; Hou, Xue, and Zhang 2018; Chordia, Subrahmanyam, and Tong 2014; Stambaugh et al. 2012).

The importance of imposing economic restrictions on the results of machine learning models in asset pricing is emphasized by Avramov, Cheng, and Metzker (2023). Generative adversarial networks and recurrent neural network with long short-term memory architecture of L. Chen et al. (2024), instrumented principal component analysis of Kelly, Pruitt, and Su (2019), conditional autoencoders of Gu et al. (2021), as well as Gu et al. (2020) offer weaker returns predictability under the economic restrictions. We address these concerns primarily by focusing on the large-cap universe. Hou, Xue, and Zhang (2018) show that most of the individual anomalies are much weaker when microcaps are eliminated from the U.S. investment universe. However, we find that the profitability of the mispricing strategy is substantial, even though it is studied on the liquid universe of stocks. Additionally, we examine short-selling constraints and transaction costs.

It is often impossible to short-sell certain stocks due to an insufficient supply of borrowable shares, but Stambaugh et al. (2012) describe how the profitability of anomalies often comes precisely from the short leg of the portfolios. Therefore, we decompose the returns on the long-short portfolios into long-only and short-only components. Short-selling constraints, however, cannot fully explain the profitability as both the long-only and short-only legs of the mispricing strategy offer a profitable investment opportunity with respect to market returns.

Moreover, we estimate the transaction costs associated with machine learning-based strategies leveraging the predictive power of anomalies internationally for the first time. Novy-Marx and Velikov (2015) study transaction costs on

[2]See Roll (1984) for a simple model decomposing stock returns into microstructure noise and changes in true prices.

a range of anomalies in the U.S. and conclude that the transaction costs are important mainly for high-turnover anomalies whose returns net of transaction costs often turn negative.[3] Our mispricing strategy remains profitable across the regions even after accounting for the transaction costs. The profitability of the strategy is, therefore, not illusory and can be capitalized by the investors.

Finally, we examine the value of international evidence for the prediction of out-of-sample returns on the anomalies. Harvey, Liu, and Zhu (2016) and Hou, Xue, and Zhang (2018) show that many anomalies cannot be replicated and many others are significant only due to the in-sample data snooping. New anomalies identified in the literature are based on the same historical datasets in the U.S., which can lead to false positive discoveries. International data provide new information with respect to the U.S. and it could, therefore, limit the number of false discoveries.[4] Using international data could also simply lead to different insights compared to the results from the U.S. as is the case with the stability of return predictability of anomalies over time, see Schwert (2003), Chordia, Subrahmanyam, and Tong (2014), or McLean and Pontiff (2016) for the U.S. evidence and Lu et al. (2017) or Ilmanen et al. (2019) for international evidence.

Another reason we use international data is to increase sample size, which in turn leads to more precise estimates under certain conditions. The central limit theorem establishes that the confidence interval around a point estimate shrinks with the square root of a number of observations. In the simplest version of the central limit theorem, it is assumed that the data are identically and independently distributed. There could be a problem in that some anomalies are specific to the U.S. as they depend on the local institutional setting. For example, accruals depend on country-specific accounting rules. The institutional uniqueness then limits the value of data outside the U.S. for predictions in the U.S. as it breaks the assumption of the identical distribution. In short, the usefulness of international data depends on the structure of drivers of stock returns around the globe. If the drivers are primarily *global*, then the larger

[3]Frazzini, Israel, and Moskowitz (2012) demonstrated that real-life transaction costs for large portfolio managers are much lower than assumed by academics. The transaction costs can be further lowered by appropriately optimized portfolio rebalancing.

[4]Note that many anomalies have been individually studied in the international markets. For examples of studies investigating cross-sectional predictability of individual signals outside the U.S. see Rouwenhorst (1998), McLean, Pontiff, and Watanabe (2009), Chui et al. (2010), Lam and Wei (2011), B. M. Barber et al. (2013), Titman et al. (2013), and Watanabe et al. (2013). The goal here is not the study of the performance of the anomalies outside of the U.S., but rather the use of international historical performance of the anomalies to better select anomalies that are likely to outperform in the future.

international sample should be beneficial. On the other hand, if the drivers are primarily *local*, then the larger international sample should have little benefit. The role of international evidence for the mispricing signal is related to a variety of risk-factor structures outside the U.S. The international evidence is likely to add little value if there is no proximity of risk-factor structures across the regions. For investigations of the risk-factor structure of international stock returns, see Rouwenhorst (1999), Griffin (2002), Griffin et al. (2010), Hou, Karolyi, and Kho (2011), Fama and French (2012), Fama and French (2017), and Bartram and Grinblatt (2018).

We find that there is only a little gain in the performance of the mispricing strategy in the U.S. when the strategy's estimation sample is extended from U.S. stocks to international stocks. The profitability of the mispricing strategy in the other regions, however, improves when we extend the estimation sample from the U.S. stocks to stocks in the respective regions. The mispricing of stocks estimated on historical data in the U.S. captures most of the predictability of stock returns outside the U.S. Taken together, our results imply that both local and global drivers of stock returns are important.

Our paper is structured as follows. We start with the data and methodology description in Section 1.1. The profitability of the mispricing strategy estimated in the U.S. is examined in Section 1.2 and the value of international evidence in Section 1.3. Finally, in Section 1.4, we examine the effect of transaction costs on the profitability of the mispricing strategy.

1.1 Data and methodology

1.1.1 Data

Our source of the accounting and market data for the U.S. is the Merged CRSP/Compustat database from the Wharton Research Data Service (WRDS). The sample spans from 1963 to 2018 period and contains all NYSE, AMEX, and NASDAQ common stocks (CRSP share code 10 or 11). The returns are adjusted for delisting following guidance as in Hou, Xue, and Zhang (2018).[5]

[5]If the delisting is on the last day of the month, returns over the month are used. The relevant delisting return is then added as a return over the next month. Delisting return (DLRET) from monthly file is used if it is not missing. $(1 + ret_{cum}) * (1 + DLRET_d) - 1$ is used if it is missing, where ret_{cum} is the cumulative return in the given month of delisting and $DLRET_d$ is the delisting return from the daily file. Lastly, the gaps are filled with $(1 + ret_{cum}) * (1 + DLRET_{avg}) - 1$, where $DLRET_{avg}$ is the average delisting return for stocks with the same first digit of delisting code (DLSTCD).

The international data are sourced from Reuters Datastream. They are filtered following Ince and Porter (2006), Griffin et al. (2010) and K.-H. Lee (2011). We manually check the names of the shares in the database for over 100 expressions describing their share class. Only the primary quotes of the ordinary shares of the companies are retained, with few exceptions where fundamental data in Datastream are linked to other share classes.[6] Real estate investment trusts (REITs) are excluded from the sample. All the international returns and financial statements are converted to U.S. dollars. The daily returns are deleted for days when the stock market is closed in a given country. The quality of data is further improved with procedures described in Tobek and Hronec (2018) and covered in Appendix 1.A. Tobek and Hronec (2018) study the implications of the choice of the fundamental database on the measurement of performance of individual fundamental anomalies. They show that the statistical significance of the individual anomalies varies across Datastream and Compustat. The research inference can, therefore, change when a different fundamental database is used. The differences across the databases are mainly due to imperfect historical fundamental coverage. Studies of the aggregated performance of the anomalies, however, do not suffer from these problems. Therefore, our analysis is not impacted.

The sample includes 23 developed countries. The countries are sorted into four regions: the U.S.; Europe (E) — Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, the Netherlands, Norway, Portugal, Spain, Sweden, Switzerland, and the United Kingdom; Japan (J); and Asia Pacific (AP) - Australia, New Zealand, Hong Kong, and Singapore.

Another important source of our data for the anomalies is the Institutional Brokers' Estimate System (I/B/E/S), which is obtained from WRDS. I/B/E/S is merged on Datastream directly as it is one of the databases provided by Thompson Reuters and Datastream includes the respective tickers in its static file. The merger with CRSP is done indirectly through CUSIPs. The databases are merged on eight-digit CUSIP and then on six-digit CUSIP if unsuccessful. The success of the merger is checked manually by comparing the quoted tickers on the exchanges with the actual names of the companies. All the variables in I/B/E/S are transformed to U.S. dollars with original Reuters exchange rates, which are provided by WRDS.

We focus on a very liquid universe of stocks. This universe covers only stocks that are both (a) within the top 95% of the overall capitalization of all stocks

[6]We follow the description in Griffin et al. (2010) on the classification of common shares.

in each region at the end of the previous month and (b) within the top 95% of the overall dollar trading volume over the previous 12 months of all stocks in each region.[7] All the stocks are required to have a price larger than \$1 (\$0.1 for Asia Pacific) at the end of the previous month. We select the size of the stock universe to provide as liquid universe as possible while retaining a reasonable number of stocks for the estimation.

Our focus on the liquid universe of stocks makes the findings more realistic. The stocks with small capitalization (micro-caps) account for only a small fraction of the overall capitalization of the market and often cannot be traded at significant volumes due to their high illiquidity.

Table 1.1 shows the average, minimum, and maximum number of stocks in the cross-section of the individual regions. There are, on average, about 1,100 stocks in the U.S. that satisfy the inclusion criteria for the liquid universe. The average number of stocks satisfying the criteria is even smaller in the other regions. The average capitalization of stocks in the liquid universe after July 1995 is \$12 billion in the U.S., \$12 billion in Europe, \$5 billion in Japan, and \$5 billion in Asia Pacific. The average size of the stocks in the sample is, therefore, quite balanced over the regions.

1.1.2 Anomalies

The sample includes 153 anomalies published in the literature. There are 93 fundamental, 11 I/B/E/S, and 49 market friction anomalies in the sample. The anomalies come almost exclusively from the top finance and accounting journals. Figure 1.1 shows the number of the published anomalies over time. It also shows the number of anomalies whose in-sample period in their respective studies has ended. The number of anomalies gradually increases over time without any apparent jumps. The full list of the anomalies is provided in Appendix 1.B. We primarily include anomalies described in Harvey, Liu, and Zhu (2016), McLean and Pontiff (2016) and Hou, Xue, and Zhang (2018). We focus on anomalies that are valid in the cross-section of stocks to be able to form long-short portfolios out of them. Any anomaly that is specific to the U.S., and which therefore cannot be constructed outside the U.S., are excluded.[8]

We update the fundamental signals for each month with financial statement

[7] Trading volume data are often not available before year 2000 for international stocks and before 1980 for NASDAQ stocks. A stricter condition on capitalization being within the top 90% of the overall capitalization of all stocks is used instead.

[8] Examples of excluded anomalies are those that are:

information from financial years ending at least six months prior. For trading information-based signals, such as market cap, we use the latest data available.[9]

Some anomalies, such as the Herfindahl Index of Hou and Robinson (2006), require the classification of industries for individual firms. The choice in the original papers is mostly with respect to the Standard Industrial Classification (SIC). We sort industries into 19 groups according to the third-level Datastream classification. The broader industry groups should make the results more robust and consistent across the data vendors. The industry classification in Datastream is available only from the static file, which means that only the latest values are available. Data vendors may slightly differ in the classification of individual firms over time because the differences between individual SIC categories are often subtle.

1.1.3 Mispricing strategy

In this subsection, we discuss the strategy that shrinks all the anomalies into a single mispricing signal ("mispricing strategy"). Lewellen et al. (2015) define the prediction problem as follows: the goal is to devise a forecasting method that predicts which stocks are likely to have the highest returns in the next month and which have the lowest based on stock characteristics (the cross-sectional anomalies). To do this, we regress monthly returns on individual stocks on their past characteristics. The future returns are then predicted from the latest available characteristics. We estimate the regressions by pooling all the available stock returns up to the date of portfolio formation. The past characteristics have to be available before the start of the measurement period of the returns. The characteristics are normalized to their cross-sectional quantiles within each region to reduce problems with outliers.

- based on quarterly fundamental data (since there is only short and problematic coverage internationally)
- connected to hand-collected data in the U.S., such as IPOs, SPOs, and mergers.
- requiring segment information and NBER data.
- institutionally specific, such as, share turnover or effective tax rate.

Some fundamental anomalies could not be implemented in Datastream as the required items are missing.

[9]We test the impact of the monthly updating of fundamental signals. Annual updating of the fundamental signals leads to identical conclusions.

To summarize, we estimate the following equation

$$r_{it} = f(x_{i,t-1,1}, x_{i,t-1,2}, ..., x_{i,t-1,M}) + \epsilon_{it}, \tag{1.1}$$

where r_{it} is the return on stock i in month t and $x_{i,t-1,1}$ is the cross-sectional quantile of a given anomaly (characteristic) for stock i available before the start of the month t. The returns are demeaned by subtracting average cross-sectional returns in every region-month. We first discuss a simpler case with linear $f()$. It is then extended to a more general structure using machine learning. The machine learning exercise follows Gu et al. (2020), who apply a suite of standard machine learning algorithms and show that they outperform the linear models in the U.S. Readers are referred to Gu et al. (2020) or any advanced machine learning textbook for a detailed theoretical description of the machine learning methods and only basic definitions are covered here.[10]

The machine learning methods have some benefits and some negatives. They provide better out-of-sample forecasts through a limitation of in-sample overfitting. They also allow for a very general interaction between the explanatory variables. This general form, however, makes the fitted models hard to estimate and the estimates hard to interpret due to the black-box approach. The intractability of the estimates is not a large concern in this study since even the linear method becomes intractable given the number of exogenous variables. Variable importance is examined in Subsections 1.3.3 and 1.3.4.

A crucial part of the application of machine learning methods is hyperparameter optimization or "tuning." Standard k-fold cross-validation does not respect temporal ordering, i.e., in $k-1$ out of k folds, predictability evaluation on the validation sample is done using observations from the future in training. Therefore we split the historical sample into subsequent mutually exclusive training and validation samples each time models are retrained. The validation sample consists of the most recent 11-year period at the point in time when the model is retrained each December. The training sample always starts with the beginning of the dataset and ends one month before the validation sample starts. It is, therefore, expanding over time.

We estimate the model using the training sample with various hyperparameters, and we measure its predictive performance using the root-mean-square error via the hyperparameter grid search on the validation sample. Hyperparameters with the best predictive performance are then selected for the

[10]See, for example, J. Friedman et al. (2001) for the textbook treatment.

estimation. As an example, the initial historical sample period is January 1963 to December 1994, and it is split into a training sample of 21 years (1963-1983) and a validation sample of 11 years (1984-1994). All models are re-estimated at the end of each year. The training sample is expanded by one year, and the length of the validation sample is kept the same.[11] The set of possible hyperparameters for individual machine learning methods is provided at the end of the specific model descriptions in the following subsections Figure 1.9 shows how the training, validation and test samples are split over time. and Table 1.7 in Appendix 1.C summarizes all methods and searched hyperparameters. The evolution of selected hyperparameters for individual models is in Appendix 1.C.

Weighted least squares

In the benchmark model, we use weighted least square estimation to determine a linear approximation of the relationship in equation (1.1). That is, we estimate weighted least squares regressions of the stock returns on the rescaled characteristics,

$$r_{it} = \beta_0 + \beta_1 x_{i,t-1,1} + \beta_2 x_{i,t-1,2} + ... + \beta_M x_{i,t-1,M} + \epsilon_{it}, \tag{1.2}$$

where the weight on individual observations is the inverse of the number of stocks in each time period and region. The weights are introduced to give equal importance to each time period. The weighting makes the moment conditions equivalent to the Fama and MacBeth (1973) regressions in Lewellen et al. (2015). The linear specification has already been applied in an international context in Jacobs and Müller (2018) and Jacobs and Müller (2020). It is therefore selected as a benchmark for the more complicated machine learning methods.[12]

Penalized weighted least squares

The linear regression model with many explanatory variables can overfit the realization of past data since it has many degrees of freedom. One way to reduce the overfitting is to introduce L1 and L2 penalties on the coefficients during

[11] The second historical sample period starts in January 1963 and ends in December 1995, and it is split into a training sample of 22 years (1963-1984) and a validation sample of 11 years (1985-1995).

[12] Capitalization-weighted regressions as in Green et al. (2017) have also been tried. The capitalization-weighting puts lower weight on small cap stocks and is more suited for value-weighted portfolios. The weighting did not outperform the selected method, and the results are therefore not reported here.

the estimation. There are two hyperparameters involved: α and λ. $\alpha \in [0, 1]$ is used to assign relative weights to the L1 and L2 penalties and λ controls the magnitude of penalization. The case with $\alpha = 1$, i.e., just L1 penalty, is denoted as the least absolute shrinkage and selection operator (LASSO) and was introduced in Tibshirani (1996). Tuned hyperparameters are α, considering values in $\{0, 0.1, \ldots, 0.9, 1\}$ and λ, considering values in $\{0.001, 0.0001\}$.

The regression tree family of methods is easy to estimate and requires a few specified hyperparameters. One such tree is depicted in Figure 1.2. The decision tree consists of nodes (the round-edged boxes) and outcomes (sharp-edged boxes). The outcomes are in percentage return per month.[13] The tree starts with a decision on whether a given stock is within the smallest 40% of stocks in the cross-section. The decision can then continue to the split based on the book-to-market ratio. The depicted tree is of depth 3, which is the maximum number of nodes in the longest branch. The tree allows for arbitrary cross-effects between the variables up to the (depth - 1) degree. We deal mainly with relatively shallow trees. The shallow trees are nonetheless able to capture various important interactions between the explanatory variables. random forest and gradient-boosting regression trees are based on a combination of individual trees. These methods cannot be easily visualized but they lead to better out-of-sample forecasting performance relative to simpler regression trees.

Random forest

Random forest is one of the most widely used ensemble tree methods. It combines forecasts from the individual decision trees that are based on subsamples of the training data. Explanatory variables are also subsampled in the individual trees to increase variety among the individual forecasts. Random forest is frequently among the top 10% of best-performing machine learning methods in various competitions and it is therefore a very robust method that is powerful in most of the settings. It requires only a few specified hyperparameters. The specification of the hyperparameters is not very important for its performance. It can therefore be used almost out-of-box. This is a large benefit with respect to neural networks where performance heavily depends on the model specification. The largest downside is that its estimations are time-consuming.

Tuned hyperparameters are *Number of trees*, considering values in {200, 300,

[13] The numbers are arbitrary and do not reflect real data.

400, 500, 600} and *Max depth of the tree*, considering values in {1, ..., 8, None}, where *None* means unlimited depth of the tree.

Other hyperparameters are fixed. The trees use randomly selected 50% of the overall training observations and the square root of the overall available explanatory variables. We employ a minimum node size of 0.1%, which is large enough to limit over-fitting but small enough to allow the method to approximate the true expected returns on stocks.[14]

Gradient boosting regression trees

The gradient boosting regression trees (GBRT) of J. H. Friedman (2001) rely on a different way of combining the regression trees than random forest. All the trees in a random forest are chosen independently, whereas they are selected in a dependent fashion in GBRT. The idea is to estimate a tree and use only a fraction of its fit for forecasts. The next iterations we conduct proceed on residuals of the dependent variable after we remove the fraction of the fitted values in the previous iteration. Shrinkage of the individual predictions guarantees that the learning can correct itself if the fitted values are selected suboptimally in some iterations. The fraction of individual predictions that is retained for the forecast is called a learning rate. The number of the learning iterations, given the learning rate, then determines how closely the particular realization of the sample from the whole population (the training sample) is over-fitted. A selection of fewer iterations reduces the risk of over-fitting (estimation error) but decreases the overall fit of the estimation (i.e., introduces an approximation error). It is therefore important to select the number of iterations with optimal estimation and approximation error trade-off. One way to do this is to rely on cross-validation. The method requires a specification of the learning rate, number of iterations (trees), and the maximum depth of the trees.

We conduct our analysis with a fast version of the gradient boosting — extreme gradient boosting (XGBOOST) of T. Chen and He (2017). The reason for this is that it is ten times faster to estimate and thus requires far less computational power. Gu et al. (2020) benchmark the different machine learning methods and only neural networks provide significantly better forecasts than GBRT. GBRT is therefore a good candidate for our empirical application and it captures most of the gains from the machine learning methods over the standard finance methods.

[14]Ignoring this parameter completely, and leaving unlimited node size, leads to almost identical results. It is thus not an important assumption.

Tuned hyperparameters are *Number of trees*, considering values in {50, 100, 200, 300, 400, 500}, *Maximum depth of the trees*, considering values in {1, ..., 9} and *learning rate*, considering values in {1%, 2.5%, 5%, 10%}.

Neural networks

Arguably the most powerful machine learning method of today is (deep) neural networks. Gu et al. (2020) show that they outperform any other method if they are optimally specified. The neural networks are a very flexible tool that encompasses many specifications.[15] The flexibility is also their largest disadvantage as it requires long experimentation and possible over-fitting of the sample.

Sequential neural networks consist of layers of neurons with information flowing between the layers in only one direction, from the input layer to the output layer. The information is fed in batches consisting of n sample points. Processing of the full training sample is called an epoch. The speed of change in the estimated parameters with new processed batches is determined through the learning rate. It is often an advantage to slow the learning rate over time to allow for the capture of finer details. We estimate the neural networks with back-propagation, along with a stochastic gradient descent version with adaptive moment estimation called Adam.[16]

In figure 1.3, we plot one of the neural network specifications corresponding to two architectures: *NN1_wide* or *NN1_narrow*. They are based on three layers. The initial layer has 150 neurons. The second hidden layer also has 150 neurons. Overall, six architectures are considered during the tuning of the hyperparameters and all of them have only one neuron in the last output layer.

- *NN1_wide*: One hidden layer with 150 neurons each.
- *NN1_narrow*: One hidden layer with 32 neurons respectively.
- *NN2_wide*: Two hidden layers with 150 neurons each
- *NN2_narrow*: Two hidden layers with 32 and 16 neurons respectively
- *NN3_wide*: Three hidden layers with 150, 150, and 100 neurons in the 1st, 2nd, and 3rd hidden layer.

[15] A linear regression is the simplest specification.
[16] See Kingma and Ba (2014).

- *NN3_narrow*: Three hidden layers with 32, 16, and 8 neurons in the 1st, 2nd, and 3rd hidden layer.

Narrow architectures are three out of the five neural network specifications used by Gu et al. (2020) and *wide* architectures are alternatives with a greater number of neurons offering higher model capacity. The input layer and all hidden layers use a rectified linear unit (ReLU) activation function while the last layer uses a linear activation. Input into each layer is batch normalized.

Additional fixed hyperparameters are a batch size of 256, maximum number of epochs equal to 25, and betas of 0.9 and 0.999 used in the Adam optimization.

Given the high model capacity of neural networks, regularization is of paramount importance. We employ dropout, early stopping, and ensembles for the purpose of regularization. For the *dropout* rate, we consider values of 0.001, 0.01, and 0.1. The meta-parameter in early stopping callback is called patience, and it determines when to stop the learning process conditional on the lack of the validation-based mean squared loss improvement in consecutive epochs. It is fixed at four. Finally, another callback called reduce learning rate on plateau is used to reduce the learning rate by a factor of two when the validation-based mean squared loss stops decreasing from one epoch to another. We produce the final forecast from an ensemble of five estimated neural networks with different initial random seeds. The combination forecast leads to a great improvement in the performance of the mispricing strategy based on neural networks.

1.1.4 Portfolio construction

The portfolios are constructed from the sorts of predicted returns in the individual regions. They are constructed as long-short and self-financing. In the long leg of the strategy, stocks are purchased in the upper quintile of the predicted next month's returns. In the short leg of the strategy, stocks short-sold in the bottom quintile of the predicted next month's returns. The quintile breakpoints are selected to provide more robust results relative to decile breakpoints. The portfolios are rebalanced every month based on signals from the end of the previous month. The portfolio returns correspond to an investable strategy that holds $1 in cash, invests $1 in the stocks that are likely to have the largest return in the next month, and shorts $1 worth of stocks that are likely to have the smallest return in the next month. The portfolios start in January 1995, unless stated otherwise.

A global strategy invests in stocks from all four regions. The global strategy is again based on stocks in the extreme quintiles of the predicted returns in the individual regions.

1.1.5 Liquidity measures

We use liquidity proxies to estimate transaction costs associated with investing into the mispricing strategy portfolios. The proxies are the Gibbs proxy of Hasbrouck (2009), the closing quoted spread proxy of Chung and Hao Zhang (2014), and the VoV(% Spread) of Fong et al. (2018). They are defined in Appendix 1.E.

We select these proxies to capture a fixed component of transaction costs and ignore the variable component that measures the price impact of larger orders. The variable component is very volatile and depends on the precise trade execution algorithm of each asset manager. The large capitalization universe of stocks reduces concerns about the variable component and it should be possible to avoid any execution costs altogether through the use of limit orders.

All of the proxies have some missing observations. The missing observations are backfilled from other proxies. The quoted spread is used first for the backfilling, followed by the VoV(% Spread), and the remaining missing observations are backfilled with the Gibbs proxy. Less than 0.02% of the observations are missing in all the three proxies and these observations are filled by 5% transaction costs.

1.2 Profitability of the mispricing strategy estimated in the U.S.

In this section, we examine the performance of the mispricing strategy worldwide. Jacobs and Müller (2018) show that the mispricing strategy estimated with least squares leads to higher returns in both absolute terms and on a risk-adjusted basis relative to the mixing of portfolios on individual anomalies. Gu et al. (2020) document that the more sophisticated machine learning methods provide higher out-of-sample predictability relative to the least squares method in the U.S. We extend the machine learning methods to the international sample to determine whether their benefits persist outside the U.S.

Table 1.2 presents the mean returns on portfolios created based on the mispricing strategy. The regressions of stock returns on their characteristics

are fit on data available up to December each year and the future stock returns are then predicted with the latest available characteristics for each of the next 12 months. The regressions are estimated with weighted least squares (WLS), penalized least squares (PWLS), gradient-boosting regression trees (GBRT), random forests (RF), and neural networks (NN). The estimates in Table 1.2 are based on the U.S. data going back to January 1963. Panel A in Table 1.2 shows the characteristics of returns on self-financing long-short portfolios. The long-short quintile portfolios invest in stocks in the top quintile of the predicted future returns and short-sell stocks in the bottom decile of the predicted returns. The reported portfolio returns are in percent per month and are from January 1995 to December 2018.

Panel A in Table 1.2 documents the striking profitability of the mispricing strategy across all regions. Diversification over the four regions (in the global columns) further reduces the maximum drawdowns and increases the Sharpe ratios for the mispricing strategy.

To understand the source of outperformance it is important to focus on the differences between these methods. The key differences between the most successful methods and more traditional methods are penalization, the interaction of predictive variables, and non-linearity. Penalization is used to solve multi-collinearity and overfitting problems. The benefit of penalization is best observed when comparing results from penalized least squares compared to the simple least squares. Gradient boosting regression trees, random forests, and neural networks further benefit from the non-linearity and possible interaction of predictive variables. The results in Panel A in Table 1.2 provide some support for the benefits of adding penalization to least squares and introducing non-linearity and variable interactions in the more complex methods.

The anomalies we use are different from anomalies used by Gu et al. (2020). Furthermore, there are no macroeconomic predictors, industry dummies, or explicit interactions between stock-level characteristics and factors. The maximum number of variables[17] we use is 153, which is smaller than the 920 variables used by Gu et al. (2020). Their variables are not filtered over time based on the publication date. Even though there is a different setup in both of these studies, the profitability of the mispricing strategy in the U.S. region is comparable with the profitability of the machine learning portfolios constructed by Gu et al. (2020). Our results, which are obtained in different empirical settings, provide support for the findings of Gu et al. (2020).

[17]Only anomalies published before the time of estimation are considered.

In Figure 1.4, we plot the cumulative returns on the neural network estimation method of the mispricing strategy in Panel A in Table 1.2. We find that there is a small drop in return profitability around 2003 in the U.S. The mispricing strategy is the least profitable in the European region on an equal-weighted basis.

1.2.1 Long-only and short-only components of the strategy

Short-selling can be connected to large costs and sometimes even outright impossible. That is why it might not be possible to replicate the returns on the mispricing strategy in practice.[18] In order to determine the role of short-selling for the strategy's y, we decompose the results of the long-short portfolios in Panel A in Table 1.2 into the long-only and short-only components in Panel B and C. We also decompose the long-short returns separately for the individual machine learning methods. The long-only component can be compared to equal-weighted and value-weighted returns on the whole market comprised of the liquid stocks, as described in Subsection 1.1.1.

Panel B in Table 1.2 documents that the mispricing strategy is more profitable than the whole market in all regions. The long-only component is responsible for most of the returns on the mispricing strategy. The short-only component then mainly serves as a hedge that increases the Sharpe ratio and lowers the maximum drawdown. The annualized returns on the long-only component of the mispricing strategy employing neural networks are about 5% a year larger than returns on the market. The other machine learning methods also outperform the market.

The more advanced machine learning methods outperform simple least squares both on the long side of the portfolio and on the short side. To conclude, the out-performance of the mispricing strategy is robust to short-selling constraints. Even short-selling-constrained investors can therefore benefit from the strategy.

[18]Short-selling constraints should not be a large issue in our liquid universe of stocks. Andrikopoulos et al. (2013) show that although some stocks cannot be short-sold in practice, focusing only on those that can be short-sold does not statistically diminish returns on eight quantitative strategies in the United Kingdom. They also show that short-selling costs are small at about 1% annually in the United Kingdom.

1.2.2 Risk-adjusted performance of the strategy

Thus far, our focus has been on the raw returns on the mispricing strategy without accounting for any risk factors. Panel D in Table 1.2 presents the performance of the long-short strategy as presented in Panel A in Table 1.2 after accounting for market returns and five Fama-French factors. The results in Panel D in Table 1.2 show that accounting for market returns have little impact on the performance of the strategy as the capital asset pricing model (CAPM) alpha is close to the mean returns for all the estimation methods. This is expected, since the portfolios are long-short and thus close to market neutral by construction.

The results are, however, very different when adjusting for the five Fama-French factors. Across the regions, the alphas are mostly insignificant for the weighted least squares portfolio but become significant at 5% level for penalized weighted least squares portfolio for all regions except Europe for the equal-weighted portfolios. Since penalization is the only difference, it clearly shows one part of the difference with respect to the traditional risk factors.

The difference between the mean returns and alphas is substantially smaller for the machine learning methods than for the linear estimation methods. This is true for the U.S. as well as internationally. The linear estimation methods therefore lead to the mispricing signal that is close to the traditional risk factors. On the other hand, due to the non-linear interactions between the predictive variables, the gradient-boosting regression trees, random forests and neural networks are able to capture the predictive relationships that linear methods are not able to.

To conclude, the profitability of the mispricing strategy is significant even on a risk-factor-adjusted basis and this holds true in all the regions. Looking at the risk-adjusted performance of the mispricing strategy when employing different underlying models further clarifies the reasons for the superiority of the machine learning methods. Penalization and the non-linear interaction of the variables lead to returns that are unrelated to the traditional risk factors.

1.3 The role of international evidence

The evidence thus far documents that the mispricing strategy trained on the past data in the U.S. is profitable out-of-sample in all the regions. Can international data outside the U.S. be used to better train the strategy?

There are some arguments for the usefulness of international data. The international data increases sample size and therefore limits the possibility for data-mining and in-sample overfitting. The larger sample size also generally provides larger power to statistical tests, which should lead to the more precise selection of truly significant strategies. One crucial requirement for the tangible benefit of the new observations is that they are independent from the original observations. The international evidence extends the sample size mainly in the year 2000. The most recent data are also the most useful as financial markets change rapidly and the older data may not be relevant anymore.

There are, however, also some problems with the suitability of international evidence. The individual global regions have very different institutional settings. Bankruptcy laws, tax laws, investor protection, and accounting standards vary widely across the regions. The institutional differences can lower the usefulness of historical data outside the respective regions. The larger estimation sample improves forecasts through consistency. The consistency, however, works only if the underlying true drivers of stock returns are uniform over the regions, which is in no way guaranteed.

1.3.1 Locally-trained mispricing strategy

Previous machine learning evidence is based on predictive regressions estimated solely on data from the U.S. In this subsection, we first investigate whether estimating the predictive regressions in the respective regions is more suitable than estimating them only on data from the U.S. In the next subsection, we explore whether combining estimation samples from the individual regions can improve the profitability of the mispricing strategy.

Panel in Table 1.3 presents statistics on the portfolio returns of mispricing strategy introduced in Subsection 1.1.3 that is estimated on local samples of stocks. That is, for example, the portfolios in Japan are formed based on predicted next month returns from predictive regressions fitted on historical data from Japan. The Table 1.3 also shows the differences in portfolio returns when using the predictive regressions estimated on data from the U.S. versus locally.

There is surprisingly only a small difference between the returns on strategies that are estimated using data from the U.S. in Table 1.2 and those that are estimated using data in the respective regions in Panel A in Table 1.3. One

explanation for the similarity is that the sample size in the U.S. is already large enough to capture the true drivers of stock returns that are globally valid.

The performance of the mispricing strategy in Asia Pacific improves when the predictive regressions are estimated in Asia Pacific for all the machine learning methods apart from neural networks. However, there are only a few liquid stocks in Asia Pacific region historically. The local estimation sample is therefore composed of stocks with relatively a smaller market cap than in the U.S. sample. This different stock universe composition might help in tailoring the fitted relationships.

The performance of the mispricing strategy in Japan is notably worse than when estimated using the U.S. data. The explanation is again simple. Japan experienced a slow eruption of an asset price bubble at the beginning of the estimation sample, in the early 1990s. The estimated relationships that are valid for this specific period fare badly out-of-sample where the stock market dynamics go back to their normal state.

1.3.2 Globally-trained mispricing strategy

Panel B in Table 1.3 shows the mean returns and other performance statistics for the mispricing strategy when the future individual stock returns are predicted from regressions estimated on historical data that are not solely from the U.S. but from the global training sample comprised of stocks from the U.S., Japan, Europe, and Asia Pacific.

We find that there is no gain from adding international stocks to the local training sample in the U.S. Historical data in the U.S. is therefore completely sufficient for future predictions in the U.S. Profitability of the mispricing strategy in Europe improves with predictions based on the global training sample relative to the training sample from the U.S. The profitability in Japan also improves with the global training sample instead of from the U.S. only sample. The largest gains in profitability from extending the training sample outside the U.S. are in the Asia Pacific region.

To conclude, table 1.3 provides mixed results on the value of international evidence. The region-specific settings are indeed an important determinant of stock return drivers. There is no gain for the U.S. investor seeking international evidence for the quantitative strategy. The larger statistical power, due to a larger sample, seems to be completely offset by the region-specific differences.

1.3.3 Variable importance

One of the disadvantages of the more complex machine learning methods is the difficulty in interpreting the resulting models due to the potentially high-dimensional and nonlinear interactions among variables. Our main goal is the superior out-of-sample performance even at the cost of the inability to fully interpret all the variable interactions in the resulting models. That being said, inspired by Sirignano et al. (2016), L. Chen et al. (2024), and Horel and Giesecke (2019), we next examine the importance of individual variables. We define the variable importance VI_j for variable j as the elasticity of predicted (region-wise demeaned) returns to changes in the individual characteristics used as predictors, as follows:

$$VI_j = \sum_{t \in T} \sum_{i \in N_t} \left| \frac{\partial \hat{r}_{it}(x_{i,t-1,1}, x_{i,t-1,2}, ..., x_{i,t-1,M})}{\partial x_j} \right|, \tag{1.3}$$

We calculate the variable importance VI_j for each characteristic in various settings. Figure 1.5 shows the variable importance across regions for the twenty five most important variables globally for the mispricing strategy as described in Subsection 1.1.3. The most important predictors are all connected to high portfolio rebalancing costs. The most important fundamental signal is sales over price, which is a measure of the value of the stocks. The second most important fundamental signal, a measure of R&D spendings, is important only in the U.S. where it was identified.

Table 1.4 shows the Spearman's rank correlation of variable importance scores across the regions under various forecasting methods. we find that there is great heterogeneity in the ranks of variable importance across the regions. This results provide insight on the limited value of extending the estimation sample from the U.S. to international stocks. More importantly, the predictions from the U.S. perform as well as the predictions from the other regions despite having only loosely connected variable importance.

We also find that there are pronounced differences in variable importance under different forecasting methods. The Spearman's rank correlation coefficient between variable importance scores using neural networks and gradient boosting regression trees is only 0.498. The dispersion of variable importance across the estimation methods does not translate to out-of-sample profitability. This suggests that there is a lot of noise behind the estimation.

1.3.4 Marginal predictive contributions of signals

Another option when interpreting results from the machine learning models is to look at the marginal relationships. The marginal predictive relationship MRj between individual signal j and (region-wise demeaned) expected returns is determined as follows:

$$MR_j(x_j) = \frac{1}{T}\sum_{t\in T}\frac{\sum_{i\in N_t}\hat{f}_{i,t}(x_{i,t-1,1},...,x_{i,t-1,j},...,x_{i,t-1,M})}{N_t} \quad \text{for } x_j \in (0,1), \tag{1.4}$$

where $\hat{f}_{i,t}(x_{i,t-1,1},...,x_{i,t-1,j},...,x_{i,t-1,M})$ is a model prediction. The value of signal x_j is evaluated on a grid ranging from 0 to 1 with a step of 0.01. The values for x_j range from 0 to 1 since variables are region-month cross-sectional quantiles of an underlying characteristic. Values of all signals other than x_j are taken at their true historical realizations. Models are trained on the same regions where forecasts are generated. A set of 100 forecasts corresponding to different values of x_j is obtained by averaging over the cross-section and over time.

Marginal predictive relationships for *size, momentum, book-to-market, asset growth, operating profitability*, and *short-term reversal* characteristics can be seen in Figures 1.6 and 1.7.

They show that machine learning models capture the well-known empirical asset pricing relationships. Our results are similar to the variable importance results of Gu et al. (2020). We find that the monotonicity of the relationship varies from model to model but the marginal relationships are similar to the relationships documented in the literature for all six reported characteristics but *asset growth* (AGr). In case of asset growth, as shown in Figure 1.7, we find its association with expected returns is missing for more complex models as can be seen by comparing the slopes for weighted least squares and penalized least squares. The zero slope for penalized least squares is due to the penalization, which is one of the two core features responsible for the superiority of machine learning methods. The second feature is non-linearity. It can be seen by comparing the marginal relationship for penalized weighted least squares model with those for gradient boosted trees, random forests or neural networks. Differences in forecasting performance after introduction of penalization and non-linearity answers how machine learning techniques outperform more standard approaches.

Figures 1.6 and 1.7 also show the effects of non-linearity and penalization internationally as well as in individual regions. These effects are however more diverse across the regions. We also document some of the well known empirical observations like the lack of a momentum anomaly or presence of book-to-market anomaly in Japan (Fama and French 2012). Other relationships are missing at the marginal level, like the momentum anomaly in Europe, e.g., Asness et al. (2013). One of the traditionally strongest signals, short-term momentum, shows a strong relationship in the U.S. as well as in all the other regions except for Asia Pacific. Despite the international diversity of the marginal relationships for anomalies, out-of-sample profitability is not fundamentally impacted. This supports the notion of a high level of noise in estimation as mentioned at the end of the Section 1.2.

1.4 Transaction costs

In this section, we describe the out-of-sample performance of the strategies after accounting for their transaction costs.

1.4.1 Transaction costs on the strategy

Panel A in Table 1.5 presents the average transaction costs on the mispricing strategy introduced in Subsection 1.1.3 using neural networks. We estimate transaction costs using the three liquidity proxies introduced in Subsection 1.1.5. All the proxies provide similar estimates of the transaction costs outside the U.S. Estimates from the Gibbs proxy are significantly higher in the U.S. than for the two other proxies. The Gibbs proxy is, however, also the noisiest proxy since it is constructed at an annual frequency. Thus, it is not very suitable for measuring the transaction costs for the most liquid stocks due to its construction relying on the auto-correlation of daily stock returns.

Panel A in Table 1.5 also shows the turnover of the mispricing strategy. The turnover is given as:

$$Turnover_t = \sum_i abs(w_{i,t} - w_{i,t-1} r_{i,t-1})/2, \tag{1.5}$$

where $w_{i,t}$ is the weight of stock i in the investment portfolio at the start of period $t-1$ and $r_{i,t-1}$ is stock return over period $t-1$ to t. The sum of all the absolute weights $w_{i,t}$ is equal to 2 since the portfolio is long-short. We find that

the turnover is close to 100% monthly in all the regions, which means that over 50% of all the held stocks have to be sold and new purchased for both the short and long leg of the strategy. The turnover can be easily reduced by staggered portfolio rebalancing but it is not a source of serious worries here due to the small average transaction costs on the liquid universe of stocks.

We select the sample of stocks to be liquid ex ante. Only about 1,000 of the most liquid U.S. stocks fulfil this criterion. These stocks should command virtually no fixed transaction costs after year 2010. The depicted costs therefore correspond to unfavourable trade executions through aggressive marketable orders. Sophisticated trade execution systems using limit orders are able to execute the strategies with much smaller transaction costs.

In figure 1.8, we map transaction costs on the mispricing strategy we estimate using neural networks in the U.S. The transaction costs are measured by using VoV(% Spread) proxy introduced in Fong et al. (2018). The figure shows that the trading costs are similar across the regions. The highest transaction costs tend to be in the Asia Pacific region and the lowest tend to be in the U.S. The transaction costs decrease significantly over time due to the advent of an electronic trading in 2000s. There are several historical episodes where the costs are heavily elevated. Two such major episodes are the Global Financial Crisis of 2007-2009 and dot-com bubble of the early 1995-2001. The costs are smaller on value-weighted portfolios relative to equal-weighted portfolios. The difference is expected because value-weighting puts a larger emphasis on more liquid stocks.

1.4.2 Performance of the strategy after transaction costs

Panel B in Table 1.5 presents the transaction costs adjusted performance of the mispricing strategy. We find that the mean returns on the strategy remain significantly positive at the 5% level. The net mean annualized returns in the U.S. are around 10% for the machine learning strategies. Sharpe ratios remain high, especially for the global strategy using neural networks, where they are close to one.

The mean returns after transaction costs for the weighted least squares method are again smaller than for the more advanced machine learning methods. The difference is even larger on a risk-adjusted basis. This difference in performance documents that the choice of appropriate forecasting method is very important for the success of investing based on the anomalies.

To conclude, the mispricing strategy remains profitable even after accounting

for the transaction costs. The profitability of the strategy can therefore be capitalized by investors.

1.5 Conclusion

We study the profitability of quantitative strategies based on previously documented anomalies around the globe. We show that synthesizing anomalies into one mispricing signal using machine learning leads to profitable investment strategy which survives on the liquid universe of stocks and after accounting for the transaction costs. The machine learning methods lead to higher (risk-adjusted) returns relative to standard methods applied in the finance literature.

We examine the role of international evidence on the precision of predictions of future stock returns. We find that out-of-sample performance in the U.S. is not improved by the inclusion of international evidence in the training sample for the mispricing strategy. Most of the predictability of the expected stock returns in all the regions can be captured solely with the U.S. training sample as the benefits of a larger estimation sample is offset by the region-specific differences.

	mean	min	max
U.S.	1100	647	1734
Japan	750	534	1079
Asia Pacific	431	226	712
Europe	691	413	1044
Global	2069	647	4058

Table 1.1. Number of stocks in cross-section
Regional decomposition the number of stocks in the sample from January 1995 to December 2018. The sample consists of stocks within the top 95% of the overall capitalization of all stocks in each region at the end of the previous month and within the top 95% of the overall dollar trading volume over the previous 12 months of all stocks in each region. All the stocks are further required to have a price larger than $1 ($0.1 for Asia Pacific) at the end of the previous month.

	Equal-weighted					Value-weighted				
	U.S.	Europe	Japan	AP	Global	U.S.	Europe	Japan	AP	Global
Panel A: Long-short Portfolios										
Weighted Least Squares										
Mean	0.513	0.694	0.563	0.434	0.560	0.328	0.728	0.518	0.515	0.476
t-stat	1.868	2.680	2.855	1.998	2.487	1.530	3.408	2.818	1.824	2.427
Sharpe Ratio	0.429	0.814	0.724	0.431	0.725	0.314	0.821	0.526	0.441	0.573
Max Drawdown	-33.84	-29.30	-22.61	-30.96	-23.21	-34.81	-24.16	-25.74	-54.95	-22.73
Penalized Weighted Least Squares										
Mean	0.651	0.798	0.809	1.048	0.807	0.496	0.630	0.614	0.761	0.623
t-stat	2.538	2.751	3.632	4.483	3.652	2.253	2.646	2.452	2.455	2.957
Sharpe Ratio	0.495	0.788	0.877	0.932	0.929	0.397	0.629	0.553	0.581	0.679
Max Drawdown	-36.13	-31.91	-25.33	-36.27	-26.05	-30.09	-28.56	-34.03	-60.19	-26.28
t-stat wrt WLS	1.104	0.785	1.317	2.888	2.184	1.283	-0.713	0.495	1.394	1.147
Gradient Boosting Regression Trees										
Mean	1.074	1.024	0.950	0.907	1.020	0.920	0.766	0.882	0.170	0.739
t-stat	3.728	3.616	4.542	4.126	4.657	3.201	3.100	3.336	0.837	3.732
Sharpe Ratio	0.870	1.100	1.108	0.974	1.415	0.728	0.749	0.754	0.175	0.954
Max Drawdown	-42.42	-31.07	-18.11	-19.11	-25.72	-45.22	-32.55	-21.42	-42.23	-25.61
t-stat wrt WLS	2.021	1.931	2.013	2.258	2.710	1.979	0.198	1.341	-1.197	1.305
t-stat wrt PWLS	1.663	1.818	0.718	-0.656	1.419	1.458	0.872	0.879	-2.004	0.636
Random Forest										
Mean	1.034	1.103	0.994	1.036	1.049	0.995	0.629	0.888	0.270	0.791
t-stat	3.387	4.017	4.096	4.534	4.590	3.349	2.374	3.455	1.126	3.663
Sharpe Ratio	0.875	1.247	1.182	1.109	1.440	0.894	0.632	0.773	0.252	1.030
Max Drawdown	-36.77	-27.01	-32.20	-20.63	-20.34	-39.88	-35.09	-21.57	-49.01	-19.23
t-stat wrt WLS	2.566	2.417	1.880	2.601	3.405	2.563	-0.459	1.292	-0.805	1.768
t-stat wrt PWLS	2.031	2.728	0.620	-0.048	1.803	1.917	-0.006	0.737	-1.454	0.944
t-stat wrt GBRT	-0.253	1.193	0.262	1.522	0.355	0.463	-1.608	0.030	0.742	0.610
Neural Networks										
Mean	1.128	0.994	1.107	1.367	1.149	1.185	0.734	0.702	0.828	1.012
t-stat	4.345	3.529	5.687	6.160	5.383	5.136	3.294	3.411	3.296	5.697
Sharpe Ratio	0.952	1.084	1.372	1.356	1.480	1.093	0.738	0.650	0.681	1.280
Max Drawdown	-25.69	-26.27	-14.78	-26.61	-17.56	-16.50	-31.48	-34.55	-44.99	-16.19
t-stat wrt WLS	5.205	2.260	2.945	4.510	6.100	4.986	0.039	0.733	1.588	3.665
t-stat wrt PWLS	5.450	2.938	3.526	1.830	5.591	4.425	0.875	0.401	0.321	3.118
t-stat wrt GBRT	0.231	-0.274	0.906	3.531	1.117	0.992	-0.260	-0.708	2.254	2.017
t-stat wrt RF	0.590	-1.242	0.405	2.148	1.022	0.839	0.780	-0.658	1.865	1.788

Table 1.2. Performance of the mispricing strategy estimated in the U.S.

The table shows returns on quintile portfolios from mispricing strategy described in Subection 1.1.3. The estimation methods are weighted least squares (WLS), penalized weighted least squares (PWLS), gradient boosting regression trees (GBRT), random forests (RF), or neural networks (NN). The regressions are rerun at the end of each December using only anomalies that have been documented by that time and using hyperparameters selected based on the most recent 11-year validation sample available. The regressions are estimated only on the past U.S. data and the future returns are predicted in all the regions. Long portfolios are constructed by buying stocks in the top quintile of the predicted next month returns and short portfolios are constructed by short-selling stocks in the bottom quintile of the predicted next month returns. The portfolios are either value-weighted or equal-weighted. Panel A presents returns on long-short portfolios. The returns on the long-short portfolios are decomposed to long-only component in Panel B and short-only component in Panel C. Panel D presents the performance of the long-short portfolio adjusted for capital asset pricing model (CAPM) market risk factor and five Fama-French factors (FF5). The out-of-sample performance is observed in the U.S., Europe, Japan, and Asia Pacific. The reported returns are for January 1995 to December 2018 period and are in percentage points. The standard errors in t-statistics are adjusted for heteroskedasticity and autocorrelation with Newey-West adjustment for up to 12 lags. *t-stat wrt WLS, PWLS, GBRT, RF* rows provide t-statistics for difference in mean returns between the two corresponding estimation methods.

	Equal-weighted					Value-weighted				
	U.S.	Europe	Japan	AP	Global	U.S.	Europe	Japan	AP	Global
Panel B: Long-only Component of the Mispricing Strategy										
Whole Market										
Mean	0.827	0.746	0.204	0.654	0.604	0.901	0.707	0.158	0.774	0.646
t-stat	2.265	1.929	0.472	1.314	1.766	2.892	2.037	0.360	1.743	2.022
Sharpe Ratio	0.459	0.466	0.121	0.298	0.408	0.599	0.463	0.098	0.425	0.479
Max Drawdown	-68.91	-63.60	-68.18	-69.02	-55.28	-62.60	-59.59	-72.67	-60.77	-57.51
Weighted Least Squares										
Mean	1.087	1.097	0.565	0.923	0.909	1.005	1.054	0.546	1.235	0.870
t-stat	3.648	2.986	1.505	1.986	3.011	3.714	3.125	1.608	3.066	3.058
Sharpe Ratio	0.717	0.701	0.344	0.438	0.667	0.761	0.708	0.364	0.681	0.680
Max Drawdown	-52.57	-63.60	-55.28	-64.28	-54.52	-45.02	-59.59	-53.92	-56.44	-52.98
t-stat wrt Mkt	1.454	2.619	2.755	1.913	2.322	0.851	2.477	2.087	2.340	2.112
Penalized Weighted Least Squares										
Mean	1.125	1.107	0.640	1.268	1.006	0.999	0.989	0.499	1.314	0.889
t-stat	3.589	2.942	1.612	2.641	3.130	3.402	2.947	1.477	3.617	3.122
Sharpe Ratio	0.747	0.715	0.379	0.564	0.729	0.750	0.690	0.325	0.682	0.720
Max Drawdown	-55.83	-62.68	-54.71	-68.24	-57.56	-51.21	-57.86	-51.97	-50.75	-53.01
t-stat wrt Mkt	1.667	2.410	2.915	3.674	3.032	0.741	1.887	1.387	2.288	1.909
t-stat wrt WLS	0.588	0.145	0.770	2.666	1.503	-0.080	-1.050	-0.388	0.538	0.263
Gradient Boosting Regression Trees										
Mean	1.405	1.136	0.712	1.096	1.108	1.328	1.044	0.714	0.892	0.991
t-stat	3.912	2.642	1.603	2.692	3.258	3.515	2.560	1.543	2.563	2.973
Sharpe Ratio	0.808	0.681	0.384	0.581	0.763	0.809	0.673	0.399	0.516	0.702
Max Drawdown	-53.46	-71.64	-58.85	-59.40	-57.99	-56.28	-68.87	-53.53	-57.08	-59.88
t-stat wrt Mkt	3.055	2.808	3.739	2.520	4.075	2.018	2.174	2.490	0.640	2.348
t-stat wrt WLS	2.070	0.353	1.160	1.205	2.064	1.589	-0.083	0.803	-1.809	0.934
t-stat wrt PWLS	1.808	0.316	0.595	-1.157	1.185	1.576	0.441	0.908	-2.334	0.784
Random Forest										
Mean	1.327	1.207	0.667	1.100	1.079	1.315	0.897	0.740	0.873	1.000
t-stat	3.900	2.877	1.510	2.709	3.309	3.674	2.095	1.528	2.493	3.008
Sharpe Ratio	0.820	0.731	0.354	0.574	0.757	0.858	0.548	0.403	0.506	0.713
Max Drawdown	-53.40	-68.86	-59.98	-59.46	-56.01	-57.71	-72.33	-59.92	-57.25	-60.16
t-stat wrt Mkt	2.668	3.308	2.796	2.449	3.748	2.087	1.059	2.601	0.540	2.399
t-stat wrt WLS	2.012	1.151	0.631	1.184	2.055	1.827	-1.157	0.835	-2.006	1.093
t-stat wrt PWLS	1.711	1.316	0.154	-1.075	0.973	1.747	-0.698	0.880	-2.394	0.837
t-stat wrt GBRT	-0.888	1.628	-0.420	0.099	-0.616	-0.104	-1.841	0.176	-0.479	0.162
Neural Networks										
Mean	1.361	1.150	0.733	1.278	1.124	1.397	1.005	0.624	1.287	1.138
t-stat	4.285	2.941	1.767	2.669	3.409	4.644	2.594	1.354	3.069	3.670
Sharpe Ratio	0.867	0.703	0.424	0.557	0.786	0.962	0.637	0.360	0.618	0.835
Max Drawdown	-51.78	-65.05	-56.26	-68.94	-56.40	-47.44	-62.53	-56.90	-63.85	-52.61
t-stat wrt Mkt	3.325	3.004	4.051	3.529	4.440	3.134	1.783	2.273	2.832	3.820
t-stat wrt WLS	4.014	0.768	1.804	2.859	3.881	3.599	-0.446	0.376	0.435	2.708
t-stat wrt PWLS	3.837	0.919	1.458	0.106	2.830	3.025	0.135	0.609	-0.162	2.293
t-stat wrt GBRT	-0.357	0.190	0.186	1.481	0.255	0.430	-0.384	-0.460	2.045	1.728
t-stat wrt RF	0.377	-1.080	0.400	1.347	0.809	0.655	0.927	-0.566	2.231	1.823

Table 1.2 Continued

	Equal-weighted					Value-weighted				
	U.S.	Europe	Japan	AP	Global	U.S.	Europe	Japan	AP	Global
Panel C: Short-only Component of the Mispricing Strategy										
Weighted Least Squares										
Mean	0.574	0.404	0.002	0.489	0.349	0.677	0.327	0.028	0.720	0.394
t-stat	1.365	0.898	0.004	0.940	0.881	1.766	0.776	0.067	1.497	1.097
Sharpe Ratio	0.298	0.227	0.001	0.210	0.216	0.427	0.194	0.017	0.339	0.271
Max Drawdown	-79.10	-76.00	-80.26	-74.83	-71.01	-73.01	-71.02	-75.00	-65.55	-69.68
Penalized Weighted Least Squares										
Mean	0.474	0.308	-0.169	0.220	0.199	0.503	0.359	-0.115	0.553	0.266
t-stat	1.143	0.687	-0.367	0.399	0.500	1.316	0.904	-0.268	1.141	0.731
Sharpe Ratio	0.234	0.165	-0.094	0.095	0.119	0.304	0.211	-0.070	0.260	0.177
Max Drawdown	-77.60	-77.13	-84.23	-72.45	-69.58	-75.86	-71.94	-73.83	-64.68	-71.39
t-stat wrt WLS	-1.320	-1.204	-1.552	-2.346	-2.478	-2.214	0.354	-1.410	-1.754	-1.746
Gradient Boosting Regression Trees										
Mean	0.331	0.113	-0.238	0.189	0.088	0.408	0.278	-0.168	0.722	0.252
t-stat	0.812	0.238	-0.519	0.347	0.218	1.152	0.692	-0.401	1.561	0.725
Sharpe Ratio	0.168	0.059	-0.132	0.083	0.053	0.259	0.161	-0.105	0.365	0.177
Max Drawdown	-76.71	-79.98	-87.35	-80.18	-71.48	-71.54	-73.19	-81.03	-64.23	-67.36
t-stat wrt WLS	-1.757	-3.043	-2.028	-2.872	-3.064	-2.181	-0.475	-1.633	0.015	-1.509
t-stat wrt PWLS	-1.254	-2.625	-0.677	-0.275	-1.518	-0.841	-1.209	-0.540	1.034	-0.218
Random Forest										
Mean	0.292	0.104	-0.327	0.064	0.030	0.320	0.268	-0.148	0.603	0.209
t-stat	0.656	0.221	-0.724	0.116	0.071	0.874	0.664	-0.372	1.252	0.585
Sharpe Ratio	0.144	0.055	-0.190	0.028	0.018	0.200	0.158	-0.093	0.295	0.143
Max Drawdown	-82.92	-77.15	-87.93	-81.12	-73.80	-77.32	-72.92	-80.36	-69.32	-70.95
t-stat wrt WLS	-2.598	-2.867	-2.949	-3.772	-4.013	-2.974	-0.559	-1.599	-0.590	-2.213
t-stat wrt PWLS	-1.899	-2.602	-1.153	-1.397	-2.364	-1.538	-1.199	-0.259	0.259	-0.937
t-stat wrt GBRT	-0.459	-0.181	-1.165	-2.008	-1.324	-1.177	-0.179	0.232	-1.024	-0.904
Neural Networks										
Mean	0.233	0.156	-0.373	-0.089	-0.025	0.212	0.271	-0.078	0.459	0.126
t-stat	0.515	0.328	-0.840	-0.154	-0.059	0.558	0.676	-0.190	0.939	0.355
Sharpe Ratio	0.109	0.083	-0.202	-0.037	-0.014	0.126	0.158	-0.046	0.218	0.084
Max Drawdown	-85.74	-79.71	-89.61	-82.61	-78.68	-80.52	-72.92	-73.10	-68.13	-71.24
t-stat wrt WLS	-4.445	-2.926	-2.965	-4.773	-5.782	-4.477	-0.607	-0.979	-1.753	-3.458
t-stat wrt PWLS	-4.082	-2.992	-3.571	-3.141	-5.723	-3.753	-1.901	0.666	-0.765	-3.551
t-stat wrt GBRT	-0.830	0.776	-1.681	-3.867	-1.991	-1.439	-0.100	1.081	-1.524	-1.925
t-stat wrt RF	-0.760	0.884	-0.378	-2.060	-1.137	-0.806	0.047	0.606	-0.744	-1.289

Table 1.2 Continued

	Equal-weighted					Value-weighted				
	U.S.	Europe	Japan	AP	Global	U.S.	Europe	Japan	AP	Global
Panel D: Performance on Risk-adjusted Basis										
Weighted Least Squares										
Mean Return	0.513	0.694	0.563	0.434	0.560	0.328	0.728	0.518	0.515	0.476
t-stat	1.868	2.680	2.855	1.998	2.487	1.530	3.408	2.818	1.824	2.427
CAPM Alpha	0.716	0.764	0.569	0.505	0.662	0.466	0.782	0.523	0.617	0.538
t-stat	2.794	3.615	3.394	2.378	3.634	2.105	4.139	2.791	2.709	2.950
FF5 Alpha	0.256	0.180	0.448	0.380	0.166	0.120	0.294	0.418	0.299	0.133
t-stat	1.303	1.278	2.877	2.178	1.278	0.701	1.861	2.267	1.340	0.898
Penalized Weighted Least Squares										
Mean Return	0.651	0.798	0.809	1.048	0.807	0.496	0.630	0.614	0.761	0.623
t-stat	2.538	2.751	3.632	4.483	3.652	2.253	2.646	2.452	2.455	2.957
CAPM Alpha	0.894	0.909	0.813	1.083	0.922	0.651	0.721	0.620	0.853	0.726
t-stat	3.476	3.768	3.916	4.550	4.683	2.524	3.462	2.632	3.422	3.878
FF5 Alpha	0.489	0.266	0.603	0.873	0.458	0.370	0.185	0.401	0.531	0.330
t-stat	2.465	1.660	3.110	3.820	3.042	1.961	1.179	1.806	1.900	2.095
Gradient Boosting Regression Trees										
Mean Return	1.074	1.024	0.950	0.907	1.020	0.920	0.766	0.882	0.170	0.739
t-stat	3.728	3.616	4.542	4.126	4.657	3.201	3.100	3.336	0.837	3.732
CAPM Alpha	1.194	1.105	0.949	0.988	1.094	0.917	0.820	0.877	0.204	0.748
t-stat	4.504	4.600	5.058	4.973	5.970	3.394	3.559	3.685	0.983	4.186
FF5 Alpha	1.158	0.490	0.754	0.690	0.848	1.018	0.386	0.708	0.024	0.602
t-stat	4.930	2.634	4.057	4.050	5.595	4.103	1.847	2.990	0.127	3.881
Random Forest										
Mean Return	1.034	1.103	0.994	1.036	1.049	0.995	0.629	0.888	0.270	0.791
t-stat	3.387	4.017	4.096	4.534	4.590	3.349	2.374	3.455	1.126	3.663
CAPM Alpha	1.199	1.180	0.993	1.119	1.136	1.034	0.645	0.883	0.307	0.810
t-stat	4.288	5.229	5.119	5.688	6.096	3.724	2.669	4.034	1.316	4.243
FF5 Alpha	0.757	0.673	0.846	0.747	0.690	0.806	0.225	0.725	0.016	0.487
t-stat	3.601	4.088	4.475	4.813	5.253	3.539	1.187	3.596	0.075	3.172
Neural Networks										
Mean Return	1.128	0.994	1.107	1.367	1.149	1.185	0.734	0.702	0.828	1.012
t-stat	4.345	3.529	5.687	6.160	5.383	5.136	3.294	3.411	3.296	5.697
CAPM Alpha	1.376	1.078	1.111	1.402	1.259	1.277	0.775	0.703	0.847	1.068
t-stat	5.680	4.631	6.121	6.591	6.777	5.378	3.416	3.247	3.586	6.285
FF5 Alpha	0.907	0.547	0.913	1.201	0.802	1.106	0.405	0.531	0.842	0.843
t-stat	5.305	3.620	5.657	6.597	6.323	5.941	2.062	2.812	3.124	5.668

Table 1.2 Continued

	Equal-weighted					Value-weighted				
	U.S.	Europe	Japan	AP	Global	U.S.	Europe	Japan	AP	Global
Panel A: Locally-trained Strategy										
Weighted Least Squares										
Mean	0.513	0.818	0.063	1.616	0.893	0.328	0.653	-0.113	0.714	0.584
t-stat	2.214	3.432	0.263	5.222	4.579	1.647	2.829	-0.403	2.473	3.111
Sharpe Ratio	0.429	0.787	0.060	1.073	1.024	0.314	0.573	-0.091	0.480	0.660
Max Drawdown	-33.84	-35.60	-62.92	-55.74	-26.74	-34.81	-33.01	-77.18	-45.19	-32.89
Diff mean wrt U.S.	0.000	0.125	-0.500	1.182	0.333	0.000	-0.074	-0.630	0.199	0.107
Diff t-stat wrt U.S.		0.724	-1.687	3.710	2.626		-0.358	-1.979	0.598	0.683
Penalized Weighted Least Squares										
Mean	0.651	0.819	0.466	1.640	0.889	0.496	0.651	0.135	0.880	0.658
t-stat	2.572	3.286	1.689	5.174	4.175	2.134	2.752	0.419	2.985	3.373
Sharpe Ratio	0.495	0.747	0.389	1.058	0.951	0.397	0.544	0.095	0.559	0.696
Max Drawdown	-36.13	-37.03	-62.61	-55.91	-27.62	-30.09	-32.61	-79.34	-47.59	-27.53
Diff mean wrt U.S.	0.000	0.021	-0.342	0.593	0.082	0.000	0.021	-0.479	0.119	0.035
Diff t-stat wrt U.S.		0.180	-1.102	2.169	1.463		0.175	-1.304	0.422	0.484
Gradient Boosting Regression Trees										
Mean	1.074	0.555	0.497	1.647	1.114	0.920	0.200	0.242	1.145	0.689
t-stat	4.464	2.429	1.830	7.497	6.523	3.731	0.934	0.822	4.289	3.890
Sharpe Ratio	0.870	0.489	0.420	1.535	1.404	0.728	0.174	0.176	0.837	0.772
Max Drawdown	-42.42	-48.38	-66.08	-29.00	-22.07	-45.22	-36.71	-74.31	-37.48	-26.04
Diff mean wrt U.S.	0.000	-0.468	-0.453	0.740	0.094	0.000	-0.566	-0.640	0.976	-0.050
Diff t-stat wrt U.S.		-1.697	-1.056	3.529	1.118		-2.364	-1.287	3.660	-0.404
Random Forest										
Mean	1.034	0.668	0.367	1.581	1.206	0.995	0.396	0.166	1.169	0.850
t-stat	4.173	2.717	1.241	8.258	7.020	4.066	1.675	0.571	4.242	4.877
Sharpe Ratio	0.875	0.563	0.299	1.675	1.584	0.894	0.342	0.124	0.884	1.029
Max Drawdown	-36.77	-47.62	-71.74	-21.65	-18.45	-39.88	-37.27	-72.45	-28.31	-21.92
Diff mean wrt USA	0.000	-0.435	-0.627	0.544	0.157	0.000	-0.234	-0.722	0.899	0.058
Diff t-stat wrt U.S.		-1.735	-1.286	2.790	2.349		-0.919	-1.619	3.107	0.536
Neural Networks										
Mean	1.128	0.925	0.755	1.315	1.321	1.185	0.700	0.418	0.711	1.101
t-stat	4.747	3.841	3.238	4.587	7.431	5.671	3.016	1.643	2.607	5.785
Sharpe Ratio	0.952	0.885	0.691	1.060	1.686	1.093	0.630	0.326	0.555	1.216
Max Drawdown	-25.69	-39.26	-47.98	-45.66	-18.19	-16.50	-35.36	-61.58	-50.49	-18.63
Diff mean wrt USA	0.000	-0.069	-0.352	-0.051	0.172	0.000	-0.034	-0.284	-0.117	0.090
Diff t-stat wrt U.S.		-0.479	-1.424	-0.151	3.674		-0.192	-1.008	-0.385	1.171

Table 1.3. Performance of the mispricing strategy estimated on the international data

The table shows returns on quintile long-short portfolios from mispricing strategy described in Subsection 1.1.3. The estimation methods are least squares, penalized least squares, random forests, gradient boosting regression trees, or neural networks. The regressions are rerun at the end of each January using only anomalies that have been published by that time and using hyperparameters selected based on the most recent 11-year validation sample available. The historical predictive regressions are estimated on individual stocks from all the four covered regions: the U.S., Japan, Europe, and Asia Pacific. Predictive regressions in Panel A are estimated on data from region where they are used for prediction. Predictive regressions in Panel B are estimated globally on data from all the regions. The long-short portfolios are constructed by buying stocks in the top quintile of the predicted next month returns and short-selling stocks in the bottom quintile of the predicted next month returns. The reported returns are for January 1995 to December 2018 period and are in percentage points. The standard errors in t-statistics are adjusted for heteroskedasticity and autocorrelation with Newey-West adjustment for up to 12 lags. Diff mean rows present difference of average monthly long-short portfolio returns between strategy estimated on the global data minus mean returns when it is estimated only on the U.S. data as in Table 1.2. Diff t-stat presents t-statistic for the significance of the difference in mean returns.

	Equal-weighted					Value-weighted				
	U.S.	Europe	Japan	AP	Global	U.S.	Europe	Japan	AP	Global
Panel B: Globally-trained Strategy										
Weighted Least Squares										
Mean	0.677	0.870	0.794	1.386	0.893	0.425	0.683	0.424	0.753	0.584
t-stat	2.446	3.165	3.454	6.278	3.944	1.907	3.017	1.922	2.935	2.767
Sharpe Ratio	0.514	0.978	0.791	1.185	1.024	0.357	0.695	0.359	0.546	0.660
Max Drawdown	-37.35	-26.27	-25.84	-26.54	-26.74	-40.62	-25.41	-48.21	-42.77	-32.89
Diff mean wrt U.S.	0.164	0.177	0.231	0.952	0.333	0.097	-0.045	-0.094	0.238	0.107
Diff t-stat wrt U.S.	1.225	1.319	1.040	4.187	2.626	0.525	-0.289	-0.417	1.124	0.683
Penalized Weighted Least Squares										
Mean	0.623	0.937	0.799	1.377	0.889	0.521	0.752	0.379	0.886	0.658
t-stat	2.116	3.130	3.385	5.637	3.643	2.255	3.360	1.484	3.458	3.152
Sharpe Ratio	0.438	0.981	0.765	1.187	0.951	0.399	0.750	0.308	0.672	0.696
Max Drawdown	-38.45	-29.32	-23.34	-27.87	-27.62	-31.57	-21.08	-49.26	-45.90	-27.53
Diff mean wrt USA	-0.028	0.139	-0.009	0.330	0.082	0.025	0.122	-0.235	0.126	0.035
Diff t-stat wrt U.S.	-0.376	2.189	-0.086	3.047	1.463	0.239	1.408	-1.655	0.740	0.484
Gradient Boosting Regression Trees										
Mean	0.922	1.134	0.885	1.772	1.114	0.646	0.779	0.608	1.173	0.689
t-stat	3.224	4.027	4.625	7.879	5.326	2.470	3.336	2.477	3.776	3.723
Sharpe Ratio	0.722	1.201	1.054	1.944	1.404	0.478	0.772	0.529	0.919	0.772
Max Drawdown	-36.12	-20.18	-23.82	-9.389	-22.07	-38.94	-28.14	-36.18	-40.84	-26.04
Diff mean wrt USA	-0.152	0.110	-0.065	0.865	0.094	-0.275	0.013	-0.274	1.004	-0.050
Diff t-stat wrt U.S.	-1.260	1.061	-0.419	5.048	1.118	-1.466	0.114	-1.239	3.678	-0.404
Random Forest										
Mean	1.032	1.127	1.027	1.825	1.206	0.795	0.799	0.587	1.266	0.850
t-stat	3.321	4.446	5.375	9.468	5.508	2.840	3.231	2.593	4.840	4.126
Sharpe Ratio	0.889	1.212	1.217	1.925	1.584	0.699	0.787	0.514	0.969	1.029
Max Drawdown	-30.95	-19.19	-21.56	-21.68	-18.45	-39.01	-23.14	-35.41	-31.47	-21.92
Diff mean wrt U.S.	-0.002	0.024	0.032	0.788	0.157	-0.200	0.170	-0.301	0.996	0.058
Diff t-stat wrt U.S.	-0.024	0.233	0.147	5.081	2.349	-1.735	1.264	-1.119	5.211	0.536
Neural Networks										
Mean	1.123	1.205	1.272	1.848	1.321	1.161	0.835	0.818	1.405	1.101
t-stat	4.037	4.563	6.777	7.964	6.083	3.913	3.409	3.998	5.196	4.971
Sharpe Ratio	0.944	1.345	1.488	1.779	1.686	0.969	0.774	0.663	1.048	1.216
Max Drawdown	-23.70	-23.12	-16.81	-23.71	-18.19	-23.71	28.61	-38.28	-33.07	-18.63
Diff mean wrt U.S.	-0.005	0.211	0.165	0.482	0.172	-0.024	0.102	0.116	0.577	0.090
Diff t-stat wrt U.S.	-0.069	2.973	2.017	5.717	3.674	-0.217	1.242	0.804	3.268	1.171

Table 1.3 Continued

	Weighted Least Squares					Gradient Boosting Regression Trees					Neural Networks				
	Global	U.S.	Europe	Japan	AP	Global	U.S.	Europe	Japan	AP	Global	U.S.	Europe	Japan	AP
Global	1.000	0.552	0.610	0.564	0.576	1.000	0.737	0.712	0.634	0.705	1.000	0.516	0.314	0.385	0.290
U.S.	0.552	1.000	0.406	0.363	0.397	0.737	1.000	0.611	0.490	0.512	0.516	1.000	0.378	0.278	0.309
Europe	0.610	0.406	1.000	0.482	0.443	0.712	0.611	1.000	0.548	0.603	0.314	0.378	1.000	0.222	0.299
Japan	0.564	0.363	0.482	1.000	0.541	0.634	0.490	0.548	1.000	0.590	0.385	0.278	0.222	1.000	0.171
AP	0.576	0.397	0.443	0.541	1.000	0.705	0.512	0.603	0.590	1.000	0.290	0.309	0.299	0.171	1.000

Table 1.4. Spearman's correlation matrices for region-specific variable importance

The table shows Spearman correlation matrices between region-specific ranks of variable importance as in equation (1.3) for the mispricing strategy. The mispricing strategy is described in Subsection 1.1.3. The strategy combines signals through predictive regressions of individual stock returns on transformed characteristics. The historical predictive regressions are estimated on data from the individual regions using weighted least squares, gradient boosting regression trees, or neural networks. Estimation sample consists of period from January 1963 to December 2018 in the U.S. and January 1990 to December 2018 in other regions.

	Equal-weighted					Value-weighted				
	U.S.	Europe	Japan	AP	Global	U.S.	Europe	Japan	AP	Global
Panel A: Transaction Costs on the Portfolios Using Neural Networks										
VoV	0.248	0.346	0.468	0.590	0.413	0.163	0.259	0.314	0.365	0.275
Gibbs	0.663	0.529	0.694	0.725	0.653	0.595	0.481	0.628	0.553	0.564
Quoted Spread	0.495	0.548	0.498	0.859	0.600	0.424	0.390	0.405	0.578	0.449
Turnover	103.90	104.20	102.70	104.10	103.70	118.00	110.00	111.10	112.90	113.00
Panel B: Net Returns on the Portfolios										
Weighted Least Squares										
Mean	0.339	0.471	0.249	0.048	0.286	0.226	0.564	0.332	0.302	0.311
t-stat	1.296	1.905	1.508	0.260	1.407	1.077	2.741	1.988	1.124	1.680
Sharpe Ratio	0.284	0.556	0.326	0.049	0.376	0.216	0.637	0.339	0.260	0.375
Max Drawdown	-39.42	-31.62	-26.14	-44.42	-29.86	-40.58	-25.45	-34.09	-59.97	-30.95
Mean	0.392	0.447	0.326	0.450	0.385	0.338	0.382	0.321	0.417	0.362
t-stat	1.549	1.546	1.535	1.856	1.767	1.541	1.637	1.329	1.337	1.749
Sharpe Ratio	0.298	0.440	0.354	0.399	0.442	0.271	0.382	0.290	0.317	0.394
Max Drawdown	-41.57	-34.63	-31.94	-49.51	-34.01	-35.58	-31.52	-40.08	-64.30	-34.50
Gradient Boosting Regression Trees										
Mean	0.809	0.636	0.426	0.475	0.618	0.751	0.473	0.543	-0.050	0.484
t-stat	2.890	2.249	2.368	2.291	2.999	2.646	1.939	2.174	-0.252	2.533
Sharpe Ratio	0.657	0.682	0.502	0.514	0.864	0.594	0.462	0.466	-0.052	0.626
Max Drawdown	-45.11	-38.11	-29.50	-23.96	-30.16	-47.00	-37.68	-29.94	-50.99	-28.08
Random Forest										
Mean	0.763	0.710	0.475	0.587	0.641	0.822	0.328	0.562	0.036	0.533
t-stat	2.558	2.592	2.239	2.684	2.972	2.809	1.245	2.294	0.157	2.576
Sharpe Ratio	0.647	0.800	0.574	0.631	0.887	0.740	0.330	0.490	0.033	0.697
Max Drawdown	-39.67	-30.90	-39.47	-23.73	-25.08	-41.70	-40.69	-23.01	-56.66	-22.41
Neural Networks										
Mean	0.880	0.648	0.639	0.776	0.736	1.022	0.475	0.388	0.463	0.736
t-stat	3.473	2.330	3.495	3.732	3.597	4.550	2.234	1.946	1.868	4.466
Sharpe Ratio	0.744	0.706	0.796	0.774	0.952	0.944	0.479	0.359	0.381	0.935
Max Drawdown	-26.97	-28.83	-18.15	-29.90	-20.34	-17.17	-32.92	-35.65	-49.58	-18.93

Table 1.5. Performance of the strategies after transaction costs
The table shows returns after transaction costs, along with transaction costs on quintile long-short portfolios from the mispricing strategy as described in Subsection 1.1.3. The estimation methods are least squares, penalized least squares, random forests, gradient boosting regression trees, or neural networks. The regressions are rerun at the end of each January using only observations from the past and only those anomalies that have been published by that time. Hyperparameters in the estimation are also selected yearly and always based on the most recent 11-year validation sample available. The historical predictive regressions are estimated on individual stocks from the U.S. The long-short portfolios are constructed by buying stocks in the top quintile of the predicted next month returns and short-selling stocks in the bottom quintile of the predicted next month returns. The reported returns are for January 1995 to December 2018 period and are in percentage points. The standard errors in t-statistics are adjusted for heteroskedasticity and autocorrelation with Newey-West adjustment for up to 12 lags. Panel A describes transaction costs and turnover on the mispricing strategy estimated using neural networks. The transaction costs are measured either with VoV(% Spread) proxy of Fong et al. (2018), average daily closing quoted spread, or Gibbs proxy of Hasbrouck (2009). The proxies are further described in Subsection 1.1.5. The transaction costs and turnover are in percentage points per month. Panel B shows portfolio returns after transaction costs on the mispricing strategy. The transaction costs in Panel B are estimated with VoV(% Spread) proxy of Fong et al. (2018). The returns are reported in percentage points per month.

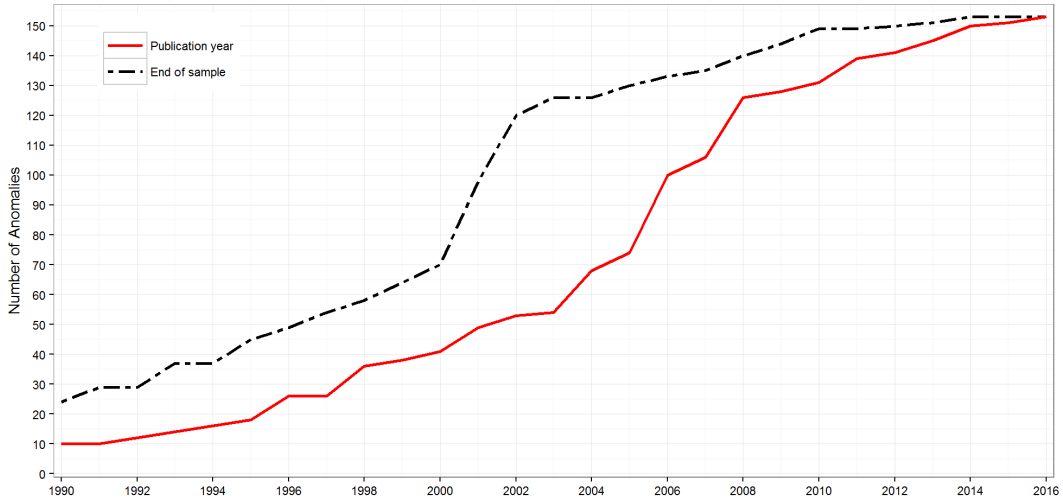


Figure 1.1. Number of the published anomalies over time
The anomalies documented in Harvey, Liu, and Zhu (2016), McLean and Pontiff (2016) and Hou, Xue, and Zhang (2018) are primarily selected. The solid line shows the cumulative number of the published anomalies over time. The dashed line shows the number of anomalies whose in-sample period in their respective studies ends before the given date.

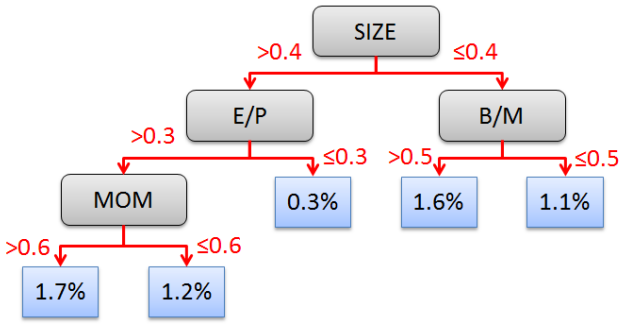


Figure 1.2. Decision tree

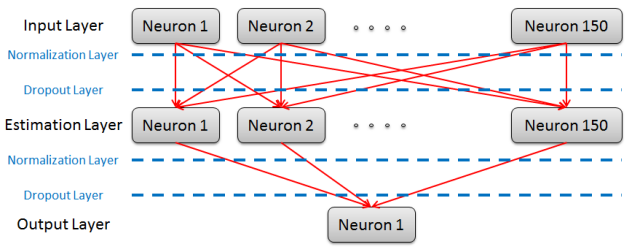


Figure 1.3. Neural network

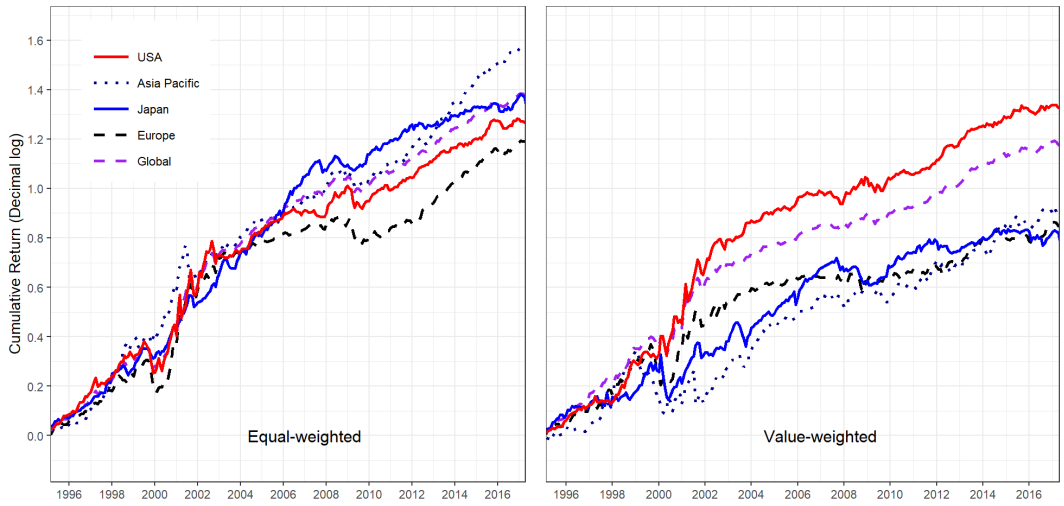


Figure 1.4. Cumulative returns on the mispricing strategy using neural networks

The figure shows cumulative returns for the mispricing strategy as shown in Panel A in Table 1.2 that is estimated on the individual stocks from the U.S. using neural networks. The returns are presented in decimal logarithms. One on the vertical scale therefore corresponds to 1000% return on the initial investment.

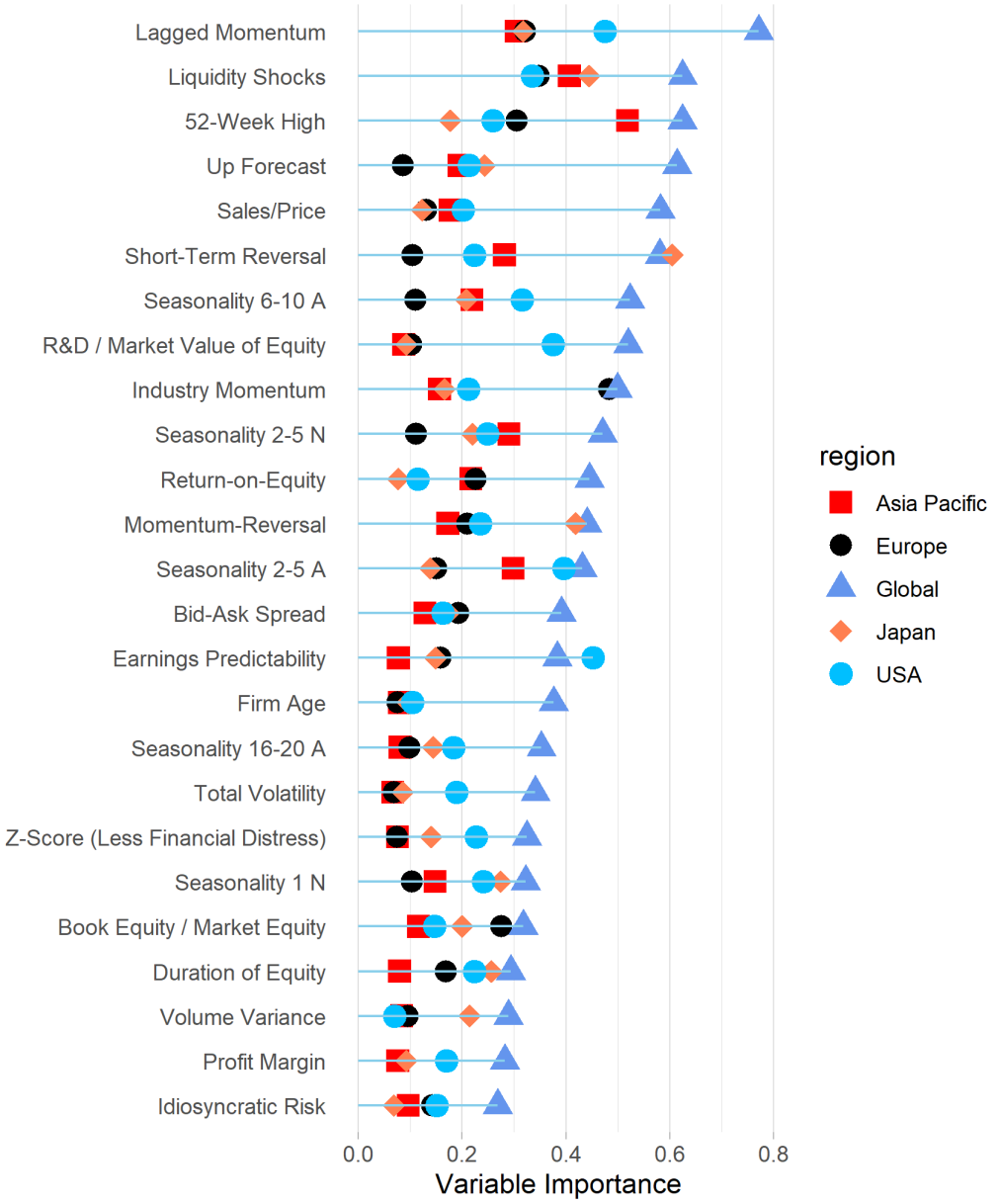


Figure 1.5. Variable importance
The figure shows variable importance as in equation (1.3) for the 25 globally most important variables in the mispricing strategy. The mispricing strategy is described in detail in Subsection 1.1.3. The strategy combines signals through predictive regressions of individual stock returns on transformed characteristics. The historical predictive regressions are estimated on data from the individual regions using neural networks. Estimation sample spans 1963 to 2018 in the U.S. and 1990 to 2018 in the other regions.

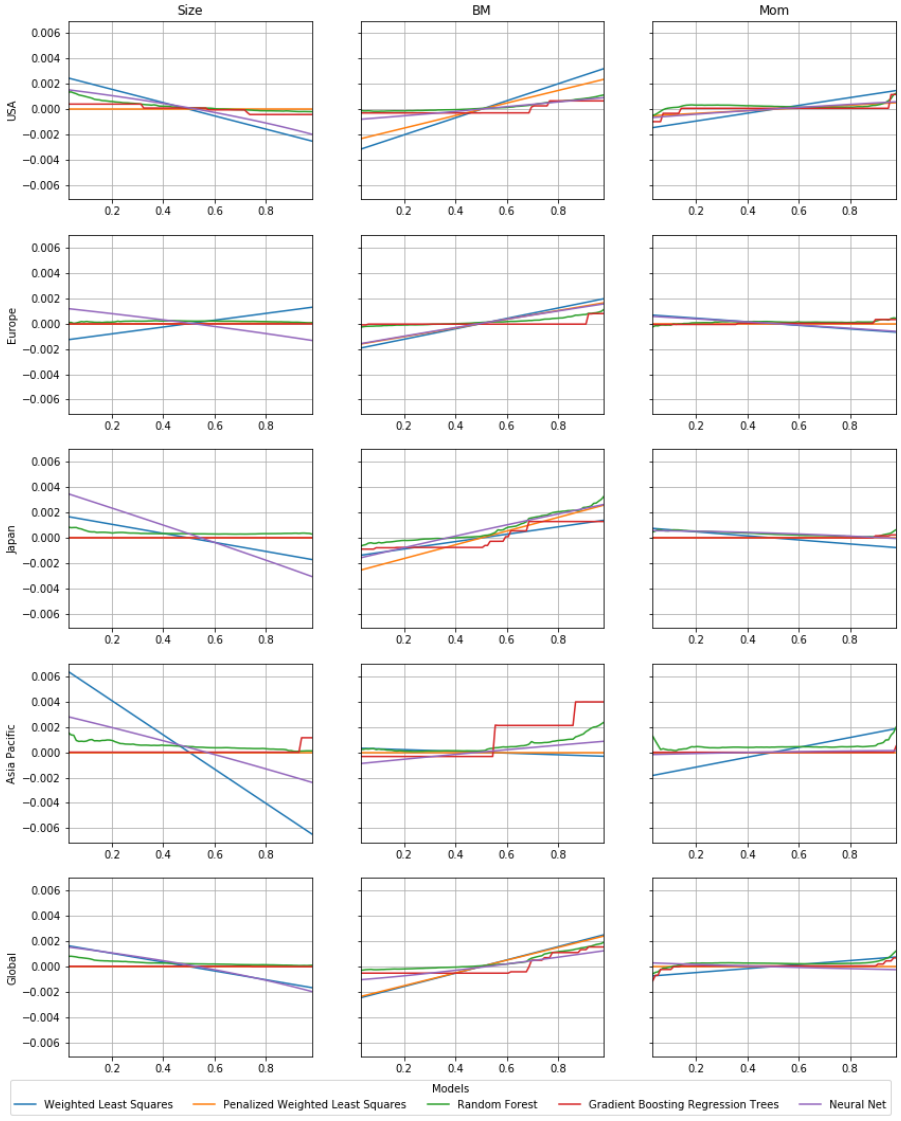


Figure 1.6. Marginal predictive contributions of Size, BM, and Mom
The figure shows the marginal predictive relationship for the mispricing strategy between size, book-to-market (BM), momentum (Mom) and (region-wise demeaned) expected returns as defined in Equation (1.4). The mispricing strategy is described in detail in Subsection 1.1.3. The methods used in the mispricing strategy are weighted least squares, penalized weighted least squares, gradient boosting regression trees, random forests, and neural networks. Predictions for individual regions are obtained using models estimated on the respective regions only. Estimation sample spans January 1963 to December 2018 in the U.S. and January 1990 to December 2018 in the other regions.

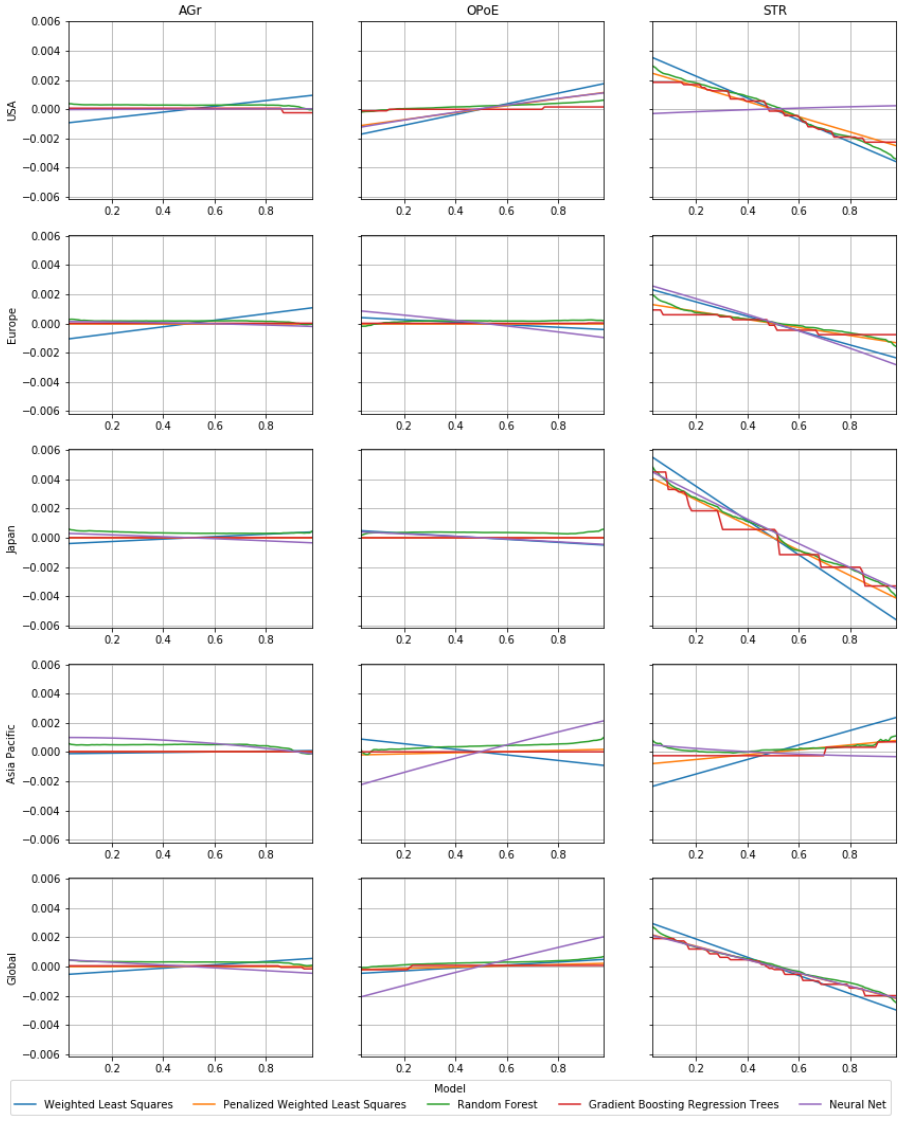


Figure 1.7. Marginal predictive contributions of AGr, OPoE, and STR
The figure shows the marginal predictive relationship for the mispricing strategy between asset growth (AGr), operating profitability (OPoE), short-term reversal (STR) and (region-wise demeaned) expected returns as defined in Equation (1.4). The mispricing strategy is described in detail in Subsection 1.1.3. The methods used in the mispricing strategy are weighted least squares, penalized weighted least squares, gradient boosting regression trees, random forests, and neural networks. Predictions for individual regions are obtained using models estimated on the respective regions only. Estimation sample spans January 1963 to December 2018 in the U.S. and January 1990 to December 2018 in the other regions.

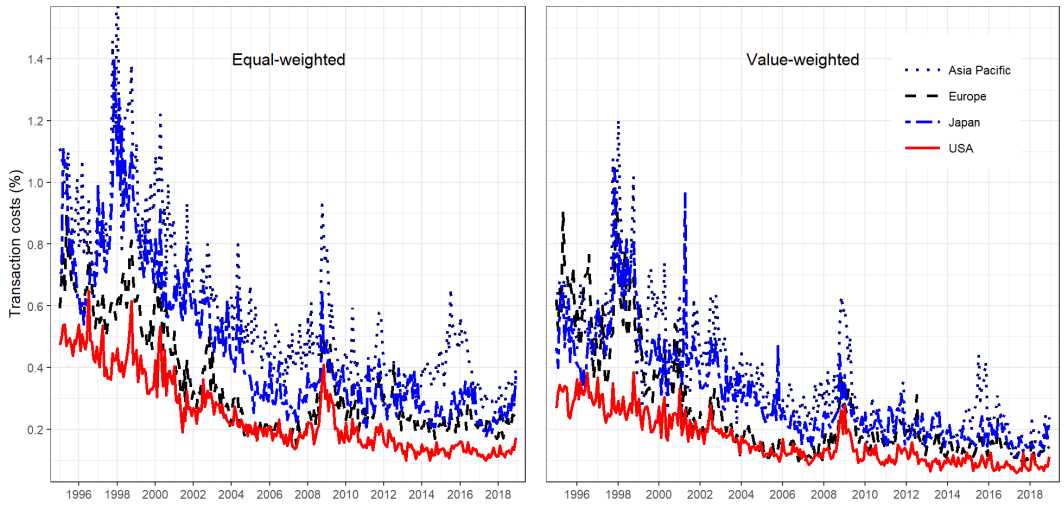


Figure 1.8. Monthly transaction costs
The figure visualizes transaction costs for the mispricing strategy using neural networks from the Panel A of Table 1.5. The transaction costs are estimated with VoV(% Spread) proxy of Fong et al. (2018).

References

Abarbanell, J. S., & Bushee, B. J. (1998). Abnormal returns to a fundamental analysis strategy. *Accounting Review*, 19–45.

Acharya, V. V., & Pedersen, L. H. (2005). Asset pricing with liquidity risk. *Journal of Financial Economics*, *77*(2), 375–410.

Alwathainani, A. M. (2009). Consistency of firms' past financial performance measures and future returns. *The British Accounting Review*, *41*(3), 184–196.

Amihud, Y. (2002). Illiquidity and stock returns: Cross-section and time-series effects. *Journal of Financial Markets*, *5*(1), 31–56.

Amihud, Y., & Mendelson, H. (1986). Asset pricing and the bid-ask spread. *Journal of Financial Economics*, *17*(2), 223–249.

Anderson, E. W., Ghysels, E., & Juergens, J. L. (2005). Do heterogeneous beliefs matter for asset pricing? *The Review of Financial Studies*, *18*(3), 875–924.

Andrew Karolyi, G. (2016). Home Bias, an Academic Puzzle. *Review of Finance*, *20*(6), 2049–2078.

Andrikopoulos, P., Clunie, J., & Siganos, A. (2013). Short-selling constraints and 'quantitative'investment strategies. *The European Journal of Finance*, *19*(1), 19–35.

Ang, A., Hodrick, R. J., Xing, Y., & Zhang, X. (2006). The cross-section of volatility and expected returns. *The Journal of Finance*, *61*(1), 259–299.

Ang, A., Chen, J., & Xing, Y. (2006). Downside risk. *The Review of Financial Studies*, *19*(4), 1191–1239.

Asness, C. S., Moskowitz, T. J., & Pedersen, L. H. (2013). Value and momentum everywhere. *The Journal of Finance*, *68*(3), 929–985.

Asparouhova, E., Bessembinder, H., & Kalcheva, I. (2010). Liquidity biases in asset pricing tests. *Journal of Financial Economics*, *96*(2), 215–237.

Avramov, D., Cheng, S., & Metzker, L. (2023). Machine learning vs. economic restrictions: Evidence from stock return predictability. *Management Science*, *69*(5), 2587–2619.

Avramov, D., Chordia, T., Jostova, G., & Philipov, A. (2013). Anomalies and financial distress. *Journal of Financial Economics*, *108*(1), 139–159.

Bali, T. G., Cakici, N., & Whitelaw, R. F. (2011). Maxing out: Stocks as lotteries and the cross-section of expected returns. *Journal of Financial Economics*, *99*(2), 427–446.

Bali, T. G., Peng, L., Shen, Y., & Tang, Y. (2013). Liquidity shocks and stock market reactions. *The Review of Financial Studies*, *27*(5), 1434–1485.

Ball, R., Gerakos, J., Linnainmaa, J. T., & Nikolaev, V. (2016). Accruals, cash flows, and operating profitability in the cross section of stock returns. *Journal of Financial Economics*, *121*(1), 28–45.

Banz, R. W. (1981). The relationship between return and market value of common stocks. *Journal of Financial Economics*, *9*(1), 3–18.

Barbee Jr, W. C., Mukherji, S., & Raines, G. A. (1996). Do sales–price and debt–equity explain stock returns better than book–market and firm size? *Financial Analysts Journal*, *52*(2), 56–60.

Barber, B., Lehavy, R., McNichols, M., & Trueman, B. (2001). Can investors profit from the prophets? security analyst recommendations and stock returns. *The Journal of Finance*, *56*(2), 531–563.

Barber, B. M., De George, E. T., Lehavy, R., & Trueman, B. (2013). The earnings announcement premium around the globe. *Journal of Financial Economics*, *108*(1), 118–138.

Barry, C. B., & Brown, S. J. (1984). Differential information and the small firm effect. *Journal of Financial Economics*, *13*(2), 283–294.

Barth, M. E., & Hutton, A. P. (2004). Analyst earnings forecast revisions and the pricing of accruals. *Review of Accounting Studies*, *9*(1), 59–96.

Bartov, E., & Kim, M. (2004). Risk, mispricing, and value investing. *Review of Quantitative Finance and Accounting*, *23*(4), 353–376.

Bartram, S. M., & Grinblatt, M. (2018). Global market inefficiencies.

Basu, S. (1977). Investment performance of common stocks in relation to their price-earnings ratios: A test of the efficient market hypothesis. *The Journal of Finance*, *32*(3), 663–682.

Belo, F., & Lin, X. (2011). The inventory growth spread. *The Review of Financial Studies*, *25*(1), 278–313.

Belo, F., Lin, X., & Bazdresch, S. (2014). Labor hiring, investment, and stock return predictability in the cross section. *Journal of Political Economy*, *122*(1), 129–177.

Bhandari, L. C. (1988). Debt/equity ratio and expected common stock returns: Empirical evidence. *The Journal of Finance*, *43*(2), 507–528.

Blitz, D., Huij, J., & Martens, M. (2011). Residual momentum. *Journal of Empirical Finance*, *18*(3), 506–521.

Blume, M. E., & Husic, F. (1973). Price, beta, and exchange listing. *The Journal of Finance*, *28*(2), 283–299.

Bondt, W. F., & Thaler, R. (1985). Does the stock market overreact? *The Journal of Finance*, *40*(3), 793–805.

Boudoukh, J., Michaely, R., Richardson, M., & Roberts, M. R. (2007). On the importance of measuring payout yield: Implications for empirical asset pricing. *The Journal of Finance*, *62*(2), 877–915.

Bradshaw, M. T., Richardson, S. A., & Sloan, R. G. (2006). The relation between corporate financing activities, analysts' forecasts and stock returns. *Journal of Accounting and Economics*, *42*(1), 53–85.

Cooper, M. J., Gulen, H., & Schill, M. J. (2008). Asset growth and the cross-section of stock returns. *The Journal of Finance*, *63*(4), 1609–1651.

Da, Z., & Warachka, M. (2011). The disparity between long-term and short-term forecasted earnings growth. *Journal of Financial Economics*, *100*(2), 424–442.

Daniel, K., & Titman, S. (2006). Market reactions to tangible and intangible information. *The Journal of Finance*, *61*(4), 1605–1643.

Datar, V. T., Naik, N. Y., & Radcliffe, R. (1998). Liquidity and stock returns: An alternative test. *Journal of Financial Markets*, *1*(2), 203–219.

Dechow, P. M., Sloan, R. G., & Soliman, M. T. (2004). Implied equity duration: A new measure of equity risk. *Review of Accounting Studies*, *9*(2-3), 197–228.

Dechow, P. M., Sloan, R. G., & Sweeney, A. P. (1995). Detecting earnings management. *The Accounting Review*, 193–225.

Diether, K. B., Malloy, C. J., & Scherbina, A. (2002). Differences of opinion and the cross section of stock returns. *The Journal of Finance*, *57*(5), 2113–2141.

Dichev, I. D. (1998). Is the risk of bankruptcy a systematic risk? *The Journal of Finance*, *53*(3), 1131–1147.

Eberhart, A. C., Maxwell, W. F., & Siddique, A. R. (2004). An examination of long-term abnormal stock returns and operating performance following r&d increases. *The Journal of Finance*, *59*(2), 623–650.

Eisfeldt, A. L., & Papanikolaou, D. (2013). Organization capital and the cross-section of expected returns. *The Journal of Finance*, *68*(4), 1365–1406.

Elgers, P. T., Lo, M. H., & Pfeiffer Jr, R. J. (2001). Delayed security price adjustments to financial analysts' forecasts of annual earnings. *The Accounting Review*, *76*(4), 613–632.

Fairfield, P. M., Whisenant, J. S., & Yohn, T. L. (2003). Accrued earnings and growth: Implications for future profitability and market mispricing. *The Accounting Review*, *78*(1), 353–371.

Fama, E. F., & French, K. R. (1992). The cross-section of expected stock returns. *The Journal of Finance*, *47*(2), 427–465.

Fama, E. F., & French, K. R. (2012). Size, value, and momentum in international stock returns. *Journal of Financial Economics*, *105*(3), 457–472.

Fama, E. F., & French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, *116*(1), 1–22.

Fama, E. F., & French, K. R. (2017). International tests of a five-factor asset pricing model. *Journal of Financial Economics*, *123*(3), 441–463.

Fama, E. F., & MacBeth, J. D. (1973). Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy*, *81*(3), 607–636.

Fong, K. Y., Holden, C. W., & Tobek, O. (2018). Are volatility over volume liquidity proxies useful for global or us research?

Francis, J., LaFond, R., Olsson, P. M., & Schipper, K. (2004). Costs of equity and earnings attributes. *The Accounting Review*, *79*(4), 967–1010.

Frankel, R., & Lee, C. M. (1998). Accounting valuation, market expectation, and cross-sectional stock returns. *Journal of Accounting and Economics*, *25*(3), 283–319.

Frazzini, A., Israel, R., & Moskowitz, T. J. (2012). Trading costs of asset pricing anomalies. *Fama-miller working paper*, 14–05.

Frazzini, A., & Pedersen, L. H. (2014). Betting against beta. *Journal of Financial Economics*, *111*(1), 1–25.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1). Springer Series in Statistics New York, NY, USA.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 1189–1232.

George, T. J., & Hwang, C.-Y. (2004). The 52-week high and momentum investing. *The Journal of Finance*, *59*(5), 2145–2176.

Green, J., Hand, J. R., & Zhang, X. F. (2017). The characteristics that provide independent information about average us monthly stock returns. *The Review of Financial Studies*.

Griffin, J. M. (2002). Are the fama and french factors global or country specific? *The Review of Financial Studies*, *15*(3), 783–803.

Griffin, J. M., Kelly, P. J., & Nardari, F. (2010). Do market efficiency measures yield correct inferences? a comparison of developed and emerging markets. *The Review of Financial Studies*, *23*(8), 3225–3277.

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical Asset Pricing via Machine Learning. *The Review of Financial Studies*, *33*(5), 2223–2273.

Gu, S., Kelly, B., & Xiu, D. (2021). Autoencoder asset pricing models. *Journal of Econometrics*, *222*(1), 429–450.

Hafzalla, N., Lundholm, R., & Matthew Van Winkle, E. (2011). Percent accruals. *The Accounting Review*, *86*(1), 209–236.

Hahn, J., & Lee, H. (2009). Financial constraints, debt capacity, and the cross-section of stock returns. *The Journal of Finance*, *64*(2), 891–921.

Harvey, C. R., Liu, Y., & Zhu, H. (2016). ... and the cross-section of expected returns. *The Review of Financial Studies*, *29*(1), 5–68.

Harvey, C. R., & Siddique, A. (2000). Conditional skewness in asset pricing tests. *The Journal of Finance*, *55*(3), 1263–1295.

Hasbrouck, J. (2009). Trading costs and returns for us equities: Estimating effective costs from daily data. *The Journal of Finance*, *64*(3), 1445–1477.

Haugen, R. A., & Baker, N. L. (1996). Commonality in the determinants of expected stock returns. *Journal of Financial Economics*, *41*(3), 401–439.

Hawkins, E. H., Chamberlin, S. C., & Daniel, W. E. (1984). Earnings expectations and security prices. *Financial Analysts Journal*, *40*(5), 24–38.

Heston, S. L., & Sadka, R. (2008). Seasonality in the cross-section of stock returns. *Journal of Financial Economics*, *87*(2), 418–445.

Hirshleifer, D., Hou, K., Teoh, S. H., & Zhang, Y. (2004). Do investors overvalue firms with bloated balance sheets? *Journal of Accounting and Economics*, *38*, 297–331.

Horel, E., & Giesecke, K. (2019). Towards explainable ai: Significance tests for neural networks. *Arxiv preprint arXiv:1902.06021*.

Hou, K., Karolyi, G. A., & Kho, B.-C. (2011). What factors drive global stock returns? *The Review of Financial Studies*, *24*(8), 2527–2574.

Hou, K., & Robinson, D. T. (2006). Industry concentration and average stock returns. *The Journal of Finance*, *61*(4), 1927–1956.

Hou, K., Xue, C., & Zhang, L. (2018). Replicating Anomalies. *The Review of Financial Studies*.

Chan, L. K., Lakonishok, J., & Sougiannis, T. (2001). The stock market valuation of research and development expenditures. *The Journal of Finance*, *56*(6), 2431–2456.

Chen, L., Pelger, M., & Zhu, J. (2024). Deep learning in asset pricing. *Management Science*, *70*(2), 714–750.

Chen, T., & He, T. (2017). Xgboost: Extreme gradient boosting.

Chordia, T., Subrahmanyam, A., & Anshuman, V. R. (2001). Trading activity and expected stock returns. *Journal of Financial Economics*, *59*(1), 3–32.

Chordia, T., Subrahmanyam, A., & Tong, Q. (2014). Have capital market anomalies attenuated in the recent era of high liquidity and trading activity? *Journal of Accounting and Economics*, *58*(1), 41–58.

Chui, A. C., Titman, S., & Wei, K. J. (2010). Individualism and momentum around the world. *The Journal of Finance*, *65*(1), 361–392.

Chung, K. H., & Zhang, H. [Hao]. (2014). A simple approximation of intraday spreads using daily data. *Journal of Financial Markets*, *17*, 94–120.

Ikenberry, D., Lakonishok, J., & Vermaelen, T. (1995). Market underreaction to open market share repurchases. *Journal of Financial Economics*, *39*(2), 181–208.

Ilmanen, A., Israel, R., Moskowitz, T., Thapar, A., & Wang, F. (2019). *Do factor premia vary over time? a century of evidence* (tech. rep.). Working Paper, AQR Capital Management.

Ince, O. S., & Porter, R. B. (2006). Individual equity return data from thomson datastream: Handle with care! *Journal of Financial Research*, *29*(4), 463–479.

Jacobs, H., & Müller, S. (2018). ... and nothing else matters? on the dimensionality and predictability of international stock returns.

Jacobs, H., & Müller, S. (2020). Anomalies across the globe: Once public, no longer existent? *Journal of Financial Economics*, *135*(1), 213–230.

Jegadeesh, N. (1990). Evidence of predictable behavior of security returns. *The Journal of Finance*, *45*(3), 881–898.

Jegadeesh, N., Kim, J., Krische, S. D., & Lee, C. (2004). Analyzing the analysts: When do recommendations add value? *The Journal of Finance*, *59*(3), 1083–1124.

Jegadeesh, N., & Titman, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of Finance*, *48*(1), 65–91.

Kelly, B., & Jiang, H. (2014). Tail risk and asset prices. *The Review of Financial Studies*, *27*(10), 2841–2871.

Kelly, B., Pruitt, S., & Su, Y. (2019). Characteristics are covariances: A unified model of risk and return. *Journal of Financial Economics*, *134*(3), 501–524.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *Arxiv preprint arXiv:1412.6980*.

Kot, H. W., & Chan, K. (2006). Can contrarian strategies improve momentum profits. *Journal of Investment Management*, *4*(1).

Kyle, A. S., & Obizhaeva, A. A. (2016). Market microstructure invariance: Empirical hypotheses. *Econometrica*, *84*(4), 1345–1404.

La Porta, R. (1996). Expectations and the cross-section of stock returns. *The Journal of Finance*, *51*(5), 1715–1742.

Lakonishok, J., Shleifer, A., & Vishny, R. W. (1994). Contrarian investment, extrapolation, and risk. *The Journal of Finance*, *49*(5), 1541–1578.

Lam, F. E. C., & Wei, K. J. (2011). Limits-to-arbitrage, investment frictions, and the asset growth anomaly. *Journal of Financial Economics*, *102*(1), 127–149.

Lee, C., & Swaminathan, B. (2000). Price momentum and trading volume. *The Journal of Finance*, *55*(5), 2017–2069.

Lee, K.-H. (2011). The world price of liquidity risk. *Journal of Financial Economics*, *99*(1), 136–161.

Lewellen, J., et al. (2015). The cross-section of expected stock returns. *Critical Finance Review*, *4*(1), 1–44.

Li, D. (2011). Financial constraints, r&d investment, and stock returns. *The Review of Financial Studies*, *24*(9), 2974–3007.

Lockwood, L., & Prombutr, W. (2010). Sustainable growth and stock returns. *Journal of Financial Research*, *33*(4), 519–538.

Loughran, T., & Wellman, J. W. (2011). New evidence on the relation between the enterprise multiple and average stock returns. *Journal of Financial and Quantitative Analysis*, *46*(6), 1629–1650.

Lu, X., Stambaugh, R. F., & Yuan, Y. (2017). *Anomalies abroad: Beyond data mining* (tech. rep.). National Bureau of Economic Research.

Lyandres, E., Sun, L., & Zhang, L. (2007). The new issues puzzle: Testing the investment-based explanation. *The Review of Financial Studies*, *21*(6), 2825–2855.

McLean, R. D., & Pontiff, J. (2016). Does academic research destroy stock return predictability? *The Journal of Finance*, *71*(1), 5–32.

McLean, R. D., Pontiff, J., & Watanabe, A. (2009). Share issuance and cross-sectional returns: International evidence. *Journal of Financial Economics*, *94*(1), 1–17.

Moskowitz, T. J., & Grinblatt, M. (1999). Do industries explain momentum? *The Journal of Finance*, *54*(4), 1249–1290.

Novy-Marx, R. (2010). Operating leverage. *Review of Finance*, *15*(1), 103–134.

Novy-Marx, R. (2012). Is momentum really momentum? *Journal of Financial Economics*, *103*(3), 429–453.

Novy-Marx, R. (2013). The other side of value: The gross profitability premium. *Journal of Financial Economics*, *108*(1), 1–28.

Novy-Marx, R., & Velikov, M. (2015). A taxonomy of anomalies and their trading costs. *The Review of Financial Studies*, *29*(1), 104–147.

Ortiz-Molina, H., & Phillips, G. M. (2014). Real asset illiquidity and the cost of capital. *Journal of Financial and Quantitative Analysis*, *49*(1), 1–32.

Palazzo, B. (2012). Cash holdings, risk, and expected returns. *Journal of Financial Economics*, *104*(1), 162–185.

Penman, S. H., Richardson, S. A., & Tuna, I. (2007). The book-to-price effect in stock returns: Accounting for leverage. *Journal of Accounting Research*, *45*(2), 427–467.

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

Pontiff, J., & Woodgate, A. (2008). Share issuance and cross-sectional returns. *The Journal of Finance*, *63*(2), 921–945.

Richardson, S. A., Sloan, R. G., Soliman, M. T., & Tuna, I. (2006). The implications of accounting distortions and growth for accruals and profitability. *The Accounting Review*, *81*(3), 713–743.

Roll, R. (1984). A simple implicit measure of the effective bid-ask spread in an efficient market. *The Journal of Finance*, *39*(4), 1127–1139.

Rouwenhorst, K. G. (1998). International momentum strategies. *The Journal of Finance*, *53*(1), 267–284.

Rouwenhorst, K. G. (1999). Local return factors and turnover in emerging stock markets. *The Journal of Finance*, *54*(4), 1439–1464.

Schwert, G. W. (2003). Anomalies and market efficiency. *Handbook of the Economics of Finance*, *1*, 939–974.

Sirignano, J., Sadhwani, A., & Giesecke, K. (2016). Deep learning for mortgage risk. *Arxiv preprint arXiv:1607.02470*.

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

Soliman, M. T. (2008). The use of dupont analysis by market participants. *The Accounting Review*, *83*(3), 823–853.

Spiess, D. K., & Affleck-Graves, J. (1995). Underperformance in long-run stock returns following seasoned equity offerings. *Journal of Financial Economics*, *38*(3), 243–267.

Stambaugh, R. F., Yu, J., & Yuan, Y. (2012). The short of it: Investor sentiment and anomalies. *Journal of Financial Economics*, *104*(2), 288–302.

Thomas, J. K., & Zhang, H. [Huai]. (2002). Inventory changes and future returns. *Review of Accounting Studies*, *7*(2), 163–187.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Titman, S., Wei, K. J., & Xie, F. (2004). Capital investments and stock returns. *Journal of Financial and Quantitative Analysis*, *39*(4), 677–700.

Titman, S., Wei, K. J., & Xie, F. (2013). Market development and the asset growth effect: International evidence. *Journal of Financial and Quantitative Analysis*, *48*(5), 1405–1432.

Tobek, O., & Hronec, M. (2018). *Does the source of fundamental data matter?* (IES Working Paper No. 15/2018). Charles University, Faculty of Social Sciences, Institute of Economic Studies. Prague.

Tuzel, S. (2010). Corporate real estate holdings and the cross-section of stock returns. *The Review of Financial Studies*, *23*(6), 2268–2302.

Watanabe, A., Xu, Y., Yao, T., & Yu, T. (2013). The asset growth effect: Insights from international equity markets. *Journal of Financial Economics*, *108*(2), 529–563.

Whited, T. M., & Wu, G. (2006). Financial constraints risk. *The Review of Financial Studies*, *19*(2), 531–559.

Zhang, X. (2006). Information uncertainty and stock returns. *The Journal of Finance*, *61*(1), 105–137.

Appendix

1.A Adjustments of returns in datastream

We apply a series of adjustments is applied on the raw returns to improve their quality. The return index (RI) is required to be larger than 0.001 on the first day of the month for precision reasons. RI is set to missing if the daily return is larger than 500% or if the price on the first day of the month is larger than $1 million. Any monthly return larger than 2000% is also set to missing. Datastream provides stale prices when there is no trade during the day or when the stock is no longer traded so that the price of the last trade is repeated until new information arrives. We therefore delete the latest observations of price with no trading. Daily returns are fixed following Tobek and Hronec (2018) when there are stale price quotes around corporate events. Monthly returns larger than 300% that revert back over the next month are set to missing following Ince and Porter (2006).[19] We winsorize 0.01% of returns in each region and years before 2000 to limit the role of outliers in returns.

[19]Specifically, returns in two consecutive months are set as missing if the return in the first month is larger than 300% and the overall return over the two months is lower than 50%.

1.B List of the anomalies

Table 1.6. List of anomalies

Fundamental	
Accruals	
Accruals	Sloan (1996)
Change in Common Equity	Richardson et al. (2006)
Change in Current Operating Assets	Richardson et al. (2006)
Change in Current Operating Liabilities	Richardson et al. (2006)
Change in Financial Liabilities	Richardson et al. (2006)
Change in Long-Term Investments	Richardson et al. (2006)
Change in Net Financial Assets	Richardson et al. (2006)
Change in Net Non-Cash Working Capital	Richardson et al. (2006)
Change in Net Non-Current Operating Assets	Richardson et al. (2006)
Change in Non-Current Operating Assets	Richardson et al. (2006)
Change in Non-Current Operating Liabilities	Richardson et al. (2006)
Change in Short-Term Investments	Richardson et al. (2006)
Discretionary Accruals	Dechow, Sloan, and Sweeney (1995)
Growth in Inventory	Thomas and Huai Zhang (2002)
Inventory Change	Thomas and Huai Zhang (2002)
Inventory Growth	Belo and Lin (2011)
M/B and Accruals	Bartov and M. Kim (2004)
Net Working Capital Changes	Soliman (2008)
Percent Operating Accrual	Hafzalla et al. (2011)
Percent Total Accrual	Hafzalla et al. (2011)
Total Accruals	Richardson et al. (2006)
Intangibles	
△ Gross Margin - △ Sales	Abarbanell and Bushee (1998)
△ Sales - △ Accounts Receivable	Abarbanell and Bushee (1998)
△ Sales - △ Inventory	Abarbanell and Bushee (1998)
△ Sales - △ SG and A	Abarbanell and Bushee (1998)
Asset Liquidity	Ortiz-Molina and Phillips (2014)
Asset Liquidity II	Ortiz-Molina and Phillips (2014)
Cash-to-assets	Palazzo (2012)
Earnings Conservatism	Francis et al. (2004)
Earnings Persistence	Francis et al. (2004)
Earnings Predictability	Francis et al. (2004)
Earnings Smoothness	Francis et al. (2004)
Earnings Timeliness	Francis et al. (2004)
Herfindahl Index	Hou and Robinson (2006)
Hiring rate	Belo, Lin, and Bazdresch (2014)
Industry Concentration Assets	Hou and Robinson (2006)
Industry Concentration Book Equity	Hou and Robinson (2006)
Industry-adjusted Organizational Capital-to-Assets	Eisfeldt and Papanikolaou (2013)
Industry-adjusted Real Estate Ratio	Tuzel (2010)
Org. Capital	Eisfeldt and Papanikolaou (2013)
RD / Market Equity	L. K. Chan et al. (2001)
RD Capital-to-assets	Li (2011)
RD Expenses-to-sales	L. K. Chan et al. (2001)
Tangibility	Hahn and H. Lee (2009)
Unexpected RD Increases	Eberhart et al. (2004)

Whited-Wu Index	Whited and Wu (2006)
Investment	
△ CAPEX - △ Industry CAPEX	Abarbanell and Bushee (1998)
Asset Growth	Cooper et al. (2008)
Change Net Operating Assets	Hirshleifer et al. (2004)
Changes in PPE and Inventory-to-Assets	Lyandres et al. (2007)
Composite Debt Issuance	Lyandres et al. (2007)
Composite Equity Issuance (5-Year)	Daniel and Titman (2006)
Debt Issuance	Spiess and Affleck-Graves (1995)
Growth in LTNOA	Fairfield et al. (2003)
Investment	Titman et al. (2004)
Net Debt Finance	Bradshaw et al. (2006)
Net Equity Finance	Bradshaw et al. (2006)
Net Operating Assets	Hirshleifer et al. (2004)
Noncurrent Operating Assets Changes	Soliman (2008)
Share Repurchases	Ikenberry et al. (1995)
Total XFIN	Bradshaw et al. (2006)
Profitability	
Asset Turnover	Soliman (2008)
Capital Turnover	Haugen and Baker (1996)
Cash-based Operating Profitability	Ball et al. (2016)
Change in Asset Turnover	Soliman (2008)
Change in Profit Margin	Soliman (2008)
Earnings / Price	Basu (1977)
Earnings Consistency	Alwathainani (2009)
F-Score	Piotroski (2000)
Gross Profitability	Novy-Marx (2013)
Labor Force Efficiency	Abarbanell and Bushee (1998)
Leverage	Bhandari (1988)
O-Score (More Financial Distress)	Dichev (1998)
Operating Profits to Assets	Ball et al. (2016)
Operating Profits to Equity	Fama and French (2015)
Profit Margin	Soliman (2008)
Return on Net Operating Assets	Soliman (2008)
Return-on-Equity	Haugen and Baker (1996)
Z-Score (Less Financial Distress)	Dichev (1998)
Value	
Assets-to-Market	Fama and French (1992)
Book Equity / Market Equity	Fama and French (1992)
Cash Flow / Market Equity	Lakonishok et al. (1994)
Duration of Equity	Dechow, Sloan, and Soliman (2004)
Enterprise Component of Book/Price	Penman et al. (2007)
Enterprise Multiple	Loughran and Wellman (2011)
Intangible Return	Daniel and Titman (2006)
Leverage Component of Book/Price	Penman et al. (2007)
Net Payout Yield	Boudoukh et al. (2007)
Operating Leverage	Novy-Marx (2010)
Payout Yield	Boudoukh et al. (2007)
Sales Growth	Lakonishok et al. (1994)
Sales/Price	Barbee Jr et al. (1996)
Sustainable Growth	Lockwood and Prombutr (2010)
Market Friction	
11-Month Residual Momentum	Blitz et al. (2011)
52-Week High	George and Hwang (2004)

Amihud's Measure (Illiquidity)	Amihud (2002)
Beta	Fama and MacBeth (1973)
Betting against Beta	Frazzini and Pedersen (2014)
Bid-Ask Spread	Amihud and Mendelson (1986)
Cash Flow Variance	Haugen and Baker (1996)
Coefficient of Variation of Share Turnover	Chordia, Subrahmanyam, and Anshuman (2001)
Coskewness	Harvey and Siddique (2000)
Downside Beta	Ang, J. Chen, and Xing (2006)
Earnings Forecast-to-Price	Elgers et al. (2001)
Firm Age	Barry and Brown (1984)
Firm Age-Momentum	X. Zhang (2006)
Idiosyncratic Risk	Ang, Hodrick, et al. (2006)
Industry Momentum	Moskowitz and Grinblatt (1999)
Lagged Momentum	Novy-Marx (2012)
Liquidity Beta 1	Acharya and Pedersen (2005)
Liquidity Beta 2	Acharya and Pedersen (2005)
Liquidity Beta 3	Acharya and Pedersen (2005)
Liquidity Beta 4	Acharya and Pedersen (2005)
Liquidity Beta 5	Acharya and Pedersen (2005)
Liquidity Shocks	Bali, Peng, et al. (2013)
Long-Term Reversal	Bondt and Thaler (1985)
Max	Bali, Cakici, and Whitelaw (2011)
Momentum	Jegadeesh and Titman (1993)
Momentum and LT Reversal	Kot and K. Chan (2006)
Momentum-Reversal	Jegadeesh and Titman (1993)
Momentum-Volume	C. Lee and Swaminathan (2000)
Price	Blume and Husic (1973)
Seasonality	Heston and Sadka (2008)
Seasonality 1 A	Heston and Sadka (2008)
Seasonality 1 N	Heston and Sadka (2008)
Seasonality 11-15 A	Heston and Sadka (2008)
Seasonality 11-15 N	Heston and Sadka (2008)
Seasonality 16-20 A	Heston and Sadka (2008)
Seasonality 16-20 N	Heston and Sadka (2008)
Seasonality 2-5 A	Heston and Sadka (2008)
Seasonality 2-5 N	Heston and Sadka (2008)
Seasonality 6-10 A	Heston and Sadka (2008)
Seasonality 6-10 N	Heston and Sadka (2008)
Share Issuance (1-Year)	Pontiff and Woodgate (2008)
Share Turnover	Datar et al. (1998)
Short-Term Reversal	Jegadeesh (1990)
Size	Banz (1981)
Tail Risk	Kelly and Jiang (2014)
Total Volatility	Ang, Hodrick, et al. (2006)
Volume / Market Value of Equity	Haugen and Baker (1996)
Volume Trend	Haugen and Baker (1996)
Volume Variance	Chordia, Subrahmanyam, and Anshuman (2001)
I/B/E/S	
Analyst Value	Frankel and C. M. Lee (1998)
Analysts Coverage	Elgers et al. (2001)
Change in Forecast + Accrual	Barth and Hutton (2004)
Change in Recommendation	Jegadeesh, J. Kim, et al. (2004)

Changes in Analyst Earnings Forecasts	Hawkins et al. (1984)
Disparity between LT and ST Earnings Growth Forecasts	Da and Warachka (2011)
Dispersion in Analyst LT Growth Forecasts	Anderson et al. (2005)
Down Forecast	B. Barber et al. (2001)
Forecast Dispersion	Diether et al. (2002)
Long-Term Growth Forecasts	La Porta (1996)
Up Forecast	B. Barber et al. (2001)

1.C Sample splitting and pool of methods

Figure 1.9 illustrates the expanding training, (fixed length) rolling validation and one year out-of-sample periods.

Table 1.7 shows list of methods considered in the main empirical analysis. The table also shows hyperparameters considered for each method.

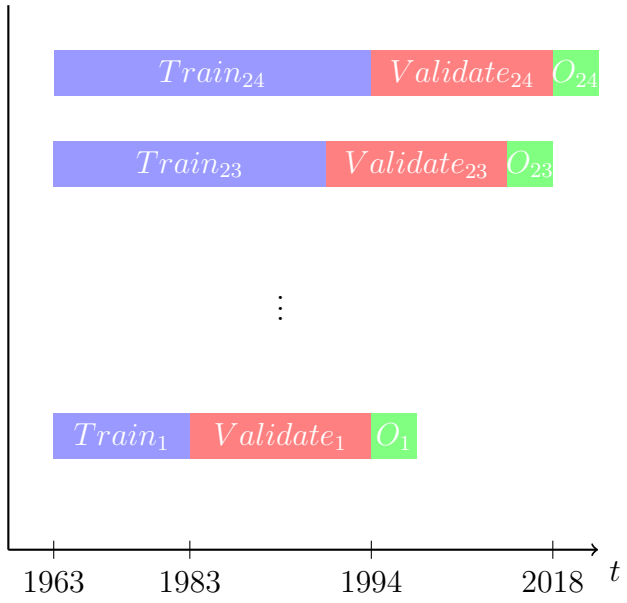


Figure 1.9. Expanding training, validation, and test setup This figure illustrates how the training set expands annually while maintaining a fixed-length validation period. Each year, the training period grows by one year while the validation period remains at 11 years. Additionally, a one-year out-of-sample test period is used. The initial training period spans 21 years (1963-1983), validation spans 11 years (1984-1994), and the test period is one year (1995). The last training period spans 44 years (1963-2006), with the final validation period from 2007 to 2017 and a one-year test period in 2018.

Method	**Hyperparameter**	**Values Considered**
WLS	-	-
Penalized WLS	α	{0, 0.1, ..., 1}
	λ	{0.001, 0.0001}
Random Forest	Number of Trees	{200, 300, ..., 600}
	Max Depth	{1, ..., 8, None}
GBRT	Number of Trees	{50, 100, ..., 500}
	Learning Rate	{1%, 2.5%, 5%, 10%}
Neural Networks	Hidden Layers	{1, 2, 3}
	Neurons per Layer	{32, 150}
	Dropout Rate	{0.001, 0.01, 0.1}

Table 1.7. List of methods and corresponding hyperparameters This table outlines the hyperparameter configurations evaluated for five distinct machine learning approaches— Weighted Least Squares (WLS), Penalized Weighted Least Squares (Penalized WLS), Random Forest, Gradient Boosted Regression Trees (GBRT), and Neural Networks used in the main empirical analysis.

1.D Optimal hyperparameters

Every time a model is trained, we perform hyperparameter optimization as described in Subsection 1.1.3. In the models we employ for the mispricing strategy, we combine signals through predictive regressions of individual stock returns on transformed characteristics. The mispricing strategy is described in detail in Subsection 1.1.3. Underlying predictive regressions are estimated using penalized weighted least squares, gradient boosted trees, random forests and neural networks. Optimal hyperparameters vary over time and their comparison for models trained and validated on the U.S. data only versus globally[20], can be seen in Figures 1.10, 1.11, 1.12, and 1.13.
Figure 1.11 shows the evolution of the selected number of trees as one of the hyperparameters for gradient-boosted trees. The selected number of trees is very similar for the U.S. only sample and the global sample.

Optimal hyperparameters in case of penalized weighted least squares and random forests, i.e. l1 and l2 mixing parameter in Figure 1.10 and the number of trees in Figure 1.12, are also similar between the U.S. only sample and the global sample. This holds strongly, especially before 2010. Looking at the architecture of the neural networks and dropout shown in Figure 1.13, selected hyperparameters differ substantially across the regions and over time.

[20]Includes the U.S. data as well.

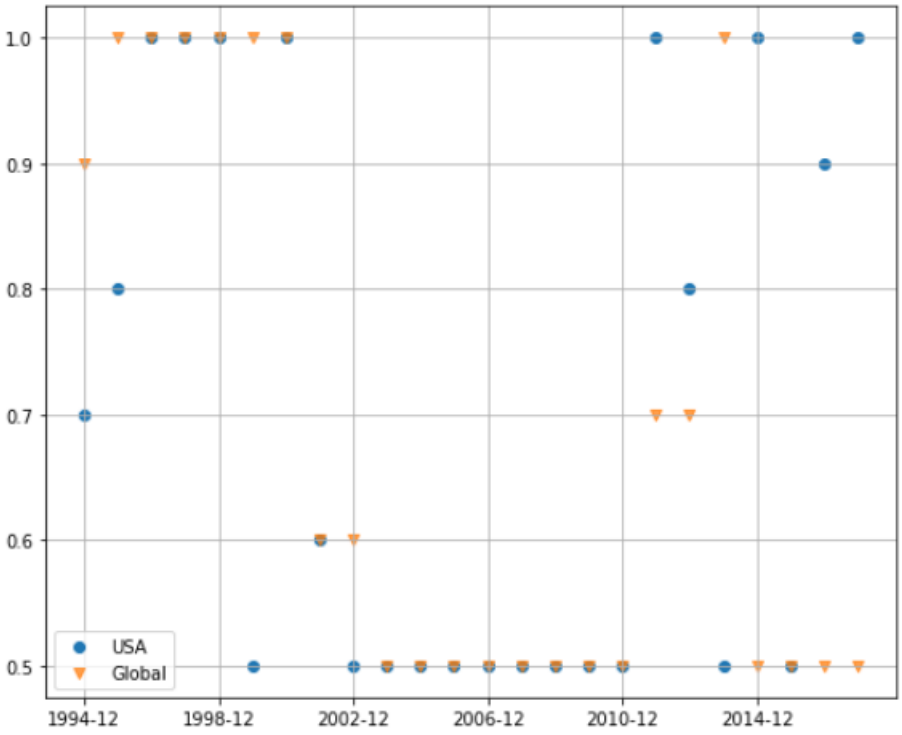


Figure 1.10. L1 and L2 mixing parameter for penalized weighted least square

The figure shows the evolution of optimal selected L1 and L2 mixing parameter as a hyperparameter for penalized weighted least squares during each tuning phase as described in Subsection 1.1.3. Penalized weighted least square are trained and cross-validated either only in the U.S. or in all the regions, i.e. globally. Training-validation-test splits are described in Subsection 1.1.3. The estimation sample is from January 1963 to December 2018 in the U.S. and from January 1990 to December 2018 elsewhere.

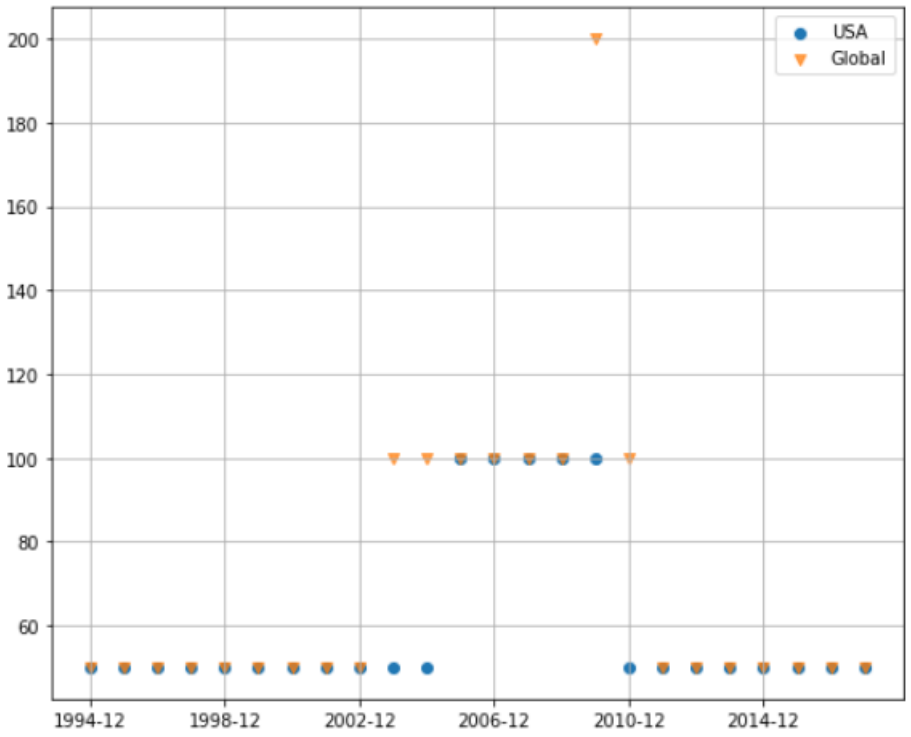


Figure 1.11. Number of trees for gradient boosting trees
The figure shows the evolution of selected optimal number of trees as a hyperparameter for gradient boosting trees during each tuning phase as described in Subsection 1.1.3. Gradient boosting trees are trained and cross-validated either only in the U.S. or in all the regions, i.e. globally. Training-validation-test splits are described in Subsection 1.1.3. The estimation sample is from January 1963 to December 2018 in the U.S. and from January 1990 to December 2018 elsewhere.

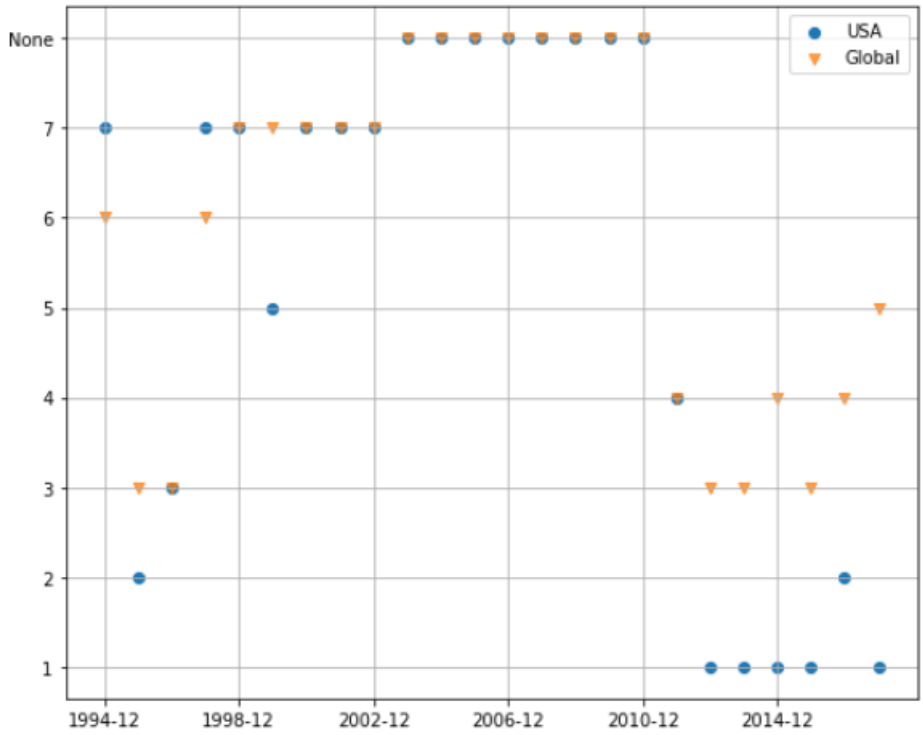


Figure 1.12. Maximum depth of the tree for random forests
The figure shows the evolution of selected optimal maximum depth of the tree as a hyperparameter for random forests during each tuning phase as described in Subsection 1.1.3. Random forests are trained and cross-validated either only in the U.S. or in all the regions, i.e. globally. Training-validation-test splits are described in Subsection 1.1.3. The estimation sample is from January 1963 to December 2018 in the U.S. and from January 1990 to December 2018 elsewhere.

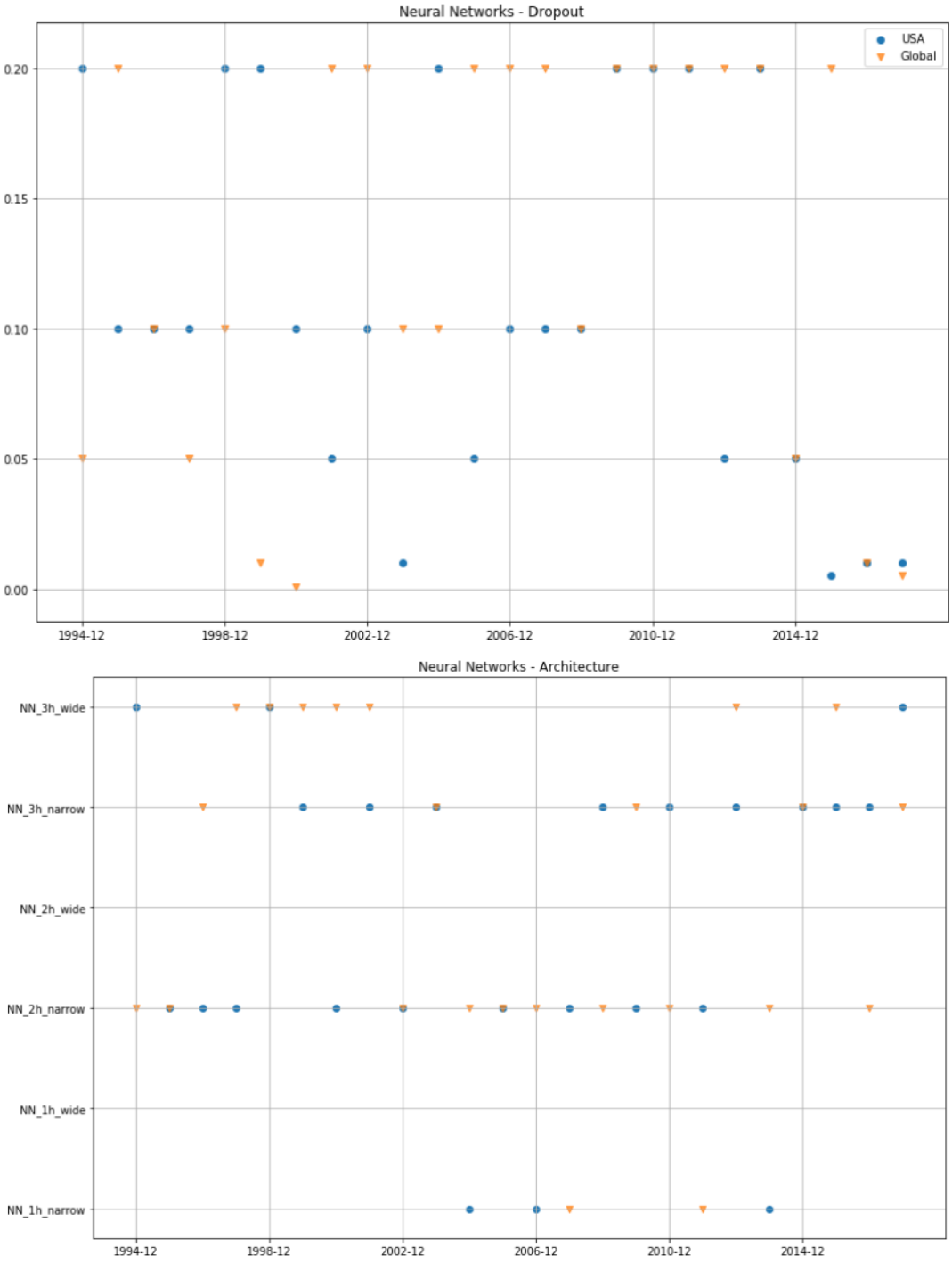


Figure 1.13. Dropout and the architecture for neural networks
The figure shows the evolution of selected optimal dropout and the architecture as hyperparameters for neural networks during each tuning phase as described in Subsection 1.1.3. Neural networks are trained and cross-validated either only in the U.S. or in all the regions, i.e. globally. Training-validation-test splits are described in Subsection 1.1.3. The estimation sample is from January 1963 to December 2018 in the U.S. and from January 1990 to December 2018 elsewhere.

1.E Definition of liquidity proxies

VoV(% Spread) Proxy

The fixed transaction costs are approximated with the VoV(% Spread) proxy introduced in Fong et al. (2018). It is defined as:

$$8 \frac{\sigma^{2/3}}{avg\ vol^{1/3}}, \tag{1.6}$$

where σ is the standard deviation of daily returns and $avg\ vol$ is the average daily trading volume in USD within a given month. The trading volume is in USD and deflated to 2000 prices. The proxy roughly measures the fixed component of the trading costs and excludes the price impact. Including the price impact would further increase the transaction costs. Fong et al. (2018) show that the price impact component is very hard to measure. It is volatile over regions, and therefore, very dependent on the execution strategy of individual asset managers. The focus is therefore solely on the fixed component of transaction costs (effective spread).

Kyle and Obizhaeva (2016) estimate a relationship between transaction costs and the size of large institutional portfolio transfers depending on the average daily trading volume and the volatility of the stocks. Their analysis is conducted on a proprietary dataset covering the 2002-2005 period. VoV(% Spread) roughly corresponds to the fixed component of their estimated transaction cost function.

Fong et al. (2018) benchmark the proxy to other proxies and find that it can be outperformed only by the closing quoted spread. The quoted spread is, however, not available for all the regions over the whole sample period.

Closing Quoted Spread

The closing quoted spread for a given month is as follows:

$$QS = \frac{1}{T} \sum_{t=1}^{T} \frac{2(ask - bid)}{ask + bid}, \tag{1.7}$$

where ask and bid are observed at the end of the trading day on each stock exchange and T is the number of days in the given month. Observations with missing or negative daily values of QS are excluded from the average. CRSP lists the best quote of bid and ask for NASDAQ stocks and the last representative quotes before the market close for NYSE and AMEX stocks. The precise definition of QS can therefore vary over the exchanges.

Chung and Hao Zhang (2014) first benchmark the QS by comparing it to high frequency effective spread estimates from the Trade and Quote (TAQ) database. They show that QS has about a 95% average cross-sectional correlation with the TAQ effective spread over the 1998 to 2009 period. Fong et al. (2018) document that it is also the best spread proxy in an international setting. One problem

with QS is that it is often missing in earlier periods and therefore has to be backfilled with other proxies.

Gibbs Proxy

Roll (1984) introduces one of the first spread proxies in the literature. He assumes that the true price of a stock follows a random walk with bid-ask jumps. That is,

$$P_t^A = P_{t-1}^A + u_t, \quad P_t^O = P_t^A + sq_t \tag{1.8}$$

$$\triangle P_t^o = s \triangle q_t + u_t,\ u_t \sim N(0, \sigma_u^2), \tag{1.9}$$

where P_t^o is the observed log price, P_t^A is the price of the underlying Brownian motion, and s is a half spread. Indicator q_t is equal to one if the last trade in the day is a buy, minus one if it is a sell, and zero if no prices are available during the day. Serial correlation of the price changes $\triangle P_t^o$ should be negative and related to the spread through the following relationship:

$$S_{roll} = 2\sqrt{-cov(\triangle P_t^o, \triangle P_{t+1}^o)}. \tag{1.10}$$

This can be contributed to the fact that

$$cov(\triangle P_t^o, \triangle P_{t+1}^o) = cov(s(q_t - q_{t-1}) + u_t, s(q_{t+1} - q_t) + u_{t+1}) = \mathbb{E}[-s^2 q_t^2] = -s^2. \tag{1.11}$$

The covariance can be positive in practice. In which case the estimate of spread is set equal to zero.

Hasbrouck (2009) proposes to extend the Roll model by estimating it with the Gibbs sampler. The idea is to estimate the equation (1.9) augmented with another dependent variable (market return) via Bayesian regression. The variables q_t are generated from the data using a Gibbs sampler.[21]

We estimate the proxy at annual frequency for each stock and calendar year. Lower frequency than annual leads to severe deterioration of the proxy's performance.

[21]Note that there is an error in the original paper in the *Journal of Finance*. The correct posterior distribution for σ_u^2 is $IG(\alpha_{prior} + \frac{n}{2}, \beta_{prior} + \frac{\sum u_t^2}{2})$.

Chapter 2

Predicting the distributions of stock returns around the globe in the era of big data and machine learning

Abstract

This paper presents a method for accurately predicting the full distribution of stock returns, given a comprehensive set of 194 stock characteristics and market variables. Such distributions, learned from rich data using a machine learning algorithm, are not constrained by restrictive model assumptions and allow the exploration of non-Gaussian, heavy-tailed data and their non-linear interactions. The method uses a two-stage quantile neural network combined with spline interpolation. The results show that the proposed approach outperforms alternative models in terms of out-of-sample losses. Furthermore, we show that the moments derived from such distributions can be useful as alternative empirical estimates in many cases, including mean estimation and forecasting. Finally, we examine the relationship between cross-sectional returns and several distributional characteristics. The results are robust to a wide range of US and international data.

This chapter was co-authored with Jozef Baruník and Ondřej Tobek and is currently R&R in the *International Journal of Forecasting*. We appreciate helpful comments from participants of the Financial Econometrics Meets Machine Learning (FinEML) conference (Lugano). I gratefully acknowledge support from the Czech Science Foundation under the 19-28231X (EXPRO) project. We appreciate helpful comments from participants at various seminars, workshops and conferences.

2.1 Introduction

Despite considerable scepticism that returns are inherently difficult to predict (Stock and Watson 2017), a substantial body of evidence has emerged over several decades showing that the first two moments of stock return distributions can be predicted to some extent using economic variables (Ang and Bekaert 2007; Ang, Hodrick, et al. 2006; McLean and Pontiff 2016). As a practical matter, mean-variance analysis provides investors with a framework for making informed choices if they accept the constraints imposed by assumptions such as multivariate normality of stock returns or a quadratic utility function. At the same time, economic agents engaged in a wide range of activities, including risk management, portfolio selection, derivatives pricing and asset pricing, have a much more complex set of preferences. They are endowed with asymmetric utilities such as disappointment aversion (Gul 1991), general preferences (Kimball 1993), quantile preferences (Manski 1988) or prospect theory (Kahneman and Tversky 1979), which require distributional information beyond the first two moments. A complete distribution forecast is necessary to fully characterise the uncertain future evolution of returns for such decision-makers. A major challenge is to develop effective forecasting techniques for a return distribution, especially given the limitations of available data.

Forecasting stock return distributions is challenging due to a number of factors, including low signal-to-noise ratio, the presence of fat tails, volatility clustering and the potentially high-dimensional nature of the relevant information sets. Worryingly, much of the existing work relies on restrictive models and assumptions, does not take into account the possible non-linear interaction of a large number of variables, and does not take into account well-known features of the data, such as non-Gaussianity and heavy tails. With rapid improvements in the accessibility and availability of large datasets and computer science algorithms, it is tempting to ask whether the problem can be significantly improved using methods that focus on learning patterns from data. In line with recent efforts to move away from exclusive reliance on models to machine learning approaches (Athey and Imbens 2019; Mullainathan and J. Spiess 2017), we propose to use machine learning to predict the full distribution of out-of-sample stock returns. We provide ample evidence that such an approach makes sense using a wide range of US and international stocks.

Our contributions can be divided into two distinct areas. First, we present a novel distribution forecasting method based on the combination of a two-stage

quantile neural network with cubic B-splines interpolation. The two-stage quantile neural network uses a substantial number of stock characteristics and market variables in two sub-networks to predict a set of τ-quantiles of returns, which are then interpolated to obtain the full distribution. This crucial step allows the derivation and use of key moments of the distribution, including mean, volatility, skewness and kurtosis. The model has been trained on US data to demonstrate out-of-sample performance from both a time perspective, using a rolling window in the US, and a cross-sectional perspective on international stocks. The performance of the proposed method is compared with that of traditional parametric models and other neural network models. The results show that the proposed approach outperforms the benchmark models in terms of both out-of-sample average quantile loss and higher out-of-sample R^2 for mean forecasts, as well as lower out-of-sample error in variance forecasts. It is crucial to highlight that utilising information from the entire distribution enhances the forecasting accuracy of the mean and variance.

Second, the paper examines how different distributional characteristics are reflected in the cross-section of stock returns. A comprehensive dataset covering both US and international stocks is used in this study, with results presented for individual regions and for the full and liquid samples. In particular, we examine the influence of individual τ-quantiles as well as mean, volatility, skewness and kurtosis derived from the predicted distribution. The profitability of long-short portfolios constructed on the basis of τ-quantile forecasts of individual stocks is analysed. It is shown that central τ-quantiles are priced into the cross-section of stock returns. There is a significant decrease in the profitability of long-short portfolios based on τ-quantiles as we move from the centre of the distribution towards the tails. We show that the mean obtained by numerical integration of the interpolated distribution forecast leads to significantly higher out-of-sample long-short portfolio returns than the mean directly predicted by the competing models. In addition, we show that no other moment studied is significantly monotonically priced in the cross-section of stock returns. This provides additional support for previously conflicting results on the role of variance, skewness and kurtosis in asset pricing.

In order to illustrate the efficacy of our methodology, we present the forecasted probability function obtained using our methods for the case of the Apple stock. Figure 2.1 illustrates the evolution of the forecasted distribution of the 22-day-ahead returns over the period 1996-2018. The distribution is obtained without any model assumptions, utilising a substantial number of characteristics.

Yet, it captures the dynamics precisely.

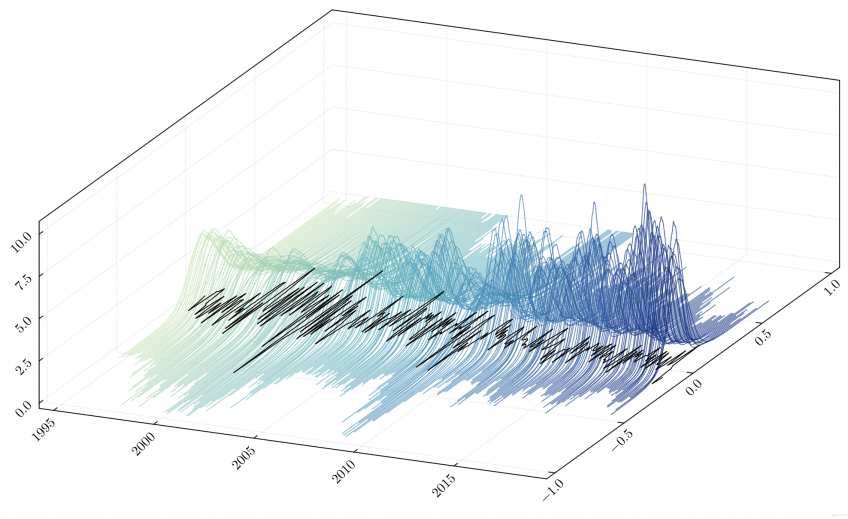


Figure 2.1. Example of forecasted probability density function over time.
This figure displays the forecasted probability density function for out-of-sample 22 days-ahead returns of an Apple stock (CUSIP 03783310) generated using our two-stage quantile neural network model.

In our methodology, we prioritise the forecasting of quantiles and subsequently approximate the distribution function rather than directly estimating a parametric distribution. In light of the stylised facts of stock returns, including the presence of fat tails, asymmetry and a domain extending from -1 to $+\infty$, it is challenging to specify a distributional form that accurately captures these characteristics. For example, the normal distribution is unable to capture the fat tails, the Student-t distribution is unable to capture the asymmetry, and the skewed Student-t distribution is unable to capture the domain restriction. Conversely, our approach approximates the distribution function in a data-driven manner, eschewing the imposition of strong parametric assumptions. Furthermore, our method is capable of estimating discrete probabilities, such as the probability mass at -1, which is of particular importance for long-horizon return forecasts. The selected methodology yields intriguing outcomes, such as the predicted discrete probability mass at 0 for certain stocks, which would be unattainable through the parametric distribution function.

The empirical results presented in this study are based on a comprehensive analysis of both U.S. and international stock data spanning the period from 1995 to 2018[1]. To ensure the robustness of our results, we focus on the full

[1]Section 2.E.6 extends the results for the U.S. to the period from 2019 to 2023.

sample as well as a restricted liquid sample. To demonstrate the generalisability of our distribution forecasting method, we train the model using only U.S. data and then apply it to international stocks and U.S. stocks in an out-of-sample manner. All requisite stock-level characteristics and market variable definitions are consistent across the United States, thereby eliminating the potential for country-specific variables (e.g., macroeconomic variables) or variables available only for a subset of stocks (e.g., options-based variables) to influence the model. This consistency allows for the transfer of predictability to the international domain.

By focusing on the aspect of predictability, we demonstrate that our distribution forecasting method exhibits a notable degree of outperformance in terms of average out-of-sample quantile loss when compared to forecasts generated by the traditional time series models as well as complex neural networks. Furthermore, the distribution forecast provides substantial benefits when forecasting the mean of stock returns. The mean forecast obtained by means of numerical integration of the interpolated distribution forecast gives rise to significantly higher out-of-sample R^2 values than the mean forecast derived directly from neural network models using mean squared error loss. With regard to volatility forecasting, the volatility forecast derived from numerical integration of the interpolated distribution forecast results in a reduction in the out-of-sample mean absolute deviation and root mean squared error in comparison to those forecasted from the GARCH model. These findings are consistent across all regions and both full and liquid samples.

In light of the distribution forecasting capability of our model, we undertake a further empirical investigation into the role of different distributional characteristics in an asset pricing application. This study examines the profitability of long-short portfolios formed based on τ-quantile forecasts of individual stocks, as well as the roles of mean, volatility, skewness, and kurtosis. Our findings indicate that, among all the quantiles, only the central quantiles are reflected in the cross-section of stock returns. Furthermore, we observe a notable decline in the profitability of long-short portfolios formed on these quantiles as we move away from the central region of the distribution towards the tails. Furthermore, we examine the profitability of long-short portfolios based on the mean versus the median forecast. Our findings indicate that the mean forecast leads to significantly higher out-of-sample long-short portfolio returns on the full sample. However, this is not the case in the liquid sample, which highlights the importance of the underlying distributional asymmetry. Finally, we demonstrate

that no other studied moment is priced in the cross-section of stock returns, providing further evidence to the previously observed conflicting results on the roles of variance, skewness, and kurtosis in asset pricing. The following section presents the results of the analysis.

2.1.1 Relation of our work to the literature

Our work relates to several strands of the literature. We contribute to stock return forecasting, which is a central topic in the asset pricing literature with a long history. In recent years, machine learning methods have become a tool of choice in this literature due to their ability to capture complex non-linear relationships, handle high-dimensional data with low signal-to-noise ratios, and provide superior sample-to-sample performance compared to traditional approaches and superior out-of-sample performance compared to more traditional approaches (Gu et al. 2020). For example, Bryzgalova et al. (2020) shows how to build better test assets by creating managed portfolios using decision trees, Dong et al. (2022) links the cross-sectional predictability of stock returns with the time-series predictability of aggregate market returns by using anomaly portfolios to predict market returns, Gu et al. (2021) uses autoencoders to build a factor asset pricing model that allows for non-linear factor loadings. Successful applications of machine learning techniques have also been demonstrated in other domains and asset classes, such as the prediction of bond returns by Bianchi et al. (2021), the ex-ante identification of good and bad performing mutual funds by Kaniel et al. (2023), the prediction of option returns by Bali, Beckmeyer, et al. (2023), or the use of convolutional neural networks to use price images to predict stock returns by J. Jiang et al. (2023). Furthermore, the robustness of many machine learning applications in stock return forecasting has been confirmed in an international context by Azevedo et al. (2023), Cakici et al. (2023) and Tobek and Hronec (2021).

The literature on forecasting the distribution of stock returns is more limited. Our work is most closely related to F. Liu (2023) in terms of forecasting the quantiles of future stock returns in the cross-section. F. Liu (2023) follows Gu et al. (2020) and compares a variety of linear models and machine learning methods in a quantile regression setting.[2] In terms of out-of-sample prediction performance, he shows that quantile neural networks significantly outperform

[2]Models compared include linear quantile regression, elastic net quantile regression, quantile random forests, quantile gradient boosted regression trees, and feed-forward neural networks.

other models. He introduces a quantile-based risk premium measure as a weighted average of five predicted quantiles, where the weights correspond to an area under an empirical step function. He also constructs a robust measure of volatility (interquartile range) and skewness from the 25%, 50% and 75% quantile forecasts, documenting that skewness is priced into the cross-section of stock returns. Our paper differs from F. Liu (2023) in several ways.

We present a novel neural network architecture for quantile forecasting and show how a list of quantile forecasts can be interpolated into a dense representation of the distribution function. Given our denser representation of the distribution, we derive moments using numerical integration, which provides a more accurate representation of the moments than directly using robust quantile-based measures. We also show how to adjust the predicted higher moments to account for the lack of extreme observations in the tails of the distribution. In addition, we provide international evidence for models trained on U.S. data only and show that our approach generalises well to international stocks. Our results are consistent with the importance of obtaining the mean prediction from the distribution[3] rather than directly using the mean squared error; however, we find no evidence that skewness is priced into the cross-section of stock returns, either in the US or internationally. We also examine the role of individual τ-quantiles in the cross-section of stock returns.

Another related paper is Yang et al. (2024), which builds on Gu et al. (2021) by using a conditional autoencoder model to approximate the distribution of future stock returns via a step function using a conditional quantile variational autoencoder (CQVAE). The authors show that the CQVAE model effectively captures the non-linear dependencies between stock returns and latent factors, leading to superior out-of-sample forecasting performance compared to traditional linear models. They also show that the mean returns estimated from the step-function approximation of the distribution function are more accurate and lead to better portfolio performance. While both our paper and Yang et al. (2024) use machine learning to predict the distribution of stock returns, there are significant differences. The CQVAE model uses latent factors learned by a variational autoencoder, while our approach uses stock characteristics and market variables directly as predictors. We go further by deriving key distributional moments such as mean, volatility, skewness and kurtosis through numerical integration, in contrast to the CQVAE model's focus on using quantile forecasts primarily for mean estimation. In addition, we validate the generalis-

[3]Even if the implementation differs.

ability and robustness of our model using both US and international stock data, demonstrating its broad applicability, whereas Yang et al. (2024) focuses solely on US stocks.

The distributions of individual stock returns are relevant to the asset pricing literature, as can be seen from the number of anomalies based on different distribution characteristics,(Hou, Xue, and Zhang 2020). However, these anomalies are typically based on historical sample estimates rather than out-of-sample forecasts, leading to significant estimation errors, especially when the persistence of tail-based distributional properties is low as in case of estimating the third and fourth moments rather than the first two moments (T.-H. Kim and White 2004). A promising alternative to distributional forecasting is the use of options to derive implied moments, (Alexiou and Rompolis 2022; An et al. 2014; T. Huang and J. Li 2019), although this method is limited by the coverage and liquidity of the options markets, especially outside the US. Instead of relying on historical sample estimates or option data, we use distributional characteristics derived from our out-of-sample distribution forecasts.

A large body of literature examines how distributional characteristics affect the cross-sectional distribution of stock returns. There is a consensus on the ambiguous relationship between realised volatility and future stock returns (Ang, Hodrick, et al. 2006; Bollerslev et al. 2020). We provide additional evidence of the absence of a clear relationship between forecast volatility and future stock returns. A similar lack of documented relationships also exists for kurtosis, although the the literature here is less extensive (Bollerslev et al. 2020).

The role of skewness, on the other hand, presents a more complex picture. Skewness is crucial in understanding individual stock returns, as Bessembinder (2018) notes that the positive skewness of individual stock returns fundamentally explains why a small fraction of stocks generate long-term stock market wealth. Theoretically, expected utility maximizing investors should require compensation for negative systematic skewness, (Harvey and Siddique 2000; Chabi-Yo et al. 2014). Alternative decision-making models lead to different implications, namely that investors may prefer positively skewed returns and thus total skewness is priced in the cross-section of stock returns, (Barberis and M. Huang 2008; Brunnermeier et al. 2007). Empirically, these relationships are also difficult to measure as results differ across time periods, frequencies, estimation methods and sample sizes. For example, (Boyer et al. 2010; Y. Zhang 2006) find a negative relationship between expected idiosyncratic skewness and return. Bali, Engle, and Murray (2016) show that the relationship between systematic skewness

and stock returns is not robust to the inclusion of other risk factors while the pricing power of idiosyncratic skewness strongly depends on the specification of the length of the estimation window and frequency of data.

On the contrary, Langlois (2020) studies differences between systematic and idiosyncratic skewness in pricing cross-section of stock returns and finds idiosyncratic skewness or total skewness is not priced in the cross-section of stock returns when controlling for other risk factors and different specifications. They document that using past, i.e. ex-post, systematic skewness leads to insignificant results. However, using predicted systematic skewness instead of past systematic skewness when forming long-short portfolio buying low predicted systematic skewness cross-sectional ranks and short-selling stocks with high predicted systematic skewness cross-sectional ranks leads to significant positive returns. Overall, the existing literature on the relationship between skewness—both systematic and idiosyncratic—and the cross-section of stock returns presents conflicting and inconclusive findings, with results highly sensitive to model specifications, sample periods, and market conditions.

In addition to asset pricing, the importance of predicting distributions is evident in risk management, where the left tail of the distribution is of particular interest. Value at Risk (VaR) is a standard risk management tool and corresponds to the τ quantile of the conditional distribution of an asset's return. Various approaches to VaR estimation, ranging from non-parametric (historical simulation) to semi-parametric to parametric, have a long history in the literature. Unlike non-parametric estimation, parametric VaR estimation requires assumptions about specific return distributions and model dynamics. The foundation of this approach is based on the GARCH volatility model of Engle (1982) and Bollerslev (1986), which is typically used together with the assumption that the residuals are distributed according to a specific distribution, typically normal or Student-t. Among semi-parametric estimation approaches, a prominent example is the CaViaR model of Engle and Manganelli (2004), which uses quantile regression to directly estimate the model dynamics without making further assumptions about the error distribution. See Nieto and Ruiz (2016) for a comprehensive review of VaR estimation methods.

While our work and the VaR estimation literature are clearly related in that both rely on accurate quantile forecasts, there are also obvious differences. First, we focus on the entire cross-section of stock returns, while the VaR estimation literature tends to work with a single asset, usually representing the entire

portfolio. This means that we can rely on a rich set of information in the form of stock-level characteristics, whereas the VaR estimation literature must rely mainly on the historical returns of the asset itself. Second, our ultimate goal is to forecast the full distribution of stock returns, not just specific 5% or 1% quantiles. As a result, we can derive any distributional risk measure, such as conditional value at risk (CVaR), skewness or kurtosis.

The rest of the paper is organised as follows. Section 2.2 introduces our distribution forecasting method, including quantile neural networks and subsequent interpolation of the quantiles into a full distribution function using splines. Section 2.3 presents the empirical results, consisting of an out-of-sample assessment of predictability and an analysis of the asset pricing implications of various distributional features.

2.2 Two-stage quantile neural networks

Let the returns $r_{i,t}$ collected over $t = 1, \ldots, T$ months and $i = 1, \ldots, N_t$ individual stocks be random variables with a distribution function[4] $F(r_{i,t})$. We are interested in forecasting a set of multiple τ quantiles $Q^{\mathcal{T}}_{r_{i,t+1}} = \left\{ Q_{r_{i,t+1}}(\tau_1|\mathcal{F}_t), \ldots, Q_{r_{i,t+1}}(\tau_K|\mathcal{F}_t) \right\}$ for $\forall \tau \in \mathcal{T} = \{\tau_1, \ldots, \tau_K\}$ where $Q_{r_{i,t}}(\tau \Big| \mathcal{F}_t) = \inf\{r_{i,t} \in \mathbb{R} : F(r_{i,t}) \geq \tau | \mathcal{F}_t\}$ of a stock return at time $t+1$ is conditional on the information set $\mathcal{F}_t$ available at time t.

Forecasting multiple τ-quantiles is possible with a multi-output feed-forward neural network model based on aggregating individual τ-quantile losses into a single loss function (as described later in this section). In addition to being computationally more efficient than training individual models for each τ, joint training for all τ quantiles allows the model to capture the relationship across the entire distribution. As a side effect, even without introducing penalisation for quantile crossing[5], the model naturally captures the monotonicity requirement of quantile predictions since there are no quantiles that are crossed in our quantile forecasts[6].

[4] In the case of the forecast distribution of stock returns, $Q_{r_{i,t}}(0) = -1$ and $Q_{r_{i,t}}(1) = \infty$.

[5] Quantile crossing refers to the situation where $Q(\tau_l) > Q(\tau_h)$ when $\tau_l < \tau_h$, violating the quantile monotonicity condition.

[6] See Appendix 2.E.5 for possible quantile crossing penalization extension.

2.2.1 Network architecture

While neural networks have been successfully used in stock return forecasting (Gu et al. 2020), quantile forecasting is challenging because we need to capture both cross-sectional variation across stocks and market-wide variation over time. One way to address this challenge is to use cross-sectional standardisation of returns to remove the time series noise. This is possible from an asset pricing perspective, where the goal is to rank individual stocks cross-sectionally to create long-short portfolios. However, since our goal is to forecast quantiles of future raw (non-standardised) stock returns, we propose a two-stage neural network architecture that exploits the predictability of standardised returns in the first stage and generates quantile forecasts of true (non-standardised) stock returns in the second stage. In this way, we allow the neural networks to exploit the information in standardised returns and to adjust for expected market volatility.

More specifically, a set of quantiles of stock returns are modelled using a large set of variables, including stock-level characteristics (e.g., size, value, or momentum) stored in a vector $x_{i,t}$, as well as market variables (e.g., market volatility) stored in a vector z_t, both of which proxy the information available at period t. We use both raw and standardised returns to address the discussion above, and we standardise the return of a stock by dividing the raw stock returns by the cross-sectional average of the stock-level volatilities at period t, such that

$$\widetilde{r}_{i,t+1} = \frac{r_{i,t+1}}{\overline{\sigma}_t}, \tag{2.1}$$

where $\overline{\sigma}_t = \frac{1}{N_t} \sum_{i=1}^{N_t} \sigma_{i,t}$ is a cross-sectional average of stock-level volatilities, $\sigma_{i,t}$ is a stock-level volatility estimate of stock i at period t, and N_t is the number of stocks in the panel at period t. Ideally, we would use $\overline{\sigma}_{t+1}$ instead of $\overline{\sigma}_{t+1}$ to rescale the standardised quantile forecasts back to the raw quantile forecasts of stock returns, but this is not known at period t when the forecast is made. Therefore, we use the cross-sectional average of stock-level volatilities at period t instead of the future value of $\overline{\sigma_{t+1}}$ to avoid look-ahead bias. $\sigma_{i,t}$ is estimated using an exponentially moving weighted average of the squared returns of stock i with a decay factor of 0.94 and is one of the input variables to the model.[7] See Appendix 2.C for a visualisation of $\overline{\sigma}$ over time in the US.

The idea behind standardising individual stock returns using the cross-

[7]The 0.94 parameter is widely used in both academia and the financial industry, an example being the RiskMetrics GARCH specification.

sectional average of stock-level volatilities rather than the stock volatility itself is to reduce the noise from individual stock volatility estimates. For the same noise-reducing reason, we standardise stock returns only by dividing by the average volatility and and not by subtracting the average return, either stock or market, because average returns are not persistent.

A stock return quantile is then modelled by the two-stage quantile neural network with two sub-networks as

$$\begin{aligned} Q^{\mathcal{T}}_{r_{i,t+1}} &= \left\{ Q_{r_{i,t+1}}(\tau_1|x_t, z_t, \overline{\sigma}_t), \ldots, Q_{r_{i,t+1}}(\tau_K|x_t, z_t, \overline{\sigma}_t) \right\} \\ &= \left[\Big(\underbrace{g^{(L)}_{W^{(L)}_{S_1}, b^{(L)}_{S_1}} \circ \ldots \circ g^{(1)}_{W^{(1)}_{S_1}, b^{(1)}_{S_1}}(x_t)}_{\text{standardized } \tau \text{ quantiles sub-network}} \Big) \times \overline{\sigma}_t \right] \times \Big(\underbrace{g^{(2)}_{W^{(2)}_{S_2}, b^{(2)}_{S_2}} \circ g^{(1)}_{W^{(1)}_{S_2}, b^{(1)}_{S_2}}(z_t)}_{\text{market volatility sub-network}} \Big) \\ &= \underbrace{\left[\Big\{ \underbrace{Q_{\widetilde{r}_{i,t+1}}(\tau_1|x_t), \ldots, Q_{\widetilde{r}_{i,t+1}}(\tau_K|x_t)}_{\text{Stage I}} \Big\} \times \overline{\sigma}_t \right] \times \widehat{\sigma}^M_t}_{\text{Stage II}}, \end{aligned} \tag{2.2}$$

where $W_S = \left(W^{(1)}_S, \ldots, W^{(L)}_S\right)$ and $b_S = \left(b^{(1)}_S, \ldots, b^{(L)}_S\right)$ are weight matrices and bias vectors from a stage S. Each weight matrix $W^{(\ell)}_S \in \mathbb{R}^{m \times n}$ contains m neurons as n column vectors $W^{(\ell)}_S = [w^{(\ell)}_{\cdot,1}, \ldots, w^{(\ell)}_{\cdot,n}]$, and $b^{(\ell)}_S$ are thresholds or activation levels that contribute to the output of a hidden layer that allows the function to be shifted.

It is important to note that, in contrast to the literature, we consider a multi-output (deep) neural network to characterise the collection of quantiles. Our network has two subnetworks, namely a standardised τ quantile network and a market volatility network in which $l \in 1, ..., L$ hidden layers transform the input data into a chain using a collection of nonlinear activation functions $g^{(1)}_S, \ldots, g^{(L)}_S$. An activation function, $g^{(\ell)}_{W^{(\ell)}_S, b^{(\ell)}_S}$, is used as a

$$g^{(\ell)}_{W^{(\ell)}_S, b^{(\ell)}_S}(u) := g_\ell\left(W^{(\ell)}_S u + b^{(\ell)}_S\right) = g_\ell\left(\sum_{i=1}^{m} W^{(\ell)}_{i,S} u + b^{(\ell)}_{i,S}\right)$$

is rectified linear units $g_\ell(u) = \max\{u, 0\}$, or $g_{\ell,S}(u) = \tanh(u)$.

Figure 2.2 details the architecture of the two-stage quantile neural network model defined above. This is our main model used for quantile forecasting, and we refer to it as the *two-stage model* throughout the paper and use the acronym 2sNN in tables with results. The first stage of the model is used to forecast quantiles of cross-sectionally standardised next month stock returns $Q^{\mathcal{T}}_{\widetilde{r}_{i,t}}$. We

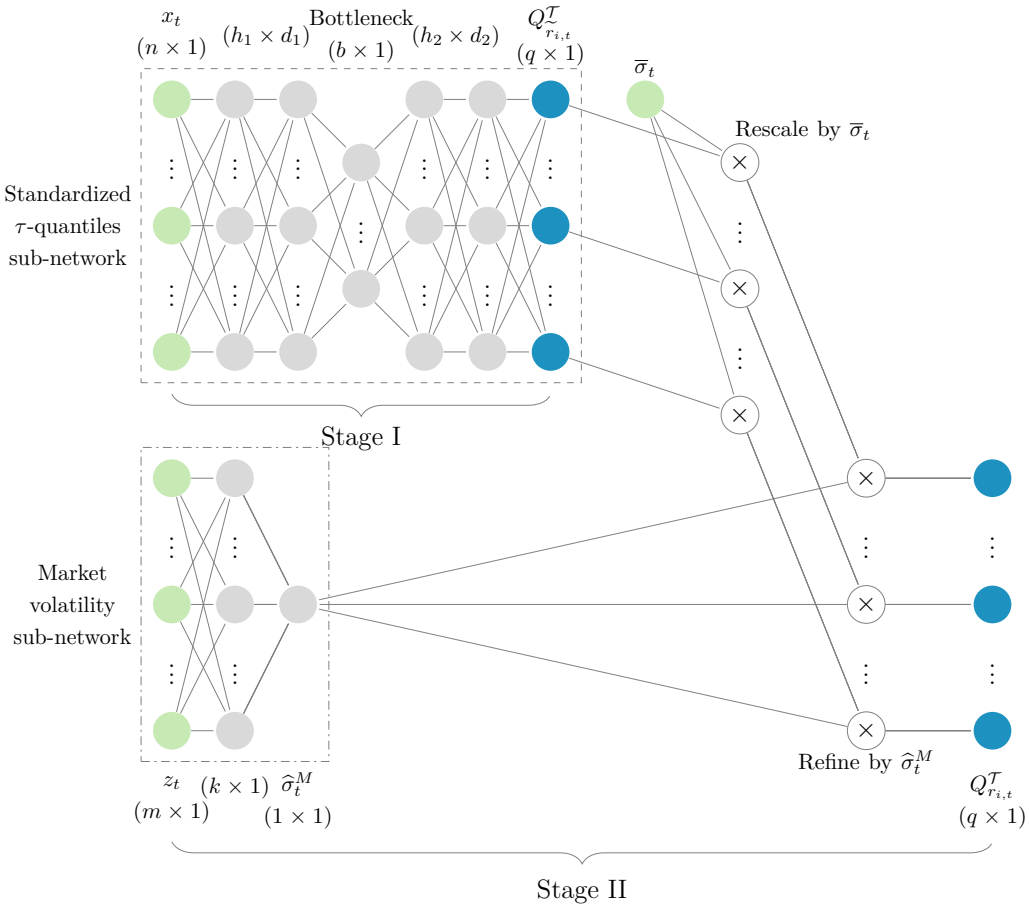


Figure 2.2. Two-stage neural network (2sNN) architecture
Stage I generates cross-sectionally standardised τ-quantile forecasts, $Q^{\mathcal{T}}_{\widetilde{r}}$, as output. Cross-sectional standardisation is performed by dividing individual future stock returns $r_{i,t+1}$ (labels) by the cross-sectional average of the individual stock standard deviation estimates, $\overline{\sigma}_{t+1}$. The standard deviation estimate is an exponential weighted moving average of volatility with a smoothing factor of 0.94. Stage II uses $\overline{\sigma}_t$, i.e. based on data from the previous period, to rescale the standardised τ-quantile forecasts generated in Stage I back to the original scale. The rescaling takes place at the multiplication nodes marked with $\times$, representing the multiplication operation between the inputs of the node. The market volatility part then generates forecasts of the next period's market volatility in order to further refine the τ-quantile forecasts produced as the output of Stage II.
The architecture uses general notations where n is the number of input features in x_t, m is the number of input features in z_t, h_1 and d_1 represent dimensions of the first set of hidden layers, h_2 and d_2 are dimensions of the second set of hidden layers, b is the bottleneck size, k is the hidden layer size in market volatility sub-network, and q is the number of quantiles predicted. The network is optimized using the quantile loss function L_τ. The forward pass generates a tensor of form $(B, q, 2)$, where B is the number of observations in the batch, q is the number of predicted quantiles, and 2 corresponds to the standardised and raw versions. Following values were used in the empirical part: $n = 176$ (input features), $m = 18$ (market features), $h_1 = h_2 = 128$ and $d_1 = d_2 = 2$ (hidden layers), $b = 4$ (bottleneck size), $k = 8$ (market volatility hidden layer), and $q = 37$ (number of quantiles).

use 37 τs in the set $\mathcal{T}$ ranging from the extreme left tail of the distribution (0.00005) to the extreme right tail of the distribution (0.99995). The second stage of the model generates quantile forecasts of the next month's raw stock returns $Q_{r_{i,t}}(\tau)$ for $\tau \in \mathcal{T}$. This stage consists of three steps. First, the standardised quantile forecasts $Q^{\mathcal{T}}_{r_{i,t}}$ from the first stage and the cross-sectional average of stock-level volatilities $\overline{\sigma}_t$ are used as inputs to rescale the standardised quantile forecasts back to the raw stock return scale $Q_{r_{i,t}}(\tau) = Q_{\tilde{r}_{i,t}}(\tau) \cdot \overline{\sigma}_t$. Second, the market volatility subnetwork generates a second scaling factor to reintroduce the time-series market-wide volatility component. This market volatility scaling factor is the same for all stocks in a given period t and is used to refine the quantile forecasts rescaled in the first step. Finally, both the standardised and raw quantile forecasts are clipped at -100% during the forward pass of the neural network. Individual steps performed during the two-stage quantile neural network forecast can be found in Algorithm 3.

In addition to the use of two stages, another key difference from standard feedforward neural network architecture is the inclusion of a bottleneck layer. Inspired by the autoencoder architecture, the bottleneck layer aims to capture the distribution hyperparameters. This is achieved by forcing the network to learn a compressed representation, as the bottleneck layer is limited to only four nodes.

The two-stage neural network takes three sets of inputs. The first set consists of 176 features representing stock-level characteristics based on 153 anomalies used in Tobek and Hronec (2021), 18 stock-level volatility estimates, and five cross-sectional averages of stock-level mean estimates. The second set of inputs consists of cross-sectional averages of the stock-level volatility characteristics used in the first set, i.e. 18 stock-level volatility estimates. The third input is a cross-sectional average of stock-level volatilities $\overline{\sigma}_t$ used for rescaling purposes. A detailed description of all features can be found in Section 2.3.1.

2.2.2 Training with quantile loss function

In order to jointly predict multiple τ-quantiles using a two-stage neural network, τ-quantile losses for individual τs must be aggregated into a single loss. The two-stage quantile neural network model also requires both standardised and raw stock returns as labels, so the aggregated quantile loss must take this into account. The loss function used during training is defined as a linear combination of the aggregated quantile losses from the first and second stage

outputs, as shown in Equation 2.3. Both coefficients in the linear combination are equal, implying equal weighting of both stages in the loss function.[8] The aggregated multi-τ-quantile loss used in the estimation is defined as

$$\mathcal{L}^{\mathcal{T}} = \frac{1}{B}\frac{1}{K}\sum_{\tau\in\mathcal{T}}\sum_{i=1}^{B}\left(\rho_\tau\left(r_{i,t} - \widehat{Q}_{r_{i,t}}(\tau)\right) + \rho_\tau\left(\widetilde{r}_{i,t} - \widehat{Q}_{\widetilde{r}_{i,t}}(\tau)\right)\right) \tag{2.3}$$

where $\widehat{Q}_{r_{i,t}}(\tau)$ is the forecast τ-quantile of the raw return of stock i at period t, $\widehat{Q}_{\widetilde{r}_{i,t}}(\tau)$ is the predicted τ-quantile of the standardised return of stock i at period t, B is the lot size and $K = 37$ is the number of τ in the $\mathcal{T}$ set. An essential part of the loss is ρ_τ, a quantile loss function defined as a piecewise linear function $\rho_\tau(\xi) = \mathbb{1}_{(\xi\geq 0)}\tau\xi + \mathbb{1}_{(\xi<0)}(\tau-1)\xi$ where $\xi = y - \widehat{Q}_y(\tau)$ is an error term between the true value and the predicted τ quantile, and $\mathbb{1}_{(.)}$ is an indicator function. A single τ quantile loss calculated for multiple observations is then calculated as a simple arithmetic mean of the individual ρ_τ losses.

The idea behind the quantile loss function is to penalise overestimation of the τ quantile by τ times the error if the error is positive and by $(\tau-1)$ times the error if the error is negative. For example, if $\tau = 0.5$, the quantile loss function is the absolute value loss function, and overprediction and underprediction are penalised equally. If $\tau = 0.05$, underprediction is penalised 19-times (0.95/0.05) more than overprediction, and minimising such a loss function leads to the estimation of the 0.05 quantile.

The multi-quantile loss function is a generalization of the standard quantile loss, allowing for the simultaneous estimation of multiple quantiles within a unified optimization framework. The aggregated loss function is defined as

$$\mathcal{L}^{\mathcal{T}}_{multi} = \frac{1}{BK}\sum_{\tau\in\mathcal{T}}\sum_{i=1}^{B}\rho_{\tau_k}\left(r_{i,t} - \widehat{Q}_{r_{i,t}}(\tau_k)\right), \tag{2.4}$$

This formulation extends classical single-quantile estimation (Koenker 2004; Y. Li and Zhu 2008) to a multi-quantile framework that leverages parameter sharing across quantile levels and implies an equal weighting of individual τ-quantile losses. In our case, this means that large stocks have the same impact as small stocks and central quantiles have the same impact as tail quantiles.

[8]This specification worked out of the box, and the hyperparameter was therefore not calibrated. In practice, the first stage has a much larger impact on the loss function because it focuses on rescaled raw returns with a scale close to one, while the second stage focuses on raw returns with a scale of about 0.1. The greater implicit focus on the first stage is desirable, as this is where most of the learning takes place, with the second stage only fine-tuning the results of the first stage.

$\mathcal{L}^{\mathcal{T}}_{multi}$ is used in training alternative multioutput quantile neural networks without two-stage architecture, specifically LNN, 1hNN and 2hNN[9]. Beyond computational efficiency, i.e. training K times fewer parameters, solving a multi-quantile loss function jointly rather than optimizing separate single-quantile losses offers several advantages. Unlike independent single-quantile models, multi-quantile optimization enforces a shared latent representation across all quantiles. This disincentivizes the overfitting of individual quantile estimates while enabling information sharing across different quantile levels. This offers an implicit regularization effect induced by shared feature extraction layers. In the single-quantile case, each model optimizes its loss independently, leading to no explicit relationship between estimated quantiles. This can lead to a violation of the monotonicity condition

$$\widehat{Q}(\tau_1 \mid X) \leq \widehat{Q}(\tau_2 \mid X) \quad \text{for all } \tau_1 < \tau_2. \tag{2.5}$$

also known as the quantile crossing problem, where $\widehat{Q}(\tau_k \mid X) > \widehat{Q}(\tau_{k+1} \mid X)$ for some $\tau_k < \tau_{k+1}$. In multi-quantile optimization, joint learning across all quantiles reduces the risk of quantile crossing because the network is exposed to the entire quantile structure simultaneously. The gradients used in optimization are computed simultaneously for all quantiles in each update step. This leads to gradient averaging, resulting in a smoother optimization landscape and improved convergence properties. Empirically, we observe that our model preserves monotonicity without explicit constraints due to this implicit ordering effect. However, when training single-quantile models separately, there is no mechanism ensuring that lower quantiles do not exceed higher quantiles. This frequently results in quantile crossing, necessitating post-processing adjustments or explicit penalization methods, as discussed in (Meinshausen 2006; Takeuchi et al. 2006). In case of encountering quantile crossing, the model can be easily extended to include the penalty term in the loss function to enforce the monotonicity condition as described in Appendix 2.E.5.

Nonetheless, extending the classical quantile regression framework to a nonparametric deep learning framework brings challenges beyond computational requirements such as ambiguity about model specification, data requirements, and more complicated interpretability. Further, multi-quantile optimization assumes that all quantiles can be modelled using a shared feature representation. This assumption holds in many applications but may be limiting, if different

[9]See Appendix 2.E for detailed specifications.

quantile levels require distinct transformations. In this case single-quantile models provide greater flexibility to learn quantile-specific patterns, but at the cost of reduced efficiency and potential inconsistencies in quantile relationships.

From a numerical point of view, the quantile loss function is a piecewise linear function that is differentiable everywhere except at zero. Gradient-based optimisation algorithms require the loss function to be differentiable in order to update the model parameters. Automatic differentiation engines, such as PyTorch's autograd (Paszke et al. 2019), are able to deal with this problem by defining subgradients for non-differentiable points. A subgradient at a point is any slope of a line that lies below the function at that point and intersects the function at that point. For the quantile loss function, the subgradient at zero can be any value between the left derivative $(\tau - 1)$ and the right derivative (τ), depending on the specific implementation of the quantile loss function.

To train the model, we minimise the loss function using stochastic gradient descent with mini-batches of size $B = 8192$ observations. Specifically, we use the Adam optimiser (Kingma and Ba 2014) with an initial learning rate of 0.0003 and $(\beta_1, \beta_2) = (0.9, 0.999)$. We adjust the number of epochs based on the number of observations available at each training period. The number of epochs is set to $100 * (A/n)$, where n is the number of observations available for training and A is the adjustment constant equal to 3000000 for the full sample and 1500000 for the fluid sample, derived from the average number of observations in each sample over time.[10] We use early stopping with a patience of 2 epochs, where 20% of the training data is used as validation to determine the stopping point. After early stopping, the model state with the lowest validation loss is taken as the final model. We use Leaky Relu (Maas et al. (2013)) as the activation function in all layers except the last layer in each subnetwork. We use batch normalisation (Ioffe and Szegedy 2015) in all layers except the last layer in both the standardised τ-quantiles subnetwork and the market volatility subnetwork. We apply dropout to all layers except the last layer in each subnetwork and use a rate of 0.2 as a result of the hyperparameter search described in Section 2.D. We also apply an L1 penalty to the weights of the first stage in each subnet. The L1 penalty for the first layer is 0.0001, and for the second layer, it is 0.00001. The first layer of the market volatility network is also regularised by an L2 penalty of 0.00001. As an additional form of regularisation, we use an ensemble of 20 networks.

The model is trained from scratch for the first training period and then

[10] See Subsection 2.3.1 for the description of the data used.

fine-tuned for subsequent training periods[11]. Fine-tuning greatly reduces the computational burden of having to re-estimate the model each period and is a natural choice given the large overlap of subsequent training samples with the widening estimation window approach. It is expected that one year of new data should not change the estimated parameters too much unless there is a novel high-information event such as the dot-com bubble or the global financial crisis, in which case fine-tuning requires more epochs to retrain.

2.2.3 Approximating probability density and its moments from quantiles using B-Splines

To approximate the probability density, let's consider a grid of τ quantile forecasts $\{Q_j(\tau), \tau_j\}_{j=1}^{K}$, where τ_j is the probability and $Q_j(\tau)$ is the corresponding quantile value, containing K grid points that can be used to obtain the cumulative distribution function $F : \mathbb{R} \to [0, 1]$. Since stock returns range from -1 to ∞, from a practical perspective we are interested in $F : [-1, \infty) \to [0, 1]$. To get a denser representation, we use cubic B-splines[12] to interpolate between adjacent quantiles by constructing piecewise polynomial functions within each interval $[Q_j, Q_{j+1}]$. Instead of interpolating the predicted quantiles directly, we create a new denser $\mathcal{T}^d$ grid with one hundred equally spaced points between every two original quantile prediction points, except the largest and smallest, i.e. the denser grid has $(K - 3) \times 100$ values, and fit the cubic B-spline to approximate the cumulative distribution function. To derive the probability density function, the derivative of the spline for the base function is calculated at each interval to obtain the new spline function representing the probability density function.

Figure 2.3 shows an example of the interpolated cumulative distribution function on the left and the implied probability density function on the right. The figure is based on τ-quantile forecasts of 22-day (monthly) returns for Microsoft in October 2008. The density function is relatively close to a Student's t-distribution with noticeably large tails during the depths of the Great Financial Crisis.

[11] Initial weights are taken from the previous training period.

[12] We use the splrep and BSpline functions from the Python scipy library, Virtanen et al. (2020).

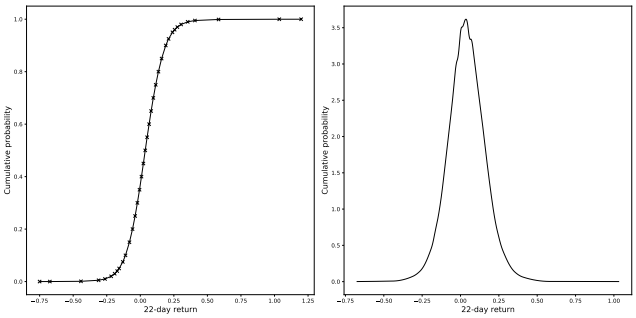


Figure 2.3. Interpolated cumulative distribution function and probability density function
This figure shows the cumulative distribution function interpolated from the inverted quantile forecasts depicted as marks (left) and derived probability density function (right) for out-of-sample 22 days-ahead returns of Microsoft stock (CUSIP 59491810) in October 2008. Quantile forecasts used to obtain the cumulative distribution function are generated using the two-stage quantile neural network model.

As the interpolated density function does not cover the extreme tails of the distribution beyond the smallest and largest predicted quantile, these tails are modelled by discrete probabilities for values at the ends of the new $\mathcal{T}^d$ grid. We use discrete probabilities implied by the estimated spline function. The density function is truncated for returns less than -1, and discrete probability is then considered for -1 returns. Excluding the extreme tails from the density function results in a slight underestimation of the variance and a more pronounced underestimation of the kurtosis of the distribution. Inferring the distribution in the tails from empirical data is problematic, so we chose this solution to avoid the problem to some extent. In cases where B-spline interpolation does not produce plausible results, we use a linear B-spline function as a fallback (see Appendix 2.E.1 for details). A complete algorithm for approximating the probability density function from the quantile forecasts is described in detail in Appendix 2.E, Algorithm 1.

Importantly, the B-spline approximation allows us to obtain moments of the probability distribution function. We first derive the zth non-central moment of a density function $f(x)$

$$m_z = \int_{-\infty}^{\infty} x^z f(x)\, dx \tag{2.6}$$

using numerical integration methods and the B-spline-based density function approximation obtained above. We use a simple tractable local linear approximation algorithm, which leads to trivial integrals of polynomial functions, and the complete algorithm is detailed in Appendix 2.E, Algorithm 2. Further, using

this approximation, we calculate central moments and rescale the moments by the integrated density function to ensure they are properly defined.

The estimated empirical quantiles from neural networks do not cover the full range of possible stock returns because it is not possible to estimate quantiles in the extreme tails of the distribution. The calculated central moments are, therefore, biased, with the bias being greater for higher moments as the tails of the distribution have a greater impact. It is, therefore, important to test how well the algorithm captures the true moments. To do this, we compare the true moments of a theoretical distribution with the output of the algorithm and propose an adjustment to reduce the measured bias. The appendix Subsection 2.E.3 describes how to adjust the moments to account for bias.

2.3 Empirical results

2.3.1 Data and features

Our data cover U.S. and international equity markets for the periods 1963 to 2018 and 1990 to 2018, respectively. We use the merged CRSP/Compustat database from the Wharton Research Data Service (WRDS) for the US sample and LSEG Datastream[13] for the international sample. Data are pre-processed using the methodology described in Tobek and Hronec (2021). There are four regions in the sample - US, Europe, Japan and Asia-Pacific, which also includes 23 developed countries.[14] We report our results for both the *full sample* and the *liquid sample*, which includes only the most liquid stocks. The liquid sample includes stocks that meet the following criteria: a price greater than 1 (or 0.1 for Asia Pacific) at the end of the previous month, a market capitalisation within the top 95%, and a trading volume within the top 95%, and a trading volume within the top 95% of the total dollar trading volume over the previous 12 months in each region. The full sample includes stocks with a price greater than 1 ($0.1 for Asia Pacific) at the end of the previous month. Market capitalisation and volume filters are not applied to the full sample. Micro-cap stocks in the full sample therefore account for only a small fraction of the capitalisation and capitalisation and trading volume of the total market. The average, minimum

[13]Formerly called Reuters or Refinitiv Datastream.

[14]US, Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, Netherlands, Norway, Portugal, Spain, Sweden, Switzerland, UK, Japan, Australia, New Zealand, Hong Kong and Singapore.

and maximum number of stocks in the cross-section in each region is shown in table Table 2.3.1.

Table 2.1. Number of stocks in the cross-section across the regions
This table shows the average, minimum and maximum number of stocks in the cross-section in each region for the full and liquid samples. The period covered is from 1973 to 2018 for the US sample and from 1990 to 2018 for the international sample.

	Full Sample				Liquid Sample			
region	USA	Europe	Japan	Asia Pacific	USA	Europe	Japan	Asia Pacific
min	3552	5098	1919	1157	873	413	534	226
mean	5136	5746	3225	2834	1167	690	749	430
max	7384	6740	3795	4553	1734	1044	1079	712

We also provide additional out-of-sample evidence by providing results for the period from 2019 to 2023 on US stocks. These results were not part of the original study and can be found in Appendix 2.E.6.

Features

The two-stage quantile neural network model uses both stock characteristics and market variables aggregated from the stock-level variables as features. There are 176 stock characteristics consisting of 153 anomalies from Tobek and Hronec (2021) plus an additional 18 stock-level volatility estimates. The 23 market-level variables consist of 18 market volatility variables and 5 market mean variables. This brings the total number of features to 194. The list of 153 anomalies is given in Appendix 2.A and includes 93 fundamental, 49 market friction and 11 I/B/E/S anomalies, mostly covered by McLean and Pontiff (2016), Hou, Xue, and Zhang (2020) and Harvey, Y. Liu, and Zhu (2016). Compared to the original data, we use weekly data during training. This means that The fundamental signals are updated every week with financial statement information from financial years ending at least 6 months earlier. Trade data information, such as market capitalisation, is taken as the most recent available. Missing characteristic values are imputed using cross-sectional medians.

Each sub-network in the two-stage quantile neural network model uses its own set of features. The standardized τ-quantile sub-network uses 176 features consisting of 153 anomalies-based stock characteristics together with 18 stock-level volatilities and 5 market mean returns. The 18 stock-level volatilities are calculated using daily exponentially weighted moving averages

(EWMA) with $\alpha \in \{0.8, 0.9, 0.94, 0.96, 0.98, 0.99\}$ smoothing factors, EWMA of negative returns with α smoothing factors of $\{0.8, 0.9, 0.94\}$ to proxy asymmetric volatility, total volatilities calculated over the past 3, 6 and 12 months, and Parkinson (1980) high-low range-based volatility. All stock-level volatility estimates are scaled by their cross-sectional mean to remove the effects of changes in market-wide average volatility. This is done to make stock-level volatilities more consistent with the other stock-level variables, which are normalised to cross-sectional empirical quantiles. The 5 additional market variables used as inputs to the standardized τ-quantile sub-network are computed as EWMAs of cross-sectional equal-weighted averages of individual stock returns $\mu_t = \alpha \cdot \left(1/N_t \sum_{i=1}^{N_t} r_{i,t-1}\right) + (1-\alpha) \cdot \mu_{t-1}$ over the number of stocks N_t with smoothing parameters $\alpha \in \{0.9, 0.94, 0.96, 0.99, 0.999\}$. These variables measure the average historical market-wide return and should help in forecasting when there is some persistent time variance in expected market mean returns. We rescale the mean variables by the same scalar as the predicted returns to reflect the fact that their scale over time is relevant in the estimation.

The market volatility sub-network uses 18 cross-sectional averages of stock-level volatilities described above. These serve as a proxy for market-wide index volatility, as they are likely to be relevant for estimating the distribution of individual stock returns, which includes both systematic and idiosyncratic volatility components.

Validation

We use the initial period from 1973 to 1994 for hyperparameter optimization. Models with varying hyperparameters are trained from 1973 to 1989 and validated from 1990 to 1994. The hyperparameters yielding the best validation performance, measured in terms of average quantile loss (see Subsection 2.3.3 for details), are then applied to train models from 1995 to 2018, generating true out-of-sample predictions. Detailed hyperparameter searches for both full and liquid samples are documented in Section 2.D together with the validation scheme.

From 1995 to 2018, using only the U.S. sample, we retrain the model annually using an expanding data window starting in 1973. To prevent data leakage into the next year, the latest date for feature calculation each year is November 29. Separate models are trained for full and liquid samples, with monthly out-of-sample predictions for both U.S. and international samples, using the

most recently trained model at each month's end. Predictions from the full sample model are applied to the full sample and, similarly, to the liquid sample. The post-original sample of 2019 to 2023 is treated accordingly.

2.3.2 Description of the estimated moments

Since our main empirical results are based on the predicted quantiles and moments of the distribution of all available stocks, we first describe the estimated moments for the US data. As the rest of the international regions show similar dynamics, we present these results in Appendix Table 2.12. Specifically, we predict 37 τ quantiles $\tau \in \{0.00005, 0.0001, 0.001, 0.005,\ 0.01, \ldots, 0.05,\ 0.075,$ $0.1, 0.15, 0.2, \ldots, 0.8, 0.85, 0.9, 0.925,\ 0.95, \ldots 0.99, 0.995, 0.999, 0.9999, 0.99995\}$, mean, standard deviation, skewness and kurtosis of all available individual stock returns.

Table 2.2 contains summary statistics for the predicted moments, and Figure 2.4 shows the distribution of the predicted mean, standard deviation, skewness and kurtosis for the full and liquid samples for all stocks across the US.

Table 2.2. Summary statistics for the moments in the U.S.
This table shows summary statistics for the forecasted first four moments for the full and liquid samples in the U.S. The moments are forecasted using the quantiles-to-moments algorithm introduced in Subsection 2.2.3.

	Full Sample			Liquid Sample		
	Mean	Median	Std. Dev.	Mean	Median	Std. Dev.
Mean	0.0068	0.0089	0.0216	0.0098	0.0092	0.0126
Std. Dev.	0.1477	0.1317	0.0743	0.1071	0.0940	0.0539
Skewness	0.8453	0.7572	0.6352	0.1800	0.1722	0.1993
Kurtosis	8.9784	7.5963	5.2384	5.4261	5.2248	2.3752

We can see that the distribution of the mean for both samples is approximately centred around zero, with the liquid sample having a slightly higher density around the mean. The mean return is positive for both samples, with the liquid sample having a slightly higher mean (0.0098) than the total sample (0.0068). The standard deviation is higher for the total sample (0.1477) than for the liquid sample (0.1071) and shows a wider spread for the total sample than for the liquid sample. Skewness is significantly higher in the full sample

(0.8453) than in the liquid sample (0.1800), reflecting more extreme positive and negative returns, while the liquid sample shows a more symmetrical distribution. Kurtosis is also higher in the full sample (8.9784) than in the liquid sample (5.4261), indicating more pronounced tails compared to the liquid sample.

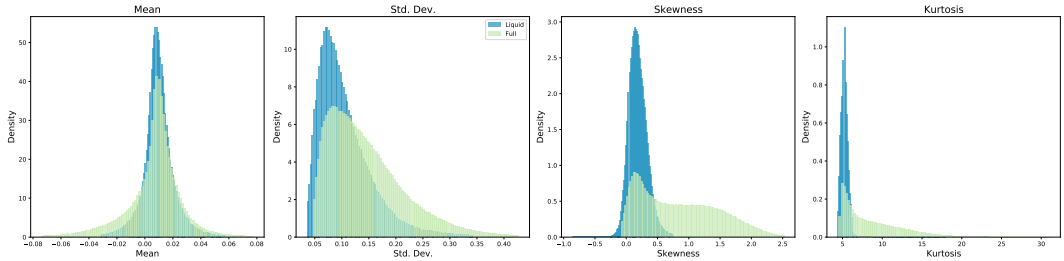


Figure 2.4. Histogram of central distribution moments
This figure shows histograms of the forecasted first four moments for the full and liquid samples in the U.S. The Moments are forecasted using the quantiles-to-moments algorithm introduced in Subsection 2.2.3. Displayed variables are trimmed to exclude the top and bottom 0.5% of their values.

Table 2.3 also includes average cross-sectional correlations between the first four predicted moments and the median of stock returns. There is a high correlation of 0.95 between the forecasted mean and the median for the liquid sample, while the correlation is lower (0.75) for the full sample. This is consistent with the fact that the full sample contains more micro-cap stocks with higher skewness, which affects the mean but not the median. There is also a high correlation between skewness and kurtosis (0.91) in the full sample. In contrast, in the liquid sample, the correlation between skewness and kurtosis is close to zero.

Table 2.3. Average cross-sectional correlation between variables in the U.S.
This table shows the average cross-sectional Spearman correlation between the forecasted first four moments and the median of the stock returns for the full and liquid samples in the U.S. Medians are forecasted using the two-stage model defined in Section 2.2 and moments are forecasted using the quantiles-to-density algorithm introduced in Subsection 2.2.3.

	Full Sample					Liquid Sample				
	Median	Mean	Variance	Skewness	Kurtosis	Median	Mean	Variance	Skewness	Kurtosis
Median	1.00	0.75	-0.63	-0.54	-0.44	1.00	0.95	-0.34	0.07	0.16
Mean	0.75	1.00	-0.22	-0.08	-0.08	0.95	1.00	-0.15	0.26	0.09
Variance	-0.63	-0.22	1.00	0.53	0.24	-0.34	-0.15	1.00	0.66	-0.30
Skewness	-0.54	-0.08	0.53	1.00	0.91	0.07	0.26	0.66	1.00	-0.05
Kurtosis	-0.44	-0.08	0.24	0.91	1.00	0.16	0.09	-0.30	-0.05	1.00

2.3.3 Predictability

Here we evaluate the out-of-sample predictions of the quantiles from our two-stage quantile neural network (2sNN). We also evaluate the predictions of the mean and variance estimated from the quantiles. This is of particular interest as we argue that the first and second moments, estimated without any assumptions about the distribution, are more predictive of the mean and variance. To benchmark our forecasting strategy, we use a linear model (LNN), the single hidden layer neural network with 32 neurons (1hNN), the double hidden layer neural network with 128 neurons (2hNN) and the commonly used GARCH. Details of the benchmark models can be found in Appendix 2.B.1 and Appendix 2.B.2.

To measure predictive performance across the distribution, we report the average loss, which is defined as the time series average of the cross-sectional τ-quantile losses averaged across all τs for each month.

$$L_{avg} = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{N_t}\sum_{i=1}^{N_t}\left(\frac{1}{K}\sum_{\tau\in\mathcal{T}} L_{\tau,t,i}\right)\right). \tag{2.7}$$

Table 2.4 shows the average out-of-sample τ-quantile losses for the full and liquid samples across all regions. The average losses of the two-stage quantile neural network are significantly smaller than all benchmark models when looking at the full sample data. Our two-stage model also significantly outperforms the 1hNN full sample model in all regions except Europe and the 2hNN full sample model in the US and Asia Pacific. The average quantile loss is also lower in Japan, but the difference is not statistically significant at the 5% level.

Predictive performance differences for the liquid sample mirror those observed for the full sample, although to a lesser extent and with fewer statistically significant differences. As expected, average out-of-sample quantile losses are generally lower for the liquid sample than for the full sample, reflecting the lower volatilities of the underlying stocks. The two-stage model consistently outperforms the GARCH model across all regions. However, when compared to alternative neural network specifications, the two-stage model only shows significant outperformance in the US region. Internationally, while the two-stage model achieves lower average quantile losses than the alternative neural network specifications, these differences do not reach statistical significance at the 5% level, except for the 1hNN model in Europe. Further, the partial outperformance of the two-stage model compared to the 2hNN model can be attributed to the

different model specifications, namely the bottleneck layer and market volatility sub-network.

Table 2.4. Out-of-sample quantile forecasts evaluation
This table shows the average out-of-sample quantile cross-section loss for quantile forecasts from the two-stage model (2sNN) and alternative models. The results are obtained separately for the full and liquid samples for individual regions covering the out-of-sample period from 1995 to 2018. Neural network models are trained annually using the expanding window on U.S. data only, while out-of-sample forecasts are generated for both the US and international samples. The average loss is calculated as the time series average of the cross-sectional averages of the average quantile loss across all $\tau \in \{0.00005, 0.0001, 0.001, 0.005,\ 0.01, \ldots, 0.05,\ 0.075,\ 0.1, 0.15, 0.2, \ldots, 0.8, 0.85, 0.9, 0.925, 0.95, \ldots 0.99, 0.995, 0.999, 0.9999, 0.99995\}$ as in (Eq. 2.7) for the two-stage NN model, as well as benchmark neural network models (the linear model (LNN), the one-layer model (1hNN), the two-layer model (2hNN)) and the benchmark GARCH model. Note that losses are multiplied by 100.0, and the t-statistics in parentheses are computed for the average loss difference between the given model specification and the two-stage benchmark model using Newey-West standard errors with 12 lags.

Specification	Variable	Full Sample				Liquid Sample			
		USA	Europe	Japan	Asia Pacific	USA	Europe	Japan	Asia Pacific
GARCH	L_{avg}	2.659	2.669	2.050	3.416	1.899	1.679	1.836	2.159
	(t-stat)	(8.47)	(14.07)	(8.15)	(15.51)	(6.38)	(4.49)	(3.05)	(6.23)
LNN	L_{avg}	2.592	2.520	2.020	3.143	1.874	1.658	1.820	2.115
	(t-stat)	(6.05)	(3.44)	(4.43)	(4.44)	(3.95)	(2.07)	(1.72)	(1.53)
1hNN	L_{avg}	2.569	2.496	2.004	3.120	1.869	1.655	1.819	2.121
	(t-stat)	3.81	(0.44)	(4.03)	(3.09)	(3.08)	(1.69)	(1.63)	(1.57)
2hNN	L_{avg}	2.564	2.494	2.001	3.123	1.870	1.654	1.823	2.128
	(t-stat)	(2.22)	(0.04)	(1.71)	(2.50)	(2.39)	(1.10)	(1.88)	(1.73)
2sNN	L_{avg}	2.550	2.494	1.992	3.106	1.858	1.649	1.813	2.106

We also examine the performance of the mean and variance forecasts. To compare our two-stage model and the forecasts obtained by the algorithms described earlier, we compute the mean forecasts directly using a two-layer hidden neural network model estimated with a mean square error loss function. We follow Gu et al. (2020) and modify the calculation of the Diebold-Mariano test statistic to compare the cross-sectional average of forecast errors instead of comparing individual errors.

Table 2.5 shows the out-of-sample R^2 of the mean and median forecasts. Our two-stage model provides the highest out-of-sample R^2. For the liquid sample and across all regions, the highest R^2 is consistently achieved by the

median forecast from the two-stage model. Forecasts using the median are not statistically different from those using the mean in the two-stage model. R^2 for the direct mean prediction from the 2 hidden layers neural network is negative and smaller than the two-stage predictions. In the case of the full sample, the highest R^2 is consistently achieved by the mean prediction derived from our model.

Table 2.5. Mean forecast evaluation
This table shows the R^2 of the mean and median forecasts of the next 22 days of stock returns for the full and liquid samples across all regions. The mean prediction based on the two-stage quantile model (2sNN - Mean) is compared with the direct mean prediction from the two-hidden-layer neural network (2hNN MSE - Mean) and the median prediction from the two-stage model (2sNN - Median). Results are obtained separately for full and liquid samples on individual regions covering the period from 1995 to 2018. Models are retrained annually using the expanding window for US data only, while out-of-sample forecasts are generated for both US and international data. Note that losses are multiplied by 100, and values in parentheses are Diebold-Mariano test statistics for the difference between the model in a row and the two-stage NN mean. Standard errors are adjusted using the Newey-West standard deviation estimator with 12 lags.

Specification	Full Sample					Liquid Sample				
	USA	Europe	Japan	Asia Pacific	Global	USA	Europe	Japan	Asia Pacific	Global
2hNN MSE - Mean	0.65	0.14	0.85	0.69	0.51	-0.68	-0.44	-0.48	-0.77	-0.54
	(-2.54)	(-3.56)	(-2.37)	(-3.39)	(-3.55)	(-1.81)	(-1.31)	(-1.84)	(-1.79)	(-1.98)
2sNN - Median	0.10	-0.04	0.30	0.06	0.09	0.37	0.32	0.47	0.55	0.50
	(-3.50)	(-4.92)	(-2.82)	(-4.83)	(-4.93)	(0.23)	(0.27)	(0.98)	(0.39)	(0.52)
2sNN - Mean	1.63	0.87	1.66	1.63	1.36	0.32	0.22	0.18	0.40	0.37

Finally, to evaluate the second-moment forecasts, we use the Mean Absolute Deviation (MAD) and the Root Mean Squared Error (RMSE) of the predicted 22-day volatility for the full and liquid samples for all regions. Table 2.6 shows a highly significant outperformance of the volatility forecast based on the two-stage model in terms of both MAD and RMSE for all regions and for both the full and liquid samples.

Table 2.6. Volatility forecast evaluation
This table shows the mean absolute deviation (MAD), root mean squared error (RMSE), and Diebold Mariano test statistic (DM) comparing the corresponding losses of the two competing models for the 22-day volatility forecasts. Volatility forecasting based on two-stage quantile forecasting (2sNN) is compared with GARCH model forecasting. Results are obtained separately for full and liquid samples on individual regions covering the period from 1995 to 2018. Neural network models are retrained annually using an expanding window on U.S. data only, while out-of-sample forecasts are generated for both US and international samples. Diebold-Mariano test statistics are computed for the difference between the GARCH and two-stage NN model volatility forecasts. Note that losses are multiplied by 100.0 and the standard errors used in the Diebold-Mariano test statistics are adjusted using the Newey-West standard deviation estimator with 12 lags.

Specification	Variable	Full Sample					Liquid Sample				
		USA	Europe	Japan	Asia Pacific	Global	USA	Europe	Japan	Asia Pacific	Global
GARCH	MAD	7.34	8.79	7.79	11.1	8.53	3.89	3.54	3.80	4.87	3.93
2sNN	MAD	4.63	5.20	3.99	6.49	4.99	2.93	2.63	2.81	3.16	2.87
GARCH	RMSE	13.14	16.24	14.63	18.70	15.50	6.93	7.10	6.72	9.52	7.36
2sNN	RMSE	7.46	8.52	5.64	10.05	8.08	4.62	4.03	3.98	4.86	4.38
	DM	(10.25)	(35.92)	(20.32)	(44.26)	(26.47)	(7.90)	(5.98)	(7.97)	(9.40)	(10.02)

Overall, we demonstrate the superior distributional forecasting performance of the two-stage model and the moments derived from it compared to alternative models. The average out-of-sample quantile loss represents the forecast over the entire distribution, while the evaluation of the mean and variance forecasts provides additional evidence derived from the forecasted quantiles.

2.3.4 Distribution and the cross-section of stock returns

In addition to forecasting, we aim to investigate whether individual distributional characteristics are priced into the cross-section of returns through the profitability of long-short portfolios formed based on forecasted τ-quantiles as well as the profitability of long-short portfolios formed using central moments – mean, volatility, skewness and kurtosis.

τ-quantiles and the cross-section of stock returns

Table 2.7 shows the average monthly returns and annualised Sharpe ratios for decile long-short equal-weighted portfolios constructed using the τ-quantile forecasts of the two-stage model. Each month, the long and short legs of the

portfolios are constructed by selecting the top and bottom 10% of stocks with the highest and lowest forecasts, respectively.

Table 2.7. Quantiles and the cross-section of stock returns in the U.S. This table shows the average monthly returns and annualised Sharpe ratios for decile long, short and long-short equal-weighted portfolios over the period 1995 to 2018. For each τ-quantile forecast, the long portfolio is formed by taking the top 10% of stocks with the highest forecast each month, The short portfolio is formed by taking the bottom 10% of stocks with the lowest forecast each month. The long-short portfolio is created by taking the difference between the long and short portfolios. Values in brackets are t-statistics adjusted for Newey-West standard errors with 12 lags. Results are obtained separately for the full and liquid samples for the US only.

	Full Sample						Liquid Sample					
	Long		Short		Long-short		Long		Short		Long-short	
τ	Mean	SR	Mean	SR	Mean	SR	Mean	SR	Mean	SR	Mean	SR
0.01	1.22 (5.62)	1.67	0.14 (0.19)	0.04	1.09 (1.52)	0.31	0.95 (5.26)	1.10	0.16 (0.21)	0.05	0.79 (1.06)	0.24
0.05	1.25 (5.66)	1.76	-0.09 (-0.12)	-0.02	1.34 (1.86)	0.39	0.96 (5.63)	1.15	0.15 (0.20)	0.04	0.81 (1.11)	0.25
0.1	1.28 (5.73)	1.79	-0.26 (-0.35)	-0.07	1.53 (2.13)	0.45	0.98 (5.72)	1.17	0.11 (0.14)	0.03	0.87 (1.20)	0.26
0.2	1.34 (6.04)	1.87	-0.57 (-0.78)	-0.16	1.91 (2.68)	0.57	1.02 (5.89)	1.21	0.04 (0.06)	0.01	0.98 (1.34)	0.30
0.3	1.46 (6.54)	1.96	-0.91 (-1.27)	-0.27	2.37 (3.36)	0.74	1.10 (6.09)	1.25	-0.04 (-0.06)	-0.01	1.14 (1.58)	0.36
0.4	1.88 (7.44)	2.06	-1.34 (-1.91)	-0.42	3.22 (4.89)	1.12	1.39 (6.31)	1.34	-0.14 (-0.20)	-0.04	1.53 (2.23)	0.54
0.5	2.88 (6.97)	1.78	-1.82 (-2.68)	-0.62	4.70 (8.25)	2.27	1.86 (5.61)	1.10	-0.35 (-0.53)	-0.12	2.21 (4.21)	1.21
0.6	3.44 (5.68)	1.37	-2.12 (-3.51)	-0.88	5.56 (9.41)	3.96	1.74 (3.69)	0.71	-0.38 (-0.88)	-0.21	2.11 (5.40)	1.34
0.7	3.48 (4.73)	1.08	-1.42 (-3.57)	-1.10	4.90 (6.57)	1.99	1.66 (2.86)	0.56	0.48 (2.50)	0.54	1.18 (2.23)	0.43
0.8	2.98 (3.72)	0.82	0.44 (1.65)	0.62	2.55 (3.16)	0.75	1.28 (1.94)	0.39	0.71 (3.95)	0.85	0.57 (0.91)	0.18
0.9	2.12 (2.57)	0.55	0.75 (3.29)	1.10	1.37 (1.69)	0.38	0.91 (1.24)	0.27	0.76 (4.29)	0.93	0.15 (0.21)	0.05
0.95	1.71 (2.11)	0.45	0.86 (4.04)	1.22	0.85 (1.07)	0.23	0.77 (1.03)	0.23	0.80 (4.62)	0.98	-0.03 (-0.04)	-0.01
0.99	1.36 (1.70)	0.36	0.97 (5.04)	1.22	0.39 (0.50)	0.11	0.60 (0.80)	0.17	0.84 (4.72)	1.01	-0.23 (-0.32)	-0.07

We show the highest average monthly returns and Sharpe ratios for the long-short portfolios formed based on the central τ-quantiles, τs of 50% or 60%, for both the full sample and the liquid sample. Average monthly returns decrease as we move towards the tails of the distribution. For the full sample, average monthly returns are not significantly different from zero for τs greater than 80% and less than 10%. For the liquid sample, the average monthly returns fall even more sharply as one moves towards the tails of the distribution, becoming not significantly different from zero for τs greater than 70% and less than 40%. The highest Sharpe ratios are achieved by the long-short portfolios with the highest average monthly returns.

These results suggest that the tail τ-quantiles in isolation are not priced into the cross-section of stock returns. On the contrary, the central τ-quantiles are associated with the highest average returns. We also provide international evidence showing similar results in Table 2.13. These results are consistent with the next empirical results documenting how the moments of the distribution do not appear to be priced in the cross-section of stock returns.

Distributional moments and the cross-section of stock returns

We then examine the profitability of long-short portfolios formed on the basis of the central moments of the distribution - mean, variance, skewness and kurtosis - forecasts estimated from our two-stage quantile model. To illustrate the importance of asymmetry in the data, we first compare the profitability of portfolios formed on the mean and median forecasts. Table 2.8 shows the average monthly returns and Sharpe ratios of the long-short decile portfolios based on the mean and median forecasts across all regions.

Table 2.8. Mean and median portfolios
This table shows average monthly returns and annualised Sharpe ratios for decile long-short equal-weighted portfolios based on the median and mean predictions from the two-stage quantile NN model (2sNN - Mean), the mean prediction from the two-hidden-layer neural network (2hNN MSE - Mean) and the median prediction from the two-stage NN model (2sNN - Median). The results are shown for the full and liquid samples and for individual regions, covering the period from 1995 to 2018. The long portfolio is formed by taking the top 10% of stocks with the highest forecast each month, and the short portfolio is formed by taking the bottom 10% of stocks with the lowest forecast each month. Values in parentheses are t-statistics adjusted for Newey-West standard errors with 12 lags. The significance of the difference in mean returns is tested using t-statistics with Newey-West standard errors with 12 lags.

Specification	Variable	Full Sample					Liquid Sample				
		USA	Europe	Japan	Asia Pacific	Global	USA	Europe	Japan	Asia Pacific	Global
2hNN MSE - Mean	Avg Ret	5.00	3.28	3.11	5.71	4.27	1.62	1.71	1.67	2.05	1.76
	SR	3.47	2.87	2.52	3.72	5.50	0.87	1.63	1.52	1.38	2.01
	t-stat	4.78	9.26	5.78	5.79	9.93	1.87	1.53	2.81	2.30	2.54
2sNN - Median	Avg Ret	4.70	3.35	3.10	5.07	4.05	2.21	2.00	2.25	2.95	2.35
	SR	2.27	2.48	1.84	3.06	3.53	1.21	1.52	1.51	1.82	2.03
	t-stat	4.26	7.14	5.62	8.75	9.16	0.84	0.19	0.10	1.07	0.75
2sNN - Mean	Avg Ret	6.44	5.29	4.03	8.05	5.95	2.35	2.02	2.27	3.12	2.44
	SR	4.34	4.49	3.31	4.99	6.92	1.70	1.88	1.92	2.13	2.83

To benchmark our forecasts, we use the mean forecasts obtained directly from the feed-forward neural network model estimated with the mean squared error loss function. The long-short portfolio based on the mean forecast of our two-stage quantile model leads to significantly higher average monthly returns and Sharpe ratios compared to the portfolios based on the mean forecasts obtained directly from the simple neural network for the full sample across all regions. This is consistent with the R^2 results in Table 2.5, as our two-stage model outperforms the direct mean forecast in terms of R^2 across all regions and also leads to higher corresponding profitability. In the case of the full sample, the average monthly returns of our mean forecast are economically

higher than the direct mean forecast in all regions and significantly higher for Japan, Asia-Pacific and globally.

Next, the average monthly returns of the mean- and median-based long-short portfolios are statistically indistinguishable for the liquid sample. This is also consistent with the R^2 results in Table 2.5, as there are no significant differences in the quality of the mean and median forecasts for the liquid sample. However, the results are different for the full sample. The difference between the mean and the median is associated with distributional asymmetry, which is more pronounced for stocks in the full sample. Mean-based long-short portfolios have significantly higher average monthly returns than median-based long-short portfolios. This is true for all regions and also globally. The difference between mean and median pricing for the full sample indicates the effect of distributional asymmetry in the cross-section of stock returns. In addition to examining the role of individual τ-quantiles and the mean, we also look at the decile portfolios and the corresponding long-short portfolios formed on the basis of the higher moments. The higher moments are obtained by the quantiles-to-moments algorithm using the two-stage model quantile forecasts as described in Subsection 2.2.3. Table 2.9 shows the average monthly returns and Sharpe ratios of the equally weighted decile long-short portfolios based on the forecast mean, median, volatility, skewness and kurtosis.

Overall, none of the higher moments appear to be monotonically priced in the cross-section of stock returns in the liquid or full sample. Average monthly returns of the long-short portfolios formed on the higher moments are mostly not significantly different from zero. There are some regional exceptions for the full sample, where volatility and skewness-based long-short portfolios have significantly positive average monthly returns in Europe and Asia Pacific. Results for the value-weighted long-short portfolios are mostly consistent with the equal-weighted portfolios and are shown in Table 2.14 in the Appendix. The results on volatility show a negative relationship between predicted volatility and future return for the liquid sample, in line with previous academic research. However, the relationship is not statistically significant, which is to be expected given the focus on large-cap stocks. The full sample results are plagued by market microstructure distortions for equal-weighted portfolios, such as bid-ask jumps, and the pattern is clearer for value-weighted portfolios. However, the effect remains statistically insignificant, which can be explained by our sample period, which focuses primarily on years more recent than the sample period of Ang, Hodrick, et al. (2006), the original study that identified volatility as priced

Table 2.9. Moments decile portfolios

This table shows average monthly returns for individual deciles (1-10) and long-short equal-weighted portfolios built at the predicted moments. Results are reported for the full and liquid samples, as well as for individual regions, covering the period from 1995 to 2018. Values in parentheses are t-statistics adjusted for Newey-West standard errors with 12 lags.

Mean	Full Sample				Liquid Sample			
Decile	USA	Europe	Japan	Asia Pacific	USA	Europe	Japan	Asia Pacific
1	-2.37 (-3.87)	-1.06 (-1.87)	-1.24 (-1.99)	-1.69 (-2.40)	-0.48 (-0.81)	-0.52 (-0.90)	-1.13 (-1.99)	-1.23 (-1.73)
2	-0.46 (-0.97)	-0.29 (-0.63)	-0.40 (-0.77)	-0.31 (-0.53)	0.33 (0.80)	0.18 (0.44)	-0.32 (-0.65)	-0.00 (-0.00)
3	0.33 (0.90)	0.04 (0.10)	-0.19 (-0.41)	0.13 (0.23)	0.64 (2.00)	0.51 (1.32)	-0.01 (-0.02)	0.48 (0.97)
4	0.68 (2.07)	0.35 (0.93)	0.10 (0.24)	0.46 (0.87)	0.75 (2.56)	0.71 (1.97)	0.16 (0.38)	0.58 (1.25)
5	1.01 (3.39)	0.57 (1.64)	0.38 (0.94)	0.87 (1.69)	0.90 (3.21)	0.72 (2.03)	0.20 (0.50)	0.68 (1.53)
6	1.28 (4.37)	0.92 (2.67)	0.70 (1.82)	1.30 (2.57)	1.00 (3.77)	0.91 (2.70)	0.34 (0.96)	1.04 (2.59)
7	1.46 (4.94)	1.22 (3.38)	0.88 (2.27)	1.68 (3.27)	1.17 (4.17)	0.94 (2.57)	0.47 (1.27)	1.10 (2.61)
8	1.82 (5.50)	1.51 (4.05)	1.27 (3.05)	2.30 (4.30)	1.20 (3.77)	1.13 (3.06)	0.65 (1.74)	1.26 (2.83)
9	2.30 (5.92)	2.05 (5.18)	1.63 (3.68)	3.05 (5.02)	1.44 (4.49)	1.37 (3.50)	0.91 (2.45)	1.52 (3.18)
10	4.07 (6.72)	4.23 (8.55)	2.79 (5.44)	6.36 (8.66)	1.87 (5.03)	1.50 (3.55)	1.14 (2.89)	1.88 (3.68)
10-1	6.44 (9.76)	5.29 (12.36)	4.03 (10.26)	8.05 (18.56)	2.35 (5.16)	2.02 (5.51)	2.27 (6.90)	3.11 (7.89)
Median								
1	-1.82 (-2.68)	-0.42 (-0.62)	-0.96 (-1.40)	-0.99 (-1.21)	-0.35 (-0.54)	-0.51 (-0.80)	-1.10 (-1.76)	-1.09 (-1.43)
2	0.16 (0.29)	0.52 (1.00)	0.08 (0.13)	1.01 (1.42)	0.34 (0.73)	0.23 (0.51)	-0.33 (-0.63)	-0.10 (-0.18)
3	0.65 (1.48)	0.29 (0.66)	0.25 (0.53)	1.28 (2.01)	0.61 (1.75)	0.50 (1.28)	-0.03 (-0.07)	0.50 (1.01)
4	0.99 (2.68)	0.50 (1.31)	0.28 (0.64)	1.14 (1.99)	0.75 (2.44)	0.69 (1.84)	0.17 (0.40)	0.48 (1.00)
5	1.19 (3.74)	0.69 (1.99)	0.36 (0.92)	1.14 (2.23)	0.83 (2.93)	0.77 (2.19)	0.29 (0.78)	0.65 (1.46)
6	1.29 (4.45)	0.91 (2.82)	0.53 (1.44)	1.08 (2.24)	1.08 (3.91)	0.91 (2.67)	0.22 (0.66)	0.93 (2.20)
7	1.44 (5.51)	1.09 (3.38)	0.82 (2.18)	1.37 (2.95)	1.13 (4.29)	0.95 (2.76)	0.46 (1.25)	1.16 (2.81)
8	1.52 (5.26)	1.36 (3.92)	1.04 (2.72)	1.82 (3.83)	1.18 (4.07)	1.03 (2.83)	0.65 (1.87)	1.39 (3.31)
9	1.80 (6.00)	1.65 (4.58)	1.34 (3.54)	2.22 (4.40)	1.41 (4.81)	1.39 (3.80)	0.92 (2.61)	1.53 (3.53)
10	2.88 (6.97)	2.94 (6.73)	2.14 (4.62)	4.08 (6.56)	1.86 (5.61)	1.49 (3.73)	1.16 (3.17)	1.86 (3.81)
10-1	4.70 (8.25)	3.35 (6.64)	3.10 (6.76)	5.07 (10.76)	2.21 (4.22)	2.00 (4.65)	2.25 (5.88)	2.96 (6.27)
Volatility								
1	1.02 (4.96)	0.75 (3.23)	0.25 (0.98)	0.76 (2.50)	0.88 (5.13)	0.81 (3.54)	0.33 (1.57)	1.00 (3.14)
2	1.11 (4.28)	0.83 (2.87)	0.36 (1.18)	1.11 (2.89)	0.98 (4.19)	0.84 (2.88)	0.26 (0.90)	0.95 (2.80)
3	1.08 (3.97)	0.92 (2.85)	0.49 (1.44)	1.07 (2.40)	1.02 (4.02)	0.85 (2.84)	0.32 (0.95)	0.91 (2.22)
4	1.09 (3.82)	0.85 (2.38)	0.59 (1.55)	0.98 (1.93)	1.10 (4.09)	0.82 (2.50)	0.27 (0.80)	0.97 (2.20)
5	1.05 (3.24)	0.85 (2.30)	0.60 (1.44)	0.95 (1.73)	0.98 (3.49)	0.81 (2.16)	0.50 (1.38)	0.97 (2.16)
6	1.05 (2.89)	0.66 (1.65)	0.62 (1.39)	0.96 (1.70)	0.96 (3.25)	0.82 (2.11)	0.36 (0.95)	0.79 (1.52)
7	1.05 (2.44)	0.57 (1.25)	0.65 (1.31)	1.08 (1.72)	0.91 (2.41)	0.74 (1.79)	0.26 (0.56)	0.62 (1.24)
8	0.86 (1.65)	0.37 (0.73)	0.83 (1.48)	1.00 (1.37)	0.84 (1.89)	0.84 (1.82)	0.31 (0.64)	0.71 (1.24)
9	0.74 (1.16)	0.62 (1.04)	0.89 (1.41)	1.61 (1.94)	0.72 (1.29)	0.71 (1.33)	0.13 (0.22)	0.62 (0.90)
10	1.05 (1.33)	3.12 (4.21)	0.61 (0.85)	4.63 (4.99)	0.44 (0.58)	0.23 (0.32)	-0.34 (-0.45)	-0.24 (-0.31)
10-1	0.03 (0.04)	2.36 (3.35)	0.35 (0.62)	3.86 (5.10)	-0.44 (-0.60)	-0.57 (-0.96)	-0.68 (-1.05)	-1.24 (-1.81)
Skewness								
1	0.85 (3.00)	0.73 (2.07)	0.31 (0.96)	0.69 (1.53)	0.61 (2.65)	0.30 (0.90)	-0.06 (-0.19)	0.44 (1.19)
2	1.08 (4.01)	0.88 (2.27)	0.37 (1.04)	0.84 (1.76)	0.75 (2.82)	0.45 (1.20)	-0.04 (-0.11)	0.46 (1.07)
3	1.05 (3.50)	0.83 (1.99)	0.48 (1.12)	0.77 (1.50)	0.77 (2.74)	0.58 (1.66)	0.06 (0.14)	0.59 (1.38)
4	1.03 (3.13)	0.61 (1.45)	0.50 (1.08)	0.67 (1.20)	0.81 (2.93)	0.79 (2.26)	0.19 (0.48)	0.70 (1.55)
5	1.04 (3.04)	0.51 (1.21)	0.55 (1.13)	0.97 (1.70)	0.84 (2.94)	0.78 (2.18)	0.20 (0.48)	0.86 (1.85)
6	0.88 (2.51)	0.59 (1.50)	0.56 (1.12)	1.31 (2.26)	1.02 (3.23)	0.84 (2.47)	0.42 (1.03)	0.74 (1.52)
7	0.90 (2.27)	0.71 (1.74)	0.67 (1.38)	1.67 (2.70)	1.11 (3.45)	1.00 (2.79)	0.38 (0.85)	1.02 (2.15)
8	0.96 (2.18)	0.99 (2.32)	0.87 (1.74)	2.20 (3.02)	1.09 (2.95)	0.97 (2.35)	0.52 (1.20)	1.01 (2.05)
9	1.23 (2.45)	1.65 (3.83)	0.98 (1.92)	2.67 (3.65)	1.03 (2.34)	1.04 (2.26)	0.54 (1.25)	0.95 (1.54)
10	1.09 (1.92)	2.03 (4.92)	0.61 (1.30)	2.37 (3.63)	0.80 (1.10)	0.72 (1.16)	0.19 (0.35)	0.54 (0.75)
10-1	0.24 (0.57)	1.30 (4.07)	0.30 (1.08)	1.68 (4.06)	0.18 (0.28)	0.42 (1.02)	0.25 (0.72)	0.11 (0.20)
Kurtosis								
1	1.08 (2.69)	0.82 (1.90)	0.48 (1.01)	0.80 (1.46)	0.76 (1.26)	0.39 (0.61)	0.03 (0.04)	0.47 (0.70)
2	0.99 (3.32)	0.93 (2.28)	0.47 (1.26)	0.68 (1.34)	0.70 (1.44)	0.67 (1.34)	-0.17 (-0.32)	0.45 (0.70)
3	1.00 (3.35)	0.76 (1.76)	0.51 (1.25)	0.69 (1.28)	0.70 (1.85)	0.63 (1.44)	0.11 (0.24)	0.34 (0.61)
4	1.05 (3.40)	0.69 (1.62)	0.49 (1.11)	0.90 (1.58)	0.80 (2.43)	0.73 (1.98)	0.23 (0.57)	0.72 (1.57)
5	0.96 (2.76)	0.54 (1.29)	0.57 (1.19)	1.07 (1.93)	0.78 (2.49)	0.90 (2.48)	0.37 (0.99)	0.76 (1.68)
6	0.99 (2.64)	0.72 (1.68)	0.70 (1.37)	1.78 (2.83)	0.91 (3.19)	0.78 (2.27)	0.43 (1.32)	0.99 (2.34)
7	0.99 (2.27)	0.84 (2.05)	0.74 (1.51)	1.94 (2.91)	1.13 (4.45)	0.90 (2.71)	0.50 (1.46)	1.00 (2.30)
8	1.02 (2.30)	1.31 (3.09)	0.91 (1.82)	2.36 (3.28)	1.07 (3.98)	0.90 (2.67)	0.37 (1.10)	0.88 (2.12)
9	1.19 (2.42)	1.49 (4.03)	0.76 (1.62)	2.41 (3.68)	1.07 (3.77)	0.94 (2.85)	0.37 (1.07)	1.02 (2.29)
10	0.83 (2.16)	1.42 (4.54)	0.27 (0.64)	1.52 (2.95)	0.92 (3.37)	0.62 (1.90)	0.16 (0.47)	0.69 (1.63)
10-1	-0.25 (-0.77)	0.60 (1.82)	-0.21 (-0.66)	0.73 (1.80)	0.15 (0.33)	0.23 (0.62)	0.14 (0.33)	0.22 (0.57)

into the cross-section of returns. The results for skewness are inconclusive for a monotonic relationship with future returns. However, there is a noticeable hill-shaped effect; stocks with low skewness have lower future returns relative to stocks with more average skewness. The low returns of the high-skewness portfolio (10) lead to a non-monotonic relationship across the decile portfolios and thus also affect the 10-1 long-short portfolio. Skewness would be more convincingly priced if 9-1 long-short portfolios were used instead.

A substantial body of research has investigated whether investors demand compensation for exposure to tail risk or higher moments, with results varying across sample periods, methodologies, and data sources. While there is general consensus in the literature on the lack of documented relationship between volatility and expected returns as well as kurtosis and expected returns, the relationship between skewness and expected returns is less clear. To compare our empirical results with evidence from the literature, see Section 2.1.1.

2.4 Conclusion

We develop a novel machine learning approach to forecasting the full distribution of stock returns, focusing on both US and international markets. Our method uses a multi-head quantile neural network to predict a set of τ quantiles based on stock characteristics and market variables. We interpolate these quantiles using cubic B-splines to derive the full distribution and its moments, such as mean, volatility, skewness and kurtosis. This approach outperforms traditional parametric models and simpler neural networks in forecasting stock returns, providing more accurate out-of-sample mean and variance forecasts.

We also examine the pricing of distributional characteristics in the cross-section of stock returns using a comprehensive dataset of US and international stocks. We find that only the central τ-quantiles are priced, with significant profitability of long-short portfolios based on these predictions. In addition, our results show that the mean obtained by numerical integration of the interpolated distribution leads to higher out-of-sample returns than direct mean forecasts from other models. We find no evidence that other moments, such as variance, skewness and kurtosis, are priced in the cross-section, adding to the ongoing debate in the asset pricing literature.

References

Abarbanell, J. S., & Bushee, B. J. (1998). Abnormal returns to a fundamental analysis strategy. *Accounting Review*, 19–45.

Acharya, V. V., & Pedersen, L. H. (2005). Asset pricing with liquidity risk. *Journal of Financial Economics*, *77*(2), 375–410.

Alexiou, L., & Rompolis, L. S. (2022). Option-implied moments and the cross-section of stock returns. *Journal of Futures Markets*, *42*(4), 668–691.

Alwathainani, A. M. (2009). Consistency of firms' past financial performance measures and future returns. *The British Accounting Review*, *41*(3), 184–196.

Amihud, Y. (2002). Illiquidity and stock returns: Cross-section and time-series effects. *Journal of Financial Markets*, *5*(1), 31–56.

Amihud, Y., & Mendelson, H. (1986). Asset pricing and the bid-ask spread. *Journal of Financial Economics*, *17*(2), 223–249.

An, B.-J., Ang, A., Bali, T. G., & Cakici, N. (2014). The joint cross section of stocks and options. *The Journal of Finance*, *69*(5), 2279–2337.

Anderson, E. W., Ghysels, E., & Juergens, J. L. (2005). Do heterogeneous beliefs matter for asset pricing? *The Review of Financial Studies*, *18*(3), 875–924.

Ang, A., & Bekaert, G. (2007). Stock return predictability: Is it there? *The Review of Financial Studies*, *20*(3), 651–707.

Ang, A., Hodrick, R. J., Xing, Y., & Zhang, X. (2006). The cross-section of volatility and expected returns. *The Journal of Finance*, *61*(1), 259–299.

Ang, A., Chen, J., & Xing, Y. (2006). Downside risk. *The Review of Financial Studies*, *19*(4), 1191–1239.

Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, *11*, 685–725.

Azevedo, V., Kaiser, G. S., & Mueller, S. (2023). Stock market anomalies and machine learning across the globe. *Journal of Asset Management*, *24*(5), 419–441.

Bali, T. G., Beckmeyer, H., Moerke, M., & Weigert, F. (2023). Option return predictability with machine learning and big data. *The Review of Financial Studies*, *36*(9), 3548–3602.

Bali, T. G., Cakici, N., & Whitelaw, R. F. (2011). Maxing out: Stocks as lotteries and the cross-section of expected returns. *Journal of Financial Economics*, *99*(2), 427–446.

Bali, T. G., Engle, R. F., & Murray, S. (2016). *Empirical asset pricing: The cross section of stock returns*. John Wiley & Sons.

Bali, T. G., Peng, L., Shen, Y., & Tang, Y. (2013). Liquidity shocks and stock market reactions. *The Review of Financial Studies*, *27*(5), 1434–1485.

Ball, R., Gerakos, J., Linnainmaa, J. T., & Nikolaev, V. (2016). Accruals, cash flows, and operating profitability in the cross section of stock returns. *Journal of Financial Economics*, *121*(1), 28–45.

Banz, R. W. (1981). The relationship between return and market value of common stocks. *Journal of Financial Economics*, *9*(1), 3–18.

Barbee Jr, W. C., Mukherji, S., & Raines, G. A. (1996). Do sales–price and debt–equity explain stock returns better than book–market and firm size? *Financial Analysts Journal*, *52*(2), 56–60.

Barber, B., Lehavy, R., McNichols, M., & Trueman, B. (2001). Can investors profit from the prophets? security analyst recommendations and stock returns. *The Journal of Finance*, *56*(2), 531–563.

Barberis, N., & Huang, M. (2008). Stocks as lotteries: The implications of probability weighting for security prices. *American Economic Review*, *98*(5), 2066–2100.

Barry, C. B., & Brown, S. J. (1984). Differential information and the small firm effect. *Journal of Financial Economics*, *13*(2), 283–294.

Barth, M. E., & Hutton, A. P. (2004). Analyst earnings forecast revisions and the pricing of accruals. *Review of Accounting Studies*, *9*(1), 59–96.

Bartov, E., & Kim, M. (2004). Risk, mispricing, and value investing. *Review of Quantitative Finance and Accounting*, *23*(4), 353–376.

Basu, S. (1977). Investment performance of common stocks in relation to their price-earnings ratios: A test of the efficient market hypothesis. *The Journal of Finance*, *32*(3), 663–682.

Belo, F., & Lin, X. (2011). The inventory growth spread. *The Review of Financial Studies*, *25*(1), 278–313.

Belo, F., Lin, X., & Bazdresch, S. (2014). Labor hiring, investment, and stock return predictability in the cross section. *Journal of Political Economy*, *122*(1), 129–177.

Bessembinder, H. (2018). Do stocks outperform treasury bills? *Journal of Financial Economics*, *129*(3), 440–457.

Bhandari, L. C. (1988). Debt/equity ratio and expected common stock returns: Empirical evidence. *The Journal of Finance*, *43*(2), 507–528.

Bianchi, D., Büchner, M., & Tamoni, A. (2021). Bond risk premiums with machine learning. *The Review of Financial Studies*, *34*(2), 1046–1089.

Blitz, D., Huij, J., & Martens, M. (2011). Residual momentum. *Journal of Empirical Finance*, *18*(3), 506–521.

Blume, M. E., & Husic, F. (1973). Price, beta, and exchange listing. *The Journal of Finance*, *28*(2), 283–299.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, *31*(3), 307–327.

Bollerslev, T., Li, S. Z., & Zhao, B. (2020). Good volatility, bad volatility, and the cross section of stock returns. *Journal of Financial and Quantitative Analysis*, *55*(3), 751–781.

Bondt, W. F., & Thaler, R. (1985). Does the stock market overreact? *The Journal of Finance*, *40*(3), 793–805.

Boudoukh, J., Michaely, R., Richardson, M., & Roberts, M. R. (2007). On the importance of measuring payout yield: Implications for empirical asset pricing. *The Journal of Finance*, *62*(2), 877–915.

Boyer, B., Mitton, T., & Vorkink, K. (2010). Expected idiosyncratic skewness. *The Review of Financial Studies*, *23*(1), 169–202.

Bradshaw, M. T., Richardson, S. A., & Sloan, R. G. (2006). The relation between corporate financing activities, analysts' forecasts and stock returns. *Journal of Accounting and Economics*, *42*(1), 53–85.

Brunnermeier, M. K., Gollier, C., & Parker, J. A. (2007). Optimal beliefs, asset prices, and the preference for skewed returns. *American Economic Review*, *97*(2), 159–165.

Bryzgalova, S., Pelger, M., & Zhu, J. (2020). Forest through the trees: Building cross-sections of stock returns. *Available At SSRN 3493458*.

Cakici, N., Fieberg, C., Metko, D., & Zaremba, A. (2023). Machine learning goes global: Cross-sectional return predictability in international stock markets. *Journal of Economic Dynamics and Control*, *155*, 104725.

Cooper, M. J., Gulen, H., & Schill, M. J. (2008). Asset growth and the cross-section of stock returns. *The Journal of Finance*, *63*(4), 1609–1651.

Da, Z., & Warachka, M. (2011). The disparity between long-term and short-term forecasted earnings growth. *Journal of Financial Economics*, *100*(2), 424–442.

Daniel, K., & Titman, S. (2006). Market reactions to tangible and intangible information. *The Journal of Finance*, *61*(4), 1605–1643.

Datar, V. T., Naik, N. Y., & Radcliffe, R. (1998). Liquidity and stock returns: An alternative test. *Journal of Financial Markets*, *1*(2), 203–219.

Dechow, P. M., Sloan, R. G., & Soliman, M. T. (2004). Implied equity duration: A new measure of equity risk. *Review of Accounting Studies*, *9*(2-3), 197–228.

Dechow, P. M., Sloan, R. G., & Sweeney, A. P. (1995). Detecting earnings management. *The Accounting Review*, 193–225.

Diether, K. B., Malloy, C. J., & Scherbina, A. (2002). Differences of opinion and the cross section of stock returns. *The Journal of Finance*, *57*(5), 2113–2141.

Dichev, I. D. (1998). Is the risk of bankruptcy a systematic risk? *The Journal of Finance*, *53*(3), 1131–1147.

Dong, X., Li, Y., Rapach, D. E., & Zhou, G. (2022). Anomalies and the expected market return. *The Journal of Finance*, *77*(1), 639–681.

Eberhart, A. C., Maxwell, W. F., & Siddique, A. R. (2004). An examination of long-term abnormal stock returns and operating performance following r&d increases. *The Journal of Finance*, *59*(2), 623–650.

Eisfeldt, A. L., & Papanikolaou, D. (2013). Organization capital and the cross-section of expected returns. *The Journal of Finance*, *68*(4), 1365–1406.

Elgers, P. T., Lo, M. H., & Pfeiffer Jr, R. J. (2001). Delayed security price adjustments to financial analysts' forecasts of annual earnings. *The Accounting Review*, *76*(4), 613–632.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica: Journal of the Econometric Society*, 987–1007.

Engle, R. F., & Manganelli, S. (2004). Caviar: Conditional autoregressive value at risk by regression quantiles. *Journal of Business & Economic Statistics*, *22*(4), 367–381.

Fairfield, P. M., Whisenant, J. S., & Yohn, T. L. (2003). Accrued earnings and growth: Implications for future profitability and market mispricing. *The Accounting Review*, *78*(1), 353–371.

Fama, E. F., & French, K. R. (1992). The cross-section of expected stock returns. *The Journal of Finance*, *47*(2), 427–465.

Fama, E. F., & French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, *116*(1), 1–22.

Fama, E. F., & MacBeth, J. D. (1973). Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy*, *81*(3), 607–636.

Francis, J., LaFond, R., Olsson, P. M., & Schipper, K. (2004). Costs of equity and earnings attributes. *The Accounting Review*, *79*(4), 967–1010.

Frankel, R., & Lee, C. M. (1998). Accounting valuation, market expectation, and cross-sectional stock returns. *Journal of Accounting and Economics*, *25*(3), 283–319.

Frazzini, A., & Pedersen, L. H. (2014). Betting against beta. *Journal of Financial Economics*, *111*(1), 1–25.

George, T. J., & Hwang, C.-Y. (2004). The 52-week high and momentum investing. *The Journal of Finance*, *59*(5), 2145–2176.

Ghysels, E., Plazzi, A., & Valkanov, R. (2016). Why invest in emerging markets? the role of conditional return asymmetry. *The Journal of Finance*, *71*(5), 2145–2192.

Goyenko, R. Y., Holden, C. W., & Trzcinka, C. A. (2009). Do liquidity measures measure liquidity? *Journal of Financial Economics*, *92*(2), 153–181.

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, *33*(5), 2223–2273.

Gu, S., Kelly, B., & Xiu, D. (2021). Autoencoder asset pricing models. *Journal of Econometrics*, *222*(1), 429–450.

Gul, F. (1991). A theory of disappointment aversion. *Econometrica: Journal of the Econometric Society*, 667–686.

Hafzalla, N., Lundholm, R., & Matthew Van Winkle, E. (2011). Percent accruals. *The Accounting Review*, *86*(1), 209–236.

Hahn, J., & Lee, H. (2009). Financial constraints, debt capacity, and the cross-section of stock returns. *The Journal of Finance*, *64*(2), 891–921.

Harvey, C. R., Liu, Y., & Zhu, H. (2016). ... and the cross-section of expected returns. *The Review of Financial Studies*, *29*(1), 5–68.

Harvey, C. R., & Siddique, A. (2000). Conditional skewness in asset pricing tests. *The Journal of Finance*, *55*(3), 1263–1295.

Haugen, R. A., & Baker, N. L. (1996). Commonality in the determinants of expected stock returns. *Journal of Financial Economics*, *41*(3), 401–439.

Hawkins, E. H., Chamberlin, S. C., & Daniel, W. E. (1984). Earnings expectations and security prices. *Financial Analysts Journal*, *40*(5), 24–38.

Heston, S. L., & Sadka, R. (2008). Seasonality in the cross-section of stock returns. *Journal of Financial Economics*, *87*(2), 418–445.

Hirshleifer, D., Hou, K., Teoh, S. H., & Zhang, Y. (2004). Do investors overvalue firms with bloated balance sheets? *Journal of Accounting and Economics*, *38*, 297–331.

Hou, K., & Robinson, D. T. (2006). Industry concentration and average stock returns. *The Journal of Finance*, *61*(4), 1927–1956.

Hou, K., Xue, C., & Zhang, L. (2020). Replicating anomalies. *The Review of Financial Studies*, *33*(5), 2019–2133.

Huang, T., & Li, J. (2019). Option-implied variance asymmetry and the cross-section of stock returns. *Journal of Banking & Finance*, *101*, 21–36.

Chabi-Yo, F., Leisen, D. P., & Renault, E. (2014). Aggregation of preferences for skewed asset returns. *Journal of Economic Theory*, *154*, 453–489.

Chan, L. K., Lakonishok, J., & Sougiannis, T. (2001). The stock market valuation of research and development expenditures. *The Journal of Finance*, *56*(6), 2431–2456.

Chordia, T., Subrahmanyam, A., & Anshuman, V. R. (2001). Trading activity and expected stock returns. *Journal of Financial Economics*, *59*(1), 3–32.

Ikenberry, D., Lakonishok, J., & Vermaelen, T. (1995). Market underreaction to open market share repurchases. *Journal of Financial Economics*, *39*(2), 181–208.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *International Conference on Machine Learning*, 448–456.

Jegadeesh, N. (1990). Evidence of predictable behavior of security returns. *The Journal of Finance*, *45*(3), 881–898.

Jegadeesh, N., Kim, J., Krische, S. D., & Lee, C. (2004). Analyzing the analysts: When do recommendations add value? *The Journal of Finance*, *59*(3), 1083–1124.

Jegadeesh, N., & Titman, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of Finance*, *48*(1), 65–91.

Jiang, J., Kelly, B., & Xiu, D. (2023). (re-) imag (in) ing price trends. *The Journal of Finance*, *78*(6), 3193–3249.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 363–391.

Kaniel, R., Lin, Z., Pelger, M., & Van Nieuwerburgh, S. (2023). Machine-learning the skill of mutual fund managers. *Journal of Financial Economics*, *150*(1), 94–138.

Kelly, B., & Jiang, H. (2014). Tail risk and asset prices. *The Review of Financial Studies*, *27*(10), 2841–2871.

Kim, T.-H., & White, H. (2004). On more robust estimation of skewness and kurtosis. *Finance Research Letters*, *1*(1), 56–73.

Kimball, M. S. (1993). Standard risk aversion. *Econometrica: Journal of the Econometric Society*, 589–611.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *Arxiv preprint arXiv:1412.6980*.

Koenker, R. (2004). Quantile regression for longitudinal data. *Journal of Multivariate Analysis*, *91*(1), 74–89.

Kot, H. W., & Chan, K. (2006). Can contrarian strategies improve momentum profits. *Journal of Investment Management*, *4*(1).

La Porta, R. (1996). Expectations and the cross-section of stock returns. *The Journal of Finance*, *51*(5), 1715–1742.

Lakonishok, J., Shleifer, A., & Vishny, R. W. (1994). Contrarian investment, extrapolation, and risk. *The Journal of Finance*, *49*(5), 1541–1578.

Langlois, H. (2020). Measuring skewness premia. *Journal of Financial Economics*, *135*(2), 399–424.

Lee, C., & Swaminathan, B. (2000). Price momentum and trading volume. *The Journal of Finance*, *55*(5), 2017–2069.

Lesmond, D. A., Ogden, J. P., & Trzcinka, C. A. (1999). A new estimate of transaction costs. *The Review of Financial Studies*, *12*(5), 1113–1141.

Li, D. (2011). Financial constraints, r&d investment, and stock returns. *The Review of Financial Studies*, *24*(9), 2974–3007.

Li, Y., & Zhu, J. (2008). L 1-norm quantile regression. *Journal of Computational and Graphical Statistics*, *17*(1), 163–185.

Liu, F. (2023). Quantile machine learning and the cross-section of stock returns. *Available at SSRN 4491887*.

Lockwood, L., & Prombutr, W. (2010). Sustainable growth and stock returns. *Journal of Financial Research*, *33*(4), 519–538.

Loughran, T., & Wellman, J. W. (2011). New evidence on the relation between the enterprise multiple and average stock returns. *Journal of Financial and Quantitative Analysis*, *46*(6), 1629–1650.

Lyandres, E., Sun, L., & Zhang, L. (2007). The new issues puzzle: Testing the investment-based explanation. *The Review of Financial Studies*, *21*(6), 2825–2855.

Maas, A. L., Hannun, A. Y., Ng, A. Y., et al. (2013). Rectifier nonlinearities improve neural network acoustic models. *Proc. icml*, *30*(1), 3.

Manski, C. F. (1988). Ordinal utility models of decision making under uncertainty. *Theory and Decision*, *25*, 79–104.

McLean, R. D., & Pontiff, J. (2016). Does academic research destroy stock return predictability? *The Journal of Finance*, *71*(1), 5–32.

Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research*, *7*, 983–999.

Moon, S. J., Jeon, J.-J., Lee, J. S. H., & Kim, Y. (2021). Learning multiple quantiles with neural networks. *Journal of Computational and Graphical Statistics*, *30*(4), 1238–1248.

Moskowitz, T. J., & Grinblatt, M. (1999). Do industries explain momentum? *The Journal of Finance*, *54*(4), 1249–1290.

Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87–106.

Nieto, M. R., & Ruiz, E. (2016). Frontiers in var forecasting and backtesting. *International Journal of Forecasting*, *32*(2), 475–501.

Novy-Marx, R. (2010). Operating leverage. *Review of Finance*, *15*(1), 103–134.

Novy-Marx, R. (2012). Is momentum really momentum? *Journal of Financial Economics*, *103*(3), 429–453.

Novy-Marx, R. (2013). The other side of value: The gross profitability premium. *Journal of Financial Economics*, *108*(1), 1–28.

Ortiz-Molina, H., & Phillips, G. M. (2014). Real asset illiquidity and the cost of capital. *Journal of Financial and Quantitative Analysis*, *49*(1), 1–32.

Palazzo, B. (2012). Cash holdings, risk, and expected returns. *Journal of Financial Economics*, *104*(1), 162–185.

Parkinson, M. (1980). The extreme value method for estimating the variance of the rate of return. *Journal of Business*, 61–65.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., ... Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems 32* (pp. 8024–8035). Curran Associates, Inc.

Penman, S. H., Richardson, S. A., & Tuna, I. (2007). The book-to-price effect in stock returns: Accounting for leverage. *Journal of Accounting Research*, *45*(2), 427–467.

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

Pontiff, J., & Woodgate, A. (2008). Share issuance and cross-sectional returns. *The Journal of Finance*, *63*(2), 921–945.

Richardson, S. A., Sloan, R. G., Soliman, M. T., & Tuna, I. (2006). The implications of accounting distortions and growth for accruals and profitability. *The Accounting Review*, *81*(3), 713–743.

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

Soliman, M. T. (2008). The use of dupont analysis by market participants. *The Accounting Review*, *83*(3), 823–853.

Spiess, D. K., & Affleck-Graves, J. (1995). Underperformance in long-run stock returns following seasoned equity offerings. *Journal of Financial Economics*, *38*(3), 243–267.

Stock, J. H., & Watson, M. W. (2017). Twenty years of time series econometrics in ten pictures. *Journal of Economic Perspectives*, *31*(2), 59–86.

Takeuchi, I., Le, Q. V., Sears, T., & Smola, A. J. (2006). Nonparametric quantile estimation. *Journal of Machine Learning Research*, *7*, 1231–1264.

Thomas, J. K., & Zhang, H. (2002). Inventory changes and future returns. *Review of Accounting Studies*, *7*(2), 163–187.

Titman, S., Wei, K. J., & Xie, F. (2004). Capital investments and stock returns. *Journal of Financial and Quantitative Analysis*, *39*(4), 677–700.

Tobek, O., & Hronec, M. (2021). Does it pay to follow anomalies research? machine learning approach with international evidence. *Journal of Financial Markets*, *56*, 100588.

Tuzel, S. (2010). Corporate real estate holdings and the cross-section of stock returns. *The Review of Financial Studies*, *23*(6), 2268–2302.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., ... SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, *17*, 261–272.

Whited, T. M., & Wu, G. (2006). Financial constraints risk. *The Review of Financial Studies*, *19*(2), 531–559.

Yang, X., Zhu, Z., Li, D., & Zhu, K. (2024). Asset pricing via the conditional quantile variational autoencoder. *Journal of Business & Economic Statistics*, *42*(2), 681–694.

Zhang, X. (2006). Information uncertainty and stock returns. *The Journal of Finance*, *61*(1), 105–137.

Zhang, Y. (2006). Individual skewness and the cross-section of average stock returns. *unpublished paper, Yale University.*

Appendix

2.A List of the anomalies

Table 2.10. List of anomalies

Fundamental	
Accruals	
Accruals	Sloan (1996)
Change in Common Equity	Richardson et al. (2006)
Change in Current Operating Assets	Richardson et al. (2006)
Change in Current Operating Liabilities	Richardson et al. (2006)
Change in Financial Liabilities	Richardson et al. (2006)
Change in Long-Term Investments	Richardson et al. (2006)
Change in Net Financial Assets	Richardson et al. (2006)
Change in Net Non-Cash Working Capital	Richardson et al. (2006)
Change in Net Non-Current Operating Assets	Richardson et al. (2006)
Change in Non-Current Operating Assets	Richardson et al. (2006)
Change in Non-Current Operating Liabilities	Richardson et al. (2006)
Change in Short-Term Investments	Richardson et al. (2006)
Discretionary Accruals	Dechow, Sloan, and Sweeney (1995)
Growth in Inventory	Thomas and H. Zhang (2002)
Inventory Change	Thomas and H. Zhang (2002)
Inventory Growth	Belo and Lin (2011)
M/B and Accruals	Bartov and M. Kim (2004)
Net Working Capital Changes	Soliman (2008)
Percent Operating Accrual	Hafzalla et al. (2011)
Percent Total Accrual	Hafzalla et al. (2011)
Total Accruals	Richardson et al. (2006)
Intangibles	
$\triangle$ Gross Margin - $\triangle$ Sales	Abarbanell and Bushee (1998)
$\triangle$ Sales - $\triangle$ Accounts Receivable	Abarbanell and Bushee (1998)
$\triangle$ Sales - $\triangle$ Inventory	Abarbanell and Bushee (1998)
$\triangle$ Sales - $\triangle$ SG and A	Abarbanell and Bushee (1998)
Asset Liquidity	Ortiz-Molina and Phillips (2014)
Asset Liquidity II	Ortiz-Molina and Phillips (2014)
Cash-to-assets	Palazzo (2012)
Earnings Conservatism	Francis et al. (2004)
Earnings Persistence	Francis et al. (2004)
Earnings Predictability	Francis et al. (2004)
Earnings Smoothness	Francis et al. (2004)
Earnings Timeliness	Francis et al. (2004)
Herfindahl Index	Hou and Robinson (2006)

Hiring rate	Belo, Lin, and Bazdresch (2014)
Industry Concentration Assets	Hou and Robinson (2006)
Industry Concentration Book Equity	Hou and Robinson (2006)
Industry-adjusted Organizational Capital-to-Assets	Eisfeldt and Papanikolaou (2013)
Industry-adjusted Real Estate Ratio	Tuzel (2010)
Org. Capital	Eisfeldt and Papanikolaou (2013)
RD / Market Equity	L. K. Chan et al. (2001)
RD Capital-to-assets	D. Li (2011)
RD Expenses-to-sales	L. K. Chan et al. (2001)
Tangibility	Hahn and H. Lee (2009)
Unexpected RD Increases	Eberhart et al. (2004)
Whited-Wu Index	Whited and Wu (2006)
Investment	
△ CAPEX - △ Industry CAPEX	Abarbanell and Bushee (1998)
Asset Growth	Cooper et al. (2008)
Change Net Operating Assets	Hirshleifer et al. (2004)
Changes in PPE and Inventory-to-Assets	Lyandres et al. (2007)
Composite Debt Issuance	Lyandres et al. (2007)
Composite Equity Issuance (5-Year)	Daniel and Titman (2006)
Debt Issuance	D. K. Spiess and Affleck-Graves (1995)
Growth in LTNOA	Fairfield et al. (2003)
Investment	Titman et al. (2004)
Net Debt Finance	Bradshaw et al. (2006)
Net Equity Finance	Bradshaw et al. (2006)
Net Operating Assets	Hirshleifer et al. (2004)
Noncurrent Operating Assets Changes	Soliman (2008)
Share Repurchases	Ikenberry et al. (1995)
Total XFIN	Bradshaw et al. (2006)
Profitability	
Asset Turnover	Soliman (2008)
Capital Turnover	Haugen and Baker (1996)
Cash-based Operating Profitability	Ball et al. (2016)
Change in Asset Turnover	Soliman (2008)
Change in Profit Margin	Soliman (2008)
Earnings / Price	Basu (1977)
Earnings Consistency	Alwathainani (2009)
F-Score	Piotroski (2000)
Gross Profitability	Novy-Marx (2013)
Labor Force Efficiency	Abarbanell and Bushee (1998)
Leverage	Bhandari (1988)
O-Score (More Financial Distress)	Dichev (1998)
Operating Profits to Assets	Ball et al. (2016)
Operating Profits to Equity	Fama and French (2015)

Profit Margin	Soliman (2008)
Return on Net Operating Assets	Soliman (2008)
Return-on-Equity	Haugen and Baker (1996)
Z-Score (Less Financial Distress)	Dichev (1998)
Value	
Assets-to-Market	Fama and French (1992)
Book Equity / Market Equity	Fama and French (1992)
Cash Flow / Market Equity	Lakonishok et al. (1994)
Duration of Equity	Dechow, Sloan, and Soliman (2004)
Enterprise Component of Book/Price	Penman et al. (2007)
Enterprise Multiple	Loughran and Wellman (2011)
Intangible Return	Daniel and Titman (2006)
Leverage Component of Book/Price	Penman et al. (2007)
Net Payout Yield	Boudoukh et al. (2007)
Operating Leverage	Novy-Marx (2010)
Payout Yield	Boudoukh et al. (2007)
Sales Growth	Lakonishok et al. (1994)
Sales/Price	Barbee Jr et al. (1996)
Sustainable Growth	Lockwood and Prombutr (2010)
Market Friction	
11-Month Residual Momentum	Blitz et al. (2011)
52-Week High	George and Hwang (2004)
Amihud's Measure (Illiquidity)	Amihud (2002)
Beta	Fama and MacBeth (1973)
Betting against Beta	Frazzini and Pedersen (2014)
Bid-Ask Spread	Amihud and Mendelson (1986)
Cash Flow Variance	Haugen and Baker (1996)
Coefficient of Variation of Share Turnover	Chordia et al. (2001)
Coskewness	Harvey and Siddique (2000)
Downside Beta	Ang, Chen, and Xing (2006)
Earnings Forecast-to-Price	Elgers et al. (2001)
Firm Age	Barry and Brown (1984)
Firm Age-Momentum	X. Zhang (2006)
Idiosyncratic Risk	Ang, Hodrick, et al. (2006)
Industry Momentum	Moskowitz and Grinblatt (1999)
Lagged Momentum	Novy-Marx (2012)
Liquidity Beta 1	Acharya and Pedersen (2005)
Liquidity Beta 2	Acharya and Pedersen (2005)
Liquidity Beta 3	Acharya and Pedersen (2005)
Liquidity Beta 4	Acharya and Pedersen (2005)
Liquidity Beta 5	Acharya and Pedersen (2005)
Liquidity Shocks	Bali, Peng, et al. (2013)
Long-Term Reversal	Bondt and Thaler (1985)

Max	Bali, Cakici, and Whitelaw (2011)
Momentum	Jegadeesh and Titman (1993)
Momentum and LT Reversal	Kot and K. Chan (2006)
Momentum-Reversal	Jegadeesh and Titman (1993)
Momentum-Volume	C. Lee and Swaminathan (2000)
Price	Blume and Husic (1973)
Seasonality	Heston and Sadka (2008)
Seasonality 1 A	Heston and Sadka (2008)
Seasonality 1 N	Heston and Sadka (2008)
Seasonality 11-15 A	Heston and Sadka (2008)
Seasonality 11-15 N	Heston and Sadka (2008)
Seasonality 16-20 A	Heston and Sadka (2008)
Seasonality 16-20 N	Heston and Sadka (2008)
Seasonality 2-5 A	Heston and Sadka (2008)
Seasonality 2-5 N	Heston and Sadka (2008)
Seasonality 6-10 A	Heston and Sadka (2008)
Seasonality 6-10 N	Heston and Sadka (2008)
Share Issuance (1-Year)	Pontiff and Woodgate (2008)
Share Turnover	Datar et al. (1998)
Short-Term Reversal	Jegadeesh (1990)
Size	Banz (1981)
Tail Risk	Kelly and H. Jiang (2014)
Total Volatility	Ang, Hodrick, et al. (2006)
Volume / Market Value of Equity	Haugen and Baker (1996)
Volume Trend	Haugen and Baker (1996)
Volume Variance	Chordia et al. (2001)
I/B/E/S	
Analyst Value	Frankel and C. M. Lee (1998)
Analysts Coverage	Elgers et al. (2001)
Change in Forecast + Accrual	Barth and Hutton (2004)
Change in Recommendation	Jegadeesh, J. Kim, et al. (2004)
Changes in Analyst Earnings Forecasts	Hawkins et al. (1984)
Disparity between LT and ST Earnings Growth Forecasts	Da and Warachka (2011)
Dispersion in Analyst LT Growth Forecasts	Anderson et al. (2005)
Down Forecast	Barber et al. (2001)
Forecast Dispersion	Diether et al. (2002)
Long-Term Growth Forecasts	La Porta (1996)
Up Forecast	Barber et al. (2001)

2.B Alternative models

We describe the benchmark models used to evaluate the performance of our two-stage model and the accuracy of our distributional moment forecasts. We compare the forecasting performance of the two-stage model with the GARCH model and three different neural network architectures, as outlined in Table 2.4. Detailed descriptions of the neural networks and the GARCH model are provided in Sections 2.B.1 and 2.B.2, respectively.

2.B.1 Feed-forward neural networks

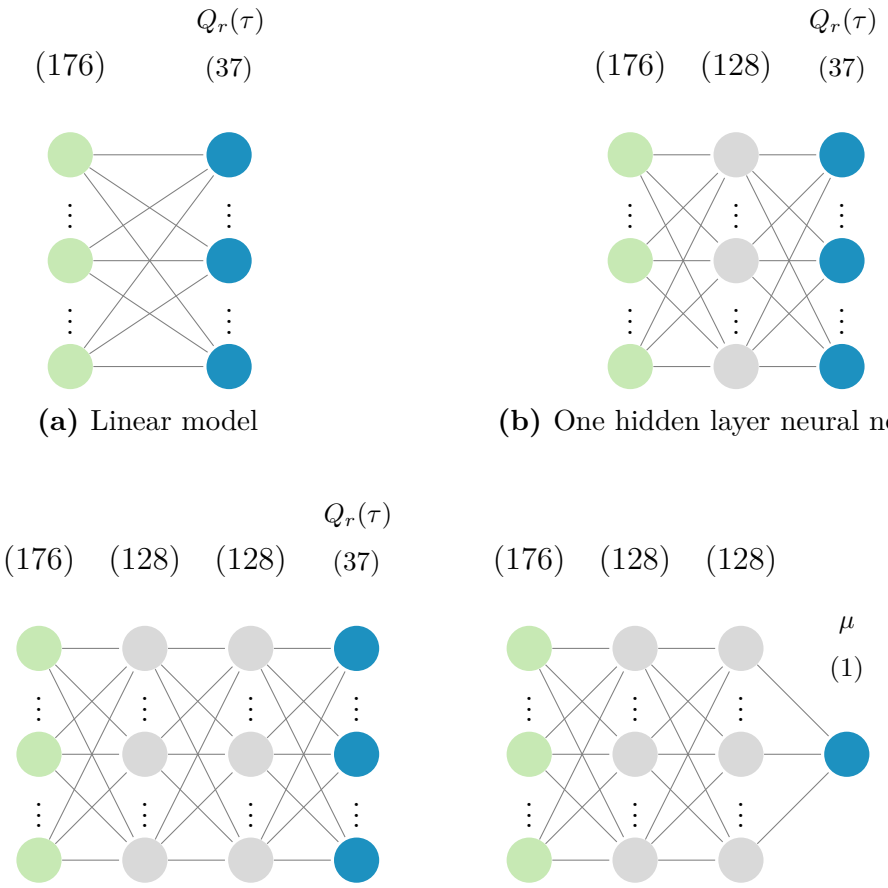
We use multi-head feed-forward neural networks with varying architectures as benchmarks for the two-stage model quantile forecasts. We consider a simple multi-output linear model without any hidden layers, neural networks with 1 hidden layer, and neural networks with 2 hidden layers as benchmarks for the two-stage model. Each model has 176 features and 37 outputs. The input features are the same as those for the standardized τ-quantiles network in the two-stage model. The 37 outputs are the τ-quantile forecasts of raw future stock returns for the same τ values as in the two-stage model. See Section 2.3.1 for the feature specification and Section 2.2 for the main model introduction.

Each neural network is trained using the aggregated multi-τ-quantile loss.

$$\mathcal{L}_{raw}^{\mathcal{T}} = \frac{1}{B}\frac{1}{K}\sum_{\tau\in\mathcal{T}}\sum_{i=1}^{B}\rho_{\tau}\left(r_{i,t} - \widehat{Q}_{r_{i,t}}(\tau)\right) \tag{2.8}$$

where $\widehat{Q}_{r_{i,t}}(\tau)$ is the forecast τ-quantile of the raw return of stock i at period t, B is the lot size and $K = 37$ is the number of τ in the $\mathcal{T}$ set.

We also use a single-head feed-forward neural network with 2 hidden layers and mean squared error loss function as a benchmark for mean forecasts comparison as in Table 2.5 and Table 2.8. This model also uses the same 176 features as the networks above. Architecture schemas for all alternative neural network models are shown in Figure 2.5.



(a) Linear model

(b) One hidden layer neural network

(c) Two hidden layers neural network

(d) Single output neural network with two hidden layers

Figure 2.5. Architectures of benchmark multi-head feed-forward neural networks

Figure 2.5a depicts the linear model, Figure 2.5b illustrates the neural network with one hidden layer, Figure 2.5c shows the neural network with two hidden layers, and Figure 2.5d presents the two hidden layers neural network with a single output node, which is used to generate mean forecasts directly using the mean squared error loss function. All networks have 176 input features; multi-output networks have 37 outputs, varying in the number of hidden layers and neurons.

Most of the hyper-parameters for all four benchmark neural networks are the same as for the two-stage model and are available in Table 2.15. The dropout rate is set to 0.2 and the learning rate to 0.0003, in line with the selected hyper-parameter for the two-stage model. Hyper-parameters that differ are the architecture in terms of the number of layers and neurons, the loss function, and other model-specific settings.

2.B.2 GARCH benchmark

In order to provide a valid benchmark for the two-stage model, we examine various specifications of the GARCH model. We utilize daily returns from all regions for the period from 1995 to 2018, corresponding to the time period and stock universe used for the main results in this paper. We base our estimations on daily returns using a rolling window of 36 months. Results are compared for both the full and liquid samples.

We consider several variants of GARCH(1,1) model with the standardized residuals distributed according to normal and t-distributions. Mean μ is either estimated or set equal to the risk-free rate at the end of the estimation window plus 5% divided by 252. Assuming these distributions, we simulate 100,000 paths of residuals for 22 days ahead and consequently derive the corresponding return paths. During this process, volatility is simulated using Equation 2.9

$$\sigma_{i,j}^2 = \omega + \alpha\epsilon_{i-1,j}^2 + \beta\sigma_{i-1,j}^2 \tag{2.9}$$

where ω, α, and β are the GARCH parameters, $\epsilon_{i-1,j}$ represents the previous day's residual, and $\sigma_{i-1,j}^2$ is the previous day's volatility. The cumulative return for the j-th simulated path is then calculated according to Equation 2.10.

$$r_j = \left(\prod_{i=1}^{22}\left(1 + \mu + \epsilon_{i,j}\sqrt{\sigma_{i,j}^2}\right)\right) - 1 \tag{2.10}$$

Finally, the variance forecast is taken as the sample standard deviation of all the cumulative returns over the 22 days, and the quantile forecast is obtained as the sample quantile of all the cumulative returns over the 22 days.

Table 2.11. Performance of various GARCH specifications
This table shows out-of-sample average quantile losses for different specifications of the GARCH model for both the full and liquid samples. The significance of the difference in average quantile losses between the GARCH(1,1) with Student's t-distribution and other specifications is captured by t-statistics adjusted for Newey-West standard errors with 12 lags. All models, unless specified, use the mean as a risk-free rate at the end of the estimation window plus 5%. GARCH(1,1) t-dist uses a t-distribution with an estimated number of degrees of freedom, and GARCH(1,1) normal dist uses a normal distribution. GARCH(1,1) fixed, t-dist models have the number of degrees of freedom fixed at 3, 4, or 5, and their auto-regressive parameters are fixed to $\alpha = 0.06$ and $\beta = 0.94$. Daily risk-free rate data are taken from the Fama-French data library.

Specification	Variable	Full Sample	Liquid Sample
GARCH(1, 1), t-dist	Avg Loss	0.02642	0.01887
GARCH(1, 1) fixed, t-dist 3 df	Avg Loss	0.02636	0.01906
	t-stat	-0.96	4.97
GARCH(1, 1) fixed, t-dist 4 df	Avg Loss	0.02641	0.01900
	t-stat	-0.17	4.25
GARCH(1, 1) fixed, t-dist 5 df	Avg Loss	0.02646	0.01901
	t-stat	1.06	4.53
GARCH(1, 1), normal dist	Avg Loss	0.02681	0.01899
	t-stat	15.25	10.30
GARCH(1, 1), t-dist, μ estimated	Avg Loss	0.02769	0.01930
	t-stat	5.22	4.23
GARCH(2, 2), t-dist	Avg Loss	0.02649	0.01893
	t-stat	2.10	1.26
EGARCH(1, 1, 1), t-dist	Avg Loss	0.02681	0.01913
	t-stat	3.38	2.25

For the t-distribution, besides estimating the number of degrees of freedom from the data, we also consider fixing the number of degrees of freedom to 3, 4, and 5. Furthermore, we try GARCH(2, 2) and EGARCH(1, 1, 1) specifications also with t-distributed standardized residuals, in both cases estimating degrees of freedom from the data.

Table 2.B.2 shows the out-of-sample average quantile losses for the considered GARCH models for both the full sample and the liquid sample. The best performance for the full sample is achieved by the GARCH(1, 1) model with t-distributed standardized residuals fixed to 4 degrees of freedom and auto-regressive parameters fixed to $\alpha = 0.06$ and $\beta = 0.94$.[15] GARCH(1, 1) specification without fixing the parameters performs similarly and achieves the best result for the liquid sample. Based on these results, we choose the GARCH(1, 1) with t-distributed standardized residuals and estimated number

[15]This specification of GARCH is widely used in practice and was coined by J.P Morgan in 1996.

of degrees of freedom to serve as the benchmark for the main empirical study in this paper.[16]

A more recent method for quantile prediction is MIDAS quantile regression introduced by Ghysels et al. (2016). MIDAS quantile regressions, in their simplest form, predict quantile as $Q = \alpha + \beta f(\gamma)$. α, β, and γ are three parameters to be estimated via non-linear OLS. $f(\gamma)$ is a weighted average of past absolute daily returns where the parameter γ determines the speed of weight decay for less recent returns. The simplest specification is quite similar to GARCH(1, 1) with its auto-regressive component. In contrast to GARCH, MIDAS quantile regression does not rely on any distributional assumptions, which, however, also means that it cannot be applied to the extreme quantiles being estimated in the present study. The range of quantities that can be estimated via MIDAS is limited by a number of observations for one stock. Three years of data for one stock allows, at most, an estimation of 0.0014 quantile, which is much larger than the minimum quantiles estimated by neural networks 0.00005.

[16]Note that the estimation of the GARCH process can fail or it can generate explosive forecasts. GARCH(1, 1) fixed, t-dist 4 df is used for these cases as it can always be computed. The results in Table 2.B.2 already reflect this for all the models.

2.C Additional empirical results

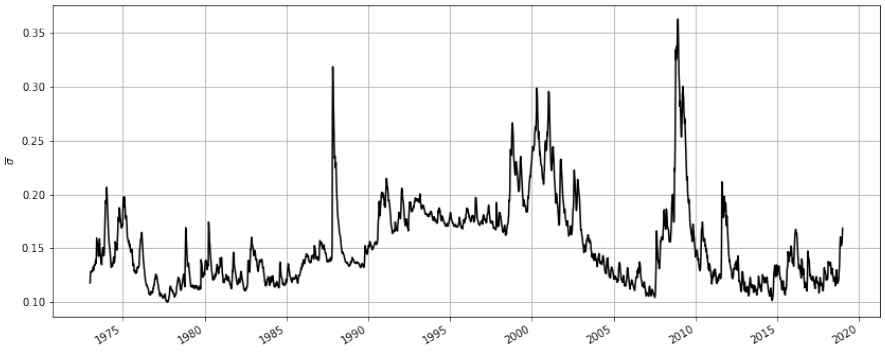


Figure 2.6. The cross-sectional average of volatilities as a scaling factor
This figure shows the evolution over time of the cross-sectional average of stock-level volatilities for the full sample of U.S. stocks. Stock-level volatilities are calculated as returns. of stock returns, exponential moving averages of squared daily returns with a smoothing factor of 0.94. The scaling factor is used to standardize stock returns to be used as a label for the standardized τ-quantile neural network or stage 1 of the two-stage model defined in Section 2.2.

Table 2.12. Average cross-sectional correlation between moments and median internationally

This table presents the average cross-sectional Spearman correlation between the forecasted first four moments and the median of the stock returns for the full and liquid samples in Europe, Japan, and the Asia Pacific. Medians are forecasted using the two-stage model defined in Section 2.2, and moments are forecasted using the quantiles-to-moments algorithm defined in Subsection 2.2.3. The predictions span the time period from 1995 to 2018.

Europe										
	Full Sample					Liquid Sample				
	Median	Mean	Variance	Skewness	Kurtosis	Median	Mean	Variance	Skewness	Kurtosis
Median	1.0	0.81	-0.58	-0.49	-0.43	1.0	0.97	-0.2	0.32	0.18
Mean	0.81	1.0	-0.25	-0.13	-0.16	0.97	1.0	-0.05	0.46	0.11
Variance	-0.58	-0.25	1.0	0.43	0.2	-0.2	-0.05	1.0	0.49	-0.47
Skewness	-0.49	-0.13	0.43	1.0	0.94	0.32	0.46	0.49	1.0	-0.1
Kurtosis	-0.43	-0.16	0.2	0.94	1.0	0.18	0.11	-0.47	-0.1	1.0
Japan										
	Full Sample					Liquid Sample				
	Median	Mean	Variance	Skewness	Kurtosis	Median	Mean	Variance	Skewness	Kurtosis
Median	1.0	0.83	-0.51	-0.39	-0.32	1.0	0.97	-0.27	0.27	0.23
Mean	0.83	1.0	-0.19	-0.02	-0.04	0.97	1.0	-0.12	0.42	0.17
Variance	-0.51	-0.19	1.0	0.31	0.04	-0.27	-0.12	1.0	0.48	-0.44
Skewness	-0.39	-0.02	0.31	1.0	0.93	0.27	0.42	0.48	1.0	0.0
Kurtosis	-0.32	-0.04	0.04	0.93	1.0	0.23	0.17	-0.44	0.0	1.0
Asia Pacific										
	Full Sample					Liquid Sample				
	Median	Mean	Variance	Skewness	Kurtosis	Median	Mean	Variance	Skewness	Kurtosis
Median	1.0	0.8	-0.65	-0.54	-0.43	1.0	0.96	-0.24	0.19	0.18
Mean	0.8	1.0	-0.29	-0.18	-0.16	0.96	1.0	-0.07	0.36	0.1
Variance	-0.65	-0.29	1.0	0.52	0.26	-0.24	-0.07	1.0	0.61	-0.42
Skewness	-0.54	-0.18	0.52	1.0	0.92	0.19	0.36	0.61	1.0	-0.1
Kurtosis	-0.43	-0.16	0.26	0.92	1.0	0.18	0.1	-0.42	-0.1	1.0

Table 2.13. Quantiles and long-short portfolios internationally
This table displays average monthly returns and annualized Sharpe ratios (SR) for decile long, short, and long-short equal-weighted portfolios over the period from 1995 to 2018. For each τ-quantile forecast, the long portfolio is formed by selecting the top 10% of stocks with the highest forecasts each month, the short portfolio by selecting the bottom 10% of stocks with the lowest forecasts, and the long-short portfolio by taking the difference between the long and short portfolios. Values in brackets are t-statistics adjusted for Newey-West standard errors with 12 lags. Results are obtained separately for full and liquid samples for Europe, Japan, and Asia Pacific.

	Europe											
	Full Sample						Liquid Sample					
	Long		Short		Long-Short		Long		Short		Long-Short	
τ	Mean	SR	Mean	SR	Mean	SR	Mean	SR	Mean	SR	Mean	SR
0.01	0.91 (3.92)	1.04	2.57 (3.61)	1.07	-1.66 (-2.41)	-0.77	0.98 (4.06)	0.99	0.06 (0.09)	0.02	0.92 (1.51)	0.39
0.05	0.86 (3.79)	1.00	2.32 (3.22)	0.96	-1.46 (-2.10)	-0.66	1.00 (4.31)	1.03	-0.01 (-0.01)	-0.00	1.01 (1.63)	0.42
0.1	0.88 (3.89)	1.02	2.13 (2.94)	0.88	-1.25 (-1.80)	-0.57	1.03 (4.53)	1.05	-0.05 (-0.07)	-0.02	1.09 (1.79)	0.46
0.2	0.95 (4.21)	1.11	1.70 (2.37)	0.71	-0.75 (-1.10)	-0.35	1.09 (4.53)	1.09	-0.14 (-0.20)	-0.05	1.23 (2.03)	0.55
0.3	1.08 (4.73)	1.23	1.19 (1.67)	0.51	-0.10 (-0.16)	-0.05	1.22 (4.76)	1.15	-0.28 (-0.38)	-0.10	1.50 (2.54)	0.70
0.4	1.69 (5.95)	1.59	0.35 (0.50)	0.16	1.35 (2.20)	0.74	1.39 (4.76)	1.12	-0.33 (-0.47)	-0.13	1.71 (3.12)	0.91
0.5	2.94 (6.73)	1.80	-0.42 (-0.62)	-0.20	3.35 (6.63)	2.48	1.49 (3.73)	0.88	-0.50 (-0.80)	-0.22	2.00 (4.65)	1.52
0.6	3.65 (6.89)	1.82	-0.67 (-1.17)	-0.37	4.32 (10.29)	3.48	1.49 (3.05)	0.70	-0.31 (-0.65)	-0.19	1.80 (5.33)	1.47
0.7	3.98 (6.39)	1.70	-0.57 (-1.69)	-0.51	4.55 (9.80)	2.55	1.29 (2.21)	0.52	0.23 (0.78)	0.21	1.06 (2.63)	0.59
0.8	4.12 (5.82)	1.63	0.17 (0.74)	0.20	3.95 (6.10)	1.74	0.97 (1.51)	0.36	0.54 (2.23)	0.53	0.43 (0.85)	0.20
0.9	3.70 (5.07)	1.46	0.45 (1.93)	0.54	3.25 (4.70)	1.41	0.70 (1.01)	0.25	0.64 (2.66)	0.64	0.06 (0.11)	0.03
0.95	3.54 (4.78)	1.40	0.57 (2.46)	0.66	2.98 (4.27)	1.31	0.54 (0.76)	0.19	0.67 (2.80)	0.68	-0.13 (-0.23)	-0.06
0.99	3.40 (4.69)	1.39	0.75 (2.98)	0.78	2.65 (3.97)	1.22	0.42 (0.58)	0.15	0.70 (2.96)	0.71	-0.28 (-0.48)	-0.12
	Japan											
	Full Sample						Liquid Sample					
	Long		Short		Long-Short		Long		Short		Long-Short	
τ	Mean	SR	Mean	SR	Mean	SR	Mean	SR	Mean	SR	Mean	SR
0.01	0.62 (2.24)	0.61	-0.04 (-0.06)	-0.01	0.66 (1.21)	0.25	0.50 (2.44)	0.43	-0.57 (-0.75)	-0.19	1.07 (1.68)	0.40
0.05	0.63 (2.22)	0.62	-0.16 (-0.22)	-0.05	0.79 (1.45)	0.30	0.53 (2.62)	0.46	-0.60 (-0.78)	-0.20	1.13 (1.76)	0.43
0.1	0.67 (2.31)	0.65	-0.25 (-0.35)	-0.08	0.92 (1.73)	0.35	0.61 (2.91)	0.53	-0.63 (-0.84)	-0.21	1.24 (1.96)	0.47
0.2	0.79 (2.63)	0.76	-0.39 (-0.55)	-0.13	1.18 (2.29)	0.46	0.71 (3.32)	0.60	-0.64 (-0.85)	-0.22	1.35 (2.16)	0.53
0.3	1.01 (3.27)	0.93	-0.53 (-0.75)	-0.18	1.54 (3.13)	0.63	0.88 (3.76)	0.71	-0.77 (-1.04)	-0.26	1.65 (2.78)	0.68
0.4	1.49 (4.33)	1.22	-0.72 (-1.03)	-0.25	2.21 (4.77)	1.01	1.08 (4.04)	0.76	-0.88 (-1.24)	-0.31	1.96 (3.72)	0.93
0.5	2.14 (4.62)	1.29	-0.96 (-1.40)	-0.35	3.10 (6.76)	1.84	1.16 (3.17)	0.67	-1.10 (-1.76)	-0.44	2.25 (5.87)	1.51
0.6	2.29 (4.09)	1.08	-1.16 (-1.92)	-0.50	3.45 (7.98)	2.96	1.05 (2.12)	0.50	-0.98 (-2.07)	-0.53	2.03 (6.33)	1.83
0.7	2.26 (3.58)	0.90	-1.03 (-2.25)	-0.70	3.28 (6.44)	1.98	0.74 (1.19)	0.30	-0.31 (-0.99)	-0.24	1.05 (2.16)	0.55
0.8	1.86 (2.69)	0.66	-0.44 (-1.30)	-0.40	2.30 (3.95)	0.98	0.34 (0.48)	0.12	0.02 (0.08)	0.02	0.32 (0.55)	0.13
0.9	1.34 (1.87)	0.44	-0.14 (-0.50)	-0.14	1.49 (2.49)	0.57	0.08 (0.11)	0.03	0.16 (0.69)	0.14	-0.08 (-0.14)	-0.03
0.95	1.11 (1.57)	0.36	-0.01 (-0.04)	-0.01	1.13 (1.94)	0.42	-0.04 (-0.05)	-0.01	0.20 (0.89)	0.18	-0.24 (-0.39)	-0.09
0.99	0.92 (1.33)	0.30	0.18 (0.73)	0.17	0.75 (1.34)	0.29	-0.15 (-0.20)	-0.05	0.24 (1.08)	0.21	-0.39 (-0.61)	-0.14
	Asia Pacific											
	Full Sample						Liquid Sample					
	Long		Short		Long-Short		Long		Short		Long-Short	
τ	Mean	SR	Mean	SR	Mean	SR	Mean	SR	Mean	SR	Mean	SR
0.01	1.01 (3.33)	0.95	3.57 (4.15)	1.11	-2.56 (-3.63)	-0.99	1.18 (3.77)	0.79	-0.50 (-0.64)	-0.14	1.68 (2.36)	0.54
0.05	1.02 (3.34)	0.98	2.94 (3.42)	0.92	-1.92 (-2.82)	-0.75	1.18 (3.74)	0.79	-0.61 (-0.77)	-0.17	1.79 (2.51)	0.58
0.1	1.05 (3.42)	1.01	2.47 (2.87)	0.78	-1.41 (-2.11)	-0.56	1.23 (3.92)	0.83	-0.67 (-0.84)	-0.19	1.90 (2.66)	0.62
0.2	1.10 (3.51)	1.05	1.64 (1.90)	0.53	-0.54 (-0.82)	-0.22	1.31 (4.20)	0.87	-0.70 (-0.88)	-0.20	2.01 (2.86)	0.68
0.3	1.28 (3.91)	1.16	0.71 (0.84)	0.23	0.57 (0.91)	0.25	1.52 (4.56)	0.96	-0.83 (-1.04)	-0.24	2.35 (3.39)	0.84
0.4	2.30 (5.37)	1.55	-0.18 (-0.21)	-0.06	2.48 (4.54)	1.24	1.81 (5.39)	1.07	-0.90 (-1.15)	-0.27	2.71 (4.65)	1.23
0.5	4.08 (6.56)	1.75	-0.99 (-1.21)	-0.35	5.07 (10.75)	3.06	1.86 (3.81)	0.76	-1.09 (-1.43)	-0.36	2.95 (6.28)	1.82
0.6	5.31 (6.81)	1.74	-1.38 (-1.82)	-0.55	6.69 (14.28)	3.55	1.54 (2.62)	0.51	-0.59 (-1.04)	-0.27	2.13 (7.48)	1.10
0.7	6.04 (7.17)	1.75	-1.33 (-2.43)	-0.82	7.37 (13.30)	2.84	1.29 (2.05)	0.39	0.43 (1.23)	0.28	0.86 (2.02)	0.31
0.8	6.02 (6.81)	1.67	-0.01 (-0.02)	-0.01	6.03 (8.91)	1.97	0.83 (1.21)	0.24	0.71 (2.16)	0.49	0.12 (0.22)	0.04
0.9	5.69 (6.24)	1.58	0.41 (1.30)	0.39	5.29 (7.41)	1.75	0.48 (0.68)	0.13	0.79 (2.46)	0.55	-0.30 (-0.51)	-0.09
0.95	5.41 (5.98)	1.55	0.56 (1.88)	0.53	4.85 (6.68)	1.67	0.27 (0.36)	0.07	0.85 (2.74)	0.59	-0.58 (-0.91)	-0.18
0.99	5.18 (5.54)	1.52	0.71 (2.19)	0.61	4.47 (5.83)	1.62	0.17 (0.22)	0.04	0.95 (3.06)	0.65	-0.78 (-1.17)	-0.24

Table 2.14. Moments value-weighted decile portfolios

This table shows average monthly returns for individual decile (1-10) and long-short value-weighted portfolios formed on the forecasted mean, median, volatility, skewness, and kurtosis. Results are reported for the full and liquid samples as well as individual regions, covering the period from 1995 to 2018. Values in brackets are t-statistics adjusted for Newey-West standard errors with 12 lags.

	Full Sample				Liquid Sample			
Mean								
Decile	USA	Europe	Japan	Asia Pacific	USA	Europe	Japan	Asia Pacific
1	-1.84 (-2.37)	-2.18 (-3.20)	-1.46 (-2.31)	-3.27 (-4.24)	-0.65 (-1.00)	-0.40 (-0.88)	-0.80 (-1.49)	-0.88 (-1.37)
2	-0.74 (-1.24)	-0.87 (-1.54)	-0.64 (-1.15)	-1.64 (-2.44)	0.28 (0.75)	0.37 (0.96)	-0.24 (-0.47)	0.17 (0.34)
3	0.10 (0.22)	-0.40 (-0.80)	-0.37 (-0.77)	-1.08 (-1.65)	0.60 (2.01)	0.44 (1.13)	-0.09 (-0.21)	0.34 (0.67)
4	0.49 (1.42)	0.48 (1.17)	-0.08 (-0.19)	-0.59 (-0.95)	0.57 (1.96)	0.66 (1.92)	0.09 (0.22)	0.50 (1.21)
5	0.74 (2.22)	0.73 (1.91)	0.28 (0.68)	0.23 (0.46)	0.86 (3.26)	0.69 (2.02)	0.19 (0.48)	0.72 (1.74)
6	0.89 (2.84)	0.94 (2.74)	0.31 (0.76)	0.73 (1.41)	0.91 (3.45)	0.75 (2.17)	0.31 (0.90)	1.10 (2.78)
7	1.28 (3.91)	1.25 (3.43)	0.62 (1.52)	1.30 (2.45)	1.05 (3.71)	0.83 (2.43)	0.45 (1.26)	1.26 (3.87)
8	1.38 (4.07)	1.46 (3.84)	0.91 (2.31)	1.68 (3.34)	1.14 (3.83)	1.09 (3.22)	0.57 (1.55)	1.32 (3.25)
9	1.84 (4.51)	1.89 (4.35)	1.24 (2.59)	2.16 (3.91)	1.17 (3.51)	1.27 (3.17)	1.08 (3.09)	1.50 (3.73)
10	2.46 (5.05)	2.36 (4.86)	1.88 (3.46)	3.50 (5.42)	2.11 (5.32)	1.45 (3.51)	1.10 (2.50)	1.85 (4.01)
10-1	4.30 (6.63)	4.54 (7.38)	3.34 (8.10)	6.77 (15.25)	2.76 (5.02)	1.85 (5.26)	1.90 (6.07)	2.73 (6.18)
Median								
Decile	USA	Europe	Japan	Asia Pacific	USA	Europe	Japan	Asia Pacific
1	-1.84 (-2.11)	-2.55 (-3.25)	-1.46 (-1.89)	-2.44 (-2.75)	-0.53 (-0.75)	-0.42 (-0.81)	-0.87 (-1.45)	-1.00 (-1.39)
2	-0.61 (-0.83)	-1.07 (-1.75)	-0.87 (-1.45)	-1.42 (-2.04)	0.11 (0.25)	0.26 (0.61)	-0.26 (-0.53)	-0.01 (-0.02)
3	-0.21 (-0.34)	-0.28 (-0.51)	-0.27 (-0.49)	-0.78 (-1.17)	0.48 (1.37)	0.48 (1.19)	-0.12 (-0.25)	0.36 (0.68)
4	0.18 (0.34)	0.09 (0.24)	-0.24 (-0.52)	-0.83 (-1.27)	0.60 (2.11)	0.59 (1.69)	-0.01 (-0.02)	0.44 (0.98)
5	0.50 (1.40)	0.44 (1.22)	0.03 (0.06)	-0.34 (-0.55)	0.70 (2.46)	0.54 (1.56)	0.29 (0.70)	0.40 (1.00)
6	0.51 (1.52)	0.40 (1.07)	-0.03 (-0.08)	-0.08 (-0.16)	0.91 (3.73)	0.69 (2.16)	0.12 (0.36)	1.10 (2.75)
7	0.84 (3.00)	0.77 (2.31)	0.26 (0.64)	0.18 (0.38)	1.06 (4.01)	0.90 (2.62)	0.41 (1.09)	1.04 (3.00)
8	0.79 (2.58)	0.77 (2.51)	0.41 (1.03)	0.92 (1.94)	1.02 (3.39)	1.11 (3.24)	0.48 (1.46)	1.39 (3.71)
9	1.21 (4.17)	1.15 (3.01)	0.67 (1.66)	1.26 (2.94)	1.27 (4.59)	1.21 (3.27)	0.91 (2.52)	1.43 (3.76)
10	1.76 (5.06)	1.58 (3.90)	1.05 (2.49)	1.82 (3.87)	1.86 (4.94)	1.43 (3.58)	1.08 (2.69)	1.85 (4.31)
10-1	3.59 (5.02)	4.13 (6.86)	2.51 (4.86)	4.26 (7.67)	2.39 (4.43)	1.85 (4.42)	1.95 (5.10)	2.86 (5.70)
Volatility								
Decile	USA	Europe	Japan	Asia Pacific	USA	Europe	Japan	Asia Pacific
1	0.84 (4.03)	0.78 (3.24)	0.29 (1.41)	0.83 (3.03)	0.81 (4.01)	0.72 (3.17)	0.23 (0.97)	0.69 (2.50)
2	0.97 (3.61)	0.72 (2.45)	0.22 (0.73)	1.03 (2.46)	0.90 (3.75)	0.73 (2.47)	0.22 (0.68)	0.97 (2.50)
3	0.93 (3.03)	0.74 (2.25)	0.28 (0.77)	1.15 (2.60)	0.87 (3.18)	0.75 (2.37)	0.25 (0.57)	1.05 (2.47)
4	1.15 (3.44)	0.71 (1.85)	0.28 (0.75)	0.84 (1.55)	1.02 (3.20)	0.71 (1.94)	0.04 (0.09)	1.24 (2.51)
5	0.83 (1.88)	0.77 (1.77)	0.20 (0.53)	0.86 (1.56)	0.98 (3.00)	0.66 (1.61)	0.41 (1.18)	1.02 (2.12)
6	1.10 (2.24)	0.74 (1.50)	0.32 (0.65)	0.81 (1.28)	0.97 (2.38)	0.67 (1.66)	0.05 (0.13)	0.73 (1.44)
7	0.83 (1.56)	0.80 (1.46)	0.41 (0.76)	0.21 (0.30)	0.80 (1.67)	0.54 (1.19)	0.13 (0.25)	0.58 (1.04)
8	0.71 (1.09)	0.01 (0.01)	0.41 (0.69)	-0.01 (-0.02)	1.00 (1.83)	0.58 (1.08)	0.46 (0.87)	0.79 (1.27)
9	0.64 (0.81)	0.05 (0.07)	0.66 (0.82)	0.06 (0.08)	0.69 (1.13)	0.81 (1.36)	0.53 (0.74)	0.78 (0.99)
10	0.16 (0.18)	-0.87 (-0.93)	-0.07 (-0.08)	1.07 (1.04)	0.32 (0.38)	0.13 (0.18)	-0.22 (-0.25)	-0.04 (-0.05)
10-1	-0.67 (-0.76)	-1.65 (-1.93)	-0.35 (-0.44)	0.24 (0.28)	-0.49 (-0.63)	-0.59 (-1.00)	-0.45 (-0.59)	-0.73 (-1.14)
Skewness								
Decile	USA	Europe	Japan	Asia Pacific	USA	Europe	Japan	Asia Pacific
1	0.79 (2.56)	0.65 (1.90)	0.22 (0.63)	0.78 (1.88)	0.55 (2.36)	0.33 (1.05)	0.06 (0.19)	0.78 (2.37)
2	1.01 (4.11)	1.05 (3.06)	0.35 (0.88)	1.13 (3.31)	0.72 (3.02)	0.44 (1.16)	-0.23 (-0.59)	0.59 (1.48)
3	0.96 (2.95)	0.97 (2.81)	0.39 (0.82)	0.82 (1.74)	0.80 (2.88)	0.61 (1.87)	0.14 (0.33)	0.88 (1.98)
4	0.81 (1.85)	0.48 (1.37)	0.19 (0.38)	0.70 (1.32)	0.77 (2.58)	0.74 (2.22)	0.23 (0.53)	0.87 (1.92)
5	0.82 (1.70)	0.60 (1.37)	0.50 (0.96)	0.40 (0.72)	0.86 (2.62)	0.78 (2.09)	0.32 (0.76)	1.14 (2.86)
6	0.80 (1.56)	0.50 (0.99)	0.08 (0.15)	0.79 (1.31)	0.99 (2.63)	0.72 (2.29)	0.61 (1.44)	1.09 (2.41)
7	0.95 (1.68)	0.11 (0.20)	0.40 (0.74)	1.90 (1.22)	1.15 (2.98)	0.88 (2.23)	0.56 (1.31)	1.26 (2.46)
8	0.81 (1.66)	0.50 (1.07)	0.39 (0.74)	0.46 (0.70)	1.18 (3.03)	0.78 (1.95)	0.74 (1.71)	1.18 (2.62)
9	0.85 (1.61)	-0.39 (-0.70)	0.42 (0.74)	0.32 (0.50)	1.31 (2.60)	1.10 (2.63)	0.60 (1.23)	1.20 (2.08)
10	0.89 (1.73)	-0.26 (-0.50)	-0.10 (-0.21)	0.35 (0.52)	0.75 (0.96)	0.71 (1.14)	0.31 (0.53)	0.83 (1.23)
10-1	0.10 (0.23)	-0.91 (-2.32)	-0.32 (-0.89)	-0.44 (-0.95)	0.19 (0.27)	0.38 (0.78)	0.25 (0.60)	0.05 (0.09)
Kurtosis								
Decile	USA	Europe	Japan	Asia Pacific	USA	Europe	Japan	Asia Pacific
1	1.05 (2.54)	0.77 (2.08)	0.44 (0.95)	0.85 (1.92)	0.50 (0.90)	0.38 (0.61)	0.11 (0.15)	0.57 (0.92)
2	0.92 (3.29)	0.80 (2.31)	0.28 (0.84)	0.88 (2.18)	0.80 (1.97)	0.71 (1.81)	0.38 (0.85)	0.85 (1.13)
3	0.74 (2.70)	0.85 (2.86)	0.25 (0.75)	0.71 (1.99)	0.81 (2.99)	0.48 (1.24)	0.30 (0.66)	0.71 (1.78)
4	0.82 (3.10)	0.68 (2.38)	0.06 (0.16)	0.66 (1.58)	0.75 (2.67)	0.60 (1.68)	0.12 (0.32)	1.03 (2.56)
5	0.80 (2.54)	0.61 (1.52)	0.26 (0.64)	0.50 (1.17)	0.93 (3.30)	0.69 (1.92)	0.06 (0.15)	0.60 (1.34)
6	0.83 (2.48)	0.42 (1.11)	0.37 (0.74)	1.10 (1.93)	0.91 (3.12)	0.76 (2.32)	0.26 (0.78)	0.85 (2.13)
7	0.77 (2.05)	0.82 (1.89)	0.28 (0.67)	0.87 (1.69)	0.90 (3.01)	0.71 (2.02)	0.14 (0.39)	1.01 (2.49)
8	0.93 (2.70)	0.16 (0.56)	0.35 (0.80)	0.79 (1.57)	1.01 (3.02)	0.76 (2.27)	0.18 (0.46)	0.82 (1.95)
9	1.05 (3.06)	-0.07 (-0.18)	0.23 (0.50)	0.75 (1.39)	1.11 (3.47)	0.81 (2.42)	0.40 (1.01)	1.21 (2.79)
10	0.94 (4.72)	-0.24 (-0.62)	-0.19 (-0.53)	0.51 (0.46)	0.80 (2.45)	0.55 (1.62)	0.22 (0.57)	1.02 (2.20)
10-1	-0.11 (-0.31)	-1.01 (-2.72)	-0.62 (-1.68)	-0.34 (-0.29)	0.29 (0.76)	0.17 (0.44)	0.11 (0.24)	0.46 (1.40)

2.D Hyperparameter search and validation

2.D.1 Hyperparameter search

We conduct a hyperparameter search using U.S. data from the period 1973 to 1994, employing weekly sampling and 22-day ahead returns, consistent with the setup for the main results in the paper. The period from 1973 to 1989 is utilized for training models with various hyperparameter values we are exploring. The period from 1990 to 1994 is used to evaluate the trained models and select the best hyperparameter values. Forecasts for the validation period are generated using the best hyperparameter values identified in the previous step. To maintain consistency with the train-test split used for the main results of the paper, we retrain the model annually during the validation period and use the most recent model for forecasting. Specifically, given a set of hyperparameter values, the model is initially trained on the 1973-1989 training sample and then rolled forward by one year four times, always using the model for the subsequent year's forecasts. We utilize the full sample given the greater number of available observations compared to the liquid sample, which provides more stable results. Lengthening the validation sample would result in a shorter training sample, which is undesirable. The training sample should cover at least one severe market-wide negative shock in equities, which occurred in 1987. This approach suggests there is some potential room for improvement in finding better hyperparameters for the liquid sample and/or for international regions.

It is not feasible to perform a full hyperparameter search for all possible hyperparameters of the neural networks. Therefore, we divide the hyperparameter search into two sequential steps and focus only on a subset of hyperparameters. The rest of the hyperparameters are fixed and are based on the optimal hyperparameters found previously in the literature, e.g., Gu et al. (2020) or Tobek and Hronec (2021). Table 2.15 shows the fixed hyperparameters and their values.

Table 2.15. Fixed hyperparameters
This table lists the hyperparameters that are fixed during the hyperparameter search. The number of epochs is set to $100 * (A/n)$, where n is the number of observations available for training and A is an adjustment constant equal to 3,000,000 for the full sample and 1,500,000 for the liquid sample, derived from the average number of observations in individual samples over time.

Hyperparameter	Value
Activation function	LeakyReLU
Batch size	8192
Batch normalization	True
Epochs	$100 * (A/n)$
Early-stopping patience	2 epochs
Early-stopping validation size	20%
Optimizer	Adam
Adam momentum	(0.9, 0.999)
Decay factor	1.0
Decay step size	100
L1 penalty for first layer	0.0001
Num. of networks in ensemble	20 liquid, 10 full sample

We start with a grid search for optimization algorithm hyperparameters, namely the learning rate (LR) and dropout rate (DR). We test values of 0.1, 0.001, and 0.0001 for LR and 0, 0.2, and 0.4 for DR, evaluating the average quantile loss for the validation sample. We use a two-stage feedforward neural network with two hidden layers, each with 128 neurons, as a starting point to perform the first round of hyperparameter search. The architecture was chosen to provide slightly higher capacity than neural networks used in Gu et al. (2020) or Tobek and Hronec (2021). Our evaluation criterion is the average quantile loss during the validation period from 1990 to 1994. Table 2.16 displays the results of the hyperparameter search.

Table 2.16. Hyperparameter search for learning rate and dropout
We conduct a hyperparameter search for the period from 1973 to 1994 using U.S. data only. We search for the optimal learning rate and dropout rate for the two-stage neural network. The rest of the hyperparameters are fixed as shown in Table 2.15. Loss is calculated as the average quantile loss for the validation period from 1990 to 1994.

Learning Rate	Dropout Rate	Loss
0.01	0.0	0.022827
0.01	0.2	0.022982
0.01	0.4	0.023135
0.001	0.0	0.022958
0.001	0.2	0.022821
0.001	0.4	0.022881
0.0001	0.0	0.022936
0.0001	0.2	0.022835
0.0001	0.4	0.022844
0.0003	0.2	**0.022778**
0.003	0.2	0.022914
0.0003	0.3	0.022815
0.0003	0.1	0.022829

The lowest average quantile loss is achieved for the learning rate of 0.0003 and dropout rate of 0.2, although dropout rates of 0.1 and 0.3 are not significantly worse than 0.2. Fixing the learning rate and dropout rate at the optimal values achieved in the first part of the hyperparameter search, we perform the second part of the hyperparameter search to find an optimal architecture for the two-stage feedforward neural network, where we try different numbers of neurons and layers in the individual stages of the network. Table 2.17 shows the results of the architecture hyperparameter search.

Table 2.17. Hyperparameter search for architecture
We search for the optimal number of neurons and layers in the individual stages of the two-stage feed-forward neural network. The loss function is the average quantile loss achieved during the validation period from 1990 to 1994.

Stage1	Stage2	Loss
2x128	8x1	**0.022778**
2x128	0x1	0.022815
2x128	16x1	0.022822
2x64	8x1	0.022829
2x256	8x1	0.022786
3x128	8x1	0.022810
1x128	8x1	0.022816
4x64	8x1	0.022872

A smaller or greater number of neurons than 8 in the second stage leads to worse results. A smaller number of neurons is preferred to prevent potential overfitting and to speed up learning. The network already needs to be heavily regularized to achieve satisfactory performance; it, therefore, should not require more learning capacity by adding more neurons.

The L1 and L2 penalties were selected in isolation as it is not computationally possible to add them as another hyper-parameter in the search. The chosen values proved to be extremely robust and to work well across all the network architectures. The L1 penalty for the first layer of the first stage and the L1/L2 penalty for the first layer of the second stage were selected based on inspection of estimated weights on the 1973-1994 training sample. The L1 penalty for the first layer of the first stage was selected using a one-layer network and inspecting how many weights are non-negative.[17] The goal was to induce some selection of input variables and to limit the effects of collinearity. The L1 penalty for the first stage is 0.0001, and it forces weights of about $100 - 150$ of the input variables to zero.[18]

The L1 and L2 penalties for the second stage were chosen solely to limit collinearity issues without any focus on variable selection. The penalty was

[17]Note that one layer neural network is just a standard quantile linear regression with L1 penalty on its weights that is estimated using stochastic gradient descent.

[18]The L1 penalty for the first stage is changed to 0.00001 for neural network specification, where raw returns are directly predicted from the first stage to reflect the smaller scale of the predicted variable.

chosen the smallest possible while still leading to no obvious collinearlity issues.[19] The L1 and L2 penalties for the second stage are the same at 0.00001. The L1 penalty for the second stage is approximately 10 times smaller than the L1 penalty for the first stage. The scale of the standardized stock returns, serving as a label for the first stage, is close to 1. The scale of the raw stock returns, serving as a label for the second stage, is roughly 10 times smaller. The L1 penalties for both stages are, therefore, consistent given the scale of the variables used as an input for the loss function.

2.D.2 Validation

Figure 2.7 illustrates the data split into training, validation, and testing sets for the U.S. and international samples. Hyperparameter optimization and model training are conducted solely on U.S. data, using international data only for out-of-sample evaluation.

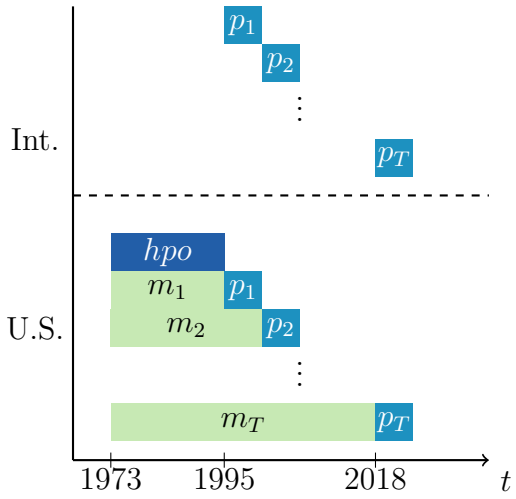


Figure 2.7. Validation sample-splitting schema
This figure shows how the data is split into training (m_i), validation (hpo) and testing (p_i) sets for the U.S. and international samples. All models are trained on the U.S. data and then used to generate out-of-sample predictions for both the U.S. and the international sample. Models are retrained each year using an expanding window of data reaching back to 1973. Predictions are generated monthly, always using the most up-to-date model available at the end of each month, e.g. p_1 predictions represent all monthly predictions during the period of year 1 and are generated using the model trained on m_1.

In order to maximize the number of observations available for training,

[19]The goal was to not have a mix of very large negative and positive weights, which are a symptom of the collinearity.

future returns are computed each week along with all the required inputs to the neural network. Future returns cover the next 22 business days[20], approximately one month of trading. This leads to approximately 4.3 times more partially overlapping observations available for training compared to using non-overlapping monthly data only[21]. Weekly sampled future returns are used only during the model training. Out-of-sample predictions are generated and evaluated using regular calendar months.

2.E Algorithms for approximating probability distribution function and moments from quantiles

Algorithm 1 contains the algorithm for deriving the probability density function from the quantiles. The algorithm is referred to as *quantiles-to-density* in the paper. Algorithm 2 contains the algorithm for deriving the moments of the probability distribution function. The algorithm is referred to as *density-to-moments* in the paper, and the algorithm below does not contain the adjustment of the moments described in Subsection 2.E.3.

2.E.1 Fallback to linear B-Splines

The liquid universe of stocks generally exhibits well-behaved density functions. Therefore, the fallback to the linear approximation of the density is rarely needed. However, it is sometimes necessary for the full sample due to the presence of infrequently traded stocks with small capitalization. There can be many days with zero trading activity for micro-cap stocks, especially in the historical period. Days with zero trading are a well-documented proxy for liquidity in the finance literature, formally introduced by Lesmond et al. (1999). See Goyenko et al. (2009) for an overview of other related liquidity proxies along with a comparison of their relative performance. This method could also be applied in bond returns forecasting where high illiquidity and zero-return observations are common. The presence of zero-return days can lead to a large probability mass around zero and, in extreme cases, to a discrete probability of a zero return. The continuous density function implied from the

[20]The count of business days here also includes holidays for simplicity.

[21]Using only monthly data during the training leads to inferior models because it can miss some reverting rare price changes.

cumulative distribution function is then not well-behaved, and the fallback to linear approximation becomes necessary.

For larger cap stocks where the density fallback was applied, the cause was manually checked. The density always exhibited a large mass around zero, indicating that the model was predicting a discrete probability of zero return in the following month. Fallback predictions for large-cap stocks were often associated with a corporate event, such as being close to official delisting or a recent trading halt. One such case is detailed in Figure 2.8, depicting forecasts for October 2014 for Idenix Pharmaceuticals, which was delisted just five days later. Idenix Pharmaceuticals received a tender takeover offer from Merck on June 9, 2014[22].

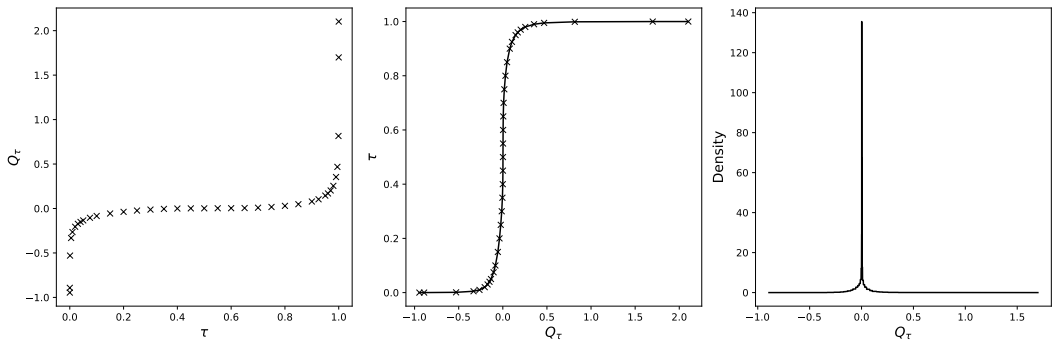


Figure 2.8. Example of density with a large mass around zero
This figure shows the derived probability density function for monthly return predictions for Idenix Pharmaceuticals in October 2014.

2.E.2 Central moments

We can derive central moments from the non-central moments using the Equation 2.11.

$$\begin{aligned} \sigma &= m_2 - m_1^2 \\ \text{skewness} &= \frac{m_3 - 3m_1\sigma - m_1^3}{\sigma^{3/2}} \\ \text{kurtosis} &= \frac{m_4 - 4m_1m_3 + 6m_1^2m_2 - 3m_1^4}{\sigma^2} \end{aligned} \tag{2.11}$$

where m_i for $i \in 1, 2, 3, 4$ are non-central moments and σ is the variance.

[22]The announcement of the acquisition is available here https://www.merck.com/news/merck-to-acquire-idenix/.

Algorithm 1 PDF

Require: $\mathbf{x}, \mathbf{y}$
$gp = 100$ ▷ Density of grid points
$d_{min} = 1e-5$ ▷ Minimum value for density function
$eps = 1e-4$ ▷ Epsilon
$z = 0$ ▷ Keep only the highest quantile with -1 return
for i in 1 to length(x) - 1 **do**
  if $x[i+1] = -1$ **then**
    $z = z + 1$
  end if
end for
$x = x[z:]$
$y = y[z:]$
if $min(y) < 0.05$ **then** ▷ Get 0.9 and 0.1 quantiles for outlier detection
  $x_{Q10} = x[y = 0.1]$
else
  $x_{Q10} = x[3]$
end if
$x_{Q90} = x[y = 0.9]$
for i in 1 to length(x) - 1 **do** ▷ Fix inconsistent quantile predictions
  if $x[i+1] < x[i] + eps$ **then**
    $x[i+1] = x[i] + eps$
  end if
end for
allocate empty vector $\tilde{\mathbf{x}}$ ▷ Create grid of denser x values
for i in 1 to length(x) - 3 **do**
  $x_l = x[i+1]$
  $x_u = x[i+2]$
  add $range(x_l, x_u, (x_u - x_l)/gp)$ to $\tilde{\mathbf{x}}$
end for
drop duplicates in $\tilde{\mathbf{x}}$
sort values of $\tilde{\mathbf{x}}$ from the smallest to the largest
fit cubic B-spline interpolation S on $\mathbf{x}$ and $\mathbf{y}$
use S to fit its first derivative on $\tilde{\mathbf{x}}$: $\mathbf{d} = S'(\tilde{\mathbf{x}})$
if $min(\mathbf{d}[(\tilde{\mathbf{x}} \geq x_{Q10}) \& (\tilde{\mathbf{x}} \leq x_{Q90})]) < d_{min}$ **then** ▷ Density fallback
  fit linear B-spline interpolation S on $\mathbf{x}$ and $\mathbf{y}$
  use S to fit its first derivative on $\tilde{\mathbf{x}}$: $\mathbf{d} = S'(\tilde{\mathbf{x}})$
end if
$\mathbf{d}[\mathbf{d} < d_{min}] = d_{min}$ ▷ Enforce minimum density
$w = min(\mathbf{d}[(\tilde{\mathbf{x}} <= x_{Q10})])$ ▷ Enforce monotonicity in the tails
$v = argmin(\mathbf{d}[(\tilde{\mathbf{x}} <= x_{Q10})])$
$\mathbf{d}[:v] = w$
$w = min(\mathbf{d}[(\tilde{\mathbf{x}} >= x_{Q90})])$
$v = argmin(\mathbf{d}[(\tilde{\mathbf{x}} >= x_{Q90})])$
$\mathbf{d}[v:] = w$
$x_{min} = min(x)$ ▷ get max and min of x
$x_{max} = max(x)$
if $min(\tilde{\mathbf{x}}) < -1$ **then** ▷ Truncate the distribution at -1 if needed
  $\mathbf{d} = \mathbf{d}[\tilde{\mathbf{x}} \geq -1]$
  $\tilde{\mathbf{x}} = \tilde{\mathbf{x}}[\tilde{\mathbf{x}} \geq -1]$
  $x_{min} = -1$
end if
$P_l = 1 - S(x_{max})$ ▷ Compute discrete probabilities at tails
$P_u = S(x_{min})$

2.E.3 Adjustment of moments

The required moment adjustments are estimated using simulated data with known precise theoretical moments. One million non-central Student's t distributions are generated with various degrees of freedom and noncentrality parameter[23]. Degrees of freedom are randomly selected from {5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 30, 40, 50, 60, 70, 80, 90, 100, 1000, 10000} with replacement and equal probability of each outcome. The noncentrality parameter is randomly selected using a uniform distribution of $[-0.5, 5]$ interval focusing mostly on positive skewness, as is observed in the empirical equities data. The scale of the generated distributions is set to 0.1 to limit the effects of the fact that stock returns cannot be smaller than -1, which is reflected in the density-to-moments algorithm but not in the theoretical distributions. Precise theoretical moments of the distribution are saved, and quantiles of the distribution are fed into the "naive" density-to-moments algorithm to compute predicted moments. Generated distributions with theoretical kurtosis larger than 20 are discarded as they would create strong leverage points and skew the estimated parameters to capture these extremes.

The values of the parameters are then estimated using OLS, where precise theoretical moments are predicted using moments produced by the "naive" density-to-moments algorithm. The simple linear regression is preferred here as it provides a satisfactory fit while being easy to interpret and more likely to generalize to other distributions.[24]

The following adjustment is proposed to decrease bias in the moments estimated from the "naive" density-to-moments algorithm.

$$\begin{aligned} k_e &= k - 3 \\ \tilde{v} &= v \times (1.0023 - 0.0021 \times s + 0.0022 \times k_e) \\ \tilde{s} &= 0.9950 \times s + 0.0261 \times s^2 + 0.0107 \times k_e \\ \tilde{k} &= 3 + 1.4185 \times k_e + 0.0466 \times k_e^2 - 0.7395 \times s \end{aligned} \tag{2.12}$$

Where v, s, k are variance, skewness, and kurtosis from the "naive" density-to-moments algorithm, $\tilde{v}, \tilde{s}, \tilde{k}$ are their adjusted counterparts, and k_e is excess

[23]Non-central Student's t distribution is defined as $\frac{Y+nc}{\sqrt{V/df}}$ where Y is a standard normal distribution, V is an independent chi-square random variable, nc is noncentrality parameter, and df are degrees of freedom. See scipy.stats.nct documentation for more details on its implementation.

[24]It is possible to further improve the fit by employing neural networks or higher order polynomials and interactions between the exogenous variables if required.

kurtosis. All of the estimated parameters are significant with t-distribution over 100 in absolute terms. The fit is high for all the regressions with R^2 over 0.9.

Variance v requires only a small adjustment, which mainly accounts for the fact that it is underestimated for heavy-tailed distributions with high kurtosis. The adjustment for skewness s is larger to account for its underestimation when either skewness or kurtosis of the distribution is high. Finally, as expected, kurtosis requires the largest adjustment due to the difficulty of capturing it with the "naive" density-to-moments algorithm.

The adjusted density-to-moments algorithm is next tested on probability distributions that are close to those actually observed in the predictions of the future return distribution. The predicted densities from neural networks have positive skewness and are heavy-tailed with positive excess kurtosis. A range of distributions is tested in Table 2.18. Normal distribution is the starting point due to its wide use in quantitative finance. t-distribution with small degrees of freedom then serves as a benchmark for a more heavy-tailed distribution. Finally, the non-central Student's t-distribution is the closest match to the distribution actually observed in the data.

Table 2.18. Adjustment fit of density-to-moments algorithm
m, v, s, k denote mean, variance, skewness, and kurtosis estimated using the "naive" (non-adjusted) density-to-moments algorithm defined in Subsection 2.E.3, m_t, v_t, s_t, k_t corresponding theoretical values and $\tilde{v}, \tilde{s}, \tilde{k}$ are generated using adjusted density-to-moments algorithm, i.e. values are adjusted using equations 2.12.

Dist	m_t	m	v_t	v	$\tilde{v}$	s_t	s	$\tilde{s}$	k_t	k	$\tilde{k}$
Normal	0.000	-0.000	1.000	0.998	1.001	0.000	-0.000	-0.000	3.000	2.980	2.971
t: df=10	0.000	-0.000	1.250	1.244	1.250	0.000	-0.000	0.009	4.000	3.852	4.242
t: df=6	0.000	-0.000	1.500	1.487	1.497	0.000	-0.000	0.023	6.000	5.161	6.282
t: df=5	0.000	-0.000	1.667	1.646	1.662	0.000	-0.001	0.035	9.000	6.360	8.293
nct: df=5, nc=1	0.119	0.119	1.919	1.892	1.914	1.266	1.116	1.200	13.321	8.385	11.164
nct: df=6, nc=3	0.345	0.345	3.072	3.035	3.076	1.832	1.686	1.823	12.991	9.691	13.330
nct: df=5, nc=4	0.476	0.476	5.698	5.593	5.736	2.718	2.372	2.644	29.832	15.827	27.104

Table 2.18 documents a good fit for lower moments and even for kurtosis post-adjustment. The computed mean is almost identical to the theoretical mean, as is the computed variance. Skewness benefits from the moment adjustment for heavy-tailed positively skewed distribution while exhibiting reasonable fit even without the adjustment. Kurtosis is the hardest to fit precisely, but the adjustment is quite precise even for the cases with the highest theoretical kurtosis. This section has, therefore, confirmed that the proposed moment

estimation technique from the quantiles of the distribution yields reasonable results for empirical analysis.

Algorithm 2 Moments

Require: $\mathbf{x}, \mathbf{y}, P_l, P_u, x_{min}, x_{max}$
$m_0, m_1, m_2, m_3, m_4 = 0, 0, 0, 0, 0$
for i in 1 to length(x) - 1 **do**
    $x_1 = \mathbf{x}[i]$
    $x_2 = \mathbf{x}[i+1]$
    $y_1 = \mathbf{y}[i]$
    $y_2 = \mathbf{y}[i+1]$
    $b = (y_2 - y_1)/(x_2 - x_1)$
    $a = y_1 - bx_1$
    $m_0 = m_0 + ax_2 + 1/2bx_2^2 - ax_1 - 1/2bx_1^2$
    $m_1 = m_1 + 1/2ax_2^2 + 1/3bx_2^3 - 1/2ax_1^2 - 1/3bx_1^3$
    $m_2 = m_2 + 1/3ax_2^3 + 1/4bx_2^4 - 1/3ax_1^3 - 1/4bx_1^4$
    $m_3 = m_3 + 1/4ax_2^4 + 1/5bx_2^5 - 1/4ax_1^4 - 1/5bx_1^5$
    $m_4 = m_4 + 1/5ax_2^5 + 1/6bx_2^6 - 1/5ax_1^5 - 1/6bx_1^6$
end for
$m_1 = m_1 + P_l x_{min} + P_u x_{max}$ ▷ Add discrete probabilities in the tails
$m_2 = m_2 + P_l x_{min}^2 + P_u x_{max}^2$
$m_3 = m_3 + P_l x_{min}^3 + P_u x_{max}^3$
$m_4 = m_4 + P_l x_{min}^4 + P_u x_{max}^4$
$m_1 = m_1/m_0$ ▷ Normalize probability measure to 1
$m_2 = m_2/m_0$
$m_3 = m_3/m_0$
$m_4 = m_4/m_0$
$variance = m_2 - m_1^2$
$skewness = (m_3 - 3m_1 var - m_1^3)/var^{3/2}$
$kurtosis = (m_4 - 4m_1 m_3 + 6m_1^2 m_2 - 3m_1^4)/var^2$

2.E.4 Two-stage quantile neural network pseudo-algorithm

Algorithm 3 provides a pseudo-algorithm for the forecasting of quantiles of stock returns via a two-stage quantile neural network.

Algorithm 3 Quantile forecast via two-stage neural network

Require: $x_{i,t}$ (Stock-level features), z_t (market-wide signals), $\sigma_{i,t}$ (stock-level volatilities)
Require: NN1 (Standardized τ-quantiles subnetwork), NN2 (Market volatility sub-network)

▷ **Stage 1: Standardized Quantile Estimation**

1: Compute the cross-sectional average of stock-level volatilities:

$$\overline{\sigma}_t \leftarrow \frac{1}{N_t} \sum_{i=1}^{N_t} \sigma_{i,t}$$

2: Compute standardized quantile forecasts using neural network NN1:

$$Q_{\widetilde{r}_{i,t+1}}(\tau_k | x_{i,t}) \leftarrow \text{NN1}(x_{i,t})$$

3: Clip standardized quantile forecasts at -100%:

$$Q_{r_{i,t+1}}(\tau_k) \leftarrow \max\left(Q_{r_{i,t+1}}(\tau_k), -1\right)$$

▷ **Stage 2: Rescaling and Market Volatility Adjustment**

4: Convert standardized quantile forecasts to raw scale:

$$Q^{'}_{r_{i,t+1}}(\tau_k) \leftarrow Q_{\widetilde{r}_{i,t+1}}(\tau_k) \cdot \overline{\sigma}_t$$

5: Compute market-wide scaling factor using neural network NN2:

$$\hat{\sigma}_t^M \leftarrow \text{NN2}(z_t)$$

6: Apply market volatility adjustment to obtain final quantile forecasts:

$$Q_{r_{i,t+1}}(\tau_k) \leftarrow Q_{r'_{i,t+1}}(\tau_k) \times \widehat{\sigma}_t^M$$

7: Clip final quantile forecasts at -100%:

$$Q_{r_{i,t+1}}(\tau_k) \leftarrow \max\left(Q_{r_{i,t+1}}(\tau_k), -1\right)$$

The first-stage neural network, NN1, maps stock-level features to standardized quantile forecasts as $\text{NN1}(x_{i,t}) = g^{(L)} \circ g^{(L-1)} \circ \cdots \circ g^{(1)}(x_{i,t})$. The second-stage neural network, NN2, maps market-wide signals to the market-wide scaling factor as $\text{NN2}(z_t) = g^{(2)} \circ g^{(1)}(z_t)$.

Each activation function $g^{(\ell)}$ follows $g^{(\ell)}(u) := g_\ell\Big(W^{(\ell)}u + b^{(\ell)}\Big)$, where $W^{(\ell)}$ are weight matrices and $b^{(\ell)}$ are bias vectors. The activation functions used are ReLU: $g_\ell(u) = \max\{u, 0\}$ or tanh: $g_\ell(u) = \tanh(u)$.

2.E.5 Addressing quantile crossing problem via penalization

Our two-stage quantile neural network approach does not explicitly enforce the non-crossing constraint between quantiles. Despite the absence of explicit constraint enforcement, the two-stage quantile neural network model naturally captures the monotonicity requirement of quantile predictions since there are no quantiles that are crossed in our quantile forecasts. If the quantile crossing problem is present in the model or its alternatives, it could be addressed by introducing a penalty term in the loss function.

$$\widehat{Q}_r(\tau_j) \leq \widehat{Q}_r(\tau_{j+1}), \quad \text{for} \quad \tau_j < \tau_{j+1}, \tag{2.13}$$

it is possible to introduce an additional penalty term in the loss function.

A simple way to enforce the non-crossing constraint is to introduce a penalty whenever a crossing occurs:

$$\mathcal{L}_{\text{cross}} = \lambda \sum_{i=1}^{B} \sum_{j=1}^{K-1} \left[\max\left(0, \widehat{Q}_{r_{i,t}}(\tau_j) - \widehat{Q}_{r_{i,t}}(\tau_{j+1})\right)\right]^2 \tag{2.14}$$

where $\widehat{Q}(X_i, \tau_j)$ is the estimated quantile function at level τ_j, λ is a hyper-parameter controlling the strength of the penalty, and the $\max(0, \cdot)$ function ensures that penalties are applied only when a crossing occurs.

The full multi-quantile loss function incorporating the penalty is:

$$\begin{aligned} \mathcal{L}^{\mathcal{T}}_{\text{multi-cross}} = & \frac{1}{BK} \sum_{k=1}^{K} \sum_{i=1}^{B} \rho_{\tau_k}\left(r_{i,t} - \widehat{Q}_{r_{i,t}}(\tau_k)\right) \\ & + \lambda \sum_{i=1}^{B} \sum_{j=1}^{K-1} \left[\max\left(0, \widehat{Q}_{y_i}(\tau_j) - \widehat{Q}_{r_{i,t}}(\tau_{j+1})\right)\right]^2 \end{aligned} \tag{2.15}$$

An interesting alternative approach to prevent quantile crossing is proposed by Moon et al. (2021), who introduce a monotonic transformation using a reparametrization technique. Rather than penalizing violations, they enforce monotonicity by construction. Their method defines each quantile in terms of the previous one plus a positive increment, which is ensured by applying the exponential function:

$$q_{\tau_j}(X) = q_{\tau_1}(X) + \sum_{k=2}^{j} \exp(h_k(X)), \tag{2.16}$$

where $q_{\tau_1}(X)$ represents the lowest quantile level being estimated, and $h_k(X)$ are unconstrained functions modelled by neural networks. By applying the exponential transformation to the outputs $h_k(X)$, the method guarantees that $\exp(h_k(X)) > 0$ for all k, thereby ensuring that $q_{\tau_j}(X) < q_{\tau_{j+1}}(X)$ for all j. This approach eliminates the need for penalty terms and hyperparameter tuning of λ, as the non-crossing constraint is satisfied by construction. The neural network can be trained using standard quantile loss functions while maintaining the monotonicity property across all quantile levels. This construction is particularly advantageous in high-dimensional settings where penalty-based approaches may struggle to enforce constraints consistently across the entire feature space.

2.E.6 2019-2023 out-of-sample results in the U.S.

To validate the robustness of our proposed two-stage neural network approach beyond the primary sample period, we conduct a comprehensive out-of-sample evaluation spanning 2019-2023. This period presents a distinctive testing ground, encompassing, for example, strong exogenous shocks such as the COVID-19 crisis. We examine the generalizability of our methodology across the main metrics reported in the main text, mainly mean return forecasting accuracy, distributional (quantile) forecasting precision, volatility prediction, and profitability of the long-short portfolios.

Table 2.19 extends our analysis to the post-sample period (2019-2023), providing evidence on the generalizability of our main findings. Consistent with the 1995-2018 results, the two-stage quantile neural network (2sNN) approach utilizing quantile regression delivers superior predictive performance, with R^2 values of 1.00% and 1.80% for the full and liquid samples, respectively. This outperforms both the two-stage NN median forecast (R^2 of 0.20% and 1.72%) and the direct mean prediction from the two-hidden-layer neural network (R^2 of 0.30% and 0.13%). While the Diebold-Mariano test statistics indicate that these differences are not statistically significant at conventional levels, the consistent pattern across sample periods reinforces the economic significance of our findings.

Table 2.19. Mean forecast evaluation in the U.S. 2019-2023. This table shows the R^2 of the mean and median forecasts of the next 22 days of stock returns for the full and liquid samples across all regions. The mean prediction based on the two-stage quantile model (2sNN) is compared with the direct mean prediction from the two-hidden-layer neural network (2hNN MSE - Mean) and the median prediction from the two-stage model (2sNN - Median). Results are obtained separately for full and liquid samples in the U.S. covering the period from 2019 to 2023. Models are retrained annually using the expanding window. Note that losses are multiplied by 100, and values in parentheses are Diebold-Mariano test statistics for the difference between the model in a row and the two-stage NN mean. Standard errors are adjusted using the Newey-West standard deviation estimator with 12 lags.

Full Sample			Liquid Sample		
2sNN Mean	2sNN Median	2hNN MSE Mean	2sNN Mean	2sNN Median	2hNN MSE Mean
1.00	0.20	0.30	1.80	1.72	0.13
	(-1.11)	(-1.49)		(-0.36)	(-1.58)

Table 2.20 presents out-of-sample volatility forecasting results for 2019-2023, complementing our main analysis. The two-stage neural network methodology demonstrates marked improvements over GARCH forecasts across both sample compositions. In the comprehensive dataset, the neural approach yields MAD and RMSE values of 4.40% and 7.39%, substantially outperforming the GARCH model's 6.65% and 12.23%. This pattern persists, although with reduced magnitude, in the liquid securities subset (MAD: 3.23% vs. 3.95%; RMSE: 5.49% vs. 7.33%). Diebold-Mariano statistics of 7.72 and 4.25 for the respective samples confirm these differences are highly significant at conventional levels.

Table 2.20. Volatility forecast evaluation in the U.S. 2019-2023. This table shows the mean absolute deviation (MAD), root mean squared error (RMSE), and Diebold Mariano test statistic (DM) comparing the corresponding losses of the two competing models for the 22-day volatility forecasts. Volatility forecasting based on two-stage quantile forecasting (2sNN) is compared with GARCH model forecasting. Results are obtained separately for full and liquid samples in the U.S. covering the period from 2019 to 2023. Neural network models are retrained annually using an expanding window. Diebold-Mariano test statistics are computed for the difference between the GARCH and 2sNN model volatility forecasts. Note that losses are multiplied by 100, and the standard errors used in the Diebold-Mariano test statistics are adjusted using the Newey-West standard deviation estimator with 12 lags.

Measure	Full Sample		Liquid Sample	
	2sNN	GARCH	2sNN	GARCH
MAD	4.40	6.65	3.23	3.95
RMSE	7.39	12.23	5.49	7.33
DM		(7.72)		(4.25)

Table 2.21. Out-of-sample quantile forecasts evaluation in the U.S. 2019-2023. This table shows the average out-of-sample quantile cross-section loss for quantile forecasts from the two-stage model and alternative models. The results are obtained separately for the full and liquid samples covering the out-of-sample period from 2019 to 2023. Neural network models are trained annually using the expanding window. The average loss is calculated as the time series average of the cross-sectional averages of the average quantile loss across all $\tau \in \{0.00005, 0.0001, 0.001, 0.005,\ 0.01, \ldots, 0.05,\ 0.075,\ 0.1, 0.15, 0.2, \ldots, 0.8, 0.85, 0.9, 0.925,\ 0.95, \ldots 0.99, 0.995, 0.999, 0.9999, 0.99995\}$ as in (Eq. 2.7) for the 2sNN model, as well as benchmark neural network models (the linear model (LNN), the one-layer model (1hNN), the two-layer model (2hNN)) and the benchmark GARCH model. Note that losses are multiplied by 100, and the t-statistics in parentheses are computed for the average loss difference between the given model specification and the two-stage benchmark model using Newey-West standard errors with 12 lags.

Specification	Variable	Full Sample	Liquid Sample
GARCH	L_{avg}	2.799	1.906
	(t-stat)	(3.42)	(3.15)
LNN	L_{avg}	2.783	1.889
	(t-stat)	(4.39)	(2.83)
1hNN	L_{avg}	2.752	1.887
	(t-stat)	(3.02)	(1.85)
2hNN	L_{avg}	2.753	1.886
	(t-stat)	(1.67)	(1.18)
2sNN	L_{avg}	2.730	1.872

Table 2.21 examines out-of-sample quantile forecasting performance during 2019-2023, reinforcing the findings from our core analysis. The two-stage neural network consistently achieves the lowest average quantile loss (L_{avg}) across both sample specifications, outperforming all benchmark models. For the full sample, our approach yields an average loss of 2.730, compared to the next best performer, 1hNN, with 2.752. This pattern holds for the liquid sample, where the two-stage model registers 1.872 against 2hNN's 1.886. The performance hierarchy remains relatively consistent across both samples, with the two-stage approach followed by increasingly simpler neural architectures, while GARCH exhibits the highest average loss (2.799 and 1.906). The statistical significance of these differences varies, with robust t-statistics when comparing against simpler models (3.42-4.39 for GARCH and LNN), but diminishing when evaluating more complex alternatives (1.67-1.18 for 2hNN).

Table 2.22. Mean and median portfolios. This table shows average monthly returns and annualised Sharpe ratios for decile long-short equal-weighted portfolios based on the median and mean predictions from the two-stage quantile NN model (2sNN), the mean prediction from the two-hidden-layer neural network (2hNN MSE - Mean) and the median prediction from the two-stage NN model (2sNN - Median). The results are shown for the full and liquid samples, covering the period from 2019 to 2023. The long portfolio is formed by taking the top 10% of stocks with the highest forecast each month, and the short portfolio is formed by taking the bottom 10% of stocks with the lowest forecast each month. Values in parentheses are t-statistics adjusted for Newey-West standard errors with 12 lags. The significance of the difference in mean returns is tested using t-statistics with Newey-West standard errors with 12 lags.

	Full Sample			Liquid Sample		
	2sNN Mean	2sNN Median	2hNN MSE Mean	2sNN Mean	2sNN Median	2hNN MSE Mean
Avg Ret	4.67	3.11	4.45	1.44	1.48	1.30
(t-stat)		(1.13)	(0.40)		(-0.09)	(0.54)
SR	2.61	1.13	3.02	0.93	0.77	0.89

Table 2.22 presents the economic value of our forecasting approaches through long-short portfolio performance during 2019-2023. For the full sample, the two-stage NN mean-based strategy delivers the highest average monthly return of 4.7%, though its Sharpe ratio (2.61) is surpassed by the 2hNN MSE mean approach (3.02). The median-based forecast generates worse performance (3.11% return, 1.13 Sharpe), though differences across strategies lack statistical significance. The liquid sample exhibits substantially diminished performance across all methods, with returns ranging from 1.30% to 1.48% and Sharpe ratios between 0.77 and 0.93. Despite this, the two-stage NN approach maintains relative efficacy, especially in less liquid securities where limits to arbitrage may preserve predictability. These results are noteworthy as they represent out-of-sample performance during the volatile period encompassing the COVID-19 pandemic and underscore robust capacity of our approach to capture distributional characteristics across varying market conditions.

Chapter 3

Portfolio selection under quantile preferences
Out-of-sample perspective

Abstract

This paper investigates the out-of-sample performance of portfolios formed by investors with quantile preferences. Unlike traditional approaches that rely on expected utility, quantile preferences focus on maximizing the τ-quantile of future portfolio returns, where $\tau \in (0, 1)$ reflects the investor's risk aversion. Lower τ values correspond to higher risk aversion. We observe significant differences in portfolio weight diversification and turnover both among portfolios optimized for different τ-quantiles and when compared to global minimum variance portfolios. For risk-averse investors, we show that constructing a global minimum variance portfolio results in a higher out-of-sample τ-quantile of portfolio returns than direct τ-quantile maximization significantly more often than in half of the considered datasets. Our findings are robust in various empirical and simulated datasets, different estimation window sizes, and both daily and monthly return frequencies. Through Monte Carlo simulations, we trace the out-of-sample underperformance of τ-quantile maximizers to small sample bias, which causes the selected portfolio weights to deviate from the optimal weights under a known closed-form solution.

This chapter was co-authored with Jozef Baruník. We gratefully acknowledge support from the Czech Science Foundation under the 19-28231X (EXPRO) project. This research was also supported by the Grant Agency of Charles University (grant no. 316521) We appreciate helpful comments from participants at various seminars, workshops and conferences.

3.1 Introduction

Every investor faces a decision on how to allocate wealth across different assets. Modern portfolio theory, pioneered by Markowitz (1952) and Roy (1952) and anchored in the expected utility framework, has dominated this domain for decades. Yet this framework exhibits well-documented shortcomings, notably its failure to account for empirically observed investor behaviours such as loss aversion, preferences for positively skewed returns, and ambiguity aversion (Barberis 2013; Gilboa and Schmeidler 1989; Kahneman and Tversky 1979; Rabin 2013). Furthermore, the empirical characteristics of asset returns—pronounced fat-tails, asymmetric distributions, and heteroskedasticity—present substantial challenges to modern portfolio theory, which rests upon assumptions of return normality (Cont 2001). As Rabin (2013) demonstrates, reasonable parametrisations of expected utility imply implausible degrees of risk aversion over larger stakes, generating what is now termed the "Rabin paradox." These substantial limitations have stimulated research into alternative frameworks better suited to capturing the nuanced complexity of investor preferences and decision-making under uncertainty.

Addressing these fundamental challenges, our paper contributes to the literature in several ways. We provide a comprehensive study of the out-of-sample performance of portfolios formed by investors with quantile preferences, an area previously unexplored in the empirical literature. Our analysis reveals that sample-based τ-quantile maximization does not necessarily translate into the highest out-of-sample τ-quantile of portfolio returns. For risk-averse investors (those maximizing low τ-quantile), forming a portfolio as a global minimum variance portfolio achieves higher out-of-sample τ-quantile of portfolio returns than τ-quantile maximization. Using Monte Carlo simulations, we demonstrate that the source of inferior τ-quantile maximizers' out-of-sample performance is a small sample bias leading to errors in the selected portfolio weights. Additionally, we document that selected portfolio weights differ across different τs as well as compared to global minimum variance portfolios in terms of weight diversification and turnover. We leverage the Mixed Integer Linear Programming (MILP) methodology developed by Benati and Rizzi (2007) to reformulate the τ-quantile maximization problem, thereby making it computationally feasible for a realistic number of assets. These findings provide novel evidence for both portfolio selection and quantile preferences literature.

These contributions are grounded in the theoretical framework of quantile

preferences, which was introduced by Manski (1988) and formally extended by Chambers (2009), Rostek (2010), de Castro and Galvao (2019), and de Castro and Galvao (2022). This framework shifts focus towards quantiles of the distribution of future payoffs rather than expected utility, aligning with experimental evidence suggesting that quantile-based decision rules better describe the behaviour of up to 50% of individuals compared to expected utility (de Castro, Galvao, et al. 2022). In a single period setting, investors with quantile preferences maximize the τ-quantile of future portfolio returns, where risk aversion is captured directly by the parameter $\tau \in [0, 1]$ rather than by the concavity of a utility function. The smaller the τ, the greater the risk aversion. This framework offers several theoretical advantages: it provides a more intuitive representation of risk preferences, naturally accommodates non-normal return distributions, and aligns with institutional risk management practices where quantile-based measures have become standard (Artzner et al. 1999; Jorion 2007).

The theoretical appeal of quantile preferences extends further as they connect multiple strands of financial economics. At $\tau = 0.5$, the framework reduces to median maximization, relating to robust statistics literature. As τ approaches zero, quantile maximization converges towards worst-case scenario minimization, connecting to robust optimization and maximin preferences (Garlappi et al. 2007; Gilboa and Schmeidler 1989). For τ values near one, the framework links with prospect theory's emphasis on potential gains (Barberis 2013) and the literature on lottery-like investments (Barberis and Huang 2008).

Despite these theoretical advantages, portfolio selection under quantile preferences presents significant challenges. Unlike mean-variance optimization with its well-established analytical solutions, quantile maximization is generally non-differentiable and non-convex (Castro et al. 2022). The non-linearity of the quantile function prevents the interchange of differentiation and quantile operators, complicating analytical derivations. From a computational perspective, the problem involves order statistics, which creates substantial complexity even for moderate numbers of assets. From the practical perspective, an investor with quantile preferences needs to select a portfolio conditional on the sample of past returns, which is non-trivial since the naive formulation explicitly contains the quantile as an order statistic, leading to high computational complexity.

We are motivated by Castro et al. (2022), who develop a theoretical model for optimal portfolio selection of an investor with quantile preferences. They formally establish various properties of such portfolios, including weight di-

versification for τ-quantiles below the median and weight concentration for τ-quantiles above the median. To demonstrate their approach, they provide an in-sample empirical example of a three-asset portfolio selection problem between stocks, bonds, and cash using grid-search to find optimal weights. However, using grid search to find optimal weights is only feasible for a small number of assets, and including nonlinear and nonconvex metric, such as quantile, into an optimization problem is not trivial as is well known from the Value at Risk (VaR) literature, (Alexander et al. 2006; Gaivoronski and Pflug 2005; Natarajan et al. 2009). Benati and Rizzi (2007) show that the mean-VaR optimization problem is NP-hard and propose a Mixed Integer Linear Programming reformulation (MILP) of the problem. Crucially, they show that MILP formulation is computationally feasible for a medium number of assets. We leverage this approach and reformulate the τ-quantile maximization problem as a MILP problem.

Our analysis of out-of-sample performance yields several important findings. We show that sample-based τ-quantile maximization does not necessarily translate into the highest out-of-sample τ-quantile of portfolio returns. For risk-averse investors maximizing low τ-quantiles, the global minimum variance portfolio frequently achieves superior out-of-sample performance. Through Monte Carlo simulations, we demonstrate that this counter-intuitive result stems from estimation error and small-sample bias in portfolio weights. These findings connect our work to the literature on estimation error in portfolio selection (DeMiguel et al. 2009; Kan and Zhou 2007) and suggest potential limitations to the practical implementation of quantile preferences.

We also document systematic differences in portfolio characteristics across the spectrum of τ values. Portfolios optimized for different quantiles exhibit varying degrees of weight diversification, concentration, and turnover, with important implications for transaction costs and rebalancing strategies. These differences are not only substantial between portfolios optimized for different τ values but also when compared to traditional benchmark models such as global minimum variance portfolios. These empirical patterns confirm and extend the theoretical predictions of Castro et al. (2022) regarding the relationship between quantile preferences and portfolio characteristics.

While our findings highlight the unique characteristics of quantile-based portfolio selection, they also situate this approach within a broader landscape of non-traditional portfolio methods. Beyond the classical mean-variance framework. For instance, Rockafellar and Uryasev (2000) proposed portfolio optimiza-

tion based on minimizing Conditional Value-at-Risk (CVaR), which provides superior handling of tail risks compared to variance-based measures. In a similar vein addressing estimation uncertainty, Goldfarb and Iyengar (2003) developed robust optimization techniques that explicitly account for parameter estimation errors. Alternative performance metrics have also inspired novel portfolio construction methods. Keating and Shadwick (2002) introduced the Omega ratio as a comprehensive performance measure considering the entire return distribution rather than just the first two moments, particularly valuable when returns exhibit non-normality. The Maximum Diversification approach proposed by Choueifaty and Coignard (2008) aims to maximize the ratio of weighted average volatilities to portfolio volatility, offering a different perspective on risk management.

Behavioural finance has likewise influenced portfolio theory. Building on the psychological foundations laid by Kahneman and Tversky (1979), the behavioural portfolio theory developed by Shefrin and Statman (2000) incorporates psychological biases and mental accounting into the investment decision process. More recent methodological innovations include the Entropy Pooling framework from Meucci (2008), which provides a flexible approach for blending prior views with empirical data, and the Hierarchical Risk Parity technique developed by Lopez de Prado (2016), which addresses numerical instability in optimization through machine learning clustering techniques. Focusing specifically on extreme market conditions, Bernard et al. (2018) explored portfolio selection based on minimizing tail dependencies, concentrating on extreme co-movements rather than standard correlation measures. Our quantile maximization framework contributes to this growing body of alternative approaches that address recognized limitations of the standard mean-variance paradigm through more nuanced risk metrics and decision criteria.

The remainder of this paper is organised as follows. Section 3.2 formally introduces quantile preferences and their theoretical foundations. Section 3.3 describes the portfolio optimization problem of an investor with quantile preferences, including our Mixed Integer Linear Programming formulation. Section 3.4 examines the out-of-sample performance using simulated data, and Section 3.5 presents empirical results from real-world datasets.

3.2 Quantile preferences

Building on our discussion in the introduction, we now define quantile preferences.

Let X and Y be random variables and $u(.)$ be a continuous and strictly increasing utility function. Given a fixed $\tau \in (0,1)$, a τ-quantile preference $\succeq^{\tau}$ of X over Y is defined as

$$X \succeq^{\tau} Y \iff Q_{\tau}[u(X)] \geq Q_{\tau}[u(Y)] \tag{3.1}$$

where $Q_{\tau}[X] = \inf\{x \in \mathbb{R} : F_X(x) \geq \tau\}$ is the τ-quantile of X, and F_X is the cumulative distribution function of X. The τ-quantile represents the threshold value that X exceeds with probability $(1-\tau)$.

Since utility function $u(.)$ in Equation 3.1 is increasing and the quantile operation is invariant with respect to monotonic transformation, quantiles of utility-transformed variables X and Y are ordered the same as quantiles of the underlying variables X and Y:

$$Q_{\tau}[u(X)] \geq Q_{\tau}[u(Y)] \iff Q_{\tau}[X] \geq Q_{\tau}[Y] \tag{3.2}$$

This invariance property established by Manski (1988) renders the utility function redundant in Equation 3.1, marking a fundamental departure from traditional risk-return frameworks. This parsimony represents a significant theoretical advantage: while expected utility theory requires specifying both a utility function and subjective probabilities, quantile preferences require only a single parameter τ that captures the essential preference characteristics.

Quantile preferences emerge naturally when we consider decision-making under uncertainty with incomplete preference specification. Chambers (2009) demonstrate that quantile preferences satisfy a set of foundational axioms, including:

1. *Monotonicity*: If $X \geq Y$ almost surely, then $X \succeq^{\tau} Y$ for all $\tau \in (0,1)$.

2. *Continuity*: The preference relation $\succeq^{\tau}$ is continuous with respect to the topology of convergence in distribution.

3. *Law invariance*: If X and Y have the same probability distribution, then $X \sim^{\tau} Y$.

4. *Comonotonic independence*: For comonotonic random variables X, Y, and Z, and $\alpha \in (0,1)$, $X \succeq^{\tau} Y$ implies $\alpha X + (1-\alpha)Z \succeq^{\tau} \alpha Y + (1-\alpha)Z$.

Rostek (2010) further establishes that quantile maximization is the only decision rule that simultaneously satisfies ordinal covariance, probabilistic sophistication, and a modified form of the betweenness property. This theoretical uniqueness strengthens the case for quantile preferences as a foundational decision framework.

In quantile preferences, the degree of risk aversion is captured directly by the parameter τ. An investor with $\succeq^{\tau_a}$ preferences is more risk-averse than an investor with $\succeq^{\tau_b}$ preferences if and only if $\tau_a < \tau_b$. This implies that the most risk-averse investor can be represented with τ approaching 0 (focusing on worst-case scenarios) and the most risk-loving investor with τ approaching 1 (focusing on best-case scenarios). This intuitive parametrization of risk preferences offers easier interpretability than approaches based on utility function curvature.

de Castro and Galvao (2022) formalise this relationship by showing that when $\tau < 0.5$, a decision-maker exhibits risk aversion in the sense of preferring a certain outcome to a risky prospect with the same expected value. When $\tau = 0.5$, the decision-maker is risk-neutral for symmetric distributions, and when $\tau > 0.5$, the decision-maker exhibits risk-seeking behaviour.

This parametrization aligns with empirical evidence from behavioural economics, which suggests that individuals often evaluate risky prospects based on specific quantiles rather than complete probability distributions (Tversky and Kahneman 1992). The parameter τ can also be interpreted as the investor's confidence level or degree of optimism/pessimism about future outcomes.

The quantile framework also provides natural connections to established risk measures in finance. For risk-averse investors with low values of τ, quantile maximization relates directly to Value-at-Risk minimization, a standard risk management tool in financial institutions. For investors with $\tau = 0.5$, quantile preferences reduce to median maximization, which has been recognized for its robustness properties in the presence of outliers or asymmetric distributions (Bassett Jr et al. 2004).

Beyond individual quantiles, an important concept in the quantile-based analysis is the quantile-preserving spread, which generalizes the mean-preserving spread concept from the expected utility theory to the quantile framework. As defined by Machina (1982), a mean-preserving spread increases risk while keeping the mean constant. Similarly, a quantile-preserving spread, introduced

by Manski (1988), is a transformation that increases dispersion while preserving a specific quantile.

Formally, random variable Y is a τ-quantile-preserving spread of random variable X if:

$$Q_\tau[Y] = Q_\tau[X] \quad \text{and} \quad \int_0^\tau (Q_\alpha[Y] - Q_\alpha[X])d\alpha \leq 0 \leq \int_\tau^1 (Q_\alpha[Y] - Q_\alpha[X])d\alpha \tag{3.3}$$

This definition implies that Y has more probability mass in the tails while maintaining the same τ-quantile as X. The mathematical expression quantifies how Y exhibits greater dispersion below the τ-quantile (the left integral condition) and above the τ-quantile (the right integral condition).

Manski (1988) demonstrates that an investor with τ-quantile preferences is indifferent between X and Y since they have the same τ-quantile, but an investor with $\tau' < \tau$ preferences would prefer X over Y, while an investor with $\tau' > \tau$ preferences would prefer Y over X. This result establishes a critical link between quantile preferences and attitudes toward dispersion in different regions of the distribution.

Given these implications of quantile preferences for investor behaviour, we now turn to the practical implementation of these concepts in portfolio management. The theoretical framework established above provides the foundation for a quantile-based approach to asset allocation that can accommodate asymmetric and fat-tailed return distributions. In the following section, we formalize the portfolio selection problem under quantile preferences and develop a computationally feasible approach to solving it for realistic investment universes.

3.3 Portfolio optimization via quantile maximization

Given a risk-aversion parameter τ, an investor with quantile preferences aims to maximize τ-quantile of portfolio returns. The portfolio choice can be formulated as the following constrained optimization problem

$$\max_{\mathbf{w}\in[0,1]^n} Q_\tau[P_\mathbf{w}], \quad \text{s.t.} \quad \sum_{i=1}^n w_i = 1. \tag{3.4}$$

$\mathbf{w} \in [0,1]^n$ is a vector of portfolio weights for N assets in the investment universe, Q_τ is the τ-quantile of portfolio returns $P_\mathbf{w}$, where $P_\mathbf{w} = \sum_{i=1}^N w_i r_i$, and r_i is

the return of asset i. The distribution of future asset returns is unknown, and the investor has to rely on the sample of past returns to form the portfolio. The closed-form solution for this optimization problem is available only under strict assumptions and specific probability distributions (Castro et al. 2022), and thus estimating the multivariate distribution of future returns provides a limited benefit. Given a sample[1] with T time periods, let $r_{t,i}$ be the return of asset i in period t. Let $P \in \mathbb{R}^T$ be a random sample of portfolio returns over T periods, i.e. for $t = 1, \ldots, T$ we have $P_t = \sum_{i=1}^{N} w_{t,i} r_{t,i}$, where $w_{t,i}$ is the weight of asset i at the beginning of the period t. The crucial difference between $P_{\mathbf{w}}$ and P is that $P_{\mathbf{w}}$ is a random variable representing population or future portfolio returns, while P is a sample of past T portfolio returns given by the weights $\mathbf{w}$ and sample asset returns. In the case of restricted short-selling, all weights need to be non-negative, resulting in the portfolio being a convex combination of returns. The sample-based optimization problem corresponding to the portfolio choice of an investor with τ-quantile preferences for a given $\tau \in (0, 1)$ is defined in Equation 3.5.

$$\begin{aligned} &\max_{w \in [0,1]^n} Q_\tau(P) \\ &\text{s.t.} \sum_{i=1}^{n} w_i = 1 \\ &p_t = \sum_{j=1}^{N} w_i r_{t,i} \quad \text{for } t = 1, \ldots, T \\ &x_i \geq 0 \quad \text{for} \quad i = 1, \ldots, N \end{aligned} \tag{3.5}$$

There are two main differences compared to the traditional mean-variance portfolio selection setting. There is no explicit risk constraint as we are used to from the portfolio selection literature[2] and we are directly maximizing a τ-quantile of portfolio returns instead of some utility from it. Both of these differences stem from the fact that risk aversion is fully captured by the parameter τ.

However, the appeal of this setting, i.e. not depending on assumptions about the shape of the utility function, does not come for free. Castro et al. (2022) use grid-search to approximate the optimal weights in this setting. For $\tau \in (0, 1)$

[1] Also referred to as estimation window throughout the text.

[2] Most of the literature is actually concerned with the risk-minimization instead of the mean-variance approach, given the amount of noise in forecasting expected returns (Clarke et al. 2006; DeMiguel et al. 2009)

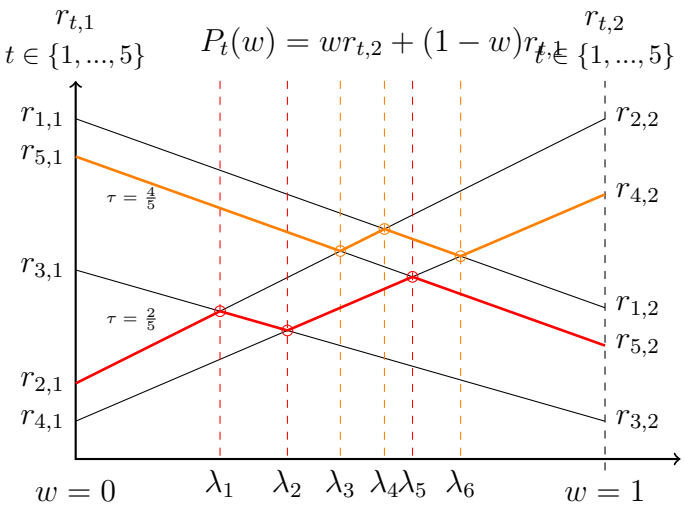


Figure 3.1. τ-quantile maximization illustration
Picture shows all possible sample portfolio returns obtained by weighting returns of asset 1 (r_1) and asset 2 (r_2) using a weight w.

and grid of values representing weights, i.e. $\Lambda_l = \{0 \leq w_1 < \ldots < w_l \leq 1\}$ they solve Equation 3.5. In case of 2 assets, it is simply $P_t = w_t r_{t,1} + (1 - w_t) r_{t,2}$. However, this is feasible only for very small N since the number of weight configurations needed to be searched grows with the power of N.

Even though closed-form solutions are available only in specific cases, it is illustrative to examine one such case as a theoretical connection between τ-quantile maximization and mean-variance optimization. Simple parametric examples below illustrate how τ-quantile maximization connects to both the global minimum variance (GMV) portfolio and mean-variance optimization under different assumptions about expected returns and risk aversion.

Consider a portfolio allocation problem with two risky assets whose returns follow a bivariate normal distribution.

$$\mathbf{r} = \begin{bmatrix} r_1 \\ r_2 \end{bmatrix} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma),$$

where the mean vector and covariance matrix are given by:

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}.$$

Let the portfolio weight vector be $\mathbf{w} = [w_1, w_2]^\top$ with the full investment constraint $w_1 + w_2 = 1$. The portfolio return $P = w_1 r_1 + w_2 r_2$ has mean

$\mu_P = w_1\mu_1 + w_2\mu_2$ and variance $\sigma_P^2 = w_1^2\sigma_1^2 + w_2^2\sigma_2^2 + 2w_1w_2\rho\sigma_1\sigma_2$.

Under normality, the τ-quantile of the portfolio return is:

$$Q_P(\tau) = \mu_P + \sigma_P\Phi^{-1}(\tau) \tag{3.6}$$

where $\Phi^{-1}(\tau)$ is the inverse of the standard normal CDF. The τ-quantile maximization problem becomes:

$$\max_{\mathbf{w}} \quad \mu_P + \sigma_P\Phi^{-1}(\tau) \quad \text{subject to} \quad w_1 + w_2 = 1 \tag{3.7}$$

When both assets have equal expected returns ($\mu_1 = \mu_2 = \mu$), the portfolio expected return $\mu_P = \mu$ becomes constant. For $\tau < 0.5$, we have $\Phi^{-1}(\tau) < 0$, making the objective function equivalent to minimizing σ_P. This leads to the global minimum variance (GMV) portfolio:

$$\mathbf{w}^*_{GMV} = \arg\min_{\mathbf{w}} \quad \mathbf{w}^\top\Sigma\mathbf{w} \quad \text{subject to} \quad w_1 + w_2 = 1 \tag{3.8}$$

with the explicit solution $w_1^* = \frac{\sigma_2^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}$ and $w_2^* = 1 - w_1^*$.

This establishes a formal equivalence between quantile maximization for risk-averse investors ($\tau < 0.5$) and the global minimum variance portfolio when expected returns are equal, explaining why we might observe similar empirical performance between these approaches.

For $\tau = 0.5$, we have $\Phi^{-1}(0.5) = 0$, which simplifies the problem to maximizing μ_P. The solution is to allocate all wealth to the asset with the highest expected return. When returns are equal, any feasible portfolio is optimal under this criterion. For the median investor, quantile maximization is, therefore, equivalent to expected return maximization without risk constraints.

For $\tau > 0.5$, we have $\Phi^{-1}(\tau) > 0$, making the objective function (with equal means) equivalent to maximizing σ_P. The solution is to allocate all wealth to the asset with a higher variance. This demonstrates that risk-seeking investors with quantile preferences will hold extremely concentrated portfolios when expected returns are equal.

When expected returns differ ($\mu_1 \neq \mu_2$), the optimization problem involves both the mean and variance components. Comparing with mean-variance optimization:

$$\max_{\mathbf{w}} \quad \mu_P - \lambda\sigma_P^2 = \mathbf{w}^\top\boldsymbol{\mu} - \lambda\mathbf{w}^\top\Sigma\mathbf{w} \tag{3.9}$$

we can establish a relationship between τ and the risk aversion parameter λ:

$$\lambda(\tau) = \frac{|\Phi^{-1}(\tau)|}{2\sigma_P} \tag{3.10}$$

This mapping has several implications. The effective risk aversion in quantile preferences varies nonlinearly with τ through $\Phi^{-1}(\tau)$, it depends on the portfolio's standard deviation, making risk aversion endogenous to portfolio choice and as τ approaches zero, the effective risk aversion approaches infinity, focusing exclusively on minimizing variance.

These results do not hold only in case of 2 assets, but extend naturally to $N > 2$. Let $\mathbf{r} = [r_1, \ldots, r_N]^\top$ be a vector of N normally distributed random variables representing asset returns, with mean vector $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_N]^\top$ and covariance matrix Σ. Let $\mathbf{w} = [w_1, \ldots, w_N]^\top$ be a vector of portfolio weights. The portfolio return is given by $P = \mathbf{w}^\top \mathbf{r}$. The portfolio return P is normally distributed with mean $\mu_P = \mathbf{w}^\top \boldsymbol{\mu}$ and variance $\sigma_P^2 = \mathbf{w}^\top \Sigma \mathbf{w}$. Thus, the portfolio return P is distributed as $P \sim \mathcal{N}(\mathbf{w}^\top \boldsymbol{\mu}, \mathbf{w}^\top \Sigma \mathbf{w})$, where $\mathbf{w}^\top \boldsymbol{\mu}$ represents the expected return of the portfolio, and $\mathbf{w}^\top \Sigma \mathbf{w}$ captures the portfolio variance, incorporating both the individual asset variances and the covariances between assets. Since the entire distribution is fully captured by its mean and variance, so is its quantile function $Q_P(\tau) = \mu_P + \sigma_P \Phi^{-1}(\tau)$, where $\mu_P = \mathbf{w}^\top \boldsymbol{\mu}$ is the portfolio mean, $\sigma_P = \sqrt{\mathbf{w}^\top \Sigma \mathbf{w}}$ is the portfolio standard deviation, and $\Phi^{-1}(\tau)$ is the quantile function of the standard normal distribution. In other words, in the case of normally distributed assets, a quantile of portfolio returns is its rescaled variance plus its mean.

Therefore, when assuming normally distributed returns with equal means[3] maximizing τ-quantile of portfolio returns for τs smaller than 0.5 results in the same portfolio weights as selected when minimizing portfolio variance. On the other side of the distribution, for τs greater than 0.5, maximizing τ-quantile corresponds to maximizing portfolio variance. Since diversification has a negative effect on the portfolio variance, maximizing variance implies full concentration in the most volatile asset.

These analytical results establish theoretical connections between quantile maximization and classical portfolio theory, showing that τ-quantile maximization encompasses various traditional strategies as special cases, depending on the value of τ and the distributional properties of asset returns. This theoretical

[3]This is common in the risk-based portfolio selection literature given the large forecasting error in the mean, (Jorion 1986; Michaud 1989).

foundation helps explain why quantile-based portfolios may exhibit different characteristics across the spectrum of τ values and why they might perform differently from traditional benchmarks in out-of-sample tests.

3.3.1 Quantile maximization via mixed integer linear programming

Computational difficulties associated with quantiles are well known in the Value at Risk (VaR) literature.[4] Benati and Rizzi (2007) propose a mean-VaR model and show it belongs to the class of NP-hard problems precisely because of the mathematical properties of VaR, i.e. non-linearity, non-convexity and presence of many local minima and maxima. Crucially, they also show that the mean-VaR optimization problem can be reformulated as a Mixed Integer Linear Programming problem which is computationally feasible for medium N and τ. [5] The mean-VaR optimization problem is Approximable-hard, and consequently problem in Equation 3.5 is also Approximable-hard, see Benati and Rizzi (2007). Following Benati and Rizzi (2007), we reformulate the problem from Equation 3.5 as a Mixed Integer Linear Programming problem shown in Equation 3.11.

$$\begin{aligned}
&\max_{w,y,Q_\tau} Q_\tau \\
&\text{s.t. } Q_\tau \leq \sum_{i=1}^{N} w_i r_{t,i} + L \cdot y_t \quad \forall t \in \{1,\dots,T\} \\
&\sum_{t=1}^{T} y_t \leq \tau T \\
&y_t \in \{0,1\} \quad \forall t \in \{1,\dots,T\} \\
&w_i \geq 0 \quad \forall i \in \{1,\dots,N\} \\
&\sum_{i=1}^{N} w_i = 1
\end{aligned} \tag{3.11}$$

where L is a sufficiently large constant. Equation 3.11 is a Mixed Integer Linear Programming problem as variable y is the (only) binary (integer) variable. This optimization problem is equivalent to Equation 3.5. It contains $2*T+N+2$ optimization constraints.

[4] VaR is just a τ-quantile of the return distribution, where τ is typically chosen as 0.05 or 0.01.

[5] It is also feasible in the case of very small T, which is, however, not of interest in most of the financial applications.

To understand the optimization problem, notice that even though Q_τ is a variable in the optimization problem when the optimum is reached, it also corresponds to the sample τ-quantile of portfolio returns P. For Q_τ to be a sample τ-quantile, we expect $\tau * 100\%$ of sample observations, or portfolio returns, to be smaller or equal to it and $(\tau - 1) * 100\%$ of sample observations to be larger than it. The first constraint includes large enough constant L, which is actually added only when $y_t = 1$, i.e. periods when the final portfolio return will be smaller than the sample τ-quantile. It is achieved by capping the number of periods $y_t = 1$ holds to the fraction of portfolio returns we expect to be smaller than Q_τ. Binary variable y_t then characterizes sample time periods where portfolio return is smaller than given τ-quantile. The last two constraints are no short-selling and full investment conditions. To solve Equation 3.11, a branch & bound algorithm implemented in multiple optimization softwares[6] can be employed.

3.4 Simulation

We use simulated data to study the composition and out-of-sample performance of portfolios formed by investors with quantile preferences. Specifically, we simulate data from a single-factor model and mimic the rolling estimation window approach from the empirical results to repeatedly form τ-quantile maximizing portfolios for different values of τ and compare them to benchmark portfolios. We focus on independent and normally distributed returns mainly since it is one of the rare cases where the τ-quantile maximization problem can be solved analytically and thus we can study the size of errors in the selected portfolio weights using a finite sample size.

We simulate a time-series of $T = 300$ monthly returns for $N = 5$ assets using a single-factor model $r_t = \beta f_t + \epsilon_t, t = 1, \dots T$, where r_t is the vector of asset returns, f_t is the vector of market factor returns, β is the vector of the factor loading, and ϵ_t is the vector of residuals. We follow Pástor (2000) and DeMiguel et al. (2009) in the simulation design. Specifically, for the market factor returns, we assume mean monthly returns of $\frac{8}{12}\%$ and monthly standard deviation of $\frac{16}{\sqrt{12}}\%$. We assume the factor loadings are uniformly distributed $f_t \sim U(0.5, 1.5)$, and residuals are independent and normally distributed with variances uniformly distributed $\sigma_n \sim U(0.1, 0.3)$. We also examine the effect

[6]We use IBM ILOG CPLEX. Other options include Gurobi, MOSEK, Matplotlib or GLPK.

of non-normality by considering residuals to be t-distributed with 3 degrees of freedom and present these results in Section 3.C. Using the simulated returns, we proceed in a standard "rolling" estimation window approach. We use 60-month long estimation windows to find the optimal portfolio weights and rebalance each month. Table 3.1 shows the normalized Herfindahl-Hirschman index and turnover of the portfolio weights for different τ levels, as well as benchmark global minimum variance and equal-weighted portfolios. The normalized Herfindahl-Hirschman index and portfolio weights turnover are defined in Section 3.A. Table 3.2 shows the standard deviation and τ-quantiles of out-of-sample portfolio returns generated by the τ-quantile maximizing portfolios and benchmark portfolios.

Table 3.1. Herfindahl-Hirschman index and turnover
This table shows the average normalized Herfindahl-Hirschman indices (HH) and average as well as median portfolio weights turnover (Δ) for the τ-quantile maximizing (qm), Equal Weighted (ew), Global Minimum Variance (gmv) and Global Minimum Variance using Ledoit-Wolf covariance (gmv_lw) portfolios. Returns data are simulated using a single-factor model with $N = 5$ assets and $T = 300$ periods. Portfolio weights are selected using 60 observations long rolling estimation window. Simulation is repeated 1000 times, and results show the average values across individual simulation runs.

strategy	ew	gmv_lw	gmv	qm0.05	qm0.1	qm0.2	qm0.3	qm0.4	qm0.5	qm0.6	qm0.7	qm0.8	qm0.9	qm0.95
HH	0.0	0.00	0.05	0.12	0.15	0.22	0.31	0.44	0.54	0.65	0.83	0.93	0.99	1.0
Δ	0.0	0.02	0.04	0.06	0.10	0.14	0.22	0.30	0.40	0.34	0.18	0.17	0.08	0.1
Δ (median)	0.0	0.01	0.03	0.00	0.00	0.00	0.07	0.05	0.11	0.03	0.00	0.00	0.00	0.0

Focusing on weight concentration measured with normalized HH index, Table 3.1 shows increasing portfolio weights concentration for τ-quantile maximizing portfolios as τ increases. In other words, given the estimation sample used to find optimal weights, the more risk-loving investor, the more concentrated portfolio, and vice versa. Even though low τ-quantile maximizers have the most diversified portfolio among all other τ-quantile maximizers, it is surprising to observe that they are still more than 2 times more concentrated than the global minimum variance portfolio, regardless of which covariance matrix estimate is used.

Further, to capture the degree to which different τs lead to weight changes from period to period, we can look at average and median portfolio weights turnover, Δ and Δ(median). We focus on the turnover based on the changes between selected portfolio weights from one period to another, ignoring the inter-period evolution of the assets themselves, to isolate the effect of the estimation sample changes on the selected weights. The highest turnover is associated

with τ-quantile maximizer for τs close to 0.5 and decreases with moving τ to either tail. In comparison with the global minimum variance portfolio, average turnover tends to be higher in the case of τ-quantile maximizers. However, average turnover does not fully capture the underlying dynamics of weight changes. Median portfolio weights turnover is significantly lower than the average one for all τ-quantile maximizing portfolios. For low (or high) enough τs, the median turnover is even lower than the median turnover for a naive global minimum variance portfolio. This indicates that changes in portfolio weights for low or high enough τ-quantile maximizers are less frequent, but when they occur, they tend to be larger. We illustrate this phenomenon in Figure 3.4.

The ultimate goal of an investor is to achieve a desired out-of-sample performance based on their preferences, even though they must rely only on in-sample data to form the portfolio. It is not clear ex-ante whether sample-based τ-quantile maximization strategies will also achieve high out-of-sample τ-quantiles. Table 3.2 shows τ-quantiles and volatility of out-of-sample portfolio returns for global minimum variance portfolio and τ-quantile maximizing portfolios.

Table 3.2. Out-of-sample performance of the simulated portfolios This table shows the standard deviation and τ-quantiles of out-of-sample portfolio returns for the τ-quantile maximizing (qm), Equal Weighted (ew), Global Minimum Variance (gmv) and Global Minimum Variance using Ledoit-Wolf covariance (gmv_lw) portfolios. Returns data are simulated using a single-factor model with $N = 5$ assets and $T = 300$ periods. Portfolio weights are selected using 60 observations long rolling estimation window. Simulation is repeated 1000 times and results show the average values across individual simulation runs. Results are shown in percentage points.

	ew	gmv_lw	gmv	qm0.05	qm0.1	qm0.2	qm0.3	qm0.4	qm0.5	qm0.6	qm0.7	qm0.8	qm0.9	qm0.95
σ	13.49	13.57	13.97	15.32	15.45	15.34	16.39	19.23	21.61	22.62	24.94	24.79	26.09	24.52
0.05	-21.38	-21.36	-22.40	-24.78	-23.65	-24.42	-25.24	-32.62	-33.28	-37.07	-38.62	-38.82	-42.13	-39.69
0.1	-17.15	-17.15	-18.31	-19.03	-19.38	-18.54	-21.91	-23.28	-25.30	-28.86	-29.74	-28.19	-32.06	-27.61
0.2	-11.37	-11.38	-10.75	-13.39	-12.12	-12.33	-14.58	-13.95	-14.97	-18.24	-21.28	-17.42	-19.33	-17.33
0.3	-7.36	-7.27	-7.79	-8.84	-8.67	-7.11	-7.54	-8.54	-10.00	-10.82	-13.56	-13.13	-12.00	-11.19
0.4	-3.44	-3.35	-3.75	-4.61	-4.11	-3.90	-4.05	-4.00	-5.62	-6.34	-6.93	-5.66	-4.27	-3.13
0.5	0.67	0.61	0.85	0.26	-0.21	0.70	1.15	0.63	-2.36	-1.89	0.29	0.70	1.40	1.92
0.6	3.70	3.53	4.22	3.52	3.79	4.78	4.04	3.91	2.86	3.76	4.92	5.94	5.82	6.16
0.7	6.99	6.75	7.83	7.69	9.41	8.70	8.84	8.39	9.39	10.35	11.11	13.56	12.66	13.29
0.8	11.45	10.75	11.77	12.74	13.63	13.07	12.86	16.17	17.44	16.87	23.03	23.15	23.15	22.35
0.9	17.93	17.68	18.34	18.54	18.97	19.56	20.09	22.82	25.47	28.36	32.10	32.21	33.06	32.66
0.95	22.30	23.52	22.70	24.46	22.57	23.89	23.89	27.36	39.36	38.25	42.01	43.48	44.39	42.80

In the ideal case of no estimation error, we would expect the τ-quantile maximizing portfolio to achieve the highest τ-quantile of out-of-sample portfolio returns. In other words, for example, we would like to see that the 0.3-quantile maximizer (column qm0.3) achieves the highest 30%-quantile (row 0.3) of out-of-sample portfolio returns among all possible strategies. Table 3.2 shows that this

is not always the case. There are multiple cases when the highest out-of-sample τ-quantile is achieved by another strategy than the corresponding τ-quantile maximizer, for example, the highest out-of-sample 10%-quantile is achieved by the qm0.2 strategy among all quantile maximizers. This points to the estimation error stemming from the in-sample optimization process even though often when the highest τ-quantile is not achieved by the very same τ-quantile maximizer, it is usually achieved by a neighbouring τ-quantile maximizer. However, crucially, the global minimum variance portfolio is the best-performing strategy in terms of out-of-sample τ-quantiles for τs smaller than 0.5. On the one hand, this is not surprising as the underlying data-generating process for the simulation leads to normally distributed returns, hence the global minimum variance portfolio is expected to perform well. On the other hand, it is perhaps surprising to see the strong out-of-sample underperformance of the τ-quantile maximizer for τs smaller than 0.5. To rule out the effect of normally distributed returns favouring the global minimum variance portfolio extensively, we also consider the case of a single-factor model with t-distributed residuals leading to non-normal returns in Section 3.C and document the same phenomenon there.

As is typical in the portfolio selection literature, the main culprit for the out-of-sample underperformance is the estimation error. It is well known that portfolios constructed using in-sample estimates are strongly affected by it, see DeMiguel et al. (2009). Maximizing quantiles via in-sample optimization is no exception. We, therefore, compare analytically obtained optimal weights with weights found by sample-based quantile maximization introduced in Section 3.3.1. In practice, the underlying data-generating process is unknown, and even if the corresponding distribution were known, the closed-form solution might not exist. Therefore, portfolio weights need to be selected based on sample returns. We leverage the fact that the τ-quantile maximization problem can be solved analytically in the case of normally distributed returns to study the estimation error in the sample-based τ-quantile maximizing portfolios. In the case of τ-quantile maximization, portfolio weights are selected during the optimization algorithm, which takes directly underlying sample returns and does not require the estimation of the covariance matrix. This contrasts with a two-step process of global minimum variance portfolio where a covariance matrix is estimated first and only then are the optimal weights calculated.

The effect of forming a portfolio via sample τ-quantile maximization instead of using analytically obtained portfolio weights can be seen in Figure 3.2. Specifically, we examine the case with only 2 assets and a sample of 1000

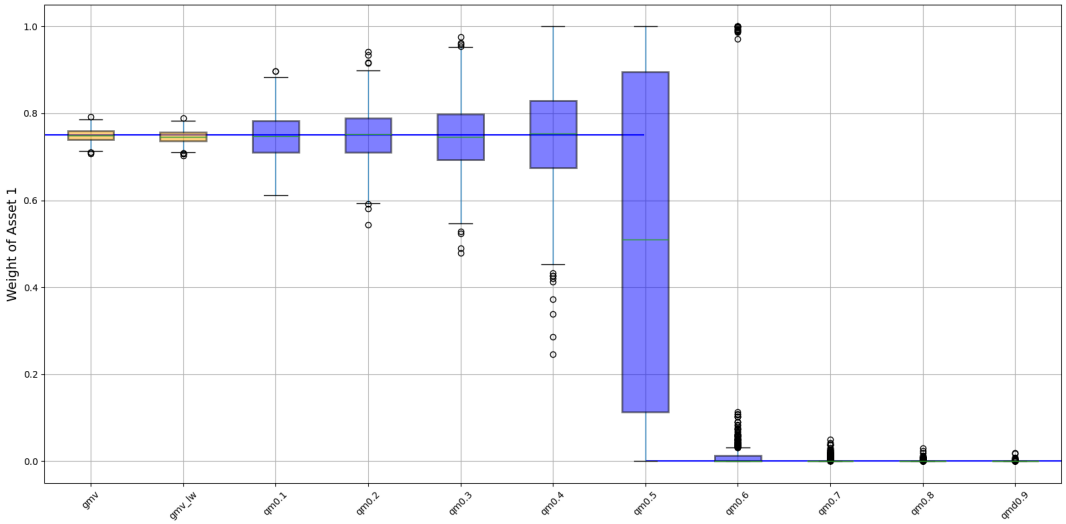


Figure 3.2. Selected portfolio weights across τs
This figure shows the distribution of weight of the first asset selected using global minimum variance portfolio with and without Ledoit-Wolf covariance matrix estimate (gmv, gmv_lw) and τ-quantile maximizing portfolios. Returns of two independent and normally distributed assets with mean 0 and variances of 0.01 and 0.03, respectively, are generated 1000 times and each sample consists of 1000 observations. The horizontal blue line represents the analytical optimal weight of the first asset given the data-generating process.

observations, where the assets are independent and normally distributed with means of 0 and standard deviations of 0.01 and 0.03, respectively. We focus on the case with only 2 assets, as finding optimal weights for τ-quantile maximizing portfolios as T grows is feasible only for small N. The optimal weight of the first asset is 0.75 and the optimal weight of the second asset is 0.25 when the optimization objective is to minimize the portfolio variance or maximize the portfolio τ-quantile for $\tau < 0.5$. For $\tau > 0.5$, the optimal weight of the more volatile asset is 1 and the optimal weight of the less volatile asset is 0, given the full investment constraint. For $\tau = 0.5$ any combination of the two assets is optimal as both assets have the same mean returns. This setup is generous in terms of the number of observations, as a sample of 1000 observations is unrealistic in most practical cases, where T is usually much smaller, around 60 or 120 monthly observations. However, even in this case, the errors in the sample-based optimal weights of the first asset with respect to the true optimal weight obtained analytically are significantly greater than corresponding errors in the case of the global minimum variance portfolio. To examine the effect of sample size on the estimation error, Figure 3.3 shows the distribution of the sample-based optimal weight of the first asset and the in-sample achieved objective value

(τ-quantile) for global minimum variance and τ-quantile maximizing portfolios in the case of different sample sizes.

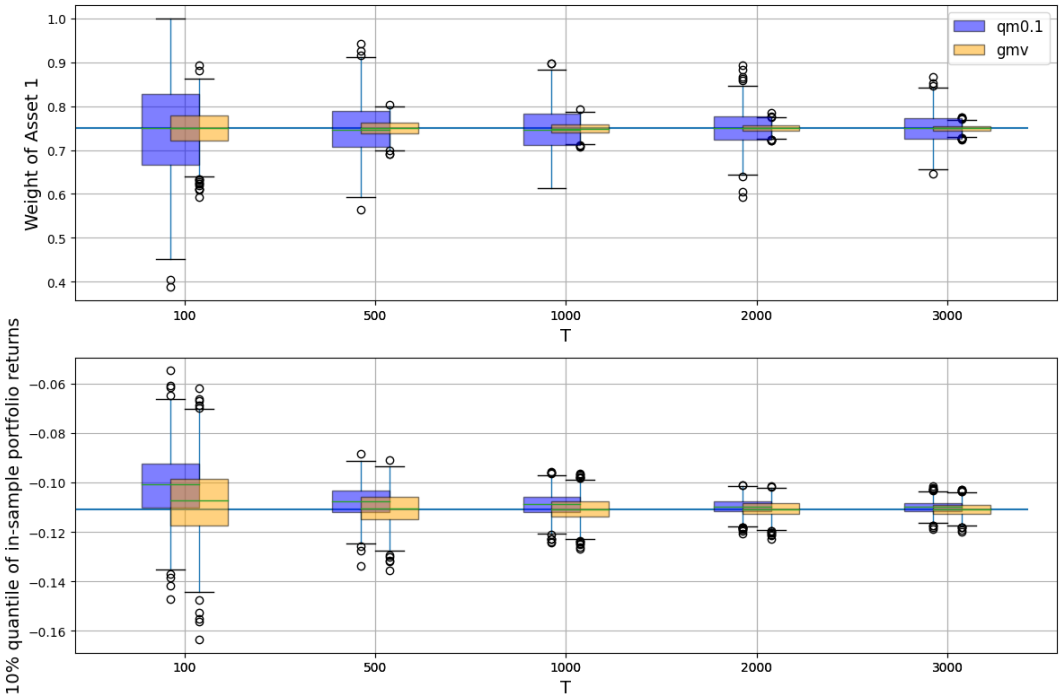


Figure 3.3. Portfolio weights & in-sample objective values
This figure shows the distribution of the sample-based optimal weight of the first asset and the in-sample achieved objective value (τ-quantile) for 1000 simulations. Returns of two independent and normally distributed assets with mean 0 and variances of 0.01 and 0.03, respectively. Portfolio weights and objective values are shown for the global minimum variance portfolio and 10%-quantile maximizing portfolio using samples with sizes of 100, 500, 1000, 2000 and 3000 observations. The horizontal blue line represents the true optimal weight of the first asset given the data-generating process.

Both the global minimum variance and τ-quantile maximizing portfolios converge to the true optimal weights and in-sample objective values as the sample size increases. However, the variation around the optimal weights is significantly greater in the case of τ-quantile maximizing portfolio than in the case of a global minimum variance portfolio for all considered sample sizes. On top of that, in-sample objective values achieved by τ-quantile maximizing portfolios are biased upwards for small sample sizes. This overestimation of in-sample portfolio returns quantiles slowly decreases as the sample size grows. Global minimum variance portfolios are less affected by the sample size, and the corresponding in-sample objective values are closer to the true optimal objective values. These results suggest that the global minimum variance approach might be more reliable for small sample sizes, whereas the quantile-based approach

may require larger sample sizes to achieve robust and accurate results. Overall, these results show that using sample-based quantile maximizing portfolios to achieve the desired out-of-sample or true τ-quantile of portfolio returns might be challenging due to estimation error and for τs smaller than 0.5, the global minimum variance portfolio might be a better practical choice.

3.5 Empirical results

In this section, we examine the performance of τ-quantile maximizing portfolios on empirical data. Section 3.5.1 describes the datasets used in the empirical evaluation, Section 3.5.2 examines the concentration and turnover of the portfolio weights chosen in comparison to portfolio selection strategies, and Section 3.5.3 presents the out-of-sample performance of the portfolios formed by the τ-quantile maximizing strategies.

3.5.1 Data

We use 10 empirical datasets with monthly and daily returns. Table 3.3 shows the datasets, the corresponding number of assets, and time periods used in the empirical evaluation. Portfolios from the French data library start in January 1996 and end in May 2024. We also form portfolios on individual stocks from the Center for Research in Security Prices (CRSP) database, ranging from January 1, 1996, to December 31, 2018. Cryptocurrency returns are from the Binance data feed and cover the period from August 1, 2017, to January 31, 2021.

Table 3.3. Datasets used in the empirical results
N stands for the number of constituents in the investment universe. D is daily frequency, and M is monthly frequency, where (D) signals that this frequency was not used for all number of constituents N. FF datasets are from the French data library and returns are value-weighted, CRSP stocks are individual stocks from the Center for Research in Security Prices database from the Wharton Research Data Service (WRDS). Crypto are cryptocurrencies from the Binance data feed.

Dataset	N	Frequency	Time Period
FF Industry Portfolios	5, 17, 30	D	Jan 1996 - May 2024
FF Portfolios - B/M x Size	6, 25	(D),M	Jan 1996 - May 2024
FF Portfolios - INV x Size	6	D,M	Jan 1996 - May 2024
FF Portfolios - Prior 12-2 x Size	6, 25	D,M	Jan 1996 - May 2024
FF Portfolios - Prior 1-0 x Size	6, 25	(D),M	Jan 1996 - May 2024
FF Europe - B/M x Size	6	M	Jan 1996 - May 2024
FF Japan - B/M x Size	6	M	Jan 1996 - May 2024
CRSP stocks	10, 20, 30	D,M	Jan 1996 - Dec 2018
Crypto	13	D	Aug 2017 - Jan 2021

In the case of monthly frequencies, portfolios are formed each period (monthly), and we consider rolling estimation windows of 36 and 60 months, corresponding to 3 and 5 years. In the case of daily frequencies, we use an estimation window of 63 days and portfolios are also formed monthly, with 21 trading days being considered a month. At any portfolio formation date d, optimal portfolio weights are obtained as a solution to the optimization problems using only the historical observations contained within the estimation window. Results shown in the main text are monthly datasets and portfolio selection is done using 60-month-long estimation windows. The remaining results, i.e. daily frequency with 63-days-long estimation windows and monthly frequency with 36-month-long estimation windows, are shown in Section 3.D and serve as robustness support for the results in the main text. In the case of individual stocks from CRSP, we use randomly selected stocks from the universe of stocks that have available returns over the whole sample period and have market capitalization greater than the cross-sectional 10% market capitalization quantile in each period. For each stock dataset with a given number of assets and frequency, we select N stocks for the universe and keep the universe the same during the whole period. Cryptocurrencies are included for the purpose of adding extremely heavy-tailed returns to the empirical evaluation and consist of the following 13 cryptocurrencies - bitcoin, ethereum, dash, eos, iota, lisk, litecoin, monero, nem, neo, stellar, xrp, zcash.

3.5.2 Weights

We start the empirical evaluation by looking at the characteristics of portfolio weights selected in different strategies. We compare the τ-quantile maximizing portfolios across various τs ($qm\tau$); and with the benchmark equal weighting, global minimum variance (gmv) and global minimum variance using Ledoit-Wolf covariance (gmv_lw) portfolios. Table 3.4 shows the normalized Herfindahl-Hirschman index and turnovers of portfolio weights for all examined strategies and datasets. We document similar dynamics in weight diversification for τs $>$ 0.5 as in the simulation results in Table 3.1. There is a decreasing diversification as we move toward the right tail. On the other side of the distribution, for τs < 0.5, the relationship between τ and diversification is less monotonous as in the simulation results. There is less diversification for τ-quantile maximizing portfolios for τs in the very left tail (0.05 and 0.1) than for τs more in the centre of the distribution (0.3 and 0.5). All datasets showcase a common weight concentration "smirk" pattern for τ-quantile maximizing portfolios, where the weights are more concentrated in the tails of the distribution, but asymmetrically more in the right tail. In terms of comparison with the global minimum variance portfolio, weight diversification of τ-quantile maximizing portfolios for 0.05 and 0.1 τs is similar to the diversification of the global minimum variance portfolio when the sample covariance matrix is used. The global minimum variance portfolio with Ledoit-Wolf covariance matrix leads to higher weight diversification than any τ-quantile maximizing strategy or the global minimum variance portfolio with sample covariance matrix as expected. With regards to the average portfolio weights turnover, the τ-quantile maximizing portfolios exhibit higher turnover than the global minimum variance portfolios. The highest turnover is observed for the τ-quantile maximizing portfolios with central τs (0.3, 0.5 and 0.7). Weight turnover decreases as we move towards the tails of the distribution. When comparing strategies used by the most risk-averse investors, the global minimum variance portfolios have lower average turnover than τ-quantile maximizing portfolios, with the global minimum variance portfolio using the Ledoit-Wolf covariance matrix having the lowest turnover of all strategies. Overall, empirical results from the weight turnover perspective are consistent with the simulation results, including different patterns for average and median turnovers.

3.5.3 Out-of-sample evaluation

After observing that the τ-quantile maximizing portfolios yield heterogeneous weights across different τ values, as well as in comparison to the global minimum variance portfolios, we proceed to evaluate the out-of-sample performance of these portfolios. Our primary focus is on assessing the τ-quantiles of the out-of-sample portfolio returns, considering τ values of $\{0.05, 0.1, 0.3, 0.5, 0.7, 0.9, 0.95\}$, corresponding to those used in τ-quantile maximization. Additionally, we compare the out-of-sample performance of these portfolios in terms of standard deviation, Sharpe ratio, and maximum drawdown. The τ-quantile maximizing portfolios are compared against benchmark portfolios, including equal weighting, global minimum variance, and global minimum variance with Ledoit-Wolf covariance matrix portfolios. Table 3.5 shows the 10%, 50% and 90% quantiles of out-of-sample portfolio returns for all examined strategies and monthly datasets using 60 months-long estimation windows.

Looking at the 10% quantile, we observe that the strategy maximizing in-sample 10%-quantile does not necessarily lead to the highest out-of-sample 10%-quantile. For the majority of datasets, the global minimum variance portfolios, either with or without the Ledoit-Wolf covariance matrix, achieve the highest out-of-sample 10% quantile of portfolio returns. This phenomenon is robust to the choice of the length of the rolling in-sample period as well as the frequency, see Section 3.D. It is also consistent with simulations results in Section 3.4, where we have shown that especially for smaller sample sizes, the global minimum variance portfolios outperform the τ-quantile maximizing portfolios for $\tau < 0.5$ in terms of the corresponding out-of-sample portfolio returns quantile. To further highlight the effect of sample noise on our results, we also observe that the 10%-quantile maximizer does not necessarily lead to the highest out-of-sample 10%-quantile even among all the other τ-quantile maximization strategies. Notably, 5%-quantile maximization strategy leads to higher out-of-sample 10%-quantile than 10%-quantile maximization strategy for a number of datasets and there is even one dataset (17 Industry Portfolios) where 10%-quantile maximization strategy leads to lower out-of-sample 10%-quantile than 30%-quantile maximization strategy. Further, most of the differences, especially between neighbouring τ-quantile maximization strategies or between low τ-quantile maximization strategies and the global minimum variance portfolios, are not statistically significant. Nevertheless, comparing the out-of-sample 10%-quantile of portfolio returns only among τ-quantile

maximizers, we observe generally decreasing out-of-sample 10%-quantile as we move further away with the τ from the 10% we are evaluating. In other words, more risk-loving investors do not achieve the same out-of-sample 10%-quantile as the more risk-averse investors. Furthermore, there is a clear benefit for an investor with τ-quantile preferences to not choose equal weights, but depending on their τ to either choose the global minimum variance portfolio or τ-quantile maximizing portfolio. Moving to the 50% quantile, there is no specific reason to believe that the global minimum variance portfolios would outperform the τ-quantile maximizing portfolios. There is no clear pattern in the sense of one strategy leading to the highest out-of-sample 50% quantile of portfolio returns, and the great majority of the differences between the strategies are not statistically significant. Looking at the 90% quantile, we observe expected results, where τ-quantile maximizing portfolios with τs in the right tail of the distribution also lead to the highest out-of-sample 90% quantile of portfolio returns. This is expected since all the benchmark portfolios are designed to minimize risk in the form of volatility or aim at perfect weight diversification. Both of these objectives go against the objective of maximizing the quantile in the far right tail of the distribution. Besides quantiles of out-of-sample portfolio returns, Table 3.8 shows the standard deviation, Sharpe ratio, and Maximum Drawdown of portfolio returns for all examined strategies and datasets. Contrary to the inferior out-of-sample performance of the τ-quantile maximizing portfolios with τ from the left tail, the global minimum variance portfolios lead to the lowest standard deviation of out-of-sample portfolio returns for all datasets and τ-quantile maximizing portfolios.

Table 3.4. Herfindahl-Hirschman index and turnover

This table shows the normalized Herfindahl-Hirschman index, mean turnover, and median turnover (Δ (median)) of portfolio weights for the τ-quantile maximizing (qm), Equal Weighted (ew), Global Minimum Variance (gmv), and Global Minimum Variance using Ledoit-Wolf covariance matrix (gmv_lw) portfolios. A rolling window sample with a length of 60 months is used. Results are in percentage points for datasets in Section 3.5.1.

	ew	gmv_lw	gmv	qm0.05	qm0.1	qm0.3	qm0.5	qm0.7	qm0.9	qm0.95
Herfindahl-Hirschman Index										
5IP	0.0	0.28	0.37	0.51	0.43	0.29	0.47	0.64	0.81	0.90
6P 2x3	0.0	0.51	0.65	0.62	0.64	0.45	0.42	0.64	0.88	0.78
6P ME INV	0.0	0.46	0.66	0.53	0.52	0.44	0.47	0.53	0.85	0.83
6P ME P10	0.0	0.54	0.73	0.69	0.62	0.61	0.41	0.62	0.77	0.91
6P ME P122	0.0	0.43	0.51	0.48	0.58	0.44	0.42	0.64	0.84	0.87
EUR 6P 2x3	0.0	0.62	0.79	0.55	0.57	0.53	0.54	0.55	0.78	0.82
JPN 6P 2x3	0.0	0.32	0.53	0.56	0.40	0.52	0.48	0.51	0.76	0.93
17IP	0.0	0.20	0.30	0.32	0.30	0.25	0.54	0.63	0.83	0.89
25P 5x5	0.0	0.34	0.53	0.47	0.47	0.33	0.31	0.42	0.76	0.81
25 ME P10	0.0	0.32	0.51	0.47	0.39	0.35	0.34	0.43	0.68	0.79
25 ME P122	0.0	0.30	0.56	0.46	0.39	0.31	0.36	0.49	0.70	0.79
30IP	0.0	0.15	0.22	0.26	0.20	0.20	0.49	0.66	0.89	0.87
48IP	0.0	0.13	0.20	0.20	0.19	0.20	0.40	0.66	0.91	0.98
CRSP5	0.0	0.20	0.46	0.33	0.36	0.39	0.59	0.72	0.92	0.97
CRSP10	0.0	0.13	0.20	0.25	0.22	0.22	0.53	0.68	0.88	0.96
CRSP20	0.0	0.08	0.14	0.20	0.18	0.21	0.43	0.65	0.79	0.89
Portfolio Turnover										
5IP	0.0	0.06	0.07	0.12	0.18	0.32	0.32	0.32	0.18	0.10
6P 2x3	0.0	0.05	0.07	0.12	0.14	0.31	0.37	0.29	0.11	0.14
6P ME INV	0.0	0.06	0.07	0.11	0.20	0.26	0.43	0.29	0.13	0.06
6P ME P10	0.0	0.05	0.04	0.09	0.14	0.36	0.47	0.23	0.08	0.01
6P ME P122	0.0	0.05	0.06	0.13	0.13	0.39	0.49	0.27	0.10	0.06
EUR 6P 2x3	0.0	0.05	0.04	0.15	0.22	0.32	0.35	0.31	0.12	0.12
JPN 6P 2x3	0.0	0.10	0.12	0.15	0.26	0.38	0.48	0.32	0.13	0.03
17IP	0.0	0.09	0.11	0.18	0.26	0.46	0.44	0.30	0.17	0.08
25P 5x5	0.0	0.10	0.13	0.19	0.28	0.52	0.66	0.41	0.23	0.12
25 ME P10	0.0	0.10	0.12	0.18	0.28	0.59	0.68	0.51	0.18	0.17
25 ME P122	0.0	0.10	0.11	0.22	0.29	0.56	0.54	0.41	0.25	0.14
30IP	0.0	0.11	0.14	0.24	0.32	0.57	0.43	0.32	0.09	0.05
48IP	0.0	0.12	0.15	0.28	0.36	0.75	0.44	0.37	0.08	0.07
CRSP5	0.0	0.05	0.05	0.09	0.11	0.30	0.24	0.18	0.13	0.08
CRSP10	0.0	0.06	0.08	0.15	0.17	0.33	0.31	0.17	0.15	0.08
CRSP20	0.0	0.08	0.12	0.21	0.29	0.48	0.28	0.21	0.14	0.11
Portfolio Turnover (median)										
5IP	0.0	0.04	0.05	0.0	0.00	0.08	0.10	0.00	0.0	0.0
6P 2x3	0.0	0.03	0.02	0.0	0.00	0.07	0.22	0.00	0.0	0.0
6P ME INV	0.0	0.03	0.01	0.0	0.00	0.00	0.22	0.00	0.0	0.0
6P ME P10	0.0	0.02	0.00	0.0	0.00	0.00	0.37	0.00	0.0	0.0
6P ME P122	0.0	0.02	0.02	0.0	0.00	0.10	0.28	0.00	0.0	0.0
EUR 6P 2x3	0.0	0.03	0.01	0.0	0.00	0.02	0.16	0.00	0.0	0.0
JPN 6P 2x3	0.0	0.07	0.07	0.0	0.00	0.14	0.27	0.00	0.0	0.0
17IP	0.0	0.06	0.08	0.0	0.00	0.38	0.20	0.00	0.0	0.0
25P 5x5	0.0	0.07	0.09	0.0	0.00	0.35	0.56	0.00	0.0	0.0
25 ME P10	0.0	0.07	0.07	0.0	0.00	0.32	0.52	0.00	0.0	0.0
25 ME P122	0.0	0.07	0.06	0.0	0.00	0.33	0.36	0.00	0.0	0.0
30IP	0.0	0.09	0.10	0.0	0.00	0.41	0.27	0.00	0.0	0.0
48IP	0.0	0.10	0.11	0.0	0.00	0.77	0.24	0.03	0.0	0.0
CRSP5	0.0	0.03	0.03	0.0	0.00	0.10	0.06	0.00	0.0	0.0
CRSP10	0.0	0.04	0.06	0.0	0.00	0.19	0.12	0.00	0.0	0.0
CRSP20	0.0	0.06	0.09	0.0	0.03	0.27	0.14	0.00	0.0	0.0

Table 3.5. Out-of-sample 10%, 50% and 90% quantiles
This table shows the 10%, 50% and 90% quantile of portfolio returns for the τ-quantile maximizing (qm), Equal Weighted (ew), Global Minimum Variance (gmv) and Global Minimum Variance using Ledoit-Wolf covariance matrix (gmv_lw) portfolios. A rolling window sample with a length of 60 months is used. Results are shown in percentage points and for datasets described in Section 3.5.1. Bold values indicate that the out-of-sample quantiles of portfolio returns are significantly different from the quantile maximization strategy targeting the given τ at the 5% significance level. Statistical significance is determined using a percentile bootstrap method described in Section 3.A.

	ew	gmv_lw	gmv	qm0.05	qm0.1	qm0.3	qm0.5	qm0.7	qm0.9	qm0.95
10% quantile										
5IP	-5.26	-4.47	-4.42	-4.49	-4.2	-4.36	-5.36	**-6.56**	-6.14	**-6.46**
6P 2x3	-5.87	-4.78	-4.56	-5.08	-5.14	-5.37	-5.89	**-7.25**	**-7.91**	**-7.61**
6P ME INV	-5.95	-4.69	-4.66	-4.84	-4.65	-4.58	-6.35	**-7.26**	**-7.68**	**-7.26**
6P ME P10	-5.95	-4.56	-4.37	-4.93	-4.87	-4.92	-6.3	**-7.67**	**-7.48**	**-7.47**
6P ME P122	-5.7	-4.79	-4.63	-5.34	-5.11	-4.87	-5.82	-6.64	**-7.94**	-7.65
EUR 6P 2x3	-4.94	-5.3	-4.97	-5.06	-5.28	-5.86	-5.88	-6.05	-6.95	-7.22
JPN 6P 2x3	-5.16	-5.2	-5.22	-4.87	-4.72	-5.25	-6.07	-5.6	-6.17	**-6.86**
17IP	-4.92	-3.65	-3.55	-3.65	-4.26	-4.39	**-7.7**	**-8.43**	**-9.5**	**-9.39**
25P 5x5	-6.07	-4.33	-4.42	-5.01	-4.92	-5.77	-6.21	**-8.19**	**-9.39**	**-9.69**
25 ME P10	-6.03	-4.73	-4.88	-5.43	-5.23	-5.31	-7.01	**-7.81**	**-8.96**	**-8.92**
25 ME P122	-6.12	-4.77	-4.74	-5.77	-5.37	-5.48	-7.75	**-8.53**	**-9.98**	**-8.95**
30IP	-4.88	-3.84	-3.83	-4.05	-4.26	-4.92	**-7.47**	**-10.68**	**-14.59**	**-12.02**
48IP	-5.06	-3.53	-3.39	-3.95	-4.54	-5.34	**-8.34**	**-11.05**	**-11.47**	**-9.89**
CRSP5	-6.5	-5.02	-4.63	-4.41	-4.37	-4.61	**-7.81**	**-8.45**	**-10.29**	**-11.41**
CRSP10	-5.28	-3.72	-3.47	-4.34	-4.84	-4.97	**-7.79**	**-7.96**	**-8.99**	**-11.35**
CRSP20	-3.73	-3.2	-3.08	-3.76	-4.06	-4.16	**-7.51**	**-7.99**	**-9.27**	**-11.34**
50% quantile										
5IP	1.41	1.45	1.44	1.33	1.35	1.35	1.22	1.25	1.27	1.25
6P 2x3	1.38	1.43	1.55	1.39	1.49	1.36	1.47	1.3	1.39	1.08
6P ME INV	1.35	1.36	1.32	1.4	1.38	1.36	1.19	1.31	1.35	1.57
6P ME P10	1.23	1.26	1.24	1.2	1.42	1.32	1.14	1.25	1.18	1.16
6P ME P122	1.23	1.43	1.41	1.4	1.27	1.44	1.31	1.34	1.33	1.14
EUR 6P 2x3	1.08	1.42	1.51	1.46	1.32	1.45	1.08	1.1	0.92	0.89
JPN 6P 2x3	0.9	0.72	0.7	0.87	0.84	0.83	0.86	1.11	0.91	0.48
17IP	1.26	1.11	1.05	1.37	1.17	1.45	1.52	1.13	0.78	0.77
25P 5x5	1.44	1.28	1.28	1.47	1.48	1.22	1.24	1.44	1.14	1.26
25 ME P10	1.32	1.31	1.33	1.25	1.4	1.27	1.33	1.17	1.04	1.26
25 ME P122	1.28	1.33	1.39	1.35	1.31	1.49	1.24	0.94	0.71	0.91
30IP	1.11	1.07	1.07	1.41	1.17	1.14	1.45	0.77	0.71	1.29
48IP	1.22	1.1	1.18	1.35	1.38	1.23	1.3	1.29	0.91	1.0
CRSP5	1.5	1.68	1.86	1.84	1.92	1.36	2.56	2.37	1.6	1.38
CRSP10	1.31	**1.05**	1.11	**1.06**	1.49	1.3	2.52	2.29	1.89	2.29
CRSP20	1.26	1.19	1.42	1.4	1.35	1.2	2.14	2.33	1.83	2.35
90% quantile										
5IP	5.83	**5.08**	**5.29**	**5.57**	5.93	5.76	6.31	6.38	6.94	6.8
6P 2x3	6.77	**6.0**	**6.08**	**5.9**	**6.41**	**6.75**	7.36	7.59	7.95	8.33
6P ME INV	**6.83**	**5.95**	**6.23**	**6.31**	**6.31**	**6.3**	7.69	7.65	8.78	8.3
6P ME P10	6.63	**5.01**	**4.82**	**5.75**	**6.11**	**6.09**	6.86	7.81	8.51	9.16
6P ME P122	6.66	**4.99**	**5.05**	**5.45**	**5.85**	**5.7**	6.62	6.9	8.24	7.51
EUR 6P 2x3	6.68	**6.13**	6.16	6.1	6.75	6.64	7.2	6.88	7.61	7.67
JPN 6P 2x3	5.69	5.71	5.56	5.9	5.68	6.17	6.51	6.3	7.0	7.57
17IP	**6.51**	**4.73**	**4.78**	**5.0**	**5.39**	**5.76**	10.12	10.65	11.47	11.12
25P 5x5	7.08	**5.64**	**5.69**	**6.44**	**6.77**	7.4	7.83	7.72	8.79	8.5
25 ME P10	**7.19**	**6.05**	**6.11**	**6.01**	**6.13**	**6.61**	**7.5**	9.42	9.74	9.43
25 ME P122	7.53	**5.62**	**5.5**	**6.08**	**5.81**	**6.48**	7.63	8.71	9.82	9.69
30IP	**6.34**	**4.8**	**4.98**	**5.52**	**5.61**	**5.75**	**11.05**	13.19	14.34	15.14
48IP	**6.38**	**4.8**	**4.79**	**5.49**	**5.28**	**7.25**	**10.99**	**12.74**	15.42	15.02
CRSP5	**7.88**	**7.01**	**6.12**	**7.33**	**7.23**	**6.74**	12.81	13.3	14.84	14.93
CRSP10	**6.14**	**4.64**	**4.49**	**5.17**	**5.29**	**6.29**	11.75	12.36	13.85	14.38
CRSP20	**6.05**	**4.19**	**4.28**	**5.29**	**5.21**	**5.66**	11.43	12.34	13.41	13.39

3.6 Conclusion

This paper explores the out-of-sample performance of portfolios formed by an investor with quantile preferences. Unlike traditional approaches that rely on expected utility, quantile preferences focus on maximizing the τ-quantile of future portfolio returns, where $\tau \in (0, 1)$ reflects the investor's risk aversion. Lower τ values correspond to higher risk aversion. Our analysis reveals that portfolios optimized using quantile preferences exhibit substantial differences in weight diversification and turnover, not only across different τ-quantiles but also in comparison to the global minimum variance portfolio. For risk-averse investors, we demonstrate that constructing a portfolio as a global minimum variance portfolio often outperforms direct τ-quantile maximization in terms of out-of-sample τ-quantile of returns. This finding is consistent across various empirical and simulated datasets, different estimation window sizes, and both daily and monthly return frequencies. The observed underperformance of τ-quantile maximization is primarily attributed to small sample bias, as evidenced by our Monte Carlo simulations. These results highlight the challenges associated with quantile-based optimization, particularly in finite samples, and suggest that traditional approaches like the global minimum variance portfolio may offer more robust out-of-sample performance for risk-averse investors. Overall, selecting τ-specific portfolio weights for different risk aversion levels remains a challenge as in practice, choosing the global minimum variance portfolio in case of risk-averse investors leads to the same portfolio weights across different τs. While quantile preferences provide a valuable perspective on investor behaviour and risk management, our findings underscore the importance of carefully considering sample size and estimation techniques when applying this framework in practice. Future research could further refine quantile-based approaches or explore hybrid models that balance the advantages of quantile preferences with the robustness of traditional portfolio optimization methods.

References

Alexander, S., Coleman, T. F., & Li, Y. (2006). Minimizing cvar and var for a portfolio of derivatives. *Journal of Banking & Finance*, *30*(2), 583–605.

Artzner, P., Delbaen, F., Eber, J.-M., & Heath, D. (1999). Coherent measures of risk. *Mathematical Finance*, *9*(3), 203–228.

Barberis, N. (2013). Thirty years of prospect theory in economics: A review and assessment. *Journal of Economic Perspectives*, *27*(1), 173–196.

Barberis, N., & Huang, M. (2008). Stocks as lotteries: The implications of probability weighting for security prices. *American Economic Review*, *98*(5), 2066–2100.

Bassett Jr, G. W., Koenker, R., & Kordas, G. (2004). Pessimistic portfolio allocation and choquet expected utility. *Journal of Financial Econometrics*, *2*(4), 477–492.

Benati, S., & Rizzi, R. (2007). A mixed integer linear programming formulation of the optimal mean/value-at-risk portfolio problem. *European Journal of Operational Research*, *176*(1), 423–434.

Bernard, C., Maj, M., & Vanduffel, S. (2018). Optimal portfolios under the threat of a crash. *Journal of Economic Dynamics and Control*, *90*, 285–301.

Castro, L. d., Galvao, A. F., Montes-Rojas, G., & Olmo, J. (2022). Portfolio selection in quantile decision models. *Annals of Finance*, *18*(2), 133–181.

Clarke, R., De Silva, H., & Thorley, S. (2006). Minimum-variance portfolios in the us equity market. *Journal of Portfolio Management*, *33*(1), 10.

Cont, R. (2001). Empirical properties of asset returns: Stylized facts and statistical issues. *Quantitative Finance*, *1*(2), 223.

de Castro, L., & Galvao, A. F. (2019). Dynamic quantile models of rational behavior. *Econometrica*, *87*(6), 1893–1939.

de Castro, L., & Galvao, A. F. (2022). Static and dynamic quantile preferences. *Economic Theory*, *73*(2), 747–779.

de Castro, L., Galvao, A. F., Noussair, C. N., & Qiao, L. (2022). Do people maximize quantiles? *Games and Economic Behavior*, *132*, 22–40.

DeMiguel, V., Garlappi, L., & Uppal, R. (2009). Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? *The Review of Financial Studies*, *22*(5), 1915–1953.

Gaivoronski, A. A., & Pflug, G. (2005). Value-at-risk in portfolio optimization: Properties and computational approach. *Journal of Risk*, *7*(2), 1–31.

Garlappi, L., Uppal, R., & Wang, T. (2007). Portfolio selection with parameter and model uncertainty: A multi-prior approach. *Review of Financial Studies*, *20*(1), 41–81.

Gilboa, I., & Schmeidler, D. (1989). Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics*, *18*(2), 141–153.

Goldfarb, D., & Iyengar, G. (2003). Robust portfolio selection problems. *Mathematics of Operations Research*, *28*(1), 1–38.

Chambers, C. P. (2009). An axiomatization of quantiles on the domain of distribution functions. *Mathematical Finance: An International Journal of Mathematics, Statistics and Financial Economics*, *19*(2), 335–342.

Choueifaty, Y., & Coignard, Y. (2008). Toward maximum diversification. *Journal of Portfolio Management*, *35*(1), 40–51.

Jagannathan, R., & Ma, T. (2003). Risk reduction in large portfolios: Why imposing the wrong constraints helps. *The Journal of Finance*, *58*(4), 1651–1683.

Jorion, P. (1986). Bayes-stein estimation for portfolio analysis. *Journal of Financial and Quantitative Analysis*, *21*(3), 279–292.

Jorion, P. (2007). *Value at risk: The new benchmark for managing financial risk* (3rd ed.). McGraw-Hill.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–291.

Kan, R., & Zhou, G. (2007). Optimal portfolio choice with parameter uncertainty. *Journal of Financial and Quantitative Analysis*, *42*(3), 621–656.

Keating, C., & Shadwick, W. F. (2002). Universal performance measures. *Journal of Performance Measurement*, *6*(3), 59–84.

Ledoit, O., & Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, *88*(2), 365–411.

Lopez de Prado, M. (2016). Hierarchical risk parity: Determining optimal diversification using cluster analysis. *Journal of Portfolio Management*, *42*(4), 59–69.

Machina, M. J. (1982). Expected utility analysis without the independence axiom. *Econometrica*, *50*(2), 277–323.

Manski, C. F. (1988). Ordinal utility models of decision making under uncertainty. *Theory and Decision*, *25*, 79–104.

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, *7*(1), 77–91.

Meucci, A. (2008). Fully flexible views: Theory and practice. *Risk*, *21*(10), 97–102.

Michaud, R. O. (1989). The markowitz optimization enigma: Is 'optimized' optimal? *Financial Analysts Journal*, *45*(1), 31–42.

Natarajan, K., Pachamanova, D., & Sim, M. (2009). Constructing risk measures from uncertainty sets. *Operations Research*, *57*(5), 1129–1141.

Pástor, L. (2000). Portfolio selection and asset pricing models. *The Journal of Finance*, *55*(1), 179–223.

Rabin, M. (2013). Risk aversion and expected-utility theory: A calibration theorem. In *Handbook of the fundamentals of financial decision making: Part i* (pp. 241–252). World Scientific.

Rockafellar, R. T., & Uryasev, S. (2000). Optimization of conditional value-at-risk. *Journal of Risk*, *2*(3), 21–41.

Rostek, M. (2010). Quantile maximization in decision theory. *The Review of Economic Studies*, *77*(1), 339–371.

Roy, A. D. (1952). Safety first and the holding of assets. *Econometrica: Journal of the Econometric Society*, 431–449.

Shefrin, H., & Statman, M. (2000). Behavioral portfolio theory. *Journal of Financial and Quantitative Analysis*, *35*(2), 127–151.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*(4), 297–323.

Appendix

3.A Evaluation

Besides using the out-of-sample quantiles of portfolio returns to evaluate the performance of the portfolios, as well as other well-established performance measures such as standard deviation, Sharpe ratio, and maximum drawdown, we also consider the Herfindahl-Hirschman index, weight turnover, and the significance of the differences in the out-of-sample quantiles of portfolio returns, which we define in the following paragraphs.

Herfindahl-Hirschman index and portfolio turnover

The (non-normalized) Herfindahl-Hirschman index is a measure of concentration and is defined as $HH_{\text{non}} = \sum_{i=1}^{N} w_i^2$ where w_i is the portfolio weight of an asset i and N is the number of assets. We mark the non-normalized Herfindahl-Hirschman index as HH_{non} to distinguish it from the normalized version which we report in the text. A high HH_{non} value corresponds to a very concentrated portfolio, while a low HH_{non} value corresponds to a very diversified portfolio in terms of weights. It is easy to see that an equally-weighted portfolio is the most diversified one in terms of weights, while a portfolio consisting of only one asset achieves the highest HH-index. To allow easier comparison between portfolios with a different number of assets we show the normalized version defined as

$$HH = \frac{HH_{non} - 1/N}{1 - 1/N}$$

for $N >= 2$

We also report the average portfolio weights turnover between two periods t and $t + 1$, defined as

$$\Delta = \frac{1}{N} \sum_{i=1}^{N}$$

where $w_{i,t}$ is the portfolio weight of an asset i at period t and N is the number of assets. This definition of turnover does not correspond to the standardly used turnover definition in the literature, as we do not consider the inter-period evolution of the assets themselves, but only the turnover based on the changes between selected portfolio weights from one period to another. This shows the effect of the estimation sample changes on the selected weights without the

effect of the previous period's asset returns on the initial portfolio weights.

Testing difference in the quantiles

Further, to evaluate the significance of the differences in the out-of-sample quantiles of portfolio returns, we use a percentile bootstrap method. First, generate bootstrap samples for each out-of-sample portfolio return series. Let $\hat{Q^*_{\tau,j}}$ be the bootstrap estimate of $Q_{\tau,j}$, where $Q_{\tau,j}$ represents the quantile at level τ associated with the j-th portfolio returns ($j = 1, 2$). Define the difference between the bootstrap estimates as

$$D^* = \hat{Q^*_{\tau,1}} - \hat{Q^*_{\tau,2}}$$

Repeat this process B times to obtain the set of differences $D^*_1, \ldots, D^*_B$. Let $l = \alpha(B/2)$[7] and $u = B - l$. Then, an approximate $1 - \alpha$ confidence interval for $Q_{\tau,1} - Q_{\tau,2}$ is given by

$$(D^*_{(l+1)}, D^*_{(u)})$$

where $D^*_{(1)} \leq \ldots \leq D^*_{(B)}$ are the ordered bootstrap differences. In order to test the null hypothesis

$$H_0 : Q_{\tau,1} = Q_{\tau,2}$$

we use the bootstrap estimates $\hat{Q^*_{\tau,1}}$ and $\hat{Q^*_{\tau,2}}$ and define

$$p^* = P(\hat{Q^*_{\tau,1}} > \hat{Q^*_{\tau,2}})$$

If the null hypothesis is true, then asymptotically p^* follows a uniform distribution. Consequently, we reject H_0 if $p^* \leq \alpha/2$ or if $p^* \geq 1 - \alpha/2$. We estimate p^* as $\hat{p}^* = \frac{A}{B}$, where A is the number of values among $D^*_1, \ldots, D^*_B$ that are greater than zero.

3.B Global minimum variance portfolio

The global minimum variance portfolio weights are selected by solving a quadratic optimization problem where the objective is to minimize the portfolio variance, subject to the constraint that the sum of the portfolio weights equals one. We further add the constraint of non-negative weights in order to reduce

[7] Rounded to the nearest integer.

the estimation error of the covariance matrix by imposing a structure on the sample-based covariance estimate, see Jagannathan and Ma (2003).

Given n assets, let $\mathbf{w} = (w_1, w_2, \ldots, w_n)^T$ represent the vector of portfolio weights, and Σ be the $n \times n$ covariance matrix of asset returns. The optimization problem is formulated as:

$$\min_{\mathbf{w}} \mathbf{w}^T \Sigma \mathbf{w} \quad \text{subject to} \quad \mathbf{1}^T \mathbf{w} = 1 \quad \text{and} \quad \mathbf{w} \geq 0$$

where $\mathbf{1}$ is an $n \times 1$ vector of ones. The solution to this problem provides the weights for the GMV portfolio, which are given by:

$$\mathbf{w}_{\text{gmv}} = \frac{\Sigma^{-1}\mathbf{1}}{\mathbf{1}^T \Sigma^{-1} \mathbf{1}}$$

The accuracy of the global minimum variance portfolio critically depends on the estimation of the covariance matrix Σ. Two commonly used methods for estimating this covariance matrix are the sample covariance estimator and the Ledoit-Wolf shrinkage estimator, Ledoit and Wolf (2004). In this approach, the sample covariance matrix $\hat{\Sigma}_{\text{sample}}$ is combined with a structured target matrix, often the identity matrix, to produce a more stable and robust estimate. The Ledoit-Wolf estimator is defined as:

$$\hat{\Sigma}_{\text{LW}} = \rho F + (1 - \rho)\hat{\Sigma}_{\text{sample}}$$

where $F = \frac{\text{Tr}(\hat{\Sigma}_{\text{sample}})}{n} I$ is the target matrix, Tr is trace, I is the identity matrix, and ρ is the shrinkage intensity. The shrinkage coefficient ρ is computed by minimizing the mean-squared error between the true covariance matrix and its estimate. This shrinkage approach mitigates estimation error and enhances the reliability of the portfolio weights, especially in high-dimensional settings where the sample covariance matrix alone may be unstable.

3.C Additional simulation results

Besides analyzing the results based on the normally distributed returns, which should favour the global minimum variance portfolio, we also simulate returns that are not normally distributed. We simulate a time series of $T = 300$ monthly returns for $N = 5$ assets using a single-factor model $r_t = \beta f_t + \epsilon_t, t = 1, \ldots T$, where r_t is the vector of asset returns, f_t is the vector of market factor returns, β is the vector of the factor loading and ϵ_t is the vector of residuals.

We follow Pástor (2000) and DeMiguel et al. (2009) in the simulation design. Specifically, for the market factor returns we assume mean monthly returns of $\frac{8}{12}\%$ and a monthly standard deviation of $\frac{16}{\sqrt{12}}\%$. We assume the factor loadings to be uniformly distributed $f_t \sim U(0.5, 1.5)$ and residuals to be t-distributed with 3 degrees of freedom. The results in Table 3.7 and Table 3.6 show the weight statistics and out-of-sample τ-quantiles for the same simulation design as in the previous case but with non-normal returns. Using the simulated returns, we proceed with a standard "rolling" estimation window approach. We use 60-month-long estimation windows to find optimal portfolio weights and rebalance each month.

Table 3.6. Herfindahl-Hirschman index and turnover
This table shows the average normalized Herfindahl-Hirschman indices (HH) and average as well as median portfolio weights turnover (Δ) for the τ-quantile maximizing (qm), Equal Weighted (ew), Global Minimum Variance (gmv) and Global Minimum Variance using Ledoit-Wolf covariance (gmv_lw) portfolios. Returns data are simulated using a single-factor model with $N = 5$ assets and $T = 300$ periods, where residuals are distributed according to the t-distribution with 3 degrees of freedom leading to non-normally distributed returns. Portfolio weights are selected using a 60 observations long rolling estimation window. Simulation is repeated 1000 times and results show the average values across individual simulation runs.

strategy	ew	gmv_lw	gmv	qm0.05	qm0.1	qm0.2	qm0.3	qm0.4	qm0.5	qm0.6	qm0.7	qm0.8	qm0.9	qm0.95
HH	0.0	0.00	0.05	0.13	0.13	0.19	0.30	0.42	0.58	0.69	0.83	0.88	0.97	0.97
Δ	0.0	0.02	0.05	0.06	0.12	0.16	0.21	0.26	0.31	0.31	0.19	0.18	0.06	0.09
Δ (med)	0.0	0.00	0.04	0.00	0.00	0.00	0.06	0.06	0.07	0.02	0.00	0.00	0.00	0.00

Table 3.7. Out-of-sample performance of the simulated portfolios This table shows the standard deviation and τ-quantiles of out-of-sample portfolio returns for the τ-quantile maximizing (qm), Equal Weighted (ew), Global Minimum Variance (gmv) and Global Minimum Variance using Ledoit-Wolf covariance (gmv_lw) portfolios. Returns data are simulated using a single-factor model with $N = 5$ assets and $T = 300$ periods where residuals are distributed according to the t-distribution with 3 degrees of freedom leading to non-normally distributed returns. Portfolio weights are selected using a 60 observations long rolling estimation window. Simulation is repeated 1000 times, and results show the average values across individual simulation runs. Results are shown in percentage points.

	ew	gmv_lw	gmv	qm0.05	qm0.1	qm0.2	qm0.3	qm0.4	qm0.5	qm0.6	qm0.7	qm0.8	qm0.9	qm0.95
σ	13.16	13.22	13.71	15.05	15.12	14.94	15.91	18.38	20.50	20.97	22.08	22.41	23.88	23.10
0.05	-20.52	-20.27	-20.13	-24.22	-22.63	-23.48	-24.06	-30.10	-32.52	-33.89	-33.98	-35.87	-37.72	-37.08
0.1	-15.39	-15.48	-16.60	-18.22	-19.10	-18.06	-18.67	-21.57	-24.47	-25.29	-26.82	-26.03	-28.71	-25.84
0.2	-10.01	-9.98	-10.52	-11.89	-11.85	-11.03	-12.08	-13.15	-13.68	-15.62	-17.58	-15.21	-15.91	-15.49
0.3	-6.05	-6.13	-6.92	-7.50	-7.10	-6.41	-6.49	-6.96	-8.78	-9.12	-9.84	-9.88	-9.13	-9.40
0.4	-2.66	-2.70	-3.34	-3.65	-3.18	-3.05	-3.06	-2.48	-4.33	-4.58	-4.04	-4.09	-2.01	-2.25
0.5	1.15	1.13	1.26	0.73	0.54	1.06	1.39	1.62	0.34	0.03	1.70	2.24	3.30	2.88
0.6	4.20	4.01	4.22	4.31	4.38	4.68	4.59	5.26	5.27	5.59	6.81	7.42	8.27	7.74
0.7	7.27	7.38	7.78	8.14	8.55	8.04	8.48	9.57	10.80	11.44	12.22	13.04	14.00	13.64
0.8	11.31	10.90	11.62	12.97	13.10	12.87	13.70	15.93	17.34	17.85	20.85	20.53	22.16	21.38
0.9	18.00	17.87	18.38	19.53	18.87	19.61	20.16	23.32	24.44	26.91	29.65	29.87	32.53	32.14
0.95	22.09	22.72	23.26	24.44	22.06	24.55	25.63	27.36	35.44	35.00	38.52	38.10	42.42	41.33

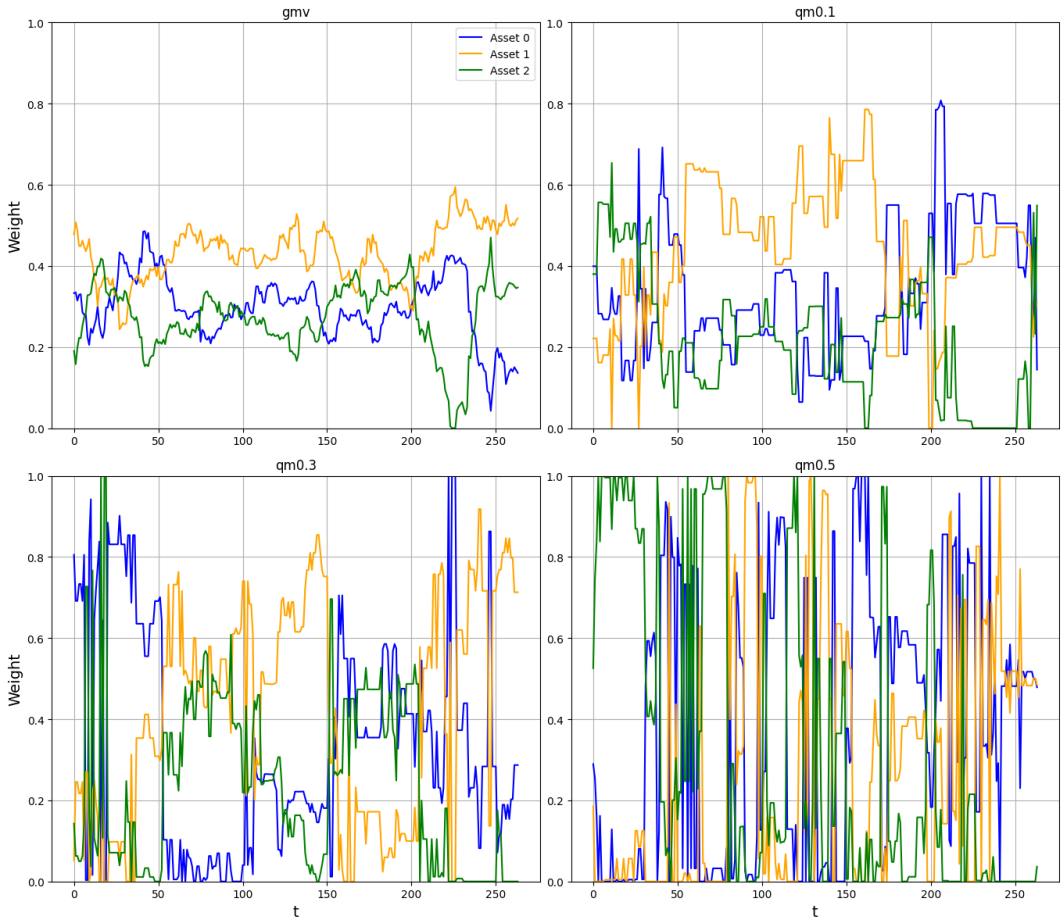


Figure 3.4. Evolution of portfolio weights over time
This figure shows an example evolution of portfolio weights over time for the $10\%, 30\%$ and 50%-quantile maximizing portfolios and global minimum variance portfolio. Returns data are simulated using a single-factor model with $N = 3$ assets and $T = 300$ periods where residuals are normally distributed. Rolling estimation window with 60 observations is used to select portfolio weights.

3.D Additional empirical results

Besides presenting quantiles of out-of-sample portfolio returns in the main text, we also document the standard deviation, Sharpe Ratio, and Maximum Drawdown of out-of-sample portfolio returns in Table 3.8. Additionally, we provide robustness checks for the results in the main text by considering a different rolling window length of 36 months instead of the 60 months, see Table 3.9, Table 3.10 and Table 3.11, as well as different frequencies of the rolling window, see Table 3.4, Table 3.5 and Table 3.8.

3.E Binomial tests for number of outperformance cases

Tables 3.15, 3.16 and 3.17 report results of binomial tests for the outperformance of the benchmark strategy Global Minimum Variance (gmv) against τ-quantile maximizing (qm) strategy with risk-aversion coefficient $\tau \in \{0.1, 0.3\}$ across different datasets. Global Minimum Variance using Ledoit-Wolf covariance matrix (gmv_lw) and Equal Weighted (ew) portfolios are compared as a reference point as well. The binomial test is used to assess whether the benchmark strategy consistently outperforms the competing strategy more than 50% of the time. $H_0 : p \leq 0.5$, where p is the probability that the benchmark strategy gmv outperforms the competing strategy. The alternative hypothesis is $H_a : p > 0.5$. Tables report the number of datasets where the benchmark strategy outperforms the competing strategy and the p-value of the binomial test. The p-value is calculated as the probability of observing the number of outperformances equal to or greater than the observed number of outperformances under the null hypothesis.

We can see that the benchmark strategy Global Minimum Variance (gmv) outperforms the τ-quantile maximizing (qm) strategies with $\tau = 0.1$ and $\tau = 0.3$ in more than 50% of the cases, and results are statistically significant at the 5% level for 60-month long estimation windows.

It is important to note that some of the datasets utilized in this analysis are not independent of one another. For instance, industry portfolios with varying numbers of industries (e.g., 5, 10, 30 industries) are derived from the same underlying data. This lack of independence between certain datasets may affect the validity of the binomial test results, as the test assumes independent observations. Consequently, the reported p-values and confidence intervals should be interpreted with caution, as the effective sample size may be smaller than the nominal count of datasets.

Table 3.8. Out-of-sample portfolio risk metrics for 60 months long estimation window

This table shows the standard deviation, Sharpe Ratio, and Maximum Drawdown for the τ-quantile maximizing (qm), Equal Weighted (ew), Global Minimum Variance (gmv) and Global Minimum Variance using Ledoit-Wolf covariance matrix (gmv_lw) portfolios. A rolling window sample with a monthly frequency and a length of 60 months is used. Results are shown in percentage points and for datasets described in Section 3.5.1. Bold values indicate that the standard deviation of out-of-sample portfolio returns is significantly different from the standard deviation of Global Minimum Variance portfolios at the 5% significance level. Statistical significance is obtained using an F-test for the equality of variances.

	ew	gmv_lw	gmv	qm0.05	qm0.1	qm0.3	qm0.5	qm0.7	qm0.9	qm0.95
Standard Deviation										
5IP	4.31	3.91	3.9	4.15	4.17	4.26	**4.69**	**5.26**	**5.51**	**5.56**
6P 2x3	**5.28**	4.44	4.46	4.82	4.89	**5.07**	**5.57**	**6.09**	**6.37**	**6.39**
6P ME INV	**5.11**	4.31	4.31	4.49	4.57	4.7	**5.58**	**5.85**	**6.47**	**6.32**
6P ME P10	**5.06**	4.02	4.01	4.24	4.25	4.44	**5.55**	**6.17**	**6.57**	**6.96**
6P ME P122	**5.14**	4.02	4.05	4.23	4.41	**4.7**	**5.14**	**5.88**	**7.14**	**6.03**
EUR 6P 2x3	5.48	4.93	4.9	5.14	5.3	5.31	**5.85**	**5.89**	**6.12**	**6.24**
JPN 6P 2x3	4.5	4.36	4.38	4.47	4.47	4.8	**5.19**	**5.37**	**5.38**	**5.9**
17IP	**4.99**	3.45	3.47	3.83	**3.91**	**4.58**	**7.15**	**8.01**	**8.91**	**8.9**
25P 5x5	**5.5**	4.31	4.29	4.78	**4.89**	**5.66**	**5.92**	**6.61**	**7.31**	**7.39**
25 ME P10	**5.64**	4.34	4.34	**4.9**	4.57	**5.05**	**6.12**	**7.11**	**7.73**	**7.87**
25 ME P122	**5.72**	4.31	4.35	4.75	4.56	**4.99**	**6.4**	**7.53**	**8.79**	**7.73**
30IP	**4.97**	3.54	3.55	**4.16**	**4.0**	**4.76**	**7.82**	**9.87**	**12.01**	**11.68**
48IP	**4.93**	3.52	3.52	**4.21**	**4.0**	**5.31**	**8.14**	**9.96**	**11.65**	**11.04**
CRSP5	**6.41**	5.24	4.9	4.92	4.99	5.6	**8.99**	**9.5**	**10.18**	**10.74**
CRSP10	**5.0**	3.53	3.47	**4.11**	3.9	**4.89**	**8.07**	**9.11**	**9.65**	**10.14**
CRSP20	**4.36**	3.26	3.32	**3.87**	**3.86**	**4.56**	**7.57**	**8.89**	**9.69**	**10.1**
Sharpe Ratio										
5IP	0.69	0.76	0.76	0.82	0.91	0.69	0.63	0.58	0.52	0.49
6P 2x3	0.58	0.68	0.69	0.66	0.75	0.64	0.59	0.44	0.46	0.38
6P ME INV	0.62	0.73	0.74	0.77	0.73	0.74	0.53	0.51	0.47	0.54
6P ME P10	0.52	0.67	0.65	0.62	0.70	0.68	0.47	0.50	0.48	0.49
6P ME P122	0.55	0.67	0.69	0.62	0.59	0.66	0.55	0.60	0.49	0.47
EUR 6P 2x3	0.50	0.68	0.72	0.63	0.62	0.61	0.52	0.42	0.32	0.35
JPN 6P 2x3	0.46	0.48	0.48	0.49	0.50	0.48	0.39	0.42	0.31	0.26
17IP	0.66	0.79	0.78	0.79	0.76	0.73	0.54	0.39	0.26	0.30
25P 5x5	0.57	0.66	0.66	0.66	0.81	0.55	0.57	0.36	0.31	0.27
25 ME P10	0.56	0.77	0.77	0.50	0.74	0.63	0.50	0.41	0.43	0.45
25 ME P122	0.59	0.73	0.72	0.61	0.68	0.75	0.49	0.43	0.28	0.42
30IP	0.64	0.69	0.67	0.82	0.68	0.63	0.69	0.35	0.15	0.37
48IP	0.66	0.76	0.77	0.75	0.78	0.68	0.63	0.38	0.56	0.50
CRSP5	0.74	0.80	0.82	0.99	0.97	0.72	0.95	0.76	0.64	0.61
CRSP10	0.72	0.80	0.80	0.68	0.89	0.69	0.93	0.80	0.72	0.68
CRSP20	0.76	0.91	0.92	0.84	0.87	0.72	0.95	0.80	0.60	0.59
Maximum Drawdown										
5IP	-46.98	-38.38	-36.97	-45.80	-37.85	-40.33	-49.70	-48.49	-48.24	-54.74
6P 2x3	-54.31	-50.59	-47.74	-57.21	-51.84	-59.44	-58.31	-58.81	-57.51	-48.28
6P ME INV	-51.86	-49.08	-48.58	-49.60	-50.21	-50.83	-51.53	-55.71	-57.97	-57.32
6P ME P10	-56.42	-42.45	-41.93	-46.99	-50.11	-42.46	-61.61	-60.98	-64.74	-64.10
6P ME P122	-55.92	-47.47	-45.93	-51.10	-56.97	-47.70	-54.18	-57.38	-54.90	-60.67
EUR 6P 2x3	-60.14	-53.70	-52.02	-56.51	-56.31	-61.57	-65.11	-62.78	-60.94	-62.40
JPN 6P 2x3	-49.93	-43.86	-43.58	-42.74	-43.01	-52.08	-58.01	-49.27	-62.76	-68.91
17IP	-51.71	-34.97	-33.67	-38.65	-38.01	-43.01	-66.19	-70.16	-76.14	-69.54
25P 5x5	-54.30	-51.81	-51.08	-54.49	-52.23	-63.40	-55.06	-62.31	-62.61	-61.43
25 ME P10	-56.61	-46.06	-46.24	-48.23	-50.53	-53.40	-63.85	-64.20	-64.86	-65.51
25 ME P122	-56.76	-47.54	-47.58	-52.29	-47.30	-49.33	-59.17	-58.20	-64.72	-64.08
30IP	-53.09	-39.07	-38.55	-43.06	-41.77	-42.58	-66.37	-85.40	-98.47	-79.20
48IP	-52.60	-39.96	-39.03	-43.40	-38.95	-43.06	-65.00	-83.24	-79.20	-79.20
CRSP5	-60.26	-60.47	-56.52	-47.92	-54.22	-45.30	-56.39	-56.35	-57.98	-57.98
CRSP10	-46.66	-36.72	-33.90	-41.03	-31.38	-49.77	-53.19	-56.35	-56.91	-56.91
CRSP20	-42.53	-30.68	-29.90	-38.17	-33.99	-40.72	-50.00	-55.13	-56.91	-56.91

Table 3.9. Herfindahl-Hirschman index and turnover of portfolio weights for 36 months long estimation window

This table shows the normalized Herchchendahl-Hirschman index, mean turnover and median turnover (Δ (median)) of portfolio weights for the τ-quantile maximizing (qm), Equal Weighted (ew), Global Minimum Variance (gmv) and Global Minimum Variance using Ledoit-Wolf covariance matrix (gmv_lw) portfolios. A rolling window sample with a length of 36 months is used. Results are shown in percentage points and for datasets described in Subsection 3.5.1.

	ew	gmv_lw	gmv	qm0.05	qm0.1	qm0.3	qm0.5	qm0.7	qm0.9	qm0.95
Herfindahl-Hirschman Index										
5IP	0.0	0.25	0.38	0.47	0.47	0.39	0.43	0.67	0.88	0.92
6P 2x3	0.0	0.43	0.61	0.60	0.65	0.51	0.52	0.69	0.84	0.92
6P ME INV	0.0	0.41	0.74	0.60	0.56	0.54	0.50	0.64	0.83	0.92
6P ME P10	0.0	0.46	0.72	0.75	0.73	0.51	0.47	0.60	0.81	0.94
6P ME P122	0.0	0.39	0.53	0.59	0.53	0.52	0.44	0.59	0.84	0.91
EUR 6P 2x3	0.0	0.52	0.74	0.62	0.60	0.63	0.55	0.60	0.77	0.86
JPN 6P 2x3	0.0	0.32	0.59	0.59	0.47	0.48	0.51	0.56	0.71	0.95
17IP	0.0	0.20	0.31	0.40	0.32	0.28	0.51	0.67	0.87	0.91
25P 5x5	0.0	0.31	0.55	0.57	0.52	0.39	0.39	0.52	0.70	0.86
25 ME P10	0.0	0.29	0.51	0.58	0.49	0.36	0.38	0.48	0.74	0.83
25 ME P122	0.0	0.28	0.55	0.51	0.47	0.37	0.39	0.51	0.76	0.88
30IP	0.0	0.13	0.24	0.33	0.29	0.24	0.49	0.70	0.87	0.97
48IP	0.0	0.11	0.20	0.23	0.23	0.23	0.46	0.66	0.85	0.97
CRSP5	0.0	0.18	0.42	0.34	0.34	0.35	0.54	0.79	0.93	0.89
CRSP10	0.0	0.11	0.21	0.24	0.25	0.27	0.50	0.75	0.91	0.90
CRSP20	0.0	0.07	0.15	0.21	0.21	0.25	0.50	0.71	0.85	0.90
Portfolio Turnover										
5IP	0.0	0.09	0.12	0.15	0.24	0.38	0.35	0.37	0.14	0.13
6P 2x3	0.0	0.08	0.12	0.15	0.20	0.40	0.43	0.33	0.17	0.13
6P ME INV	0.0	0.08	0.11	0.16	0.24	0.45	0.47	0.34	0.19	0.13
6P ME P10	0.0	0.08	0.08	0.11	0.25	0.44	0.50	0.32	0.17	0.08
6P ME P122	0.0	0.08	0.12	0.16	0.26	0.52	0.52	0.38	0.21	0.12
EUR 6P 2x3	0.0	0.09	0.08	0.11	0.24	0.32	0.45	0.39	0.16	0.14
JPN 6P 2x3	0.0	0.12	0.15	0.19	0.27	0.40	0.50	0.42	0.18	0.10
17IP	0.0	0.13	0.17	0.20	0.32	0.60	0.54	0.34	0.14	0.16
25P 5x5	0.0	0.15	0.19	0.25	0.39	0.61	0.76	0.45	0.26	0.17
25 ME P10	0.0	0.14	0.20	0.20	0.37	0.70	0.73	0.48	0.25	0.15
25 ME P122	0.0	0.14	0.18	0.23	0.38	0.78	0.68	0.52	0.31	0.18
30IP	0.0	0.17	0.23	0.29	0.39	0.74	0.52	0.35	0.18	0.11
48IP	0.0	0.18	0.24	0.33	0.45	0.80	0.57	0.37	0.22	0.12
CRSP5	0.0	0.06	0.07	0.10	0.17	0.38	0.35	0.21	0.18	0.07
CRSP10	0.0	0.09	0.12	0.17	0.28	0.48	0.38	0.26	0.14	0.08
CRSP20	0.0	0.11	0.18	0.29	0.37	0.64	0.37	0.31	0.13	0.14
Portfolio Turnover (median)										
5IP	0.0	0.07	0.09	0.0	0.00	0.18	0.19	0.0	0.0	0.0
6P 2x3	0.0	0.05	0.04	0.0	0.00	0.10	0.22	0.0	0.0	0.0
6P ME INV	0.0	0.05	0.02	0.0	0.00	0.17	0.23	0.0	0.0	0.0
6P ME P10	0.0	0.05	0.03	0.0	0.00	0.14	0.38	0.0	0.0	0.0
6P ME P122	0.0	0.04	0.07	0.0	0.00	0.28	0.35	0.0	0.0	0.0
EUR 6P 2x3	0.0	0.06	0.03	0.0	0.00	0.00	0.17	0.0	0.0	0.0
JPN 6P 2x3	0.0	0.09	0.09	0.0	0.00	0.11	0.25	0.0	0.0	0.0
17IP	0.0	0.09	0.12	0.0	0.00	0.51	0.24	0.0	0.0	0.0
25P 5x5	0.0	0.11	0.12	0.0	0.00	0.40	0.67	0.0	0.0	0.0
25 ME P10	0.0	0.10	0.13	0.0	0.00	0.65	0.58	0.0	0.0	0.0
25 ME P122	0.0	0.10	0.11	0.0	0.00	0.71	0.57	0.0	0.0	0.0
30IP	0.0	0.14	0.19	0.0	0.00	0.70	0.30	0.0	0.0	0.0
48IP	0.0	0.15	0.18	0.0	0.11	0.79	0.31	0.0	0.0	0.0
CRSP5	0.0	0.04	0.05	0.0	0.00	0.18	0.09	0.0	0.0	0.0
CRSP10	0.0	0.06	0.09	0.0	0.00	0.40	0.20	0.0	0.0	0.0
CRSP20	0.0	0.09	0.14	0.0	0.17	0.61	0.15	0.0	0.0	0.0

Table 3.10. Out-of-sample 10%, 50% and 90% quantiles of portfolio returns for 36 months long estimation window This table shows the 10%, 50% and 90% quantile of portfolio returns for the τ-quantile maximizing (qm), Equal Weighted (ew), Global Minimum Variance (gmv) and Global Minimum Variance using Ledoit-Wolf covariance matrix (gmv_lw) portfolios. A rolling window sample with a length of 36 months is used. Results are shown in percentage points and for datasets described in Section 3.5.1. Bold values indicate that the out-of-sample quantiles of portfolio returns are significantly different from the quantile maximization strategy targeting the given τ at the 5% significance level. Statistical significance is determined using a percentile bootstrap method described in Section 3.A.

	ew	gmv_lw	gmv	qm0.05	qm0.1	qm0.3	qm0.5	qm0.7	qm0.9	qm0.95
					10% quantile					
5IP	-5.34	-4.22	-4.31	-4.62	-4.46	-4.92	-5.28	**-6.86**	**-7.62**	**-7.42**
6P 2x3	-5.76	-4.49	-4.64	-4.83	-5.01	-5.17	-5.95	**-7.46**	**-8.08**	**-8.14**
6P ME INV	-5.97	-4.62	-4.56	-4.94	-4.83	-5.02	**-6.7**	**-7.49**	**-7.61**	**-7.35**
6P ME P10	-6.19	-4.79	-4.75	-5.06	-4.9	-5.66	-6.87	**-8.12**	**-7.97**	**-7.69**
6P ME P122	-5.84	-4.86	-5.0	-5.49	-5.76	-5.38	-5.38	-6.92	**-7.83**	**-7.62**
EUR 6P 2x3	-5.12	-5.06	-5.09	-5.42	-4.86	-5.22	-5.76	-6.48	-7.17	-7.35
JPN 6P 2x3	-5.32	-5.22	-4.88	-5.28	-5.26	-5.53	-6.65	-6.17	-6.68	**-6.92**
17IP	-4.85	-3.89	-3.98	-4.03	-4.87	-4.49	**-7.79**	**-9.15**	**-9.22**	**-10.02**
25P 5x5	-5.95	-4.5	-4.54	-5.09	-4.73	-5.14	-6.38	**-8.26**	**-9.78**	**-9.97**
25 ME P10	-6.16	-4.85	-4.94	-5.05	-5.36	-5.25	-6.44	**-8.54**	**-9.54**	**-9.21**
25 ME P122	-6.02	-5.09	-4.95	-5.85	-5.49	-6.02	-6.45	**-8.31**	**-8.28**	**-8.56**
30IP	-4.58	-3.6	-3.76	-4.52	-4.18	-5.48	**-8.85**	**-10.5**	**-13.6**	**-13.64**
48IP	-4.86	-3.51	-3.37	-4.24	-4.45	-6.0	**-9.69**	**-9.59**	**-13.05**	**-12.7**
CRSP5	-7.62	-5.41	-5.26	-4.85	-5.68	-5.9	**-9.06**	**-9.77**	**-12.51**	**-12.7**
CRSP10	-5.72	-4.74	-4.08	-4.56	-5.62	-5.86	**-8.73**	**-9.64**	**-13.1**	**-12.53**
CRSP20	-4.03	-3.63	-3.64	-4.22	-4.3	-5.8	**-8.74**	**-9.9**	**-11.11**	**-11.72**
					50% quantile					
5IP	1.35	1.27	1.31	1.32	1.32	1.09	1.19	1.09	1.09	1.07
6P 2x3	1.38	1.36	1.24	1.16	1.29	1.24	1.39	1.31	1.32	1.02
6P ME INV	1.33	1.38	1.2	1.39	1.41	1.23	1.42	1.34	1.32	1.46
6P ME P10	1.25	1.27	1.33	1.28	1.28	1.35	1.27	1.34	1.2	1.28
6P ME P122	1.29	1.3	1.31	1.22	1.13	1.54	1.09	1.2	0.91	1.12
EUR 6P 2x3	0.96	1.25	1.25	1.19	1.05	1.45	0.84	0.66	0.75	0.62
JPN 6P 2x3	0.8	0.65	0.59	0.58	0.61	0.88	0.83	0.85	0.62	0.92
17IP	1.07	0.92	0.98	0.98	1.02	1.01	1.14	0.43	0.71	0.34
25P 5x5	1.44	1.23	1.15	1.29	1.41	1.43	1.29	1.45	1.34	1.22
25 ME P10	1.32	1.24	1.16	1.23	1.14	1.24	1.33	1.02	1.13	1.0
25 ME P122	1.27	1.29	1.19	1.22	1.09	1.6	1.29	1.01	0.45	0.78
30IP	1.06	0.94	0.77	1.04	1.04	1.03	1.21	1.1	0.86	1.15
48IP	1.11	0.94	0.83	1.19	0.97	1.12	1.04	0.82	1.0	1.62
CRSP5	1.39	1.29	1.31	1.52	1.59	1.28	2.51	1.85	1.69	1.08
CRSP10	1.22	0.91	0.9	1.29	1.22	1.08	1.92	2.01	0.9	1.06
CRSP20	1.06	0.99	1.22	1.45	1.35	0.84	2.2	1.33	1.57	0.91
					90% quantile					
5IP	**5.88**	**5.31**	**5.43**	**5.7**	**5.85**	**5.79**	6.46	6.74	7.29	7.34
6P 2x3	6.79	**6.11**	**6.23**	**6.42**	**6.42**	6.98	7.31	8.23	8.17	8.3
6P ME INV	7.08	**5.98**	**5.93**	**6.37**	**6.28**	**6.44**	7.37	7.7	8.43	9.04
6P ME P10	**7.22**	**6.1**	**5.88**	**6.58**	**6.22**	**7.06**	7.71	8.57	9.18	9.39
6P ME P122	**7.14**	**5.71**	**5.77**	**6.19**	**5.96**	6.8	7.26	8.08	8.47	8.09
EUR 6P 2x3	**6.61**	**6.34**	**6.33**	6.69	6.56	6.61	6.86	7.11	7.7	7.3
JPN 6P 2x3	5.92	5.69	5.48	5.69	6.31	5.97	6.5	6.5	7.04	7.32
17IP	**6.49**	**4.87**	**4.88**	**5.0**	**5.18**	**6.1**	9.9	11.13	11.96	12.03
25P 5x5	**7.1**	**5.64**	**5.7**	**6.43**	**6.8**	**6.65**	7.84	8.23	9.08	9.02
25 ME P10	**7.23**	**6.01**	**5.96**	**6.29**	**6.41**	**7.07**	**7.29**	8.96	9.3	10.24
25 ME P122	7.52	**5.56**	**5.54**	**6.46**	**6.42**	7.0	8.27	9.2	10.01	9.42
30IP	**6.33**	**4.8**	**4.75**	**5.0**	**5.8**	**6.95**	**10.93**	13.7	16.05	15.8
48IP	**6.37**	**4.74**	**4.82**	**5.46**	**5.97**	**6.8**	**11.83**	**13.46**	15.93	16.09
CRSP5	**9.17**	**6.88**	**6.06**	**7.21**	**7.77**	**8.54**	14.76	14.19	15.07	16.65
CRSP10	**6.66**	**5.09**	**4.81**	**6.56**	**5.71**	**6.46**	12.98	14.26	14.68	14.76
CRSP20	**6.4**	**4.32**	**4.36**	**4.89**	**4.94**	**7.18**	13.5	14.08	14.68	14.33

Table 3.11. Out-of-sample portfolio risk metrics for 36 months long estimation window

This table shows the standard deviation, Sharpe Ratio, and Maximum Drawdown for the τ-quantile maximizing (qm), Equal Weighted (ew), Global Minimum Variance (gmv) and Global Minimum Variance using Ledoit-Wolf covariance matrix (gmv_lw) portfolios. A rolling window sample with a length of 36 months is used. Results are shown in percentage points and for datasets described in Section 3.5.1. Bold values indicate that the standard deviation of out-of-sample portfolio returns is significantly different from the standard deviation of Global Minimum Variance portfolios at the 5% significance level. Statistical significance is determined using an F-test for the equality of variances.

	ew	gmv_lw	gmv	qm0.05	qm0.1	qm0.3	qm0.5	qm0.7	qm0.9	qm0.95
Standard Deviation										
5IP	4.3	3.86	3.88	4.04	4.15	**4.39**	**4.86**	**5.69**	**6.2**	**6.14**
6P 2x3	**5.23**	4.37	4.39	4.69	4.83	4.85	**5.54**	**6.33**	**6.86**	**7.04**
6P ME INV	**5.14**	4.22	4.19	4.53	4.43	**4.8**	**5.55**	**6.14**	**6.64**	**6.55**
6P ME P10	**5.48**	4.39	4.37	4.64	4.79	**5.31**	**5.94**	**6.74**	**7.1**	**7.33**
6P ME P122	**5.49**	4.31	4.32	4.67	4.79	**5.05**	**5.52**	**6.74**	**7.26**	**7.38**
EUR 6P 2x3	5.41	4.88	4.81	**5.59**	5.15	5.13	**5.78**	**5.88**	**6.17**	**5.97**
JPN 6P 2x3	4.66	4.45	4.42	4.62	4.66	4.97	**5.35**	**5.31**	**5.55**	**6.03**
17IP	**4.92**	3.49	3.52	3.9	**4.07**	**4.56**	**7.48**	**8.36**	**9.24**	**9.54**
25P 5x5	**5.45**	4.29	4.24	4.69	**5.01**	**4.92**	**5.8**	**6.73**	**7.7**	**8.32**
25 ME P10	**5.68**	4.38	4.43	4.88	4.68	**5.13**	**5.95**	**7.44**	**7.44**	**8.33**
25 ME P122	**5.67**	4.27	4.23	4.67	**4.93**	**5.41**	**6.5**	**8.19**	**8.31**	**8.26**
30IP	**4.89**	3.58	3.63	**4.08**	**4.34**	**5.0**	**8.17**	**10.28**	**12.74**	**13.0**
48IP	**4.84**	3.5	3.48	**4.16**	**4.43**	**5.31**	**8.77**	**9.77**	**12.03**	**12.29**
CRSP5	**6.89**	5.43	4.99	5.2	5.27	**6.56**	**10.85**	**11.42**	**11.91**	**12.48**
CRSP10	**5.24**	3.84	3.83	**4.43**	**4.43**	**5.74**	**10.09**	**11.44**	**13.42**	**13.45**
CRSP20	**4.46**	3.38	3.42	3.84	**4.13**	**5.98**	**10.18**	**11.29**	**13.11**	**13.3**
Sharpe Ratio										
5IP	0.64	0.70	0.69	0.76	0.76	0.54	0.61	0.46	0.42	0.38
6P 2x3	0.59	0.67	0.69	0.59	0.60	0.59	0.58	0.36	0.32	0.30
6P ME INV	0.60	0.73	0.72	0.66	0.74	0.65	0.56	0.41	0.43	0.57
6P ME P10	0.53	0.70	0.71	0.65	0.57	0.64	0.48	0.45	0.47	0.49
6P ME P122	0.56	0.67	0.68	0.62	0.50	0.73	0.50	0.54	0.35	0.43
EUR 6P 2x3	0.43	0.62	0.68	0.49	0.62	0.64	0.45	0.29	0.24	0.17
JPN 6P 2x3	0.35	0.43	0.41	0.40	0.47	0.46	0.22	0.34	0.20	0.23
17IP	0.63	0.66	0.64	0.55	0.55	0.69	0.52	0.33	0.34	0.26
25P 5x5	0.59	0.65	0.67	0.65	0.69	0.62	0.54	0.30	0.32	0.26
25 ME P10	0.55	0.72	0.70	0.54	0.53	0.66	0.42	0.36	0.36	0.33
25 ME P122	0.58	0.68	0.66	0.59	0.65	0.67	0.52	0.45	0.49	0.41
30IP	0.62	0.57	0.55	0.42	0.68	0.63	0.58	0.52	0.34	0.37
48IP	0.65	0.66	0.69	0.71	0.62	0.49	0.38	0.45	0.38	0.47
CRSP5	0.69	0.66	0.62	0.79	0.89	0.48	0.68	0.50	0.52	0.43
CRSP10	0.71	0.70	0.67	0.84	0.60	0.45	0.65	0.52	0.34	0.42
CRSP20	0.74	0.74	0.75	0.72	0.77	0.41	0.66	0.50	0.50	0.41
Maximum Drawdown										
5IP	-46.98	-39.34	-38.18	-37.96	-39.62	-43.64	-49.99	-71.55	-77.96	-77.96
6P 2x3	-54.31	-51.00	-47.80	-48.41	-53.50	-57.27	-60.08	-61.49	-65.10	-65.79
6P ME INV	-51.86	-49.25	-48.58	-51.46	-50.14	-49.27	-53.45	-60.97	-57.50	-57.46
6P ME P10	-56.42	-43.89	-41.98	-43.90	-47.18	-44.32	-58.41	-66.92	-67.75	-64.74
6P ME P122	-55.92	-50.30	-47.35	-50.66	-50.15	-45.09	-53.39	-59.70	-64.44	-63.23
EUR 6P 2x3	-60.14	-53.79	-51.32	-66.68	-54.88	-59.94	-63.64	-64.01	-66.42	-65.51
JPN 6P 2x3	-50.22	-49.54	-47.10	-53.46	-49.72	-54.46	-66.34	-59.11	-61.47	-68.48
17IP	-51.71	-37.47	-36.56	-46.15	-35.58	-36.75	-68.68	-74.85	-75.38	-80.04
25P 5x5	-54.30	-53.13	-52.56	-51.45	-54.61	-55.88	-50.72	-67.22	-75.86	-77.25
25 ME P10	-56.61	-45.00	-45.56	-47.93	-46.77	-51.57	-57.59	-57.58	-56.19	-68.14
25 ME P122	-56.76	-50.47	-48.94	-48.57	-47.45	-48.07	-57.45	-62.32	-63.77	-67.28
30IP	-53.09	-40.64	-42.41	-42.53	-41.85	-39.43	-63.46	-78.78	-94.23	-96.37
48IP	-52.60	-41.67	-40.49	-35.59	-42.31	-51.90	-63.07	-78.84	-79.20	-81.46
CRSP5	-60.26	-64.22	-61.06	-53.94	-39.91	-54.47	-74.50	-80.09	-67.36	-69.16
CRSP10	-46.66	-39.16	-37.18	-28.99	-39.01	-54.59	-76.15	-80.69	-91.74	-84.40
CRSP20	-42.53	-31.24	-29.05	-32.80	-35.55	-70.41	-74.70	-79.07	-91.74	-84.40

Table 3.12. Herfindahl-Hirschman index and turnover of portfolio weights for 63 days long estimation window

This table shows the normalized Herchchendahl-Hirschman index, mean turnover and median turnover (Δ (median)) of portfolio weights for the τ-quantile maximizing (qm), Equal Weighted (ew), Global Minimum Variance (gmv) and Global Minimum Variance using Ledoit-Wolf covariance matrix (gmv_lw) portfolios. A rolling window sample with daily frequency and a length of 63 days is used. Results are shown in percentage points and for datasets described in Section 3.5.1.

	ew	gmv_lw	gmv	qm0.05	qm0.1	qm0.3	qm0.5	qm0.7	qm0.9	qm0.95
Herfindahl-Hirschman Index										
5IP d	0.0	0.19	0.26	0.51	0.43	0.37	0.48	0.68	0.83	0.85
6P 2x3 d	0.0	0.27	0.32	0.68	0.62	0.47	0.50	0.67	0.80	0.86
6P ME INV d	0.0	0.24	0.28	0.68	0.60	0.47	0.50	0.69	0.78	0.85
6P ME P122 d	0.0	0.28	0.33	0.61	0.58	0.45	0.49	0.70	0.82	0.87
EUR 6P 2x3 d	0.0	0.12	0.16	0.64	0.55	0.50	0.49	0.65	0.81	0.86
JPN 6P 2x3 d	0.0	0.16	0.22	0.57	0.51	0.47	0.49	0.59	0.78	0.83
17IP d	0.0	0.19	0.24	0.38	0.32	0.27	0.37	0.63	0.85	0.86
25P 5x5 d	0.0	0.19	0.23	0.51	0.45	0.37	0.40	0.56	0.73	0.79
30IP d	0.0	0.12	0.16	0.27	0.24	0.22	0.36	0.68	0.90	0.93
CRSP5 d	0.0	0.11	0.20	0.28	0.27	0.30	0.44	0.78	0.92	0.95
CRSP10 d	0.0	0.09	0.17	0.23	0.22	0.24	0.37	0.74	0.90	0.90
CRSP20 d	0.0	0.08	0.15	0.20	0.19	0.19	0.35	0.69	0.88	0.89
Crypto d	0.0	0.32	0.56	0.49	0.45	0.30	0.28	0.48	0.69	0.79
Portfolio Turnover										
5IP d	0.0	0.01	0.01	0.04	0.04	0.04	0.04	0.04	0.03	0.03
6P 2x3 d	0.0	0.01	0.01	0.04	0.04	0.04	0.05	0.04	0.04	0.03
6P ME INV d	0.0	0.01	0.01	0.04	0.04	0.05	0.05	0.04	0.04	0.04
6P ME P122 d	0.0	0.01	0.02	0.04	0.04	0.05	0.05	0.04	0.03	0.03
EUR 6P 2x3 d	0.0	0.01	0.01	0.04	0.05	0.05	0.05	0.04	0.03	0.03
JPN 6P 2x3 d	0.0	0.01	0.02	0.05	0.05	0.05	0.05	0.05	0.04	0.04
17IP d	0.0	0.02	0.02	0.05	0.05	0.06	0.06	0.05	0.04	0.04
25P 5x5 d	0.0	0.02	0.02	0.06	0.06	0.07	0.07	0.05	0.05	0.05
30IP d	0.0	0.02	0.02	0.06	0.06	0.06	0.06	0.04	0.03	0.03
CRSP5 d	0.0	0.01	0.01	0.03	0.03	0.04	0.04	0.03	0.02	0.02
CRSP10 d	0.0	0.01	0.02	0.04	0.04	0.05	0.05	0.04	0.03	0.03
CRSP20 d	0.0	0.02	0.03	0.05	0.05	0.06	0.06	0.04	0.03	0.04
Crypto d	0.0	0.08	0.08	0.15	0.18	0.37	0.44	0.31	0.16	0.12
Portfolio Turnover (median)										
5IP d	0.0	0.00	0.00	0.0	0.0	0.00	0.00	0.0	0.0	0.0
6P 2x3 d	0.0	0.00	0.00	0.0	0.0	0.00	0.00	0.0	0.0	0.0
6P ME INV d	0.0	0.00	0.00	0.0	0.0	0.00	0.00	0.0	0.0	0.0
6P ME P122 d	0.0	0.00	0.00	0.0	0.0	0.00	0.00	0.0	0.0	0.0
EUR 6P 2x3 d	0.0	0.00	0.00	0.0	0.0	0.00	0.00	0.0	0.0	0.0
JPN 6P 2x3 d	0.0	0.00	0.00	0.0	0.0	0.00	0.00	0.0	0.0	0.0
17IP d	0.0	0.00	0.00	0.0	0.0	0.00	0.00	0.0	0.0	0.0
25P 5x5 d	0.0	0.00	0.00	0.0	0.0	0.00	0.00	0.0	0.0	0.0
30IP d	0.0	0.00	0.00	0.0	0.0	0.00	0.00	0.0	0.0	0.0
CRSP5 d	0.0	0.00	0.00	0.0	0.0	0.00	0.00	0.0	0.0	0.0
CRSP10 d	0.0	0.00	0.00	0.0	0.0	0.00	0.00	0.0	0.0	0.0
CRSP20 d	0.0	0.00	0.00	0.0	0.0	0.00	0.00	0.0	0.0	0.0
Crypto d	0.0	0.05	0.04	0.0	0.0	0.23	0.27	0.0	0.0	0.0

Table 3.13. Out-of-sample 10%, 50% and 90% quantiles of portfolio returns for 63 days-long estimation window

This table shows the 10%, 50% and 90% quantile of portfolio returns for the τ-quantile maximizing (qm), Equal Weighted (ew), Global Minimum Variance (gmv) and Global Minimum Variance using Ledoit-Wolf covariance matrix (gmv_lw) portfolios. A rolling window sample with daily frequency and a length of 63 days is used. Results are shown in percentage points and for datasets described in Section 3.5.1. Bold values indicate that the out-of-sample quantiles of portfolio returns are significantly different from the quantile maximization strategy targeting the given τ at the 5% significance level. Statistical significance is determined using a percentile bootstrap method described in Section 3.A.

	ew	gmv_lw	gmv	qm0.05	qm0.1	qm0.3	qm0.5	qm0.7	qm0.9	qm0.95
					10% quantile					
5IP d	**-1.23**	-1.08	-1.08	-1.09	-1.12	-1.18	**-1.35**	**-1.58**	**-1.73**	**-1.74**
6P 2x3 d	**-1.34**	-1.09	-1.08	-1.13	-1.1	**-1.18**	**-1.44**	**-1.67**	**-1.72**	**-1.71**
6P ME INV d	**-1.34**	-1.11	-1.11	-1.13	-1.15	-1.17	**-1.45**	**-1.67**	**-1.71**	**-1.71**
6P ME P122 d	**-1.34**	-1.09	-1.09	-1.1	-1.09	**-1.2**	**-1.48**	**-1.77**	**-1.81**	**-1.84**
EUR 6P 2x3 d	**-1.13**	-0.98	-0.98	-1.0	-0.98	-1.03	**-1.18**	**-1.4**	**-1.48**	**-1.5**
JPN 6P 2x3 d	-1.4	-1.31	-1.31	-1.34	-1.32	**-1.41**	**-1.46**	**-1.57**	**-1.66**	**-1.67**
17IP d	**-1.28**	**-0.92**	**-0.91**	-0.97	-0.99	**-1.17**	**-1.71**	**-2.17**	**-2.46**	**-2.39**
25P 5x5 d	**-1.33**	-0.98	-0.98	-1.01	-1.04	**-1.19**	**-1.57**	**-1.93**	**-2.07**	**-2.04**
30IP d	**-1.25**	**-0.9**	**-0.9**	-0.99	-1.01	**-1.21**	**-1.9**	**-2.78**	**-3.28**	**-3.3**
CRSP5 d	**-1.52**	-1.29	-1.29	-1.37	-1.36	**-1.49**	**-1.82**	**-2.7**	**-3.11**	**-3.17**
CRSP10 d	**-1.53**	**-1.18**	**-1.17**	-1.25	-1.29	**-1.47**	**-2.14**	**-3.65**	**-4.06**	**-4.1**
CRSP20 d	**-1.34**	**-1.04**	**-1.03**	-1.14	-1.13	**-1.43**	**-2.28**	**-3.61**	**-4.02**	**-4.07**
Crypto d	-5.14	-4.22	-4.03	-4.74	-4.41	-4.9	**-5.93**	**-6.66**	**-6.38**	**-6.7**
					50% quantile					
5IP d	0.08	0.08	0.08	0.08	0.08	0.09	0.09	0.09	0.08	0.08
6P 2x3 d	0.07	0.06	0.06	0.07	0.07	0.06	0.07	0.07	0.06	0.07
6P ME INV d	0.06	0.07	0.07	0.07	0.07	0.06	0.07	0.07	0.06	0.07
6P ME P122 d	0.07	**0.06**	**0.06**	**0.05**	0.06	0.07	0.09	0.08	0.07	0.07
EUR 6P 2x3 d	0.07	0.08	0.08	0.08	0.08	0.09	0.09	0.07	0.06	0.07
JPN 6P 2x3 d	0.04	0.05	0.05	0.05	0.06	0.05	0.06	0.04	0.03	0.02
17IP d	0.08	0.07	0.07	0.08	0.08	0.08	0.1	0.07	0.08	0.08
25P 5x5 d	0.08	0.07	0.07	0.06	0.07	0.09	0.09	0.08	0.08	0.07
30IP d	0.08	0.07	0.07	**0.06**	0.06	0.08	0.11	0.1	0.04	0.02
CRSP5 d	0.06	0.05	0.05	0.06	0.04	0.04	0.06	0.0	**0.0**	**0.0**
CRSP10 d	0.06	0.07	0.06	0.06	0.07	0.06	0.06	0.05	0.0	**0.0**
CRSP20 d	0.07	0.07	0.06	0.06	0.07	0.06	0.06	0.04	0.03	0.0
Crypto d	0.21	0.28	0.19	0.15	0.17	0.23	0.06	-0.03	-0.17	-0.14
					90% quantile					
5IP d	**1.27**	**1.12**	**1.12**	**1.15**	**1.14**	**1.21**	**1.34**	**1.57**	1.67	1.69
6P 2x3 d	**1.31**	**1.14**	**1.14**	**1.15**	**1.15**	**1.21**	**1.39**	1.6	1.66	1.66
6P ME INV d	**1.31**	**1.15**	**1.14**	**1.14**	**1.16**	**1.22**	**1.41**	1.6	1.67	1.67
6P ME P122 d	**1.32**	**1.11**	**1.1**	**1.1**	**1.1**	**1.22**	**1.45**	1.72	1.8	1.81
EUR 6P 2x3 d	**1.16**	**1.04**	**1.04**	**1.06**	**1.04**	**1.11**	**1.2**	1.43	1.49	1.5
JPN 6P 2x3 d	**1.42**	**1.33**	**1.32**	**1.35**	**1.36**	**1.4**	**1.5**	1.64	1.71	1.68
17IP d	**1.31**	**0.94**	**0.92**	**1.02**	**1.04**	**1.21**	**1.78**	**2.26**	2.51	2.48
25P 5x5 d	**1.31**	**1.01**	**1.01**	**1.03**	**1.05**	**1.22**	**1.56**	**1.86**	2.0	2.01
30IP d	**1.29**	**0.94**	**0.94**	**1.04**	**1.06**	**1.26**	**2.06**	**2.94**	3.42	3.46
CRSP5 d	**1.56**	**1.36**	**1.35**	**1.42**	**1.42**	**1.52**	**1.8**	**2.77**	3.2	3.19
CRSP10 d	**1.57**	**1.23**	**1.22**	**1.31**	**1.34**	**1.47**	**2.24**	**3.73**	4.18	4.13
CRSP20 d	**1.42**	**1.07**	**1.06**	**1.18**	**1.21**	**1.44**	**2.3**	**3.85**	4.26	4.17
Crypto d	**5.42**	**4.88**	**4.72**	**4.98**	**4.79**	**5.54**	6.29	6.82	7.32	7.49

Table 3.14. Out-of-sample portfolio risk metrics for 63 days long estimation window

This table shows standard deviation, Sharpe Ratio, and Maximum Drawdown for the τ-quantile maximizing (qm), Equal Weighted (ew), Global Minimum Variance (gmv) and Global Minimum Variance using Ledoit-Wolf covariance matrix (gmv_lw) portfolios. A rolling window sample with daily frequency and a length of 63 days is used. Results are shown in percentage points and for datasets described in Section 3.5.1. Bold values indicate that the standard deviation of out-of-sample portfolio returns is significantly different from the standard deviation of Global Minimum Variance portfolios at the 5% significance level. Statistical significance is obtained using an F-test for the equality of variances.

	ew	gmv_lw	gmv	qm0.05	qm0.1	qm0.3	qm0.5	qm0.7	qm0.9	qm0.95
Standard Deviation										
5IP d	**1.17**	1.03	1.02	1.05	1.04	**1.1**	**1.26**	**1.49**	**1.62**	**1.62**
6P 2x3 d	**1.24**	1.09	1.08	1.11	1.09	**1.13**	**1.31**	**1.51**	**1.56**	**1.57**
6P ME INV d	**1.24**	1.1	1.1	1.12	1.1	**1.14**	**1.29**	**1.45**	**1.49**	**1.5**
6P ME P122 d	**1.26**	1.07	1.06	1.08	1.08	**1.18**	**1.42**	**1.6**	**1.68**	**1.69**
EUR 6P 2x3 d	**1.07**	0.96	0.95	0.97	0.96	**1.01**	**1.13**	**1.33**	**1.4**	**1.43**
JPN 6P 2x3 d	**1.25**	1.18	1.18	**1.22**	1.2	**1.28**	**1.34**	**1.42**	**1.48**	**1.49**
17IP d	**1.25**	0.9	0.89	**0.96**	**0.98**	**1.14**	**1.61**	**2.01**	**2.19**	**2.19**
25P 5x5 d	**1.26**	1.01	1.01	1.02	**1.04**	**1.18**	**1.5**	**1.79**	**1.88**	**1.87**
30IP d	**1.23**	0.89	0.89	**0.96**	**0.97**	**1.16**	**1.83**	**2.69**	**3.03**	**3.03**
CRSP5 d	**1.49**	1.23	1.22	**1.3**	**1.3**	**1.41**	**1.89**	**3.02**	**3.36**	**3.41**
CRSP10 d	**1.47**	1.1	1.09	**1.17**	**1.2**	**1.32**	**2.24**	**3.71**	**4.22**	**4.23**
CRSP20 d	**1.32**	0.98	0.97	**1.06**	**1.08**	**1.29**	**2.26**	**3.63**	**4.16**	**4.18**
Crypto d	**4.92**	4.34	4.29	**4.6**	4.37	**4.87**	**5.76**	**6.24**	**7.04**	**7.2**
Sharpe Ratio										
5IP d	0.13	0.14	0.14	0.15	0.15	0.14	0.13	0.11	0.09	0.10
6P 2x3 d	0.11	0.14	0.14	0.14	0.15	0.15	0.10	0.06	0.07	0.08
6P ME INV d	0.12	0.14	0.14	0.15	0.15	0.15	0.10	0.09	0.08	0.09
6P ME P122 d	0.12	0.11	0.11	0.10	0.13	0.13	0.13	0.10	0.11	0.10
EUR 6P 2x3 d	0.11	0.13	0.13	0.15	0.15	0.16	0.12	0.08	0.08	0.08
JPN 6P 2x3 d	0.06	0.07	0.08	0.06	0.08	0.08	0.08	0.05	0.04	0.03
17IP d	0.13	0.14	0.14	0.14	0.14	0.15	0.12	0.09	0.09	0.09
25P 5x5 d	0.13	0.14	0.14	0.15	0.16	0.16	0.13	0.09	0.09	0.10
30IP d	0.13	0.14	0.15	0.13	0.14	0.15	0.16	0.12	0.09	0.08
CRSP5 d	0.10	0.12	0.12	0.10	0.12	0.07	0.05	0.06	0.04	0.03
CRSP10 d	0.13	0.13	0.13	0.10	0.13	0.09	0.10	0.07	0.07	0.05
CRSP20 d	0.14	0.15	0.15	0.12	0.17	0.10	0.10	0.11	0.11	0.07
Crypto d	0.14	0.18	0.21	0.16	0.13	0.15	0.12	0.05	0.11	0.12
Maximum Drawdown										
5IP d	-51.17	-40.02	-38.91	-42.76	-35.25	-41.85	-54.78	-66.81	-84.61	-82.43
6P 2x3 d	-58.47	-48.73	-49.17	-52.90	-52.67	-52.12	-63.39	-81.76	-78.73	-77.14
6P ME INV d	-57.49	-55.12	-54.21	-56.00	-58.16	-54.37	-55.49	-69.58	-70.92	-73.83
6P ME P122 d	-59.16	-54.07	-53.99	-53.93	-55.58	-52.23	-62.00	-72.87	-67.43	-73.99
EUR 6P 2x3 d	-63.02	-64.37	-64.82	-64.87	-64.24	-62.68	-59.15	-64.14	-66.88	-71.52
JPN 6P 2x3 d	-54.46	-57.64	-57.52	-57.23	-58.04	-53.05	-58.18	-70.91	-75.21	-74.55
17IP d	-56.05	-41.11	-39.95	-43.20	-40.35	-42.05	-60.51	-71.78	-76.41	-73.54
25P 5x5 d	-59.08	-53.33	-53.29	-57.32	-58.32	-52.86	-63.33	-84.00	-82.19	-79.44
30IP d	-57.06	-41.38	-40.63	-43.39	-40.68	-48.32	-62.75	-85.95	-97.73	-97.01
CRSP5 d	-53.39	-48.04	-50.36	-54.98	-50.35	-56.33	-71.69	-85.65	-96.54	-95.93
CRSP10 d	-62.43	-41.58	-41.15	-40.88	-38.61	-46.35	-66.32	-86.55	-97.44	-98.93
CRSP20 d	-60.09	-49.81	-50.15	-54.16	-29.89	-52.45	-67.20	-87.11	-86.09	-95.09
Crypto d	-92.29	-87.16	-81.48	-81.91	-89.19	-90.11	-95.99	-98.68	-98.55	-97.82

Table 3.15. Binomial tests of outperformance for 60 months long estimation window

This table shows the number of datasets where the benchmark Global Minimum Variance (gmv) strategy outperforms τ-quantile maximizing (qm) strategy, the Global Minimum Variance using Ledoit-Wolf covariance matrix (gmv_lw) and Equal Weighted (ew) strategy. Outperformance is defined as a higher τ-quantile of out-of-sample portfolio returns achieved compared to the non-benchmark strategy and as a lower standard deviation of out-of-sample portfolio returns compared to the benchmark strategy. The null hypothesis that the probability of benchmark outperformance is equal to 0.5 is tested against the alternative hypothesis that the probability of benchmark outperformance is greater than 0.5 using a one-sided binomial test. Values in parentheses are p-values. Results in this table summarize outperformances reported in Table 3.5 for 60-month-long estimation windows. The number of total datasets is 16.

	gmv_lw	ew	qm0.1	qm0.3
0.1	13	14	12	13
	(0.011)	(0.002)	(0.038)	(0.011)
std	7	16	16	16
	(0.773)	(0.0)	(0.0)	(0.0)

Table 3.16. Binomial tests of outperformance for 36 months long estimation window

This table shows the number of datasets where the benchmark Global Minimum Variance (gmv) strategy outperforms τ-quantile maximizing (qm) strategy, the Global Minimum Variance using Ledoit-Wolf covariance matrix (gmv_lw) and Equal Weighted (ew) strategy. Outperformance is defined as a higher τ-quantile of out-of-sample portfolio returns achieved compared to the non-benchmark strategy and as a lower standard deviation of out-of-sample portfolio returns compared to the benchmark strategy. The null hypothesis that the probability of benchmark outperformance is equal to 0.5 is tested against the alternative hypothesis that the probability of benchmark outperformance is greater than 0.5 using a one-sided binomial test. Values in parentheses are p-values. Results in this table summarize outperformances reported in Table 3.10 for 36-month-long estimation windows. The number of total datasets is 16.

	gmv_lw	ew	qm0.1	qm0.3
0.1	7	16	15	16
	(0.773)	(0.0)	(0.0)	(0.0)
std	9	16	16	16
	(0.402)	(0.0)	(0.0)	(0.0)

Table 3.17. Binomial tests of outperformance for 63 days long estimation window

This table shows the number of datasets where the benchmark Global Minimum Variance (gmv) strategy outperforms τ-quantile maximizing (qm) strategy, the Global Minimum Variance using Ledoit-Wolf covariance matrix (gmv_lw) and Equal Weighted (ew) strategy. Outperformance is defined as a higher τ-quantile of out-of-sample portfolio returns achieved compared to the non-benchmark strategy and as a lower standard deviation of out-of-sample portfolio returns compared to the benchmark strategy. The null hypothesis that the probability of benchmark outperformance is equal to 0.5 is tested against the alternative hypothesis that the probability of benchmark outperformance is greater than 0.5 using a one-sided binomial test. Values in parentheses are p-values. Results in this table summarize outperformances reported in Table 3.13 for 63 days long estimation windows. The number of total datasets is 13.

	gmv_lw	ew	qm0.1	qm0.3
0.1	13	13	13	13
	(0.0)	(0.0)	(0.0)	(0.0)
std	13	13	13	13
	(0.0)	(0.0)	(0.0)	(0.0)

Chapter 4

Conclusion

This thesis explores novel approaches to stock returns forecasting, empirical asset pricing, and portfolio selection, emphasizing the integration of machine learning techniques and quantitative methods to enhance predictive accuracy and investment decision-making.

Chapter 1 provides a comprehensive analysis of out-of-sample returns on 153 equity anomalies, establishing that machine learning-based aggregation of these anomalies into a single mispricing signal offers consistent profitability on a global scale. The findings underscore that U.S.-based evidence plays a crucial role in predicting international returns, while past performance in non-U.S. regions falls short in forecasting success within the U.S. These insights highlight the importance of region-specific evidence in quantitative strategy selection and its implications for global investment frameworks.

Chapter 2 introduces an advanced method for predicting the full distribution of stock returns, utilizing a two-stage quantile neural network combined with spline interpolation. This approach transcends traditional model limitations by accommodating non-Gaussian, heavy-tailed data and uncovering complex, non-linear interactions among stock characteristics and market variables. The superior performance of this method, in terms of out-of-sample losses, offers investors robust empirical tools for mean estimation and forecasting, applicable across diverse datasets from both U.S. and international markets.

Chapter 3 delves into the portfolio selection process for investors with quantile preferences, revealing substantial variation in portfolio diversification and turnover based on different τ-quantiles. The analysis shows that risk-averse investors may achieve higher out-of-sample τ-quantiles of returns by constructing global minimum variance portfolios rather than by directly maximizing quantiles

based on sample data. Monte Carlo simulations attribute the underperformance of quantile maximization strategies to small sample bias, which disrupts the alignment of portfolio weights with the theoretically optimal configuration.

Overall, the application of machine learning and quantile-specific analyses reveals previously unexplored dimensions of risk and return, providing valuable insights for both academic research and practical investment strategies. Future research could extend this work, especially the third chapter, by methodological improvement of the quantile maximization strategies to achieve better out-of-sample performance in line with the investor's preferences.

Chapter 5

Responses to opponents' reports

I appreciate the time and effort of all three opponents in reviewing my thesis. Their feedback has been instrumental in improving this work. Below, I provide my responses to their comments and suggestions. The comments of the opponents are in *italics*.

5.1 Dr. Maria Grith PhD

5.1.1 Chapter 2

"This chapter is well constructed, with the literature, methodology, and implementation well motivated and explained. It demonstrates a solid foundation of theoretical concepts and programming, and I believe it has a strong chance of acceptance by the International Journal of Forecasting. Below, I highlight a few points that may deserve attention."

1. *"Multiple Quantile Forecasting.*
   I suggest providing a more detailed treatment of the multi-quantile loss function in Equation 3*. Beyond computational efficiency, what are the key differences between solving a multi-quantile loss function and multiple single-quantile loss functions without any penalization? Classical references for multi-quantile regression include Li Li and Zhu (2008) and Koenker (2004). Extending their framework to a nonparametric setting presents particular challenges that should be articulated clearly, and a thorough discussion of the proposed implementation is essential."*
   Response: Thank you for this suggestion. I have extended the discussion

on multi-quantile loss minimization in Subsection 2.2.2 and provided a more detailed treatment of the multi-quantile loss function in Equation 2.3.

2. *"Quantile crossing.*
   The author of the thesis claims that the monotonicity requirement of quantile predictions is not violated in their application, even without explicit penalization for crossing. To broaden the generalizability of the results, it would be beneficial to outline potential approaches for cases where the monotonicity requirement is not met in practice. In related work, Moon et al. (2021) have estimated multiple conditional quantiles using neural networks while maintaining the non-crossing property. It would be insightful to compare the current implementation with their findings."
   Response: I have added a discussion on quantile crossing in Subsection 2.2.2, highlighting the advantages of multi-quantile optimization, such as implicit regularization, feature sharing, and gradient averaging, as well as its limitations compared to single-quantile training.

   Thank you for pointing me to Moon et al. (2021); I was not aware of this paper, and I find their approach elegant. While a direct comparison would be interesting, a full-scale implementation and systematic evaluation of their method relative to ours would be best suited for a separate study. I have included their reference and a brief discussion in Appendix 2.E.5.

3. *"Multi-output Feed-forward Neural Network Model. The network setup is interesting, and it is evident that a lot of thought went into it. There are a few points that I would like to mention:*

   - *An interesting name abbreviation for the method/algorithm.*
   - *The challenge lies in visualizing the entire system intuitively. Two potential improvements can be suggested:*
     (a) *A pseudo-algorithm summarizing the steps on page 78 could be created.*
     (b) *Figure 3.2 should be refined to include general notations instead of numbers, incorporating the objective function and standardization. It is also essential to clarify the necessity of the second network.*
   - *On page 81, there are many hyperparameters. Are all of them necessary? An essential discussion regarding performance on NN concerns the dropouts: Do sparse networks behave better than non-sparse ones in applications?*

- *Similarly to the point above, is the bottleneck layer necessary? Are similar results obtained with less intricate choices?*
- *On a practical level, when trained to minimize a multi-output loss function, many neural networks behave as if the networks are being trained independently. Is this observation applicable in the current context?"*

Response: Thank you for these comments. I have addressed each point below.

- **Abbreviation.**: I did not come up with an interesting name, but I have started using the acronym 2sNN across the tables.
- **A pseudo-algorithm.**: I have added a pseudo-algorithm summarizing the method in Appendix 2.E.4 and referenced it in the main text to improve clarity.
- **Figure improvement**: I have included a more general notation in Figure 2.2.
- **Clarification of the necessity of the second network.** The market volatility sub-network (the second network) is incorporated to provide time-series market-wide volatility as an additional input for quantile predictions. Beyond stock-level volatilities, it captures broader market conditions, such as periods of high or low volatility. Table 2.17 presents the results of the architecture hyperparameter search, indicating that the second network adds value. While not strictly necessary—since the NNh2 model performs reasonably well without it—it improves overall model performance. A brief discussion on this has been added to Appendix 2.D.
- **Necessity of all hyperparameters**: The hyperparameters for the two-stage quantile neural network (2sNN) were selected based on empirical validation and prior findings in financial deep learning (Gu et al. 2020; Tobek and Hronec 2021). Given the impracticality of searching all possible hyperparameter values, a structured two-step search process was implemented, as described in Section 2.D.1.
  - Learning rate & dropout rate: Selected based on validation loss (see Table 2.16). The final choice of 0.0003 for the learning rate and 0.2 for the dropout rate provided the best balance between

stability and generalization. The choice of non-zero dropout rates is consistent with the model specification in previous studies on deep learning in asset pricing (Gu et al. 2020). While some degree of sparsity enhances robustness, excessive regularization can degrade model performance, particularly in small-sample settings.

- Neurons & layers: The architecture search (see Table 2.17) indicated that models with fewer than two hidden layers in Stage I or fewer than eight neurons in Stage II underperformed. Increasing beyond 256 neurons did not yield additional improvements, suggesting diminishing returns.
- L1/L2 regularization: The selected penalties (0.0001 for Stage I, 0.00001 for Stage II) were chosen to prevent overfitting while maintaining interpretability. The hyperparameter search showed that further increases in L1/L2 penalties unnecessarily restricted model capacity.
- Ensemble size - The choice of 20 models for the liquid sample and 10 models for the full sample balances computational efficiency with variance reduction.

The necessity of each hyperparameter was evaluated through validation loss minimization, ensuring that only those parameters contributing to improved generalization and stability were retained.

- **Bottleneck layer justification.** The inclusion of a bottleneck layer in the two-stage quantile network (2sNN) is motivated by the autoencoder architecture, where it provides additional regularization by forcing the network to learn a compressed representation of the input data. Alternative architectures tested during the hyperparameter search are shown in Table 2.17, along with their corresponding validation losses, which support the choice of the bottleneck layer for 2sNN. This decision is further validated by the results in Table 2.4 and other tables assessing predictability, where the 2sNN model outperforms alternative architectures over the out-of-sample period. I have added a brief discussion on this to Section 2.3.
- **Single vs. multi-output loss function.**: Thank you for this comment. For a general discussion on this topic and specific changes

in the main text, please refer to my response to the comment on Multi-Quantile Loss Minimization above.

I have also included an out-of-sample quantile loss comparison between single-output and multi-output 2sNN models in Table 5.1 below. The comparison considers selected τ-quantiles (20%, 50%, 80%) using the full sample and the U.S. region from 1995 to 2018. The results show that while the multi-output 2sNN model performs slightly better than its single-output counterpart, the differences are not statistically significant for $\tau = 0.5$ and $\tau = 0.8$. Overall, this suggests that the multi-output model behaves differently from the single-output alternative, but the performance improvements are not statistically significant at the median and upper quantiles.

Table 5.1. Comparison of single-output vs. multi-output 2sNN models
This table presents the average out-of-sample quantile loss for 20%, 50%, and 8% quantile forecasts, comparing the multi-output two-stage model (2sNN) with its single-output alternative (2sNN-single). Both models share the same architecture, but the single-output version predicts one quantile at a time using a standardized τ-quantile layer. The results are shown for the out-of-sample period from 1995 to 2018 for the U.S. full sample only. Loss values are scaled by a factor of 100 for readability. The t-statistics (in parentheses) are computed for the loss difference between the multi-output and the single-output models using Newey-West standard errors with 12 lags.

	0.2	0.5	0.8
2sNN single	3.36	4.68	3.82
2sNN	3.24	4.66	3.79
(t-stat)	(-1.94)	(-1.06)	(-0.61)

4. *"Literature. I was intrigued to learn how the results in this chapter contribute to the debate on "previously observed conflicting results on the roles of variance, skewness, and kurtosis in asset pricing." However, I could not find a significant discussion in the related literature. It would be insightful to illustrate which studies find similar results and how the current approach sheds light on the ability of moments to forecast returns."*
Response: Thank you for this suggestion. I have expanded the literature review in Section 2.1.1 to provide a more detailed discussion of empirical evidence on the pricing power of higher moments. Since our study finds that higher moments are not priced, the comparison with previous literature is incorporated in the literature review section, which outlines studies that both support and contradict the pricing power of higher moments.

While direct study-to-study comparisons, including replications, would be the most effective way to resolve conflicting findings, differences in methodologies, datasets, and sample periods make this challenging and beyond the scope of this paper. Nonetheless, our study contributes to this debate by providing robust out-of-sample evidence across different regions and datasets, offering further insights into the role of higher moments in asset pricing.

5.1.2 Chapter 3

"The topic discussed in this chapter is understudied, and the author offers significant empirical contributions that could guide future research in this area. This section is less polished and refined regarding ideas than the other chapters. Nevertheless, the implementation and explanations are transparent. Below, I summarize a few points that may support the work for potential publication."

Response: Thank you for recognizing the gap in the literature and the empirical contributions of this chapter. I appreciate your constructive feedback, which will help refine the ideas and improve the clarity of the presentation.

1. *"τ-Quantile Maximization and Global Minimum Variance Portfolio It is essential to highlight the connection between these two problems through some straightforward parametric examples, demonstrating a theoretical foundation for any parallels that may exist."*

   Response: Thank you for this suggestion—I agree that it enhances the clarity of the chapter. I have added a parametric example in Section 3.3 to illustrate the connection between τ-quantile maximization, the global minimum variance portfolio, and mean-variance optimization under different scenarios.

2. *"Small Sample Bias. The authors note that small sample bias results in inaccuracies in the selected portfolio weights. A key question is how to address this bias. While this is a complex issue, some empirical insights have been gained. For example, a potential avenue for further research is to investigate whether regularizing the τ-quantile portfolio weights toward the global minimum variance portfolio weights improves estimation outcomes for agents with quantile preferences."*

   Response: I fully agree. We have explored several approaches to address this issue, including τ-specific regularization in the objective function,

maximizing quantiles of demeaned returns to reduce noise from historical mean estimates, and regularizing τ-quantile portfolio weights toward the global minimum variance portfolio weights. However, none of these attempts (not reported) proved successful when tested out-of-sample. This remains a very interesting avenue for future research.

5.2 Prof. Jan Hanousek Ph.D.

5.2.1 Chapter 1

"The first paper aims to show significant profitability gains in long-short portfolios formed using predictions derived from machine learning techniques. Compared to other existing studies, the author examines the usefulness of global data (outside of the US markets) in expanding possible gains in managed portfolios. I found the paper interesting, yet I have several questions and comments:"

1. *"Several other papers have presented the idea of analysing returns on anomalies outside the US. For example, the authors refer to Jacobs and Muller (2018, 2020) in the chapter. Authors referred to the main difference using a broader set of assets, including small-cap stocks, and a more global view. It would be worth more discussion, maybe extending the model by using weights stemming from links or dynamic correlation of the regional markets and the US to shed more light on differences in the results."*
   Response: Thank you for the suggestion. In this paper, we chose to rely on a list of published anomalies to ensure full transparency in variable selection and facilitate replicability. While the dynamic correlation between regional markets and the US could serve as a useful predictive signal, to the best of our knowledge, it is not a documented pricing anomaly. For this reason, we did not incorporate it into our analysis.

2. *"Similar comment/suggestion to comparing the existing results: Authors should try to disentangle the effect of small-cap vs. large liquid assets and regional vs. global anomalies. In other words, what would be the results if authors kept only liquid assets and used global anomalies, and what would be the effect if they used small-cap and regional anomalies? Maybe I misunderstood the described differences since the authors use the universe of the most liquid stocks later in the chapter. "*

Response: The first paper focuses exclusively on the most liquid stocks. Excluding small-cap stocks ensures results that are more relevant to investors, reduces the impact of microstructure noise, and presents a higher hurdle for both predictability and profitability. However, the second chapter—Predicting the distributions of stock returns around the globe in the era of big data and learning—includes analyses on both liquid and full samples.

Regarding global vs. regional anomalies, this paper considers only global anomalies, meaning those that can be calculated for stocks across all regions. This approach allows us to directly assess whether expanding the estimation sample from U.S. stocks to an international setting improves predictive regression-based investment strategies for U.S. investors.

3. *"It would be beneficial to use a diagram/scheme of the methods used, their order and their variations; the same would apply to the training/tuning setup. Let me note that comparing the out-of-the-sample predictions is difficult when the training sample grows in each step. Also, an exciting approach could be to accept the dynamic and changing features of financial markets and their anomalies and to use moving windows for the training, say the last five years."*

   Response: I have added Table 1.7, which summarizes all methods used along with their corresponding hyperparameters. Additionally, Figure 1.9 now illustrates the training, validation, and test splits.

   We use a validation approach based on a growing training sample, which aligns with established methods in the literature, such as Gu et al. (2020). Machine learning algorithms generally require a large amount of data to learn underlying patterns and to avoid overfitting. This setup fully adheres to the out-of-sample principle, as the model is trained only on data available at the time of prediction, ensuring valid out-of-sample comparisons. I appreciate the suggestion of considering moving windows, as it raises an interesting question about how far back historical evidence remains relevant. I find this an intriguing idea for future research.

4. *"Technical notes on fundamental data availability. It was unclear in the paper if the authors controlled the availability of the fundamental financials used as additional explanatory variables. I mean that the variables for the firm financials are available with a delay. The minimum delay is*

determined by the date of the quarterly announcements (referring to the previous quarter results)"

Response: We ensure that fundamental financial data is incorporated with an appropriate lag to account for reporting delays. Specifically, at each monthly observation point, we use only financial statement information from fiscal years ending at least six months prior to the current month. This implementation applies to all 93 fundamental anomalies. This procedure is explicitly stated on page 13: "We update the fundamental signals for each month with financial statement information from financial years ending at least six months prior." Further, in a separate paper not included in this thesis - Tobek and Hronec (2018), we also examine the effect of the choice of the source of fundamental data on the performance of the anomalies.

5. *"Overall, and this applies to all chapters, machine learning training/tuning and the issue of overfitting should be discussed a bit more."*

   Response: Our validation approach and out-of-sample design are detailed in Section 1.1.3 and strictly adhere to the no look-ahead principle. Additionally, in Chapter 2, we have incorporated new out-of-sample results for the period 2018–2023, demonstrating that our approach and considered models generalize well to unseen data, including turbulent periods such as the COVID-19 pandemic.

6. *"As concluding remarks, the authors said: "Most of the predictability of the expected stock returns in all the regions can be captured solely with the U.S. training sample as the benefits of a larger estimation sample is offset by the region-specific differences." This statement can be grounds for future research, analyzing which factors offset the benefits and what selection of certain regional anomalies can be helpful from the information and portfolio returns point of view."*

   Response: Thank you for the remark. I share the same view and agree that regional specificities remain underexplored in the literature.

7. *"Minor suggestions:*

   - *The tables presented should have a better legend. For example, in Panel D, I can only guess that the second line below each row is the t-statistics. Similarly, it could bring a better outcome by presenting formal tests of differences between methods and approaches, both*

means and coefficients."

Response: I have improved Table 1.2. It contains t-statistics for testing both if mean returns are significantly different from zero as well as testing if the difference in mean returns between each strategy and previously compared strategies is significantly different from zero.

- *"Some references are incomplete, e.g., Tobek, O., & Hronec, M. 2018. Does source of fundamental data matter?"*

  Response: I have corrected the mentioned reference and revised other incomplete references throughout the text.

5.2.2 Chapter 2

"This chapter proposed an alternative method for predicting/modeling the distribution of returns. Contrary to the classical approach, the authors used a forecasting method combining a two-stage quantile neural network with cubic B-splines interpolation. Their approach has the advantage of predicting the other distributional characteristics, such as skewness and kurtosis, or even estimating the entire distribution of the returns. Some of my comments and suggestions are common to the entire thesis, so let me mention only those specifically appearing in this chapter and not already mentioned above:"

1. *"My concern about training setup and possible overfitting discussion is still valid here. Moreover, the long period for training and estimation requires an assumption of distribution stability over a more extended period. Which is not generally accepted. There is, for example, observed and accepted increased investor demand for lottery- like returns, Kumar (2009) or Blau et al. (2016), to name a few articles about changes in investor preferences, suggesting instability of the studied distribution over a long period."*

   Response: I acknowledge the concerns regarding overfitting and the assumption of distributional stability. Below, I address these issues in detail.

   Our model is retrained annually using an *expanding window approach*, incorporating all available data up to that point. This ensures that while historical data informs model training, the most recent information is always included, allowing the model to adapt to evolving market conditions. Importantly, the feature set (information set) always remains up to date

— older data stays in the training set, but new feature values continuously shape the model's conditional quantile estimates, reflecting changes in market conditions, investor sentiment, and economic shifts.

To mitigate overfitting risks, the model incorporates multiple *regularization mechanisms*, including *L1* and *L2* penalties, dropout layers, and batch normalization. Additionally, early stopping based on validation loss ensures training ceases when generalization performance deteriorates.

A natural alternative to the expanding window approach would be a *rolling window* approach, where only a fixed-length window of past data is retained. However, this has several drawbacks in asset pricing applications. A rolling window approach discards valuable historical data, which can be problematic given that many asset pricing anomalies exhibit long-term persistence (Hou et al. 2020). Retaining all data via an expanding window allows the model to learn both persistent and time-varying relationships. Further, a rolling window places excessive weight on short-term trends and may lead to excessive model instability. Given the noisiness of financial returns, a short rolling window may force the model to chase short-lived patterns rather than learn robust long-term relationships. Expanding windows are widely used in financial machine learning. For example, Gu et al. (2020) apply expanding windows to return prediction models, leveraging the full history of available data while allowing feature values to dynamically update. A potential middle ground could involve information weighting schemes, where older observations are discounted instead of being completely removed. This would maintain flexibility while preventing overreliance on outdated information, though such an approach requires a calibration.

With regard to the assumption of distribution stability, our model does not assume distributional stationarity. Instead, it learns time-varying relationships between predictors and conditional quantiles, allowing it to adapt to changes in market dynamics. Moreover, the two-stage quantile network architecture explicitly accounts for market-wide shifts in risk preferences, particularly through the second-stage volatility adjustment.

Finally, to further address concerns about overfitting, we have added a post-original sample, i.e. true out-of-sample results to Section 2.E.6 for the period of 2018-2023 for the U.S. region. These results show the

generalizability of our model to unseen data, thus hopefully addressing the concern of overfitting.

2. *"When authors present their estimates of distributional characteristics, one can ask what their error is. For example, are the numbers for the entire sample and subsamples statistically different?"*
Response: A key limitation of using neural networks for distributional forecasting is the lack of well-established methods for constructing confidence intervals or estimation errors. Unlike parametric models, where confidence intervals can be derived using asymptotic properties or bootstrapping, neural networks lack an analytical likelihood function and rely on highly non-linear optimization processes, making traditional error estimation techniques inapplicable.

   As far as I know, several techniques have been proposed to estimate uncertainty in neural networks, but each has its limitations. Bootstrap Aggregation (Bagging) is computationally expensive and does not directly provide individual confidence intervals (Efron and Tibshirani 1994), Bayesian Neural Networks (BNNs) are theoretically appealing but suffer from scalability issues in large-scale settings (Blundell et al. 2015), Monte Carlo Dropout (MC-Dropout) provides empirical distributions over forecasts but is an approximation of Bayesian inference (Gal and Ghahramani 2016), and Deep Ensembles are one of the most practical methods for uncertainty estimation, yet lack formal hypothesis testing interpretations (Lakshminarayanan et al. 2017). Overall, these methods offer potential avenues for future returns.

   In terms of subsample analysis, we explicitly report results for different geographic regions (US, Europe, Japan, Asia-Pacific) and liquidity subsamples (full sample vs. liquid sample). This allows us to assess the robustness of our approach across diverse datasets. Further, we have added Section 2.E.6 with previously not-used data from 2018-2023 for the U.S. region, which provides a sub-sample analysis.

3. *"As before, I would love to see the comparisons, maybe using confidence intervals across subsamples and parameters or a table with a composite summary of rejected hypotheses of equality. It also carries our interesting interpretations."*
Response: We conduct a number of formal statistical tests to evaluate

differences in forecast quality, such as Newey-West adjusted t-statistics for quantile loss differences between models (Table 2.4) and differences in Diebold-Mariano test statistics for differences in predictive accuracy (Table 2.6).

We, however, only test differences between models, not between subsamples as we are interested in whether the performances of individual models differ and not necessarily if the performance of the same model differs across subsamples as it is expected, e.g. different performance across regions and, crucially on liquid and full sample. We report test statistics in corresponding tables.

For example, we test for significant differences in quantile forecasting accuracy across models using Newey-West adjusted t-statistics (Table 2.4), and we formally compare volatility forecast errors across models using Diebold-Mariano test statistics (Table 2.6). Tests we perform and test statistics we obtain are always described in the table captions.

Subsample results for the period 2018-2023 and the U.S. region are also presented in Section 2.E.6.

4. *"Minor comments:*
Repeating words or forgotten text. Example: page 76, middle of the page: 'A large body of literatur distribution of stock returns. of stock returns.' "
Response: Thank you for pointing out this error. I have corrected it and other grammatical inconsistencies throughout the thesis as well.

5.2.3 Chapter 3

" The fourth chapter builds on the results and techniques introduced in the previous parts. In particular, this section focuses on the out-of-sample performance of portfolios chosen by investors, maximizing the τ-quantile of the future returns. Additional comments to this chapter:"

1. *"The authors find that "A portfolio often yields higher out-of-sample τ-quantile returns compared to approaches to direct quantile maximization." Using a formal test (based on the multinomial distribution associated with the Monte Carlo simulation), authors should be specific if "often" means a statistically significant finding."*
Response: Thank you for the suggestion. I have added results of the

binomial tests testing if the fraction of times the benchmark Global Minimum Variance portfolio outperforms the τ-quantile maximizing portfolio in case of $\tau \in \{0.1, 0.3\}$ is statistically significantly different from 0.5. Results are shown in Appendix 3.E.

2. *"The technical setup and interesting results would deserve a deeper statistical analysis of the Monte Carlo study's results, like various tests of (linear) trends, increasing/decreasing patterns, differences, etc., since they provide attractive, non- technical, and practical results."*
Response: I appreciate the suggestion for a deeper statistical analysis of the Monte Carlo study's results. While such an analysis could provide additional insights, I have chosen to focus on the primary objective of the Monte Carlo simulations—demonstrating that the empirical results of the paper are not unexpected and align with the theoretical expectations from a simple simulation setup. Including more detailed statistical tests might shift attention away from the main idea and introduce complexity that is not essential for conveying the core message. The underlying reason for the observed outperformance is traced to the small-sample bias in estimating the τ-quantile maximizing portfolio weights, which is discussed in detail in Section 3.4.

5.3 Dr Rodrigo Hizmeri PhD

5.3.1 Introduction

"The introduction (labelled Chapter 1) describes accurately the content of the subsequent chapters. However, I found it a bit confusing when we refer to the first chapter as being actually the second chapter. For instance, Martin writes '[...] In the first chapter, we demonstrate significant profitability gains [...]'. I think it would be easier if the introductory chapter is unnumbered or, alternatively, we refer to papers as Chapter 2, 3, and 4."
Response: I have renumbered the chapters to avoid confusion. The introductory chapter is now unnumbered, and the subsequent chapters are labelled 1, 2, and 3.

5.3.2 Chapter 1

"As this chapter is published, I will only provide minor comments that relate to clarifying questions or typos, with the aim of improving the exposition of the thesis: "

1. *"On page 15, you write: 'The characteristics are normalized to their cross-sectional quantiles within each region to reduce problems with outliers'. What do you exactly mean by normalized to their cross-section quantiles? Do you winsorize the characteristics each month at the 1% level or at the IQR level? The latter will have larger consequences on the data then the former, and additionally it is far less common."*
   Response: Given a specific characteristic, each month in each region all stocks are ranked based on this characteristic. The stock with the highest value is assigned value of 1, all the other stocks are assigned values between 0 and 1 based on their rank. Below is the code implementation of this procedure in Python:
   ```
   data[signal] = data.groupby(["date", "region"])[signal].rank(pct=True)
   ```

2. *" On page 23, you mistyped the word "use". It should read: 'The maximum number of variables we use is 153, [...].' "*
   Response: Fixed.

3. *"At the end of the same paragraph Page*23*, it should read: Our results, which are obtained in different empirical settings provide support for the findings of Gu et al.* 2020*."*
   Response: Fixed.

4. *"Amihud's citation should have the journal's name in capital letters. Ditto for: Asness et al.* 2013*, Banz* 1981*, Barth and Hutton* 1984*, Basu* 1977*, Bhandari* 1988*, among others. "*
   Response: Fixed.

5.3.3 Chapter 2

"The paper considers a two-stage quantile neural network approach to forecast the distribution of returns. The authors retrieve the moments using numerical integration from a smooth distribution and show that only the central moments are priced in the cross-section of stock returns. The paper motivates the use of quantile preferences due to agents having a more complex set of preferences

that require information from higher-order moments. While I completely agree with this assertion, it is a bit counterintuitive to see that lower quantiles, which should be related to risk-aversion are not compensated in the cross-section."

Response: Thank you for this comment. We do not necessarily motivate the use of quantile preferences[1]in this paper, on the contrary, we motivate the usefulness and importance of distributional forecasts for agents with complex preferences[2], where quantile preferences are only one of the examples.

Regarding the observation that lower quantiles are not compensated in the cross-section of stock returns, I acknowledge that this result may appear counterintuitive given the expectation that risk-averse agents would demand compensation for exposure to left-tail risk. However, this finding aligns with some existing empirical evidence in the literature, which has shown conflicting or weak pricing effects for tail risk and higher moments.

To address this, I have expanded the literature review in Section 2.1.1 to include a more detailed discussion of prior studies on the pricing of tail risk and higher moments, such as those by Ang et al. (2006), Bali, Engle, and Murray (2016), Bollerslev et al. (2020), and Langlois (2020). These studies suggest that the pricing of left-tail risk may be sample-specific or diminished in certain periods, which could explain our results over the 1995-2018 period.

1. *"The chapter would benefit from a more comprehensive discussion regarding the economic intuition behind finding that central τ-quantiles are priced in the cross-section of stock returns, while extreme quantiles are not. Can we still consider an agent that is risk-averse or has quantile preferences?"*
   Response: The finding that central τ-quantiles are priced in the cross-section of stock returns while extreme quantiles are not suggests a non-trivial relationship between investor preferences, risk aversion, and market structure. Below, we outline several key economic interpretations and discuss whether such results align with the behaviour of a risk-averse agent or an agent with quantile preferences. We have also expanded the discussion of these points in Section 2.1.1 and refer to it in Section 2.3.4 for comparison.

   Our quantile forecasts and, subsequently, distributional moment forecasts do not separate systematic and idiosyncratic components of risk. Theoretically, only systematic risk should be priced, as idiosyncratic risk can be

[1]Even though the theme of going beyond the first two moments into quantiles is a bonding link to the 3rd chapter.

[2]E.g. disappointment aversion, general preferences, prospect theory, etc.

diversified away in case of no limits to arbitrage. Our forecasts are total forecasts, and therefore if extreme quantiles or higher moments are more representative of idiosyncratic risk, rather than systematic risk, their lack of pricing in the cross-section would not be surprising.

From a quantile preferences perspective, it is important to distinguish between the quantile of individual stock returns and quantile of portfolio returns. In this paper, we focus on forecasting quantiles of marginal distributions of individual stock returns only. However, investors with quantile preferences maximize a τ-quantile of their future wealth, which is a function of the joint distribution of all assets in their portfolio. Therefore our results revolving around marginal distributions are only part of the story needed for these comprehensive empirical asset-pricing anomalies.

2. *"How do you reconcile your findings of skewness and kurtosis with previous literature showing that left-tail risk and skewness are important priced risk factors."*

   Response: I have expanded the literature review covering higher moments and tail risk in Section 2.1.1.

   Prior literature often relies on realized return-based measures of skewness and tail risk, whereas we rely on predicted measures. Studies such as Bali, Cakici, and Whitelaw (2011) and Boyer et al. (2010) measure skewness using past realized return distributions. Our approach instead forecasts skewness and kurtosis from quantile-based predictions, which incorporate a broader set of return distributional information but may differ from ex-post realized skewness metrics. Further, our model includes realized moments like skewness, kurtosis, and other higher-moment measures as features in predicting conditional quantiles directly capturing the predictive relationship between these moments and future returns. Using predicted instead of past realized skewness measures can lead to different conclusions as in the case of Langlois (2020). Further, the distinction between total, systematic/coskewness and idiosyncratic skewness needs to be made as Langlois (2020) document that only predicted systematic skewness is priced, while the role of idiosyncratic skewness is not clear. This contrasts with our setup, where we are focused on predicted total skewness and do not attempt any decomposition.

   A more comprehensive analysis requiring the implementation of methodolo-

gies from conflicting studies is required to disentangle conflicting evidence and is a valuable avenue for future research.

3. *"I would expect an investor with quantile preferences should display a monotonically decreasing pricing kernel. If this is true, then more negative quantiles should have a higher compensation than higher quantiles, unless the pricing kernel is U-shaped and increasing in the region where the mean is. I think the paper would benefit by adding a discussion along these lines."*

   Response: I have expanded the discussion in Section 2.3.4 accordingly. See my response to the previous comment for a more detailed explanation.

4. *"You found that only the central moment extracted from the proposed approach is priced in the cross-section. Why the other moments are uninformative? Are they too noisy? If you create a double sorting portfolio first sorting on an standard expected mean, and then sorting on your mean, do you still see the same results?"*

   Response: With regard to why other moments than the first moment are uninformative, the short answer is I don't know. For the longer discussion, see the response to the previous comments. I consider the question of why higher moments (specifically skewness based on the literature) and left-tail quantiles do not appear to be priced to be one of the most interesting avenues for future research, deserving a separate study.

   In Table 5.2, you can find the results of the conditional double sorting on mean derived from quantile forecasts and direct mean-squared-error-based forecasts from NNh2 model which uses the same features as the Standardized *tau*-quantile sub-network of the two-stage quantile neural network model. Double sorting is performed conditionally, where stocks are first sorted on the direct standard expected mean forecast from the NNh2 model in the columns and subsequently sorted on mean from a distributional forecast of 2sNN in the rows. The table shows the incremental pricing power of the mean forecast derived from the distributional forecast compared to the direct standard expected mean forecast. This phenomenon is consistent across regions and both full and liquid samples.

5. *"Given that the mean value is the only priced moment, it would be crucial to clearly state the contribution of this Chapter in relation to the previous Chapter, where you show that sorting on predictive premia through Machine*

Learning methods is also priced in the cross-section. Methodologically is very clear, the aim here is forecast the full distribution, but the application somewhat diminish the methodological contribution."

Response: There are multiple additional contributions of this chapter compared to the previous one. First, the chapter introduces a novel methodology for forecasting the entire distribution of stock returns instead of focusing on a single distributional moment - the mean as is the case in the first chapter. Second, even in terms of mean forecasting, we show that using a mean forecast derived from the distributional forecast is superior both in terms of predictive accuracy and economic performance. In this sense, the chapter builds on the first one where the focus is on the role of international evidence and the comparison of different machine learning methods for predicting stock returns. Third, this chapter provides robust empirical evidence based on international data as well as full and liquid-only samples on the pricing of higher moments in the cross-section of stock returns. Even though our empirical results show that higher moments are not priced in the cross-section of stock returns, we believe that this is an important contribution to the existing literature which contains conflicting evidence.

6. *"There are many typos throughout the paper. I will highlight some of them, but I would suggest going carefully through the manuscript.*

   (a) There are many repeated words, for instance:

   - *Page 74: ... of tail-based distributional of tail-based distributional properties.*
   - *Page 74: ... and liquidity of the and liquidity of options markets,*

   ...

   - *Page 74: ... of stock returns. Of stock returns.*

   (b) Page 77: We would use $\overline{\sigma_{t+1}}$ intead of $\overline{\sigma_{t+1}}$ to

   ...

   (c) Many references need to have journal's name in capitals."

Response: Thank you for catching these errors. I have corrected the

typos, journal names and math notation. I have also thoroughly proofread the thesis to fix other similar issues.

Table 5.2. Double sorting on MSE mean and quantile mean

This table presents average monthly stock returns (in percentages) for portfolios formed by double-sorting based on forecasts from two neural network models: a two-hidden-layer neural network optimized with MSE (2hNN MSE - Mean) and a two-stage quantile neural network model (2sNN - Mean). The double-sorting procedure first categorizes stocks into quintiles (Q1-Q5) based on the 2hNN MSE forecasts (columns), then further sorts those stocks based on the 2sNN forecasts (rows). Q1 represents the lowest forecast returns, while Q5 represents the highest forecast returns. Results are presented for four geographic regions (USA, Europe, Japan, Asia Pacific) from 1995 to 2018, with separate panels for Full Sample and Liquid Sample within each region. Values in parentheses are t-statistics adjusted using Newey-West standard errors with 12 lags.

USA

	Full Sample						Liquid Sample					
	Q1	Q2	Q3	Q4	Q5	Q5-Q1	Q1	Q2	Q3	Q4	Q5	Q5-Q1
Q1	-2.64 (-4.18)	-0.40 (-0.92)	0.25 (0.66)	0.48 (1.37)	0.91 (2.22)	3.54 (8.66)	-0.57 (-0.87)	0.58 (1.48)	0.76 (2.33)	0.66 (1.91)	1.13 (2.82)	1.70 (4.11)
Q2	-1.11 (-1.56)	0.42 (1.25)	0.80 (3.05)	1.13 (4.02)	1.63 (5.52)	2.73 (4.77)	-0.01 (-0.02)	0.85 (2.94)	0.80 (3.14)	1.09 (3.79)	1.23 (3.50)	1.24 (4.02)
Q3	-0.34 (-0.74)	0.72 (2.51)	1.01 (4.43)	1.32 (4.94)	1.89 (5.71)	2.23 (7.23)	0.37 (0.84)	0.84 (3.55)	1.10 (4.63)	1.08 (4.45)	1.31 (4.11)	0.94 (3.19)
Q4	0.42 (1.02)	1.22 (4.40)	1.26 (5.49)	1.64 (5.88)	2.37 (5.90)	1.96 (5.89)	0.76 (2.09)	0.70 (2.91)	1.11 (4.85)	1.30 (4.56)	1.53 (4.48)	0.77 (3.50)
Q5	1.20 (2.98)	1.83 (5.06)	1.84 (5.66)	2.11 (6.04)	3.39 (6.18)	2.18 (6.57)	1.14 (2.86)	1.34 (4.64)	1.41 (4.49)	1.47 (4.70)	1.88 (5.45)	0.74 (3.00)
Q5-Q1	3.84 (7.28)	2.23 (6.43)	1.59 (5.24)	1.63 (5.86)	2.48 (5.35)		1.71 (3.81)	0.77 (3.04)	0.65 (3.37)	0.81 (3.48)	0.75 (2.60)	

Europe

	Full Sample						Liquid Sample					
	Q1	Q2	Q3	Q4	Q5	Q5-Q1	Q1	Q2	Q3	Q4	Q5	Q5-Q1
Q1	-1.72 (-3.68)	-0.36 (-1.02)	0.03 (0.09)	0.46 (1.53)	0.71 (2.26)	2.44 (7.78)	-0.87 (-1.53)	0.21 (0.63)	0.35 (1.12)	0.71 (2.28)	0.76 (1.94)	1.64 (4.57)
Q2	-0.87 (-2.13)	0.04 (0.13)	0.49 (1.77)	0.79 (2.77)	1.29 (3.99)	2.16 (6.90)	-0.30 (-0.67)	0.32 (1.04)	0.59 (2.06)	0.69 (2.29)	0.96 (2.68)	1.26 (3.88)
Q3	-0.47 (-1.17)	0.25 (0.81)	0.68 (2.53)	1.00 (3.32)	1.49 (4.55)	1.95 (6.85)	-0.01 (-0.03)	0.31 (1.06)	0.73 (2.64)	0.83 (2.70)	1.08 (3.04)	1.10 (4.99)
Q4	-0.12 (-0.33)	0.57 (1.96)	0.82 (2.80)	1.21 (3.86)	1.91 (5.32)	2.03 (6.92)	0.51 (1.40)	0.67 (2.22)	0.79 (2.75)	1.00 (3.32)	1.21 (3.26)	0.71 (2.92)
Q5	0.87 (2.44)	0.94 (2.99)	1.22 (3.63)	1.70 (4.89)	2.95 (6.57)	2.08 (6.89)	0.52 (1.21)	0.85 (2.51)	0.85 (2.75)	1.14 (3.63)	1.52 (3.72)	1.00 (4.05)
Q5-Q1	2.60 (8.79)	1.31 (7.11)	1.19 (7.93)	1.24 (7.36)	2.24 (8.52)		1.40 (4.25)	0.64 (3.30)	0.50 (2.39)	0.42 (2.70)	0.76 (3.68)	

Japan

	Full Sample						Liquid Sample					
	Q1	Q2	Q3	Q4	Q5	Q5-Q1	Q1	Q2	Q3	Q4	Q5	Q5-Q1
Q1	-1.42 (-2.48)	-0.49 (-1.20)	-0.23 (-0.63)	0.06 (0.17)	0.46 (1.27)	1.88 (5.38)	-1.27 (-2.32)	-0.56 (-1.36)	-0.25 (-0.68)	0.10 (0.29)	0.41 (0.99)	1.69 (4.03)
Q2	-0.79 (-1.52)	-0.36 (-0.99)	0.05 (0.16)	0.50 (1.60)	1.02 (2.83)	1.81 (5.89)	-0.65 (-1.34)	-0.18 (-0.53)	0.22 (0.75)	0.11 (0.39)	0.62 (1.85)	1.27 (4.06)
Q3	-0.44 (-0.93)	-0.03 (-0.10)	0.29 (0.99)	0.67 (2.27)	1.36 (3.54)	1.80 (6.71)	-0.33 (-0.75)	0.21 (0.61)	0.24 (0.86)	0.43 (1.47)	0.71 (2.26)	1.04 (4.16)
Q4	0.02 (0.05)	0.33 (1.04)	0.59 (1.98)	0.92 (2.78)	1.68 (4.04)	1.65 (6.34)	-0.06 (-0.14)	0.27 (0.87)	0.34 (1.13)	0.68 (2.38)	0.83 (2.44)	0.89 (3.76)
Q5	0.59 (1.38)	1.00 (2.55)	1.05 (2.78)	1.44 (3.58)	2.48 (5.16)	1.89 (6.52)	0.05 (0.13)	0.46 (1.31)	0.61 (2.08)	0.73 (2.38)	1.21 (3.06)	1.16 (5.27)
Q5-Q1	2.02 (7.25)	1.49 (6.93)	1.27 (6.54)	1.37 (6.61)	2.02 (7.33)		1.33 (4.05)	1.02 (4.07)	0.87 (4.24)	0.63 (3.02)	0.80 (3.18)	

Asia Pacific

	Full Sample						Liquid Sample					
	Q1	Q2	Q3	Q4	Q5	Q5-Q1	Q1	Q2	Q3	Q4	Q5	Q5-Q1
Q1	-2.68 (-4.17)	-0.45 (-0.93)	-0.00 (-0.01)	0.32 (0.74)	0.74 (1.48)	3.41 (8.08)	-1.26 (-1.82)	0.09 (0.16)	0.13 (0.30)	0.18 (0.36)	0.72 (1.38)	1.98 (4.69)
Q2	-1.26 (-2.41)	-0.12 (-0.28)	0.51 (1.39)	0.80 (1.99)	1.47 (3.23)	2.72 (8.14)	-0.72 (-1.20)	0.33 (0.95)	0.25 (0.69)	0.61 (1.50)	0.89 (1.85)	1.61 (4.52)
Q3	-0.84 (-1.63)	0.11 (0.28)	0.63 (1.65)	1.15 (2.94)	1.75 (3.83)	2.59 (7.66)	-0.03 (-0.08)	0.54 (1.58)	0.74 (2.16)	0.91 (2.47)	1.31 (3.04)	1.35 (5.51)
Q4	-0.40 (-0.86)	0.71 (1.87)	1.01 (2.84)	1.37 (3.46)	2.21 (4.70)	2.61 (8.35)	0.57 (1.44)	0.72 (1.79)	0.85 (2.55)	1.00 (2.61)	1.56 (3.29)	1.00 (3.97)
Q5	0.46 (1.03)	1.24 (3.08)	1.66 (3.99)	2.00 (4.19)	3.27 (6.33)	2.81 (8.37)	0.46 (0.94)	1.10 (2.76)	1.14 (3.17)	1.08 (2.53)	1.92 (4.08)	1.46 (5.49)
Q5-Q1	3.14 (7.89)	1.69 (6.49)	1.67 (6.73)	1.67 (6.95)	2.53 (8.63)		1.72 (3.58)	1.01 (2.63)	1.01 (4.01)	0.90 (2.82)	1.20 (3.69)	

5.3.4 Chapter 3

1. *"The theoretical motivation of the paper could be improved."*
   Response: I have substantially revised the introduction in Section 2.1 to strengthen the theoretical foundations and explicitly position our contributions within the broader literature on decision-making under uncertainty. I have also added text on how quantile preferences connect to multiple strands of financial economics. Further, I have also added a parametric example connecting the τ-quantile maximization with the mean-variance optimization as well as the global minimum variance portfolio in Section 3.3. I believe these revisions substantially strengthen the theoretical motivation of our paper.

2. *"The estimation window seems somewhat arbitrary. Are the results robust to different windows? It is worth commenting on this point."*
   Response: We consider multiple estimation windows in terms of length for both frequencies we study. Specifically, 36 and 60-month-long estimation windows are considered for monthly frequencies (corresponding to 3 and 5 years). For daily frequencies, we consider 63 days long estimation windows (corresponding to 3 months). We find that the results are robust to different estimation windows (both reported and not reported). Description of different estimation windows can be found in Section 3.5.

3. *"The main finding is that global minimum variance portfolios often outperform quantile -based portfolios for risk-averse investors. One should expect that if quantile indexes the attitude towards risk of the investor, extreme quantiles should increase the coefficient of absolute risk aversion, and therefore outperform the mean-variance portfolio. I think it is worth discussing this result. The discussion of the smirk pattern is a vague attempt that requires some work."*
   Response: Thank you for this comment. I fully agree with this expectation on in-sample data or in an oracle setting. However, one of the key findings of this paper is that looking at out-of-sample performance, this is not the case. Based on the simulation results we think this discrepancy is caused by the estimation error and small-sample bias in quantile maximization. As shown in DeMiguel et al. (2009), empirical portfolio optimization is highly sensitive to estimation errors, and quantile maximization is particularly vulnerable because While mean-variance optimization primarily

relies on the estimation of means and covariances (which, despite biases, are still relatively stable), quantile-based optimization involves extreme order statistics, which are far more sensitive to sampling variability. Portfolio weights optimized from small samples may then be unstable, leading to portfolios that fail to provide true downside protection out-of-sample.

With regards to the empirical "smirk" pattern detected across datasets, it only describes a generally observed phenomenon of different concentrations of the portfolio weights across quantiles.

4. *"Some robustness should also help strengthen the results of the paper: Performance after transaction costs, I could not find a discussion of the statistical significance of the performance differences between strategies, did I overlook? Do you really need cryptocurrencies? I feel crypto looks a bit disconnected from the main analysis."*
Response: Thank you for these comments, I have updated the thesis to better address these points. Below are my responses.

We did not include transaction costs in the analysis, as one of our main empirical design choices was to focus on universes of assets used in the portfolio selection literature (e.g. FF portfolios) and extend the analysis with universes containing more heavy-tailed assets like individual stocks and cryptocurrencies. I am not sure about the accuracy of transaction cost estimates for such a diverse set of assets. However, even though we do not include transaction costs directly, we report the portfolio turnovers which can be used as a proxy for transaction costs. The portfolio turnover results, both on the simulated data as well as empirical, show heterogeneity across quantiles with the highest portfolio turnover for central τ-quantile maximizing portfolios and lower for the extreme quantile maximizing portfolios (see Table 3.4 and Table 3.1). Further, low τ-quantile maximizing portfolios have higher turnover than the global minimum variance portfolios indicating that the performance difference would be even larger if transaction costs were included.

The motivation for including cryptocurrencies is to have extremely heavy-tailed assets considered in the analysis, as it is not clear ex-ante how the quantile-maximizing portfolio selection would perform in such an environment compared to traditional min-variance portfolios. The reasoning behind this inclusion can be found in Section 3.5.1.

In terms of statistical significance, we use a percentile bootstrap method to test the statistical differences in the out-of-sample quantiles of portfolio returns. The methodology is described in Appendix 3.A, and significant results are reported in the corresponding tables, with descriptions in the caption.

5. *"The presentation must be improved. Many tables would not fit the page size."*

   Response: Thank you for pointing this out. I have adjusted the tables to fit the page size and improved the presentation of the results overall.

6. *"Minor comments:*

   (a) The literature review could be more comprehensive. For instance, relating this work to other non-traditional portfolio selection methods.

   (b) There are several typos and somewhat akward phrasings thoughtought the chapter. For instance, approximable-hard seems odd."

   Response: Thank you for these comments. I have added references to other non-traditional portfolio selections with brief descriptions to the Introduction of the paper. I did not go into too much detail as I feel that the main focus of the chapter should remain on the core topic of how an investor with quantile preferences would construct their portfolio and how it would perform compared to traditional benchmark portfolios.

   With regards to the typos and awkward phrasings, I have proofread the chapter and improved the language throughout.

References

Ang, A., Hodrick, R. J., Xing, Y., & Zhang, X. (2006). The cross-section of volatility and expected returns. *The Journal of Finance*, *61*(1), 259–299.

Bali, T. G., Cakici, N., & Whitelaw, R. F. (2011). Maxing out: Stocks as lotteries and the cross-section of expected returns. *The Journal of Financial Economics*, *99*(2), 427–446.

Bali, T. G., Engle, R. F., & Murray, S. (2016). *Empirical asset pricing: The cross section of stock returns*. John Wiley & Sons.

Blau, B. M., Bowles, T. B., & Whitby, R. J. (2016). Gambling preferences, options markets, and volatility. *Journal of Financial and Quantitative Analysis*, *51*(2), 515–540.

Blundell, C., Cornebise, J., Kavukcuoglu, K., & Wierstra, D. (2015). Weight uncertainty in neural network. *International Conference on Machine Learning*, 1613–1622.

Bollerslev, T., Li, S. Z., & Zhao, B. (2020). Good volatility, bad volatility, and the cross section of stock returns. *Journal of Financial and Quantitative Analysis*, *55*(3), 751–781.

Boyer, B., Mitton, T., & Vorkink, K. (2010). Expected idiosyncratic skewness. *The Review of Financial Studies*, *23*(1), 169–202.

DeMiguel, V., Garlappi, L., & Uppal, R. (2009). Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? *The Review of Financial Studies*, *22*(5), 1915–1953.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. Chapman; Hall/CRC.

Gal, Y., & Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. *International Conference on Machine Learning*, 1050–1059.

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, *33*(5), 2223–2273.

Hou, K., Xue, C., & Zhang, L. (2020). Replicating anomalies. *The Review of Financial Studies*, *33*(5), 2019–2133.

Koenker, R. (2004). Quantile regression for longitudinal data. *Journal of Multivariate Analysis*, *91*(1), 74–89.

Kumar, A. (2009). Who gambles in the stock market? *The Journal of Finance*, *64*(4), 1889–1933.

Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, *30*.

Langlois, H. (2020). Measuring skewness premia. *Journal of Financial Economics*, *135*(2), 399–424.

Li, Y., & Zhu, J. (2008). L 1-norm quantile regression. *Journal of Computational and Graphical Statistics*, *17*(1), 163–185.

Moon, S. J., Jeon, J.-J., Lee, J. S. H., & Kim, Y. (2021). Learning multiple quantiles with neural networks. *Journal of Computational and Graphical Statistics*, *30*(4), 1238–1248.

Tobek, O., & Hronec, M. (2018). *Does the source of fundamental data matter?* (IES Working Paper No. 15/2018). Charles University, Faculty of Social Sciences, Institute of Economic Studies. Prague.

Tobek, O., & Hronec, M. (2021). Does it pay to follow anomalies research? machine learning approach with international evidence. *Journal of Financial Markets*, *56*, 100588.

The last page no longer left empty.