UNIVERSITY OF CALIFORNIA SAN DIEGO

# Essays in Financial Economics

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Economics

by

Tjeerd De Vries

Committee in charge:

Professor Allan Timmermann, Co-Chair
Professor Alexis Akira Toda, Co-Chair
Professor James D. Hamilton
Professor Yixiao Sun
Professor Rossen Valkanov

2024

The Dissertation of Tjeerd De Vries is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2024

# DEDICATION

To my parents and sister.

TABLE OF CONTENTS

Dissertation Approval Page ..... iii

Dedication ..... iv

Table of Contents ..... v

List of Figures ..... ix

List of Tables ..... xi

Acknowledgements ..... xiii

Vita ..... xiv

Abstract of the Dissertation ..... xv

Chapter 1 A Tale of Two Tails: A Model-free Approach to Estimating Disaster Risk Premia and Testing Asset Pricing Models ..... 1

1.1 Introduction ..... 1
1.1.1 Related Literature ..... 5
1.2 Empirical Estimates of Quantile Difference ..... 7
1.2.1 Notation ..... 8
1.2.2 Methodology and Econometric Model ..... 10
1.2.3 Data and Estimation ..... 14
1.3 Equity Premium Puzzle and SDF Implications ..... 15
1.3.1 Equity Premium Puzzle ..... 15
1.3.2 Driver of Disaster Risk Premia: Insurance or Beliefs? ..... 19
1.3.3 Predicting the Equity Premium ..... 21
1.3.4 Pricing Kernel Monotonicity and Stochastic Dominance ..... 22
1.3.5 Belief Recovery ..... 23
1.4 QR and Robust Estimation of Disaster Risk ..... 24
1.4.1 QR in the Conditional Lognormal Model ..... 24
1.4.2 QR versus Nonparametric SDF Estimation ..... 26
1.5 Disaster Risk and SDF Volatility ..... 28
1.5.1 A Bound on the SDF Volatility ..... 29
1.5.2 Quantile Predictability in the Left-Tail ..... 31
1.5.3 Distribution Bound in Asset Pricing Models ..... 32
1.5.4 Data and Empirical Estimation of the Distribution Bound ..... 36
1.5.5 Unconditional Evidence of Disaster Risk ..... 37
1.6 A Model-Free Lower Bound on Disaster Risk Premia ..... 39
1.6.1 Approximating the Quantile Difference ..... 39
1.6.2 A Lower Bound on Disaster Risk Premia ..... 41
1.6.3 Calculating the Lower Bound ..... 44
1.6.4 Tightness of the Lower Bound: In-sample Evidence ..... 44
1.6.5 Tightness of the Lower Bound: Out-of-sample Evidence ..... 46
1.6.6 Robustness of the QR Estimates ..... 47
1.7 Conclusion ..... 47

1.8 Acknowledgements . . . . . 49

Chapter 2 Scale Economies, Bargaining Power, and Investment Performance: Evidence from Pension Plans . . . . . 54
2.1 Introduction . . . . . 54
2.2 Data and Summary Statistics . . . . . 61
2.2.1 Asset Allocation . . . . . 63
2.2.2 Investment Management Cost . . . . . 64
2.3 Empirical Hypotheses . . . . . 65
2.3.1 Theories of Asset Management . . . . . 66
2.3.2 Plan Size and Choice of Investment Management Style . . . . . 69
2.3.3 Economies of Scale in Investment Management Costs . . . . . 70
2.3.4 Plan Size and Return Performance . . . . . 72
2.4 Investment Management Mandates . . . . . 73
2.4.1 Internal versus External Management . . . . . 75
2.4.2 Active versus Passive Management . . . . . 79
2.4.3 Asset Allocation Decisions . . . . . 81
2.5 Investment Management Costs . . . . . 82
2.6 Investment Performance and Plan Size . . . . . 87
2.6.1 Return Regressions . . . . . 87
2.6.2 Risk-adjusted Return Performance . . . . . 90
2.7 Investment Management Style, Costs, and Return Performance . . . . . 91
2.7.1 A Matching Estimator . . . . . 91
2.7.2 Cost Estimates . . . . . 94
2.7.3 Return Performance . . . . . 95
2.8 Conclusion . . . . . 98
2.9 Acknowledgements . . . . . 99

Chapter 3 Robust Asset-Liability Management . . . . . 116
3.1 Introduction . . . . . 116
3.1.1 Related literature . . . . . 118
3.2 Problem statement . . . . . 120
3.2.1 Model setup . . . . . 120
3.2.2 Problem . . . . . 121
3.2.3 Assumptions . . . . . 121
3.3 Robust asset-liability management . . . . . 123
3.3.1 Robust immunization . . . . . 123
3.3.2 Robust immunization with principal components . . . . . 127
3.3.3 Relation to existing literature . . . . . 128
3.3.4 Implementation . . . . . 130
3.4 Evaluation: static hedging . . . . . 133
3.4.1 Experimental design . . . . . 133
3.4.2 Results . . . . . 136
3.5 Evaluation: dynamic hedging . . . . . 140
3.5.1 Implementing dynamic hedging . . . . . 141
3.5.2 Experimental design . . . . . 143
3.5.3 Results . . . . . 144
3.6 Conclusion . . . . . 145

3.7 Acknowledgments . . . 145

Appendix A Appendix to Chapter 1 . . . 146
A.1 Proofs . . . 146
A.1.1 Decomposing the Equity Premium . . . 146
A.1.2 Stochastic Dominance and Pricing Kernel Monotonicity . . . 146
A.1.3 Proof of Proposition 1.4.1 . . . 148
A.1.4 Proof of Proposition 1.5.1 . . . 151
A.1.5 Distribution Bound when SDF and Return are Jointly Normal . . . 152
A.1.6 Minimizer of Distribution Bound with Normal SDF . . . 154
A.1.7 Distribution Bound when SDF and Return are Log-normal . . . 154
A.1.8 Distribution Bound with Pareto Distribution . . . 156
A.1.9 Derivation of Gâteaux Derivative . . . 160
A.1.10 Proof of Proposition 1.6.4 . . . 161
A.1.11 Formulas for market moments . . . 166
A.2 Risk-Neutral Quantile Regression: Robustness and Departure from Conditional Log-normality . . . 167
A.2.1 Linear versus Non-linear Model: Out-of-Sample Forecasting Accuracy . . . 167
A.2.2 Additional Evidence Against the Lognormal Assumption . . . 167
A.3 Estimating the Risk-Neutral Quantile Function . . . 169
A.3.1 Data Description . . . 169
A.3.2 Estimation Procedure . . . 171
A.4 Verifying Assumption 1.6.2(ii) in Representative Agent Models . . . 172
A.4.1 Log utility . . . 172
A.4.2 CRRA utility . . . 172
A.4.3 CARA utility . . . 172
A.4.4 HARA utility . . . 173
A.4.5 Lower Bound in the Data for CRRA utility . . . 173
A.5 Disaster Probability in Representative Agent Models . . . 173
A.5.1 Log Utility . . . 173
A.5.2 CRRA Utility . . . 175
A.5.3 Exponential utility . . . 177
A.6 Lower Bound in the Data and Robustness . . . 178
A.6.1 Lower Bound in the Data . . . 178
A.6.2 Robustness of the Lower Bound and Risk-neutral Quantile . . . 178
A.6.3 Measurement Error Bias in Quantile Regression . . . 180
A.7 Additional Figures . . . 183

Appendix B Appendix to Chapter 2 . . . 185
B.1 CEM data . . . 185
B.2 Asset Allocation . . . 186
B.2.1 Asset Class Frequency and Geographic Coverage . . . 186
B.2.2 Asset and Sub-asset Classes . . . 187
B.2.3 Evolution in Asset Allocation . . . 189
B.2.4 Asset Management Mandate . . . 191
B.2.5 Asset Allocation and Size: Nonparametric Estimates . . . 193
B.3 Cost Data . . . 195
B.3.1 Cost Components . . . 195

B.3.2 Variation in Costs by Investment Management Mandate . . . . . 196
B.3.3 Management Costs by Country of Domicile . . . . . 200
B.4 Returns, Benchmarks and Risk-Adjustments . . . . . 200
B.4.1 Benchmarks . . . . . 200
B.4.2 Asset Class Return Performance . . . . . 201
B.4.3 Risk Adjustment Regressions . . . . . 202
B.4.4 Construction of risk factors . . . . . 202
B.4.5 Factor regression results . . . . . 203
B.4.6 Factor exposures in policy-adjusted returns . . . . . 204
B.4.7 Performance in Sub-Asset Classes . . . . . 205

Appendix C Appendix to Chapter 3 . . . . . 236
C.1 Space of cumulative discount rates . . . . . 236
C.2 Proof of main results . . . . . 237
C.3 Generic full column rank of $A_+$ . . . . . 245
C.4 No-arbitrage term structure model . . . . . 248
C.4.1 Model and bond price formula . . . . . 248
C.4.2 Data . . . . . 250
C.4.3 Parameter estimates . . . . . 250
C.5 Miscellaneous results . . . . . 253
C.5.1 Bias in the estimated yield curve . . . . . 253
C.5.2 Approximating forward rate changes . . . . . 254
C.5.3 Key rate duration matching . . . . . 256
C.5.4 Sign test . . . . . 256

Bibliography . . . . . 258

LIST OF FIGURES


Figure 1.1. **Effect of jumps on physical and risk-neutral quantile functions.** ..... 10

Figure 1.2. **Lorenz curve and Gini coefficient of the conditional equity premium.** 18

Figure 1.3. **Disaster risk premia for 30-day returns at the 5th percentile** ..... 20

Figure 1.4. **Estimated equity premium** ..... 22

Figure 1.5. **Disaster risk premia at the 5th percentile** ..... 29

Figure 1.6. **HJ and distribution bound in disaster risk model without and with jumps.** ..... 35

Figure 1.7. **Physical/risk-neutral CDF and distribution bound for monthly S&P500 returns** ..... 38

Figure 1.8. **Lower bound on disaster risk premium at 5th percentile** ..... 48

Figure 2.1. **Asset allocation over time** ..... 109

Figure 2.2. **Sub-asset class allocation over time for U.S. plans** ..... 110

Figure 2.3. **Asset allocation by management style and plan size** ..... 111

Figure 2.4. **Median cost by asset management style** ..... 113

Figure 2.5. **Relation between log Cost and log AUM** ..... 114

Figure 3.1. **Basis functions of robust immunization.** ..... 132

Figure 3.2. **Goodness-of-fit of forward rate change approximation.** ..... 135

Figure 3.3. **Return error for different holding periods.** ..... 137

Figure 3.4. **Comparison of best specifications** ..... 139

Figure 3.5. **Distribution of absolute return error** ..... 144

Figure A.1. **Ordinal dominance curve with and without first-order stochastic dominance** ..... 148

Figure A.2. **HJ and distribution bound for heavy tailed returns** ..... 158

Figure A.3. **Out-of-sample quantile forecasting loss** ..... 168

Figure A.4. **Lower bound with CRRA utility for 90-day returns** ..... 174

Figure A.5. **Out-of-sample forecast using risk-adjusted quantile with VIX benchmark** ... 181

Figure A.6. **Out-of-sample forecast using risk-neutral quantile with VIX benchmark** ... 182

Figure A.7. **Bias in QR resulting from measurement error** ... 183

Figure A.8. **Highest existing risk-neutral moment for 30-day returns** ... 184

Figure B.1. **Total AUM by asset class and year for U.S. and non-U.S. plans** ... 225

Figure B.2. **Sub-asset class allocation over time for non-U.S. plans** ... 226

Figure B.3. **Frequency of internal and external active management in 2019** ... 227

Figure B.4. **Nonparametric estimates of the relation between plan size and AUM allocation.** ... 228

Figure B.5. **Investment management costs by mandate for stocks and fixed income holdings.** ... 229

Figure B.6. **Evolution in investment management costs by mandate** ... 230

Figure B.7. **Evolution of stock investment management costs by mandate** ... 231

Figure B.8. **Evolution of fixed income investment management costs by mandate** 232

Figure B.9. **Median management costs by mandate in public sub-asset classes** .. 233

Figure B.10. **Policy-adjusted gross returns** ... 234

Figure B.11. **Policy return box Plots** ... 235

Figure C.1. **Key rate perturbations** ... 257

LIST OF TABLES


Table 1.1. **Tail correlations (in %) of physical and risk-neutral quantile function in asset pricing models** . . . . . 13

Table 1.2. **Risk-neutral quantile regression** . . . . . 50

Table 1.3. **OLS estimates of conditional equity premium** . . . . . 51

Table 1.4. **Sample bounds and bootstrap result** . . . . . 51

Table 1.5. **Quantile regression with lower bound** . . . . . 52

Table 1.6. **Quantile regression with model-free quantile forecast** . . . . . 53

Table 2.1. **Small and large plans' investment allocation by sub-asset class and management structure in 2019** . . . . . 100

Table 2.2. **Asset allocation regression for internal vs. external management** . . . . 101

Table 2.3. **Significance test for the difference in APE for size and cost spread** . . 102

Table 2.4. **Asset allocation regression for passive vs. active management** . . . . . 103

Table 2.5. **Asset allocation regression** . . . . . 104

Table 2.6. **Economies of scale for cost among different investment mandates** . . . 105

Table 2.7. **Regression of policy- and risk-adjusted returns on plan characteristics** 107

Table 2.8. **Effect of asset management style on cost and returns using matching** 108

Table 3.1. **Return error (%) for 30-day holding period**. . . . . 138

Table 3.2. $\ell^1$ **norm of investment shares**. . . . . 140

Table A.1. **Expanding quantile prediction with risk-neutral quantile** . . . . . 170

Table A.2. **Summary statistics of lower bound** . . . . . 178

Table A.3. **Quantile regression using Lower Bound and VIX** . . . . . 179

Table B.1. **Number of participants per asset class and year (Panel A) and by frequency of participation (Panel B)** . . . . . 207

Table B.2. **AUM allocation by asset class in 2009 and 2019** . . . . . 208

Table B.3. **Aggregate Asset Allocation for U.S. (Panel A) and non-U.S. (Panel B) plans** . . . . . 209

Table B.4. **Small and large plans' investment allocation by sub-asset class and management structure in 2009** ..... 210

Table B.5. **Plans' relative allocation to multiple investment mandates** ..... 212

Table B.6. **Frequency of internal and external active management** ..... 214

Table B.7. **Monotonicity test of asset allocation and size** ..... 215

Table B.8. **Regression of cost on plan characteristics** ..... 216

Table B.9. **Economies of scale at the sub-asset class level** ..... 217

Table B.10. **Average scaled investment management costs by asset class and country in 2009 and 2019** ..... 219

Table B.11. **Summary statistics for asset class returns** ..... 220

Table B.12. **Regression of net returns on risk factors** ..... 221

Table B.13. **Regressions of policy-adjusted gross returns on a single risk factor** .. 222

Table B.14. **Regression of sub-asset class returns on plan characteristics** ..... 223

Table C.1. **Parameter estimates** ..... 252

ACKNOWLEDGEMENTS

I would like to express my gratitude to my advisors, Allan Timmermann and Alexis Akira Toda. They taught me how to do research and how to write a paper. I benefited enormously from their continued guidance and support, and I hope to emulate their standards in the future.

I would also like to thank the other members of my committee, namely James Hamilton, Yixiao Sun, and Ross Valkanov, who each provided great feedback during various stages of my research.

Before the PhD, I was fortunate to have Arco van Oord as a supervisor, whose insights and enthusiasm for econometrics and finance furthered my interest in research. I am also grateful to have attended lectures of Mark Veraar (TU Delft) and Andrej Zlatoš (UCSD), who teach beautiful mathematics.

I am also thankful to have made new friends that made the time in San Diego so much more enjoyable, including Edo, Victor, Alec, Davide, Anindo, Vivian, Nikolay, Carlos, Giampaolo, Jordan, Francesco, and others.

Finally, I would like to thank my parents and sister, whose unconditional support and love made my academic career possible.

Chapter 1, in full, is currently being prepared for submission for publication of the material. The dissertation author is the sole author of this material.

Chapter 2, in full, is currently being prepared for submission for publication of the material. De Vries, Tjeerd; Kalfa, Yanki; Timmermann, Allan; Wermers, Russ. "Scale Economies, Bargaining Power, and Investment Performance: Evidence from Pension Plans". The dissertation author is a primary author of this material.

Chapter 3, in full, is currently being prepared for submission for publication of the material. De Vries, Tjeerd; Toda, Alexis Akira. "Robust Asset-Liability Management". The dissertation author is a primary author of this material.

VITA

| | |
|---|---|
| 2016 | Bachelor of Science in Econometrics and Operations Research, Erasmus University Rotterdam, the Netherlands |
| 2018 | Master of Science in Applied Mathematics, Delft University of Technology, the Netherlands |
| 2024 | Doctor of Philosophy in Economics, University of California San Diego |

ABSTRACT OF THE DISSERTATION

Essays in Financial Economics

by

Tjeerd De Vries

Doctor of Philosophy in Economics

University of California San Diego, 2024

Professor Allan Timmermann, Co-Chair
Professor Alexis Akira Toda, Co-Chair

This thesis considers several questions in financial economics. A common theme is the estimation of conditional tail risk and portfolio formation of large institutional investors, specifically pension funds.

Chapter 1 introduces a model-free methodology to assess the impact of disaster risk on the market return. Using S&P500 returns and the risk-neutral quantile function derived from option prices, I employ quantile regression to estimate local differences between the conditional physical and risk-neutral distributions. The results indicate substantial disparities primarily in the left-tail,

reflecting the influence of disaster risk on the equity premium. These differences vary over time and persist beyond crisis periods. On average, the bottom 5% of ex-ante returns contribute to 17% of the equity premium, shedding light on the Peso problem. I also find that disaster risk increases the stochastic discount factor's volatility. Using a lower bound observed from option prices on the left-tail difference between the physical and risk-neutral quantile functions, I obtain similar results, reinforcing the robustness of my findings.

Chapter 2 explores the relation between the size of a defined benefit pension plan and its choice of active vs. passive management, internal vs. external management, and public vs. private markets. We find positive scale economies in pension plan investments; large plans have stronger bargaining power over their external managers in negotiating fees as well as having access to higher (pre-fee)-performing funds, relative to small plans. Using matching estimators, we find that internal management is associated with significantly lower costs than external management, reinforcing the enhanced bargaining power of large pension plans that have fixed-cost advantages in setting up internal management.

In chapter 3, we analyze how financial institutions can hedge their balance sheets against interest rate risk when they have long-term assets and liabilities. Using the perspective of functional and numerical analysis, we propose a model-free bond portfolio selection method that generalizes classical immunization and accommodates arbitrary liability structure, portfolio constraints, and perturbations in interest rates. We prove the generic existence of an immunizing portfolio that maximizes the worst-case equity with a tight error estimate and provide a solution algorithm. Numerical evaluations using empirical and simulated yield curves from a no-arbitrage term structure model support the feasibility and accuracy of our approach relative to existing methods.

# Chapter 1

# A Tale of Two Tails: A Model-free Approach to Estimating Disaster Risk Premia and Testing Asset Pricing Models

## 1.1 Introduction

Disaster risk has emerged as a pervasive and influential concept in asset pricing, offering a prominent explanation of the equity premium puzzle, as well as other asset pricing puzzles.[1] Little is known, however, about the quantitative properties of disaster risk and evidence for it is often inferred indirectly, such as from the historically high equity premium. Nevertheless, a high equity premium does not necessarily arise due to disaster risk, and the literature has yet to reach an unambiguous conclusion regarding its ability to explain asset pricing puzzles (see, e.g., Julliard and Ghosh (2012)). Ross (2015) refers to disaster risk as dark matter and summarizes the concept as follows: "It is unseen and not directly observable but it exerts a force that can change over time and that can profoundly influence markets".

In this paper, I propose a model-free methodology to measure and track disaster risk in S&P500 returns through time. My results unequivocally show that disaster risk is pervasive and is a primary determinant of the equity premium. In establishing these results, I confront two critical challenges that have hindered inference so far about disaster risk. Firstly, to estimate disaster risk in a model-free manner, the literature often estimates the stochastic discount factor (SDF), defined as the ratio of risk-neutral to physical density. Disaster risk is then thought of as the tendency

[1] See, for example, Rietz (1988), Barro (2006, 2009), Drechsler and Yaron (2011), Gabaix (2012), Wachter (2013), Constantinides and Ghosh (2017), Isoré and Szczerbowicz (2017), Farhi and Gourio (2018), Seo and Wachter (2019) and Schreindorfer (2020).

of the SDF to take large values in the left-tail of the return distribution. However, this approach faces scrutiny due to the potential for erratic results when estimating the density ratio in the tails, thereby complicating robust inference. Secondly, it is crucial to account for changing conditioning information. Typically, estimation of the physical density involves pooling historical returns, while the risk-neutral density relies on forward-looking option prices. This disparity in information sets can lead to inconsistent estimates of conditional disaster risk.

To address these challenges, I consider an approach that avoids the need for density estimation. Starting from the absence of arbitrage opportunities, a risk-neutral distribution exists that can be identified from option prices without assuming any model (Breeden and Litzenberger, 1978). However, the conditional physical distribution, which describes the actual evolution of the market return, remains unobserved. To proceed, I use quantile regression (QR) to estimate

$$\underbrace{Q_{t,\tau}(R_{m,t\to N})}_{\text{Unobserved}} = \beta_0(\tau) + \beta_1(\tau)\underbrace{\widetilde{Q}_{t,\tau}(R_{m,t\to N})}_{\text{Observed}} \qquad \tau \in (0,1), \tag{1.1.1}$$

where $Q_{t,\tau}$ and $\widetilde{Q}_{t,\tau}$ represent the physical and risk-neutral $\tau$-quantiles, respectively, of the market return $R_{m,t\to N}$, from period $t$ to $t+N$. The parameters in (1.1.1) can be estimated using quantile regression, with the observed time series of returns, $\{R_{m,t\to N}\}_{t=1}^{T}$, as the dependent variable and $\{\widetilde{Q}_{t,\tau}\}_{t=1}^{T}$ as the regressor. Importantly, both $R_{m,t\to N}$ and $\widetilde{Q}_{t,\tau}$ are conditioned on the same information set.

In general, quantile regression estimates the best linear approximation to the physical quantile function, from which $R_{m,t\to N}$ is drawn. But because the risk-neutral quantile function is a highly non-linear transformation of state variables, the estimation accommodates non-linear dependence between the physical quantile function and the state variables it depends on. Hence, any deviation from the risk-neutral benchmark, $[\beta_0(\tau), \beta_1(\tau)] = [0, 1]$, signifies a local difference between the physical and risk-neutral measures at the $\tau$-quantile. Since the equity premium is determined by these differences, it is natural to define *disaster risk premia* as the difference between $Q_{t,\tau}$ and $\widetilde{Q}_{t,\tau}$ in the left-tail, i.e. for values of $\tau$ close to zero.

Based on the QR estimates, two key findings emerge: (i) the risk-neutral benchmark cannot be rejected in the right-tail ($\tau \geq 0.7$) but it is rejected in the left-tail ($\tau \leq 0.3$); and (ii) the in-sample and out-of-sample explanatory power of the risk-neutral quantile is significantly higher in the right-tail compared to the left-tail. Both findings suggest that disaster risk is the main driver of the equity premium.

Building on these results, I estimate the conditional Lorenz curve and Gini coefficient associated with the equity premium. These statistics summarize how much the conditional equity premium is driven by the lowest returns (disaster), akin to its interpretation of wealth inequality in labor economics. I find that the Lorenz curve is always concave, and the Gini coefficients are far above zero in every time period, thus showing that disaster risk is a pervasive feature of the data. On average, I find that ex-ante returns below the 5th percentile contribute to 17% of the total equity premium.

While this result demonstrates that disaster risk is an important driver of expected returns, it also adds nuance to the degree of disaster risk necessary to explain the equity premium. In particular, previous papers attribute about 90% of the equity premium to returns below the 5th percentile (Barro, 2009; Backus, Chernov, and Martin, 2011; Beason and Schreindorfer, 2022). The results differ since I account for conditioning information embedded in the risk-neutral quantile function, whereas unconditional estimates of disaster risk tend to overestimate this risk, since the physical distribution acquires fatter tails when averaging out state variables.

Comparing the physical and risk-neutral quantile functions over time also sheds light on the role of risk aversion and forward-looking beliefs in jointly determining disaster risk premia. Particularly during crises, the value of an insurance contract that hedges against disaster risk increases, resulting in a downward movement in the left-tail of the risk-neutral quantile function. Simultaneously, investors often revise their beliefs about the likelihood of another disaster, frequently assigning a higher probability to such an event. This effect drives down the left-tail of the physical quantile function, creating an ambiguous overall impact on disaster risk premia. However, the quantile regression estimates indicate that the risk-neutral quantile function decreases propor-

tionally more, highlighting the greater influence of risk aversion in determining disaster risk premia.

Given that the equity premium is primarily driven by disaster risk premia, the discussion above implies that the left-tail of the risk-neutral quantile function can predict the equity premium. An OLS regression of the equity premium against the 5% risk-neutral quantile shows preliminary evidence of forecasting ability, especially out-of-sample. In line with theoretical expectations, a decline in the 5% risk-neutral quantile is associated with a substantial increase in the equity premium. Notably, during the 2008 financial crisis and the 2020 Covid-19 crisis, monthly estimates of the equity premium reached peaks of around 5%.

Besides the equity premium puzzle, the QR estimates shed light on the role of first-order stochastic dominance and the pricing kernel puzzle. Specifically, I find that $\widetilde{Q}_{t,\tau} < Q_{t,\tau}$ holds across most of the distribution, except in the far right-tail, where $\widetilde{Q}_{t,\tau} > Q_{t,\tau}$ frequently occurs. This violation of stochastic dominance raises questions in asset pricing models using the expected utility framework, as it suggests that a representative investor exhibits negative risk aversion. Furthermore, I show that a violation of stochastic dominance implies that the pricing kernel is not monotonic, thereby confirming the pricing kernel puzzle while accounting for conditioning information, and without the need to estimate a density ratio.

To further understand the influence of disaster risk on the pricing kernel, I introduce a *distribution bound* on the SDF volatility that is closely related to the Hansen and Jagannathan (1991) bound. The distribution bound summarizes the risk-return trade-off of an asset paying out one dollar when the market return falls below a certain threshold. I show that disaster risk makes the risk-return trade off highly favorable by going short in an asset paying one dollar in case of a disaster. The price of such an asset is high because investors are willing to pay a significant premium to insure against disaster risk, but the risk is limited since the actual probability of a disaster is comparatively low. The Sharpe ratio associated to this investment therefore dominates the Sharpe ratio of a direct investment in the market portfolio. Specifically, in the data, the Sharpe ratio on selling an asset that pays out one dollar if the return falls below the 5th percentile is 30% in monthly units, while the Sharpe ratio of investing in the market portfolio is only 13%. I also show

that models which do not embed a source of disaster risk, such as conditional lognormal models, cannot rationalize this finding.

I conclude by proposing a model-free lower bound on disaster risk premia to assess the robustness of my earlier findings. This lower bound is observed from option prices and is inspired by recent bounds on the equity premium (Martin, 2017; Chabi-Yo and Loudis, 2020). Using quantile regression, I show that the lower bound explains a substantial proportion of the fluctuation in disaster risk premia over time. Moreover, the lower bound relaxes the assumption of a time-homogeneous relation between the physical and risk-neutral quantile functions. Empirically, the lower bound closely aligns with the disaster risk estimates derived from quantile regression, further strengthening the robustness of my earlier findings.

### 1.1.1 Related Literature

My approach, which uses quantile regression to estimate local dispersions between the physical and risk-neutral distribution, is related to a larger body of literature that estimates the pricing kernel from returns and option data (Aït-Sahalia and Lo, 2000; Jackwerth, 2000; Rosenberg and Engle, 2002; Beare and Schmidt, 2016; Linn, Shive, and Shumway, 2018; Cuesdeanu and Jackwerth, 2018). However, estimating the pricing kernel from returns and options can be challenging, especially in the tails of the distribution, where the ratio of densities that defines the pricing kernel can become unstable. In addition, using historical returns to estimate the physical density can lead to inconsistent results (Linn, Shive, and Shumway, 2018). Beason and Schreindorfer (2022) apply a similar methodology to decompose the unconditional equity premium.

In contrast, QR can be used to draw inference on the pricing kernel indirectly, by leveraging the *observed* realized return and risk-neutral distribution, which avoids the estimation of a density ratio. Furthermore, QR can account for changes in the shape and scale of the underlying SDF over time due to changing conditional information, while the approach of Cuesdeanu and Jackwerth (2018) renders an estimate of the SDF that only allows the normalizing constant to be time-varying, since the shape and scale are time invariant (see Section 1.4.1).

I use the QR estimates to shed light on disaster risk, by analyzing the Lorenz curve associated to the equity premium. Effectively, quantiles decompose the equity premium state-by-state, which is conceptually different from prior literature such as Schneider (2019) and Chabi-Yo and Loudis (2023), whose decompositions rely on averages across the return distribution. This state-by-state decomposition allows one to diagnose more precisely which part of the return distribution contributes to the equity premium. Unlike Schneider (2019), I also incorporate time series data, which enhances estimation of the physical measure. Chabi-Yo and Loudis (2023) is more closely aligned with my methodology since they decompose the market return into three distinct components. However, the estimation of the physical measure differs significantly, as I rely on quantile regression rather than non-linear weighted least squares. Notably, quantile regression proves to be more robust, especially in the tails. Beason and Schreindorfer (2022) also employ a state-by-state decomposition but provide unconditional estimates, potentially leading to an overestimation of disaster risk. Specifically, my methodology shows that the worst 5% of returns contribute only 17% to the equity premium, while Beason and Schreindorfer (2022) report an estimate of 91.5%.

My approach to infer conditional disaster risk is also related to the high-frequency literature. Bollerslev and Todorov (2011) and Bollerslev, Todorov, and Xu (2015) use semimartingale theory to dissect which part of the equity premium is coming from downside/upside risk. This approach is very general but limited to short horizons. Furthermore, it separates the contribution from diffusion and jump risk, while the QR approach can also be used for any part of the return distribution. The QR approach does not require the specification of underlying state variables driving the economy, as these variables are implicitly embedded within the risk-neutral quantile function. A similar rationale has been applied by Andersen, Bondarenko, Todorov, and Tauchen (2015) in a high-frequency context, leveraging derivative prices to gain insights into qualitative features of latent state variables.

Complementary to the QR estimates, I derive a nonparametric bound on the SDF volatility closely related to the bound of Hansen and Jagannathan (1991). They argue that the SDF is necessarily volatile and use this observation to screen asset pricing models. Several papers have

built on this insight using higher-order moment bounds (Snow, 1991; Almeida and Garcia, 2012; Liu, 2021) and entropy bounds (Stutzer, 1995; Bansal and Lehmann, 1997; Alvarez and Jermann, 2005b; Backus, Chernov, and Zin, 2014). These bounds all provide a measure of how much the risk-neutral distribution differs from the physical distribution. Unlike the distribution bound, all of these measures are global in that they rely on averages over the entire state space. The distribution bound in this paper is a function rather than a single statistic and can be considered an intermediate approach between a single bound and a complete estimate of the SDF.

The last part of this paper is also related to the growing literature on using options to estimate forward-looking equity premia (Martin, 2017; Martin and Wagner, 2019; Chabi-Yo and Loudis, 2020). However, unlike those papers that focus on the conditional expectation of excess returns, this paper uses option data to bound conditional return quantiles. Since the obtained bound does not require any parameter estimation and provides information on the entire distribution, it complements the recovery literature (Ross, 2015; Schneider and Trojani, 2019). Furthermore, the observed time variation in the approximation for the left-tail quantile, as documented in this paper, aligns with the concept of time-varying disaster risk proposed in various models by Gabaix (2012), Wachter (2013), Constantinides and Ghosh (2017), Isoré and Szczerbowicz (2017), Farhi and Gourio (2018) and Seo and Wachter (2019).

The rest of this paper is organized as follows. Section 1.2 presents the main empirical results from the quantile regressions and its consequences for the equity premium and SDF are discussed in Section 1.3. Section 1.4 provides further evidence on the robustness of QR to estimate disaster risk relative to extant approaches. Section 1.5 introduces the distribution bound, discusses its use in asset pricing models, and presents estimates of the distribution bound from empirical data. Building on the results of Sections 1.2 and 1.3, Section 1.6 establishes a model-free lower bound on disaster risk premia. Finally, Section 1.7 concludes.

## 1.2 Empirical Estimates of Quantile Difference

This section documents empirical estimates of the conditional difference between the physical and risk-neutral quantile functions. I first discuss the notation and then consider an example

to clarify the idea and motivate the methodology.

### 1.2.1 Notation

Let $R_{m,t\to N}$ denote the market return from period $t$ to $t+N$, where $N$ typically represents 30-, 60-, or 90-days. The risk-free rate over the same period is denoted by $R_{f,t\to N}$, which is assumed to be known at time $t$. In the absence of arbitrage, there exists a positive random variable $M_{t\to N}$ such that, conditional on the investor's information at time $t$,

$$\mathbb{E}_t\left[M_{t\to N}R_{m,t\to N}\right]=1. \tag{1.2.1}$$

The random variable $M_{t\to N}$ is referred to as the stochastic discount factor (SDF) and the expectation in (1.2.1) is calculated under the *physical* probability measure $\mathbb{P}_t$, which is the actual distribution of the market return, i.e. $R_{m,t\to N}\sim\mathbb{P}_t$. The SDF can potentially depend on many state variables, but these are suppressed from the notation for brevity. It is convenient to restate (1.2.1) in terms of risk-neutral probabilities:

$$\widetilde{\mathbb{E}}_t\left(R_{m,t\to N}\right)=1/\mathbb{E}_t\left[M_{t\to N}\right]=R_{f,t\to N},$$

where the expectation is calculated under the *risk-neutral* measure $\widetilde{\mathbb{P}}_t$ induced by $M_{t\to N}$. Finally, $F_t(x)\coloneqq\mathbb{P}_t(R_{m,t\to N}\leq x)$ denotes the physical CDF of the market return conditional on the investor's information available at time $t$, $f_t(\cdot)$ denotes the conditional probability density function (PDF) and $Q_{t,\tau}$ denotes the conditional $\tau$-quantile. As before, a tilde superscript refers to the risk-neutral measure, so that

$$\widetilde{F}_t(\widetilde{Q}_{t,\tau})=\widetilde{\mathbb{P}}_t\left(R_{m,t\to N}\leq\widetilde{Q}_{t,\tau}\right)=\tau,\qquad\forall\tau\in(0,1).$$

The physical and risk-neutral quantiles depend on the underlying random variable $R_{m,t\to N}$ (i.e., $\widetilde{Q}_{t,\tau}\coloneqq\widetilde{Q}_{t,\tau}(R_{m,t\to N})$), but I typically omit this dependence as the underlying random variable always refers to the market return.

To clarify my approach of using quantiles to analyze disaster risk, I consider the following

asset pricing model that will be used several times in the paper.

**Example 1.2.1** (Disaster risk)**.** Consider the disaster risk model analyzed in Backus, Chernov, and Martin (2011). The SDF process is given by

$$\log M_{t\to N} = \log \beta - \gamma \log G_{t\to N},$$

where $\beta$ is a time discount factor, $\gamma$ is the coefficient of relative risk aversion, and $G_{t\to N} = C_{t+N}/C_t$ is consumption growth in period $t+N$. Consumption growth follows a two-component structure:

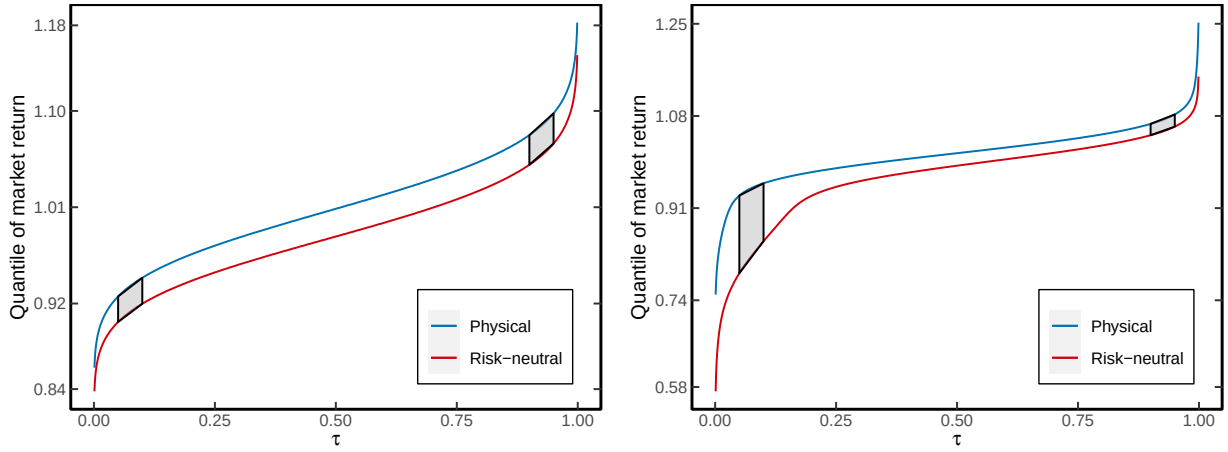
$$\log G_{t\to N} = z_{1,t+N} + z_{2,t+N}, \quad z_{1,t+N} \sim \mathcal{N}(\mu, \sigma^2),$$

and $z_{2,t+N}$ is a Poisson mixture of normals to capture jumps representing rare shocks to consumption growth that are large in magnitude. The number of jumps, $j$, take on nonnegative integer values with probability $e^{-\omega}\omega^j/j!$, and conditional on $j$, the jump term is normal: $z_{2,t+N}|j \sim \mathcal{N}(j\theta, j\delta^2)$. Backus, Chernov, and Martin (2011) show that the risk-neutral distribution of consumption growth in a representative agent model is again a normal mixture with parameters:

$$\widetilde{\mu} = \mu - \gamma\sigma^2, \quad \widetilde{\omega} = \omega e^{-\gamma\theta+(\gamma\delta)^2}, \quad \widetilde{\theta} = \theta - \gamma\delta^2. \tag{1.2.2}$$

In this setup, risk aversion amplifies the jump frequency ($\widetilde{\omega} > \omega$ if $\theta < 0$) as well as the jump size ($\widetilde{\theta} < \theta$). If the model is calibrated such that $\theta \ll 0$, then $z_{2,t+N}$ can be interpreted as a disaster shock if a jump takes place ($j \geq 1$).

Figures 1.1a and 1.1b illustrate the impact of jumps on the physical and risk-neutral quantile functions. Specifically, in the absence of jumps, the market return follows a lognormal distribution and Figure 1.1a shows that the difference between the physical and risk-neutral quantile functions is approximately equal in both tails. However, when jumps are introduced, this difference is almost entirely concentrated in the left-tail. This result is driven by the impact of jumps on the risk-neutral distribution, and the requirement that $\theta < 0$ is crucial to drive a wedge between the physical and risk-neutral measures in the left-tail (see (1.2.2)). The question is whether these distinct shape restrictions on the physical and risk-neutral distribution are supported by the data.




**(a)** Without jumps **(b)** With jumps

**Figure 1.1. Effect of jumps on physical and risk-neutral quantile functions.** The left panel displays the physical and risk-neutral quantile functions without jumps ($\omega = 0$), while the right panel illustrates the quantile functions with jumps ($\omega = 1.4$). In both cases, the mean of the disaster shock ($\theta$) is set to $-0.0074$. The market return is defined as a levered claim on the consumption asset. The trapezoids represent the difference in quantile functions at the 10th and 90th percentiles.

### 1.2.2 Methodology and Econometric Model

Building on the discussion in Example 1.2.1, it is of interest to estimate the quantile difference between the physical and risk-neutral measures. The disaster risk model predicts that these differences are significant in the left-tail while negligible in the right-tail. This is because investors' marginal utility of wealth in states associated with a disaster is very high. Consequently, the following excess return,

$$\underbrace{\widetilde{\mathbb{E}}_t\left[\mathbb{1}\left(R_{m,t\to N} \leq x\right)\right]}_{\text{price}} - \underbrace{\mathbb{E}_t\left[\mathbb{1}\left(R_{m,t\to N} \leq x\right)\right]}_{\text{expected payoff}} = \widetilde{F}_t(x) - F_t(x),$$

can be rather high for thresholds $x$ in the left-tail. The price is high because investors are willing to pay for insurance against disaster risk, even though the actual probability of a disaster can be significantly lower. This effect drives a wedge between the physical and risk-neutral CDFs in the left-tail. For analytical and estimation convenience, it proves more fruitful to consider the inverse of the CDFs (i.e., the quantile functions). Therefore, I consider $Q_{t,\tau} - \widetilde{Q}_{t,\tau}$, and refer to these differences in the left-tail as *disaster risk premia*.

While the conditional risk-neutral distribution and its quantile function can be inferred from option prices without specific modeling assumptions (Breeden and Litzenberger, 1978), the same cannot be said for the physical distribution, unless strong assumptions are made about the martingale component of the SDF (Ross, 2015; Borovička, Hansen, and Scheinkman, 2016). The information available about the conditional physical distribution is limited to a single realization of the market return, as $R_{m,t\to N}$ follows $\mathbb{P}_t$ conditional on time $t$. Consequently, the primary challenge in measuring disaster risk premia lies in the unobservable nature of $Q_{t,\tau}$, which has made model-free inference challenging thus far.

**Risk-Neutral Quantile Regression**

In order to overcome this difficulty, I assume the following model for the physical quantile function

$$\underbrace{Q_{t,\tau}(R_{m,t\to N})}_{\text{Unobserved}} = \beta_0(\tau) + \beta_1(\tau)\underbrace{\widetilde{Q}_{t,\tau}(R_{m,t\to N})}_{\text{Observed}}, \qquad \forall \tau \in (0,1). \tag{1.2.3}$$

If the world is risk-neutral, $[\beta_0(\tau), \beta_1(\tau)] = [0,1]$ for all $\tau$. Departures from risk-neutrality at a specific percentile $\tau$ are reflected by point estimates of $[\beta_0(\tau), \beta_1(\tau)]$ that are far from the $[0,1]$ benchmark. Given a sample of $T$ observations $\{R_{m,t\to N}, \widetilde{Q}_{t,\tau}\}_{t=1}^{T}$, the unknown parameters in (1.2.3) can be estimated by quantile regression (Koenker and Bassett, 1978):

$$[\widehat{\beta}_0(\tau), \widehat{\beta}_1(\tau)] = \underset{(\beta_0,\beta_1)\in\mathbb{R}^2}{\arg\min} \sum_{t=1}^{T} \rho_\tau(R_{m,t\to N} - \beta_0 - \beta_1 \widetilde{Q}_{t,\tau}), \tag{1.2.4}$$

where $\rho_\tau(\cdot)$ is the check function from quantile regression

$$\rho_\tau(x) = \begin{cases} \tau x, & \text{if } x \geq 0 \\ (\tau - 1)x & \text{if } x < 0. \end{cases}$$

Even if the world is not risk-neutral, the model in (1.2.3) can still be correctly specified, as is the case for conditional lognormal models (see Section 1.4.1). When the model is misspec-

ified, the estimation in (1.2.4) remains meaningful as QR finds the best linear approximation to the conditional quantile function (Angrist, Chernozhukov, and Fernández-Val, 2006).[2] Since the risk-neutral quantile itself is a highly non-linear transformation of state variables, the model can accommodate non-linear dependence between the physical quantile function and state variables driving the economy. The benefit of using the risk-neutral quantile function as a regressor is that it does not require the econometrician to take a stand on the state variables driving the physical distribution. Furthermore, both $R_{m,t\to N}$ and $\widetilde{Q}_{t,\tau}$ are conditioned on the same information set, thus avoiding the mismatched information critique of Linn, Shive, and Shumway (2018).

In addition, theory often suggests tantalizing links between the tails of the physical and risk-neutral distribution. Table 1.1 presents correlations between $Q_{t,\tau}$ and $\widetilde{Q}_{t,\tau}$ for both left and right tails in different asset pricing models. In most models, these correlations are nearly one, indicating a strong positive relation that can be modeled by (1.2.3). Only for $\tau = 0.3$, the correlation is notably lower at 41% in the Campbell and Cochrane (1999) model and -67% in the Drechsler and Yaron (2011) model.

In Appendix A.2.1, I consider non-linear specifications as alternatives to the linear model in (1.2.3). Broadly speaking, I find that the linear model outperforms all non-linear models when predicting the physical quantile function out-of-sample. Based on this evidence, and the close linear approximation suggested by asset pricing models, I use the linear specification throughout most of the paper.

*Remark* 1. An alternative to QR is nonparametric estimation of the SDF as proposed by Aït-Sahalia and Lo (2000), Jackwerth (2000) and Rosenberg and Engle (2002). This method can infer the quantile difference from the estimated SDF but relies on pooled historical returns, which can be problematic for forward-looking distribution estimation (Linn, Shive, and Shumway, 2018). More recently, Linn, Shive, and Shumway (2018) and Cuesdeanu and Jackwerth (2018) proposed an estimator of the SDF that accounts for forward-looking information. However, this method

[2]This result is analogous to OLS, which finds the best linear approximation to the conditional *expectation* function, even if the model is misspecified.

**Table 1.1. Tail correlations (in %) of physical and risk-neutral quantile function in asset pricing models**

| Percentile | 0.05 | 0.1 | 0.2 | 0.3 | 0.7 | 0.8 | 0.9 | 0.95 |
|---|---|---|---|---|---|---|---|---|
| Lognormal | | | | | | | | |
| Campbell and Cochrane (1999) | 96.94 | 94.49 | 83.61 | 40.88 | 86.51 | 94.25 | 97.27 | 98.26 |
| Bansal and Yaron (2004) | 99.97 | 99.97 | 99.98 | 99.98 | 99.99 | 99.99 | 99.99 | 100.00 |
| Disaster | | | | | | | | |
| Drechsler and Yaron (2011) | 99.90 | 99.44 | 94.67 | -67.16 | 96.88 | 98.75 | 99.45 | 99.67 |
| Wachter (2013) | 95.40 | 99.63 | 99.57 | 98.98 | 99.71 | 99.88 | 99.94 | 99.97 |
| Constantinides and Ghosh (2017) | 99.86 | 99.72 | 99.21 | 97.58 | 85.96 | 94.68 | 97.32 | 97.90 |

*Note*: This table reports the correlation between $Q_{t,\tau}$ and $\widetilde{Q}_{t,\tau}$ in conditional lognormal models and models that embed a source of conditional disaster risk. The correlations at various percentiles are obtained by simulating $10^6$ draws of the ergodic distribution of states in each model.

presents challenges such as non-convex optimization, the objective function might be undefined due to the small number of existing risk-neutral moments (see Figure A.8), ambiguity in basis function selection, and the inability to account for shape changes in the SDF leading to incorrect conditional inference. QR, on the other hand, avoids these issues, as shown in more detail in Section 1.4.1.

**Measures of Fit**

Based on the quantile regression (1.2.4), I consider two measures of fit to evaluate how well the risk-neutral quantile locally approximates the physical distribution. The first in-sample measure, $R^1(\tau)$, is defined as[3]

$$R^1(\tau) \coloneqq 1 - \frac{\min_{b_0,b_1} \sum_{t=1}^{T} \rho_\tau(R_{m,t\to N} - b_0 - b_1 \widetilde{Q}_{t,\tau})}{\min_{b_0} \sum_{t=1}^{T} \rho_\tau(R_{m,t\to N} - b_0)}. \tag{1.2.5}$$

This measure of fit was proposed by Koenker and Machado (1999) and is a clean substitute for the OLS $R^2$. I also consider an out-of-sample measure of fit

$$R^1_{oos}(\tau) \coloneqq 1 - \frac{\sum_{t=w}^{T} \rho_\tau(R_{m,t\to N} - \widetilde{Q}_{t,\tau})}{\sum_{t=w}^{T} \rho_\tau(R_{m,t\to N} - \overline{Q}_{t,\tau})}, \tag{1.2.6}$$

where $\overline{Q}_{t,\tau}$ is the historical rolling quantile of the market return from time $t-w+1$ to $t$, and $w$ is the rolling window length. Notice that (1.2.6) is a genuine out-of-sample metric since no parameter

[3]It is well known that $b_0$ in the denominator of (1.2.5) equals the in-sample $\tau$-quantile.

estimation is used. In the equity premium literature, Campbell and Thompson (2008) stress the importance of out-of-sample predictability; (1.2.6) is analogous to their out-of-sample $R^2$.

### 1.2.3 Data and Estimation

To estimate the quantile regression in (1.2.4), I require data on the market return and the risk-neutral distribution over time. I use overlapping returns on the S&P500 index from WRDS over the period 2003–2021 to represent the market return. I calculate the market return over a horizon of 30-, 60-, and 90-days. Second, over the same horizon, I use put and call option prices on the S&P500 on each day $t$ from OptionMetrics to estimate the risk-neutral quantile function based on the Breeden and Litzenberger (1978) formula:

$$\widetilde{F}_t\left(\frac{K}{S_t}\right) = R_{f,t\to N}\frac{\partial}{\partial K}\text{Put}_t(K), \tag{1.2.7}$$

where $\text{Put}_t(K)$ denotes the time $t$ price of a European put option on the S&P500 index with stock price $S_t$, strike price $K$ and expiration date $t+N$. This formula is model-free and only requires a no-arbitrage assumption. Due to the lack of a continuum of option prices, interpolation of different maturity options and missing data for option prices far in– and out-of-the money, it is a nontrivial exercise to obtain accurate estimates of $\widetilde{F}_t$ (and hence $\widetilde{Q}_{t,\tau}$) from (1.2.7). A detailed description of my approach that overcomes these issues is described in Appendix A.3.2, which is based on Filipović, Mayerhofer, and Schneider (2013).[4] Finally, I obtain the risk-free rate from Kenneth French's website.[5]

Table 1.2 shows the QR estimates of (1.2.4). The point estimates are close to the $[0,1]$ benchmark in the right-tail ($\tau \geq 0.7$), but not in the left-tail ($\tau \leq 0.3$).[6] Additionally, the joint restriction that $[\beta_0(\tau), \beta_1(\tau)] = [0,1]$ is rejected for all $\tau \leq 0.2$, at all horizons. In contrast, the null hypothesis is never rejected for $\tau \geq 0.8$. The fact that the risk-neutral distribution provides a good approximation of the physical distribution in the right-tail is confirmed by the measures of

[4]This approach uses a kernel density and adds several correction terms to approximate the risk-neutral density. I follow Barletta and Santucci de Magistris (2018) and use a principal components step to avoid overfitting in the tails.

[5]See http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html#Research

[6]Because the risk-neutral quantile function is estimated, there is a concern for attenuation bias due to measurement error. Unreported simulations show that this bias is very small in a setting that mimics the empirical application.

fit, $R^1(\tau)$ and $R^1_{oos}(\tau)$, which are also shown in Table 1.2. Specifically, both in- and out-of-sample, the risk-neutral quantile fits the physical distribution much better in the right-tail.

*Remark* 2. The standard errors for the quantile regression in Table 1.2 are obtained by the smooth extended tapered block bootstrap (SETBB) of Gregory, Lahiri, and Nordman (2018), which is robust to heteroscedasticity and weak dependence.[7] This robustness is important in the estimation, since I use overlapping returns which creates time dependence in the error term, akin to the overlapping observation problem in OLS (Hansen and Hodrick, 1980). SETBB also renders an estimate of the covariance matrix between $\widehat{\beta}_0(\tau)$ and $\widehat{\beta}_1(\tau)$, which can be used to test joint restrictions on the coefficients.[8]

## 1.3 Equity Premium Puzzle and SDF Implications

Building on the estimates in Table 1.2, this section shows that the conditional equity premium is driven by disaster risk, and that disaster risk is a pervasive feature of the data, which poses a new challenge to asset pricing models. I further comment on two implications of Table 1.2 that relate to properties of the SDF that have previously received attention in the literature.

### 1.3.1 Equity Premium Puzzle

The results in Table 1.2 show that the physical distribution is close to risk-neutral in the right-tail, but not in the left-tail. Investors in the market portfolio thus get compensated for bearing downside risk, but not upside risk. This result has important repercussions for explanations of the

[7] It may seem counterintuitive that the standard errors decrease in the tails, which are generally harder to estimate. However, since the regressor $\widetilde{Q}_{t,\tau}$ changes with $\tau$, there is an opposing effect that can cause the standard errors to decrease in the tails. This happens if $\widetilde{Q}_{t,\tau}$ is more variable in the tails, akin to the intuition in OLS that more variability in the regressor decreases the standard error. In the data, $\widetilde{Q}_{t,\tau}$ is much more variable in the tails.

[8] I use the `QregBB` function from the eponymous *R*-package, available on the author's Github page: https://rdrr.io/github/gregorkb/QregBB/man/QregBB.html. The only user required input for this method is the block length in the bootstrap procedure.

equity premium puzzle. To see this, consider the following decomposition of the equity premium[9]

$$\begin{aligned}\mathbb{E}_t\left[R_{m,t\to N}\right] - R_{f,t\to N} &= \int_0^1 \left(Q_{t,\tau} - \widetilde{Q}_{t,\tau}\right) \mathrm{d}\tau \\ &= \underbrace{\int_0^{\underline{\tau}} \left(Q_{t,\tau} - \widetilde{Q}_{t,\tau}\right) \mathrm{d}\tau}_{\text{disaster risk}} + \int_{\underline{\tau}}^1 \left(Q_{t,\tau} - \widetilde{Q}_{t,\tau}\right) \mathrm{d}\tau, \end{aligned} \tag{1.3.1}$$

where $\underline{\tau}$ is a percentile close to zero. The first term on the right-hand side aggregates the local difference between the risk-neutral and physical quantiles in the left-tail, which I define as the contribution of disaster risk. The results in Table 1.2 show that these differences are the primary determinant for the equity premium, as in the right-tail we have $Q_{t,\tau} \approx \widetilde{Q}_{t,\tau}$. The latter finding is consistent with the modeling assumption in (time-varying) disaster risk models that shocks to the market return are negative conditional on a disaster occurring (see, e.g., the condition $\theta < 0$ in Example 1.2.1). Hence, an asset pricing model seeking to explain the (conditional) equity premium of the market return must embed a source of disaster risk.

To illustrate the pervasiveness of disaster risk in the data, I consider the Lorenz curve associated with the conditional equity premium

$$L_t(x) \coloneqq \frac{\int_0^x \left(Q_{t,\tau} - \widetilde{Q}_{t,\tau}\right) \mathrm{d}\tau}{\mathbb{E}_t\left[R_{m,t\to N}\right] - R_{f,t\to N}} \overset{(1.3.1)}{=} \frac{\int_0^x \left(Q_{t,\tau} - \widetilde{Q}_{t,\tau}\right) \mathrm{d}\tau}{\int_0^1 \left(Q_{t,\tau} - \widetilde{Q}_{t,\tau}\right) \mathrm{d}\tau}, \quad 0 \le x \le 1.$$

The Lorenz curve summarizes the proportion of the equity premium contributed by the bottom $x\%$ of returns, akin to its interpretation in labor economics to summarize wealth inequality. Since $Q_{t,\tau}$ is unobserved, I use instead the inferred value, $\widehat{Q}_{t,\tau} = \widehat{\beta}_0(\tau) + \widehat{\beta}_1(\tau)\widetilde{Q}_{t,\tau}$, with the estimated parameters coming from the QR estimates in (1.2.4).

Figure 1.2a shows the average Lorenz curve in the data, together with the Lorenz curve implied by various asset pricing models.[10] In the data, the Lorenz curve is quite concave, thus showing that the majority of the equity premium is contributed by the left-tail. At the same time, my estimation adds nuance to the degree of disaster risk influencing the equity premium.

[9] See Appendix A.1.1 for a derivation.

[10] I thank Beason and Schreindorfer (2022) for making the code to simulate from these models publicly available.

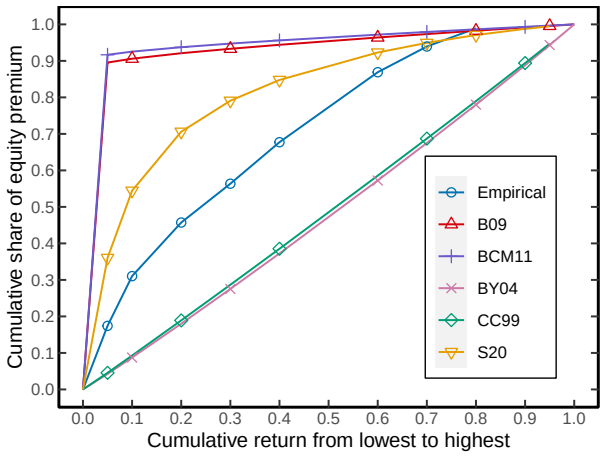
Specifically, while the disaster risk models of Barro (2009) and Backus, Chernov, and Martin (2011) attribute approximately 90% of the equity premium to the lowest 5% of returns, empirical estimates suggest this proportion is only around 17%. These findings also deviate substantially from the nonparametric estimates of Beason and Schreindorfer (2022), who report that 91.5% of the equity premium is driven by the bottom 5% of returns. Our results differ because I account for conditioning information, while Beason and Schreindorfer (2022) employ an unconditional approach. Using unconditional averages can inflate the tails of the physical distribution (Chabi-Yo, Garcia, and Renault, 2008), leading to an overestimation of disaster risk.

On the other hand, the models of Campbell and Cochrane (1999) and Bansal and Yaron (2004) are even more misspecified since the Lorenz curve in these models is slightly convex, thus attributing more than 50% of the equity premium to upside returns. The model of Schreindorfer (2020) matches the Lorenz curve best, even though it also overestimates the contribution of disaster risk to the equity premium.

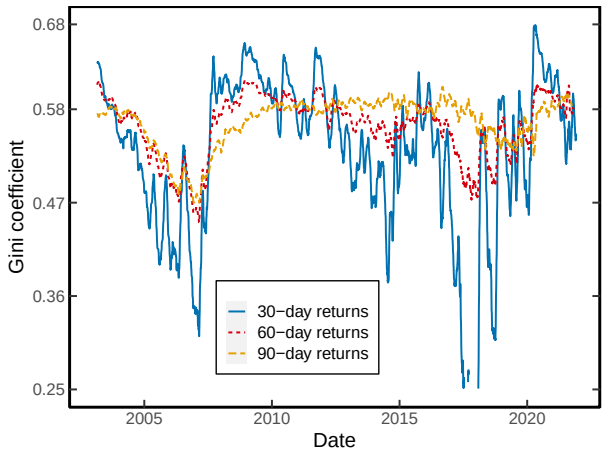
I also consider the Gini coefficient derived from the Lorenz curve

$$G_t = 2 \int_0^1 L_t(\tau)\, \mathrm{d}\tau - 1.$$

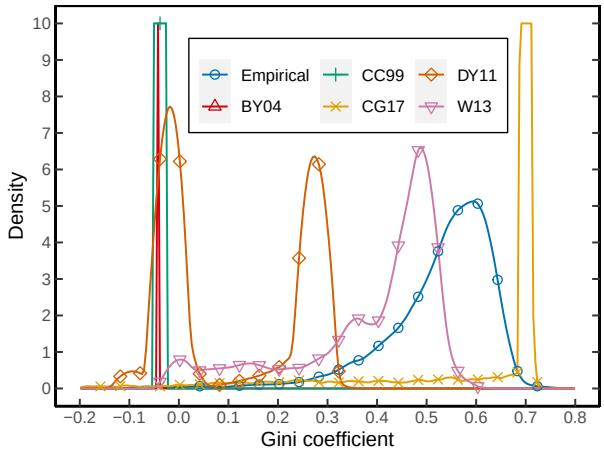
By construction, the Gini coefficient is between -1 and 1, and a value closer to 1 indicates that a bigger proportion of the equity premium is coming from the left-tail. In contrast, a value of 0 suggests that the equity premium is evenly distributed across the return distribution, while negative values imply that the right-tails contribute more to the equity premium than the left-tails. Figure 1.2b shows the time series of conditional Gini coefficients for the various return horizons. For 30-day returns, the Gini coefficient mostly hovers between 0.33 and 0.68. At longer horizons, the Gini coefficients exhibit less variability and typically range between 0.47 to 0.6. These coefficients are also countercyclical, peaking during periods associated with economic downturns, such as the 2008 financial crisis and the Covid-19 crisis. Overall, the Gini coefficients consistently exhibit strong positive values, highlighting the pervasiveness of conditional disaster risk in the data, which extends beyond crisis periods.




**(a)** Average Lorenz curve




**(b)** Conditional Gini coefficient




**(c)** Ergodic distribution of Gini coefficients

**Figure 1.2. Lorenz curve and Gini coefficient of the conditional equity premium.** This figure presents the Lorenz curve and Gini coefficient associated with the conditional equity premium in both empirical data and asset pricing models. Panel (a) displays the time-averaged Lorenz curve estimated from 30-day returns (Empirical) alongside Lorenz curves implied by the unconditional asset pricing models of Barro (2009) (B09), Backus, Chernov, and Martin (2011) (BCM11), and Schreindorfer (2020) (S20), as well as the average Lorenz curve from conditional asset pricing models by Campbell and Cochrane (1999) (CC99) and Bansal and Yaron (2004) (BY04). Panel (b) depicts the estimated Gini coefficient over time for different return horizons and is smoothed using a 30-day rolling window. Panel (c) shows the ergodic distribution of Gini coefficients estimated from 30-day returns (Empirical) and those implied by the conditional asset pricing models of CC99, BY04, Drechsler and Yaron (2011) (DY11), Wachter (2013) (W13), and Constantinides and Ghosh (2017) (CG17). Model parameters are calibrated on a monthly frequency, and the ergodic distribution is derived from 10,000 state draws.

Finally, I analyze the ergodic distribution of Gini coefficients in time-varying asset pricing models and compare it to the distribution implied by the data.[11] Figure 1.2c displays these

[11] In the model, I obtain the distribution of Gini coefficients from the state distribution. In the data, I rely on the time series average. If the data are generated by the model and the system is ergodic, Birkhoff's theorem implies that the state and time averages are equal almost everywhere.

distributions and shows that many asset pricing models have difficulty in matching the empirical distribution. The conditional lognormal models of Campbell and Cochrane (1999) and Bansal and Yaron (2004) imply negative Gini coefficients, with minimal variation among different states, contrary to what the data indicate. The models of Drechsler and Yaron (2011), Wachter (2013), Constantinides and Ghosh (2017) all incorporate a source of disaster risk, but they also have difficulty to match the empirics. In particular, the models of Drechsler and Yaron (2011) and Wachter (2013) embed too little disaster risk, while the model of Constantinides and Ghosh (2017) overestimates the impact of disaster risk.

### 1.3.2 Driver of Disaster Risk Premia: Insurance or Beliefs?

Disaster risk premia have two components: an insurance effect and a forward-looking beliefs effect (under rational expectations). To see this, consider again the short position in a derivative security that pays one dollar if the market return is below a threshold, denoted by $x$, in the left-tail. The excess return on such an investment can be interpreted as

$$\underbrace{\widetilde{\mathbb{E}}_t\left[\mathbb{1}\left(R_{m,t\to N}\leq x\right)\right]}_{\text{price of insurance}}-\underbrace{\mathbb{E}_t\left[\mathbb{1}\left(R_{m,t\to N}\leq x\right)\right]}_{\text{forward looking belief}}=\widetilde{F}_t(x)-F_t(x).$$

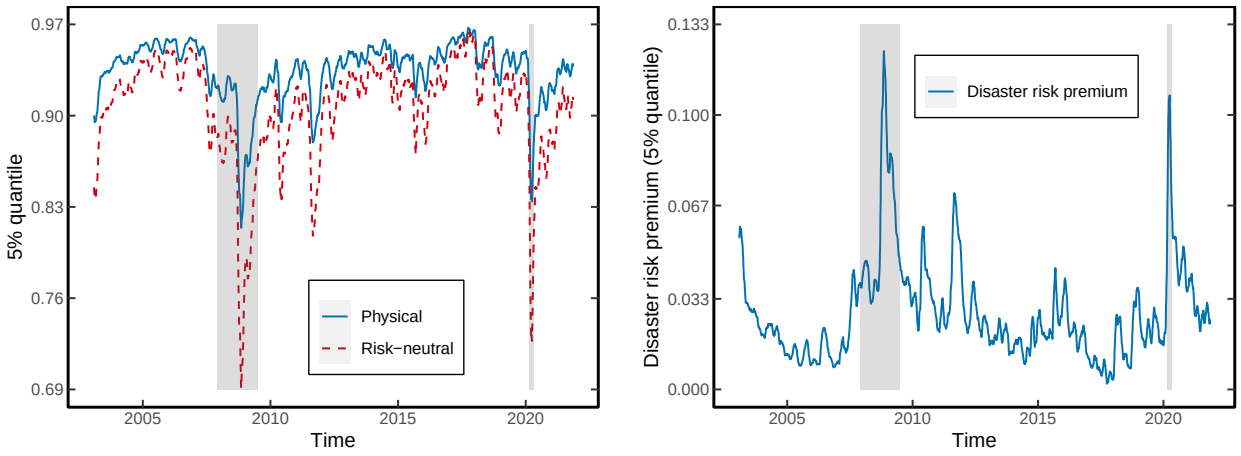
During a crisis, the price of this insurance security tends to rise. This effect can occur in the disaster risk model (Example 1.2.1), if risk aversion increases when a disaster hits, leading to an increase in $\widetilde{F}_t(x)$ and a subsequent decrease in $\widetilde{Q}_{t,\tau}$. Simultaneously, investors may believe that the actual probability of a disaster increases during a crisis. This belief drives up $F_t(x)$ and, consequently, pushes down $Q_{t,\tau}$.

Building on this discussion, it is not immediately clear what the net effect is on disaster risk premia ($Q_{t,\tau}-\widetilde{Q}_{t,\tau}$), as both $Q_{t,\tau}$ and $\widetilde{Q}_{t,\tau}$ tend to decrease during periods of heightened market uncertainty. Figure 1.3a illustrates this effect for 30-day returns and $\tau=0.05$. Notably, during the global financial crisis and Covid-19 crisis, both the physical and risk-neutral quantile functions exhibit significant drops.

To shed light on the net effect on disaster risk premia during crises, Figure 1.3b displays the

evolution of disaster risk premia over time. The most significant change occurs during the peak of the global financial crisis and the Covid-19 crisis. In these turbulent periods, disaster risk premia consistently rise, suggesting that the insurance effect is more substantial than the forward-looking beliefs effect.[12] Because of these large increases, disaster risk is a more important driver of the equity premium, which clarifies the countercyclical Gini coefficients in Figure 1.2b.

The downward fluctuations in the risk-neutral quantile function can be particularly pronounced, plummeting to as low as 63% during crisis periods. In contrast, the physical quantile function only drops to 78%, suggesting that a monthly loss of 22% or more had a 5% probability. To put this in perspective, this probability is 14 times higher than the estimate obtained from historical monthly S&P500 returns (from 1926 to 2021). This calculation shows that historical estimates can diverge significantly from forward-looking beliefs. Furthermore, the time fluctuations in the physical quantile function lend empirical support to the notion of time-varying disaster risk, as proposed in various models such as Gabaix (2012), Wachter (2013), Constantinides and Ghosh (2017), Isoré and Szczerbowicz (2017), Farhi and Gourio (2018) and Seo and Wachter (2019).




**(a)** 5% quantile

**(b)** Disaster risk premium at 5% quantile

**Figure 1.3. Disaster risk premia for 30-day returns at the 5th percentile**. Panel (a) shows the physical and risk-neutral quantile functions over time at $\tau = 0.05$. The physical quantile function is estimated from the quantile regression in (1.2.4). Panel (b) shows the associated disaster risk premium, $Q_{t,\tau} - \widetilde{Q}_{t,\tau}$. Both panels are smoothed using a 30-day moving window. The two shaded bars denote the Great Recession period (Dec 2007 – June 2009) and Covid-19 crisis (Feb 2020 – April 2020).

[12] I find similar results for 60- and 90-day returns.

### 1.3.3 Predicting the Equity Premium

The previous results establish that, in times of heightened market uncertainty, the equity premium is driven more by disaster risk. This observation suggests a strong link between $\mathbb{E}_t\left[R_{m,t\to N}\right] - R_{f,t\to N}$ and the tail of the risk-neutral distribution, which motivates the predictive regression

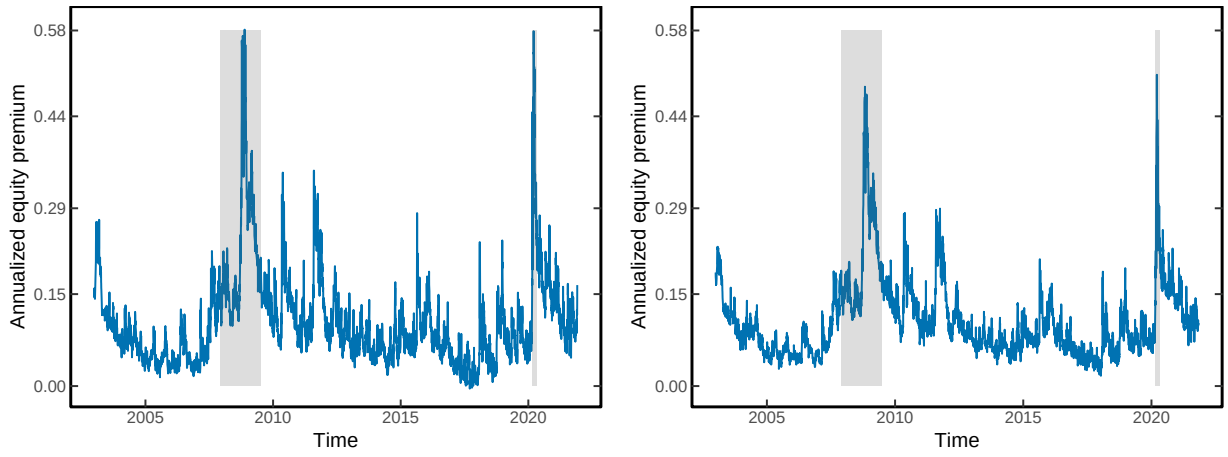
$$R_{m,t\to N} - R_{f,t\to N} = \beta_0 + \beta_1 \widetilde{Q}_{t,\tau} + \varepsilon_{t\to T}, \tag{1.3.2}$$

where $\widetilde{Q}_{t,\tau}$ is evaluated at $\tau = 0.05$.

Table 1.3 shows the results at several return horizons. In all cases, the coefficient is negative, consistent with previous findings that the equity premium increases under market uncertainty. Following Welch and Goyal (2008), the table also reports the out-of-sample $R^2$, denoted by $R^2_{oos}$, which compares the predictions of (1.3.2) to a rolling average of excess returns. Precisely, I estimate (1.3.2) using the sub-sample covering 2003–2012, and fix the estimated parameters to predict excess returns over the out-of-sample period 2013–2021. Encouragingly, $R^2_{oos}$ is always positive and statistically significant according to the Diebold and Mariano (1995) test, thus suggesting that the left-tail of the risk-neutral quantile function outperforms the historical mean benchmark. These values are also substantially higher compared to the $R^2_{oos}$ reported by Welch and Goyal (2008) using various valuation ratios, or Martin (2017) using SVIX.[13]

Figure 1.4 shows the estimated equity premium over time for 30- and 60-day returns. The panels are annualized to make them comparable. Both panels display considerable variation in the equity premium over time and large values during the global financial crisis and Covid-19 crisis. In these periods, Figure 1.4a suggests that the annualized equity premium peaks at 58%, which is substantial relative to more conventional estimates based on dividend-price ratios. On the other hand, the estimates around the 2008 financial crisis are in line with Martin (2017, Figure IV).

[13]The latter is not directly comparable however, since SVIX does not require parameter estimation.




**(a)** 30-day returns

**(b)** 60-day returns

**Figure 1.4. Estimated equity premium.** This figure shows the estimated equity premium based on (1.3.2) for 30-day returns (Panel 1.4a) and 60-day returns (Panel 1.4b). In both cases, the equity premium is converted to annual units. The two shaded bars signify the Great Recession period (Dec 2007 – June 2009) and Covid-19 crisis (Feb 2020 – April 2020).

### 1.3.4 Pricing Kernel Monotonicity and Stochastic Dominance

Besides the equity premium puzzle, the QR estimates in Table 1.2 also provide insights into other asset pricing anomalies, such as pricing kernel monotonicity. Pricing kernel monotonicity refers to the property that $M_{t\to N}(R_{m,t\to N}) \coloneqq \mathbb{E}\left[M_{t\to N}|R_{m,t\to N}\right]$ is a decreasing function of the market return. Asset pricing models that link the SDF to the marginal rate of substitution imply that the pricing kernel is indeed a decreasing function. Empirically, there is suggestive evidence that the pricing kernel is not monotonic, which is puzzling as it contradicts that a representative investor is risk-averse (see Aït-Sahalia and Lo (1998), Jackwerth (2000), Rosenberg and Engle (2002), Bakshi, Madan, and Panayotov (2010), Beare and Schmidt (2016) and Cuesdeanu and Jackwerth (2018)). However, a formal statistical test that can detect violations of monotonicity is challenging as one needs uniform confidence bands for the estimated SDF, which requires tools from empirical process theory (see, e.g., Beare and Schmidt (2016)).

I consider a different approach based on stochastic dominance. Proposition A.1.1 in the Appendix shows that pricing kernel monotonicity implies that the physical distribution is first-order stochastic dominant (FOSD) over the risk-neutral distribution, i.e., $F_t(x) \leq \widetilde{F}_t(x)$ for all $x$. The latter condition can be rephrased as $F_t(\widetilde{Q}_{t,\tau}) \leq \tau$ for all $\tau \in (0,1)$. A violation of stochastic

dominance, and hence pricing kernel monotonicity, is thus implied if there is statistical evidence that $F_t(\widetilde{Q}_{t,\tau}) > \tau$ for a single $\tau$. To investigate this possibility, let[14]

$$\begin{aligned} \text{Hit}_{t\to N} &= \mathbb{1}\left(R_{m,t\to N} < \widetilde{Q}_{t,\tau}\right) - \tau, \\ \overline{\text{Hit}} &= \frac{1}{T}\sum_{t=1}^{T} \text{Hit}_{t\to N}. \end{aligned} \tag{1.3.3}$$

Hence, $\overline{\text{Hit}}$ provides an estimate of $\mathbb{E}(F_t(\widetilde{Q}_{t,\tau}) - \tau)$ which ought to be negative for all $\tau$ under FOSD.[15] The "$\overline{\text{Hit}}$" column in Table 1.2 reports the value of (1.3.3), which is positive for $\tau = 0.95$ at the 30- and 60-day horizon. However, these estimates are not significant at the conventional levels and a violation of FOSD cannot be concluded.

Since $F_t(x) \leq \widetilde{F}_t(x)$ if and only if $Q_{t,\tau} > \widetilde{Q}_{t,\tau}$, it follows that violations of stochastic dominance can also be identified directly from the quantile function. Based on the QR estimates (1.2.4), consider the predicted quantile function $\widehat{Q}_{t,\tau} = \widehat{\beta}_0(\tau) + \widehat{\beta}_1(\tau)\widetilde{Q}_{t,\tau}$. The last column in Table 1.2 displays the time series average of instances where $\widehat{Q}_{t,\tau} > \widetilde{Q}_{t,\tau}$. Broadly speaking, for all horizons, violations of stochastic dominance are infrequent, except far in the right-tail. At $\tau = 0.95$, stochastic dominance is frequently violated, consistent with a non-monotonic pricing kernel.[16] In representative agent models, this result is puzzling as it contradicts the assumption of decreasing marginal utility of wealth (see Proposition A.1.2 in the Appendix).

### 1.3.5 Belief Recovery

A recent literature asks to what extent Arrow prices can be used to learn about the underlying probability distribution of the data, or the subjective probabilities used by investors. Since Arrow prices are confounded by risk aversion, it is impossible to identify the underlying probabilities from Arrow prices alone, unless one imposes additional restrictions (Ross, 2015; Borovička, Hansen, and Scheinkman, 2016; Bakshi, Chabi-Yo, and Gao, 2018; Qin, Linetsky, and Nie, 2018; Jackwerth and Menner, 2020). For example, Ross (2015) uses the Perron-Frobenius theorem to

[14] The $\text{Hit}_{t\to N}$ function was first introduced by Engle and Manganelli (2004) in a different context.

[15] $\overline{\text{Hit}}$ also yields another measure of the difference between $F_t$ and $\widetilde{F}_t$. Consistent with the quantile regression estimates, the Hit statistic shows that $F_t$ and $\widetilde{F}_t$ are similar in the right-tail, but different in the left-tail.

[16] The most significant violations occur during two major financial crises: the 2008 financial crisis and the 2020 Covid-19 crisis.

recover investors' beliefs, which agrees with the underlying physical measure under rational expectations.

Complementary to this insight, the QR estimates in Table 1.2 show that the right-tail of the physical distribution can approximately be recovered from the right-tail of the risk-neutral distribution, which aligns with the investor's belief under rational expectations. In contrast, the left-tail of the physical distribution cannot be recovered even though the risk-neutral quantile serves as a conservative lower bound. In Section 1.6, I propose a more stringent lower bound to recover the left-tail of the physical distribution as well from option data.

## 1.4 QR and Robust Estimation of Disaster Risk

Section 1.3.1 demonstrated that the conditional lognormal assumption is inconsistent with the observed disaster risk premia in the market. At the same time, Figure 1.2a showed that disaster risk models tend to overestimate the magnitude of disaster risk in the data. These conclusions heavily rely on the accuracy of QR in providing estimates of the physical quantile function.

In this section, I compare QR to nonparametric SDF methods for estimating disaster risk. Foreshadowing the results, I show that QR is more robust and argue that the SDF approach tends to overestimate disaster risk. These results help explain the current disagreement about the extent of disaster risk in the data, and provide further support for QR to estimate this risk.

### 1.4.1 QR in the Conditional Lognormal Model

To convey the intuition, it is convenient to work with a discretized version of the Black and Scholes (1973b) model. There is a riskless asset that offers a certain return, $R_{f,t\to N} \equiv R_f = e^{r_f N}$, and a risky asset with return

$$R_{m,t\to N} = \exp([\mu_t - \frac{1}{2}\sigma_t^2]N + \sigma_t\sqrt{N}Z_{t+N}), \tag{1.4.1}$$

where $\mu_t$ represents the conditional mean return, $\sigma_t$ is the conditional volatility, and $Z_{t+N}$ is a random shock that follows a standard normal distribution. In this setup, $M_{t\to N} \coloneqq \exp(-[r_f + \xi_t^2/2]N - \xi_t\sqrt{N}Z_{t+N})$ is a valid SDF with conditional Sharpe ratio

$$\xi_t = \frac{\mu_t - r_f}{\sigma_t}. \tag{1.4.2}$$

Hence, under risk-neutral measure, the conditional distribution of $R_{m,t\to N}$ is given by

$$\log \widetilde{R}_{m,t\to N} \sim \mathcal{N}\left((r_f - \frac{1}{2}\sigma_t^2)N, \sigma_t^2 N\right). \tag{1.4.3}$$

Notice that $\sigma_t$ is implicitly observed from the risk-neutral distribution, but $\mu_t$ is unobserved with mean $\mu \coloneqq \mathbb{E}[\mu_t]$ and variance $\sigma_\mu^2 \coloneqq \mathrm{Var}(\mu_t) < \infty$. The following result characterizes the limiting behavior of the QR estimates (1.2.4) in the lognormal model when the variance of the equity premium is small. A convenient way to model this is by means of a drifting sequence $\sigma_\mu^T \to 0$ as $T \to \infty$, which captures the intuition that the volatility of the equity premium is much smaller than the return volatility.

**Proposition 1.4.1** (QR in Lognormal Model)**.** *In the lognormal model described above with return observations $\{R_{m,t\to N}\}_{t=1}^T$ and risk-neutral quantile functions $\{\widetilde{Q}_{t,\tau}\}_{t=1}^T$, the following hold.*

*(i) Suppose that conditional on time $t$, $\mu_t$ follows a normal distribution $\mu_t \sim \mathcal{N}(\mu, \sigma_\mu^2)$, independent of $\sigma_t$. Let $Q_{t,\tau}(\sigma_t, \sigma_\mu)$ denote the physical quantile function of $R_{m,t\to N}$ conditional on $\sigma_t$ only. Then, for all $\tau \in \mathcal{I} \coloneqq$ a closed subset of $[\varepsilon, 1-\varepsilon]$ for $0 < \varepsilon < 1$, the physical quantile function satisfies*

$$\begin{aligned} Q_{t,\tau}(\sigma_t, \sigma_\mu) &= \exp\left[(\mu - \frac{1}{2}\sigma_t^2)N + \left(\sqrt{\sigma_\mu^2 N^2 + \sigma_t^2 N}\right)\Phi^{-1}(\tau)\right] \\ &= \widetilde{Q}_{t,\tau} e^{(\mu - r_f)N}\left(1 + \mathcal{O}\left(\sigma_\mu N\right)\right), \end{aligned}$$

*where $\Phi^{-1}(\tau)$ denotes the quantile function of the standard normal distribution.*

*(ii) Consider a drifting sequence for $\sigma_\mu$, denoted by $\sigma_\mu^T \to 0$ as $T \to \infty$. Then, under Assumption*

*A.1.4 in the Appendix, the estimated parameters in the quantile regression*

$$\left[\widehat{\beta}_0(\sigma_\mu^T;\tau),\widehat{\beta}_1(\sigma_\mu^T;\tau)\right] = \underset{(\beta_0,\beta_1)\in\mathbb{R}^2}{\arg\min}\sum_{t=1}^{T}\rho_\tau(R_{m,t\to N}-\beta_0-\beta_1\widetilde{Q}_{t,\tau}),$$

*satisfy*

$$\left[\widehat{\beta}_0(\sigma_\mu^T;\tau),\widehat{\beta}_1(\sigma_\mu^T;\tau)\right] = \left[0, e^{(\mu-r_f)N}\right] + o_p(1). \tag{1.4.4}$$

*Furthermore, the quantile forecast based on the QR estimates satisfies*

$$\widehat{\beta}_0(\sigma_\mu^T;\tau) + \widehat{\beta}_1(\sigma_\mu^T;\tau)\widetilde{Q}_{t,\tau} = Q_{t,\tau} + o_p(1). \tag{1.4.5}$$

*Proof.* See Appendix A.1.3. ■

Proposition 1.4.1((i)) shows that the risk-neutral quantile function is a good predictor of $Q_{t,\tau}(\sigma_t;\sigma_\mu)$ when $\sigma_\mu$ is small, and the difference between the two functions is governed by the unconditional equity premium $e^{(\mu-r_f)N}$. In this case, Proposition 1.4.1((ii)) suggests that the QR estimates are almost constant across $\tau$ and close to $[0, e^{(\mu-r_f)N}]$. This result obtains without assuming that $\mu_t$ follows a normal distribution. The wedge between $Q_{t,\tau}(\sigma_t;\sigma_\mu)$ and $\widetilde{Q}_{t,\tau}$ not explained by the equity premium can be attributed to uncertainty about $\mu_t$, which increases the variance of the physical distribution. The assumption that $\sigma_\mu$ is small relative to $\sigma_t$ accords with empirical findings of Martin (2017, Table I), who finds that $2.4\% \leq \sigma_\mu \leq 4.6\%$, whereas $\sigma_t$ hovers around 20%. Unreported simulations show that the approximation in (1.4.4) obtains closely when the model is calibrated to match these stylized facts. As a result, the physical quantile forecast based on the QR estimates in (1.4.5) is also highly accurate.

### 1.4.2 QR versus Nonparametric SDF Estimation

Because of the availability of closed-form expressions in the lognormal model, it is instructive to compare the QR approach to alternative methods for estimating the physical distribution. Since the SDF represents the Radon–Nikodym derivative of the risk-neutral and physical measures, it is possible to obtain the physical quantile function from the estimated SDF. There is a substantial literature on how to estimate the SDF in a forward-looking manner (see Remark 1). For this

comparison, I consider the state-of-the-art SDF estimator proposed by Cuesdeanu and Jackwerth (2018) (CJ).

After some algebra, the SDF in the Black-Scholes model can be expressed as a function of the market return:

$$M_{t\to N} = \exp\left(-\frac{N}{2}\left[\mu_t + r_f + \frac{r_f^2 - \mu_t^2}{\sigma_t^2}\right]\right)(R_{m,t\to N})^{-\xi_t/\sigma_t}, \tag{1.4.6}$$

where $\xi_t$ is the conditional Sharpe ratio (1.4.2). CJ project the unobserved SDF in (1.4.6) on the market return and estimate an SDF of the form

$$\widehat{M}_{t\to N} = C_t g(R_{m,t\to N}),$$

where $C_t$ is a time-varying constant, and $g(\cdot)$ is an unknown function that can be estimated by choosing a sieve basis. Since $g(\cdot)$ is *time-homogeneous*, it is evident that changes in the shape of the true SDF in (1.4.6) are not captured by the estimated SDF. Specifically, in times when the Sharpe ratio is high, the physical and risk-neutral measures exhibit more distinct differences, as the true SDF becomes steeper. Because the estimated SDF does not account for these shape changes, it leads to a severe underestimation of the physical quantile function in the left-tail. Proposition 1.4.1 demonstrates that the QR approach does not suffer from this limitation.

To illustrate this discussion, I simulate returns from the lognormal model and estimate the physical quantile function at the 5th percentile using QR and the SDF estimate of CJ. Since the conditional (physical) quantile function is known analytically in the lognormal model, I evaluate the forecast accuracy using the quantile error ratio, $\widehat{Q}_{t,\tau}/Q_{t,\tau}$, where $\widehat{Q}_{t,\tau}$ is the predicted physical quantile based on QR or the SDF estimate. Panel 1.5a displays the empirical density of error ratios obtained by simulating 1,000 returns. In line with Proposition 1.4.1((ii)), the error ratio corresponding to QR is symmetric and closely centered around one. In contrast, when the physical quantile is inferred from the estimated SDF, the error density is biased and exhibits fat tails since the estimated SDF cannot change shape. Consequently, in periods of high disaster risk premia, the

CJ method severely *underestimates* $Q_{t,\tau}$.

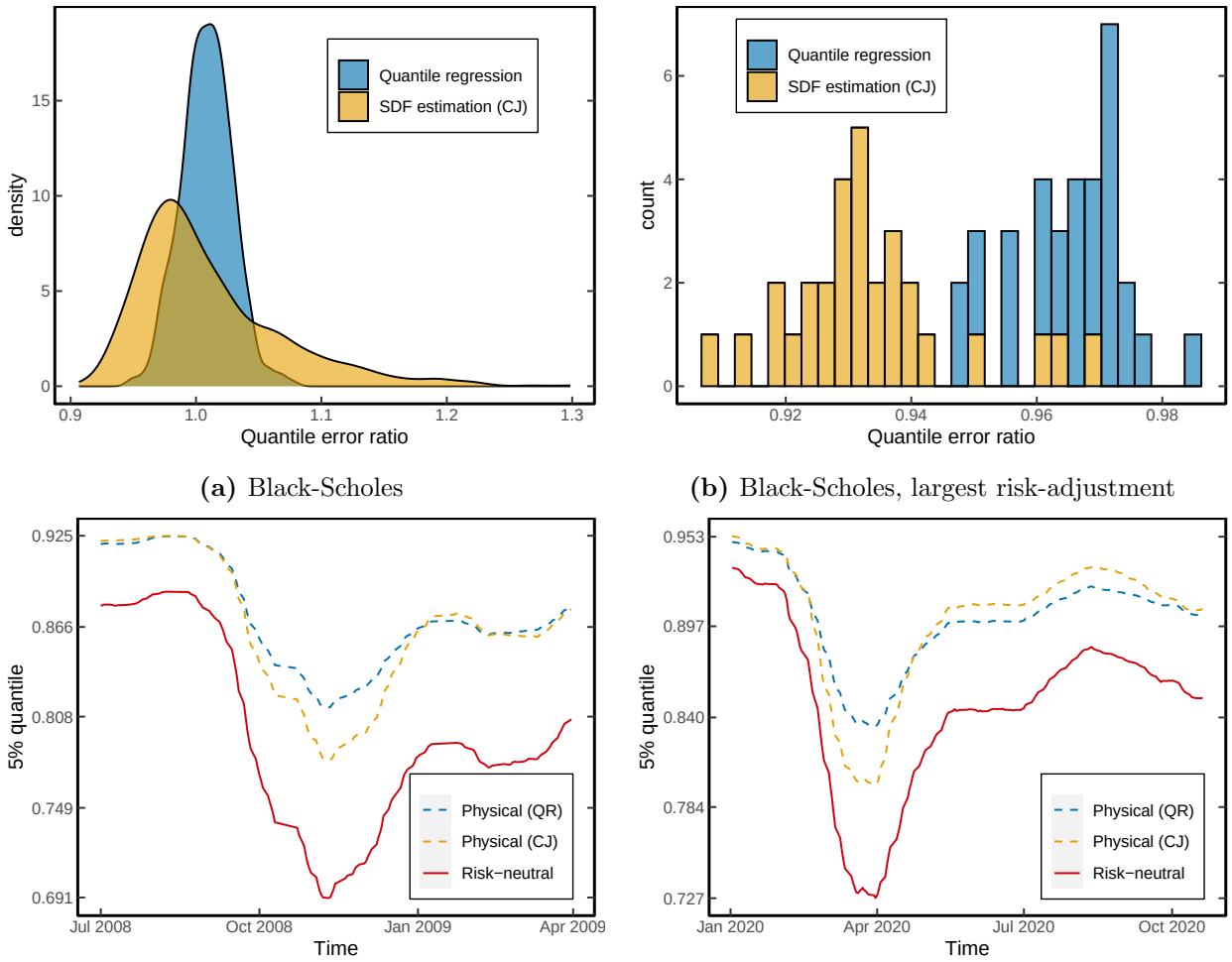
Panel 1.5b presents the histogram of error ratios conditioned on the 30 largest values of $Q_{t,\tau} - \widetilde{Q}_{t,\tau}$, clearly illustrating the downward bias in the SDF method. On average, the predicted physical quantile is 7% lower than its actual value when disaster risk premia are high. The QR approach is less affected by this bias because it can capture changes in the shape of the SDF. The computational benefits of QR are also notable, as the computation of the physical quantile forecast takes less than a second. On the other hand, the SDF method requires more than 20 minutes to complete the same task.[17]

The bottom panels of Figure 1.5 further illustrate the difference between QR and CJ using the 30-day return data from Section 1.2.3, particularly during the 2008 financial crisis and the Covid-19 crisis. At the height of both crises, both methods predict increases in disaster risk premia as $\widehat{Q}_{t,\tau} - \widetilde{Q}_{t,\tau}$ rises significantly. As mentioned earlier, the SDF approach implies that disaster risk premia increase less relative to the QR approach, as the shape of the SDF remains constant over time. However, it is worth noting that while the QR approach performs well when returns are conditional lognormal, Appendix A.2.2 demonstrates that the quantile forecasts based on QR contradict (1.4.5), casting further doubt on the validity of the conditional lognormal assumption in the data.

## 1.5 Disaster Risk and SDF Volatility

Section 1.3.1 demonstrated that the physical and risk-neutral distributions locally differ most in the left-tail. In this section, I show that these local differences imply that the SDF must be highly volatile; an observation that is closely related to the Hansen and Jagannathan (1991) bound. Furthermore, I use this insight to argue that the left-tail of the physical distribution cannot be too predictable, which clarifies the low explanatory power in Table 1.2.

[17] Moreover, the optimization problem required to implement the sieve estimation did not converge, as the maximum number of iterations were exceeded. This problem occurs due to the large number of parameters to estimate, and because the optimization problem is not convex (see Remark 1).




**(a)** Black-Scholes

**(b)** Black-Scholes, largest risk-adjustment

**(c)** 2008 financial crisis

**(d)** Covid-19 crisis

**Figure 1.5. Disaster risk premia at the 5th percentile**. Panel (a) shows the quantile error ratio, $\widehat{Q}_{t,\tau}/Q_{t,\tau}$, in a conditional Black and Scholes (1973b) model for $\tau = 0.05$, where $\widehat{Q}_{t,\tau}$ is the predicted physical quantile based on QR or the SDF estimate of CJ. Volatility is generated according to an AR(1)-model with mean value 0.2, standard deviation 0.03 and a persistency of 0.9. The mean of the physical distribution follows $\mu_t \sim \mathcal{N}(0.07, 0.02^2)$, the risk-free rate equals $r_f = 0.01$, the time horizon is one-year, and the number of observations is 1,000. Panel (b) shows the histogram of error ratios conditioned the 30 events for which $Q_{t,\tau} - \widetilde{Q}_{t,\tau}$ is maximal. The bottom panels illustrate the difference between the predicted physical quantile obtained from QR, and the predicted quantile coming from the SDF estimate of CJ, during the global financial crisis and the Covid-19 crisis. Both estimates are based on 30-day returns, using the data from Section 1.2.3. The bottom panels are smoothed using a 30-day rolling window.

### 1.5.1 A Bound on the SDF Volatility

For ease of notation, I define $\phi_t(\tau) := F_t(\widetilde{Q}_{t,\tau})$, which can be interpreted as the ordinal dominance curve of the measures $\mathbb{P}_t$ and $\widetilde{\mathbb{P}}_t$ (Hsieh and Turnbull, 1996). Furthermore, let

$$\aleph_t^+ := \{M_{t\to N} : M_{t\to N} \geq 0 \text{ and } \mathbb{E}_t\left[M_{t\to N} R_{m,t\to N}\right] = 1\},$$

which is the space of all nonnegative conditional SDFs. The volatility bound on the SDF can now be stated as follows.

**Proposition 1.5.1** (Distribution bound)**.** *Assume no-arbitrage, then for any* $M_{t\to N} \in \aleph_t^+$*, we have*

$$\frac{\sigma_t(M_{t\to N})}{\mathbb{E}_t\left[M_{t\to N}\right]} \geq \frac{|\tau - \phi_t(\tau)|}{\sqrt{\phi_t(\tau)(1-\phi_t(\tau))}} \qquad \forall \tau \in (0,1). \tag{1.5.1}$$

*If a risk-free asset exists, then* $\mathbb{E}_t\left[M_{t\to N}\right] = 1/R_{f,t\to N}$ *and* (1.5.1) *simplifies to*

$$\sigma_t(M_{t\to N}) \geq \frac{1}{R_{f,t\to N}} \frac{|\tau - \phi_t(\tau)|}{\sqrt{\phi_t(\tau)(1-\phi_t(\tau))}} \qquad \forall \tau \in (0,1).$$

*The bound can be further rewritten in terms of the conditional CDFs only*

$$\sigma_t(M_{t\to N}) = \frac{1}{R_{f,t\to N}} \frac{\left|\widetilde{F}_t(x) - F_t(x)\right|}{F_t(x)(1-F_t(x))} \qquad \forall x \in (0,\infty). \tag{1.5.2}$$

*Proof.* See Appendix A.1.4. ■

If $\mathbb{P}_t = \widetilde{\mathbb{P}}_t$, agents are risk-neutral and the dominance curve evaluates to $\phi_t(\tau) = \tau$. In that case the distribution bound degenerates to zero. Proposition 1.5.1 makes precise the sense in which any local difference between the physical and risk-neutral distribution leads to a volatile SDF. Compare this to the classical Hansen and Jagannathan (1991) (HJ) bound:

$$\sigma_t(M_{t\to N}) \geq \frac{1}{R_{f,t\to N}} \frac{|\mathbb{E}_t\left[R_{m,t\to N}\right] - R_{f,t\to N}|}{\sigma_t(R_{m,t\to N})}. \tag{1.5.3}$$

The lower bound in (1.5.3) shows that any excess return leads to a volatile SDF. Essentially, (1.5.3) uses three sources of information: (i) the mean of the physical distribution (ii) the mean of the risk-neutral distribution (iii) the variance of the physical distribution. The lower bound in (1.5.3) is also a global measure of distance between $\mathbb{P}_t$ and $\widetilde{\mathbb{P}}_t$, since the mean and volatility are averages across the whole distribution.

In contrast, the bound in (1.5.2) compares the physical and risk-neutral distribution at every point $x$, which is a *local* measure of distance between $\mathbb{P}_t$ and $\widetilde{\mathbb{P}}_t$. To clarify this local interpretation,

consider the following decomposition of the (scaled) equity premium

$$\begin{aligned}\frac{\mathbb{E}_t\left[R_{m,t\to N}\right]-R_{f,t\to N}}{R_{f,t\to N}} &= -\,\mathbb{COV}_{\mathrm{t}}[R_{m,t\to N},M_{t\to N}]\\ &= \int_0^\infty \mathbb{COV}_{\mathrm{t}}[\mathbb{1}\left(R_{m,t\to N}\leq x\right),M_{t\to N}]\,\mathrm{d}x, \end{aligned} \tag{1.5.4}$$

where the first equation follows since the SDF prices the market return (1.2.1), and the second equation is a consequence of Hoeffding's identity (see Lemma A.1.5.). Equation (1.5.4) shows that $\mathbb{COV}_{\mathrm{t}}[\mathbb{1}\left(R_{m,t\to N}\leq x\right),M_{t\to N}]$ locally measures the dependence between the SDF and market return. In other words, it quantifies how the SDF's variability relates to the market return's variability at different quantiles.

To explain the equity premium and disaster risk premia, the SDF must exhibit sufficient variability. Since the distribution bound can be derived from applying the Cauchy-Schwarz inequality to $\mathbb{COV}_{\mathrm{t}}[\mathbb{1}\left(R_{m,t\to N}\leq x\right),M_{t\to N}]$, it is expected to yield sharper bounds on the SDF volatility than the HJ bound if, for example, there is high tail dependence between the SDF and market return such as in the disaster risk model.[18]

### 1.5.2 Quantile Predictability in the Left-Tail

The bound presented in Proposition 1.5.1 sheds light on the seemingly "low" explanatory power observed in the left-tail quantile regressions in Table 1.2. For tractability, it is more convenient to show this for CDFs instead of quantile functions, but the intuition remains the same. Specifically, suppose one could predict $F_t(x)$ at some $x$ in the left-tail, then this prediction can be exploited by going short in an asset that pays $\mathbb{1}\left(R_{m,t\to N}\leq x\right)$. The profit and risk associated to

[18] See McNeil, Frey, and Embrechts (2015, Chapter 7.2.4) for a formal definition of tail dependence.

this investment are, respectively

$$\begin{aligned}
&\frac{1}{R_{f,t\to N}}\left(\widetilde{\mathbb{E}}_t\left[\mathbb{1}\left(R_{m,t\to N}\leq x\right)\right]-\mathbb{E}_t\left[\mathbb{1}\left(R_{m,t\to N}\leq x\right)\right]\right)\\
&\qquad=\frac{1}{R_{f,t\to N}}\left(\widetilde{F}_t(x)-F_t(x)\right), \qquad (1.5.5)\\
&\sigma_t(\mathbb{1}\left(R_{m,t\to N}\leq x\right))=\sqrt{F_t(x)(1-F_t(x))}.
\end{aligned}$$

Although such binary state payoffs do not exist in reality, they can be replicated closely by a portfolio of put options. In consequence, high predictability of $F_t(x)$ in the left-tail would render too good a Sharpe ratio; a near-arbitrage opportunity. Following the reasoning in Ross (2005, Chapter 5), a crude upper bound on the SDF volatility imposes limitations on the degree of predictability in the left-tail by the distribution bound in Proposition 1.5.1. This argument breaks down in the right-tail since (1.5.5) is roughly zero, and high predictability would not imply counterfactually high SDF volatility.

### 1.5.3 Distribution Bound in Asset Pricing Models

The estimated Gini coefficients in Section 1.3.1 demonstrate that conditional disaster risk is a pervasive feature of the data. This section complements those findings using the unconditional version of the distribution bound in Proposition 1.5.1:

$$\frac{\sigma(M)}{\mathbb{E}\left[M\right]}\geq\frac{\tau-\phi(\tau)}{\sqrt{\phi(\tau)(1-\phi(\tau))}}, \qquad (1.5.6)$$

where $\sigma(M)$ represents the unconditional SDF volatility, and $\phi(\tau)=F(\widetilde{Q}_\tau)$. In this context, $F(\cdot)$ denotes the unconditional physical CDF, and $\widetilde{Q}_\tau$ is the unconditional risk-neutral quantile function of the market return. The main benefit of using unconditional distributions is that they can be estimated without running the risk-neutral quantile regressions. Moreover, the bound in (1.5.6) only requires the estimation of distribution functions, whereas existing approaches typically use unconditional density functions to estimate disaster risk (see, e.g. Beason and Schreindorfer (2022)).

The subsequent examples demonstrate that the HJ bound is always stronger than the distribution bound in models that do not embed a source of disaster risk. In contrast, models that

incorporate disaster risk can generate distribution bounds that exceed the HJ bound in the left-tail. Since I use unconditional distributions, the time subscripts will be omitted from the notation.

**Example 1.5.1** (CAPM)**.** The Capital Asset Pricing Model (CAPM) specifies the SDF as

$$M = \alpha - \beta R_m,$$

where $R_m$ denotes the return on the market portfolio. In this case $M \notin \aleph^+$, since the SDF can become negative. However, this probability is very small over short time horizons or we can think of $M$ as an approximation to $M^* \coloneqq \max(0, M) \in \aleph^+$. Since the HJ bound is derived by applying the Cauchy-Schwarz inequality to $\mathbb{COV}(R_m, M)$, the inequality binds if $M$ is a linear combination of $R_m$. Hence, under CAPM, the HJ bound is (weakly) stronger than the distribution bound regardless of the distribution of $R_m$.

**Example 1.5.2** (Joint normality)**.** Suppose that $M$ and $R_m$ are jointly normally distributed and denote the mean and variance of $R_m$ by $\mu_R$ and $\sigma_R^2$ respectively. The normality assumption violates no-arbitrage since $M$ can be negative, but could be defended as an approximation over short time horizons when the variance is small (see Example 1.5.3). In Appendix A.1.5, I prove that

$$\left|\mathbb{COV}\left(\mathbb{1}\left(R_m \leq \widetilde{Q}_\tau\right), M\right)\right| = f_R(\widetilde{Q}_\tau)\left|\mathbb{COV}(R_m, M)\right|, \tag{1.5.7}$$

where $f_R(\cdot)$ is the marginal density of $R_m$.[19] This identity gives an explicit expression for the weighting factor in Hoeffding's identity (1.5.4). In Appendix A.1.5, I also derive an explicit expression for the relative efficiency between the distribution and HJ bound, defined by

$$\frac{\text{HJ bound}}{\text{distribution bound}} = \frac{\sqrt{\phi(\tau)(1-\phi(\tau))}}{\sigma_R f_R(\widetilde{Q}_\tau)}. \tag{1.5.8}$$

To see that the HJ bound is always stronger than the distribution bound, minimize (1.5.8) with respect to $\tau$. Appendix A.1.6 shows that the minimizer $\tau^*$ satisfies $\widetilde{Q}_{\tau^*} = \mu_R$. For this choice,

[19] Notice that this is the marginal density under physical measure $\mathbb{P}$.

$\phi(\tau^*) = \mathbb{P}(R_m \leq \widetilde{Q}_{\tau^*}) = 1/2$ and $f_R(\widetilde{Q}_{\tau^*}) = 1/\sqrt{2\pi\sigma_R^2}$. Therefore, (1.5.8) can be bounded by

$$\frac{\sqrt{\phi(\tau)(1-\phi(\tau))}}{\sigma_R f(\widetilde{Q}_\tau)} \geq \frac{\sqrt{2\pi}}{2} \approx 1.25.$$

Hence, the HJ bound is always stronger in a model where the SDF and market return are jointly normal.

**Example 1.5.3** (Joint lognormality)**.** Let $Z_R$ and $Z_M$ be standard normal random variables with correlation $\rho$ and consider the specification

$$\begin{aligned} R_m &= e^{(\mu_R - \frac{\sigma_R^2}{2})N + \sigma_R\sqrt{N}Z_R} \\ M &= e^{-(r_f + \frac{\sigma_M^2}{2})N + \sigma_M\sqrt{N}Z_M}, \end{aligned}$$

where $N$ governs the time scale in annual units. Simple algebra shows that the no-arbitrage condition, $\mathbb{E}[MR_m] = 1$, is satisfied when $\mu_R - r_f = -\rho\sigma_R\sigma_M$. It is difficult to find an analytical solution for the relative efficiency between the HJ and distribution bound in this case, but linearization leads to a closed form expression which is quite accurate in simulations. The details are described in Appendix A.1.7, where I show that

$$\min_{\tau\in(0,1)} \frac{\text{HJ bound}}{\text{distribution bound}} \approx \frac{1}{2}\sqrt{\frac{2\pi\sigma_R^2 N}{\exp(\sigma_R^2 N) - 1}}. \tag{1.5.9}$$

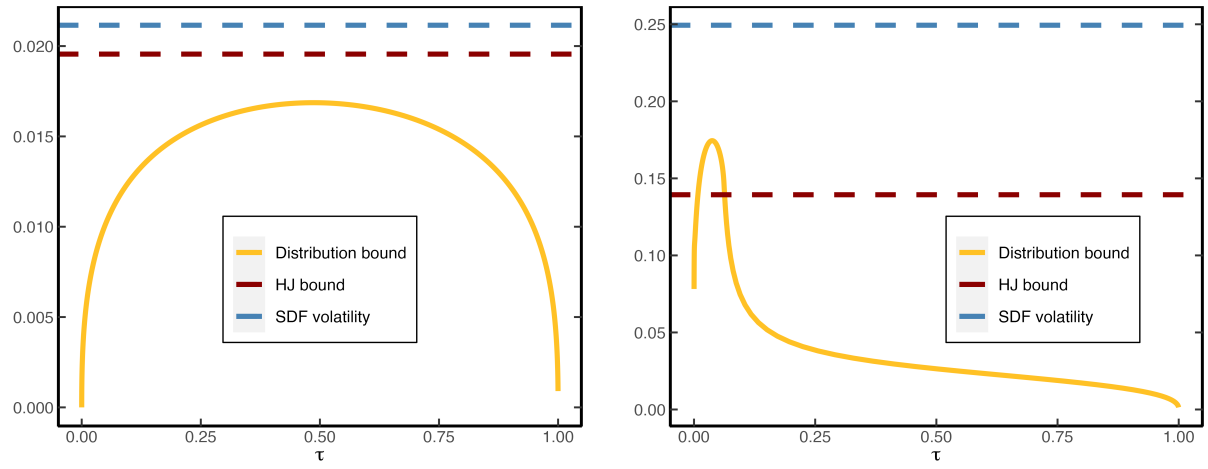
This expression is independent of $\mu_R$. An application of l'Hôspital's rule reveals that the relative efficiency converges to $\sqrt{2\pi}/2$ if $N \to 0^+$.[20] The ratio in (1.5.9) is less than 1 if $\sigma_R \geq 0.92$ and $N = 1$. Since the annualized market return volatility is about 16%, the HJ bound is stronger than the distribution bound under any reasonable parameterization if the SDF and market return are lognormal.

**Example 1.2.1** (Continued)**.** The disaster risk model discussed in Section 1.2.1 is calibrated according to the results in Backus, Chernov, and Martin (2011, Table II). The market return in this model is considered as a levered claim on consumption growth, i.e. an asset that pays dividends

[20] This is the same relative efficiency in Example 1.5.2, which is unsurprising as the linearization becomes exact in the limit as $N \to 0^+$.

proportional to $G^{\lambda}_{t\to N}$. Here $\lambda$ governs the variability of the claim to equity. I convert the model implied volatility bounds to monthly units, to facilitate the comparison with the empirical bounds obtained in Section 1.5.5.

The distribution bound, HJ bound and SDF volatility are depicted in Panels 1.6a (without jumps) and 1.6b (with jumps). Consistent with Example 1.5.3, the distribution bound in the model without jumps never exceeds the HJ bound because both the market return and SDF follow lognormal distributions. The distribution bound with jumps has a sharp peak at $\tau = 0.037$, after which it steadily decreases. Interestingly, there is a range of $\tau$ values for which the distribution bound is stronger than the HJ bound.[21] This result can be understood from the physical and risk-neutral quantile functions in Figure 1.1b. The risk-neutral quantile function displays a heavy left-tail, owing to the implied disaster risk embedded in the SDF. Consequently, it is extremely profitable to sell digital put options which pay out in case of a disaster. These put options must have high Sharpe ratios as their prices are high (insurance against disaster risk), but the actual probability of a disaster event occurring is low enough that the risk associated with selling such insurance is limited.




**(a)** Without jumps **(b)** With jumps

**Figure 1.6. HJ and distribution bound in disaster risk model without and with jumps.** Panels (a) and (b) show the HJ and distribution bound for the disaster risk model (Example 1.2.1) without and with jumps, respectively. The bounds and true SDF volatility are reported in monthly units. Parameters are calibrated according to Backus, Chernov, and Martin (2011, Table II).

[21] In Appendix A.1.8, I show that the distribution bound can also exceed the HJ bound when returns follow the Pareto distribution.

### 1.5.4 Data and Empirical Estimation of the Distribution Bound

To further illustrate the presence of disaster risk in the data, I estimate the distribution bound (1.5.1) empirically, using the same 30-day S&P500 returns as discussed in Section 1.2.3. However, in this case, I use non-overlapping returns that cover the period 1996–2021.[22] These returns are sampled at the middle of each month, resulting in a total of 312 observations. Over this period, the Sharpe ratio is 13%, and the HJ bound therefore implies that the monthly SDF is quite volatile.

The distribution bound consists of three unknowns that need to be estimated: (i) the physical distribution ($F$); (ii) the risk-neutral quantile function ($\widetilde{Q}_\tau$), and; (iii) the risk-free rate ($R_f$). To estimate the unconditional risk-free rate, denoted by $\widehat{R}_f$, I rely on the historical average of monthly interest rates. Next, to obtain an estimate of the physical distribution, I employ a kernel (CDF) estimator, given by:

$$\widehat{F}(x) \coloneqq \frac{1}{T}\sum_{t=1}^{T} \Phi\left(\frac{x - R_{m,t\to N}}{h}\right), \tag{1.5.10}$$

where $\Phi(\cdot)$ is the Epanechnikov kernel and $h$ is the bandwidth determined by cross-validation. This choice of estimator ensures that the distribution bound is a smooth function of $\tau$, which reduces the impact of outliers relative to the discontinuous empirical CDF.

Finally, I apply the procedure outlined in Section 1.2.3 to estimate $\widetilde{F}_t$ (the conditional risk-neutral CDF). Subsequently, I average the conditional distributions to estimate the unconditional CDF:

$$\widehat{\widetilde{F}}(x) \coloneqq \frac{1}{T}\sum_{t=1}^{T} \widetilde{F}_t(x).$$

Under appropriate assumptions about the distribution of returns, $\widehat{\widetilde{F}}$ converges to $\widetilde{F}$ as $T \to \infty$. An estimate of the unconditional risk-neutral quantile function can then be obtained from

$$\widehat{\widetilde{Q}}(\tau) \coloneqq \inf\left\{x \in \mathbb{R} : \tau \leq \widehat{\widetilde{F}}(x)\right\}. \tag{1.5.11}$$

Finally, based on the physical CDF (1.5.10) and risk-neutral quantile function (1.5.11), I estimate

[22] I use non-overlapping returns in this section to facilitate testing and to make the results comparable to other nonparametric bounds, which are typically estimated based on non-overlapping returns (see e.g. Liu (2021)).

the distribution bound by

$$\widehat{\theta}(\tau) \coloneqq \frac{\left|\tau - \widehat{\phi}(\tau)\right|}{\sqrt{\widehat{\phi}(\tau)(1-\widehat{\phi}(\tau))}\widehat{R}_f}, \qquad \tau \in [\varepsilon, 1-\varepsilon] \subseteq (0,1), \tag{1.5.12}$$

where $\widehat{\phi}(\tau) \coloneqq \widehat{F}(\widehat{\widetilde{Q}}(\tau))$ is the estimated ordinal dominance curve and $\varepsilon$ is a small positive number.

### 1.5.5 Unconditional Evidence of Disaster Risk

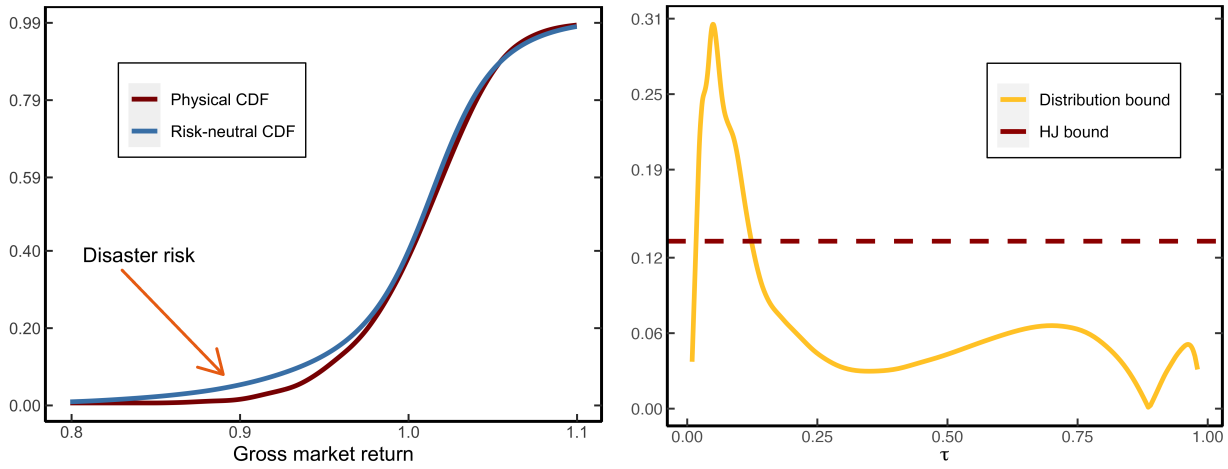
Figure 1.7a illustrates the estimated physical and risk-neutral measures, which differ most in the left-tail. The distribution bound shows that this difference leads to a volatile SDF, which is shown in Figure 1.7b. The lower bound on the SDF volatility implied by the distribution bound is much stronger than the HJ bound in the left-tail. This finding aligns with empirical evidence documenting that high Sharpe ratios can be attained by selling out-of-the money put options (see Broadie, Chernov, and Johannes (2009) and the references therein). The supremum of the distribution bound occurs around the 5th percentile, implying that the monthly SDF volatility must exceed 31%. This value is more than twice the level indicated by the sample HJ bound. Moreover, the shape of the distribution bound is quite similar to the distribution bound implied by the disaster risk model in Figure 1.6b.[23]

The graphical evidence suggests that the distribution bound renders a stronger bound on the SDF volatility than the HJ bound. To test this hypothesis more formally, I fix a priori the probability level at 0.037 ($\tau = 0.037$), which renders the sharpest bound on the SDF volatility in the disaster risk model (Example 1.2.1). At this probability level, the distribution bound is 26% in the data, which is roughly double the level implied by the HJ bound.

To see whether this difference is statistically significant, I consider the following test statistic

$$\mathcal{T} \coloneqq \widehat{\theta}(0.037) - \frac{\left|\bar{R}_m - \widehat{R}_f\right|}{\widehat{\sigma}\widehat{R}_f}. \tag{1.5.13}$$

[23]The non-monotonicity in the right-tail of the distribution bound occurs because $\widetilde{F}(x) > F(x)$, for $x$ large enough. That is, the physical distribution does not first-order stochastically dominates the risk-neutral distribution. This result is consistent with the positive $\overline{\text{Hit}}$ estimates in Table 1.2.




**(a)** **(b)**

**Figure 1.7. Physical/risk-neutral CDF and distribution bound for monthly S&P500 returns**. Panel (a) shows the unconditional phyisical and risk-neutral CDF for monthly S&P500 returns, over the period 1996-2021. Panel (b) shows the distribution bound as function of $\tau$, together with the HJ bound.

The first term on the right denotes the estimated distribution bound (1.5.12) evaluated at the 3.7th percentile, using the entire time series of returns $\{R_{m,t\to N}\}$. The second term denotes the estimated HJ bound, using $\bar{R}_m$ and $\widehat{\sigma}$ as the respective sample mean and standard deviation of $\{R_{m,t\to N}\}$. A value of $\mathcal{T} > 0$ indicates that the distribution bound is stronger than the HJ bound. To test this restriction, consider the null and alternative hypothesis:

$$H_0 : \mathcal{T} \leq 0 \tag{1.5.14}$$
$$H_1 : \mathcal{T} > 0.$$

Since the distribution of (1.5.13) is difficult to characterize, I use stationary bootstrap to approximate the $p$-value under the null hypothesis. The stationary bootstrap is used to generate time indices from which we recreate (with replacement) bootstrapped returns $\{R^{\star}_{m,t\to N}\}$ (Politis and Romano, 1994). The same bootstrapped time indices are used to re-estimate the physical CDF and risk-neutral quantile function. I repeat the bootstrap exercise 100,000 times and for each bootstrap sample, I calculate the test statistic $\mathcal{T}^{\star}$. Finally, the empirical $p$-value is obtained as the fraction of times $\mathcal{T}^{\star} \leq 0$. The last column in Table 1.4 shows that the $p$-value is 7.5%, which provides preliminary evidence that the distribution bound significantly exceeds the HJ bound in the left-tail.

*Remark* 3. When the HJ bound is stronger than the distribution bound, many of the bootstrap samples may not include disaster shocks. Over the entire sample period, there are only two instances where returns were less than -20%: in September 2008 and February 2020. When considering bootstrap samples that include both of these months, the $p$-value is only 3.6%. In contrast, the $p$-value increases to 22% for bootstrap samples that exclude these months. These findings underscore the sensitivity of the test to the presence of disaster shocks. Overall, the results suggest that, unconditionally, the SDF needs to be highly volatile to be consistent with local differences between the physical and risk-neutral measure in the left-tail.

## 1.6 A Model-Free Lower Bound on Disaster Risk Premia

The previous findings indicate that the risk-neutral quantile function is not a good approximation of the physical quantile function in the left-tail. In this section, I derive a lower bound on disaster risk premia observed from option prices. This lower bound does not require parameter estimation and relaxes the assumption of a time-homogeneous linear relation between the physical and risk-neutral quantiles in (1.2.3).

### 1.6.1 Approximating the Quantile Difference

To analyze the difference between $Q_{t,\tau}$ and $\widetilde{Q}_{t,\tau}$, I use some elementary tools from functional analysis. The quantile function can be regarded as a map $\varphi$ between normed spaces, taking as input a distribution function and returning the quantile function: $\varphi(F_t) = F_t^{-1} = Q_{t,\tau}$. Expanding $\varphi$ around the observed risk-neutral CDF yields

$$Q_{t,\tau} - \widetilde{Q}_{t,\tau} = \varphi(F_t) - \varphi(\widetilde{F}_t) = \varphi'_{\widetilde{F}_t}(F_t - \widetilde{F}_t) + o\left(\left\|F_t - \widetilde{F}_t\right\|\right), \tag{1.6.1}$$

where $\|\cdot\|$ is a norm on a suitable linear space[24] and $\varphi'_{\widetilde{F}_t}(F_t - \widetilde{F}_t)$ is the Gâteaux derivative of $\varphi$ at $\widetilde{F}_t$ in the direction of $F_t$:

$$\begin{aligned}\varphi'_{\widetilde{F}_t}(F_t - \widetilde{F}_t) &\coloneqq \lim_{\lambda \downarrow 0} \frac{\varphi\left[(1-\lambda)\widetilde{F}_t + \lambda F_t\right]}{\lambda} \\ &= \frac{\partial}{\partial \lambda} \varphi\left((1-\lambda)\widetilde{F}_t + \lambda F\right)\Big|_{\lambda=0}. \end{aligned} \tag{1.6.2}$$

Heuristically, the Gâteaux derivative can be thought of as measuring the change in the quantile function when the risk-neutral distribution is moved in the direction of the physical distribution. Appendix A.1.9 shows that the Gâteaux derivative is given by

$$\varphi'_{\widetilde{F}_t}(F_t - \widetilde{F}_t) = \frac{\tau - F_t(\widetilde{Q}_{t,\tau})}{\widetilde{f}_t(\widetilde{Q}_{t,\tau})} = \frac{\tau - \phi_t(\tau)}{\widetilde{f}_t(\widetilde{Q}_{t,\tau})}, \tag{1.6.3}$$

where $\phi_t(\tau) = F_t(\widetilde{Q}_{t,\tau})$ denotes the conditional ordinal dominance curve. I proceed under the working hypothesis that the remainder term in (1.6.1) is "small" in the sup-norm, $\|g\|_\infty = \sup_x |g(x)|$.

**Assumption 1.6.1.** *The remainder term in* (1.6.1) *can be neglected.*

*Remark* 4. The assumption implies that the first order approximation in (1.6.1) is accurate. The condition that $||F_t - \widetilde{F}_t||_\infty$ is small can be understood as excluding near-arbitrage opportunities, since the distribution bound in Proposition 1.5.1 shows that substantial pointwise differences between $F_t(\cdot)$ and $\widetilde{F}_t(\cdot)$ lead to a very volatile SDF.

I combine (1.6.1) and (1.6.3) in conjunction with Assumption 1.6.1 to obtain the approximation

$$Q_{t,\tau} - \widetilde{Q}_{t,\tau} \approx \underbrace{\frac{\tau - F_t(\widetilde{Q}_{t,\tau})}{\widetilde{f}_t(\widetilde{Q}_{t,\tau})}}_{\text{risk-adjustment}}. \tag{1.6.4}$$

The second term on the right can be thought of as a risk-adjustment term to capture the unobserved wedge between $Q_{t,\tau}$ and $\widetilde{Q}_{t,\tau}$. The approximation in (1.6.4) contains the terms $\widetilde{Q}_{t,\tau}$ and $\widetilde{f}_t(\widetilde{Q}_{t,\tau})$, which are directly observed at time $t$ using the Breeden and Litzenberger (1978) formula

[24] Formally, the space can be defined as $\{\Delta : \Delta = c(F - G), F, G \in \mathbb{D}, c \in \mathbb{R}\}$ and $\mathbb{D}$ is the space of distribution functions (Serfling, 2009). See van der Vaart (2000, Section 20.1) and Serfling (2009, p. 217) for further details about the approximation.

in (1.2.7). However, $F_t(\cdot)$ is unknown and hence (1.6.4) cannot be used directly to approximate $Q_{t,\tau}$.

### 1.6.2 A Lower Bound on Disaster Risk Premia

To make further progress, I show that the numerator term, $\tau - F_t(\widetilde{Q}_{t,\tau})$, can be lower bounded with option data under economically motivated constraints. This bound, combined with the approximation in (1.6.4), will then imply a lower bound on disaster risk premia.

I start from the observation that the SDF in representative agent models can be expressed as a function of the market return (Chabi-Yo and Loudis, 2020):

$$\frac{\mathbb{E}_t\left[M_{t\to N}\right]}{M_{t\to N}} = \frac{\frac{u'(W_t x_0)}{u'(W_t x)}}{\widetilde{\mathbb{E}}_t\left[\frac{u'(W_t x_0)}{u'(W_t x)}\right]} \qquad \text{with } x = R_{m,t\to N} \text{ and } x_0 = R_{f,t\to N}, \tag{1.6.5}$$

where $W_t$ is the agent's wealth at time $t$ and $u(x)$ represents the agent's utility function. Define

$$\zeta(x) \coloneqq \frac{u'(W_t R_{f,t\to N})}{u'(W_t x)} \quad \text{and} \quad \theta_k = \frac{1}{k!}\left(\frac{\partial^k \zeta(x)}{\partial x^k}\right)_{x=R_{f,t\to N}}. \tag{1.6.6}$$

Notice that $\zeta(\cdot)$ is simply the inverse of the intertemporal marginal rate of substitution (IMRS) and $\theta_k$ are the coefficients of its Taylor expansion around $R_{f,t\to N}$. I make the following assumptions about the market return and the IMRS of the representative agent.

**Assumption 1.6.2.** *In the representative agent model, it holds that (i)* $\widetilde{\mathbb{E}}_t\left[R^3_{m,t\to N}\right] < \infty$*; and (ii)* $\zeta^{(4)}(x) \leq 0$.

Assumption 1.6.2(i) allows for fat tails in the risk-neutral distribution as long as the third moment exists. This assumption relaxes the implicit assumption made by Chabi-Yo and Loudis (2020) that infinitely many moments exist. Figure A.8 in the Appendix illustrates that the risk-neutral distribution frequently exhibits a finite number of moments, some of which may not exceed 4, particularly in turbulent market conditions. Chabi-Yo and Loudis (2020) present sufficient conditions for 1.6.2(ii) to hold, which relate to the sign of the fifth derivative of the utility function of the representative agent. Specifically, for common utility functions such as CRRA or HARA

utility, parameter restrictions are needed to ensure that 1.6.2(ii) holds.[25]

I need one additional assumption to bound disaster risk premia. To state this assumption and the resulting lower bound, I use the following notation for high-order risk-neutral moments and truncated high-order risk-neutral moments, respectively.

$$\begin{aligned}\widetilde{\mathbb{M}}_{t\to N}^{(n)} &\coloneqq \widetilde{\mathbb{E}}_t\left[\left(R_{m,t\to N}-R_{f,t\to N}\right)^n\right]\\ \widetilde{\mathbb{M}}_{t\to N}^{(n)}[k_0] &\coloneqq \widetilde{\mathbb{E}}_t\left[\mathbb{1}\left(R_{m,t\to N}\leq k_0\right)\left(R_{m,t\to N}-R_{f,t\to N}\right)^n\right].\end{aligned} \tag{1.6.7}$$

**Assumption 1.6.3.** *In the representative agent model, the following holds:*

*(i)* $(-1)^{k-1}\theta_k \geq \frac{1}{R_{f,t\to N}^k}$ *for* $k=1,2,3$

*(ii)* $\widetilde{\mathbb{M}}_{t\to N}^{(3)} \leq 0$.

Chabi-Yo and Loudis (2020, Table 6) provide empirical evidence that 1.6.3((i)) holds with equality when estimating the conditional equity premium. Assumption 1.6.3((ii)) is a very mild restriction on risk-neutral skewness, which is almost always negative at every date and time horizon. This empirical fact is well known.[26]

The following two propositions show how option data can be employed to establish bounds on the difference between the physical and risk-neutral measures in the left-tail.

**Proposition 1.6.4** (Lower Bound on CDF)**.** *Suppose Assumptions 1.6.2 and 1.6.3 hold, and assume that the risk-neutral density exists. Then,*

$$\tau - F_t\left(\widetilde{Q}_{t,\tau}\right) \geq \frac{\sum_{k=1}^3 \frac{(-1)^{k-1}}{R_{f,t\to N}^k}\left(\tau\widetilde{\mathbb{M}}_{t\to N}^{(k)} - \widetilde{\mathbb{M}}_{t\to N}^{(k)}[\widetilde{Q}_{t,\tau}]\right)}{1+\sum_{k=1}^3 \frac{(-1)^{k-1}}{R_{f,t\to N}^k}\widetilde{\mathbb{M}}_{t\to N}^{(k)}} =: \mathrm{CLB}_{t,\tau}, \tag{1.6.8}$$

[25]For example, for CRRA utility, the risk aversion coefficient cannot be too large. See Appendix A.4 for a detailed discussion.

[26]Chabi-Yo and Loudis (2020) argue that all odd risk-neutral moments should be negative, since they expose the investor to unfavorable market conditions.

*for all $\tau \leq \tau'$, where $\tau'$ is defined implicitly by*

$$\widetilde{Q}_{t,\tau'} = \min\left(R_{f,t\to N} - \sqrt{\widetilde{\mathbb{VAR}}_t(R_{m,t\to N})}, \widetilde{Q}_{t,\tau^*}\right),$$

*and $\widetilde{Q}_{t,\tau^*}$ is defined in Theorem A.1.12.*

*Proof.* See Appendix A.1.10. ■

**Proposition 1.6.5** (Lower Bound on Disaster Risk Premia)**.** *Consider the same assumptions in Proposition 1.6.4 and assume additionally that Assumption 1.6.1 holds. Then, for all $\tau \leq \tau'$*

$$Q_{t,\tau} - \widetilde{Q}_{t,\tau} \geq \overbrace{\frac{\mathrm{CLB}_{t,\tau}}{\widetilde{f}_t(\widetilde{Q}_{t,\tau})}}^{\text{risk-adjustment}} =: \mathrm{LB}_{t,\tau}. \tag{1.6.9}$$

*Proof.* By Assumption 1.6.1, the approximation in (1.6.4) holds, which in combination with Proposition 1.6.4 renders

$$\begin{aligned} Q_{t,\tau} - \widetilde{Q}_{t,\tau} &\overset{(1.6.4)}{\approx} \frac{\tau - F_t(\widetilde{Q}_{t,\tau})}{\widetilde{f}_t(\widetilde{Q}_{t,\tau})} \\ &\overset{(1.6.8)}{\geq} \frac{1}{\widetilde{f}_t(\widetilde{Q}_{t,\tau})}\left(\frac{\sum_{k=1}^3 \frac{(-1)^{k+1}}{R_{f,t\to N}^k}\left(\tau\widetilde{\mathbb{M}}_{t\to N}^{(k)} - \widetilde{\mathbb{M}}_{t\to N}^{(k)}[\widetilde{Q}_{t,\tau}]\right)}{1 + \sum_{k=1}^3 \frac{(-1)^{k+1}}{R_{f,t\to N}^k}\widetilde{\mathbb{M}}_{t\to N}^{(k)}}\right). \end{aligned}$$ ■

Proposition 1.6.4 provides a bound on the physical CDF that requires no parameter estimation and relies solely on time $t$ information. This result complements recent work on belief recovery. Ross (2015) demonstrated CDF recovery under the assumption of transition independence, but subsequent research has questioned this assumption (Borovička, Hansen, and Scheinkman, 2016; Qin, Linetsky, and Nie, 2018; Jackwerth and Menner, 2020). In contrast, Proposition 1.6.4 establishes a lower bound on the left-tail of the physical distribution using a different set of mild economic constraints. Additionally, Section 1.2.3 showed that the right-tail of $F_t$ can be approximately recovered from the risk-neutral distribution due to the minimal need for risk-adjustment. These findings suggest the potential for approximate recovery of $F_t$ using option prices.

I will test this hypothesis using the lower bound on disaster risk premia in Proposition

1.6.5. Specifically, a tight lower bound in (1.6.9) would enable direct inference on both the physical distribution ($Q_{t,\tau}$) and disaster risk premia ($Q_{t,\tau} - \widetilde{Q}_{t,\tau}$). While Section 1.2.2 proposed the quantile model (1.2.3) to estimate $Q_{t,\tau}$, it can be criticized for having time-homogeneous coefficients. Proposition 1.6.5 relaxes that assumption. Furthermore, the lower bound in (1.6.9) is not prone to the historical sample bias critique of Welch and Goyal (2008). Alternatively, one can estimate a disaster risk model to infer $Q_{t,\tau}$, but this approach is also susceptible to misspecification concerns and faces challenges in estimation due to the scarcity of disaster events in the data (Julliard and Ghosh, 2012; Martin, 2013).

### 1.6.3 Calculating the Lower Bound

Before assessing how tight the lower bound is in Proposition 1.6.5, I outline the procedure to calculate it, which depends on $\text{CLB}_{t,\tau}$ and $\widetilde{f}_t(\widetilde{Q}_{t,\tau})$. Both functions can be derived from $\widetilde{Q}_{t,\tau}$, which is estimated using the same data and procedure of Section 1.2.3. To see that $\widetilde{f}_t(\widetilde{Q}_{t,\tau})$ can be derived from $\widetilde{Q}_{t,\tau}$, notice that $\frac{d}{d\tau}\widetilde{Q}_t(\tau) = 1/\widetilde{f}_t(\widetilde{Q}_{t,\tau})$. The latter term can thus be approximated by[27]

$$\frac{1}{\widetilde{f}_t(\widetilde{Q}_{t,\tau})} \approx \frac{\widetilde{Q}_t(\tau+h) - \widetilde{Q}_t(\tau-h)}{2h},$$

where $h$ is the bandwidth of the $\tau$-grid. Second, to calculate $\text{CLB}_{t,\tau}$ in (1.6.8), I use $\widetilde{Q}_{t,\tau}$, as well as the formula for high-order risk-neutral moments in Appendix A.1.11.

Given the evidence in Table 1.2 that $Q_{t,\tau} > \widetilde{Q}_{t,\tau}$ in the left-tail, Proposition 1.6.5 has nontrivial content in the data if $\text{LB}_{t,\tau} \geq 0$. Appendix Table A.2 contains summary statistics of $\text{LB}_{t,\tau}$, which show that the lower bound is always positive, right-skewed, more pronounced in the right-tail and economically meaningful in magnitude, with outliers that can spike up to 29%.

### 1.6.4 Tightness of the Lower Bound: In-sample Evidence

To test whether the lower bound in Proposition 1.6.5 is tight, I form *excess quantile returns*: $R_{m,t\to N} - \widetilde{Q}_{t,\tau}$. Since $\widetilde{Q}_{t,\tau}$ is observed at time $t$, it follows that $Q_{t,\tau}(R_{m,t\to N} - \widetilde{Q}_{t,\tau}) =$

[27] I slightly abuse notation to emphasize that the derivative is taken w.r.t. $\tau$, so that $\widetilde{Q}_t(\tau+h)$ denotes $\widetilde{Q}_{t,\tau+h}$.

$Q_{t,\tau}(R_{m,t\to N}) - \widetilde{Q}_{t,\tau}$. Subsequently, I use QR to estimate the model

$$Q_{t,\tau}(R_{m,t\to N}) - \widetilde{Q}_{t,\tau}(R_{m,t\to N}) = \beta_0(\tau) + \beta_1(\tau)\mathrm{LB}_{t,\tau},$$
$$[\widehat{\beta}_0(\tau), \widehat{\beta}_1(\tau)] = \underset{(\beta_0,\beta_1)\in\mathbb{R}^2}{\arg\min} \sum_{t=1}^{T} \rho_\tau(R_{m,t\to N} - \widetilde{Q}_{t,\tau} - \beta_0 - \beta_1 \mathrm{LB}_{t,\tau}). \tag{1.6.10}$$

Regression (1.6.10) is a quantile analogue of the mean excess return regressions of Welch and Goyal (2008). Under the null hypothesis that the lower bound is tight, it holds that

$$H_0: \quad [\beta_0(\tau), \beta_1(\tau)] = [0, 1]. \tag{1.6.11}$$

Less restrictive, one can test whether $\beta_0(\tau) = 0$ and $\beta_1(\tau) > 0$, which implies that the statistical "factor" $\mathrm{LB}_{t,\tau}$ explains the conditional quantile wedge.[28]

Table 1.5 presents the results of regression (1.6.10). The null hypothesis of a tight lower bound in (1.6.11) is not rejected for $\tau = 0.2$, but it is rejected for $\tau \in \{0.05, 0.1\}$ across all horizons. When the null hypothesis is rejected, the $\beta_1(\tau)$-coefficient exceeds 1, consistent with the theory that $\mathrm{LB}_{t,\tau}$ represents a lower bound on disaster risk premia. In all cases, the lower bound is economically meaningful, since $\beta_1(\tau)$ is significantly different from 0, while $\beta_0(\tau) = 0$ can never be rejected. The explanatory power of the regression in Table 1.5 is modest, as shown by the $R^1(\tau)$ measure-of-fit:

$$R^1(\tau) = 1 - \frac{\min_{b_0,b_1} \sum \rho_\tau(R_{m,t\to N} - b_0 - b_1 \mathrm{LB}_{t,\tau})}{\min_{b_0} \sum \rho_\tau(R_{m,t\to N} - b_0)}. \tag{1.6.12}$$

But, following the reasoning of Section 1.5.2, the predictive power in the left-tail cannot be too big, for otherwise near-arbitrage opportunities exist.

I also directly test the predictive power of the lower bound in estimating the physical

[28] For example, if we start with a quantile factor model $Q_{t,\tau} = \widetilde{Q}_{t,\tau} + \beta(\tau)\mathrm{LB}_{t,\tau}$, the model has one testable implication for the data: the intercept in a quantile regression of $R_{m,t\to N} - \widetilde{Q}_{t,\tau}$ on $\mathrm{LB}_{t,\tau}$ should be zero. Quantile factor models have recently been proposed by Chen, Dolado, and Gonzalo (2021).

quantile function. To this end, I use the following model-free quantile forecast:

$$\widehat{Q}_{t,\tau} \coloneqq \widetilde{Q}_{t,\tau} + \text{LB}_{t,\tau}. \tag{1.6.13}$$

To evaluate the accuracy of this forecast, I use QR to estimate the model

$$Q_{t,\tau} = \beta_0(\tau) + \beta_1(\tau)\widehat{Q}_{t,\tau}. \tag{1.6.14}$$

An accurate forecast would imply the joint restriction

$$H_0 : \ \beta_0(\tau) = 0, \quad \beta_1(\tau) = 1. \tag{1.6.15}$$

Table 1.6 summarizes the estimates of (1.6.14) for several percentiles. The results compare favorably to the risk-neutral estimates in Table 1.2. First, the point estimates are closer to the $[0, 1]$ benchmark. Second, the Wald test on the joint restriction in (1.6.15) is never rejected except for $\tau = 0.05$ at the 60-day horizon. Third, the in-sample explanatory power is higher. The same conclusion applies when comparing the predictive results to the expanding quantile regression from Table A.1 in the Appendix. Collectively, these findings suggest that $\widehat{Q}_{t,\tau}$ can be considered as a good lower bound on the physical quantile function in the left-tail.

### 1.6.5 Tightness of the Lower Bound: Out-of-sample Evidence

Given that the in-sample results from Table 1.6 suggest that $\widehat{Q}_{t,\tau}$ is a good lower bound for $Q_{t,\tau}$, it is natural to assess its out-of-sample performance by using $\widehat{Q}_{t,\tau}$ to directly predict $Q_{t,\tau}$, which does not require any parameter estimation.

To assess the out-of-sample performance, I use the $R^1_{oos}(\tau)$ measure of fit defined in (1.2.6) with $\widehat{Q}_{t,\tau}$ instead of $\widetilde{Q}_{t,\tau}$. Table 1.6 shows that $\widehat{Q}_{t,\tau}$ improves upon the historical rolling quantile out-of-sample in all cases. In particular, this outperformance is most pronounced at the 5th percentile, which is expected since option data are known to provide useful information about extreme downfalls in the stock market (Bates, 2008; Bollerslev and Todorov, 2011). In Appendix A.6.2, I

run a battery of robustness tests which show that, out-of-sample, $\text{LB}_{t,\tau}$ better predicts the conditional quantile function than other benchmarks such as the risk-neutral quantile or the VIX index. The latter result is particularly encouraging since the VIX predictor uses in-sample information.

### 1.6.6 Robustness of the QR Estimates

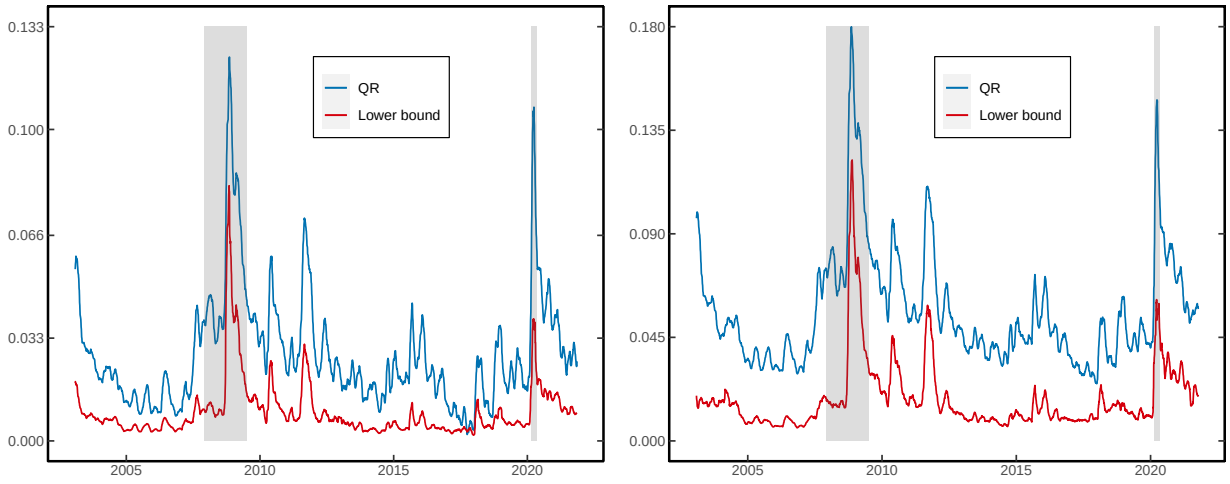
The in- and out-of-sample results support $\text{LB}_{t,\tau}$ as a robust lower bound for disaster risk premia. It is instructive to compare this lower bound to the disaster risk premia reported in Figure 1.3b, which are inferred from the quantile regression in (1.2.4). Consistent with the theory, the estimated disaster risk premium at $\tau = 0.05$ exceeds the lower bound in 99% of cases for 30-day returns and 99.9% for 60-day returns. When violations of the lower bound occur, the differences are typically small.

Figures 1.8a and 1.8b show the lower bound for 30- and 60-day returns, respectively, alongside the disaster risk premium estimated from the quantile regression (1.2.4). In both cases, there is a substantial correlation between the lower bound and the disaster risk premium obtained from the QR estimates. Especially during the global financial crisis and the Covid-19 crisis, both methods predict significant increases in the disaster risk premium. Outside these crisis periods, the lower bound is more conservative. Overall, the model-free lower bound corroborates the robustness of the estimated disaster risk premium in Figure 1.3b.

## 1.7 Conclusion

I use return and option data on the S&P500 in combination with quantile regression to estimate local differences between the conditional risk-neutral and physical quantile functions. Empirically, these differences are substantial in the left-tail, whereas in the right-tail, they are barely discernible. Therefore, the lion's share of the equity premium is driven by downside returns, which is model-free evidence for disaster risk.

By tracking these quantile differences over time, the results also demonstrate that disaster risk is time-varying, pervasive, and a driving force behind much of the equity premium, even outside




**(a)** 30-day returns **(b)** 60-day returns

**Figure 1.8. Lower bound on disaster risk premium at 5th percentile.** Panel (a) shows the lower bound on the disaster risk premium for 30-day returns, at $\tau = 0.05$. QR denotes the estimated disaster risk premium from the quantile regression (1.2.4). The right panel shows a similar graph for 60-day returns. Both figures are smoothed using a 30-day rolling window. The two shaded bars signify the Great Recession period (Dec 2007 – June 2009) and Covid-19 crisis (Feb 2020 – April 2020).

crisis periods. Additionally, my findings show that disaster risk is more nuanced than previous literature suggests. Much of the disagreement can be attributed to the incorporation of conditioning information. While prior research primarily focused on unconditional estimation, my approach accounts for conditioning information embedded in the risk-neutral quantile function, which is crucial to obtain accurate estimates of disaster risk.

To build on this finding, I show that disaster risk makes the SDF highly volatile. In particular, option strategies involving a short position in an asset that pays one dollar in case of a disaster exhibit substantially higher Sharpe ratios compared to a direct investment in the market portfolio. The data reveal that such investment strategies yield a monthly Sharpe ratio of 30%, more than doubling the Sharpe ratio of the market return.

Finally, I suggest a model-free lower bound on disaster risk premia observed from option prices. This lower bound serves as a good predictor of the quantile wedge, exhibiting spikes during crises and significant fluctuations over time. Furthermore, the lower bound closely aligns with estimates of disaster risk premia based on quantile regression, thereby reinforcing the robustness

of my findings.

## 1.8 Acknowledgements

Chapter 1, in full, is currently being prepared for submission for publication of the material. The dissertation author is the sole author of this material.

**Table 1.2. Risk-neutral quantile regression**

| Horizon | $\tau$ | $\widehat{\beta}_0(\tau)$ | $\widehat{\beta}_1(\tau)$ | Wald test ($p$-value) | $R^1(\tau)$[%] | $R^1_{oos}(\tau)$[%] | $\overline{\text{Hit}}$[%] | $\widehat{Q}_{t,\tau} > \widetilde{Q}_{t,\tau}$[%] |
|---|---|---|---|---|---|---|---|---|
| 30 days* | 0.05 | 0.43 (0.208) | 0.56 (0.223) | 0.00 | 6.28 | 6.11 | -2.67 (0.676) | 99.88 |
| | 0.1 | 0.45 (0.201) | 0.54 (0.209) | 0.01 | 3.45 | 1.01 | -3.56 (1.089) | 98.52 |
| | 0.2 | 0.69 (0.284) | 0.30 (0.290) | 0.02 | 0.55 | 0.89 | -3.73 (1.719) | 90.98 |
| | 0.3 | 1.02 (0.357) | -0.02 (0.360) | 0.00 | 0.00 | 2.49 | -5.51 (2.147) | 99.58 |
| | 0.4 | 1.17 (0.237) | -0.16 (0.237) | 0.00 | 0.03 | 1.75 | -7.32 (2.357) | 97.25 |
| | 0.6 | -0.45 (0.216) | 1.44 (0.213) | 0.00 | 4.62 | 4.19 | -8.05 (2.468) | 99.93 |
| | 0.7 | -0.18 (0.162) | 1.18 (0.159) | 0.03 | 7.79 | 7.47 | -5.84 (2.220) | 99.95 |
| | 0.8 | -0.09 (0.141) | 1.09 (0.137) | 0.19 | 12.44 | 12.50 | -3.24 (1.886) | 99.95 |
| | 0.9 | 0.03 (0.113) | 0.97 (0.108) | 0.96 | 20.41 | 21.88 | -0.04 (1.235) | 55.85 |
| *(Obs. 4333) | 0.95 | 0.12 (0.120) | 0.89 (0.114) | 0.57 | 27.07 | 31.31 | 0.27 (0.863) | 22.41 |
| 60 days** | 0.05 | 0.45 (0.303) | 0.54 (0.343) | 0.00 | 3.12 | 13.14 | -3.33 (0.875) | 100.00 |
| | 0.1 | 0.58 (0.263) | 0.41 (0.283) | 0.00 | 1.79 | 3.50 | -5.57 (1.320) | 100.00 |
| | 0.2 | 0.78 (0.336) | 0.21 (0.345) | 0.01 | 0.38 | -0.03 | -6.60 (2.351) | 99.95 |
| | 0.3 | 0.93 (0.434) | 0.07 (0.438) | 0.00 | 0.01 | -0.12 | -7.81 (3.012) | 99.47 |
| | 0.4 | 0.36 (0.325) | 0.65 (0.323) | 0.02 | 0.25 | 2.34 | -8.48 (3.439) | 99.79 |
| | 0.6 | -0.65 (0.342) | 1.64 (0.333) | 0.02 | 5.57 | 4.60 | -7.68 (3.465) | 99.77 |
| | 0.7 | -0.31 (0.266) | 1.30 (0.256) | 0.05 | 8.41 | 7.65 | -7.34 (3.260) | 99.91 |
| | 0.8 | -0.08 (0.183) | 1.08 (0.174) | 0.07 | 12.70 | 12.23 | -5.53 (2.683) | 100.00 |
| | 0.9 | 0.04 (0.147) | 0.96 (0.138) | 0.58 | 21.66 | 22.79 | -1.94 (1.707) | 92.86 |
| **(Obs. 4312) | 0.95 | 0.04 (0.135) | 0.96 (0.126) | 0.90 | 31.07 | 34.19 | 0.43 (1.046) | 13.73 |
| 90 days*** | 0.05 | 0.60 (0.405) | 0.37 (0.478) | 0.01 | 2.90 | 15.63 | -2.95 (1.102) | 100.00 |
| | 0.1 | 0.59 (0.321) | 0.40 (0.356) | 0.00 | 3.46 | 3.84 | -6.36 (1.495) | 100.00 |
| | 0.2 | 0.57 (0.516) | 0.43 (0.534) | 0.03 | 0.83 | 1.93 | -7.53 (2.896) | 100.00 |
| | 0.3 | 0.62 (0.637) | 0.39 (0.643) | 0.04 | 0.17 | -0.52 | -8.42 (3.668) | 99.84 |
| | 0.4 | 0.42 (0.468) | 0.60 (0.463) | 0.02 | 0.22 | -1.76 | -9.52 (4.199) | 99.77 |
| | 0.6 | -0.84 (0.426) | 1.82 (0.413) | 0.01 | 6.37 | 3.81 | -11.60 (4.542) | 99.98 |
| | 0.7 | -0.46 (0.307) | 1.45 (0.293) | 0.02 | 10.45 | 8.87 | -9.43 (4.056) | 100.00 |
| | 0.8 | -0.23 (0.204) | 1.23 (0.192) | 0.10 | 15.47 | 16.54 | -6.66 (3.189) | 100.00 |
| | 0.9 | -0.02 (0.170) | 1.02 (0.157) | 0.79 | 23.18 | 27.92 | -1.12 (1.971) | 100.00 |
| ***(Obs. 4291) | 0.95 | 0.08 (0.153) | 0.93 (0.139) | 0.86 | 32.14 | 39.88 | -0.06 (1.366) | 52.37 |

*Note*: This table reports the QR estimates of (1.2.4) over the sample period 2003–2021 at different horizons, using overlapping returns. Standard errors are shown in parentheses and based on SETBB with a block length equal to the prediction horizon. *Wald test* denotes the $p$-value of the joint restriction $[\beta_0(\tau), \beta_1(\tau)] = [0, 1]$. $R^1(\tau)$ denotes the goodness of fit measure (1.2.5). $R^1_{oos}(\tau)$ is the out-of-sample goodness of fit (1.2.6), using a rolling window of size 10 times the prediction horizon. $\overline{\text{Hit}}$ refers to the sample expectation defined in (1.3.3) and standard errors are reported in parentheses, which are obtained by stationary bootstrap based on 10,000 bootstrap samples. The last column indicates the time series average of the event that $\widehat{Q}_{t,\tau} > \widetilde{Q}_{t,\tau}$, where $\widehat{Q}_{t,\tau} = \widehat{\beta}_0(\tau) + \widehat{\beta}_1(\tau)\widetilde{Q}_{t,\tau}$.

**Table 1.3. OLS estimates of conditional equity premium**

| | Full sample | | | | Sub-sample | | |
|---|---|---|---|---|---|---|---|
| Horizon | $\widehat{\beta}_0$ | $\widehat{\beta}_1$ | $R^2$[%] | Obs | $R^2$[%] | $R^2_{oos}$[%] | $p$-value DM |
| 30 days | 0.13 (0.089) | -0.14 (0.096) | 1.79 | 4333 | 10.44 | 1.82 | 0.00 |
| 60 days | 0.17 (0.120) | -0.18 (0.137) | 2.74 | 4312 | 18.16 | 3.30 | 0.00 |
| 90 days | 0.20 (0.150) | -0.22 (0.178) | 3.58 | 4291 | 26.67 | 4.20 | 0.00 |
| Period | 2003–2021 | | | | 2013–2021 | | |

*Note*: This table reports the OLS estimates of (1.3.2) for 30-, 60- and 90-day returns. Standard errors are shown in parentheses and calculated using stationary bootstrap, with an average block length equal to the return horizon. $R^2_{oos}$ denotes the out-of-sample $R^2$ using the historical rolling mean of excess returns. The window length is equal to 5 years. *p-value DM* denotes the $p$-value of the Diebold and Mariano (1995) test that the risk-neutral quantile exhibits equal out-of-sample forecasting accuracy as the rolling mean. The "*Period*" row indicates the specific time periods used for estimation.

**Table 1.4. Sample bounds and bootstrap result**

| Sample size | HJ bound | distribution bound | $p$-value |
|---|---|---|---|
| 312 | 0.133 | 0.260 | 0.075 |

*Note*: This table reports the HJ and distribution bound for monthly S&P500 returns over the period 1996–2021. The distribution bound is evaluated at $\tau = 0.0374$. The final column denotes the $p$-value of the null hypothesis in (1.5.14). The $p$-value is obtained from 100,000 bootstrap samples and counts the fraction of times that $\mathcal{T}^\star \leq 0$.

**Table 1.5. Quantile regression with lower bound**

| Horizon | $\tau$ | $\widehat{\beta}_0(\tau)$ | $\widehat{\beta}_1(\tau)$ | Wald test ($p$-value) | $R^1(\tau)$[%] | Obs |
|---|---|---|---|---|---|---|
| 30 days | 0.05 | -0.01 (0.005) | 4.43 (0.349) | 0.00 | 6.03 | 4333 |
| | 0.1 | -0.01 (0.006) | 2.17 (0.450) | 0.03 | 3.18 | |
| | 0.2 | -0.01 (0.006) | 1.33 (0.400) | 0.02 | 0.41 | |
| 60 days | 0.05 | -0.01 (0.013) | 5.53 (0.571) | 0.00 | 3.60 | 4312 |
| | 0.1 | -0.02 (0.011) | 3.25 (0.540) | 0.00 | 2.23 | |
| | 0.2 | -0.02 (0.009) | 1.50 (0.398) | 0.27 | 0.48 | |
| 90 days | 0.05 | -0.02 (0.032) | 6.37 (1.113) | 0.00 | 4.91 | 4291 |
| | 0.1 | -0.02 (0.018) | 3.05 (0.528) | 0.00 | 4.43 | |
| | 0.2 | -0.02 (0.019) | 1.36 (0.626) | 0.69 | 1.46 | |

*Note*: This table reports the QR estimates of (1.6.10) over the sample period 2003-2021 at different horizons, using overlapping returns. Standard errors are shown in parentheses and calculated using SETBB with a block length equal to the prediction horizon. *Wald test* denotes the $p$-value of the joint restriction $[\beta_0(\tau), \beta_1(\tau)] = [0, 1]$. $R^1(\tau)$ denotes the goodness-of-fit measure (1.6.12).

**Table 1.6. Quantile regression with model-free quantile forecast**

| Horizon | $\tau$ | $\widehat{\beta}_0(\tau)$ | $\widehat{\beta}_1(\tau)$ | Wald test ($p$-value) | $R^1(\tau)$[%] | $R^1_{oos}(\tau)$[%] | Obs |
|---|---|---|---|---|---|---|---|
| 30 days | 0.05 | 0.29 (0.249) | 0.70 (0.265) | 0.06 | 6.28 | 9.94 | 4333 |
| | 0.1 | 0.28 (0.250) | 0.72 (0.260) | 0.18 | 3.57 | 4.02 | |
| | 0.2 | 0.57 (0.381) | 0.43 (0.388) | 0.29 | 0.58 | 2.53 | |
| 60 days | 0.05 | 0.30 (0.382) | 0.71 (0.426) | 0.02 | 3.40 | 17.81 | 4312 |
| | 0.1 | 0.38 (0.352) | 0.61 (0.373) | 0.13 | 2.35 | 9.22 | |
| | 0.2 | 0.44 (0.487) | 0.56 (0.498) | 0.21 | 0.57 | 4.28 | |
| 90 days | 0.05 | 0.36 (0.520) | 0.64 (0.602) | 0.05 | 4.26 | 21.98 | 4291 |
| | 0.1 | 0.31 (0.482) | 0.70 (0.521) | 0.06 | 4.19 | 13.22 | |
| | 0.2 | 0.23 (0.696) | 0.78 (0.710) | 0.48 | 0.70 | 5.99 | |

*Note*: This table reports the QR estimates of (1.6.14) over the sample period 2003-2021. Standard errors are shown in parentheses and calculated using the SETBB, with block length equal to the prediction horizon. *Wald test* gives the $p$-value of the Wald test on the joint restriction: $\widehat{\beta}_0(\tau) = 0, \widehat{\beta}_1(\tau) = 1$. $R^1(\tau)$ denotes the in-sample goodness-of fit criterion (1.2.5). $R^1_{oos}(\tau)$ is the out-of-sample goodness-of fit, using a rolling window size equal to 10 times the return horizon.

# Chapter 2

# Scale Economies, Bargaining Power, and Investment Performance: Evidence from Pension Plans

## 2.1 Introduction

During recent decades, the professional asset management industry has undergone significant structural changes. The competitive landscape, influenced by both passive and active managers, has led to a substantial reduction in fees. Advancements in technology and increased availability of information have also played a role in this fee reduction (Blake, Rossi, Timmermann, Tonks, and Wermers, 2013). Furthermore, both active and passive managers have refined their investment offerings, focusing on specialization in their investment strategies. Simultaneously, major institutional investors like pension plans and endowments have expanded their allocations to alternative asset classes, including hedge funds, private debt, private equity, and real assets.

Defined-benefit (DB) pensions continue to play a significant role in the global financial market, with the total assets under management (AUM) of DB pensions experiencing substantial growth. Notably, state and local government DB plans in the U.S. have seen their AUM increase from $1.4 trillion in 1995 to $5.1 trillion in 2020, while private-sector DB plans in the U.S. have grown from $1.5 trillion to $3.4 trillion over the same period (Investment Company Institute, 2021, p. 177). Moreover, the DB landscape now includes several very large pension plans, such as CalSTRS, one of the world's largest pension funds, with total assets exceeding $314.8 billion as of May 31, 2022.[1]

The confluence of the above-noted shifts in the asset management industry with the in-

[1]See https://www.calstrs.com/investment-portfolio

creased bulk of the largest DB plans brings several new issues to light, such as a potential increase in the bargaining power of DB plans in their interactions with their external money managers. Simply put, the negotiating power of very large DB plans, of late, may bring substantial changes in the balance of power between large DB plans and their investment managers.

To explore these issues, our study conducts a granular analysis of the DB industry, with an emphasis on the interaction of DB plan size with fees, asset allocation, and investment performance. For example, one economically important trend is that large DB plans are increasingly managing assets "in-house," to cut fees while potentially maintaining a reasonable level of performance (Beath, Flynn, Jethalal, and Reid, 2022).[2] A key issue that we explore is whether such in-house management brings greater bargaining power to plans when they negotiate fees and shop for the best investment managers for external management services—and, whether such bargaining power mainly resides with the largest pension plans due to the fixed costs of establishing and maintaining internal management.

Our study brings a contrast to past research on funds with small-scale investors. In retail mutual fund markets, individual investors are usually considered as "atomistic" agents who have no individual (or collective) bargaining power. The seminal paper of Berk and Green (2004) (henceforth, BG) presents a model based on this assumption as well as the usual assumption that "alpha" generation by fund managers exhibits diseconomies of scale. Predictions from their model include that (1) skilled investment managers collect all of the rents from their alpha-generating efforts, (2) flows from atomistic investors occur at each period (in reaction to updated information about manager skills)—either into or out of each fund until management fees equal expected pre-fee alphas, and (3) all investors, being atomistic, obtain the exact same zero expected alpha, net-of-fees.

In the pension plan market that we examine, we propose that the only key assumption from the BG model that can be accepted without further investigation is the presence of pre-fee scale diseconomies in fund-level alpha generation. While retail markets may leave all bargaining power in the hands of asset managers, the situation is different for the largest pension plans. Given

[2] As an important example, CalSTRS recently stated that in-house management and co-management with external managers has been instrumental to their cost savings (see link). However, remains unclear whether the choice of in-house management—or, the threat thereof—leads to greater negotiating power with external managers to obtain better pre-fee performance and/or lower fees.

their substantial scale, these plans have the potential to negotiate favorable terms with external managers, including lower fees and access to well-performing managers. As a result, the economics of the largest plans may diverge from the BG model's zero expected alpha (net of fees) assumption, primarily due to their enhanced bargaining power. Accordingly, we explore the relationship between DB plan size, bargaining power, and the ability of large plans to capture value from external managers, whether through lower fees or higher pre-fee alpha generation by these managers.[3]

We also investigate the impact of scale in pension plans on asset allocation trends. For example, it is unclear whether the bargaining power of large plans results in a greater use of external active managers (presumably at lower fee levels) or a greater tendency to internally manage assets (either actively or passively). As another example, as large plans move assets to internal management, it is important to assess whether any potential reduction in their internal active management skills, compared to external managers, outweighs the cost savings gained through internal passive management for different asset classes.[4] Thus, our paper provides a unique view into scale economies of plans and the associated bargaining power at the level of plan asset classes.[5]

The modeling framework proposed by Gârleanu and Pedersen (2018) (GP; henceforth), has parallels to the empirical setting of our analysis. In their model, investors incur a fixed search cost to identify skilled external asset managers who, in turn, incur a fixed cost from acquiring information about asset returns that enables them to outperform passive investments. Investment management fees in the GP model are determined through Nash bargaining, leaving a natural mechanism through which plan size (as a proxy for bargaining power) matters for fees as well as for net-of-fee returns when some investors are not atomistic in size. Further, information acquisition costs can be expected to be higher in the less transparent private asset markets than in public asset markets. This is consistent with an equilibrium in which investment management costs are

[3]Of course, for internal management to pose a "threat" to external managers, there must be a large mass of plans that stand ready to manage internally—and, a finite mass of smaller plans, such that external managers do not retain all bargaining power, unlike the infinitely deep supply of capital assumed in BG. In this vein, we note that, while some large plans may choose not to spend the fixed costs of setting up internal management, the mere threat to do so gives them negotiating power with external managers.

[4]That is, an important issue is whether internally managing a greater share of a plan's assets in a particular asset class leads to a different mix of active vs. passive management in that asset class, as well as other asset classes held by the plan. The role of pension plan scale in internalizing active vs. passive management can be expected to depend on the relative fixed costs of creating and maintaining an internal management organization for each, within a given asset class.

[5]For example, large pension plans may be more capable of actively managing private asset classes, where they might directly exert their size to obtain more favorable investments.

relatively high in private asset markets, and the largest plans benefit disproportionately from their higher ability to engage with skilled managers, either due to their enhanced ability to overcome fixed search costs and/or to negotiate lower investment management fees once they identify skilled managers.

Two major trends drive our inquiry into scale-related performance in the pension fund industry. First, pensions have increasingly moved toward passive management of their public market exposures in both equities and (to a lesser extent) fixed-income investments. Second, pensions have increasingly turned to private equity funds or to direct investments in real assets as a source of diversification and higher long-term returns. These developments can be partially attributed to the shrinking number of publicly traded stocks that are available in the U.S., which is particularly relevant to large pension plans. In the face of these trends, as we will show, many large plans have assigned a greater role to internal asset management. Both of these trends are consistent with a shrinking level of fixed costs in setting up an internal asset management organization across all asset classes, but especially so in public market securities.

Our inquiry exploits a unique database to explore several dimensions of the pension plan sector, including both cross-sectional and time-series aspects. Our data is sourced from CEM, a Toronto-based private consulting company that collects information from a diverse range of pension plans. Each year, CEM gathers data on these plans' asset allocations as percentages within major asset classes (e.g., public equities, fixed income, hedge funds, private equity, public debt, private debt, and real assets), asset subclasses (e.g., small-cap U.S. equities or infrastructure investments), and, within each subclass, their choice between active and passive management and between internal and external management. The CEM database uniquely includes data on AUM, gross returns, and investment costs for each subclass/active-passive combination. Additionally, the CEM staff routinely apply a battery of checks to obtain the most precise data possible.[6] We believe that the CEM data allows a closer look at the above questions than has been possible with prior studies.

With this CEM database, we find that large pension plans tend to invest a greater share of their plan assets in less-liquid sectors of the market, as well as sectors of the market where

[6]From our discussions with CEM, it is apparent that CEM researchers maintain frequent contact with their "subscribers" in cases where data looks suspect in order to maintain the data integrity.

scale-related bargaining power can be expected to be especially important in achieving net-of-fee alphas, such as private equity investments (see also Dyck and Pomorski (2016)). Further, large plans tend to use internal management to a greater degree, particularly in public asset classes where the fixed-cost of establishing and maintaining internal investment management is lower.

Further, we identify two major shifts in the asset allocation of U.S. pension plans. First, the share of (publicly-traded) stocks and fixed income assets has declined from nearly 90% in the early 1990s to 70% at the end of our sample (2019), while allocations to non-traditional asset classes such as private equity, hedge funds, and real assets increase significantly over time.[7] Second, within traditional asset classes such as equity and bonds, we see large shifts toward more specialized mandates. For instance, there has been a transition from broad or all U.S. equities to funds focusing on large, medium, and small market capitalization segments in the equity space. Similarly, we observe a move from general U.S. bond allocations towards more specialized strategies targeting high-yield and credit objectives in the fixed-income sector. By far the biggest shift is toward international and global assets, which become more prominent over time, particularly in stock allocations.

These shifts in asset allocation are consistent with a decrease in the fixed costs of managing investments, and this decrease varies widely across asset classes and sub-asset classes (as well as within an asset class or sub-asset class, between active and passive management).[8] When we examine the tendency of pension plans to manage assets internally, we find that plan size is of key importance. Larger plans are significantly more likely to manage assets internally in all asset classes except for hedge funds and multi-assets.

Upon further examination, we observe that the size of a pension plan negatively correlates with its tendency to employ active management, particularly for public securities, such as equities. Larger pensions increasingly harness the substantial economies of scale offered by passive management in public securities. This shift towards passive management is more pronounced over time,

[7]Hedge fund holdings, on average across plans, increase from 1% in 2003 to 6% in 2019. Private equity holdings also increase to 9% in 2019 from 4% in 2000; allocations to real assets increase to 10% in 2019 from 4% in the early 1990s.

[8]The shift from public, broad asset classes to private or more-specialized public asset classes is consistent with Blake, Rossi, Timmermann, Tonks, and Wermers (2013), who find that investment managers have moved to more specialized sub-asset-classes in seeking to provide value to pension sponsors. This is consistent with the increased competition in broad public asset classes that might be expected from a larger amount of aggregate investment money chasing a diminishing number of public securities.

given the rapid decrease in fixed costs associated with passive management. This trend aligns with the diminishing capacity of larger plans to extract alpha from public securities markets, especially equities, as mentioned above. For the share of equities and fixed income that are managed externally, large plans are more likely to manage equities passively while preferring to manage fixed income investments actively, relative to smaller plans. Among internally managed assets, we find no significant association between plan size and choice of active vs. passive management—suggesting, perhaps, that diseconomies-of-scale in active management offset incremental cost reductions as the size of internal management grows.

Subsequently, we examine investment costs for our sample of pension plans by investment management mandate. Here, we find that the median cost of internally and passively managing stocks is 1-3 bps per year throughout the sample period, while internally and actively managed stocks bear a median cost that fluctuates between 5 and 11 bps per year. The median cost for externally and passively managed stocks hovers between 4 and 8 bps per year, while the median cost of external active management is noticeably higher—between 32 and 48 bps per year. For fixed income holdings, we observe similar patterns. Moreover, we find that external passive management costs have been decreasing over time for stocks and fixed income, converging toward the lower level of internal passive management costs. In contrast, we find no evidence of a convergence in the costs between internal and external active management for these asset classes or, indeed, for the private asset classes—consistent with a change in the composition of external actively managed mandates—i.e., a move toward more specialized strategies.

We find strong evidence of significant economies of scale in investment management costs, and document that these follow a power law as a function of the amount of assets invested by a plan. The associated concave relation between investment management costs and plan holdings are particularly strong for public asset classes. Conversely, for the more labor-intensive private asset classes we find that it is more difficult to reduce average costs as plan size increases. We also find evidence of bigger economies of scale in fees for passively managed than for actively managed investments.

Plans' choice of management style (internal versus external and active versus passive) is likely to be endogenous in the sense that it depends on plan size and asset class characteristics. To account for such confounding effects (and, notably, control for plan size) and get a more direct

estimate of the impact on plans' costs and return performance, we use a difference-in-differences approach that matches plans switching management style (e.g., from external to internal management) with similar plans that retain the same management style. We find strong evidence that plans' management costs unequivocally decrease when they switch from external to internal or from active to passive management, whereas costs increase when switching from internal to external or from passive to active management.

Our results for plan performance are as follows. First, for public asset classes, we find a modest association between plan size and net return performance, with the largest decile of plans (sorted by AUM) outperforming the smallest decile by about 20 bps per year. Next, we find that plan size matters more for alternative asset classes, where the top decile of plans outperforms the smallest decile of plans by about 200 bps per year.[9] Using our matching approach, we find significant evidence that plans' gross and (particularly) net return performance, controlling for plan size, improves among alternative asset classes following a switch from external to internal management, whereas return performance instead deteriorates following the reverse switch from internal to external management. For stocks and fixed income accounts, return performance (both gross and net of costs) improves for both transitions, i.e., from external to internal and from internal to external management. We attribute this to mean reversion in returns since poor prior-year returns is likely to cause a switch in management style. Finally, we find an insignificant effect on return performance for plans switching between active and passive management, consistent with managers setting costs so that the marginal plan is indifferent between active and passive management.

Our paper builds on prior research that finds a positive relation between total plan size and performance. Specifically, Dyck and Pomorski (2011) document that larger plans allocate more to asset classes where their scale is more likely to provide bargaining power with respect to the fees charged by external asset managers, specifically, private equity and real estate. Our analysis generalizes these findings in several ways. First, our empirical analysis focuses on the endogenous choice of plans between internal and external management of their assets, as well as their endogenous choice between active and passive management. Our empirical model allows for

[9]These numbers are based on policy-adjusted returns. We explain in detail how these are constructed in Section 2.6.

separately measuring the probability of a plan to employ internal (or active) management, followed by a measurement of the impact of scale on the level of internal (or active) management—conditional on choosing such mandates. Second, we present results from a matching estimator that allows us to estimate the effect on costs and return performance of plans' choice of internal versus external or active versus passive management, after controlling for plan size and other confounders. Third, we exploit the sub-asset class granularity of our data, and document a power-law relation between size and investment management costs (within a sub-asset class) which more precisely indicates economies of scale in all asset classes, as well as at the plan level, and not just the scale economies in private equity and real estate documented by Dyck and Pomorski (2011). Fourth, we show that economies of scale in costs differ significantly across passive and active mandates, while they are similar for internally and externally managed accounts. Fifth, compared to Dyck and Pomorski (2011), our dataset extends the time series by ten years, enabling us to investigate the time trends in management choices. These trends are critically influenced by the growing scale of DB plans relative to the markets in which they invest, as well as time-series changes in the fixed-costs required to set up internal management.

The remainder of the paper proceeds as follows. Section 2.2 introduces the main features of our data from CEM with additional details provided in Appendix B.1. Section 2.3 develops a set of hypotheses that we set out to test empirically in the subsequent analysis. Section 2.4 covers the determinants of internal versus external and active versus passive investment management decisions. Section 2.5 provides a detailed analysis of the cost data, and Section 2.6 analyzes gross and net-of-cost return performance and how it relates to plan characteristics. Finally, Section 2.7 analyses how cost and return performance relates to investment management mandates (or "styles") and Section 2.8 concludes.

## 2.2 Data and Summary Statistics

We obtain our data from CEM Benchmarking, a Toronto-based company that uses detailed annual surveys to collect data on public and private pension sponsors domiciled both in the U.S. and in a number of other developed-market countries. A key advantage of this dataset is its highly detailed fee/cost data, separated by sub-asset class, as well as by active vs. passive mandates and by internal vs. external management within each sub-asset class. In total, the CEM Benchmarking

database covers 613 U.S. and 524 non-U.S. plans (CEM "PlanIDs") that participated in the survey at some point during our 29-year sample period from 1991 to 2019.[10]

CEM plan surveys in the U.S. and the U.K. are primarily collected from defined benefit (henceforth, DB) pension plans and other similar capital investment pools. Apart from these regions, the type of plans for which the survey is collected is country-specific, such as industry-based DB pools in the Netherlands, buffer funds in Sweden, insurance-backed retirement funds in Finland, or defined contribution plans in Australia. Even though reporting to CEM is voluntary, previous research has found no evidence of self-reporting bias related to performance (Bauer, Cremers, and Frehen, 2010).[11] The self-reported data are checked by CEM for internal (same year) consistency, year-over-year consistency, and outlier reporting. CEM data is biased toward larger plans, yet plans contained in the database are broadly distributed across size (total plan AUM). The aggregate AUM covered by CEM in 2019 is $9.04 trillion, with U.S. plans accounting for $3.81 trillion, and non-U.S. plans holding the remaining $5.23 trillion (using 2019 exchange rates). Some plans only report results for a few years—in some cases only for a single year. However, while roughly 500 plans report to CEM for three or fewer years, 317 plans report to CEM for at least 10 years. This fact, coupled with the large cross-section of plans surveyed by CEM each year (at least since 1999), allows us to analyze a representative sample of worldwide pension plans.[12] Further details on the CEM database, and the mechanism used to collect data from plans, are contained in the Appendix.[13]

The CEM survey collects data on four categories of variables, separately for passively vs. actively managed, and, in turn, for internally vs. externally managed assets within each of six major asset classes (and their corresponding sub-asset classes), namely: stocks, fixed income, hedge funds

[10]The CEM dataset has been used in the past by French (2008), who shows that pension plans shift from active to passive management over time, and Andonov, Kok, and Eichholtz (2013), who document scale-economies for pension plan costs in real estate investments. Broeders, van Oord, and Rijsbergen (2016) looked at scale benefits for Dutch pension plans, using different proprietary data.

[11]From discussions with CEM, the primary reason for funds to leave the survey is turnover in direct contacts with clients, i.e., the personnel of a particular pension plan changes. High-fee plans, predominantly small plans, are less likely to participate in the survey which can be very labor intensive to complete.

[12]Details are provided in Appendix Table B.1. That said, our sample is especially reflective of North American plans. In our empirical results, we point out when differences exist between the early years of our sample and later years—which contain a higher proportion (relative to early years) of plans domiciled outside of North America.

[13]For comparison, according to the Investment Company Institute (2021), in 2019, there were $54.9 trillion of total net assets invested in worldwide regulated open-end funds, with the U.S. accounting for $25.9 trillion, or nearly half, of these investments. The Center for Retirement Research at Boston College (CRR) estimates that U.S. public pension plans held $4.1 trillion of assets in 2019. See https://publicplansdata.org/.

and multi-asset class (jointly), private equity, private debt, and real assets. Included for each of four potential mandate choices within each asset class (e.g., internal active) is the dollar value of assets (using exchange rates for foreign plans), internal management costs or external management fees (AUM-based as well as performance-related), and asset returns, measured both gross and net of fees.[14] A full list of variables is contained in Appendix B.2.2—B.4.

### 2.2.1 Asset Allocation

Figure 2.1 shows the proportion of investments allocated to each of the six major asset classes for U.S. (top panel) and non-U.S. (bottom panel) DB plans. For U.S. plans, the average plan allocation to public equities (stocks) varies between 50 and 60% from the beginning of our sample (1991) until the Global Financial Crisis (2007-2008), after which it drops below 50% of portfolio holdings. These plans increasingly allocate to alternative asset classes by the end of our sample (2019)—from less than 8% in 1991 to almost 28% in 2019. Non-U.S. plans show a similar pattern of asset allocation over time, albeit with lower levels of stock investments.

Even larger shifts have taken place during our sample period in the sub-asset classes that comprise the six main asset classes. Figure 2.2a shows that the allocation of U.S. plans to broad-based U.S. stock strategies (U.S. Broad/All) is 86% at the beginning of our sample, dropping to 18% by 2019. In turn, these U.S. plans allocated more to international stock strategies, such as ACWI ex U.S., EAFE, emerging markets, and global (12% in 1991 versus 58% in 2019), and allocated more to specialized market capitalization strategies such as small, medium and large cap stocks. Trends in fixed-income investments show a similar movement from broad-based to more specialized mandates. Allocations between alternative sub-asset classes exhibit a trend toward greater allocations to hedge funds (Figure 2.2c), LBOs (Figure 2.2d), private credit (Figure 2.2e), and natural resources and infrastructure (Figure 2.2f).

Subsequently, we present results for small and large pension plans, defined as plans below the 30th and above the 70th percentile in total plan AUM each year, in Figure 2.3. This figure includes bar charts for the asset allocation by management mandate within asset classes for the year 2019, with similar results in 1999 and 2009. The three asset classes, stocks, fixed income, and real assets, encompass all four management mandates: internal passive (IP), external passive (EP),

[14]For each asset class, data is subdivided into several sub-asset classes such as U.S. large cap stocks or emerging market stocks, as shown in, for example, Appendix Table B.4.

internal active (IA), and external active (EA). Passive mandates are not available for the remaining alternative asset classes: private debt, private equity, and hedge funds. Large plans exhibit a higher fraction of internally managed assets, both for active and passive mandates, particularly in publicly-traded fixed income and stocks. These asset classes are associated with the lowest fixed and variable internal management costs, making them more conducive for setting up internal asset management.

Table 2.1 reports small and large plans' choice of investment management mandate in the form of the share of plans' AUM within individual sub-asset classes allocated to each of the four management mandates (IP, EP, IA, and EA). Large plans make far greater use of internal active management than small plans. This holds both in public asset classes and even more so among the four private asset classes. Differences can be very large, e.g., with 58% of large plans' assets in global equity being managed internally and actively, versus only 1% for small plans. Small plans also make far greater use of external active investment management than large plans.

### 2.2.2 Investment Management Cost

We now turn to the time series evolution in investment management cost. We measure the aggregate investment management cost in an asset or sub-asset class as the sum of AUM-based fees and performance-related fees, and report it as a percentage of AUM allocated by a particular DB plan to that asset or sub-asset class during a particular year. Further, we scale the median value of this cost by the "grand average" cost averaged across asset classes, plans and years in our data.[15] This generates a new measure of cost, *scaled cost*, which is expressed as a percentage of the average cost. While this scaling does not show the cost level in bps/year, it allows us to interpret time trends in management cost as well as compare costs across different asset classes and management mandates.[16]

First, consider investment management costs for stock holdings (Figures 2.4a and 2.4c). For passively managed accounts (Figure 2.4a), median costs increase over time from 5% to 8% of average costs when internally managed, yet decline from 18% to 9% when externally managed. Hence, by the end of our sample, the cost of internal and external passive management converge, suggesting that passive management has increasingly become a "commoditized" investment management service.

[15] We use this transformation to preserve confidentiality of cost *levels*, as requested by CEM.

[16] For example, a *scaled cost* of 100% implies that the median costs are equal to the average costs in our sample while a value of 50% implies a median cost of half the average cost. Appendix Table B.10 reports scaled costs by asset class and country-of-domicile for plans for selected years.

Active equity management costs are far higher, at 22% of average costs for internally managed accounts, and rising from 81% to 110% of average costs for externally managed accounts. In this case, we do not find any evidence of convergence. Part of the reason for this non-convergence appears to be that plans choose to internalize the active management of the least specialized, lower-cost sub-asset classes (e.g., broad cap U.S. stocks) and conversely externalize the high-cost segments (e.g., emerging market stocks and small cap stocks) which require more specialized knowledge to manage actively. Moreover, as we show subsequently, external fund managers generate positive net-of-fee return performance— especially for the largest pension plans in our sample, serving to reduce pressures on their investment managers to reduce active management fees. We find very similar results for fixed income allocations (Figures 2.4b and 2.4d).

In summary, passively managed assets have become largely commoditized, resulting in lower costs, especially for large plans that have transitioned to internal management. Smaller plans, although still reliant on external passive management, benefit from lower fees due to economies of scale in this domain. External active management fees have, in general, remained durably higher than internal active management costs, due to an increasing specialization of external active managers and the alpha benefits such specialization brings to pension plans.

## 2.3 Empirical Hypotheses

Our paper examines investment management costs and return performance as a function of pension plan scale and plans' choice of investment management style (internal vs. external and active vs. passive). Our primary focus is on understanding how economies of scale affect investment costs, specifically the fixed costs associated with different allocations, such as those to active or passive management, as well as internal or external management.[17]

That is, a key theme of our paper is the role of fixed costs in investment management and its impact on economies of scale for pension plans. We highlight the bargaining power that large plans gain because of their ability to manage investments internally, potentially avoiding the higher costs associated with external managers. In this section, we formulate testable hypotheses drawing from theories of asset management, considering scale economies, uncertainty in active management

[17]Fixed costs include the costs of setting up a management "shop", such as the costs of office space, datasets, and human capital, both for internal and external management—but, also, the search costs of plans in locating skilled active external managers.

skills, the cost of information acquisition, and heterogeneity in investors' abilities to identify skilled managers

### 2.3.1 Theories of Asset Management

**Berk and Green (2004)**

A useful starting point for our inquiry is the seminal paper of Berk and Green (2004) (BG). In mutual fund markets, BG propose an equilibrium model that starts with the assumption of homogeneous diseconomies of scale among funds in their investments in financial markets. Given the implied absence of any differential bargaining power of (atomistic) mutual fund investors in their model, mutual funds, in equilibrium, grow to a size at which their diseconomies result in zero expected net-of-fee alphas.[18]

In our setting that includes some very large pension plans as well as a finite set of small plans, we can expect important deviations from this idealized BG "no-frictions equilibrium" outcome. Specifically, the differential bargaining power of individual plans cannot be dismissed, and brings many interesting features to the competitive market for investment managers who cater to such plans.[19] That is, the industrial organization of the market for delegated investment management in the pension fund industry is far more complex than that modeled by BG for the mutual fund industry, with outcomes depending on issues such as the relative bargaining power of plans and managers.[20]

For example, the median plan in the CEM database (based on U.S. equity dollar allocations), in 2019, contains U.S. equity investments totalling $1.01 billion, while the $10^{th}$ and $90^{th}$ percentile plans oversee $107.0 million and $9.25 billion, respectively. While U.S.-domiciled equity mutual funds exhibit a similar dispersion in size, most investors in mutual funds have a relatively small

[18]That is, open-end mutual fund managers allow their funds to grow to a size that leaves zero expected alphas, net-of-fees, to rational atomistic investors in their funds—but that maximizes fund manager fee income. So, in BG's implied setting of a limited number of truly skilled asset managers, all of the expected rents (ex-ante alphas) accrue to investment managers, since the infinite pool of investors (supply of capital) competes away any net-of-fee performance through their inflows to funds rationally inferred to be overseen by skilled managers.

[19]BG allow for a limited role for the fixed costs of active investment managers, i.e., in modeling the decision of such managers to continue operations or to shut down. In our setting, the fixed-costs of investment management, both active and passive, as well as internal vs. external management, are central to the ongoing choices made by pension plans. This distinction implies that there exist scale economies at the plan level which affect both investment costs and allocations to external vs. internal management, abstracting from the choice of investment managers to discontinue their operations.

[20]Prior papers on pension plan choice of investment managers (and their dismissal) do not focus on the role of plan scale in such manager choice and negotiated fees (see, e.g., Blake, Rossi, Timmermann, Tonks, and Wermers (2013), Rossi, Blake, Timmermann, Tonks, and Wermers (2018) and Beath, Flynn, Jethalal, and Reid (2022)).

investment and can be considered "atomistic"—that is, they have an insufficient ownership fraction to incentivize or to empower them to negotiate fees with their fund managers.[21]

Thus, while net-of-fee scale diseconomies may be relatively homogeneous among mutual fund investors, mostly due to their limited (collective) negotiating power, such diseconomies can be expected to be much more diverse among pension plans. Small plans might be expected to hold little power to negotiate with their investment managers due to their high fixed costs of search and internal management (per unit of AUM), and, accordingly, may face qualitatively similar net-of-fee diseconomies of scale as investors in mutual funds; in contrast, large plans might use their bulk to reduce diseconomies, or to potentially reverse them and to realize positive scale economies in investment management fees.

**Gârleanu and Pedersen (2018)**

Addressing some of these limitations of the BG model, Gârleanu and Pedersen (2018), henceforth GP, develop a general equilibrium model for assets and asset management in the presence of fixed costs that pose a friction for all investors (in our setting, for DB plans). The GP model introduces delegated investment management with uninformed and informed managers, where the latter receive a signal that is correlated with returns on a risky asset, as in Grossman and Stiglitz (1980). Importantly, the true manager type (informed vs. uninformed) is unobserved by investors, and a fixed search cost must be paid to help identify skilled investment managers.[22]

Our pension plan setting shares similarities with GP's deviation from the BG model. Most importantly, we observe wide heterogeneity in asset allocations among pension plans, impacting their capacity and motivation to cover fixed costs associated with external manager search or internal management setup. Large plans with billions of dollars to invest and many experienced professionals can better handle the fixed costs of internal management and are expected to be more

[21] Among U.S.-domiciled domestic equity mutual funds, the median fund manages $514 million, while funds at the $10^{th}$ and $90^{th}$ percentiles, respectively, manage $38.7 million and $6.8 billion, respectively, at the end of December 2019. For comparison, the median amount invested in mutual funds by U.S. households was $200,000 in 2021 (Investment Company Institute, 2021). We recognize that fiduciaries of large defined contribution (DC) plans—some of which hold greater than $1 billion in AUM—might hold some bargaining power with their investment managers (see, for example, Sialm, Starks, and Zhang (2015)). However, large DC plans hold levels of AUM that tend to pale in comparison with that of large DB plans.

[22] Investors have the option of either investing their money directly (passively) and, thus, foregoing the search cost, or searching for an informed manager who will charge a fixed investment fee for actively managing investor assets. The size of this fee is modeled through Nash bargaining between the manager and investor. This feature of the GP model suggests that investors' bargaining power should matter to their choice of investment mandate as well as to investment alphas and fees.

capable of identifying skilled managers.[23] Conversely, small plans will neither have the incentive to undertake costly search, nor to establish internal management, leading to distinct choices between external and internal management. Thus, the choice between external and internal management will be indicative of the fixed costs of internal vs. external management, especially among large pension plans. Small plans can be expected to choose the "corner solution" of no internal management.

The model of Gârleanu and Pedersen (2018) can also be used to compare outcomes in private asset markets, such as real estate and private equity, which display high search and information costs, versus more transparent and lower information-cost asset markets for publicly traded securities, such as stocks and bonds. Specifically, in asset markets with lower search costs for locating skilled active managers, as well as lower information acquisition costs for such managers, the increased competition among active managers both reduces the average active "alpha" (before fees), and applies pressure on active managers to reduce fees. In the face of these shrinking fees, we can expect to see more specialization in active management, as we have described in Section 2.2.2. With these developments, it naturally follows that passive management gains market share among less-specialized investment strategies relative to active management, when search costs are low and information acquisition is less costly.

Conversely, search costs tend to be much higher in the market for managing private assets, as well as less efficient public-market assets, and only investors with the capacity for undertaking a sophisticated search process (i.e., low search costs relative to AUM) might hire active managers in these markets. Information acquisition costs (paid by active investment managers) also tend to be higher in these markets, and prices are less efficient due to the higher cost of entry and the resulting weaker competition among informed managers. To cover their higher information acquisition fixed costs, investment managers also charge higher fees in private asset markets. In equilibrium, we would expect larger pension plans to be more willing to engage with skilled managers in private markets, in part because of their enhanced ability to locate skilled managers as well as the negotiating power that large plans possess. That is, the market for private investments can be expected to be less important for small pension plans, as they are unable to bargain for positive

[23]To be sure, large plans are more likely to have access to the most skilled managers due to the greater fee income that they potentially bring, which can compound the advantages that their greater manager search capabilities bring. In this paper, we focus on the bargaining power possessed by large plans due to their enhanced ability to "internalize" investment management.

expected risk-adjusted returns.

In turn, informed managers in private asset markets will tend to earn higher fees due to the high fixed-costs of search and internalization by pension plans. Similarly, to compensate investors for the higher cost of searching for managers of private asset classes, we expect to find higher abnormal returns, net of fees in private markets—among those large plans that have bargaining power.

### 2.3.2 Plan Size and Choice of Investment Management Style

Based on the discussion of the key factors determining pension plans' choice of investment management mandate, asset allocation decisions, and investment performance as well as costs, we now articulate a set of hypotheses that discipline our subsequent empirical analysis. Our first hypothesis involves plan size and the corresponding choice of investment management mandate (internal vs. external management, and active vs. passive management).

We believe that large plans have much stronger incentives for internal investment management due to the fixed costs of asset management in all asset classes.[24] Consequently, large plans are expected to allocate a higher proportion of assets internally within asset classes where their investments are larger, and the cost advantage of internal management over external management is more pronounced.

A second dimension is the choice of active vs. passive management. In the Gârleanu and Pedersen (2018) model, small investors with high manager search costs typically choose passive investment due to the significant fixed cost of searching for skilled managers. Conversely, large investors, with a more favorable cost-benefit profile due to their higher search capacity and assets, are inclined to seek out active managers. This leads to the expectation that small plans would allocate more of their assets passively, while large plans would favor active management. This tendency might be more pronounced in private asset classes, whereas the situation in public asset markets may differ. The combined effect of a smaller scope for generating abnormal returns and larger diseconomies of scale in the highly competitive stock and bond markets may incentivize large plans to make greater use of internal active rather than external active management in order to

[24] We recognize that fixed costs are likely smaller in some asset classes and strategies than others in our discussion to follow. For example, fixed costs in managing U.S. equities, passively, is likely to be lower than other asset classes/strategies.

reduce investment management costs.

As plans grow in size, they tend to explore alternative asset classes as they exhaust opportunities in the crowded public equity and fixed income markets. Plan size also plays a role in the available choices within each asset class. Larger plans can leverage their size to negotiate more favorable terms, including lower fees in alternative asset classes that might be less accessible to smaller plans. Consequently, we anticipate a positive relation between plan size and allocations to private asset classes, while expecting a negative association between plan size and investments in public stocks and bonds. We summarize these relations between investment choices and plan size in the following hypothesis:

**Hypothesis I** (Plan size and investment management)**.** *Large pension plans, relative to other plans:*

*(i) manage a bigger fraction of their assets internally, measured across all asset classes.*

*(ii) have a higher probability of switching from external active management to internal active management in public asset classes.*

*(iii) allocate a larger fraction of their portfolios to private asset classes and, correspondingly, a smaller fraction of their portfolios to public assets (stocks, bonds).*

### 2.3.3 Economies of Scale in Investment Management Costs

Economies of scale matter in investment management because many costs, such as legal, data, and computing expenses, are either fixed or do not increase proportionally with assets under management. This suggests an inverse relation between a plan's holdings in a specific asset class and the average costs of managing it, meaning that larger plans typically experience lower costs and fees per dollar invested compared to smaller plans. Still, larger plans may also face higher costs due to the need for additional personnel and increased transaction expenses when dealing with larger investment amounts. Investment management costs can also vary depending on the labor-intensity of different investments, influenced by factors such as asset class liquidity and transparency.

To better understand scale economies in investment management costs, we examine the power law framework developed by Gabaix (2009, 2016), positing that dollar management costs,

$\text{Cost}^{\$}$, follow a power law as a function of AUM:[25]

$$\text{Cost}^{\$} \propto \text{AUM}^{\beta}. \tag{2.3.1}$$

Power law coefficients $\beta < 1$ are consistent with economies of scale in investment management costs, and the smaller is $\beta$, the bigger the cost economies of scale. Conversely, $\beta > 1$ suggests diseconomies of scale since increasing AUM by a certain factor leads to disproportionately higher management costs.

We use the posited relation in (2.3.1) to formulate a set of hypotheses on economies of scale in investment management. Our most basic hypothesis is that costs grow less than proportionately with assets under management, i.e., $\beta < 1$. Our next cost hypothesis is that investment management costs vary systematically across public and private asset classes. Specifically, we would expect greater cost economies of scale for public asset classes such as stocks and fixed income ($\beta^{public}$) that are traded in transparent and liquid markets than for private asset classes ($\beta^{private}$) which typically involve more labor-intensive (less computerized) processes that are harder to scale up.

Scale economies in costs are also likely to be linked to management mandate, so we analyze the cost-size relation at the asset class level for the four different mandates, namely, Internal Passive (IP), Internal Active (IA), External Passive (EP), and External Active (EA).[26] Passive investment management has largely become commoditized in a way that facilitates scaling more easily than the labor intensive active investment management process. Moreover, besides lower per-dollar human-capital costs, large passive management funds can implement trading strategies that enhance their returns, such as securities lending and favorable per-dollar trading terms with prime brokers, relative to smaller passive funds. Hence, our third cost hypothesis is that passive investment management lends itself more easily to scaling than active management, in part because it is associated with lower market impact.

For both internal and external management to coexist within a specific asset class, and to align with the empirical observation that not all plans exclusively manage their assets either

[25]Two variables $X$ and $Y$ are said to be related via a power law if $Y = cX^{\beta}$, where $c$ is an arbitrary constant. Gabaix (2009, 2016) suggests that power laws are ubiquitous among economic variables such as firm or city size, income, and wealth. While these power laws typically hold primarily in the tails of the distribution, we find the assumption plausible across the entire distribution (see Figure 2.5).

[26]For private assets, we focus on active management mandates only, since the vast majority of such assets are actively managed.

internally or externally, we propose our fourth cost-scale hypothesis. This hypothesis tests whether economies of scale for both management mandates (internal vs. external) are equal, i.e., that there is an equilibrium where pension plans optimally decide whether to use internal or external investment management for a given asset class. Additionally, this equilibrium assumes that identical scaling technologies are applied in both internal and external asset management.

**Hypothesis II** (Economies of scale in investment management costs)**.** *In the context of the power law relation in (2.3.1), the following holds:*

*(i) Pension plans' investment management costs display significant economies of scale and exhibit a concave relation to AUM:* $\beta < 1$.

*(ii) Economies of scale in the cost of investment management are greater for publicly traded assets than for private asset classes:* $\beta^{public} < \beta^{private}$.

*(iii) For each asset class, and for both internally and externally managed accounts, passive investment management offers better economies of scale than active management:* $\beta^{IP} \leq \beta^{IA}$ *and* $\beta^{EP} \leq \beta^{EA}$.

*(iv) For each asset class and management mandate (active or passive), the economies of scale cost parameter is identical for internally and externally managed assets:* $\beta^{IP} = \beta^{EP}$, *and* $\beta^{IA} = \beta^{EA}$.

### 2.3.4 Plan Size and Return Performance

Our final set of hypotheses is concerned with how return performance, both gross and net of fees, varies across plan size, investment mandate, and asset class. Plan size can have both a positive and a negative impact on investment performance. In particular, large plans have more resources to search for skilled managers and monitor their return performance on a continual basis, allowing them to better reduce the challenge of plan scale in generating higher gross returns (before fees). Conversely, AUM can have a negative effect on gross returns as managers with more money to invest run out of ideas. Importantly, though, this mechanism is most relevant for externally managed assets.

Plan size will further impact investment performance net of fees positively if there are sizeable economies of scale in the cost of investment management. That is, the existence of significant

fixed costs in asset management gives large plans a distinct advantage, especially in active management and in private asset classes. Large plans can also be assumed to possess more bargaining power which they can use to negotiate more favorable terms with external managers. Moreover, because large plans are more likely to have internal asset management capabilities, they can use this as a credible "threat" or reservation point in negotiations with external managers.

Asset classes matter to this relation because of the large differences in acquiring information and managing investments in public and private asset classes and even within these broad categories. Information costs are generally much higher for private assets such as real estate, private equity, and venture capital. Competition among managers of private assets is also not as fierce as that for public asset classes such as stocks and fixed income which offer passive investment products that help bound how high investment management fees can go.

**Hypothesis III** (Plan size and return performance)**.**

*(i) The largest plans earn positive investment returns both before and after fees, i.e., gross and net return performance is an increasing function of plan size.*

*(ii) Net-of-fee returns are particularly strongly positively related to plan size for private asset classes.*

We next set out to test these hypotheses more formally, beginning with plans' choice of investment management styles (Section 2.4), moving on to investment management costs (Section 2.5), and finishing with return performance (Section 2.6).

## 2.4 Investment Management Mandates

This section examines the impact of plan, manager, and asset characteristics, including plan size (AUM), investment management costs, and plan domicile, on the choice of investment management mandate. Specifically, it assesses whether plans opt for internal or external asset management, and whether they favor active or passive investment management. Our analysis performs a set of regressions that use as dependent variable the proportion of investments in asset class $A$, in a given year, $t$, that is managed by plan $i$ in a certain strategy, denoted $\omega_{iAt}$ and defined in more detail below. For example, $\omega_{iAt}$ can denote the proportion of investments in asset class

$A$ that are managed internally. We regress this proportion on a set of covariates, $x_{iAt}$, as well as asset-class and time fixed effects, $c_A$ and $\lambda_{At}$:

$$\omega_{iAt} = c_A + \lambda_{At} + \beta_A' x_{iAt} + \epsilon_{iAt}. \tag{2.4.1}$$

In practice, internal management involves substantial fixed-cost investments, including hiring compliance staff and traders, IT system setup, database subscriptions, and hiring skilled investment analysts. Many plans, especially smaller ones, allocate zero assets to internal management due to these fixed costs. Similarly, it is uncommon for plans or external managers to manage alternative asset classes passively. The panel regression in (2.4.1) does not account for the presence of many "zeros" in the data. It focuses on estimating plan choices between management mandates (internal vs. external or passive vs. active) at the intensive margin. However, this approach may introduce model misspecification because variables like plan size and management costs likely influence both the *extent* to which a plan manages assets internally and whether it chooses internal management for *any* of its assets.

To deal with the large number of zeros and to obtain an estimate that accounts for plans' choice along both the intensive and extensive margins, we use the Cragg (1971) estimator. This estimator consists of two equations, namely (i) a selection equation that estimates the probability that a plan's allocation choice lies on the boundary (e.g., zero internal management); and (ii) an outcome equation that estimates the effect of a variable on the proportion of assets managed internally for plans with at least some internal management in that asset class. More formally, the regression model we estimate takes the form:

$$\omega_{iAt} = s_{iAt} h^*_{iAt},$$

$$s_{iAt} = 1\left[\gamma' x_{s,iAt} + \varepsilon_{iAt} > 0\right], \tag{2.4.2a}$$

$$h^*_{iAt} = \exp\left(\lambda_{At} + \beta' x_{o,iAt} + e_{iAt}\right), \tag{2.4.2b}$$

where $s_{iAt}$ is a selection indicator that depends on $x_{s,iAt}$ (covariates influencing selection) and $h^*_{iAt}$ denotes the choice or outcome variable that depends on $x_{o,iAt}$. If the selection indicator equals

zero, the dependent variable $\omega_{iAt}$ will also take a value of zero and, hence, lie on the boundary.[27]

Assuming that the error terms $\varepsilon_{iAt}$ and $e_{iAt}$ in (2.4.2a) and (2.4.2b) are independent normal random variables with marginal distributions $\varepsilon_{iAt} \sim N(0,1)$ and $e_{iAt}|x_{o,iAt} \sim N(0,\sigma^2)$, the conditional expectation of $\omega_{iAt}$ given the variables $x_{s,iAt}, x_{o,iAt}$ simplifies to

$$\mathbb{E}\left(\omega_{iAt}|x_{s,iAt},x_{o,iAt}\right) = \Phi\left(\gamma' x_{s,iAt}\right)\exp\left(\lambda_{At} + \beta' x_{o,iAt} + \frac{\sigma^2}{2}\right), \tag{2.4.3}$$

where $\Phi(\cdot)$ is the CDF of the standard normal distribution.

To gauge the effect of changing a single variable, $x$, on the expected value of $\omega_{iAt}$, we examine the average partial effect (APE) of $x$:

$$\mathrm{APE}_x(x_{s,iAt},x_{o,iAt};\gamma,\beta) = \left.\frac{\partial \mathbb{E}\left(\omega|x_s,x_o\right)}{\partial x}\right|_{x_s=x_{s,iAt},x_o=x_{o,iAt}}. \tag{2.4.4}$$

Since the expectation in (2.4.3) depends on both the selection and outcome equations, the APE in (2.4.4) accounts for both the intensive and extensive margin effects of changing $x$ and so depends on both $\gamma$ and $\beta$. Letting $\widehat{\gamma}$ and $\widehat{\beta}$ denote the maximum likelihood estimates, we can compute the sample APE as

$$\widehat{\mathrm{APE}}_x = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathrm{APE}_x(x_{s,iAt},x_{o,iAt};\widehat{\gamma},\widehat{\beta}). \tag{2.4.5}$$

Intuitively, $\widehat{\mathrm{APE}}_x$ captures the average effect of changing $x$ while holding all other variables constant.

### 2.4.1 Internal versus External Management

To examine the determinants of plans' decision on managing investments in a given asset class internally or externally, we estimate models for the proportion of plan $i$'s allocation to asset class $A$ that is internally managed in year $t$, $\omega_{iAt}^{internal} \coloneqq \mathrm{AUM}_{iAt}^{internal}/\mathrm{AUM}_{iAt}$, where $\mathrm{AUM}_{iAt}^{internal}$ and $\mathrm{AUM}_{iAt}$ refer to the internally managed and total AUM of plan $i$ in asset class $A$ of year $t$.

We consider the following variables. First, to capture plan size, we include $\log(\mathrm{AUM}_{it-1})$, the logarithm of the total dollar value of plan $i$'s assets under management (AUM) in year $t-1$.[28]

[27] This model is more flexible than a standard Tobin (1958) model, since the variables determining selection (extensive margin) can be different from the variables driving the outcome (intensive margin) equation. Moreover, since $\gamma$ and $\beta$ are decoupled, the effect of a variable on the selection and outcome equations can also be different.

[28] Plan AUM is typically measured at the end of the year.

Second, we include the lagged spread in the cost of external versus internal management in asset class $A$ measured in basis points ($\text{CostSpread}^{E-I}_{iAt-1}$). Third, we include a dummy that takes a value of one for non-U.S. plans and is zero otherwise ($\text{nonUS}_i$) and a dummy that takes a value of one for private plans and is zero otherwise ($\text{Private}_i$). Finally, we include asset class fixed effects, $c_A$, and year fixed effects, $\lambda_{At}$, leading to the model:

$$\begin{aligned}\omega^{internal}_{iAt} = c_A + \lambda_{At} + \beta_{1,A}\log(\text{AUM})_{it-1} \\ + \beta_{2,A}\text{CostSpread}^{E-I}_{iAt-1} + \beta_{3,A}\text{Private}_i + \beta_{4,A}\text{nonUS}_i + \epsilon_{iAt}. \end{aligned} \tag{2.4.6}$$

Table 2.2 reports our regression results. To retain a parsimonious specification for the Cragg estimator, we include only the log-size and cost spread between external and internal in the selection equation (2.4.2a) whereas in the outcome equation (2.4.2b) we further include time fixed effects and the dummies for whether a plan is private or public and domiciled inside or outside the U.S. Our estimates of average partial effects are shown in columns to the right of the panel estimates in the table.

Across all asset classes, our estimates show that larger plans employ internal management to a significantly greater extent than smaller plans, consistent with Hypothesis I((i)). For instance, our panel estimates in Panel A of Table 2.2 indicate that a 10% increase in plan size is associated with roughly a one percent increase in the proportion of the plan's stock portfolio that is managed internally (0.83% and 1.14% for the panel and Cragg estimates, respectively). A 10% increase in plan size is associated with a comparable but slightly bigger increase in the proportion of the plan's fixed income portfolio that is internally managed (1.10% and 1.77%).

For alternative asset classes we continue to see a significant association between plan size and the proportion of those asset classes that is managed internally, but the effects are generally not as strong as for stocks or fixed income, with the exception of private debt.

Our Cragg estimates on log-size are notably larger than the corresponding panel estimates for both stocks and fixed income. This finding can be attributed to the fact that plan size increases both the proportion of assets managed internally for plans already using internal investment management and the likelihood of plans transitioning from *no* internal management to *some* internal

management. This highlights the importance of explicitly accounting for selection effects.

Panel B in Table 2.2 verifies this point by reporting estimates from the Cragg selection regression. The table quantifies the effect of lagged AUM and the cost spread on the probability that plans manage at least some of their investments in a given class internally. The first row of estimates shows that plan AUM in a given asset class is a highly significant determinant of the probability that a plan manages some of its assets internally within the asset class. All coefficient estimates on log-size are positive, so larger plans are significantly more likely to manage some of their assets internally, regardless of asset class. In contrast, the external-minus-internal cost spread appears to be a far less important determinant of plans' decision on whether to employ internal asset management and this variable is only statistically significant for one asset class (Hedge funds and multi assets).

The lower panel in Table 2.2 illustrates the importance of these estimates by reporting the probability that a plan manages some of its assets internally as we vary the plan size from the 10th through the 50th and 90th percentiles of the 2019 AUM distribution. We keep the cost spread at its average value in these calculations, although this is not important given that the cost spread does not have a big effect on the results. For stock holdings, we find that small plans (in the 10th percentile of the AUM distribution) have a 13% chance of managing some of their stock portfolio internally. This rises to 34% for medium-sized plans and to nearly 66% for plans in the 90th percentile of the size distribution. Hence, large plans are five times more likely to manage some of their stock holdings internally than small plans. Similarly, large plans are almost three times more likely to manage some of their fixed income holdings internally than small plans (72% versus 29%).

Small plans rarely manage private assets internally, except for real assets (12.82%). Specifically, the Cragg probability estimates vary from 0.50% to 9% for plans located at the 10th percentile of the size distribution. These probability estimates rise notably to between one-tenth (10.59% for hedge funds) to one-half (52.56% for private debt) for the largest plans, i.e., those in the 90th percentile of the size distribution.

These estimates are all consistent with Hypothesis I((i)). Moreover, our estimates are also consistent with relatively modest fixed costs of setting up internal management shops in stocks and bonds, as compared to doing so for alternative asset classes (such as private equity) that require

more specialized skills and knowledge, as well as more costly connections to external sources of information. Consequently, it is rare for small plans to manage their alternative assets internally.

To summarize, our findings suggest that plans' decision to overcome the hurdle of managing at least some of their investments in a given asset class internally is mainly determined by plan size, whereas the cost spread (external versus internal) is not as important (of course, interpreted with caution due to the endogeneity of internal management as a function of costs). Conversely, for plans that have decided to manage some of their assets internally, the cost spread is important for how large a proportion of their assets they manage internally.[29]

We now turn to the estimation of average partial effects. Equation (2.4.3) shows that the average partial effect is a nonlinear function of the covariates. The effect of changing plan size or the cost spread therefore depends on the initial value from which the variable is changed. Specifically, consider the APE when a specific covariate, $x_{iAt}$, takes on the value $\xi$:

$$\widehat{\text{APE}}_x(\xi) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\text{APE}_x(\widetilde{x}_{s,iAt}, \widetilde{x}_{o,iAt}; \widehat{\gamma}, \widehat{\beta}), \tag{2.4.7}$$

where $\widetilde{x}_{o,iAt} = [x_{o,iAt} \backslash x_{iAt}, \xi]$ is the vector of variables in the outcome equation with $x_{iAt} = \xi$ and $\widetilde{x}_{s,iAt}$ is defined similarly. For example, (2.4.7) can used to calculate the APE when $\log(\text{AUM})_{it-1}$ is set at its 10th percentile. To examine if nonlinearities are economically important, we evaluate partial effects at the 10th and 90th percentiles and test if the difference between the two estimates is statistically significant.

Results from this analysis applied to the size and cost spread variables are presented in Panel A of Table 2.3. For the size variable, the APE is larger by an order of magnitude for the largest plans (90th percentile) than for the smaller ones (10th percentile). For example, going from a plan with a small stock portfolio (10th percentile) to a plan with a large stock portfolio (90th percentile), our estimates suggest that the two plans will increase the proportion of their internally managed stocks by 0.38% and 3.19%, respectively. Moreover, these differences are statistically

[29]For stocks, as the cost spread increases by 100 bps, the allocation to internal management is predicted to increase by 12.2% (15.3%) based on the Cragg (panel) estimates. For fixed income this effect is bigger at 30.3% (23.8%). The estimated coefficients on the cost spread have the wrong sign and fail to be significant for private equity and private debt. Conversely, the effect is positive for hedge funds and real assets, but insignificant. However, the choice of internal management is endogenous to cost differences, and the composition of externally managed assets can change when internal management is employed, thus making a clean interpretation of the cost spread coefficient difficult.

significant for the public asset classes, and positive for all private asset classes. Hence, big plans are disproportionately more likely to move public assets from external to internal management as they grow larger.

For the cost spread, plans that pay the highest costs (90th percentile) for management of their fixed income assets are significantly more likely to move assets from external to internal management than the plans that pay the lowest costs (10th percentile). This makes good sense since the low-cost group has a weaker incentive to switch from external to internal management as they already pay relatively low costs. For the other asset classes, the APE of the cost differential fails to be significant.

### 2.4.2 Active versus Passive Management

We next use our framework to examine the determinants of plans' decisions to use active or passive management in different asset classes. Let $\omega_{iAt}^{active} \coloneqq \text{AUM}_{iAt}^{active}/\text{AUM}_{iAt}$ be the fraction of investments in asset class $A$ that is actively managed by plan $i$ in year $t$. We use this as our dependent variable in a set of panel regressions

$$\begin{aligned}\omega_{iAt}^{active} = c_A + \lambda_{At} + \beta_{1,A}\log(\text{AUM})_{it-1} \\ + \beta_{2,A}\text{CostSpread}_{iAt-1}^{A-P} + \beta_{3,A}\text{Private}_i + \beta_{4,A}\text{NonUS}_i + \epsilon_{iAt},\end{aligned} \tag{2.4.8}$$

with all variables previously defined, except $\text{CostSpread}_{iAt-1}^{A-P}$, which denotes the basis point spread between the cost of active and passive management for plan $i$ in asset class $A$ at time $t-1$. Because the vast majority of plans do not use passive management in alternative asset classes, we only have sufficient data to report estimates for stocks, fixed income and real assets.

Table 2.4 reports estimates from the panel regressions in (2.4.8) as well as for the Cragg estimator using the same format as in the previous subsection. For stock portfolios, we find that larger plans manage a significantly higher proportion of their stock holdings passively, as both the panel and Cragg estimates of the coefficients on log-size are negative and statistically significant. For every 10% increase in plan size, the proportion of stock holdings managed passively increases by about 0.4%. Also, a higher spread in the cost of managing stocks actively rather than passively is associated with a large and highly significant negative effect on the proportion of stock holdings

managed actively. Specifically, the Cragg estimate suggests that raising this cost spread by 100 bps is associated with a 33% increase in the proportion of plans' stock holdings that are passively managed. Private plans also manage significantly more of their stock portfolios actively than public plans do.

For fixed income holdings, we find some evidence that larger plans manage a slightly higher proportion of their assets actively, as the Cragg APE estimate of log-size is significantly positive. However, the effect is small and the panel estimate is insignificant. The Cragg APE estimate on the cost spread is highly significant and negative, suggesting that higher active management costs, measured relative to passive management costs, lead plans to significantly increase the proportion of their passively managed fixed income holdings.[30]

Plans can mitigate the higher costs of active investment by transitioning from external active management to internal active management, which is typically more cost-effective. This strategy is likely to be particularly appealing for the largest plans with the greatest capacity for overcoming the fixed costs of setting up internal management. To explore if this holds in our data, we examine the proportion of plans' actively managed stock and bond portfolios that are managed internally in the right panel of Table 2.4.

For both stock and fixed income holdings, we find that the estimated coefficient on log-size is positive and highly significant. Hence, an increase in plan AUM is associated with a significantly higher allocation to that part of plans' actively managed portfolios that is managed internally, consistent with Hypothesis I((ii)). A higher spread in the costs of active versus passive management also leads to plans internalizing more of their actively managed stock and fixed income portfolios.

Table 2.4 also reports estimates on dummies for whether a plan is private or public and whether a plan resides in the U.S. or outside the U.S.. Our Cragg estimates show that private plans tend to allocate approximately 4.8% more of their stock holdings to active management than their public peers while we find no significant difference between U.S. and non-U.S. plans. For fixed income holdings, we find that non-U.S. plans manage 9.8% less of their fixed income holdings actively than U.S. plans.

While many plans mix different management mandates at the asset class level (e.g., external

[30]For real assets, we find a negative relation between plan size (AUM) and the proportion of actively managed assets, but the estimated effect is small and insignificant. This finding reflects that very few plans in our dataset manage real assets passively and, for those that do, predominantly in one sub-asset class, namely REITs.

active and external passive management of their stock portfolio), they mostly choose a single management mandate at the sub-asset class level. In other words, it is common to find asset managers that employ internal passive management for their U.S. Large cap portfolio and external active management for their emerging market stock portfolios, but it is rare to see managers that simultaneously employ different management mandates for their U.S. large cap portfolio. In those cases where plans mix multiple management mandates for a particular sub-asset class, this tends to be done exclusively by the largest plans.

Finally, we consider again the APE estimates evaluated at different percentiles of the size and cost spread distribution. Our estimates are presented in Panel B of Table 2.3. Most of the differences in APE estimates are statistically insignificant. Plans paying the highest costs for active equity management (90th percentile) are more likely to move their stock holdings from active to passive management than are plans paying the lowest costs (10th percentile)—a point that is driven by externally managed holdings. Interestingly, the rightmost column in the table shows that, for fixed income investments, the largest plans (90th percentile) are considerably more likely than smaller plans (10th percentile) to switch from external active management to internal active management as they grow larger.

### 2.4.3 Asset Allocation Decisions

To examine if plans' asset allocation decisions are consistent with Hypothesis I((iii)), we conduct a set of panel regressions that use as dependent variable the weight of asset class $A$ for plan $i$ in year $t$, $\omega_{iAt} = \text{AUM}_{iAt}/\text{AUM}_{it}$, which we model by:

$$\begin{aligned}\omega_{iAt} = c_A + \lambda_{At} + \beta_{1,A}\log(\text{AUM}_{it-1}) + \beta_{2,A}\text{Cost}_{iAt-1} \\ + \beta_{3,A}\text{Private}_i + \beta_{4,A}\text{nonUS}_i + \beta_{5,A}\text{LiabilityRetiree}_{it} + \epsilon_{iAt}.\end{aligned} \tag{2.4.9}$$

The list of regressors is similar to that adopted earlier with two exceptions. First, our cost variable ($\text{Cost}_{iAt-1}$) is now the lagged per-dollar cost for plan $i$ in asset class $A$ measured as a fraction of AUM and denoted in bps. Second, we also control for liability-related effects on asset allocation decisions by including $\text{LiabilityRetiree}_{it}$, the fraction of a plan's total liabilities owed to retirees.

We do so because plans are likely to consider their liability structure when deciding how much to allocate to asset classes with different risk characteristics. For example, more mature plans may allocate a larger fraction of their portfolio to fixed income. Because only a subset of plans report data on liabilities, including this as a covariate results in a substantial decline in sample size. We therefore report in Table 2.5 results both with (Panel B) and without (Panel A) this variable included.

We find evidence largely consistent with our empirical prediction as larger plans allocate significantly less of their portfolios to stocks, fixed income, hedge funds, and private debt. Conversely, they allocate a significantly greater share of their investments toward private equity and real assets. These findings hold regardless of whether we estimate our panel regressions on the larger sample (top panel), or on the smaller sample that controls for plan liabilities (bottom panel).[31]

Investment management costs are also an important driver of plans' asset allocation decisions. We obtain negative and highly significant estimates on the cost variable for five of the six asset classes with the sixth (private debt) being insignificant. Coefficient estimates vary greatly across asset classes; by far the highest estimate is obtained for fixed income (-20.10) and stocks (-7.11) with smaller estimates for hedge funds and multi assets (-1.65) and, in particular, real assets (-0.44) and private equity (-0.11). In contrast, the estimates of the LiabilityRetiree variable in the bottom panel of Table 2.5 are statistically insignificant across all asset class specifications, except for stock investments.

## 2.5 Investment Management Costs

Our above results indicate that plan size plays an important role in determining the choice of investment management mandate. In concert with these choices, plan size is likely to be a key determinant of investment management fees/costs, as larger plans benefit from internal management scale economies and possess greater bargaining power to negotiate external management fees. This section explores the role of plan size in determining investment management costs across different asset classes and investment management mandates. Our focus is on how larger plans can use the threat of internal management to establish bargaining power with external managers, particularly in asset classes with relatively low fixed costs of setting up such management.

[31] Our estimates are consistent with the findings reported in Dyck and Pomorski (2011).

Taking logs in the power law equation in (2.3.1), we obtain a linear relation between the log-cost and log-AUM whose slope measures the economies of scale coefficient, $\beta$. To see if this is a suitable characterization of the cost-size relation in our data, Figure 2.5 provides log-log plots of AUM versus costs for stocks and fixed income portfolios across the four investment management mandates. These plots suggest that the power law provides a good approximation to the cost-size relation. The slope is notably flatter for passively managed portfolios than for active ones consistent with larger economies of scale (lower $\beta$) for passive than for active management of both stock and fixed income accounts.

Generalizing the power law relation in (2.3.1) to allow for additional determinants of costs, we examine the following model[32]

$$\text{Cost}^{\$}_{iats} = (\text{AUM}_{iats})^{\beta_{As}} \exp\left(c_{As} + \lambda_{Ats} + \gamma_{1,As}\text{Private}_i + \gamma_{2,As}\text{nonUS}_i\right) \exp(\varepsilon_{iats}), \tag{2.5.1}$$

where $\text{Cost}^{\$}_{iats}$ ($\text{AUM}_{iats}$) is the dollar cost (AUM) of plan $i$ in sub-asset class $a$ at time $t$ for mandate $s$, $c_{As}$ is a constant that varies across asset classes $A$ and mandate $s$, $\lambda_{Ats}$ is a time fixed effect for asset class $A$ and mandate $s$, $\text{Private}_i$ is a dummy equal to one if plan $i$ is private and $\text{nonUS}_i$ is a dummy equal to one if plan $i$ is domiciled outside the U.S. Taking logs in (2.5.1), we obtain the following panel model which allows us to estimate the power law coefficient, $\beta_{As}$.[33]

$$\log(\text{Cost}^{\$}_{iats}) = c_{As} + \lambda_{Ats} + \beta_{As}\log(\text{AUM}_{iats}) + \gamma_{1,As}\text{Private}_i + \gamma_{2,As}\text{nonUS}_i + \varepsilon_{iats}. \tag{2.5.2}$$

We estimate this model at the sub-asset class level to leverage the granularity of the data provided by CEM, significantly increasing the sample size compared to simply using less-granular asset class level data. Notice, also, that we impose homogeneity in the power-law coefficient within each asset

[32] Bikker (2017) uses different cost functions to show that average costs are decreasing in size and that investment costs are $U$-shaped. Related to this, Alserda, Bikker, and Lecq (2018) find large economies of scale for administrative costs, and diseconomies of scale for investment costs.

[33] We include time fixed effects but not plan fixed effects in (2.5.2). Because AUM varies a lot across plans and is highly persistent, including plan fixed effects would make it difficult to estimate the size-cost relationship. For example, a high-profile pension plan with hundreds of billions of dollars in AUM is likely to face very different investment costs compared to a much smaller plan with a few hundred million dollars in AUM and plan fixed effects are likely to capture this.

class (across its sub-asset classes) so that information from all sub-asset classes is used to estimate the economies of scale parameter for the associated asset class.

The top panel in Table 2.6 shows estimates of (2.5.2) obtained for the different management mandates at the asset class level. First, consider the two public asset classes, stocks and fixed income, for which we have sufficient data to consider all four management mandates. Across both asset classes and for all four management mandates, our estimates of $\beta$ are less than unity and, consistent with Hypothesis II((i)), we reject the null hypothesis of no economies of scale, $\beta_{1,As} = 1$.

Turning to the importance of investment mandate for scale economies, our estimates of $\beta_{As}$ are around 0.75 for passively managed stocks and fixed income assets but closer to 0.90 for actively managed accounts in these asset classes. This suggests that economies of scale are much higher for passively managed than for actively managed public assets. This result seems intuitive, since it is much easier to scale-up an index investment than an active strategy (consistent with the conjecture of Berk and Green (2004)). Our finding that passive management lends itself better to scaling than active management is also consistent with Hypothesis II((iii)) and seems highly plausible.[34] Our estimates of the power law coefficients are very similar within active or within passive management, regardless of whether assets are managed internally or externally. The choice of passive versus active management is thus more important to economies of scale than is the decision for whether to manage assets internally or externally.

Turning to the four alternative asset classes, passive management is uncommon, so we only report estimates for internal active and external active mandates.[35] Table 2.6 shows that the estimates of $\beta$ are generally higher than those obtained for stocks and fixed income, averaging 0.95 and ranging from 0.91 to 1.01. This finding is consistent with Hypothesis II((ii)), suggesting somewhat *lower* scale advantages in unit investment costs for alternative asset classes, compared with publicly traded assets.[36]

[34]Passive investment management relies heavily on computer algorithms that are easy to scale up. Passive portfolios may venture into more sub-asset classes as they grow in size in order to limit any adverse market impact, but this is unlikely to raise costs by much. Conversely, active investment management is more labor intensive, and more adversely affected by the market impact of trading and, thus, more difficult to scale up.

[35]For hedge funds and multi assets, there are only 140 observations of internal active management, so we do not report IA estimates for this case.

[36]This finding is consistent with the far more labor-intensive process of managing specialized asset classes such as private equity. For these asset classes, there is generally no reliable public price that aggregates market information in the same way as for stocks and fixed income, making scaling more difficult and passive management infeasible. The main exception is REITS within the real asset class, but again we do not have a sufficient number of data points on this sub-asset class to conduct a meaningful analysis.

We next consider, in columns two and three of Table 2.6, investment management cost differences between private vs. public and U.S. vs. non-U.S. plans, respectively. There is evidence that private plans incur higher costs than public plans in the internal and external management of stocks and fixed income assets. We find very little evidence of notable differences in private and public plans' costs of passively managing stocks or fixed-income, as well as managing alternative assets, either internally or externally. Non-U.S. plans pay significantly higher costs, on average, than U.S. plans for both internal and external passive management of stocks and fixed income assets, but pay lower fees for management of these asset classes in external active accounts. Among the alternative asset classes, non-U.S. plans pay significantly higher fees for internal active management of private debt and real assets but they incur significantly lower costs for external active management of private debt as compared to their U.S. peers.

To formally test Hypothesis II that there are statistically significant differences in scale economies between internal/external and passive/active management, respectively, we estimate a model that pools observations across the four management mandates $s$:

$$\begin{aligned}\log(\text{Cost}^{\$}_{iats}) = c_{As} + \lambda_{Ats} + \beta_{1,As}\text{Dummy}_s + \beta_{2,As}\log(\text{AUM}_{iats}) \\ + \beta_{3,As}\text{Dummy}_s \times \log(\text{AUM}_{iats}) + \beta_{4,As}\text{Private}_i \\ + \beta_{5,As}\text{nonUS}_i + \varepsilon_{iats}, \qquad (2.5.3)\end{aligned}$$

where each of the dummy variables $\text{Dummy}_s$ equals one if $s \in \{\text{IA, EA, EP}\}$. The fourth investment management mandate (IP) is treated as the benchmark so all effects are measured relative to this case. For example, for internally managed assets $\text{Dummy}_s = 1$ if $s = \text{IA}$ and zero otherwise so this dummy allows us to estimate the differential impact of internal active management on cost relative to the benchmark of internal passive management. We can test the null hypothesis of no scale differences between internal passive and internal active management by examining the significance of $\beta_{3,As}$.

We present the results of these tests in the bottom three rows of Table 2.6. For stocks and fixed income, we cannot reject the null hypothesis of equal returns to scale for internal and external passive management, in line with Hypothesis II((iv)). Moreover, we cannot reject the

null hypothesis that cost economies of scale are identical across internal and and external active mandates for three of five asset classes, the two exceptions being fixed income and real assets. For fixed income assets, internal active management is associated with significantly higher scale economies than external active management ($\beta^{\text{IA}} = 0.84$ versus $\beta^{\text{EA}} = 0.94$), while for real assets internal active management has weaker scale economies than external asset management ($\beta^{\text{IA}} = 1.01$ versus $\beta^{\text{EA}} = 0.92$). Hence the empirical evidence is mixed in relation to Hypothesis II((iv)).

Finally, in the bottom row we report p-values for a one-sided test of equal economies of scale in passive and active management for stock and fixed income portfolios against the alternative that cost economies are bigger for passively managed than for actively managed accounts. Consistent with Hypothesis II((iii)) we reject the null hypothesis for both stocks and fixed income, which indicates that larger plans, in particular, can achieve significant cost economies by switching from active to passive management.

In summary, our results demonstrate that scale economies in asset management costs vary along two important margins: (i) management mandate (IP, EP, IA, EA); and (ii) asset class. To help quantify the economic importance of variation in costs along these margins, the right panel of Table 2.6 reports management costs for small, medium, and large plans, represented by the $10^{\text{th}}$, $50^{\text{th}}$, and $90^{\text{th}}$ percentiles of the (2019) AUM distribution for a given mandate and asset class combination. These columns summarize the economic effect on costs of the full set of coefficient estimates from our analysis.

Several important points emerge. First, internal passive management leads to substantial cost savings for both stocks and fixed income investments, with external passive management being roughly twice as costly as internal passive management. Second, internal active management costs are lower than external active management costs by an order of magnitude both for publicly traded assets (stocks and fixed income) and also for private asset classes, especially private equity.

Third, there are particularly strong economies of scale across stocks and fixed income accounts, as demonstrated by the significantly lower per-dollar unit cost of plans in the 90th percentile, compared with plans in the 10th percentile of the size distribution. Economies of scale are generally far smaller for actively managed private asset classes, regardless of whether these are managed internally or externally.

We also estimate (2.5.2) separately for each *sub-asset class*, using only those sub-asset classes

that contain a sufficiently large number of observations to allow us to obtain accurate estimates. In Appendix Table B.9, we find that the cost economy of scale estimates are in line with those obtained for the broader asset classes. Economies of scale are notably larger (i.e., $\beta$ estimates are lower) for passive management of EAFE and U.S. broad stock mandates, as well as for inflation-indexed bonds. In turn, scale economies are much lower for diversified private equity, real estate, and REIT accounts.

## 2.6 Investment Performance and Plan Size

Next, we examine how plan characteristics, such as plan size, affects investment performance. As we have seen, plan size is a key determinant of costs. In this section, we explore whether plan size also influences the ability of plans to identify the best-performing asset managers and their bargaining power for net return performance after costs – a crucial question for plan beneficiaries.

A unique feature of our data is that it contains "policy returns" for each plan/sub-asset class/mandate (e.g., an internal active mandate) combination. Policy returns are negotiated targets between fund managers and plan sponsors, and can be used as a simple form of risk-adjustment.[37] Specifically, let $r_{iat}$ be the return of plan $i$ in sub-asset class $a$ during year $t$, while $r^P_{iat}$ is the associated policy return for the same plan, sub-asset class, and time period. The policy-adjusted return, $\widetilde{r}_{iat}$, is then[38]

$$\widetilde{r}_{iat} = r_{iat} - r^P_{iat}. \tag{2.6.1}$$

### 2.6.1 Return Regressions

We examine the relation between plan characteristics and investment performance through a set of panel regressions that use policy-adjusted returns as the dependent variable. These regressions are estimated separately for each of our asset classes using plan-year sub-asset class returns as the unit of observation, and thus take the form:

[37] This simple, but powerful method for risk-adjusting is especially important for our sample, where many plans are represented for only one or a few years. In subsection 2.6.2, we explore robustness of our results to using a more conventional risk-adjustment approach based on plans' exposure to a set of common risk factors.

[38] Appendix B.4 reports summary statistics for raw returns and policy-adjusted returns.

$$\widetilde{r}_{iat} = c_a + \lambda_{At} + \beta_{1,A} \log(\text{AUM}_{iat-1}) + \beta_{2,A}\text{Private}_i + \beta_{3,A}\text{nonUS}_i + \beta'_{4,A} x_{iat} + \epsilon_{iat}, \tag{2.6.2}$$

where $\widetilde{r}_{iat}$ denotes the policy-adjusted gross or net return on plan $i$'s holdings in sub-asset class $a \in A$ in year $t$, $c_a$ denotes a sub-asset class fixed effect, and $\lambda_{At}$ is an asset-class time fixed effect.[39] Although we use returns at the sub-asset class level, we impose that the coefficient estimates, $\beta_A$, are the same within a particular asset class, $A$. $x_{iat}$ contains a set of control variables that include $\omega_{iat}^{\text{External}}$ (the share of external management), $\omega_{iat}^{\text{Active}}$ (the share of active management), and a dummy equal to one if plan $i$ pays a performance fee at time $t$ in sub-asset class $a$ ($\text{Perform}_{iat}$). To not confound the impact of plan size with the choice of external and active management on returns, we include the latter as controls. Then, in Section 2.7, we use a matching approach to rigorously estimate the effect of external and active management on returns.

Panel A in Table 2.7 presents our estimates from regression (2.6.2), applied separately to gross returns (top) and net returns (bottom). This allows us to examine whether differences in investment performance are explained by differences across plans in costs and fees, as well as differences in pre-fee alphas.

First, consider the relation between plan size and return performance for the two public asset classes. Large plans generate significantly higher policy-adjusted gross returns for stocks than small plans, and this effect is more pronounced for net returns, consistent with Hypothesis III. Specifically, going from a plan ranked in the 10th percentile to a plan ranked in the 90th percentile of the AUM of 2019 plan stock holdings increases the expected policy-adjusted gross and net returns by 26 and 41 bps/year, respectively (see top of Panel B). This is consistent with larger plans exploiting their ability to identify skilled managers and their bargaining power to retain some of the alpha, as implied by Hypothesis III((i)).

The size-return relation is stronger among alternative asset classes, particularly for private equity investments. Moreover, the coefficients on size are bigger for net returns than for gross returns, consistent with large plans not only earning higher gross returns in these asset classes than

[39] $\lambda_{At}$ can help absorb omitted risk factors provided that plan exposures to such factors are relatively homogeneous.

smaller plans, but also paying lower management costs. Differences in the net return performance of large and small plans are economically large. Specifically, going from a plan in the 10th percentile of the 2019 AUM asset class distribution to a plan in the 90th percentile is associated with increases in mean net policy-adjusted returns of 109 bps/year (hedge funds and multi asset), 419 bps (private equity), 79 bps (private debt), and 122 bps (real assets).[40] As a robustness check, we also compute the increase in policy-adjusted gross and net return based on portfolio sorts. In particular, we form equally-weighted portfolio returns based on the bottom 30th and upper 70th size percentile within each year.[41] We then calculate the time series average return for portfolios sorted on size. The bottom rows of panel B in Table 2.7 show that the size effect is broadly similar to the estimates based on the panel regression, particularly after accounting for the fact that the portfolios do not go as far out in the tail of the plan size distribution as the panel estimates.

Because we have fewer data points on the alternative asset classes, we also explore a specification that pools return data across all plans, alternative asset classes and years and imposes homogeneous slope coefficients:

$$\widetilde{r}_{iat} = c_a + \lambda_t + \beta_1 \log(\text{AUM}_{iat-1}) + \beta_2 \text{Private}_i + \beta_3 \text{nonUS}_i + \beta_4' x_{iat} + \epsilon_{iat}. \tag{2.6.3}$$

By assuming that the coefficients are the same across alternative asset classes, this specification uses far more data points which can increase the precision of our estimates. Results are shown in the "Alt" column of Table 2.7. We find a significantly positive relation between plans' log-AUM and policy-adjusted gross and net returns. Again, the coefficient on size is larger for net returns than for gross returns, consistent with some of the higher net returns earned by the largest plans stemming from their ability to better exploit economies of scale and reduce costs consistent with Hypotheses III((i))-((ii)).

Given the significantly positive association between policy-adjusted net returns and log-size observed for five out of six asset classes, we would also expect to find a positive and significant

[40] For gross returns the magnitude is somewhat smaller, as we find increases of 26 bps (stocks), 2 bps (fixed income), 71 bps (hedge funds and multi assets), 303 bps (private equity), 75 bps (private debt), and 76 bps (real assets) per year.

[41] We use the 30th and 70th percentile instead of the 10th and 90th percentile to increase the number of observations within a year.

association between log-AUM and plans' total portfolio performance (i.e., the overall performance of a pension plan). We explore if this relation holds by estimating the following panel model for plan-level total portfolio returns:

$$\widetilde{r}_{it} = \lambda_t + \beta_1 \log(\text{AUM}_{it-1}) + \beta_2 \text{Private}_i + \beta_3 \text{nonUS}_i + \beta_4' x_{it} + \epsilon_{it}, \tag{2.6.4}$$

where $\widetilde{r}_{it}$ is the policy-adjusted return on plan $i$'s total assets in year $t$, gross or net of costs. The "Total portfolio" column in Table 2.7 shows that larger plans obtain modestly higher policy-adjusted gross and net returns. For example, moving from the 10th to the 90th percentile plan as ranked by total AUM is associated with an increase in policy-adjusted net total-portfolio returns of 23 bps per annum.

### 2.6.2 Risk-adjusted Return Performance

Policy returns constitute a natural benchmark against which to measure individual plans' return performance. However, it is more common to measure investment performance by adjusting for plans' exposure to a small set of risk factors. Such an approach is not feasible here, however, because most plans have short return records in the CEM database.

As an alternative to conducting plan-level risk adjustments, for each asset class, we form equal-weighted portfolios that comprise up to 29 years of annual plan-level returns. We refer to this equal-weighted aggregate return for asset class $A$ in year $t$ as $\bar{r}_{At}$ and use it to estimate the following risk-adjustment regression:

$$\bar{r}_{At} - r_{ft} = \alpha_A + \beta_A' \text{F}_{At} + \epsilon_{At}, \tag{2.6.5}$$

where $r_{ft}$ is the risk-free rate and $\text{F}_{At}$ refers to the risk factors used for asset class $A$. We consider both the Fama and French (1993) three-factor model and the seven-factor model of Fung and Hsieh (2001) which includes the market excess return, a bond trend, currency trend, commodity trend, size spread, bond market and credit spread. The risk factor regressions provide a very good fit for stocks, fixed income, and hedge funds and a somewhat poorer fit for plans' returns in the three remaining alternative asset classes. Appendix B.4.4 provides further details.

Repeating the earlier analysis from Table 2.7 on the plan-year risk-adjusted returns, we find results that are broadly in line with those obtained for the policy-adjusted returns. The last four columns in Table 2.7 shows results for stocks, fixed income, alternative assets and the total portfolio. We find that the largest plans continue to produce significantly higher risk-adjusted returns on the alternative asset classes both on a gross and net of cost basis. Using risk-adjusted returns also leads to higher coefficient estimates on the size variable for fixed income and total portfolio returns.

## 2.7 Investment Management Style, Costs, and Return Performance

As we have described throughout this paper, pension plans must decide whether to manage their investments within a given sub-asset class internally or externally, as well as actively or passively. This decision reflects a variety of plan characteristics, particularly plan size (AUM) and sub-asset class, with some sub-asset classes lending themselves more easily to passive and internal management than others. In this section, we analyze how such decisions affect plan performance in the form of management costs and benchmark-adjusted returns. In doing so, we recognize that the estimation of the effect of plans' decisions on management styles and the resulting expected performance is endogenous. For example, a large plan might use its resources to identify a genuinely skilled external active manager, and bargain for lower fees rather than switching to lower-cost internal active management or even passive management. Plans may also switch management style because of external shocks affecting all plans within a given sub-asset class.

### 2.7.1 A Matching Estimator

We believe that a "gold standard" for estimation of the effect of plans' management style decisions on plan performance comes from the literature on treatment effects and matching estimators. The idea is to compare the performance in a given asset class of two otherwise similar plans where one plan switches from, say, external to internal management, while the other plan continues to manage its assets externally. Key to this type of matching estimator is, first, to obtain an accurate match, and, second, that there are a sufficient number of cases (switches) to allow us to accurately estimate any performance differences between the two types of plans. Provided

that these conditions hold, the advantage of the resulting estimator is that, under a set of well-understood conditions, it controls for differences in any confounding factors that can vary across plans and asset classes.

Specifically, to account for the endogeneity of plans' asset management decisions and the presence of confounding factors, we adopt the difference-in-differences estimator recently proposed by Imai, Kim, and Wang (2021). The chief advantage of this estimator is that it can handle unbalanced panels such as ours and datasets with a small time-series dimension. It also allows units to switch treatment status over time. All of these are features we observe in the CEM data.

We use the effect of management style on plan cost as our lead example, but our analysis is parallel for policy-adjusted returns. First consider the effect of internal management on the cost (in bps) of plan $i$ in sub-asset class $a$ at time $t$, $\text{Cost}_{iat}$. Using the potential outcomes framework of Imbens and Rubin (2015), define the average effect of switching from external to internal management on management costs

$$\begin{aligned}\Delta C^{ex\to in} := \mathbb{E}\Big(&\text{Cost}_{iat}(\text{Internal}_{iat}=1,\text{Internal}_{iat-1}=0)\\&-\text{Cost}_{iat}(\text{Internal}_{iat}=0,\text{Internal}_{iat-1}=0)|\text{Internal}_{iat}=1,\text{Internal}_{iat-1}=0\Big),\end{aligned}\tag{2.7.1}$$

where $\text{Cost}_{iat}(\text{Internal}_{iat}=1,\text{Internal}_{iat-1}=0)$ is the potential cost outcome of a plan switching from external management at time $t-1$ to internal management at time $t$, whereas $\text{Cost}_{iat}(\text{Internal}_{iat}=0,\text{Internal}_{iat-1}=0)$ denotes the potential cost for the same plan not switching management style.[42]

To account for unobserved confounding variables such as bargaining power, we rely on the following parallel trend assumption

$$\begin{aligned}&\mathbb{E}\left(\text{Cost}_{iat}(\text{Internal}_{iat}=0,\text{Internal}_{iat-1}=0)-\text{Cost}_{iat-1}|\text{Internal}_{iat}=1,\text{Internal}_{iat-1}=0,x_{iat}\right)\\=&\mathbb{E}\left(\text{Cost}_{iat}(\text{Internal}_{iat}=0,\text{Internal}_{iat-1}=0)-\text{Cost}_{iat-1}|\text{Internal}_{iat}=0,\text{Internal}_{iat-1}=0,x_{iat}\right).\end{aligned}\tag{2.7.2}$$

In our analysis, $x_{iat}$ contains the following time-varying (potentially confounding) controls:

[42] In the language of the treatment effect literature, Equation (2.7.1) represents the average treatment effect on the treated.

- $\text{AUM}_{iat}$: total AUM allocated by plan $i$ to sub-asset class $a$ at time $t$
- $\text{Active}_{iat}$: an indicator for whether plan $i$ manages sub-asset class $a$ actively at time $t$
- $\text{Private}_i$: an indicator for whether plan $i$ is private
- $\text{nonUS}_i$: an indicator for whether plan $i$ is domiciled in the U.S.
- sub-asset class $a$ at time $t$.

The parallel trend assumption (2.7.2) stipulates that the change in management cost is equal between the treatment group (plans switching from external to internal management) and the control group (plans continuing with external management) in the absence of treatment ($\text{Internal}_{iat} = 0$), once we condition on $x_{iat}$. These observed variables allow us to control for differences in costs induced by plan size (captured by $\text{AUM}_{iat}$), active vs. passive management (captured by $\text{Active}_{iat}$), public vs. private plans (captured by $\text{Private}_i$), U.S. vs. non-U.S. plans (captured by $\text{nonUS}_i$), and sub-asset class heterogeneity. We include the latter as a control to impose that plans can only be matched within the same sub-asset class because of the large heterogeneity in costs (and potential confounding variables) across different sub-asset classes.[43] Intuitively, our matching approach can therefore be thought of as providing an estimate of the "pure" effect on cost of choosing internal management as opposed to external management after controlling for plan size, differences across sub-asset class, and other plan characteristics.

We also consider estimating the reverse effect of a switch from internal to external management:

$$\begin{aligned}\Delta C^{in\to ex} := \mathbb{E}\Big(&\text{Cost}_{iat}(\text{External}_{iat} = 1, \text{External}_{iat-1} = 0) \\ &- \text{Cost}_{iat}(\text{External}_{iat} = 0, \text{External}_{iat-1} = 0) | \text{External}_{iat} = 1, \text{External}_{iat-1} = 0\Big),\end{aligned}$$

[43] Note that the parallel trend assumption rules out potentially time-varying omitted confounding variables that affect both management costs and the choice of external management. Once the set of potentially confounding variables is specified, treatment and control units are matched based on their propensity score, which is a measure of how similar the plans are along the variables contained in $x_{iat}$. Finally, an estimate of the effect of internal management on cost is obtained by forming an average between treatment and control units in the matched set. See Imai, Kim, and Wang (2021, Equation 18). We use the `PanelMatch` package in `R` developed by these authors to carry out the estimation.

where $\text{Cost}_{iat}(\text{External}_{iat} = 1, \text{External}_{iat-1} = 0)$ is the potential cost outcome of a plan switching from internal management at time $t-1$ to external management at time $t$, whereas $\text{Cost}_{iat}(\text{External}_{iat} = 0, \text{External}_{iat-1} = 0)$ denotes the cost for the same plan that sticks with internal management.

### 2.7.2 Cost Estimates

We begin our analysis by applying the matching estimator to the cost data. Results from the matching estimator are shown in Panel A of Table 2.8. We find that switching from external to internal management (top row) is associated with substantial cost savings, especially in private asset classes. For stocks and fixed income, a change from external to internal management leads to a decrease in cost of 3 bps/year and 5 bps/year, respectively, while, in private markets, cost savings are 320 bps/year (private equity), 26 bps/year (private debt), and 47 bps/year (real assets).[44] Aggregating across all alternative asset classes, we obtain an estimate of 56 bps/year in cost savings. When pooling all asset classes ("All") we obtain a cost savings estimate a little over 4 bps/year. This estimate is closer to the savings for the public asset classes, reflecting their importance in plans' portfolios.

In the data, there are also a number of plans that switch from internal to external management. We analyze the effect on cost of this reverse switch using the same methodology. Panel A of Table 2.8 shows that costs significantly increase when plans switch from internal to external management. Management costs increase by 7 bps/year for stocks, 5 bps/year for fixed income, and by 24–96 bps/year in the alternative asset classes. Across the alternative asset classes, a switch from internal to external asset management is associated with a 54 bps/year increase in costs – essentially mirroring the cost savings estimate (55 bps/year) for the reverse external-to-internal switch. Interestingly, the total portfolio-level increase in costs associated with a switch from internal to external management is somewhat higher than the cost savings associated with the reverse switch (18 bps versus 4 bps).

In summary, our matching estimates indicate that switching from external to internal asset management leads to modest cost savings for public asset classes, but very large cost savings for private asset classes, while the reverse shift from internal to external asset management leads to modestly higher costs for stocks and bonds and significantly higher management costs for alternative

[44] We omit hedge funds and and multi-assets since our sample contains too few plans in these asset classes that switch between external and internal management.

assets. We note that the smaller cost savings for public asset classes are multiplied by the large allocations of DB plans to them.

We next consider the effect on costs of switching between active and passive management. We limit our analysis to the three asset classes (stocks, fixed income, and real assets) for which we have a sufficiently large number of transitions to facilitate accurate estimation. Our estimates are shown in the three rightmost columns of Table 2.8.[45] We find that switching from active to passive management reduces costs by around 9 bps/year for stocks and real assets and by 2 bps/year for fixed income, consistent with the low overall level of fees for this asset class. All estimates are statistically significant.

Conversely, a switch from passive to active management leads to a significant increase in costs. As before, the estimated effect is most pronounced for stocks (15 bps/year) and real assets (16 bps/year), and smaller for fixed income (5 bps/year).

### 2.7.3 Return Performance

To analyze the effect of internal versus external and active versus passive management on return performance, we adopt our matching methodology to policy-adjusted returns. In this case, we use as conditioning variables the previous year's AUM in sub-asset class $a$ ($\text{AUM}_{iat-1}$), as well as the current sub-asset class for matching to ensure that treated and control observations have similar covariate values.[46]

Panel B of Table 2.8 shows that a switch from external to internal management leads to a significant increase in gross and net return performance for all asset classes, except private debt, where the estimate is insignificant due to a very small number of transitions. In all cases, the effect is larger for net returns due to the associated decrease in cost. For stocks and fixed income, a switch from external to internal management is accompanied by an increase in net policy-adjusted returns of 108 bps/year and 47 bps/year, respectively. For two of the three private asset classes, the effect is more pronounced, with increases in policy-adjusted net returns of 255 bps/year (private equity) and 194 bps/year (real assets). For alternative asset classes as a whole, switching from external to internal asset management is associated with an increase in net returns of 198 bps/year. Similarly,

[45] We use the same set of potentially confounding variables $x_{iat}$ as for the internal/external estimates, except we replace $\text{Active}_{iat}$ with $\text{External}_{iat}$ to control for heterogeneity in costs associated with external management.

[46] I.e., we impose that treatment and control units are only matched within the same sub-asset class, as in Section 2.7.1.

when pooling all asset classes, the estimated effect on net portfolio-level returns from switching from external to internal management is 93 bps/year. We note that these yearly increases in return performance, for plans that move from external to internal management, are consistent with the need for such plans to justify the fixed costs of setting up internal management.

The bottom rows of Panel B in Table 2.8 also show the effect of changing from internal to external management. For the alternative asset classes, we find a *decrease* in return performance, both gross and net of fees ranging from -578 bps/year to -60 bps/year for net return performance in private equity and real assets, respectively. Overall, across all alternative asset classes, we find a decline in net return performance of 202 bps/year which, again, mirrors the estimate (198 bps) from the reverse switch.

The deterioration in return performance from a switch toward external management does not carry over to public asset classes. Instead, we find that a switch from internal to external management raises average net policy-adjusted returns by 139 bps/year for stocks, and by 26 bps/year for fixed income, although the latter estimate is insignificant.

For the total portfolio, our estimate suggests a small and statistically insignificant improvement of 22 bps/year in net return performance following a switch from internal to external asset management. This is substantially lower than the improvement of 93 bps/year observed for the reverse switch (external to internal management).

These results suggest the following. First, for private asset classes, we observe significant improvements in return performance (both net and gross of costs) associated with a switch from external to internal management while, conversely, return performance is significantly decreased following the reverse switch from internal to external management. Hence, return performance in alternative asset classes has benefited significantly from plans switching from external to internal asset management. A significant driver of this is cost savings, but we also observe large improvements in gross returns following the adoption of internal management for private equity and real assets.

Next, for public asset classes, and stocks in particular, a switch in management style (from external to internal or internal to external) is likely driven by poor prior-year performance. If such switches are driven by poor return performance in year $t-1$, mean-reversion in performance would dictate that we should expect to see improved performance in year $t$. Consistent with this

explanation, we find underperformance of -140 bps/year in the year prior to the switch for plans that move from internal to external management. For comparison, plans in the control group have a slight outperformance of 50 bps during this period. This suggests that the parallel trend assumption in (2.7.2) is violated for plans that switch from internal to external management.[47] Moreover, if we extend the set of conditioning variables to include net returns in year $t-1$, the estimate of the effect on net stock returns declines from 139 bps/year to 12 bps/year.

The fact that we do not observe a similar effect for the alternative asset classes may reflect that the improvement in return performance associated with internal management is much larger and, also, that asset values are not marked-to-market in the same way as for the public asset classes. Cost savings from internal asset management are also much larger for the alternative asset classes, as we have seen.

Last, we apply our matching methodology to estimate the effect of passive and active management on policy-adjusted returns.[48] The three rightmost columns in the top rows of Table 2.8 (Panel B) show that a switch from active to passive management does not significantly increase gross or net return performance with signs being both negative (stocks) and positive (fixed income, real assets). A switch from passive to active management (bottom rows) is, however, associated with a significantly negative policy-adjusted net return of -36 bps/year for stock investments. Policy-adjusted return estimates for a switch from passive to active management are also negative both gross and net of costs for fixed income and real assets, but fail to be statistically significant.

Our findings suggest that changes from active to passive or from passive to active management are, on the whole, not associated with a significant effect on plans' gross or net policy-adjusted return performance. This is consistent with an equilibrium in which competition is so strong among active managers (in public asset classes in particular) that fees are set so that, for the marginal plan that decides to switch, a very small change in performance can be expected.[49]

[47] We see no similar violation of this assumption for fixed income.

[48] The set of controls used is identical to those listed in Section 2.7.2.

[49] Notice that this finding does not contradict our earlier result that large plans tend to get better return performance, particular net of costs. The matching estimates in Table 2.8 control for plan size and are based on the population of plans that switch between active and passive management. Hence, they characterize the return effect for the marginal plan that switches management style.

## 2.8 Conclusion

This paper explores the relation between pension plan size and allocations to active vs. passive management, internal vs. external management, and public vs. private market investments. Consistent with fixed costs being important in setting up internal investment management capabilities, large plans *internally* manage a significantly greater proportion of assets than their smaller peers. Similarly, taking advantage of their greater ability to identify internal and external investment opportunities in the less transparent markets for private assets, large plans also allocate more of their holdings to asset classes such as private equity and real assets and less to (public) stocks and fixed income.

Our results indicate a strong role for economic scale in pension plan fees and investment performance: investment management costs follow a power law with cost economies being particularly strong for passively managed accounts and public asset classes. Hence, large plans pay significantly lower fees per dollar invested than their smaller peers. While large plans' better ability to identify skilled external managers and negotiate lower fees has only translated into modestly higher net-of-cost return performance in the highly competitive public asset markets (stocks and fixed income), we find strong evidence that larger plans earn economically large and significant abnormal returns in the markets for private assets (again, compared to their smaller peers). Private markets are less transparent and so allow the largest plans to benefit from their comparative advantage in searching for skilled managers.

The scale disadvantages in investment management costs that we identify for smaller plans indicate that these plans may perform best when they embrace passive management which is widely available in public asset markets. For private asset classes, passive management is generally not an option (other than for special cases such as REITS) and fixed costs are too high to be covered by small plans which, consequently, rely almost entirely on external active management and have to accept the higher management fees typically charged for this service. Conversely, large plans have the ability to manage private assets internally and negotiate lower external investment management fees. This helps explain why plan size (scale) is particularly important in determining investment performance in private asset markets and why private asset classes have become particularly important for large plans in recent years.

## 2.9 Acknowledgements

Chapter 2, in full, is currently being prepared for submission for publication of the material. De Vries, Tjeerd; Kalfa, Yanki; Timmermann, Allan; Wermers, Russ. "Scale Economies, Bargaining Power, and Investment Performance: Evidence from Pension Plans". The dissertation author is a primary author of this material.

**Table 2.1. Small and large plans' investment allocation by sub-asset class and management structure in 2019**. This table shows the share (in %) of AUM allocated to the four management mandates: Internal Passive (IP), External Passive (EP), Internal Active (IA), and External Active (EA) for the given sub-asset classes. The share is calculated as follows: $\omega_{ats} = \frac{\text{AUM}_{ats}}{\text{AUM}_{at}}$, where $\text{AUM}_{ats} = \sum_i \text{AUM}_{iats}$, and $\text{AUM}_{at} = \sum_s \sum_i \text{AUM}_{iats}$, where $i$ denotes plan $i$, $a$ indicates the sub-asset class, $t$ denotes the year 2019, and $s$ denotes one of the four mandates. The shares are calculated separately for small and large plans, defined by the bottom and top 30th percentile of AUM in 2019 respectively. For small and large plans, rows sum up to 100%.

| | Small Plans (in %) | | | | Large Plans (in %) | | | |
|---|---|---|---|---|---|---|---|---|
| **Stocks** | IP | EP | IA | EA | IP | EP | IA | EA |
| ACWI x. US | 1.66 | 27.51 | | 70.82 | 2.77 | 31.00 | 2.77 | 63.45 |
| EAFE | | 19.95 | 1.44 | 78.62 | 18.53 | 14.94 | 11.99 | 54.55 |
| Emerging | | 13.50 | | 86.50 | 15.02 | 8.65 | 13.54 | 62.79 |
| Global | | 24.74 | 1.14 | 74.12 | 15.19 | 6.00 | 58.51 | 20.29 |
| Other | | | | | 25.74 | 0.62 | 26.75 | 46.90 |
| U.S. Broad | 14.17 | 57.90 | 2.06 | 25.87 | 34.06 | 32.08 | 8.55 | 25.31 |
| U.S. Large Cap | | 51.88 | 11.45 | 36.67 | 32.05 | 31.53 | 20.02 | 16.40 |
| U.S. Mid Cap | | 34.60 | | 65.40 | 27.54 | 6.15 | 25.06 | 41.25 |
| U.S. Small Cap | | 18.50 | | 81.50 | 19.66 | 4.87 | 13.25 | 62.22 |
| **Fixed Income** | | | | | | | | |
| Bundled LDI | | 1.61 | 37.56 | 60.83 | 28.22 | 45.20 | 2.66 | 23.92 |
| Cash | | | 54.70 | 45.30 | | | | 100.00 |
| Convertibles | | | | 100.00 | | | | 100.00 |
| EAFE | | | | | 86.88 | | | 13.12 |
| Emerging | | | | 100.00 | 7.51 | 6.21 | 23.91 | 62.37 |
| Global | | 0.60 | | 99.40 | 8.84 | 0.63 | 82.76 | 7.77 |
| High Yield | | | 5.87 | 94.13 | | 3.59 | 23.03 | 73.37 |
| Inflation Indexed | 24.63 | 48.74 | 9.87 | 16.77 | 40.47 | 11.61 | 41.33 | 6.60 |
| Long Bonds | 0.32 | 21.33 | 5.46 | 72.88 | 18.54 | 0.58 | 14.46 | 66.43 |
| Other | | 14.44 | 14.54 | 71.02 | 72.48 | 0.88 | 7.01 | 19.63 |
| U.S. | | 13.18 | 2.75 | 84.07 | 6.27 | 10.37 | 46.22 | 37.14 |
| **Hedge & multi ass.** | | | | | | | | |
| Funded TAA | | | 6.05 | 93.95 | | | 58.27 | 41.73 |
| Hedge Fund | | | | 100.00 | | | | 100.00 |
| Risk Parity | | | | 100.00 | | | 28.19 | 71.81 |
| **Private Equity** | | | | | | | | |
| Div. Private Eq. | | | 0.08 | 99.92 | | | 18.86 | 81.14 |
| LBO | | | | 100.00 | | | 0.27 | 99.73 |
| Other | | | | 100.00 | | | 26.81 | 73.19 |
| Venture Capital | | | | 100.00 | | | 0.70 | 99.30 |
| **Private Credit** | | | | | | | | |
| Mortgages | | | 1.98 | 98.02 | | | 67.24 | 32.76 |
| Credit | | | 10.49 | 89.51 | | | 31.45 | 68.55 |
| **Real Assets** | | | | | | | | |
| Commodities | | 18.43 | | 81.57 | 19.70 | 1.82 | 58.20 | 20.28 |
| Infrastructure | | | | 100.00 | | | 61.39 | 38.61 |
| Nat. Resource | | | | 100.00 | | | 46.70 | 53.30 |
| Other | | | | 100.00 | | | 28.42 | 71.58 |
| Real Estate | | | 2.62 | 97.38 | | | 39.67 | 60.33 |
| REIT | | 6.60 | | 93.40 | 2.53 | 3.59 | 77.54 | 16.33 |

**Table 2.2. Asset allocation regression for internal vs. external management.** **Panel A** of this table presents estimates of the regression (2.4.6): $\omega_{iAt}^{internal} = c_A + \lambda_{At} + \beta_{1,A}\log(\text{AUM})_{it-1} + \beta_{2,A}\text{CostSpread}_{iAt-1}^{E-I} + \beta_{3,A}\text{Private}_i + \beta_{4,A}\text{nonUS}_i + \epsilon_{iAt}$, where $\omega_{iAt}^{internal}$ is the share of AUM that is internally managed by plan $i$ in asset class $A$ at time $t$, $\lambda_{At}$ is a time fixed effect, $\text{AUM}_{it-1}$ denotes the lagged AUM of sponsor $i$, $\text{CostSpread}_{iAt-1}^{E-I}$ is plan $i$'s cost spread between external and internal management in asset class $A$ at time $t-1$, $\text{Private}_i$ is a dummy equal to one if plan $i$ is private, and $\text{nonUS}_i$ is a dummy equal to one if the plan is domiciled outside the U.S. For each asset class we also use the Cragg estimator with a point mass at 0. We report the average partial effects of Cragg estimates in columns named "Cragg APE". In case a plan is fully internal (external) we impute the external (internal) cost as the median cost from plans that are similar in size, where size in a given year is either small (bottom 30th percentile), medium (between 30th and 70th percentile), or large (top 70th percentile) of total AUM. Robust standard errors are in parenthesis and clustered by plan. Boldface coefficients are statistically significant at the 5% level. The asset class "Private Debt" does not include time fixed effects due to the small sample size, and estimation for "Hedge Funds" start in 2000 due to lack observation prior to 2000. **Panel B** presents the results of the Cragg selection equation (2.4.2a). The bottom part of panel B shows the probability of allocating at least some portion of investments internally. We fix the the cost spread at the mean cost spread across time and sponsors, and show the different probabilities based on size, using the 10th, 50th, and 90th percentile of AUM in 2019 in a given asset class. All coefficients and standard errors are multiplied by 100.

**Panel A: Panel and Cragg Estimation**

| | Stocks | | Fixed Income | | Hedge & Multi ass. | | Private Equity | | Private Debt | | Real Assets | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Panel | Cragg APE | Panel | Cragg APE | Panel | Cragg APE | Panel | Cragg APE | Panel | Cragg APE | Panel | Cragg APE |
| $\log(\text{AUM}_{it-1})$ | **8.29** (0.824) | **11.24** (2.461) | **10.96** (0.997) | **17.66** (2.647) | **0.82** (0.397) | 0.68 (0.454) | **2.34** (0.787) | **2.32** (0.825) | **11.25** (2.360) | **14.31** (5.034) | **6.80** (0.990) | **6.40** (1.540) |
| $\text{CostSpread}_{iAt-1}^{E-I}$ | **15.25** (5.245) | 12.23 (6.946) | **23.81** (4.784) | **30.29** (8.660) | -0.00 (0.189) | 0.17 (0.420) | -0.09 (0.061) | -0.23 (0.240) | -0.01 (0.087) | -0.95 (1.729) | 1.03 (0.623) | 2.66 (1.576) |
| $\text{Private}_i$ | -0.29 (2.067) | 1.65 (4.260) | -3.36 (2.845) | 2.36 (6.810) | 0.43 (0.851) | -0.89 (1.377) | 1.81 (1.535) | 3.50 (2.906) | 6.54 (7.181) | 15.27 (12.758) | 0.29 (2.469) | 2.21 (3.204) |
| $\text{NonUS}_i$ | **13.99** (2.097) | **13.06** (4.596) | **23.60** (2.759) | **21.91** (5.387) | 0.74 (1.232) | -0.83 (1.196) | **14.45** (2.512) | **11.91** (4.490) | **22.23** (7.275) | 1.88 (9.156) | **27.79** (2.964) | **17.44** (3.814) |
| Obs | 7205 | 7205 | 7222 | 7222 | 1944 | 1944 | 4322 | 4322 | 1055 | 1055 | 5676 | 5676 |
| $R^2$ | 0.26 | | 0.29 | | 0.06 | | 0.18 | | 0.30 | | 0.24 | |

**Panel B: Cragg Selection Estimates**

| | Stocks | Fixed Income | Hedge & Multi ass. | Private Equity | Private Debt | Real Assets |
|---|---|---|---|---|---|---|
| $\log(\text{AUM}_{it-1})$ | **38.19** (4.018) | **28.49** (3.309) | **33.07** (8.783) | **23.59** (5.709) | **34.59** (8.935) | **28.08** (4.227) |
| $\text{CostSpread}_{iAt-1}^{E-I}$ | 36.77 (23.892) | 21.52 (18.451) | **-13.33** (5.741) | -2.30 (1.712) | -2.32 (4.065) | 0.20 (2.617) |
| | $\Pr(\omega_{iAt}^{internal} > 0 \mid X = x)$ | | | | | |
| Plan size | | | | | | |
| $10^{\text{th}}$ percentile | 13.22 | 28.85 | 0.50 | 5.60 | 9.26 | 12.82 |
| $50^{\text{th}}$ percentile | 34.35 | 48.97 | 2.51 | 12.54 | 24.86 | 27.08 |
| $90^{\text{th}}$ percentile | 66.18 | 72.12 | 10.59 | 26.06 | 52.56 | 49.71 |

**Table 2.3. Significance test for the difference in APE for size and cost spread.** This table shows the APE (2.4.7) estimated for size ($\log \text{AUM}_{it-1}$) and external minus internal cost spread (Panel A). In Panel B we calculate two different cost spreads. For "Active Overall", we define the cost spread to be active minus passive for the given asset class, and for " Active Int. vs. Ext", we define the cost spread as Active external minus active internal. We set $\log(\text{AUM})_{it-1}$ and $\text{CostSpread}_{iAt-1}$ to their $10^{\text{th}}$ and $90^{\text{th}}$ percentile values in 2019 respectively. Then we test if the estimated APEs are equal using a $\chi^2$-test. We report the difference of the calculated APEs. Panel A uses the proportion of plans' holdings within each asset class that is internally managed as the dependent variable. Panel B uses either the proportion of plans' holdings within each asset class that is actively managed (columns 1-3) or the proportion of internally managed assets that is managed actively as the dependent variable (columns 4-5). The columns show the calculation for each asset class. All coefficients and standard errors are multiplied by 100. We compute the standard errors of the computed APEs using the Delta method and standard errors are reported in parenthesis. Boldface coefficients are significant at the 5% level.

| | Panel A: Internal vs. External Management | | | | | |
|---|---|---|---|---|---|---|
| | Stocks | Fixed Income | Hedge multi ass. | Private Equity | Private Debt | Real Asset |
| log(AUM) | | | | | | |
| Percentile: 10 | **3.75** | **6.47** | 0.44 | **0.41** | **6.57** | **2.55** |
| | (0.875) | (1.060) | (0.241) | (0.208) | (1.910) | (0.499) |
| Percentile: 90 | **31.90** | **43.80** | 0.69 | 1.70 | 20.10 | **6.22** |
| | (9.860) | (10.500) | (0.796) | (1.090) | (10.800) | (2.190) |
| Difference | **28.15** | **37.33** | 0.25 | 1.29 | 13.53 | 3.67 |
| Cost Spread | | | | | | |
| Percentile: 10 | **9.40** | **21.50** | 0.19 | -0.09 | -0.96 | **1.35** |
| | (4.640) | (5.230) | (0.281) | (0.104) | (1.770) | (0.554) |
| Percentile: 90 | 12.20 | **28.60** | 0.16 | -0.08 | -0.92 | 1.82 |
| | (7.610) | (8.740) | (0.474) | (0.081) | (1.630) | (0.986) |
| Difference | 2.8 | **7.10** | -0.04 | 0.01 | 0.03 | 0.47 |
| Obs | 7205 | 7222 | 1944 | 4322 | 1055 | 5676 |

| | Panel B: Active vs. Passive Management | | | | |
|---|---|---|---|---|---|
| | Active Overall | | | Active Int. vs. Ext. | |
| | Stocks | Fixed Income | Real Asset | Stocks | Fixed Income |
| log(AUM) | | | | | |
| Percentile: 10 | **-4.60** | 2.76 | -0.07 | **3.80** | **5.87** |
| | (0.720) | (1.630) | (0.353) | (0.114) | (0.105) |
| Percentile: 90 | **-4.16** | **2.08** | 0.18 | **17.30** | **38.30** |
| | (1.340) | (0.624) | (0.231) | (0.789) | (0.993) |
| Difference | 0.44 | -0.68 | 0.25 | 13.5 | **32.43** |
| Cost Spread | | | | | |
| Percentile: 10 | **-30.30** | **-19.40** | -0.11 | 1.26 | **18.50** |
| | (4.470) | (5.570) | (0.326) | (3.980) | (4.380) |
| Percentile: 90 | **-39.70** | **-25.80** | -0.11 | 1.15 | **24.60** |
| | (8.310) | (9.740) | (0.376) | (4.510) | (7.20) |
| Difference | **-9.40** | -6.4 | -0.01 | -0.11 | **6.10** |
| Obs | 7206 | 7210 | 4395 | 7012 | 7090 |

**Table 2.4. Asset allocation regression for passive vs. active management**. The left panel shows the regression of the fraction of actively managed assets over total assets for stocks, fixed income and real assets as shown in equation (2.4.8). In the four rightmost columns the dependent variable is defined as the proportion of actively managed assets that is internally managed: $\text{AUM}_{iAt}^{Active,Internal}/\text{AUM}_{iAt}^{Active}$. For each asset class the regression specification is as follows. We include the log of year t - 1 AUM per sponsor, and the cost spread between active and passive at year $t-1$. For the rightmost columns the cost spread is defined as the difference between active external minus active internal cost. For cases where a plan is fully internal (external) we impute the external (internal) cost as the median cost for plan that is similar size (small ($30^{\text{th}}$ percentile in total AUM), medium (between $30^{\text{th}}$ and $70^{\text{th}}$ percentile in total AUM), or large ($70^{\text{th}}$ percentile in total AUM)) in that given year. We include a dummy that equals 1 if the plan is private ($\text{Private}_i$), and a dummy that equals 1 if the plan is located outside of the U.S. ($\text{nonUS}_i$). Lastly, we control for time fixed effects. For each asset class we run two regressions, namely a time fixed effects panel regression and a hurdle regression with a point a mass at 1. The column heading "Panel" denotes the fixed effects regression estimates. We report the average partial effects (APE) of Cragg estimates in columns starting with "Cragg". Robust standard errors are in parenthesis and clustered by plan. Boldface coefficients are statistically significant at the 5% level. All coefficients and standard errors are multiplied by 100.

| | Active Allocation | | | | | | Active Internal Allocation | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Stocks | | Fixed Income | | Real Assets | | Stocks | | Fixed Income | |
| | Panel | Cragg APE | Panel | Cragg APE | Panel | Cragg APE | Panel | Cragg APE | Panel | Cragg APE |
| $\log(\text{AUM}_{it-1})$ | **-3.20** | **-4.27** | 0.08 | **2.45** | -0.11 | 0.04 | **6.43** | **8.25** | **10.22** | **16.81** |
| | (0.675) | (0.827) | (0.658) | (1.231) | (0.160) | (0.318) | (0.833) | (2.333) | (1.036) | (2.675) |
| $\text{CostSpread}_{iAt-1}$ | **-20.47** | **-32.87** | **-7.36** | **-22.23** | -0.01 | -0.11 | **8.10** | 1.50 | **17.52** | **27.02** |
| | (4.019) | (5.663) | (3.586) | (9.363) | (0.193) | (0.349) | (3.865) | (4.988) | (4.260) | (7.701) |
| $\text{Private}_i$ | 3.13 | **4.79** | -1.94 | 2.01 | -0.15 | 0.18 | 1.03 | 3.52 | -4.27 | -0.55 |
| | (1.997) | (2.263) | (2.041) | (2.561) | (0.845) | (0.729) | (2.072) | (4.718) | (2.824) | (6.661) |
| $\text{NonUS}_i$ | **4.64** | 1.56 | **-8.93** | **-9.75** | 0.40 | **1.47** | **14.94** | **16.10** | **24.28** | **24.54** |
| | (2.098) | (2.227) | (2.139) | (2.678) | (0.739) | (0.664) | (2.230) | (6.019) | (2.828) | (5.217) |
| Constant | 0.57 | | -7.73 | | -0.42 | | **-42.72** | | **-59.85** | |
| | (5.874) | | (6.038) | | (2.089) | | (5.948) | | (7.976) | |
| Obs | 7206 | 7206 | 7210 | 7210 | 4395 | 4395 | 7012 | 7012 | 7090 | 7090 |
| $R^2$ | 0.10 | | 0.04 | | 0.01 | | 0.18 | | 0.27 | |

**Table 2.5. Asset allocation regression**. This table reports estimates of the regression (2.4.9): $\omega_{iAt} = c_A + \lambda_{At} + \beta_{1,A}\log(\text{AUM}_{it-1}) + \beta_{2,A}\text{Cost}_{iAt-1} + \beta_{3,A}\text{Private}_i + \beta_{4,A}\text{nonUS}_i + \beta_{5,A}\text{LiabilityRetiree}_{it} + \epsilon_{iAt}$, where $\omega_{iAt}$ denotes plan $i$'s proportion of assets allocated to asset class $A$ at time $t$, $c_A$ is the asset class fixed effect, $\lambda_{At}$ denotes the time fixed effect, $\log(\text{AUM}_{it-1})$ denotes the log of plan $i$'s total AUM at time $t-1$, $\text{Cost}_{iAt-1}$ denotes the cost (in bps) of plan $i$ in asset class $A$ at time $t-1$, $\text{Private}_i$ denotes a dummy variable equal to one if plan $i$ is not a public plan, $\text{nonUS}_i$ is a dummy variable equal to one if the plan is domiciled outside the U.S., and $\text{LiabilityRetiree}_{it}$ denotes the fraction of plan $i$'s total liabilities owed to retirees in year $t$. **Panel A** excludes $\text{LiabilityRetiree}_{it}$ as a regressor. All coefficient estimates and standard errors are multiplied by 100. The robust standard errors are in parenthesis and clustered by plan. Boldface coefficients are statistically significant at the 5 percent level.

| **Panel A** | Stocks | Fixed Income | Hedge & multi ass. | Private Equity | Private Debt | Real Assets |
|---|---|---|---|---|---|---|
| $\log(\text{AUM}_{it-1})$ | **-2.08** | **-0.87** | **-1.85** | **0.44** | **-0.40** | **0.65** |
| | (0.339) | (0.324) | (0.271) | (0.157) | (0.168) | (0.130) |
| $\text{Cost}_{iAt-1}$ | **-7.11** | **-20.10** | **-1.65** | **-0.11** | 0.01 | **-0.44** |
| | (3.109) | (4.341) | (0.430) | (0.029) | (0.012) | (0.097) |
| $\text{Private}_i$ | **-2.90** | **5.94** | -1.73 | -0.33 | -0.99 | **-2.03** |
| | (0.920) | (0.858) | (0.940) | (0.434) | (0.589) | (0.294) |
| $\text{nonUS}_i$ | **-3.38** | **2.66** | **-3.75** | **-2.09** | 0.12 | **1.66** |
| | (0.930) | (0.934) | (0.939) | (0.413) | (0.656) | (0.379) |
| Obs | 7206 | 7219 | 2611 | 4413 | 1073 | 5677 |
| $R^2$ | 0.24 | 0.12 | 0.15 | 0.20 | 0.10 | 0.31 |
| **Panel B** | | | | | | |
| $\log(\text{AUM}_{it-1})$ | **-2.50** | -0.65 | **-1.84** | **0.60** | **-0.43** | **0.72** |
| | (0.423) | (0.388) | (0.336) | (0.157) | (0.176) | (0.182) |
| $\text{Cost}_{iAt-1}$ | **-8.92** | **-15.49** | **-1.37** | **-0.09** | 0.01 | **-0.58** |
| | (3.558) | (4.274) | (0.467) | (0.027) | (0.010) | (0.123) |
| $\text{Private}_i$ | **-4.46** | **8.78** | -1.74 | -0.42 | -0.96 | **-2.51** |
| | (1.088) | (0.977) | (1.061) | (0.440) | (0.629) | (0.370) |
| $\text{nonUS}_i$ | **-2.48** | **2.57** | **-4.17** | **-2.19** | -0.03 | **1.65** |
| | (1.210) | (1.169) | (1.126) | (0.408) | (0.681) | (0.493) |
| $\text{LiabilityRetiree}_{it}$ | **-6.00** | 2.48 | 1.64 | 1.11 | 1.90 | -0.56 |
| | (3.026) | (3.354) | (2.434) | (1.184) | (1.680) | (1.028) |
| Obs | 4435 | 4440 | 1757 | 2774 | 782 | 3499 |
| $R^2$ | 0.27 | 0.16 | 0.14 | 0.22 | 0.11 | 0.30 |

**Table 2.6. Economies of scale for cost among different investment mandates**. The *regression* panel of this table shows estimates of the model (2.5.2): $\log(\text{Cost}^{\$}_{iats}) = c_{As} + \lambda_{Ats} + \beta_{As}\log(\text{AUM}_{iats}) + \gamma_{1,As}\text{Private}_i + \gamma_{2,As}\text{nonUS}_i + \varepsilon_{iats}$, where $\text{Cost}^{\$}_{iats}$ is the cost (in dollars) of plan $i$ in sub-asset class $a$ at time $t$ for mandate $s$, $c_{As}$ is a constant that varies with asset class $A$ and mandate $s$, $\lambda_{Ats}$ is the time fixed effect for asset class $A$ and mandate $s$, $\log(\text{AUM}_{iats})$ is the log of total AUM of plan $i$ in sub-asset class $a$ at time $t$ for mandate $s$, $\text{Private}_i$ is a dummy equal to one if plan $i$ is private and $\text{nonUS}_i$ is a dummy equal to one if plan $i$ is located outside the U.S. For stocks and fixed income, we estimate the panel separately for the following mandates $s$: Internal Passive (IP), Internal Active (IA), External Passive (EP) and External Active (EA). The boldface coefficients on log(AUM) are significantly different from one at the 5% level and boldface coefficients on the other coefficients are significantly different from zero. Robust standard errors are reported in parenthesis and are clustered by plan. The *size percentile* columns show $\widehat{\text{Cost}}^{\$}_{iats}/\text{AUM}_{iats}$ in bps, where $\widehat{\text{Cost}}^{\$}_{iats}$ is predicted based on the *regression* panel. We set $\text{Private}_i$ and $\text{nonUS}_i$ equal to zero and use the $10^{\text{th}}$, $50^{\text{th}}$ and $90^{\text{th}}$ percentile of $\text{AUM}_{iats}$ in 2019 to obtain the fraction of cost relative to AUM. The bottom panel shows $p$-values of the null hypotheses that returns to scale are the same for different mandates, where a boldface $p$-value indicates a rejection of the null hypothesis.

| | Regression | | | | | Size percentile | | |
|---|---|---|---|---|---|---|---|---|
| | $\log(\text{AUM}_{iats})$ | $\text{Private}_i$ | $\text{nonUS}_i$ | Obs | $R^2$ | 10% | 50% | 90% |
| Stocks | | | | | | | | |
| IP | **0.76** (0.037) | 0.25 (0.157) | **0.93** (0.120) | 2294 | 0.70 | 2.67 | 1.48 | 0.85 |
| EP | **0.75** (0.015) | -0.01 (0.051) | **0.24** (0.055) | 11239 | 0.62 | 5.39 | 2.94 | 1.65 |
| IA | **0.89** (0.027) | **0.46** (0.167) | 0.22 (0.148) | 3552 | 0.70 | 9.36 | 7.25 | 5.62 |
| EA | **0.88** (0.007) | **0.04** (0.021) | **-0.28** (0.023) | 25799 | 0.86 | 62.66 | 49.98 | 39.11 |
| Fixed Income | | | | | | | | |
| IP | **0.80** (0.047) | -0.09 (0.210) | **0.39** (0.175) | 1269 | 0.69 | 2.94 | 1.51 | 1.00 |
| EP | **0.79** (0.024) | 0.11 (0.071) | **0.26** (0.074) | 4125 | 0.63 | 4.57 | 2.84 | 1.92 |
| IA | **0.84** (0.021) | **0.51** (0.124) | **0.25** (0.103) | 5293 | 0.72 | 4.09 | 2.77 | 2.03 |
| EA | **0.94** (0.010) | 0.00 (0.036) | **-0.18** (0.040) | 17544 | 0.76 | 27.75 | 23.98 | 20.92 |
| Hedge & Multi ass. | | | | | | | | |
| EA | **0.95** (0.018) | 0.09 (0.062) | -0.03 (0.064) | 4801 | 0.78 | 146.87 | 133.21 | 120.66 |
| Private Equity | | | | | | | | |
| IA | 1.01 (0.035) | 0.19 (0.215) | 0.37 (0.241) | 768 | 0.78 | 18.00 | 18.49 | 19.02 |
| EA | **0.93** (0.015) | **-0.08** (0.039) | 0.02 (0.050) | 8480 | 0.86 | 382.93 | 312.52 | 268.04 |
| Private Debt | | | | | | | | |
| IA | 0.95 (0.064) | -0.39 (0.274) | **0.76** (0.286) | 411 | 0.79 | 12.25 | 10.13 | 8.64 |
| EA | 0.94 (0.036) | -0.18 (0.147) | **-0.62** (0.139) | 1377 | 0.75 | 188.03 | 165.91 | 146.75 |
| Real Assets | | | | | | | | |
| IA | 1.01 (0.032) | 0.00 (0.138) | **0.49** (0.135) | 2211 | 0.74 | 11.58 | 11.79 | 11.98 |
| EA | **0.92** (0.011) | -0.06 (0.036) | -0.07 (0.037) | 12117 | 0.79 | 161.87 | 136.15 | 115.65 |

Hypothesis Testing ($p$-value)

| Null hypothesis | Stocks | Fixed Income | Private Equity | Private Debt | Real Assets |
|---|---|---|---|---|---|
| $\beta^{\text{IP}} = \beta^{\text{EP}}$ | 0.90 | 0.46 | | | |
| $\beta^{\text{IA}} = \beta^{\text{EA}}$ | 0.19 | **0.00** | 0.33 | 0.79 | **0.01** |
| $\beta^{\text{P}} = \beta^{\text{A}}$ | **0.00** | **0.00** | | | |

**Table 2.7. Regression of policy- and risk-adjusted returns on plan characteristics**. This table shows estimates of model (2.6.2): $\widetilde{r}_{iat} = c_a + \lambda_{At} + \beta_{1,A}\log(\text{AUM}_{iat-1}) + \beta_{2,A}\text{Private}_i + \beta_{3,A}\text{nonUS}_i + \beta'_{4,A}x_{iat} + \epsilon_{iat}$, where $\widetilde{r}_{iat}$ denotes the policy-adjusted **gross** (top) and **net** (bottom) return; $\log(\text{AUM}_{iat-1})$ denotes plan $i$'s total AUM in sub-asset class $a$ at time $t-1$; $\text{Private}_i$ is a dummy for whether a plan is private; $\text{nonUS}_i$ is a dummy for whether plan $i$ is domiciled in the U.S.; $x_{iat}$, a set of controls including the fraction of external and active management, as well as a performance fee dummy. The column "Alt." estimates (2.6.3) and pools the alternative asset classes: Hedge & multi assets, Private equity, Private debt and Real assets. The column *Total portfolio* uses plan-level aggregate returns $r_{it}$ from portfolios and estimates (2.6.4). Portfolios are constructed as weighted averages (by AUM) of asset class investments per sponsor in a given year. The risk-adjusted return estimates only include U.S. plans and are only shown for stocks, fixed income, alternative, and total portfolio. Risk-adjusted returns are estimated at the asset class level instead of sub-asset class level. Robust standard errors are reported in parentheses and clustered by sponsor. Boldface coefficients are statistically significant at the 5% level. The "Panel reg." rows show the effect on policy-adjusted gross and net returns of moving from the bottom $10^{\text{th}}$ percentile to the upper $90^{\text{th}}$ percentile in plan size in 2019 based on the panel estimates in the upper panel. The "Portfolio sort" rows show the effect on policy-adjusted gross and net returns of moving from the bottom $30^{\text{th}}$ percentile to the upper $70^{\text{th}}$ percentile in plan size within a year. Portfolios are constructed as equal-weighted average returns of sub-asset classes within a year and for a given plan size.

| | Policy-Adjusted Returns | | | | | | | | Risk-Adjusted Returns | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Panel A** | Stocks | Fixed income | Hedge & multi ass. | Private equity | Private debt | Real assets | Alt. | Total portfolio | Stocks | Fixed income | Alt. | Total portfolio |
| **Gross** | | | | | | | | | | | | |
| $\log(\text{AUM}_{iat-1})$ | **0.06** | 0.00 | 0.18 | **0.63** | 0.12 | **0.16** | **0.29** | **0.04** | -0.03 | **0.17** | **0.36** | **0.14** |
| | (0.026) | (0.031) | (0.093) | (0.133) | (0.123) | (0.080) | (0.059) | (0.020) | (0.048) | (0.054) | (0.127) | (0.039) |
| $\text{Private}_i$ | **0.17** | 0.06 | 0.52 | -0.21 | -0.41 | 0.36 | 0.17 | **0.11** | 0.16 | **1.22** | -0.63 | **0.40** |
| | (0.075) | (0.064) | (0.326) | (0.495) | (0.377) | (0.240) | (0.207) | (0.052) | (0.129) | (0.141) | (0.400) | (0.103) |
| $\text{nonUS}_i$ | -0.14 | **-0.23** | **-0.68** | **2.19** | 0.01 | 0.02 | **0.45** | -0.02 | | | | |
| | (0.096) | (0.086) | (0.323) | (0.563) | (0.418) | (0.253) | (0.218) | (0.063) | | | | |
| Obs | 22879 | 18042 | 2762 | 4984 | 1015 | 8819 | 17580 | 7181 | 3897 | 3721 | 6298 | 4907 |
| $R^2$ | 0.07 | 0.05 | 0.20 | 0.21 | 0.18 | 0.08 | 0.09 | 0.19 | 0.00 | 0.02 | 0.01 | 0.01 |
| **Net** | | | | | | | | | | | | |
| $\log(\text{AUM})_{iat-1}$ | **0.09** | 0.03 | **0.28** | **0.86** | 0.14 | **0.25** | **0.43** | **0.06** | -0.00 | **0.19** | **0.46** | **0.16** |
| | (0.026) | (0.031) | (0.091) | (0.126) | (0.123) | (0.084) | (0.061) | (0.019) | (0.047) | (0.054) | (0.135) | (0.041) |
| $\text{Private}_i$ | 0.14 | 0.06 | 0.46 | 0.23 | -0.28 | 0.46 | 0.34 | 0.10 | 0.13 | **1.20** | -0.57 | **0.37** |
| | (0.075) | (0.063) | (0.325) | (0.481) | (0.374) | (0.244) | (0.208) | (0.051) | (0.129) | (0.139) | (0.412) | (0.109) |
| $\text{nonUS}_i$ | -0.09 | **-0.22** | **-0.64** | **2.15** | 0.20 | 0.24 | **0.56** | 0.09 | | | | |
| | (0.095) | (0.083) | (0.324) | (0.534) | (0.417) | (0.258) | (0.215) | (0.060) | | | | |
| Obs | 22878 | 18042 | 2762 | 4986 | 1015 | 8819 | 17582 | 7181 | 3897 | 3721 | 6300 | 4907 |
| $R^2$ | 0.06 | 0.04 | 0.15 | 0.19 | 0.13 | 0.06 | 0.06 | 0.13 | 0.00 | 0.02 | 0.01 | 0.01 |
| **Panel B** | | | | | | | | | | | | |
| **Panel reg.** | Mean return increase: moving from the $10^{\text{th}}$ to $90^{\text{th}}$ plan size percentile in 2019 | | | | | | | | | | | |
| **Gross** | 0.26 | 0.02 | 0.71 | 3.03 | 0.75 | 0.76 | 1.39 | 0.18 | | | | |
| **Net** | 0.41 | 0.13 | 1.09 | 4.19 | 0.79 | 1.22 | 2.02 | 0.23 | | | | |
| **Portfolio sort** | Mean return increase: moving from the $30^{\text{th}}$ to $70^{\text{th}}$ plan size percentile | | | | | | | | | | | |
| **Gross** | 0.21 | 0.02 | 0.39 | 3.33 | 0.10 | 0.90 | | | | | | |
| | (0.379) | (0.177) | (0.983) | (2.056) | (0.804) | (0.581) | | | | | | |
| **Net** | 0.39 | 0.14 | 0.79 | **4.25** | 0.43 | **1.26** | | | | | | |
| | (0.380) | (0.176) | (0.923) | (2.073) | (0.790) | (0.592) | | | | | | |

**Table 2.8. Effect of asset management style on cost and returns using matching.** This table shows the effect of switching from external to internal management (Internal), internal to external management (External), active to passive management (Passive) and passive to active management (Active) on cost (Panel A) and policy-adjusted returns (Panel B) for different asset classes. The asset classes "Alt" and "All" pool observations across the alternative asset classes and all asset classes respectively. In Panel A, the effect is estimated using the following controls: $\text{AUM}_{iat}$, total AUM allocated by plan $i$ to sub-asset class $a$ at time $t$; $\text{Private}_i$, an indicator denoting whether plan $i$ is private; $\text{nonUS}_i$, an indicator denoting whether plan $i$ is domiciled in the U.S.; and sub-asset class $a$ at time $t$ to ensure that plans in the treated group are matched with plans in the control group that invest in the same sub-asset class. To estimate the effect of internal/external (passive/active) management, we also use the control: $\text{Active}_{iat}$ ($\text{External}_{iat}$), an indicator denoting if plan $i$ manages sub-asset class $a$ actively (externally) at time $t$. In Panel B, the effect of management style is estimated separately for gross and net returns, using the controls: $\text{AUM}_{iat-1}$; and sub-asset class $a$ at time $t$. Robust standard errors are reported in parentheses and boldface coefficients are significant at the 5% level. "Treated units" denotes the number of plans that switch management style by asset class.

| **Panel A: Cost (in bps)** | | Stocks | Fixed income | Private equity | Private debt | Real assets | Alt. | All | | Stocks | Fixed income | Real assets |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Internal | **-2.97** | **-5.38** | **-320.00** | **-26.00** | **-47.06** | **-55.69** | **-4.28** | Passive | **-9.49** | **-1.97** | **-9.43** |
| | | (0.567) | (1.054) | (65.806) | (8.598) | (14.945) | (11.006) | (1.574) | | (0.509) | (0.385) | (0.860) |
| | External | **7.37** | **5.01** | **96.34** | **24.43** | **39.16** | **54.32** | **17.59** | Active | **14.70** | **4.88** | **16.12** |
| | | (0.482) | (0.670) | (5.981) | (4.660) | (3.140) | (2.695) | (0.769) | | (0.625) | (0.342) | (1.366) |
| | Treated units | | | | | | | | | | | |
| | Internal | 212 | 186 | 36 | 11 | 99 | 145 | 545 | Passive | 720 | 298 | 44 |
| | External | 210 | 181 | 36 | 14 | 83 | 133 | 526 | Active | 551 | 260 | 28 |
| | Obs | 25184 | 19101 | 5609 | 1136 | 9808 | 16553 | 60839 | | 25184 | 19101 | 9808 |
| **Panel B: Returns (in bps)** | | | | | | | | | | | | |
| **Gross** | Internal | **105.06** | **39.56** | **230.77** | -5.09 | **197.60** | **192.89** | **89.47** | Passive | -24.19 | 4.54 | 36.67 |
| | | (23.734) | (13.770) | (109.327) | (64.413) | (65.048) | (51.383) | (17.420) | | (19.824) | (13.871) | (41.692) |
| **Net** | Internal | **107.81** | **47.26** | **254.72** | 32.50 | **193.64** | **197.72** | **93.07** | Passive | -13.42 | 7.54 | 47.64 |
| | | (23.866) | (13.999) | (109.474) | (61.681) | (66.086) | (51.949) | (17.591) | | (19.825) | (13.861) | (41.539) |
| | Treated units | 202 | 157 | 30 | 8 | 95 | 133 | 494 | | 687 | 279 | 39 |
| **Gross** | External | **146.02** | **34.79** | **-523.39** | **-107.85** | -28.01 | **-167.79** | **36.70** | Active | -21.53 | -5.79 | -17.42 |
| | | (21.293) | (15.751) | (122.537) | (47.167) | (53.532) | (46.874) | (16.338) | | (16.290) | (14.958) | (49.319) |
| **Net** | External | **139.23** | 25.84 | **-578.24** | **-113.89** | -60.28 | **-202.59** | 22.03 | Active | **-36.17** | -11.37 | -34.82 |
| | | (21.296) | (15.686) | (122.262) | (47.779) | (53.931) | (47.025) | (16.371) | | (16.307) | (14.956) | (49.285) |
| | Treated units | 197 | 152 | 28 | 12 | 86 | 126 | 476 | | 541 | 255 | 26 |
| | Obs | 24814 | 18700 | 5446 | 1078 | 9514 | 16039 | 59564 | | 24814 | 18700 | 9514 |

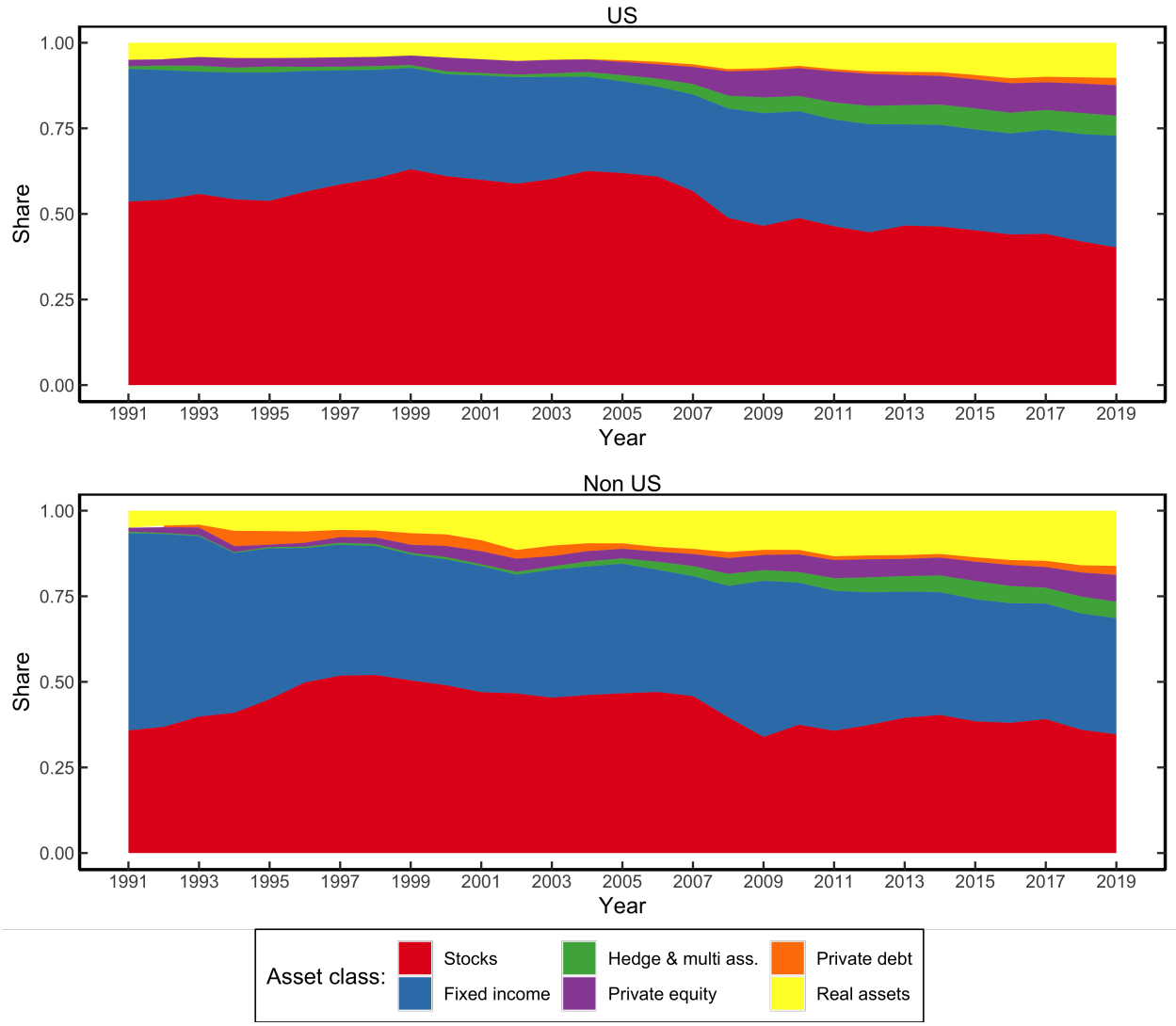



**Figure 2.1. Asset allocation over time.** This figure shows the share of total AUM allocated to each of the six asset classes within a year. The shares are reported separately for U.S. plans (top panel) and non-U.S. plans (bottom panel).

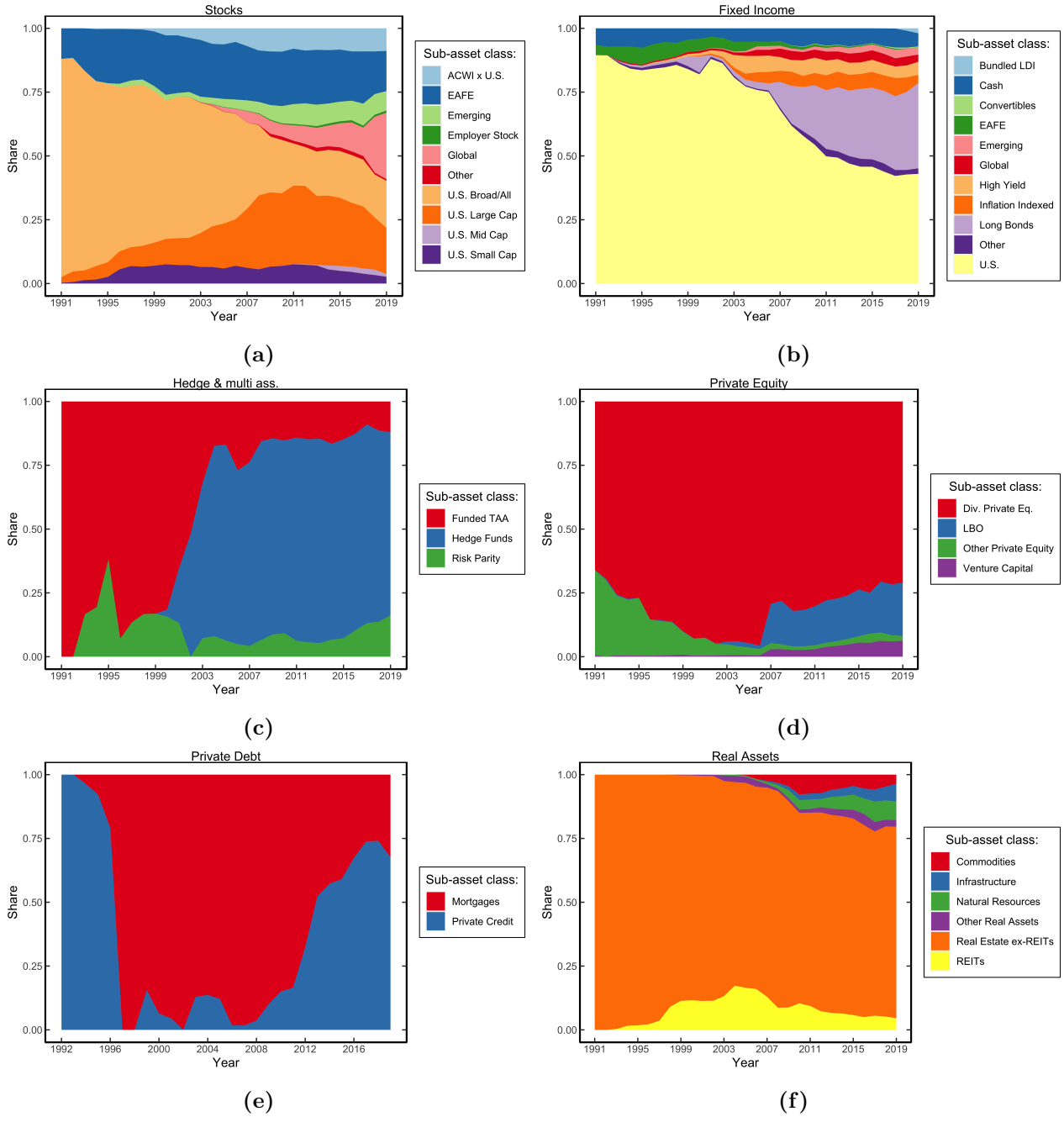
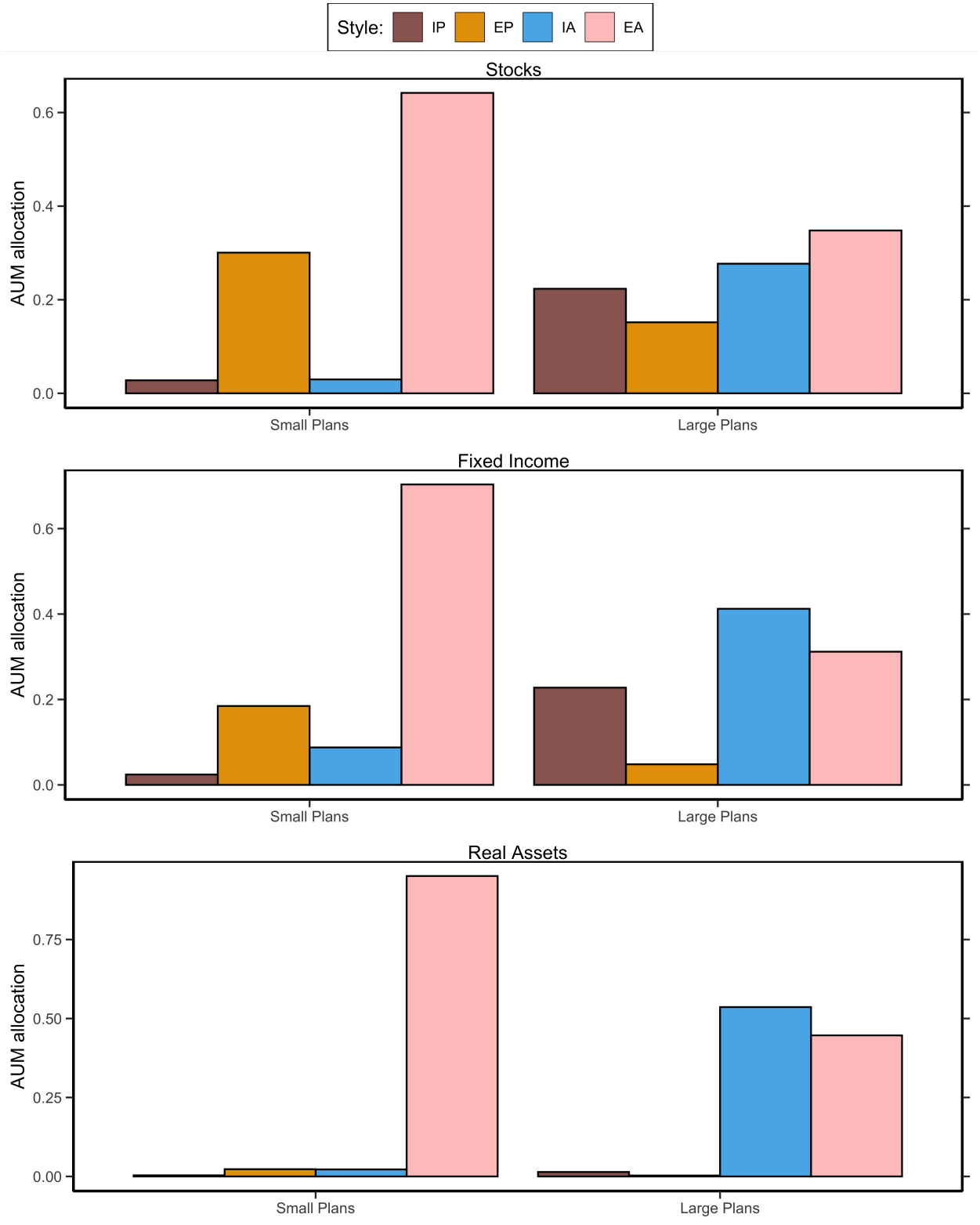



**Figure 2.2. Sub-asset class allocation over time for U.S. plans.** This figure shows the share of total AUM allocated to each sub-asset class for a given year and asset class for U.S. plans only.

**Figure 2.3. Asset allocation by management style and plan size.** This figure shows the share of total AUM allocated to the four management styles: Internal Passive (IP), External Passive (EP), Internal Active (IA) and External Active (EA). The shares are calculated in 2019 for the asset classes: Stocks, Fixed Income and Real Assets. Within each year, we also distinguish by small and large plans, which are defined by the bottom 30 and top 70 percentile relative to the total AUM within an asset class.


Style: IP EP IA EA
Stocks
AUM allocation
0.6
0.4
0.2
0.0
Small Plans
Large Plans
Fixed Income
AUM allocation
0.6
0.4
0.2
0.0
Small Plans
Large Plans
Real Assets
AUM allocation
0.75
0.50
0.25
0.00
Small Plans
Large Plans

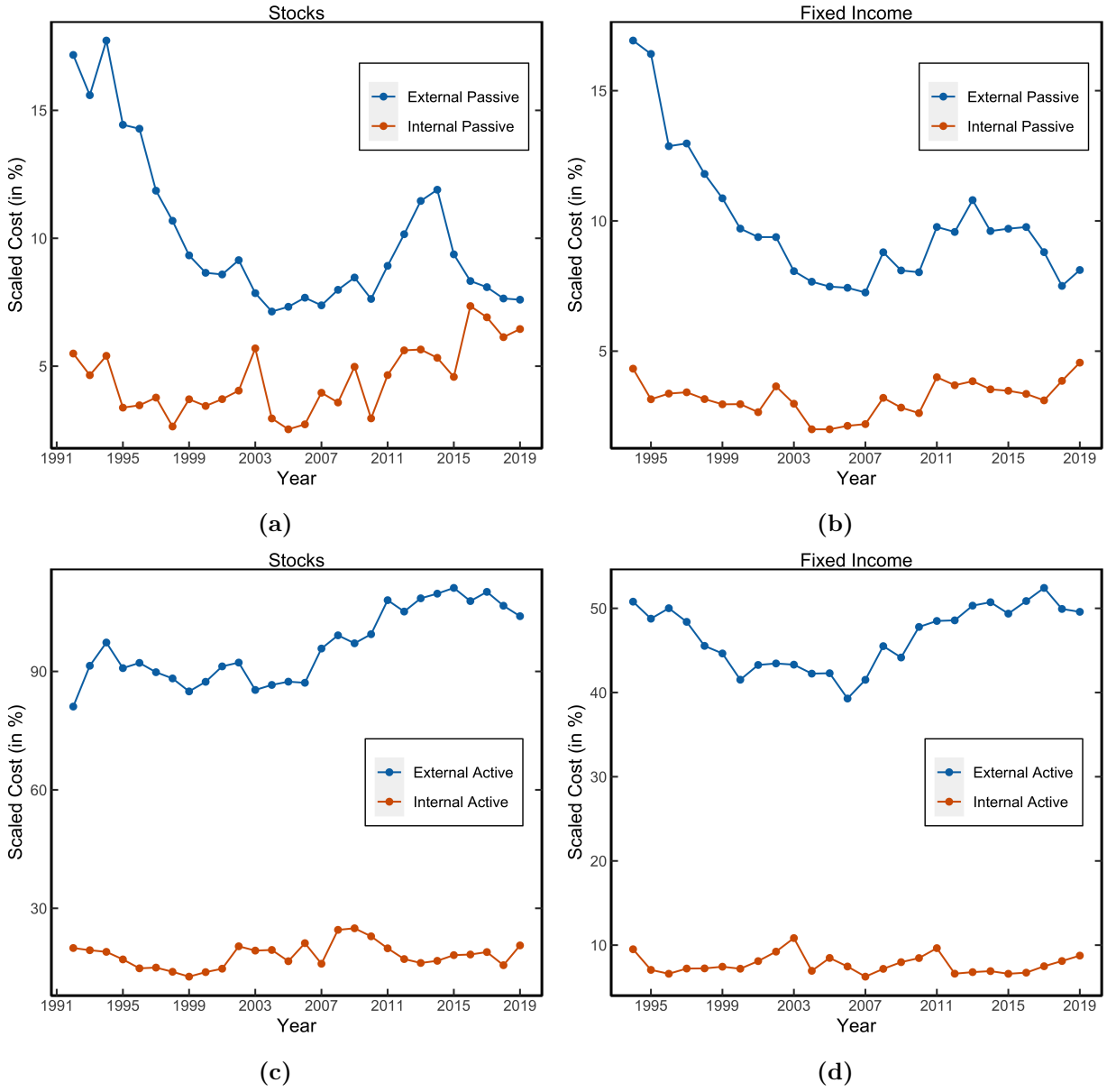
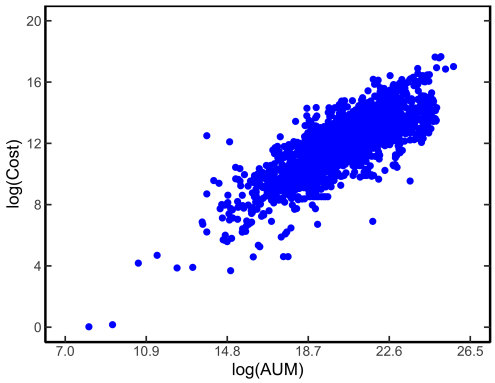
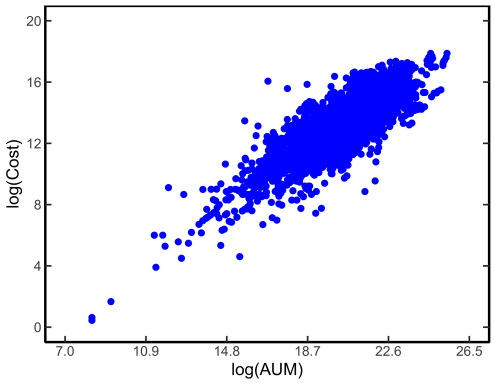



**Figure 2.4. Median cost by asset management style**. The figure shows a time series plot of the (scaled) median cost across plans by asset management style for the public asset classes. The four asset management styles considered are: Internal Passive, External Passive, Internal Active, and External Active management. We only include years that have enough plan observations for each asset class and style. Median cost are scaled by the average cost across years and plans.

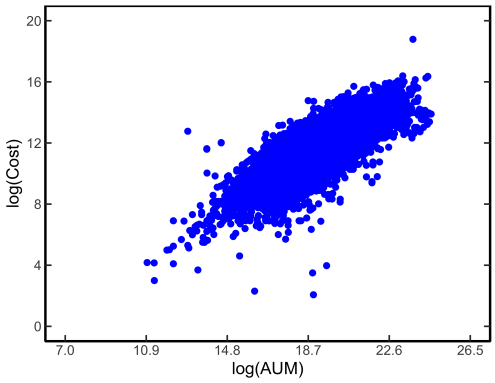
**Figure 2.5. Relation between log Cost and log AUM**. This figure shows a scatter plot of $\log(\text{AUM}_{iats})$ versus $\log(\text{Cost}^{\$}_{iats})$, where $\text{AUM}_{iats}$ (resp. $\text{Cost}^{\$}_{iats}$) denotes the dollar AUM holdings (resp. dollar cost) of plan $i$ in sub-asset class $a$ at time $t$ for asset management style $s$. The asset management styles we consider are: Internal Passive, Internal Active, External Passive, and External Active. In each panel and for a given style, observations are pooled across plans, sub-asset classes, and years over the sample period 1991–2019.



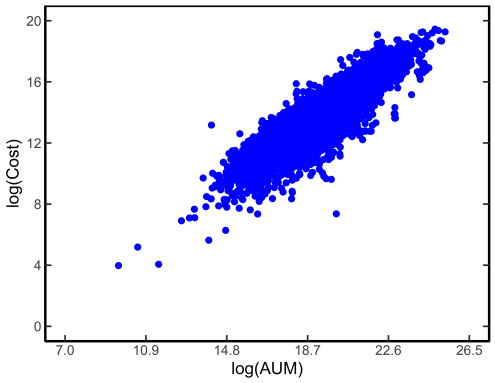

**(a)** Stocks, Internal Passive



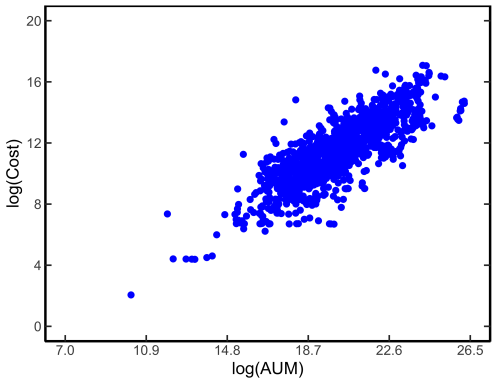

**(b)** Stocks, Internal Active



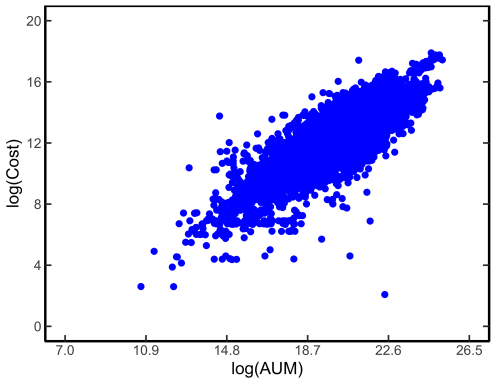

**(c)** Stocks, External Passive



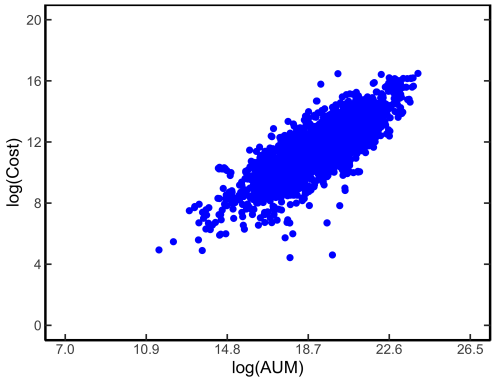

**(d)** Stocks, External Active



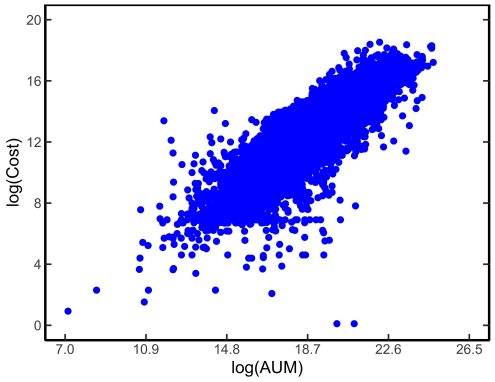

**(e)** Fixed income, Internal Passive




**(f)** Fixed income, Internal Active




**(g)** Fixed income, External Passive




**(h)** Fixed income, External Active

# Chapter 3

# Robust Asset-Liability Management

## 3.1 Introduction

Many financial institutions have long-term commitments. For instance, insurance companies promise annuities or life insurance payments to customers; (defined-benefit) pension plans promise predetermined pension payments to retirees; or commercial banks may make long-term loans at fixed interest rates and thus commit to receiving certain future cash flows in exchange of funding the projects with short-term deposits. In such circumstances, it becomes crucial for financial institutions to effectively manage their assets and liabilities to hedge against interest rate risk. The recent gilt market crisis in the UK showcases the importance of liability-driven investing strategies and the risk associated with interest rate changes, which eventually led to an £65 billion emergency intervention by the Bank of England.[1] Even more recently, Silicon Valley Bank and First Republic Bank collapsed as a result of increased interest rates and the subsequent decline in value of long-term bonds and mortgages.[2,3]

If zero-coupon bonds of all maturities were to exist, any deterministic future cash flow can be replicated by these bonds (which is called a "dedication" strategy), and the problem becomes trivial, at least theoretically. However, in practice dedication is infeasible due to market incompleteness: there are fewer bonds available for trade than the number of payment dates of the liability, or the long-term liability could have a longer maturity than the government bond with longest maturity. Thus, in general, one can only hope to hedge against interest rate risk approximately. The question

[1] https://www.bankofengland.co.uk/speech/2022/november/sarah-breeden-speech-at-isda-aimi-boe-on-nbfi-and-leverage

[2] https://www.ft.com/content/f9a3adce-1559-4f66-b172-cd45a9fa09d6

[3] https://www.economist.com/finance-and-economics/2023/05/03/what-the-first-republic-deal-means-for-americas-banks

of fundamental practical importance is how to achieve this goal given the set of bonds available for trade.

In this article, we propose a new method to construct a hedging portfolio that maximizes equity (asset minus liability) under the most adversarial interest rate shock. This so-called *maxmin* problem originates in the work of Fisher and Weil (1971), who show that a portfolio that matches value and duration (weighted average time to payment) is maxmin against parallel shocks to the forward rate. In that and subsequent works, the liability is assumed to be a zero-coupon bond and no-shortsale constraints are imposed (or implicitly assumed not to bind). These restrictions are undesirable in practice because most liabilities pay out over time and shortsales are essential when liabilities have very long maturities (like pensions). Furthermore, there is no systematic analysis of the existence, uniqueness, and optimality of the solution method as well as explicit or tight error estimates.

Our approach overcomes these shortcomings using techniques from functional and numerical analysis. First we argue that the most general formulation of the maxmin problem is intractable because the objective function is not convex and the space has infinite dimension. To make the problem manageable, we approximate the objective function using the Gateaux differential with respect to basis functions that approximate yield curve shifts. This allows us to recast the maxmin problem as a saddle point (minmax) problem where the inner maximization is a large linear programming problem and the outer minimization is a small convex programming problem, which is computationally tractable. We prove that a robust immunizing portfolio generically exists (Proposition 3.3.1) and its solution achieves the smallest error order and maximizes the worst-case equity (Theorem 3.3.3). This maxmin result is significantly different from the existing literature because both the liability structure and bond portfolio constraint are arbitrary and the guaranteed equity bound is tight. When the majority of forward rate changes are captured by a small number of principal components such as the level of the overall interest rate, we improve this guaranteed equity bound by incorporating moment matching (e.g., duration matching) in the portfolio constraint (Theorem 3.3.5). We also propose particular basis functions (transformation of Chebyshev polynomials) that are motivated by approximation theory.

An alternative approach to asset-liability management, referred to as classical immunization (see, e.g., Redington (1952)), involves matching the interest rate sensitivity of asset and liability.

A common measure of interest rate sensitivity is duration, and matching the duration of asset and liability makes equity insensitive to small interest rate changes. Although classical immunization is intuitive and elegant, by assumption it only allows for small parallel shifts in the yield curve. Furthermore, when there are multiple bonds, it is not obvious how to construct the portfolio because there are infinitely many linear combinations that achieve the same duration. Extensions such as high-order duration matching (which are designed to allow non-infinitesimal or non-parallel shifts in the yield curve) result in unstable portfolio weights and extreme leverage, leading to poor performance (Mantilla-Garcia, Martellini, Milhau, and Ramirez-Garrido, 2022). Our approach contains classical immunization and its extensions as a special case by choosing a monomial basis and imposing only a value matching constraint. In simulation, we show that our preferred robust immunization method that combines moment matching and a Chebyshev polynomial basis does not suffer from extreme leverage and significantly outperforms existing methods.

The simulation exercise uses historical yield curve data to evaluate the change in equity resulting from instantaneous yield curve shocks. A hedging method's success is measured by its ability to minimize these equity changes. Indeed, we find that robust immunization generates approximation errors that are an order of magnitude smaller than the existing approaches and has lower downside risk, in line with our maxmin result. This numerical experiment has a static flavor, since we only consider one-time perturbations. In a separate simulation based on a no-arbitrage term structure model, we consider the dynamic properties of robust immunization, allowing for portfolio rebalancing every three months. Over a 10-year period of rebalancing, robust immunization achieves an approximation error at least 24% lower in the 1% worst-case scenario compared to existing methods. Because our approach is model-free, we expect our proposed method to be useful for practitioners in asset-liability management.[4]

### 3.1.1 Related literature

When inputs to a problem such as beliefs, information, or shocks are complicated, it is common to optimize against the worst case scenario, i.e., solve the *maxmin* problem (Gilboa and Schmeidler, 1989; Bergemann and Morris, 2005; Du, 2018; Brooks and Du, 2021). In the context

[4]This statement is similar to the fact that the Black and Scholes (1973a) option pricing model has been hugely successful precisely because the model requires only a few assumptions, namely the absence of arbitrage and the stock price following a geometric Brownian motion, and no assumptions on investor preferences are required.

of asset-liability management, Redington (1952, p. 290) considers the Taylor expansion of assets minus liabilities in response to a small change in the (constant) interest rate and anticipates the importance of convexity to guarantee the portfolio value. Fisher and Weil (1971) formalize this idea and show that if the liability is a zero-coupon bond and a bond portfolio matches the value and duration, then the portfolio value can never fall below liabilities under any parallel shift to the forward rate. Bierwag and Khang (1979) show that when the investor has a fixed budget to invest in bonds, then classical immunization (duration matching) is maxmin in the sense that it maximizes the worst possible rate of return under any parallel shift to the forward rate. Fong and Vasicek (1984) consider any perturbation to the forward curve such that the slope of the forward curve is bounded by some constant and derive a lower bound on the portfolio return over the investment horizon that is proportional to it. The constant of proportionality is a measure of interest rate risk and is called "$M$-squared". Minimization of $M$-squared renders a portfolio that minimizes the likelihood of a deviation from liabilities. Zheng (2007) considers perturbations to the forward rate that are Lipschitz continuous, derives the maximum deviation of the bond value, and applies it to a portfolio choice problem.

Several classical books and papers such as Macaulay (1938), Hicks (1939, pp. 184-188), and Samuelson (1945) discovered that the average time to payment ("duration") of a bond captures the interest rate sensitivity of the bond with respect to parallel shifts in the yield curve. Redington (1952) suggested matching the duration of the asset and liability ("immunization") to hedge against interest rate risk. Chambers, Carleton, and McEnally (1988), Nawalkha and Lacey (1988), and Prisman and Shores (1988) use polynomials to approximate the yield curve and discuss immunization using high-order duration measures. Ho (1992) introduced the concept of "key rate duration", which is the bond price sensitivity with respect to local shifts in the yield curve at certain key rates (e.g., 10-year yield). Litterman and Scheinkman (1991) use principal component analysis (PCA) to identify common factors that affect bond returns and find that the three factors called *level*, *slope*, and *curvature* explain a large fraction of the variations in returns. Using these factors, Willner (1996) defines level, slope, and curvature durations and shows how they can be used for asset-liability management. See Sydyak (2016) for a review of this literature. In a recent paper, Onatski and Wang (2021) argue that PCA based on the yield curve is prone to spurious analysis since bond yields are highly persistent. As a result, Crump and Gospodinov (2022) show

that PCA tends to favor a much lower dimension of the factor space than the true dimension, which can lead to large costs in bond portfolio management. We further discuss our contribution relative to the literature in Section 3.3.3.

## 3.2 Problem statement

### 3.2.1 Model setup

Time is continuous and denoted by $t \in [0, T]$, where $T > 0$ is the planning horizon. There are finitely many bonds indexed by $j = 1, \ldots, J$, where $J \geq 2$. The cumulative payout of bond $j$ is denoted by the (weakly) increasing function $F_j : [0, T] \to \mathbb{R}_+$. For instance, if bond $j$ is a zero-coupon bond with face value normalized to 1 and maturity $t_j$, then

$$F_j(t) = \begin{cases} 0 & \text{if } 0 \leq t < t_j, \\ 1 & \text{if } t_j \leq t \leq T. \end{cases} \tag{3.2.1}$$

Similarly, if bond $j$ continuously pays out coupons at rate $c_j$ and has zero face value, then $F_j(t) = c_j t$ for $0 \leq t \leq T$.

The fund manager seeks to immunize future cash flows against interest rate risk by forming a portfolio of bonds $j = 1, \ldots, J$. Let $F : [0, T] \to \mathbb{R}_+$ be the cumulative cash flow to be immunized and $y : [0, T] \to \mathbb{R}$ be the yield curve, which the fund manager takes as given. The present discounted value of cash flows is given by the Riemann-Stieltjes integral

$$\int_0^T \mathrm{e}^{-ty(t)} \, \mathrm{d}F(t). \tag{3.2.2}$$

Because the expression $ty(t)$ appears elsewhere, it is convenient to introduce the notation $x(t) \coloneqq ty(t)$. Note that by the definition of the instantaneous forward rate, we have

$$x(t) = \int_0^t f(u) \, \mathrm{d}u, \tag{3.2.3}$$

where $f(u)$ is the instantaneous forward rate at term $u$. Because $x$ is the integral of forward rates, we refer to it as the *cumulative discount rate*. Using $x$, we can rewrite the present discounted value

of cash flows (3.2.2) as

$$P(x) \coloneqq \int_0^T \mathrm{e}^{-x(t)}\,\mathrm{d}F(t), \tag{3.2.4}$$

which is a functional of $x$. The price $P_j(x)$ of bond $j$ can be defined analogously. The fund manager's problem is to approximate $P(x)$ using a linear combination of bonds $\{P_j(x)\}_{j=1}^J$ in a way such that the approximation is robust against perturbations to the yield curve $y$ (and hence the cumulative discount rate $x$).

### 3.2.2 Problem

We now formulate the fund manager's problem. Let $\mathcal{Z} \subset \mathbb{R}^J$ and $\mathcal{H}$ be the sets of admissible portfolios and perturbations to the cumulative discount rate, respectively. We consider the following maxmin problem:

$$\sup_{z\in\mathcal{Z}} \inf_{h\in\mathcal{H}} \left[\sum_{j=1}^J z_j P_j(x+h) - P(x+h)\right]. \tag{3.2.5}$$

Here, the objective function $\sum_{j=1}^J z_j P_j(x+h) - P(x+h)$ represents the difference between assets and liabilities, or "equity". The interpretation of the maxmin problem (3.2.5) is as follows. Given the portfolio $z \in \mathcal{Z}$, nature chooses the most adversarial perturbation $h \in \mathcal{H}$ to minimize equity. The fund manager chooses the portfolio $z$ that guarantees the highest equity under the worst possible perturbation.

### 3.2.3 Assumptions

The maxmin problem (3.2.5) is not tractable because we have not yet specified the admissible sets $\mathcal{Z}, \mathcal{H}$ and the objective function is nonlinear (not even convex) in $h$. We thus impose several assumptions to make progress.

**Assumption 1** (Discrete payouts)**.** *The bonds and liability pay out on finitely many dates, whose union is denoted by* $\{t_n\}_{n=1}^N \subset (0, T]$.

Assumption 1 always holds in practice. Under this assumption, each $F_j$ is a step function with discontinuities at points contained in $\{t_n\}_{n=1}^N$, and integrals of the form (3.2.4) reduce to summations.

**Assumption 2** (Portfolio constraint)**.** *The set of admissible portfolios $\mathcal{Z} \subset \mathbb{R}^J$ is nonempty and closed. Furthermore, all $z \in \mathcal{Z}$ satisfy value matching:*

$$P(x) = \sum_{j=1}^{J} z_j P_j(x). \tag{3.2.6}$$

Value matching (3.2.6) is merely a normalization to make the initial equity (assets minus liabilities) equal to 0. This assumption is common in the immunization literature (see, for example, Bierwag and Khang (1979)).

We now specify the space of cumulative discount rates and their perturbations. Let $C^r[0,T]$ be the vector space of $r$-times continuously differentiable functions on $[0,T]$, with the convention that $C^0[0,T] = C[0,T]$ denote the space of continuous functions. We let the space of forward rates be the Banach space of continuous functions $C[0,T]$ endowed with the supremum norm denoted by $\|\cdot\|_\infty$.[5] Since by definition the cumulative discount rate is the integral of the forward rate, if $f$ is continuous, then $x : [0,T] \to \mathbb{R}$ defined by (3.2.3) is continuously differentiable with $x(0) = 0$. We define the space of cumulative discount rates by

$$\mathcal{X} = \left\{ x \in C^1[0,T] : x(0) = 0 \right\}. \tag{3.2.7}$$

Lemma C.1.1 in the Appendix shows that $\mathcal{X}$ is a Banach space endowed with the norm $\|x\|_{\mathcal{X}} = \sup_{t\in[0,T]} |x'(t)|$. The set of admissible perturbations is a subset $\mathcal{H} \subset \mathcal{X}$. The next assumption allows us to approximate any element $x \in \mathcal{X}$.

**Assumption 3.** *There exists a countable basis $\{h_i\}_{i=1}^{\infty}$ of $\mathcal{X}$ such that for each $1 \le I \le N$, the $I \times N$ matrices $H = (h_i(t_n))$ and $G = (h_i'(t_n))$ have full row rank.*

We refer to each $h_i$ as a basis function. Assumption 3 says that the basis functions $\{h_i\}$ and their derivatives $\{h_i'\}$ are linearly independent when evaluated on the payout dates. We impose this assumption to avoid portfolio indeterminacy. In practice, we can always ensure that $H$ and $G$ have full row rank by removing certain basis functions if necessary. A typical example satisfying Assumption 3 is to let $h_i$ be a polynomial of degree $i$ with $h_i(0) = 0$ (Lemma C.1.2).

[5] As we use several different norms in this paper, we use subscripts to distinguish them. An example is the $\ell^p$ norm on $\mathbb{R}^J$ for $p = 1, 2$, which we denote by $\|\cdot\|_p$.

## 3.3 Robust asset-liability management

In this section we solve the maxmin problem (3.2.5) in the limit when the admissible set of perturbations $\mathcal{H}$ shrinks to $\{0\}$. In practice, the resulting portfolio solution is expected to provide a good hedge against the worst-case interest rate shock when the change in interest rates is small.

### 3.3.1 Robust immunization

As the set of cumulative discount rates $\mathcal{X}$ forms an infinite-dimensional vector space, we employ tools from functional analysis to analyze how prices change in response to perturbations in the discount rate, denoted by $h \in \mathcal{X}$. These perturbations can take various forms, such as a parallel shift, characterized by a constant function $h(t) \equiv c \in \mathbb{R}$, or a linear shift represented by $h(t) = ct$. We assess the price change following an arbitrary shift in the cumulative discount rate by using the Gateaux differential of $P(x)$:[6]

$$\delta P(x; h) \coloneqq \lim_{\alpha \to 0} \frac{1}{\alpha}(P(x + \alpha h) - P(x)) = -\int_0^T \mathrm{e}^{-x(t)} h(t) \, \mathrm{d}F(t). \tag{3.3.1}$$

*Remark.* The operator $h \mapsto \delta P(x; h)$ defined by (3.3.1) is a bounded linear operator from $\mathcal{X}$ to $\mathbb{R}$ (Lemma C.1.3), which is called the Fréchet derivative and denoted by $P'(x)$. Thus by definition $P'(x)h = \delta P(x; h)$. In broad terms, $P'(x)h$ quantifies the first-order impact on price change when the cumulative discount rate curve is perturbed by $h$.

Our approach to constructing a maxmin solution is based on assessing the sensitivity of assets and liabilities to perturbations in specific directions $h$. Specifically, given the basis functions $\{h_i\}_{i=1}^I$ and bonds $j = 1, \ldots, J$, we define the sensitivity matrix $A = (a_{ij}) \in \mathbb{R}^{I \times J}$, where each element $a_{ij}$ represents the sensitivity of bond $j$ (with $F = F_j$) to a perturbation evaluated at $h = h_i$. The exact expression for $a_{ij}$ is given by

$$a_{ij} \coloneqq -\frac{P_j'(x)h_i}{P(x)} = -\frac{\delta P_j(x; h_i)}{P(x)} = \frac{1}{P(x)} \int_0^T \mathrm{e}^{-x(t)} h_i(t) \, \mathrm{d}F_j(t). \tag{3.3.2}$$

Division by $P(x)$ is merely a normalization to make $a_{ij}$ dimensionless. Similarly, we define the

[6]Note that we can interchange the order of integration and differentiation using the dominated convergence theorem.

sensitivity vector $b = (b_i) \in \mathbb{R}^I$ of liabilities by

$$b_i \coloneqq -\frac{P'(x)h_i}{P(x)} = -\frac{\delta P(x; h_i)}{P(x)} = \frac{1}{P(x)} \int_0^T \mathrm{e}^{-x(t)} h_i(t) \, \mathrm{d}F(t). \tag{3.3.3}$$

If $h \in \operatorname{span}\{h_i\}_{i=1}^I$, so $h = \sum_{i=1}^I w_i h_i$ for some $w \in \mathbb{R}^I$, then under Assumption 2 we obtain

$$\lim_{\alpha \to 0} \frac{1}{\alpha P(x)} \left[ \sum_{j=1}^J z_j P_j(x + \alpha h) - P(x + \alpha h) \right] = - \langle w, Az - b \rangle ,$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product. Hence, the change in equity following an infinitesimal perturbation in the discount rate is governed by the assets and liabilities' Fréchet derivative. If the portfolio is chosen such that $Az = b$, i.e., the Fréchet derivatives of assets and liabilities are matched, then the worst-case equity is insensitive to small perturbations in the yield curve. We will use this insight to construct the maxmin solution in Theorem 3.3.3. Before doing so, we present several auxiliary results. In the discussion below, it is convenient to introduce notation for the value matching constraint, which is always assumed to hold (Assumption 2). Specifically, set $h_0 \equiv 1$ and define $a_{0j}$ using (3.3.2). Define the $1 \times J$ vector $a_0 \coloneqq (a_{0j})$ and the $(I+1) \times J$ matrix and $(I+1) \times 1$ vector

$$A_+ \coloneqq \begin{bmatrix} a_0 \\ A \end{bmatrix} \quad \text{and} \quad b_+ \coloneqq \begin{bmatrix} 1 \\ b \end{bmatrix}. \tag{3.3.4}$$

In what follows, longer proofs are relegated to Appendix C.2.

**Proposition 3.3.1** (Minmax)**.** *Suppose Assumptions 1–3 hold, $I \geq J - 1$, and $A_+$ in* (3.3.4) *has full column rank. Define the $I \times N$ matrix $G = (h_i'(t_n))$ and the set*

$$\mathcal{W} \coloneqq \left\{ w \in \mathbb{R}^I : G'w \in [-1, 1]^N \right\}. \tag{3.3.5}$$

*Then there exists $(z^*, w^*) \in \mathcal{Z} \times \mathcal{W}$ that achieves the minmax value*

$$V_I(\mathcal{Z}) \coloneqq \inf_{z \in \mathcal{Z}} \sup_{w \in \mathcal{W}} \langle w, Az - b \rangle . \tag{3.3.6}$$

*Furthermore, $V_I(\mathcal{Z}) \geq 0$, and $z \in \mathcal{Z}$ achieves $V_I(\mathcal{Z}) = 0$ if and only if $A_+ z = b_+$.*

The matrix $G = (h_i'(t_n))$ can be thought of as the $I \times N$ matrix of perturbations to forward rates. The set $\mathcal{W}$ in (3.3.5) thus characterizes the span of perturbations to the forward rate that are bounded in absolute value by one. Proposition 3.3.1 assumes that $A_+$ in (3.3.4) has full column rank. If the cumulative payouts of bonds $\{F_j\}$ and the basis functions $\{h_i\}$ are linearly independent, the matrix $A_+$ generically has full column rank and therefore a solution $(z, w) \in \mathcal{Z} \times \mathcal{W}$ to the minmax problem (3.3.6) generically exists. Appendix C.3 makes this statement precise.

The solution $z$ to the minmax problem (3.3.6) depends on the basis functions $\{h_i\}_{i=1}^I$ only through its span and it is immaterial how we parameterize these functions. Although this result is obvious, we note it as a proposition.

**Proposition 3.3.2** (Basis invariance)**.** *Let everything be as in Proposition 3.3.1 and $\mathcal{Z}^*$ be the set of solutions $z^* \in \mathcal{Z}$ to the minmax problem* (3.3.6)*. Then $V_I(\mathcal{Z})$ and $\mathcal{Z}^*$ depend on the basis functions $\{h_i\}_{i=1}^I$ only through its span.*

For any bond portfolio $z \in \mathcal{Z}$, define the portfolio share $\theta = (\theta_j) \in \mathbb{R}^J$ by

$$\theta_j \coloneqq z_j P_j(x)/P(x). \tag{3.3.7}$$

Under Assumption 2, the portfolio share $\theta$ satisfies $\sum_{j=1}^J \theta_j = 1$. Therefore the $\ell^1$ norm $\|\theta\|_1 = \sum_{j=1}^J |\theta_j|$ satisfies $\|\theta\|_1 = 1$ if and only if $\theta_j \geq 0$ for all $j$, and $\|\theta\|_1 > 1$ is equivalent to $\theta_j < 0$ for some $j$. Thus $\|\theta\|_1$ can be interpreted as a measure of leverage, which we refer to as the *gross leverage*.

To state our main result, we consider the following set of admissible perturbations to the cumulative discount rate for any $\Delta > 0$:

$$\mathcal{H}_I(\Delta) \coloneqq \left\{ h \in \operatorname{span}\{h_i\}_{i=1}^I : (\forall n) \left|h'(t_n)\right| \leq \Delta \right\}. \tag{3.3.8}$$

Because $h$ is a perturbation to the cumulative discount rate, which is the integral of the forward rate, choosing $h \in \mathcal{H}_I(\Delta)$ amounts to allowing the forward rates to change by at most $\Delta$ while spanned by the first $I$ basis functions. The following theorem is our main theoretical result.

**Theorem 3.3.3** (Robust immunization)**.** *Let everything be as in Proposition 3.3.1 and $\mathcal{H}_I(\Delta)$ be*

*as in* (3.3.8)*. Then the guaranteed equity satisfies*

$$\lim_{\Delta \downarrow 0} \frac{1}{\Delta} \sup_{z \in \mathcal{Z}} \inf_{h \in \mathcal{H}_I(\Delta)} \left[ \sum_{j=1}^{J} z_j P_j(x+h) - P(x+h) \right] = -P(x) V_I(\mathcal{Z}). \tag{3.3.9}$$

*Letting* $z^* \in \mathcal{Z}$ *be the solution to the minmax problem* (3.3.6) *and* $\theta = (\theta_j) \in \mathbb{R}^J$ *be the corresponding portfolio share defined by* (3.3.7)*, then*

$$\sup_{h \in \mathcal{H}_I(\Delta)} \left| P(x+h) - \sum_{j=1}^{J} z_j^* P_j(x+h) \right| \leq \Delta P(x) \left( V_I(\mathcal{Z}) + \frac{1}{4} \Delta T^2 \mathrm{e}^{\Delta T} (1 + \|\theta\|_1) \right). \tag{3.3.10}$$

Theorem 3.3.3 has several implications. First, (3.3.9) shows that, to the first order, the guaranteed equity is exactly $-\Delta P(x) V_I(\mathcal{Z})$ when forward rates (hence yields) are perturbed by at most $\Delta$ within the span of the basis functions. The minmax value $V_I(\mathcal{Z})$ has a natural interpretation and is the answer to the following question: "if forward rates change by at most one percentage point, what is the largest percentage point decline in the portfolio value?" The maxmin formula (3.3.9) provides an exact characterization of the worst-case outcome, and the number $V_I(\mathcal{Z})$ can be solved as the minmax value (3.3.6).

Second, the error estimate (3.3.10) shows that the solution $z^* \in \mathcal{Z}$ to the minmax problem (3.3.6) achieves the lower bound in (3.3.9), to the first order. In this sense $z^*$ is an optimal portfolio, which we refer to as the *robust immunizing portfolio*. Clearly, this immunizing portfolio is independent of $\Delta > 0$ as the minmax problem (3.3.6) does not involve $\Delta$. In addition, the minmax value (3.3.6) satisfies the following comparative statics.

**Proposition 3.3.4** (Monotonicity of minmax value)**.** *Let everything be as in Proposition 3.3.1. If* $I < I'$ *and* $\mathcal{Z} \subset \mathcal{Z}'$*, then* $V_I(\mathcal{Z}) \leq V_{I'}(\mathcal{Z})$ *and* $V_I(\mathcal{Z}) \geq V_I(\mathcal{Z}')$*.*

The result $V_I(\mathcal{Z}) \leq V_{I'}(\mathcal{Z})$ is obvious because the more basis functions we use, the more freedom nature has to select adversarial perturbations. The result $V_I(\mathcal{Z}) \geq V_I(\mathcal{Z}')$ is also obvious because the larger the set of admissible portfolios is, the more freedom the fund manager has to select portfolios.

### 3.3.2 Robust immunization with principal components

So far we have put no structure on the basis functions $\{h_i\}_{i=1}^I$ beyond Assumption 3. The set of admissible perturbations (3.3.8) depends only on span $\{h_i\}_{i=1}^I$ and the particular order or parameterization does not matter. However, in practice there could be some factor structure in the forward rate. For instance, a typical shift to the forward curve might be decomposed into the sum of a parallel shift and a nonparallel shift of a smaller size. In this section we formalize this idea and extend Theorem 3.3.3 to a setting where the perturbation in a particular direction (principal component) could be larger.

For any $\Delta_1, \Delta_2 > 0$, consider the following admissible set of perturbations:

$$\begin{aligned}&\mathcal{H}_I(\Delta_1, \Delta_2)\\ &\qquad = \left\{h \in \text{span}\,\{h_i\}_{i=1}^I : (\exists\alpha)(\forall n)\left|\alpha h_1'(t_n)\right| \le \Delta_1, \left|h'(t_n) - \alpha h_1'(t_n)\right| \le \Delta_2\right\}. \quad (3.3.11)\end{aligned}$$

Choosing $h \in \mathcal{H}_I(\Delta_1, \Delta_2)$ amounts to perturbing the forward rate in the direction spanned by the first component ($h_1'$) by a magnitude at most $\Delta_1$, and then perturbing in an arbitrary direction spanned by the first $I$ basis functions by a magnitude at most $\Delta_2$. Thus setting $\Delta_1 \gg \Delta_2$ captures the idea that $h_1$ is the first principal component. In this setting, we can generalize Theorem 3.3.3 as follows.

**Theorem 3.3.5** (Robust immunization with principal components)**.** *Let everything be as in Proposition 3.3.1 and suppose the set*

$$\mathcal{Z}_1 \coloneqq \left\{ z \in \mathcal{Z} : \sum_{j=1}^J a_{1j} z_j = b_1 \right\} \quad (3.3.12)$$

*is nonempty, where $a_{1j}$ and $b_1$ are defined by* (3.3.2) *and* (3.3.3) *with $i = 1$. Let $\mathcal{H}_I(\Delta_1, \Delta_2)$ be as in* (3.3.11)*. Then the guaranteed equity satisfies*

$$\lim \frac{1}{\Delta_2} \sup_{z \in \mathcal{Z}} \inf_{h \in \mathcal{H}_I(\Delta_1, \Delta_2)} \left[ \sum_{j=1}^J z_j P_j(x+h) - P(x+h) \right] = -P(x) V_I(\mathcal{Z}_1), \quad (3.3.13)$$

*where the limit is taken over $\Delta_1, \Delta_2 \to 0$, $\Delta_1/\Delta_2 \to \infty$, and $\Delta_1^2/\Delta_2 \to 0$. Letting $z^* \in \mathcal{Z}_1$ be the*

*solution to the minmax problem* (3.3.6) *with portfolio constraint* $\mathcal{Z}_1$*, we have*

$$\sup_{h\in\mathcal{H}_I(\Delta_1,\Delta_2)}\left|P(x+h)-\sum_{j=1}^{J}z_j^*P_j(x+h)\right|\le\Delta_2P(x)\left(V_I(\mathcal{Z}_1)+O(\Delta_2+\Delta_1^2/\Delta_2)\right). \tag{3.3.14}$$

The value added of Theorem 3.3.5 relative to Theorem 3.3.3 can be explained as follows. Comparing to (3.3.11) to (3.3.8) and applying the triangle inequality

$$\left|h'(t)\right|\le\left|\alpha h_1'(t)\right|+\left|h'(t)-\alpha h_1'(t)\right|,$$

we obtain $\mathcal{H}_I(\Delta_1,\Delta_2)\subset\mathcal{H}_I(\Delta_1+\Delta_2)$. Therefore to first-order, the maximum portfolio return loss can be bounded as

$$\underbrace{\Delta_2V_I(\mathcal{Z}_1)}_{\text{Theorem 3.3.5}}\le\underbrace{(\Delta_1+\Delta_2)V_I(\mathcal{Z})}_{\text{Theorem 3.3.3}}.$$

Thus if $\Delta_1\gg\Delta_2$ in typical situations (see Figure 3.2), then imposing the constraint $\mathcal{Z}_1$ in (3.3.12) improves the performance.[7]

*Remark.* Theorem 3.3.5 can be further generalized if we allow larger perturbations spanned by the first few basis functions. For instance, if we use the first two basis functions, we can define $\mathcal{H}_I(\Delta_1,\Delta_2,\Delta_3)$ analogously to (3.3.11) by incorporating the constraints $|\alpha_ih_i'(t_n)|\le\Delta_i$ for $i=1,2$ and $|h'(t_n)-\alpha_1h_1'(t_n)-\alpha_2h_2'(t_n)|\le\Delta_3$. The portfolio constraint (3.3.12) then becomes

$$\mathcal{Z}_2\coloneqq\left\{z\in\mathcal{Z}:\sum_{j=1}^{J}a_{ij}z_j=b_i\text{ for }i=1,2\right\}, \tag{3.3.15}$$

and the maxmin formula (3.3.13) involves $V_I(\mathcal{Z}_2)$.

### 3.3.3 Relation to existing literature

In this section we discuss in some detail how Theorem 3.3.3 is related to the existing literature. The following corollary shows that when $I=J-1$ and there is no portfolio constraint beyond value matching, the immunizing portfolio can be solved explicitly.

[7] On the other hand, if $\Delta_1\sim\Delta_2$, then imposing the constraint $\mathcal{Z}_1$ worsens the performance because $V_I(\mathcal{Z}_1)\ge V_I(\mathcal{Z})$ by Proposition 3.3.4.

**Corollary 3.3.6** (Robust immunization with $I = J-1$)**.** *Let everything be as in Proposition 3.3.1 and suppose that the only portfolio constraint is value matching* (3.2.6)*, so the set of admissible portfolios is*

$$\mathcal{Z}_0 \coloneqq \left\{ z \in \mathbb{R}^J : P(x) = \sum_{j=1}^{J} z_j P_j(x) \right\}. \tag{3.3.16}$$

*If $I = J-1$ and the square matrix $A_+$ in* (3.3.4) *is invertible, then the unique solution to* (3.3.6) *is $z^* = A_+^{-1} b_+$, with $V_I(\mathcal{Z}) = 0$.*

*Proof.* Immediate from the proof of Proposition 3.3.1. ■

*Remark* 5. The special case of Corollary 3.3.6 with $I = J-1 = 1$ and $h_1(t) = t$ reduces to classical immunization that matches the bond value and duration. To see this, recall that the duration of the cash flow $F$ is defined by the weighted average time to payment

$$D = \frac{\int_0^T t\mathrm{e}^{-ty(t)}\,\mathrm{d}F(t)}{\int_0^T \mathrm{e}^{-ty(t)}\,\mathrm{d}F(t)}.$$

Using the definition $x(t) = ty(t)$ and (3.3.1), the duration can be rewritten as

$$D = \frac{\int_0^T t\mathrm{e}^{-x(t)}\,\mathrm{d}F(t)}{\int_0^T \mathrm{e}^{-x(t)}\,\mathrm{d}F(t)} = -\frac{P'(x)h_1}{P(x)} = b_1,$$

where $h_1(t) = t$ and we have used (3.3.3). A similar calculation implies that the duration of the immunizing portfolio is

$$-\frac{\sum_{j=1}^J z_j P_j'(x) h_1}{\sum_{j=1}^J z_j P_j(x)} = -\frac{\sum_{j=1}^J z_j P_j'(x) h_1}{P(x)} = \sum_{j=1}^J a_{1j} z_j$$

using value matching (3.2.6) and (3.3.2). Therefore if $z = A_+^{-1} b_+$, so $A_+ z = b_+$, the duration is matched. By the same argument, setting $I = J-1$ and $h_i(t) = t^i$ reduces to high-order duration matching ($I = J-1 = 2$ is convexity matching).

*Remark.* Proposition 3.3.4 explains why high-order duration matching ($I = J-1$, no portfolio constraint, and $h_i(t) = t^i$) does not necessarily have good performance (Mantilla-Garcia, Martellini, Milhau, and Ramirez-Garrido, 2022). When $I = J-1$, as we increase $I$, both the number of basis functions $I$ and the set of admissible portfolios $\mathcal{Z}$ expand. Because increasing $I$ makes $V_I(\mathcal{Z})$ larger

but expanding $\mathcal{Z}$ makes it smaller, the combined effect could go either way.

In addition to the setting in Corollary 3.3.6, if the liability pays out on a single date and the immunizing portfolio does not involve shortsales, we can obtain the following global result.

**Proposition 3.3.7** (Guaranteed funding)**.** *Let everything be as in Corollary 3.3.6 and suppose that the liability pays out on a single date. If* $z^* = A_+^{-1} b_+ \geq 0$*, then for all* $h \in \operatorname{span}\{h_i\}_{i=1}^I$ *we have*

$$\sum_{j=1}^{J} z_j^* P_j(x+h) \geq P(x+h). \tag{3.3.17}$$

*Remark.* Our maxmin result (Theorems 3.3.3 and 3.3.5) is quite different from the existing literature such as Fisher and Weil (1971) and Bierwag and Khang (1979). To the best of our knowledge, in this literature it is always assumed that the liability pays out on a single date and the portfolio does not involve shortsales ($z \geq 0$) yet this constraint does not bind. Under these assumptions, Proposition 3.3.7 shows that the immunizing portfolio always funds the liability, which generalizes the result of Fisher and Weil (1971) (who proved (3.3.17) for $I = J-1 = 1$ and $h_1(t) = t$). However, this result is quite restrictive because liabilities could be paid out over time and shortsales are essential when the maturity of the liability is very long (such as pensions). Our maxmin result (3.3.9) accommodates arbitrary liability structures and portfolio constraints.

### 3.3.4 Implementation

To implement robust immunization, we need to choose the basis functions $\{h_i\}_{i=1}^I$. For each $i$, it is natural to choose $h_i$ such that $h_i$ is a polynomial of degree $i$ with $h_i(0) = 0$, for Assumption 3 then holds (Lemma C.1.2). By basis invariance (Proposition 3.3.2), any choice of such a basis will result in the same immunizing portfolio.

However, we suggest using Chebyshev polynomials because they enjoy good approximation properties (Trefethen, 2019, Ch. 2–4). To be more specific, let $T_n : [-1,1] \to \mathbb{R}$ be the $n$-degree Chebyshev polynomial defined by $T_n(\cos\theta) = \cos n\theta$ and setting $x = \cos\theta$. We map $[0,T]$ to $[-1,1]$ using $t \mapsto x = 2t/T - 1$, and define $g_i : [0,\infty) \to \mathbb{R}$ by

$$g_i(t) = T_{i-1}(2t/T - 1) \tag{3.3.18}$$

so that we can allow any (continuous) perturbations to the forward rate for $t \in [0, T]$. Then our basis functions for perturbing the cumulative discount rate (the integral of the forward rate) can be defined by

$$h_i(t) = \int_0^t g_i(u)\,\mathrm{d}u \tag{3.3.19}$$

for each $i$. The following lemma provides explicit formulas for the basis functions (3.3.19).

**Lemma 3.3.8** (Chebyshev basis for forward rate)**.** *Let $T_n$ be the $n$-degree Chebyshev polynomial defined by $T_n(\cos\theta) = \cos n\theta$ and setting $x = \cos\theta$. For each $i$, the basis function $h_i$ in* (3.3.19) *can be expressed as*

$$h_1(t) = t, \tag{3.3.20a}$$

$$h_2(t) = \frac{1}{4}T\left((2t/T-1)^2 - 1\right), \tag{3.3.20b}$$

*and for $i \geq 3$,*

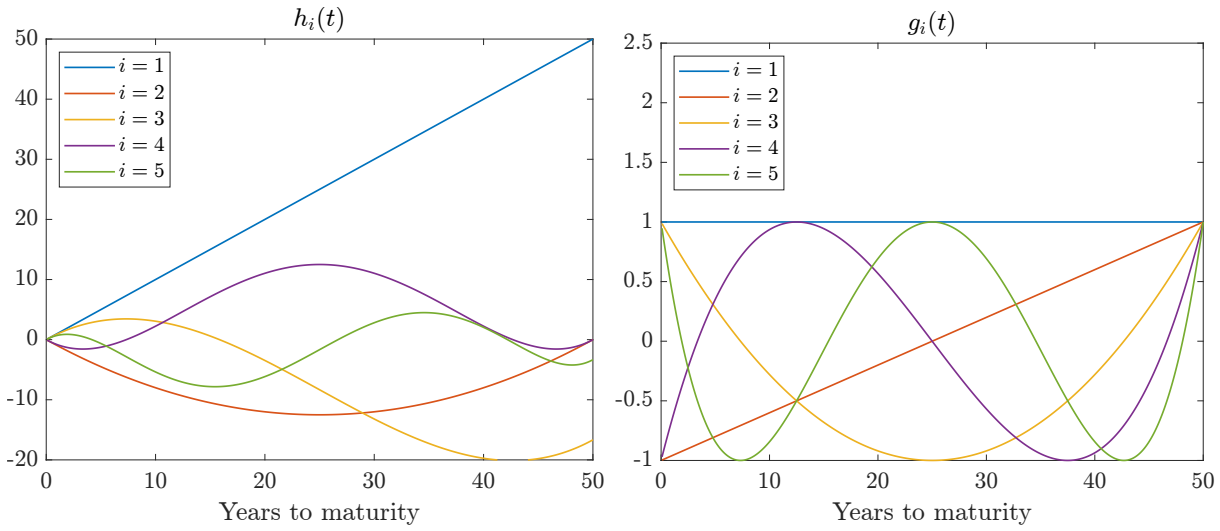
$$h_i(t) = \frac{1}{4}T\left(\frac{T_i(2t/T-1)}{i} - \frac{T_{i-2}(2t/T-1)}{i-2} + \frac{2(-1)^i}{i(i-2)}\right). \tag{3.3.20c}$$

Figure 3.1a shows the graphs of the first few basis functions (3.3.20) for $T = 50$ years. Figure 3.1b shows the graphs of $g_i = h_i'$ in (3.3.18), which are the rows of the matrix $G$ in Proposition 3.3.1.

We now describe the algorithm to implement robust immunization in practice. Although the underlying theory (which heavily relies on functional and numerical analysis) may not be familiar to practitioners, the implementation only requires little more than basic linear algebra and linear programming.

**Robust Immunization.**

(i) Let $\mathbf{t} = (t_1, \ldots, t_N)$ be the $1 \times N$ vector of asset/liability payout dates and $T = t_N$ be the planning horizon. Let $\mathbf{y} = (y_1, \ldots, y_N)$ be the $1 \times N$ vector of yields, $\mathbf{f} = (f_1, \ldots, f_N)$ be the $1 \times N$ vector of liabilities, and $\mathbf{F} = (f_{jn})$ be the $J \times N$ matrix of bond payouts.




**(a)** Cumulative discount rate.

**(b)** Forward rate.

**Figure 3.1. Basis functions of robust immunization**.

(ii) Let $I \geq J-1$, define the basis functions by (3.3.20), evaluate at each $t_n$, and construct the $I \times N$ matrix of basis functions $\mathbf{H} = (h_i(t_n))$ and their derivative $\mathbf{G} = (h_i'(t_n)) = (g_i(t_n))$. Define the $1 \times N$ vector of zero-coupon bond prices $\mathbf{p} = \exp(-\mathbf{y} \odot \mathbf{t})$, where $\odot$ denotes entry-wise multiplication (Hadamard product).

(iii) Define the $I \times J$ matrix $A$, $I \times 1$ vector $b$, and $1 \times J$ vector $a_0$ by

$$A \coloneqq (\mathbf{H}\,\mathrm{diag}(\mathbf{p})\mathbf{F}')/(\mathbf{p}\mathbf{f}'), \qquad b \coloneqq \mathbf{H}\,\mathrm{diag}(\mathbf{p})\mathbf{f}'/(\mathbf{p}\mathbf{f}'), \qquad a_0 \coloneqq \mathbf{p}\mathbf{F}'/(\mathbf{p}\mathbf{f}'),$$

where $\mathrm{diag}(\mathbf{p})$ denotes the diagonal matrix with diagonal entries given by $\mathbf{p}$. Define the $(I+1) \times J$ matrix $A_+$ and $(I+1) \times 1$ vector $b_+$ by

$$A_+ \coloneqq \begin{bmatrix} a_0 \\ A \end{bmatrix} \quad \text{and} \quad b_+ \coloneqq \begin{bmatrix} 1 \\ b \end{bmatrix}.$$

(iv) If $I = J-1$ and there are no portfolio constraints, calculate the immunizing portfolio as $z^* = A_+^{-1} b_+$. Otherwise, solve the minmax problem (3.3.6).

Note that the inner maximization in (3.3.6) is a linear programming problem with $I$ variables

and $2N$ inequality constraints, which is straightforward to solve numerically even when $N$ is large (a few hundred in typical applications). The outer minimization is a convex minimization problem with $J$ variables, which is also straightforward to solve numerically.

## 3.4 Evaluation: static hedging

In this section we evaluate the performance of robust immunization and other existing methods using a numerical experiment in a static setting.

### 3.4.1 Experimental design

#### Data and yield curve model

We obtain daily U.S. Treasury nominal yield curve data from November 25, 1985 to September 2022 from the Federal Reserve.[8] We denote the days by $s = 1, \dots, S$, where $S = 9{,}201$ is the sample size. These daily yield curves are estimated using the methodology of Gürkaynak, Sack, and Wright (2007), who assume that the instantaneous forward rate at term $t$ is specified by the Svensson (1994) model

$$f(t) = \beta_0 + \beta_1 \exp(-t/\tau_1) + \beta_2(t/\tau_1)\exp(-t/\tau_1) + \beta_3(t/\tau_2)\exp(-t/\tau_2), \tag{3.4.1}$$

where $\beta_0, \beta_1, \beta_2, \beta_3 \in \mathbb{R}$ and $\tau_1, \tau_2 > 0$ are parameters. The functional form (3.4.1) allows for two humps in the forward curve that are governed by the parameters $\tau_1$ and $\tau_2$. Integrating the forward rate in (3.4.1), we obtain the cumulative discount rate

$$x(t) = \beta_0 t - \beta_1\tau_1 \exp(-t/\tau_1) - \beta_2(t+\tau_1)\exp(-t/\tau_1) - \beta_3(t+\tau_2)\exp(-t/\tau_2).$$

Note that the parameters in (3.4.1) change over time, but we suppress the time subscript $s$ for notional clarity. Our data set includes the estimated parameters $(\beta_0, \beta_1, \beta_2, \beta_3, \tau_1, \tau_2)$ for each day, with which we can evaluate the forward curve (and hence the yield and cumulative discount curves) at arbitrary term $t \geq 0$.

*Remark.* The estimated parameters of Gürkaynak, Sack, and Wright (2007) go back all the way to 1961, but we only use their data beyond 11/25/1985 when bonds with a maturity of 30 years

[8] https://www.federalreserve.gov/data/nominal-yield-curve.htm

were introduced in the market. The authors caution against extrapolation of the forward rate beyond the maximum available bond maturity. Anticipating our empirical application, we need to obtain forward rates with a maturity up to 50 years. Since extrapolation is still necessary in this case, we extrapolate the forward rate by a constant beyond the 30-year maturity. This approach is motivated by no-arbitrage arguments which stipulate that the long term forward rate is constant (Dybvig, Ingersoll, and Ross, 1996). In Appendix C.5.1, we show how the constant forward rate assumption affects our estimate of the yield curve.

**Approximating forward rate changes by basis functions**

Our theory is based on the assumption that changes in the forward rate can be approximated by the basis functions. To evaluate this assumption, we regress the $d$-day ahead forward rate changes on the basis functions $g_i$ in (3.3.18) and calculate a goodness-of-fit measure denoted by $R^2$ (see Appendix C.5.2 for details).

The left panel of Figure 3.2 shows this goodness-of-fit measure $R^2$ for various horizons $d$. The goodness-of-fit seems to be independent of $d$ except when $I = 1$. The first basis function (constant) explains between 50 and 65% of variations in the forward rate changes, and the first two basis functions (constant and linear) explain about 80%. This result shows that it can be important to account for principal components in constructing the robust immunization portfolio, as in Theorem 3.3.5. The right panel shows the unexplained component $1 - R^2$ as we include more basis functions. We can see that setting $I = 10$ captures about 99.9% ($1 - R^2 < 10^{-3}$) of variations in the forward rate changes.

**Cash flow and immunization methods**

We now turn to the immunization design. We suppose that the future cash flows of the liability are equal to $1/(12T)$ every month for $T = 50$ years (so the cumulative cash flow is normalized to 1), and the bonds available for trade are zero-coupon bonds with face value 1 and years to maturity being a subset of $\{1, 5, 10, 20, 30\}$. We intentionally choose a long maturity of 50 years for the cash flows because it is of interest to study how the yield curve at the long end affects the performance of the immunization methods.

We consider three immunization methods. The first method is high-order duration matching (`HD`) explained in Remark 5, which is a special case of robust immunization by setting $I = J - 1$

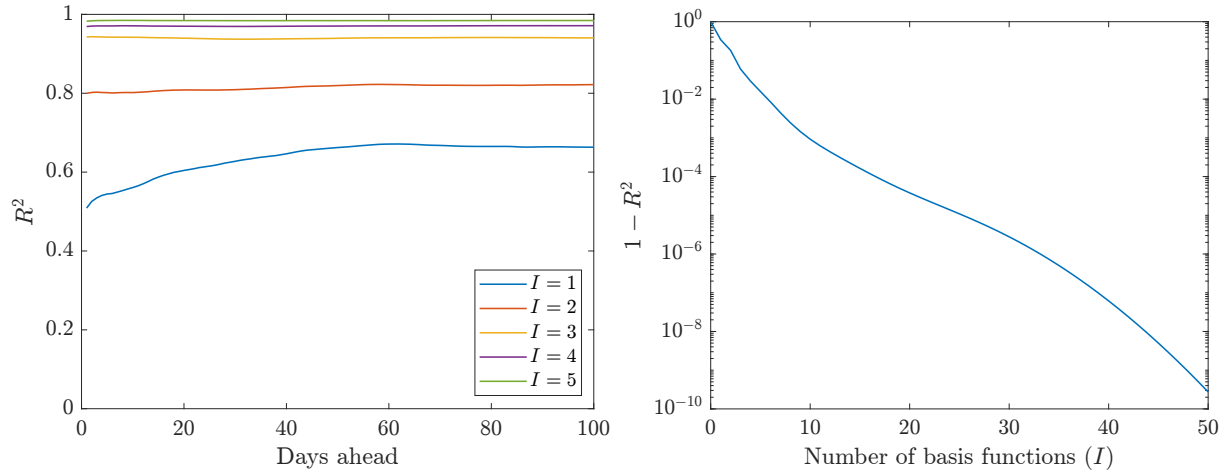



**Figure 3.2. Goodness-of-fit of forward rate change approximation**.

Note: The left panel shows $R^2$ for each $d$-day ahead change in the forward rate using the basis functions $\{g_i\}_{i=1}^I$ in (3.3.18) as regressors. The right panel shows the combined $1 - R^2$ as we increase the number of basis functions $I$. See Appendix C.5.2 for details.

and $h_i(t) = t^i$. By basis invariance (Proposition 3.3.2), we can choose any polynomial basis, so we use the Chebyshev functions in Lemma 3.3.8 with $T = 50$. The second method is key rate duration matching (`KRD`) proposed by Ho (1992) and explained in Appendix C.5.3. In short, this method is designed to match the liability and portfolio sensitivity to interest rate changes at pre-specified maturities. The third method is our proposed robust immunization method (`RI`) with the Chebyshev basis for the forward rate in Lemma 3.3.8. Motivated by the right panel of Figure 3.2, we set the number of basis functions to $I = 10$. For the portfolio constraint, motivated by Theorem 3.3.5 and the left panel of Figure 3.2, we consider value matching only ($\mathcal{Z}_0$ in (3.3.16)), value- and duration matching ($\mathcal{Z}_1$ in (3.3.12)) and value-, duration- and convexity matching. We refer to these methods as `RI`(0), `RI`(1) and `RI`(2) respectively.[9]

For each method, we consider immunizing the cash flows with $J = 2, 3, 4, 5$ bonds. $J = 2$ corresponds to using the 1- and 30-year zero-coupon bonds, and we add the 5-, 10-, and 20-year bond for $J = 3, 4, 5$, respectively. Note that for `HD`, $J = 2$ is simply classical immunization with duration matching; $J = 3$ is duration and convexity matching. For `KRD`, we use the longer maturity bonds to match the key rates and we use the remaining shortest maturity bond to match value. For example, in case $J = 3$, we use the 30- and 5-year bond to match the 30- and 5-year key rate of liabilities and we use the remaining 1-year bond to match the value of liabilities.

[9]The `RI`(2) method is defined only when $J \geq 3$ because otherwise the portfolio constraint $\mathcal{Z}_2$ is generally empty.

**Return error**

Suppose that on day $s$, the fund manager immunizes future cash flows with a bond portfolio $z_s = (z_{sj})$ constructed by the `HD`, `KRD`, and `RI` methods. Motivated by the error estimate (3.3.10), we evaluate each method using the absolute return error on day $s + d$ defined by

$$\frac{1}{P(x_s)} \left| P(x_{s+d}) - \sum_{j=1}^{J} z_{sj} P_j(x_{s+d}) \right|, \tag{3.4.2}$$

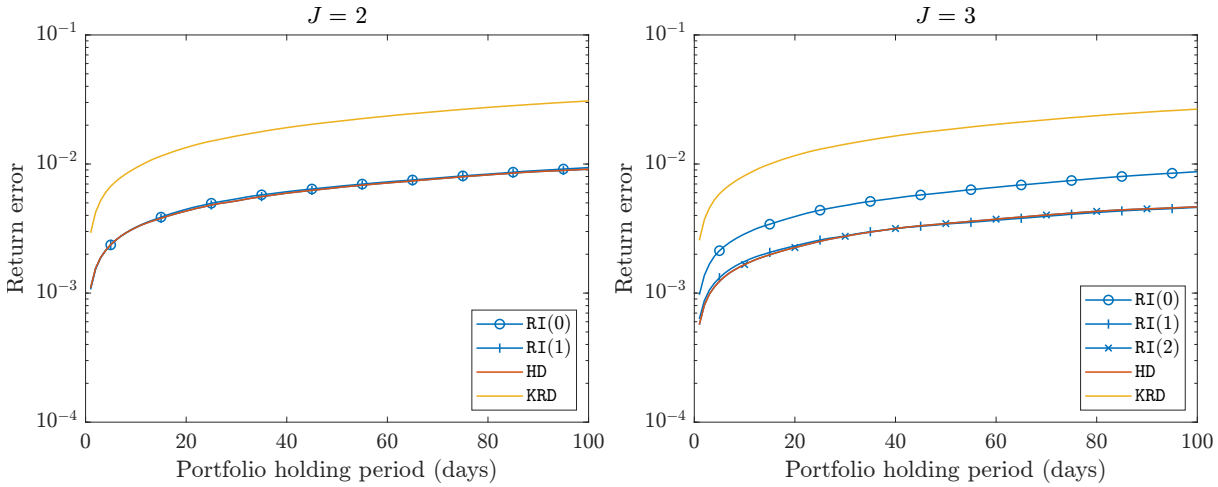
where $x_s(t)$ is the cumulative discount rate on day $s$ for term $t$ and we consider the portfolio holding period of $d = 1, \ldots, 100$ days.[10] The performance measure (3.4.2) can be understood as the return error if after forming the immunizing portfolio on day $s$, the yield curve instantaneously shifts to that of day $s + d$. In this sense the return error (3.4.2) is a performance measure of static hedging. We address dynamic hedging in Section 3.5.

### 3.4.2 Results

Figure 3.3 shows the return error defined by (3.4.2) averaged over the sample period. The return error worsens with longer portfolio holding periods ($d$) for all bond quantities and methods because of greater yield curve fluctuations. When there are only two bonds ($J = 2$, Figure 3.3a), by construction the `HD` and `RI`(1) method agree and they achieve the lowest return error. When there are three bonds ($J = 3$, Figure 3.3b), by construction the `HD` and `RI`(2) methods agree, and they achieve the lowest return error, with `RI`(1) close behind. When there are four bonds ($J = 4$, Figure 3.3c), `RI`(1) clearly outperforms all other methods. Finally, in case of five bonds ($J = 5$, Figure 3.3d), `RI`(1) and `RI`(2) are the best performing methods with `RI`(2) being slightly more accurate over short horizons whereas `RI`(1) is more accurate over longer holding periods. Overall, the lowest error is achieved by `RI`(1) with four bonds. Turning to the existing approaches in the literature, we see that `HD` does well only for $J \leq 3$, while the performance of `KRD` is only comparable to robust immunization in case of using five bonds.

Figure 3.3 presents only average return errors. To evaluate the performance of each method under adversarial circumstances, Table 3.1 presents the mean, 95- and 99 percentiles of the return

[10] We also considered the relative pricing error $\frac{1}{P(x_{s+d})} \left| P(x_{s+d}) - \sum_{j=1}^{J} z_{sj} P_j(x_{s+d}) \right|$ but it makes no material difference because $P(x_s)$ and $P(x_{s+d})$ have the same order of magnitude.




**(a)** 2 bonds. **(b)** 3 bonds.

**(c)** 4 bonds. **(d)** 5 bonds.

**Figure 3.3. Return error for different holding periods**.

Note: The figure presents the return error in portfolio value defined by (3.4.2) over various holding periods, averaged over the entire sample period. `RI`(0): robust immunization with a value matching; `RI`(1): robust immunization with value and duration matching; `RI`(2): robust immunization with value, duration, and convexity matching; `HD`: high-order duration matching; `KRD`: key rate duration matching. The panels in the figure show the error for different number of bonds $J$ used to construct the immunizing portfolio.

error for a portfolio holding period of 30 days. According to this table, the performance of the `HD` method is non-monotonic, which performs best when $J = 3$ but deteriorates when $J \geq 4$. The performance of the `KRD` method monotonically improves with $J$, but it is accurate only when $J = 5$. In contrast, `RI`(1) and `RI`(2) perform well with any number of bonds and their return errors are an order of magnitude lower compared to `HD` and `KRD` when $J \geq 4$.

We can summarize the findings in Figure 3.3 and Table 3.1 as follows: (i) Regardless of the number of bonds, one of the robust immunization (`RI`) methods achieves the lowest return

**Table 3.1. Return error (%) for 30-day holding period**.

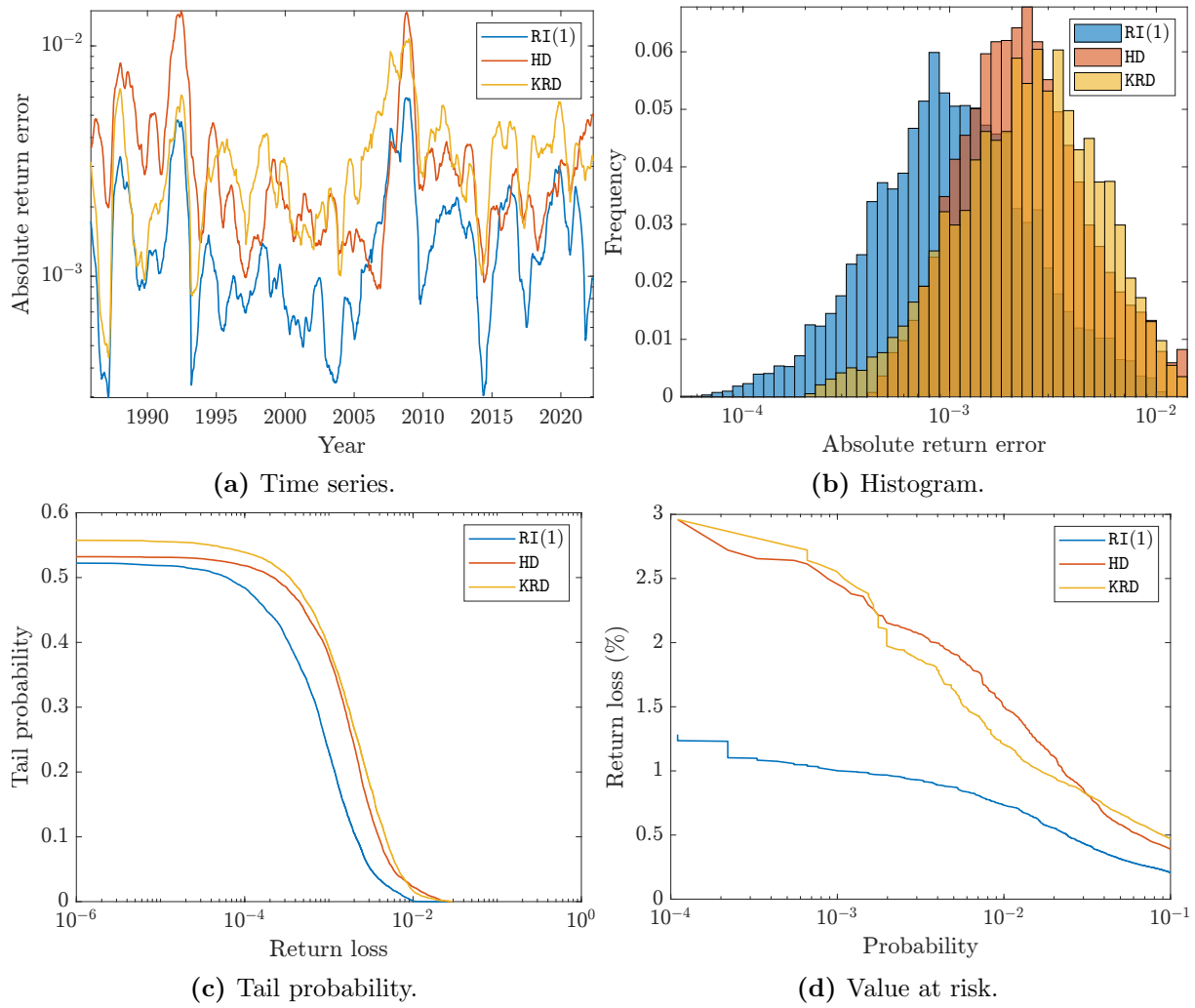
| Method: | RI(0) | RI(1) | RI(2) | HD | KRD |
|---|---|---|---|---|---|
| Mean | | | | | |
| $J = 2$ | 0.54 | **0.52** | - | **0.52** | 1.65 |
| $J = 3$ | 0.48 | **0.28** | **0.28** | **0.28** | 1.42 |
| $J = 4$ | 0.54 | **0.12** | 0.19 | 1.02 | 0.85 |
| $J = 5$ | 0.44 | **0.19** | **0.19** | 1.9 | 0.28 |
| $95^{\text{th}}$ percentile | | | | | |
| $J = 2$ | 1.66 | **1.60** | - | **1.60** | 4.3 |
| $J = 3$ | 1.32 | 0.87 | **0.86** | **0.86** | 3.67 |
| $J = 4$ | 1.52 | **0.43** | 0.58 | 3.44 | 2.2 |
| $J = 5$ | 1.37 | 0.64 | **0.54** | 6.2 | 0.91 |
| $99^{\text{th}}$ percentile | | | | | |
| $J = 2$ | 2.38 | **2.49** | - | **2.49** | 6.85 |
| $J = 3$ | 2.19 | 1.84 | **1.77** | **1.77** | 5.99 |
| $J = 4$ | 2.53 | **0.85** | 1.17 | 7.62 | 3.59 |
| $J = 5$ | 2.42 | 1.07 | **0.91** | 15.15 | 1.49 |

Note: See Figure 3.3 caption. The best performing method is indicated in **bold**.

error, and generally RI(1) (matching value and duration) or RI(2) (matching value, duration, and convexity) is the best. (ii) The performance of the HD method is non-monotonic in $J$, performing best with $J = 3$ but poorly with $J \geq 4$. (iii) The performance of KRD is poor for $J \leq 4$ and good for $J = 5$.

We next compare the performance of the best specification for each method. For example, we set $J = 3$ for HD and $J = 5$ for KRD, and we consider RI(1) for robust immunization with $J = 4$ bonds. Figure 3.4a shows the time series plot of the return error for each immunization method. We see that RI(1) is dominating the other methods almost uniformly over the entire sample period. Furthermore, KRD meaningfully outperforms HD only before 1990. Figure 3.4b shows the histogram of the absolute return errors (3.4.2). We can see that large return errors tend to be less frequent with RI(1). To see this formally, Figure 3.4c plots the survival probability of return losses (defined analogously to (3.4.2) but without taking absolute values) above various thresholds. The fact that RI(1) has lower tail (survival) probability than other methods implies that losses are less likely. Figure 3.4d plots the value at risk (VaR) of each method. The value at risk is the quantile of the return distribution and hence the graph plots the size of the return loss corresponding to the

specified loss probability. RI(1) uniformly has the lowest value at risk. These findings are consistent with Theorem 3.3.5 because the robust immunization method is designed to maximize the return error under the most adversarial perturbation to the cumulative discount rate.




**(a)** Time series. **(b)** Histogram.

**(c)** Tail probability. **(d)** Value at risk.

**Figure 3.4. Comparison of best specifications**

Note: The figures compare the best specification for each method using 4 bonds for RI(1), 3 bonds for HD and 5 bonds for KRD. Return errors are averaged at every time period over each $d$-day ahead forecast, where $d = 1, \ldots, 100$. The time series plot in Figure 3.4a shows the 180-day moving average for visibility.

We also test more formally whether the absolute return errors of RI(1) dominate HD and KRD. To do so, we use the nonparametric sign test which can be used to test whether the median absolute return error is the same for both methods. More details about this test are described in Appendix C.5.4, where we show that the 30-day return error for RI(1) is significantly better than the best performing HD and KRD method.

**Leverage**

To shed light on the observation that the performance in Figure 3.3 is non-monotonic for `HD`, Table 3.2 shows the gross leverage ($\ell^1$ norm) of the portfolio shares $\|\theta\|_1 = \sum_{j=1}^{J} |\theta_j|$. The leverage for `HD` portfolios is rather pronounced for $J = 4, 5$ compared to `RI`, both in median and in the right tail. Mantilla-Garcia, Martellini, Milhau, and Ramirez-Garrido (2022) show that levered portfolios can lead to poor out-of-sample hedging, which can explain the poor performance of `HD` for $J = 4, 5$ in Figure 3.3.

**Table 3.2. $\ell^1$ norm of investment shares**.

| Method: | `RI`(0) | `RI`(1) | `RI`(2) | `HD` | `KRD` |
|---|---|---|---|---|---|
| Median | | | | | |
| $J = 2$ | 1 | 1 | 1 | 1 | 1 |
| $J = 3$ | 1 | 1 | 1 | 1 | 1 |
| $J = 4$ | 1 | 1 | 1 | 5.49 | 1 |
| $J = 5$ | 1 | 1 | 1.09 | 15.55 | 1 |
| $95^{\text{th}}$ percentile | | | | | |
| $J = 2$ | 1 | 1 | 1 | 1 | 1 |
| $J = 3$ | 1.02 | 1 | 1.27 | 1.27 | 1 |
| $J = 4$ | 1 | 1 | 1.18 | 12.12 | 1 |
| $J = 5$ | 1 | 1 | 1.91 | 30.07 | 1.01 |
| $99^{\text{th}}$ percentile | | | | | |
| $J = 2$ | 1 | 1 | 1 | 1 | 1 |
| $J = 3$ | 1.05 | 1 | 1.31 | 1.31 | 1 |
| $J = 4$ | 1 | 1 | 1.17 | 13.69 | 1 |
| $J = 5$ | 1 | 1 | 2.12 | 33.6 | 1.05 |

Note: This table shows the $\ell^1$ norm of the investment shares, $\|\theta\|_1$, for robust immunization with a value matching constraint (`RI`(0)), robust immunization with a value- and duration matching constraint (`RI`(1)), robust immunization with a value-, duration- and convexity matching constraint (`RI`(2)), high-order duration matching (`HD`) and key rate duration matching (`KRD`).

## 3.5 Evaluation: dynamic hedging

Although the static hedging experiment in Section 3.4 may be informative, it only addresses the performance of various immunization methods under a one-shot instantaneous change in the yield curve. In practice, the fund manager will rebalance the portfolio over time, in which case the yield curve as well as the bond maturities change. In this section, to evaluate the performance of various immunization methods under practical situations, we conduct a dynamic hedging

experiment using simulated yield curves.

### 3.5.1 Implementing dynamic hedging

Let $\{s_n\}_{n=0}^N$ be the portfolio rebalancing dates (with the normalization $s_0 = 0$) and assume that the coupon payment dates of the liability are contained in this set. For simplicity let $s_n = n\Delta$ with $\Delta > 0$ so the dates are evenly spaced, although this is inessential. The liability pays $f_s \geq 0$ at time $s > 0$. The fund manager can use $J$ zero-coupon bonds with face value 1 and maturities $\{t_j\}_{j=1}^J$ to hedge the liability. We introduce the following notations:

$$\begin{aligned} x_s(t) &= \text{cumulative discount rate for term } t \text{ at time } s, \\ P_s &= \text{present value of liability at time } s, \\ V_s &= \text{net asset value (NAV) of fund at time } s, \\ z_s = (z_{sj}) &= \text{immunizing portfolio at time } s, \\ C_s &= \text{cash position at time } s, \\ R_s &= \text{gross short rate at time } s. \end{aligned}$$

We now describe how to calculate these quantities recursively. At time $s$, the present value of the liability (after coupon payment) is

$$P_s \coloneqq \sum_{n:s_n>s} \mathrm{e}^{-x_s(s_n-s)} f_{s_n}.$$

Note that at time $s$, the remaining term of the $n$-th payment is $s_n - s$ and we only retain future payments in the sum. Let $s^- = s - \Delta$ denote the previous rebalancing period. The NAV of the fund consists of the present value of the bond and cash positions carried over from the previous period minus the current liability payment, which is

$$V_s \coloneqq \underbrace{R_{s^-} C_{s^-}}_{\text{cash}} + \underbrace{\sum_{j=1}^{J} z_{s^-j} \mathrm{e}^{-x_s(t_j-\Delta)}}_{\text{bond}} - \underbrace{f_s}_{\text{liability}}.$$

Here, note that the cash position earns a (predetermined) gross return $R_{s^-}$, and the zero-coupon

bonds have shorter maturities $t_j - \Delta$ because time has passed. The equity (asset minus liability) is therefore

$$\begin{aligned} E_s &\coloneqq V_s - P_s \\ &= R_{s^-} C_{s^-} + \sum_{j=1}^{J} z_{s^- j} \mathrm{e}^{-(x+h)(t_j - \Delta)} - f_s - \sum_{n: s_n > s} \mathrm{e}^{-(x+h)(s_n - s)} f_{s_n} \\ &= R_{s^-} C_{s^-} - f_s + \sum_{j=1}^{J} z_{s^- j} \mathrm{e}^{-(x+h)(t_j - \Delta)} - \sum_{n: s_n - \Delta - s^- > 0} \mathrm{e}^{-(x+h)(s_n - \Delta - s^-)} f_{s_n}, \end{aligned} \tag{3.5.1}$$

where $x = x_{s^-}$ denotes the cumulative discount rate at $s^-$ and $h = x_s - x_{s^-}$ denotes the perturbation in the cumulative discount rate. As an illustration, consider the robust immunization method introduced in Section 3.3. The fund manager's problem at time $s^-$ is to maximize the worst case equity, where the equity is defined by $E_s$ in (3.5.1). Shifting $s^-$ to $s$, the time $s$ objective function is then

$$E_{s+\Delta}(z, x+h) \coloneqq R_s C_s - f_{s+\Delta} + \sum_{j=1}^{J} z_{sj} \mathrm{e}^{-(x+h)(t_j - \Delta)} - \sum_{n: s_n - \Delta - s > 0} \mathrm{e}^{-(x+h)(s_n - \Delta - s)} f_{s_n},$$

where $x = x_s$ is the current cumulative discount rate. Because $f_{s+\Delta}$ is predetermined and $C_s$ is determined by the budget constraint and hence independent of the perturbation $h$, the dynamic hedging problem reduces to the static hedging problem discussed in Section 3.3 except that *all payments need to be treated as if their maturities are reduced by* $\Delta$. This modification takes into account the passage of time and hence the reduction in bond maturities by the next rebalancing date. For example, if the time to rebalancing is one month, a 1-year zero coupon bond is treated as if it is an 11-month bond.

Given the current cumulative discount rate $x_s$, it is straightforward to apply various immunizing methods to bonds and liability with maturities reduced by $\Delta$. Suppose the new (time $s$) immunizing portfolio $z_s = (z_{sj})$ is chosen. Then the cash position is the difference between the NAV and portfolio value, which is

$$C_s = V_s - \sum_{j=1}^{J} z_{sj} \mathrm{e}^{-x_s(t_j)}.$$

Note that although we reduce the maturities by $\Delta$ to form the portfolio, we use the actual maturities to evaluate the portfolio value and define the cash position. Initializing at $V_0 = P_0$ (100% funding), we can implement dynamic hedging by repeating this procedure for $s = \Delta, 2\Delta, \ldots$. We evaluate the quality of the hedge at time $s$ using the absolute return error

$$\frac{1}{P_{s^-}} \left| V_s - P_s \right|. \tag{3.5.2}$$

### 3.5.2 Experimental design

In Section 3.4, we used the parsimonious Svensson (1994) model fitted to the historical yield curve data to evaluate the performance of static hedging. Unlike static hedging, where we only consider changes to the yield curve over short horizons, in dynamic hedging the yield curve changes over long horizons have a large impact on portfolio performance. This feature makes it problematic to use historical data for performance evaluation. For instance, suppose that a particular portfolio selection method over-weights in long-term bonds. Because historical yields have been trending downwards during the 1985–2022 period, this method may appear to have a good performance. However, the opposite is true had the yields been trending upwards.

For this reason, in our dynamic hedging experiment, we only use simulated yield curves generated from a no-arbitrage term structure model. Specifically, we apply the Ang, Bekaert, and Wei (2008) 3-factor regime switching model. By simulating yields from this (stationary) regime-switching model, we can evaluate the performance of various immunization methods under a wide variety of yield curves. A more detailed description of the model, as well as the data used to estimate the model is provided in Appendix C.4.[11]

We implement the dynamic hedging approach using the same liability and zero-coupon bonds from the static problem in Section 3.4.1. We use all 5 bonds for immunization for `RI`(1) and `KRD`. In contrast, we only use 3 bonds for `HD` since the performance for $J > 3$ is comparatively worse relative to the other methods (see Figure 3.3). Since we estimate the yield curve model of Ang, Bekaert, and Wei (2008) based on quarterly data, we assume that the immunizing portfolio is rebalanced every quarter. We analyze the performance over a period of 10 years and repeat the

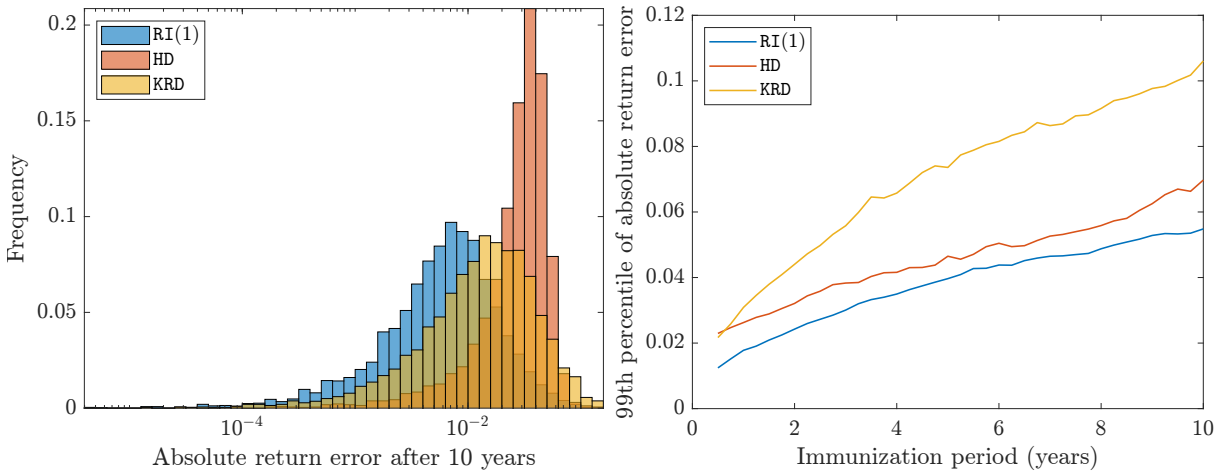
[11] We chose to estimate the model ourselves instead of using the parameters reported in Ang, Bekaert, and Wei (2008, Table III) to better reflect the evolution in yields over the last decade.

simulation 5,000 times.

### 3.5.3 Results

The results are summarized in Figure 3.5. The left panel shows the histogram of absolute return errors at the end of the 10-year period across all simulations.[12] Overall, it is clear that RI(1) is the superior method, since it has more mass in the left tail where the absolute return error is small. Also, the MSE is four times smaller compared to KRD, which comes second best. The worst performing method is HD, which has a much higher bias than KRD, but smaller variance.

The right panel of Figure 3.5 sheds light on the maxmin property by showing the 99th percentile of the absolute return error for each method throughout the 10-year period across all simulations. We see that RI(1) strictly dominates the other methods in the maxmin sense as well, consistently maintaining an absolute return error below that of the competing methods. Due to increased uncertainty, the percentiles are naturally increasing over time. Looking at the other two methods, we find that KRD compares poorly to HD because of outliers in the right tail, especially at the end of the immunization period.




**(a)** Histogram of return error. **(b)** 99th percentile of return error.

**Figure 3.5. Distribution of absolute return error**

The left panel shows the histogram of absolute return errors calculated at the end of the 10-year immunization period. The right panel shows the 99th percentile of the absolute return error throughout the 10-year immunization period, calculated across all 5,000 simulations.

[12] I.e. the absolute return error in (3.5.2) evaluated at $s = 40$.

## 3.6 Conclusion

This paper uses techniques from functional and numerical analysis to study the classical portfolio immunization problem. The goal is to construct a portfolio that protects a financial institution against interest rate risk. We use the concept of a Fréchet derivative to find a portfolio that hedges against general perturbations to the cumulative discount rate. Subsequently, we present a maxmin result that proves existence of an immunizing portfolio which maximizes the worst-case equity loss and we provide a solution algorithm. This maxmin portfolio, which we refer to as robust immunization, contains duration and convexity matching as a special case. In our empirical applications, we show that a judicious choice of basis functions for the discount rate leads to a robust immunization method that outperforms existing approaches in the static and dynamic case.

## 3.7 Acknowledgments

Chapter 3, in full, is currently being prepared for submission for publication of the material. De Vries, Tjeerd; Toda, Alexis Akira. "Robust Asset-Liability Management". The dissertation author is a primary author of this material.

# Appendix A

# Appendix to Chapter 1

## A.1 Proofs

This section contains proofs and detailed calculations of results used in the main paper.

### A.1.1 Decomposing the Equity Premium

For any atomless integrable random variable $X$ with CDF $F(\cdot)$ and quantile function $Q = F^{-1}$, we have

$$\mathbb{E}(X) = \int_{\mathbb{R}} x \, \mathrm{d}F(x) = \int_0^1 Q(\tau) \, \mathrm{d}\tau.$$

The second identity holds by the change of variables formula for the Lebesgue-Stieltjes integral. In case $F$ has a density, the formula follows from a simple substitution $x \to Q(\tau)$. Hence,

$$\mathbb{E}_t \left[ R_{m,t\to N} \right] - R_{f,t\to N} = \mathbb{E}_t \left[ R_{m,t\to N} \right] - \widetilde{\mathbb{E}}_t \left( R_{m,t\to N} \right) = \int_0^1 \left( Q_{t,\tau} - \widetilde{Q}_{t,\tau} \right) \mathrm{d}\tau.$$

### A.1.2 Stochastic Dominance and Pricing Kernel Monotonicity

In this section I provide more details on the relation between stochastic dominance and pricing kernel monotonicity. To begin with, recall that the physical distribution is first-order stochastic dominant (FOSD) over the risk-neutral distribution if and only if $F_t(x) \leq \widetilde{F}_t(x)$, or $\widetilde{Q}_{t,\tau} \leq Q_{t,\tau}$. The definition is also equivalent to $F_t(\widetilde{Q}_{t,\tau}) \leq \tau$ for all $\tau \in (0,1)$, which follows from the substitution $x \to \widetilde{Q}_{t,\tau}$.

To see the connection with pricing kernel monotonicity, recall from Beare and Schmidt (2016) that pricing kernel monotonicity is equivalent to $\phi_t(\tau) \coloneqq F_t(\widetilde{Q}_{t,\tau})$ being a convex function for

all $\tau$.[1] Figure A.1 shows two different ordinal dominance curves (ODCs); the blue line corresponds to a situation where FOSD holds and the pricing kernel is monotonic (hence convex), whereas the yellow line shows a scenario where FOSD does not hold and convexity automatically fails. The geometric argument for why non-monotonicity is implied by a failure of FOSD is conveyed by the figure: if FOSD fails, the yellow line must cross the 45-degree line for some $\tau \in (0,1)$, which automatically implies that the ODC is non-convex since the ODC has to satisfy $\phi_t(1) = 1$, because the physical and risk-neutral measures are equivalent. The proposition below thus follows.

**Proposition A.1.1.** *If the pricing kernel is a monotonically decreasing function of the market return, the physical measure first-order stochastically dominates the risk-neutral measure. Conversely, a violation of FOSD implies a violation of pricing kernel monotonicity.*

A violation of FOSD is puzzling from the viewpoint of expected utility maximization. In this framework, the SDF is given by $u'(R_{m,t\to N})/\mathbb{E}_t(u'(R_{m,t\to N}))$, where $u(\cdot)$ is a utility function and the initial endowment is normalized to one for simplicity. The following proposition shows that a sufficient (but not necessary) condition for FOSD to hold is that $u'(\cdot)$ is non-increasing; a rather ubiquitous assumption in asset pricing models.

**Proposition A.1.2.** *In the expected utility framework, a sufficient condition for the physical measure to first-order stochastically dominate the risk-neutral measure is that $u'(\cdot)$ is non-increasing.*

*Proof.* Using the SDF to change from physical to risk-neutral measure, it follows that FOSD is equivalent to

$$\begin{aligned} & F_t(x) \leq \widetilde{F}_t(x) \\ \iff & \mathbb{E}_t\left[\mathbb{1}\left(R_{m,t\to N} \leq x\right)\right] \leq \mathbb{E}_t\left[\frac{u'(R_{m,t\to N})}{\mathbb{E}_t\left[u'(R_{m,t\to N})\right]}\mathbb{1}\left(R_{m,t\to N} \leq x\right)\right] \\ \iff & 0 \leq \mathbb{COV}_{\mathrm{t}}(\mathbb{1}\left(R_{m,t\to N} \leq x\right), u'(R_{m,t\to N})). \end{aligned}$$

By Lemma A.5.1, the covariance above is nonnegative if $u'(\cdot)$ is non-increasing. ■

[1] Beare and Schmidt (2016) actually consider the reverse function $\phi_t(\tau) = \widetilde{F}_t(Q_{t,\tau})$, so that pricing kernel monotonicity is equivalent to $\phi_t(\cdot)$ being concave.

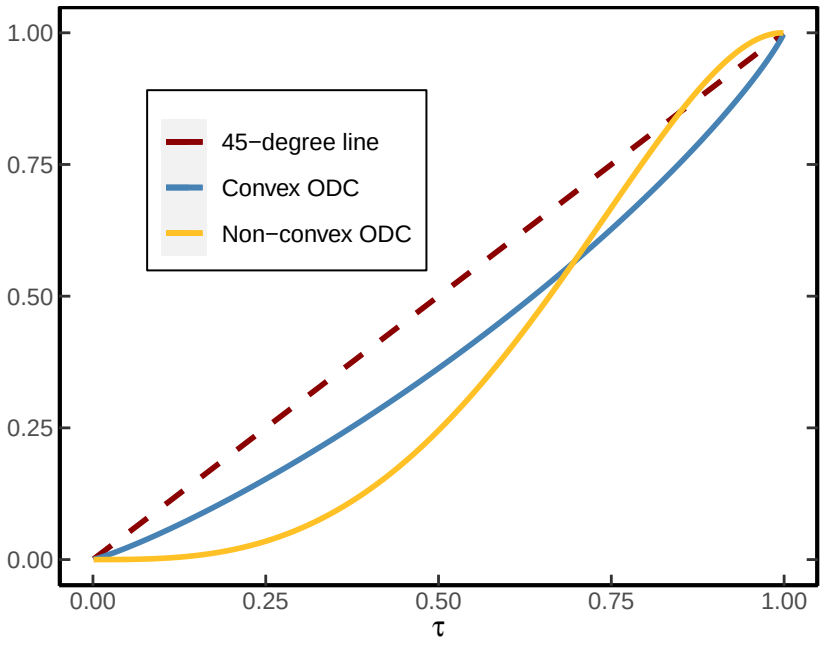



**Figure A.1. Ordinal dominance curve with and without first-order stochastic dominance.** This figure shows two different ordinal dominance curves. The blue ODC corresponds to a situation where the physical measure FOSD the risk-neutral measure, whereas the yellow line shows a situation where FOSD fails.

### A.1.3 Proof of Proposition 1.4.1

I separately show ((i)) and ((ii)) of Proposition 1.4.1. To prove these results, I use the following lemma.

**Lemma A.1.3.** *In the lognormal model, the physical and risk-neutral quantile functions conditional on $\mu_t, \sigma_t$ are given by, respectively*

$$Q_{t,\tau} = \exp\left[(\mu_t - \frac{1}{2}\sigma_t^2)N + \sigma_t\sqrt{N}\Phi^{-1}(\tau)\right] \tag{A.1.1}$$

$$\widetilde{Q}_{t,\tau} = \exp\left[(r_f - \frac{1}{2}\sigma_t^2)N + \sigma_t\sqrt{N}\Phi^{-1}(\tau)\right], \tag{A.1.2}$$

*where $\Phi^{-1}(\cdot)$ denotes the quantile function of the standard normal distribution. If $\mu_t \sim \mathcal{N}(\mu, \sigma_\mu^2)$ and independent from $\sigma_t$, the physical quantile function conditional on $\sigma_t$, but not $\mu_t$, equals*

$$Q_{t,\tau}(\sigma_t, \sigma_\mu) = \exp\left[(\mu - \frac{1}{2}\sigma_t^2)N + \left(\sqrt{\sigma_\mu^2 N^2 + \sigma_t^2 N}\right)\Phi^{-1}(\tau)\right]. \tag{A.1.3}$$

*Proof.* The quantile function of a random variable $X$ such that $\log X \sim \mathcal{N}(a, b^2)$, is given by $\exp(a+$

$b\Phi^{-1}(\tau)$). Therefore, the quantile functions conditional on $\mu_t, \sigma_t$ in (A.1.1) and (A.1.2) follow immediately from the conditional lognormal assumption. In (A.1.3), the function is conditioned on $\sigma_t$, but not $\mu_t$. Since $\mu_t$ is assumed to be normally distributed and independent from $\sigma_t$, it follows that

$$(\mu_t - \frac{1}{2}\sigma_t^2)N + \sigma_t\sqrt{N}Z_{t+N}|\sigma_t \sim \mathcal{N}\left((\mu - \frac{1}{2}\sigma_t^2)N, \sigma_\mu^2 N^2 + \sigma_t^2 N\right).$$

The expression in (A.1.3) can now be obtained again using the general formula of the lognormal quantile function. ■

*Proof of Proposition 1.4.1((i)).* Recall that $\sqrt{a^2+b^2} \leq \sqrt{a^2+b^2+2ab} = a+b$, provided $a, b \geq 0$. This inequality shows that

$$\begin{aligned}
&\exp\left[\left(\sqrt{\sigma_\mu^2 N^2 + \sigma_t^2 N} - \sigma_t\sqrt{N}\right)\Phi^{-1}(\tau)\right] \\
&\leq \exp\left[\left(\sqrt{\sigma_\mu^2 N^2 + \sigma_t^2 N} - \sigma_t\sqrt{N}\right)\left|\Phi^{-1}(\tau)\right|\right] \\
&\leq \exp\left(\sigma_\mu N\left|\Phi^{-1}(\tau)\right|\right) \\
&= 1 + \mathcal{O}\left(\sigma_\mu N\right),
\end{aligned}$$

uniformly in $\tau \in \mathcal{I}$ and the support of $\sigma_t$. In combination with Lemma A.1.3, it follows that

$$\begin{aligned}
Q_{t,\tau}(\sigma_t, \sigma_\mu) &= \widetilde{Q}_{t,\tau} e^{(\mu - r_f)N} \exp\left[\left(\sqrt{\sigma_\mu^2 N^2 + \sigma_t^2 N} - \sigma_t\sqrt{N}\right)\Phi^{-1}(\tau)\right] \\
&= \widetilde{Q}_{t,\tau} e^{(\mu - r_f)N}\left(1 + \mathcal{O}\left(\sigma_\mu N\right)\right).
\end{aligned}$$ ■

In order to prove Proposition 1.4.1((ii)), I need additional regularity conditions stated in Assumption A.1.4 below. The following notation for the quantile empirical process will be used:

$$\begin{aligned}
L_{T,\tau}(\beta, \sigma_\mu) &\coloneqq \frac{1}{T}\sum_{t=1}^{T}\rho_\tau(R_{m,t\to N} - \beta_0 - \beta_1\widetilde{Q}_{t,\tau}) \\
L_\tau(\beta, \sigma_\mu) &\coloneqq \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\rho_\tau(R_{m,t\to N} - \beta_0 - \beta_1\widetilde{Q}_{t,\tau})\right].
\end{aligned}$$

**Assumption A.1.4.** *In the lognormal model, assume additionally that*

*(i)* $\mathbb{E}\left[R_{m,t\to N}\right]$ *and* $\mathbb{E}\left[\widetilde{Q}_{t,\tau}\right]$ *are finite,*

*(ii)* $L_\tau(\beta, 0)$ *has an identifiably unique minimum* $\beta^*$ *at* $\sigma_\mu = 0$*, i.e., for all* $\varepsilon > 0$

$$\inf_{\|\beta-\beta^*\|>\varepsilon} L_\tau(\beta, 0) - L_\tau(\beta^*, 0) > 0.$$

*(iii) as* $T \to \infty$*, for any compact set* $\mathcal{B}$ *and sequence* $b_T \searrow 0$*,*

$$\sup_{\beta\in\mathcal{B}} \left\| L_\tau(\beta, \sigma_\mu^T) - L_\tau(\beta, 0) \right\| = o(1) \text{ (Uniform continuity).} \tag{A.1.4a}$$

$$\sup_{\sigma_\mu \le b_T} \sup_{\beta\in\mathcal{B}} \| L_{T,\tau}(\beta, \sigma_\mu) - L_\tau(\beta, \sigma_\mu) \| = o_p(1) \text{ (Uniform LLN).} \tag{A.1.4b}$$

*Proof of Proposition 1.4.1((ii)).* Consider the population minimization problem of quantile regression at $\sigma_\mu = 0$

$$[\beta_0^*(0;\tau), \beta_1^*(0;\tau)] \coloneqq \underset{(\beta_0,\beta_1)\in\mathbb{R}^2}{\arg\min}\ L_\tau(\beta, 0). \tag{A.1.5}$$

Assumptions A.1.4((i),(ii)) ensure that the objective function is well defined and the solution in (A.1.5) is unique for all $\tau \in \mathcal{I}$. At $\sigma_\mu = 0$, $Q_{t,\tau} = e^{(\mu-r)N}$, so that $[\beta_0^*(0;\tau), \beta_1^*(0;\tau)] = [0, e^{(\mu-r)N}]$. To ease notation in the following derivation, I write $\widehat{\beta}(\sigma_\mu^T) \coloneqq \arg\min_\beta L_{T,\tau}(\beta, \sigma_\mu^T)$ and $\beta^*(0) = \arg\min_\beta L_\tau(\beta, 0)$. It then follows that for every $\varepsilon > 0$ there exists a $\delta > 0$ such that

$$\begin{aligned}
&\mathrm{P}\left( \left\| \widehat{\beta}(\sigma_\mu^T) - \beta^*(0) \right\| > \varepsilon \right) \\
&\le \mathrm{P}\left( L_\tau(\widehat{\beta}(\sigma_\mu^T), 0) - L_\tau(\beta^*(0), 0) > \delta \right) \\
&= \mathrm{P}\left( L_\tau(\widehat{\beta}(\sigma_\mu^T), 0) - L_{T,\tau}(\widehat{\beta}(\sigma_\mu^T), \sigma_\mu^T) + L_{T,\tau}(\widehat{\beta}(\sigma_\mu^T), \sigma_\mu^T) - L_\tau(\beta^*(0), 0) > \delta \right) \\
&\le \mathrm{P}\left( L_\tau(\widehat{\beta}(\sigma_\mu^T), 0) - L_{T,\tau}(\widehat{\beta}(\sigma_\mu^T), \sigma_\mu^T) + L_{T,\tau}(\beta^*(0), \sigma_\mu^T) - L_\tau(\beta^*(0), 0) > \delta \right) \\
&\le \mathrm{P}\left( 2 \sup_{\beta\in\mathcal{B}} \left\| L_\tau(\beta, 0) - L_{T,\tau}(\beta, \sigma_\mu^T) \right\| > \delta \right).
\end{aligned}$$

The second line follows from identification and the second to last line from the minimization property of $\widehat{\beta}(\sigma_\mu^T)$. Therefore, it suffices to show that

$$\sup_{\beta\in\mathcal{B}} \left\| L_\tau(\beta, 0) - L_{T,\tau}(\beta, \sigma_\mu^T) \right\| = o_p(1).$$

This claim follows from

$$\begin{aligned}
&\sup_{\beta\in\mathcal{B}} \left\| L_\tau(\beta,0) - L_{T,\tau}(\beta,\sigma_\mu^T) \right\| \\
\leq &\sup_{\beta\in\mathcal{B}} \left\| L_\tau(\beta,0) - L_\tau(\beta,\sigma_\mu^T) \right\| + \left\| L_\tau(\beta,\sigma_\mu^T) - L_{T,\tau}(\beta,\sigma_\mu^T) \right\| \\
\leq &\sup_{\beta\in\mathcal{B}} \left\| L_\tau(\beta,0) - L_\tau(\beta,\sigma_\mu^T) \right\| + \sup_{\sigma_\mu\leq b_T} \sup_{\beta\in\mathcal{B}} \left\| L_\tau(\beta,\sigma_\mu) - L_{T,\tau}(\beta,\sigma_\mu) \right\|.
\end{aligned}$$

The first term is $o(1)$ by (A.1.4a) and the second term is $o_p(1)$ by (A.1.4b), which completes the proof. The claim in (1.4.5) easily follows from (1.4.4).

■

### A.1.4 Proof of Proposition 1.5.1

*Proof.* Starting from the definition of the risk-neutral quantile function, it follows that

$$\begin{aligned}
\tau &= \widetilde{\mathbb{P}}_t \left[ R_{m,t\to N} \leq \widetilde{Q}_{t,\tau} \right] \\
&= \widetilde{\mathbb{E}}_t \left[ \mathbb{1}\left( R_{m,t\to N} \leq \widetilde{Q}_{t,\tau} \right) \right] \\
&= \frac{1}{\mathbb{E}_t\left[M_{t\to N}\right]} \mathbb{E}_t \left[ M_{t\to N} \mathbb{1}\left( R_{m,t\to N} \leq \widetilde{Q}_{t,\tau} \right) \right] \\
&= \frac{1}{\mathbb{E}_t\left[M_{t\to N}\right]} \left( \mathbb{COV}_{\mathrm{t}} \left( M_{t\to N}, \mathbb{1}\left( R_{m,t\to N} \leq \widetilde{Q}_{t,\tau} \right) \right) + \mathbb{E}_t\left[M_{t\to N}\right] \mathbb{E}_t \left[ \mathbb{1}\left( R_{m,t\to N} \leq \widetilde{Q}_{t,\tau} \right) \right] \right) \\
&= \frac{1}{\mathbb{E}_t\left[M_{t\to N}\right]} \mathbb{COV}_{\mathrm{t}} \left( M_{t\to N}, \mathbb{1}\left( R_{m,t\to N} \leq \widetilde{Q}_{t,\tau} \right) \right) + \underbrace{\mathbb{E}_t \left[ \mathbb{1}\left( R_{m,t\to N} \leq \widetilde{Q}_{t,\tau} \right) \right]}_{=\phi_t(\tau)}.
\end{aligned} \tag{A.1.6}$$

Rearranging then yields

$$\frac{1}{\mathbb{E}_t\left[M_{t\to N}\right]} \mathbb{COV}_{\mathrm{t}} \left( M_{t\to N}, \mathbb{1}\left( R_{m,t\to N} \leq \widetilde{Q}_{t,\tau} \right) \right) = \tau - \phi_t(\tau).$$

Using Cauchy-Schwarz renders the inequality

$$\begin{aligned}
\frac{1}{\mathbb{E}_t\left[M_{t\to N}\right]} \sigma_t(M_{t\to N}) \sigma_t \left( \mathbb{1}\left( R_{m,t\to N} \leq \widetilde{Q}_{t,\tau} \right) \right) &\geq |\tau - \phi_t(\tau)| \\
\frac{\sigma_t(M_{t\to N})}{\mathbb{E}_t\left[M_{t\to N}\right]} &\geq \frac{|\tau - \phi_t(\tau)|}{\sigma_t\left( \mathbb{1}\left( R_{m,t\to N} \leq \widetilde{Q}_{t,\tau} \right) \right)}.
\end{aligned} \tag{A.1.7}$$

Finally, since $\mathbb{1}\left(R_{m,t\to N} \leq \widetilde{Q}_{t,\tau}\right)$ is a Bernoulli random variable, it follows that

$$\sigma_t\left(\mathbb{1}\left(R_{m,t\to N} \leq \widetilde{Q}_{t,\tau}\right)\right) = \sqrt{\phi_t(\tau)(1-\phi_t(\tau))}. \tag{A.1.8}$$

Proposition 1.5.1 now follows after substituting (A.1.8) into (A.1.7). The bound formulated in terms of the CDFs in (1.5.2) follows from the substitution $\widetilde{Q}_{t,\tau} \to x$. ■

### A.1.5 Distribution Bound when SDF and Return are Jointly Normal

In this Section I derive (1.5.7) and (1.5.8) , when $M$ and $R_m$ are jointly normal. First consider (1.5.8). The proof of the distribution bound in Proposition 1.5.1 gives the following identity

$$\frac{|\tau - \phi(\tau)|}{R_f} = \left|\mathbb{COV}\left(\mathbb{1}\left(R_m \leq \widetilde{Q}_\tau\right), M\right)\right|.$$

Standard SDF properties also yield the well known result

$$\frac{|\mathbb{E}(R_m) - R_f|}{R_f} = |\mathbb{COV}(R_m, M)|.$$

These results, combined with (1.5.7) prove (1.5.8), since

$$\begin{aligned}\frac{\text{HJ bound}}{\text{distribution bound}} &= \frac{\frac{\left|\mathbb{E}[R_m]-R_f\right|}{\sigma_R R_f}}{\frac{|\tau-\phi(\tau)|}{\sqrt{\phi(\tau)(1-\phi(\tau))}R_f}} \\ &\overset{(1.5.7)}{=} \frac{\sqrt{\phi(\tau)(1-\phi(\tau))}}{\sigma_R f_R(\widetilde{Q}_\tau)},\end{aligned}$$

where $f_R(\widetilde{Q}_\tau)$ is the marginal density of $R_m$.

Finally, I make use of the following covariance identities to prove (1.5.7).

**Lemma A.1.5** (Hoeffding)**.** *For any square integrable random variable $X$ and $Z$ with marginal CDFs $F_X, F_Z$ and joint CDF $F_{X,Z}$, it holds that*

$$\mathbb{COV}\left[\mathbb{1}\left(Z \leq z\right), X\right] = -\int_{-\infty}^{\infty}\left[F_{X,Z}(x,z) - F_X(x)F_Z(z)\right]\mathrm{d}x \tag{A.1.9}$$

$$\mathbb{COV}\left[Z, X\right] = -\int_{-\infty}^{\infty}\mathbb{COV}\left[\mathbb{1}\left(Z \leq z\right), X\right]\mathrm{d}z. \tag{A.1.10}$$

*Proof.* See Lehmann (1966). ■

I also need a relation for the bivariate normal distribution. Suppose that $X, Z$ are jointly normal with correlation $\rho$, mean $\mu_X, \mu_Z$ and variance $\sigma_X^2, \sigma_Z^2$, then

$$\frac{\partial \Phi_2(x, z; \rho, \mu_X, \mu_Z, \sigma_X^2, \sigma_Z^2)}{\partial \rho} = \sigma_X \sigma_Z \phi_2(x, z; \rho, \mu_X, \mu_Z, \sigma_X^2, \sigma_Z^2), \tag{A.1.11}$$

where $\Phi_2(\cdot)$ denotes the bivariate normal CDF and $\phi_2(\cdot)$ denotes the bivariate normal PDF (Sungur, 1990). We can now prove a covariance identity for jointly normal random variables.

**Proposition A.1.6.** *Suppose $R_m$ and $M$ are jointly normal with correlation $\rho$, then*

$$-\mathbb{COV}\left[\mathbb{1}\left(R_m \leq x\right), M\right] = \mathbb{COV}\left[R_m, M\right] \phi_R(x), \tag{A.1.12}$$

*where $\phi_R(\cdot)$ is the marginal density of $R_m$.*

*Proof.* To lighten notation, I suppress the dependence on $\mu_R, \mu_M, \sigma_R^2, \sigma_M^2$ in the joint CDF and PDF. We then have

$$\begin{aligned}
-\mathbb{COV}\left[\mathbb{1}\left(R_m \leq x\right), M\right] &= \int_{-\infty}^{\infty} \Phi_2(x, m; \rho) - \Phi_2(x, m; 0) \, \mathrm{d}m \\
&= \int_{-\infty}^{\infty} \int_0^{\rho} \sigma_R \sigma_M \phi_2(x, m; y) \, \mathrm{d}y \, \mathrm{d}m \\
&= \sigma_R \sigma_M \rho \phi_R(x) \\
&= \mathbb{COV}\left[R_m, M\right] \phi_R(x),
\end{aligned}$$

where, in the first line, I use (A.1.9) together with $F_R(r) F_M(m) = \Phi_2(r, m; 0)$, the second line follows from (A.1.11) and the third line follows from Fubini's theorem to swap the order of integration and $\int_{-\infty}^{\infty} \phi_2(x, m; y) \, \mathrm{d}m = \phi_R(x)$. ■

*Remark* 6. The second covariance identity in (A.1.10) shows that $\mathbb{COV}\left[\mathbb{1}\left(R_m \leq x\right), M\right]$ is a measure of local dependence. In case of joint normality (A.1.12), the weight is given by the marginal PDF. For other distributions, the weighting factor is more complicated, but sometimes can be given an explicit form using a local Gaussian representation (see Chernozhukov, Fernández-Val, and Luo (2018)).

### A.1.6 Minimizer of Distribution Bound with Normal SDF

This section shows that the relative efficiency between the HJ bound and distribution bound is minimized when $\widetilde{Q}_\tau = \mu_R$. To see this, write $x = \widetilde{Q}_\tau$, and use $F(\cdot)$ to denote the physical CDF of $R_m$. I also drop the $R$ subscript for $f$ to avoid notational clutter. Consider

$$\Gamma(x) = \frac{F(x)(1-F(x))}{f(x)^2}.$$

Minimizing $\Gamma(x)$ is equivalent to minimizing (1.5.8) and first order conditions imply that the optimal $x^*$ satisfies

$$[f(x^*) - 2F(x^*)f(x^*)]f(x^*)^2 - 2f(x^*)f'(x^*)[F(x^*)(1-F(x^*))] = 0. \tag{A.1.13}$$

Since $f, F$ are the respective PDF and CDF of the normal random variable $R_m$, it follows that $f'(\mu_R) = 0$ and $F(\mu_R) = 1/2$. As a result, (A.1.13) holds when $\widetilde{Q}_{\tau^*} = x^* = \mu_R$.

### A.1.7 Distribution Bound when SDF and Return are Log-normal

This section provides a closed form approximation for the relative efficiency between the HJ bound and distribution bound under joint lognormality. The result depends on Stein's Lemma (Casella and Berger, 2002, Lemma 3.6.5):[2]

**Lemma A.1.7** (Stein's Lemma)**.** *If $X_1, X_2$ are bivariate normal, $g : \mathbb{R} \to \mathbb{R}$ is differentiable and $\mathbb{E}\,|g'(X_1)| < \infty$, then*

$$\mathbb{COV}\left(g(X_1), X_2\right) = \mathbb{E}\left[g'(X_1)\right]\mathbb{COV}(X_1, X_2).$$

To prove the approximation, we approximate $M$ by a first order Taylor expansion, which gives

$$\widehat{M} = e^{-(r_f + \frac{\sigma_M^2}{2})N} + Z_M \sigma_M \sqrt{N} e^{-(r_f + \frac{\sigma_M^2}{2})N}.$$

[2] I use the form of Stein's Lemma reported in Cochrane (2005, p. 163), which follows from Stein's lemma as reported in Casella and Berger (2002).

Notice that $\widehat{M} = M + o_p(\sqrt{N})$. Consequently, by Stein's Lemma

$$\begin{aligned}
&\mathbb{COV}(R_m, M) \approx \mathbb{COV}(R_m, \widehat{M}) = \sigma_M \sqrt{N} e^{-(r_f + \frac{\sigma_M^2}{2})N} \mathbb{COV}(R_m, Z_M) \\
&= \sigma_M \sqrt{N} e^{-(r_f + \frac{\sigma_M^2}{2})N} \mathbb{E}\left[\sigma_R \sqrt{N} \exp\left(\left[\mu_R - \frac{\sigma_R^2}{2}\right] N + \sigma_R \sqrt{N} Z_R\right)\right] \mathbb{COV}(Z_R, Z_M) \\
&= \sigma_M \sigma_R N e^{-(r_f + \frac{\sigma_M^2}{2})N} e^{\mu_R N} \mathbb{COV}(Z_R, Z_M).
\end{aligned}$$

By Proposition A.1.6,

$$\begin{aligned}
&\mathbb{COV}(\mathbb{1}\left(\log R_m \leq x\right), M) \approx \mathbb{COV}\left(\mathbb{1}\left(\log R_m \leq x\right), \widehat{M}\right) \\
&= \sigma_M \sqrt{N} e^{-(r_f + \frac{\sigma_M^2}{2})N} \mathbb{COV}\left(\mathbb{1}\left(\log R_m \leq x\right), Z_M\right) \\
&= \sigma_M \sqrt{N} e^{-(r_f + \frac{\sigma_M^2}{2})N} \mathbb{COV}\left(\mathbb{1}\left((\mu_R - \sigma_R^2/2)N + \sigma_R \sqrt{N} Z_R \leq x\right), Z_M\right) \\
&= -\sigma_M \sqrt{N} e^{-(r_f + \frac{\sigma_M^2}{2})N} f\left(x\right) \mathbb{COV}\left(Z_R, Z_M\right).
\end{aligned}$$

Here, $f$ is the density of a normal random variable with mean $(\mu_R - \sigma_R^2/2)N$ and variance $N\sigma_R^2$. As a result,

$$\left| \frac{\mathbb{E}\left[R_m\right] - e^{N r_f}}{\tau - \phi(\tau)} \right| \approx \frac{\sigma_R \sqrt{N} e^{\mu_R N}}{f(x)}. \tag{A.1.14}$$

The same reasoning in Example 1.5.2 implies that the relative efficiency between the HJ and distribution bound can be approximated by

$$\frac{\text{HJ bound}}{\text{distribution bound}} = \frac{\frac{\left|\mathbb{E}[R_m] - R_f\right|}{\sigma(R_m) R_f}}{\frac{|\tau - \phi(\tau)|}{\sqrt{\phi(\tau)(1-\phi(\tau))} R_f}} \tag{A.1.15}$$

$$\overset{\text{(A.1.14)}}{\approx} \frac{\sqrt{\mathbb{P}(r \leq x) \cdot (1 - \mathbb{P}(r \leq x))}}{\sigma(R_m)} \times \frac{\sigma_R \sqrt{N} e^{\mu_R N}}{f(x)}, \tag{A.1.16}$$

where $r = \log R$ and $x = \log \widetilde{Q}_\tau$. Using the same reasoning as in Example 1.5.2, the expression on the right hand side of (A.1.15) is minimized by choosing $x = \log \widetilde{Q}_\tau^*$ s.t. $\mathbb{P}(R_m \leq \widetilde{Q}_\tau^*) = 1/2$. In that case the relative efficiency equals

$$\frac{\sqrt{2\pi\sigma_R^2} \sqrt{N} e^{\mu_R N}}{2\sqrt{[\exp(\sigma_R^2 N) - 1] \exp(2\mu_R N)}} = \frac{1}{2} \sqrt{\frac{2\pi\sigma_R^2 N}{\exp(\sigma_R^2 N) - 1}}.$$

### A.1.8 Distribution Bound with Pareto Distribution

This section derives an explicit expression of the distribution bound when the return and SDF follow the Pareto distribution.

**Example A.1.1** (Pareto distribution)**.** Let $U \sim \mathbf{Unif}[0,1]$ (Uniform distribution on [0,1]) and consider the following specification:

$$M = AU^{\alpha},\ R_m = BU^{-\beta} \qquad \text{with} \qquad \alpha, \beta, A, B > 0. \tag{A.1.17}$$

A random variable $X \sim \mathbf{Par}(C, \zeta)$ follows a Pareto distribution with scale parameter $C > 0$ and shape parameter $\zeta > 0$ if the CDF is given by

$$\mathbb{P}(X \leq x) = \begin{cases} 1 - (x/C)^{-\zeta} & x \geq C \\ 0 & x < C. \end{cases}$$

The assumption (A.1.17) implies that returns follow a Pareto distribution, both under the physical and risk-neutral measures. This fact allows me to obtain an explicit expression for the distribution bound. I summarize these properties in the Proposition below.

**Proposition A.1.8.** *Let the SDF and return be given by* (A.1.17)*. Then,*

*(i) Under $\mathbb{P}$, the distribution of returns is Pareto: $R_m \sim \mathbf{Par}\left(B, \frac{1}{\beta}\right)$.*

*(ii) Under $\widetilde{\mathbb{P}}$, the distribution of returns is Pareto: $R_m \sim \mathbf{Par}\left(B, \frac{\alpha+1}{\beta}\right)$.*

*(iii) The Sharpe ratio on the asset return is given by*

$$\frac{\mathbb{E}[R_m] - R_f}{\sigma(R_m)} = \frac{\frac{B}{1-\beta} - \frac{\alpha+1}{A}}{\sqrt{\frac{B^2}{1-2\beta} - \left(\frac{B}{1-\beta}\right)^2}}. \tag{A.1.18}$$

*(iv) The distribution bound is given by*

$$\frac{1}{R_f} \frac{|\tau - \phi(\tau)|}{\sqrt{\phi(\tau)(1-\phi(\tau))}} = \frac{A}{1+\alpha} \frac{\left|\tau - 1 + (1-\tau)^{\frac{1}{\alpha+1}}\right|}{\sqrt{(1-(1-\tau)^{\frac{1}{\alpha+1}})(1-\tau)^{\frac{1}{\alpha+1}}}}.$$

*(v) If $\beta \nearrow \frac{1}{2}$, the HJ bound converges to* 0.

*Proof.* See the end of this section. ■

Proposition A.1.8((iv)) shows that the distribution bound is independent of the Pareto tail index $\beta$. Properties ((iv)) and ((v)) provide some intuition when the distribution bound is stronger than the HJ bound. Namely, heavier tails of the distribution of $R_m$ (as measured by $\beta$) lead to a lower Sharpe ratio. However, the distribution bound is unaffected by $\beta$ since it only depends on the tail index $\alpha$. Therefore, when $\beta$ gets close to $1/2$, the HJ bound is rather uninformative whereas the distribution bound may fare better. Moreover, no additional restrictions on the parameter space are necessary to calculate the distribution bound, while the HJ bound requires $\beta < 1/2$.[3]

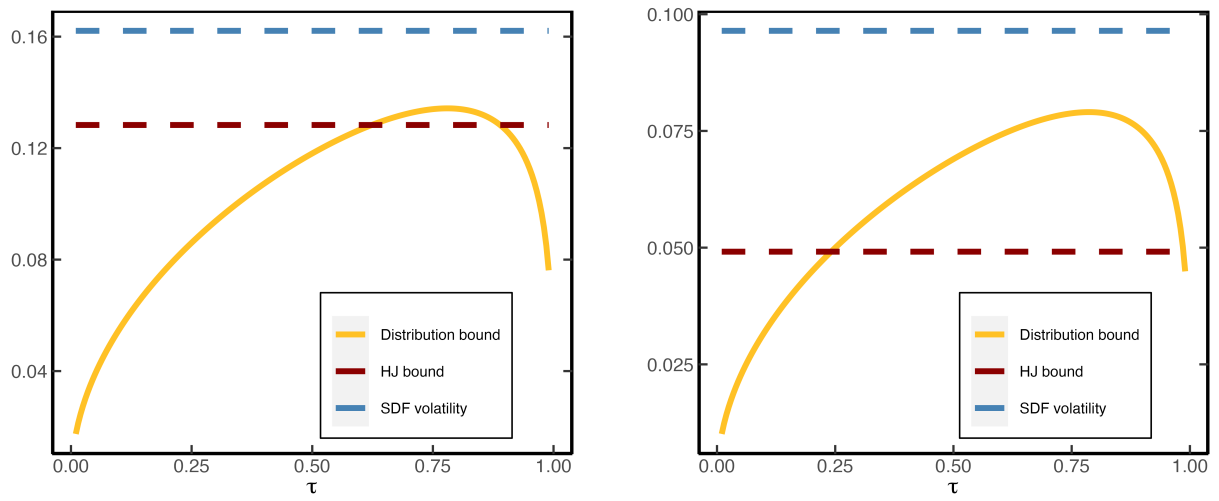
Figure A.2 shows two instances of the distribution and HJ bound using different parameter calibrations. Both calibrations are targeted to match an equity premium of 8% and risk-free rate of 0%, but in Panel A.2b, the distribution of returns has a fatter tail compared to Panel A.2a. In both calibrations, the distribution bound has a range of values for which it is stronger than the HJ bound. In line with Proposition A.1.8, we see that the range is larger in Panel (b), since the HJ bound is less informative owing to the heavier tails of $R_m$. However, the distribution bound attains its maximum in the right-tail since that is the region where the physical and risk-neutral measure differ most. This result is inconsistent with the empirical results from Table 1.2, which indicate that the physical and risk-neutral measure are nearly identical in the right-tail.

*Proof of Proposition A.1.8.* (i) The distribution of returns is Pareto, since

$$\begin{aligned}\mathbb{P}(R_m \leq x) &= \mathbb{P}\left(U^{-\beta} \leq x/B\right) \\ &= \mathbb{P}\left(U \geq (x/B)^{-\frac{1}{\beta}}\right) = 1 - \left(\frac{x}{B}\right)^{-\frac{1}{\beta}}, \qquad x \geq B.\end{aligned}$$

(ii) Since $R_f M$ is the Radon-Nikodym derivative that induces a change of measure from $\mathbb{P}$ to $\widetilde{\mathbb{P}}$,

[3] The latter restriction is not unreasonable for asset returns, since typical tail index estimates suggest $\beta \in [1/4, 1/3]$ (Danielsson and de Vries, 2000).




**(a)** $\beta = 0.33$ **(b)** $\beta = 0.45$

**Figure A.2. HJ and distribution bound for heavy tailed returns.** Both panels plot the distribution bound, HJ bound and true SDF volatility for the Pareto model (A.1.17). In Panel (b), the distribution of returns has a fatter tail compared to Panel (a). Panel (a) uses the parameters $[A, \alpha, B, \beta] = [1.19, 0.19, 0.72, 0.33]$. Panel (b) uses the parameters $[A, \alpha, B, \beta] = [1.11, 0.11, 0.59, 0.45]$. Both calibrations imply an equity premium of 8% and (net) risk-free rate of 0%.

it follows that

$$\begin{aligned}
\widetilde{\mathbb{P}}(R_m \leq x) &= R_f \mathbb{E}\left[M \mathbb{1}\left(R_m \leq x\right)\right] \\
&= R_f \int_0^1 A u^\alpha \mathbb{1}\left(B u^{-\beta} \leq x\right) \mathrm{d}u \\
&= R_f A \int_0^1 u^\alpha \mathbb{1}\left(u \geq \left(\frac{x}{B}\right)^{-\frac{1}{\beta}}\right) \mathrm{d}u \\
&= \frac{R_f A}{\alpha + 1}\left(1 - \left(\frac{x}{B}\right)^{-\frac{\alpha+1}{\beta}}\right) \\
&= 1 - \left(\frac{x}{B}\right)^{-\frac{\alpha+1}{\beta}}.
\end{aligned}$$

The last line follows from (A.1.21) below.

(iii) Routine calculations show that the mean and variance of $R_m$ are given by (provided $\beta < 1/2$)

$$\mathbb{E}\left[R_m\right] = \frac{B}{1-\beta} \qquad \sigma^2(R_m) = \frac{B^2}{1-2\beta} - \left(\frac{B}{1-\beta}\right)^2. \tag{A.1.19}$$

Likewise, the distribution of the SDF follows from

$$\mathbb{P}\left(M \leq x\right) = \mathbb{P}\left(AU^{\alpha} \leq x\right) = \left(\frac{x}{A}\right)^{\frac{1}{\alpha}}, \quad 0 \leq x \leq A.$$

In this case, $M$ is said to have a Pareto lower tail. The expectation is given by

$$\mathbb{E}\left[M\right] = \frac{A}{\alpha + 1}.$$

The constraint $\mathbb{E}\left[MR_m\right] = 1$ forces

$$\frac{AB}{\alpha - \beta + 1} = 1. \tag{A.1.20}$$

In addition from $\mathbb{E}\left[M\right] = \frac{1}{R_f}$ it follows that

$$\frac{A}{\alpha + 1} = \frac{1}{R_f}. \tag{A.1.21}$$

The Sharpe ratio can now be computed from (A.1.19) and (A.1.21).

(iv) It is straightforward to show that the quantiles of a $\mathbf{Par}\left(C, \zeta\right)$ distribution are given by

$$Q_\tau = C(1-\tau)^{-1/\zeta}.$$

It therefore follows that the risk-neutral quantile function is equal to

$$\widetilde{Q}_\tau = B(1-\tau)^{-\frac{\beta}{\alpha+1}}.$$

As a result

$$\begin{aligned}
\mathbb{P}(R_m \leq \widetilde{Q}_\tau) &= \mathbb{P}\left(R_m \leq B(1-\tau)^{-\frac{\beta}{\alpha+1}}\right) \\
&= 1 - \left(\frac{B}{B(1-\tau)^{\frac{-\beta}{\alpha+1}}}\right)^{\frac{1}{\beta}} \\
&= 1 - (1-\tau)^{\frac{1}{\alpha+1}}.
\end{aligned}$$

Hence, the distribution bound evaluates to

$$\frac{1}{R_f}\frac{|\tau-\phi(\tau)|}{\sqrt{\phi(\tau)(1-\phi(\tau))}}=\frac{A}{1+\alpha}\frac{\left|\tau-1+(1-\tau)^{\frac{1}{\alpha+1}}\right|}{\sqrt{(1-(1-\tau)^{\frac{1}{\alpha+1}})(1-\tau)^{\frac{1}{\alpha+1}}}}.$$

(v) The HJ bound, as given by the Sharpe ratio in (A.1.18), goes to 0 as $\beta \nearrow 1/2$ since $\sigma(R_m) \nearrow \infty$. ■

### A.1.9 Derivation of Gâteaux Derivative

In this Section I derive (1.6.3). For ease of exposition, I drop the time subscripts. For $\lambda \in [0,1]$, define $\widetilde{F}_\lambda \coloneqq (1-\lambda)\widetilde{F}+\lambda F$. The following (trivial) identity will prove helpful[4]

$$\tau = \widetilde{F}_\lambda \widetilde{F}_\lambda^{-1}. \tag{A.1.22}$$

To further simplify notation, write $q(\lambda) \coloneqq \widetilde{F}_\lambda^{-1}$. Then (A.1.22) becomes

$$\tau = (1-\lambda)\widetilde{F}(q(\lambda))+\lambda F(q(\lambda)).$$

Applying the implicit function theorem, we obtain

$$q'(\lambda) = -\frac{-\widetilde{F}(q(\lambda))+F(q(\lambda))}{(1-\lambda)\widetilde{f}(q(\lambda))+\lambda f(q(\lambda))}.$$

Plug in $\lambda = 0$ to get

$$q'(0) = -\frac{-\widetilde{F}(q(0))+F(q(0))}{\widetilde{f}(q(0))}. \tag{A.1.23}$$

Notice that

$$\widetilde{F}_\lambda\big|_{\lambda=0} = \widetilde{F} \implies q(\lambda)\big|_{\lambda=0} = q(0) = \widetilde{F}^{-1}. \tag{A.1.24}$$

Substitute (A.1.24) into (A.1.23) to obtain

$$q'(0) = -\frac{-\widetilde{F}(\widetilde{F}^{-1})+F(\widetilde{F}^{-1})}{\widetilde{f}(\widetilde{F}^{-1})} = \frac{\tau-F(\widetilde{F}^{-1})}{\widetilde{f}(\widetilde{F}^{-1})}. \tag{A.1.25}$$

[4]This "equality" may actually only be an inequality for some $\tau$, but this is immaterial to the argument.

Notice that $q'(0)$ is exactly equal to the Gâteaux derivative from the definition in (1.6.2), since

$$\frac{\partial}{\partial\lambda}\varphi\left[(1-\lambda)\widetilde{F}+\lambda F\right]\Big|_{\lambda=0}=\frac{\partial}{\partial\lambda}q(\lambda)\Big|_{\lambda=0}=q'(0).$$

### A.1.10 Proof of Proposition 1.6.4

In the proofs that follow, I repeatedly use Taylor's theorem with integral remainder, which is stated for completeness.

**Lemma A.1.9** (Taylor's theorem)**.** *Let $\zeta^{(3)}(\cdot)$ be absolutely continuous on the closed interval between $a$ and $x$, then*

$$\zeta(x)=\sum_{k=0}^{3}\frac{\zeta^{(k)}(a)}{k!}(x-a)^k+\int_a^x\frac{\zeta^{(4)}(t)}{3!}(x-t)^3\,\mathrm{d}t.$$

The proof of Proposition 1.6.4 proceeds in several stages, by first proving an infeasible lower bound on $\tau-F_t(\widetilde{Q}_{t,\tau})$, which is later refined into a feasible lower bound under additional assumptions. Before doing so, I collect several results about the SDF in representative agent models.

**Lemma A.1.10.** *Assume a representative agent model with SDF given by* (1.6.5)*, then*

$$\tau-F_t(\widetilde{Q}_{t,\tau})=-\frac{\widetilde{\mathbb{COV}}_{\mathrm{t}}\left[\mathbb{1}\left(R_{m,t\to N}\leq\widetilde{Q}_{t,\tau}\right),\zeta(R_{m,t\to N})\right]}{\widetilde{\mathbb{E}}_t\left[\zeta(R_{m,t\to N})\right]},\tag{A.1.26}$$

*where $\zeta(\cdot)$ is defined in* (1.6.6)*.*

*Proof.* Use the reciprocal of the SDF to pass from physical to risk-neutral measure

$$\begin{aligned}F_t(\widetilde{Q}_{t,\tau})&=\mathbb{E}_t\left[\mathbb{1}\left(R_{m,t\to N}\leq\widetilde{Q}_{t,\tau}\right)\right]=\widetilde{\mathbb{E}}_t\left[\mathbb{1}\left(R_{m,t\to N}\leq\widetilde{Q}_{t,\tau}\right)\frac{\mathbb{E}_t\left[M_{t\to N}\right]}{M_{t\to N}}\right]\\&=\widetilde{\mathbb{COV}}_{\mathrm{t}}\left[\mathbb{1}\left(R_{m,t\to N}\leq\widetilde{Q}_{t,\tau}\right),\frac{\mathbb{E}_t\left[M_{t\to N}\right]}{M_{t\to N}}\right]+\tau.\end{aligned}\tag{A.1.27}$$

Rearranging the above and using the definition of $\zeta(\cdot)$ in (1.6.6), as well as (1.6.5), we obtain (A.1.26). ■

**Lemma A.1.11.** *Under Assumption 1.6.2,*

$$\widetilde{\mathbb{E}}_t\left[\zeta(R_{m,t\to N})\right] \leq \sum_{k=0}^{3} \theta_k \widetilde{\mathbb{M}}_{t\to N}^{(k)} = 1 + \sum_{k=1}^{3} \theta_k \widetilde{\mathbb{M}}_{t\to N}^{(k)},$$

*where $\zeta(x)$ is the IMRS defined in* (1.6.6).

*Proof.* In the integral of Lemma A.1.9, substitute $s = (t-a)/(x-a)$ to get

$$\begin{aligned}\zeta(x) &= \sum_{k=0}^{3} \frac{\zeta^{(k)}(a)}{k!}(x-a)^k + (x-a)^4 \int_0^1 \frac{\zeta^{(4)}(a+s(x-a))}{3!}(1-s)^3\,\mathrm{d}s \\ &\leq \sum_{k=0}^{3} \frac{\zeta^{(k)}(a)}{k!}(x-a)^k,\end{aligned}$$

since $\zeta^{(4)}(x) < 0$ by Assumption 1.6.2(ii). Using this result with $a = R_{f,t\to N}$ and taking expectations, we obtain

$$\widetilde{\mathbb{E}}_t\left[\zeta(R_{m,t\to N})\right] \leq \sum_{k=0}^{3} \theta_k \widetilde{\mathbb{M}}_{t\to N}^{(k)}. \qquad \blacksquare$$

Under Assumption 1.6.2, the difference between the physical and risk-neutral distribution in the left-tail can be bounded as follows.

**Theorem A.1.12** (Infeasible Lower Bound)**.** *Let Assumption 1.6.2 hold and assume that the risk-neutral CDF is absolutely continuous with respect to Lebesgue measure. Define $\tau^*$ so that $G(\widetilde{Q}_{t,\tau^*}) = \widetilde{\mathbb{E}}_t\left(G(R_{m,t\to N})\right)$, where*

$$G(R_{m,t\to N}) \coloneqq \int_{R_{f,t\to N}}^{R_{m,t\to N}} \zeta^{(4)}(t)(R_{m,t\to N} - t)^3\,\mathrm{d}t.$$

*Then for all $\tau \leq \tau^*$,*

$$\tau - F_t\left(\widetilde{Q}_{t,\tau}\right) \geq \frac{\sum_{k=1}^{3} \theta_k \left(\tau \widetilde{\mathbb{M}}_{t\to N}^{(k)} - \widetilde{\mathbb{M}}_{t\to N}^{(k)}[\widetilde{Q}_{t,\tau}]\right)}{1 + \sum_{k=1}^{3} \theta_k \widetilde{\mathbb{M}}_{t\to N}^{(k)}}, \tag{A.1.28}$$

*where $\widetilde{\mathbb{M}}_{t\to N}^{(k)}, \widetilde{\mathbb{M}}_{t\to N}^{(k)}[\widetilde{Q}_{t,\tau}]$ are defined in* (1.6.7).

*Proof of Theorem A.1.12.* By Taylor's theorem,

$$
\begin{aligned}
&-\widetilde{\mathbb{COV}}_{\mathrm{t}}\left[\mathbb{1}\left(R_{m,t\to N}\leq\widetilde{Q}_{t,\tau}\right),\zeta(R_{m,t\to N})\right]=\sum_{k=1}^{3}\theta_k\left(\tau\widetilde{\mathbb{M}}_{t\to N}^{(k)}-\widetilde{\mathbb{M}}_{t\to N}^{(k)}[\widetilde{Q}_{t,\tau}]\right)\\
&-\widetilde{\mathbb{COV}}_{\mathrm{t}}\left[\mathbb{1}\left(R_{m,t\to N}\leq\widetilde{Q}_{t,\tau}\right),\frac{1}{3!}\int_{R_{f,t\to N}}^{R_{m,t\to N}}\zeta^{(4)}(t)(R_{m,t\to N}-t)^3\,\mathrm{d}t\right] \qquad \text{(A.1.29)}\\
&\geq\sum_{k=1}^{3}\theta_k\left(\tau\widetilde{\mathbb{M}}_{t\to N}^{(k)}-\widetilde{\mathbb{M}}_{t\to N}^{(k)}[\widetilde{Q}_{t,\tau}]\right).
\end{aligned}
$$

The last line follows from Lemma A.1.13 below. Hence,

$$
\begin{aligned}
\tau-F_t(\widetilde{Q}_{t,\tau})&=-\frac{\widetilde{\mathbb{COV}}_{\mathrm{t}}\left[\mathbb{1}\left(R_{m,t\to N}\leq\widetilde{Q}_{t,\tau}\right),\zeta(R_{m,t\to N})\right]}{\widetilde{\mathbb{E}}_t\left[\zeta(R_{m,t\to N})\right]}\\
&\geq\frac{\sum_{k=1}^{3}\theta_k\left(\tau\widetilde{\mathbb{M}}_{t\to N}^{(k)}-\widetilde{\mathbb{M}}_{t\to N}^{(k)}[\widetilde{Q}_{t,\tau}]\right)}{1+\sum_{k=1}^{3}\theta_k\widetilde{\mathbb{M}}_{t\to N}^{(k)}},
\end{aligned}
$$

where the first identity follows from Lemma A.1.10 and the inequality follows from (A.1.29) and Lemma A.1.11. ■

*Remark* 7. The condition that $\tau\leq\tau^*$ is sufficient but not necessary, as the proof of Theorem A.1.12 shows. Furthermore, the proof also shows that $\tau^*>0$ exists regardless of the utility function. In practice, however, $\tau^*$ is unknown since $G(\cdot)$ depends on the unknown utility function of the representative agent. Appendix A.4.5 shows that $\tau^*\approx 0.5$ in the data for CRRA utility and different levels of risk aversion. In light of this result, it seems that $\tau\in\{0.05,0.1,0.2\}$ is sufficiently conservative for the lower bound to hold, and I use these values in the empirical application in Section 1.6.4.

**Lemma A.1.13.** *Suppose that Assumption 1.6.2 holds. In addition, define $\tau^*$ so that $G(\widetilde{Q}_{t,\tau^*})=\widetilde{\mathbb{E}}_t\left(G(R_{m,t\to N})\right)$, where*

$$
G(R_{m,t\to N})\coloneqq\int_{R_{f,t\to N}}^{R_{m,t\to N}}\zeta^{(4)}(t)(R_{m,t\to N}-t)^3\,\mathrm{d}t.
$$

*Then for all $\tau\leq\tau^*$,*

$$
\widetilde{\mathbb{COV}}_{\mathrm{t}}\left[\mathbb{1}\left(R_{m,t\to N}\leq\widetilde{Q}_{t,\tau}\right),\int_{R_{f,t\to N}}^{R_{m,t\to N}}\zeta^{(4)}(t)(R_{m,t\to N}-t)^3\,\mathrm{d}t\right]\leq 0. \qquad \text{(A.1.30)}
$$

*Proof.* If $\zeta^{(4)} \equiv 0$, then (A.1.30) trivially holds. Hence, assume that $\zeta^{(4)}$ is not identically equal to zero. First we show that $G(R_{m,t\to N})$ is increasing on $(0, R_{f,t\to N})$, since by Leibniz' rule

$$G'(R_{m,t\to N}) = -3\int_{R_{m,t\to N}}^{R_{f,t\to N}} \zeta^{(4)}(t)(R_{m,t\to N} - t)^2 \,\mathrm{d}t \geq 0.$$

The inequality follows since $\zeta^{(4)}(t) < 0$ by Assumption 1.6.2(ii). Temporarily write $K = \widetilde{Q}_{t,\tau}$ to ease notation and consider

$$\Gamma(K) = \widetilde{\mathbb{COV}}_{\mathrm{t}}\left[\mathbb{1}\left(R_{m,t\to N} \leq K\right), \int_{R_{f,t\to N}}^{R_{m,t\to N}} \zeta^{(4)}(t)(R_{m,t\to N} - t)^3 \,\mathrm{d}t\right].$$

By Leibniz' rule again, we get

$$\Gamma'(K) = \widetilde{f}_t(K)\left(G(K) - \widetilde{\mathbb{E}}_t\left(G(R_{m,t\to N})\right)\right).$$

Since $G(R_{f,t\to N}) = 0$, $G(R_{m,t\to N}) \leq 0$ and $G(R_{m,t\to N})$ is increasing on $(0, R_{f,t\to N})$, we know that $\Gamma'(K) \leq 0$ for all $K \leq K^* < R_{f,t\to N}$, where $K^*$ is defined such that $G(K^*) = \widetilde{\mathbb{E}}_t\left(G(R_{m,t\to N})\right)$. To complete the proof, define $\tau^*$ so that it satisfies $\widetilde{Q}_{\tau^*} = K^*$. ■

*Remark* 8. The bound in (A.1.28) is infeasible since $\{\theta_k\}_{k=1}^3$ is unknown.[5] However, Chabi-Yo and Loudis (2020) show that these unknowns relate to the coefficient of relative risk aversion, relative prudence and relative temperance of the representative agent. Based on this observation and using results from the expected utility literature (Eeckhoudt and Schlesinger, 2006), the authors propose an additional restriction on $\theta_k$ that allows me to prove the feasible lower bound in Proposition 1.6.4.

*Proof of Proposition 1.6.4.* Using Assumption 1.6.3((i)) and 1.6.3((ii)), we get $\theta_2\widetilde{\mathbb{M}}_{t\to N}^{(2)} \leq -1/R_{f,t\to N}^2\widetilde{\mathbb{M}}_{t\to N}^{(2)}$ and $\theta_3\widetilde{\mathbb{M}}_{t\to N}^{(3)} \leq 1/R_{f,t\to N}^3\widetilde{\mathbb{M}}_{t\to N}^{(3)}$, from which it follows that

$$1 + \sum_{k=1}^{3} \theta_k\widetilde{\mathbb{M}}_{t\to N}^{(k)} \leq 1 - \frac{1}{R_{f,t\to N}^2}\widetilde{\mathbb{M}}_{t\to N}^{(2)} + \frac{1}{R_{f,t\to N}^3}\widetilde{\mathbb{M}}_{t\to N}^{(3)}. \tag{A.1.31}$$

[5] In Appendix A.5, I use comparative statics for common utility functions to analyze the tail difference between the physical and risk-neutral distribution.

Second, recall that for $K > 0$

$$\widetilde{F}_t(K)\widetilde{\mathbb{M}}^{(k)}_{t\to N} - \widetilde{\mathbb{M}}^{(k)}_{t\to N}[K] = -\widetilde{\mathbb{COV}}_{\mathrm{t}}\left[\mathbb{1}\left(R_{m,t\to N} \leq K\right), \left(R_{m,t\to N} - R_{f,t\to N}\right)^k\right].$$

If $k = 1, 3$, then Chebyshev's sum inequality A.5.1 implies that

$$\Gamma(K) \coloneqq \widetilde{\mathbb{COV}}_{\mathrm{t}}\left[\mathbb{1}\left(R_{m,t\to N} \leq K\right), \left(R_{m,t\to N} - R_{f,t\to N}\right)^k\right] \leq 0.$$

Hence under Assumption 1.6.3((i)),

$$\theta_k\left(\widetilde{F}_t(K)\widetilde{\mathbb{M}}^{(k)}_{t\to N} - \widetilde{\mathbb{M}}^{(k)}_{t\to N}[K]\right) \geq \frac{1}{R^k_{f,t\to N}}\left(\widetilde{F}_t(K)\widetilde{\mathbb{M}}^{(k)}_{t\to N} - \widetilde{\mathbb{M}}^{(k)}_{t\to N}[K]\right) \quad \text{for } k = 1, 3. \tag{A.1.32}$$

If $k = 2$, we obtain from Leibniz' rule

$$\Gamma'(K) = \widetilde{f}_t(K)\left[(K - R_{f,t\to N})^2 - \widetilde{\mathbb{VAR}}_t(R_{m,t\to N})\right]. \tag{A.1.33}$$

It follows that (A.1.33) is positive if $K \leq R_{f,t\to N} - \sqrt{\widetilde{\mathbb{VAR}}_t(R_{m,t\to N})} =: K^{**}$. Combining (A.1.32) and (A.1.33), we get for $K \leq K^{**}$

$$\theta_k\left(\widetilde{F}_t(K)\widetilde{\mathbb{M}}^{(k)}_{t\to N} - \widetilde{\mathbb{M}}^{(k)}_{t\to N}[K]\right) \geq \frac{(-1)^{k+1}}{R^k_{f,t\to N}}\left(\widetilde{F}_t(K)\widetilde{\mathbb{M}}^{(k)}_{t\to N} - \widetilde{\mathbb{M}}^{(k)}_{t\to N}[K]\right). \tag{A.1.34}$$

Collecting the results from (A.1.31) and (A.1.34) and using the general upper bound (A.1.28) from Theorem A.1.12, it follows that

$$\begin{aligned}
\tau - F_t\left(\widetilde{Q}_{t,\tau}\right) &\overset{\text{(A.1.28)}}{\geq} \frac{\sum_{k=1}^3 \theta_k\left(\tau\widetilde{\mathbb{M}}^{(k)}_{t\to N} - \widetilde{\mathbb{M}}^{(k)}_{t\to N}[\widetilde{Q}_{t,\tau}]\right)}{1 + \sum_{k=1}^3 \theta_k\widetilde{\mathbb{M}}^{(k)}_{t\to N}} \\
&\geq \frac{\sum_{k=1}^3 \frac{(-1)^{k-1}}{R^k_{f,t\to N}}\left(\tau\widetilde{\mathbb{M}}^{(k)}_{t\to N} - \widetilde{\mathbb{M}}^{(k)}_{t\to N}[\widetilde{Q}_{t,\tau}]\right)}{1 + \sum_{k=1}^3 \frac{(-1)^{k-1}}{R^k_{f,t\to N}}\widetilde{\mathbb{M}}^{(k)}_{t\to N}},
\end{aligned}$$

for all $\tau$ such that $\widetilde{Q}_{t,\tau} \leq \min(K^*, K^{**})$, where $K^*$ is defined in Theorem A.1.12. ■

*Remark* 9. The bound only holds for quantiles far enough in the left-tail. Compared to Theorem A.1.12, the additional condition needed for the bound to hold is that $\widetilde{Q}_{t,\tau} \leq R_{f,t\to N} -$

$\sqrt{\widetilde{\mathbb{VAR}}_t(R_{m,t\to N})}$, which covers a wide range of quantiles in the left-tail, since in the data $\sqrt{\widetilde{\mathbb{VAR}}_t(R_{m,t\to N})}$ is in the order of $10^{-3}$ for 90-day returns, whereas the risk-free rate is typically around 1.[6]

### A.1.11 Formulas for market moments

This Section presents formulas for the (un)truncated risk-neutral moments of the excess market return. I use a slight abuse of notation and write $\widetilde{Q}(\tau) \coloneqq \widetilde{Q}_\tau(R_{m,t\to N})$, to emphasize that the integrals below are taken with respect to $\tau$.

**Proposition A.1.14.** *Any risk-neutral moment can be computed from the risk-neutral quantile function, since*

$$\widetilde{\mathbb{E}}_t\left[(R_{m,t\to N} - R_{f,t\to N})^n\right] = \int_0^1 [\widetilde{Q}_\tau(R_{m,t\to N} - R_{f,t\to N})]^n \,\mathrm{d}\tau = \int_0^1 [\widetilde{Q}(\tau) - R_{f,t\to N}]^n \,\mathrm{d}\tau. \quad \text{(A.1.35)}$$

*Moreover, any truncated risk-neutral moment can be calculated by*

$$\widetilde{\mathbb{E}}_t\left[(R_{m,t\to N} - R_{f,t\to N})^n \mathbb{1}\left(R_{m,t\to N} \leq k_0\right)\right] = \int_0^{\widetilde{F}_t(k_0)} [\widetilde{Q}(\tau) - R_{f,t\to N}]^n \,\mathrm{d}\tau.$$

*Proof.* For any random variable $X$ and integer $n$ such that the $n$-th moment exists, we have

$$\mathbb{E}\left[X^n\right] = \int_0^1 [Q_X(\tau)]^n \,\mathrm{d}\tau.$$

This follows straightforward from the substitution $x = Q(\tau)$. Now use that for any constant $a \in \mathbb{R}$, $Q_{X-a}(\tau) = Q_X(\tau) - a$ to derive (A.1.35). The truncated formula follows similarly. ■

*Remark* 10. Frequently I use $k_0 = \widetilde{Q}_\tau$, in which case the truncated moment formula reduces to

$$\widetilde{\mathbb{E}}_t\left[(R_{m,t\to N} - R_{f,t\to N})^n \mathbb{1}\left(R_{m,t\to N} \leq \widetilde{Q}_\tau\right)\right] = \int_0^\tau [\widetilde{Q}(p) - R_{f,t\to N}]^n \,\mathrm{d}p.$$

[6] At the 30- and 60-day horizon, the risk-neutral standard deviation is even smaller.

## A.2 Risk-Neutral Quantile Regression: Robustness and Departure from Conditional Lognormality

### A.2.1 Linear versus Non-linear Model: Out-of-Sample Forecasting Accuracy

This section explores alternative specifications to the linear quantile model presented in (1.2.3), focusing only on 30-day returns. The findings for longer time horizons are very similar and omitted for parsimony. Specifically, I consider the addition of higher-order terms to the linear model, such as:

$$Q_{t,\tau} = \beta_0(\tau) + \beta_1(\tau)\widetilde{Q}_{t,\tau} + \beta_2(\tau)\widetilde{Q}^2_{t,\tau}. \tag{A.2.1}$$

To evaluate the performance of the non-linear model in (A.2.1) vs. the linear model in (1.2.3), I recursively estimate the model parameters based on an expanding window, starting at January 2, 2003. The first sub-sample ends at August 15, 2012 and I increase the sample size on a monthly basis. For each sub-sample, I calculate the out-of-sample forecasting accuracy using the formula:

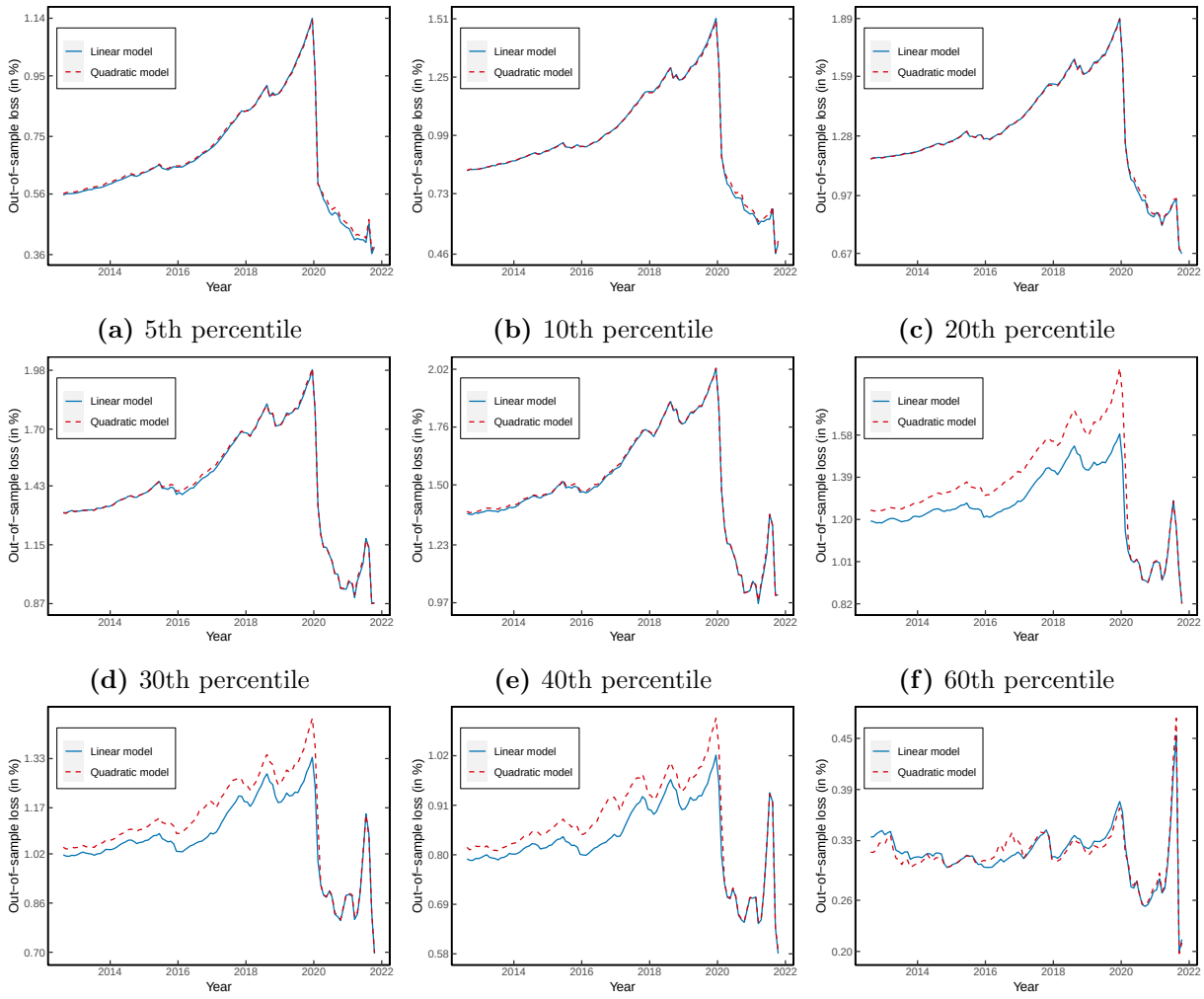
$$\frac{1}{\#t}\sum_t \rho_\tau(R_{m,t\to N} - \widehat{Q}_{t,\tau}), \tag{A.2.2}$$

where $\widehat{Q}_{t,\tau}$ is the predicted physical quantile based on the parameters estimated from the sub-sample. The summation includes all dates that are at least one month ahead of the end of the sub-sample period.

Figure A.3 shows the out-of-sample loss at various percentiles. In most cases, the linear model outperforms the quadratic model, with some exceptions observed at the 95th percentile during specific periods. These results continue to hold when adding other non-linear terms, such as cubic, exponential or logarithmic factors. Additionally, I find that the risk-neutral quantile function exhibits a high correlation with higher-order terms. Consequently, the non-linear model tends to produce quantile forecasts that closely resemble those generated by the linear model.

### A.2.2 Additional Evidence Against the Lognormal Assumption

Table 1.2 already indicates evidence against the lognormal model since the QR estimates in the left- and right-tail are rather different, in contradiction with (1.4.4). To further assess the




**(a)** 5th percentile **(b)** 10th percentile **(c)** 20th percentile

**(d)** 30th percentile **(e)** 40th percentile **(f)** 60th percentile

**(g)** 70th percentile **(h)** 80th percentile **(i)** 95th percentile

**Figure A.3. Out-of-sample quantile forecasting loss**. These figures show the out-of-sample loss of forecasting the physical quantile function, based on an expanding window. The loss at different percentiles is calculated by (A.2.2).

implications of the lognormal model, I analyze the accuracy of the physical quantile forecast in (1.4.5) out-of-sample. Specifically, I use QR based on the first $t_0$ observations to estimate the model

$$Q_{t,\tau}(R_{m,t\to N}) = \widehat{\beta}_{0,t_0}(\tau) + \widehat{\beta}_{1,t_0}(\tau)\widetilde{Q}_{t,\tau}, \tag{A.2.3}$$

where the $t_0$-subscript in $\beta_{\cdot,t_0}$ refers to the fact that the coefficients are estimated using observations up to time $t_0$. Using an expanding window to estimate $\beta_{\cdot,t_0}$, the model produces dynamic quantile forecasts of the form

$$\widehat{Q}^{\text{logn}}_{t,\tau} = \widehat{\beta}_{0,t}(\tau) + \widehat{\beta}_{1,t}(\tau)\widetilde{Q}_{t,\tau}. \tag{A.2.4}$$

In the lognormal case, Proposition 1.4.1((ii)) suggests that $Q_{t,\tau}(R_{m,t\to N}) \approx \widehat{Q}^{\text{logn}}_{t,\tau}$. This approximation can be tested using the joint restriction

$$H_0: \quad [\beta_0(\tau), \beta_1(\tau)] = [0, 1],$$

in the quantile regression

$$\min_{\beta_0,\beta_1\in\mathbb{R}} \sum_t \rho_\tau \left( R_{m,t\to N} - \beta_0 - \beta_1 \widehat{Q}^{\text{logn}}_{t,\tau} \right).$$

The results are summarized in Table A.1 and show that the point estimates are quite far from the $[0, 1]$ benchmark. The Wald test on the joint restriction tends to reject $H_0$ far enough in the tail, but for $\tau = 0.2$ the null hypothesis is never rejected due to the large standard errors. Additionally, the $R^1(\tau)$ statistic shows that the explanatory power is low relative to Table 1.2, even though the sample sizes are different. Hence, the results are incompatible with (1.4.4) and (1.4.5) and provide evidence against the conditional lognormal assumption, which is in line with evidence from the literature (see e.g. Martin (2017, Result 4)).

## A.3 Estimating the Risk-Neutral Quantile Function

### A.3.1 Data Description

To estimate the risk-neutral quantile function at each time point, I use daily option prices from OptionMetrics, covering the period from January 1, 1996, to December 31, 2021. These options include European Put and Call options on the S&P500 index. The option contracts provide data on the highest closing bid, lowest closing ask price, and the price of the forward contract on the underlying security. To approximate the unobserved option price, I use the midpoint between the bid and ask prices. Additionally, I gather daily risk-free rate data from Kenneth French's website.[7] Finally, I obtain stock price data on the closing price of the S&P500 from WRDS.

I implement an additional data cleaning procedure for the option data before estimating the martingale measure. Firstly, I exclude all observations where the highest closing bid price is

[7] See http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html#Research

**Table A.1. Expanding quantile prediction with risk-neutral quantile**

| Horizon | $\tau$ | $\widehat{\beta}_0(\tau)$ | $\widehat{\beta}_1(\tau)$ | Wald test ($p$-value) | $R^1(\tau)$[%] | Obs |
|---|---|---|---|---|---|---|
| 30 days | 0.05 | 0.54 (0.185) | 0.42 (0.193) | 0.00 | 4.36 | 3804 |
| | 0.1 | 0.59 (0.205) | 0.39 (0.212) | 0.01 | 2.39 | |
| | 0.2 | 0.83 (0.332) | 0.15 (0.338) | 0.04 | 0.28 | |
| 60 days | 0.05 | 0.55 (0.310) | 0.39 (0.329) | 0.06 | 1.84 | 3753 |
| | 0.1 | 0.80 (0.339) | 0.16 (0.352) | 0.03 | 0.22 | |
| | 0.2 | 0.87 (0.416) | 0.11 (0.425) | 0.10 | 0.25 | |
| 90 days | 0.05 | 0.78 (0.335) | 0.11 (0.358) | 0.01 | 0.88 | 3702 |
| | 0.1 | 0.74 (0.376) | 0.21 (0.395) | 0.08 | 1.34 | |
| | 0.2 | 0.73 (0.481) | 0.26 (0.491) | 0.31 | 0.52 | |

*Note*: This table reports the QR estimates of (A.2.4) using an expanding window based on an initial 500 observations. The sample period is 2003-2021. *Wald test* denotes the $p$-value of the joint restriction $[\beta_0(\tau), \beta_1(\tau)] = [0, 1]$. Standard errors are reported in parentheses and calculated using the SETBB with a block length equal to the prediction horizon. $R^1(\tau)$ denotes the goodness of fit measure (1.2.5).

zero. Additionally, I remove option prices that violate no-arbitrage bounds. Subsequently, I filter out option prices with maturities less than 7 days or greater than 500 days. Following this cleaning process, I retain 23,264,113 option-day observations.

For the quantile regression application, I exclude all observations before 2003. During the period from 1996 to 2003, there are many days with insufficient option data to estimate $\widetilde{Q}_{t,\tau}$ at the 30-, 60-, and 90-day horizons. I also discard days in the post-2003 period when I cannot estimate the risk-neutral quantile, although this is a rare occurrence, accounting for approximately 2% of the total days. Most of these instances are concentrated at the beginning of the sample period.

### A.3.2 Estimation Procedure

There is a substantial literature on how to extract the martingale measure from option prices. I use the `RND Fitting Tool` application on MATLAB, which is developed by Barletta and Santucci de Magistris (2018).[8]. The tool is based on the orthogonal polynomial expansion of Filipović, Mayerhofer, and Schneider (2013). In short, the idea is to approximate the conditional risk-neutral density function by an expansion of the form

$$\widetilde{f}_t(x) \approx \phi(x) \left[ 1 + \sum_{k=1}^{K} \sum_{i=0}^{k} c_k w_{i,k} x^k \right],$$

where $\phi(x)$ is an arbitrary density and the polynomial term serves to tilt the density function towards the risk-neutral distribution. Further details about the estimation of the coefficients $w_{i,k}$ and $c_k$ can be found in Filipović, Mayerhofer, and Schneider (2013).

For my purpose, I need to choose the kernel function $\phi(\cdot)$, the estimation method for $c_k$ and the degree of the expansion $K$. I follow the recommendation of Barletta and Santucci de Magistris (2018) and use the double beta distribution for the kernel and principal component analysis to estimate $c_k$. This is the most robust method for S&P500 options. To avoid overfitting, I use $K = 3$ if the number of option data is less than 70, $K = 6$ if the number is less than 100 and $K = 8$ otherwise. This choice renders a good approximation for most time periods.

I interpolate the estimated risk-neutral densities for a given time horizon. Occasionally, there are no two interpolation points. In such cases, I drop the observations to avoid negative density estimates due to extrapolation. Since the RND Fitting Tool is designed for an equal number of put and call options, I use Put-Call parity to convert in-the-money call prices to put prices and vice versa. Subsequently, I use Black-Scholes implied volatilities to interpolate the Call-Put option price curve near the forward price. This transformation ensures that the risk-neutral density does not have a discontinuity for strike prices that are close to being at-the-money (Figlewski, 2010). Finally, I integrate the density function and take the inverse to obtain the risk-neutral quantile

[8]The application can be downloaded from the author's GITHUB page: https://github.com/abarletta/rndfittool

function:

$$\widetilde{Q}_{t,\tau} \coloneqq \inf\left\{x \in \mathbb{R} : \tau \leq \widetilde{F}_t(x)\right\}, \quad \text{where } \widetilde{F}_t(x) = \int_0^x \widetilde{f}_t(y)\,\mathrm{d}y.$$

## A.4 Verifying Assumption 1.6.2(ii) in Representative Agent Models

The proof of Theorem A.1.12 relies on Assumption 1.6.2(ii). This section derives parameter restrictions for common utility functions that are needed so that Assumption 1.6.2(ii) is satisfied. Most of these restrictions resemble those of Chabi-Yo and Loudis (2020). I also illustrate the lower bound with actual data assuming CRRA utility.

### A.4.1 Log utility

In this case $u(x) = \log x$. It follows that $\zeta(x) = x/R_{f,t\to N}$. Clearly $\zeta^{(4)}(x) = 0$ and Assumption 1.6.2 holds.

### A.4.2 CRRA utility

More generally, consider $u(x) = \frac{x^{1-\gamma}}{1-\gamma}$ for $\gamma \geq 0$. It follows that $\zeta(x) = (\frac{x}{R_{f,t\to N}})^{\gamma}$ and hence

$$\zeta^{(4)}(x) = \frac{1}{R^{\gamma}_{f,t\to N}}\gamma(\gamma-1)(\gamma-2)(\gamma-3)x^{\gamma-4}.$$

Part (ii) of Assumption 1.6.2 holds if $\gamma \in [0,1]$, but also if $\gamma \in [2,3]$. Notice that the additional restrictions in the feasible lower bound in Proposition 1.6.4 cannot be accommodated by this model. To see this, observe that $\theta_2 \leq -1/R^2_{f,t\to N}$ implies that $\gamma(\gamma-1)/2 \leq -1/R^2_{f,t\to N}$, which cannot hold for any reasonable interest rate. This failure illustrates that a representative agent model with CRRA utility is misspecified in that it cannot produce a sizable risk-premium on skewness.[9]

### A.4.3 CARA utility

In this case, $u(x) = 1 - e^{-\gamma x}$ and $\zeta(x) = e^{\gamma^*(x - R_{f,t\to N})}$, where $\gamma^* = W_t\gamma$. Since $\zeta^{(4)} > 0$, Assumption 1.6.2 does not hold.

[9] See in particular Chabi-Yo and Loudis (2020, Equation (A.5)), which shows that $\theta_2$ is related to the risk-premium on market skewness.

### A.4.4 HARA utility

The utility function is given by $u(x) = \frac{1-\gamma}{\gamma}\left(\frac{ax}{1-\gamma}+b\right)^{\gamma}$, where $a > 0$ and $\frac{ax}{1-\gamma}+b > 0$. Successive differentiation renders

$$\zeta^{(4)}(x) = \frac{-\gamma(\gamma+1)(\gamma+2)(aW_t)^4\left(\frac{aW_t x}{1-\gamma}+b\right)^{-\gamma-3}\left(\frac{aW_t R_{f,t\to N}}{1-\gamma}+b\right)^{\gamma-1}}{(1-\gamma)^3}.$$

We see that $\gamma \in [0,1)$ is a sufficient condition for $\zeta^{(4)}(x) \leq 0$.

### A.4.5 Lower Bound in the Data for CRRA utility

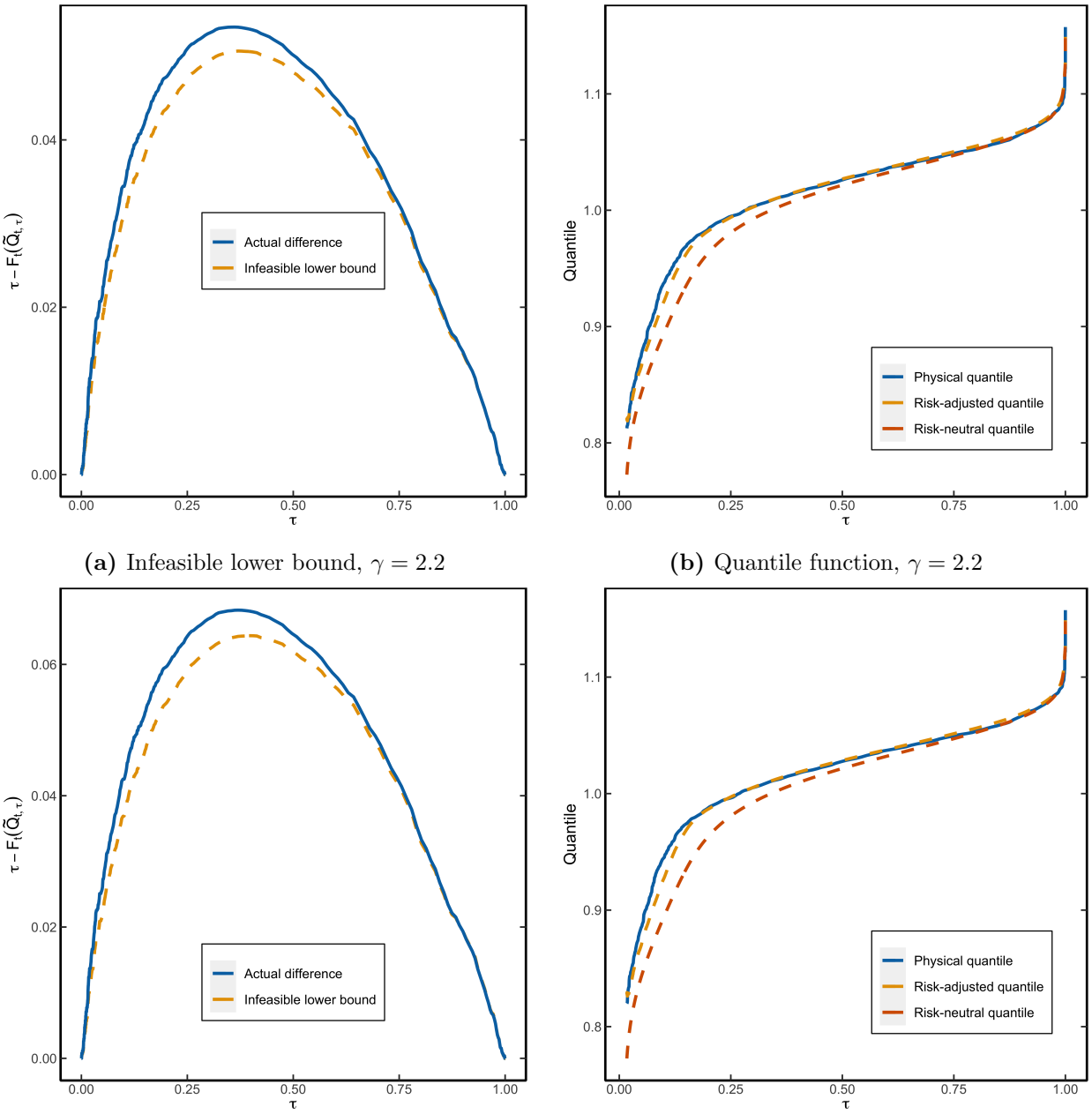
Figure A.4 illustrates the infeasible lower bound as well as the quantile approximation for CRRA utility with different levels of risk aversion. The risk-neutral distribution is obtained from option data over a 90-day horizon on October 28, 2015. Panels A.4a and A.4c show the infeasible lower bound from Theorem A.1.12 when risk aversion is 2.2 and 2.9 respectively. Consistent with the theorem, the infeasible lower bound is below $\tau - F_t(\widetilde{Q}_{t,\tau})$ in the left-tail, and seems to hold for a large range of $\tau$'s, in particular for all $\tau \leq 0.5$. The right panels show the quantile approximation (1.6.9) based on the infeasible lower bound. We see that the risk-adjusted quantile approximation comes much closer to the physical quantile relative to the risk-neutral quantile function.

## A.5 Disaster Probability in Representative Agent Models

In this section, I derive results regarding conditional tail probabilities in representative agent models. I demonstrate how these probabilities can be computed using common utility functions and analyze their sensitivity to changes in underlying parameters (comparative statics). These results do not hinge on specific assumptions about the market return distribution and extend existing findings in the literature, which often rely on log-normality assumptions.

### A.5.1 Log Utility

Chabi-Yo and Loudis (2020, Remark 1) show that their bounds on the equity premium equal the bounds of Martin (2017) when the representative agent has log preferences. Here, I derive the analogous result for the subjective crash probability of a log investor reported by Martin (2017,




**(a)** Infeasible lower bound, $\gamma = 2.2$

**(b)** Quantile function, $\gamma = 2.2$

**(c)** Infeasible lower bound, $\gamma = 2.9$

**(d)** Quantile function, $\gamma = 2.9$

**Figure A.4. Lower bound with CRRA utility for 90-day returns**. This figure shows the lower bound on $\tau - F_t(\widetilde{Q}_{t,\tau})$ as well as the quantile approximation $Q_{t,\tau} \approx \widetilde{Q}_{t,\tau} + \text{LB}_{t,\tau}$ in a representative agent model with CRRA utility function, $u(x) = x^{1-\gamma}/(1-\gamma)$, for $\gamma \in \{2.2, 2.9\}$. The left panels show the infeasible lower bound $\text{CLB}_{t,\tau}$, and the true risk-adjustment, $\tau - F_t(\widetilde{Q}_{t,\tau})$. The right panels show the physical, risk-neutral and risk-adjusted quantile functions. The risk-adjusted quantile function uses the infeasible lower bound. The risk-neutral distribution is coming from option data on the S&P500 on October 28, 2015 with a maturity of 90 days.

Result 2). In our notation, Martin (2017) shows that

$$\mathbb{P}_t\left(R_{m,t\to N} < \alpha\right) = \alpha\left[\text{Put}_t'(\alpha S_t) - \frac{\text{Put}_t(\alpha S_t)}{\alpha S_t}\right], \tag{A.5.1}$$

where $\text{Put}'_t$ is the derivative of the put option price curve seen as a function of the strike. Under log preferences and using (A.1.27), it follows that

$$\begin{aligned}
&\mathbb{P}_t(R_{m,t\to N} < \widetilde{Q}_{t,\tau}) = \tau + \frac{1}{R_{f,t\to N}}\widetilde{\mathbb{COV}}_{\text{t}}\left[\mathbb{1}\left(R_{m,t\to N} \leq \widetilde{Q}_{t,\tau}\right), R_{m,t\to N}\right] \\
&= \tau + \frac{1}{R_{f,t\to N}}\left(\widetilde{\mathbb{E}}_t\left[\mathbb{1}\left(R_{m,t\to N} \leq \widetilde{Q}_{t,\tau}\right) R_{m,t\to N}\right] - \widetilde{\mathbb{E}}_t(R_{m,t\to N})\widetilde{\mathbb{E}}_t\left(\mathbb{1}\left(R_{m,t\to N} \leq \widetilde{Q}_{t,\tau}\right)\right)\right) \\
&= \frac{1}{R_{f,t\to N}}\widetilde{\mathbb{E}}_t\left[\mathbb{1}\left(R_{m,t\to N} \leq \widetilde{Q}_{t,\tau}\right) R_{m,t\to N}\right]. \qquad \text{(A.5.2)}
\end{aligned}$$

The result now follows upon substituting $\widetilde{Q}_\tau = \alpha$, since Martin (2017) shows that (A.5.2) equals the right hand side of (A.5.1).

### A.5.2 CRRA Utility

I now consider the case in which the representative agent has constant relative risk aversion (CRRA) utility, $u(x) = x^{1-\gamma}/(1-\gamma)$, where $\gamma$ is the relative risk aversion parameter. First, I show that the excess market return is non-decreasing in $\gamma$ *regardless* of the distribution of the market return.[10] Next, I extend the argument to show that the difference between the physical and risk-neutral measures increases at every point within their support. The proofs rely on the following lemma, which is a special case of the FKG inequality (Hsu and Varadhan, 1999, Theorem 1.3).

**Lemma A.5.1** (Chebyshev sum inequality)**.** *Let $X$ be a random variable and let $g, h$ both be non-increasing or non-decreasing. Then,*

$$\mathbb{E}\left(g(X)h(X)\right) \geq \mathbb{E}\left(g(X)\right)\mathbb{E}\left(h(X)\right).$$

*The inequality is reversed if one is non-increasing and the other is non-decreasing.*

*Proof.* Let $X_1, X_2$ be IID copies of $X$ and assume that $g, h$ are non-decreasing. It follows that

$$\left(g(X_1) - g(X_2)\right)\left(h(X_1) - h(X_2)\right) \geq 0. \qquad \text{(A.5.3)}$$

Taking expectations on both sides completes the proof. The same proof goes through if $g, h$ are non-increasing. If one is non-increasing and the other is non-decreasing, the inequality in (A.5.3)

[10] Cochrane (2005) derives this result when the distribution is lognormal.

is reversed. ■

**Proposition A.5.2.** *Assume that a representative investor has CRRA utility, with* $\gamma \geq 0$ *and* $\mathbb{E}_t\left[R_{m,t\to N}^{\gamma+1}\log R_{m,t\to N}\right] < \infty$. *Then,* $\mathbb{E}_t\left[R_{m,t\to N}\right] - R_{f,t\to N}$, *is non-decreasing in* $\gamma$.

*Remark* 11. I suppress the dependence of the physical expectation on $\gamma$ in the notation for readability.

*Proof.* According to Chabi-Yo and Loudis (2020, Equation (53)), we have

$$\mathbb{E}_t\left[R_{m,t\to N}\right] - R_{f,t\to N} = \frac{\widetilde{\mathbb{E}}_t\left[R_{m,t\to N}^{\gamma+1}\right]}{\widetilde{\mathbb{E}}_t\left[R_{m,t\to N}^{\gamma}\right]} - R_{f,t\to N} =: g(\gamma).$$

It is enough to show that $g'(\gamma) \geq 0$ for $\gamma \geq 0$. Taking first order conditions, we need to show that

$$\widetilde{\mathbb{E}}_t\left[R_{m,t\to N}^{\gamma+1}\log R_{m,t\to N}\right]\widetilde{\mathbb{E}}_t\left[R_{m,t\to N}^{\gamma}\right] \geq \widetilde{\mathbb{E}}_t\left[R_{m,t\to N}^{\gamma+1}\right]\widetilde{\mathbb{E}}_t\left[R_{m,t\to N}^{\gamma}\log R_{m,t\to N}\right]. \tag{A.5.4}$$

Introduce another probability measure $\mathbb{P}^*$, defined by

$$\mathbb{E}_t^*\left[Z\right] \coloneqq \frac{\widetilde{\mathbb{E}}_t\left[ZR_{m,t\to N}^{\gamma}\right]}{\widetilde{\mathbb{E}}_t\left[R_{m,t\to N}^{\gamma}\right]}. \tag{A.5.5}$$

We can rewrite (A.5.4) into

$$\mathbb{E}_t^*\left[R_{m,t\to N}^{\gamma}\log R_{m,t\to N}\right] \geq \mathbb{E}_t^*\left[R_{m,t\to N}^{\gamma}\right]\mathbb{E}_t^*\left[\log R_{m,t\to N}\right]. \tag{A.5.6}$$

Inequality (A.5.6) now follows from Lemma A.5.1. ■

I mimic the steps above to show that the physical distribution differs more from the risk-neutral distribution at every point in the support, whenever risk aversion is increasing. As before, the dependence of the physical measure on $\gamma$ is omitted.

**Proposition A.5.3.** *Assume that a representative investor has CRRA utility, with* $\gamma \geq 0$ *and* $\mathbb{E}_t\left[R_{m,t\to N}^{\gamma}\log R_{m,t\to N}\right] < \infty$, *then* $F_t(x)$ *is non-increasing in* $\gamma$. *In particular,* $\tau - F_t(\widetilde{Q}_{t,\tau})$ *is non-decreasing in* $\gamma$.

*Proof.* I start from the relation

$$F_t(x) = \widetilde{\mathbb{E}}_t \left[ \frac{R^{\gamma}_{m,t\to N}}{\widetilde{\mathbb{E}}_t \left[ R^{\gamma}_{m,t\to N} \right]} \mathbb{1} \left( R_{m,t\to N} \leq x \right) \right].$$

From first order conditions, we need to show that

$$\widetilde{\mathbb{E}}_t \left[ \log(R_{m,t\to N}) \mathbb{1} \left( R_{m,t\to N} \leq x \right) R^{\gamma}_{m,t\to N} \right] \widetilde{\mathbb{E}}_t \left[ R^{\gamma}_{m,t\to N} \right] \leq \\ \widetilde{\mathbb{E}}_t \left[ R^{\gamma}_{m,t\to N} \mathbb{1} \left( R_{m,t\to N} \leq x \right) \right] \widetilde{\mathbb{E}}_t \left[ \log(R_{m,t\to N}) R^{\gamma}_{m,t\to N} \right].$$

Using the same change of measure as in (A.5.5), we obtain the equivalent statement

$$\mathbb{E}^*_t \left[ \log(R_{m,t\to N}) \mathbb{1} \left( R_{m,t\to N} \leq x \right) \right] \leq \mathbb{E}^*_t \left[ \mathbb{1} \left( R_{m,t\to N} \leq x \right) \right] \mathbb{E}^*_t \left[ \log R_{m,t\to N} \right].$$

This inequality holds, since $\log(y)$ and $\mathbb{1}\left(y \leq x\right)$ are respectively increasing and non-increasing in $y$, hence the result follows from Lemma A.5.1. Using the substitution $x \to \widetilde{Q}_{t,\tau}$, it follows that $\tau - F_t(\widetilde{Q}_{t,\tau})$, is non-decreasing in $\gamma$. ■

### A.5.3 Exponential utility

Here, I assume that the representative agent has exponential utility, $u(x) = 1 - e^{-\gamma^* x}$, where $\gamma^*$ is the absolute risk aversion. According to Chabi-Yo and Loudis (2020, Equation (55)), the following expression for the equity premium obtains

$$\mathbb{E}_t \left[ R_{m,t\to N} \right] - R_{f,t\to N} = \frac{\widetilde{\mathbb{E}}_t \left[ R_{m,t\to N} e^{\gamma R_{m,t\to N}} \right]}{\widetilde{\mathbb{E}}_t \left[ e^{\gamma R_{m,t\to N}} \right]} - R_{f,t\to N},$$

where $\gamma = \gamma^* W_t$ is relative risk aversion and $W_t$ represents the agent's wealth at time $t$. Since there is a one-to-one relation between $\gamma$ and $\gamma^*$, it follows from the results in Section A.5.2 that the equity premium is increasing in $\gamma^*$, and so is the distance between the physical and risk-neutral distribution, as measured by: $\tau - F_t(\widetilde{Q}_{t,\tau})$.

## A.6 Lower Bound in the Data and Robustness

### A.6.1 Lower Bound in the Data

In the empirical application, I calculate the lower bound, $\text{LB}_{t,\tau} = \text{CLB}_{t,\tau}/\widetilde{f}_t(\widetilde{Q}_{t,\tau})$, for 30-, 60-, and 90-day returns. Summary statistics of $\text{LB}_{t,\tau}$ are presented in Table A.2. The lower bound is right-skewed and is most significant at the 5th and 10th percentile. Moreover, over the 30-day horizon, it can reach as high as 25% and maintains an average of approximately 1% in the far left-tail.

**Table A.2. Summary statistics of lower bound**

| Horizon | $\tau$ | Mean | Median | Std. dev. | Min | Max |
|---|---|---|---|---|---|---|
| 30 days | 0.05 | 0.92 | 0.63 | 1.07 | 0.08 | 24.38 |
| | 0.1 | 0.70 | 0.45 | 0.87 | 0.06 | 12.22 |
| | 0.2 | 0.47 | 0.25 | 0.74 | 0.04 | 10.93 |
| 60 days | 0.05 | 1.81 | 1.31 | 1.67 | 0.10 | 19.23 |
| | 0.1 | 1.71 | 1.19 | 1.66 | 0.25 | 19.89 |
| | 0.2 | 1.14 | 0.69 | 1.50 | 0.12 | 23.57 |
| 90 days | 0.05 | 2.65 | 2.02 | 2.02 | 0.02 | 18.63 |
| | 0.1 | 2.86 | 2.12 | 2.32 | 0.04 | 24.47 |
| | 0.2 | 1.97 | 1.22 | 2.33 | 0.26 | 28.92 |

*Note*: This table reports summary statistics of the lower bound, $\text{LB}_{t,\tau} = \text{CLB}_{t,\tau}/\widetilde{f}_t(\widetilde{Q}_{t,\tau})$, in (1.6.13) at different time horizons and different quantile levels over the sample period 2003-2021. All statistics are in percentage point.

### A.6.2 Robustness of the Lower Bound and Risk-neutral Quantile

The lower bound, $\text{LB}_{t,\tau}$, tries to capture the difference between the physical and risk-neutral quantile functions in the left-tail. What are some other measures that are available at a daily frequency and contain information about the quantile wedge? One candidate is the VIX index, which is defined as

$$\text{VIX}_t^2 = \frac{2R_{f,t\to N}}{N}\left[\int_0^{F_t}\frac{1}{K^2}\text{Put}_t(K)\,\mathrm{d}K + \int_{F_t}^{\infty}\frac{1}{K^2}\text{Call}_t(K)\,\mathrm{d}K\right],$$

where $F_t$ is the forward price on the S&P500, and $\text{Put}_t(K)$ (resp. $\text{Call}_t(K)$) is the put (resp. call) option price on the S&500 with strike $K$. Martin (2017) shows that VIX measures risk-neutral

entropy

$$\text{VIX}_t^2 = \frac{2}{N}\widetilde{L}_t\left(\frac{R_{m,t\to N}}{R_{f,t\to N}}\right),$$

where entropy is defined as $\widetilde{L}_t(X) \coloneqq \log \widetilde{\mathbb{E}}_t[X] - \widetilde{\mathbb{E}}_t[\log X]$. Entropy, just like variance, is a measure of spread in the distribution. However, entropy places more weight on left-tail events than variance, since entropy places more weight on out-of-the money put options. As such, VIX is a natural candidate to explain potential differences between $Q_{t,\tau}$ and $\widetilde{Q}_{t,\tau}$. Second, the Chicago Board Options Exchange provides daily data on VIX at the 30-day horizon.

Table A.3 shows the result of the quantile regression

$$Q_{t,\tau}(R_{m,t\to N}) - \widetilde{Q}_{t,\tau}(R_{m,t\to N}) = \beta_0(\tau) + \beta_1(\tau)\text{LB}_{t,\tau} + \beta_{\text{VIX}}(\tau)\text{VIX}_t. \tag{A.6.1}$$

We see that $\beta_{\text{VIX}}$ is marginally significant in the left-tail. In contrast, $\beta_1(\tau)$ is even more significant compared to Table 1.5. Furthermore, the explanatory power of the model that only includes VIX is lower compared to the model that only includes $\text{LB}_{t,\tau}$ (see Table 1.5).

**Table A.3. Quantile regression using Lower Bound and VIX**

| | $\widehat{\beta}_0(\tau)$ | $\widehat{\beta}_1(\tau)$ | $\widehat{\beta}_{\text{VIX}}(\tau)$ | $R^1(\tau)$[%] | $R^1(\tau)$[%] (VIX only) |
|---|---|---|---|---|---|
| $\tau = 0.05$ | -0.20 (1.889) | 10.09 (0.319) | -0.30 (0.130) | 6.34 | 5.51 |
| $\tau = 0.1$ | -0.35 (1.313) | 5.06 (0.302) | -0.22 (0.089) | 3.41 | 2.84 |
| $\tau = 0.2$ | -0.28 (0.955) | 3.62 (0.256) | -0.25 (0.068) | 0.61 | 0.18 |

*Note*: This table reports the QR estimates of (A.6.1) over the 30-day horizon. The sample period is 2003-2021, standard errors are shown in parentheses and calculated using SETBB with a block length equal to the forecast horizon. $R^1(\tau)$ denotes the goodness-of-fit measure (1.2.5). The last column denotes the goodness-of-fit in the model that only uses VIX as covariate. The standard error and point estimate of $\beta_0$ is multiplied by 100 for readability.

As a second robustness check, I consider how well the direct quantile forecast, $\widehat{Q}_{t,\tau} = \widetilde{Q}_{t,\tau} + \text{LB}_{t,\tau}$, compares to the VIX forecast. Since $\widehat{Q}_{t,\tau}$ does not require any parameter estimation, this exercise is a measure of out-of-sample performance. However, VIX does not directly measure $Q_{t,\tau}$

and hence I use an expanding window to obtain the VIX benchmark: $\widehat{Q}^{\text{VIX}}_{t,\tau} \coloneqq \widehat{\beta}_0(\tau) + \widehat{\beta}_1(\tau)\text{VIX}_t$. Finally, I use the following out-of-sample metric to compare both forecasts

$$R^1_{oos}(\tau) = 1 - \sum_{t=500}^{T} \rho_\tau(R_{m,t\to N} - \widehat{Q}_{t,\tau}) / \sum_{t=500}^{T} \rho_\tau(R_{m,t\to N} - \widehat{Q}^{\text{VIX}}_{t,\tau}).$$

Notice that $R^1_{oos}(\tau) > 0$, if $\widehat{Q}_{t,\tau}$ attains a lower error than $\widehat{Q}^{\text{VIX}}_{t,\tau}$. This exercise is more ambitious, since $\widehat{Q}^{\text{VIX}}_{t,\tau}$ makes use of in-sample information. Nonetheless, Figure A.5 shows that $\widehat{Q}_{t,\tau}$ outperforms the VIX predictor at all percentiles.

Figure A.6 performs a similar exercise in the right-tail, but using $\widetilde{Q}_{t,\tau}$ instead of $\widehat{Q}_{t,\tau}$, since Table 1.2 shows that the risk-neutral quantile is a good approximation to $Q_{t,\tau}$ in the right-tail. We see that $\widetilde{Q}_{t,\tau}$ outperforms $\widehat{Q}^{\text{VIX}}_{t,\tau}$ at all quantile levels. Hence, the risk-neutral approximation in the right-tail is more accurate than using the in-sample VIX measure.

### A.6.3 Measurement Error Bias in Quantile Regression

In the empirical application, we have to estimate $\widetilde{Q}_{t,\tau}, \widetilde{f}(\cdot)$ and $\text{CLB}_{t,\tau}$ from market data. Therefore, the estimated coefficients in the quantile regression are biased due to measurement error in the covariate. I present simulation evidence which shows that the bias is small in finite samples.

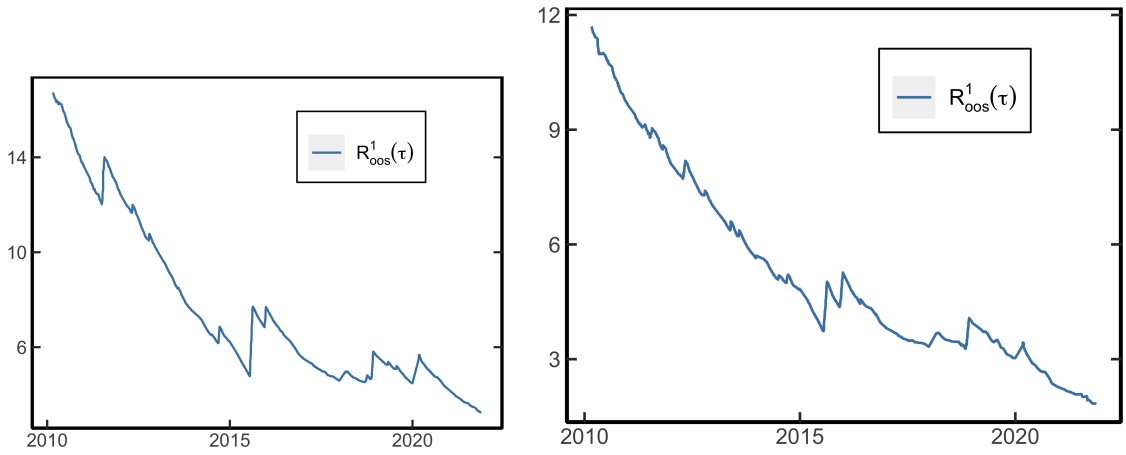
The setup is as follows. I simulate returns and option prices according to a discretized version of the Black and Scholes (1973b) model:

$$\begin{aligned} R_{m,t\to N} &= \exp\left((\mu_t - \frac{1}{2}\sigma_t^2)N + \sigma_t\sqrt{N}Z_{t+N}\right), \quad Z_{t+N} \sim \mathcal{N}(0,1) \qquad \text{(A.6.2)} \\ \sigma_t &\sim \textbf{Unif}\left[0.05, 0.35\right] \\ \mu_t &\sim \textbf{Unif}\left[-0.02, 0.2\right]. \end{aligned}$$

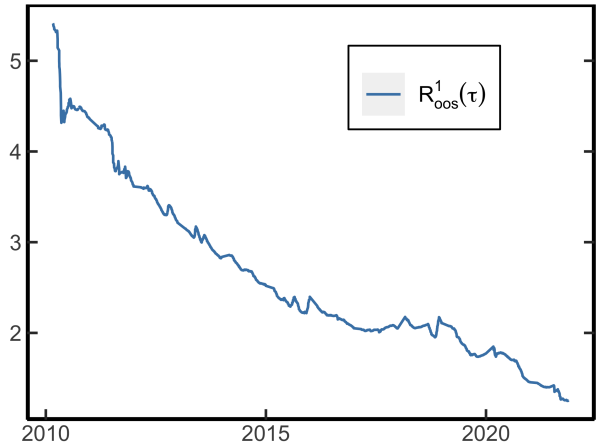
The return distribution under risk-neutral dynamics is given by

$$\widetilde{R}_{m,t\to N} = \exp\left((r_t - \frac{1}{2}\sigma_t^2)N + \sigma_t\sqrt{N}Z_{t+N}\right) \qquad \text{(A.6.3)}$$

$$r_t \sim \textbf{Unif}\left[0, 0.03\right]. \qquad \text{(A.6.4)}$$




**(a)** $\tau = 0.05$ **(b)** $\tau = 0.1$

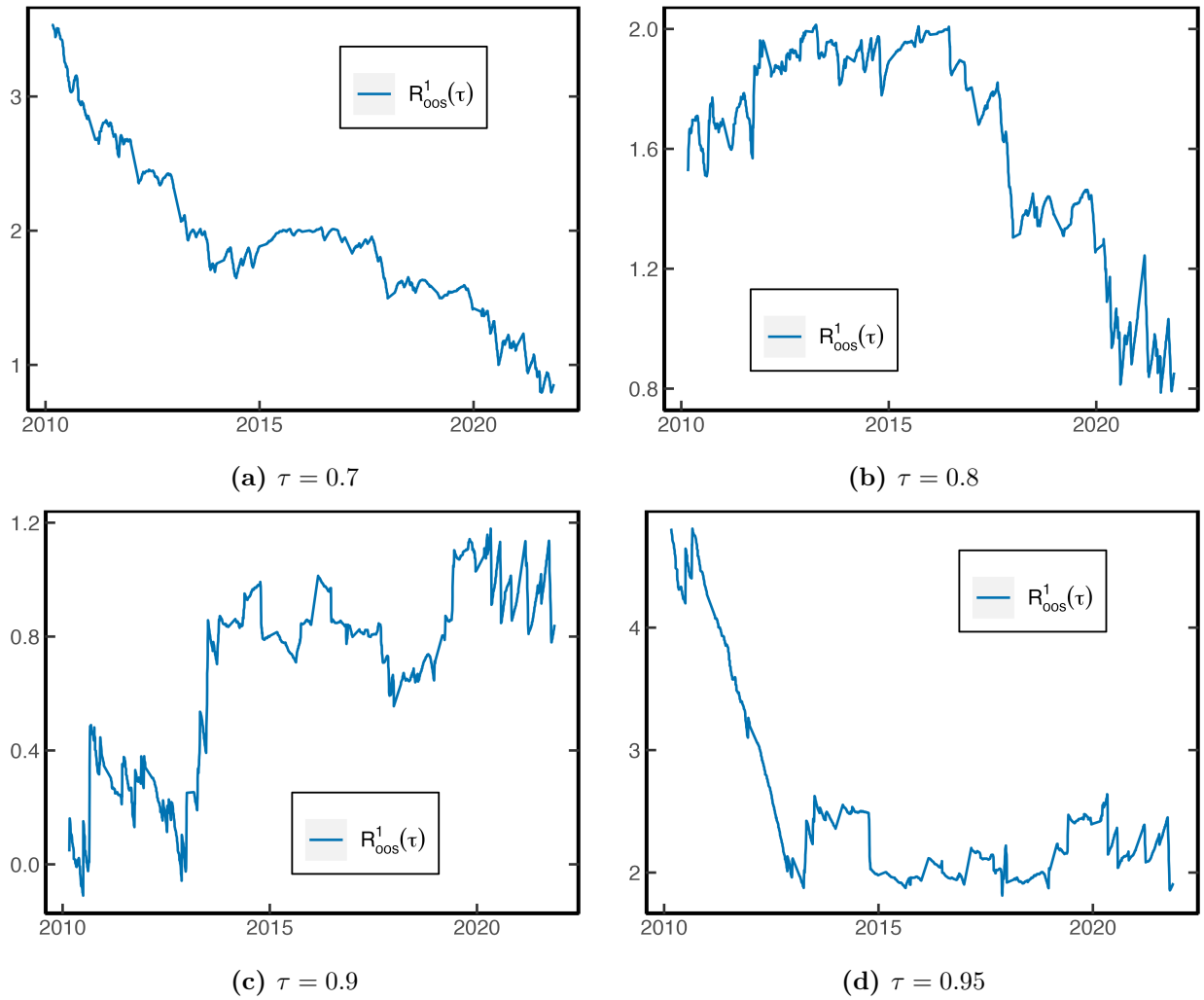



**(c)** $\tau = 0.2$

**Figure A.5. Out-of-sample forecast using risk-adjusted quantile with VIX benchmark.** This figure shows the cumulative out-of sample $R^1(\tau)$, defined as $R^1_{oos}(\tau) = 1 - \sum_{t=500}^{T} \rho_\tau(R_{m,t\to N} - \widehat{Q}_{t,\tau})/\sum_{t=500}^{T} \rho_\tau(R_{m,t\to N} - \widehat{Q}^{\text{VIX}}_{t,\tau})$, where $\widehat{Q}_{t,\tau} = \widetilde{Q}_{t,\tau} + \text{LB}_{t,\tau}$, $\widehat{Q}^{\text{VIX}}_{t,\tau} = \widehat{\beta}_0(\tau) + \widehat{\beta}_1(\tau) \cdot \text{VIX}_t$, and $\widehat{\beta}_0(\tau), \widehat{\beta}_1(\tau)$ are the regression estimates from a quantile regression of $R_{m,t\to N}$ on $\text{VIX}_t$, using data only up to time $t$. The horizon is 30 days and the QR estimates are dynamically updated using an expanding window over the period 2003–2021. The initial sample uses 500 observations.

I calculate the lower bound assuming a return horizon of 90 days. As in the empirical application, I assume that options with an exact 90-day maturity are not available, but instead we observe options with maturity 85 and 97 days. I generate a total of 1,000 options every time period with maturities randomly sampled from 85 and 97 days.[11] These numbers are roughly consistent with the latter part of the empirical sample. The procedure is repeated for a total of 1,000 time periods. For the entire sample, I compare the estimated and analytical lower bound, which are

[11] So on average there will 500 options with maturity 85 days and 500 with maturity 97 days.




**(a)** $\tau = 0.7$ **(b)** $\tau = 0.8$

**(c)** $\tau = 0.9$ **(d)** $\tau = 0.95$

**Figure A.6. Out-of-sample forecast using risk-neutral quantile with VIX benchmark.** This figure shows the cumulative out-of sample $R^1(\tau)$, defined as $R^1_{oos}(\tau) = 1 - \sum_{t=500}^{T} \rho_\tau(R_{m,t\to N} - \widetilde{Q}_{t,\tau}) / \sum_{t=500}^{T} \rho_\tau(R_{m,t\to N} - \widehat{Q}^{\text{VIX}}_{t,\tau})$, where $\widehat{Q}^{\text{VIX}}_{t,\tau} = \widehat{\beta}_0(\tau) + \widehat{\beta}_1(\tau) \cdot \text{VIX}_t$, and $\widehat{\beta}_0(\tau), \widehat{\beta}_1(\tau)$ are the regression estimates from a quantile regression of $R_{m,t\to N}$ on $\text{VIX}_t$, using data only up to time $t$. The horizon is 30 days and the QR estimates are dynamically updated using an expanding window over the period 2003–2021. The initial sample uses 500 observations.

given by respectively

$$\text{LB}^e_{t,\tau} \coloneqq \widehat{\widetilde{Q}}_{t,\tau} + \frac{\widehat{\text{CLB}_{t,\tau}}}{\widehat{\widetilde{f}_t(\widetilde{Q}_{t,\tau})}}$$
$$\text{LB}^a_{t,\tau} \coloneqq \widetilde{Q}_{t,\tau} + \frac{\text{CLB}_{t,\tau}}{\widetilde{f}_t(\widetilde{Q}_{t,\tau})}.$$

The hats signify that the risk-neutral quantile, PDF and CDF lower bound are estimated from the available options at time $t$, using the procedure in Appendix A.3.2. The terms in $\text{LB}^a_{t,\tau}$ are

obtained from the known analytical expression of the risk-neutral distribution (recall (A.6.3)). I then use QR to estimate the models

$$Q_{t,\tau} = \widehat{\beta}_0(\tau) + \widehat{\beta}_{1,e}(\tau)\mathrm{LB}^e_{t,\tau}$$
$$Q_{t,\tau} = \widehat{\beta}_0(\tau) + \widehat{\beta}_{1,a}(\tau)\mathrm{LB}^a_{t,\tau}.$$

I use the ratio $\widehat{\beta}_{1,e}/\widehat{\beta}_{1,a}$ to measure the relative bias in the sample. This experiment is repeated 500 times to get a distribution of the relative bias. Figure A.7 shows boxplots of the bias for several quantiles. We see that the relative bias is very small and centered around 1. Hence, the error in measurement problem resulting from estimating the lower bound is limited in this case.

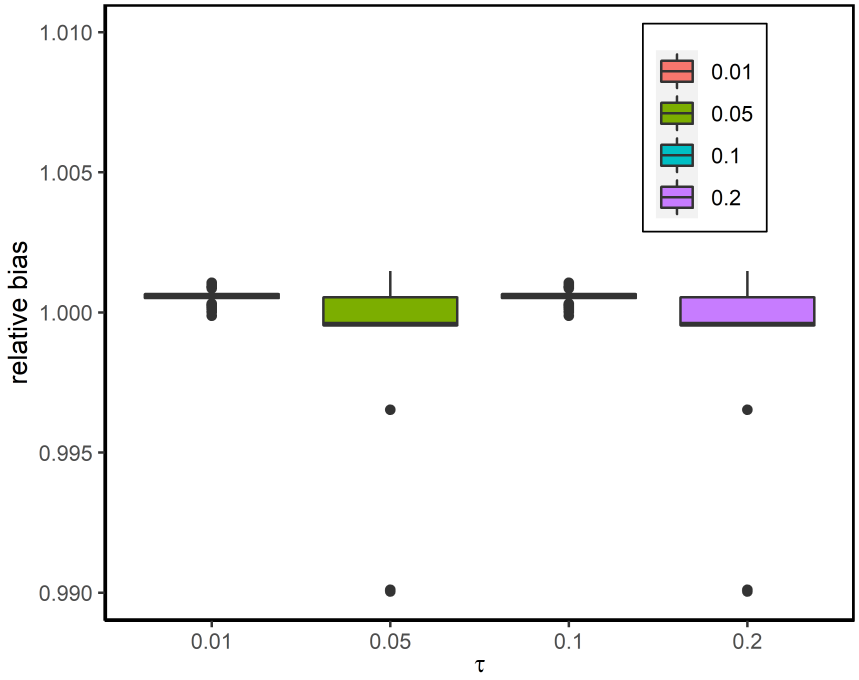



**Figure A.7. Bias in QR resulting from measurement error.** This boxplot shows the relative bias in the quantile regression estimate as a result of measurement error.

## A.7 Additional Figures

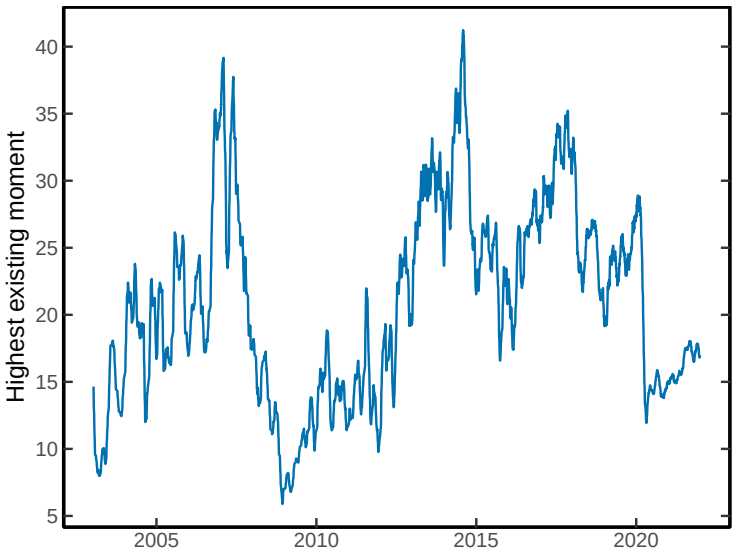



**Figure A.8. Highest existing risk-neutral moment for 30-day returns.** This figure shows $p_t^* := \sup\{p : \widetilde{\mathbb{E}}_t \left(R_{m,t\to N}^p\right) < \infty\}$ over time, where $R_{m,t\to N}$ represents the 30-day return. $p_t^*$ is calculated from the moment formula of Lee (2004), $p_t^* = \frac{1}{2\beta_R} + \frac{\beta_R}{8} + \frac{1}{2}$, where $\beta_R = \limsup_{x\to\infty} \frac{\sigma_{\text{IV}}^2(x)}{|x|/N}$, $\sigma_{\text{IV}}(x)$ is the implied volatility at log-moneyness $x = \log(K/(e^{rN}S_0))$, and $N = 30/365$ is the time horizon. $\beta_R$ is estimated from the call option with highest available strike price. The figure is smoothed using a 30-day moving average.

# Appendix B

# Appendix to Chapter 2

## B.1 CEM data

This appendix discusses details of the CEM data collected from annual surveys sent out to a large sample of international pension plans. To participate in the survey (and to receive its results), CEM requires plans to report data on asset returns and costs by sub-asset class. Each of these sub-asset classes are further split into active/passive and internal/external management styles. CEM classifies internally managed holdings and returns as internal if the buy/sell decision is made within the pension fund organization. In addition, plans are asked to report policy returns, benchmarks, policy weights, and the number of external mandates—all within each sub-asset class—a unique feature of the CEM database. Other questions in the survey pertain to governance, operations and support costs as well as information such as the number of active plan members, the type of investments being offered and the percentage of the plan's liabilities due to retirees. Only a small number of variables are constructed by CEM themselves, such as a plan's asset volatility, which is computed using CEM's internal model.

A benefit of the CEM database is that there are no systematic biases in reporting related to performance. After consultation with CEM, it appears that plans' decision to report in a specific year is unrelated to their investment performance.[1] This conclusion is also reached in a study by Bauer, Cremers, and Frehen (2010). However, most of the pension funds that provide data to CEM are typically larger in size, compared to the average pension plan. Our data show that the average plan size in the United States in 2019 was approximately $25 billion, and the maximum AUM

[1] An important incentive for plans to participate in the CEM survey is to compare their performance and fees, as well as asset allocations, with those of other pension plans.

recorded was $376 billion, which included 10 sponsors with over $100 billion in AUM. Notably, the eight largest U.S. sponsors in our data are among the top 10 largest DB plans nationwide.

According to our database, U.S. domiciled DB plans held a total of $3.81 trillion in AUM in 2019, compared to a total AUM of $8.1 trillion in aggregate across all U.S. DB plans (Investment Company Institute, 2021, p. 177). Of the total AUM, public plans contributed $2.54 trillion, while private plans contributed $1.27 trillion. Hence, our sample covers approximately 38% of the total AUM in the U.S. public sector, which amounted to $6.68 trillion in 2019. Moving outside the U.S., our coverage of AUM includes $1.61 trillion in Canada, $2.42 trillion in Europe (including the UK), and $1.2 trillion in the rest of the world.[2]

## B.2 Asset Allocation

### B.2.1 Asset Class Frequency and Geographic Coverage

To track pension plans' asset allocation, CEM groups each plan's holdings data into six major asset classes, namely stocks, fixed income, hedge funds and multi-asset, private equity, private debt, and real assets. These broad asset classes are further divided into sub-asset classes, as described in Section B.2.2 below.

Panel A in Table B.1 reports the total number of plans in the survey as well as the number of plans reporting holdings within each asset class for each year.[3] The total number of plans participating in the survey ranges from 123 in 1991 to 448 in 2012 and ends at 308 in 2019 (see Panel B).

The most frequently held asset classes are, by far, stocks and fixed income, followed by real assets and private equity, hedge funds, and private debt - the latter being distinctly less common than the other asset classes. Prior to 2000, it was uncommon for plans to hold private debt or hedge fund investments, but these asset classes have been increasingly embraced by plans during the latter years in our sample particularly after 2010, in the case of private debt.

Table B.2 shows the coverage for all countries in our database at two points in time, 2009 and 2019. In 2009, plans domiciled in countries such as Australia, South Korea, Sweden, New Zealand, France, the UK, and Denmark are included. Plans domiciled in China, Saudi Arabia,

[2]CEM provides exchange rates for all countries and years, which allows us to convert foreign currency denominated AUM to U.S. dollar AUM.

[3]From our discussions with CEM, a plan always reports its holdings and returns for every asset class.

Switzerland, Germany, the Emirates or South Africa show up in the survey at some point during our sample. Still, the Netherlands, Canada, and the U.S. account for more than 70% of total AUM throughout our sample.

### B.2.2 Asset and Sub-asset Classes

The CEM database contains information about cost, returns and allocation at the sub-asset class level. In this section we provide details about each of these individual sub-asset classes.

**Stocks**

- U.S. Stocks: U.S. small, mid and large cap stocks. This category also includes U.S. 130/30 type investment strategies.
- Europe: Stock investments in the following countries: Austria, Belgium, Denmark, Finland, France, Germany, Ireland, Italy, Netherlands, Norway, Portugal, Spain, Sweden, Switzerland and United Kingdom.
- Asia-Pacific: Stock investments in Australia, Hong Kong, Japan, New Zealand and Singapore.
- EAFE: Mandates invested primarily in Europe, Australasia, and the Far East (EAFE). Countries in this category include Australia, Austria, Belgium, Denmark, Finland, France, Germany, Hong Kong, Ireland, Israel, Italy, Japan, Luxembourg, Netherlands, New Zealand, Norway, Portugal, Singapore, Spain, Sweden, Switzerland and the United Kingdom
- Emerging: Emerging markets and any other countries not explicitly listed in the above categories.
- ACWI x U.S.: MSCI All Country World Index excluding the United States.
- Global: Mandates invested on a global basis.

**Fixed Income**

- U.S. fixed income: Mainly U.S. Treasury notes or U.S. mortgage backed securities.
- Long Bonds: Dedicated strategies where a manager has a mandate to invest in long bonds. Typically these bonds are due to mature between 10 and 30 years in the future.

- High Yield Bonds: Bonds issued by entities that do not meet the criteria for receiving investment-grade ratings from a major credit rating agency High yield mandates are included in this category as well.
- Bundled LDI: External mandates which blend fixed income and derivatives to generate returns aimed at hedging plan liabilities.
- Cash & Equivalents: Cash managed as a separate asset class, including cash underlying derivative positions.

**Hedge Funds & Multi-asset**

- Hedge Funds: Funded absolute return strategies, i.e. strategies that are equity market neutral.
- Funded Global TAA: Fully funded long-only segregated asset pool dedicated to tactical asset allocation.
- Risk parity: Portfolios aiming to distribute the overall portfolio risk evenly across various asset classes within a diversified portfolio. The portfolio is diversified while meeting return expectations through the use of leverage.

**Private Equity**

- Venture Capital.
- LBO and Energy partnerships.
- Other private equity: Unlisted equity investments in turnarounds, start-ups, mezzanine, distressed financing and energy partnerships.
- Diversified: All private equity investments if the plan does not distinguish between the above categories.

**Private Credit**

- Direct lending, non-traded loans, leveraged loans, distressed bank loan/debt products, mezzanine and other private debt or private credit arrangements.
- Mortgages: Direct mortgages, not including mortgage-backed securities. Mortgage-backed securities are treated as fixed income.

**Real Assets**

- Commodities: Actual physical investments in commodities (crude oil, sugar, copper etc.), commodity funds or products that may invest in an index like the S&P GSCI. Derivative exposures that are fully backed by cash (not just the margin requirement) are also included in this category.

- REITs: Real estate investment trust (REIT) is a type of entity that possesses and often manages income-generating real estate properties. These properties can encompass various forms of commercial real estate, including office buildings, apartment complexes, warehouses, shopping centers, hotels, and more.

- Real Estate: Direct real estate holdings, segregated real estate holdings, and more. nternal real estate management refers to in-house staff making decisions to buy or sell individual properties. Any other approach is considered an external real estate holding. This category also includes joint ventures.

- Infrastructure: Local distribution networks for utilities like electricity, water, and gas, as well as specific transportation assets like toll roads, airports, bridges, and tunnels. Internal infrastructure management indicates that in-house personnel are responsible for deciding when to acquire or divest these assets.

- Other Real Assets: Investments in real assets other than the classes described above.

### B.2.3 Evolution in Asset Allocation

Figure B.1 shows the time-series evolution in total AUM by asset class, aggregated across all U.S. (left panel) and non-U.S. (right panel) plans in our sample. We note a marked shift towards greater coverage of plans outside the U.S. during our sample.

U.S. plans' investments in equities rise steadily from around $252 billion at the beginning of the sample to $1.5 trillion at the end. This increase reflects the cumulative effect of high equity returns during our sample along with increased inflows to equity investments for existing plans and the increased number of plans included in the CEM survey. Fixed income investments rise from $182 billion in 1991 to more than $1.24 trillion in 2019. The remaining asset classes all start at low

levels in the early sample but rise steadily, ending at levels that exceed $300 billion in 2019 for real assets and private equity and just below $250 billion for hedge funds and multi-asset.

For non-U.S. plans, a very different allocation pattern emerges with stocks and fixed income holdings following almost identical paths, both ending near $1.7 trillion in 2019. Among the alternative asset classes, real assets are relatively more important for non-U.S. plans than for their U.S. counterparts, although the ranking at the end of the sample is the same as that for U.S. plans.

Supplementing Figure B.1, Table B.3 reports the time-series evolution in asset allocation for U.S. and non-U.S. plans during our sample. For U.S. plans (Panel A), stock holdings account for a little over half (55%) of total asset values in the early sample, peaking at a share of 63% in 1999, before declining to 40% in 2019. Fixed income holdings account for 35-40% of overall portfolio values in the 1990s, before falling to a range of 25-30% between 1999 and 2007 and retaining a fairly steady portfolio weight averaging 32% from 2008 to 2019.

Hedge fund and multi-asset holdings rise from roughly 1% in 2004 to more than 4% in 2009. In the last five years of the sample, plans hold around 6% of their assets in hedge funds. Allocations to private equity start out around 2% at the beginning of the sample, rise to 4% in 2000, before doubling to 8% in 2010 and remaining in the 8-9% range for the rest of our sample. Private debt holdings account for less than 0.1% of AUM prior to 2003 but rise modestly to end up at 2% in 2019. Finally, real assets hover around 4% during the nineties, rise to around 7% over the next decade and end up at 10% in the last years in our sample.

For non-U.S. plans (Panel B), we observe similar patterns. At the beginning of the sample, the vast majority of plan assets is allocated to stocks and fixed income. In contrast to the U.S. sample, however, fixed income takes up most of the investments (57%), followed by stock holdings (36%) in the early part of the sample. Over time, alternative asset classes become more prominent, with hedge fund and multi-asset holdings accounting for 4.8% of total assets at the end of our sample. Private equity (7.8%) and real assets (16%) in particular also comprise a significant portion of total assets. At the end of our sample, stock holdings are the major source of non-US plans' asset allocation (34.7%), closely followed by fixed income (33.9%).

Figure B.2 shows the investment shares of sub-asset classes for non-U.S. plans. For stocks, we see an increase in the allocation to "Global" and "Emerging Market" equities. This is also true for "Global" fixed income allocations. In private equity, we see increased portfolio weights on

limited buyouts venture capital. Since the Global Financial Crisis, we also see an increase in the allocation to private credit in lieu of mortgages. In the real asset class, allocations to infrastructure increase whereas there is a divestment from real estate.

In summary, stocks and fixed income account for more than 90% of the total value of pension plans' asset holdings in the early nineties. This share has declined to about 70% at the end of our sample, with real assets, private equity and hedge fund investments accounting for most of the increased allocation to alternative asset classes. While stocks and bonds thus remain by far the most important asset classes, alternative assets are clearly gaining significant ground, having nearly tripled their share of pension plans' portfolios from roughly 10% to close to 30% during our nearly 30-year sample.

### B.2.4 Asset Management Mandate

In each sub-asset class, the AUM of a sponsor are managed according to their *asset management mandate* (or style). CEM provides information at the sub-asset class level about the following management mandates:

- Internally managed: the buy-sell decisions for the underlying assets (e.g., individual stocks, bonds, property) are made within the organization. This also includes wholly-owned subsidiaries.

- Externally managed: the buy-sell decision for the underlying assets are made by third-party entities, such as money managers.

- Passively managed (or indexed): designed to either replicate broad capital market benchmarks (e.g., the S&P 500 for U.S. stocks) or dedicated to matching liability requirements.

- Actively managed: assets given to an external manager to manage according to a set of objectives and constraints.

- Limited partnerships: investments in funds with a predetermined lifespan, where assets are sold, and invested capital is returned upon reaching the investment horizon.

- Co-investments: minority investments directly made into an operational company in conjunction with a financial sponsor or another private equity investor, typically in the context of a

leveraged buyout, recapitalization, or growth capital transaction.

- Fund of Funds: Investments in funds whose holdings consist primarily of other funds.

**Empirical Evidence**

Table B.4 shows how small (bottom $30^{\text{th}}$ percentile in AUM) and large (top $30^{\text{th}}$ percentile in AUM) plans allocated sub-asset classes to the four management styles in 2009 and thus complements Table 2.1 in the main text that shows similar evidence for 2019. External active management is dominant for small plans, particularly in the private asset classes but also for most sub-asset classes in stocks and fixed income. For stocks and fixed income, some plans also use external passive management, particularly for ACWI ex U.S., Other, U.S. Broad stocks, and inflation indexed and long bonds. In contrast, large plans use internal allocation far more often than the smaller plans. This holds both among stock and fixed income investments and involves both internal active and internal passive management. Among the holdings in the private asset classes, internal active management plays an important role for the private equity "other" assets, mortgages (private debt), commodities, infrastructure, real estate and REIT investments.

In results not reported here, we find that, across all asset classes, non-U.S. plans manage a significantly higher portion of their investments internally compared to their U.S. peers. Differences are particularly large for fixed income, private debt, and real assets in which the proportion of internally managed assets for non-U.S. plans exceeds that of U.S. plans by more than 20%.

In some cases, plans use multiple investment management styles to allocate their holdings within a particular sub-asset class. For those plans that do not adopt a single management style for all of their holdings in a particular sub-asset class, Table B.5 shows the allocation share to the six possible pair-wise combinations of the four investment management styles. The table covers only the largest plans because this usage of multiple investment management styles within a single sub-asset class is extremely uncommon among smaller plans. For stock accounts, combinations of external active and external passive as well as combinations of external active and internal active management are the most widespread pairs, but external active combined with internal passive management is also not uncommon. Among fixed income investments as well as investments in the private asset classes, combinations of external active and internal active management is the most common pairing.

Table B.6 shows statistics on the number of sub-asset classes per plan/year that are internal and external actively managed. We find that external active management is more common than internal active management. Additionally, on average, a greater amount of AUM is allocated to internal active management in comparison to external active management. These trends are consistent across all asset classes, and lend support to the hypothesis that only big plans have the expertise and resources to set up internal teams. Furthermore, internal management tends to be utilized exclusively for a select few specialized sub-asset classes. Figure B.3 shows a bar chart of the number of sub-asset classes that are actively managed, either internally or externally by a specific plan. For stocks and fixed income, the number of sub-asset classes that are internally managed is always lower than the externally managed assets, with the exception of plans that invest in a single fixed income sub-asset class.

### B.2.5 Asset Allocation and Size: Nonparametric Estimates

Our panel regressions in equation (2.4.9) of the main text assume a linear relation between plans' asset allocation and their AUM. To avoid invalid inference due to possible model misspecification and examine how good an approximation the linear model provides, we adopt a nonparametric approach that allows for a more flexible specification of the relation between a plan's weight in asset class $A$ at time $t$, $\omega_{iAt}$ and plan characteristics, $x_{iAt}$:

$$\omega_{iAt} = \theta(\widetilde{x}_{iAt}) + \epsilon_{iAt}, \tag{B.2.1}$$

where $\widetilde{x}_{iAt} \coloneqq x_{iAt} - (1/N)\sum_i x_{iAt}$ denotes the vector of cross-sectionally demeaned plan characteristics and $\theta(\cdot)$ is an unknown function of plan characteristics. We apply cross sectional demeaning to deal with potential time fixed effects such as trends. To estimate the unknown function $\theta(\cdot)$, we use the pooled kernel estimator

$$\widehat{\theta}(x) = \left[\iota^\top W_H(x)\iota\right]^{-1} \iota^\top W_H(x)\Omega_A, \tag{B.2.2}$$

where $W_H(x)$ is a weighting matrix with bandwidth $H$ and $\Omega_A$ stacks plan-level asset allocations $\omega_{iAt}$ in an $(n\sum_i^n T_i \times 1)$ vector, with $n$ denoting the number of plans and $T_i$ the number of time

series observations of plan $i$.[4]

Figure B.4 shows the nonparametric weight estimates for the individual asset classes as a function of the lagged value of log AUM. The relation is declining for stocks, fixed income, hedge funds/multi assets and private debt, whereas we find an increasing relation for private equity and real assets. All of this is consistent with the linear regression estimates from Table 2.5. Specifically, stock holdings decline from 53% to 48% as we move from small to large plans. Similarly, fixed income allocations decline from 37% for the smallest plans to 33% for the largest plans, consistent with large plans choosing to hold a greater fraction of their investments in alternative asset classes.

The plots in Figure B.4 show only mild deviations from linearity. A particularly critical form of misspecification from the linear modeling assumption in our panel regressions would be the presence of a non-monotonic relation between plan size and AUM allocations. To test more formally whether the relations in Figure B.4 are monotonic, we use the monotonic relation test of Patton and Timmermann (2010). The monotonic relation is specified to be either positive ("+") or negative ("–") as specified for the different asset classes in Table B.7. Under the null hypothesis, there is no positive (rep. negative) monotonic relation between plan size and allocation to a given asset class. Conversely, there is a monotonic relation between lagged plan size and allocation to a given asset class under the alternative. For example, if we specify a negative (decreasing) relation under the alternative, small $p$-values indicate that larger plans allocate a smaller amount of their investments to a given asset class.

To implement the test, each year ($t$) we sort plans by AUM, keeping only those plans that also report holdings the following year ($t+1$). We then form equal weighted quartile portfolios for the size-sorted plans going from the smallest to the largest plans. We conduct these tests only for those asset classes for which we have a sufficient number of observations, leading us to drop private debt. The results are reported in Table B.7. We find significant evidence of a monotonically decreasing relation between plan size and allocations to stocks and hedge funds and multi-asset mandates. Furthermore, the test also provides evidence for a monotonically increasing relation between plan size and allocations to private equity and real assets. Only for fixed income do we fail to reject the null of no monotonic relation between plan size and allocation.

[4]Our analysis uses the product kernel of a standard normal density and picks the bandwidth for each covariate as $h = b\widehat{\sigma}_x n^{-1/6}$, where $\widehat{\sigma}_x$ is the sample standard deviation of $\widetilde{x}_{iAt}$ and $b$ is a tuning parameter (we set $b = 2$).

## B.3 Cost Data

CEM collects detailed cost data at the sub-asset class level. In general, all costs—internal and external to the pension plan—related to management of plan assets are included in the survey.[5]

### B.3.1 Cost Components

We list the various cost components that a plan reports to CEM below:

**Internal investment costs**

- Compensation, benefits and direct expenditures associated with the staff overseeing internal portfolios. If staff is responsible for multiple asset classes, the cost is split according to the estimated time allocation
- Consulting, research, legal, trading systems and other third party costs.
- General operating expenses, including rent, utilities, IT services, investment accounting, financial control, and human resources. These costs are also allocated based on usage.

**External investment costs**

- Base fees remitted to third-party managers including investment management fees, manager-of-manager fees, commitment fees and fees netted from returns.
- Performance fees paid to (third-party) managers.
- Costs associated with balanced mandates, proportionally allocated based on actual holdings.
- Compensation, benefits and direct expenses for staff members primarily responsible for selecting, monitoring, and overseeing external managers.
- Third-party investment management fees prior to any deductions for rebates. These rebates constitute the limited partners' portion of specific fee income realized by the partner in connection with the fund, such as fees related to break-ups, monitoring, and funding.

**Limited partnership costs**

[5]For our empirical results, we proxy the plan's cost by average cost relative to AUM in a specific sub-asset class and year. This measure of cost also includes performance fees.

- Unreturned Invested Capital: Contributed capital less contributed capital attributable to realized investments less the aggregate amount of write-downs, if any, with respect to unrealized investments. This is often the amount on which fees are based after the investment period ends.

- Percentage fee on unreturned invested capital (post investment period): Private equity management fees are typically paid as a percentage of the committed amount during the investment period and as a percentage of unreturned invested capital after the investment period ends.

- Rebate percentage: the limited partners' share of certain fee income realized by the General Partner in connection with the fund such as fees for break-up, monitoring and funding.

**Other expenses**

- Oversight of the fund, including expenses such as staff salaries, direct costs (e.g., travel, director fees, director's insurance, etc.), and unallocated overhead related to the supervision of fund assets.

- Trustee and custodial costs.

- Consulting costs for manager searches, scenario testing, system consulting, and internal or external costs for performance measurement.

- Legal fees related to the entire fund which includes, among others, fiduciary insurance and printing.

- Fund of Funds Costs: top-layer management fees levied by the fund-of-funds manager as the manager base fees. It also includes the expenses incurred in the underlying funds. In cases where this data is unavailable, CEM applies a standard default.

### B.3.2 Variation in Costs by Investment Management Mandate

Investment management mandate is a key determinant of costs, but there is considerable heterogeneity in how much individual plans pay in fees. We present several figures that illustrate this heterogeneity. As in the main text, we scale all cost figures by the grand average cost, averaged

across plans, asset classes and years. Hence, all cost are expressed in percentage units relative to the average cost in our sample.[6]

We begin by presenting box plots in Figure B.5, with the median and interquartile range of 2019 plan-level costs for public asset classes and the four management mandates represented in our sample, scaled (to maintain proprietary data confidentiality required by CEM) by the grand-average cost, i.e., costs averaged across asset classes, across plans, and over time. For both stocks and fixed income, the cost ranges are low and narrow for passively managed accounts (IP and EP). Internal active (IA) management costs are a little higher, on average, than passive management fees and slightly more dispersed among stock and fixed income accounts. Median costs grow notably bigger, and cost ranges wider, for external actively (EA) managed accounts, which charge far higher fees than all other account types. We note (in unreported tests) that this holds across all sub-asset classes and throughout our sample.

Figure B.6 presents box-and-whisker plots displaying how total costs evolve over time for the active, passive, internal, and external management styles. Costs are aggregated across asset classes on a value-weighted basis. The scaled median internal management cost (top left panel) fluctuates around 14% of average costs with no discernible trend. Internal management costs are very homogeneous across plans. For example, the 95$^{\text{th}}$ percentile of scaled internal management costs is at most 62% of average costs.

In sharp contrast, scaled median external management costs are trending up starting at around 100% in 1999 to around 133% of average costs in 2019. Differences in external management costs across plans are also far higher than what we see for internal management costs with 95% bands ranging from 22% to nearly 450% of average costs towards the end of our sample.

Median passive management costs have declined modestly from around 18% of average costs in the early sample to close to 9% of average costs per year at the end. Differences in passive management costs across plans are also very modest, with the 95% bands ranging from 2% to 22% of average costs at the end of the sample.

In contrast, median active management costs rise from close to two-thirds of average costs in the early sample to about 150% of average costs at the end of the sample. The spread in active

[6]We implement this scaling to preserve confidentiality of the cost levels. However, this transformation of costs still allows us to compare cost across different asset classes and years.

investment management costs is also very large, with the 95% confidence band going from close to zero to nearly 400% of average costs at the end of the sample.

The plots in Figure B.6 show management costs aggregated across different asset classes whose weights are shifting over time. To isolate the impact of shifts in the weights of individual asset classes, Figures B.7–B.8 plot investment management costs for individual asset classes segregated by internally vs. externally managed assets and passively vs. actively managed assets. Hence, these plots show both the time-series variation and the degree of heterogeneity in management costs by asset class and management style.

We begin by examining equity investment cost. In most years, plans' passive, internal management costs for stocks amount to less than 7% of average costs, whereas internal active costs are somewhat higher, varying in the range 10-45% of average costs. In both cases, there is no discernible time-series trend in public market investment management costs. Fees for externally-managed stock portfolios (right panels in Figure B.7) are notably higher, with a (scaled) median annual cost that varies between 9% and 22% of average costs for passive management and active management fees between 56% and 130% of average costs. Overall, we find a far greater degree of variation in the costs of externally managed stock portfolios than for internally managed ones.

Figure B.8 shows similar plots for fixed income investments. For internal passively managed fixed income portfolios, median costs fluctuate between 4% and 9% of average costs, with three-quarters of plans paying less than 11% of average costs in most years and always less than 18%. The costs for actively managed internal portfolios are similar. The costs of externally managed fixed income portfolios fluctuate at a higher level, around 11% of average costs for passive portfolios, and 44% for actively managed portfolios. Again, a trend in these fees is notably absent with year-to-year variation more likely to reflect shifts in the composition of our sample of plans.

**Variations at the Sub-asset Class Level**

Figure B.9 provides further granularity by plotting median costs for the most important sub-asset classes in our sample. First consider management costs for U.S. Large and Small cap stocks. Median passive management costs are declining over time whereas active costs for internal and external management are steady around 67% of average costs for active large cap and 11% of average costs for internal management. Median external active management costs for small cap

portfolios are around 130% of average costs versus 44% for internal active management.

Median costs for passively managed EAFE mandates converge to approximately 11% of average costs and we see a similar trend for passive management of Broad stocks whose median cost converges to around 7% of average costs. Median active management costs have maintained their gap between external and internal management of about 67% of average costs. Median internal management cost for EAFE is around 44% and 22% of average costs for Broad stocks. External active management costs for EAFE mandates amount to 110% and 90% of average costs for Broad stocks.

Median internal passive costs for U.S. fixed income fluctuate between 2% and 4% of average costs in most years. External passive management costs start considerably higher but trend downward, converging towards internal passive management costs at the end of the sample. Median internal active management costs are around 11% of average costs versus 44% for external active management costs.

The last panels in Figure B.9 show that median internal passive management costs for Canadian fixed income have been fluctuating around 11% of average costs, and external passive management costs converged to 11% of average costs toward the end of our sample. Median internal active management costs is around 11% of average costs without any considerable trend; external active management start above 44% and decrease to approximately 38% of average costs.

Supplementing these figures, Table B.8 shows regressions of costs (in bps) by sub-asset class on dummies such as external, and active. Across all sub-asset classes, external investment management is significantly costlier than internal management and active management is costlier than passive management. External investment management is disproportionately costly in the private sub-asset classes and for specialized sub-asset classes such as emerging market stocks and bonds and high yield bonds.

Table B.8 does not control for plan size. To highlight the importance of plan size, Table B.9 presents regression results of the power law in investment costs at the sub-asset class level, as discussed in Equation (2.5.2) in the main text. The table shows that economies of scale are higher (lower $\beta$ estimates) for passively managed EAFE and U.S. broad stocks, and for inflation-indexed bonds. For the alternative asset classes we find lower scales of economy for the cost of managing diversified private equity, real estate, and REITS.

### B.3.3 Management Costs by Country of Domicile

Investment management costs depend not only on investment management style and asset class but also on country-of-domicile for the investment plan. To illustrate this, Table B.10 presents plans' mean cost per asset class by country-of-domicile in 2009 and 2019, again measured relative to the grand average cost figure. Across countries and at both points in time, management costs are lowest for fixed income balances, followed by stock portfolios. Private credit and real asset accounts fall in the middle in most countries with hedge funds and, particularly, private equity management costs being much higher.[7] The table also shows interesting geographical variation in costs, with surprisingly similar costs of managing stocks and fixed income assets in the U.S., Canada, and the Netherlands and relatively low costs of managing public assets in countries such as Australia and Sweden. The cost of managing private equity and hedge fund assets is quite similar across domiciles, while conversely we see bigger geographical differences in the cost of managing real assets, probably due to the very heterogeneous nature of this asset class.

## B.4 Returns, Benchmarks and Risk-Adjustments

### B.4.1 Benchmarks

The CEM database contains a detailed list of returns, policy weights and return benchmarks, all available at the sub-asset class level. We state the definition of these variables below.

- Returns: Actual full-year returns for a specific sub-asset class. Returns are categorized as gross returns and net returns (net of cost).
- Policy weights: Weights that reflect plans' long-term policy, normal or target asset mix such as 60% stocks and 40% bonds. Policy weights add to 100% and are provided at year-end levels.
- Benchmarks: Broad investable capital market indexes (for example, the S&P500 for U.S. stocks) used to gauge asset class performance. If multiple benchmarks apply for an asset class, each benchmark is weighted accordingly (e.g., 60% S&P 500 and 40% Russell 3000). Our data sample contains a total of 15,101 different policy benchmarks, which also includes

[7] As we show in the paper, these broad cost estimates conceal a lot of variation related to changes in investment management styles (active versus passive, external versus internal).

some esoteric benchmarks tailored to specialized investments such as the Dow Jones Brookfield Global Infrastructure Index or the KOSPI 3-year average return.

- Total policy return: Returns that track the policy mix and/or benchmark changes through the year.

### B.4.2 Asset Class Return Performance

Table B.11 reports summary statistics for gross-of-fee returns grouped by asset class, averaged across all plan-years. Over our sample period, private equity holdings earned the highest mean return (15.9% per annum), followed by stock holdings (10.8%), real assets (8.4%), and private debt (7.8%). Hedge funds & multi assets (7.1%) and fixed income (7.0%) earned the lowest average sample returns. Volatility estimates, reported on the diagonal of Panel B in Table B.11, show that private equity is by far the most volatile asset class (22.3% per annum), followed by stocks (16.3%), real assets (11.6%), hedge funds (10.5%) and private debt (9.7%). Fixed income holdings, unsurprisingly, record the lowest volatility (6.9%).

While expected return performance is clearly an important driver of plans' asset allocation decisions, it is by no means the only explanation for the increased importance of alternative asset classes over our sample. The possibility of reducing portfolio-level return volatility by diversifying across asset classes has also been a key determinant of these decisions.

To better understand the extent to which pension plans gained from diversification across asset classes, Panel B in Table B.11 reports the average correlation across our six asset classes. Stock returns are positively correlated with returns on all other asset classes and have the lowest correlation with real assets (0.201) and fixed income (0.278) and the highest correlation with hedge funds and multi assets (0.858). Fixed income returns, on the other hand, are negatively correlated with returns on both real assets and private equity, though insignificantly so. The correlation between fixed income returns and returns on hedge funds and multi asset (0.547) or returns on private debt (0.598) is much stronger.

These correlation estimates are sufficiently low to imply clear diversification benefits from adding alternative asset classes to the plans' public asset holdings, with the possible exception of hedge funds and multi assets whose returns were highly correlated with both stock and fixed income returns during our sample.

### B.4.3 Risk Adjustment Regressions

An alternative to studying policy-adjusted returns is to correct for plans' return exposures to a small set of the most important risk factors. With less than 30 annual return observations per asset class, we need to choose the risk factors judiciously, in many cases eliminating factors whose coefficient estimates are insignificant. In particular, we consider the following risk factors for the individual asset classes:

- Stocks: The Fama and French (1993) three factor model: market excess return (Market), small minus big (SMB), and high minus low book-to-market ratios (HML).
- Fixed Income: U.S. Aggregate Bond Index, U.S. Corporate Index, U.S. High Yield Index, Global Diversified Index, U.S. Long Treasury (1–3 years).
- Hedge Funds and multi asset: The seven factor model of Fung and Hsieh (2001) which includes the market excess return (Market), a bond trend, currency trend, commodity trend, size spread, bond market and credit spread factor.
- Private Equity, Private debt and Real assets: A subset of the seven factor model that includes the market excess return, size spread, and bond factor. For these asset classes we also include lags of each factor to account for staleness in returns.

### B.4.4 Construction of risk factors

We next describe the construction of the Fung and Hsieh (2001) risk factors that are used in Section 2.6.2 of the main text. Three factors are obtained from Hsieh's website:[8] the Bond trend-following factor, the Currency trend-following factor and the commodity trend-following factor. In addition, we construct the following factors ourselves (following instructions on Hsieh's website):

- Equity Market Factor: Constructed using monthly S&P 500 returns.
- Size spread factor: Russel 2000 index monthly return - S&P500 monthly return.
- Bond Market Factor: Monthly changes in the 10-year treasury constant maturity yield (month end-to-month end).[9]

[8] https://faculty.fuqua.duke.edu/~dah7/HFRFData.htm
[9] Available at the Federal Reserve Bank of St. Louis: https://fred.stlouisfed.org/series/DGS10

- Credit Spread Factor: Monthly changes in Moody's Baa yield less 10-year treasury constant maturity yield (month end-to-month end).[10]

All series are annualized from their underlying monthly values. Since the risk factors are most appropriate for U.S. plans, we only construct portfolios based on U.S. plans.[11]

### B.4.5 Factor regression results

The "portfolio" columns in Table B.12 show estimates from regressing equal-weighted asset-class returns on the risk factors as in Equation (2.6.5) of the main text.[12] For the stock portfolio, the market excess return factor obtains a highly significant loading of 0.95 which is close to unity, both in an economic and statistical sense.[13] The size factor is also significant but the coefficient is an order of magnitude smaller than the market factor. The book-to-market factor is insignificant. Overall, these three factors generate an $R^2$ of 0.95, suggesting that most of the time-series variation in plans' (aggregate) stock returns is explained by the market factor. At -0.69%, the average plan alpha is negative but statistically insignificant.

For the fixed income portfolios, the Bloomberg U.S. Aggregate Bond Index, a credit risk factor and a term structure variable all generate highly significant and positive estimates. The time-series $R^2$ (0.97) is even higher for the fixed income portfolio than for the stock portfolio (0.95). After adjusting for these risk factors, the average fixed income portfolio generates a positive and statistically significant alpha of 65 bps.

For hedge funds and multi asset mandates, the market, size spread factor and bond market factors obtain statistically significant coefficients which explain 92% of the time-series variation in average returns. For the private equity portfolio, the market equity excess return and its lagged value both obtain significant coefficients as does the concurrent bond market factor. These factors explain 74% of the variation in returns. Finally, risk factors explain a notably smaller fraction of the time-series variation for private credit and real assets with $R^2$ values of 0.40 and 0.44, respectively. For these asset classes, only the equity market return or its lag generate statistically significant

[10] Available at the Federal Reserve Bank of St. Louis: https://fred.stlouisfed.org/series/DBAA and https://fred.stlouisfed.org/series/DGS10

[11] The risk factor regressions for the alternative asset classes that use the Fung and Hsieh (2001) factors have fewer time series observations since the factor data only go back to 1994.

[12] These regressions use excess returns net of costs, but the results are nearly identical if instead we use gross excess returns.

[13] Betas on the market return significantly higher than unity would be consistent with plans applying leverage.

coefficients, in both cases with values that are quite small (0.22 and 0.27, respectively). Average alpha estimates for the alternative asset classes tend to be greater in absolute terms, though only statistically significant for one of the four alternative asset classes.

The "pooled" columns in Table B.12 show estimates of the factor loadings using pooled panel regressions on individual plan-year observations. The risk factors retain even stronger statistical power over individual plans' returns in the pooled panel regressions. Unsurprisingly, however, the explanatory power of the risk factors over individual plans' returns is somewhat lower than for the aggregate regressions. This is to be expected because of idiosyncratic variation in individual plans' returns around their benchmarks due to active management.

These results suggest that traditional risk-adjustment methods work particularly well for the two most liquid asset classes (stocks and fixed income) as well as for hedge funds, but do a worse job at tracking performance in the most illiquid asset classes in our sample such as private equity, private debt, and real assets.

### B.4.6 Factor exposures in policy-adjusted returns

In Figure B.10, we show box and whisker plots of the policy-adjusted gross returns. These are roughly centered around zero for all asset classes. Our approach of risk-adjusting returns by subtracting the plan-specific policy benchmark returns (Figure B.11) can be criticized on the grounds that some plans could earn abnormal returns by deviating from their policy targets. To address this concern, we next examine whether significant exposures to systematic risk factors remain after subtracting policy returns from plan returns.

To obtain meaningful estimates, we require that the plans have ten or more annual return observations, and we limit our analysis to stocks and fixed income. Moreover, we only include a single risk factor to reduce the number of parameters estimated for each asset class. Our single-factor regressions for individual plans' policy-adjusted returns thus take the form

$$\widetilde{r}_{iAt} = \alpha_{iA} + \beta'_{iA} F_{At} + \epsilon_{iAt}. \tag{B.4.1}$$

For stocks, we use the market excess return while for fixed income portfolios we use the Bloomberg U.S. aggregate bond Index excess return as the single risk factor. Table B.13 summarizes the

results. Across 199 plans with the required number of observations on stock returns, the mean and median values of $\beta_{iA}$ are -0.0003 and 0.003, respectively, while the mean and median values of $\alpha_{iA}$ are both 0.003. Interquartile ranges are quite narrow: -0.001 to 0.0082 (alpha estimates) and -0.023 to 0.0252 (beta estimates).

For fixed income, we find similar results: Across 203 plans with at least 10 annual observations, the mean and median estimate of $\alpha_{iA}$ is 0.003 and 0.002 (interquartile range of -0.0013 to 0.0064), respectively, while the mean and median estimate of $\beta_{iA}$ is -0.004 and 0.002 (interquartile range of -0.0929 to 0.1038).

These results show that the policy-adjustment procedure succeeds in capturing the vast amount of systematic risk exposures in plans' returns and that the plans choose market risk exposures that are very close to those laid out in the policy benchmarks for both stock and fixed income holdings.

### B.4.7 Performance in Sub-Asset Classes

To help further pin down the relation between variables such as AUM and investment performance in different markets, we estimate panel regressions of policy-adjusted net returns on individual sub-asset classes. We only consider those sub-asset classes for which we have sufficiently many observations to obtain reasonably precise parameter estimates. Table B.14 presents results for a set of sub-asset classes chosen on the basis that they have at least 1,000 observations.

Our estimates show that a plan's log-AUM is significantly positively related to policy-adjusted net returns for the EAFE, U.S. Large Cap, Global, and Emerging categories but fails to be significantly related to stock investments such as U.S. small cap or ACWI ex U.S. Moreover, the economic effect of AUM can be quite large: the increases in average annual returns associated with moving from a plan in the 10th percentile to a plan in the 90th percentile of holdings in a given sub-asset class are 50 bps (EAFE), 54 bps (U.S. Large Cap), 67 bps (Global), and 62 bps (Emerging).

Examining the log-AUM coefficient estimates more closely, we see that they are bigger for the net return regressions than for gross returns for all sub asset classes with exception of U.S. Small caps. This suggests that the largest plans' better performance in these sub-asset classes, as compared to their smaller peers, is, at least in part, driven by their ability to reduce costs.

We also find a significant relation between log-AUM and policy-adjusted *net* returns for fixed income (Canada, Global, Inflation Indexed), hedge funds, private equity (Diversified private equity, Other) and real assets (REITs). The coefficients for Hedge funds, Diversified private equity and Other private equity are very large (0.46, 0.97 and 1.29 respectively), so that moving from a plan in the tenth percentile to a plan in the 90th percentile of the (2019) size distribution in these sub-asset classes is associated with increases in mean returns of 192 bps, 403 bps, and 537 bps, respectively. Once again, the coefficient estimates on log-AUM tend to be notably higher for net returns than for gross returns, consistent with bigger cost savings for the largest plans also in these sub-asset classes.

Although many of the estimates on the private plan dummy are quite large and positive, we only find two instances (U.S. Broad or All stocks and Diversified private equity) for which private plans appear to produce policy-adjusted net returns whose means are significantly different from those of public plans.

## Tables

**Table B.1. Number of participants per asset class and year (Panel A) and by frequency of participation (Panel B)**. Panel A presents the total number of observations (plans) per asset class and year. Panel B presents the time series frequency of unique plans in the CEM database.

Panel A: Total number of observations by asset class and year

| | Stocks | Fixed Income | Hedge & multi ass. | Private Equity | Private Debt | Real Assets | Total Public | Total Private | Total |
|---|---|---|---|---|---|---|---|---|---|
| 1991 | 122 | 122 | 17 | 69 | | 100 | 33 | 90 | 123 |
| 1992 | 163 | 162 | 30 | 85 | 2 | 129 | 31 | 132 | 163 |
| 1993 | 216 | 216 | 34 | 112 | 2 | 160 | 55 | 164 | 219 |
| 1994 | 265 | 267 | 40 | 137 | 6 | 201 | 76 | 192 | 268 |
| 1995 | 294 | 297 | 49 | 139 | 9 | 223 | 96 | 201 | 297 |
| 1996 | 292 | 295 | 44 | 134 | 10 | 210 | 91 | 204 | 295 |
| 1997 | 271 | 272 | 32 | 130 | 10 | 201 | 95 | 177 | 272 |
| 1998 | 285 | 285 | 28 | 140 | 12 | 201 | 99 | 186 | 285 |
| 1999 | 304 | 304 | 25 | 144 | 16 | 207 | 113 | 191 | 304 |
| 2000 | 284 | 285 | 29 | 148 | 17 | 203 | 111 | 174 | 285 |
| 2001 | 293 | 293 | 42 | 154 | 18 | 201 | 116 | 177 | 293 |
| 2002 | 273 | 273 | 56 | 145 | 18 | 184 | 107 | 166 | 273 |
| 2003 | 277 | 277 | 68 | 149 | 21 | 191 | 107 | 170 | 277 |
| 2004 | 285 | 285 | 78 | 158 | 28 | 210 | 107 | 178 | 285 |
| 2005 | 297 | 298 | 91 | 157 | 37 | 218 | 115 | 183 | 298 |
| 2006 | 289 | 289 | 105 | 161 | 36 | 218 | 109 | 180 | 289 |
| 2007 | 354 | 356 | 150 | 213 | 43 | 266 | 121 | 235 | 356 |
| 2008 | 334 | 337 | 156 | 209 | 41 | 261 | 113 | 224 | 337 |
| 2009 | 334 | 335 | 157 | 215 | 36 | 257 | 113 | 222 | 335 |
| 2010 | 346 | 346 | 172 | 226 | 45 | 267 | 118 | 228 | 346 |
| 2011 | 373 | 374 | 206 | 252 | 56 | 313 | 113 | 262 | 375 |
| 2012 | 446 | 445 | 253 | 304 | 80 | 380 | 202 | 246 | 448 |
| 2013 | 443 | 443 | 265 | 308 | 97 | 380 | 199 | 247 | 446 |
| 2014 | 420 | 419 | 255 | 294 | 103 | 367 | 204 | 218 | 422 |
| 2015 | 359 | 360 | 209 | 267 | 109 | 314 | 146 | 215 | 361 |
| 2016 | 343 | 345 | 204 | 256 | 113 | 303 | 143 | 202 | 345 |
| 2017 | 347 | 350 | 200 | 260 | 144 | 311 | 152 | 198 | 350 |
| 2018 | 331 | 334 | 196 | 249 | 152 | 303 | 145 | 189 | 334 |
| 2019 | 305 | 308 | 176 | 236 | 159 | 281 | 134 | 174 | 308 |

Panel B: Total number of plan count by frequency of observation

| # of obs | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Plan Count | 240 | 124 | 134 | 65 | 54 | 39 | 45 | 59 | 29 | 31 |
| # of obs | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| Plan Count | 21 | 24 | 29 | 21 | 22 | 15 | 16 | 17 | 16 | 17 |
| # of obs | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | |
| Plan Count | 15 | 7 | 12 | 18 | 14 | 9 | 15 | 12 | 17 | |

**Table B.2. AUM allocation by asset class in 2009 and 2019**. This table shows total AUM allocated to Stocks, Fixed Income and Other assets in billions of USD for all countries in the CEM database. Other assets bundles the asset classes: Private Equity, Private Debt and Real Assets. AUM (%) denotes the share of total AUM per country, which is defined by $\text{Share}_{At} = \sum_i \text{AUM}_{iAt} / \sum_i \sum_A \text{AUM}_{iAt}$, where $\text{AUM}_{iAt}$ denotes total AUM of plans in country $i$ in asset class $A$ in year $t$.

| Year | 2009 | | | | 2019 | | | |
|---|---|---|---|---|---|---|---|---|
| | Stocks | Fixed Income | Other assets | AUM(%) | Stocks | Fixed Income | Other assets | AUM(%) |
| U.S. | 1132.77 | 801.62 | 501.33 | 57.62 | 1525.22 | 1236.51 | 1032.26 | 42.17 |
| Canada | 254.81 | 203.64 | 164.88 | 14.73 | 497.31 | 371.9 | 767.06 | 17.82 |
| Australia | 32.93 | 29.6 | 12.42 | 1.77 | 77.01 | 52.13 | 40.86 | 1.89 |
| Belgium | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| China | 0 | 0 | 0 | 0 | 101.86 | 48.7 | 108.67 | 2.88 |
| Denmark | 4.35 | 15.67 | 3.29 | 0.55 | 0 | 0 | 0 | 0 |
| Emirates | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Finland | 30.68 | 53.16 | 21.85 | 2.5 | 60.18 | 55.5 | 50.08 | 1.84 |
| France | 16.02 | 14.83 | 1.49 | 0.76 | 0 | 0 | 0 | 0 |
| Germany | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Netherlands | 146.67 | 243.74 | 132.29 | 12.35 | 447.34 | 615.54 | 349.73 | 16.4 |
| New Zealand | 4.1 | 5.5 | 4.24 | 0.33 | 17.99 | 6.51 | 7.18 | 0.35 |
| Other USD | 0.08 | 0.04 | 0.02 | 0 | 44.21 | 25.85 | 11.5 | 0.9 |
| Saudi Arabia | 0 | 0 | 0 | 0 | 12.13 | 8.77 | 7.86 | 0.32 |
| South Africa | 0 | 0 | 0 | 0 | 15.89 | 4.51 | 2.48 | 0.25 |
| South Korea | 51.52 | 199.14 | 10.81 | 6.18 | 228.25 | 304.15 | 69.26 | 6.65 |
| Sweden | 37.89 | 44.48 | 8.34 | 2.14 | 68.65 | 54.33 | 34.75 | 1.75 |
| Switzerland | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| UK | 29.03 | 8.1 | 7.26 | 1.05 | 223.84 | 207.7 | 177.85 | 6.77 |

**Table B.3. Aggregate Asset Allocation for U.S. (Panel A) and non-U.S. (Panel B) plans**. This table shows the share of total AUM dedicated to each of the six asset classes during each of the years in our sample: $\omega_{At} = \sum_i \text{AUM}_{iAt} / \sum_i \sum_A \text{AUM}_{iAt}$, where $i$ indicates plans, $t$ indicates year, $A$ indicates the asset class, estimated separately for U.S. and non-U.S. plans.

| | Panel A: U.S. Plans | | | | | | Panel B: Non-U.S. Plans | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Stocks | Fixed Income | Hedge & multi ass. | Private Equity | Private Debt | Real Assets | Stocks | Fixed Income | Hedge & multi ass. | Private Equity | Private Debt | Real Assets |
| 1991 | 53.63 | 38.76 | 0.73 | 1.91 | | 4.98 | 35.81 | 57.64 | 0.43 | 1.14 | | 4.97 |
| 1992 | 54.09 | 37.91 | 1.33 | 1.82 | 0.09 | 4.76 | 36.86 | 56.32 | 0.31 | 1.69 | 0.49 | 4.34 |
| 1993 | 55.83 | 35.72 | 1.72 | 2.54 | 0.07 | 4.12 | 39.85 | 52.72 | 0.29 | 2.24 | 0.82 | 4.08 |
| 1994 | 54.25 | 37.03 | 1.42 | 2.79 | 0.07 | 4.43 | 40.97 | 46.66 | 0.25 | 1.75 | 4.50 | 5.87 |
| 1995 | 53.80 | 37.48 | 1.76 | 2.46 | 0.05 | 4.44 | 44.93 | 44.08 | 0.35 | 0.70 | 3.98 | 5.95 |
| 1996 | 56.46 | 35.30 | 1.18 | 2.64 | 0.03 | 4.39 | 49.84 | 39.23 | 0.40 | 1.15 | 3.32 | 6.06 |
| 1997 | 58.61 | 33.34 | 1.04 | 2.76 | 0.01 | 4.24 | 51.81 | 38.25 | 0.57 | 1.68 | 2.08 | 5.61 |
| 1998 | 60.32 | 31.76 | 1.11 | 2.68 | 0.01 | 4.13 | 52.01 | 37.67 | 0.59 | 1.90 | 2.05 | 5.78 |
| 1999 | 63.06 | 29.59 | 0.83 | 2.78 | 0.01 | 3.73 | 50.44 | 36.75 | 0.52 | 2.35 | 3.37 | 6.57 |
| 2000 | 61.05 | 29.75 | 0.84 | 4.04 | 0.02 | 4.30 | 49.01 | 36.77 | 0.66 | 3.25 | 3.39 | 6.91 |
| 2001 | 59.96 | 30.48 | 0.71 | 4.02 | 0.03 | 4.80 | 46.99 | 36.81 | 0.55 | 3.81 | 3.20 | 8.63 |
| 2002 | 58.84 | 31.11 | 0.72 | 3.98 | 0.03 | 5.31 | 46.63 | 34.68 | 0.83 | 3.83 | 2.57 | 11.47 |
| 2003 | 60.19 | 29.80 | 1.03 | 3.96 | 0.07 | 4.95 | 45.39 | 37.35 | 0.96 | 3.03 | 3.06 | 10.21 |
| 2004 | 62.54 | 27.58 | 1.35 | 3.62 | 0.11 | 4.81 | 46.15 | 37.52 | 1.51 | 2.98 | 2.30 | 9.53 |
| 2005 | 61.94 | 26.78 | 1.83 | 3.86 | 0.49 | 5.10 | 46.65 | 37.87 | 1.55 | 2.80 | 1.60 | 9.54 |
| 2006 | 60.90 | 26.28 | 2.34 | 4.17 | 0.74 | 5.57 | 47.04 | 35.68 | 2.37 | 2.94 | 1.37 | 10.59 |
| 2007 | 56.64 | 28.24 | 3.05 | 5.00 | 0.74 | 6.32 | 45.83 | 35.04 | 2.94 | 3.52 | 1.51 | 11.16 |
| 2008 | 48.83 | 31.93 | 3.76 | 7.09 | 0.66 | 7.71 | 39.65 | 38.39 | 3.52 | 4.60 | 1.78 | 12.06 |
| 2009 | 46.47 | 32.88 | 4.59 | 7.90 | 0.66 | 7.42 | 33.92 | 45.62 | 3.08 | 4.47 | 1.49 | 11.43 |
| 2010 | 48.74 | 31.17 | 4.42 | 8.14 | 0.66 | 6.75 | 37.47 | 41.51 | 3.10 | 5.15 | 1.31 | 11.45 |
| 2011 | 46.35 | 31.13 | 5.02 | 9.05 | 0.65 | 7.72 | 35.71 | 40.90 | 3.66 | 5.31 | 1.09 | 13.33 |
| 2012 | 44.42 | 31.46 | 5.33 | 9.32 | 0.73 | 8.32 | 37.43 | 38.73 | 4.38 | 5.28 | 1.13 | 13.05 |
| 2013 | 46.38 | 29.44 | 5.59 | 8.75 | 0.95 | 8.46 | 39.50 | 36.87 | 4.51 | 5.05 | 1.11 | 12.96 |
| 2014 | 46.16 | 29.57 | 5.89 | 8.37 | 1.06 | 8.57 | 40.34 | 35.88 | 4.85 | 5.25 | 1.05 | 12.64 |
| 2015 | 45.05 | 29.37 | 6.11 | 8.45 | 1.29 | 9.38 | 38.48 | 35.64 | 5.36 | 5.62 | 1.29 | 13.61 |
| 2016 | 43.85 | 29.40 | 6.03 | 8.58 | 1.47 | 10.32 | 38.04 | 34.93 | 5.00 | 6.15 | 1.46 | 14.41 |
| 2017 | 43.89 | 30.24 | 5.72 | 8.04 | 1.61 | 9.88 | 39.11 | 33.78 | 4.59 | 6.06 | 1.79 | 14.66 |
| 2018 | 41.74 | 31.18 | 6.09 | 8.56 | 1.80 | 10.04 | 36.02 | 33.96 | 4.93 | 7.04 | 2.11 | 15.94 |
| 2019 | 40.00 | 32.43 | 5.83 | 8.88 | 2.14 | 10.22 | 34.66 | 33.91 | 4.82 | 7.83 | 2.64 | 16.13 |

**Table B.4. Small and large plans' investment allocation by sub-asset class and management structure in 2009**. This table shows the share (in %) of AUM allocated to the four management mandates: Internal Passive (IP), External Passive (EP), Internal Active (IA), and External Active (EA) for the given sub-asset classes. The share is calculated as follows: $\omega_{ats} = \frac{\text{AUM}_{ats}}{\text{AUM}_{at}}$, where $\text{AUM}_{ats} = \sum_i \text{AUM}_{iats}$, and $\text{AUM}_{at} = \sum_s \sum_i \text{AUM}_{iats}$, where $i$ denotes plan $i$, $a$ indicates the sub-asset class, $t$ denotes the year 2009, and $s$ denotes one of the four mandates. The shares are calculated separately for small and large plans, defined by the bottom and top 30th percentile of AUM in 2009 respectively. For small and large plans, rows sum up to 100%.

| | Small Plans (in %) | | | | Large Plans (in %) | | | |
|---|---|---|---|---|---|---|---|---|
| **Stocks** | IP | EP | IA | EA | IP | EP | IA | EA |
| ACWI x. U.S. | | 39.93 | | 60.07 | | 16.78 | 7.26 | 75.96 |
| EAFE | | 8.97 | | 91.03 | 18.06 | 15.95 | 11.25 | 54.74 |
| Emerging | | 27.82 | | 72.18 | 10.76 | 4.82 | 12.57 | 71.85 |
| Global | | 11.41 | | 88.59 | 7.29 | 3.16 | 55.39 | 34.16 |
| Other | 7.79 | 92.21 | | | 16.94 | 0.76 | 39.98 | 42.32 |
| U.S. Broad | 0.89 | 45.84 | | 53.27 | 24.48 | 30.76 | 12.74 | 32.02 |
| U.S. Large Cap | | 29.26 | | 70.74 | 31.95 | 18.50 | 13.92 | 35.62 |
| U.S. Mid Cap | | | | | | | | |
| U.S. Small Cap | | 25.04 | | 74.96 | 13.19 | 14.65 | 7.73 | 64.43 |
| **Fixed Income** | | | | | | | | |
| Bundled LDI | | | | | | | | |
| Cash | | | 12.51 | 87.49 | | | 42.89 | 57.11 |
| Convertibles | | | | | | | | 100.00 |
| EAFE | | | | | | 13.05 | | 86.95 |
| Emerging | | | | 100.00 | | | 25.74 | 74.26 |
| Global | | | | 100.00 | 1.60 | 0.51 | 77.56 | 20.33 |
| High Yield | | | | 100.00 | 1.32 | 0.01 | 8.48 | 90.19 |
| Inflation Indexed | 39.55 | 42.23 | 1.71 | 16.51 | 30.87 | 5.74 | 35.43 | 27.97 |
| Long Bonds | 1.04 | 49.66 | 2.18 | 47.12 | 7.08 | 2.23 | 15.00 | 75.69 |
| Other | | 2.52 | | 97.48 | 80.74 | 0.23 | 8.93 | 10.10 |
| U.S. | | 17.73 | | 82.27 | 2.32 | 6.13 | 41.47 | 50.09 |
| **Hedge & multi ass.** | | | | | | | | |
| Funded TAA | | | | 100.00 | | | 1.08 | 98.92 |
| Hedge Fund | | | | 100.00 | | | | 100.00 |
| Risk Parity | | | | | | | | 100.00 |
| **Private Equity** | | | | | | | | |
| Div. Private Eq. | | | | 100 | | | 7.57 | 92.43 |
| LBO | | | | | | | 0.76 | 99.24 |
| Other | | | | | | | 38.68 | 61.32 |
| Venture Capital | | | | 100 | | | 0.08 | 99.92 |
| **Private Debt** | | | | | | | | |
| Mortgages | | | | 100 | | | 87.10 | 12.90 |
| Credit | | | | | | | 28.87 | 71.13 |
| **Real Assets** | | | | | | | | |
| Commodities | | | | 100.00 | 11.86 | 8.27 | 40.13 | 39.74 |
| Infrastructure | | | | 100.00 | | | 64.60 | 35.40 |
| Nat. Resource | | | | 100.00 | | | 15.76 | 84.24 |
| Other | | | 11.52 | 88.48 | | | 14.93 | 85.07 |
| Real Estate | | | 6.55 | 93.45 | | | 29.96 | 70.04 |
| REIT | | | | 100.00 | 5.22 | 1.55 | 48.25 | 44.98 |

**Table B.5. Plans' relative allocation to multiple investment mandates**. This table shows the 2019 allocation share to different pairs of management mandates for large plans that utilize more than one management style within the same sub-asset class. Large plans belong to the top 30th percentile by AUM. The total number of plans are indicated in parentheses and rows sum to 100%.

| Sub-Asset class | EA&EP | EA&IA | EA&IP | IA&IP | EP&IP | EP&IA |
|---|---|---|---|---|---|---|
| **Stocks** | | | | | | |
| ACWI X U.S. | 68.19 (10) | 5.12 (1) | 26.69 (3) | | | |
| EAFE | 70.94 (10) | 14.96 (3) | 14.09 (2) | | | |
| Emerging | 32.07 (19) | 34.10 (5) | 17.66 (4) | 16.17 (1) | | |
| Global | 16.75 (11) | 55.11 (13) | 18.26 (1) | 9.89 (1) | | |
| Other | | 100.00 (2) | | | | |
| U.S. Broad | 60.45 (12) | 9.70 (2) | 10.41 (2) | | 19.45 (2) | |
| U.S. Large Cap | 57.14 (6) | 9.73 (2) | 13.9 (4) | 19.23 (2) | | |
| U.S. Mid Cap | | | 100.00 (2) | | | |
| U.S. Small Cap | 27.43 (4) | 17.42 (4) | 41.01 (3) | 14.14 (2) | | |
| **Fixed Income** | | | | | | |
| Bundled LDI | | | | 100.00 (1) | | |
| Cash | | 100.00 (12) | | | | |
| Emerging | 2.95 (2) | 76.75 (10) | 7.62 (2) | 12.68 (2) | | |
| Global | | 36.49 (3) | 6.19 (1) | 54.82 (2) | 2.50 (1) | |
| High Yield | 11.73 (1) | 88.27 (9) | | | | |
| Inflation Index | 64.29 (4) | | 8.98 (1) | 2.09 (1) | | 24.64 (2) |
| Long | 25.47 (1) | 74.53 (3) | | | | |
| Other | | 7.11 (3) | 92.89 (1) | | | |
| U.S. | 58.71 (7) | 24.87 (4) | 3.21 (1) | 6.21 (1) | | 7.00 (1) |
| **Hedge & Multi Ass.** | | | | | | |
| Funded TAA | | 100.00 (3) | | | | |
| Risk Parity | | 100.00 (3) | | | | |
| **Private Equity** | | | | | | |
| LBO | | 100.00 (1) | | | | |
| Other | | 100.00 (4) | | | | |
| VC | | 100.00 (3) | | | | |
| Div. PE | | 100.00 (17) | | | | |
| **Private Debt** | | | | | | |
| Private Credit | | 100.00 (13) | | | | |
| **Real Assets** | | | | | | |
| Commodities | 11.77 (1) | 12.97 (3) | 11.20 (1) | 64.06 (2) | | |
| Infrastructure | | 100.00 (22) | | | | |
| Nat. Resource | | 100.00 (11) | | | | |
| Other | | 100.00 (1) | | | | |
| Real Estate | | 100.00 (33) | | | | |
| REITs | 31.14 (3) | 64.87 (3) | 3.99 (1) | | | |

**Table B.6. Frequency of internal and external active management.** This table shows the mode and the mean of how often each plan employs internal (IA) –and external active (EA) management for sub-asset classes in a given asset class for the years 1999, 2009 and 2019. The mode and the mean are calculated across plans within a given year and asset class. Avg. AUM denotes the average AUM (in millions U.S. dollar) allocated to IA or EA management within each asset class.

| | | **Stocks** | | | **Fixed income** | | | **Hedge & multi ass.** | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Year | Style | Mode | Mean | Avg. AUM | Mode | Mean | Avg. AUM | Mode | Mean | Avg. AUM |
| 1999 | IA | 1 | 1.70 | 2841.78 | 1 | 1.74 | 1933.32 | 1 | 1.00 | 911.38 |
| | EA | 3 | 2.77 | 726.49 | 2 | 2.00 | 560.07 | 1 | 1.00 | 662.12 |
| 2009 | IA | 1 | 2.23 | 2444.96 | 1 | 2.05 | 2458.06 | 1 | 1.00 | 69.89 |
| | EA | 3 | 3.41 | 781.43 | 2 | 2.51 | 901.38 | 1 | 1.46 | 728.44 |
| 2019 | IA | 1 | 2.44 | 3906.48 | 1 | 2.25 | 3292.93 | 1 | 1.07 | 2723.24 |
| | EA | 4 | 3.18 | 1503.44 | 2 | 2.95 | 1415.54 | 1 | 1.61 | 1487.49 |

| | | **Private equity** | | | **Private credit** | | | **Real assets** | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Year | Style | Mode | Mean | Avg. AUM | Mode | Mean | Avg. AUM | Mode | Mean | Avg. AUM |
| 1999 | IA | 1 | 1.05 | 459.16 | 1 | 1.00 | 611.12 | 1 | 1.19 | 512.69 |
| | EA | 1 | 1.20 | 374.62 | 1 | 1.00 | 1751.71 | 1 | 1.18 | 400.05 |
| 2009 | IA | 1 | 1.09 | 733.78 | 1 | 1.00 | 2492.83 | 1 | 1.45 | 1605.56 |
| | EA | 1 | 1.72 | 688.94 | 1 | 1.00 | 323.78 | 1 | 1.96 | 532.30 |
| 2019 | IA | 1 | 1.29 | 2930.15 | 1 | 1.13 | 2581.10 | 1 | 1.79 | 4379.81 |
| | EA | 1 | 2.13 | 1270.12 | 1 | 1.38 | 606.23 | 2 | 2.96 | 836.72 |

**Table B.7. Monotonicity test of asset allocation and size**. This table tests the monotonic relation between asset allocation and size for different asset classes. *p-value* $\mu_1 = \mu_4$ tests whether the mean on the first quartile portfolio (smallest plans) equals the mean on the fourth quartile portfolio (largest plans). *p-value MR test* denotes the $p$-value of the null hypothesis that $\min(\mu_i - \mu_{i-1}) \leq 0$ (positive relation) or $\min(\mu_{i-1} - \mu_i) \leq 0$ (negative relation). *Relation* signifies whether we test for a positive ("+") or negative ("–") monotonic relation. Portfolios are constructed as follows: we sort plans into quartiles based on size and use an equal weighted average of plans within a quartile and asset class. For a given year, we only include plans that also show up in the next's year database. At the end of the next year, the AUM allocation is calculated for each of the portfolios.

| | Stocks | Fixed income | Hedge & multi ass. | Private equity | Real assets |
|---|---|---|---|---|---|
| Aum allocation | | | | | |
| p-value: $\mu_1 = \mu_4$ | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| p-value MR test | 0.014 | 0.123 | 0.000 | 0.000 | 0.006 |
| Relation | – | – | – | + | + |

**Table B.8. Regression of cost on plan characteristics.** This table shows regression estimates of the model: $\text{Cost}_{iat} = c_a + \beta_{1,a}\text{External}_{iat} + \beta_{2,a}\text{Active}_{iat} + \beta_{3,a}\text{Private}_i + \beta_{4,a}\text{nonUS}_i + \varepsilon_{iat}$, where $\text{Cost}_{iat}$ is the cost (in bps) of plan $i$ in sub-asset class $a$ at time $t$, $\text{External}_{iat}$ ($\text{Active}_{iat}$) is a dummy equal to one if plan $i$ manages sub-asset class $a$ externally (actively) at time $t$, $\text{Private}_i$ is a dummy equal to one if plan $i$ is private, and $\text{nonUS}_i$ is a dummy equal to one if the plan is domiciled outside the U.S. We only keep those sub-asset classes that have 1,000 observations or more. Robust standard errors are clustered by sponsor and reported in parentheses. Boldface coefficients are significant at the 5% level.

| | External | Active | Private | nonUS | Obs | $R^2$ |
|---|---|---|---|---|---|---|
| Stocks | | | | | | |
| Canada | **0.13** | **0.17** | 0.02 | | 2615 | 0.25 |
| | (0.014) | (0.016) | (0.011) | | | |
| EAFE | **0.29** | **0.41** | **0.05** | 0.01 | 5769 | 0.24 |
| | (0.028) | (0.015) | (0.014) | (0.017) | | |
| U.S. Broad or All | **0.19** | **0.33** | **0.04** | 0.01 | 5413 | 0.38 |
| | (0.018) | (0.013) | (0.010) | (0.011) | | |
| U.S. Large Cap | **0.13** | **0.31** | **0.04** | 0.00 | 2509 | 0.39 |
| | (0.026) | (0.017) | (0.014) | (0.014) | | |
| U.S. Small Cap | **0.35** | **0.53** | **0.15** | -0.25 | 3288 | 0.01 |
| | (0.065) | (0.043) | (0.064) | (0.414) | | |
| Global | **0.28** | **0.41** | **0.05** | -0.01 | 2849 | 0.27 |
| | (0.037) | (0.023) | (0.018) | (0.019) | | |
| Emerging | **0.39** | **0.52** | **0.07** | **-0.05** | 3770 | 0.28 |
| | (0.040) | (0.024) | (0.021) | (0.022) | | |
| ACWI x U.S. | **0.32** | **0.46** | **0.09** | | 1215 | 0.49 |
| | (0.087) | (0.020) | (0.020) | | | |
| Fixed Income | | | | | | |
| Canada | **0.10** | **0.10** | **0.02** | | 2326 | 0.40 |
| | (0.009) | (0.010) | (0.008) | | | |
| Cash | **0.05** | | -0.11 | 0.12 | 5372 | 0.00 |
| | (0.008) | | (0.175) | (0.206) | | |
| U.S. | **0.15** | **0.13** | **0.05** | 0.05 | 4406 | 0.10 |
| | (0.011) | (0.014) | (0.010) | (0.029) | | |
| Other | **0.41** | **0.23** | 0.03 | -0.05 | 1379 | 0.09 |
| | (0.039) | (0.050) | (0.054) | (0.049) | | |
| Long Bonds | **0.07** | **0.11** | 0.01 | **-0.02** | 1651 | 0.36 |
| | (0.012) | (0.009) | (0.010) | (0.009) | | |
| Global | **0.24** | **0.23** | **0.06** | -0.03 | 1108 | 0.19 |
| | (0.028) | (0.034) | (0.027) | (0.024) | | |
| Inflation Indexed | **0.08** | **0.09** | 0.02 | -0.02 | 1870 | 0.21 |
| | (0.010) | (0.010) | (0.011) | (0.011) | | |
| High Yield | **0.27** | **0.24** | 0.01 | 0.04 | 2006 | 0.04 |
| | (0.055) | (0.070) | (0.023) | (0.027) | | |
| Emerging | **0.45** | **0.38** | 0.03 | 0.03 | 1299 | 0.15 |
| | (0.061) | (0.057) | (0.040) | (0.047) | | |
| Hedge & Multi ass. | | | | | | |
| Funded TAA | **0.60** | | -0.03 | 0.14 | 1262 | 0.01 |
| | (0.118) | | (0.185) | (0.196) | | |
| Hedge Funds | | | 0.12 | 0.09 | 2630 | 0.00 |
| | | | (0.083) | (0.083) | | |
| Private Equity | | | | | | |
| Diversified | **5.28** | | **-0.70** | 0.52 | 4680 | 0.02 |
| | (0.321) | | (0.305) | (0.338) | | |
| Other | **3.33** | | -0.90 | 0.35 | 1347 | 0.03 |
| | (0.750) | | (0.532) | (0.736) | | |
| Real Assets | | | | | | |
| Real Estate ex-REITs | **1.22** | | -0.10 | **-0.25** | 6416 | 0.08 |
| | (0.077) | | (0.071) | (0.085) | | |
| REITs | **0.40** | **0.31** | 0.04 | **-0.11** | 1825 | 0.12 |
| | (0.028) | (0.035) | (0.033) | (0.032) | | |
| Infrastructure | **2.68** | | -0.58 | **-1.35** | 1582 | 0.02 |
| | (0.248) | | (0.427) | (0.655) | | |

**Table B.9. Economies of scale at the sub-asset class level**. This table shows estimates of the model: $\log(\text{Cost}^{\$}_{iats}) = c_{as} + \lambda_{ats} + \beta_{1,as}\log(\text{AUM}_{iats}) + \beta_{2,as}\text{Private}_i + \beta_{3,as}\text{nonUS}_i + \varepsilon_{iats}$, where $\text{Cost}^{\$}_{iats}$ is the (dollar) cost of plan $i$ in sub-asset class $a$ at time $t$ for asset mandate $s$, $c_{as}$ is a constant that varies with sub-asset class $a$ and mandate $s$, $\lambda_{ats}$ is the time fixed effect for sub-asset class $a$ in investment management mandate $s$, $\log(\text{AUM}_{iats})$ is the log of total AUM of plan $i$ in sub-asset class $a$ at time $t$ for mandate $s$, $\text{Private}_i$ is a dummy equal to one if plan $i$ is private and $\text{nonUS}_i$ is a dummy equal to one if plan $i$ is located outside the US. For stock and fixed income, we estimate the panel separately for the following styles $s$: Internal Passive (IP), Internal Active (IA), External Passive (EP) and External Active (EA). Robust standard errors are clustered by plan. The boldface coefficients on log(AUM) are significantly different from one at the 5% level and boldface coefficients on the other covariates are significantly different from zero. We only include sub-asset classes that have more than 400 observations.

| | $\log(\text{AUM}_{iats})$ | $\text{Private}_i$ | $\text{nonUS}_i$ | Obs | $R^2$ |
|---|---|---|---|---|---|
| EAFE (Stocks) | | | | | |
| IP | **0.74** (0.054) | 0.33 (0.299) | 0.51 (0.260) | 956 | 0.69 |
| EP | **0.76** (0.033) | 0.05 (0.121) | -0.14 (0.148) | 3999 | 0.67 |
| IA | 0.94 (0.062) | 0.29 (0.330) | **0.48** (0.215) | 1049 | 0.68 |
| EA | **0.90** (0.010) | **0.06** (0.026) | **-0.08** (0.029) | 10503 | 0.93 |
| U.S. Broad/All (Stocks) | | | | | |
| IP | **0.77** (0.045) | 0.26 (0.208) | **0.93** (0.191) | 1780 | 0.74 |
| EP | **0.75** (0.026) | 0.07 (0.070) | **0.47** (0.082) | 6888 | 0.68 |
| IA | **0.87** (0.039) | **0.50** (0.183) | **0.63** (0.171) | 2077 | 0.67 |
| EA | **0.95** (0.015) | **0.13** (0.044) | -0.09 (0.054) | 8155 | 0.88 |
| Inflation Indexed (Fixed Income) | | | | | |
| IP | 0.94 (0.079) | -0.44 (0.404) | **0.85** (0.334) | 1072 | 0.72 |
| EP | **0.78** (0.052) | 0.19 (0.150) | **0.33** (0.146) | 1384 | 0.66 |
| IA | **0.76** (0.056) | -0.07 (0.234) | **0.72** (0.230) | 954 | 0.67 |
| EA | **0.88** (0.049) | 0.01 (0.199) | 0.05 (0.221) | 1107 | 0.39 |
| Diversified Private Equity | | | | | |
| IA | 1.07 (0.038) | **0.54** (0.186) | **0.64** (0.216) | 682 | 0.83 |
| EA | **0.93** (0.008) | **-0.16** (0.029) | -0.04 (0.034) | 5042 | 0.96 |
| Other Private Equity | | | | | |
| IA | 0.91 (0.063) | 0.04 (0.260) | 0.19 (0.319) | 578 | 0.72 |
| EA | **0.96** (0.019) | -0.10 (0.058) | **-0.14** (0.066) | 1189 | 0.93 |
| Real Estate ex-REITs (Real Assets) | | | | | |
| IA | 1.04 (0.043) | -0.23 (0.189) | 0.06 (0.133) | 1936 | 0.73 |
| EA | **0.95** (0.010) | **-0.10** (0.037) | **-0.20** (0.039) | 7129 | 0.94 |
| REITs (Real Assets) | | | | | |
| IA | 0.97 (0.056) | 0.40 (0.301) | **0.46** (0.232) | 604 | 0.78 |
| EA | **0.88** (0.024) | 0.01 (0.079) | **-0.32** (0.075) | 1734 | 0.80 |

**Table B.10. Average scaled investment management costs by asset class and country in 2009 and 2019**. This table shows the average investment management costs measured relative to the grand average cost. The grand average cost ($\overline{\text{Cost}}$) is calculated as $\overline{\text{Cost}} = \frac{1}{N|A|(T-t+1)} \sum_{i=1}^{N} \sum_{A=1}^{|A|} \sum_{t=1991}^{T} \text{Cost}_{iAt}$, where $i$ indicates plan sponsors, $A$ indicates asset class, $t$ indicates year. We calculate the scaled cost ($\overline{\text{Cost}}_{At}$) separately for each country, asset class, and time period as $\overline{\text{Cost}}_{At} = \frac{1}{N} \sum_{i=1}^{N} \text{Cost}_{iAt} / \overline{\text{Cost}}$.

| | 2009 | | | | | | 2019 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Country | Stock | Fixed Income | Hedge & Multi Ass. | Private Equity | Private Credit | Real Asset | Stock | Fixed Income | Hedge & Multi Ass. | Private Equity | Private Credit | Real Asset |
| U.S. | 79 | 48 | 534 | 1342 | 239 | 327 | 67 | 44 | 427 | 931 | 483 | 367 |
| Australia | 67 | 21 | 645 | 1207 | | 267 | 25 | 22 | 233 | 719 | 195 | 153 |
| Canada | 73 | 28 | 567 | 1424 | 62 | 237 | 76 | 31 | 383 | 922 | 313 | 311 |
| China | | | | | | | 34 | 30 | 318 | 763 | 471 | 196 |
| Denmark | 82 | 17 | 476 | 981 | | 189 | | | | | | |
| Finland | 51 | 16 | 683 | 1174 | | 237 | 63 | 46 | 613 | 945 | 195 | 235 |
| France | 38 | 24 | 0 | 3149 | | 24 | | | | | | |
| Netherlands | 56 | 33 | 524 | 1110 | 48 | 224 | 36 | 30 | 500 | 999 | 97 | 214 |
| New Zealand | 94 | 20 | 499 | 1399 | | 164 | 59 | 57 | 211 | 632 | | 190 |
| Other USD | 113 | 115 | 833 | 1394 | | 193 | 103 | 77 | 414 | 910 | 725 | 311 |
| Saudi Arabia | | | | | | | 73 | 31 | 386 | 931 | 474 | 449 |
| South Africa | | | | | | | 72 | 16 | 536 | 469 | | 135 |
| South Korea | 41 | 13 | | 9511 | | 71 | 35 | 4 | 619 | 378 | | 200 |
| Sweden | 29 | 13 | 375 | 1228 | | 92 | 37 | 9 | 329 | 1006 | 293 | 127 |
| UK | 22 | 14 | 570 | 1169 | | 343 | 54 | 42 | 306 | 1046 | 337 | 279 |

**Table B.11. Summary statistics for asset class returns**. This table reports summary measures for returns on the six asset classes. Panel A presents summary statistics on the mean and Sharpe ratio. Mean returns are computed as the average return of the asset class across years and plan sponsors: $\bar{r}_A = \frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N} r_{iAt}$, where $r_{iAt}$ is the gross return of plan $i$ in asset class $A$ at time $t$. Standard deviations of returns, reported on the diagonal in Panel B, are computed as follows: $\sqrt{(1/T)\sum_{t=1}^{T}(\bar{r}_{At} - \bar{r}_A)^2}$, where $\bar{r}_{At} = (1/N)\sum_{i=1}^{N} r_{iAt}$. The Sharpe Ratio is computed as the ratio of the mean excess return over the standard deviation of excess returns. In panel A, summary statistics are reported separately for all plans ("All"), and for passively managed assets ("Passive") and actively managed assets ("Active"). Because all Private Equity and Private Debt assets are actively managed, we do not provide any summary statistics for them in the "Passive" subheading. The asset class "Hedge & multi ass." includes hedge funds and multi-assets, hence also includes passively managed assets. The lower triangle of Panel B presents pairwise correlations between mean returns across aggregate asset classes. Boldface correlations are statistically significant at the 5% level.

**Panel A: Summary Statistics**

| | All | | Passive | | Active | |
|---|---|---|---|---|---|---|
| | Mean | Sharpe | Mean | Sharpe | Mean | Sharpe |
| | 0.108 | 0.525 | 0.100 | 0.460 | 0.108 | 0.502 |
| | 0.070 | 0.669 | 0.067 | 0.598 | 0.065 | 0.572 |
| | 0.071 | 0.546 | 0.129 | 1.263 | 0.068 | 0.450 |
| | 0.159 | 0.631 | | | 0.156 | 0.644 |
| | 0.077 | 0.636 | | | 0.078 | 0.553 |
| | 0.084 | 0.537 | 0.054 | 0.220 | 0.086 | 0.531 |

**Panel B: Correlation Matrix**

| | Stocks | Fixed Income | Hedge & multi ass. | Private Equity | Private Debt | Real Assets |
|---|---|---|---|---|---|---|
| Stocks | 0.163 | | | | | |
| Fixed Income | 0.278 | 0.069 | | | | |
| Hedge & multi ass. | **0.858** | **0.547** | 0.105 | | | |
| Private Equity | **0.412** | -0.050 | **0.382** | 0.223 | | |
| Private Debt | 0.306 | **0.598** | **0.528** | -0.006 | 0.097 | |
| Real Assets | 0.201 | -0.161 | 0.124 | **0.528** | 0.204 | 0.116 |

**Table B.12. Regression of net returns on risk factors**. This table shows estimates of alphas and betas (factor loadings) from regressions of U.S. plans' annual average net returns on risk factors for different asset classes (see Equation (2.6.5) in the main text). In the column "Portfolio", the returns are constructed as an equally weighted average over the individual plans' net returns in a specific year and asset class. In the column "pooled", returns are pooled across all U.S. plans. Cluster robust standard errors are reported in parentheses for pooled regression and robust standard errors are reported in parentheses for the portfolio regressions. Boldface coefficients are significant at the 5% level.

**Public asset classes**

| | Stocks | | | Fixed Income | |
|---|---|---|---|---|---|
| Factors | Portfolio | Pooled | Factors | Portfolio | Pooled |
| $\alpha$ | -0.69 | **-0.01** | $\alpha$ | **0.65** | **0.01** |
| | (0.812) | (0.001) | | (0.235) | (0.001) |
| Market | **0.95** | **0.95** | Bond Index | **0.78** | **0.79** |
| | (0.047) | (0.004) | | (0.081) | (0.040) |
| SMB | **0.12** | **0.13** | Corp. Index | **0.44** | **0.42** |
| | (0.045) | (0.008) | | (0.095) | (0.033) |
| HML | 0.02 | **0.02** | High Yield Index | 0.04 | **0.05** |
| | (0.048) | (0.006) | | (0.044) | (0.015) |
| | | | Global Div. Index | 0.01 | **0.01** |
| | | | | (0.011) | (0.005) |
| | | | Long Treasury | **0.24** | **0.23** |
| | | | | (0.034) | (0.023) |
| $R^2$ | 0.95 | 0.92 | | 0.97 | 0.58 |
| Obs | 29 | 4860 | | 26 | 4617 |

**Alternative asset classes**

| | Hedge & Multi Ass. | | Private Equity | | Private Credit | | Real Assets | |
|---|---|---|---|---|---|---|---|---|
| Factors | Portfolio | Pooled | Portfolio | Pooled | Portfolio | Pooled | Portfolio | Pooled |
| $\alpha$ | **-9.01** | **-0.10** | 2.60 | 0.02 | 5.46 | **0.08** | 2.44 | **0.03** |
| | (2.894) | (0.012) | (3.217) | (0.010) | (3.624) | (0.010) | (3.502) | (0.006) |
| Market | **0.47** | **0.46** | **0.19** | **0.21** | **0.22** | **0.11** | 0.15 | **0.14** |
| | (0.035) | (0.017) | (0.073) | (0.021) | (0.072) | (0.027) | (0.083) | (0.014) |
| $\text{Market}_{t-1}$ | | | **0.27** | **0.27** | 0.09 | 0.02 | **0.27** | **0.24** |
| | | | (0.048) | (0.017) | (0.124) | (0.043) | (0.130) | (0.019) |
| SizeSpread | **-0.17** | **-0.16** | -0.06 | **-0.11** | -0.20 | **0.15** | 0.20 | **0.14** |
| | (0.072) | (0.027) | (0.179) | (0.035) | (0.219) | (0.043) | (0.117) | (0.026) |
| $\text{SizeSpread}_{t-1}$ | | | **-0.30** | **-0.18** | 0.15 | **0.21** | **0.26** | **0.24** |
| | | | (0.083) | (0.055) | (0.096) | (0.041) | (0.102) | (0.020) |
| BondMarket | **0.11** | **0.13** | **0.88** | **0.62** | -0.49 | **-0.16** | 0.01 | 0.06 |
| | (0.030) | (0.012) | (0.273) | (0.067) | (0.297) | (0.068) | (0.268) | (0.037) |
| $\text{BondMarket}_{t-1}$ | | | **-0.73** | **-0.46** | 0.43 | 0.07 | 0.00 | -0.06 |
| | | | (0.270) | (0.057) | (0.303) | (0.060) | (0.251) | (0.035) |
| Bond trend | **-0.03** | **-0.03** | | | | | | |
| | (0.008) | (0.003) | | | | | | |
| Currency trend | 0.02 | **0.01** | | | | | | |
| | (0.010) | (0.004) | | | | | | |
| Commodity trend | 0.00 | 0.00 | | | | | | |
| | (0.011) | (0.004) | | | | | | |
| CreditSpread | 0.11 | **0.12** | | | | | | |
| | (0.073) | (0.027) | | | | | | |
| $R^2$ | 0.92 | 0.58 | 0.74 | 0.22 | 0.40 | 0.16 | 0.44 | 0.15 |
| Obs | 26 | 1801 | 25 | 2522 | 25 | 353 | 25 | 3105 |

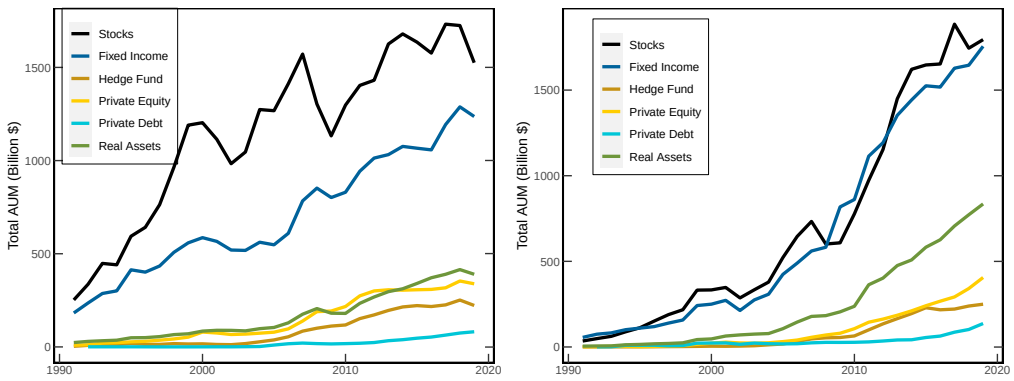
**Table B.13. Regressions of policy-adjusted gross returns on a single risk factor.** This table shows summary statistics of plan-level policy-adjusted gross returns regressed on a single factor (see (B.4.1)), where $\alpha$ denotes a plan's "alpha" and $\beta$ denotes the factor loading. For stocks we use the excess market return factor, and for fixed income we use the U.S. aggregate bond index factor. We require plans to have at least 10 years of observations to be included in the regression and only consider U.S. plans.

| | Stocks | | Fixed Income | |
|---|---|---|---|---|
| | $\alpha$ | $\beta$ | $\alpha$ | $\beta$ |
| Min. | -0.0367 | -0.3373 | -0.0214 | -1.3565 |
| 1st Qu. | -0.0010 | -0.0234 | -0.0013 | -0.0929 |
| Median | 0.0034 | 0.0031 | 0.0020 | 0.0015 |
| Mean | 0.0038 | -0.0003 | 0.0029 | -0.0042 |
| 3rd Qu. | 0.0082 | 0.0252 | 0.0064 | 0.1038 |
| Max. | 0.0319 | 0.2257 | 0.0248 | 1.4038 |
| # of Plans | 199 | | 203 | |

**Table B.14. Regression of sub-asset class returns on plan characteristics.** This table shows estimates of the model: $\widetilde{r}_{iat} = \lambda_{at} + \beta_{1,a} \log(\text{AUM}_{iat-1}) + \beta_{2,a}\text{Private}_i + \beta_{3,a}\text{nonUS}_i + \beta'_{4,a}x_{iat} + \varepsilon_{iat}$, where $\widetilde{r}_{iat}$ denotes the policy-adjusted **net** return, $\lambda_{at}$ is a time fixed effect, $\text{AUM}_{iat-1}$ is plan $i$'s total AUM allocated to sub-asset class $a$ at time $t-1$, $\text{Private}_i$ is a dummy equal to one if plan $i$ is private, $\text{nonUS}_i$ is a dummy equal to one if plan $i$ is domiciled outside the U.S., and $x_{iat}$ is a vector of controls that include $\text{External}_{iat}$ and $\text{Active}_{iat}$. Both controls are dummy variables equal to one if sub-asset class $a$ is managed externally and actively by plan $i$, respectively. For comparison, the first column reports results when running the same regression using **gross** returns. We keep only those sub-asset classes that have 1,000 observations or more. Robust standard errors are clustered by sponsor and reported in parentheses. Boldface coefficients are significant at the 5% level.

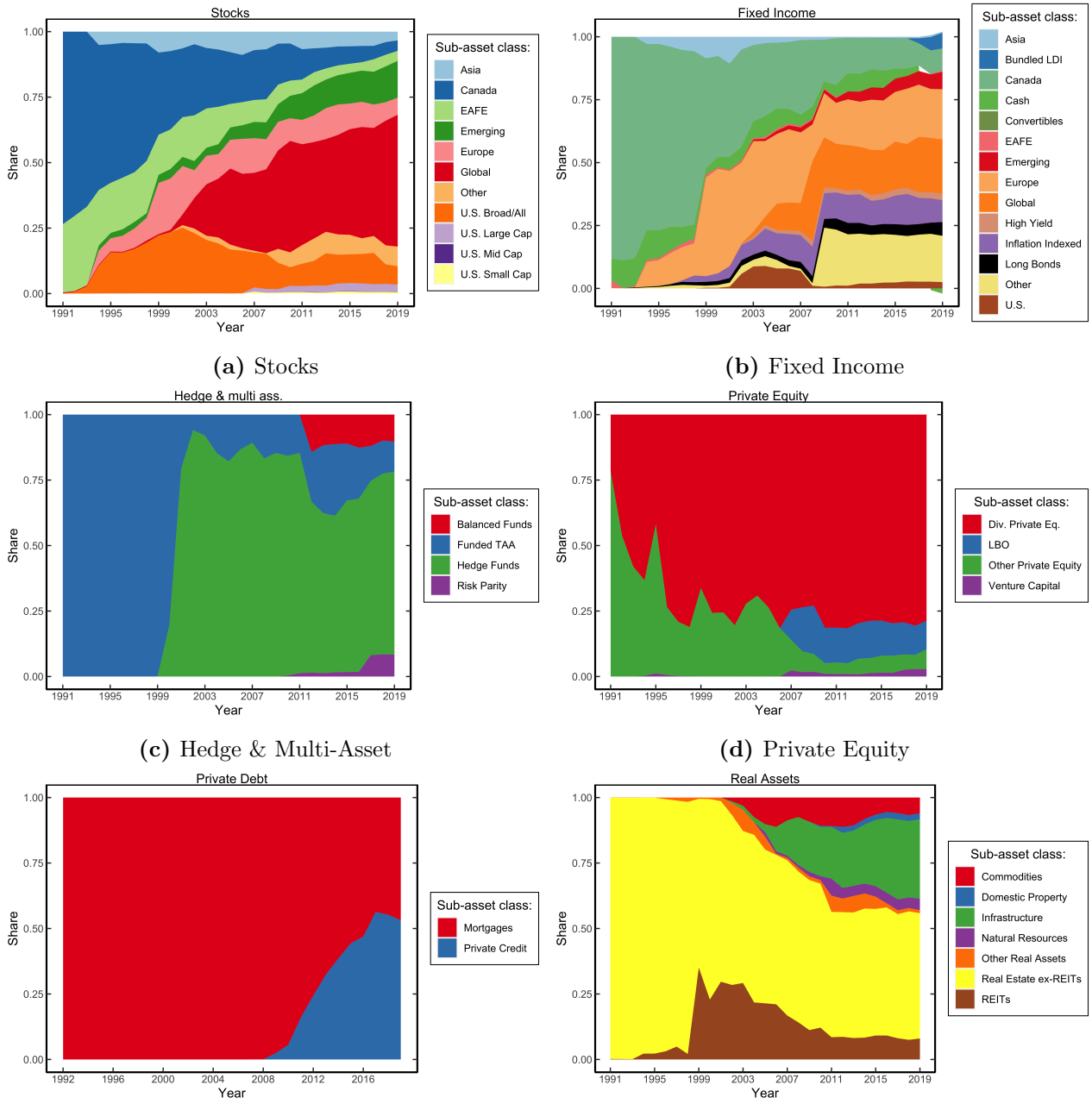
| | **Gross** | **Net** | | | | |
|---|---|---|---|---|---|---|
| Stocks | log(AUM) | log(AUM) | Private | nonUS | Obs | $R^2$ |
| Canada | -0.07 | -0.04 | 0.20 | | 2568 | 0.35 |
| | (0.079) | (0.078) | (0.174) | | | |
| EAFE | 0.08 | **0.12** | 0.00 | -0.31 | 5571 | 0.20 |
| | (0.053) | (0.052) | (0.162) | (0.186) | | |
| U.S. Broad or All | 0.12 | 0.13 | **0.28** | -0.11 | 5209 | 0.11 |
| | (0.074) | (0.074) | (0.137) | (0.218) | | |
| U.S. Large Cap | 0.11 | **0.13** | 0.22 | **0.45** | 2439 | 0.09 |
| | (0.060) | (0.060) | (0.123) | (0.220) | | |
| U.S. Small Cap | **0.28** | 0.18 | 0.26 | 0.86 | 3142 | 0.10 |
| | (0.113) | (0.180) | (0.316) | (0.481) | | |
| Global | **0.12** | **0.16** | 0.13 | -0.29 | 2698 | 0.07 |
| | (0.057) | (0.057) | (0.228) | (0.256) | | |
| Emerging | 0.10 | **0.15** | -0.21 | -0.25 | 3560 | 0.08 |
| | (0.057) | (0.057) | (0.213) | (0.198) | | |
| ACWI x U.S. | 0.14 | 0.20 | 0.12 | | 1173 | 0.20 |
| | (0.122) | (0.123) | (0.335) | | | |
| Fixed Income | | | | | | |
| Canada | 0.03 | **0.05** | -0.05 | | 2270 | 0.12 |
| | (0.026) | (0.025) | (0.078) | | | |
| Cash | -0.02 | 0.01 | 0.07 | **-0.40** | 5977 | 0.01 |
| | (0.043) | (0.045) | (0.114) | (0.180) | | |
| U.S. | -0.03 | -0.01 | -0.04 | 0.46 | 4277 | 0.30 |
| | (0.052) | (0.051) | (0.107) | (0.405) | | |
| Other | -0.40 | -0.37 | -1.14 | -2.19 | 1195 | 0.03 |
| | (0.449) | (0.449) | (1.260) | (1.636) | | |
| Long Bonds | 0.06 | 0.08 | -0.09 | **-0.38** | 1594 | 0.05 |
| | (0.056) | (0.056) | (0.201) | (0.156) | | |
| Global | **0.30** | **0.33** | 0.51 | -0.51 | 1020 | 0.15 |
| | (0.102) | (0.102) | (0.431) | (0.426) | | |
| Inflation Indexed | **0.13** | **0.14** | -0.11 | 0.16 | 1754 | 0.03 |
| | (0.062) | (0.062) | (0.227) | (0.185) | | |
| High Yield | 0.15 | 0.17 | -0.05 | 0.43 | 1897 | 0.22 |
| | (0.096) | (0.095) | (0.242) | (0.307) | | |
| Emerging | -0.13 | -0.05 | 0.30 | **-0.54** | 1224 | 0.26 |
| | (0.105) | (0.109) | (0.244) | (0.238) | | |
| Hedge & multi ass. | | | | | | |
| Funded TAA | -0.10 | 0.08 | -0.45 | 0.07 | 1123 | 0.14 |
| | (0.415) | (0.288) | (0.711) | (0.709) | | |
| Hedge Funds | **0.37** | **0.46** | 0.48 | -0.40 | 2406 | 0.17 |
| | (0.101) | (0.099) | (0.402) | (0.395) | | |
| Private Equity | | | | | | |
| Diversified | 0.31 | **0.79** | **1.22** | **2.57** | 4212 | 0.22 |
| | (0.215) | (0.190) | (0.620) | (0.609) | | |
| Other | **1.22** | **1.29** | 4.21 | 3.45 | 1176 | 0.07 |
| | (0.583) | (0.579) | (3.067) | (2.078) | | |
| Real Assets | | | | | | |
| Real Estate ex-REITs | 0.19 | 0.29 | 0.49 | **0.72** | 6067 | 0.07 |
| | (0.157) | (0.159) | (0.356) | (0.292) | | |
| REITs | 0.22 | **0.28** | 0.50 | 0.27 | 1686 | 0.06 |
| | (0.131) | (0.131) | (0.389) | (0.351) | | |
| Infrastructure | 0.05 | 0.56 | 0.92 | 0.79 | 1443 | 0.11 |
| | (0.383) | (0.378) | (0.816) | (0.813) | | |

# Figures




**(a)** U.S. Plans **(b)** Non-U.S. Plans

**Figure B.1. Total AUM by asset class and year for U.S. and non-U.S. plans**. This figure presents total AUM (in billion dollars) allocated to stocks, fixed income, hedge fund and multi assets, private equity, private debt, and real assets for U.S. and non-U.S. plans. Total AUM is defined as $\text{AUM}_{At} = \sum_i \text{AUM}_{iAt}$, where $\text{AUM}_{iAt}$ indicates the AUM of plan $i$ in asset class $A$ at time $t$.




**(a)** Stocks

**(b)** Fixed Income

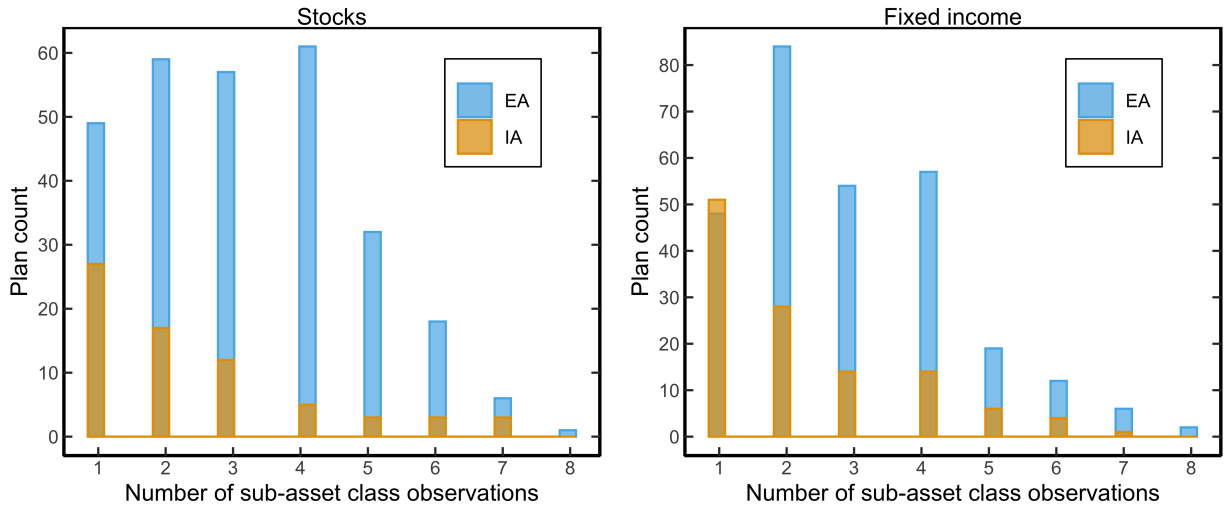
**(c)** Hedge & Multi-Asset

**(d)** Private Equity

**(e)** Private Debt

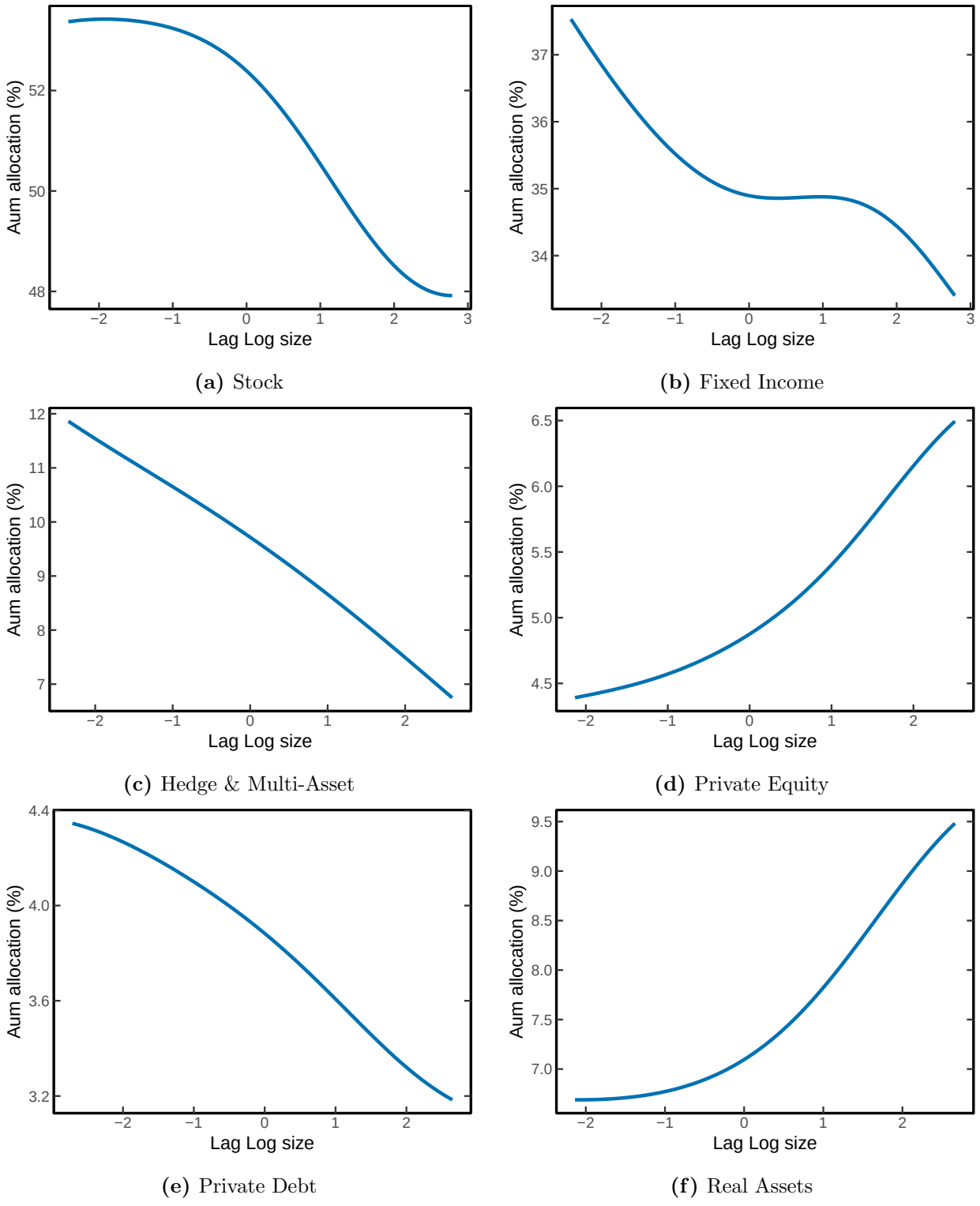
**(f)** Real Assets

**Figure B.2. Sub-asset class allocation over time for non-U.S. plans.** This figure shows the share of total AUM allocated to each sub-asset class for a given year and asset class for plans domiciled outside the U.S.




**(a)** **(b)**

**Figure B.3. Frequency of internal and external active management in 2019**. This figure shows a histogram of the number of sub-asset class observations by plan for internal active (IA) and external active (EA) management in 2019 for stocks and fixed income.




**(a)** Stock

**(b)** Fixed Income

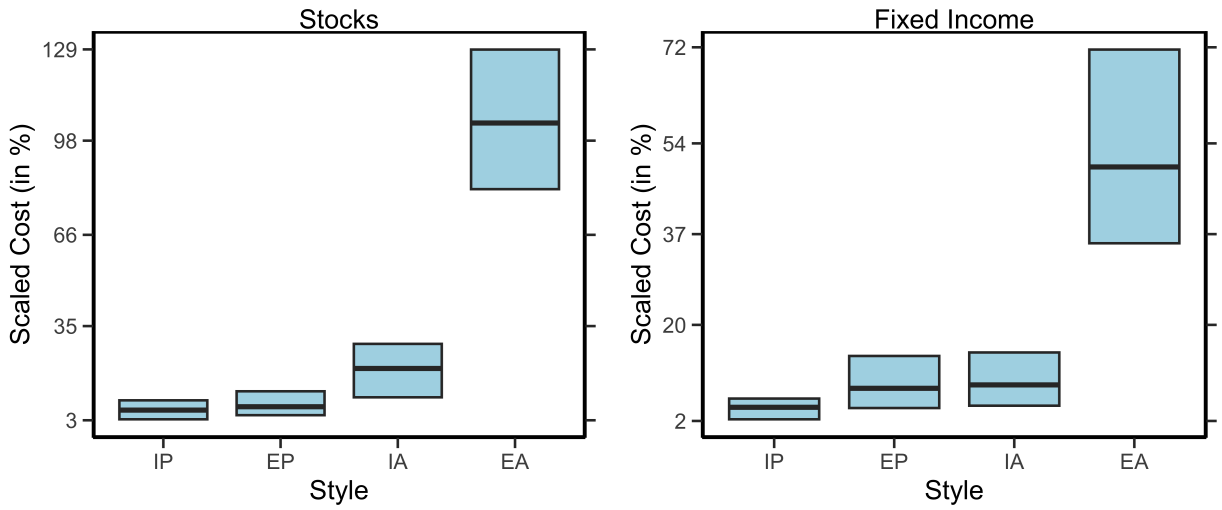
**(c)** Hedge & Multi-Asset

**(d)** Private Equity

**(e)** Private Debt

**(f)** Real Assets

**Figure B.4. Nonparametric estimates of the relation between plan size and AUM allocation.** This figure shows the pooled kernel estimate of AUM allocation ($\omega_{iAt}$) on $\log(\text{AUM}_{iAt-1})$ for different asset classes, over the sample period 1991–2019. The values of $\log(\text{AUM}_{iAt-1})$ are cross sectionally demeaned to account for time trends.




**(a)** **(b)**

**Figure B.5. Investment management costs by mandate for stocks and fixed income holdings.** The figure shows boxplots of scaled cost by management mandate for public asset classes in 2019. The different type of management styles include: Internal Passive (IP), External Passive (EP), Internal Active (IA) and External Active (EA). Cost are scaled by the average cost across plans, years, and asset classes.

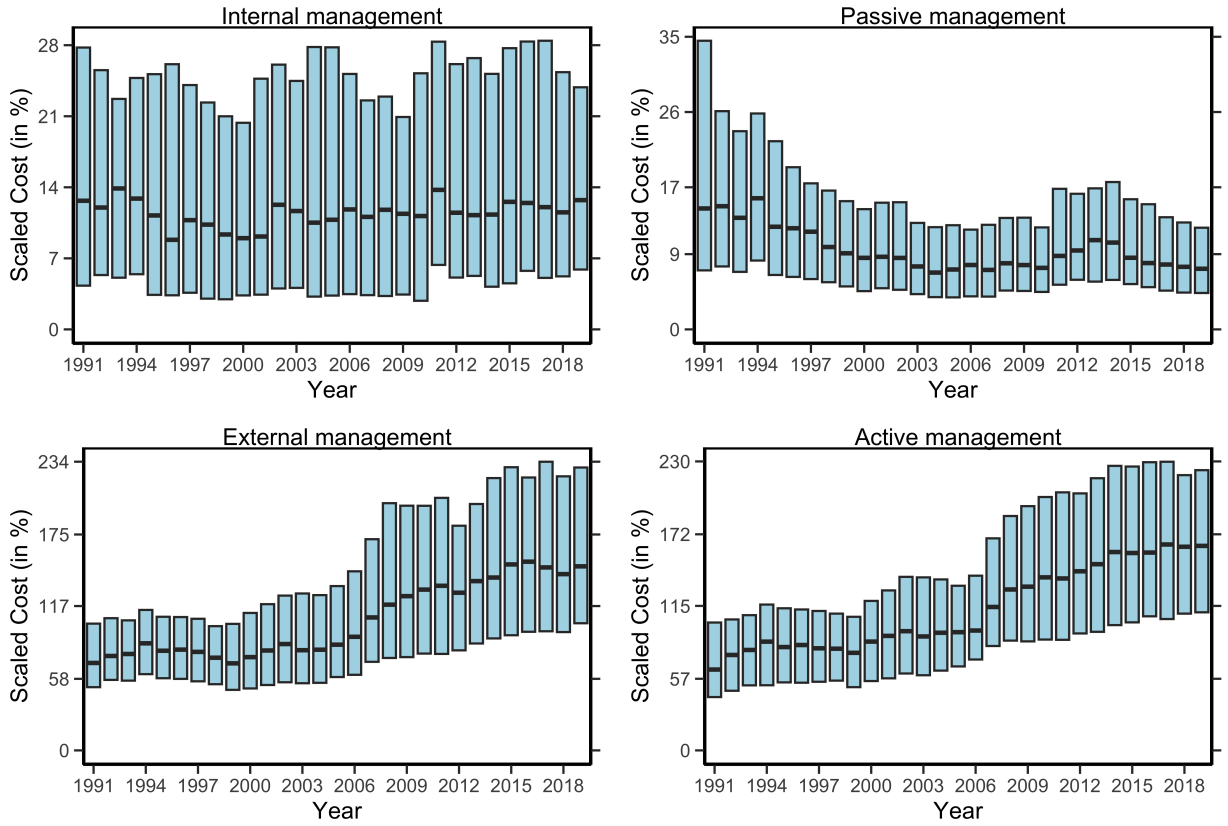



**Figure B.6. Evolution in investment management costs by mandate.** The figure shows box plots of total (scaled) cost for internal, external, passive, and active management across plans over the sample period 1991–2019. Costs are averaged over the asset classes (by AUM) to get a plan level measure. Finally, we divide the cost by the average cost computed across plans, years, and asset classes.

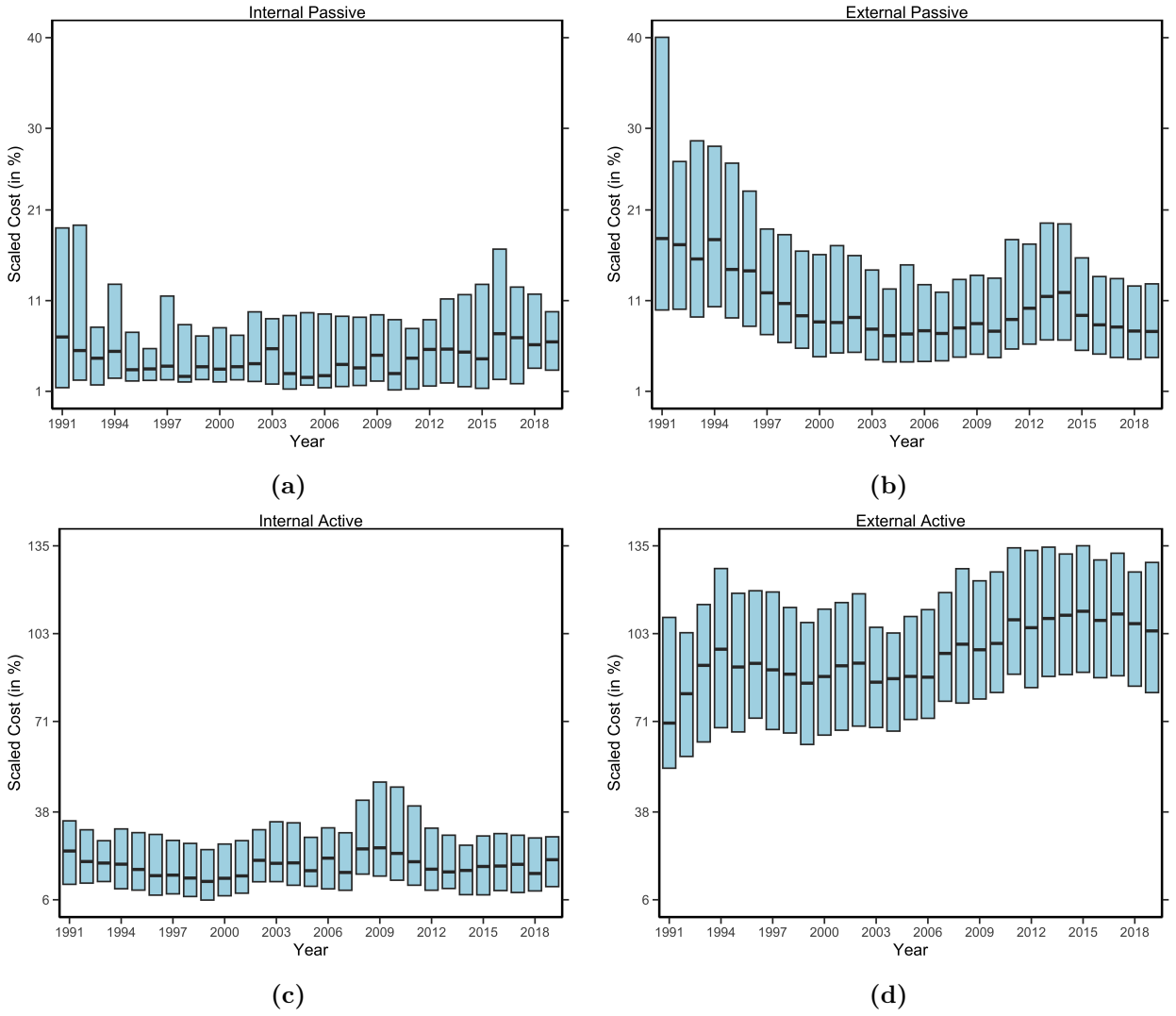



**Figure B.7. Evolution of stock investment management costs by mandate.** The figure shows box plots of scaled cost in stock investments for the mandates: Internal Passive, External Passive, Internal Active and External Active. Cost are defined as the weighted average (by AUM) of all costs attributed to a particular investment style for a specific plan/year. The cost are scaled by the average cost across years, asset classes, and plans.

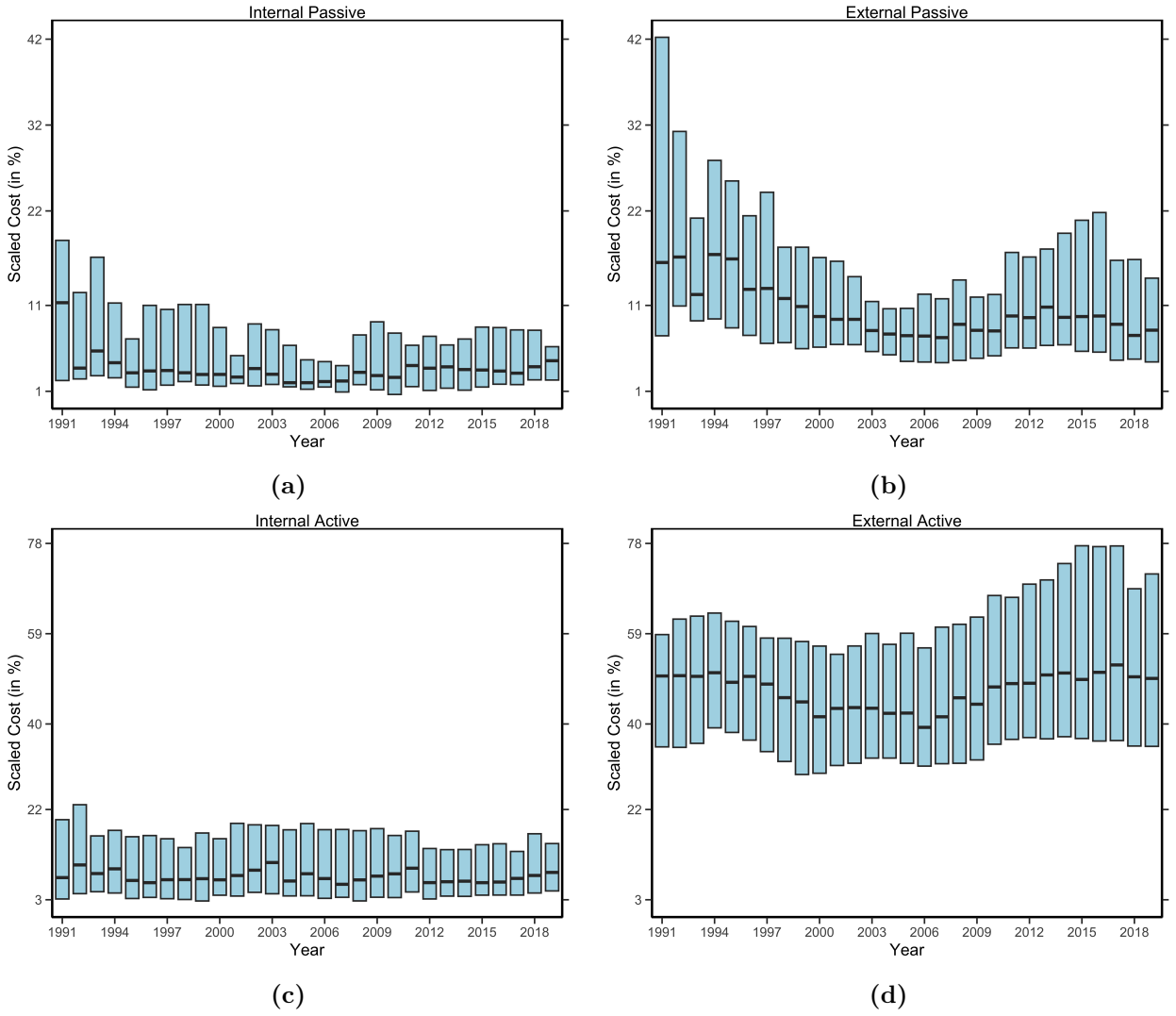



**Figure B.8. Evolution of fixed income investment management costs by mandate.** The figure shows boxplots of scaled cost in fixed income investments for the mandates: Internal Passive, External Passive, Internal Active and External Active. Cost are defined as the weighted average (by AUM) of all costs attributed to a particular investment style for a specific plan/year. The cost are scaled by the average cost across years, asset classes, and plans.

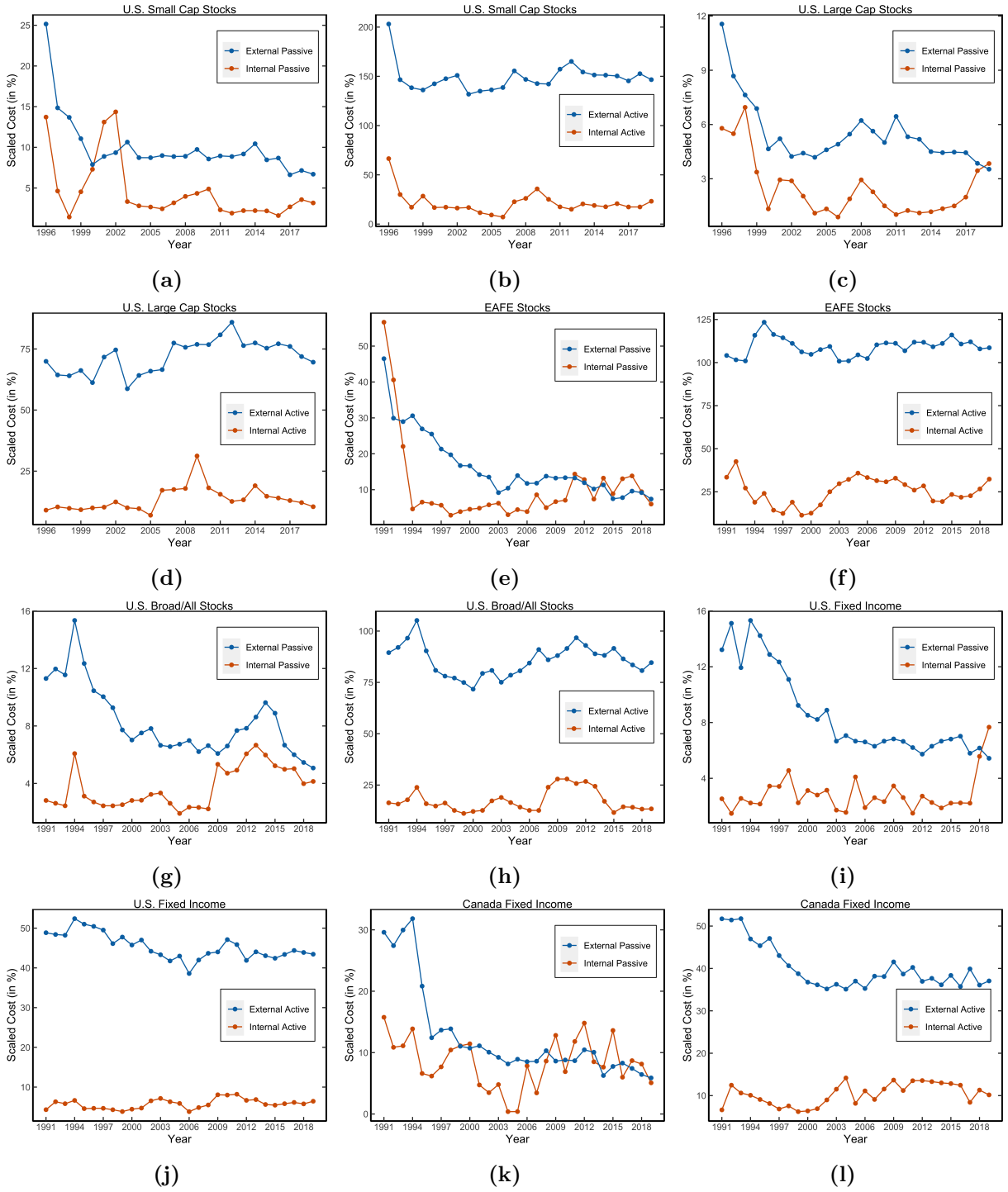
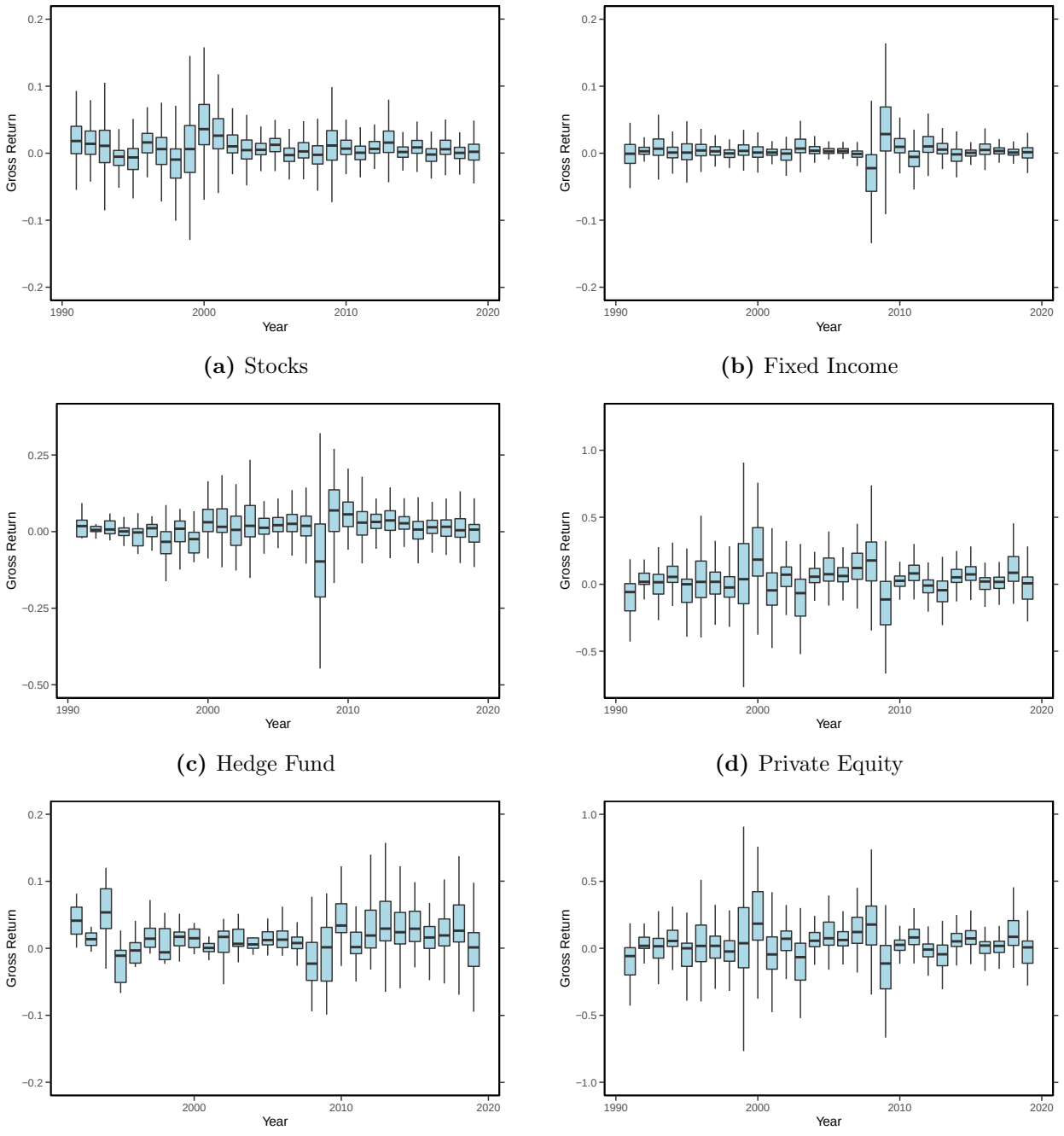



**Figure B.9. Median management costs by mandate in public sub-asset classes.** This figure shows median (scaled) investment management costs at the sub-asset class level for four different management mandates: Internal Passive, External Passive, Internal Active, External Active. Median costs represent the median of average cost across plans for a given year. We only include sub-asset classes that have enough time series observations for all management mandates. Finally, we scale the costs by the average cost across years, asset classes, and plans.




**(a)** Stocks

**(b)** Fixed Income

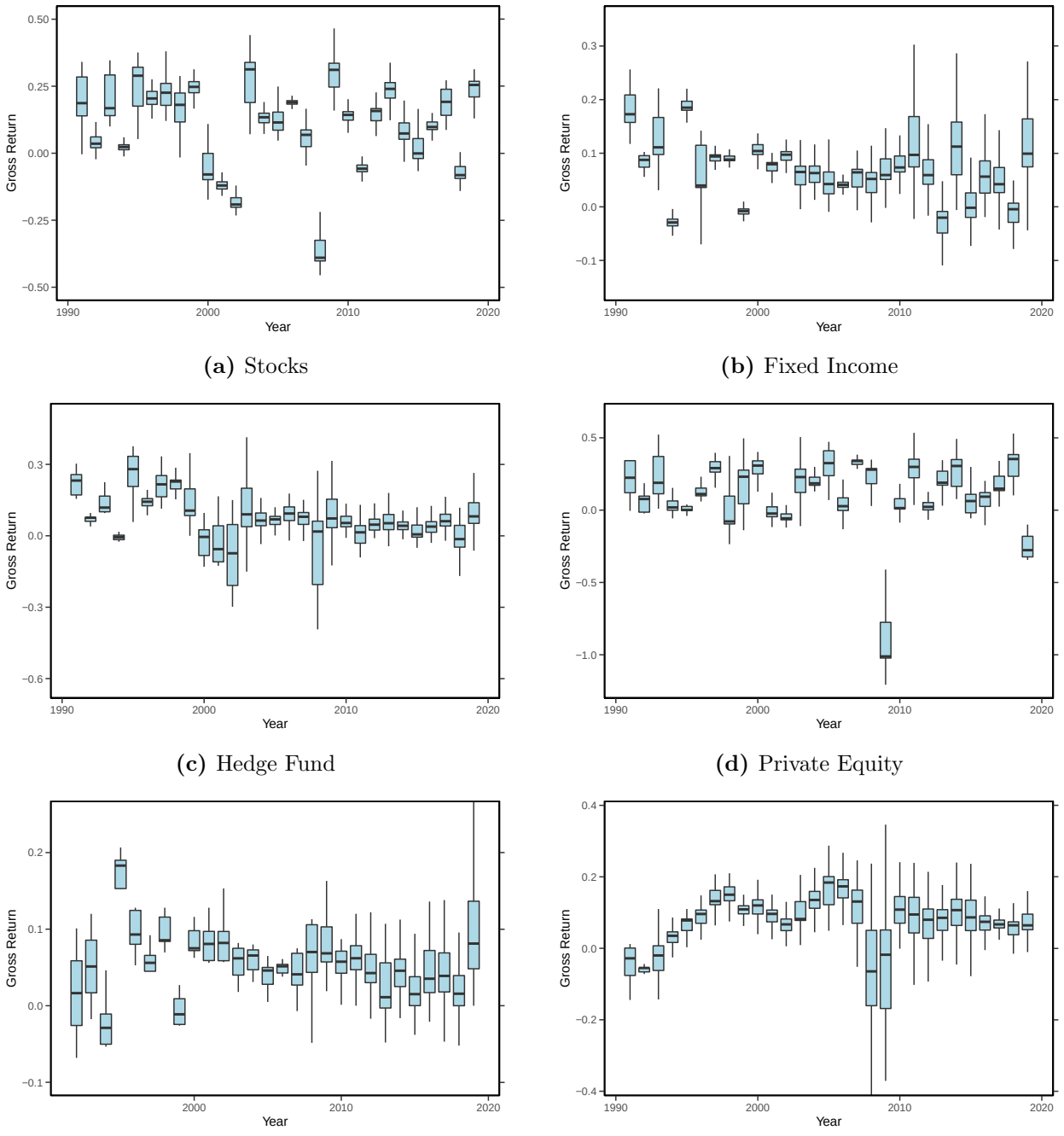
**(c)** Hedge Fund

**(d)** Private Equity

**(e)** Private Debt

**(f)** Real Assets

**Figure B.10. Policy-adjusted gross returns**. This figure shows box plots of gross policy-adjusted returns pooled across plans in a given year for different asset classes.




**(a)** Stocks

**(b)** Fixed Income

**(c)** Hedge Fund

**(d)** Private Equity

**(e)** Private Debt

**(f)** Real Assets

**Figure B.11. Policy return box Plots**. This figure presents the time series box plots for policy gross returns across plans and asset classes.

# Appendix C

# Appendix to Chapter 3

## C.1 Space of cumulative discount rates

**Lemma C.1.1.** *Let* $\mathcal{X} = \left\{x \in C^1[0,T] : x(0) = 0\right\}$ *be the vector space of continuously differentiable functions on* $[0,T]$ *with* $x(0) = 0$*. For* $x \in \mathcal{X}$*, define* $\|x\|_{\mathcal{X}} = \sup_{t\in[0,T]} |x'(t)|$*. Then* $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ *is a Banach space.*

*Proof.* Write $\|\cdot\| = \|\cdot\|_{\mathcal{X}}$ to simplify notation. Let us first show that $\|\cdot\|$ is a norm on $\mathcal{X}$. Since $x \in \mathcal{X}$ is continuously differentiable, $x'$ is continuous, so $\|x\| = \sup_{t\in[0,T]} |x'(t)| \in [0,\infty)$. Clearly $0 \in \mathcal{X}$ and $\|0\| = 0$. If $\|x\| = 0$, then $x'(t) = 0$ for all $t \in [0,T]$. Then $x(t) = x(0) + \int_0^t x'(u)\,\mathrm{d}u = 0$ because $x(0) = 0$, so $x = 0$. For any $\alpha \in \mathbb{R}$ and $x \in \mathcal{X}$, we have

$$\|\alpha x\| = \sup_{t\in[0,T]} \left|\alpha x'(t)\right| = |\alpha| \sup_{t\in[0,T]} \left|x'(t)\right| = |\alpha|\,\|x\|\,.$$

For anly $x, y \in \mathcal{X}$, we have

$$\|x+y\| = \sup_{t\in[0,T]} \left|x'(t) + y'(t)\right| \leq \sup_{t\in[0,T]} \left|x'(t)\right| + \sup_{t\in[0,T]} \left|y'(t)\right| = \|x\| + \|y\|\,.$$

Therefore $\|\cdot\|$ is a norm. To show that $\mathcal{X}$ is complete, let $\{x_n\}_{n=1}^{\infty} \subset \mathcal{X}$ be a Cauchy sequence with respect to the norm $\|\cdot\|$. Then by the definition of $\|\cdot\|$, $\{x_n'\}_{n=1}^{\infty}$ is Cauchy in $C[0,T]$, so there exists $f \in C[0,T]$ such that $\|x_n' - f\|_{\infty} \to 0$ as $n \to \infty$, where $\|\cdot\|_{\infty}$ denotes the supremum norm in $C[0,T]$. Define $x(t) = \int_0^t f(u)\,\mathrm{d}u$. Then clearly $x$ is continuously differentiable and $x(0) = 0$, so

$x \in \mathcal{X}$. Furthermore,

$$\|x_n - x\| = \sup_{t\in[0,T]} \left|x_n'(t) - x'(t)\right| = \sup_{t\in[0,T]} \left|x_n'(t) - f(t)\right| = \left\|x_n' - f\right\|_\infty \to 0,$$

so we have $x_n \to x$ in $\mathcal{X}$. Therefore $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ is a Banach space. ■

**Lemma C.1.2** (Polynomial basis)**.** *Suppose Assumption 1 holds and $h_i$ is a polynomial of degree $i$ with $h_i(0) = 0$. Then Assumption 3 holds.*

*Proof.* Since $h_i$ is a polynomial of degree $i$ with $h_i(0) = 0$, without loss of generality we may assume $h_i(t) = t^i$. Then $h_i'(t) = it^{i-1}$. By the Stone-Weierstrass theorem (Folland, 1999, p. 139), $\operatorname{span}\{h_i'\}_{i=1}^\infty$ is dense in $C[0,T]$. Since $\|x\|_{\mathcal{X}} = \|x'\|_\infty$, it follows that $\operatorname{span}\{h_i\}_{i=1}^\infty$ is dense in $\mathcal{X}$. By Assumption 1, we can choose $I$ distinct points $\{t_{n_j}\}_{j=1}^I$. Consider the $I \times I$ submatrix of $H$ defined by $\widetilde{H} = (h_i(t_{n_j})) = (t_{n_j}^i)$. Dividing the $j$-th column by $t_{n_j} > 0$, $\widetilde{H}$ reduces to a Vandermonde matrix, which is invertible. Therefore $H$ has full row rank. The same argument applies to $G$. ■

**Lemma C.1.3.** *Fix $x \in \mathcal{X}$ and define $T : \mathcal{X} \to \mathbb{R}$ by*

$$Th = \delta P(x; h) = -\int_0^T \mathrm{e}^{-x(t)} h(t)\,\mathrm{d}F(t).$$

*Then $T$ is a bounded linear operator.*

*Proof.* Clearly $T$ is a linear operator. If $h \in \mathcal{X}$, then $\|h\|_{\mathcal{X}} = \sup_{t\in[0,T]} |h'(t)| < \infty$. Since $h(0) = 0$, we obtain

$$|h(t)| = \left|\int_0^t h'(u)\,\mathrm{d}u\right| \le \int_0^t \|h\|_{\mathcal{X}}\,\mathrm{d}u = t\,\|h\|_{\mathcal{X}}.$$

Therefore

$$|Th| \le \int_0^T \mathrm{e}^{-x(t)}\,|h(t)|\,\mathrm{d}F(t) \le \|h\|_{\mathcal{X}} \int_0^T t\mathrm{e}^{-x(t)}\,\mathrm{d}F(t),$$

so $T$ is a bounded linear operator with $\|T\| \le \int_0^T t\mathrm{e}^{-x(t)}\,\mathrm{d}F(t)$. ■

## C.2 Proof of main results

*Proof of Proposition 3.3.1.* Let us first show that $\mathcal{W}$ in (3.3.5) is compact, convex, and contains 0 in the interior. Clearly $0 \in \mathcal{W}$. Since $w \mapsto G'w$ is linear (hence continuous) and $G'0 = 0$ is an

interior point of $[-1,1]^N$, 0 is an interior point of $\mathcal{W}$. Since $\mathcal{W}$ is defined by weak linear inequalities, it is closed and convex. Let us show compactness. By Assumption 3, $H$ has full row rank, and so does $G$. Take $n_1, \ldots, n_I$ such that the $I \times I$ matrix $\widetilde{G} \coloneqq (g_{i,n_j})$ is invertible. Define

$$\widetilde{\mathcal{W}} \coloneqq \left\{ w \in \mathbb{R}^I : \widetilde{G}' w \in [-1,1]^I \right\} = (\widetilde{G}')^{-1}[-1,1]^I.$$

Since $\widetilde{\mathcal{W}}$ is defined by a subset of inequalities that define $\mathcal{W}$, clearly we have $\mathcal{W} \subset \widetilde{\mathcal{W}}$. Furthermore, $\widetilde{\mathcal{W}}$ is compact because it is the image of the compact set $[-1,1]^I$ under the linear (hence continuous) map $(\widetilde{G}')^{-1} : \mathbb{R}^I \to \mathbb{R}^I$. Therefore $\mathcal{W} \subset \widetilde{\mathcal{W}}$ is compact.

Next, let us show that the minmax problem (3.3.6) has a solution $(z^*, w^*) \in \mathcal{Z} \times \mathcal{W}$. Since $\mathcal{W}$ is nonempty and compact and $w \mapsto \langle w, Az - b \rangle$ is linear (hence continuous),

$$M(z) \coloneqq \max_{w \in \mathcal{W}} \langle w, Az - b \rangle \tag{C.2.1}$$

exists. The maximum theorem (Berge, 1963, p. 116) implies that $M$ is continuous. Furthermore, since $0 \in \mathcal{W}$, we have $M(z) \geq 0$ and hence $V_I(\mathcal{Z}) = \inf_{z \in \mathcal{Z}} M(z) \geq 0$. Let $\|\cdot\|_2$ denote the $\ell^2$ (Euclidean) norm. Since 0 is an interior point of $\mathcal{W}$, there exists $\epsilon > 0$ such that $w \in \mathcal{W}$ whenever $\|w\|_2 \leq \epsilon$. If $Az \neq b$, setting $w = \epsilon \frac{Az-b}{\|Az-b\|_2}$, we obtain

$$M(z) \geq \left\langle \epsilon \frac{Az - b}{\|Az - b\|_2}, Az - b \right\rangle = \epsilon \|Az - b\|_2 . \tag{C.2.2}$$

Note that the lower bound (C.2.2) is valid even if $Az = b$.

To bound (C.2.2) from below, let us show that

$$\|Az - b\|_2 = \|A_+ z - b_+\|_2 \tag{C.2.3}$$

when $z \in \mathcal{Z}$. Using the definition (3.3.4), it suffices to show that $a_0 z - 1 = 0$ if $z \in \mathcal{Z}$. But since by Assumption 2 value matching holds, dividing (3.2.6) by $P(x)$ and using (3.3.2) for $i = 0$ (hence $h_0 \equiv 1$), we obtain

$$1 = \frac{1}{P(x)} \sum_{j=1}^{J} z_j P_j(x) = \sum_{j=1}^{J} a_{0j} z_j = a_0 z,$$

which implies (C.2.3). Define $m \coloneqq \min_{\|z\|_2=1} \|A_+ z\|_2$, which is achieved because $\|z\|_2 = 1$ is a nonempty compact set and $z \mapsto \|A_+ z\|_2$ is continuous. Since by assumption $A_+$ has full column rank, we have $A_+ z = 0$ only if $z = 0$, so $m > 0$. Therefore it follows from (C.2.2) and (C.2.3) that for any $z \in \mathcal{Z}$,

$$M(z) \geq \epsilon \|Az - b\|_2 = \epsilon \|A_+ z - b_+\|_2 \geq \epsilon(m \|z\|_2 - \|b_+\|_2) \to \infty \tag{C.2.4}$$

as $\|z\|_2 \to \infty$, so we may restrict the minimization of $M(z)$ to a compact subset of $\mathcal{Z}$. Since $M(z)$ is continuous, the minmax value $V_I(\mathcal{Z})$ is achieved.

Finally, let us show that $z \in \mathcal{Z}$ achieves $V_I(\mathcal{Z}) = 0$ if and only if $A_+ z = b_+$. If $A_+ z = b_+$, then $Az = b$ so clearly $M(z) = 0$ and $V_I(\mathcal{Z}) = 0$. If $V_I(\mathcal{Z}) = 0$, then for any $z \in \mathcal{Z}$ with $M(z) = V_I(\mathcal{Z}) = 0$, (C.2.2) and (C.2.3) imply $\|A_+ z - b_+\|_2 = 0$ and therefore $A_+ z = b_+$. ■

*Proof of Proposition 3.3.2.* Suppose that $\operatorname{span}\{\widetilde{h}_i\}_{i=1}^I = \operatorname{span}\{h_i\}_{i=1}^I$. Since $\{h_i\}_{i=1}^I$ span $\{\widetilde{h}_i\}_{i=1}^I$, there exists an $I \times I$ matrix $C = (c_{ij})$ such that $\widetilde{h}_i = \sum_{j=1}^I c_{ij} h_j$. Since $\{h_i\}_{i=1}^I$ are linearly independent, $C$ is unique. Since $\{\widetilde{h}_i\}_{i=1}^I$ also span $\{h_i\}_{i=1}^I$, $C$ must be invertible. Then $\widetilde{H} = CH$, $\widetilde{A} = CA$, $\widetilde{b} = Cb$, $\widetilde{G} = CG$, so setting $w = C'\widetilde{w}$, we obtain

$$\widetilde{M}(z) \coloneqq \sup_{\widetilde{w}:\widetilde{G}'\widetilde{w} \in [-1,1]^N} \left\langle \widetilde{w}, \widetilde{A}z - \widetilde{b} \right\rangle = \sup_{w:G'w \in [-1,1]^N} \langle w, Az - b \rangle =: M(z).$$

Therefore the minimizers of $M$ and $\widetilde{M}$ agree and the conclusion holds. ■

To prove Theorem 3.3.3, we recall Taylor's theorem with the integral form for the remainder term.

**Lemma C.2.1** (Taylor's theorem)**.** *Let $f \in C^{n+1}[0,1]$, so $f : [0,1] \to \mathbb{R}$ is $n+1$ times continuously differentiable. Then*

$$f(1) = \sum_{k=0}^{n} \frac{f^{(k)}(0)}{k!} + \int_0^1 f^{(n+1)}(s) \frac{(1-s)^n}{n!} \, \mathrm{d}s. \tag{C.2.5}$$

*Proof.* For $n = 0$, (C.2.5) is obvious from the fundamental theorem of calculus:

$$f(1) - f(0) = \int_0^1 f'(s) \, \mathrm{d}s.$$

Suppose (C.2.5) holds for some $n-1$ and consider $n$. Using integration by parts and the induction hypothesis, we obtain

$$\begin{aligned}\int_0^1 f^{(n+1)}(s)\frac{(1-s)^n}{n!}\,\mathrm{d}s &= \left[f^{(n)}(s)\frac{(1-s)^n}{n!}\right]_0^1 + \int_0^1 f^{(n)}(s)\frac{(1-s)^{n-1}}{(n-1)!}\,\mathrm{d}s \\ &= -\frac{f^{(n)}(0)}{n!} + \left(f(1) - \sum_{k=0}^{n-1}\frac{f^{(k)}(0)}{k!}\right) \\ &= f(1) - \sum_{k=0}^{n}\frac{f^{(k)}(0)}{k!},\end{aligned}$$

so (C.2.5) holds for $n$. ■

*Proof of Theorem 3.3.3.* For any $x, h \in \mathbb{R}$, define $f : [0,1] \to \mathbb{R}$ by $f(s) = \mathrm{e}^{-x-sh}$. Applying Lemma C.2.1 for $n = 1$, we obtain

$$\mathrm{e}^{-x-h} = \mathrm{e}^{-x} - \mathrm{e}^{-x}h + \int_0^1 (1-s)\mathrm{e}^{-x-sh}h^2\,\mathrm{d}s.$$

Setting $x = x(t)$ and $h = h(t)$ for $x, h \in \mathcal{X}$ and integrating both sides on $[0,T]$ with respect to $F$, we obtain

$$\begin{aligned}\int_0^T \mathrm{e}^{-x(t)-h(t)}\,\mathrm{d}F(t) = \int_0^T \mathrm{e}^{-x(t)}\,\mathrm{d}F(t) - \int_0^T \mathrm{e}^{-x(t)}h(t)\,\mathrm{d}F(t) \\ + \int_0^T\int_0^1 (1-s)\mathrm{e}^{-x(t)-sh(t)}h(t)^2\,\mathrm{d}s\,\mathrm{d}F(t).\end{aligned}$$

Using the definition of $P$ and $P'$, we obtain

$$P(x+h) = P(x) + P'(x)h + \int_0^T\int_0^1 (1-s)\mathrm{e}^{-x(t)-sh(t)}h(t)^2\,\mathrm{d}s\,\mathrm{d}F(t). \tag{C.2.6}$$

A similar equation holds for each $P_j$. Hence for any $z = (z_j) \in \mathbb{R}^J$ we have

$$P(x+h) - \sum_{j=1}^{J} z_j P_j(x+h) = E_0 + E_1 + E_2, \tag{C.2.7}$$

where

$$E_0 \coloneqq P(x) - \sum_{j=1}^{J} z_j P_j(x), \tag{C.2.8a}$$

$$E_1 \coloneqq \left( P'(x) - \sum_{j=1}^{J} z_j P_j'(x) \right) h, \tag{C.2.8b}$$

$$E_2 \coloneqq \int_0^T \int_0^1 (1-s) \mathrm{e}^{-x(t)-sh(t)} h(t)^2 \,\mathrm{d}s \,\mathrm{d}\left( F(t) - \sum_{j=1}^{J} z_j F_j(t) \right). \tag{C.2.8c}$$

Since $\mathcal{Z}$ satisfies value matching by Assumption 2, we have $E_0 = 0$ by (C.2.8a). Inspection of Assumption 3, (3.3.8), and (3.3.5) reveals that any $h \in \mathcal{H}_I(\Delta)$ can be expressed as $h = \Delta \sum_{i=1}^{I} w_i h_i$ for some $w \in \mathcal{W}$. Using (C.2.8b), (3.3.2), and (3.3.3), we obtain

$$E_1 = \left( P'(x) - \sum_{j=1}^{J} z_j P_j'(x) \right) h = \Delta P(x) \left\langle w, Az - b \right\rangle. \tag{C.2.9}$$

To bound $E_2$, note that the last integral in (C.2.6) is nonnegative because $1 - s \geq 0$ on $s \in [0,1]$ and $F$ is increasing. Furthermore, it can be bounded above by

$$\int_0^T \int_0^1 (1-s) \mathrm{e}^{-x(t)+\|h\|_\infty} \|h\|_\infty^2 \,\mathrm{d}s \,\mathrm{d}F(t) = \frac{1}{2} \|h\|_\infty^2 \,\mathrm{e}^{\|h\|_\infty} P(x).$$

Therefore $E_2$ in (C.2.8c) can be bounded as

$$-\frac{1}{2} \|h\|_\infty^2 \,\mathrm{e}^{\|h\|_\infty} \sum_{z_j \geq 0} z_j P_j(x) \leq E_2 \leq \frac{1}{2} \|h\|_\infty^2 \,\mathrm{e}^{\|h\|_\infty} \left( P(x) - \sum_{z_j < 0} z_j P_j(x) \right). \tag{C.2.10}$$

Using (3.2.6) and (3.3.7), we obtain

$$\begin{aligned} P(x) - \sum_{z_j<0} z_j P_j(x) = \sum_{z_j \geq 0} z_j P_j(x) &= \frac{1}{2} \left( P(x) + \sum_{j=1}^{J} |z_j| \, P_j(x) \right) \\ &= \frac{1}{2} P(x) \left( 1 + \sum_{j=1}^{J} |\theta_j| \right) = \frac{1}{2} P(x)(1 + \|\theta\|_1). \end{aligned} \tag{C.2.11}$$

Noting that $\|h\|_\infty \le \Delta T$ for $h \in \mathcal{H}_I(\Delta)$, it follows from (C.2.10) and (C.2.11) that

$$|E_2| \le \frac{1}{4}\Delta^2 T^2 \mathrm{e}^{\Delta T} P(x)(1 + \|\theta\|_1). \tag{C.2.12}$$

Combining (C.2.7), $E_0 = 0$, (C.2.9), and (C.2.12), we obtain

$$\begin{aligned}\langle w, Az - b\rangle - \frac{1}{4}\Delta T^2 \mathrm{e}^{\Delta T}(1 + \|\theta\|_1) \\ \le \frac{1}{\Delta P(x)}\left[P(x+h) - \sum_{j=1}^{J} z_j P_j(x+h)\right] \\ \le \langle w, Az - b\rangle + \frac{1}{4}\Delta T^2 \mathrm{e}^{\Delta T}(1 + \|\theta\|_1). \end{aligned} \tag{C.2.13}$$

Since by (3.3.7) $\theta_j$ is proportional to $z_j$, there exists some constant $c(x) > 0$ that depends only on $x$ such that $\|\theta\|_1 \le c(x)\|z\|_2$. Therefore maximizing (C.2.13) over $w \in \mathcal{W}$, it follows from the definition of $M(z)$ in (C.2.1) that

$$\begin{aligned} M(z) - \frac{1}{4}\Delta T^2 \mathrm{e}^{\Delta T}(1 + c(x)\|z\|_2) \\ \le \frac{1}{\Delta P(x)} \sup_{h \in \mathcal{H}_I(\Delta)} \left[P(x+h) - \sum_{j=1}^{J} z_j P_j(x+h)\right] \\ \le M(z) + \frac{1}{4}\Delta T^2 \mathrm{e}^{\Delta T}(1 + c(x)\|z\|_2). \end{aligned} \tag{C.2.14}$$

Let $m, \epsilon > 0$ be as in the proof of Proposition 3.3.1 and take $\bar{\Delta} > 0$ such that $\epsilon m = \frac{1}{4}\bar{\Delta} T^2 \mathrm{e}^{\bar{\Delta} T} c(x)$. Then if $0 < \Delta < \bar{\Delta}$, by (C.2.4) both sides of (C.2.14) grow to infinity as $\|z\|_2 \to \infty$. Therefore when we take the infimum of (C.2.14) as well as $M(z)$ with respect to $z \in \mathcal{Z}$, we may restrict it to some compact subset $\mathcal{Z}' \subset \mathcal{Z}$. Therefore there exists a constant $c' > 0$ such that

$$M(z) - c'\Delta \le \frac{1}{\Delta P(x)} \sup_{h \in \mathcal{H}_I(\Delta)} \left[P(x+h) - \sum_{j=1}^{J} z_j P_j(x+h)\right] \le M(z) + c'\Delta$$

for all $z \in \mathcal{Z}'$ and $\Delta \in (0, \bar{\Delta})$. Taking the infimum over $z \in \mathcal{Z}$ (which is achieved in $\mathcal{Z}'$) and letting $\Delta \to 0$, by the definition of $V_I(\mathcal{Z})$ in (3.3.6), we obtain (3.3.9).

To show the error estimate (3.3.10), let $z^* \in \mathcal{Z}$ be a solution to the minmax problem (3.3.6).

It follows from (C.2.13) that

$$\frac{1}{\Delta P(x)}\left|P(x+h)-\sum_{j=1}^{J} z_j^* P_j(x+h)\right| \le |\langle w, Az^*-b\rangle| + \frac{1}{4}\Delta T^2 \mathrm{e}^{\Delta T}(1+\|\theta\|_1).$$

Taking the supremum over $w \in \mathcal{W}$ and noting that $\mathcal{W}$ is symmetric ($w \in \mathcal{W}$ implies $-w \in \mathcal{W}$), it follows from the definition of $V_I(\mathcal{Z})$ in (3.3.6) that (3.3.10) holds. ■

*Proof of Proposition 3.3.4.* For each $I$, let $M_I(z) = \sup_{w\in\mathcal{W}_I} \langle w, A_I z - b_I\rangle$, where $A_I, b_I$ denote the matrix $A$ and vector $b$ defined by (3.3.2) and (3.3.3) and $\mathcal{W}_I$ denotes the set $\mathcal{W}$ defined by (3.3.5). Suppose $I < I'$. Letting $0_N$ denote the zero vector of $\mathbb{R}^N$, we have $\mathcal{W}_I \times \{0_{I'-I}\} \subset \mathcal{W}_{I'}$, so

$$\begin{aligned} M_I(z) &= \sup_{w\in\mathcal{W}_I} \langle w, A_I z - b_I\rangle = \sup_{w\in\mathcal{W}_I\times\{0_{I'-I}\}} \langle w, A_{I'} z - b_{I'}\rangle \\ &\le \sup_{w\in\mathcal{W}_{I'}} \langle w, A_{I'} z - b_{I'}\rangle = M_{I'}(z). \end{aligned}$$

Taking the infimum over $z \in \mathcal{Z}$, we obtain $V_I(\mathcal{Z}) \le V_{I'}(\mathcal{Z})$. Similarly,

$$V_I(\mathcal{Z}) = \inf_{z\in\mathcal{Z}} M_I(z) \ge \inf_{z\in\mathcal{Z}'} M_I(z) = V_I(\mathcal{Z}').$$ ■

*Proof of Theorem 3.3.5.* Because the proof is similar to that of Theorem 3.3.3, we only provide a sketch.

By assumption, $\mathcal{Z}_1$ in (3.3.12) is nonempty, and it is clearly closed. Hence by Proposition 3.3.1 the minmax value $V_I(\mathcal{Z}_1)$ defined by (3.3.6) is achieved by some $z^* \in \mathcal{Z}_1$. Inspection of Assumption 3, (3.3.11), and (3.3.5) reveals that any $h \in \mathcal{H}_I(\Delta_1, \Delta_2)$ can be expressed as $h = \Delta_1 v h_1 + \Delta_2 \sum_{i=1}^{I} w_i h_i$ for some $w \in \mathcal{W}$ and $v \in \mathbb{R}$ with $|v| \le \min_n 1/|h'(t_n)| =: \bar{v} \in (0,\infty)$. Applying a similar argument to the derivation of (C.2.13), we obtain

$$\begin{aligned} &\frac{1}{P(x)}\left[P(x+h)-\sum_{j=1}^{J} z_j P_j(x+h)\right] \\ &\qquad = \Delta_1 v (Az-b)_1 + \Delta_2 \langle w, Az-b\rangle + O(\Delta_1^2+\Delta_2^2), \end{aligned}$$

where $(Az-b)_1$ denotes the first entry of the vector $Az-b$. Maximizing both sides over $h \in$

$\mathcal{H}_I(\Delta, \Delta_1)$, we obtain

$$\sup_{h \in \mathcal{H}_I(\Delta,\Delta_1)} \frac{1}{P(x)} \left[ P(x+h) - \sum_{j=1}^{J} z_j P_j(x+h) \right] \\ = \Delta_1 \bar{v} \, |(Az-b)_1| + \Delta_2 M(z) + O(\Delta_1^2 + \Delta_2^2),$$

where $M(z)$ is defined by (C.2.1). Dividing both sides by $\Delta_2 > 0$ and letting $\Delta_2 \to 0$, $\Delta_1/\Delta_2 \to \infty$, and $\Delta_1^2/\Delta_2 \to 0$, the objective function remains finite only if $(Az-b)_1 = 0$, which is equivalent to $z \in \mathcal{Z}_1$. Under this condition, we have

$$\frac{1}{\Delta} \sup_{h \in \mathcal{H}_I(\Delta_1,\Delta_2)} \frac{1}{P(x)} \left[ P(x+h) - \sum_{j=1}^{J} z_j P_j(x+h) \right] = M(z) + O(\Delta_2 + \Delta_1^2/\Delta_2).$$

Minimizing over $z \in \mathcal{Z}_1$ and letting $\Delta_2 \to 0$, we obtain (3.3.13). The proof of (3.3.14) is similar. ■

*Proof of Proposition 3.3.7.* Suppose that the liability has maturity $s$ with face value 1. Then the value of the liability is

$$P(x) = \int_0^T \mathrm{e}^{-x(t)} \, \mathrm{d}F(t) = \mathrm{e}^{-x(s)}.$$

Let $z^* = A_+^{-1} b_+$ be the immunizing portfolio and assume $z^* \geq 0$. Take any perturbation $h \in \operatorname{span}\{h_i\}_{i=1}^I$ and write $h = \sum_{i=1}^I w_i h_i$. Then the funding ratio is

$$\phi(w) \coloneqq \frac{\sum_{j=1}^J z_j^* P_j(x+h)}{P(x+h)} = \sum_{j=1}^{J} z_j^* \int_0^T \mathrm{e}^{-x(t)+x(s)-h(t)+h(s)} \, \mathrm{d}F_j(t).$$

Since $z^* \geq 0$ and the exponential function is convex, $\phi(w)$ is convex in $w \in \mathbb{R}^I$.

Let us show that $\nabla\phi(0) = 0$. To this end we compute

$$\begin{aligned} \frac{\partial \phi}{\partial w_i}(0) &= \sum_{j=1}^{J} z_j^* \int_0^T \mathrm{e}^{-x(t)+x(s)}(-h_i(t) + h_i(s)) \, \mathrm{d}F_j(t) \\ &= \mathrm{e}^{x(s)} \sum_{j=1}^{J} z_j^* \left( - \int_0^T \mathrm{e}^{-x(t)} h_i(t) \, \mathrm{d}F_j(t) + h_i(s) \int_0^T \mathrm{e}^{-x(t)} \, \mathrm{d}F_j(t) \right) \\ &= \mathrm{e}^{x(s)} \left( -P(x) \sum_{j=1}^{J} a_{ij} z_j^* + h_i(s) \sum_{j=1}^{J} z_j^* P_j(x) \right), \end{aligned} \tag{C.2.15}$$

where the last line uses (3.3.2) and (3.2.4) for each bond $j$. Using value matching (3.2.6) and the fact that the liability is a zero-coupon bond, we obtain

$$h_i(s)\sum_{j=1}^{J} z_j^* P_j(x) = h_i(s)P(x) = \mathrm{e}^{-x(s)}h_i(s) = \int_0^T \mathrm{e}^{-x(t)}h_i(t)\,\mathrm{d}F(t) = P(x)b_i, \tag{C.2.16}$$

where the last equality uses (3.3.3). Combining (C.2.15) and (C.2.16), we obtain

$$\nabla\phi(0) = b - Az^* = 0. \tag{C.2.17}$$

Since $\phi$ is convex, it follows that $\phi(w) \geq \phi(0) = 1$ for all $w$, which implies (3.3.17). ■

*Proof of Lemma 3.3.8.* For $i = 1$, since $T_0(x) = 1$, we have $g_1(t) = 1$ and hence (3.3.19) implies (3.3.20a). For $i = 2$, since $T_1(x) = x$, we have $g_2(t) = \min\{2t/T - 1, 1\}$. Integrating this expression gives (3.3.20b). Suppose $i \geq 3$. Letting $x = \cos\theta$, we can evaluate the integral of Chebyshev polynomials as

$$\begin{aligned}
\int_{-1}^{x} T_n(x)\,\mathrm{d}x &= \int_{\pi}^{\theta} \cos n\theta(-\sin\theta)\,\mathrm{d}\theta \\
&= \frac{1}{2}\int_{\pi}^{\theta} (\sin(n-1)\theta - \sin(n+1)\theta)\,\mathrm{d}\theta \\
&= \frac{1}{2}\left[\frac{\cos(n+1)\theta}{n+1} - \frac{\cos(n-1)\theta}{n-1}\right]_{\pi}^{\theta} \\
&= \frac{1}{2}\left(\frac{T_{n+1}(x)}{n+1} - \frac{T_{n-1}(x)}{n-1} - \frac{2(-1)^n}{(n+1)(n-1)}\right).
\end{aligned}$$

Therefore for $i \geq 3$ we have

$$h_i(t) = \frac{1}{4}T\left(\frac{T_i(2t/T-1)}{i} - \frac{T_{i-2}(2t/T-1)}{i-2} + \frac{2(-1)^i}{i(i-2)}\right),$$

which is (1.3.3) ■

## C.3 Generic full column rank of $A_+$

This appendix shows that the matrix $A_+$ in (3.3.4) generically has full column rank, which makes Proposition 3.3.1 applicable.

**Proposition C.3.1.** *Let $I \geq J-1$, $\{h_i\}_{i=1}^I$ be the basis functions, and set $h_0 \equiv 1$. Suppose that there exist $\{m_i\}_{i=1}^J \subset \{0,1,\ldots,I\}$ with $m_1 = 0$ and $\{\tau_j\}_{j=1}^J \subset (0,T]$ such that (i) at date $\tau_j$, bond $j$ makes a lump-sum payout $f_j \coloneqq F_j(\tau_j) - F_j(\tau_j-) > 0$, and (ii) the $J \times J$ matrix $\widetilde{H} = (h_{m_i}(\tau_j))$ is invertible. Then there exists a closed set $S \subset \mathbb{R}^J$ with Lebesgue measure 0 such that the matrix $A_+$ in* (3.3.4) *has full column rank whenever $(f_1,\ldots,f_J) \notin S$.*

*If in addition all bonds are zero-coupon bonds, then $A_+$ has full column rank.*

We need the following lemma to prove Proposition C.3.1.

**Lemma C.3.2.** *Let $A, B$ be $N \times N$ matrices and define $f : \mathbb{R}^N \to \mathbb{R}$ by $f(x) = \det(A\,\mathrm{diag}(x) + B)$, where* $\mathrm{diag}(x)$ *denotes the diagonal matrix with diagonal entries $x_1,\ldots,x_N$. If* $\det A \neq 0$*, then for any $c \in \mathbb{R}$ the set*

$$f^{-1}(c) \coloneqq \big\{x \in \mathbb{R}^N : f(x) = c\big\}$$

*is closed and has Lebesgue measure* 0*.*

*Proof.* Since

$$\begin{aligned}\det(A\,\mathrm{diag}(x) + B) &= \det(A(\mathrm{diag}(x) + A^{-1}B)) \\ &= \det(A) \times \det(\mathrm{diag}(x) + A^{-1}B),\end{aligned}$$

without loss of generality we may assume that $A$ is the identity matrix. Let $B = (b_{mn})$. That $f^{-1}(c)$ is closed is obvious because $f$ is continuous.

Let us show by induction on the dimension $N$ that $f^{-1}(c)$ is a null set. If $N = 1$, then $f(x) = x_1 + b_{11}$, so $f^{-1}(c) = \{c - b_{11}\}$ is a singleton, which is a null set. Suppose the claim holds when $N = n-1$ and consider $n$. Let $B_n$ be the $n \times n$ matrix obtained from the first $n$ rows and columns of $B$, and let

$$f_n(x_1,\ldots,x_n) = \det(\mathrm{diag}(x_1,\ldots,x_n) + B_n).$$

Clearly $f_n$ is affine in each variable $x_1,\ldots,x_n$. Using the Laplace expansion along the $n$-th column, it follows that

$$f_n(x_1,\ldots,x_n) = (x_n + b_{nn}) f_{n-1}(x_1,\ldots,x_{n-1}) + g_{n-1}(x_1,\ldots,x_{n-1})$$

for some function $g_{n-1}$ that is affine in each variable $x_1, \ldots, x_{n-1}$.

Define the sets $f_{n-1}^{-1}(0) \subset \mathbb{R}^{n-1}$ and $G \subset \mathbb{R}^n$ by

$$f_{n-1}^{-1}(0) \coloneqq \{(x_1, \ldots, x_{n-1}) : f_{n-1}(x_1, \ldots, x_{n-1}) = 0\},$$
$$G \coloneqq \left\{(x_1, \ldots, x_n) : (x_1, \ldots, x_{n-1}) \notin f_{n-1}^{-1}(0), x_n = (c - g_{n-1})/f_{n-1} - b_{nn}\right\}.$$

Then $f_n^{-1}(c) \subset (f_{n-1}^{-1}(0) \times \mathbb{R}) \cup G$. By the induction hypothesis, $f_{n-1}^{-1}(0)$ has measure 0 in $\mathbb{R}^{n-1}$. Since $G$ is the graph of a Borel measurable function, by Fubini's theorem it has measure 0. Therefore $f_n^{-1}(c)$ is a null set. ■

*Proof of Proposition C.3.1.* Define $\mathbf{h} : [0, T] \to \mathbb{R}^I$ by $\mathbf{h}(t) = (h_0(t), h_1(t), \ldots, h_I(t))'$. Let the $j$-th column vector of $A_+$ be $\mathbf{a}_j = (a_{0j}, \ldots, a_{Ij})'$. By assumption, bond $j$ pays $f_j > 0$ at $\tau_j \in (0, T]$, so it follows from (3.3.2) that

$$\mathbf{a}_j = \frac{1}{P(x)} \int_{[0,T]\setminus\{\tau_j\}} \mathrm{e}^{-x(t)} \mathbf{h}(t) \, \mathrm{d}F_j(t) + \frac{1}{P(x)} \mathrm{e}^{-x(\tau_j)} f_j \mathbf{h}(\tau_j) =: \mathbf{p}_j f_j + \mathbf{q}_j. \tag{C.3.1}$$

Collecting (C.3.1) into a matrix, we can write $A_+ = P \operatorname{diag}(f) + Q$, where $P, Q$ are matrices with $j$-th column vectors $\mathbf{p}_j, \mathbf{q}_j$ and $f = (f_1, \ldots, f_J)$. To show that $A_+$ generically has full column rank, let $\widetilde{A}_+$ be the $J \times J$ matrix obtained by taking its $m_i$-th row for $i = 1, \ldots, J$. Define $\widetilde{P}, \widetilde{Q}$ similarly. Then $\widetilde{A}_+ = \widetilde{P} \operatorname{diag}(f) + \widetilde{Q}$. Since $\mathbf{p}_j = \mathrm{e}^{-x(\tau_j)} \mathbf{h}(\tau_j)/P(x)$, we obtain

$$\det \widetilde{P} = P(x)^{-J} \left( \prod_{j=1}^{J} \mathrm{e}^{-x(\tau_j)} \right) \det \widetilde{H} \neq 0.$$

Therefore by Lemma C.3.2, $\widetilde{A}_+$ is generically invertible, so $A_+$ has generically full column rank.

If in addition all bonds are zero-coupon bonds, then (C.3.1) reduces to $\mathbf{a}_j = \mathrm{e}^{-x(\tau_j)} f_j \mathbf{h}(\tau_j)/P(x)$, where $\tau_j$ is the maturity. Then $A_+ = P \operatorname{diag}(f)$, which has full column rank because $\det \widetilde{P} \neq 0$ and $f_j > 0$ for all $j$. ■

The fact that the set of $(f_1, \ldots, f_J)$ for which $A_+$ has rank deficiency is contained in a closed set with Lebesgue measure 0 implies that the set of rank deficiency is nowhere dense (has empty interior). In this sense the rank deficiency of $A_+$ is "rare". The following example shows

that the zero-coupon bond assumption in Proposition C.3.1 is essential.

**Example C.3.1.** Suppose $I = J - 1 = 1$ and the basis function is $h_1(t) = t$. Bond 1 is a zero-coupon bond with face value $f_1 > 0$ and maturity $t_1$. Bond 2 pays $f_n > 0$ at time $t_n$, where $n = 2, 3$. To simplify notation, write $x(t_1) = x_1$ etc. The determinant of the matrix $A_+$ is

$$\begin{aligned}\det A_+ &= P(x)^{-2} \det \begin{bmatrix} f_1 \mathrm{e}^{-x_1} & f_2 \mathrm{e}^{-x_2} + f_3 \mathrm{e}^{-x_3} \\ f_1 \mathrm{e}^{-x_1} t_1 & f_2 \mathrm{e}^{-x_2} t_2 + f_3 \mathrm{e}^{-x_3} t_3 \end{bmatrix} \\ &= P(x)^{-2} f_1 \mathrm{e}^{-x_1} \left( f_2 \mathrm{e}^{-x_2} (t_2 - t_1) + f_3 \mathrm{e}^{-x_3} (t_3 - t_1) \right).\end{aligned}$$

Therefore for any $t_2 < t_1 < t_3$ and $f_3 > 0$, we have $\det A_+ = 0$ if and only if

$$(f_1, f_2) \in \left\{ (f_1, f_2) \in \mathbb{R}^2_{++} : f_2 = f_3 \mathrm{e}^{x_2 - x_3} \frac{t_3 - t_1}{t_1 - t_2} \right\}. \tag{C.3.2}$$

The closure of the rank deficiency set (C.3.2) is a ray in $\mathbb{R}^2$ and has measure 0.

## C.4 No-arbitrage term structure model

The no-arbitrage term structure model of Ang, Bekaert, and Wei (2008) features multiple factors, regime switching, and closed-form solutions for bond prices, which is convenient for simulating yield curves. This appendix summarizes their model and presents parameter estimates based on our yield curve data.

### C.4.1 Model and bond price formula

The equation numbers follow that of Ang, Bekaert, and Wei (2008). The model has three factors denoted by $X_t = (q_t, f_t, \pi_t)'$. The dynamics of factors follows the regime-dependent VAR process

$$X_{t+1} = \mu(s_{t+1}) + \Phi X_t + \Sigma(s_{t+1}) \varepsilon_{t+1}, \tag{2}$$

where

$$\mu(s_t) = \begin{bmatrix} \mu_q \\ \mu_f(s_t) \\ \mu_\pi(s_t) \end{bmatrix}, \qquad \Phi = \begin{bmatrix} \Phi_{qq} & 0 & 0 \\ \Phi_{fq} & \Phi_{ff} & 0 \\ \Phi_{\pi q} & \Phi_{\pi f} & \Phi_{\pi\pi} \end{bmatrix}, \qquad \Sigma(s_t) = \begin{bmatrix} \sigma_q & 0 & 0 \\ 0 & \sigma_f(s_t) & 0 \\ 0 & 0 & \sigma_\pi(s_t) \end{bmatrix}, \quad (3)$$

and $\varepsilon$ is IID $N(0, I_3)$. The regime $s_t$ is a finite-state Markov chain taking values denoted by $k = 1, \ldots, K$ with transition probability matrix $\Pi = (p_{kk'})$. The real short rate is given by

$$\widehat{r}_t = \delta_0 + \delta_1' X_t. \quad (4)$$

The regime-dependent price of risk is denoted by $\lambda(s_t) = (\lambda_f(s_t), \lambda_\pi(s_t))'$. Furthermore, define

$$\gamma_t = \gamma_0 + \gamma_1 q_t = \gamma_0 + \gamma_1 e_1' X_t, \quad (6)$$

where $e_n$ denotes the $n$-th unit vector.

With this notation, the price of zero-coupon bonds can be obtained in closed form (Ang, Bekaert, and Wei, 2008, Proposition B). For each maturity $n$, the nominal zero-coupon bond price in regime $i$ and factor $X$ is given by

$$P_n(i, X) = \exp(A_n(i) + B_n X), \quad (B1)$$

where the scalar $A_n(i)$ and the $M \times 1$ vector $B_n$ can be computed as follows.

Let $M = 3$ be the number of factors and $M_1 = 2$ be the number of non-$q$ factors. Partition $B_n$ as $B_n = [B_{nq}; B_{nx}]$, where $B_{nq}$ is a scalar and $B_{nx}$ is $2 \times 1$. Similarly, let $\Sigma_x(i)$ be the lower $2 \times 2$ block of $\Sigma(i)$.

First, define $A_0(i) = 0$ and $B_0 = 0$. Then define $\{(A_n, B_n)\}_{n=1}^{\infty}$ recursively by

$$\begin{aligned} A_{n+1}(i) = -\delta_0 - B_{nq}\sigma_q\gamma_0 + \log \sum_j p_{ij} \exp\Big( & A_n(j) + (B_n - e_M)'\mu(j) \\ & - (B_{nx} - e_{M_1})'\Sigma_x(j)\lambda(j) + \frac{1}{2}(B_n - e_M)'\Sigma(j)\Sigma(j)'(B_n - e_M)\Big), \end{aligned} \quad (B2.a)$$

$$B_{n+1} = -\delta_1 + \Phi'(B_n - e_M) - B_{nq}\sigma_q\gamma_1 e_1. \quad (B2.b)$$

### C.4.2 Data

We use end of the quarter yield data from Liu and Wu (2021) for the period of 1985:Q4 to 2022:Q4; a total of 149 quarterly observations.[1] The authors use a nonparametric approach to estimate the yield curve up to the 30-year maturity, which allows us to infer the long end of the yield curve consistently over time. The inflation data for the same period are obtained from the Bureau of Labor Statistics, from the *CPI for All Urban Consumers series* (seasonally adjusted).

In our dynamic hedging experiment, we need to infer the yields up to a maturity of 50 years. Estimating the model of Ang, Bekaert, and Wei (2008), we sometimes find counterfactual steep declines in the yield curve for long maturities, depending on the maturities used for estimation. To mitigate this issue, we incorporate 50-year yields in the estimation, treating them as equivalent to observed 30-year yields. This inclusion proves essential for generating yield curves that remain relatively "flat" over long horizons, thereby preventing the possibility of counterfactual steep declines at the long end of the yield curve. Additionally, we incorporate 1-year yields in the estimation to capture short-run dynamics.[2]

### C.4.3 Parameter estimates

We consider the benchmark model $\mathrm{IV}^{\mathrm{C}}$ of Ang, Bekaert, and Wei (2008, §I.B.4). This model has four regimes. There are two state variables denoted by $s^f, s^\pi$, which both take values in $\{1, 2\}$. The combined state $s$ thus takes four values

$$s = 1 \coloneqq (s^f = 1, s^\pi = 1),$$
$$s = 2 \coloneqq (s^f = 1, s^\pi = 2),$$
$$s = 3 \coloneqq (s^f = 2, s^\pi = 1),$$
$$s = 4 \coloneqq (s^f = 2, s^\pi = 2).$$

[1] https://sites.google.com/view/jingcynthiawu/yield-data

[2] Unlike Ang, Bekaert, and Wei (2008), we do not use additional yield data as overidentifying restrictions.

We also impose the following restrictions consistent with Ang, Bekaert, and Wei (2008):

$$\begin{aligned}
\delta_0 &= 0.0077, && \text{(mean of nominal short rate)} \\
\delta_1 &= (1, 1, \delta_\pi)', \\
\Phi_{fq} &= 0, \\
\mu_q &= 0, \\
\gamma_0 &= 0, \\
\lambda_\pi(s_t) &= 0.
\end{aligned}$$

We estimate the model using maximum likelihood using the parameters from Ang, Bekaert, and Wei (2008, Table III) as starting values. Table C.1 below summarizes the resulting parameter estimates.

**Table C.1. Parameter estimates**

Note: This table shows parameter estimates from the regime switching model of Ang, Bekaert, and Wei (2008).

| Real short rate | | $\delta_1$ | | |
|---|---|---|---|---|
| | $\delta_0$ | $q$ | $f$ | $\pi$ |
| | 0.008 | 1.000 | 1.000 | -0.199 |
| Companion Form $\Phi$ | | $q$ | $f$ | $\pi$ |
| | $q$ | 0.962 | 0.000 | 0.000 |
| | $f$ | 0.000 | 0.969 | 0.000 |
| | $\pi$ | -0.139 | 0.246 | 0.178 |
| Moments of $X_t$ | | | | |
| | | | Regime 1 | Regime 2 |
| | | $\mu_q \times 100$ | 0.000 | 0.000 |
| | | $\mu_f(s_t^f) \times 100$ | -0.621 | -0.020 |
| | | $\mu_\pi(s_t^\pi) \times 100$ | -0.789 | 0.726 |
| | | $\sigma_q \times 100$ | 0.054 | 0.054 |
| | | $\sigma_f(s_t^f) \times 100$ | 0.400 | 0.108 |
| | | $\sigma_\pi(s_t^\pi) \times 100$ | 0.048 | 0.624 |
| Prices of Risk | | | $\lambda_f(s_t^\pi)$ | |
| | | $\gamma_1$ | Regime 1 | Regime 2 |
| | | -84.137 | -19.734 | 0.051 |
| Transition Probabilities $\Pi$ | | | | |
| | $s_{t+1}=1$ | $s_{t+1}=2$ | $s_{t+1}=3$ | $s_{t+1}=4$ |
| $s_t=1$ | 0.744 | 0.174 | 0.037 | 0.045 |
| $s_t=2$ | 0.685 | 0.216 | 0.052 | 0.047 |
| $s_t=3$ | 0.001 | 0.001 | 0.354 | 0.645 |
| $s_t=4$ | 0.000 | 0.000 | 0.020 | 0.980 |

## C.5 Miscellaneous results

### C.5.1 Bias in the estimated yield curve

In our empirical application in Section 3.4, we assume that the forward rate is constant beyond the 30-year maturity, $f(t) = f(30)$ for all $t \geq 30$. As a result, the inferred date $s$ yield curve with term $t \geq 30$ satisfies[3]

$$\begin{aligned}\widehat{y}_s(t) &:= \frac{1}{t}\int_0^t f_s(u)\,\mathrm{d}u = \frac{1}{t}\int_0^{30} f_s(u)\,\mathrm{d}u + \frac{1}{t}\int_{30}^t f_s(30)\,\mathrm{d}u \\ &= \frac{1}{t}\int_0^{30} f_s(u)\,\mathrm{d}u + f_s(30) - \frac{30}{t}f_s(30) \\ &= f_s(30) + O\left(\frac{1}{t}\right).\end{aligned}$$

Taking unconditional expectations and comparing to the true (unobserved) yield, we obtain

$$\begin{aligned}\mathbb{E}\left[\widehat{y}_s(t) - y_s(t)\right] &= \mathbb{E}\left[f_s(30) - y_s(t)\right] + O\left(\frac{1}{t}\right) \\ &= \mathbb{E}\left[f_s(30) - f_s(t)\right] + \mathbb{E}\left[f_s(t) - y_s(t)\right] + O\left(\frac{1}{t}\right). \qquad \text{(C.5.1)}\end{aligned}$$

Under integrability conditions on $y_s(t)$ and a mild stationarity assumption on bond returns, Alvarez and Jermann (2005a, Proposition 5) show that

$$\mathbb{E}\left[\lim_{t\to\infty} f_s(t)\right] = \mathbb{E}\left[\lim_{t\to\infty} y_s(t)\right]. \qquad \text{(C.5.2)}$$

Using the dominated convergence theorem and (C.5.2) in (C.5.1), we get

$$\mathbb{E}\left[\widehat{y}_s(t)\right] = \mathbb{E}\left[y_s(t)\right] + \mathbb{E}\left[f_s(30) - f_s(t)\right] + o(1).$$

Hence, on average we estimate the correct yield plus a bias term that reflects the average gap between the 30-year forward rate and long forward rate.

[3] Throughout we ignore the approximation error coming from misspecification of the forward rate model.

### C.5.2 Approximating forward rate changes

In this appendix we evaluate the goodness-of-fit of approximating forward rate changes by basis functions. Let $I$ be the number of basis functions to include, $d$ be the number of days ahead, and $\{t_n\}_{n=1}^N$ be the set of terms (in years) to evaluate forward rates, where we set $t_n = n/12$ and $N = 360$ so that it corresponds to a 30-year horizon at monthly interval. We use the following procedure.

(i) For each day $s$ and term $t_n$, calculate the $d$-day ahead change in the forward rate $f_{s+d}(t_n) - f_s(t_n)$ by evaluating (3.4.1).

(ii) Estimate

$$f_{s+d}(t_n) - f_s(t_n) = \sum_{i=1}^{I} \gamma_{is} g_i(t_n) + \epsilon_s(t_n) \tag{C.5.3}$$

by ordinary least squares (OLS), where $g_i$ is the basis function for the forward rate in (3.3.18). Let $\widehat{\gamma}_{is}$ be the OLS estimator.

(iii) Calculate the goodness-of-fit measure

$$R^2 := \frac{\sum_{s=1}^{S} \sum_{n=1}^{N} \left( \sum_{i=1}^{I} \widehat{\gamma}_{is} g_i(t_n) \right)^2}{\sum_{s=1}^{S} \sum_{n=1}^{N} (f_{s+d}(t_n) - f_s(t_n))^2}. \tag{C.5.4}$$

The goodness-of-fit measure (C.5.4) is similar to the conventional $R^2$ in OLS, except that we use "0" as the benchmark instead of the sample mean because $g_1 \equiv 1$ is already constant. The following proposition shows that $R^2$ in (C.5.4) can be computed efficiently.

**Proposition C.5.1** (Efficient calculation of $R^2$)**.** *Define the $S \times N$ matrix $C = (c_{sn})$ by $c_{sn} = f_{s+d}(t_n) - f_s(t_n)$ and the $I \times N$ matrix $G = (g_i(t_n))$. Then*

$$R^2 = \frac{\mathrm{tr}(G'(GG')^{-1}GC'C)}{\mathrm{tr}(C'C)}, \tag{C.5.5}$$

*where* tr *denotes the trace (sum of diagonal entries) of the square matrix.*

*Proof.* The $n$-th diagonal entry of the $N \times N$ matrix $C'C$ is $\sum_{s=1}^{S} c_{sn}^2$. Therefore the denominator

of (C.5.4) is

$$\sum_{s=1}^{S}\sum_{n=1}^{N}(f_{s+d}(t_n) - f_s(t_n))^2 = \sum_{n=1}^{N}\sum_{s=1}^{S} c_{sn}^2 = \mathrm{tr}(C'C),$$

which is the denominator of (C.5.5).

Stacking (C.5.3) into an $N \times 1$ vector and using $G = (g_i(t_n))$, we obtain

$$c_s = G'\gamma_s + \epsilon_s,$$

where $c_s = (c_{sn})_{n=1}^N$, $\gamma_s = (\gamma_{is})_{i=1}^I$, and $\epsilon_s = (\epsilon_s(t_n))_{n=1}^N$. Therefore the OLS estimator is $\widehat{\gamma}_s = (GG')^{-1}G'c_s$ and the $N \times 1$ vector of fitted values is

$$\widehat{c}_s \coloneqq G'\widehat{\gamma}s = G'(GG')^{-1}Gc_s.$$

Stacking this vector for $s = 1, \ldots, S$ and taking the transpose, we can define the $S \times N$ matrix of fitted values $\widehat{C} = (\widehat{c}_{sn})$ by

$$\widehat{C} \coloneqq CG'(GG')^{-1}G.$$

By the same argument as the case with the denominator and using the property $\mathrm{tr}(AB) = \mathrm{tr}(BA)$, the numerator of (C.5.4) becomes

$$\begin{aligned}\sum_{s=1}^{S}\sum_{n=1}^{N}\widehat{c}_{sn}^2 &= \mathrm{tr}(\widehat{C}'\widehat{C}) = \mathrm{tr}(\widehat{C}\widehat{C}') \\ &= \mathrm{tr}(CG'(GG')^{-1}GG'(GG')^{-1}GC') \\ &= \mathrm{tr}(CG'(GG')^{-1}GC') \\ &= \mathrm{tr}(G'(GG')^{-1}GC'C),\end{aligned}$$

which is the numerator of (C.5.5). ■

### C.5.3 Key rate duration matching

This appendix explains the key rate duration matching method of Ho (1992). The key rate duration of a bond with yield curve $y$ and yield change $\Delta$ at time to maturity $t$ is defined by

$$\mathrm{KRD}(y, t, \Delta) \coloneqq \frac{P(y_-) - P(y_+)}{2\Delta P(y)},$$

where $y_\pm$ denotes the yield curve after changing $y(t)$ to $y(t) \pm \Delta$ at a specific term $t$ and linearly interpolating between the adjacent terms. Following the literature, we set the shift to $\Delta = 0.01$ (100 basis points).

Figure C.1 illustrates the procedure for a set of key rates on December 2, 2016. Key rate duration matching amounts to matching the key rate of liabilities at maturities $\{t_j\}_{j=1}^{J}$ using a portfolio of zero-coupon bonds with the same maturities.[4]

### C.5.4 Sign test

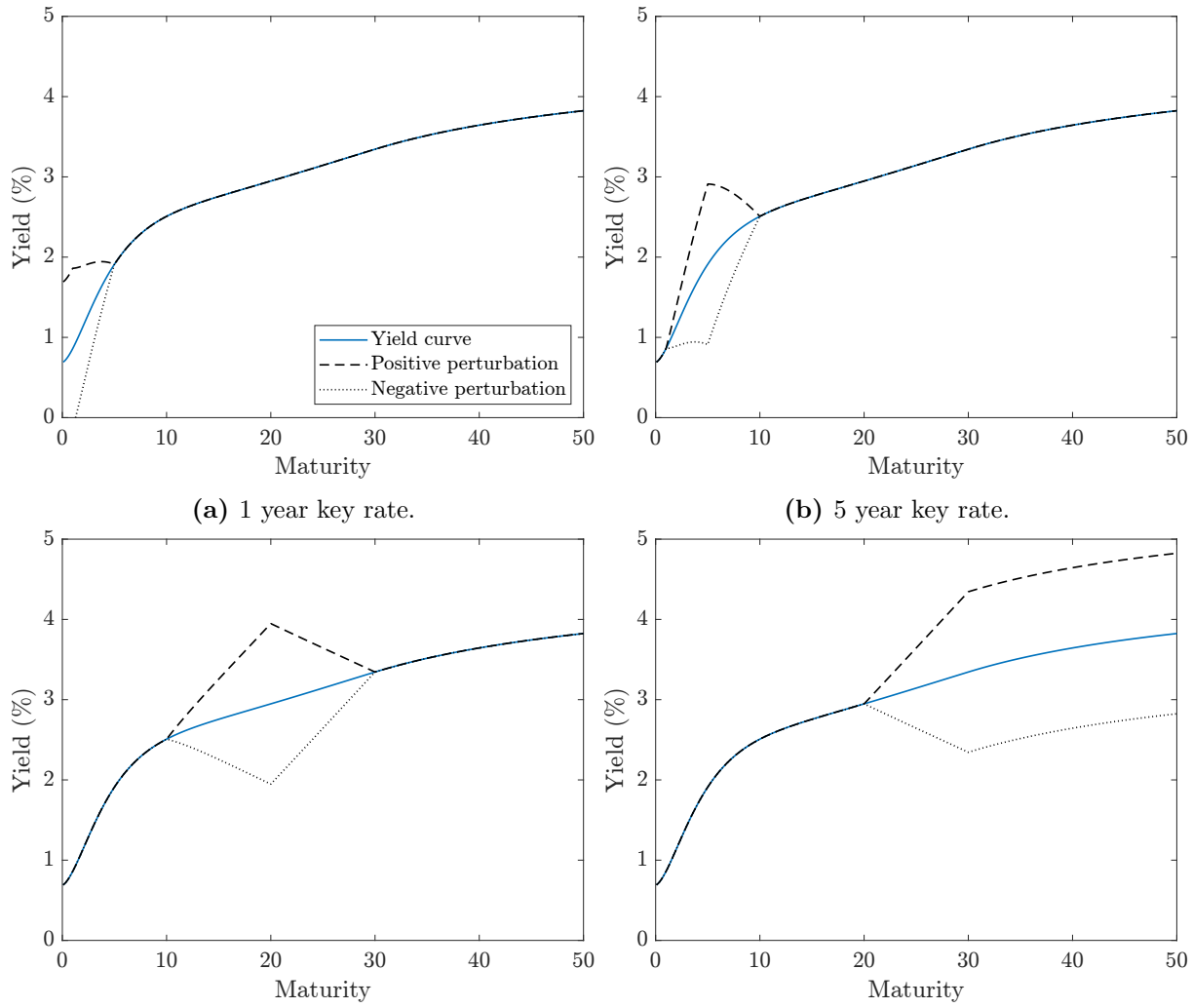
In this section, we test whether the absolute return error of `RI`(1) is significantly better compared to `HD` or `KRD`. For `RI`(1) and `KRD`, we use 5 bonds and for `HD` we use 3 bonds since the performance with more bonds is comparatively worse. Subsequently, we calculate the 30-day absolute return error (3.4.2) for non-overlapping sample periods, starting at November 25, 1985. This procedure renders a total of 304 return error observations. Let us denote the return errors for each method by $e_{\texttt{RI(1)}}, e_{\texttt{HD}}$ and $e_{\texttt{KRD}}$. Under the (one-sided) null and alternative hypothesis, we have

$$H_0 : \mathrm{P}(e_{\texttt{RI(1)}} > e_{\texttt{HD}}) \geq 0.5 \quad \text{vs.} \quad H_1 : \mathrm{P}(e_{\texttt{RI(1)}} > e_{\texttt{HD}}) < 0.5 \tag{C.5.6a}$$

$$H_0 : \mathrm{P}(e_{\texttt{RI(1)}} > e_{\texttt{KRD}}) \geq 0.5 \quad \text{vs.} \quad H_1 : \mathrm{P}(e_{\texttt{RI(1)}} > e_{\texttt{KRD}}) < 0.5. \tag{C.5.6b}$$

The test statistic for the sign test counts the number of positive differences between $e_{\texttt{RI(1)}}$ and the error term of the alternative method. Under $H_0$, this test statistic follows a binomial distribution with success probability $p = 0.5$. Using the normal approximation to the binomial distribution,

[4] The key rate duration of a zero-coupon bond with maturity $t$ is equal to $t$ and zero otherwise. Since we use linear interpolation after a key rate perturbation to keep the yield curve continuous, the key rate for a zero-coupon bond with maturity $t$ is not exactly equal to $t$ in our application.




**(a)** 1 year key rate.

**(b)** 5 year key rate.

**(c)** 20 year key rate.

**(d)** 30 year key rate.

**Figure C.1. Key rate perturbations**

Note: The figures show positive and negative perturbations to the yield curve due to a 1% change in the respective key rate. We linearly interpolate the yields after a change in the key rate to ensure that the yield curve remains continuous. The true yield curve (in blue) is calculated on December 2, 2016.

we find $Z$-scores of $-5.22$ and $-9.12$ corresponding to the hypotheses (C.5.6a) and (C.5.6b). Both test scores are sufficient to reject $H_0$ under any conventional significance level.

# Bibliography


Yacine Aït-Sahalia and Andrew W. Lo. Nonparametric estimation of state-price densities implicit in financial asset prices. *Journal of Finance*, 53(2):499–547, 1998. doi:10.1111/0022-1082.215228.

Yacine Aït-Sahalia and Andrew W. Lo. Nonparametric risk management and implied risk aversion. *Journal of Econometrics*, 94(1):9–51, 2000. doi:10.1016/S0304-4076(99)00016-0.

Caio Almeida and René Garcia. Assessing misspecified asset pricing models with empirical likelihood estimators. *Journal of Econometrics*, 170(2):519–537, 2012. doi:10.1016/j.jeconom.2012.05.020.

Gosse A.G. Alserda, Jacob A. Bikker, and Fieke S.G. Van Der Lecq. X-efficiency and economies of scale in pension fund administration and investment. *Applied Economics*, 50(48):5164–5188, 2018. doi:10.1080/00036846.2018.1486011.

Fernando Alvarez and Urban J. Jermann. Using asset prices to measure the persistence of the marginal utility of wealth. *Econometrica*, 73(6):1977–2016, 2005a. doi:10.1111/j.1468-0262.2005.00643.x.

Fernando Alvarez and Urban J. Jermann. Using asset prices to measure the persistence of the marginal utility of wealth. *Econometrica*, 73(6):1977–2016, 2005b. doi:10.1111/j.1468-0262.2005.00643.x.

Torben G. Andersen, Oleg Bondarenko, Viktor Todorov, and George Tauchen. The fine structure of equity-index option dynamics. *Journal of Econometrics*, 187(2):532–546, 2015. doi:10.1016/j.jeconom.2015.02.037.

Aleksandar Andonov, Nils Kok, and Piet Eichholtz. A global perspective on pension fund investments in real estate. *Journal of Portfolio Management*, 39(5):32–42, 2013. doi:10.3905/jpm.2013.39.5.032.

Andrew Ang, Geert Bekaert, and Min Wei. The term structure of real rates and expected inflation. *Journal of Finance*, 63(2):797–849, April 2008. doi:10.1111/j.1540-6261.2008.01332.x.

Joshua Angrist, Victor Chernozhukov, and Iván Fernández-Val. Quantile regression under misspecification, with an application to the U.S. wage structure. *Econometrica*, 74(2):539–563, 2006. doi:10.1111/j.1468-0262.2006.00671.x.

David Backus, Mikhail Chernov, and Ian Martin. Disasters implied by equity index options. *Journal*

*of Finance*, 66(6):1969–2012, 2011. doi:10.1111/j.1540-6261.2011.01697.x.

David Backus, Mikhail Chernov, and Stanley Zin. Sources of entropy in representative agent models. *Journal of Finance*, 69(1):51–99, 2014. doi:10.1111/jofi.12090.

Gurdip Bakshi, Dilip Madan, and George Panayotov. Returns of claims on the upside and the viability of U-shaped pricing kernels. *Journal of Financial Economics*, 97(1):130–154, 2010. doi:10.1016/j.jfineco.2010.03.009.

Gurdip Bakshi, Fousseni Chabi-Yo, and Xiaohui Gao. A recovery that we can trust? Deducing and testing the restrictions of the recovery theorem. *Review of Financial Studies*, 31(2):532–555, 2018. doi:10.1093/rfs/hhx108.

Ravi Bansal and Bruce N. Lehmann. Growth-optimal portfolio restrictions on asset pricing models. *Macroeconomic Dynamics*, 1(2):333–354, 1997. doi:10.1017/S1365100597003039.

Ravi Bansal and Amir Yaron. Risks for the long run: A potential resolution of asset pricing puzzles. *Journal of Finance*, 59(4):1481–1509, 2004. doi:10.1111/j.1540-6261.2004.00670.x.

Andrea Barletta and Paolo Santucci de Magistris. Analyzing the risks embedded in option prices with *rndfittool*. *Risks*, 6(2):1–15, 2018. doi:10.3390/risks6020028.

Robert J. Barro. Rare disasters and asset markets in the twentieth century. *Quarterly Journal of Economics*, 121(3):823–866, 2006. doi:10.1162/qjec.121.3.823.

Robert J. Barro. Rare disasters, asset prices, and welfare costs. *American Economic Review*, 99 (1):243–264, 2009. doi:10.1257/aer.99.1.243.

David S. Bates. The market for crash risk. *Journal of Economic Dynamics and Control*, 32(7): 2291–2321, 2008. doi:10.1016/j.jedc.2007.09.020.

Rob Bauer, K. J. Martijn Cremers, and Rik Frehen. Pension fund performance and costs: Small is beautiful. Available at SSRN: https://ssrn.com/abstract=965388, 2010.

Brendan K. Beare and Lawrence D. W. Schmidt. An empirical test of pricing kernel monotonicity. *Journal of Applied Econometrics*, 31(2):338–356, 2016. doi:10.1002/jae.2422.

Tyler Beason and David Schreindorfer. Dissecting the equity premium. *Journal of Political Economy*, 130(8):2203–2222, 2022. doi:10.1086/720396.

Alexander Beath, Chris Flynn, Rashay Jethalal, and Michael Reid. A case for scale: How the world's largest institutional investors leverage scale to deliver real outperformance. *CEM Benchmarking White Paper*, 2022.

Claude Berge. *Topological Spaces*. Oliver & Boyd, Edinburgh, 1963.

Dirk Bergemann and Stephen Morris. Robust mechanism design. *Econometrica*, 73(6):1771–1813, November 2005. doi:10.1111/j.1468-0262.2005.00638.x.

Jonathan B. Berk and Richard C. Green. Mutual fund flows and performance in rational markets. *Journal of Political Economy*, 112(6):1269–1295, 2004. doi:10.1086/424739.

Gerald O. Bierwag and Chulsoon Khang. An immunization strategy is a minimax strategy. *Journal of Finance*, 34(2):389–399, May 1979. doi:10.2307/2326978.

Jacob A. Bikker. Is there an optimal pension fund size? A scale-economy analysis of administrative and investment costs. In *Pension Fund Economics and Finance*, pages 9–40. Routledge, 2017.

Fischer Black and Myron Scholes. The pricing of options and corporate liabilities. *Journal of Political Economy*, 81(3):637–654, 1973a. doi:10.1086/260062.

Fischer Black and Myron Scholes. The pricing of options and corporate liabilities. *Journal of Political Economy*, 81(3):637–654, 1973b. doi:10.1086/260062.

David Blake, Alberto G. Rossi, Allan Timmermann, Ian Tonks, and Russ Wermers. Decentralized investment management: Evidence from the pension fund industry. *Journal of Finance*, 68(3): 1133–1178, 2013. doi:10.1111/jofi.12024.

Tim Bollerslev and Viktor Todorov. Tails, fears, and risk premia. *Journal of Finance*, 66(6): 2165–2211, 2011. doi:10.1111/j.1540-6261.2011.01695.x.

Tim Bollerslev, Viktor Todorov, and Lai Xu. Tail risk premia and return predictability. *Journal of Financial Economics*, 118(1):113–134, 2015. doi:10.1016/j.jfineco.2015.02.010.

Jaroslav Borovička, Lars Peter Hansen, and José A. Scheinkman. Misspecified recovery. *Journal of Finance*, 71(6):2493–2544, 2016. doi:10.1111/jofi.12404.

Douglas T. Breeden and Robert H. Litzenberger. Prices of state-contingent claims implicit in option prices. *Journal of Business*, 51(4):621–651, 1978. doi:10.1086/296025.

Mark Broadie, Mikhail Chernov, and Michael Johannes. Understanding index option returns. *Review of Financial Studies*, 22(11):4493–4529, 2009. doi:10.1093/rfs/hhp032.

Dirk W.G.A. Broeders, Arco van Oord, and David R. Rijsbergen. Scale economies in pension fund investments: A dissection of investment costs across asset classes. *Journal of International Money and Finance*, 67:147–171, 2016. doi:10.1016/j.jimonfin.2016.04.003.

Benjamin Brooks and Songzi Du. Optimal auction design with common values: An informationally robust approach. *Econometrica*, 89(3):1313–1360, May 2021. doi:10.3982/ECTA16297.

John Y. Campbell and John H. Cochrane. By force of habit: A consumption-based explanation of aggregate stock market behavior. *Journal of Political Economy*, 107(2):205–251, 1999.

John Y. Campbell and Samuel B. Thompson. Predicting excess stock returns out of sample: Can anything beat the historical average? *Review of Financial Studies*, 21(4):1509–1531, 2008. doi:10.1093/rfs/hhm055.

George Casella and Roger L. Berger. *Statistical Inference*. Duxbury/Thomson Learning. Belmont, CA, 2002.

Fousseni Chabi-Yo and Johnathan Loudis. The conditional expected market return. *Journal of Financial Economics*, 137(3):752–786, 2020. doi:10.1016/j.jfineco.2020.03.009.

Fousseni Chabi-Yo and Johnathan A. Loudis. A decomposition of conditional risk premia and implications for representative agent models. *Management Science*, 2023. doi:10.1287/mnsc.2022.01663.

Fousseni Chabi-Yo, René Garcia, and Eric Renault. State dependence can explain the risk aversion puzzle. *Review of Financial Studies*, 21(2):973–1011, 2008. doi:10.1093/rfs/hhm070.

Donald R. Chambers, Willard T. Carleton, and Richard W. McEnally. Immunizing default-free bond portfolios with a duration vector. *Journal of Financial and Quantitative Analysis*, 23(1): 89–104, March 1988. doi:10.2307/2331026.

Liang Chen, Juan J. Dolado, and Jesús Gonzalo. Quantile factor models. *Econometrica*, 89(2): 875–910, 2021. doi:10.3982/ECTA15746.

Victor Chernozhukov, Iván Fernández-Val, and Siyi Luo. Distribution regression with sample selection, with an application to wage decompositions in the UK, 2018. URL https://arxiv.org/abs/1811.11603.

John H. Cochrane. *Asset Pricing*. Princeton University Press, 2005.

George M. Constantinides and Anisha Ghosh. Asset pricing with countercyclical household consumption risk. *Journal of Finance*, 72(1):415–460, 2017. doi:10.1111/jofi.12471.

John G. Cragg. Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica*, 39(5):829–844, 1971.

Richard K. Crump and Nikolay Gospodinov. On the factor structure of bond returns. *Econometrica*, 90(1):295–314, January 2022. doi:10.3982/ECTA17943.

Horatio Cuesdeanu and Jens Carsten Jackwerth. The pricing kernel puzzle in forward looking data. *Review of Derivatives Research*, 21(3):253–276, 2018. doi:10.1007/s11147-017-9140-8.

Jon Danielsson and Casper G. de Vries. Value-at-risk and extreme returns. *Annales d'Économie et de Statistique*, (60):239–270, 2000. doi:10.2307/20076262.

Francis X. Diebold and Robert S. Mariano. Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):134–144, 1995. doi:10.1198/073500102753410444.

Itamar Drechsler and Amir Yaron. What's vol got to do with it. *Review of Financial Studies*, 24 (1):1–45, 2011. doi:10.1093/rfs/hhq085.

Songzi Du. Robust mechanisms under common valuation. *Econometrica*, 86(5):1569–1588, Septem-

ber 2018. doi:10.3982/ECTA14993.

Philip H. Dybvig, Jonathan E. Ingersoll, and Stephen A. Ross. Long forward and zero-coupon rates can never fall. *Journal of Business*, 69(1):1–25, January 1996. doi:10.1086/209677.

Alexander Dyck and Lukasz Pomorski. Is bigger better? Size and performance in pension plan management. Rotman School of Management Working Paper No. 1690724, Available at SSRN: https://ssrn.com/abstract=1690724, 2011.

Alexander Dyck and Lukasz Pomorski. Investor scale and performance in private equity investments. *Review of Finance*, 20(3):1081–1106, 2016. doi:10.1093/rof/rfv030.

Louis Eeckhoudt and Harris Schlesinger. Putting risk in its proper place. *American Economic Review*, 96(1):280–289, 2006. doi:10.1257/000282806776157777.

Robert F. Engle and Simone Manganelli. CAViaR: Conditional autoregressive value at risk by regression quantiles. *Journal of Business & Economic Statistics*, 22(4):367–381, 2004. doi:10.1198/073500104000000370.

Eugene F. Fama and Kenneth R. French. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33:23–49, 1993. doi:10.1016/0304-405X(93)90023-5.

Emmanuel Farhi and François Gourio. Accounting for macro-finance trends: Market power, intangibles, and risk premia. *Brookings Papers on Economic Activity*, (2):147–250, 2018. doi:10.1353/eca.2018.0024.

Stephen Figlewski. Estimating the implied risk-neutral density for the US market portfolio. In *Volatility and Time Series Econometrics: Essays in Honor of Robert Engle*, chapter 15, pages 323–353. Oxford University Press, 03 2010. doi:10.1093/acprof:oso/9780199549498.003.0015.

Damir Filipović, Eberhard Mayerhofer, and Paul Schneider. Density approximations for multivariate affine jump-diffusion processes. *Journal of Econometrics*, 176(2):93–111, 2013. doi:10.1016/j.jeconom.2012.12.003.

Lawrence Fisher and Roman L. Weil. Coping with the risk of interest-rate fluctuations: Returns to bondholders from naïve and optimal strategies. *Journal of Business*, 44(4):408–431, October 1971. doi:10.1086/295402.

Gerald B. Folland. *Real Analysis: Modern Techniques and Their Applications*. John Wiley & Sons, Hoboken, NJ, 2 edition, 1999.

Gifford H. Fong and Oldrich A. Vasicek. A risk minimizing strategy for portfolio immunization. *Journal of Finance*, 39(5):1541–1546, December 1984. doi:10.1111/j.1540-6261.1984.tb04923.x.

Kenneth R. French. Presidential address: The cost of active investing. *Journal of Finance*, 63(4): 1537–1573, 2008. doi:10.1111/j.1540-6261.2008.01368.x.

William Fung and David A. Hsieh. The risk in hedge fund strategies: Theory and evidence from

trend followers. *Review of Financial Studies*, 14(2):313–341, 2001. doi:10.1093/rfs/14.2.313.

Xavier Gabaix. Power laws in economics and finance. *Annual Review of Economics*, 1(1):255–294, 2009. doi:10.1146/annurev.economics.050708.142940.

Xavier Gabaix. Variable rare disasters: An exactly solved framework for ten puzzles in macro-finance. *Quarterly Journal of Economics*, 127(2):645–700, 2012. doi:10.1093/qje/qjs001.

Xavier Gabaix. Power laws in economics: An introduction. *Journal of Economic Perspectives*, 30 (1):185–206, 2016. doi:10.1257/jep.30.1.185.

Nicolae Gârleanu and Lasse H. Pedersen. Efficiently inefficient markets for assets and asset management. *Journal of Finance*, 73(4):1663–1712, 2018. doi:10.1111/jofi.12696.

Itzhak Gilboa and David Schmeidler. Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics*, 18(2):141–153, 1989. doi:10.1016/0304-4068(89)90018-9.

Karl B. Gregory, Soumendra N. Lahiri, and Daniel J. Nordman. A smooth block bootstrap for quantile regression with time series. *Annals of Statistics*, 46(3):1138–1166, 2018. doi:10.1214/17-AOS1580.

Sanford J. Grossman and Joseph E. Stiglitz. On the impossibility of informationally efficient markets. *American Economic Review*, 70(3):393–408, 1980.

Refet S. Gürkaynak, Brian Sack, and Jonathan H. Wright. The U.S. Treasury yield curve: 1961 to the present. *Journal of Monetary Economics*, 54(8):2291–2304, November 2007. doi:10.1016/j.jmoneco.2007.06.029.

Lars Peter Hansen and Robert J. Hodrick. Forward exchange rates as optimal predictors of future spot rates: An econometric analysis. *Journal of Political Economy*, 88(5):829–853, 1980. doi:10.1086/260910.

Lars Peter Hansen and Ravi Jagannathan. Implications of security market data for models of dynamic economies. *Journal of Political Economy*, 99(2):225–262, 1991. doi:10.1086/261749.

John R. Hicks. *Value and Capital*. Oxford University Press, Oxford, UK, 1939.

Thomas S. Y. Ho. Key rate durations: Measures of interest rate risks. *Journal of Fixed Income*, 2 (2):29–44, 1992. doi:10.3905/jfi.1992.408049.

Fushing Hsieh and Bruce W. Turnbull. Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *Annals of Statistics*, 24(1):25–40, 1996. doi:10.1214/aos/1033066197.

Elton P. Hsu and S. R. Srinivasa Varadhan. *Probability Theory and Applications*. American Mathematical Society, 1999.

Kosuke Imai, In Song Kim, and Erik H. Wang. Matching methods for causal inference with

time-series cross-sectional data. *American Journal of Political Science*, 67(3):587–605, 2021. doi:10.1111/ajps.12685.

Guido W. Imbens and Donald B. Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015.

Investment Company Institute. Investment company fact book, 2021. URL https://www.ici.org/system/files/2021-05/2021_factbook.pdf.

Marlène Isoré and Urszula Szczerbowicz. Disaster risk and preference shifts in a New Keynesian model. *Journal of Economic Dynamics and Control*, 79:97–125, 2017. doi:10.1016/j.jedc.2017.04.001.

Jens Carsten Jackwerth. Recovering risk aversion from option prices and realized returns. *Review of Financial Studies*, 13(2):433–451, 2000. doi:10.1093/rfs/13.2.433.

Jens Carsten Jackwerth and Marco Menner. Does the Ross recovery theorem work empirically? *Journal of Financial Economics*, 137(3):723–739, 2020. doi:10.1016/j.jfineco.2020.03.006.

Christian Julliard and Anisha Ghosh. Can rare events explain the equity premium puzzle? *Review of Financial Studies*, 25(10):3037–3076, 2012. doi:10.1093/rfs/hhs078.

Roger Koenker and Gilbert Bassett. Regression quantiles. *Econometrica*, 46(1):33–50, 1978. doi:10.2307/1913643.

Roger Koenker and José A. F. Machado. Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association*, 94(448):1296–1310, 1999. doi:10.1080/01621459.1999.10473882.

Roger W. Lee. The moment formula for implied volatility at extreme strikes. *Mathematical Finance*, 14(3):469–480, 2004. doi:10.1111/j.0960-1627.2004.00200.x.

Erich L. Lehmann. Some concepts of dependence. *Annals of Mathematical Statistics*, 37(5):1137–1153, 1966. doi:10.1214/aoms/1177699260.

Matthew Linn, Sophie Shive, and Tyler Shumway. Pricing kernel monotonicity and conditional information. *Review of Financial Studies*, 31(2):493–531, 2018. doi:10.1093/rfs/hhx095.

Robert B. Litterman and José Scheinkman. Common factors affecting bond returns. *Journal of Fixed Income*, 1(1):54–61, 1991. doi:10.3905/jfi.1991.692347.

Yan Liu. Index option returns and generalized entropy bounds. *Journal of Financial Economics*, 139(3):1015–1036, 2021. doi:10.1016/j.jfineco.2020.08.011.

Yan Liu and Jing Cynthia Wu. Reconstructing the yield curve. *Journal of Financial Economics*, 142(3):1395–1425, 2021. doi:10.1016/j.jfineco.2021.05.059.

Frederick R. Macaulay. *Some Theoretical Problems Suggested by the Movements of Interest Rates,*

*Bond Yields and Stock Prices in the United States since 1856*. National Bureau of Economic Research, 1938.

Daniel Mantilla-Garcia, Lionel Martellini, Vincent Milhau, and Hector Enrique Ramirez-Garrido. Improving interest rate risk hedging strategies through regularization. *Financial Analysts Journal*, 78(4):18–36, August 2022. doi:10.1080/0015198X.2022.2095193.

Ian Martin. Consumption-based asset pricing with higher cumulants. *Review of Economic Studies*, 80(2):745–773, 2013. doi:10.1093/restud/rds029.

Ian Martin. What is the expected return on the market? *Quarterly Journal of Economics*, 132(1): 367–433, 2017. doi:10.1093/qje/qjw034.

Ian Martin and Christian Wagner. What is the expected return on a stock? *Journal of Finance*, 74(4):1887–1929, 2019. doi:10.1111/jofi.12778.

Alexander J. McNeil, Rüdiger Frey, and Paul Embrechts. *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton University Press, 2015.

Sanjay K. Nawalkha and Nelson J. Lacey. Closed-form solutions of higher-order duration measures. *Financial Analysts Journal*, 44(6):82–84, 1988. doi:10.2469/faj.v44.n6.82.

Alexei Onatski and Chen Wang. Spurious factor analysis. *Econometrica*, 89(2):591–614, March 2021. doi:10.3982/ECTA16703.

Andrew J. Patton and Allan Timmermann. Monotonicity in asset returns: New tests with applications to the term structure, the CAPM, and portfolio sorts. *Journal of Financial Economics*, 98(3):605–625, 2010.

Dimitris N. Politis and Joseph P. Romano. The stationary bootstrap. *Journal of the American Statistical Association*, 89(428):1303–1313, 1994. doi:10.1080/01621459.1994.10476870.

Eliezer Z. Prisman and Marilyn R. Shores. Duration measures for specific term structure estimations and applications to bond portfolio immunization. *Journal of Banking & Finance*, 12(3):493–504, 1988. doi:10.1016/0378-4266(88)90011-8.

Likuan Qin, Vadim Linetsky, and Yutian Nie. Long forward probabilities, recovery, and the term structure of bond risk premiums. *Review of Financial Studies*, 31(12):4863–4883, 2018. doi:10.1093/rfs/hhy042.

Frank M. Redington. Review of the principles of life-office valuations. *Journal of the Institute of Actuaries*, 78(3):286–340, 1952. doi:10.1017/S0020268100052811.

Thomas A. Rietz. The equity risk premium a solution. *Journal of Monetary Economics*, 22(1): 117–131, 1988. doi:10.1016/0304-3932(88)90172-9.

Joshua V. Rosenberg and Robert F. Engle. Empirical pricing kernels. *Journal of Financial Economics*, 64(3):341–372, 2002. doi:10.1016/S0304-405X(02)00128-9.

Stephen A. Ross. *Neoclassical Finance*. Princeton University Press, 2005.

Stephen A. Ross. The recovery theorem. *Journal of Finance*, 70(2):615–648, 2015. doi:10.1111/jofi.12092.

Alberto G. Rossi, David Blake, Allan Timmermann, Ian Tonks, and Russ Wermers. Network centrality and delegated investment performance. *Journal of Financial Economics*, 128(1):183–206, 2018. doi:10.1016/j.jfineco.2018.02.003.

Paul A. Samuelson. The effect of interest rate increases on the banking system. *American Economic Review*, 35(1):16–27, March 1945.

Paul Schneider. An anatomy of the market return. *Journal of Financial Economics*, 132(2):325–350, 2019. doi:10.1016/j.jfineco.2018.10.015.

Paul Schneider and Fabio Trojani. (Almost) model-free recovery. *Journal of Finance*, 74(1):323–370, 2019. doi:10.1111/jofi.12737.

David Schreindorfer. Macroeconomic tail risks and asset prices. *Review of Financial Studies*, 33 (8):3541–3582, 2020. doi:10.1093/rfs/hhz105.

Sang Byung Seo and Jessica A. Wachter. Option prices in a model with stochastic disaster risk. *Management Science*, 65(8):3449–3469, 2019. doi:10.1287/mnsc.2017.2978.

Robert J. Serfling. *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, 2009.

Clemens Sialm, Laura T. Starks, and Hanjiang Zhang. Defined contribution pension plans: Sticky or discerning money? *Journal of Finance*, 70(2):805–838, 2015. doi:10.1111/jofi.12232.

Karl N. Snow. Diagnosing asset pricing models using the distribution of asset returns. *Journal of Finance*, 46(3):955–983, 1991. doi:10.1111/j.1540-6261.1991.tb03773.x.

Michael Stutzer. A Bayesian approach to diagnosis of asset pricing models. *Journal of Econometrics*, 68(2):367–397, 1995. doi:10.1016/0304-4076(94)01656-K.

Engin A. Sungur. Dependence information in parameterized copulas. *Communications in Statistics - Simulation and Computation*, 19(4):1339–1360, 1990. doi:10.1080/03610919008812920.

Lars E. O. Svensson. Estimating and interpreting forward interest rates: Sweden 1992–1994. techreport 4871, National Bureau of Economic Research, September 1994.

Oleg Sydyak. Interest rate risk management and asset liability management. In Pietro Veronesi, editor, *Handbook of Fixed-Income Securities*, chapter 7, pages 119–146. John Wiley & Sons, 2016. doi:10.1002/9781118709207.ch7.

James Tobin. Estimation of relationships for limited dependent variables. *Econometrica*, 26(1): 24–36, 1958. doi:10.2307/1907382.

Lloyd N. Trefethen. *Approximation Theory and Approximation Practice*. Society for Industrial and Applied Mathematics, Philadelphia, PA, extended edition, 2019. doi:10.1137/1.9781611975949.

Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 2000.

Jessica A. Wachter. Can time-varying risk of rare disasters explain aggregate stock market volatility? *Journal of Finance*, 68(3):987–1035, 2013. doi:10.1111/jofi.12018.

Ivo Welch and Amit Goyal. A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies*, 21(4):1455–1508, 2008. doi:10.1093/rfs/hhm014.

Ram Willner. A new tool for portfolio managers: Level, slope, and curvature durations. *Journal of Fixed Income*, 6(1):48–59, 1996. doi:10.3905/jfi.1996.408171.

Harry Zheng. Macaulay durations for nonparallel shifts. *Annals of Operations Research*, 151: 179–191, 2007. doi:10.1007/s10479-006-0115-7.