The London School of Economics and Political Science

Essays in Asset Pricing

Ran Shi

Thesis submitted to the Department of Finance of the London School of Economics and Political Science for the degree of

Doctor of Philosophy

April 2022

Acknowledgements

I am deeply indebted to my advisors, Ian Martin and Kathy Yuan, for their invaluable advice and generous support. I am also enormously grateful to Dimitri Vayanos for his feedback on research. They guide me to be an economist; their work fosters my passion in financial economics.

I benefited a lot from conversations with Thummim Cho, Vicente Cuñat, Christian Julliard, Péter Kondor, Dong Lou, Igor Makarov, Martin Oehmke, Cameron Peng, Christopher Polk, and Walker Ray in completing this thesis. I would like to thank them for their help.

I thank all the other people in the finance department at LSE for creating an extraordinary learning environment.

I dedicated this thesis to my family, who makes it happen.

Declaration of Authorship

I certify that the thesis I have presented for examination for the PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

I warrant that this authorisation does not, to the best of my belief, infringe the rights of any third party.

I declare that my thesis consists of 66,953 words.

Statement of conjoint work

I confirm that Chapter 2 is jointly co-authored with Ian Martin and I contribute 50% of the work. I confirm that Chapter 3 is jointly co-authored with Jiantao Huang and I contribute 50% of the work.

Abstract

Essays in Asset Pricing

by Ran Shi

This thesis contains three chapters studying asset prices from different financial markets to understand the economic forces driving their movements and recover economic variables of interest.

In Chapter 1, I develop and estimate a model to quantify the effects of financial constraints, arbitrage capital, and hedging demands on asset prices and their deviations from frictionless benchmarks. Using foreign exchange derivatives data, I find that financial constraints and hedging demands contribute to 46 and 35 percent variation in the deviations from covered interest parity of the one-year maturity. While arbitrage capital fluctuation explains the remaining 19 percent variation on average, it periodically stabilizes prices when the other two forces exert disproportionately large impacts. The model features general financial constraints and produces a nonparametric arbitrage profit function. I unveil the shapes and dynamics of financial constraints from estimates of this function.

In Chapter 2 (co-authored with Ian Martin), we propose a framework to compute sharp bounds of the crash probability of an individual stock using option prices. Empirical tests suggest that these bounds are close to the exact forward-looking crash probabilities. Out of sample, either the lower or upper bound outperforms combinations of stock characteristics in terms of forecasting stock-specific crash events. Applying the framework to study the equity of global systemically important banks (G-SIBs) gives rise to forward-looking fragility and stability measures of the global financial system.

In Chapter 3 (co-authored with Jiantao Huang), we develop a transparent Bayesian approach to quantify uncertainty in linear stochastic discount factor (SDF) models. We show that, for a Bayesian decision maker, posterior model probabilities increase with maximum in-sample Sharpe ratios and decrease with model dimensions. Entropy of posterior probabilities represents model uncertainty. We apply our approach to quantify the time series of model uncertainty in North American, European, and Asian Pacific equity markets. Model uncertainty is countercyclical in these markets before the 2008 financial crisis, but remains high afterwards. It predicts investors' asset allocation decisions across equity and fixed-income funds. In survey data, investors tend to be more pessimistic about equity performance during periods of high model uncertainty.

Contents

List of Figures

List of Tables

Chapter 1

A Quantitative Model of Limited Arbitrage in Currency Markets: Theory and Estimation

Modern finance theory and practice build heavily on the assumption of no arbitrage. One of the textbook no-arbitrage conditions is covered interest rate parity (CIP): risk-free rates are the same for all countries after exchange rate risk is fully hedged. Before the 2008 global financial crisis, this condition broadly holds in the data.[1] After the crisis, significant and persistent CIP violations have emerged for all major currency pairs.[2] A failure of the no-arbitrage assumption has become a new normal in one of the largest financial markets in the world.

Existing limits-to-arbitrage theory provides guidance to understanding this phenomenon. First, hedging demand imbalances in the foreign exchange (FX) forwards and swaps markets can cause "price pressures", misaligning forward premiums or currency swap rates. Second, arbitrageurs such as trading desks of global FX dealer banks face binding limits to arbitrage. As a result, they do not have the insatiable appetites to "arbitrage away" the deviations.[3]

With all the valuable theoretical perspectives, we still do not understand how quantitatively important each economic force is for CIP deviations. More broadly speaking, existing limits-to-arbitrage models generally fall short in their potential to be directly mapped to data and offer quantitative answers.

[1] Frenkel and Levich (1975, 1977) attribute CIP deviations to transactions costs. Taylor (1987) confirms the CIP condition using high-frequency data within time windows of approximately one minute. Working with tick-by-tick data, Akram, Rime, and Sarno (2008, 2009) find that most profitable deviations last less than five minute and the CIP condition holds on average in their sample period from February 13 to September 30, 2004.

[2] Baba and Packer (2009) analyze large CIP violations during the global financial crisis. Ivashina, Scharfstein, and Stein (2015) study short-maturity CIP deviations during the Eurozone sovereign crisis, emphasizing their role as a barometer of wholesale dollar funding conditions. Du, Tepper, and Verdelhan (2018) establish the new post-crisis benchmark of CIP deviations around 25 basis points and investigate the causes.

[3] Du, Tepper, and Verdelhan (2018) and Borio, McCauley, McGuire, and Sushko (2016) provide suggestive evidence linking CIP deviations to both hedging demands and dealers' limited arbitrage capacity.

This paper aims to bridge the gap. I incorporate hedging demands and arbitrage limits in a parsimonious model. Yields on arbitrage opportunities such as CIP deviations are direct equilibrium outcomes from the model. The model further distinguishes two determinants of arbitrage limits: arbitrageurs' capital and financial constraints. Arbitrageurs' optimal trading decisions endogenously determine their capital accumulation. These decisions are made under exogenously specified financial constraints that may arise from agency concerns or regulatory requirements.

The key innovation of the model is that it allows for a general form of financial constraints and provides identifying conditions for estimating these constraints. I show that arbitrageurs' capital accumulates faster when arbitrage yields are large. In the meantime, significant price gaps show up after adverse shocks that drain the arbitrage capital. How arbitrageurs' capital returns respond to yields on their arbitrage opportunities reveals the shape and dynamics of binding financial constraints. I exploit this equilibrium relationship to back out financial constraints from market price data.

I estimate hedging demands based on two additional equilibrium outcomes from the model. On the one hand, financial constraints determine the maximized arbitrage profits. In equilibrium, arbitrageurs "arbitrage alongside the margin" in the sense that their arbitrage positions equal marginal increases in arbitrage profits regarding yields on their arbitrage opportunities. I compute arbitrage positions based on estimates of financial constraints leveraging this equilibrium outcome. On the other hand, equilibrium arbitrage yields such as CIP deviations must equate external hedging demands with these arbitrage positions. After computing the arbitrage positions, I can estimate the hedging demand functions for different currency pairs.

I propose a variance decomposition scheme using the estimated model by computing counterfactual CIP deviations after holding different model ingredients constant. According to my decomposition, on average, 46 percent of variation in one-year CIP deviations of G6 currencies against the US dollar is due to changing financial constraints (throughout the paper, the term "G6 currencies" refers to the euro, yen, pound, Canadian dollar, Australian dollar, and Swiss Franc.) Demands for dollars in forward markets due to FX risk management practices of exporters and global bond investors (hedging demands) explain another 35 percent. Fluctuations in arbitrageurs' capital account for the remaining 19 percent.

Four sets of empirical findings emerge from analyzing the estimated model. First, the importance of financial constraints and hedging demands in explaining CIP deviations varies across currency pairs. Hedging demands account for 56 percent variation in the Canadian dollar CIP deviations, but less than 30 percent in the context of yen and euro. Varying financial constraints fill the vacancy left by hedging demands for these two currencies, explaining almost 60 percent variation in their basis against the dollar.

Second, arbitrageurs' capital plays a unique role in influencing the deviations. In 2009-2019, it contributes to a limited fraction of variation in CIP deviations (19 percent on average, as pointed out earlier). However, it can stabilize the basis during periods of significant variation in financial constraints or hedging demands. According to the variance decomposition results, shutting down hedging demand or financial friction variation always reduces variation in CIP deviations. However, holding arbitrageurs' capital constant can increase fluctuations in the currency basis for certain periods. During these periods, arbitrage capital counterbalances the other two forces and dampens variation in CIP deviations. For example, in 2013-2014, the one-year Canadian dollar basis is overwhelmingly driven by hedging demands. If arbitrage capital remains constant, (counterfactual) variation in one-year Canadian dollar CIP deviations would double.

Third, *shapes* of financial constraints change dramatically before and after 2014. These shapes can help understand arbitrageurs' internal capital allocation decisions in response to regulatory reforms. In 2009-2013, arbitrageurs can build CIP arbitrage positions that are at least four times larger than their equity capital without significantly downsizing other investment positions. However, in 2014-2019, building the same size of arbitrage positions will force the arbitrageurs to decrease standard investment positions by 40 percent. More importantly, a hard leverage cap of around seven (times the equity capital) emerges for the same period. This finding appears to be consistent with the fact that the supplementary leverage ratio (SLR) requirement was finalized in the third quarter of 2014. Overall, the shape of constraints after 2014 can be interpreted as a risk-weighted capital requirement plus a hard leverage ratio requirement.

Finally, bilateral net exports and net foreign security purchases are dominant forces explaining currency hedging demands.[4] For security purchases, all impacts on hedging demands come from net bond purchases (as oppose to equity). The estimated demand functions suggest that *forward* dollar demands increase in the US net exports and net bond purchases. In other words, exporters and investors holding foreign bonds hedge more when they need to repatriate more future incomes denominated in foreign currencies.

I now describe the model to provide intuitions on its mechanism and estimation. In the model, competitive arbitrageurs trade with hedgers; trading determines equilibrium arbitrage yields such as CIP deviations. Hedgers exchange specific currencies for dollars forward. I build an optimizing foundation for this forward dollar demand in a two-country setting. Hedgers from each country are subject to endowment shocks denominated in foreign currencies. They manage their exchange rate risk using forward contracts. Differences in their foreign endowments create hedging demand imbalances described by the demand functions specified in the model.

Competitive arbitrageurs maximize (additive) discounted log utilities over their lifetime consumption stream.[5] They profit from multiple arbitrage opportunities by absorbing demand imbalances in FX derivatives markets for different currency pairs. In the meantime, they have access to standard investment opportunities: one risk-free and one risky

[4] Net foreign security purchases are the difference between domestic (US) residents' purchases of foreign securities and foreign residents' purchases of domestic securities. As a clarification, this measure does not necessarily represent portfolio flows as "round-trip" trades can occur between international investors from different countries.

[5] Simplified version of the model lasts two periods; the full model comes with infinite time horizon. I also extend the theory for general constant relative risk-aversion (CRRA) utilities.

asset. Arbitrage limits arise from financial constraints on both their arbitrage positions and investment positions. The constraints induce a trade-off between deploying capital to their standard investment business and diverting the resource to arbitrage activities. Sizable arbitrage positions come at the cost of cutting back routine investment positions.

The model features an agnostic view regarding the specific form of financial constraints.[6] This setup encompasses standard specifications of frictions such as the margin requirements, leverage ratio requirements, Value-at-Risk rules, transaction costs, and credit/debt/funding valuation adjustments. This generality in modeling choice allows me to reach robust theoretical conclusions and perform nonparametric estimation of the constraints. This approach is particularly helpful given the sophisticated nature of post-crisis financial regulations. In this new era, numerous regulatory constraints exist and many of them can be binding at the same time.

Even without the dedication to a certain form of financial constraints, the model still yields strong predictions. Most importantly, the model argues that the absolute values of present arbitrage yields (e.g., CIP deviations) predict future returns on arbitrageurs' capital,[7] and this predictive relationship is convex. This increasing and convex function (of capital returns in response to arbitrage yields) reveals the form of financial constraints. For example, under margin requirements, the function equals zero when deviations are small and increases linearly after a threshold. However, with Value-at-Risk rules, arbitrageurs' capital returns smoothly respond to all levels of deviations regardless how small they are. The function is defined in the same way as the profit function in standard production theories, thus named the *arbitrage profit function.* A more convex arbitrage profit function implies a higher hurdle for arbitrageurs to covert CIP deviations into sizable arbitrage profits.

Empirical evidence supports the prediction. Average CIP deviations among G6 currencies predict monthly and quarterly returns on arbitrageurs' capital, after controlling for common time-series return predictors. One basis point increase in the average deviations forecasts at least two percentage point increase in the (annualized) returns. Two statistical tests, one parametric and another semi-parametric confirm that the predictive relationship is convex.

Model estimation takes two steps, both relying on equilibrium outcomes of the model. The first step is to estimate the arbitrage profit function, which characterizes the increasing and convex response of arbitrageurs' capital returns to CIP deviations. The model allows this function to change across time, reflecting varying stringency of financial constraints.

[6] In the model, arbitrageurs solve dynamic consumption and portfolio choice problems with one risk-free asset, one risky asset, and multiple riskless arbitrage opportunities, under general position constraints: a bounded, closed, and convex set. The specification of constraints is the same as Cvitanić and Karatzas (1992), who solve the Merton model of one risk-free asset and multiple risky assets under general position constraints. One theoretical contribution is that I characterize arbitrageurs' optimality conditions using a more accessible primal-dual approach.

[7] Though CIP basis can be positive (e.g., Australian dollars against US dollars) or negative (e.g., euro or yen against US dollars), arbitrageurs can always profit from it by properly switching legs of their positions. Thus, absolute values of the deviations contribute to arbitrage profits. I will stop mentioning "the absolute values" later on in the introduction for the ease of exposition.

I develop a new statistical procedure to estimate both the functional form and time-series variation of arbitrage profit functions in the model. The functional form reveals how the (binding) financial constraints look like collectively; the time-series variation describes dynamics of the constraints. I compute equilibrium arbitrage positions using the these estimates, which can be inferred from market prices once the arbitrage profit function is known.

The second step is estimating parameters in hedging demand functions. In equilibrium, CIP deviations in the model are such that arbitrage positions equal hedging demands. Plugging-in the inferred arbitrage positions enable demand estimation without using position data in FX derivatives markets. To resolve the endogeneity concerns about CIP deviations and latent demands (unobservable drivers of hedging demands), I construct an instrumental variable estimator for demand elasticities. The instruments for the deviations of one currency are observable hedging demand drivers of *other* currencies. The identification strategy is motivated by the fact that FX arbitrageurs such as global dealer banks can profit from multiple currency basis simultaneously. For a specific currency, hedging demand drivers of other currencies shift arbitrage profits. Arbitrage positions exploiting the CIP deviations of this particular currency change accordingly. The instruments effectively become "supply shifters" (if we interpret arbitrageurs as suppliers of "arbitrage services") that are uncorrelated with latent demands.

Broader contribution of this paper is twofold. Typically, there is a separation of theory and empirics in the limits-to-arbitrage literature. I aim to partially bridge the gap by building a model that synthesizes necessary ingredients in the existing theory and maps directly to the data to quantify the economic forces at work. The backbone of my model is close to Gabaix and Maggiori (2015), who study real imbalances in the currency spot markets absorbed by "financiers" facing commitment problems (which translate into quadratic position limits). My model focuses on hedging demand imbalances in FX forwards and swaps markets. It specifies financial frictions in a general format. I further quantify demand imbalances and financial frictions to explain empirical patterns and facilitate counterfactual exercises.

The second contribution is methodological: the estimation framework can be applied to other violations of the no-arbitrage condition in today's financial markets. I demonstrate how to back out the financial constraints and arbitrage positions from arbitrageurs' capital returns and arbitrage yields. The flexible nonparametric arbitrage profit function approach is particularly useful in light of the numerous regulatory reforms after 2008. Unlike the demand system approach to asset pricing (Koijen and Yogo, 2019), my methodology for estimating demand function parameters does not rely on position data (though high-quality position data can help discipline my estimation), but instead draws inferences using price data based on arbitrageurs' optimality conditions.

Literature. The key ingredients of my model, demand shocks and the limited arbitrage capacity, follows the standard limits-to-arbitrage literature. Examples for demand shocks in the FX derivatives markets include i. hedging demands of currency carry traders

(Du, Tepper, and Verdelhan (2018) offer suggestive evidence based on the association between cross-sectional variation in CIP deviations and average interest rate differentials); ii. financial institutions' FX risk management practices (Puriya and Bräuning (2021) identify this driving force for short-maturity FX forward contracts). Price impacts of demand shocks have also been investigated in markets of commodity futures by Acharya, Lochstoer, and Ramadorai (2013), options by Gârrleanu, Pedersen, and Poteshman (2008), long-term interest rate swaps by Klingler and Sundaresan (2019), government bonds by Greenwood and Vayanos (2010), public equities by Coval and Stafford (2007) and Lou (2012). My empirical findings on the importance of hedging demands from exporters and foreign currency long-term bond investors in driving forward dollar demands and determining CIP deviations complement this literature.

The literature on financial constraints is enormous. In the field of international finance, see Gabaix and Maggiori (2015) for their impacts on spot exchange rates. In financial economics, Gârleanu and Pedersen (2011) and Gromb and Vayanos (2002, 2018) are examples of theories examining violations of the law of one price in light of margin or collateral constraints. Andersen, Duffie, and Song (2019) explain CIP deviations in light of debt-overhang costs to equity holders of derivatives dealers. In models such as Kyle and Xiong (2001) and Kondor and Vayanos (2019), aggregate arbitrage capital endogenously creates risk-aversion dynamics, inducing commonality in asset prices in response to arbitrage capital fluctuations. My contribution to the literature is that I characterize equilibrium outcomes which are robust to assumptions about the financial constraints, and develop empirical methods for estimating functional form of the constraints.

Vayanos and Vila (2021) is an example of quantitative limits-to-arbitrage models for government bond markets, calibrated to predictive regression coefficients. Jermann (2020) presents and calibrates a model featuring holding costs for long-term bond to explain negative swap spreads after the financial crisis. My empirical approach distinguishes from their exercises by directly estimating the equilibrium conditions using asset price data. To my knowledge, this paper presents the first fully estimated limits-to-arbitrage model, which explains not just the level but also the variation in deviations from the law of one price.

The paper is structured as follows. Section 1.1 briefly reviews the definition and measurements of CIP deviations to provide additional backgrounds. Section 1.2 presents a simplified version of the model to illustrate key intuitions. Section 1.3 introduces the full model and characterizes its equilibrium outcomes. Section 1.4 describes additional data and measurements, tests the model's main prediction, enriches the model to map it to data, and describes estimation methodologies. Section 1.5 performs quantitative exercises using the estimated model. Section 1.6 concludes. All proofs are in the Appendix.

1.1 CIP deviations and their measures

At time t, the CIP deviation for currency i of maturity τ is b such that the following equation holds:

$$\exp\left(r^{\$}_{t\to(t+\tau)}\tau\right) = \exp\left(r^{i}_{t\to(t+\tau)}\tau + b\tau\right)\frac{F_{t\to(t+\tau)}}{E_t}, \tag{1.1}$$

where $r^{\$}_{t\to(t+\tau)}$ and $r^{i}_{t\to(t+\tau)}$ represent risk-free rates of the US dollar and currency i from time t to $(t+\tau)$; $F_{t\to(t+\tau)}$ is the forward price of currency i in dollars maturing at time $(t+\tau)$; E_t is the spot price.

Following Du, Tepper, and Verdelhan (2018), I focus on two measures of CIP deviations using different derivative contracts: FX forwards and cross-currency basis swaps (currency or basis swaps in short). From FX forward contracts, observable currency forward prices $F_{t\to(t+\tau)}$ (thus observable forward premiums $F_{t\to(t+\tau)}/E_t$) can be plugged in to the equation above. The two risk-free rates $r^{\$}_t$ and r^{i}_t are commonly measured by overnight index swap (OIS) rates for different countries. I call CIP deviations calculated from equation (1.1) using these variables the forward-OIS bases.

A more direct measure of CIP deviations comes from the currency swap contracts. In a currency swap contract, two parties (namely Alice and Bob) exchange currencies at spot rates upfront and pay each other back effectively with floating rate bonds. Specifically, Alice, receiving £100 from Bob initially, will pay Bob back with (cashflows of) a pound floating rate bond (face value = £100); Bob, receiving \$135 (let $E_t = 1.35$ be the GBP/USD spot rate) from Alice at beginning of the contract, will pay Alice back with a dollar floating rate bond (face value = \$135). Currency swap contracts quote b such that Bob pays the the dollar floating rates $\{r^{\$}_{t\to(t+\Delta t)}, r^{\$}_{(t+\Delta t)\to(t+2\Delta t)}, \ldots\}$, and Alice pays *adjusted* pound floating rates $\{r^{£}_{t\to(t+\Delta t)}, r^{£}_{(t+\Delta t)\to(t+2\Delta t)}, \ldots\} + b$. The payments are usually made on a quarterly basis (i.e., $\Delta t = 0.25$). Back to equation (1.1), we can interpret this quoted currency swap rate as CIP deviations defined through treating $r^{\$}_{t\to(t+\tau)}$ and $r^{i}_{t\to(t+\tau)}$ as interest rate swap rates (swapping the two floating rates). Du, Tepper, and Verdelhan (2018) and Augustin, Chernov, Schmid, and Song (2020) describe detailed trading arrangements justifying this conclusion.

Throughout this paper, I focus on one-year currency swap rates and use forward-OIS implied one-year CIP deviations for validation. At this maturity, both the FX forwards and currency swaps have high trading volumes and low bid-ask spreads. I collect the FX forward/spot prices, OIS rates, and currency swap rates from Bloomberg. Table 1.1 reports summary statistics of the two deviation measures; Figure 1.12 in the Appendix plots these two measures for G6 currencies. Overall, currency swap rates offer more conservative and less volatile measures of CIP deviations compared with the forward-OIS bases at the one-year maturity.

1.2 A simple model in a two-period deterministic economy

To begin with, I present a simple model with no uncertainty to illustrate key insights of the model. As a preview, the model features an agnostic view about the specific constraints arbitrageurs face, predicts a convex relationship between arbitrageurs' investment return and their arbitrage profit (from CIP deviations), and determines CIP deviations as tractable equilibrium outcomes.

The economy lasts for two periods: today and tomorrow.[8] There are two types of agents: arbitrageurs and hedgers. Each type composes an identical continuum of measure one.

Arbitrageurs. Each arbitrageur is endowed with initial capital k today. They choose their consumptions and derive utilities from them as follows:

$$\log(y) + \frac{1}{1+\rho}\log(y'). \tag{1.2}$$

The subjective time discount rate $1/(1+\rho)$ belongs to the interval (0, 1), i.e., $\rho > 0$; consumptions are y today and y' tomorrow.

The arbitrageurs can invest in a risk-free asset earning a net return r ($r > 0$), or simply store their capital safely with zero net return. I assume that the amount of capital stored cannot be negative.[9] Arbitrageurs can also profit from a riskless arbitrage opportunity, yielding b per unit of position they enter. For currency markets, we can interpret b as CIP deviations, which is either positive (e.g., Australian dollars) or negative (e.g., yen).

By consuming y today, arbitrageurs are saving (or equivalently, investing) $s = (k - y)$ amount of capital to fund their future consumption. Denote by π_0 and π the "weights" of their investment positions in the risk-free asset and the arbitrage opportunity, their absolute positions are $\pi_0 s$ and πs accordingly. To earn the risk-free rate of return, capital is needed: $(1+r)\pi_0 s$ units of capital next period come at a cost of $\pi_0 s$ today. In comparison, harvesting the arbitrage profits $\pi s b$ next period requires no capital today. As a result, the arbitrageurs' capital next period k' is given by

$$k' = s + \pi_0 s(1+r) - \pi_0 s + \pi s b - 0 = s\left[1 + \pi_0 r + \pi b\right]. \tag{1.3}$$

According to equation (1.3), arbitrageurs store $(1 - \pi_0)s$ units of capital (after investing $\pi_0 s$ in the risk-free asset). As the economy lasts for only two periods, arbitrageurs consume all their capital tomorrow, i.e., $y' = k'$.

Replacing y' in problem (1.2) with $s\left[1 + \pi_0 r + \pi b\right]$, we can see that arbitrageurs are

[8] Notation-wise, variables tomorrow come with prime superscripts.

[9] This claim rules the possibility that arbitrageurs can raise fund by paying a gross interest rate of one (the storage yield), for this itself leads to another riskless arbitrage within the model: borrowing at cost one, investing in the risk-free asset yielding $(1 + r)$. The storage technology is needed in the model because, after introducing arbitrage limits later in the paper, arbitrageurs need to devote capital to arbitrage positions. The capital buttressing their arbitrage activities is "stored" in the sense that they cannot generate a return as high as r. We can treat the zero storage yield here as a normalization.

solving two separate problems:

$$\underset{y,\, s=k-y}{\text{maximize}} \quad \log(y) + \frac{1}{1+\rho}\log(s) \quad \text{and} \quad \underset{\pi_0 \leq 1,\, \pi \in \mathbb{R}}{\text{maximize}} \quad \log(1+\pi_0 r + \pi b). \tag{1.4}$$

The restriction $\pi_0 \leq 1$ appearing in the second problem is due to the assumption that capital stored is nonnegative, that is, $(1-\pi_0)s \geq 0$.

Hedgers. Hedgers use currency forwards or swaps to manage their foreign exchange exposures. Under the context of CIP deviations, I assume that their (aggregate) demand for selling foreign currencies (say, pounds) in exchange for dollars *in forward markets* is

$$q(b) = \gamma_0 - \gamma b, \quad \gamma > 0. \tag{1.5}$$

These *forward dollar demands* are downward sloping with regard to the CIP deviation b. This is because, according to equation (1.1), a smaller b for the pound is equivalent to higher forward price of pounds against dollars. It propels hedgers' willingness to sell pounds for dollars forward, creating higher forward dollar demands.[10]

The equilibrium arbitrage yield. Arbitrageurs "take the opposite side" against hedgers' demands: their arbitrage positions are effectively *forward dollar supplies.* When hedgers sell pounds for dollars forward (positive forward dollar demand, $q > 0$), pressing GBP/USD forward price to drop below the no-arbitrage benchmark, a positive CIP deviation emerges.[11] In response, arbitrageurs can take advantage of this opportunity by borrowing pounds (yielding $-r^{£}$), swapping pounds to dollars (yielding $r^{£} + b - r^{\$}$), and lending dollars (yielding $+r^{\$}$). They supply dollars in the currency forward markets because of the need to payback the dollars they received at the beginning of the swap contract. A more simplistic view is that with $b > 0$, the forward price of GBP/USD is relatively low, arbitrageurs tend to offer (supply) dollars to buy pounds forward. Their total supply of dollars $\pi(b)s$ is positive,[12] in which $\pi(b)$ solves (1.4) for a given b. The

[10] Appendix 1.7.3 provides an optimization foundation to the reduced form specification (1.5). I build a two-country currency-risk hedging model, in which US hedgers manage their currency exposures through selling pounds and UK hedgers conduct the opposite trade in pound-dollar forward market. Their hedging needs do not necessarily cancel out, which give rise to the (net) demands specified in (1.5), representing hedging demand *imbalances* in currency markets.

The hedging demand $q(b)$ can take either positive or negative signs. According to the micro-foundation in Appendix 1.7.3, US hedgers offer forward pounds for dollars while UK hedgers seek opposite trades. When the US hedgers' demand exceeds its UK counterpart, $q(b)$ is positive. The net effect is a positive demand for forward dollars. This demand becomes negative when the UK hedgers hedge more. In other words, a negative $q(b)$ can be interpreted as the net demand for foreign currencies (by selling dollars forward).

When the demand q is negative, hedgers are selling forward dollars in exchange for pounds, causing negative forward dollar demands. As b becomes smaller (thus a higher forward GBP/USD price F or, to put it differently, a lower dollar forward price), hedgers tend to sell less dollar: q still increases as b decreases.

[11] Without frictions, arbitrageurs absorb hedging demands imbalances "with ease" and equilibrium deviations always equal zero. This ideal outcome As we will show later in Proposition 2, when there are arbitrage limits, a positive forward dollar demand, $q > 0$, is equivalent to both $\gamma_0 > 0$ and the equilibrium CIP deviations b^* to be $0 < b^* \leq \gamma_0/\gamma$.

[12] I provide rigorous theoretical arguments for this through Lemma 2 in the Appendix.

equilibrium deviations solve the following equation:

$$\pi(b)s = q(b). \tag{1.6}$$

Of note, all the analysis goes through in the same way when there is a negative forward dollar demand, i.e., $q < 0$.[13]

The equilibrium b that solves (1.6) is such that

$$b = \frac{\gamma_0}{\gamma + \pi(b)/b \times s}. \tag{1.7}$$

The frictionless benchmark. Without friction, the second maximization problem in (1.4) commands $\pi_0 = 1$ and $|\pi(b)| \to \infty$ whenever $|b| > 0$. This implies that $\pi(b)/b \to \infty$ for any b around a neighborhood of zero. According to equation (1.7), the equilibrium deviation is zero. Absence of arbitrage in the model is a result of "hyper-elastic" arbitrage positions in response to arbitrage yields.

Limits to arbitrage. Arbitrage limits that cause CIP deviations must prevent $\pi(b)/b$ from going to infinity whenever b deviates from zero. I assume that they arise from the following position constraint on π_0 and π:

$$(\pi_0,\, \pi) \in \mathcal{C}, \tag{1.8}$$

in which $\mathcal{C}$ is a subset of $(-\infty, 1] \times \mathbb{R}$ (domains defined in the second problem of (1.4)), outside of which the combination of π_0 and π becomes infeasible. Under this assumption, the second problem of (1.4) is equivalent to maximize $(\pi_0 r + \pi b)$ subject to the condition that $(\pi_0,\, \pi) \in \mathcal{C}$. The outcome from solving this problem represents the optimal return on investment for the arbitrageurs, denoted by

$$S_{\mathcal{C}}(r,\, b) = \sup_{(\pi_0,\, \pi) \in \mathcal{C}} \{\pi_0 r + \pi b\}. \tag{1.9}$$

$S_{\mathcal{C}}$ is often named the *support function* of the set $\mathcal{C}$. It works the same way as the profit function in the standard production theory, when the set $\mathcal{C}$ is a production set (Mas-Colell, Whinston, and Green, 1995, Chapter 5.B-5.C, p. 128-143). We can call this function the *arbitrage profit function.* As I will illustrate soon, this function defines the optimal investment return per unit of capital for arbitrageurs.

The trade-off arbitrageurs face is fully characterized by the position constraint. When they extend their positions to take advantage of an arbitrage opportunity, they have to cut positions on the risk-free asset (i.e., put a fraction $(1 - \pi_0)$ of their capital inefficiently in storage). Facing this trade-off, arbitrageurs optimally allocate their capital such that they enjoy a (net) return of $S_{\mathcal{C}}(r,\, b)$ per unit of savings. As a result, in equilibrium,

[13] With negative forward dollar demand, i.e., $q < 0$, Proposition 2 presented later commands a negative b. Arbitrageurs will borrow dollars (yielding $-r^{\$}$), swap dollars for pounds (yielding $r^{\$} - r^{£} - b$), and lend pounds (yielding $r^{£}$). Their optimal arbitrage position $\pi(b)s$ is negative, which implies a negative supply of forward dollars (demanding dollars forward). This negative supply is due to the fact that arbitrageurs will receive forward dollars and return pounds at the end of their swap contracts.

$k' = s[1 + S_{\mathcal{C}}(r,\, b)]$. Now I enumerate four assumptions about the constraint $\mathcal{C}$ and one assumption about the arbitrageurs' positions.

Assumption 1. *$\mathcal{C}$ is a subset of* $[0,\, 1] \times \mathbb{R}$.

The assumption that $\pi_0 \leq 1$ reiterates $(1 - \pi_0)s \geq 0$, that is, "negative storage" is not allowed – arbitrageurs cannot borrow money at a zero net interest rate. Assuming $\pi_0 \geq 0$ forbids arbitrageurs from borrowing at the rate r and then storing the proceeds (to further enlarge their arbitrage positions after exhausting all their initial capital k).

Assumption 2. *$\mathcal{C}$ is bounded, closed, and convex.*

The boundedness assumption is straightforward, under which $S_{\mathcal{C}}(r,\, b) < \infty$. Closeness of the set $\mathcal{C}$ means that for any unattainable combination $(\pi_0,\, \pi)$ (falling in the complement of $\mathcal{C}$, an open set), a small neighbor of this combined position is still infeasible for the arbitrageurs to take on: unachievable positions do not suddenly become feasible. Convexity of $\mathcal{C}$ means that convex combinations of feasible position pairs are still available to the arbitrageurs.

Assumption 3. *"Going all in" on the risk-free asset is allowed for the arbitrageurs, that is,* $(1,\, 0) \in \mathcal{C}$.

From this assumption, we have $S_{\mathcal{C}}(r,\, b) \geq r$, the optimal return per unit of savings invested is at least r. Thus, taking advantage of arbitrage opportunities benefits the arbitrageurs, although this activity may require inefficient storage of arbitrage capital.

Assumption 4. *When the arbitrage yield b equals zero, the arbitrage position is zero, that is,* $\pi(0) = 0$.

This is a behavioral assumption about the arbitrageurs. When $b = 0$, the total arbitrage profit is always zero and arbitrage positions π can take any value. Assumption 4 restricts the positions to zero. We can interpret this assumption as arbitrageurs regard the riskless arbitrage opportunity as simply nonexistent whenever its yield equals zero.

Three examples illustrate the set $\mathcal{C}$ under these assumptions and characterize arbitrageurs' optimal choices.

Example 1 (Margin requirements). Margin requirements as highlighted in Gârleanu and Pedersen (2011) can prevent arbitrageurs from building up a large derivative position to "arbitrage away" the opportunities such as CIP deviations. Following their convention (of symmetric margins[14]), I let $\mathcal{C}$ be $\{0 \leq \pi_0 \leq 1,\, \pi \in \mathbb{R} : \pi_0 + m|\pi| \leq 1\}$. Under this specification, arbitrageurs need to post collaterals into margin accounts for their derivative positions: for one unit increase in the notional value, m units of capital are occupied, thus not available for investing in the risk-free asset.[15] With margin requirements,

[14] Extending the characterization to asymmetric margin requirements changes the constrain to $\pi_0 + m^+\pi^+ + m^-\pi^- \leq 1$ where m^+ and m^- apply to long (π^+) and short (π^-) legs of derivative contracts respectively.

[15] An implicit assumption here is that capital posted in the margin account are "stored" using the one-to-one storage technology. In practice, money in the margin account earns interest. Then we could interpret this implicit assumption as a normalization argument, that is, all prices r and b will be normalized by the margin account compensation rate.

the arbitrage position is

$$\pi(b) = \frac{\mathrm{sgn}(b)}{m} I_{\{|b| \geq mr\}}.^{16}$$

Arbitrageurs behave in an "all-or-nothing" manner: they remain inactive when the arbitrage yield is small; otherwise, they build arbitrage positions to the fullest capacity. Panel (A) of Figure 1.1 shows the shape of $\mathcal{C}$ and plots $\pi(b)$.

Example 2 (Costs and adjustments). Now consider the case that a total arbitrage position of value πs will incur a cost or adjustment of $C(\pi s,\, s)$.[17] Adopting a standard specification for adjustment cost functions in investment theory (e.g., Hayashi (1982)), I assume $C(\pi s,\, s) = c(\pi)s$, that is, the cost function is (positively) homogeneous of degree one. As a result, the budget constraint (1.3) is now $k' = s\left[1 + \pi_0 r + \pi b - c(\pi)\right]$. Define $\hat{\pi}_0 = \pi_0 - c(\pi)/r$, the optimization problem of (1.9) becomes maximizing $\hat{\pi}_0 r + \pi b$ subject to the condition that $\mathcal{C} = \{0 \leq \hat{\pi}_0 \leq 1,\, \pi \in \mathbb{R} : \hat{\pi}_0 + c(\pi)/r \leq 1\}$. To make it more specific, I let the function $c(\pi)$ be quadratic with regard to $|\pi|$, that is, $c(\pi) = G|\pi| + (1/2)g\pi^2$ ($G \geq 0,\, g \geq 0$). I present $\mathcal{C}$ and the optimal arbitrage position $\pi(b)$ in Panel (B) of Figure 1.1 under this specification. Similar to margin requirements, there is still a region of inaction for the arbitrageurs: they do not respond when $|b| < G$. However, when arbitrage yields are moderately large, that is, when $|b|$ is greater than G but still smaller than $\sqrt{G^2 + 2gr}$, arbitragers gradually increase their positions, until exhausting all their capital. Clearly, if $g = 0$, meaning that the quadratic term $(1/2)g\pi^2$ disappears from the cost function $c(\pi)$, this example collapses to the one under margin requirements where $m = G/r$.

Example 3 (Value-at-Risk constraints). Another family of constraints arbitrageurs can face result from Value-at-Risk (VaR) calculations as highlighted by Adrian and Shin (2014) in the study of bank leverage. Under this rule, arbitrageurs need enough equity capital to cover their VaR_α, defined as

$$\inf\left\{V > 0 : \mathbb{P}\left[\text{change in asset value} \leq -V\right] \leq 1 - \alpha\right\},$$

based on a pre-specified small threshold α. Arbitrageurs adjust their investment positions to abide by the rule. As an illustration, I consider a simple Gaussian VaR setting, under which changes in r and b are both normal; for simplicity, I further assume that these changes are independent. In summary, $\Delta r \sim \mathcal{N}(\mu_{\Delta r},\, \sigma_{\Delta r})$, $\Delta b \sim \mathcal{N}(\mu_{\Delta b},\, \sigma_{\Delta b})$ and $\Delta r \perp \Delta b$. Under this setting, $\mathrm{VaR}_\alpha = z_{(1-\alpha)}\sqrt{\pi_0^2 s^2 \sigma_{\Delta r}^2 + \pi^2 s^2 \sigma_{\Delta b}^2}$, where $\sigma_{\Delta r}$ and $\sigma_{\Delta b}$

[16] The signum function $\mathrm{sgn}(b)$ equals -1 when $b < 0$, 0 when $b = 0$, and 1 when $b > 0$.

[17] In a standard (I, K) type of investment theory presented as early as by Lucas (1967), adjustment costs are relate to *both* the investment I (πs here) and the capital stock K (s here). This is because the relative size of I given K may help determine the cost. The costs or adjustments may be due to funding value adjustments as demonstrated in Andersen, Duffie, and Song (2019), which is an implicit debt-overhang cost to equity holders. Arbitrageurs' effective funding costs and (opportunity) costs of collaterals for different currency pairs may also render CIP arbitrage less profitable (Augustin, Chernov, Schmid, and Song, 2020). And, as many may argue, counterparty credit risk adjustments may also plague the seemingly riskless CIP arbitrage, for most currency derivatives are not centrally cleared (though unfavorable evidence provided in Du, Tepper, and Verdelhan (2018)). The cost function here can also be interpreted as (credit) risk adjustments.

can be calibrated from historical data, $z_{(1-\alpha)}$ is the $[100(1-\alpha)]$th percentile of the standard normal distribution. Thus the VaR constraint $\text{VaR}_\alpha \leq s$ yields $z^2_{(1-\alpha)}\left(\sigma^2_{\Delta r}\pi_0^2 + \sigma^2_{\Delta b}\pi^2\right) \leq 1$. As a normalization, we can let $z^2_{(1-\alpha)}\sigma^2_{\Delta r} = 1$ and define $v = \sigma_{\Delta b}/\sigma_{\Delta r}$, then the set $\mathcal{C}$ becomes $\left\{0 \leq \pi_0 \leq 1,\, \pi \in \mathbb{R} : \pi_0^2 + v^2\pi^2 \leq 1\right\}$. Under this VaR constraint, arbitrageurs choose their arbitrage positions

$$\pi(b) = \frac{b}{v\sqrt{r^2v^2 + b^2}}.$$

This setting features smooth arbitrage responses to the magnitude of arbitrage yeilds, in sharp contrast to the outcomes under margin requirements. Panel (C) of Figure 1.1 illustrates the set $\mathcal{C}$ as well as the function $\pi(b)$.

Panel (D) of Figure 1.1 compares the arbitrage profit function $S_{\mathcal{C}}$ for the three examples. Of note, $S_{\mathcal{C}}$ is positively homogeneous of degree one (e.g., Molchanov and Molinari (2018, p. 75)), thus the plot shows $S_{\mathcal{C}}(1, b/r)$ as a function of b/r for cleaner demonstration. $S_{\mathcal{C}}(1,\, b/r)$ reaches its minimum value of one at $b = 0$. This is when the arbitrage opportunity does not exist, so the investment return must be r. When $|b|$ deviates from zero, $S_{\mathcal{C}}(1, b/r)$ will never decrease.

The three special cases of $\mathcal{C}$ exemplify the benefits of developing a theory *without* taking a strong stand on the form of the constraint. Different shapes of $\mathcal{C}$ lead to distinctive arbitrage responses, which translate into peculiar (in)elasticities of $\pi(b)$, the supply of forward dollars. From equation (1.7), equilibrium arbitrage yields thus differ. I summary theoretical results based on this agnostic view of arbitrage limits in propositions below.

Proposition 1. *The optimal behavior of arbitrageurs imposes the following equilibrium conditions:*

$$\frac{1}{1+\rho}\left(\frac{y'}{y}\right)^{-1}\left[1 + S_{\mathcal{C}}(r,\, b)\right] = 1$$

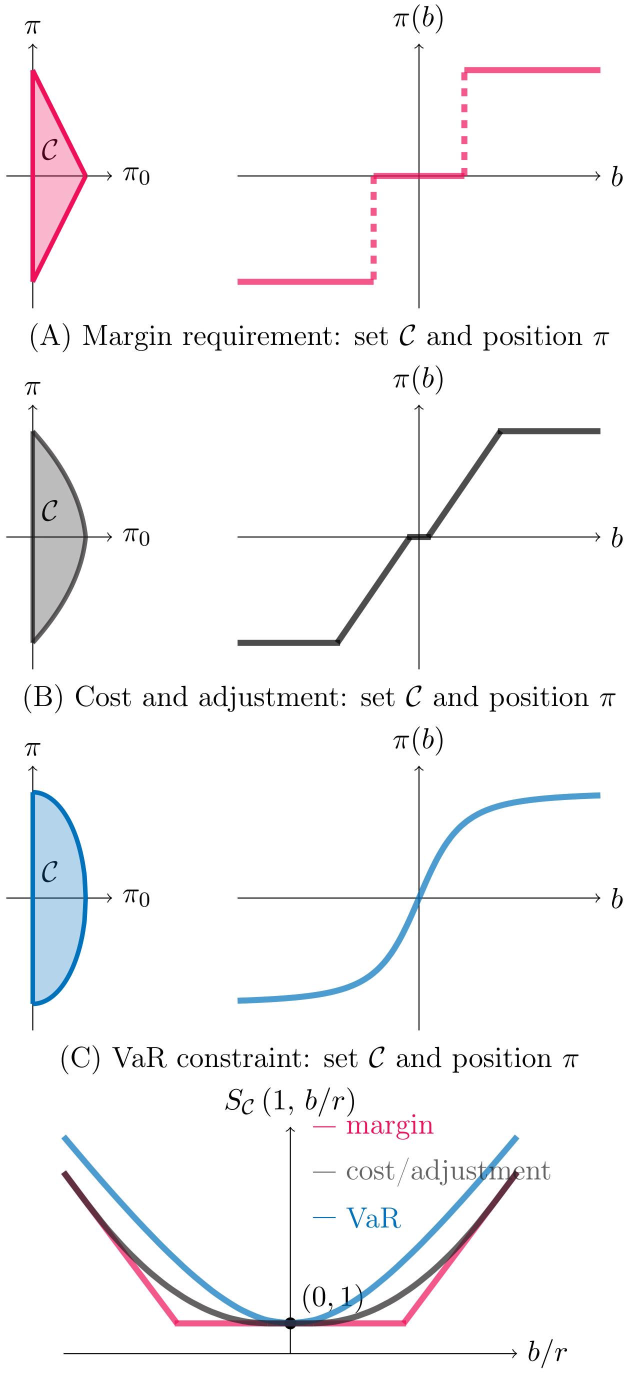
for their consumption growth (the Euler equation) and

$$\frac{1}{r}\left(\frac{k'-k}{k}\right) = \left(\frac{1}{2+\rho}\right)S_{\mathcal{C}}\left(1,\, \frac{b}{r}\right) - \left(\frac{1+\rho}{2+\rho}\right)\frac{1}{r}$$

for their capital accumulation. All else equal, the net return on arbitrageurs' capital $[(k'-k)/k]$*: i. increases in* $|b|$*;*[18] *ii. is a convex function of* b.

The consumption Euler equation is standard under the log utility, in which $(1 + S_{\mathcal{C}})$ acts as the return on intertemporal savings. Motivated by this equation, we can conceptualize arbitrageurs' decision problem as a two-stage one: they first optimize $\pi_0 r + \pi b$ subject to the constraint (1.8), which gives the optimal return $S_{\mathcal{C}}$; next, they choose their consumption plan y (and thus s, k', and y') according to the Euler equation, taking $S_{\mathcal{C}}$ as given. Arbitrage limits only prevent them from responding insatiably to arbitrage profits.

[18] Strictly speaking, all increasing or decreasing statements henceforward refer to nondecreasing or nonincreasing respectively. I avoid invoking the latter terms for conceptual simplicity, disregarding mathematical rigor.



(A) Margin requirement: set $\mathcal{C}$ and position π

(B) Cost and adjustment: set $\mathcal{C}$ and position π

(C) VaR constraint: set $\mathcal{C}$ and position π

(D) Arbitrage profit function

Figure 1.1: Examples of arbitrage constraints $\mathcal{C}$, arbitrage positions, and arbitrageurs' optimal investment returns.

Their intertemporal saving behavior remains optimal under any predetermined position constraints.

Specifications of $\mathcal{C}$ directly affects how arbitrageurs' capital accumulation responds to arbitrage yields. For example, under VaR constraints, a nonzero b lifts arbitrageurs' investment return above r, regardless how small $|b|$ is. However, with margin requirements, there exists a region around zero, within which no value of b increases the arbitrageurs' investment return.

Now we turn to optimal arbitrage positions and the equilibrium arbitrage yield. The proposition below summarizes the results.

Proposition 2. *If $\mathcal{C}$ is such that the support function $S_{\mathcal{C}}$ is differentiable, the arbitrageurs' optimal arbitrage positions are*

$$\pi(b) = \frac{\partial S_{\mathcal{C}}(r,\, b)}{\partial b}.$$

Furthermore, if $\mathcal{C}$ is such that $S_{\mathcal{C}}$ is twice differentiable, the equilibrium deviation b^* that solves $\pi(b)s = q(b)$ uniquely exists and

i. if $\gamma_0 \geq 0$, $0 \leq b^* \leq \gamma_0/\gamma$ and $q(b^*) \geq 0$; otherwise, $\gamma_0/\gamma \leq b^* \leq 0$ and $q(b^*) \leq 0$;

ii. $|b^*|$ decreases as the arbitrageurs' initial capital k increases.

From Proposition 2, we know that optimal arbitrage positions can be derived from the arbitrage profit function $S_{\mathcal{C}}$.[19] This function $S_{\mathcal{C}}$, on the other hand, reflects how arbitrageurs' investment return responds to arbitrage yields (Proposition 1). These observations lay foundations for identifying the forward dollar supply by the arbitrageurs. If we can measure the capital return $(k' - k)/k$, a nonparametric regression of this return on the arbitrage yield (e.g., CIP deviations) reveals the $S_{\mathcal{C}}$, which in turn gives us $\pi(b)$. We will revisit this idea later in the full quantitative model in Section 1.3.

The sign of equilibrium deviations is determined only by the hedgers' demand $q(b)$. Negative CIP deviations indicate that there is a net demand for foreign currencies ($q(b) < 0$) while positive deviations imply a net demand for dollars ($q(b) > 0$), in currency forwards and swaps markets. The largest possible absolute deviation in equilibrium $|b^*|$ is always less than $|\gamma_0|/\gamma$, which is the outcome when no arbitrage force exists to absorb the hedging demand imbalances. In this equilibrium, b^* is such that $q = 0$.

The log-utility assumption brings up wealth effects, thus arbitrageurs' capital k have major impacts on their absolute arbitrage positions, which equals πs.[20] Larger capital stock increases arbitrage capacity, leading to smaller arbitrage yields in equilibrium.

1.3 A quantitative equilibrium model of limited arbitrage

In this section, I develop a quantitative model of limited arbitrage in currency market by enriching the simple model of Section 1.2. The extension comes from four dimensions: (i) a risky project is now available to the arbitrageurs; (ii) multiple (instead of only one) riskless arbitrage opportunities exist; (iii) the model is dynamic in which arbitrageurs optimize their discounted life-time utility; (iv) time-varying external hedging demand exists for

[19] At points that the partial derivative $\partial S_{\mathcal{C}}(r,\, b)/\partial b$ is not well-defined, indeterminacy can arise and $\pi(b)$ falls into a closed convex set, namely the *subdifferential* of $S_{\mathcal{C}}$. See Bertsekas (2009, p. 182-186) for further expositions.

[20] As shown in the Appendix, in equilibrium, arbitrageurs' savings s is proportional to their initial capital endowment k, due to the log-utility assumption.

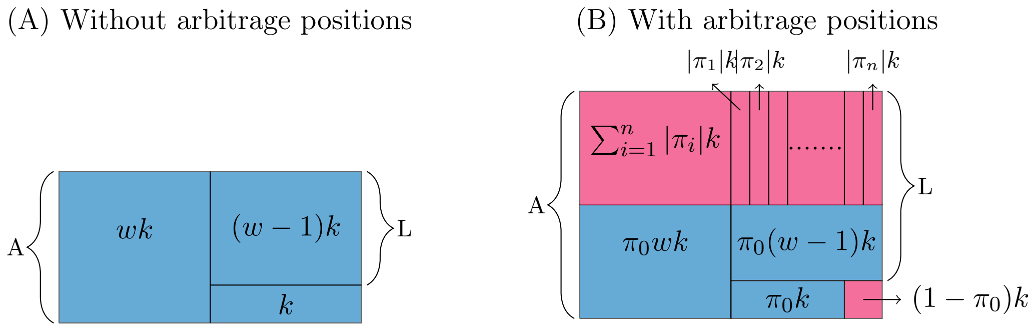


Figure 1.2: Arbitrageurs' balance sheet with and without arbitrage positions (A: asset, L: liability).

each arbitrage opportunity. The risky project and random hedging demands make the model stochastic. I characterize equilibrium outcomes of the model, test their predictions, and use the equilibrium conditions to estimate the model.

1.3.1 Model setup

Time is continuous, going from zero to infinity. As before, there are two groups of agents: arbitrageurs and hedgers, both of a continuum of mass one.

Of note, throughout the rest of the paper, I will omit the time subscripts whenever it does not cause confusion.

Arbitrageurs. Arbitrageurs maximize a utility function

$$\mathbb{E}_t\left[\int_0^\infty e^{-\rho s}\log\left(y_{t+s}\right)\,\mathrm{d}s\right], \tag{1.10}$$

in which $\rho > 0$ is the instantaneous time discount rate, and y_t is their rate of consumption at date t. At date 0, they are endowed with $k_0 > 0$ amount of capital.

As before, with date-t capital k_t at hand, arbitrageurs can borrow or save at a risk-free rate r_t, or safely store their capital (with zero net return). They can also profit from multiple riskless arbitrage opportunities, yielding at rate b_{it}, $i = 1, \ldots, n\,(n \geq 1)$, per unit of position they build up. In currency markets, these arbitrage yields are CIP deviations for different currencies.

Arbitrageurs now have access to a risky project, the net return of which follows a diffusion process $\mathrm{d}\widetilde{r}_t = (\mu_t \mathrm{d}t + \sigma_t dz_t)$ where $\{z_t\}_{t=0}^{\infty}$ is a standard Brownian motion on a complete probability space. In other words, the date-t expected rate of return of this risky project is μ_t and its volatility being σ_t. In the context of currency markets, large dealer banks play an essential role in FX arbitrage. If we treat them as the arbitrageurs, this risky project represents the a consolidated portfolio of their business activities (e.g., consumer financing, commercial banking, investment banking, security brokerage and trading, asset management, etc.), in addition to FX arbitrage.

Capital accumulation without arbitrage opportunities. Ignoring the arbitrage opportunities for now, with date-t capital k_t, arbitrageurs choose their positions in the risk project and the risk-free asset. Denote by w_t the ratio of risky project investments to total capital, their investment return within the time interval $[t,\, t+\mathrm{d}t]$ is

$$\mathrm{d}r(w_t) = w_t(\mathrm{d}\widetilde{r}_t) + (1-w_t)(r_t\mathrm{d}t),$$

where $\mathrm{d}\widetilde{r}_t = (\mu_t\mathrm{d}t + \sigma_t dz_t)$ by assumption. Their capital evolves according to $k_{t+\mathrm{d}t} = k_t[1+\mathrm{d}r(w_t)] - y_t\mathrm{d}t$, that is,

$$\frac{\mathrm{d}k}{k} = r\mathrm{d}t + w(\mu - r)\mathrm{d}t + w\sigma\mathrm{d}z - \frac{y}{k}\mathrm{d}t.$$

Following the literature (e.g., He and Krishnamurthy (2013); Brunnermeier and Sannikov (2014)), I expect $w_t > 1$, which means arbitrageurs build up leveraged positions in the risky project, funded by risk-free debt. Their balance sheet is illustrated in Panel (A) of Figure 1.2.

Capital accumulation with arbitrage opportunities. With arbitrage opportunities, arbitrageurs' date-t problem can be thought of as making two sets of decisions. On the one hand, they choose the amount of capital, denoted by $\pi_{0t}k_t$ ($\pi_0 \leq 1$), to support their "normal lines of business", that is, borrowing at the risk-free rate and making leveraged investment in the risky project. This investment, costing $\pi_{0t}k_t$ initially, leads to $\pi_{0t}k_t[1+\mathrm{d}r(w_t)]$ amount of capital at date $(t+\mathrm{d}t)$, where $\mathrm{d}r(w_t)$ follows the same definition above. The risk exposure w_t is chosen optimally under the standard risk-return trade-off. On the other hand, arbitrageurs also decide the size of arbitrage positions relative to their capital, denoted by the vector $\boldsymbol{\pi}_t = (\pi_{1t}, \ldots, \pi_{nt})^\top$, for each of the n arbitrage opportunities. Total arbitrage profits at date $(t+\mathrm{d}t)$ are $\left(\sum_{i=1}^{n} \pi_{it}b_{it}\mathrm{d}t\right)k_t$, or $(\boldsymbol{\pi}_t^\top \boldsymbol{b}_t\mathrm{d}t)k_t$ for simplicity, where $\boldsymbol{b}_t = (b_{1t}, \ldots, b_{nt})^\top$. These arbitrage profits come at zero cost at date-t. We can write down arbitrageurs' total capital at date $(t+\mathrm{d}t)$ as

$$k_{t+\mathrm{d}t} = k_t + \pi_{0t}k_t[1+\mathrm{d}r(w_t)] - \pi_{0t}k_t + (\boldsymbol{\pi}_t^\top \boldsymbol{b}_t\mathrm{d}t)k_t - 0 - y_t\mathrm{d}t,$$

which extends equation (1.3) under the new dynamic stochastic environment with multiple arbitrage opportunities. Simplifying the equation above, arbitrageurs' capital evolves according to

$$\frac{\mathrm{d}k}{k} = \pi_0\left[r\mathrm{d}t + w(\mu - r)\mathrm{d}t + w\sigma\mathrm{d}z\right] + \boldsymbol{\pi}^\top \boldsymbol{b}\mathrm{d}t - \frac{y}{k}\mathrm{d}t. \tag{1.11}$$

Panel (B) of Figure 1.2 illustrates the structure of arbitrageurs' balance sheet after building up arbitrage positions. Their original balance-sheet composition are colored in blue and arbitrage positions are colored in red. Taking advantage of arbitrage opportunities potentially leads to downsizing the normal business. In doing so, arbitrageurs are effectively setting a fraction $(1-\pi_0)$ of their capital aside in storage, earning zero net returns. We can also view this amount of capital as necessary capital buffers to support arbitrage positions (thus also colored in red). In the context of major FX dealer banks, $(1-\pi_0)k$

represents the amount of bank capital deployed to their trading desks dedicated to CIP arbitrage. The choice regarding π_0 can also be interpreted as resource allocation decisions in internal capital markets.

To sum up, arbitrageurs now choose i. π_0 that determines the size of their "conventional" investment as well as its leverage w, ii. arbitrage positions $\boldsymbol{\pi}$ in each of the arbitrage opportunities, iii. consumption rate y, to maximize their utility (1.10) subject to the capital accumulation equation (1.11).

Without arbitrage limits, arbitrageurs will choose $\pi_0 = 1$ and $|\pi_i| \to \infty$ for any $i \in \{1, \ldots, n\}$ such that $b_i \neq 0$.

Arbitrage limits arise from financial constraints defined by the set $\mathcal{C}$. Combinations of π_0 and $\boldsymbol{\pi}$ must fall within $\mathcal{C}$. Arbitrageurs face the trade-off between chasing larger arbitrage profits and downsizing their normal business operations, under this constraint. Extending equation (1.9), the arbitrage profit function (the support function of $\mathcal{C}$) is now defined as

$$S_{\mathcal{C}}(r, \boldsymbol{b}) = \sup_{(\pi_0, \boldsymbol{\pi}) \in \mathcal{C}} \{\pi_0 r + \boldsymbol{\pi}^\top \boldsymbol{b}\}.$$

Time-varying financial constraint. I allow for time variation in the set $\mathcal{C}$ to reflect changing financial constraints. Specifically, I assume

$$(\pi_{0t}, \boldsymbol{\pi}_t) \in \mathcal{C}_t. \tag{1.12}$$

Arbitrage profit functions can be defined for each $\mathcal{C}_t$ accordingly.

I collect assumptions in the previous section about the constraint set and present a summarized one below:

Assumption 5. *At any time t, $\mathcal{C}_t$ is a bounded, closed and convex subset of $[0, 1] \times \mathbb{R}^n$, which is nonempty with $(1, \mathbf{0}_n) \in \mathcal{C}_t$; the arbitrage position $\pi_{it} = 0$ when $b_{it} = 0$.*

Hedgers. I extend the hedging demand specification of (1.5) to each of the n arbitrage opportunities by assuming

$$\boldsymbol{q}_t = \boldsymbol{\gamma}_{0,t} - \gamma \boldsymbol{b}_t, \quad \gamma > 0, \tag{1.13}$$

where elements in $\boldsymbol{q}_t = (q_{1t}, \ldots, q_{nt})^\top$ are external (net) hedging demands. Following the interpretations in Section 1.2 in the context of CIP arbitrage, q_{it} represents the demand for forward dollars via the exchange of currency i. The vector $\boldsymbol{\gamma}_{0,t} = (\gamma_{01,t}, \ldots, \gamma_{0n,t})^\top$ captures the fundamental hedging demand differences among currencies, due to trade imbalances or cross-border investment. The positive scaler γ indicates that hedging demands for forward dollars are always decreasing in the CIP deviations. Appendix 1.7.3 further discusses micro-foundation of this specification.

The equilibrium arbitrage yield. At time t, the equilibrium arbitrage yield $\boldsymbol{b}_t^*$ is a vector such that

$$\boldsymbol{\pi}_t k_t = \boldsymbol{q}_t. \tag{1.14}$$

where $\boldsymbol{q}_t$ is defined by (1.13); $\boldsymbol{\pi}_t$ are (part of) the solutions to the arbitrageurs' problem: choosing $\{y_t,\, w_t,\, \pi_{0t},\, \boldsymbol{\pi}_t\}$ to maximize the utility function (1.10) subject to the capital accumulation equation (1.11) under the constraint (1.12).[21]

Before characterizing equilibrium outcomes, I add three remarks to finish describing the model setup. First, the model does not consider risky arbitrage. As a result, it is not suitable for investigating many intriguing pricing phenomena such as stock market anomalies. For CIP arbitrage, instead of treating it as risky payoffs, standard practices apply valuation adjustments, modify margin requirements, or resort to VaR calculations to address risk-related concerns (e.g., the counterparty risk and the mark-to-market valuation risk). Financial constraints $\mathcal{C}_t$ in the current model encompass these scenarios, as illustrated by examples discussed in the previous section. In addition, CIP arbitrage does not involve convergence trading and is not subject to the (endogenous) risk induced by random arbitrage horizons in models such as Kondor (2009). Thus, this riskless arbitrage model tends to be a good fit for studying CIP deviations and other "near-arbitrage" bases, such as the positive gap between the interest on excess reserve rate and the reverse repo rate.

Second, the log-utility assumption, although inducing myopic behaviors, is not restrictive for analyzing riskless arbitrage. The intuition is that arbitrageurs cannot exploit riskless arbitrage opportunities to hedge against future shocks to their assets and financial constraints. In other words, they only adjust arbitrage positions in response to contemporaneous shocks. On the other hand, arbitrageurs do adjust risk exposures through their positions on the risky project, taking into consideration their changing investment opportunities. Appendix 1.7.1 presents and characterizes equilibrium outcomes of the same model under general CRRA utility specifications to clarify these points.

Third, the risky project in the model prevents us from carrying over model solutions of Section 1.2 directly. To see this more clearly, in equation (1.4) of the previous section, arbitrageurs' log investment return is $\log(1+\pi_0 r+\pi b)$. Maximizing it under the constraint $(\pi_0,\, \pi) \in \mathcal{C}$, it is almost trivial to see that, in equilibrium, $\pi_0 r + \pi b = S_{\mathcal{C}}(r,\, b)$. Under the full model here, arbitrageurs' instantaneous (expected) log investment return is $\mathbb{E}\log[1 + \pi_0 \mathrm{d}r(w) + \boldsymbol{\pi}^\top \boldsymbol{b}\mathrm{d}t]$ (ignoring consumption, long-horizon log investors effectively maximize expected log returns period by period). Maximizing it under the constraint (1.12) is not straightforward. I develop theoretical tools to solve this type of problems building on the concept of convex conjugacy (also see Appendix 1.7.1 for details). These tools also apply to the "true" dynamic setting under CRRA utilities. As a preview, the results are surprisingly simple: a multivariate generalization $\pi_0 r + \boldsymbol{\pi}^\top \boldsymbol{b} = S_{\mathcal{C}}(r,\, \boldsymbol{b})$ still holds in equilibrium and risk exposures of arbitrageurs are adjusted through changing w (although π_0 also affects the size of their risky positions, it is completely pinned down by $\mathcal{C}$).

[21] Assumption 8 in Appendix 1.7.1 summarizes additional technical assumptions regarding investment opportunities $(r_t,\, \mu_t,\, \sigma_t)$, financial constraints $\mathcal{C}_t$, and external hedging demands $\gamma_{0,t}$.

1.3.2 Equilibrium characterization

This section characterizes the equilibrium outcomes. I first present arbitrageurs' optimal choices and their capital dynamics in equilibrium. Then I show the equation that equilibrium arbitrage yields must satisfy and analyze its properties. Again, most time subscripts are suppressed.

Recall that arbitrageurs' choice variables include their consumption rate y, the amount of capital used for their "standard business" π_0, their risky asset weight w, and the vector $\boldsymbol{\pi}$ determining their arbitrage positions. I start presenting their choices of π_0 and $\boldsymbol{\pi}$ in the following proposition.

Proposition 3. *If $\mathcal{C}$ is such that $S_{\mathcal{C}}$ is differentiable, equilibrium arbitrage positions are given by*

$$\pi_i = \frac{\partial S_{\mathcal{C}}(r,\, \boldsymbol{b})}{\partial b_i}, \quad \textit{for all } i = 1, \ldots, n.$$

In equilibrium, the fraction of capital maintained for investment opportunities other than arbitrage is given by

$$\pi_0 = \frac{\partial S_{\mathcal{C}}(r,\, \boldsymbol{b})}{\partial r}.$$

Optimal arbitrage positions are not affected by the profile of the risky project (i.e., its risk and return captured by μ and σ), but determined fully by the risk-free rate r in combination with arbitrage yields $\boldsymbol{b}$. The shape of $\mathcal{C}$ determines the specific functional form of $\boldsymbol{\pi}$ with regard to "prices" $(r,\, \boldsymbol{b})$ via the arbitrage profit function $S_{\mathcal{C}}$. Examples of this function are available in Figure 1.1 discussed in the previous section.

According to Proposition 3, arbitrageurs have to set aside $(1 - \partial S_{\mathcal{C}}(r,\, \boldsymbol{b})/\partial r)$ fraction of their total equity capital to support their optimal choice of arbitrage positions. This choice is again not affected by the risk and return characteristics defined by μ and σ. The remaining fraction will be used for building up risky asset positions of size $w(\partial S_{\mathcal{C}}(r,\, \boldsymbol{b})/\partial r)$.

In Appendix 1.7.1, I show that the optimal π_0 and $\boldsymbol{\pi}$ given by Proposition 3 do not change for general CRRA utility functions. Proposition 4 below provides arbitrageurs' optimal choices of y and w. Its generalization for CRRA utility functions yields more complicated results, which are provide in Proposition 7 of Appendix 1.7.1.

Proposition 4. *In equilibrium, arbitrageurs' optimal rate of consumption y is such that $y = \rho k$; their position on the risky project $\pi_0 w$ equals $(\mu - r)/\sigma^2$, that is,*

$$w = \frac{\mu - r}{\sigma^2} \left(\frac{\partial S_{\mathcal{C}}(r,\, \boldsymbol{b})}{\partial r} \right)^{-1}.$$

Arbitrageurs' choice of total risky asset exposure $(\pi_0 w)$ exhibits the behavior of classical "Mertonian" demand (Merton, 1973).[22] The rate of consumption $y = \rho k$ is a standard

[22] Under the current log utility case, this quantity equals the myopic mean-variance efficient demand $(\mu - r)/\sigma^2$ (Proposition 4 above). In Proposition 7 of the Appendix 1.7.1, I extend the result for general CRRA utilities which account for both intertemporal hedging and endogenous dynamic risk aversion. Its form still complies with the "Mertonian" demands under intertemporal settings.

result under the log utility.[23] We can always interpret their choices as a two-stage sequence. First, given the "price vector" $(r,\, \boldsymbol{b})$ and knowing their constraints $\mathcal{C}$, arbitrageurs nail down the size of their arbitrage positions $\boldsymbol{\pi}$ and set aside $(1-\pi_0)$ fraction of their total capital in support of these arbitrage activities. Second, with the remaining $\pi_0 k$ amount of capital ready for use, arbitrageurs solve the standard consumption-saving problem with assets defined by the triplet $(r,\, \mu,\, \sigma)$.

I now present the dynamics of arbitrageurs' capital in equilibrium, which serves as the key identifying equation for quantitative analysis.

Proposition 5. *In equilibrium, the arbitrageurs' capital evolves according to the following rule:*

$$\frac{\mathrm{d}k}{k} = \left[S_{\mathcal{C}}(r,\, \boldsymbol{b}) - \rho + \lambda^2\right] \mathrm{d}t + \lambda \mathrm{d}z, \tag{1.15}$$

where $\lambda = (\mu - r)/\sigma$ is the Sharpe ratio of the risky project available to arbitrageurs.

Proposition 5 allows for intuitive interpretations. To see this, plugging the two equilibrium conditions $y = \rho k$ and $\pi_0 w = \lambda/\sigma$ from Proposition 4 into the budget constraint (1.11), we have

$$\frac{\mathrm{d}k}{k} = \Big[\Big(\pi_0 r + \boldsymbol{\pi}^\top \boldsymbol{b}\Big) - \rho + \lambda^2 \Big] \mathrm{d}t + \lambda \mathrm{d}z.$$

Comparing the equation above with equation (1.15) in Proposition 5, we can see that, the equilibrium π_0 and $\boldsymbol{\pi}$ are such that

$$\pi_0 r + \boldsymbol{\pi}^\top \boldsymbol{b} = S_{\mathcal{C}}(r,\, \boldsymbol{b}).$$

The result indicates that when solving the infinite horizon optimization problem, arbitrageurs still behave as if they were solving the simple problem of maximizing $(\pi_0 r + \boldsymbol{\pi}^\top \boldsymbol{b})$ subject to $(\pi_0,\, \boldsymbol{\pi}) \in \mathcal{C}$, when choosing π_0 and $\boldsymbol{\pi}$ each period. This optimization problem is a multivariate extension of solutions to the simple model presented in Section 1.2 in which only one arbitrage opportunity exists.

Another way of looking at the dynamics of arbitrageurs' capital is through the view of Euler equations or, equivalently, stochastic discount factors (SDF). Let us define $\Lambda_t = e^{-\rho t}/y_t$. Then, under the log utility, optimal intertemporal choices of the arbitrageurs enforce that $\mathrm{d}\Lambda/\Lambda$ is an SDF, pricing the risky asset(s) available to them. As $y = \rho k$, $\Lambda_t = e^{-\rho t}/(\rho y_t)$, Proposition 5 indicates that

$$\frac{\mathrm{d}\Lambda}{\Lambda} = -S_{\mathcal{C}}(r,\, \boldsymbol{b})\mathrm{d}t - \lambda \mathrm{d}z.$$

The risk premium of the risky project $(\mu - r)\mathrm{d}t$ equals $\mathbb{E}[(-\mathrm{d}\Lambda/\Lambda)\mathrm{d}\tilde{r}]$, the opposite of its return covariance with this specific SDF defined by the consumption (or capital) of the arbitrageurs. In other words, the consumption Euler equation holds in the model. Constraints on the arbitrage positions do not render arbitrageurs' intertemporal consumption

[23] For general CRRA utilities, the equilibrium ratio y/k varies according to the state of the economy. Proposition 7 of the Appendix 1.7.1 presents the general result.

and portfolio choice suboptimal. When $\boldsymbol{b}$ is a vector of zeros, that is, no arbitrage opportunity exists, $S_{\mathcal{C}}(r, \boldsymbol{b}) = r$.[24] The SDF takes the conventional form of $(-r\mathrm{d}t - \lambda \mathrm{d}z)$ in continuous time. Riskless arbitrage opportunities effectively serve as a "booster technology" to ramp up the risk-free rate available to the arbitrageurs.

The next proposition shows the system of equations determining the equilibrium level of arbitrage yields. It also discusses sufficient conditions for the existence and uniqueness of the equilibrium.

Proposition 6. *The equilibrium arbitrage yields* $\boldsymbol{b}$ *under* (1.14) *solves*

$$\frac{\partial S_{\mathcal{C}}(r, \boldsymbol{b})}{\partial \boldsymbol{b}} k = \boldsymbol{\gamma}_0 - \gamma \boldsymbol{b}. \tag{1.16}$$

If the scalar $\gamma > 0$ *and the set* $\mathcal{C}$ *satisfying Assumption 5 also guarantees that the arbitrage profit function* $S_{\mathcal{C}}$ *is twice continuously differentiable, a unique solution exists in a ball in* $\mathbb{R}^n$ *centered at zero with a radius* $\|\boldsymbol{\gamma}_0\|_2/\gamma$.[25]

Proof of existence relies on verifying sufficient conditions for the Brouwer fixed-point theorem; uniqueness is a result of the implicit function theorem. Appendix 1.7.1 contains details of the proof. The radius $\|\boldsymbol{\gamma}_0\|_2/\gamma$ corresponds to the norm of arbitrage yields without arbitrageurs, that is, when $\boldsymbol{b} = \boldsymbol{\gamma}_0/\gamma$. Under this scenario, the vector $\boldsymbol{b}$ adjusts such that all hedging demand imbalances equal zero. Arbitrageurs help reduce the overall equilibrium arbitrage yields in the sense that their norms become smaller than $\|\boldsymbol{\gamma}_0\|_2/\gamma$. Arbitrage forces dampen the influences of the "raw hedging demands" $\boldsymbol{\gamma}_0$ (hedging demands when no CIP deviations exist) on $\boldsymbol{b}$: the response of equilibrium $\boldsymbol{b}$ to changes in $\boldsymbol{\gamma}_0$ is less than $1/\gamma$ whenever there are arbitrageurs taking advantage of the arbitrage opportunities induced by demand imbalances.

1.4 Empirics: testing model predictions and estimating the model

1.4.1 Data

Beyond CIP deviation measures described early on, I assemble data from several other sources. I collect the trade-weighted broad dollar index, the VIX index, Fed fund rates, treasury yields, euro implied volatilities (CBOE EuroCurrency ETF Volatility Index) from Federal Reserve Economic Data (FRED). I download the yield curve of RefCorp strips from Bloomberg (for calculating the dollar convenience yields following Longstaff (2004)).

I also collect bilateral trade data from IMF Direction of Trade Statistics; bilateral portfolio transaction data as well as cross-border bank claim data from the US Treasury

[24] Recall that Assumption 5 requires $\mathcal{C} \subset [0, 1] \times \mathbb{R}^n$, under which π_0 is always smaller than one.

[25] Of note, $\mathcal{C}$, $\boldsymbol{\gamma}_0$, k and r all vary across time in the model. The equilibrium condition holds one by one at each time point. The existence and uniqueness results thus apply only to each time period for a given collection of $\{\mathcal{C}, \boldsymbol{\gamma}_0, k, r\}$. The proposition is silent on the Markovian equilibrium under which we are interested in the property of a mapping from the state of the economy to equilibrium arbitrage yeilds $\boldsymbol{b}$ such that the equation in this proposition always holds. I leave this exploration for future research.

International Capital (TIC) System; bilateral foreign direct investment data from the US Bureau of Economic Analysis.

I create a measure of arbitrageurs' capital in currency markets. It is motivated by the fact that these markets are predominantly dealer-intermediated. I consider 49 global dealer banks which are participants of semi-annual foreign exchange turnover surveys (FXS) by local monetary authorities in New York, London, Tokyo, Toronto, Sydney, Singapore, and Hong Kong. Table 1.13 of Appendix 1.7.4 lists names of their holding companies. The equity capital of these dealer banks' holding companies is my intended measure of arbitrageurs' capital. Their fundamental and price data come from Compustat and CRSP,[26] I use Bloomberg to access their five-year credit default swap (CDS) rates.

1.4.2 Supporting evidence of the model

Without committing to specific financial constraints, the model still yields a strong prediction: an increasing arbitrage yield (e.g., the CIP deviation) should predict higher returns on arbitrageurs' capital, and as the former goes up, the latter should go up increasingly fast (a convex relationship). The equilibrium outcome stated in equation (1.15) illustrates this point. On the left-hand side of this equation are arbitrageurs' capital returns next period, and on the right-hand, the function $S_{\mathcal{C}}$, which is increasing and convex in $\boldsymbol{b}$, the arbitrage yields. It is worthwhile reiterating that the convexity of this function is a direct result of Assumption 6 that $\mathcal{C}$ is always convex.

Arbitrageurs' capital returns in currency markets

The most direct measure of the 49 FX dealer banks' equity capital is the book equity (BE) of their holding companies. As suggested by the theory, if these banks are indeed *the* arbitrageurs of FX derivatives markets, CIP deviations should predict returns on their equity capital. I compute for each bank their book equity returns (growth of book equities next quarter divided by present book equity levels) and estimate the following panel regressions:

$$\frac{1}{\tau}\text{return}_{i,t+\tau} = \alpha_i + \beta\,\bar{b}_t + \epsilon_{i,t+\tau},$$

where returns on the left hand side are annualized by dividing $\tau = 0.25$ (a quarter), the subscript i denotes banks and t stands for quarters. The independent variable $\bar{b}_t$ is the cross-sectional average of absolute one-year basis swap rates for EUR, JPY, GBP, AUD, CAD, and CHF (namely, the G6 currencies) against the dollar. Sample periods begin from March 2009 and end at December 2019.[27] The first two columns in Table 1.2 show the regression results. Overall, average CIP deviations significantly predict these banks' book equity growth: one standard deviation increase in the deviations is associated with around 1.6 percentage points increase in FX dealer banks' book equity. This finding is robust to

[26] All variables are calculated (or derived) based on data from database name ©CRSP daily stock, Center for Research in Security Prices (CRSP®) The University of Chicago Booth School of Business.

[27] I use this sample period to avoid tumultuous periods of the global financial crisis and the COVID-19 pandemic.

measurements of CIP deviations using either the currency swap rates or the forward-OIS implied basis.

There is a major drawback in using the book equity measure: it is an accounting variable that is observable only quarterly. This drawback will become more pronounced as later model testing and estimation include nonparametric procedures. To circumvent this issue, a the potential surrogate measure, the market equity (ME), becomes particularly attractive. This measure comes from high quality real-time market price data. It is also worthwhile noting that, for the 49 FX dealer banks under study, their average and median market-to-book (MB) ratios equal 1.10 and 1.05 respectively (during the sample period of 2009-2019).[28] The time-series standard deviation of market-to-book ratios averaged across these banks is 0.13. These features partially motivate the use of market equity.

An additional motivation for using the market equity measure (at least in the context of testing the predictive relationship here) is the fact that CIP deviations *do not* predict changes in the market-to-book ratios. As the equation

$$\frac{\mathrm{BE}_{t+1}}{\mathrm{BE}_t} \times \frac{\mathrm{MB}_{t+1}}{\mathrm{MB}_t} = \frac{\mathrm{ME}_{t+1}}{\mathrm{ME}_t}$$

holds by definition, if a predictor does not predict the ratio $\mathrm{MB}_{t+1}/\mathrm{MB}_t$ which stands for "returns" on the book-to-market ratio, it must simultaneously predict (or fail to predict) book and market equity returns. The third and fourth columns of Table 1.2 verify this conclusion by regressing $(\mathrm{MB}_{t+1}/\mathrm{MB}_t - 1)$ on $\bar{b}_t$. The slope coefficients are statistically indistinguishable from zero. Since we have already seen from the same table that $\bar{b}_t$ predict book equity returns of the 49 FX dealers, this variable should also predict their market equity returns.

Now I redo the panel regression using market equity returns. Both measures of CIP deviations are considered. For comparison, I still consider quarterly observations of quarterly returns. The last two columns of Table 1.2 document the results. Average CIP deviations also significantly predict these banks' market equity returns. The slope coefficients are larger: one standard deviation increase in the deviations is associated with six percentage points increase in expected market equity returns. The larger regression coefficient (compared with the case for book equity returns) is mainly due to the fact that market equity returns are more volatile than book equity returns: annualized time-series volatilities are 28.8% for the former and 9.8% for the later.

From now on, I will use returns on market equity of the 49 FX dealer banks' holding companies to measure arbitrageurs' capital returns in the model. Using market returns raises concerns about confounding effects of other return predictors, such as valuations ratios and volatilities. In the following section, I will further investigate the predictive relationship and try to mitigate these concerns by adjusting for potential return predictors.

[28] Market-to-book ratios of bank equities are around one not only during the post crisis period under study. In fact, these ratios have been close to one until the mid 1990s. During the exceptional period of 1996-2008, MB ratios of banks were over two. Explanations to these patterns are beyond the scope of this paper. Interested readers may refer to papers such as Calomiris and Nissim (2014); Atkeson et al. (2019).

CIP deviations predict arbitrageurs' capital returns

I document additional evidence on the predictive power of CIP deviations on FX dealer banks' capital returns. Columns headed with "FXS" in Table 1.3 present results from the following time-series regressions using quarterly observations:

$$\frac{1}{\tau}\text{return}_{t+\tau} = \beta_0 + \beta\,\overline{b}_t + \epsilon_{t+\tau},$$

where the dependent variable is one-quarter-ahead ($\tau = 0.25$) value- or equal-weighted stock returns of the 49 dealer banks' holding companies. Returns are annualized by dividing τ. The independent variable $\overline{b}_t$ is still the cross-sectional average of absolute one-year basis swap rates of G6 currencies against the dollar. Sample periods begin from March 2009 and end at December 2019. Regression coefficients are statistically significant in these columns of Table 1.3. On average, one basis point increase in the average CIP deviations predicts around two percentage points increase in the returns of arbitrageurs' capital.

A set of placebo tests are included Table 1.3. The same predictive regressions are repeated for returns of five exchange-traded funds (ETFs) tracking the S&P500 index (SPY), the global financial sector (IXG), the US financial sector (IYF), US broker-dealers and securities exchanges (IAI), and US insurance companies (KIE). Average CIP deviations *do not* predict placebo outcomes, except for returns of the ETF tracking the global financial sector. This unique positive finding is not surprising as the 49 FX dealer banks are likely to be essential constituents of the fund. These placebo tests suggest that the 49 global dealer banks under consideration do play special roles in CIP arbitrage: they tend to be *the* arbitrageurs both in my model and in reality.

Table 1.4 assembles additional results for the same time-series regression using daily and monthly observations. Regression coefficients are remarkably stable for the main outcome variable: value-weighted equity returns on the 49 FX dealers banks' holding companies. One basis point increase in the average CIP deviations is still associated with around two percentage point increase in these returns. Placebo test results remain consistently negative (again, except for the ETF tracking the global financial sector). For monthly observations, five hedge fund index returns are also included for placebo tests: one global composite index from BarclaysHedge, four indices from Hedge Fund Research tracking global composite, relative value arbitrage, global-macro, and macro-currency strategies. CIP deviations do not predict the composite hedge fund return indices.[29]

Table 1.5 presents results from the adjusted version of the predictive regression:

$$\frac{1}{\tau}\text{return}_{t+\tau} = \beta_0 + \beta\,\overline{b}_t + \phi \cdot \text{control}_t + \epsilon_{t+\tau},$$

in which control variables include earnings yields and dividend yields averaged across the

[29] All results till now focus on quarterly returns. Table 1.14 in Appendix 1.7.4 also confirms the predictive relationship (as well as negative results from placebo tests) for *monthly* returns. The regression coefficients remain stable (around two) for the main outcome variable, though adjusted R^2s drop for monthly returns.

49 FX dealer banks' holding companies. Quarterly returns are annualized by dividing $\tau = 0.25$. The effective fed fund rates, as well as the VIX index are also incorporated. Results for both monthly and daily observations are reported. Average G6 currency CIP deviations still demonstrate significant predictive power, but the magnitude is reduced by a half after controlling for the banks' earnings yields, which also strongly predict the returns. Table 1.15 reports results from repeating the same exercise for monthly returns. All results remain largely unchanged, except for the declined adjusted R^2s. To sum up, this set of time-series regressions suggest that CIP deviations predict arbitrageurs' capital returns, both before and after controlling for common return predictors. In addition, banks' earnings yields also emerge as an important return predictor.

The predictive relationship is convex

I now further investigate whether the predictive relationship is convex, as suggested by equation (1.15) in Proposition (5). As the term $S_{\mathcal{C}}(r, \boldsymbol{b})$ contains both the risk-free rate and the CIP deviations, I rewrite equation (1.15) as follows:

$$\frac{1}{\mathrm{d}t}\frac{\mathrm{d}k/k}{r} = S_{\mathcal{C}}\left(1, \frac{\boldsymbol{b}}{r}\right) - \frac{\rho}{r} + \frac{\lambda^2}{r} + \frac{\lambda \mathrm{d}z}{\mathrm{d}t},$$

after dividing both sides by $r\mathrm{d}t$ and leveraging the property of $S_{\mathcal{C}}$ that it is positively homogeneous of degree one. This motivates the following regression specification

$$\frac{1}{\tau}\left(\frac{\text{return}_{t+\tau}}{r_t}\right) = S_0\left(\frac{\bar{b}_t}{r_t}\right) + \phi \cdot \text{control}_t + \varepsilon_{t+\tau},$$

in which r_t denotes the risk-free rate; $\tau = 0.25$ denotes the time interval of one quarter; controls include the reciprocal of r_t (as suggested by the theory in which $-\rho/r$ shows up), the earnings yield that emerges as a strong return predictor in the previous section, as well as the VIX index; the function $S_0(\cdot)$ captures the functional form of $S_{\mathcal{C}}(1, \cdot)$.

To begin with, I contrast the parametric configuration of $S_0(x) = \beta_0 + \beta x$ and $S_0(x) = \beta_0 + \beta x^2$ in Table 1.6.[30] The slope coefficient β is significantly positive *only* under the quadratic specification. Estimates of these coefficients are stable across daily and monthly observations. To mitigate concerns about low risk-free rates creating large dependent variables to the extent that some may become "outliers", I redo the same regressions using robust estimators based on the Huber loss function. Robust estimators confirm that β is only significant under the quadratic specification, suggesting a convex predictive relationship.

Next, I estimate the equation using semi-parametric techniques. The nonparametric component S_0 is expanded to shape-constrained B-spline basis (Eilers and Marx, 1996). Table 1.7 reports the estimation results and tests for the significance of $S_0(\bar{b}_t/r_t)$ using both daily and monthly observations. I consider three types of configurations for S_0:

[30] I do not incorporate the linear and quadratic terms simultaneously due to the potential multicollinearity concern: correlation between $\bar{b}/r$ and $(\bar{b}/r)^2$ is 0.9 in the sample.

i. both convex and increasing (as suggested by theory), ii. only increasing, and iii. no restrictions. Under all conditions, regression coefficients for $1/r_t$ are significantly negative, as suggest by the theory. The nonparametric term $S_0(\bar{b}_t/r_t)$ cannot be ignored. Figure 1.5 plots the estimated $S_0(x)$ (which equals $S_{\mathcal{C}}(1, x)$) under the three specifications. Convex patterns consistently show up even without imposing the convexity constraint. A kink exists around $x = 3$, before which S_0 increases slowly and after which the function shoots up, indicating substantial arbitrage profits when $|b| > 3r$. As the median level of r is 16 basis points, arbitrageurs appear to enjoy large arbitrage profits after CIP deviations exceed 48 basis points.

This semi-parametric estimation implicitly adopts one crucial assumption: the financial constraint $\mathcal{C}$ does not change across time. Next, I will account for the dynamics of financial constraints and formally estimate the model.

1.4.3 Quantitative specifications, identification, and estimation

In this section I enrich the theoretical model presented above with additional assumptions to map it to data. The main goal is to quantify $\mathcal{C}_t$ (financial constraints) and $(\gamma_{0,t}, \gamma)$ (hedging demands and the elasticity parameter) in equation (1.16) of Proposition 6. To achieve this goal, I adopt a two-step estimation strategy. First, I estimate $\mathcal{C}_t$ using the equilibrium capital accumulation equation (1.15). Then, knowing $\mathcal{C}_t$ and thus the function $S_{\mathcal{C}_t}$, I compute the equilibrium arbitrage positions on the left hand side of equation (1.16), and then estimate hedging demands. I begin with an assumption simplifying the financial constraints.

Separating shapes and dynamics of financial constraints

Time-varying financial constraints $\{\mathcal{C}_t\}_{t\geq 0}$ is a series of sets satisfying Assumption 5. Estimating a sequence of random sets is challenging (if not impossible). To make progress, I adopt a simplifying assumption about the financial constraints by separating their shapes and dynamics.

Assumption 6. *There exists a constant set $\mathcal{C}_0$, such that $\mathcal{C}_t = \{(\pi_0, \alpha_t \boldsymbol{\pi}) : (\pi_0, \boldsymbol{\pi}) \in \mathcal{C}_0\}$ for a sequence $\alpha_t > 0$.*

This assumption implies that at any time, financial constraints defined by the set $\mathcal{C}_t$ is derived from a "baseline" $\mathcal{C}_0$ by shifting the largest possible arbitrage positions. The time series $\{\alpha_t\}$ serves the role of "shifters", which captures variation in the financial constraints. When $\alpha_t > 1$, $\mathcal{C}_t$ subsumes the baseline specification $\mathcal{C}_0$, larger arbitrage positions become feasible conditional on the same amount of capital dedicated to arbitrage activities (π_0 fixed). When $0 < \alpha_t < 1$, $\mathcal{C}_t$ shrinks, and arbitrageurs tend to cut back their arbitrage positions.

Under Assumption 6, the shape and dynamics of financial constraints each has a concrete characterization: the set $\mathcal{C}_0$ for the shape and the sequence $\{\alpha_t\}$ for the dynamics.

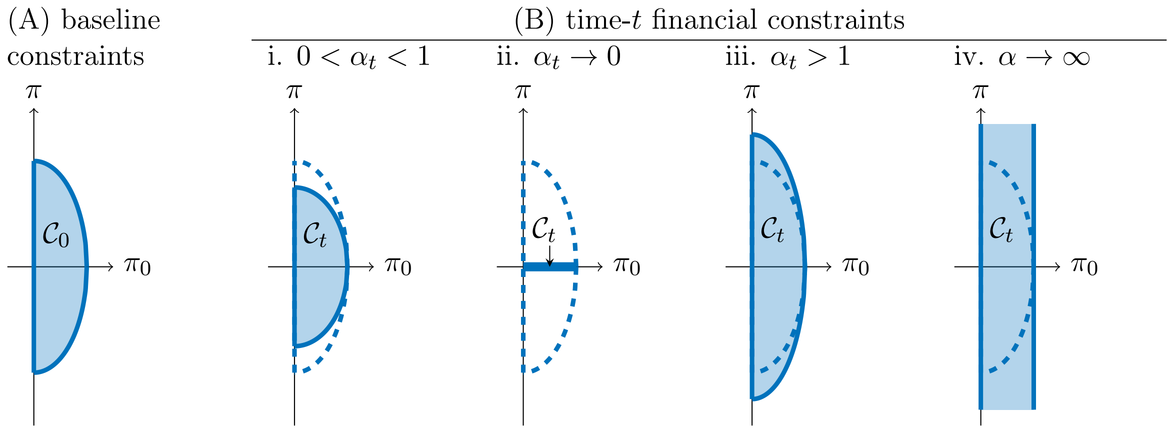


Figure 1.3: Dissecting the shape and dynamics of financial constraints

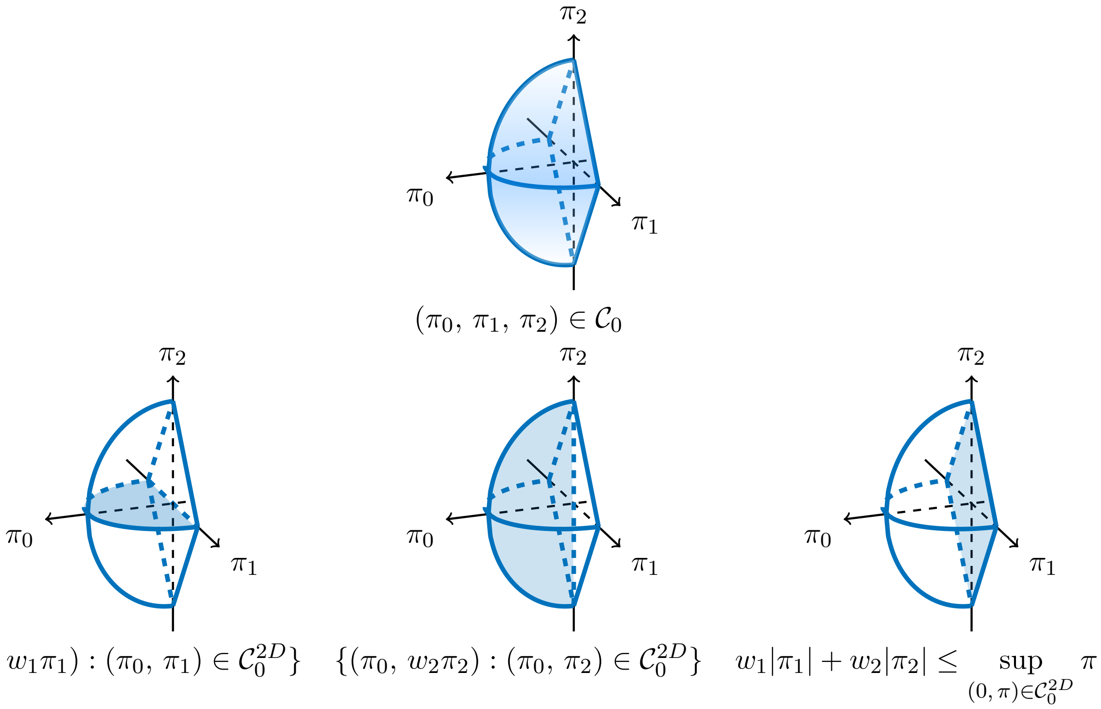
I discuss this dissection of financial constraints intuitively through Figure 1.3.[31] The first plot to the left of Figure 1.3 shows the baseline constraints defined by the set $\mathcal{C}_0$. This set determines the shape of financial constraints. In this illustration, it represents a VaR condition applied to the arbitrage positions (see examples from the previous section for details). $\mathcal{C}_0$ can also depict other types of constraints or combinations of multiple constraints. The other four plots in Figure 1.3 describe how a sequence of sets $\{\mathcal{C}_t\}$ is generated from combining $\mathcal{C}_0$ and $\{\alpha_t\}$. The time series $\{\alpha_t\}$ translates to the dynamics of financial constraints, according to Assumption 6. Plot (B)-i. and (B)-iii of Figure 1.3 illustrate how the financial constraints become tighter or loser from their baseline level according to the value α_t ($\mathcal{C}_0$ boundaries outlined in the dashed curves for comparison). Plot (B)-ii. and (B)-iv of Figure 1.3 are two extreme cases. Under the first scenario, α_t goes to zero and the set $\mathcal{C}_t$ collapses to a line segment: no arbitrage activities are allowed. All hedging demand imbalances have to be counterbalanced by large arbitrage yields. Under the second scenario, α_t becomes infinitely large, and the constraints morph into a band spanning to infinity: no limits to arbitrage exist. This corresponds to the frictionless benchmark, under which CIP deviations must always be zero.

The arbitrage profit function for time-t financial constraints $\mathcal{C}_t$ under Assumption 6 is given by the following lemma.

Lemma 1. *Under Assumption 6,* $S_{\mathcal{C}_t}(r, \boldsymbol{b}) = S_{\mathcal{C}_0}(r, \alpha_t \boldsymbol{b})$.

Lemma 1 translates Assumption 6 on sets $\{\mathcal{C}_t\}$ into properties of the arbitrage profit function. The baseline set $\mathcal{C}_0$ determines the functional form of $S_{\mathcal{C}_0}$ (shape); the series $\{\alpha_t\}$ induce time variation to the financial constraints, as well as arbitrage profit functions (dynamics). Quantifying the financial constraints is equivalent to estimating both the function $S_{\mathcal{C}_0}$ and the sequence α_t.

[31] For the ease of exposition, the illustrations cover the case of one arbitrage opportunity, while the intuitions easily carry over to higher dimensions.



$(\pi_0,\, \pi_1,\, \pi_2) \in \mathcal{C}_0$

$\{(\pi_0,\, w_1\pi_1) : (\pi_0,\, \pi_1) \in \mathcal{C}_0^{2D}\}$ $\{(\pi_0,\, w_2\pi_2) : (\pi_0,\, \pi_2) \in \mathcal{C}_0^{2D}\}$ $w_1|\pi_1| + w_2|\pi_2| \leq \sup_{(0,\,\pi)\in\mathcal{C}_0^{2D}} \pi$

Figure 1.4: The (baseline) financial constraint $\mathcal{C}_0$ and its two-dimensional generator $\mathcal{C}_0^{2D}$ under Assumption 7.

Parameterization

I introduce a simplifying assumption and parameterize three components of the model to facilitate estimation, summarized by four items in this section.

Item 1: reducing the dimension of financial constraints. I begin by simplifying the (baseline) shape of financial constraints, defined via the function $S_{\mathcal{C}_0}$. The challenge to estimate this object comes from the "curse of dimensionality": as a rule of thumb, estimating a function of dimension d generally requires a sample size that is an exponential of d (Stone, 1982). I adopt the following simplifying assumption to sidestep this challenge.

Assumption 7. *$S_{\mathcal{C}_0}(1,\, \boldsymbol{b}) = S_0(\bar{b})$ where $\bar{b} = \sum_{i=1}^{n} w_i |b_i|$ and $\sum_{i=1}^{n} w_i = 1$.*

Interpretation of the assumption is straightforward. It treats the weighted average of CIP deviations as a measure of overall arbitrage yields accessible to arbitrageurs. I use over-the-counter FX derivatives trading volume to construct these weights. The derivatives include FX forwards, FX swaps, and currency swaps. The trading volume data also come from semi-annual FX surveys of local monetary authorities in New York, London, Tokyo, Toronto, Sydney, Singapore, and Hong Kong. Figure A1 in the Appendix plot the volume shares of G6 currencies and the remainder, beginning from the year 2009. Though I suppress time subscripts here, these weights can vary across time when used for aggregating CIP deviations at different time (at time t, $\bar{b}_t = \sum_{i=1}^{n} w_{it}|b_{it}|$).

Figure 1.4 demonstrates implications from Assumption 7 in detail. The function $S_0(x)$

defines a support function

$$S_{\mathcal{C}_0^{2D}}(x,\, y) = xS_0(y/x)$$

for a set $\mathcal{C}_0^{2D}$ in $\mathbb{R}^2$. Combing this two-dimensional set with a vector of weights further generates the financial constraint $\mathcal{C}_0$ in $\mathbb{R}^{n+1}$. The top plot in Figure 1.4 illustrates a three-dimensional set $\mathcal{C}_0$, the configuration of which satisfies Assumption 7. The three plots at the bottom show its intersections with three planes $\{(\pi_0,\, \pi_1,\, \pi_2) : \pi_i = 0\}$, $(i = 0,\, 1,\, 2)$. In the π_0-π_1 (sub)space, the intersection is indeed a set defined as $\{(\pi_0,\, w_1\pi_1) : (\pi_0,\, \pi_1) \in \mathcal{C}_0^{2D}\}$. The same rule holds for the intersection in the π_0-π_2 (sub)space. The weighted average term in Assumption 7 is reflected directly in the π_1-π_2 intersection (see the diamond shape in the plot at the bottom right corner).

Combining Assumption 7 with Lemma 1,

$$S_{\mathcal{C}_t}(r_t,\, \boldsymbol{b}_t) = S_{\mathcal{C}_0}(r_t,\, \alpha_t\boldsymbol{b}_t) = r_t S_{\mathcal{C}_0}\left(1,\, \frac{\alpha_t\boldsymbol{b}_t}{r_t}\right) = r_t S_0\left(\alpha_t\frac{\bar{b}_t}{r_t}\right). \tag{1.17}$$

Substituting this result into equation (1.15), and dividing both sides by $r_t\mathrm{d}t$, we have

$$\frac{1}{\mathrm{d}t}\frac{\mathrm{d}k_t/k_t}{r_t} = \left[S_0\left(\alpha_t\frac{\bar{b}_t}{r_t}\right) - \frac{\rho}{r_t} + \frac{\lambda_t^2}{r_t}\right] + \frac{\lambda_t}{r_t}\frac{\mathrm{d}z_t}{\mathrm{d}t}. \tag{1.18}$$

I now introduce two additional parameterization schemes for objects in equation (1.18): the Shape-ratios of arbitrageurs' risky project λ_t and the dynamics of financial constraints α_t.

Item 2: parameterizing Shape-ratios. I parametrize the whole term $(\lambda_t^2/r_t - \rho/r_t)$ as the linear combination of variables that may predict the arbitrageurs' capital return $\mathrm{d}k_t/k_t$ other than the CIP deviations, that is, $\lambda_t^2/r_t - \rho/r_t = \boldsymbol{\phi}^\top\boldsymbol{v}_t$. The vector $\boldsymbol{v}_t$ include variables such as earnings yields for the 49 dealer banks, and the VIX index, which are potential predictors of the capital return $\mathrm{d}k_t/k_t$. The reciprocal of r_t is also included as suggested by theory.

Item 3: parameterizing the dynamics of financial constraints. I parameterize the positive process α_t as $\exp(\boldsymbol{\delta}^\top\boldsymbol{u}_t)$ where $\boldsymbol{u}_t$ is a vector containing variables that may drive the time-series variation in financial constraints. It includes the dollar index, quarterly lagged volatilities of average CIP deviations, changes in dealer banks' CDS, the TED spread (three-month dollar LIBOR rates minus the three-month Treasury bill rates), the implied volatility of euro, the VIX index, and the dollar convenience yield (the three-month RefCorp bond yield minus the three-month treasury yield).[32]

[32] The dollar index captures risk-bearing capacity of global banks as argued by Avdjiev, Du, Koch, and Shin (2019). Past volatilities of CIP deviations may affect VaR calculations involving FX derivatives positions. Bank CDS rates determine funding value adjustments as illustrate by Andersen, Duffie, and Song (2019). The TED spread measures credit risk in the banking sector. I add the implied volatility of euro and the VIX index as additional controls for risk appetite in currency markets and, more broadly, global financial markets. The measurement of dollar convenience yields follows Longstaff (2004); Augustin et al. (2020) find that swap dealers' effective funding rates are related to convenience yields.

Under these parameterization schemes, we can now write equation (1.18) as

$$\frac{1}{\tau}\left(\frac{\text{return}_{t+\tau}}{r_t}\right) = \left[S_0\left(\exp(\boldsymbol{\delta}^\top \boldsymbol{u}_t)\frac{\overline{b}_t}{r_t}\right) + \boldsymbol{\phi}^\top \boldsymbol{v}_t\right] + \varepsilon_{t+\tau}, \tag{1.19}$$

after replacing dt by τ, in which the error term $\varepsilon_{t+\tau}$ are future shocks to arbitrageurs' capital. Estimating equation (1.19) yields the function $S_0(\cdot)$ as well as vectors $\boldsymbol{\delta}$ and $\boldsymbol{\phi}$. According to equation (1.17), knowledge regarding the function S_0 and the vector $\boldsymbol{\delta}$ (which translates into α_t) fully reveals $S_{\mathcal{C}_t}$ (the arbitrage profit function determined by time-varying financial constraints). In equilibrium, arbitrage positions can be calculated as

$$\pi_{it} = \frac{\partial S_{\mathcal{C}_t}(r_t,\, \boldsymbol{b}_t)}{\partial b_{it}} = \frac{r_t \partial S_0(\alpha_t \overline{b}_t / r_t)}{\partial b_{it}} = \alpha_t w_{it} \text{sgn}(b_{it}) S_0'\left(\frac{\alpha_t \overline{b}_t}{r_t}\right), \quad \alpha_t = \exp(\boldsymbol{\delta}^\top \boldsymbol{u}_t). \tag{1.20}$$

Now shifting attention to equation (1.16) and writing it in an element-wise manner, we have

$$\pi_{it} k_t = \gamma_{0,it} - \gamma b_{it}.$$

The left-hand side of this equation becomes observable if we know S_0 and $\boldsymbol{\delta}$, according to equation (1.20). I now introduce parameterization for hedging demand intercepts $\gamma_{0,it}$ on the right-hand side.

Item 4: parameterizing hedging demands. Hedgers' demands are further specified as follows (optimization foundation for the hedging demands in Appendix 1.7.3 helps motivate the specification):

$$\gamma_{0,it} = \boldsymbol{\beta}_i^\top \boldsymbol{x}_{it} + \ell_{it},$$

where $\boldsymbol{x}_{it}$ is a vector of observable hedging demand drivers including bilateral net exports, net foreign direct investment flows, net security purchases (long-term bonds and equities), changes in net cross-border bank claims, and interest rate differentials (all calculated as domestic, the US, minus foreign, country i); an intercept term of constant one is also included in $\boldsymbol{x}_{it}$; $\ell_{it} \sim \mathcal{N}(0,\, \sigma_\ell^2)$ captures unobservable components of hedging demands (or, in extension, liquidity-driven demands for forward dollars which I do not model explicitly in the micro-foundation section of Appendix 1.7.3). This specification implies that

$$\begin{aligned} \pi_{it} k_t &= \boldsymbol{\beta}_i^\top \boldsymbol{x}_{it} - \gamma b_{it} + \ell_{it} \\ &= \overline{\boldsymbol{\beta}}^\top \boldsymbol{x}_{it} + \sum_{j=1}^{n-1} \boldsymbol{\eta}_j^\top \left(I[j=i] \times \boldsymbol{x}_{jt}\right) - \gamma b_{it} + \ell_{it}. \end{aligned} \tag{1.21}$$

The second equation in (1.21) adopts the transformation $\boldsymbol{\beta}_i = \overline{\boldsymbol{\beta}} + \boldsymbol{\eta}_i$ where $\sum_{i=1}^n \boldsymbol{\eta}_i = 0$. As a result, the vector $\overline{\boldsymbol{\beta}}$ is the cross-sectional average of $\boldsymbol{\beta}_i$, which accounts for mean responses of hedging demands to observables in the model for all currencies in the sample. Under this specification, estimating $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n$ is equivalent to estimating $\overline{\boldsymbol{\beta}}$, $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_{n-1}$.

Identification and estimation

Under the current parameterization scheme, model estimation takes two steps. First, I estimate equation (1.19) to find the triplet $\{S_0(\cdot), \boldsymbol{\delta}, \boldsymbol{\phi})\}$; these estimates allow me to calculate arbitrage positions π_{it} according to equation (1.20). Second, knowing π_{it} as well as k_t, I estimate vectors $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n$ and the "semi-elasticity" parameter γ from equation (1.21).[33]

Step 1: estimating financial constraints. The first step relies on the following identification assumption

$$\mathbb{E}\left[\varepsilon_{t+\tau} \mid \boldsymbol{b}_t,\, \boldsymbol{u}_t,\, \boldsymbol{v}_t\right] = 0,$$

in equation (1.19), which holds according to the theory. Specifically, this condition argues that the current CIP deviations ($\boldsymbol{b}_t$), dynamics of financial constraints (determined by $\boldsymbol{u}_t$), as well as drivers of arbitrageurs' expected capital returns ($\boldsymbol{v}_t$) do not affect shocks to arbitrageurs' *future* realized capital returns. This argument *does not* preclude the possibility that current (or even past) shocks to arbitrageurs' capital affect these variables. Arbitrageurs in the model do respond to contemporaneous shocks and adjust their arbitrage positions. Moreover, since they are global dealer banks, these shocks can even have "real" impacts through trade finance (Xu, 2020) and cross-border capital flows (Amiti, McGuire, and Weinstein, 2019), thus affecting the hedging demands. This two-sided influence complicates equilibrium CIP deviations and can induce co-movement between the arbitrage yields and contemporaneous shocks to arbitrageurs' capital. However, these relationships should not apply to future unexpected shocks, as the identification condition commands.

The identifying condition can be violated if, for example, additional unobservable risk premium drivers exist. To be more concrete, this corresponds to the case that the $\boldsymbol{\phi}^\top \boldsymbol{u}_t$ term in equation (1.19) should in fact be $(\boldsymbol{\phi}^\top \boldsymbol{u}_t + \ell_t^u)$ where ℓ_t^u is the unobservable component. This term must be correlated with $\boldsymbol{b}_t$ through its impact on k_t (recall that $\boldsymbol{b}_t$ must solve the equilibrium condition (1.16) at time t). Given the fact that powerful return predictors are usually difficult to find beyond valuation ratios and volatility measures (which I have included in the vector $\boldsymbol{v}_t$), this concern might not be of primary importance.

The current framework is in fact flexible enough to incorporate additional controls that drive arbitrageurs' equity returns. Future research could help improve the current estimation when new dealer bank equity return predictors are identified, which will be added into the vector $\boldsymbol{v}_t$.

If a meaningful unobserved risk premium driver does exist, my estimation may exaggerate the response of arbitrageurs' capital returns to arbitrage yields. This is because higher ℓ_t^u is equivalent to higher expected capital returns, and is associated with lower current capital valuations. According to result [ii] of Proposition 2, this leads to higher (absolute) CIP deviations. In other words, $\text{cov}(\ell_t^u, \bar{b}_t) > 0$. As a result, the estimated

[33] I call the parameter γ "semi-elasticity" because b is related to logarithms of forward prices F according to equation 1.1, the initial definition of CIP deviations. In addition, hedging demands in the model can be interpreted as forward dollar demands as discussed in Section 1.2. Of course, $(-\gamma)$ should be the proper semi-elasticity.

response of returns to CIP deviations will subsume both the direct effects from arbitrage profits and (positively) confounded effects through ℓ_t^u. The escalated level of arbitrage profit functions map to a more lenient view of financial constraints: all else equal, relaxed financial constraints allows arbitrageurs to build more aggressive arbitrage positions and reap larger arbitrage profits. Under such scenario, I interpret my estimates of the financial constraints as conservative ones (i.e., supersets) that must contain the truth at each time period.

With the identifying condition $\mathbb{E}\left[\varepsilon_{t+\tau} \mid \boldsymbol{b}_t, \boldsymbol{u}_t, \boldsymbol{v}_t\right] = 0$, I estimate equation (1.19) using semi-parametric nonlinear least squares. The algorithm for estimating this equation is described in Appendix 1.7.2.

Table 1.8 reports estimation results from the first step. According to Table 1.8, increases in dollar index, lagged currency swap volatility, and implied volatility of euro are significantly associated with tightening financial constraints. I also consider equal weighting ($\bar{b}_t = \sum_{i=1}^{n} |b_{it}|/n$) for robustness and results remain largely unchanged.

I plot in Figure 1.6 the times series of α_t (adjusted by sample mean) based on the estimates of $\boldsymbol{\delta}$ in Table 1.8 under the volume-weighting scheme. Smaller α_t indicates tighter financial constraints. According to Figure 1.6, arbitrageurs appear to face toughest constraints during the year 2015-2016. The sharp tightening begins from the middle of 2014. Perhaps not coincidentally, the Volcker rule regulating proprietary trading becomes in effect during the second quarter of 2014. In addition, the supplementary leverage ratio requirement is finalized during the third quarter of this year. The estimated dynamics of financial constraints seems to delineate FX Dealer banks' responses to these regulatory reforms. Another interesting period is the first quarter of 2017, witnessing extremely tight constraints. It is during the same quarter that the liquidity coverage ratio (LCR) requirement reaches its full effects.

Step 2: estimating hedging demands. In the second step, I estimate hedging demand parameters. I calculate π_{it} using equation (1.20) based on estimates of $S_0(\cdot)$ and $\boldsymbol{\delta}$ from the first step. Now the goal is to estimate parameters $\boldsymbol{\beta}_i$ and γ in equation (1.21). Since the left-hand side of this equation are now observable, a standard panel-data linear regression can generate estimators for $(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n, \gamma)$. The main issue with this estimation is that unobservable hedging demands ℓ_{it} will affect the equilibrium deviations b_{it}, thus contaminating the ordinary least-square estimator of γ.

To address this issue, I propose an instrumental variable (IV) for b_{it} based on the following assumption:

$$\mathbb{E}\left[\ell_{it} \mid \boldsymbol{x}_{i't}\right] = 0, \quad i' \neq i.$$

This condition states that unobservable hedging demands for a particular currency are not related to observable hedging demand drivers of *other* currencies. In other words, bilateral trade and portfolio flows between the UK and US, which may drive hedging demands for pounds, should not affect hedging demands for yen. If this condition is satisfied, we can instrument b_{it} using estimators $\widehat{b}_{it}$ from the following (first-stage) regression:

$$b_{it} = \boldsymbol{\psi}^{\top} \boldsymbol{z}_{it} + \overline{\boldsymbol{\phi}}^{\top} \boldsymbol{x}_{it} + \sum_{j=1}^{n-1} \boldsymbol{\xi}_j^{\top} \left(I[j = i] \times \boldsymbol{x}_{jt} \right) + e_{it},$$

because the right-hand side instrumental vector $\boldsymbol{z}_{it} = \sum_{i' \neq i} w_{i't} \boldsymbol{x}_{i't}^{(-\iota)}$ is not associated with ℓ_{it}. The weights are calculated from the volume of FX derivatives, which also appear in Assumption 7 and equation (1.20). The superscript "$(-\iota)$" for $\boldsymbol{x}$ means that the constant one for intercepts is excluded from this vector.

This instrument should not be a weak one in theory ($\boldsymbol{\psi} \neq \mathbf{0}$), as it directly affects levels of CIP deviations for other currencies (i.e., the vector $\boldsymbol{b}_{-i}$). Changes to $\boldsymbol{b}_{-i}$ will affect arbitrageurs' equilibrium arbitrage positions not only for the involved currencies ($\boldsymbol{\pi}_{-i}$), but also for currency i (π_i). If we conceptualize arbitrageurs as "suppliers" of arbitrage services, this instrument is effectively a supply shifter in the tradition of Berry, Levinsohn, and Pakes (1995). The volume-weights reflect the belief that demands for derivative contacts on dominant currencies should have larger impacts on arbitrageurs' optimal positions, transmitting more pronounced "supply" shocks.

The exclusion restriction of the proposed instruments can be invalid when there are common shocks to both observable hedging demand drivers $\boldsymbol{x}_{1t}, \ldots, \boldsymbol{x}_{nt}$, and latent hedging demands $\ell_{1t}, \ldots, \ell_{nt}$, thus relating $\boldsymbol{x}_{i't}$ to ℓ_{it}. A necessary outcome of this scenario is that $\boldsymbol{x}_{1t}, \ldots, \boldsymbol{x}_{nt}$ present a strong factor structure. In the data, leading principle components of variables in these vectors never explain more than 40% total variation (40% for bilateral net exports as the highest, 23% for bilateral changes in net bank claims as the lowest). This exploratory analysis provides suggestive evidence favoring the identification condition.

If the identifying condition is indeed violated, then my estimate of the γ parameter is likely to be downward biased. Adversarial shocks under tumultuous market conditions suppressing *all* bilateral trades and portfolio investments (the observables) tend to be associated with dollar shortages, boosting demands for spot dollars (via synthetic dollar funding) and dampening the need for forward dollars. If we interpret the unobservables absorbed by ℓ as forward dollar demands due to liquidity needs, the instrument constructed using $\boldsymbol{x}_{-i}$ will be positively correlated with ℓ_i in equation (1.21): they both drop in bad times. Estimates of $-\gamma$ (a negative object in theory) will be inflated by the instrument, which is equivalent to downward biased γ estimates.

Table 1.9 reports the second-step demand estimation results. In these estimations, I normalize arbitrageurs' capital to one at the beginning of the sample (January 2009). The key parameter of interest is γ. The OLS estimation of γ is negative, suggesting that this simple approach is mired by unobservable demand drivers. IV estimations yield γ estimates of around 1.4. Weak IV test statistics for the first-stage regression exceed theory cutoffs calculated following Stock and Yogo (2005).

Interpreting the number $\gamma = 1.4$ relies on estimates of $\overline{\boldsymbol{\beta}}$, the components of which are significantly positive for net purchases of long-term securities (only long-term bonds, not equities) and net exports. This finding itself is intuitive. Higher (US) net exports indicate that US exporters expect more foreign-currency receivables. To hedge these cash flows against currency risk, they sell foreign currencies forward in exchange for dollar. As a result, increased net exports indicate higher forward dollar demands. Similar reasoning apply to net foreign asset purchases which generate foreign-currency denominated cash flows (and capital gains) in the future. The coefficient is around five for net long-term bond purchases, which is 3.6 times of γ. This suggests that one basis point increase in the CIP deviations is equivalent to $1/3.6 \approx 0.28$ billion decrease in this variable in terms of impacts on hedging demands. Similarly, the coefficient for net exports is around ten ($\approx 7 \times \gamma$). Thus, one basis point increase in the CIP deviations tends to have the same impact on hedging demands as $1/7 \approx 0.14$ billion decrease in net exports.[34]

1.5 Quantitative analysis

1.5.1 Model-implied CIP deviations

With estimates of the function $S_0(\cdot)$, α_t (determined by the vector $\boldsymbol{\delta}$), $\boldsymbol{\gamma}_{0,t} = [\gamma_{01,t}, \ldots, \gamma_{0n,t}]^\top$ (each element determined by $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n$ respectively), and the parameter γ, the equilibrium condition (1.16) becomes a pricing system: at time-t, CIP deviations $\boldsymbol{b}$ solves

$$\frac{r_t \partial S_0 \left(\alpha_t \boldsymbol{w}_t^\top \boldsymbol{b}\right)}{\partial \boldsymbol{b}} k_t = \boldsymbol{\gamma}_{0,t} - \gamma \boldsymbol{b}, \tag{1.22}$$

where $\boldsymbol{w}_t$ contain weights calculated from FX derivatives trading volumes, k_t is measured by market equity of the 49 FX dealer banks. I solve for model-implied $\boldsymbol{b}$ each month from this equation and compare it with data.

Denote by $\hat{b}$ the model-implied CIP deviations for a specific currency and by b the true data (one-year CIP deviations measured using currency swap rates). Figure 1.7 contrast CIP deviations solved from equation (1.22) against data. Overall, the model-implied CIP deviations track the data well. Panel (A) of Table 1.10 reports the means and standard deviations of b and $\hat{b}$ as well as their correlations. Sample periods are January 2009 to December 2019. Overall, moments of model-implied CIP deviations closely track ones from the data for G6 currencies.

If we decompose the variance of observed data b as

$$\sigma^2[b] = \mathrm{cov}[b,\, b] = \mathrm{cov}[b,\, \hat{b}] + \mathrm{cov}[b,\, b - \hat{b}],$$

the ratio $\mathrm{cov}[b,\, \hat{b}]/\sigma^2[b]$ measures the fraction of total variance in the data that the model accounts for. This quantity is equivalent to the slope coefficient of the following regression

$$\hat{b}_t = \beta_0 + \beta b_t + \varepsilon_t,$$

[34] Since I do not rule out correlations between $\boldsymbol{x}_{it}$ and ℓ_{it} in equation (1.21), $\boldsymbol{\beta}$ estimates cannot be treated as causally identified. These results should be interpreted with caution. Calculation here may be illustrative, but can at least help better understand the model.

The last two columns of Table 1.10 report estimates of β in this regression and their standard errors. For euro, yen, pounds, and Canadian dollars, model-implied CIP deviations account for at least 57 percent of total variation in the data. For Australian dollar and Swiss Franc, the model-implied CIP deviations explain over 30 percent of observed variation. The relative poor performance for CHF is mainly due to low variation in the model-implied quantities. The correlation between b and $\hat{b}$ is 0.59 for CHF but the variance of $\hat{b}$ is 46 percent lower. Overall, the model explains around 57 percent of variation in one-year CIP deviations of G6 currencies.

Out-of-sample analysis: sample splitting. I repeat the two-step model estimation exercise using the 2009-2015 subsample, and treat the 2016-2019 subsample as testing data. Adopting a common "trick" facing the bias-variance trade-off (when performing out-of-sample prediction tasks), I choose a more parsimonious hedging demand specification, which only includes net exports and net bond purchases. Using parameters estimated from the first subsample, I solve for CIP deviations according to equation 1.22. Figure 1.8 shows the out-of-sample prediction results. Levels of predictions align well with the data in the testing sample.

Out-of-sample analysis: additional currencies. I further check model performance by applying it to four currencies *not* used in the first-step model estimation: the Swedish krona (SEK), Norwegian krone (NOK), New Zealand dollar (NZD) and Hong Kong dollar (HKD). As an approximation, when solving for equilibrium CIP deviations from equation (1.22) for these new currency pairs, I ignore all off-diagonal elements in the partial differentiation on the left-hand side. I use $\overline{\boldsymbol{\beta}}$ estimates from Table 1.9 to computer their hedging demands (instead of finding $\boldsymbol{\beta}_i$ for each currency). I compare model outcomes with data in Figure 1.9. Although no information regarding these currencies is used for estimation, CIP deviations solved from the model still align well with data. Panel (B) of Table 1.10 compares moments for the model-implied ones with data and repeat the regression analysis above. The model tracks data moments well. On average, model-implied quantities explain over 30 percent variation in the data.

Restoring CIP deviations back to their pre-crisis levels. With equation (1.22), we can investigate counterfactual CIP deviations when arbitrageurs are facing tighter or loser financial constraints. I conduct this exercise via replacing α_t in equation (1.22) by $(c_\alpha \alpha_t)$ and resolve for equilibrium CIP deviations. A larger constant c_α indicates loser financial constraints. Table 1.11 reports time-series average of counterfactual CIP deviations as well as their standard deviations for different c_α. One particularly interesting observation from Table 1.11 is that loosening the financial constraints by allowing for 2.5 times larger arbitrage legs (recalling the illustration in Figure 1.3) can restore the post-crisis CIP deviations back to their pre-crisis levels (of around five basis points).

1.5.2 Shapley-value decomposition of the model-implied CIP deviations

To determine the relative contribution of (the dynamics of) financial constraints (α_t), hedging demands $\boldsymbol{\gamma}_{0,t}$, and arbitrageurs' capital (k_t) to time-series variation in CIP deviations, I adapt a Shapley decomposition (see Shorrocks (2013) for its application in linear models) to the equilibrium pricing function. For each of the three forces, Shapley decomposition determines its marginal contribution to total variation in model-implied CIP deviations. This decomposition scheme is especially useful as the three economic forces interact with each other to determine equilibrium CIP deviations nonlinearly through equation (1.22). Conceptually, the three economic forces are teammates who cooperate on a task – producing variation in $\boldsymbol{b}$. The Shapley decomposition calculates their "wages" for finishing the task in an efficient, fair, and easy-to-interpret manner.

I begin by adapting the Shapley decomposition to my equilibrium model. Equation (1.22) defines an implicit function $\boldsymbol{b} = L(\alpha,\, k,\, \boldsymbol{\gamma}_0)$ that maps the three variables to the equilibrium CIP deviations. For variable $v \in \{\alpha,\, k,\, \boldsymbol{\gamma}_0\}$, I compute

$$I_v = \sum_{V \subset \{\alpha, k, \boldsymbol{\gamma}_0\} \backslash \{v\}} \frac{|V|}{6} \left\{ \sigma^2[L(V,\, v)] - \sigma^2[L(V,\, \overline{v})] \right\},$$

where $\sigma^2[L(V,\, v)]$ denotes the variance of counterfactual CIP deviations calculated from the implicit function, holding $\{\alpha_t,\, k_t,\, \boldsymbol{\gamma}_{0,t}\} \backslash \{V, v\}$ constant (as its sample average) while allowing both v and variables in V to vary; $\sigma^2[L(V,\, \overline{v})]$ denotes the variance calculated similarly holding both $\{\alpha_t,\, k_t,\, \boldsymbol{\gamma}_{0,t}\} \backslash \{V, v\}$ and the variable of interest v constant (only variables in V are allow to change across time). For each v, the identity sums across all configurations excluding itself. Under this decomposition scheme, the variance of model-implied CIP deviations satisfies

$$\sigma^2[\hat{b}] = I_\alpha + I_k + I_{\boldsymbol{\gamma}_0}.$$

Of note, for the vector $\boldsymbol{\gamma}_0$, when computing counterfactual CIP deviations of currency i, only its ith element is held constant when needed.

Table 1.12 reports the fraction of variation in model-implied CIP deviations ($I_v/\sigma^2[\hat{b}]$) that can be attributed to each of the three drivers. On average, financial constraints are responsible for 46.4 percent of variation in model-implied CIP deviations. Hedging demands and arbitragers' capital explain the other 38.0 and 15.6 percents.

For variation in the data, consider the following equation

$$\sigma^2[b] = \sigma^2[\hat{b}] + \text{cov}[b - \hat{b},\, \hat{b}] + \text{cov}[b,\, b - \hat{b}].$$

Since $\sigma[\hat{b}]/\sigma[b] \approx 1$ for most currencies according to Table 1.10, ratios above also approximate the fraction of variation in the data that can be attributed to each of the three economic forces.

These impacts differentiate across currencies. For euro and yen, the dynamics of financial constraints plays a crucial role in driving CIP deviations, accounting for 60-70 percent CIP deviations in the model. For commodity currencies including Canadian dollars and Australian dollars, hedging demands account for approximately 70 and 40 percent variation respectively. Arbitrageurs' capital dynamics exerts substantial impacts

(30 percent) only on pound-dollar CIP deviations.

To further investigate the dynamics of variance attribution, I perform the Shapley decomposition on a four-year rolling-window basis. Figure 1.10 presents the results. The most striking pattern from plots in Figure 1.10 is that arbitrageurs' capital can stabilize the CIP basis when financial constraints or hedging demands exert disproportionately large impacts. That is, under the counterfactual settings of holding arbitrageurs' capital constant, fluctuations in CIP deviations can increase. For example, in 2013-2014, Canadian dollar basis is overwhelmingly driven by hedging demands. If arbitrage capital remains constant, (counterfactual) variation in Canadian dollar CIP deviations would double.

One limitation to the Shapley decomposition due to the fact that arbitrageurs' capital is endogenously determined according to equation (1.15). Thus counterfactual CIP deviations lead to alternative dynamics of k_t, the variation of which further generates feedbacks to the equilibrium basis. I do not account for this interaction in my current decomposition exercise. Failing to do so may exaggerate influences of arbitrageurs' capital. A potential channel is that higher CIP deviations due to relatively low levels of (contemporaneous) k_t help replenish arbitrageurs' capital in the future, enabling arbitrageurs to better absorb future financial and hedging demand shocks.

1.5.3 The shape of financial constraints

The (basline) shape of financial constraints, namely $\mathcal{C}_0$, can be recovered from estimates of the function $S_0(x)$ as follows

$$\bigcap_{0<\theta<\pi/2} \{(x,\, y) : x + y\tan\theta \leq S_0(\tan\theta)\}\,.$$

Intuitively, $\mathcal{C}_0$ is a set containing all points "inside" the envelope of half planes $x+y\tan\theta \leq S_0(\tan\theta)$ for varying θ. Layering the half planes will unveil the shape of of financial constraints, a procedure similar to tomography: the shape of an object can be reconstructed from its shadows when light beams shine on it from many different angles. angles.[35]

One particularly interesting exercise would be figuring out how the baseline shapes of financial constraints morph across time. This shape-shifting variation can capture additional dynamics of financial constraints beyond the series α_t. To make progress, I reestimate model (1.19) for subsample periods of 2009-2013 and 2015-2019 and compare the recovered shape estimates.

Figure 1.11 presents the estimated S_0 functions as well as the recovered sets $\mathcal{C}_0$. The top and bottom panels correspond to results for 2009-2013 and 2014-2019 respectively. White areas enclosed by blue half planes are the sets $\mathcal{C}_0$. The x-axis corresponds to π_0 (fractions of equity capital deployable to support routine business) and y-axis is for π (arbitrage

[35] Rigorously speaking, the set "$\mathcal{C}_0$" recovered from S_0 using this procedure is the two-dimensional generator $\mathcal{C}_0^{2D}$ under Assumption 7. Readers may revisit Figure 1.4 for illustration. I will use the two notations interchangeably here. I also restrict the range of θ such that the recovered set is in the first quadrant. According to Assumption 7, $\mathcal{C}_0$ is symmetric to the horizontal axis, thus its shape in the fourth quadrant is trivial.

positions). Shapes of these sets contain important information regarding arbitrageurs' internal capital allocation decisions. Let us shift our focus to the bottom right corner of $\mathcal{C}_0$ in Panel (A). The pattern suggests that, in 2009-2013, arbitrageurs can build arbitrage positions that are almost three times of their equity capital without the need to curtail other investment positions. If arbitrage positions are four times larger than the equity capital, their routine investments will shrink about 10 percent. Going beyond this level, increased π leads to sharp decreases in π_0 and the response is almost linear, indicating pronounced balance sheet costs.

Panel (B) of Figure 1.11 suggests that during 2014-2019, the balance sheet space becomes more costly. Arbitrage positions quickly translate into downsized routine investments. For an arbitrage position that is five times of arbitrageurs' equity capital, the size of normal business position (π_0) is reduced by more than one half (compared with 10 percent during 2009-2013). A hard leverage cap of around seven emerges in this period. This outcome appears to be consistent with the fact that the supplementary leverage ratio (SLR) requirement was finalized in the third quarter of 2014.

1.6 Conclusion

Most existing limits-to-arbitrage models lack the potential to be mapped to data directly, thus the valuable insights they offer are hard to quantify. This paper attempts to partially bridge the gap by developing a quantitative model of limited arbitrage with a special focus on deviations from covered interest rate parity (CIP) conditions.

The model and its estimation methods can be a useful framework for understanding other "anomalous" pricing phenomena in today's financial markets, such as the IOER-RRP arbitrage (interest rates on excess reserves being greater than the over overnight reverse repo rates), the CDS-bond basis (the difference between credit spreads and credit default swap rates of the same bond), and negative swap spreads (thirty-year Treasury yields exceeding the corresponding swap rates). Common research questions arise in response to these phenomena. For example, who are the main arbitrageurs in these markets? What types of constraints they face (that are binding)? What are the main drivers of demands for the involved derivatives contracts? What explains time-series variation of the underlying arbitrage opportunities? The current paper illustrates how to use the framework to answer such questions.

My main contribution is to combine potential drivers of price dislocations such as hedging demands, financial constraints, and arbitrageurs' capital in a parsimonious equilibrium model. The model is flexible enough to incorporate existing knowledge about these economic forces and estimate their influences on asset prices (and their deviations from frictionless benchmarks). The key innovation is a general specification of financial constraints, and the theoretical and econometric tools developed for unveiling their shapes and capturing their dynamics.

Table 1.1: Summary statistics of one-year CIP deviation measures for G6 currencies against the dollar.

	currency swap rates					forward-OIS bases				
	mean	s.d.	median	min	max	mean	s.d.	median	min	max
EUR	−28.59	16.22	−26.60	−107.75	−17.00	−37.65	19.29	−35.97	−81.87	−23.24
JPY	−34.43	14.91	−30.75	−82.38	−23.00	−53.06	22.50	−53.28	−109.38	−33.54
GBP	−9.93	11.43	−7.62	−77.07	−1.88	−13.49	12.93	−10.05	−55.89	−3.64
CAD	−10.56	10.98	−11.50	−32.88	−3.60	−8.73	12.20	−4.57	−74.27	−0.59
AUD	14.35	6.67	13.50	−4.12	18.90	13.51	14.43	13.93	−53.19	21.04
CHF	−26.52	13.10	−24.75	−80.75	−16.00	−54.21	23.53	−50.18	−102.76	−34.72

Table 1.2: Predictive regressions: book equity and market equity returns of global dealer banks on one-year basis swap rates

This table presents results from the following panel regressions

$$\frac{1}{\tau}\text{return}_{i,t+\tau} = \alpha_i + \beta \bar{b}_t + \varepsilon_{i,t+\tau},$$

for quarterly observations. The dependent variables are one-quarter-ahead net returns on the book equity (BE), market equity (ME), or (artificially defined "returns" on) market-to-book ratio (MB) of 49 dealer banks surveyed by FX committees of New York, London, Tokyo, Toronto, Sydney, Singapore and Hong Kong. Variables are collected for their holding companies. All returns are annualized (divided by $\tau = 0.25$) net ones in percentage points. The subscript i represents banks and t denotes quarters. The independent variable $\bar{b}_t$ is the cross-sectional average of absolute one-year basis swap rates or forward-OIS bases for EUR, JPY, GBP, AUD, CAD, and CHF against the dollar. Sample periods begin from January 2009 and end at December 2019. Specifications with and without ($\alpha_i = \alpha$ for all $i = 1, \dots, 49$) bank fixed effects are both included. Sample periods begin from March 2009 and end at December 2019. Numbers in parentheses are Driscoll-Kraay standard errors robust to general forms of serial correlations and cross-sectional correlations among banks (Driscoll and Kraay, 1998).

	$(\text{BE}_{i,t+\tau}/\text{BE}_{i,t} - 1)\%$		$(\text{MB}_{i,t+\tau}/\text{MB}_{i,t} - 1)\%$		$(\text{ME}_{i,t+\tau}/\text{ME}_{i,t} - 1)\%$	
Panel A: CIP deviations measured by currency swap rates						
$\bar{b}$ (b.p.)	0.244	0.227	0.627	0.639	0.875	0.871
	(0.114)	(0.108)	(0.426)	(0.420)	(0.397)	(0.395)
const.	0.62		−14.59		−14.53	
	(2.88)		(10.22)		(9.66)	
Bank f.e.	✗	✓	✗	✓	✗	✓
N obs.	1713	1713	1713	1713	1713	1713
adj.-R^2 (%)	1.0	4.0	1.5	1.4	2.9	2.4
Panel B: CIP deviations measured by forward-OIS implied bases						
$\bar{b}$ (b.p.)	0.168	0.158	0.604	0.617	0.780	0.785
	(0.082)	(0.078)	(0.356)	(0.352)	(0.341)	(0.338)
const.	2.01		−15.02		−13.66	
	(2.76)		(9.28)		$(8, 72)$	
Bank f.e.	✗	✓	✗	✓	✗	✓
N obs.	1713	1713	1713	1713	1713	1713
adj.-R^2 (%)	0.7	3.7	1.9	1.8	3.1	2.7

Table 1.3: Predictive regressions: quarterly returns of FX committee surveyed (FXS) dealer banks on one-year basis swap rates and placebo tests

This table presents results from the following time-series regressions:

$$\frac{1}{\tau}\text{return}_{t+\tau} = \beta_0 + \beta\bar{b}_t + \epsilon_{t+\tau},$$

for quarterly observations. The dependent variables are one-quarter-ahead value- or equal-weighted equity returns of 49 dealer banks participating FX surveys (FXS) conducted by local monetary authority at New York, London, Tokyo, Toronto, Sydney, Singapore and Hong Kong. Variables are collected for their holding companies. Additional placebo tests use returns from five ETFs tracking the S&P500 index (SPY), the global financial sector (IXG), the US financial sector (IYF), US broker-dealers and securities exchanges (IAI), and US insurance companies (KIE). All returns are net ones in percentage, as well as annualized (divided by $\tau = 0.25$ as shown in the regression specification). The independent variable $\bar{b}_t$ is the cross-sectional average of absolute one-year one-year basis swap rates or forward-OIS bases for EUR, JPY, GBP, AUD, CAD, and CHF against the dollar. Sample periods begin from January 2009 and end at December 2019. Numbers in parentheses are Newey-West standard errors under automatic bandwidth selection.

ret. (p.p.)	FXS (vw)	FXS (ew)	ETF-SPY (S&P500)	ETF-IXG (Gl. Fin.)	ETF-IYF (US Fin.)	ETF-IAI (US B&D)	ETF-KIE (US Insur.)
Panel A: CIP deviations measured by currency swap rates							
$\bar{b}$ (b.p.)	1.98	1.67	0.61	1.52	1.10	1.43	1.20
	(0.87)	(0.77)	(0.44)	(0.72)	(0.67)	(0.82)	(0.76)
const.	−32.0	−26.6	3.7	−20.4	−7.4	−17.2	−7.1
	(19.7)	(17.5)	(10.4)	(16.7)	(15.7)	(20.6)	(17.7)
N obs.	44	44	44	44	44	44	44
R^2-adj. (%)	6.2	4.5	0.6	4.9	2.5	4.0	2.1
Panel B: CIP deviations measured by forward-OIS implied bases							
$\bar{b}$ (b.p.)	1.50	1.28	0.40	1.16	0.79	0.90	0.78
	(0.64)	(0.66)	(0.26)	(0.54)	(0.45)	(0.53)	(0.49)
const.	−28.4	−23.9	6.9	−17.5	−3.5	−10.4	−0.6
	(17.4)	(17.3)	(8.1)	(15.3)	(13.6)	(18.5)	(15.5)
N obs.	43	43	43	43	43	43	43
R^2-adj. (%)	13.5	10.6	1.6	11.1	5.6	4.7	3.6

Table 1.4: Predictive regressions: quarterly returns of FX committee surveyed (FXS) dealer banks on one-year basis swap rates and placebo tests

This table presents results from the following time-series regressions:

$$\frac{1}{\tau}\text{return}_{t+\tau} = \beta_0 + \beta \bar{b}_t + \epsilon_{t+\tau},$$

for daily and monthly observations. The dependent variables are one-quarter-ahead value- or equal-weighted equity returns of 49 dealer banks surveyed by FX committees of New York, London, Tokyo, Toronto, Sydney, Singapore and Hong Kong. Variables are collected for their holding companies. Additional placebo tests use returns from five ETFs tracking the S&P500 index (SPY), the global financial sector (IXG), the US financial sector (IYF), US broker-dealers and securities exchanges (IAI), and US insurance companies (KIE). For monthly observations, five hedge fund index returns are also included: one global composite index from BarclaysHedge (BCH), four indices from Hedge Fund Research (HFR) tracking global composite, relative value arbitrage, global-macro, and macro-currency strategies. All returns are net ones in percentage, as well as annualized (divided by $\tau = 0.25$ as shown in the regression specification). The independent variable $\bar{b}_t$ is the cross-sectional average of absolute one-year basis swap rates for EUR, JPY, GBP, AUD, CAD, and CHF against the dollar. Sample periods begin from January 2009 and end at December 2019. Numbers in parentheses are Newey-West standard errors under automatic bandwidth selection.

Panel A: daily observations

ret. (p.p.)	FXS (vw)	FXS (ew)	ETF-SPY (S&P500)	ETF-IXG (Gl. Fin.)	ETF-IYF (US Fin.)	ETF-IAI (US B&D)	ETF-KIE (US Insur.)
$\bar{b}$ (b.p.)	2.46	2.25	0.53	1.93	1.27	1.47	1.30
	(0.71)	(0.71)	(0.30)	(0.62)	(0.51)	(0.65)	(0.54)
const.	−39.8	−36.4	4.9	−28.5	−10.5	−15.2	−8.7
	(13.8)	(13.4)	(7.1)	(12.4)	(10.9)	(14.5)	(12.0)
N obs.	2859	2859	2761	2761	2761	2761	2761
R^2-adj. (%)	12.0	10.3	2.3	9.6	5.9	5.6	5.3

Panel B: monthly observations

ret. (p.p.)	FXS (vw)	FXS (ew)	ETF-SPY (S&P500)	ETF-IXG (Gl. Fin.)	ETF-IYF (US Fin.)	ETF-IAI (US B&D)	ETF-KIE (US Insur.)
$\bar{b}$ (b.p.)	2.18	1.97	0.46	1.69	1.03	1.27	1.06
	(0.79)	(0.79)	(0.36)	(0.74)	(0.64)	(0.80)	(0.72)
const.	−34.3	−30.9	6.7	−23.5	−4.8	−10.7	−3.1
	(16.3)	(15.6)	(9.2)	(16.2)	(14.6)	(19.6)	(16.1)
N obs.	132	132	132	132	132	132	132
R^2-adj. (%)	9.7	7.9	1.0	7.1	3.3	3.6	2.8
ret. (p.p.)			BCH (Gl. Com.)	HFR (Gl. Com.)	HFR (Re. Val.)	HFR (Macro)	HFR (Macro. Cur)
$\lvert b \rvert$ (b.p.)			0.27	0.21	0.17	−0.11	0.12
			(0.16)	(0.14)	(0.12)	(0.10)	(0.12)
const.			0.1	0.4	2.6	4.0	−1.5
			(4.0)	(3.5)	(2.9)	(2.6)	(2.5)
N obs.			132	132	132	132	132
R^2-adj. (%)			1.9	1.2	1.1	0.4	1.5

Table 1.5: Predictive regressions: quarterly returns of FX committee surveyed dealer banks on one-year basis swap rates adjusted by controls

This table presents results from the following time-series regressions:

$$\frac{1}{\tau}\text{return}_{t+\tau} = \beta_0 + \beta\bar{b}_t + \phi \cdot \text{control}_t + \epsilon_{t+\tau}$$

for daily and monthly observations. The dependent variable is the one-quarter-ahead value-weighted equity return of 49 dealer banks surveyed by FX committees of New York, London, Tokyo, Toronto, Sydney, Singapore and Hong Kong. All returns are net ones in percentage, as well as annualized (divided by $\tau = 0.25$ as specified in the regression equation). The independent variable $\bar{b}_t$ is the cross-sectional average of absolute one-year basis swap rates for EUR, JPY, GBP, AUD, CAD, and CHF against the dollar. Control variables include the average smoothed earnings yield (E/P) and dividend yield (D/P) for the 49 dealer banks, the effective Fed fund rate (FFR), and the CBOE volatility index (VIX). Sample periods begin from January 2009 and end at December 2019. Numbers in parentheses are Newey-West standard errors under automatic bandwidth selection.

ret. (p.p.)	Daily observations			Monthly observations		
$\bar{b}$ (b.p.)	2.46	1.18	1.90	2.18	1.00	1.69
	(0.71)	(0.49)	(0.57)	(0.79)	(0.55)	(0.68)
E/P		17.1			16.6	
		(3.5)			(4.4)	
D/P			3.99			3.96
			(2.33)			(2.74)
FFR		5.26	−1.34		4.53	−1.28
		(4.61)	(5.19)		(5.75)	(6.48)
VIX		0.00	2.85		0.47	3.28
		(0.70)	(1.09)		(0.96)	(1.40)
const.	−39.8	−157.7	−92.2	−34.3	−157.7	−96.9
	(13.8)	(23.8)	(23.5)	(16.3)	(29.8)	(32.1)
N obs.	2859	2859	2859	132	132	132
R^2-adj. (%)	12.0	43.1	27.2	9.7	42.4	27.8

Table 1.6: Testing predictions from Proposition 1: linear regressions

This table presents results from the following time-series regressions:

$$\frac{1}{\tau}\left(\frac{\text{return}_{t+\tau}}{r_t}\right) = \beta_0 + \beta_1 X_t + \psi \times \left(\frac{1}{r_t}\right) + \phi \cdot \text{control}_t + \varepsilon_{t+\tau}, \quad X_t = \frac{\bar{b}_t}{r_t} \text{ or } \left(\frac{\bar{b}_t}{r_t}\right)^2$$

using both daily and monthly observations. The notation "return$_{t+\tau}$" denotes one-quarter-ahead value-weighted equity returns of 49 dealer banks surveyed by FX committees of New York, London, Tokyo, Toronto, Sydney, Singapore and Hong Kong. All returns are net ones in percentage, as well as annualized (divided by $\tau = 0.25$ as shown in the regression specification). The cross-sectional average of absolute one-year basis swap rates for EUR, JPY, GBP, AUD, CAD, and CHF against the dollar is denoted by $|b|$. The effective Fed fund rate is denoted by r. The independent variable X is the time-series of either $|b|/r$ or its square $(|b|/r)^2$. Another independent variable of interest is the inverse of the effective Fed fund rate $(1/r_t)$, inspired by the capital accumulation formula in Proposition 1. Control variables include the smoothed earnings yield (E/P) averaged across the 49 dealer banks, and the CBOE volatility index (VIX). Sample periods begin from January 2009 and end at December 2019. Robust regressions use the Huber loss function to accommodate potential outliers. Numbers in parentheses are Newey-West standard errors under automatic bandwidth selection.

	Daily observations				Monthly observations			
ret./r	OLS		Robust		OLS		Robust	
$\bar{b}/r$	138.5		104.1		117.1		84.9	
	(109.7)		(102.3)		(113.9)		(103.0)	
$(\bar{b}/r)^2$		33.4		30.0		24.7		27.2
		(14.0)		(12.5)		(8.9)		(3.63)
$1/r$	−0.27	−0.21	−0.15	−0.11	−0.19	−0.18	−0.12	−0.12
	(0.18)	(0.11)	(0.12)	(0.06)	(0.18)	(0.10)	(0.12)	(0.07)
E/P	79.9	88.2	68.8	76.2	59.4	64.0	69.5	79.6
	(26.7)	(27.3)	(27.0)	(27.8)	(31.3)	(33.2)	(17.6)	(18.2)
VIX	0.91	−0.19	1.45	0.53	7.79	7.51	5.05	2.52
	(7.07)	(6.56)	(4.62)	(4.43)	(8.90)	(8.13)	(6.03)	(4.96)
const.	−625.9	−621.9	−545.8	−559.3	−599.3	−566.4	−599.7	−604.6
	(133.2)	(132.9)	(177.5)	(186.4)	(149.4)	(153.8)	(89.3)	(84.8)
N obs.	2859	2859	2859	2859	132	132	132	132
R^2-adj. (%)	28.7	32.6	–	–	22.5	30.3	–	–

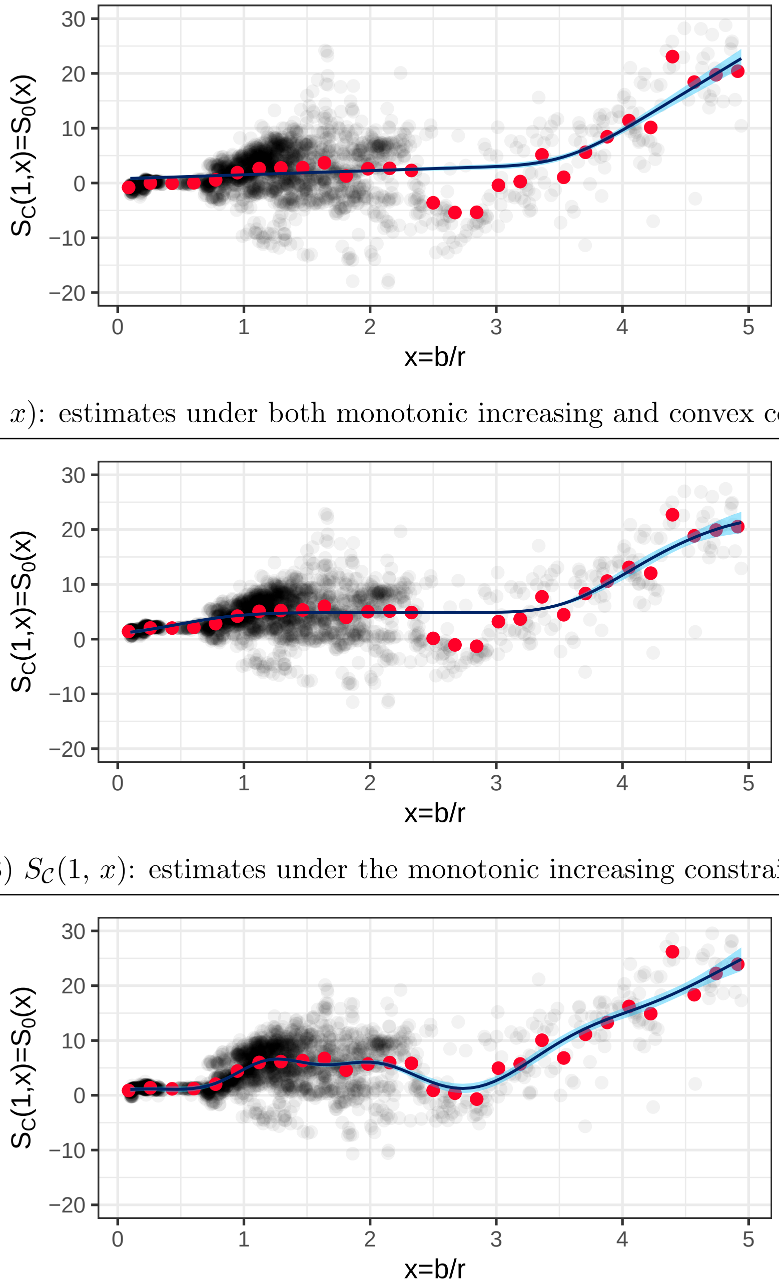
Table 1.7: Testing predictions from Proposition 1: semi-parametric regressions

This table presents results from the following semi-parametric regressions:

$$\frac{1}{\tau}\left(\frac{\text{return}_{t+\tau}}{r_t}\right) = S_0\left(\frac{\bar{b}_t}{r_t}\right) + \psi \times \left(\frac{1}{r_t}\right) + \phi \cdot \text{control}_t + \epsilon_{t+\tau}$$

using both daily and monthly observations. The notation "return$_{t+\tau}$" denotes one-quarter-ahead value-weighted equity returns of 49 dealer banks surveyed by FX committees of New York, London, Tokyo, Toronto, Sydney, Singapore and Hong Kong. All returns are net ones in percentage, as well as annualized (divided by $\tau = 0.25$ as shown in the regression specification). The cross-sectional average of absolute one-year basis swap rates for EUR, JPY, GBP, AUD, CAD, and CHF against the dollar is denoted by $|b|$. The effective Fed fund rate is denoted by r. Out of robustness concerns, only observations with $|b|/r$ falling within the their sample $5\% - 95\%$ IQR are considered. Another independent variable of interest is the inverse of the effective Fed fund rate $(1/r_t)$, inspired by the capital accumulation formula in Proposition 1. Control variables include the smoothed earnings yield (E/P) averaged across the 49 dealer banks, and the CBOE volatility index (VIX). Sample periods begin from January 2009 and end at December 2019. Semi-parametric estimation of the model uses shape-constrained B-splines basis for the functional term s. Numbers in parentheses are standard errors calculated from parametric block bootstrap procedures (that is, residuals of the fitted models are re-sampled). Block sizes are ninety for daily observations and three for monthly observations. The table also presents specification tests of whether the functional term should be included ($S_0 \equiv 0$ or not) by showing the test statistics, their (approximate) theoretical distributions, and test p-values.

ret./r	Daily observations			Monthly observations		
$1/r$	-0.30	-0.39	-0.54	-0.22	-0.22	-0.31
	(0.10)	(0.10)	(0.09)	(0.10)	(0.09)	(0.10)
E/P	15.8	18.0	55.6	2.3	0.66	20.7
	(38.8)	(39.7)	(37.5)	(46.9)	(48.1)	(50.9)
VIX	6.64	4.59	-2.30	9.73	10.47	6.24
	(5.95)	(6.06)	(5.27)	(7.17)	(6.88)	(7.19)
Test $H_0 : S_0 \equiv 0$ v.s. $H_1 : S_0 \not\equiv 0$						
F-stat	189	189	133	13.9	7.37	6.07
Appr. dist.	$F(3, 2538)$	$F(3, 2538)$	$F(8, 2538)$	$F(1, 116)$	$F(2, 116)$	$F(4, 116)$
p-value	$< 10^{-6}$	$< 10^{-6}$	$< 10^{-6}$	3×10^{-4}	9×10^{-4}	3×10^{-4}
Shape constraints for $f(\cdot)$:						
increasing	✓	✓	✗	✓	✓	✗
convex	✓	✗	✗	✓	✗	✗
N obs.	2541	2541	2541	119	119	119
R^2-adj. (%)	28.0	28.7	40.2	18.7	18.6	21.6



(A) $S_{\mathcal{C}}(1,\, x)$: estimates under both monotonic increasing and convex constraints

(B) $S_{\mathcal{C}}(1,\, x)$: estimates under the monotonic increasing constraint

(C) $S_{\mathcal{C}}(1,\, x)$: estimates without shape constraints

Figure 1.5: Estimates of $S_{\mathcal{C}}(1,\, x)$ under different configurations

Table 1.8: First-step model estimation: the financial constraints

This table presents results from the following semi-parametric regressions:

$$\frac{1}{\tau}\left(\frac{\text{return}_{t+\tau}}{r_t}\right) = \left[S_0\left(\exp(\boldsymbol{\delta}^\top \boldsymbol{u}_t)\frac{\bar{b}_t}{r_t}\right) + \boldsymbol{\phi}^\top \boldsymbol{v}_t\right] + \varepsilon_{t+\tau}$$

using daily observations. Model parameters are $\boldsymbol{\delta}$ and $\boldsymbol{\phi}$. The functional form of $S_0(\cdot)$ is treated as unknown and also estimated. The notation "return$_{t+\tau}$" denotes one-quarter-ahead value-weighted equity returns of 49 dealer banks surveyed by FX committees of New York, London, Tokyo, Toronto, Sydney, Singapore and Hong Kong. All returns are net ones in percentage, as well as annualized (divided by $\tau = 0.25$ as shown in the regression specification). The cross-sectional average of absolute one-year basis swap rates for EUR, JPY, GBP, AUD, CAD, and CHF against the dollar is denoted by $\bar{b}_t$. Both volume and equal weighted results are reported. The first set of independent variables in vector $\boldsymbol{u}_t$ are the dollar index, quarterly lagged volatilities of average CIP deviations, changes in dealer banks' CDS, the TED spread (three-month dollar LIBOR rates minus the three-month Treasury bill rates), the implied volatility of euro, the VIX index, and the three-month dollar convenience yield (the RefCorp bond yield minus the treasury yield). The second set of variables in vector $\boldsymbol{v}_t$ are reciprocals of the Fed fund rates ($1/r$), the earnings yields for the 49 dealer banks (E/P), and the VIX index. Numbers in parentheses are standard errors from parametric bootstrap procedures.

	weighted by volume	equally weighted
$\boldsymbol{\delta}$: dynamics of financial constraint		
dollar index	−0.018	−0.013
	(0.009)	(0.005)
lagged vol $\bar{b}_t$ (b.p.)	−0.037	−0.036
	(0.035)	(0.029)
Δ bank CDS (%)	0.096	0.047
	(0.358)	(0.294)
TED spread (%)	3.84	2.49
	(2.12)	(1.38)
ivol euro	−0.156	−0.124
	(0.079)	(0.040)
VIX	0.059	0.040
	(0.039)	(0.030)
\$ conv. yield (%)	−0.933	−0.522
	(0.218)	(0.150)
$\boldsymbol{\phi}$: return controls		
$1/r$	−0.275	−0.376
	(0.099)	(0.111)
E/P	74.5	76.1
	(21.0)	(20.8)
VIX	0.389	1.88
	(5.523)	(5.37)
N obs.	2541	2541
Deviance R^2 (%)	44.6	43.8

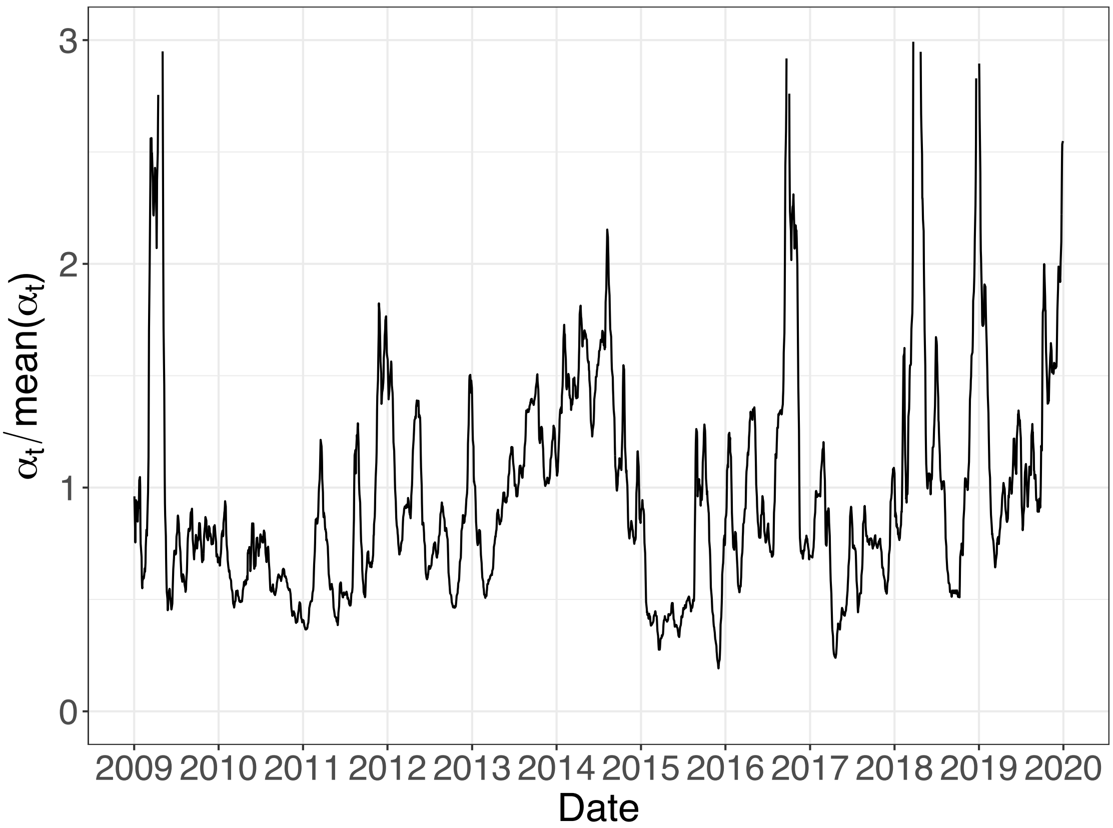


Figure 1.6: The dynamics of financial constraints: $\alpha_t = \exp(\boldsymbol{\delta}^\top \boldsymbol{u}_t)$

Table 1.9: Second-step model estimation: hedging demands

This table presents results from the following panel regressions:

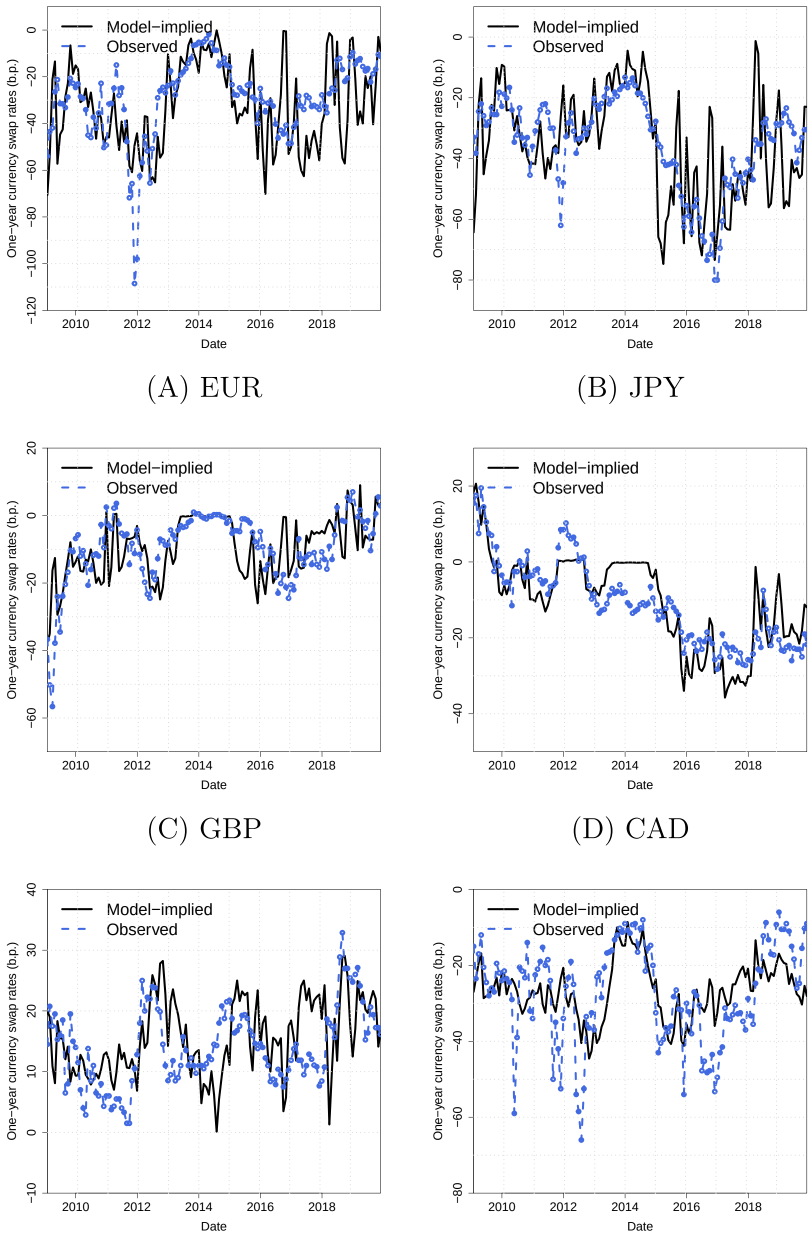
$$\pi_{it}k_t = \overline{\boldsymbol{\beta}}^\top \boldsymbol{x}_{it} + \sum_{j=1}^{n-1} \boldsymbol{\eta}_j^\top \left(I[j = i] \times \boldsymbol{x}_{jt}\right) - \gamma b_{it} + \ell_{it}$$

using monthly observations. π_{it} in the dependent variable is arbitrage positions calculated based on the first-step estimation. k_t is the total market equity of the 49 FX dealer banks, normalized to one on January 2009. The independent variables in $\boldsymbol{xit}$ include a constant one (for the intercept), bilateral net foreign direct investment, net purchases of long-term securities (sovereign and local government bonds, corporate bonds, equities), changes in net bank claims of deposits and short-term securities, and net exports (all in billions). Interest rate differentials (calculated using three-month inter-bank rates, in basis points) are also included. All "net" terms are calculated as "US minus foreign" (taking the US perspective). b_{it} stands for one-year CIP deviations for currency i at time t. The instrumental variable for b_{it} is $\boldsymbol{z}_{it} = \sum_{i' \neq i} w_{i't}\boldsymbol{x}_{i't}^{(-\iota)}$, weighted average of $\boldsymbol{x}_{i't}$ ($i' \neq i$) vectors excluding the constant one (thus the "$-\iota$" superscript). The first-stage regression is then

$$b_{it} = \boldsymbol{\psi}^\top \boldsymbol{z}_{it} + \overline{\boldsymbol{\phi}}^\top \boldsymbol{x}_{it} + \sum_{j=1}^{n-1} \boldsymbol{\xi}_j^\top \left(I[j = i] \times \boldsymbol{x}_{jt}\right) + e_{it},$$

Currencies under consideration are EUR, JPY, GBP, AUD, CAD, and CHF. The sample period is January 2009-December 2019. Numbers in parentheses are Driscoll-Kraay standard errors robust to general forms of serial correlations and correlations among currency pairs (Driscoll and Kraay, 1998).

	OLS	IV	
γ	0.18	1.31	1.43
	(0.22)	(0.51)	(0.45)
Weak IV test:			
Cragg-Donald Statistic		28.7	23.5
theory cutoff (5% relative bias)		18.4	18.4
$\overline{\boldsymbol{\beta}}$:			
net direct investment flows	0.92	0.47	0.37
	(0.64)	(0.79)	(0.85)
net purchase of long-term securities	−0.77	3.39	
	(0.70)	(1.59)	
• bond			5.12
			(1.64)
• equity			−1.03
			(2.15)
net change in bank claims	−0.49	−0.11	−0.10
	(0.33)	(0.33)	(0.32)
bilateral net exports	4.01	11.00	10.00
	(4.78)	(4.63)	(4.06)
$r^{\text{foreign}} - r^{\$}$ (%)	−0.03	−0.07	−0.06
	(0.03)	(0.03)	(0.03)
const.	−19.5	−45.0	−48.7
	(5.9)	(12.9)	(11.5)
N obs	784	784	784
R^2-adj. (%)	54.6	55.9	55.9



(A) EUR

(B) JPY

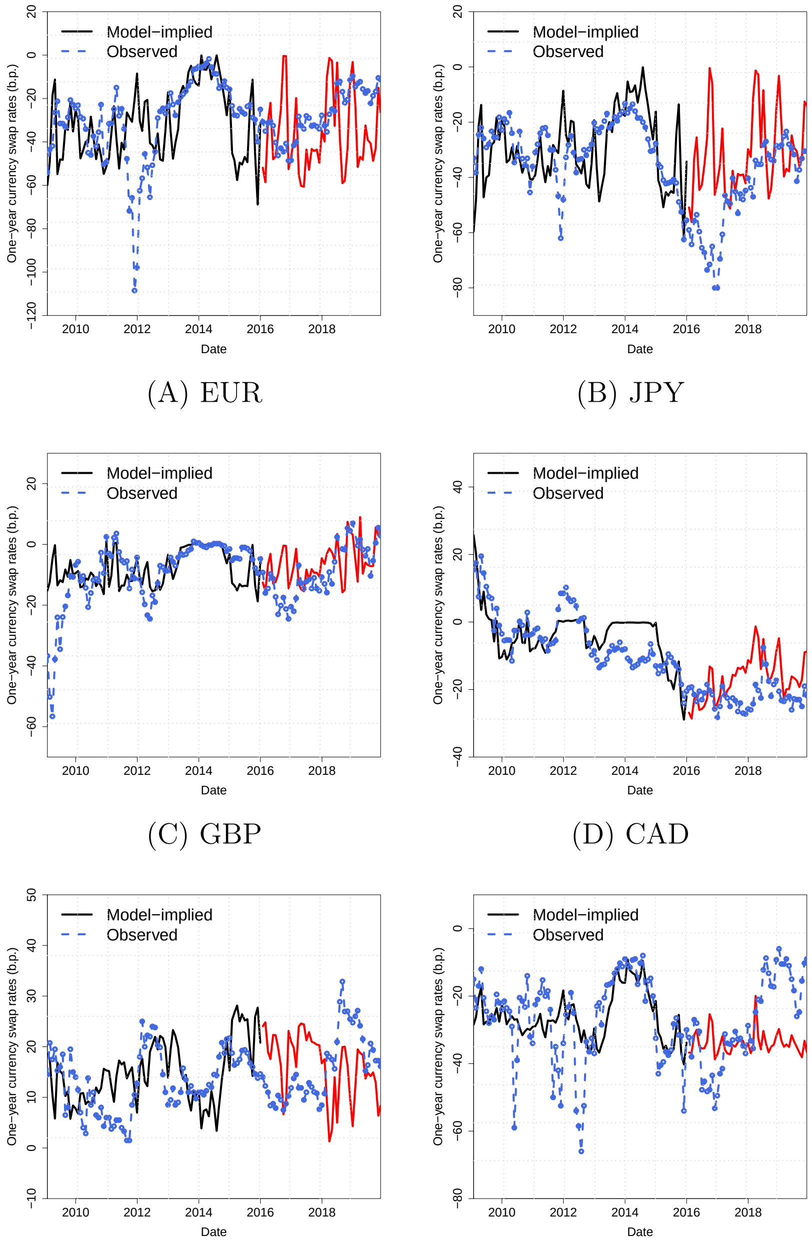
(C) GBP

(D) CAD

(E) AUD

(F) CHF

Figure 1.7: One-year CIP deviations for G6 currencies: model-implied and observations from currency swaps



(A) EUR

(B) JPY

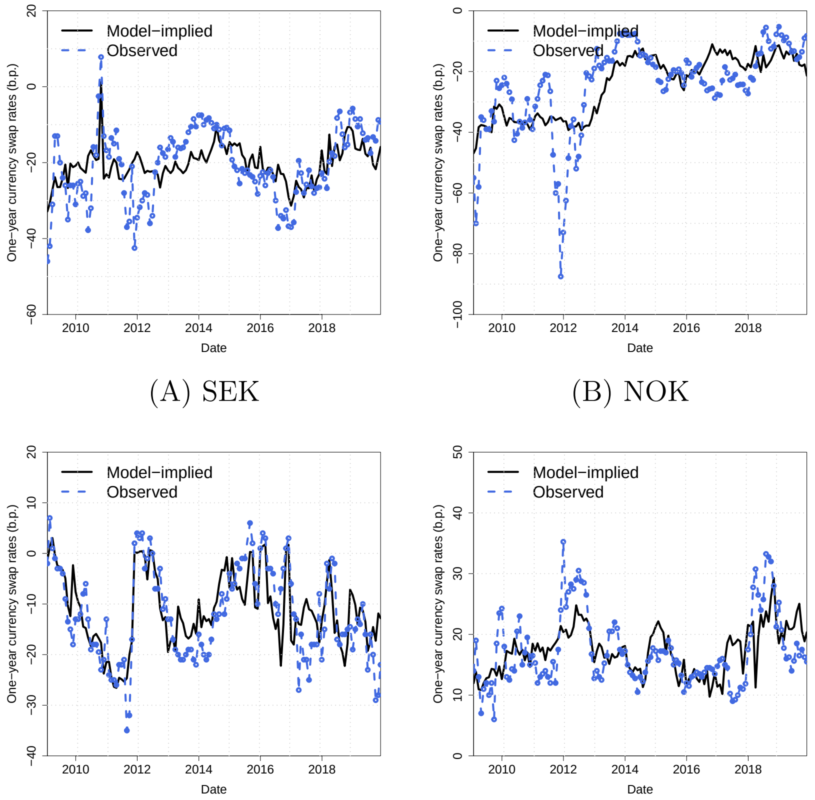
(C) GBP

(D) CAD

(E) AUD

(F) CHF

Figure 1.8: One-year CIP deviations for G6 currencies: model-implied (in-sample 2009-2015 in black, out-of-sample 2016-2019 in red) and observations from currency swaps



(A) SEK

(B) NOK

(C) HKD

(D) NZD

Figure 1.9: One-year CIP deviations for currencies not used for estimation: model-implied and observations from currency swaps

Table 1.10: Model-implied CIP deviations

This table documents CIP deviations solved from the estimated equilibrium equation (1.22), denoted by $\hat{b}$ and compares it with b, the true data (one-year currency swap rates). The parameter $\beta = \text{cov}[\hat{b}, b]/\sigma^2[b]$ measures the fraction of variation in the data explained by the model, estimated from regressing $\hat{b}$ on b.

Currency	$\mathbb{E}[b]$	$\sigma[b]$	$\mathbb{E}[\hat{b}]$	$\sigma[\hat{b}]$	$\text{Corr}[b, \hat{b}]$	β	(s.e.)
Panel A: G6 currencies used for estimation							
EUR	-29.2	16.8	-32.4	18.6	0.56	0.62	(0.13)
JPY	-35.1	14.9	-35.4	17.8	0.61	0.73	(0.12)
GBP	-9.3	10.3	-9.7	9.0	0.59	0.52	(0.08)
CAD	-10.8	11.1	-10.6	12.2	0.83	0.91	(0.10)
AUD	14.0	6.3	15.6	6.3	0.40	0.40	(0.09)
CHF	-26.8	13.0	-26.6	7.4	0.57	0.32	(0.07)
Panel B: currencies not used for model estimation							
SEK	-20.6	9.5	-20.4	4.7	0.58	0.29	(0.06)
NOK	-24.7	14.6	-24.6	10.1	0.69	0.47	(0.08)
HKD	-12.1	9.1	-11.2	7.3	0.75	0.60	(0.07)
NZD	17.2	5.9	17.3	3.9	0.54	0.35	(0.07)

Table 1.11: Model-implied CIP deviations: counterfactual time-series average

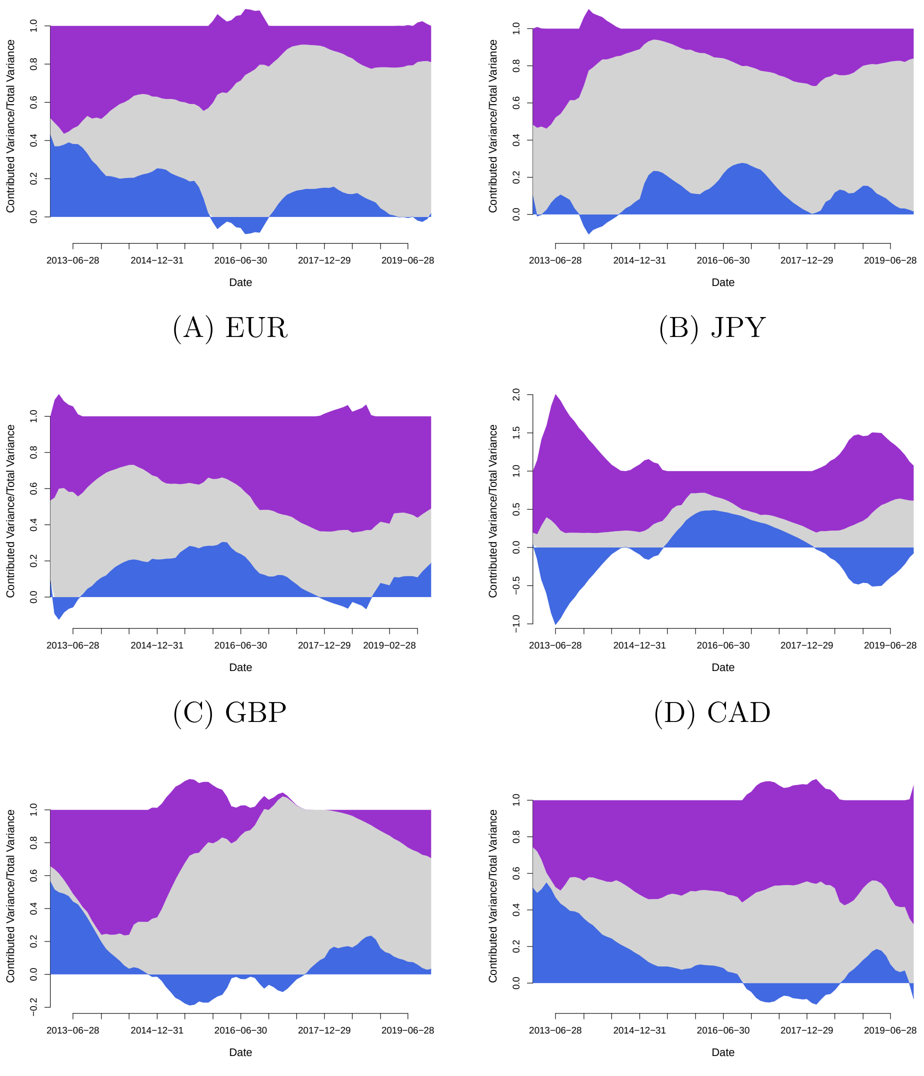
This table reports counterfactual time-series average CIP deviations from 2009 to 2019. The constant c_α relaxes ($c_\alpha < 1$) or tightens ($c_\alpha > 1$) the financial constraint. Numbers in parentheses are standard errors for these (counterfactual) sample mean statistics.

Currency	$c_\alpha = 0.5$	$c_\alpha = 1$	$c_\alpha = 1.5$	$c_\alpha = 2$	$c_\alpha = 2.5$
EUR	−51.4	−32.2	−19.2	−11.2	−6.1
	(15.3)	(17.3)	(15.1)	(11.8)	(7.1)
JPY	−50.2	−35.3	−24.1	−16.2	−8.5
	(15.3)	(17.2)	(17.1)	(14.8)	(8.2)
GBP	−14.0	−9.7	−6.5	−4.2	−2.6
	(11.2)	(9.0)	(7.5)	(6.1)	(4.5)
CAD	−13.7	−10.7	−7.5	−6.3	−4.5
	(13.7)	(12.2)	(14.3)	(9.4)	(7.3)
AUD	21.0	15.6	11.8	7.6	5.2
	(5.3)	(6.3)	(10.3)	(6.5)	(5.7)
CHF	−30.8	−26.6	−22.6	−19.0	−14.2
	(5.4)	(6.9)	(8.0)	(8.4)	(6.6)

Table 1.12: Shapley decomposition of model-implied CIP deviations

The table reports the full-sample Shapley decomposition results for G6 currencies. The decomposition quantifies marginal contribution from each of the three economic forces to variation in model-implied CIP deviations.

Currency	financial constraints	hedging demands	arbitrage capital
EUR	0.553	0.287	0.160
JPY	0.596	0.187	0.217
GBP	0.415	0.294	0.291
CAD	0.235	0.558	0.207
AUD	0.519	0.318	0.163
CHF	0.432	0.470	0.098
Avg.	0.458	0.352	0.190



(A) EUR

(B) JPY

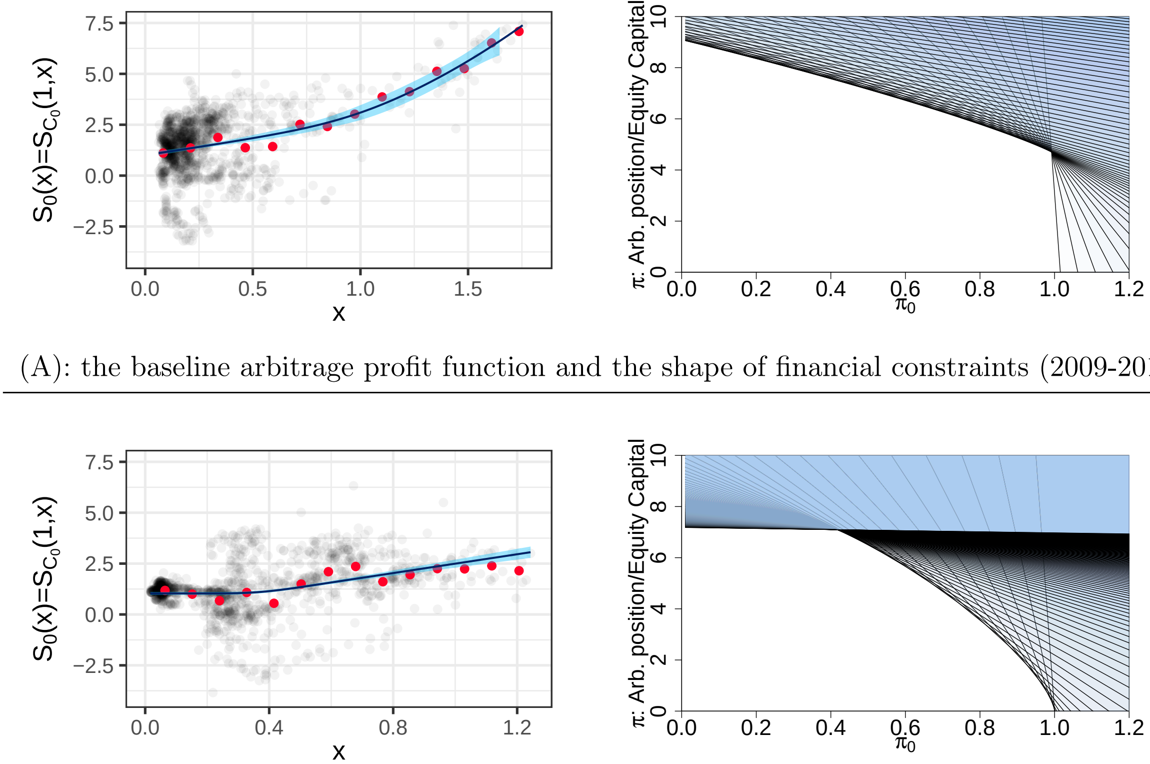
(C) GBP

(D) CAD

(E) AUD

(F) CHF

Figure 1.10: Rolling window Shapley decomposition of variation in CIP deviations (gray: financial constraints, purple: hedging demands, blue: arbitrageurs' capital)



(A): the baseline arbitrage profit function and the shape of financial constraints (2009-2013)

(B): the baseline arbitrage profit function and the shape of financial constraints (2014-2019)

Figure 1.11: Estimates of the baseline arbitrage profit function and shapes of financial constraints

1.7 Appendices

1.7.1 Proofs

Proof of Proposition 1

I begin the proof by stating the following lemma:

Lemma 2. *Let the pair* (π_0^*, π^*) *be such that* $\pi_0^* r + \pi^* b = S_{\mathcal{C}}(r, b)$*, then if* $b > 0$*,* $\pi^* \geq 0$*; if* $b < 0$*,* $\pi^* \leq 0$*.*

Proof of Lemma 2. By Assumption 3, $(1, 0) \in \mathcal{C}$, as a result $S_{\mathcal{C}}(r, b) \geq r$. If $b > 0$ and $\pi^* < 0$ or $b < 0$ and $\pi^* > 0$, then $S_{\mathcal{C}}(r, b) < \pi_0 r \leq r$ (it is alway the case that $\pi_0 \leq 1$), a contradiction, thus the lemma holds.

This lemma says that when $b > 0$, arbitrageurs profit from selling dollars forward[36], i.e., $\pi^* \geq 0$ (positive forward dollar supplies). In the same vein, when $b < 0$, they will buy dollar forward, i.e., $\pi^* \leq 0$ (negative forward dollar supplies).

Now we prove Proposition 1. The arbitrageurs' optimization problem is equivalent to

$$\underset{y,\, s=k-y}{\text{maximize}} \quad \log(y) + \frac{1}{1+\rho}\log(k-y) \quad \text{and} \quad \underset{(\pi_0,\, \pi)\in\mathcal{C}}{\text{maximize}} \quad \pi_0 r + \pi b. \tag{1.23}$$

The first order condition with regard to y and s commands $s = y/(1+\rho)$. Combining this condition with $s = k - y$, we have $y = [(1+\rho)/(2+\rho)]k$ and $s = k/(2+\rho)$. By the definition of $S_{\mathcal{C}}$ and the boundedness of $\mathcal{C}$, optimal combination of π_0 and π must be such that $\pi_0 r + \pi b = S_{\mathcal{C}}(r, b) < \infty$. As a result,

$$k' = s + \pi_0 sr + \pi sb = [1 + S_{\mathcal{C}}(r, b)]\, s = \frac{1 + S_{\mathcal{C}}(r, b)}{2+\rho}k.$$

Noticing the fact that support functions are positively homogeneous of degree one (e.g., Molchanov and Molinari (2018, p. 75-76)), divide both sides of the equation above by $r > 0$ and then minus $1/r$ yield the capital accumulation equation. Substituting k by $[(2+\rho)/(1+\rho)]y$ and k' by y', we have

$$y' = \frac{1 + S_{\mathcal{C}}(r, b)}{1+\rho}y,$$

which agrees with the consumption Euler equation in the proposition.

Since $\mathcal{C}$ is convex, the support function $S_{\mathcal{C}}$ is subadditive and (thus) convex (e.g., Molchanov and Molinari (2018, p. 75-76)). It follows directly that $(k' - k)/k$ is a convex function of b.

To prove that arbitrageurs' capital gain return $(k' - k)/k$ is increasing in $|b|$, we only need to show that, with r fixed, if $b \geq 0$ ($b \leq 0$), $S_{\mathcal{C}}(r, b)$ increases (decreases) in b.

[36] They can offer pounds for dollars to earn the favorable rate $r^{\pounds} + b$ ($b > 0$) in cross-currency swap markets now and return dollars to reclaim pounds later, or simply buy pounds forward (with dollars) in FX swap/forward markets. Either way, they are supplying forward dollars.

Consider $b' \geq b \geq 0$,

$$S_{\mathcal{C}}(r,\, b') = \pi_0^{*'} r + \pi^{*'} b' \geq \pi_0^{*} r + \pi^{*} b' = \pi_0^{*} r + \pi^{*} b + \pi^{*}(b' - b) \geq S_{\mathcal{C}}(r,\, b),$$

where the pair $(\pi_0^{*'},\, \pi_0^{*'}) \in \mathcal{C}$ maximize $\pi_0 r + \pi b'$, and the pair $(\pi_0^{*},\, \pi_0^{*}) \in \mathcal{C}$ maximize $\pi_0 r + \pi b$. The last inequality above holds because $\pi^{*} \geq 0$ when $b \geq 0$ (Lemma 2). Using the same notation, for $b \leq b' \leq 0$,

$$S_{\mathcal{C}}(r,\, b) = \pi_0^{*} r + \pi^{*} b \geq \pi_0^{*'} r + \pi^{*'} b = \pi_0^{*'} r + \pi^{*'} b' + \pi^{*'}(b - b') \geq S_{\mathcal{C}}(r,\, b').$$

The last inequality above follows from the fact that $\pi^{*'} \leq 0$ when $b' \leq 0$ (Lemma 2).

Q.E.D.

Proof of Proposition 2

The arbitrageurs' optimal positions π_0^{*} and π^{*} are such that $\pi_0^{*} r + \pi^{*} b = S_{\mathcal{C}}(r,\, b)$. Since $S_{\mathcal{C}}$ is positively homogeneous of degree one, and is differentiable (by assumption), we can apply Euler's homogeneous function theorem (Mas-Colell, Whinston, and Green, 1995, Theorem M.B.2, p. 929), which implies that $\pi(b) = \pi^{*} = \partial S_{\mathcal{C}}(r,\, b)/\partial b$.

Then we turn to the existence and uniqueness result. Plugging the result for $\pi(b) = \partial S_{\mathcal{C}}(r,\, b)/\partial b$ and the expression (1.5) for $q(b)$ into equation (1.6), we have

$$b = \frac{\gamma_0}{\gamma + \dfrac{\partial S_{\mathcal{C}}(r,\, b)}{b \partial b} s}. \tag{1.24}$$

Noticing that, by assumption, $\pi(0) = \partial S_{\mathcal{C}}(r,\, 0)/\partial b = 0$ and $S_{\mathcal{C}}$ is twice differentiable,

$$\frac{\partial S_{\mathcal{C}}(r,\, b)}{b \partial b} = \frac{1}{b - 0}\left(\frac{\partial S_{\mathcal{C}}(r,\, b)}{\partial b} - \frac{\partial S_{\mathcal{C}}(r,\, 0)}{b \partial b}\right) = \left.\frac{\partial^2 S_{\mathcal{C}}(r,\, b)}{\partial b^2}\right|_{b = \hat{b} \in [0,\, b]}.$$

Due to the convexity of the support function $S_{\mathcal{C}}$, the condition $\partial^2 S_{\mathcal{C}}(r,\, b)/\partial b^2 \geq 0$ holds against any values of b for which $S_{\mathcal{C}}$ is well-defined, thus the right hand side of equation (1.24), namely $F(b)$, uniformly falls within the interval $[0,\, \gamma_0/\gamma]$ if $\gamma_0 \geq 0$ or $[\gamma_0/\gamma,\, 0]$ if $\gamma_0 < 0$. Since $S_{\mathcal{C}}$ is twice differentiable, $\pi(b) = \partial S_{\mathcal{C}}(r,\, b)/\partial b$ is continuous, and so is the function $F(b)$ (notice that its denominator is always positive as $\gamma > 0$). By Brouwer's fixed point theorem (Mas-Colell, Whinston, and Green, 1995, Theorem M.I.1, p. 952), the equation $F(b) = b$ admits a solution b^{*} in $[0,\, \gamma_0/\gamma]$ if $\gamma_0 \geq 0$ or $[\gamma_0/\gamma,\, 0]$ if $\gamma_0 < 0$, thus the existence result.

The uniqueness result follows naturally from the monotonicity of $\pi(b)s - q(b)$ in b. Since $\pi'(b)s - q'(b) = \partial^2 S_{\mathcal{C}}(r,\, b)/\partial b^2 s + \gamma > 0$ as long as $\gamma > 0$, $\pi(b)s - q(b)$ monotonically increases, thus the solution b^{*} to $\pi(b)s - q(b) = 0$ is unique.

Next, we prove the conclusion that $|b^{*}|$ is decreasing in the arbitrageurs' initial capital

k. From the proof of Proposition 1, the arbitrageurs' saving $s = k/(2+\rho)$, thus

$$k = \frac{(\rho+2)(\gamma_0 - \gamma b)}{\pi(b)}. \tag{1.25}$$

The right-hand side function of b in equation (1.25), denote by $G(b)$, has a derivative

$$G'(b) = -(\rho+2)\frac{\gamma\pi(b) + (\gamma_0 - \gamma b)\pi'(b)}{[\pi(b)]^2},$$

in which $\pi'(b) \geq 0$ for all b and $\gamma > 0$ by the assumption. When $\gamma_0 \geq 0$, $\gamma_0 - \gamma b \geq 0$ and $b \geq 0$, which implies $\pi(b) \geq 0$ from Lemma 2. As a result, $G'(b) \leq 0$: an increase in k will leads to a smaller $b^* \geq 0$ such that $G(b^*) = k$. When $\gamma_0 < 0$, $\gamma_0 - \gamma b \leq 0$ and $b \leq 0$, indicating $\pi(b) \leq 0$ from Lemma 2. Then $G'(b) \geq 0$: a increase in k will require a larger $b^* \leq 0$ such that $G(b^*) = k$. Summing up the conclusions, $|b^*|$ decreases in k.

Q.E.D.

Proofs (and possible generalizations) of propositions and lemmas in Section 1.3

Here I consider the general case: replacing the log utility with a power utility function $u(y) = (y^{1-\gamma} - 1)/(1-\gamma)$. The log utility specification is the special case when $\gamma = 1$. The arbitrageurs' problem is to maximize

$$J_t = \mathbb{E}_t \left[\int_0^\infty e^{-\rho s} u\left(y_{t+s}\right) \mathrm{d}s \right],$$

under the budget constraint (as defined in equation (1.11))

$$\frac{\mathrm{d}k}{k} = \pi_0 \left[r\mathrm{d}t + w(\mu - r)\mathrm{d}t + w\sigma\mathrm{d}z \right] + \boldsymbol{\pi}^\top \boldsymbol{b}\mathrm{d}t - \frac{y}{k}\mathrm{d}t,$$

and position constraint

$$(\pi_0,\, \boldsymbol{\pi}) \in \mathcal{C},$$

by choosing $(y,\, w,\, \pi_0, \boldsymbol{\pi})$. It is worthwhile mentioning that all time-t subscripts are omitted here. For example, the constraint $(\pi_0,\, \boldsymbol{\pi}) \in \mathcal{C}$ indeed represents $(\pi_{0t},\, \boldsymbol{\pi}_t) \in \mathcal{C}_t$.

The Hamilton-Jacobi-Bellman (HJB) equation for arbitrageurs' optimization problem is

$$\rho J = \sup_{y,\, w,\, \pi_0,\, \boldsymbol{\pi}} \left\{ u(y) + \mathcal{D}J \right\}, \text{ where } (\pi_0,\, \boldsymbol{\pi}) \in \mathcal{C}. \tag{1.26}$$

The value function $J(k,\, K,\, \boldsymbol{s})$ is defined for the capital k of each arbitrageur, as well as the aggregate capital $K = \int_{[0,\,1]} k\mathrm{d}i$ of all arbitrageurs (Kyle and Xiong, 2001; Kondor and Vayanos, 2019). K is effectively an additional state variable because it determines the equilibrium arbitrage yield vector $\boldsymbol{b}$. Given that arbitrageurs are identical and of mass one, $K = k$ in equilibrium. The vector $\boldsymbol{s} \in \mathbb{R}^p$ incorporates all other state variables such as ones that determine i. the hedging demands; ii. time variation of the constraint; iii. the triplet $(r_t,\, \mu_t,\, \sigma_t)$ characterizing arbitrageurs' other investment opportunities. I make clear the assumptions about $\boldsymbol{s}$ as follows:

Assumption 8. *The dynamics of* $\boldsymbol{s} = (s_1,\, \ldots,\, s_p)^\top$ *is written as a vector Itô process defined in a complete probability space, that is,*

$$\mathrm{d}\boldsymbol{s} = P(\boldsymbol{s})\mathrm{d}t + Q(\boldsymbol{s})\mathrm{d}\boldsymbol{z}_s,$$

where $\{\boldsymbol{z}_s\}$ *is a p-dimensional vector of independent standard Brownian motions; the vector-valued function* $P : \mathbb{R}^p \mapsto \mathbb{R}^p$ *is such that* $\sup_{\boldsymbol{s}} \|P(\boldsymbol{s})\|_2 < \infty$*; the matrix-valued function* $Q : \mathbb{R}^p \mapsto \mathbb{R}^{p\times p}$ *is such that* $Q(\boldsymbol{s})Q(\boldsymbol{s})^\top$ *is positive definite with a finite dominant eigenvalue for all* $\boldsymbol{s}$*.*

The time-varying elements of the model $\boldsymbol{\gamma}_{0,t}$ *(hedging demand intercepts),* $\mathcal{C}_t$ *(financial constraints), and* $(r_t,\, \mu_t,\, \sigma_t)$ *(investment opportunities beyond riskless arbitrage) all relate to* $\boldsymbol{s}$ *as follows:*

i. $\boldsymbol{\gamma}_{0,t} = \boldsymbol{\gamma}_0(\boldsymbol{s})$ *where the mapping* $\gamma_0 : \mathbb{R}^p \mapsto \mathbb{R}^n$ *is continuously differentiable;*

ii. $\mathcal{C}_t$ *is such that its support function* $S_{\mathcal{C}_t}(x) = S_0(\boldsymbol{s}, x)$ *for all* $x \in \mathrm{dom}(S_{\mathcal{C}_t})$ *where* S_0 *is continuously differentiable in* $\boldsymbol{s}$*;*

iii. $r_t = r(\boldsymbol{s})$*,* $\mu_t = \mu(\boldsymbol{s})$*,* $\sigma_t = \sigma(\boldsymbol{s})$ *where the three mappings* $r : \mathbb{R}^p \mapsto \mathbb{R}$*,* $\mu : \mathbb{R}^p \mapsto \mathbb{R}$ *and* $\sigma : \mathbb{R}^p \mapsto \mathbb{R}^+$ *are all continuously differentiable.*

Under the assumption above, I calculate the infinitesimal generator for the value function $J(k,\, K,\, \boldsymbol{s})$ as follows:

$$\begin{aligned}
\mathcal{D}J =& J_k \frac{\mathbb{E}\left[\mathrm{d}k\right]}{\mathrm{d}t} + \frac{1}{2} J_{kk} \frac{\mathbb{E}\left[\mathrm{d}k\mathrm{d}k\right]}{\mathrm{d}t} + J_{k\boldsymbol{s}}^\top \frac{\mathbb{E}\left[\mathrm{d}k\mathrm{d}\boldsymbol{s}\right]}{\mathrm{d}t} + J_{kK} \frac{\mathbb{E}\left[\mathrm{d}k\mathrm{d}K\right]}{\mathrm{d}t} \\
&+ \underbrace{J_K \frac{\mathbb{E}\left[\mathrm{d}K\right]}{\mathrm{d}t} + J_{\boldsymbol{s}}^\top \frac{\mathbb{E}\left[\mathrm{d}\boldsymbol{s}\right]}{\mathrm{d}t} + \frac{1}{2}\mathrm{tr}\left\{ J_{\boldsymbol{ss}} \frac{\mathbb{E}\left[\mathrm{d}\boldsymbol{s}\mathrm{d}\boldsymbol{s}^\top\right]}{\mathrm{d}t} \right\} + \frac{1}{2} J_{KK} \frac{\mathbb{E}\left[\mathrm{d}K\mathrm{d}K\right]}{\mathrm{d}t} + J_{K\boldsymbol{s}}^\top \frac{\mathbb{E}\left[\mathrm{d}K\mathrm{d}\boldsymbol{s}\right]}{\mathrm{d}t}}_{\text{constant w.r.t. } (y,\, w,\, \pi_0,\, \boldsymbol{\pi})} \\
=& J_k \left\{ k\pi_0 \left[r + w(\mu - r)\right] + k\boldsymbol{\pi}^\top \boldsymbol{b} - y \right\} + \frac{1}{2} J_{kk} k^2 \pi_0^2 w^2 \sigma^2 \\
&+ \sum_{j=1}^{p} J_{ks_j} k\pi_0 w\sigma \frac{\mathbb{E}\left[\mathrm{d}z\mathrm{d}s_j\right]}{\mathrm{d}t} + J_{kK}^\top k\pi_0 w K \Pi_0 W \sigma^2 + \text{constant},
\end{aligned}$$

where $\Pi_0 = \int_{[0,\,1]} \pi_0 \mathrm{d}i$ and $w = \int_{[0,\,1]} w \mathrm{d}i$ aggregate positions π_0 and w of all arbitrageurs. In equilibrium, $\Pi_0 = \pi_0$ and $W = w$.

For the ease of exposition below, I introduce two definitions first.

Definition 1. *(Bertsekas, 2009, Chapter. 1, p. 7, Properness of a function) A proper function* f *is one such that* $f(x) < \infty$ *for at least one* x *in its domain and* $f(x) > -\infty$ *for all* x *in its domain.*

Definition 2. *(Bertsekas, 2009, Chapter. 1, p. 83, Conjugate functions) Consider a real-valued function f, the conjugate function of f is the function $f^\star$ defined by $f^\star(y) = \sup\{x^\top y - f(x)\}$.*

Now I begin to list and prove a set of lemmas.

Lemma 3. *Define the indicator function for $\mathcal{C}$:*

$$I_\mathcal{C}(x) = \left\{ \begin{array}{ll} 0, & x \in \mathcal{C} \\ +\infty, & x \notin \mathcal{C} \end{array} \right. . \tag{1.27}$$

$I_\mathcal{C}$ is a proper closed convex function.

Proof. First, it is always true that $I_\mathcal{C} \geq 0 > -\infty$, and, as long as $\mathcal{C}$ is nonempty (true by assumption), $I_\mathcal{C} = 0 < \infty$, for $x \in \mathcal{C}$. As a result, $I_\mathcal{C}$ is proper.

Second, consider the epigraph of $I_\mathcal{C}$, defined as $\{(x,\, \alpha) : I_\mathcal{C}(x) \leq \alpha\}$. By definition, this set is $\mathcal{C} \times [0,\, \infty)$, which is convex as long as $\mathcal{C}$ is convex. Thus, $I_\mathcal{C}$ must be convex (Bertsekas, 2009, Chapter. 1, p. 8, Definition 1.1.3).

Third, consider the set $\{x : I_\mathcal{C}(x) \leq \alpha\}$, which equals $\mathcal{C}$ (a closed set by assumption) when $\alpha \geq 0$ and $\emptyset$ (always closed) otherwise. Thus, $I_\mathcal{C}$ is a closed function. □

Lemma 4. *The indicator function of the set $\mathcal{C}$ is such that $I_\mathcal{C}(x) = \sup_y \left\{x^\top y - S_\mathcal{C}(y)\right\}$, that is, $I_\mathcal{C} = S_\mathcal{C}^\star$.*

Proof. First, noticing that

$$S_\mathcal{C}(y) = \sup_{x \in \mathcal{C}} x^\top y = \sup \left\{x^\top y - I_\mathcal{C}(x)\right\},$$

that is, $S_\mathcal{C}$ is the conjugate of $I_\mathcal{C}$, or simply $S_\mathcal{C} = I_\mathcal{C}^\star$.

Next, since $I_\mathcal{C}$ is a proper closed convex function, by the Conjugacy Theorem (Bertsekas, 2009, Chapter. 1, p. 85-86), $I_\mathcal{C}^{\star\star} = I_\mathcal{C}$, that is, the conjugate function of $I_\mathcal{C}^\star$ is $I_\mathcal{C}$ itself. Thus, $S_\mathcal{C}^\star = I_\mathcal{C}$. □

Lemma 5. *The HJB equation of* (1.26) *under the constraint* $(\pi_0,\, \boldsymbol{\pi}) \in \mathcal{C}$ *is equivalent to*

$$\rho J = \inf_{\boldsymbol{\nu}} \sup_{y,\, w,\, \widehat{\boldsymbol{\pi}}} \left\{u(y) + \mathcal{D}J + S_\mathcal{C}(\boldsymbol{\nu}) - \widehat{\boldsymbol{\pi}}^\top \boldsymbol{\nu}\right\}, \tag{1.28}$$

without any constraints, where

$$\widehat{\boldsymbol{\pi}} = \begin{pmatrix} \pi_0 \\ \boldsymbol{\pi} \end{pmatrix}$$

is a vector concatenating π_0 and $\boldsymbol{\pi}$, $\boldsymbol{\nu} = (\nu_0, \nu_1, \ldots, \nu_n)^\top$ is a vector of $(n+1)$ dimensions.

Proof. Under the definition of indicator functions introduced in (1.27), the problem of (1.26) is equivalent to

$$\rho J = \sup_{y,\, w,\, \widehat{\boldsymbol{\pi}}} \left\{u(y) + \mathcal{D}J - I_\mathcal{C}(\widehat{\boldsymbol{\pi}})\right\}.$$

This equivalence is easy to understand. When $\widehat{\boldsymbol{\pi}} \in \mathcal{C}$, the optimization problem is exactly the original one. Otherwise, the indicator function penalizes the objective function so harshly that regardless how carefully the choice variables are picked, the outcome is always $-\infty$.

From Lemma 4, $-I_{\mathcal{C}}(\widehat{\boldsymbol{\pi}}) = \inf_{\boldsymbol{\nu}} \left\{ S_{\mathcal{C}}(\boldsymbol{\nu}) - \widehat{\boldsymbol{\pi}}^{\top}\boldsymbol{\nu} \right\}$. Thus

$$\begin{aligned}
\rho J &= \sup_{y,\, w,\, \widehat{\boldsymbol{\pi}}} \left\{ u(y) + \mathcal{D}J + \inf_{\boldsymbol{\nu}} \left\{ S_{\mathcal{C}}(\boldsymbol{\nu}) - \widehat{\boldsymbol{\pi}}^{\top}\boldsymbol{\nu} \right\} \right\}, \\
&= \sup_{y,\, w,\, \widehat{\boldsymbol{\pi}}} \inf_{\boldsymbol{\nu}} \left\{ u(y) + \mathcal{D}J + S_{\mathcal{C}}(\boldsymbol{\nu}) - \widehat{\boldsymbol{\pi}}^{\top}\boldsymbol{\nu} \right\} \qquad (1.29) \\
&= \inf_{\boldsymbol{\nu}} \sup_{y,\, w,\, \widehat{\boldsymbol{\pi}}} \left\{ u(y) + \mathcal{D}J + S_{\mathcal{C}}(\boldsymbol{\nu}) - \widehat{\boldsymbol{\pi}}^{\top}\boldsymbol{\nu} \right\}.
\end{aligned}$$

The last equation follows from the fact that the function $\{u(y) + \mathcal{D}J + S_{\mathcal{C}}(\boldsymbol{\nu}) - \widehat{\boldsymbol{\pi}}^{\top}\boldsymbol{\nu}\}$ as a whole is concave in $(y,\, w,\, \widehat{\boldsymbol{\pi}})$ with fixed $\boldsymbol{\nu}$ and convex in $\boldsymbol{\nu}$ with fixed $(y,\, w,\, \widehat{\boldsymbol{\pi}})$, satisfying the saddle point property. □

Now I prove the Proposition 3. From Lemma 5, the initial maximization problem of (1.28) leads to the following first-order condition with regard to w:

$$J_k k \pi_0 (\mu - r) + J_{kk} k^2 \pi_0^2 w \sigma^2 + \sum_{j=1}^{p} J_{k s_j} k \pi_0 \sigma \frac{\mathbb{E}\left[\mathrm{d}z \mathrm{d}s_j\right]}{\mathrm{d}t} + J_{kK} k \pi_0 K \Pi_0 W \sigma^2 = 0. \qquad (1.30)$$

For elements in $\widehat{\boldsymbol{\pi}}$, the first order condition with regard to π_0 is

$$J_k k \left[r + w(\mu - r) \right] + J_{kk} k^2 \pi_0 w^2 \sigma^2 + \sum_{j=1}^{p} J_{k s_j} k w \sigma \frac{\mathbb{E}\left[\mathrm{d}z \mathrm{d}s_j\right]}{\mathrm{d}t} + J_{kK} k w K \Pi_0 W \sigma^2 - \nu_0 = 0. \qquad (1.31)$$

Performing the calculation of (1.31)-[(1.30)×w/π_0], for both sides of the two equations, we have

$$J_k k r = \nu_0. \qquad (1.32)$$

For $\pi_1, \ldots, \pi_n$ in the vector $\widehat{\boldsymbol{\pi}}$, Lemma 5 commands choosing π_i to maximize $(J_k k b_i - \nu_i)\pi_i$. As long as $J_k k b_i$ does not equal ν_i, the maximized objective function reaches infinity. Thus, in equilibrium,

$$J_k k b_i = \nu_i, \qquad (1.33)$$

for all $i = 1, \ldots, n$.

Now consider the equation (1.29) shown in the proof of Lemma 5. The initial minimization problem with regard to $\boldsymbol{\nu}$,

$$\inf_{\boldsymbol{\nu}} \left\{ u(y) + \mathcal{D}J + S_{\mathcal{C}}(\boldsymbol{\nu}) - \widehat{\boldsymbol{\pi}}^{\top}\boldsymbol{\nu} \right\},$$

will only yield two possible outcomes: $-\infty$ or $\{u(y)+\mathcal{D}J\}$. In equilibrium, this outcome cannot be $-\infty$, thus $\boldsymbol{\nu}$ must be such that $S_{\mathcal{C}}(\boldsymbol{\nu}) - \widehat{\boldsymbol{\pi}}^\top \boldsymbol{\nu} = 0$, that is

$$\pi_0 \nu_0 + \sum_{i=1}^{n} \pi_i \nu_i = S_{\mathcal{C}}(\nu_0,\, \nu_1,\, \ldots,\, \nu_n). \tag{1.34}$$

Noticing that $S_{\mathcal{C}}$ is positively homogeneous of degree one, divide both sides of equation (1.34) by $J_k k$

$$\pi_0 \frac{\nu_0}{J_k k} + \sum_{i=1}^{n} \pi_i \frac{\nu_i}{J_k k} = S_{\mathcal{C}}\left(\frac{\nu_0}{J_k k},\, \frac{\nu_1}{J_k k},\, \ldots,\, \frac{\nu_n}{J_k k}\right),$$

and combine the result above with equation (1.32) as well as the set of equations (1.33),

$$\pi_0 r + \boldsymbol{\pi}^\top \boldsymbol{b} = S_{\mathcal{C}}(r,\, \boldsymbol{b}).$$

Proposition 3 then follows from the equation above as well as Euler's homogeneous function theorem (Mas-Colell, Whinston, and Green, 1995, Theorem M.B.2, p. 929).

Q.E.D.

For Proposition 4, I state the following generalized version (for CRRA utility functions) and then present its proof.

Proposition 7. *Under Assumption 8, in equilibrium, there exist a function $g(K,\, \boldsymbol{s}) : \mathbb{R}\times\mathbb{R}^p \mapsto \mathbb{R}$ of the aggregate capital and the state variables, such that arbitrageurs' optimal rates of consumption y satisfy*

$$\log\left(\frac{y}{k}\right) = \frac{1}{\gamma}\left[\log\rho - (1-\gamma) g(K,\, \boldsymbol{s})\right];$$

their total positions on the risky project ($\pi_0 w$) are

$$\frac{1}{A(K,\, \boldsymbol{s})}\left(\frac{\mu - r}{\sigma^2} + \sum_{j=1}^{p} \lambda_j(K,\, \boldsymbol{s})\beta_j\right),$$

where

$$\beta_j = \frac{\operatorname{cov}[\mathrm{d}\widetilde{r},\, \mathrm{d}s_j]}{\operatorname{var}[\mathrm{d}\widetilde{r}]},$$

is the regression coefficient of the changes in the j-th state variable, namely $\mathrm{d}s_j$, regressed on the risky project return $\mathrm{d}\widetilde{r}$; functions A and λ_j are

$$A(K,\, \boldsymbol{s}) = \gamma - (1-\gamma)\frac{\partial g(K,\, \boldsymbol{s})}{\partial K} K,$$

$$\lambda_j(K,\, \boldsymbol{s}) = (1-\gamma)\frac{\partial g(K,\, \boldsymbol{s})}{\partial s_j}.$$

Proof. From Lemma 5, the maximization problem yields the following first-order-condition for y:

$$u'(y) = J_k.$$

Homogeneity of $u'(y) = y^{-\gamma}$ implies that the value function is of the following format:

$$J(k,\, K,\, \boldsymbol{s}) = \frac{1}{\rho} u\left(kG(K,\, \boldsymbol{s})\right), \quad \text{where } G(K,\, \boldsymbol{s}) = \exp(g(K,\, \boldsymbol{s})).$$

For the special case of log utility ($\gamma = 1$), the specification still holds and

$$J(k,\, K,\, \boldsymbol{s}) = \frac{1}{\rho} \log(k) + g(K,\, \boldsymbol{s}).$$

Noticing that

$$J_k = \frac{1}{\rho} u'(kG)G = \frac{1}{\rho} k^{-\gamma} G^{1-\gamma},$$

the equation $u'(y) = J_k$ is equivalent to $\rho y^{-\gamma} = k^{-\gamma} G^{1-\gamma}$. Taking logarithm and rearranging terms,

$$\log\left(\frac{y}{k}\right) = \frac{1}{\gamma}\left[\log \rho - (1-\gamma) g(K,\, \boldsymbol{s})\right].$$

From (1.30),

$$\pi_0 w = -\frac{J_k}{kJ_{kk}}\left(\frac{\mu - r}{\sigma^2}\right) - \sum_{j=1}^{p} \frac{J_{ks_j}}{kJ_{kk}} \underbrace{\frac{\mathbb{E}\left[(\sigma \mathrm{d}z)\mathrm{d}s_j\right]}{\mathbb{E}\left[(\sigma \mathrm{d}z)(\sigma \mathrm{d}z)\right]}}_{=\beta_j} - \frac{J_{kK}}{kJ_{kk}} K \Pi_0 W.$$

Second order derivatives of the value function are given by

$$J_{kk} = -\frac{\gamma}{\rho} k^{-\gamma-1} G^{1-\gamma} = -\frac{\gamma J_k}{k}, \quad J_{ks_j} = \frac{1-\gamma}{\rho} k^{-\gamma} G^{-\gamma} \frac{\partial G}{\partial s_j}, \quad J_{kK} = \frac{1-\gamma}{\rho} k^{-\gamma} G^{-\gamma} \frac{\partial G}{\partial K}.$$

Plugging all expressions above to the equation for $\pi_0 w$:

$$\begin{aligned}
\pi_0 w &= \frac{\mu - r}{\gamma \sigma^2} + \sum_{j=1}^{p} \frac{1-\gamma}{\gamma} \frac{\partial G}{G \partial s_j} \beta_j + \frac{1-\gamma}{\gamma} \frac{\partial G}{G \partial K} K \Pi_0 W \\
&= \frac{\mu - r}{\gamma \sigma^2} + \sum_{j=1}^{p} \frac{1-\gamma}{\gamma} \frac{\partial g}{\partial s_j} \beta_j + \frac{1-\gamma}{\gamma} \frac{\partial g}{\partial K} K \Pi_0 W \\
&= \frac{\mu - r}{\gamma \sigma^2} + \sum_{j=1}^{p} \frac{\lambda_j(K,\, \boldsymbol{s})}{\gamma} \beta_j + \frac{1-\gamma}{\gamma} \frac{\partial g}{\partial K} K \Pi_0 W
\end{aligned}$$

Noticing that the aggregate positions Π_0 and W equal π_0 and w respectively, then

$$\left(1 - \frac{1-\gamma}{\gamma} \frac{\partial g}{\partial K} K\right) \pi_0 w = \frac{1}{\gamma}\left(\frac{\mu - r}{\sigma^2} + \sum_{j=1}^{p} \lambda_j(K,\, \boldsymbol{s}) \beta_j\right),$$

that is,

$$\pi_0 w = \frac{1}{A(K,\, \boldsymbol{s})}\left(\frac{\mu - r}{\sigma^2} + \sum_{j=1}^{p} \lambda_j(K,\, \boldsymbol{s}) \beta_j\right).$$

□

Proposition 4 is the special case of the results above when $\gamma = 1$. Results collected here in Proposition 7 have natural interpretations. The consumption-to-wealth ratio y/k is a function of the state variables $\boldsymbol{s}$ and aggregate capital K when $\gamma \neq 1$. It equals the constant ρ for the log utility case.

Arbitrageurs' demand for the risky project (proportional to their capital) is "Mertonian", both the myopic mean-variance demand and state-variable hedging demands (driven by the betas) appear when $\gamma \neq 1$. The ratio $-\lambda_j/A$ can be interpreted as the risk premium of the risky project due to its exposure to the risk factor s_j. With the log utility, $\lambda_j(K, \boldsymbol{s}) = 1$ for all $j = 1, \dots, p$, and all hedging demands disappear.

Since the aggregate capital also becomes an endogenous state-variable, a dynamic risk-aversion function $A(K, \boldsymbol{s})$ emerges and replaces the constant relative risk-aversion parameter γ, similar to the exposition of Kondor and Vayanos (2019). With the log utility specification, the dynamic risk-aversion $A(K, \boldsymbol{s})$ equals one. Proof of Proposition 4 thus follows through.

Q.E.D.

Next I present and prove a generalized version of Proposition 5.

Proposition 8. *In equilibrium, the arbitrageurs' capital evolves according to the following rule:*

$$\frac{\mathrm{d}k}{k} = \left[\lambda\hat{\sigma} - \frac{y}{k} + S_{\mathcal{C}}(r, \boldsymbol{b})\right]\mathrm{d}t + \hat{\sigma}\mathrm{d}z \tag{1.35}$$

where

$$\hat{\sigma} = \frac{1}{\mathrm{d}t}\mathbb{E}\left[\left(\frac{\mathrm{d}k}{k}\right)^2\right] = \frac{1}{A(K, \boldsymbol{s})}\left(\lambda + \sum_{j=1}^{p}\sigma\lambda_j(K, \boldsymbol{s})\beta_j\right);$$

$$\frac{y}{k} = \exp\left\{\frac{1}{\gamma}\left[\log\rho - (1-\gamma)g(K, \boldsymbol{s})\right]\right\};$$

functions $\lambda_j(K, \boldsymbol{s})$, $j = 1, \dots, p$, and $A(K, \boldsymbol{s})$ are defined as in Proposition 7; $\lambda = (\mu - r)/\sigma$ is the Sharpe ratio of the risky project available to arbitrageurs.

Proof. Plugging results from Proposition 7 into the dynamic budget constraint of arbitrageurs, we have

$$\begin{aligned}
\frac{\mathrm{d}k}{k} &= \pi_0\left[r\mathrm{d}t + w(\mu - r)\mathrm{d}t + w\sigma\mathrm{d}z\right] + \boldsymbol{\pi}^\top\boldsymbol{b}\mathrm{d}t - \frac{y}{k}\mathrm{d}t \\
&= \left(\pi_0 r + \boldsymbol{\pi}^\top\boldsymbol{b}\right)\mathrm{d}t + \pi_0 w(\mu - r)\mathrm{d}t - \frac{y}{k} + \pi_0 w\sigma\mathrm{d}z \\
&= S_{\mathcal{C}}(r, \boldsymbol{b})\mathrm{d}t + \frac{1}{A(K, \boldsymbol{s})}\left(\lambda^2 + \sum_{j=1}^{p}\lambda\sigma\lambda_j(K, \boldsymbol{s})\beta_j\right)\mathrm{d}t - \frac{y}{k}\mathrm{d}t + \frac{1}{A(K, \boldsymbol{s})}\left(\lambda + \sum_{j=1}^{p}\sigma\lambda_j(K, \boldsymbol{s})\beta_j\right)\mathrm{d}z,
\end{aligned}$$

in which

$$\frac{y}{k} = \exp\left\{\frac{1}{\gamma}\left[\log\rho - (1-\gamma)g(K, \boldsymbol{s})\right]\right\}$$

in equilibrium. The proposition follows through. □

With $\gamma = 1$ (the log-utility case), functions $A = 1$ and $\lambda_j = 0$, as a result, $\hat{\sigma} = \lambda$.

Also, the ratio y/k is the constant ρ in equilibrium under the log utility as in Proposition 4. Plugging these quantities back to equation (1.35), we have

$$\frac{\mathrm{d}k}{k} = \left[\lambda^2 - \rho + S_{\mathcal{C}}(r,\, \boldsymbol{b})\right] \mathrm{d}t + \lambda \mathrm{d}z,$$

completing the proof for Proposition 5.

Q.E.D.

I now prove Proposition 6. First, I show that it is alway the case that $k > 0$. From Proposition 5, arbitrageurs' date-t capital in equilibrium is

$$k_t = k_0 \exp\left\{ \int_0^t \left[\frac{1}{2}\lambda_s^2 - \rho + S_{\mathcal{C}}(r_s,\, \boldsymbol{b}_s)\right] \mathrm{d}s + \int_0^t \lambda_s \mathrm{d}z_s \right\},$$

which is greater than zero as long as $k_0 > 0$ (by model assumption).

Now consider the closed ball $B(0,\, \|\boldsymbol{\gamma}_0\|_2/\gamma)$ in $\mathbb{R}^n$ and an arbitrary vector $\boldsymbol{b}$ in this ball. For any fixed r and $k > 0$, since

$$\begin{aligned}
\frac{\partial S_{\mathcal{C}}(r,\, \boldsymbol{b})}{\partial \boldsymbol{b}} &= \frac{\partial S_{\mathcal{C}}(r,\, \mathbf{0})}{\partial \boldsymbol{b}} + \frac{\partial^2 S_{\mathcal{C}}(r,\, \boldsymbol{b}^*)}{\partial \boldsymbol{b} \partial \boldsymbol{b}^\top} \boldsymbol{b} \\
&= \boldsymbol{\pi}(\mathbf{0}) + \boldsymbol{H}_{\mathcal{C}}(r,\, \boldsymbol{b}^*)\boldsymbol{b}, \\
&= \boldsymbol{H}_{\mathcal{C}}(r,\, \boldsymbol{b}^*)\boldsymbol{b} \qquad\qquad (\boldsymbol{\pi}(\mathbf{0}) = \mathbf{0} \text{ By Assumption 6})
\end{aligned}$$

for some $\boldsymbol{b}^*$ (as a function of $\boldsymbol{b}$) such that $\boldsymbol{b}^* \in B(0,\, \boldsymbol{b}) \subset B(0,\, \|\boldsymbol{\gamma}_0\|_2/\gamma)$, where $\boldsymbol{H}_{\mathcal{C}} = \partial^2 S_{\mathcal{C}}/\partial \boldsymbol{b} \partial \boldsymbol{b}^\top$ defines the Hessian matrix of $S_{\mathcal{C}}$, we have $\boldsymbol{b} = (\gamma + \boldsymbol{H}_{\mathcal{C}}(r,\, \boldsymbol{b}^*)k)^{-1} \boldsymbol{\gamma}_0$. Convexity of the support function $S_{\mathcal{C}}$ commands that $\boldsymbol{H}_{\mathcal{C}}$ is positive semi-definite everywhere. Let

$$\boldsymbol{H}_{\mathcal{C}}(r,\, \boldsymbol{b}^*) = \boldsymbol{\Gamma}^\top \mathrm{diag}(d_1, \ldots, d_n) \boldsymbol{\Gamma}, \quad d_1 \geq d_2 \geq \cdots \geq d_n \geq 0$$

be its eigen-decomposition (for $\boldsymbol{H}_{\mathcal{C}}(r,\, \boldsymbol{b}^*)$ is real-valued and symmetric, this decomposition must exist), then

$$(\gamma + \boldsymbol{H}_{\mathcal{C}}(r,\, \boldsymbol{b}^*)k)^{-1} = \boldsymbol{\Gamma} \mathrm{diag}\left(\frac{1}{\gamma + d_1 k}, \ldots, \frac{1}{\gamma + d_n k}\right) \boldsymbol{\Gamma}^\top$$

and

$$\|\, (\gamma + \boldsymbol{H}_{\mathcal{C}}(r,\, \boldsymbol{b}^*)k)^{-1} \boldsymbol{\gamma}_0 \|_2 \leq \frac{1}{\gamma + d_n k} \|\boldsymbol{\gamma}_0\|_2 \leq \frac{1}{\gamma} \|\boldsymbol{\gamma}_0\|_2.$$

Thus $(\gamma + \boldsymbol{H}_{\mathcal{C}}(r,\, \boldsymbol{b}^*(\boldsymbol{b}))k)^{-1} \boldsymbol{\gamma}_0$ is a continuous mapping from $B(0,\, \|\boldsymbol{\gamma}_0\|_2/\gamma)$ to itself. By Brouwer's fixed point theorem (Mas-Colell, Whinston, and Green, 1995, Theorem M.I.1, p. 952), the original equation, which finds the fixed point of this mapping, admits a solution.

Uniqueness of the solution is due to the fact that the Jocabian matrix of function $(\partial S_{\mathcal{C}}(r,\, \boldsymbol{b})/\partial \boldsymbol{b})k + \gamma \boldsymbol{b}$ is

$$J(r,\, \boldsymbol{b}) = H_{\mathcal{C}}(r,\, \boldsymbol{b}) + \gamma,$$

the determinant of which equals $\prod_{i=1}^{n}(d_i + \gamma) > 0$ everywhere. Thus, by the implicit function theorem (Mas-Colell, Whinston, and Green, 1995, Theorem M.E.1, p. 941-942), within the ball $B(0, \|\boldsymbol{\gamma}_0\|_2/\gamma)$, equation $(\partial S_{\mathcal{C}}(r, \boldsymbol{b})/\partial \boldsymbol{b})k + \gamma \boldsymbol{b} = \boldsymbol{\gamma}_0$ admits a unique solution.

Q.E.D.

I finish this section by proving Lemma 1. For $(\pi_0^*, \boldsymbol{\pi}^*) \in \mathcal{C}_t$ such that $S_{\mathcal{C}_t}(r, \boldsymbol{b}) = \pi_0^* r + \boldsymbol{\pi}^{*\top}\boldsymbol{b}$, we have that

$$S_{\mathcal{C}_t}(r, \boldsymbol{b}) = \pi_0^* r + \left(\frac{\boldsymbol{\pi}^*}{\alpha_t}\right)^{\top} (\alpha_t \boldsymbol{b}) \leq S_{\mathcal{C}_0}(r, \alpha_t \boldsymbol{b})$$

because $(\pi_{0t}^*, \boldsymbol{\pi}_t^*/\alpha_t) \in \mathcal{C}_0$. For any $(\pi_0, \boldsymbol{\pi}) \in \mathcal{C}_0$, since $(\pi_0, \alpha_t \boldsymbol{\pi}) \in \mathcal{C}_t$, it must be that

$$S_{\mathcal{C}_t}(r, \boldsymbol{b}) \geq \pi_0 r + (\alpha_t \boldsymbol{\pi})^{\top} \boldsymbol{b} = \pi_0 r + \boldsymbol{\pi}^{\top}(\alpha_t \boldsymbol{b}).$$

Since the inequality above holds for an arbitrary pair of $(\pi_0, \boldsymbol{\pi}) \in \mathcal{C}_0$, $S_{\mathcal{C}_t}(r, \boldsymbol{b}) \geq S_{\mathcal{C}_0}(r, \alpha_t \boldsymbol{b})$. Combining results above, $S_{\mathcal{C}_t}(r, \boldsymbol{b}) = S_{\mathcal{C}_0}(r, \alpha_t \boldsymbol{b})$, which is the conclusion of Lemma 1.

Q.E.D.

1.7.2 Algorithmic details for the first-step estimation

The statistical model is equivalent to

$$\mathbb{E}\left[y_t\right] = \boldsymbol{\lambda}^\top \boldsymbol{z}_t + S\left(x_t \alpha_t\right),$$

$$\alpha_t = \exp\left(\boldsymbol{\delta}^\top \boldsymbol{u}_t\right).$$

The semi-parametric nonlinear least square problem to solve is

$$\underset{\boldsymbol{\lambda}, \boldsymbol{\delta}, S(\cdot)}{\text{minimize}} \quad \sum_{t=1}^{T} \left[y_t - \boldsymbol{\lambda}^\top \boldsymbol{z}_t - S\left(x_t \alpha_t\right)\right]^2.$$

To start the algorithm, initialize $\boldsymbol{\delta}$ with a guess $\boldsymbol{\delta}^{(0)}$. At iteration i,

- Treating $\boldsymbol{\delta}^{(i)}$ as known, calculate $\alpha_t^{(i)}$. Then fit the semi-parametric model

$$\underset{\boldsymbol{\lambda}, S}{\text{minimize}} \quad \sum_{t=1}^{T} \left[y_t - \boldsymbol{\lambda}^\top \boldsymbol{z}_t - S\left(x_t \alpha_t^{(i)}\right)\right]^2$$

to find $\boldsymbol{\lambda}^{(i)}$, $S^{(i)}(\cdot)$ and the residuals $\varepsilon_t^{(i)}$;

- Define $\hat{y}_t^{(i)} = y_t - \boldsymbol{\lambda}^{(i)\top} \boldsymbol{z}_t$, solve the following problem

$$\underset{\boldsymbol{\delta}}{\text{minimize}} \quad L = \sum_{t=1}^{T} \left[\hat{y}_t^{(i)} - S^{(i)}\left(x_t \exp\left(\boldsymbol{\delta}^\top \boldsymbol{u}_t\right)\right)\right]^2.$$

Specifically, consider the Taylor expansion at $x_t \alpha_t^{(i)}$

$$\begin{aligned} L \approx & \sum_{t=1}^{T} \left[\hat{y}_t^{(i)} - S^{(i)}\left(x_t \alpha_t^{(i)}\right) - S^{(i)\prime}\left(x_t \alpha_t^{(i)}\right) x_t \left(\exp\left(\boldsymbol{\delta}^\top \boldsymbol{u}_t\right) - \alpha_t^{(i)}\right)\right]^2 \\ = & \sum_{t=1}^{T} \left[\hat{y}_t^{(i)} - S^{(i)}\left(x_t \alpha_t^{(i)}\right) + S^{(i)\prime}\left(x_t \alpha_t^{(i)}\right) x_t \alpha_t^{(i)} - S^{(i)\prime}\left(x_t \alpha_t^{(i)}\right) x_t \exp\left(\boldsymbol{\delta}^\top \boldsymbol{u}_t\right)\right]^2. \end{aligned}$$

Define

$$w_t = S^{(i)\prime}\left(x_t \alpha_t^{(i)}\right) x_t$$

$$\tilde{y}_t^{(i)} = \frac{1}{w_t} \underbrace{\left[\hat{y}_t^{(i)} - S^{(i)}\left(x_t \alpha_t^{(i)}\right)\right]}_{\varepsilon_t^{(i)}} + \alpha_t^{(i)},$$

the approximate problem becomes

$$\underset{\boldsymbol{\delta}}{\text{minimize}} \quad L = \sum_{t=1}^{T} w_t^2 \left[\tilde{y}_t^{(i)} - \exp(\boldsymbol{\delta}^\top \boldsymbol{u}_t)\right]^2.$$

which is a weighted nonlinear least square problem. Solve the problem to get $\boldsymbol{\delta}^{(i+1)}$.

- Start iteration $i + 1$

The algorithm iterates the above loop until convergence.

1.7.3 Microfoundation of (net) hedging demands

There are two countries, country d (domestic, the US) and f (foreign, the UK). Each country issues its own currency. We call the domestic (d) currency "dollar" and foreign (f) currency "pound". The exchange rate (pounds against dollars) at date t is E_t. In other words, one pound is exchanged for E_t amount of dollars (in practice, the GBP/USD currency pair). I assume that the dynamics of this exchange rate follows a Geometric Brownian motion:[37]

$$\frac{\mathrm{d}E}{E} = \mu_e \mathrm{d}t + \sigma_e \mathrm{d}z_e,$$

where μ_e measures the expected rate of appreciation for pounds, σ_e captures its volatility, the process $\{z^e\}$ is a standard Brownian motion.

There is a continuum of mass one identical hedgers in each country, namely d-hedgers and f-hedgers. Hedgers are exposed to currency risks due to their endowments abroad. Specifically, j-hedgers' ($j \in \{d,\, f\}$) endowments at time t is D_t^j from abroad, denominated in foreign currencies (i.e., D_t^f is denominated in dollars, and D_t^d is denominated in pounds). These endowments can be interpreted as cash flows from each country's Balance of Payments (BOP) items, such as export receivables, (changes in) direct or portfolio investment, as well as returns received from existing asset positions abroad. I assume that these endowments, denominated in dollars, satisfy multi-factor structures:

$$\begin{aligned} D^d E &= \boldsymbol{\lambda}_d^\top \boldsymbol{x} + \lambda_{d,0}, \\ D^f &= \boldsymbol{\lambda}_f^\top \boldsymbol{x} + \lambda_{f,0}, \end{aligned}$$

in which the vector of factors, denoted by $\boldsymbol{x}$, is a multivariate Itô process:

$$\mathrm{d}\boldsymbol{x} = \boldsymbol{\mu}(\boldsymbol{x})\mathrm{d}t + \boldsymbol{\sigma}(\boldsymbol{x})\mathrm{d}\boldsymbol{z}_x.$$

Elements in the vector $\{\boldsymbol{z}_x\}$ are standard Brownian motions. Functions $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are such that $\mathbb{E}\left[\mathrm{d}\boldsymbol{x}\right] = \boldsymbol{\mu}(\boldsymbol{x})\mathrm{d}t$ and $\mathbb{E}\left[\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{x}^\top\right] = \boldsymbol{\sigma}(\boldsymbol{x})\boldsymbol{\sigma}(\boldsymbol{x})^\top \mathrm{d}t$ are well defined in a complete probability space.

Hedgers of type j maximize mean-variance utilities over instantaneous wealth changes in $[t,\, t + \mathrm{d}t]$ denominated in their home currency

$$\mathbb{E}\left[\mathrm{d}W^j\right] - \frac{A^j}{2}\mathrm{var}\left[\mathrm{d}W^j\right], \quad j \in \{d,\, f\}, \tag{1.36}$$

where $\mathrm{d}W^j$ represents these instantaneous wealth changes. The parameter A^j captures the level of risk-aversion of the type-j hedgers who face the trade-off between the mean and variance of wealth change $\mathrm{d}W^j$. Agents optimizing (1.36) can be interpreted as overlapping generations who are born at date t, manage their wealth from t to $(t + \mathrm{d}t)$, consume everything and then die at time $(t + \mathrm{d}t)$. If their preferences over final consumptions are characterized by the von Neumann-Morgenstern expected utility $\mathbb{E}[u(\cdot)]$, the risk-aversion

[37] Again, all time subscripts are omitted whenever there is no confusion caused.

parameter A^j in problem (1.36) can be regarded as $A^j = -u''(W^j)/u'(W^j)$.[38] I assume that

$$A^d = \frac{A^f}{E} = A,$$

just to guarantee that the risk aversion parameters are invariant against the exchange rate.

In the baseline setting, $\mathrm{d}W^d = \mathrm{d}\left(D^d E\right)$ (d-hedgers in the US repatriating pound endowments) and $\mathrm{d}W^f = \mathrm{d}\left(D^f/E\right)$ (f-hedgers in the UK repatriating dollar endowments). Hedgers cannot manage their wealth changes from time t to $(t + \mathrm{d}t)$.

Beyond the baseline setting, I allow hedgers to alter their currency risk exposures using forward contracts.[39] These contracts, signed at time t, allow hedges to exchange F units of dollars for one pound at time $(t + \mathrm{d}t)$. Taking CIP deviations as given, the forward price F satisfy

$$F \exp\left(r^f \mathrm{d}t + b\mathrm{d}t\right) = E \exp\left(r^d \mathrm{d}t\right),$$

where r^f and r^d are pound and dollar risk-free rates. For d-hedgers managing their pound exposures, they can sign a forward contract exchanging h^d pounds for $h^d F$ dollars. As a result, their wealth changes is now

$$\begin{aligned}
\mathrm{d}W^d =& \left[h^d F + \left(D^d + \mathrm{d}D^d - h^d\right)(E + \mathrm{d}E)\right] - D^d E \\
=& h^d E \left(\frac{F}{E} - 1 - \frac{\mathrm{d}E}{E}\right) + \mathrm{d}\left(D^d E\right) \\
=& h^d E \left[\left(r^d - r^f - b - \mu_e\right)\mathrm{d}t - \sigma_e \mathrm{d}z_e\right] + \mathrm{d}\left(D^d E\right). \qquad (1.37)
\end{aligned}$$

Choosing h^d to maximize (1.36) for $j = d$ under (1.37) yields the following first-order condition:

$$E\left(r^d - r^f - b - \mu_e\right) - A^d \left\{h^d E^2 \sigma_e^2 - \frac{E\sigma_e}{\mathrm{d}t} \mathbb{E}\left[\mathrm{d}z_e \mathrm{d}\left(D^d E\right)\right]\right\} = 0,$$

[38] This is, of course, also due to the fact that shocks to hedgers' endowment factors $\boldsymbol{x}$ and spot exchange rate return $\mathrm{d}E/E$ are all Gaussian.

[39] In practice, other forward-like derivative contracts such as foreign exchange swaps (FX swaps in short) and cross-currency basis swaps (currency swaps in short) can serve similar purposes, though preferred by agents of different business models and cash flow durations.

from which we solve for h^d as

$$\begin{aligned}
h^d &= \frac{r^d - r^f - \mu_e}{EA^d\sigma_e^2} - \frac{b}{EA^d\sigma_e^2} + \frac{1}{E\sigma_e \mathrm{d}t}\mathbb{E}\left[\mathrm{d}z_e \mathrm{d}\left(D^d E\right)\right] \\
&= \frac{r^d - r^f - \mu_e}{EA\sigma_e^2} - \frac{b}{EA\sigma_e^2} + \frac{D^d}{\sigma_e^2 \mathrm{d}t}\mathbb{E}\left[\frac{\mathrm{d}E}{E}\left(\frac{\mathrm{d}D^d}{D^d} + \frac{\mathrm{d}E}{E} + \frac{\mathrm{d}D^d}{D^d}\frac{\mathrm{d}E}{E}\right)\right] \\
&= \frac{r^d - r^f - \mu_e}{EA\sigma_e^2} - \frac{b}{EA\sigma_e^2} + \frac{D^d}{\sigma_e^2 \mathrm{d}t}\mathbb{E}\left[\frac{\mathrm{d}E}{E}\frac{\mathrm{d}D^d}{D^d} + \sigma_e^2 \mathrm{d}t\right] \\
&= -\frac{\mu_e + r^f - r^d}{EA\sigma_e^2} - \frac{b}{EA\sigma_e^2} + \underbrace{\frac{\mathrm{cov}\left[\mathrm{d}E/E,\, \mathrm{d}D^d/D^d\right]}{\mathrm{var}\left[\mathrm{d}E/E\right]}}_{\beta_d} D^d + D^d. \qquad (1.38)
\end{aligned}$$

The equation above conveys straightforward intuitions. Consider $(D^d - h^d)$, which represents the d-hedgers' *unhedged* pound exposure. We can also treat the quantity as if it is a pure speculative position on GBP/USD. This term increases in $(\mu_e + r^f - r^d)$, which is the (expected) excess return from a GBP/USD carry trade (borrowing dollars, exchanging for pounds in spot markets, then lending pounds). This excess return over the variance (scaled by the risk-aversion parameter) is the canonical mean-variance portfolio demand.

The d-hedgers' pure (unhedged) pound exposure $(D^d - h^d)$ increases when β_d, the regression coefficient of d-hedgers' endowment growth rates on the currency returns, decreases. Lower β_d makes the exchange rate E itself a better hedge against a future drop in d-hedgers' endowments, thus incentives hedgers to take on more the currency risk.

Hedged position h^d decreases in the CIP deviations. Recall that h^d represents the quantity of pounds d-hedgers are selling forward. As higher b translates to relatively lower forward pound price: selling pounds for dollar forward becomes less favorable, thus a smaller hedged position.

Similarly, for f-hedgers hedging against USD/GBP exchange risk, they will sell h^f units of dollar for h^f/F units of pounds forward. The resulting wealth change is

$$\begin{aligned}
\mathrm{d}W^f =& \frac{h^f}{F} + \left(D^f + \mathrm{d}D^f - h_t^f\right)\left(\frac{1}{E} + \mathrm{d}\left(\frac{1}{E}\right)\right) - \frac{D^f}{E} \\
=& \frac{h^f}{E}\left[\frac{E}{F} - 1 - E\mathrm{d}\left(\frac{1}{E}\right)\right] + \mathrm{d}\left(\frac{D^f}{E}\right) \\
=& \frac{h^f}{E}\left[\left(r^f - r^d + b + \mu_e - \sigma_e^2\right)\mathrm{d}t + \sigma_e dz_e\right] + \mathrm{d}\left(\frac{D^f}{E}\right). \qquad (1.39)
\end{aligned}$$

f-hedgers will choose h^f to maximize (1.36) for $j = f$ under their budget constraint (1.39), will lead to the following first-order condition:

$$\frac{1}{E}\left(r^f - r^d + b + \mu_e - \sigma_e^2\right) - A^f\left\{\left(\frac{1}{E}\right)^2 h^f\sigma_e^2 + \frac{\sigma_e}{E\mathrm{d}t}\mathbb{E}\left[\mathrm{d}z_e\mathrm{d}\left(\frac{D^f}{E}\right)\right]\right\} = 0.$$

From the equation above, we can solve for h^f:

$$\begin{aligned}
h^f &= \frac{r^f - r^d + \mu_e - \sigma_e^2}{(A^f/E)\sigma_e^2} + \frac{b}{(A^f/E)\sigma_e^2} - \frac{E}{\sigma_e \mathrm{d}t}\mathbb{E}\left[\mathrm{d}z_e \mathrm{d}\left(\frac{D^f}{E}\right)\right] \\
&= \frac{r^f - r^d + \mu_e - \sigma_e^2}{A\sigma_e^2} + \frac{b}{A\sigma_e^2} - \frac{D^f}{\sigma_e^2 \mathrm{d}t}\mathbb{E}\left[\frac{\mathrm{d}E}{E}\left(\frac{\mathrm{d}D^f}{D^f} - \frac{\mathrm{d}E}{E} + \frac{\mathrm{d}E}{E}\frac{\mathrm{d}E}{E} - \frac{\mathrm{d}D^f}{D^f}\frac{\mathrm{d}E}{E}\right)\right] \\
&= \frac{r^f - r^d + \mu_e - \sigma_e^2}{A\sigma_e^2} + \frac{b}{A\sigma_e^2} - \frac{D^f}{\sigma_e^2 \mathrm{d}t}\mathbb{E}\left[\frac{\mathrm{d}E}{E}\frac{\mathrm{d}D^f}{D^f} - \sigma_e^2 \mathrm{d}t\right] \\
&= \frac{\mu_e + r^f - r^d}{A\sigma_e^2} - \frac{1}{A} + \frac{b}{A\sigma_e^2} - \underbrace{\frac{\operatorname{cov}\left[\mathrm{d}E/E,\, \mathrm{d}D^f/D^f\right]}{\operatorname{var}\left[\mathrm{d}E/E\right]}}_{\beta_f} D^f + D^f. \qquad (1.40)
\end{aligned}$$

The f-hedgers optimal choice of h^f delivers similar intuitions as the d-hedgers'. Now that h^f represents the amount of dollars f-hedgers are selling forward for pounds, thus a long position on pounds, it increases in the (expected) GBP/USD risk premium, and decreases in β_f as defined above (a higher β_f means pounds do not offer protection against f-hedgers' endowment risk).

Based on equation (1.38) and (1.40), we can calculate the net demand for dollars in forward markets, in dollar terms. Since d-hedgers sell h^d units of pounds for dollars and f-hedgers sell h^f units of dollar for pounds, the net forward dollar demand is

$$h^d E - h^f = \underbrace{-\frac{2(\mu_e + r^f - r^d)}{A\sigma_e^2} + \frac{1}{A} + D^d E(1+\beta_d) - D^f(1-\beta_f)}_{\gamma_0} - \underbrace{\frac{2}{A\sigma_e^2}}_{\gamma>0} b.$$

This expression agrees with the specification of hedgers' demand given in equation (1.5). Plugging in the assumptions that

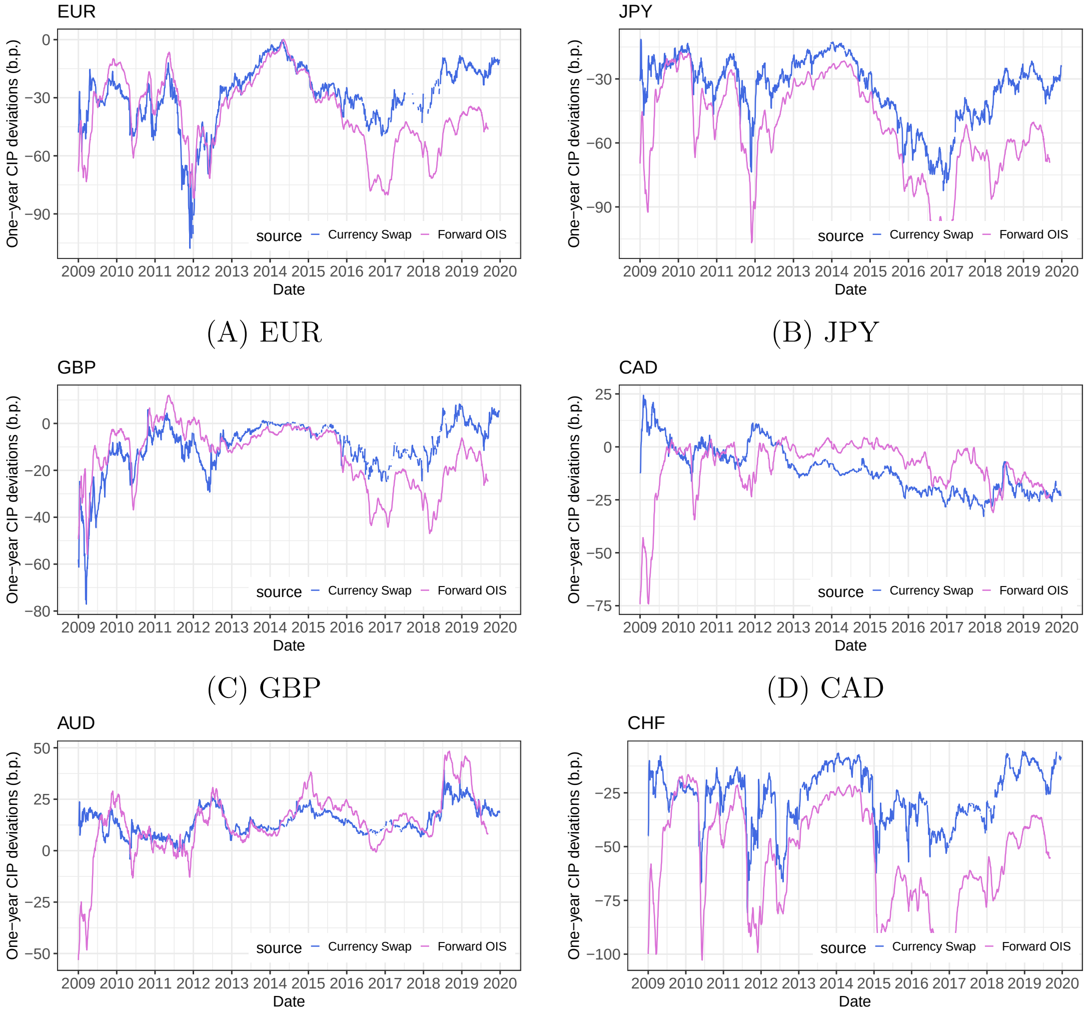
$$\begin{aligned}
D^d E &= \boldsymbol{\lambda}_d^\top \boldsymbol{x} + \lambda_{d,0}, \\
D^f &= \boldsymbol{\lambda}_f^\top \boldsymbol{x} + \lambda_{f,0},
\end{aligned}$$

we have

$$\begin{aligned}
&h^d E - h^f \\
&= \underbrace{\left[\boldsymbol{\lambda}_d^{(o)}(1+\beta_d) - \boldsymbol{\lambda}_f^{(o)}(1-\beta_f)\right]^\top}_{\boldsymbol{\beta}^\top} \boldsymbol{x}^{(o)} + \\
&= \underbrace{\left[\boldsymbol{\lambda}_d^{(u)}(1+\beta_d) - \boldsymbol{\lambda}_f^{(u)}(1-\beta_f)\right]^\top \boldsymbol{x}^{(u)} + \lambda_{d,0}(1+\beta_d) - \lambda_{f,0}(1-\beta_f) + \frac{1}{A} - \frac{2(\mu_e + r^f - r^d)}{A\sigma_e^2}}_{\ell} - \gamma b,
\end{aligned}$$

where $\boldsymbol{x}^{(o)}$ denotes observable components in $\boldsymbol{x}$ ($\boldsymbol{\lambda}_d^{(o)}$ and $\boldsymbol{\lambda}_f^{(o)}$ being loadings for the observable factors) and $\boldsymbol{x}^{(u)}$ denotes the unobservables ($\boldsymbol{\lambda}_d^{(u)}$ and $\boldsymbol{\lambda}_f^{(u)}$ being their loadings). Thus the net hedging demands have three parts: the linear combination of observable factors $\boldsymbol{\beta}^\top \boldsymbol{x}^{(o)}$, the latent unobservable demand ℓ, and the downward sloping response term $-\gamma b$.

1.7.4 Additional tables and figures



(A) EUR

(B) JPY

(C) GBP

(D) CAD

(E) AUD

(F) CHF

Figure 1.12: One-year CIP deviations for G6 currencies: currency swap rates and forward-OIS basis

Table 1.13: Dealer banks surveyed by foreign exchange committees, October 2004-April 2020

Australia and New Zealand Banking Group Limited	Bank of America Corporation
Bank of China	Bank of East Asia Limited
Bank of Montreal	Bank of New York Mellon Corporation
Bank of Nova Soctia	Barclays Plc
BNP Paribas SA	Canadian Imperial Bank of Commerce
China Bank of Communications	Citigroup Inc
Commerzbank AG	Commonwealth Bank of Australia
Crédit Agricole Corporate and Investment Bank	Credit Suisse Group AG
DBS bank Ltd	Deutsche Bank AG
Goldman Sachs Group Inc	Hang Seng Bank Limited
HSBC Holdings	Industrial and Commercial Bank of China
The ING Group	JP Morgan Chase & Co
Lloyds Banking Group Plc	Macquarie Bank Limited
Mizuho Bank Limited	Morgan Stanley
Mitsubishi UFJ Financial Group	National Australia Bank
National Bank of Canada	NatWest Group Plc
Nomura Holdings Inc	Oversea-Chinese Banking Corporation Limited
Resona Holdings Inc	Royal Bank of Canada
Shinsei Bank Limited	Skandinaviska Enskilda Banken AB
Société Générale SA	Standard Chartered Plc
State Street Corporation	Sumitomo Mitsui Financial Group Inc
Sumitomo Mitsui Trust Holdings Inc	Toronto-Dominion Bank
UBS AG	UniCredit SpA
United Overseas Bank Limited	Wells Fargo & Co
Westpac Banking Corporation	

Table 1.14: Predictive regressions: monthly returns of FX committee surveyed (FXS) dealer banks on one-year basis swap rates and placebo tests

This table presents results from the following linear regressions:

$$\frac{1}{\tau}\text{return}_{t+\tau} = \beta_0 + \beta|b_t| + \epsilon_{t+\tau},$$

for daily and monthly observations. The dependent variables are one-month-ahead value- or equal-weighted equity returns of 49 dealer banks surveyed by FX committees of New York, London, Tokyo, Toronto, Sydney, Singapore and Hong Kong. Additional placebo tests use returns from five ETFs tracking the S&P500 index (SPY), the global financial sector (IXG), the US financial sector (IYF), US broker-dealers and securities exchanges (IAI), and US insurance companies (KIE). For monthly observations, five hedge fund index returns are also included: one global composite index from BarclaysHedge (BCH), four indices from Hedge Fund Research (HFR) tracking global composite, relative value arbitrage, global-macro, and macro-currency strategies. All returns are net ones in percentage, as well as annualized (divided by $\tau = 1/12$ as shown in the regression specification). The independent variable $|b_t|$ is the cross-sectional average of absolute one-year basis swap rates for EUR, JPY, GBP, AUD, CAD, and CHF against the dollar. Sample periods begin from January 2009 and end at December 2019. Numbers in parentheses are Newey-West standard errors under automatic bandwidth selection.

Panel A: daily observations							
ret. (p.p.)	FXS (vw)	FXS (ew)	ETF-SPY (S&P500)	ETF-IXG (Gl. Fin.)	ETF-IYF (US Fin.)	ETF-IAI (US B&D)	ETF-KIE (US Insur.)
$\|b\|$ (b.p.)	2.17	2.03	0.41	1.85	1.18	1.30	1.11
	(0.82)	(0.81)	(0.41)	(0.78)	(0.70)	(0.80)	(0.73)
const.	−33.3	−31.0	7.8	−26.2	−8.4	−10.6	−4.7
	(15.2)	(15.1)	(8.9)	(14.7)	(13.5)	(17.0)	(14.5)
N obs.	2859	2859	2761	2761	2761	2761	2761
R^2-adj. (%)	3.5	3.4	0.3	2.8	1.4	1.6	1.1
Panel B: monthly observations							
ret. (p.p.)	FXS (vw)	FXS (ew)	ETF-SPY (S&P500)	ETF-IXG (Gl. Fin.)	ETF-IYF (US Fin.)	ETF-IAI (US B&D)	ETF-KIE (US Insur.)
$\|b\|$ (b.p.)	1.71	1.58	0.35	1.66	0.96	1.23	0.97
	(0.63)	(0.64)	(0.40)	(0.67)	(0.59)	(0.78)	(0.67)
const.	−23.4	−21.7	9.3	−22.1	−3.4	−8.6	−1.7
	(13.0)	(12.9)	(9.4)	(13.4)	(13.4)	(18.3)	(14.2)
N obs.	132	132	132	132	132	132	132
R^2-adj. (%)	1.8	1.7	−0.5	1.7	0.3	0.7	0.2
ret. (p.p.)			BCH (Gl. Com.)	HFR (Gl. Com.)	HFR (Re. Val.)	HFR (Macro)	HFR (Macro. Cur)
$\|b\|$ (b.p.)			0.13	0.08	0.09	−0.17	0.04
			(0.14)	(0.14)	(0.10)	(0.10)	(0.11)
const.			3.5	3.6	4.7	5.3	0.2
			(3.5)	(3.3)	(2.6)	(2.9)	(2.5)
N obs.			132	132	132	132	132
R^2-adj. (%)			−0.5	−0.6	−0.4	0.0	−0.7

Table 1.15: Predictive regressions: monthly returns of FX committee surveyed dealer banks on one-year basis swap rates adjusted by controls

This table presents results from the following linear regressions:

$$\frac{1}{\tau}\text{return}_{t+\tau} = \beta_0 + \beta \bar{b}_t + \phi \cdot \text{control}_t + \epsilon_{t+\tau}$$

for daily and monthly observations. The dependent variable is the one-month-ahead value-weighted equity return of 49 dealer banks surveyed by FX committees of New York, London, Tokyo, Toronto, Sydney, Singapore and Hong Kong. All returns are net ones in percentage, as well as annualized (divided by $\tau = 0.25$ as specified in the regression equation). The independent variable $\bar{b}_t$ is the cross-sectional average of absolute one-year basis swap rates for EUR, JPY, GBP, AUD, CAD, and CHF against the dollar. Control variables include the average smoothed earnings yield (E/P) and dividend yield (D/P) for the 49 dealer banks, the effective Fed fund rate (FFR), and the CBOE volatility index (VIX). Sample periods begin from January 2009 and end at December 2019. Numbers in parentheses are Newey-West standard errors under automatic bandwidth selection.

ret. (p.p.)	Daily observations			Monthly observations		
$\bar{b}$ (b.p.)	2.17	1.15	1.81	1.71	1.02	1.53
	(0.82)	(0.64)	(0.67)	(0.63)	(0.56)	(0.59)
E/P		14.7			10.8	
		(8.9)			(7.6)	
D/P			1.28			0.41
			(1.85)			(2.22)
FFR		5.03	0.26		3.17	0.20
		(6.99)	(6.27)		(6.59)	(6.92)
VIX		−0.44	1.95		−0.34	1.42
		(1.11)	(1.80)		(1.44)	(1.79)
const.	−33.3	−128.2	−65.1	−23.4	−93.4	−47.4
	(15.2)	(68.3)	(36.7)	(13.0)	(53.5)	(36.2)
N obs.	2859	2859	2859	132	132	132
R^2-adj. (%)	3.5	10.8	6.1	1.8	3.6	1.0

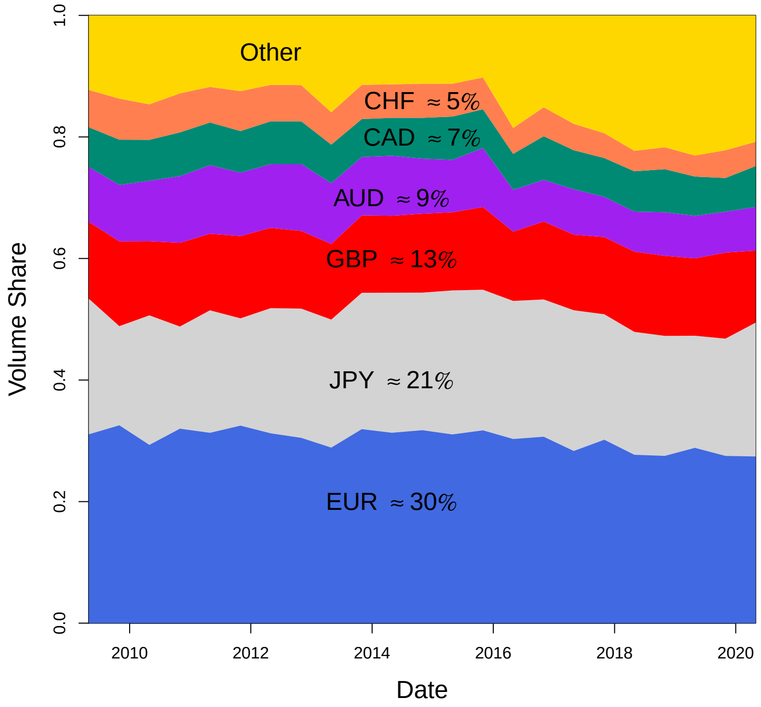


Figure 1.13: FX derivatives trading volume shares

Chapter 2

Option-Implied Bounds for the Crash Probability of a Stock

This chapter is joint work with Dr. Ian Martin.

Using option prices, we develop a theoretical framework to bound the expectation of a payoff that is contingent on the return of a stock. Both the lower and upper bounds are available from this framework and they are sharp in theory. We apply this framework to calculating forecasting bounds for stock crash probabilities. The crash probability bounds appear to be tight throughout empirical tests.[1] We show further that these bounds can forecast crash events out-of-sample, and they outperform combinations of various stock characteristics documented in the existing literature that are related to crash risk.[2]

The bounds are constructed from (and only from) current security prices, thus no historical data are needed. This feature enables real-time forecasting, which can offer timely insights into the downside risk of an individual stock. In theory, they can bound the probability of a single-stock crash event any time in the future (of course, in reality, their forecasting horizons are limited by maturities of option contracts). Their timeliness is only constrained by "freshness" of security prices.

Compared with interval forecasts generated from statistical procedures, these option-implied bounds are probabilistic with guaranteed coverage in theory. In other words, these forecasting bounds have 100% "confidence levels" in the sense of traditional statistical inference. As Keynes has noted in his 1921 book *A Treatise on Probability*: "*Many probabilities, which are incapable of numerical measurement, can be placed nevertheless between numerical limits.*" It is exactly the same notion that motivates these forecasting bounds with guaranteed coverage.

These forecasting bounds are robust in that they do not impose any distributional

[1] To clarify, sharpness refers to the result that the bounds cannot be further improved without leveraging additional data (for example, high-quality basket option prices in deep markets, which are not available in practice) or making additional assumptions (which can be fragile). Tightness describes the fact that both the lower and upper bounds are close to the true crash probability. The former is a theoretical property and the latter is an empirical phenomenon.

[2] See, for example, Chen, Hong, and Stein (2001); Boyer, Mitton, and Vorkink (2009) and Greenwood, Shleifer, and You (2019).

assumptions on the stock returns. Their sharpness remains when stock returns contain components like stochastic volatilities and jumps. This important feature is a strength of our methodology, since stock returns exhibit complicated time-varying distributional patterns (Andersen, Bollerslev, Diebold, and Ebens, 2001; Andersen, Benzoni, and Lund, 2002).

The methodological framework consists of two steps: recovering risk-neutral (marginal) distributions of stock and market returns from option prices first (applying the well-known approach of Breeden and Litzenberger (1978)), then bounding the physical (in contrast to the risk-neutral) expectations using theories of copula functions (namely, the Fréchet-Hoeffding bounds). In short, the method can be branded as "Breeden-Litzenberger meets Fréchet-Hoeffding".

The underlying theory guarantees sharpness of the proposed bounds. We demonstrate that, in order to get sharper results, one needs solid knowledge regarding risk-neutral correlations between single-stock and market returns, that is, the market prices of correlation risks. This type of knowledge is elusive. Practically, it is difficult to observe correlation prices that are not mired by microstructural issues, as no deep markets of rainbow options on both single-stock prices and the market index exist. Theoretically, it is dangerous to impose *ad-hoc* assumptions on correlation risks and their prices: the 2008 financial crisis offers hard lessons on the mispricing of correlation risks.

We compute the crash probability bounds for stocks belonging to the S&P 500 index. These bounds target on the probabilities of 5%, 10% and 20% crashes in one month, one quarter, six months and one year. They demonstrate significant variations across firms. For example, the time-series averaged probabilities of a 20% crash in one year can be as low as within [4.6%, 13.0%] for some firm, and as high as within [34.0%, 55.7%] for another firm.

These forecasting bounds also vary significantly across time. In Figure 2.1, we plot the monthly forecasting bounds for the probability of a crash that is worse than 20% over the one-year horizon for two companies: Cisco and AIG. Cisco had the largest market capitalization on March, 2000 during the dot-com bubble. According to Figure 2.1, its crash probabilities started climbing up during the second half of 2000 and peaked in early 2001, being over 35% to 50%. In reality, its market capitalization had dropped more than 70% by the end of 2001. The other company in Figure 2.1 is AIG, which was in the eye of the tornado during the subprime crisis. Its crash probability started surging rapidly from the late 2007, almost one year ahead of the paramount outbreak of the crisis in October, 2008. The crash probability peaked at over 55% to 80% during the crisis.

We examine the tightness of these bounds by regressing the future crash event indicators directly on the crash probability bounds.[3] If the bounds are tight, they should both be very close to the true crash probability. Since the crash probability is the expectation of the crash event indicator, tight bounds will deliver intercepts of zero and slopes of one

[3] A crash event is defined as the stock return being smaller than a certain threshold, say, the gross return of a stock being smaller than 0.80 is a 20% crash event. A crash event indicator is a binary variable that equals one if the crash does happen and zero otherwise.

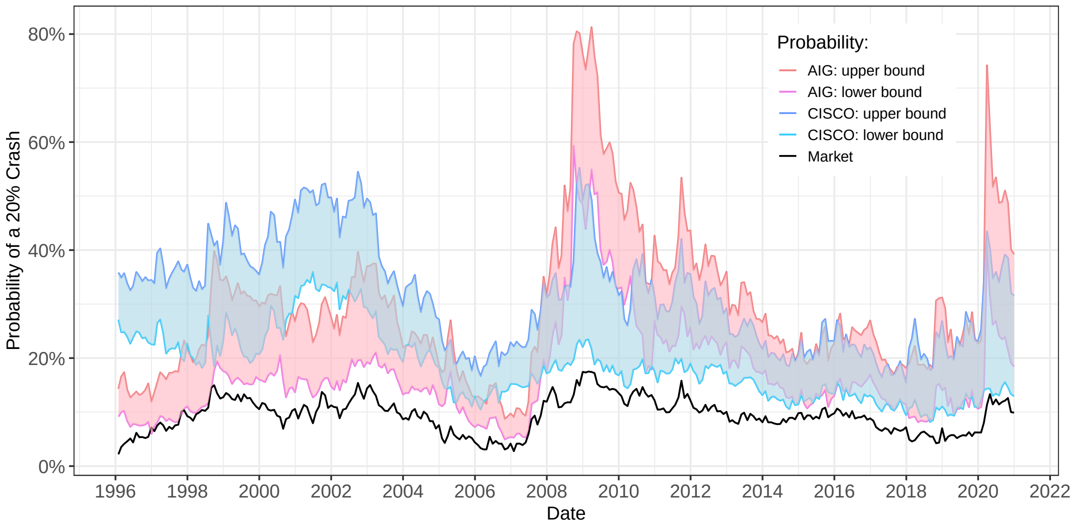


Figure 2.1: Monthly forward-looking probabilities of a crash (one-year net returns being less than -20%): Cisco stock, AIG stock, and the market (S&P 500). The forecasting bounds for stocks are based on the approach presented in this paper. The point forecast for the crash probability of the market is based on the approach in Martin (2017).

consistently throughout the regressions. These hypotheses are strongly supported by the data. Across different regression settings, the intercepts are mostly not significantly different from zero. On the contrary, the slopes are highly significant and, more importantly, they are extremely close to one, especially for the lower bound.

We then include stock characteristics, which have been reported in the previous literature to be related to crashes, into the regressions. The regression slopes for our bounds are still significantly different from zero, suggesting that they continue explaining variation in the crash probability. The slopes for the lower bound are still very close to one. The adjusted-R^2s in most of these regressions drop after including these characteristics. These evidence further suggest that the theory-motivated bounds drive out stock characteristics in terms of explaining variation in the crash probability.

Out-of-sample predictive performance of the bounds is evaluated against the combination of over ten stock characteristics. We design a procedure to emulate an avid "data-snooper" in order to compete against our single forecasting variable (either the lower or the upper bound). In doing so, we split the dataset into a training and a testing sample. The stock characteristics are combined through linear regressions, as well as logistic regressions, by fitting them to the training sample. In addition, when fitting these models, we add-in a "machine learning" flavor by using ℓ_1 penalty (as known as the LASSO in the statistics literature, see, for example, Tibshirani (1996)) to select "best" possible models through cross-validation (in the training sample). The predictive power of forecasting bounds is then compared with this pure data-mining procedure. The theory motivated bounds consistently outperform predictors extracted from stock characteristics across all forecasting horizons.

These crash probability bounds is then applied to study the declines in the equity of global systemically important banks (G-SIBs). We first compute the crash probability bounds for the majority of G-SIBs across different time. Then we aggregate these bounds across banks based on simple probability inequalities to created fragility and stability measures of the global banking system. These two measures help illustrate the many potential applications of these bounds in terms of creating macroeconomic indicators.

Related Literature. Asset pricing bounds have been derived for contingent claims in the context of incomplete markets or market imperfections (Cochrane and Saá-Requejo, 2000; Bernardo and Ledoit, 2000; Constantinides et al., 2008). Instead of bounding contingent claim prices that are determined by risk-neutral expectations, this paper establishes a new robust framework to compute bounds for physical expectations of contingent claims.

A large literature proposes methods to recover risk-neutral probabilities from option prices. An incomplete list includes Breeden and Litzenberger (1978); Jackwerth and Rubinstein (1996); Aït-Sahalia and Lo (1998); Rubinstein (1994). While the starting point of our derivation relies on the insights of Breeden and Litzenberger (1978), the major challenge of bounding the physical expectations are addressed by the new approaches introduced in this paper.

A few papers have attempted to forecast crashes in the stock market. Chen, Hong, and Stein (2001) use characteristics such as (detrended) trading volume and past returns to forecast negative skewness in the cross-section of individual stocks. Greenwood et al. (2019) use characteristics to forecast crashes at the industry level conditional on observing past price surge. Bates (1991) finds evidences from put option prices that forecast the stock market crash of 1987 (Black Monday). There is also a (downside/skewness/tail) risk literature that measures related objects as this paper does, but their main focus are how these risks manifest themselves into the cross-section of expected stock returns (see, for example, Ang, Chen, and Xing (2006); Boyer, Mitton, and Vorkink (2009); Kelly and Jiang (2014)). Martin (2017) has a section on recovering the physical crash probability of the market (which we have adapted to include in Figure 2.1), but going from the market crash to individual stock crash is not straightforward.

The rest of this paper is organized as follows. Section 2.1 introduces the underlying theory and method, as well as discusses theoretical properties of the bounds. Section **??** provides details of our data sample. Section 2.3 presents our empirical results. Section **??** concludes. All proofs for theoretical results are available in the Appendix.

2.1 Theory

2.1.1 Physical and risk-neutral expectations

To begin with, consider an investor who chooses to invest fully into the market of risky assets. The investor derives utility from her terminal wealth with a power utility function

$u(x) = u^{1-\gamma}/(1-\gamma)$. The investor's portfolio choice problem is then[4]

$$\underset{\boldsymbol{w}}{\text{maximize}} \quad \mathbb{E}\left[u\left(\boldsymbol{w}^{\top}\boldsymbol{R}\right)\right], \quad \sum\nolimits_{i}^{n} w_i = 1,$$

where the random vector $\boldsymbol{R} = [R_1, \ldots, R_n]^{\top}$ concatenates gross returns on all risky assets; the choice variables in $\boldsymbol{w} = [w_1, \ldots, w_n]^{\top}$ capture the investor's portfolio weights. The first-order conditions for this problem are

$$\mathbb{E}\left[\left(\boldsymbol{w}^{\star\top}\boldsymbol{R}\right)^{-\gamma} R_i\right] = \lambda \quad \text{for all } i,$$

where λ is a Lagrangian multiplier; the superscript * represents solutions to the optimization problem. By assumption, this investor chooses to invest fully in the market, thus the market return, denoted by R_m, is such that $R_m = \boldsymbol{w}^{\star\top}\boldsymbol{R}$. Plugging R_m into the first-oder conditions, a direct implication is that $R_m^{-\gamma}/\lambda$ is a stochastic discount factor (SDF) because

$$\mathbb{E}[MR_i] = 1 \quad \text{for all } i,$$

where

$$M = R_m^{-\gamma}/\lambda$$

is the SDF of this specific marginal investor.

Assuming that there is no arbitrage, for any random payoff of interest X, the *risk-neutral* expectation of X must satisfy the following equation

$$\frac{1}{R_f}\mathbb{E}^*[X] = \mathbb{E}[MX],$$

where R_f is the risk-free rate. The equation holds because both sides calculate today's price of a claim to the random payoff X.[5] This equation, combined with the SDF induced by the investor's marginal behavior, can be used to determine physical expectations for the random payoff of interest.[6] To be specific, as $M\lambda R_m^{\gamma} \equiv 1$,

$$\mathbb{E}[X] = \mathbb{E}[M\lambda R_m^{\gamma} X] = \lambda\mathbb{E}[M(R_m^{\gamma} X)] = \frac{\lambda}{R_f}\mathbb{E}^*[R_m^{\gamma} X].$$

Plugging in a special constant payoff $X = 1$ to the equation above,

$$1 = \frac{\lambda}{R_f}\mathbb{E}^*[R_m^{\gamma}].$$

[4] For clarity and the ease of exposition, we suppress unnecessary time subscripts because our theory and its derived methodologies are static in nature. It is worthwhile noting that all expectations can be regarded as conditional ones evolving in time.

[5] Here an implicit assumption is that the power-utility investor is marginal across all markets, including the option markets. In the mean time, she still chooses to hold the market portfolio according to our initial assumption.

[6] Strictly speaking, the expectation operator here is taken under the investor's subjective probabilities instead of the "objective" probabilities of an oracle.

Dividing the two equations we just have,

$$\mathbb{E}[X] = \frac{\mathbb{E}^*[R_m^\gamma X]}{\mathbb{E}^*[R_m^\gamma]}. \tag{2.1}$$

There are many potential applications of equation (2.1) in terms of characterizing the behaviors of individual stock returns. For example, for a given constant q, if we let $X = \boldsymbol{I}(R_i \leq q)$, where $\boldsymbol{I}(\cdot)$ equals one if the event in the parentheses is true and zero if not, equation (2.1) implies that

$$\mathbb{P}[R_i \leq q] = \frac{\mathbb{E}^*[R_m^\gamma \boldsymbol{I}(R_i \leq q)]}{\mathbb{E}^*[R_m^\gamma]}, \tag{2.2}$$

because $\mathbb{P}[R_i \leq q] = \mathbb{E}[\boldsymbol{I}(R_i \leq q)]$. Equation (2.2) shows that one can fully recover the physical probability distribution of a particular stock return, as perceived by the power-utility investor who is holding the market, from risk-neutral distributions. Physical distributions computed this way are forward-looking, which make them useful for real-time forecasting. A direct application is forecasting the crash probability of a stock, as we will demonstrate throughout the rest of this paper.

Another case is that one can simply let X be the return on a stock, that is, $X = R_i$ for some i, the expected return of this stock, $\mathbb{E}[R_i]$, can be given as

$$\mathbb{E}[R_i] = \frac{\mathbb{E}^*[R_m^\gamma R_i]}{\mathbb{E}^*[R_m^\gamma]}. \tag{2.3}$$

A special case of equation (2.3) is when the relative risk aversion parameter γ equals one. Since $\mathbb{E}^*[R_m] = R_f$, equations (2.3) becomes

$$\mathbb{E}[R_i] = \frac{1}{R_f}\mathbb{E}^*[R_m R_i], \tag{2.4}$$

where the market portfolio, R_m, by assumption, is equivalent to the optimal portfolio return for the log investor (because the investor *chooses* to invest fully in the market). This return is also called the growth-optimal (portfolio) return (Latane, 1959; Breiman, 1960). Martin and Wagner (2019), based on this special case of equation (2.4), derive an equation for the expected return of a stock in terms of the risk-neutral variances.

The assumptions and results above will hold through the rest of this paper.

2.1.2 Recovering the risk-neutral marginal distributions

Option prices carry rich information about the risk-neutral distributions of stock or market index returns. The static replication logic of Breeden and Litzenberger (1978) can recover these risk-neutral distributions from prices of options with various strike prices. We outline our implementation of the Breeden-Litzenberger method to begin the methodological framework. Let the constant S_0 be the price of the underlying (e.g., a stock or the market index) today and the random variable S be the underlying asset price at maturity.

Assume that the underlying asset does not pay dividends, then $S = RS_0$ where $R \sim Q$ is the gross return on the underlying asset; Q is the risk-neutral distribution of this gross return. Our goal is to recover Q from the prices of European options on this asset, which can be expressed as

$$\text{put}(K) = \frac{1}{R_f} \int_0^\infty \max(K - xS_0, 0)\, \mathrm{d}Q(x), \tag{2.5}$$

where put(K) denotes the price of a put option with a strike price K. Integrating (2.5) by parts yields: $\int_0^{\frac{K}{S_0}} Q(x)\, \mathrm{d}x = R_f \text{put}(K)/S_0$. Taking derivatives with regard to K,

$$Q\left(\frac{K}{S_0}\right) = R_f \text{put}'(K).$$

Combine the result above with the put-call parity, i.e., call(K) − put$(K) = S_0 - K/R_f$, where call(K) represents the corresponding call option price with a strike K,

$$Q\left(\frac{K}{S_0}\right) = R_f \text{call}'(K) + 1.$$

Throughout our execution, we only use the prices of out-of-the-money options, that is,

$$Q\left(\frac{K}{S_0}\right) = \begin{cases} R_f \text{put}'(K), & K \leq R_f S_0 \\ R_f \text{call}'(K) + 1, & K > R_f S_0 \end{cases}, \tag{2.6}$$

because these contracts are much more liquid ones.

In reality, prices of options are only available at a limited number of strikes. Differentiation might be inaccurate when available strikes (with observable option prices) are not dense enough. To overcome this difficulty, We fit nonparametric shape-constrained models for option prices as functions of their moneyness. This procedure has two benefits: 1) it rules out arbitrage across different strikes at a given maturity horizon; 2) it smoothly interpolates between strikes to enable stable numerical differentiation when implementing (2.6). Technical details of this approach are available in the Appendix.

2.1.3 Bounds on physical expectations from option prices

Denote by Q_{mi} the *joint* cumulative distribution function (CDF) of market and single stock returns (R_m, R_i) under the risk-neutral probability, and by Q_m, Q_i the *marginal* CDFs of R_m, R_i, also under the risk-neutral probability. Consider a very general specification for the payoff X: let $X = h(R_i)$, where $h : \mathbb{R}_+ \mapsto \mathbb{R}$ is an arbitrary continuous function.[7] That is, we would like to investigate the physical expectation of any well-behaved payoffs contingent on the return of a stock. Equation (2.1) can be rewritten more explicitly as

$$\mathbb{E}[h(R_i)] = \frac{\int x^\gamma h(y)\, \mathrm{d}Q_{mi}(x, y)}{\int x^\gamma\, \mathrm{d}Q_m(x)}. \tag{2.7}$$

[7] Strictly speaking, h is continuous almost everywhere.

Equation (2.7) simply tells us that if we can fully characterize the risk-neutral joint distribution Q_{mi},[8] we can evaluate the expectation of any well-behaved payoff contingent on the return R_i. As illustrated through equation (2.2) and (2.3), we can then calculate the forward-looking probability of a crash (let $h(R_i) = \boldsymbol{I}(R_i \leq q)$) or the expected return (let $h(R_i) = R_i$) of a stock.

As described in the earlier section, we can use prices of market index options to recover the risk-neutral marginal Q_m and prices of options for stock i to recover the risk-neutral marginal Q_i.

It is, however, almost impossible to fully characterize the risk-neutral *joint* distribution Q_{mi}. Widely traded index and equity options with liquid market places are almost exclusively written on one underlying asset. To recover the risk-neutral joint distribution, one needs to observe the prices of many options written on *both* the stock index and the stock of interest.[9] In addition, these options have to vary not only in terms of strike prices, but also in terms of their contingent payoff formats.[10] Needless to say, all these options have to be smoothly traded with high volume and deep market to avoid microstructural issues. In reality, these options are rare (if not nonexistent) and thinly traded, making it infeasible to back out the whole risk-neutral joint distribution of the market return and the stock return. This poses difficulties to the exact evaluation of $\mathbb{E}[h(R_i)]$ using equation (2.7).

Although no exact answer can be given to $\mathbb{E}[h(R_i)]$ under the current framework, sharp bounds can be obtained for it. Bounding this expectation term relies on dissecting the joint distribution into two parts: the marginals and the dependence structure. The marginals can both be regarded as known, as a result of applying the Breeden-Litzenberger approach to option prices. Then minimizing/maximizing across all possible dependence structures can bound the integral in the numerator of (2.7). To proceed and formalize these arguments, we will introduce some basic probability theories on *copula functions* first.[11]

Definition 3. *(A two-dimensional copula, or briefly, a copula) A two-dimensional copula is a function $C : [0,1]^2 \mapsto [0,1]$ with the following properties:*

1. *C is grounded: $C(x,0) = C(0,y) = 0$ for any (x,y) in its domain;*
2. *$C(x,1) = x$ and $C(1,y) = y$ for any (x,y) in its domain;*
3. *C is two-increasing: for all rectangles $B = [x_1, y_1] \times [x_2, y_2] \subset [0,1]^2$, the "volume" of B, which is defined by*

$$V_H(B) = C(x_2, y_2) - C(x_2, y_1) - C(x_1, y_2) + C(x_1, y_1)$$

is non-negative, that is, $V_H(B) \geq 0$.

[8] As a by-product, the risk-neutral marginal $Q_m(x) = \lim_{y\to\infty} Q_{mi}(x,y)$ will be known to us

[9] Ross (1976b) illustrates the use of numerous options to recover the risk-neutral joint densities; Martin (2018) refines Ross's argument and points out the limitations in practice.

[10] Martin (2018) illustrates the use of butterfly option strategies contingent on $(R_m + \alpha R_i)$ with finely varying α to recover the joint risk-neutral distribution in theory.

[11] See Nelsen (2007) for a theoretical monologue on this topic.

Copula functions play a central role in the distribution theory of multivariate random variables. They are the joint CDFs of two random variables whose marginals are both uniform on $[0, 1]$. In general, for any two random variables, treating their marginals as given, their joint distribution is uniquely determined by a copula function. This conclusion is due to a theorem by Sklar (1959).

Theorem 1. *(Sklar, 1959) Let Q be the joint CDF for the random vector (X, Y) with marginal CDFs F_X and F_Y. Then there exists a copula C, such that for all $x, y \in \mathbb{R}$,*

$$Q(x, y) = C(F_X(x),\, F_Y(y)).$$

The Sklar's theorem formalizes the idea of dissecting the dependence structure from the marginals. It applies to any joint distribution. With the two marginals fixed, any joint distribution is uniquely defined by a copula function, which "glues" together the two marginals. Based on the Sklar's theorem, there exists a copula function $C(\cdot, \cdot)$ such that the joint risk-neutral distribution between the market return and the stock return can be expressed as $Q_{mi} = C(Q_m(x), Q_i(y))$. A simple change of variable gives the following[12]

$$\int x^\gamma h(y)\, \mathrm{d}Q_{mi}(x, y) = \int_{[0,1]^2} \left[Q_m^{-1}(u)\right]^\gamma h\left(Q_i^{-1}(v)\right)\, \mathrm{d}C(u, v). \tag{2.8}$$

Now we can formalized the idea of bounding the term $\mathbb{E}[h(R_i)]$, which is equivalent to bounding the integral in (2.8), because the denominator in (2.7) can be treated as known (The marginal Q_m is recovered from the index option prices). Let $\mathcal{C}$ be the set of all two-dimensional copula functions, our bounds are defined as follows:

$$\min_{C\in\mathcal{C}} \int_{[0,1]^2} k\,(u, v)\, \mathrm{d}C(u, v) \le \int x^\gamma h(y)\, \mathrm{d}Q_{mi}(x, y) \le \max_{C\in\mathcal{C}} \int_{[0,1]^2} k\,(u, v)\, \mathrm{d}C(u, v), \tag{2.9}$$

where the integrand $k(u, v)$ is fully specified as

$$k\,(u, v) = \left[Q_m^{-1}(u)\right]^\gamma h\left(Q_i^{-1}(v)\right). \tag{2.10}$$

The integrand term absorbs all the information about the risk-neutral marginals, leaving the copula function the only unknown ingredient. This integrand is completely specified because the marginals Q_m and Q_i (as well as their inverses) are recovered from the option prices.

Calculating the two bounds in (2.9) is an optimization problem defined within a functional space (the space of all copula functions, as denoted by $\mathcal{C}$), which makes it difficult to solve. However, if the integrand $k(u,\ v)$, as defined in (2.10), is also two-increasing like

[12] Strictly speaking, the inverse notations for the two marginal CDFs (which might not be continuous) in (2.8) should be treated as generalized inverse distribution functions. That is, the notations Q_j^{-1} in equation (2.8) carry the following definition: $Q_j^{-1}(p) = \inf\{x \in \mathbb{R} : Q_j(x) \ge p\}$, for $j \in \{m, i\}$.

the copula function,[13] Corollary 2.2 of Tchen (1980) simplifies the problem as follows:

$$\min_{C\in\mathcal{C}} \int_{[0,1]^2} k\left(u,v\right)\, \mathrm{d}C(u,v) = \int_{[0,1]^2} k\left(u,v\right)\, \mathrm{d}\left(\min_{C\in\mathcal{C}} C(u,v)\right), \tag{2.11}$$

$$\max_{C\in\mathcal{C}} \int_{[0,1]^2} k\left(u,v\right)\, \mathrm{d}C(u,v) = \int_{[0,1]^2} k\left(u,v\right)\, \mathrm{d}\left(\max_{C\in\mathcal{C}} C(u,v)\right), \tag{2.12}$$

where $\min_{C\in\mathcal{C}} C(u,v)$ and $\max_{C\in\mathcal{C}} C(u,v)$ are defined *point-wise* for any $(u,v)\in[0,1]^2$. Sufficient conditions for this simplification to hold (i.e., for $k(u,v)$ to be two-increasing) is summarized in Assumption 9.

Assumption 9. *The payoff function h defined on $[0,\,\infty)$ satisfies the following two conditions:*

1. *it does not cross the x-axis, that is for any R and R' in $[0,\,\infty)$, $h(R)h(R')\geq 0$;*
2. *it is an increasing function.*

If we are interested in evaluating $\mathbb{E}[R_i]$, that is $h(x)=x$, Assumption 9 is satisfied. For crash probabilities, $h(x)=\boldsymbol{I}(x\leq q)$, which is decreasing. We can apply equations (2.11) and (2.12) to $-h$, for which Assumption 9 still holds. An exception arises when we are interested in log returns of stocks $\mathbb{E}[\log R_i]$. Since $h(x)=\log x$ always crosses the x-axis, violating Assumption 9, the simplification we presented does not apply. As a result, no analytical bounds are available for the expectation. In Section 2.5, we will provide a numerical solution to bounding $\mathbb{E}[h(R_i)]$ for any well behaved payoff function h.

Point-wise bounds for copulas appearing in equations (2.11) and (2.12) are characterized by the following theorem.

Theorem 2. *(Fréchet-Hoeffding theorem) If $C(u,v)$ is a copula, then*

$$\max(u+v-1,0)\leq C(u,v)\leq \min(u,v), \quad (u,v)\in[0,1]^2.$$

One can easily verify that both the lower bound $\max(u+v-1,0)$ and the upper bound $\min(u,v)$ are themselves copula functions according to Definition 3.

The Fréchet-Hoeffding theorem has broad implications, which is not limited to our specific application (bounding $\mathbb{E}[h(R_i)]$ according to equation (2.7)). Consider, for example, the problem of finding feasible domains of risk-neutral correlations between stock and market returns, $\mathrm{corr}^*[R_m,\,R_i]$, with known marginals Q_i and Q_m. This is equivalent to bounding $\mathbb{E}^*[R_mR_i]$: the lower bound is achieved when $\max(u+v-1,0)$ is the copula function connecting the two marginals; the upper bound is reached when the copula function is $\min(u,v)$.

As an illustration, in Figure 2.2, we plot the maximum and minimum achievable risk-neutral correlations between one-year returns of three stocks and the market. We selectively present three companies: CISCO, AIG (ones chosen for the motivating plot in Figure 2.1), as well as the cruise line company Carnival (demonstrating paramount crash risk

[13] See the definition of this concept in bullet point 3. of Definition 3.

during the COVID-19 period), all of which have long histories belonging to the S&P500 index. Risk-neutral marginals at each time point are fixed by option prices. We can see that $\mathrm{corr}^*[R_i,\, R_m]$ is always greater than minus one and smaller than one: marginal distributions contain information on dependence structures. With fixed marginals, the conventional wisdom that correlations can be *any* number between minus one and one does not hold. It is also worthwhile noticing that the possible range of risk-neutral correlations becomes much smaller when stock crash risks are mounting (e.g., around -0.6 to 0.4 for Carnival in the early stage of the pandemic).

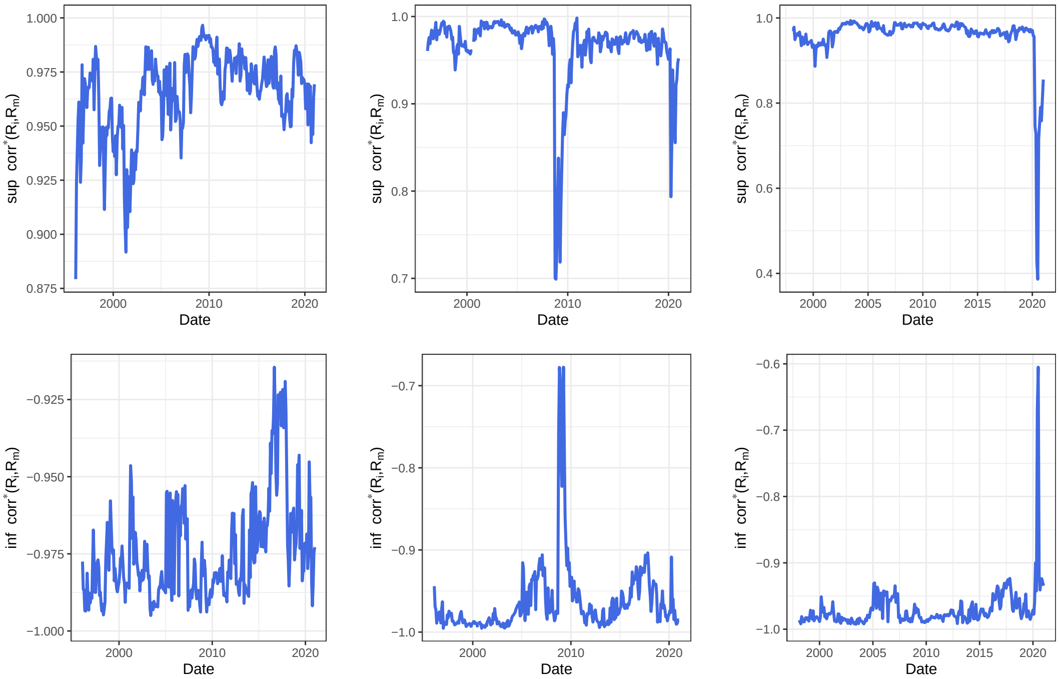


Figure 2.2: Bounds for the risk-neutral correlations between one-year returns $\mathrm{corr}^*[R_m,\, R_i]$: the case of CISCO, AIG, and Carnival (marginal distributions determined by option prices).

Substituting the Fréchet-Hoeffding lower bound into the right hand side of (2.11) and the upper bound into the right hand side of (2.12), we have the following result.

Proposition 9. *(bounds for general payoffs contingent on a stock return) Under Assumption 9,*

$$\frac{\mathbb{E}^*\left[R_m^\gamma h\left(Q_i^{-1}\left(1-Q_m(R_m)\right)\right)\right]}{\mathbb{E}^*\left[R_m^\gamma\right]} \leq \mathbb{E}[h(R_i)] \leq \frac{\mathbb{E}^*\left[R_m^\gamma h\left(Q_i^{-1}\left(Q_m(R_m)\right)\right)\right]}{\mathbb{E}^*\left[R_m^\gamma\right]}.$$

Bounds for $\mathbb{E}[h(R_i)]$ are sharp in the sense that

1. *the lower bound is achieved if $Q_i(R_i)+Q_m(R_m)=1$, that is, if the risk-neutral stock and market returns are countermonotonic; and*
2. *the upper bound is achieved if $Q_i(R_i)=Q_m(R_m)$, that is, if the risk-neutral stock and market returns are comonotonic*[14]*.*

[14] Two random variables are said to be countermonotonic if one is a monotonically decreasing transfor-

Result 9 lays foundations of our methodological framework. Sharpness of the bounds is due to the fact that both the lower and upper bounds for copula functions in the Fréchet-Hoeffding theorem are themselves copulas, summarized by conditions under which these bounds are achieved.

As our primary interest is on crash probabilities, we can let $h(x) = -\boldsymbol{I}(x \leq q)$ in Proposition 9 (the minus sign here is added to guarantee that h satisfies Assumption 9). Bounds for the crash probabilities of stocks are presented in Proposition 10.

Proposition 10. *(bounds for stock crash probabilities) We can bound the crash probability of a stock $\mathbb{P}[R_i \leq q]$ by*

$$\frac{\mathbb{E}^*\left[R_m^\gamma \boldsymbol{I}\left(R_m \leq q_l\right)\right]}{\mathbb{E}^*\left[R_m^\gamma\right]} \leq \mathbb{P}[R_i \leq q] \leq \frac{\mathbb{E}^*\left[R_m^\gamma \boldsymbol{I}\left(R_m \geq q_u\right)\right]}{\mathbb{E}^*\left[R_m^\gamma\right]},$$

where $q_l = Q_m^{-1}(Q_i(q))$ and $q_u = Q_m^{-1}(1 - Q_i(q))$ and

1. *the lower bound is achieved when the risk-neutral stock and market returns are comonotonic and the upper bound is achieved when the two risk-neutral returns are countermonotonic;*
2. *the lower bound is always smaller than the upper bound.*

As most stocks typically move with, rather than against, the market, we anticipate that comonotonicity is closer to the truth than countermonotonicity. Hence, a priori, we would expect that the lower bound is more likely to be tighter (i.e., closer to the "true" crash probability) than the upper bound. Our empirical results in Section 2.3.1 confirm this intuition, showing that the lower bounds do indeed track the forward-looking crash probabilities better in the panel of S&P 500 stocks.

Under our framework, calculating $\mathbb{E}^*\left[R_m^\gamma \boldsymbol{I}\left(R_i \leq q\right)\right] / \mathbb{E}^*\left[R_m^\gamma\right]$ gives us $\mathbb{P}[R_i \leq q]$, according to equation (2.2). This calculation is infeasible without knowing the price of correlation risks between the market and stock returns. Bounds provided in Proposition 10 sidestep this obstacle and provide approximate answers: Knowing q_l and q_u, knowledge regarding the (risk-neutral) market return distributions is sufficient for us to compute the bounds.

The two quantiles q_l and q_u define tail regions for calculating the two bounds. By definition, they are such that

$$\mathbb{P}^*[R_m \leq q_l] = \mathbb{P}^*[R_i \leq q] = \mathbb{P}^*[R_m \geq q_u].$$

Panel (A) of Figure 2.3 illustrates this point in detail. From prices of option on stock i, we know the risk-neutral marginal $Q_i(\cdot)$ and thus $\mathbb{P}^*[R_i \leq q]$, the risk-neutral crash probability of this stock. "Inverting" this probability with the marginal Q_m (available from market

mation of the other ($R_i = Q_i^{-1}(1 - Q_m(R_m))$ here as $Q_i^{-1}(1 - Q_m(\cdot))$ is a decreasing function); they are said to be comonotonic if one is a monotonically increasing transformation of the other ($R_i = Q_i^{-1}(Q_m(R_m))$ here as $Q_i^{-1}(Q_m(\cdot))$ is an increasing function.

index options) gives us the two quantiles q_l and q_u. Recall from equation (2.2), the lower and upper bounds in Proposition 10 are effectively $\mathbb{P}[R_m \leq q_l]$ and $\mathbb{P}[R_m \geq q_u]$, that is, under the "true" probability, q_l and q_u are such that

$$\mathbb{P}[R_m \leq q_l] \leq \mathbb{P}[R_i \leq q] \leq \mathbb{P}[R_m \geq q_u].$$

These inequalities are in stark contrast to the risk-neutral case under which the three probabilities are equal.

Panel (B) of the same figure further shows how we can find q_l and q_u directly from option prices. According to equation (2.6), slopes of single-stock put option prices (the black solid line) as a function of option strikes (the blue dotted curve) defines $Q_i(q)$. Shifting this line in parallel to the point that it becomes tangent to the *index put* options (also as a function of strike prices, shown as a blue solid curve) nails down q_l. Similarly, a red solid line tangent to *index call* options, which has an opposite slope $(-Q_i(q))$, defines q_u.

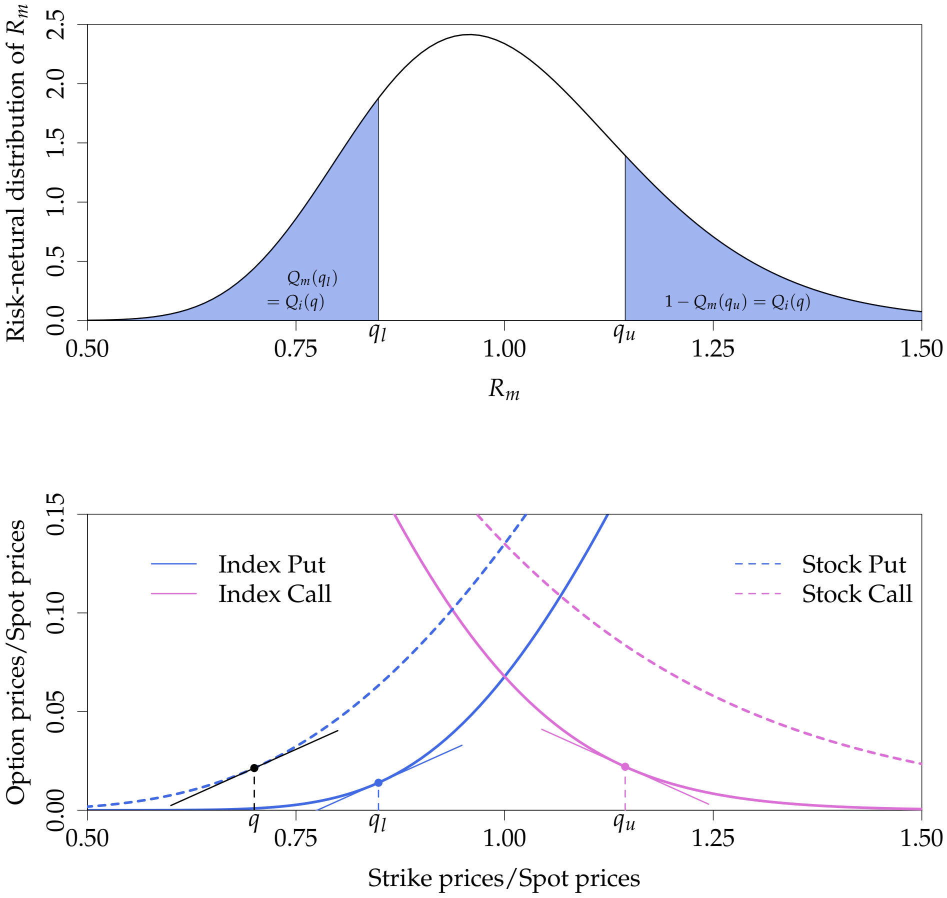


Figure 2.3: Defining the crash probability bounds: the two quantiles q_l and q_u in Proposition 10.

With q_l and q_u determined from our earlier discussions, now a natural follow-up question is: can the bounds in Proposition 10 relate directly to option prices? Corollary 1 expounds on this issue.

Corollary 1. *(crash probability bounds and the valuation of "power" contracts) If $\gamma = 1$, the lower bound in Proposition 10 can be written as*

$$\mathbb{P}[R_m \leq q_l] = q_l \left[\mathrm{put}'_m(K_l) - \frac{\mathrm{put}_m(K_l)}{K_l} \right]$$

and the upper bound can be written as

$$\mathbb{P}[R_m \geq q_u] = q_u \left[\frac{\mathrm{call}_m(K_u)}{K_u} - \mathrm{call}'_m(K_u) \right],$$

in which $K_l = q_l S_{m0}$, $K_u = q_u S_{m0}$, and S_{m0} is the spot price of the market index.

More generally, for any $\gamma > 0$, we can evaluate the three risk-neutral expectations in Proposition 10 as follows:

$$\mathbb{E}^*\left[R_m^\gamma\right] = R_f^\gamma + \frac{R_f}{S_{m0}^\gamma} \left[\int_0^F \gamma(\gamma-1) K^{\gamma-2} \mathrm{put}(K)\,\mathrm{d}K + \int_F^\infty \gamma(\gamma-1) K^{\gamma-2} \mathrm{call}(K)\,\mathrm{d}K \right]$$

$$\mathbb{E}^*\left[R_m^\gamma \boldsymbol{I}\left(R_m \leq q_l\right)\right] = R_f q_l^\gamma \left[\mathrm{put}'_m(K_l) - \gamma \frac{\mathrm{put}_m\left(K_l\right)}{K_l} \right] + \frac{R_f}{S_{m0}^\gamma} \int_0^{K_l} \gamma(\gamma-1) K^{\gamma-2} \mathrm{put}_m\left(K\right)\,\mathrm{d}K$$

$$\mathbb{E}^*\left[R_m^\gamma \boldsymbol{I}\left(R_m \geq q_u\right)\right] = R_f q_u^\gamma \left[\gamma \frac{\mathrm{call}_m(K_u)}{K_u} - \mathrm{call}'(K_u) \right] + \frac{R_f}{S_{m0}^\gamma} \int_{K_u}^{\infty} \gamma(\gamma-1) K^{\gamma-2} \mathrm{call}_m\left(K\right)\,\mathrm{d}K$$

where $F = R_f S_{m0}$ is the forward price of the market index.

Our evaluation of $\mathbb{E}\left[R_m^\gamma\right]$ is due to a well-known result in Carr and Madan (2001). Results for the case of $\gamma = 1$, that is, the log utility case are the same as findings in Martin (2017) for the tail probabilities of the market. The difference between gradients of option prices as a function of strikes (e.g., $\mathrm{put}'(K)$) and option prices divided by strikes (e.g., $\mathrm{put}'(K)/K$) reveals objective tail probabilities. The differences are determined by the convexity of option prices at the two critical strike prices, $K_l = q_l S_{m0}$ and $K_u = q_u S_{m0}$.

For an arbitrary γ, *all* options with strikes falling into the tail regions $[0,\, K_l]$ and $[K_u,\, \infty)$ matter, in addition to option price convexities at K_l and K_u. Integrals with respect to call and put options represent prices of option portfolios. When $\gamma > 1$, all else equal, higher out-of-the-money option prices translate into greater *physical* probability of crash, echoing the conventional wisdom. This is untrue when $0 < \gamma < 1$, that is, when $\gamma(\gamma - 1) < 0$.

To further understand the theoretical properties of bounds in Proposition 10, we now resort to a simple example, under which the risk-neutral distribution of $(R_i,\, R_m)$ is jointly log-normal.

Example 1. Assume that, under the risk-neutral measure,

$$\begin{bmatrix} \log R_i \\ \log R_m \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_i \\ \mu_m \end{bmatrix}, \begin{bmatrix} \sigma_m^2 & \rho\sigma_m\sigma_i \\ \rho\sigma_m\sigma_i & \sigma_i^2 \end{bmatrix} \right),$$

where $\mu_i + \frac{1}{2}\sigma_i^2 = \mu_m + \frac{1}{2}\sigma_m^2 = \log R_f$. Relevant results regarding this example are summarized in the following corollary.

Corollary 2. *Define $p^* = \mathbb{P}^*[R_i \leq q]$ and denote by $\Phi(\cdot)$ the CDF of standard normal distribution, our bounds under Proposition 10 for the log-normal example are*

$$\Phi\left(\Phi^{-1}(p_i^*) - \gamma\sigma_m\right) \leq \mathbb{P}[R_i \leq q] \leq \Phi\left(\Phi^{-1}(p_i^*) + \gamma\sigma_m\right). \tag{2.13}$$

The exact expression of $\mathbb{P}[R_i \leq q]$ according to equation (2.2) *for the example is*

$$\mathbb{P}[R_i \leq q] = \Phi\left(\Phi^{-1}(p_i^*) - \gamma\rho\sigma_m\right). \tag{2.14}$$

In Corollary 2, letting the correlation ρ be minus one or one in equation (2.14), we have the same bounds as ones derived directly from Proposition 10. As the domain for ρ being in $[-1, 1]$ is a result of the Cauchy-Schwartz inequality, one may consider bounding the crash probability using this inequality. Specifically, for any $h(\cdot)$, as $\mathbb{E}^*[R_m^\gamma h(R_i)] = \text{cov}^*[R_m^\gamma, h(R_i)] + \mathbb{E}^*[R_m^\gamma]\mathbb{E}^*[h(R_i)]$, knowing the risk-neutral marginals (and thus, all marginal moments), the expectation can then be bounded without resorting to Proposition 9 and 10, because $|\text{cov}^*[R_m^\gamma, h(R_i)]| \leq \sqrt{\text{var}^*[h(R_i)]\text{var}^*[R_m^\gamma]}$.

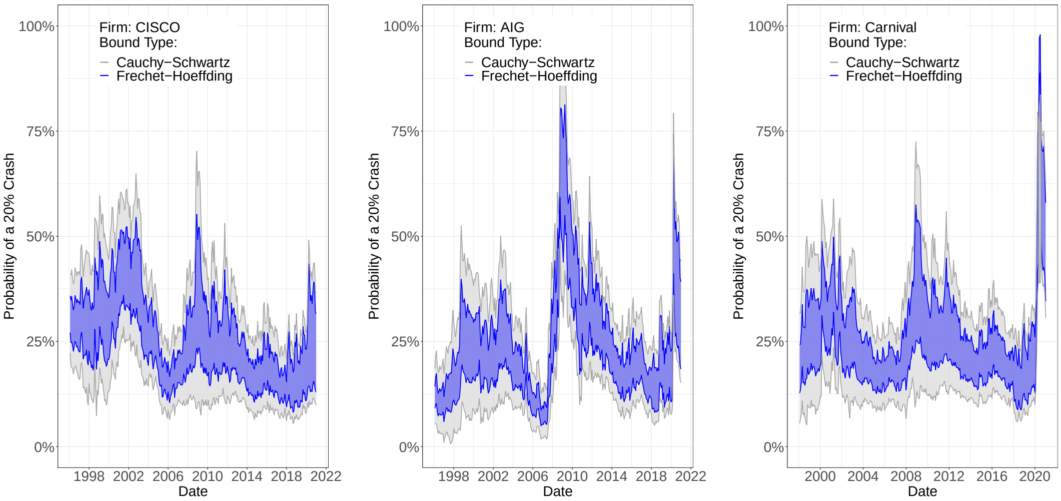


Figure 2.4: Comparing the proposed crash probability bounds with bounds derived from the Cauchy-Schwartz inequality: forward-looking crash probabilities of CISCO, AIG, and Carnival at one-year horizon (marginal distributions recovered from option prices).

Figure 2.4 compares the behaviors of our proposed bounds and the ones based on the Cauchy-Schwartz inequality, with marginal distributions recovered from market prices of options. We selectively present three companies: CISCO, AIG (ones chosen for the motivating plot in Figure 2.1), as well as the cruise line company Carnival (demonstrating paramount crash risk during the COVID-19 period), all of which have long histories belonging to the S&P500 index. Overall, widths of our bounds are around 40-50% narrower than those computed using the Cauchy-Schwartz inequality.

The example helps illustrate two variables that determine the widths of our bounds. The first is the risk-aversion parameter γ. We can see its influences clearly from inequalities in (2.13) of the example: increased γ widens the bounds. We summarize general theoretical properties regarding this parameter in Corollary 3.

Corollary 3. *The effects of risk-aversion parameter γ on the crash probability bounds in Proposition 10 are summarized as follows:*

1. *when $\gamma = 0$, the lower and the upper bounds agree and they both equal $\mathbb{P}^*[R_i \leq q]$, the risk-neutral crash probability;*
2. *for any $0 < \gamma < \infty$, as γ increases, the lower bound decreases and the upper bound increases;*
3. *as $\gamma \to \infty$, for any q such that $0 < Q_i(q) < 1$, the lower bound goes to 0 and the upper bound goes to 1, the option-implied bounds become trivial.*

The second variable determining widths of our bounds is the price of market volatilities (i.e., the risk-neutral variance of market returns). According to the log-normal example, it appears that our bounds tend to become wider when the market volatility is higher. As high market volatility regimes quest for more accurate answers to crash probabilities, wider bounds during these periods pose serious concerns regarding the usefulness of our framework. However, this is not what we observe empirically: our bounds narrow when (if not before) crisis emerges (as can be seen from Figure 2.1), providing more definitive guidance to the forward-looking crash probabilities.

The reason behind the above phenomenon is that, in addition to the risk-aversion parameter and prices of market volatilities, the *range* of (risk-neutral) correlations between R_i and R_m, $\text{corr}^*[R_i, R_m]$, also determines the widths of our bounds. The conventional wisdom is that $\text{corr}^*[R_i, R_m]$ can take any value from minus one to one, which simply reiterates the Cauchy-Schwartz inequality. This is not true once marginals of (R_i, R_m) are taken as given (from option prices). The fallacy directly points out the slackness of bounds derived using the Cauchy-Schwartz inequality. On the contrary, our proposed bounds are sharp as we account for the implications of marginal distributions on the dependence structure using the Fréchet-Hoeffding inequality.

To sum up, bounds derived using our approach are sharp without additional data (e.g., rainbow options on *both* individual stocks and the market) or assumptions (regarding the dependence structure). They have the potential to be useful forward-looking measures of single-stock crash probabilities. We now turn to the data to evaluate their performance.

2.2 Data

We focus on firms included in the S&P 500 index. The index constituent data are from Compustat. S&P 500 index options and equity options data are from OptionMetrics. The sample is monthly from January 1996 to December 2020. For the last trading day of each month, we query OptionMetrics for the volatility surfaces of equity options. Firms under consideration are ones that have been included into the S&P 500 index before the start of the calendar year. The monthly volatility surface data of S&P 500 index option (SPX) are also collected. We keep the volatility surface data with time-to-maturity of one month, three months, six months and one year. Risk-free rates are linearly interpolated from the yield curves, provided also by OptionMetrics. Prices, returns, share volumes and shares outstanding of individual stocks are monthly from CRSP[15]. All other firm characteristics that are used in the remaining parts of this paper come from Compustat. Table 2.1 provides summary statistics of the sample counts.

[Table 2.1 about here]

Additional concerns may rise because 1) equity pays dividends and 2) equity options are American options. Our choice of using volatility surface data helps mitigate these concerns. For all options (including the S&P 500 index option), OptionMetrics accommodates the dividend effects by projecting dividend yields from historical dividend records. For American options of single stocks, OptionMetrics computes the implied volatilities through a proprietary binomial tree algorithm which already accounts for the early-exercise premia. Given the relatively short maturity horizon under consideration and the use of only out-of-money options, the European and American implied volatility tend to be close. Thus, we take the volatility surfaces as measures of European implied volatility, following Carr and Wu (2009) and Martin and Wagner (2019).

When recovering risk-neutral marginals according to Section 2.1.2, the lack of observable deep out-of-the-money options prices leaves the tail behaviors undetermined. We extrapolate a flat volatility smile outside the range of observed strikes following existing literature (e.g., Carr and Wu (2009) for index options). Theoretical literatures on the asymptotic behavior of the volatility surface provide guidance on this approach (e.g., Hodges (1996); Rogers and Tehranchi (2010)). In light of these theories, we tend to assign heavier tails to the risk-neutral marginals.

2.3 Empirical tests

2.3.1 In-sample tests

At the end of each month, we compute the forward-looking bounds for the probabilities of (gross) stock returns being less than $q = 80\%$, 90% or 95%. The forecasting horizons under consideration are $\tau = 1, 3, 6$ or 12 months. For stock i, we denote by $\text{ProbLower}_{i,t}(\tau, q)$ the lower bound and by $\text{ProbUpper}_{i,t}(\tau, q)$ the upper bound, both forecasted over horizon τ conditioning on information at time t. The risk-aversion parameter is fixed as 1.5 throughout our execution.[16]

[15] All variables are calculated (or derived) based on data from database name ©CRSP Daily Stock, Center for Research in Security Prices (CRSP®), The University of Chicago Booth School of Business.

[16] The main regression results and out-of-sample forecasts are robust against any specification from one to three. If the risk-aversion parameter is too large, the bounds become loose, according to the theoretical property in Result 3.

We report in Table 2.2 and 2.3 the summary statistics of these forecasting bounds.

[Table 2.2 and 2.3 about here]

Since the bounds are computed for each firm at the end of each month, the two tables focus on different aspects. Table 2.2 summarizes variation across firms by first averaging the bounds for each firm along the time dimension. Table 2.3 summarizes variation across time by first averaging the bounds for each month across all firms. Results in these tables suggest pronounced variation both across time and cross-sectionally. Across all three cases, there are more cross-sectional variation than time-series variation on average, especially for lower bound.

Denote by $R_{i,t\to t+\tau}$ the τ period ahead gross returns when the current time is t. To test whether these forecasting bounds are tight or not, we first run linear regressions with the following specification: for the lower bound

$$\boldsymbol{I}(R_{i,t\to t+\tau} \leq q) = \alpha + \beta \text{ ProbLower}_{i,t}(\tau, q) + \varepsilon_{i,t+\tau},$$

or for the upper bound

$$\boldsymbol{I}(R_{i,t\to t+\tau} \leq q) = \alpha + \beta \text{ ProbUpper}_{i,t}(\tau, q) + \varepsilon_{i,t+\tau},$$

with $q = 0.80, 0.90$ and 0.95. Notice that the expectations of $\boldsymbol{I}(R_{i,t\to t+\tau} \leq q)$, that is, $\mathbb{E}[\boldsymbol{I}(R_{i,t\to t+\tau} \leq q)]$ should be the true crash probability $\mathbb{P}[R_{i,t\to t+\tau} \leq q]$.[17] If the bounds are tight, which means that they are close to the true crash probability, the parameter β should be close to one and α should be approximately zero.

The regression results are shown in Table 2.4. The standard errors in parentheses are two-way cluster following Petersen (2009). The standard errors in square brackets are from block bootstrap procedures according to Martin and Wagner (2019) with 2500 simulations. The intercept parameters are mostly not significantly different from zero, specially over longer maturity horizons. The slope parameters are always significant, meaning that the bounds can explain variation in the crash probability. Most importantly, the estimates agree reasonably well with the a prior belief that $\beta \approx 1$ and $\alpha \approx 0$. The slope coefficients for the lower bounds almost equal one perfectly for all four maturity horizons. The same coefficients are around 0.7 to 0.9 when the forecasting horizon is one month for the upper bound. The smallest case is when the forecasting horizon is one year, under which the slope coefficients for the upper bound are around 0.5 to 0.6. These slightly smaller regression coefficients suggest that the upper bound might not be as tight as the lower bound: our prior would be that the price of correlation risks between the stock and the market returns tend to be positive (positive risk-neutral correlations).

[Table 2.4 about here]

[17] Both the probability and the expectation here should be conditional. We omit the time subscript for simplicity.

Next, we adjust for stock characteristics in the linear regression tests. Eleven characteristics are under consideration, which are rolling-window stock beta (five-year window), momentum (past twelve month return excluding the last month), (log) size, book-to-market ratio, past volatility (last one month), gross profitability over book asset, leverage (debt-to-asset), solvency (cash or cash equivalent to current liability), turnover, detrended turnover, and short interest. These characteristics are ones that have been reported to be related to expected returns (Fama and French, 1993; Jegadeesh and Titman, 1993) and crashes (Chen et al., 2001; Greenwood et al., 2019). Table 2.5, 2.6 and 2.7 report the regression outcomes for a crash of over 20%, 10% and 5% separately. Through these tests, the crash probability bounds stay consistently significant. Most importantly, these multivariate regressions including stock characteristics are have smaller adjusted R^2 compared with the univariate case including only crash probability bounds. These evidences suggest that the option-implied bounds drive out characteristics in terms of explaining variation in the crash probabilities.

[Table 2.5, 2.6 and 2.7 about here]

2.3.2 Out-of-sample forecasts

The crash probability bounds rely on no free parameters, as they are observable in real time together with the security prices. This feature makes them natural candidates for out-of-sample forecasting. In this section, we demonstrate that they do perform well forecasting crash events at the stock level, which can be done by simply thresholding the forward-looking crash probability bounds.

To pose some real challenges, we design a procedure to emulate an avid “data-snooper”. In doing so, we split the dataset into a training and a testing sample. The stock characteristics are combined through linear regressions, as well as logistic regressions, by fitting them to the training sample. In addition, when fitting these models, we select the “best” possible models through cross-validation using the LASSO.

The predictive power of forecasting bounds is then compared with the pure data-mining procedure. The forecasting target is the events of individual stock crashes. The performance measure is the ROC curve and the area under the curve (AUC) measure, which balance the type-I and type-II forecasting errors.

We report in Table 2.8 the AUC statistics for the option-implied bounds and the statistical procedures. The results in Table 2.8 confirm that simply thresholding the option-implied crash probability bounds outperform the statistical procedure combining various stock characteristics. The lower bounds consistently dominate the statistical procedures. For all approaches, the forecasting power in general is more pronounced when q becomes smaller (AUC becomes larger). Out-of-sample forecasting accuracy in general declines as the forecasting horizon becomes longer. To visualize the performances of different approach, we also plot the ROC curves for a crash over 20% in Figure 2.5.

[Table 2.8 about here]

[Figure 2.5 about here]

2.4 Application

2.4.1 Fragility and stability measures of the global banking system

Baron, Verner, and Xiong (2020) report the link between the large declines in bank equity and macroeconomic downturn, as well as the predicative power of large bank equity declines on banking crisis. Since the bank equity returns can only be calculated *ex post*, the forward-looking equity crash bounds that we have introduced in this paper is a clear *ex ante* alternative to consider.

In this section, we demonstrate a specific application of the derived bounds to monitoring the crash probability of global systemically important banks (G-SIBs), and construct two global banking fragility measures from the crash probability bounds.

The G-SIBs under consideration are listed in Table 2.9. Twenty-one G-SIBs that have been included by the Financial Stability Board since 2011 are considered. These chosen G-SIBs all have their stocks traded in the US stock market or have issued American depositary receipts. Banks that are not traded in the US stock market are not considered in this study. Table 2.9 also reports the time periods during which option data are available for the equity of these banks.

We compute the forward-looking crash probability bounds for the equity value of these global banks using the method described in Section 2.1. Then we define two aggregate measures using these bounds for individual banks to monitor their overall fragility. For a given set of banks, the probability of *at least* one crash can be bounded from below as follows:

$$\begin{aligned}\mathbb{P}\left[\text{at least one crash}\right] &= \mathbb{P}\left[\cup_i\{R_i \leq q\}\right] \\ &\geq \max_i \mathbb{P}[R_i \leq q] \\ &\geq \max_i \inf \mathbb{P}[R_i \leq q],\end{aligned}$$

where q is a pre-specified return level (for example, 80%) to define a crash (large equity decline), $\inf \mathbb{P}[R_i \leq q]$ is the lower bound for the crash probability derived from option data for individual bank i. We define the quantity, $\max_i \inf \mathbb{P}[R_i \leq q]$, as the *fragility* measure: the higher this measure is, the more likely that a banking crisis will emerge. On the other hand, for the same set of banks, the probability that all of them are facing large equity value declines is bounded from above because

$$\begin{aligned}\mathbb{P}\left[\text{all crash}\right] &= \mathbb{P}\left[\cap_i \{R_i \leq q\}\right] \\ &\leq \min_i \mathbb{P}[R_i \leq q] \\ &\leq \min_i \sup \mathbb{P}[R_i \leq q],\end{aligned}$$

then the probability of *no* system-wide crash is bounded from below:

$$1 - \mathbb{P}\left[\text{all crash}\right] \geq 1 - \min_i \sup \mathbb{P}[R_i \leq q],$$

where $\sup \mathbb{P}[R_i \leq q]$ is the upper bound for the crash of bank i. We define $1 - \min_i \sup \mathbb{P}[R_i \leq q]$ as a measure of *stability*: if this quantity is large, the probability of a full-scale melt down of the global banking system will be small.

We show in Figure 2.6 these two measures over the one-year horizon. The crash events is defined by specifying q as 0.80 (a 20% crash), 0.70 (a 30% crash) and 0.60 (a 40% crash). The time scale is monthly from January 1996 to December 2017. The two measures move in opposite direction as expected. The fragility measure surges and the stability measure plunges during the period of subprime crisis. The stability measure *begins* to decline from mid 2007, which predates the onset of the crisis.

[Figure 2.6 about here]

2.5 Bounds for general contingent payoffs

For a general function h, there is no guarantee that $x^\gamma h(y)$ is two-increasing, which prevents us from using the Fréchet and Hoeffding bounds for solving the optimization problem defined in (2.9).

Hofer and Iacò (2014) propose an algorithm to solve this problem. For any well-behaved function k in (2.9),[18]

$$\max_{C \in \mathcal{C}} \int_{[0,1]^2} k\left(u, v\right) \, \mathrm{d}C(u, v) \approx \max_{\pi \in \mathcal{P}_n} \frac{1}{n} \sum_{i=1}^{n} k\left(\frac{i}{n+1}, \frac{\pi(i)}{n+1}\right), \quad \text{as } n \text{ is large enough,} \tag{2.15}$$

where $\mathcal{P}_n$ is the set of all possible permutations (the total number of which is $n!$) of the set $\{1, \ldots, n\}$; $\pi \in \mathcal{P}_n$ selects one specific permutation of $\{1, \ldots, n\}$, that is, π is a one-to-one mapping from the n-elements to themselves. The right-hand side of (2.15) is a canonical problem in combinatorial optimization called the linear assignment problem.[19] Algorithmic solution to this problem is due to Kuhn (1955), which is called the Hungarian algorithm. This algorithm reduces the complexity of solving the right-hand side optimization problem in (2.15) from $O(n!)$ (brute-force searching) to $O(n^3)$.

[18] Here, it means that: 1) the function guarantees that the integral is finite; 2) the function is continuous almost everywhere within its domain.

[19] See Burkard et al. (2012) for an extensive coverage on this topic

In terms of the lower bounds, one can apply the Hungarian algorithm to the integral involving $-k(u,v)$ to get an upper bound. The negative of the upper bound for $\int_{[0,1]^2}[-k(u,v)]\,\mathrm{d}C(u,v)$ minimizes the original integral over the space of copula functions.

Applying the Sklar's theorem and the methods presented above gives us the following result:

Proposition 11. *Let h be a function that is continuous almost everywhere. Define function $k(u,v)$ on $[0,1]^2$ as in equation (2.10). Let $\pi_{\min}$ be a permutation of $\{1,\dots,n\}$ that minimizes $\sum_{k=1}^{n} k\left(\frac{i}{n+1},\frac{\pi(i)}{n+1}\right)$, and $\pi_{\max}$ be a permutation of $\{1,\dots,n\}$ that maximizes $\sum_{k=1}^{n} k\left(\frac{i}{n+1},\frac{\pi(i)}{n+1}\right)$, then*

$$\frac{1}{nC}\sum_{i=1}^{n+1} k\left(\frac{i}{n+1},\frac{\pi_{\min(i)}}{n+1}\right) \le \mathbb{E}[h(R_i)] \le \frac{1}{nC}\sum_{i=1}^{n+1} k\left(\frac{i}{n+1},\frac{\pi_{\max(i)}}{n+1}\right),$$

holds approximately for n being large enough, where the constant C equals $\int_0^1 \left[Q_m^{-1}(u)\right]^\gamma \,\mathrm{d}u$.

Proposition 11 is valid for a very general class of functions. One simple but important example is when $h(R_i) = \boldsymbol{I}(q_1 \le R_i \le q_2)$, that is, when we are interested in evaluating the probability that stock i's return falls in the interval $[q_1, q_2]$.

2.6 Conclusion

This paper has proposed a new framework to derived bounds for the expectation of a payoff that is contingent on an individual stock return. The bounds are computed directly from option prices and are forward-looking by nature. The sharpness of these bounds are theoretically guaranteed. The framework is general enough and may be of interest by itself.

Applying this framework, we compute bounds for the stock crash probabilities. Through panel regressions, we show that these crash probability bounds are close to the true crash probabilities. Out-of-sample analysis shows that they perform well in forecasting crash events and consistently outperform the combination of various stock characteristics.

At the micro-level, these bounds can provide real-time monitoring of crash risks at the firm level. Compared with a point forecast, having a forecasting bounds can be beneficial in that forecasting uncertainties are quantified and sensitivity analysis based on crash probabilities can be guided.

These crash probability bounds are potentially useful for constructing forward-looking macroeconomic indicators through thoughtful aggregation. For example, when applied to the study of G-SIBs, the maximum of the lower bounds and the minimum of the upper bounds can be used to construct fragility and stability measures of the global banking system.

Table 2.1: Sample Summary

This table summarizes the data sources. The sample period is monthly from January 1996 to December 2017. For each months, we query the OptionMetrics database for firms that have been included to the S&P 500 index before the beginning of that year. The data we use are the implied volatility surface data from which we collect the (implied) option premiums and corresponding strike prices. The time horizons, i.e., the date to maturity are one month, two months, six months and one year. The total number of firm-month pairs, unique firms, unique months, and the median and mean of average number of firms per month are reported. .

Maturity	1	3	6	12
Total no. of observations	128,259	127,540	126,472	126,470
No. of sample months	264	264	264	264
No. of sample firms	1033	1032	1029	1029
Median no. of firms/month	485	484	480	480
Average no. of firms/month	486	483	479	479

Table 2.2: Variation of crash probability bounds across firms

This table presents the summary statistics of time-series averages of crash probability bounds. The sample period is monthly from January 1996 to December 2017. The crash probabilities under consideration are $\mathbb{P}_t[R_i \leq q]$ for $q = 0.80, 0.90, 0.95$. The time-series averages are taken for each firm separately. To rule out the impact of outliers, the time-series average is taken and reported only for firms with over four-year observations. Maturity horizons are one month, two months, six months and one year.

	Lower bound				Upper bound			
Maturity	1	3	6	12	1	3	6	12
No. of firms	700	699	695	695	700	699	695	695
Panel A: $q = 0.80$, down by over 20%								
Min.	0.003	0.017	0.031	0.046	0.012	0.042	0.078	0.130
1st Qu.	0.015	0.051	0.083	0.110	0.032	0.097	0.172	0.271
Median	0.025	0.072	0.110	0.141	0.046	0.130	0.214	0.317
3rd Qu.	0.038	0.098	0.141	0.174	0.064	0.165	0.257	0.364
Max.	0.145	0.240	0.292	0.340	0.196	0.344	0.440	0.557
Mean	0.031	0.080	0.116	0.146	0.053	0.138	0.220	0.320
Std.dev.	0.023	0.040	0.047	0.050	0.031	0.056	0.068	0.074
Panel B: $q = 0.90$, down by over 10%								
Min.	0.031	0.066	0.091	0.103	0.051	0.121	0.187	0.244
1st Qu.	0.077	0.147	0.182	0.200	0.114	0.229	0.312	0.397
Median	0.105	0.182	0.217	0.232	0.147	0.272	0.357	0.441
3rd Qu.	0.135	0.217	0.252	0.265	0.184	0.316	0.398	0.480
Max.	0.295	0.360	0.379	0.402	0.358	0.468	0.534	0.618
Mean	0.112	0.186	0.220	0.234	0.155	0.276	0.357	0.439
Std.dev.	0.047	0.054	0.052	0.050	0.057	0.067	0.066	0.065
Panel C: $q = 0.95$, down by over 5%								
Min.	0.104	0.166	0.174	0.170	0.139	0.240	0.292	0.330
1st Qu.	0.191	0.248	0.265	0.260	0.247	0.348	0.410	0.473
Median	0.223	0.280	0.295	0.290	0.282	0.385	0.446	0.508
3rd Qu.	0.257	0.310	0.323	0.318	0.320	0.420	0.478	0.540
Max.	0.389	0.426	0.425	0.436	0.456	0.536	0.583	0.647
Mean	0.227	0.282	0.296	0.290	0.286	0.386	0.446	0.507
Std.dev.	0.050	0.047	0.044	0.044	0.056	0.054	0.052	0.053

Table 2.3: Variation of crash probability bounds over time

This table presents summary statistics of the cross-sectional averages of crash probability bounds. The sample period is monthly from January 1996 to December 2017. The sample period is monthly from January 1996 to December 2017. The crash probabilities under consideration are $\mathbb{P}_t[R_i \leq q]$ for $q = 0.80, 0.90, 0.95$. The cross-sectional averages are take year by year. Maturity horizons are one month, two months, six months and one year.

	Lower bound				Upper bound			
Maturity	1	3	6	12	1	3	6	12
No. of months	264	264	264	264	264	264	264	264
Panel A: $q = 0.80$, down by over 20%								
Min.	0.003	0.031	0.061	0.103	0.012	0.052	0.110	0.189
1st Qu.	0.014	0.049	0.086	0.120	0.026	0.083	0.146	0.238
Median	0.019	0.061	0.100	0.133	0.034	0.105	0.186	0.288
3rd Qu.	0.038	0.097	0.131	0.158	0.065	0.176	0.273	0.379
Max.	0.126	0.176	0.197	0.216	0.245	0.410	0.517	0.633
Mean	0.028	0.075	0.111	0.142	0.049	0.131	0.212	0.313
Std.dev.	0.022	0.033	0.033	0.029	0.037	0.066	0.079	0.088
Panel B: $q = 0.90$, down by over 10%								
Min.	0.039	0.122	0.173	0.197	0.057	0.168	0.249	0.316
1st Qu.	0.076	0.152	0.194	0.215	0.101	0.215	0.293	0.375
Median	0.092	0.168	0.205	0.224	0.126	0.246	0.331	0.418
3rd Qu.	0.135	0.211	0.236	0.243	0.189	0.327	0.414	0.497
Max.	0.219	0.280	0.298	0.302	0.371	0.508	0.598	0.691
Mean	0.106	0.181	0.216	0.232	0.148	0.270	0.352	0.436
Std.dev.	0.041	0.038	0.030	0.023	0.062	0.071	0.072	0.075
Panel C: $q = 0.95$, down by over 5%								
Min.	0.139	0.240	0.262	0.256	0.170	0.298	0.351	0.395
1st Qu.	0.193	0.260	0.279	0.276	0.234	0.335	0.396	0.457
Median	0.212	0.269	0.288	0.286	0.263	0.364	0.425	0.491
3rd Qu.	0.250	0.292	0.302	0.298	0.326	0.428	0.491	0.556
Max.	0.319	0.360	0.360	0.349	0.450	0.560	0.639	0.719
Mean	0.221	0.278	0.293	0.289	0.279	0.382	0.443	0.505
Std.dev.	0.038	0.026	0.021	0.019	0.059	0.057	0.059	0.064

Table 2.4: Tightness of the crash probability bounds: linear regression tests

This table reports the results from regressing the indicator function of realized equity returns being less than a threshold, q, on the option-implied bounds, $\text{ProbLower}_{i,t}(\tau, q)$ and $\text{ProbUpper}_{i,t}(\tau, q)$, for firms belonging to the S&P 500 index. The data are monthly from January 1996 to December 2017. The return horizons, denoted by τ, are one month, three month, six months, and one year. Results in Panel A, B, C are from the linear regressions,

$$\boldsymbol{I}(R_{i,t\to t+\tau} \leq q) = \alpha + \beta \ \text{ProbLower}_{i,t}(\tau, q) + \varepsilon_{i,t+\tau},$$

or

$$\boldsymbol{I}(R_{i,t\to t+\tau} \leq q) = \alpha + \beta \ \text{ProbUpper}_{i,t}(\tau, q) + \varepsilon_{i,t+\tau},$$

with $q = 0.80, 0.90$ and 0.95. Values in parentheses are standard errors with two-way clustering following Petersen (2009). Values in square brackets are standard errors from block bootstrap using 2500 bootstrap samples following Martin and Wagner (2019). Adjusted R^2s are also reported.

	Lower bound				Upper bound			
Maturity	1	3	6	12	1	3	6	12
Panel A: $q = 0.80$, down by over 20%								
α	−0.005	−0.014	−0.021	−0.045	−0.011	−0.017	−0.017	−0.051
	(0.002)	(0.005)	(0.008)	(0.009)	(0.003)	(0.007)	(0.012)	(0.016)
	[0.002]	[0.009]	[0.015]	[0.025]	[0.004]	[0.008]	[0.022]	[0.036]
β	1.034	1.152	1.193	1.105	0.703	0.680	0.602	0.519
	(0.136)	(0.105)	(0.097)	(0.085)	(0.104)	(0.076)	(0.074)	(0.065)
	[0.129]	[0.162]	[0.177]	[0.170]	[0.117]	[0.094]	[0.151]	[0.145]
R^2-Adj.	6.88%	6.84%	5.90%	5.31%	6.67%	6.08%	4.46%	3.99%
Panel B: $q = 0.90$, down by over 10%								
α	−0.023	−0.039	−0.063	−0.068	−0.031	−0.041	−0.044	−0.068
	(0.007)	(0.011)	(0.015)	(0.018)	(0.009)	(0.017)	(0.026)	(0.033)
	[0.008]	[0.011]	[0.039]	[0.060]	[0.010]	[0.026]	[0.046]	[0.076]
β	1.121	1.165	1.257	1.195	0.861	0.791	0.717	0.634
	(0.093)	(0.079)	(0.081)	(0.081)	(0.084)	(0.079)	(0.086)	(0.087)
	[0.082]	[0.121]	[0.203]	[0.211]	[0.069]	[0.143]	[0.194]	[0.213]
R^2-Adj.	7.05%	5.02%	4.27%	3.45%	7.04%	4.64%	3.22%	2.60%
Panel C: $q = 0.95$, down by over 5%								
α	−0.037	−0.049	−0.080	−0.021	−0.045	−0.042	−0.034	−0.009
	(0.016)	(0.021)	(0.026)	(0.028)	(0.022)	(0.037)	(0.046)	(0.051)
	[0.017]	[0.029]	[0.045]	[0.059]	[0.026]	[0.043]	[0.087]	[0.107]
β	1.129	1.144	1.239	1.051	0.923	0.813	0.717	0.578
	(0.083)	(0.078)	(0.087)	(0.092)	(0.090)	(0.104)	(0.112)	(0.108)
	[0.088]	[0.088]	[0.145]	[0.178]	[0.092]	[0.118]	[0.214]	[0.202]
R^2-Adj.	3.99%	2.54%	2.42%	1.77%	4.10%	2.35%	1.71%	1.26%

Table 2.5: Tightness of the crash probability bounds: linear regression tests adding characteristics for a 20% crash

This table reports the results from regressing the indicator function of realized equity returns being less than 0.80 on the option-implied bounds as well as other characteristics for firms belonging to the S&P 500 index. The data are monthly from January 1996 to December 2017. The return horizons are one month, three month, six months, and one year. Values in parentheses are standard errors with two-way clustering following Petersen (2009). Adjusted R^2s are also reported.

	Lower bound				Upper bound			
Maturity	1	3	6	12	1	3	6	12
intercept	0.009	0.077	0.079	0.093	−0.0004	0.072	0.094	0.094
	(0.024)	(0.051)	(0.068)	(0.069)	(0.025)	(0.053)	(0.069)	(0.069)
bounds	0.809***	0.948***	1.277***	1.054***	0.535***	0.452***	0.451***	0.335***
	(0.116)	(0.126)	(0.142)	(0.126)	(0.109)	(0.082)	(0.084)	(0.069)
beta	0.0003	−0.008	−0.040	−0.049*	0.007	0.016	0.016	0.018
	(0.007)	(0.017)	(0.026)	(0.027)	(0.007)	(0.016)	(0.024)	(0.024)
mom	−0.008*	−0.007	0.003	0.001	−0.006	−0.005	0.007	0.006
	(0.004)	(0.011)	(0.015)	(0.015)	(0.004)	(0.011)	(0.015)	(0.015)
logsize	−0.0001	−0.003	−0.001	−0.003	−0.0003	−0.005	−0.005	−0.006
	(0.001)	(0.003)	(0.004)	(0.004)	(0.001)	(0.003)	(0.004)	(0.004)
bm	−0.010***	−0.021***	−0.042***	−0.042***	−0.009***	−0.017**	−0.036***	−0.036***
	(0.003)	(0.007)	(0.009)	(0.009)	(0.003)	(0.007)	(0.010)	(0.010)
past_vol	0.022	0.031	−0.039	0.023	0.018	0.056**	0.027	0.064
	(0.014)	(0.029)	(0.039)	(0.039)	(0.012)	(0.022)	(0.039)	(0.039)
gross_prof	−0.010**	−0.019*	−0.037**	−0.038**	−0.010**	−0.017	−0.031*	−0.030*
	(0.004)	(0.011)	(0.017)	(0.017)	(0.004)	(0.011)	(0.018)	(0.018)
debt/asset	−0.001	−0.011	−0.042**	−0.048**	0.001	−0.008	−0.037*	−0.040**
	(0.004)	(0.011)	(0.019)	(0.019)	(0.004)	(0.012)	(0.020)	(0.020)
cce/cliab.	−0.002**	−0.005**	−0.010***	−0.011***	−0.002*	−0.004*	−0.008**	−0.009**
	(0.001)	(0.002)	(0.003)	(0.003)	(0.001)	(0.002)	(0.003)	(0.003)
turnover	0.029	−0.017	0.012	0.072	0.020	−0.003	0.050	0.070
	(0.050)	(0.108)	(0.133)	(0.134)	(0.050)	(0.110)	(0.135)	(0.134)
d_turnover	0.023	0.141	0.281	0.165	0.038	0.088	0.143	0.096
	(0.060)	(0.133)	(0.184)	(0.183)	(0.063)	(0.134)	(0.186)	(0.183)
shortint.	0.002*	0.007***	0.013***	0.013***	0.002**	0.008***	0.015***	0.015***
	(0.001)	(0.002)	(0.004)	(0.004)	(0.001)	(0.002)	(0.004)	(0.004)
R^2-Adj.	5.0%	5.3%	5.2%	4.9%	4.9%	4.8%	4.0%	3.9%

Notes:

***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

Table 2.6: Tightness of the crash probability bounds: linear regression tests adding characteristics for a 10% crash

This table reports the results from regressing the indicator function of realized equity returns being less than 0.90 on the option-implied bounds as well as other characteristics for firms belonging to the S&P 500 index. The data are monthly from January 1996 to December 2017. The return horizons are one month, three month, six months, and one year. Values in parentheses are standard errors with two-way clustering following Petersen (2009). Adjusted R^2s are also reported.

	Lower bound				Upper bound			
Maturity	1	3	6	12	1	3	6	12
intercept	−0.003	0.092	0.063	0.167*	−0.023	0.085	0.108	0.180*
	(0.051)	(0.076)	(0.096)	(0.100)	(0.055)	(0.080)	(0.098)	(0.100)
bounds	0.941***	1.074***	1.381***	1.110***	0.730***	0.672***	0.642***	0.445***
	(0.096)	(0.128)	(0.136)	(0.134)	(0.094)	(0.092)	(0.103)	(0.096)
beta	−0.0005	−0.036	−0.065*	−0.061*	0.020	0.012	0.021	0.027
	(0.017)	(0.028)	(0.035)	(0.034)	(0.017)	(0.026)	(0.032)	(0.032)
mom	−0.011	−0.0003	0.012	0.014	−0.008	0.006	0.022	0.022
	(0.012)	(0.018)	(0.019)	(0.019)	(0.012)	(0.018)	(0.020)	(0.020)
logsize	0.001	−0.003	−0.0004	−0.006	0.0001	−0.005	−0.006	−0.010*
	(0.003)	(0.004)	(0.005)	(0.005)	(0.003)	(0.004)	(0.005)	(0.005)
bm	−0.017**	−0.039***	−0.064***	−0.066***	−0.015**	−0.037***	−0.061***	−0.062***
	(0.007)	(0.010)	(0.011)	(0.011)	(0.007)	(0.010)	(0.012)	(0.012)
past_vol	0.064	0.060	−0.010	0.081*	0.043	0.047	0.006	0.078
	(0.040)	(0.054)	(0.047)	(0.047)	(0.035)	(0.036)	(0.047)	(0.048)
gross_prof	−0.014	−0.028*	−0.060***	−0.054**	−0.013	−0.024	−0.049**	−0.045**
	(0.010)	(0.016)	(0.020)	(0.021)	(0.010)	(0.016)	(0.022)	(0.022)
debt/asset	−0.012	−0.051***	−0.084***	−0.098***	−0.009	−0.048***	−0.081***	−0.090***
	(0.010)	(0.017)	(0.024)	(0.024)	(0.010)	(0.018)	(0.025)	(0.026)
cce/cliab.	−0.003*	−0.009***	−0.015***	−0.016***	−0.003	−0.007**	−0.012***	−0.013***
	(0.002)	(0.003)	(0.005)	(0.005)	(0.002)	(0.003)	(0.005)	(0.005)
turnover	−0.007	−0.051	−0.061	0.007	−0.023	−0.063	−0.080	−0.041
	(0.106)	(0.151)	(0.163)	(0.165)	(0.106)	(0.153)	(0.165)	(0.165)
d_turnover	0.032	0.251	0.340	0.150	0.066	0.254	0.315	0.195
	(0.149)	(0.204)	(0.234)	(0.236)	(0.153)	(0.199)	(0.239)	(0.236)
shortint.	0.005**	0.010***	0.018***	0.018***	0.006***	0.012***	0.020***	0.020***
	(0.002)	(0.003)	(0.004)	(0.004)	(0.002)	(0.003)	(0.004)	(0.004)
R^2-Adj.	6.1%	4.8%	4.4%	3.8%	6.1%	4.4%	3.4%	3.0%

Notes:

***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

Table 2.7: Tightness of the crash probability bounds: linear regression tests adding characteristics for a 5% crash

This table reports the results from regressing the indicator function of realized equity returns being less than 0.95 on the option-implied bounds as well as other characteristics for firms belonging to the S&P 500 index. The data are monthly from January 1996 to December 2017. The return horizons are one month, three month, six months, and one year. Values in parentheses are standard errors with two-way clustering following Petersen (2009). Adjusted R^2s are also reported.

	Lower bound				Upper bound			
Maturity	1	3	6	12	1	3	6	12
intercept	0.029	0.066	0.051	0.217*	−0.007	0.068	0.127	0.261**
	(0.073)	(0.097)	(0.114)	(0.117)	(0.079)	(0.103)	(0.119)	(0.121)
bounds	0.908***	1.002***	1.259***	0.881***	0.777***	0.662***	0.593***	0.326***
	(0.112)	(0.158)	(0.155)	(0.150)	(0.107)	(0.121)	(0.132)	(0.121)
beta	0.001	−0.041	−0.057	−0.041	0.028	0.012	0.028	0.032
	(0.028)	(0.032)	(0.037)	(0.037)	(0.027)	(0.031)	(0.036)	(0.035)
mom	−0.004	0.007	0.015	0.018	0.0002	0.014	0.025	0.025
	(0.019)	(0.022)	(0.021)	(0.021)	(0.019)	(0.022)	(0.022)	(0.021)
logsize	−0.0005	−0.0003	0.00003	−0.006	−0.001	−0.002	−0.004	−0.008
	(0.004)	(0.005)	(0.006)	(0.006)	(0.004)	(0.005)	(0.006)	(0.006)
bm	−0.027***	−0.051***	−0.079***	−0.081***	−0.025**	−0.050***	−0.077***	−0.078***
	(0.010)	(0.011)	(0.012)	(0.012)	(0.010)	(0.012)	(0.013)	(0.013)
past_vol	0.113**	0.110*	0.067	0.135***	0.071	0.076*	0.053	0.124**
	(0.054)	(0.063)	(0.051)	(0.051)	(0.046)	(0.044)	(0.050)	(0.051)
gross_prof	−0.026*	−0.032	−0.055***	−0.047**	−0.025*	−0.028	−0.045**	−0.039*
	(0.014)	(0.020)	(0.021)	(0.022)	(0.014)	(0.020)	(0.022)	(0.023)
debt_asset	−0.033**	−0.070***	−0.114***	−0.123***	−0.029**	−0.066***	−0.108***	−0.114***
	(0.015)	(0.021)	(0.026)	(0.027)	(0.015)	(0.021)	(0.027)	(0.028)
cce/cliab.	−0.005*	−0.011***	−0.016***	−0.016***	−0.004	−0.009**	−0.013**	−0.013**
	(0.003)	(0.004)	(0.005)	(0.005)	(0.003)	(0.004)	(0.005)	(0.005)
turnover	−0.061	−0.060	−0.094	−0.054	−0.087	−0.091	−0.143	−0.104
	(0.147)	(0.168)	(0.177)	(0.178)	(0.146)	(0.170)	(0.179)	(0.178)
d_turnover	−0.028	0.205	0.253	0.115	0.033	0.251	0.302	0.179
	(0.219)	(0.232)	(0.258)	(0.262)	(0.220)	(0.224)	(0.260)	(0.258)
shortint.	0.007**	0.014***	0.019***	0.018***	0.008***	0.015***	0.020***	0.019***
	(0.003)	(0.004)	(0.004)	(0.004)	(0.003)	(0.004)	(0.005)	(0.005)
R^2-Adj.	3.8%	2.8%	2.9%	2.4%	3.8%	2.5%	2.2%	1.9%

Notes:

***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

Table 2.8: Area under the curve (AUC) statistics of out-of-sample forecasting using option-implied bounds and characteristic-based procedures

This table reports AUCs of forecasting whether a stock's net return will be less than -20% using the option-implied bounds, as well as statistical procedures. Stocks of firms belonging to the S&P 500 index are considered. The data are monthly from January 1996 to December 2017. The return horizons are one month, three month, six months, and one year. The two statistical procedures under consideration are logistic regression and linear regression, both fine tuned by the LASSO variable selection method. The training sample consists the first half of the data (1996-2006) and the testing sample consists the rest. AUC are all calculated from the testing sample.

Maturity	1	3	6	12
Panel A: $q = 0.80$, down by 20%				
Lower Bound	0.871	0.771	0.723	0.721
Upper Bound	0.874	0.764	0.703	0.695
Char.Logistic-Lasso	0.802	0.720	0.675	0.676
Char.Linear-Lasso	0.832	0.729	0.680	0.679
Panel B: $q = 0.90$, down by 10%				
Lower Bound	0.760	0.680	0.657	0.644
Upper Bound	0.760	0.673	0.626	0.605
Char.Logistic-Lasso	0.732	0.656	0.625	0.626
Char.Linear-Lasso	0.733	0.639	0.616	0.614
Panel C: $q = 0.95$, down by 5%				
Lower Bound	0.645	0.609	0.610	0.593
Upper Bound	0.646	0.604	0.578	0.585
Char.Logistic-Lasso	0.636	0.597	0.585	0.584
Char.Linear-Lasso	0.629	0.587	0.574	0.572

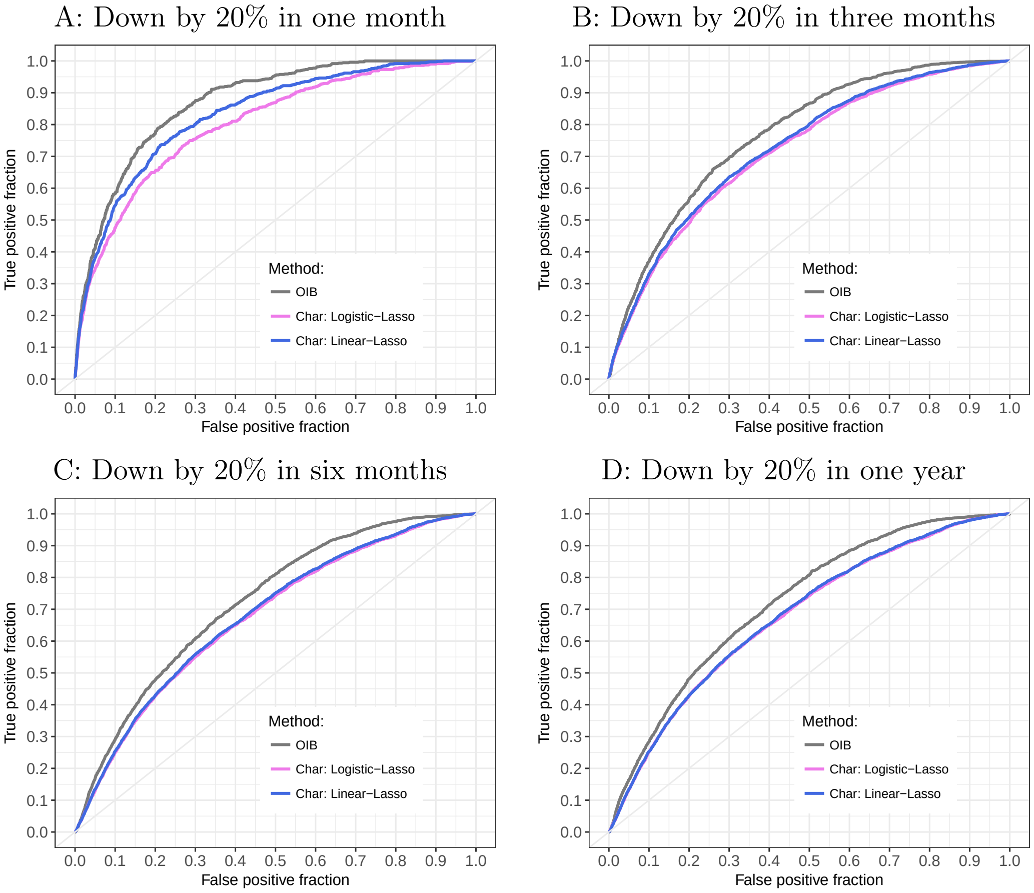


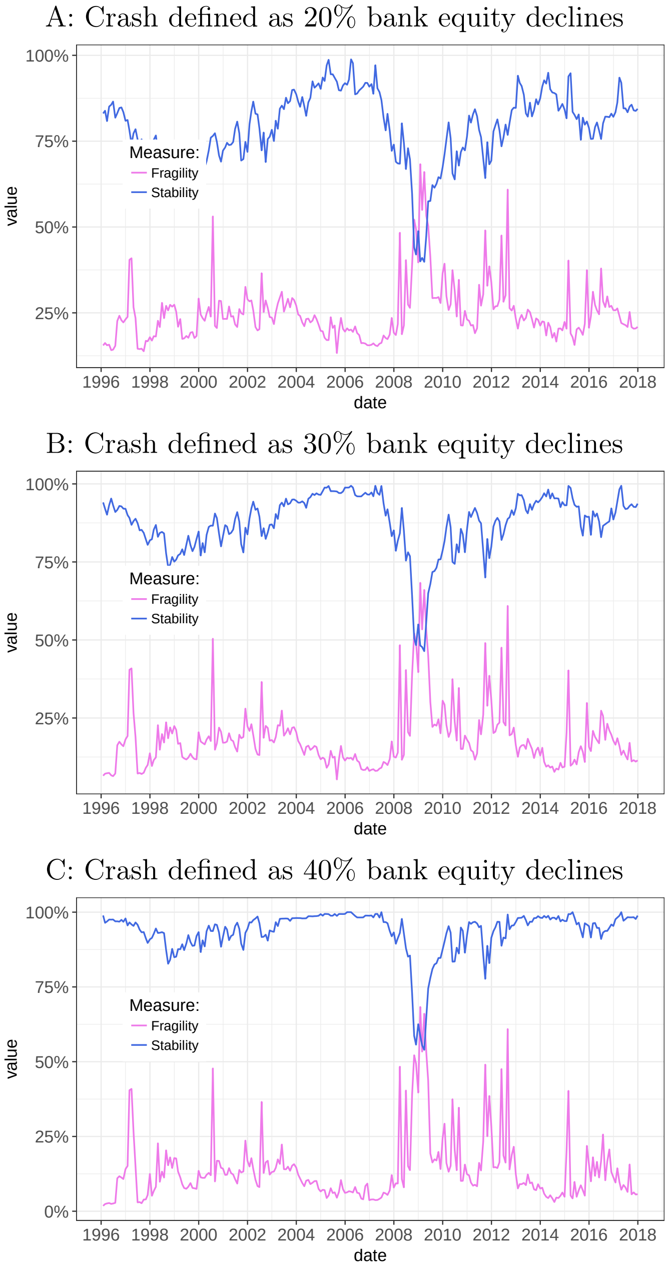
Figure 2.5: ROC curves of out-of-sample forecasting of crash events: option-implied bounds (OIC) v.s. characteristics-based procedures

The figures report ROC curves of forecasting whether a stock's net return will be less than −20% using the option-implied bounds, as well as statistical procedures. Stocks of firms belonging to the S&P 500 index are considered. The data are monthly from January 1996 to December 2017. The return horizons are one month, three month, six months, and one year. The two statistical procedures under consideration are logistic regression and linear regression, both fine tuned by the LASSO variable selection method. The training sample consists the first half of the data (1996-2006) and the testing sample consists the rest. ROC curves are all based on the testing sample.

Table 2.9: G-SIBs under consideration

This table lists the twenty-one global systemically important banks (G-SIBs) under study, as subset of all the G-SIBs ever defined by the Financial Stability Board since 2011. The third column reports the period considered as a G-SIB for each bank. The forth column reports the time window within which the option data on each bank's equity are available. To be considered in this analysis, a G-SIB must be either publicly traded in the US or have issued American depositary receipts.

Bank name	Country	FSB G-SIB period	Option sample period
Mizuho FG	Japan	2011-present	2008.10-2017.12
Sumitomo Mitsui FG	Japan	2011-present	2011.06-2017.12
Mitsubishi UFJ FG	Japan	2011-present	1998.02-2017.12
Deutsche Bank	Germany	2011-present	2001.11-2017.12
ING Bank	Netherlands	2011-present	1997.07-2017.12
BBVA	Spain	2012-2015	1998.10-2017.12
Santander	Spain	2011-present	1997.11-2017.12
Credit Suisse	Switzerland	2011-present	2005.03-2017.12
UBS	Switzerland	2011-present	2000.07-2014.12
Barclays	United Kingdom	2011-present	2007.11-2017.12
HSBC	United Kingdom	2011-present	1999.12-2017.12
Lloyds	United Kingdom	2011-2012	2008.10-2017.12
Royal Bank of Canada	Canada	2017-present	2000.10-2017.12
Bank of America	United States	2011-present	1996.01-2017.12
Bank of New York Mellon	United States	2011-present	1996.01-2017.12
Citi	United States	2011-present	1996.01-2017.12
Goldman Sachs	United States	2011-present	1999.08-2017.12
JP Morgan Chase	United States	2011-present	1996.01-2017.12
Morgan Stanley	United States	2011-present	1996.01-2017.12
State Street	United States	2011-present	1996.01-2017.12
Wells Fargo	United States	2011-present	1996.01-2017.12



A: Crash defined as 20% bank equity declines

B: Crash defined as 30% bank equity declines

C: Crash defined as 40% bank equity declines

Figure 2.6: Fragility and stability measures of the global banking system

The figures present the two measures constructed from the crash probability bounds for the global systemically important banks (G-SIBs) introduced in Section 2.4.1. Both measures are based on one-year crash probability bounds. At the end of each month from 1996-2017, the (cross-sectional) maximum of the lower bounds among the G-SIBs is the fragility measure, and one minus the minimum of the upper bounds is the stability measure. Each panel corresponds to a specific choice of q in defining a crash event.

2.7 Appendices

2.7.1 Proofs

Proof of Proposition 9

Proof. When the function h satisfies Assumption 9, the integrand

$$k(u,v) = \left[Q_m^{-1}(u)\right]^{\gamma} h\left(Q_i^{-1}(v)\right)$$

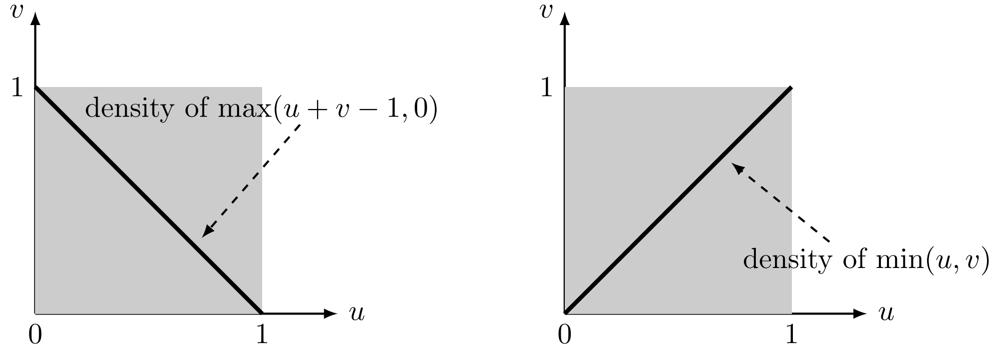
is two-increasing. Based on Corollary 2.2 of Tchen (1980) and the Fréchet-Heoffding (F-H) theorem,

$$\int_{[0,1]^2} \left[Q_m^{-1}(u)\right]^{\gamma} h\left(Q_i^{-1}(v)\right)\, \mathrm{d}u\, \mathrm{d}v \geq \int_{[0,1]^2} \left[Q_m^{-1}(u)\right]^{\gamma} h\left(Q_i^{-1}(v)\right)\, \mathrm{d}\left(\max(u+v-1,\ 0)\right),$$

and

$$\int_{[0,1]^2} \left[Q_m^{-1}(u)\right]^{\gamma} h\left(Q_i^{-1}(v)\right)\, \mathrm{d}u\, \mathrm{d}v \leq \int_{[0,1]^2} \left[Q_m^{-1}(u)\right]^{\gamma} h\left(Q_i^{-1}(v)\right)\, \mathrm{d}\left(\min(u,v)\right).$$

The probability densities of the F-H lower bound, $\max(u+v-1,\ 0)$, and the F-H upper bound, $\min(u,v)$, are uniformly distributed along the two diagonals of the square $[0,1]^2$ in $\mathbb{R}^2$, illustrated as follows:



Integrating the right hand sides of the two inequalities above (with regard to these two densities), the numerators of the lower and upper bounds are

$$\int_0^1 \left[Q_m^{-1}(u)\right]^{\gamma} h\left(Q_i^{-1}(1-u)\right)\, \mathrm{d}u \overset{R_m=Q_m^{-1}(u)}{=\!=} \int_0^{\infty} R_m^{\gamma} h(Q_i^{-1}(1-Q_m(R_m)))\, \mathrm{d}R_m,$$

$$\int_0^1 \left[Q_m^{-1}(u)\right]^{\gamma} h\left(Q_i^{-1}(u)\right)\, \mathrm{d}u \overset{R_m=Q_m^{-1}(u)}{=\!=} \int_0^{\infty} R_m^{\gamma} h(Q_i^{-1}(Q_m(R_m)))\, \mathrm{d}R_m,$$

which deliver bounds summarized in the proposition.

The lower bound is achieved when the copula function linking Q_m and Q_i is $\max(u+v-1,\ 0)$, that is, the joint risk-neutral CDF of $(Q_m(R_m), Q_i(R_i))$ is $\max(u+v-1,0)$. This implies that $Q_m(R_m)+Q_i(R_i) \equiv 1$. Similarly, the upper bound is a achieved when the joint risk-neutral CDF of $(Q_m(R_m), Q_i(R_i))$ is $\min(u,v)$, that is, when $Q_i(R_i) = Q_m(R_m)$. □

Proof of Proposition 10.

Proof. In Proposition 9, let $h(R_i) = -\boldsymbol{I}(R_i \le q)$, which satisfies Assumption 9, then

$$\begin{aligned}\mathbb{P}[R_i \le q] &= -\,\mathbb{E}[h(R_i)] \\ &\ge -\,\frac{\mathbb{E}^*\left[R_m^\gamma h(Q_i^{-1}(Q_m(R_m)))\right]}{\mathbb{E}^*\left[R_m^\gamma\right]} \\ &= \frac{\mathbb{E}^*\left[R_m^\gamma \boldsymbol{I}(Q_m(R_m) \le Q_i(q))\right]}{\mathbb{E}\left[R_m^\gamma\right]} \\ &= \frac{\mathbb{E}^*\left[R_m^\gamma \boldsymbol{I}\left(R_m \le q_l\right)\right]}{\mathbb{E}\left[R_m^\gamma\right]} \qquad \left(q_l = Q_m^{-1}\left(Q_i\left(q\right)\right) \text{ by definition}\right).\end{aligned}$$

For the inequality in the second step, "=" is achieved if and only if R_m and R_i are comonotonic such that $Q_i(R_i) = Q_m(R_m)$. Similarly, we have

$$\begin{aligned}\mathbb{P}[R_i \le q] &= -\mathbb{E}[h(R_i)] \\ &\le -\frac{\mathbb{E}^*\left[R_m^\gamma h(Q_i^{-1}(1 - Q_m(R_m)))\right]}{\mathbb{E}^*\left[R_m^\gamma\right]} \\ &= \frac{\mathbb{E}^*\left[R_m^\gamma \boldsymbol{I}\left(1 - Q_m(R_m) \le Q_i(q)\right)\right]}{\mathbb{E}^*\left[R_m^\gamma\right]} \\ &= \frac{\mathbb{E}^*\left[R_m^\gamma \boldsymbol{I}\left(R_m \ge q_u\right)\right]}{\mathbb{E}^*\left[R_m^\gamma\right]} \qquad \left(q_u = Q_m^{-1}\left(1 - Q_i\left(q\right)\right) \text{ by definition}\right).\end{aligned}$$

Again, the inequality in the second step can be strictly equal if and only if R_m and R_i are coutermonotonic satisfying $Q_i(R_i) + Q_m(R_m) = 1$.

The upper bound is always greater than the lower bound. A bridge between them is the *risk-neutral* crash probability. Specifically, by the continuous version of Chebyshev's sum inequality[20],

$$\frac{\mathbb{E}^*\left[R_m^\gamma \boldsymbol{I}\left(R_m \le q_l\right)\right]}{\mathbb{E}\left[R_m^\gamma\right]} \le \frac{\mathbb{E}^*\left[R_m^\gamma\right]\mathbb{E}^*\left[\boldsymbol{I}\left(R_m \le q_l\right)\right]}{\mathbb{E}^*\left[R_m^\gamma\right]} = Q_m(q_l) = Q_i(q) = \mathbb{P}^*[R_i \le q],$$

$$\frac{\mathbb{E}^*\left[R_m^\gamma \boldsymbol{I}\left(R_m \ge q_u\right)\right]}{\mathbb{E}^*\left[R_m^\gamma\right]} \ge \frac{\mathbb{E}^*\left[R_m^\gamma\right]\mathbb{E}^*\left[\boldsymbol{I}\left(R_m \ge q_u\right)\right]}{\mathbb{E}^*\left[R_m^\gamma\right]} = 1 - Q_m(q_u) = Q_i(q) = \mathbb{P}^*[R_i \le q].$$

□

Proof of Corollary 1

Proof. A proof for the case of $\gamma = 1$ follows from Martin (2017). Here we prove the general results for any $\gamma > 0$ and then recast the proof for the $\gamma = 1$ case in light of our new

[20] For functions f and g which are integrable over $[0, 1]$, both non-increasing or both non-decreasing, then

$$\int_0^1 f(x)g(x)\,\mathrm{d}x \ge \int_0^1 f(x)\,\mathrm{d}x \int_0^1 g(x)\,\mathrm{d}x.$$

And if one is non-increasing and the other is non-decreasing, the inequality above is reversed. Let $f(x) = [Q_m^{-1}(x)]^\gamma$ (non-decreasing), then specifying $g(x) = \boldsymbol{I}(x \le Q_i(q))$ yields the first inequality below and $g(x) = \boldsymbol{I}(x \ge Q_i(q))$ the second.

general results.

First, the Carr-Madan formula (Carr and Madan, 2001) says that for any smooth function $g(\cdot)$,

$$g(S) = g(F)+g'(F)(S-F)+\int_0^F g''(K)\max\{K-S,\,0\}\,\mathrm{d}K+\int_F^\infty g''(K)\max\{S-K,\,0\}\,\mathrm{d}K.$$

Let S_{m0} and $F = S_{m0}R_f$ be the spot and forward level of the market index, the function $g(S)$ be S^γ. Treating S, a random variable, as the level of market index next period, taking the risk-neutral expectations on both sides of the equation above (changing orders of integrals when needed), we have

$$\begin{aligned}\mathbb{E}^*\left[S^\gamma\right] =& S_{m0}^\gamma R_f^\gamma + \gamma S_{m0}^{\gamma-1} R_f^{\gamma-1}(\mathbb{E}^*[S] - F) \\ &+ \int_0^F \gamma(\gamma-1)K^{\gamma-2}R_f\mathrm{put}(K)\,\mathrm{d}K + \int_F^\infty \gamma(\gamma-1)K^{\gamma-2}R_f\mathrm{call}(K)\,\mathrm{d}K.\end{aligned}$$

Dividing both sides by S_{m0}^γ and noticing that $R_m = S/S_{m0}$ and that $\mathbb{E}^*[S] = F$, we have the first equation.

Next, noticing that

$$\mathbb{E}^*\left[R_m^\gamma \boldsymbol{I}(R_m \le q_l)\right] = \frac{\mathbb{E}^*\left[S^\gamma \boldsymbol{I}(S \le K_l)\right]}{S_{m0}^\gamma} = \frac{R_f}{S_{m0}^\gamma}\int_0^{K_l} K^\gamma \mathrm{put}''(K)\,\mathrm{d}K$$

where the second equation comes from (2.6) following the Breeden-Litzenberger static replication logic (Breeden and Litzenberger, 1978). Integrating the last integral by parts and using the fact that put(0) = put′(0) = 0, we have

$$\begin{aligned}\int_0^{K_l} K^\gamma \mathrm{put}''(K)\,\mathrm{d}K &= K^\gamma \mathrm{put}'(K)\Big|_0^{K_l} - \int_0^{K_l}\gamma K^{\gamma-1}\mathrm{put}'(K)\,\mathrm{d}K \\ &= K_l^\gamma \mathrm{put}'(K_l) - \left(\gamma K^{\gamma-1}\mathrm{put}(K)\Big|_0^{K_l} - \int_0^{K_l}\gamma(\gamma-1)K^{\gamma-2}\mathrm{put}(K)\,\mathrm{d}K\right) \\ &= K_l^\gamma \mathrm{put}'(K_l) - \left(\gamma K_l^{\gamma-1}\mathrm{put}(K_l) - \int_0^{K_l}\gamma(\gamma-1)K^{\gamma-2}\mathrm{put}(K)\,\mathrm{d}K\right).\end{aligned}$$

Plugging the expression back to the equation for $\mathbb{E}^*\left[R_m^\gamma \boldsymbol{I}(R_m \le q_l)\right]$ yields the second equation.

Finally, as

$$\mathbb{E}^*\left[R_m^\gamma \boldsymbol{I}(R_m \ge q_u)\right] = \frac{R_f}{S_{m0}^\gamma}\int_{K_u}^\infty K^\gamma R_f \mathrm{call}''(K)\,\mathrm{d}K,$$

following the same logic, we integrate the right-hand side integral by parts

$$\begin{aligned}\int_{K_u}^{\infty} K^{\gamma}\text{call}''(K)\,\mathrm{d}K &= K^{\gamma}\text{call}'(K)\Big|_{K_u}^{\infty} - \int_{K_u}^{\infty}\gamma K^{\gamma-1}\text{call}'(K)\,\mathrm{d}K \\ &= -K_u^{\gamma}\text{call}'(K_u) - \left(\gamma K^{\gamma-1}\text{call}(K)\Big|_{K_u}^{\infty} - \int_{K_u}^{\infty}\gamma(\gamma-1)K^{\gamma-2}\text{call}(K)\,\mathrm{d}K\right) \\ &= -K_u^{\gamma}\text{call}'(K_u) + \left(\gamma K_u^{\gamma-1}\text{call}(K_l) + \int_{K_u}^{\infty}\gamma(\gamma-1)K^{\gamma-2}\text{call}(K)\,\mathrm{d}K\right)\end{aligned}$$

where the second and third equations rely on the fact that $\text{call}'(\infty) = 0$ and $\text{call}(\infty) = 0$ respectively. Again, multiplying the last formula by R_f/S_{m0}^{γ} leads to the third equation.

Letting $\gamma = 1$ in the three equations of the general results and calculating the ratios of expectations for the bounds generates the formulas for the special case. □

Proof of Corollary 2

Proof. We prove results related to the log-normal example here. First, we state a simple fact that, for a log-normal random variable X such that $\log X \sim (\mu, \sigma^2)$,

$$\mathbb{E}\left[X\boldsymbol{I}(X \leq q)\right] = \Phi\left(\frac{\log q - \mu - \sigma^2}{\sigma}\right)\mathbb{E}[X], \quad \mathbb{E}\left[X\boldsymbol{I}(X \geq q)\right] = \Phi\left(\frac{\mu + \sigma^2 - \log q}{\sigma}\right)\mathbb{E}[X]$$

Noticing that R_m^{γ} is also log-normal, i.e., $\log(R_m^{\gamma}) \sim \mathcal{N}(\gamma\mu_m, \gamma^2\sigma_m^2)$, the lower bound according to Proposition 10 is

$$\begin{aligned}\frac{\mathbb{E}^*[R_m^{\gamma}\boldsymbol{I}(R_m \leq q_l)]}{\mathbb{E}^*[R_m^{\gamma}]} =& \frac{\mathbb{E}^*\left[R_m^{\gamma}\boldsymbol{I}\left(R_m^{\gamma} \leq \left[Q_m^{-1}(Q_i(q))\right]^{\gamma}\right)\right]}{\mathbb{E}^*[R_m^{\gamma}]} \\ =&\Phi\left(\frac{\gamma\log\left[Q_m^{-1}(Q_i(q))\right] - \gamma\mu_m - \gamma^2\sigma_m^2}{\gamma\sigma_m}\right),\end{aligned}$$

while $\log\left[Q_m^{-1}(Q_i(q))\right] = \frac{\log q - \mu_i}{\sigma_i}\sigma_m + \mu_m$ due to the log-normality of both R_i and R_m. Plugging this equation in and simplifying the terms, we have

$$\frac{\mathbb{E}^*[R_m^{\gamma}\boldsymbol{I}(R_m \leq q_l)]}{\mathbb{E}^*[R_m^{\gamma}]} = \Phi\left(\frac{\log q - \mu_i}{\sigma_i} - \gamma\sigma_m\right).$$

For the upper bound

$$\begin{aligned}\frac{\mathbb{E}^*[R_m^{\gamma}\boldsymbol{I}(R_m \geq q_u)]}{\mathbb{E}^*[R_m^{\gamma}]} =& \frac{\mathbb{E}^*\left[R_m^{\gamma}\boldsymbol{I}\left(R_m^{\gamma} \geq \left[Q_m^{-1}(1 - Q_i(q))\right]^{\gamma}\right)\right]}{\mathbb{E}^*[R_m^{\gamma}]} \\ =&\Phi\left(\frac{\gamma\mu_m + \gamma^2\sigma_m^2 - \gamma\log\left[Q_m^{-1}(1 - Q_i(q))\right]}{\gamma\sigma_m}\right).\end{aligned}$$

Given that $\log\left[Q_m^{-1}(1 - Q_i(q))\right] = \frac{\mu_i - \log q}{\sigma_i}\sigma_m + \mu_m$, plugging into the expression above for the upper bound,

$$\frac{\mathbb{E}^*[R_m^\gamma \boldsymbol{I}(R_m \geq q_u)]}{\mathbb{E}^*[R_m^\gamma]} = \Phi\left(\frac{\log q - \mu_i}{\sigma_i} + \gamma\sigma_m\right).$$

The risk-neutral crash probability $\mathbb{P}^*[R_i \leq q]$ is simply $\Phi\left(\frac{\log q - \mu_i}{\sigma_i}\right)$, thus,

$$\Phi^{-1}(p^*) = \frac{\log q - \mu_i}{\sigma_i}.$$

Plugging this equation back to the expressions for the lower and upper bounds leads to equation (2.13).

Now we turn our attention to results in equation (2.14). Noticing that,

$$[\log R_m^\gamma \mid \log R_i = r_i] \sim \mathcal{N}\left(\gamma\mu_m + \gamma\rho\frac{\sigma_m}{\sigma_i}(r_i - \mu_i),\, \gamma^2(1-\rho^2)\sigma_m^2\right),$$

we have

$$\begin{aligned}
\mathbb{E}^*\left[R_m^\gamma \mid R_i \leq q\right] =& \mathbb{E}^*\left[R_m^\gamma \mid \log R_i \leq \log q\right] \\
=& \exp\left(\gamma\mu_m - \gamma\rho\frac{\sigma_m}{\sigma_i}\mu_i + \frac{1}{2}\gamma^2(1-\rho^2)\sigma_m^2\right)\mathbb{E}^*\left[\exp\left(\gamma\rho\frac{\sigma_m}{\sigma_i}r_i\right) \,\middle|\, r_i \leq \log q\right] \\
=& \exp\left(\gamma\mu_m - \gamma\rho\frac{\sigma_m}{\sigma_i}\mu_i + \frac{1}{2}\gamma^2(1-\rho^2)\sigma_m^2\right)\frac{\int_{-\infty}^{\log q}\exp\left(\gamma\rho\frac{\sigma_m}{\sigma_i}x\right)\phi(x\,;\,\mu_i,\,\sigma_i^2)\,\mathrm{d}x}{\int_{-\infty}^{\log q}\phi(x\,;\,\mu_i,\,\sigma_i^2)\,\mathrm{d}x},
\end{aligned}$$

where $\phi(x;\mu,\,\sigma^2)$ represents the probability density function of $\mathcal{N}(\mu,\,\sigma^2)$. On the one hand,

$$\begin{aligned}
&\int_{-\infty}^{\log q}\exp\left(\gamma\rho\frac{\sigma_m}{\sigma_i}x\right)\phi(x\,;\,\mu_i,\,\sigma_i^2)\,\mathrm{d}x \\
=&\int_{-\infty}^{\log q}\frac{1}{\sqrt{2\pi}\sigma_i}\exp\left(\gamma\rho\frac{\sigma_m}{\sigma_i}x - \frac{x^2}{2\sigma_i^2} - \frac{\mu_i^2}{2\sigma_i^2} + \frac{\mu_i x}{\sigma_i^2}\right)\mathrm{d}x \\
=&\int_{-\infty}^{\log q}\frac{1}{\sqrt{2\pi}\sigma_i}\exp\left\{-\frac{(x-\mu_i-\gamma\rho\sigma_m\sigma_i)^2}{2\sigma_i^2} + \frac{(\mu_i+\gamma\rho\sigma_m\sigma_i)^2-\mu_i^2}{2\sigma_i^2}\right\}\mathrm{d}x \\
=&\exp\left(\frac{2\mu_i\gamma\rho\sigma_m\sigma_i + \gamma^2\rho^2\sigma_m^2\sigma_i^2}{2\sigma_i^2}\right)\Phi\left(\frac{\log q - \mu_i - \gamma\rho\sigma_m\sigma_i}{\sigma_i}\right)
\end{aligned}$$

on the other hand,

$$\int_{-\infty}^{\log q}\phi(x\,;\,\mu_i,\,\sigma_i^2)\,\mathrm{d}x = \Phi\left(\frac{\log q - \mu_i}{\sigma_i}\right).$$

Thus, we can simplify terms for calculating $\mathbb{E}^*\left[R_m^\gamma \mid R_i \leq q\right]$ as follows,

$$\begin{aligned}
&\mathbb{E}^*\left[R_m^\gamma \mid R_i \leq q\right] \\
=&\exp\left\{\gamma\mu_m - \gamma\rho\frac{\sigma_m}{\sigma_i}\mu_i + \frac{1}{2}\gamma^2(1-\rho^2)\sigma_m^2 + \gamma\mu_i\rho\frac{\sigma_m}{\sigma_i} + \frac{1}{2}\gamma^2\rho^2\sigma_m^2\right\}\frac{\Phi\left(\frac{\log q - \mu_i - \gamma\rho\sigma_m\sigma_i}{\sigma_i}\right)}{\Phi\left(\frac{\log q - \mu_i}{\sigma_i}\right)} \\
=&\exp\left(\gamma\mu_m + \frac{1}{2}\gamma^2\sigma_m^2\right)\frac{\Phi\left(\frac{\log q - \mu_i}{\sigma_i} - \gamma\rho\sigma_m\right)}{\Phi\left(\frac{\log q - \mu_i}{\sigma_i}\right)} \\
=&\mathbb{E}^*\left[R_m^\gamma\right]\frac{\Phi\left(\Phi^{-1}(p^*) - \gamma\rho\sigma_m\right)}{\mathbb{P}^*[R_i \leq q]}.
\end{aligned}$$

As a result, under our framework,

$$\mathbb{P}[R_i \leq q] = \frac{\mathbb{E}^*[R_m^\gamma \boldsymbol{I}(R_i \leq q)]}{\mathbb{E}^*[R_m^\gamma]} = \frac{\mathbb{E}^*[R_m^\gamma \mid R_i \leq q] \times \mathbb{P}^*[R_i \leq q]}{\mathbb{E}^*[R_m^\gamma]} = \Phi\left(\Phi^{-1}(p^*) - \gamma\rho\sigma_m\right).$$

□

Proof of Corollary 3

Proof. First, when $\gamma = 0$, the bounds become

$$\mathbb{P}^*[R_m \leq q_l] \leq \mathbb{P}[R_i \leq q] \leq \mathbb{P}^*[R_m \geq q_u].$$

By definition, both the lower and upper bounds equal $\mathbb{P}^*[R_i \leq q]$.

Second, define $\psi(x) = \boldsymbol{I}(Q_m(x) \leq Q_i(q))$, which is a decreasing function. The lower bound is then $\mathbb{E}^*[R_m^\gamma\psi(R_m)]/\mathbb{E}^*[R_m^\gamma]$. Differentiating with regard to $\gamma \in (0, \infty)$:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\gamma}\left\{\frac{\mathbb{E}^*[R_m^\gamma\psi(R_m)]}{\mathbb{E}^*[R_m^\gamma]}\right\} &= \frac{\mathbb{E}^*[R_m^\gamma\log(R_m)\psi(R_m)]\mathbb{E}^*[R_m^\gamma] - \mathbb{E}^*[R_m^\gamma\psi(R_m)]\mathbb{E}^*[R_m^\gamma\log(R_m)]}{\{\mathbb{E}^*[R_m^\gamma]\}^2} \\
&= \frac{1}{\{\mathbb{E}^*[R_m^\gamma]\}^2}\iint\left(x^\gamma\log(x)\psi(x)y^\gamma - x^\gamma\psi(x)y^\gamma\log(y)\right)\,\mathrm{d}Q_m(x)\,\mathrm{d}Q_m(y) \\
&\leq \frac{1}{\{\mathbb{E}^*[R_m^\gamma]\}^2}\left[\iint_{x\geq y} x^\gamma y^\gamma\psi(y)\log\left(\frac{x}{y}\right)\,\mathrm{d}Q_m(x)\,\mathrm{d}Q_m(y)\right. \\
&\qquad\left. + \iint_{x\leq y} x^\gamma y^\gamma\psi(x)\log\left(\frac{x}{y}\right)\,\mathrm{d}Q_m(x)\,\mathrm{d}Q_m(y)\right] \\
&= \frac{1}{\{\mathbb{E}^*[R_m^\gamma]\}^2}\left[\iint_{x\leq y} x^\gamma y^\gamma\psi(x)\log\left(\frac{y}{x}\right)\,\mathrm{d}Q_m(x)\,\mathrm{d}Q_m(y)\right. \\
&\qquad\left. + \iint_{x\leq y} x^\gamma y^\gamma\psi(x)\log\left(\frac{x}{y}\right)\,\mathrm{d}Q_m(x)\,\mathrm{d}Q_m(y)\right] \\
&= 0.
\end{aligned}$$

The only inequality is due to that fact that when $x \geq y$, $\psi(x) \leq \psi(y)$. Thus the lower bound is decreasing with regard to the risk-aversion parameter γ. Similar technique can be applied to $\psi(x) = \boldsymbol{I}(Q_m(x) \geq 1 - Q_i(q))$, an increasing function, which leads to the conclusion that the upper bound is increasing with regard to γ.

Third, notice that the lower bound is such that

$$\frac{\mathbb{E}^* \left[R_m^\gamma \boldsymbol{I}\left(R_m \leq q_l\right)\right]}{\mathbb{E}\left[R_m^\gamma\right]} \leq \frac{q_l^\gamma}{\mathbb{E}^*[R_m^\gamma]}.$$

To show that the lower bound converges to zero, we only need that $\mathbb{E}^*[(R_m/q_l)^\gamma] \to \infty$ as $\gamma \to \infty$. This holds if $\mathbb{P}^*[R_m/q_l > 1] > 0$. If this condition does not hold, $R_m \leq q_l = Q_m^{-1}(Q_i(q))$ with probability one, which violates the assumption that $Q_i(q) < 1$. Thus, $\mathbb{E}^*[(R_m/q_l)^\gamma] \to \infty$ and the lower bound converges to zero as $\gamma \to \infty$.

To show that the upper bound goes to one as $\gamma \to \infty$, note that

$$1 = \frac{\mathbb{E}^*[R_m^\gamma \boldsymbol{I}(R_m < q_u)] + \mathbb{E}^*[R_m^\gamma \boldsymbol{I}(R_m \geq q_u)]}{\mathbb{E}^*[R_m^\gamma]}.$$

As a result, we only need to show that $\frac{\mathbb{E}^*[R_m^\gamma \boldsymbol{I}(R_m < q_u)]}{\mathbb{E}^*[R_m^\gamma]} \to 0$ as $\gamma \to \infty$. Again, this is satisfied when $\mathbb{P}^*[R_m/q_u > 1] > 0$. If this is untrue, $R_m \leq q_u = Q_m^{-1}(1 - Q_i(q))$ with probability one, thus, $1 - Q_i(q) = 1$, violating the assumption that $Q_i(q) > 0$. □

2.7.2 Details about recovering risk-neutral marginals

This section presents implementation details about recovering the risk-neutral marginal distributions from the option prices. At a specific date, let $x_i = K_i$, $i = 1, 2, \ldots$, be the available strike prices of a certain option contract (on a specific underlying with a given maturity); let $y_i = \text{put}(K_i)$ if $K_i \leq R_f S_0$ (i.e., out-of-the-money put prices), and $y_i = \text{call}(K_i) + K_i/R_f - S_0$ if $K_i > R_f S_0$ (i.e., put prices implied by the out-of-the-money call prices under the put-call parity). Treating the (x_i, y_i) pairs as observables, the following nonparametric shape-constrained model is fitted:

$$\min_{f \in \mathcal{F}} \left\{ \sum_i [y_i - f(x_i)]^2 + \frac{1}{2}\lambda \|f\|_2^2 \right\}$$

where $\mathcal{F} = \{f \in \mathcal{C}(\mathbb{R}_+) : f > 0, f' > 0, f'' > 0\}$ is the set of continuous functions defined on $\mathbb{R}_+$ that are both monotonically increasing and convex (to rule out arbitrage opportunities along the moneyness dimension). This nonparametric fitting is implemented via the shape-constrained B-spline basis approach of Pya and Wood (2015). The tuning parameter λ is chosen via standard generalized cross-validation procedures following Pya and Wood (2015).

Based on the nonparametric fitting outcomes, smooth relationship between option prices and strike prices is obtained with refined details. Arbitrage is also ruled out along side the moneyness dimension. Taking derivatives according to (2.6) generates the risk-neutral marginal distribution.

2.7.3 Additional Tables and Figures

Table 2.10: Summary statistics for *realized* correlations

	mean	quantiles				
		5%	25%	50%	75%	95%
Panel A: Average across time (resulting $N = 1055$)						
Correlation	0.508	0.290	0.412	0.497	0.609	0.748
Spearman-ρ	0.498	0.268	0.404	0.489	0.600	0.723
Kendall-τ	0.373	0.194	0.295	0.359	0.452	0.569
Panel B: Average across firms (resulting $T = 312$)						
Correlation	0.475	0.260	0.400	0.480	0.566	0.656
Spearman-ρ	0.466	0.257	0.396	0.473	0.556	0.637
Kendall-τ	0.347	0.186	0.288	0.350	0.418	0.487

Chapter 3

Model Uncertainty in the Cross Section

This chapter is joint work with Jiantao Huang.

Introduction

Recent literature has provided a wide spectrum of real and financial uncertainty measures.[1] They display pronounced time-series variation, and their innovations are associated with business cycle fluctuations through impacts on firms' investment and hiring activities.

A relatively understudied direction is measures of uncertainty that matters for firms' investors. Uncertainty has ambiguous implications for investors' asset allocation decisions. The conventional wisdom of "flight-to-safety" or "flight-to-liquidity" claims that investors respond to large uncertainty shocks by curtailing risk exposures or hoarding liquid assets.[2] However, uncertainty may also arise endogenously during periods of "Schumpeterian growth," when new technologies lead to unforeseeable disruptive industry dynamics. Instead of seeking safety or liquidity, investors respond by chasing glamour stocks in search of a new El Dorado: examples include the "railway mania" in the mid 1840s and the "tech bubble" in the late 1990s.[3]

This paper creates an uncertainty measure for the equity market and examines its implications to investors' asset allocation decisions. Existing equity market uncertainty measures focus on volatilities (as well as their unexpected innovations) of aggregate market returns. These measures do not take into account a fundamental challenge equity investors

[1] Bloom (2009) constructs uncertainty shocks using jumps in the (price) of stock market volatilities. Ludvigson, Ma, and Ng (2021) and Jurado, Ludvigson, and Ng (2015) construct and compare real and financial uncertainty indices. Baker, Bloom, and Davis (2016) develop economic policy uncertainty indices based on news coverage. Manela and Moreira (2017) use textual analysis and machine learning methods to extrapolate the VIX index back in history.

[2] The underlying drivers include institutional redemption pressures (Vayanos, 2004), preferences featuring robustness concerns (Caballero and Krishnamurthy, 2008), and asymmetric information (Guerrieri and Shimer, 2014)

[3] This argument is related to the literature on "growth options", see, for example, Abel (1983); Pástor and Veronesi (2006, 2009).

face: identifying factors that determine the cross section of expected returns. This challenge is more demanding nowadays as investors cohabit with the “factor zoo:” too many factors have been proposed. If we interpret existing financial uncertainty measures as “time-series uncertainty,” a missing dimension is the cross section: in addition to uncertainties about which direction the market is going, there are also enormous uncertainties about which stocks and factors will outperform.

We attempt to bridge this gap by creating a cross-sectional uncertainty measure. Specifically, we take the perspective of a Bayesian investor adopting the linear stochastic discount factor (SDF) models to price assets. Investors are not clairvoyant as they do not know the “true” model. Instead, they learn *both* model parameters and specifications through Bayesian updating.

Our key innovation is a generalized g-prior along the line of Zellner (1986), from which Bayesian investors update their posterior beliefs. As originally exposited in Zellner (1986), the g-prior is a natural prior choice in a sequential decision-making setup. We revisit Zellner’s original arguments and tailor his g-prior for linear SDF models. Our prior specification naturally rules out extremely high Sharpe ratio investment opportunities, guaranteeing the absence of (near) arbitrage as pointed out in Ross (1976a); Cochrane and Saá-Requejo (2000); Kozak, Nagel, and Santosh (2018).

Drawing inferences based on our prior, a Bayesian investor then comes up with well-defined posterior probabilities of asset pricing models. These derived posteriors address the indeterminacy issue induced by “flat” priors as emphasized by Chib, Zeng, and Zhao (2020). Most importantly, the posterior model probabilities have intuitive closed-form solutions; they increase with model-implied (in-sample) Sharpe ratios and decrease with model dimensions. The result crystallizes two competing forces for an asset pricing model to be (optimally) chosen by a Bayesian investor: higher in-sample profits (on paper or in back tests) and model simplicity.

We define cross-sectional uncertainty regarding linear SDF models as the entropy of posterior model probabilities. The intuition is straightforward. Suppose that there are only two candidate factor models, and we are uncertain about which one is true. One extreme case is that the first model dominates the other with a high posterior probability, i.e., 99%. Under this scenario, entropy is close to its lower bound zero (and we are clearly facing low uncertainty). On the contrary, if the two models’ posterior probabilities are 50-50, the entropy reaches its maximum (picking a model boils down to the exercise of coin tossing). To sum up, the higher the entropy measure is, the more uncertain Bayesian investors are about factor models.

We then examine the behaviors and implications of our cross-sectional model uncertainty measure. We document four sets of empirical findings, summarized as follows.

First, we measure uncertainty regarding 14 popular factor strategies in the US stock market. Model uncertainty displays considerable time-series variation and exhibits countercyclical behaviours, as in Figure 3.1. Particularly, model uncertainty increases *before* stock market crashes and peaks under tumultuous market conditions. It reaches its upper bound at the bust of the dot-com bubble and the 2008 global financial crisis. In other words, posterior model probabilities are almost equalized during these two periods: all models are wrong (or right, which does not make any difference). Under extreme market conditions, investors do not only face higher second-moment (volatility) and third-moment (skewness) risk but they are also confronted with higher (if not the highest) model uncertainty, i.e., they are incredibly uncertain about which model can help navigate them out of the storm.

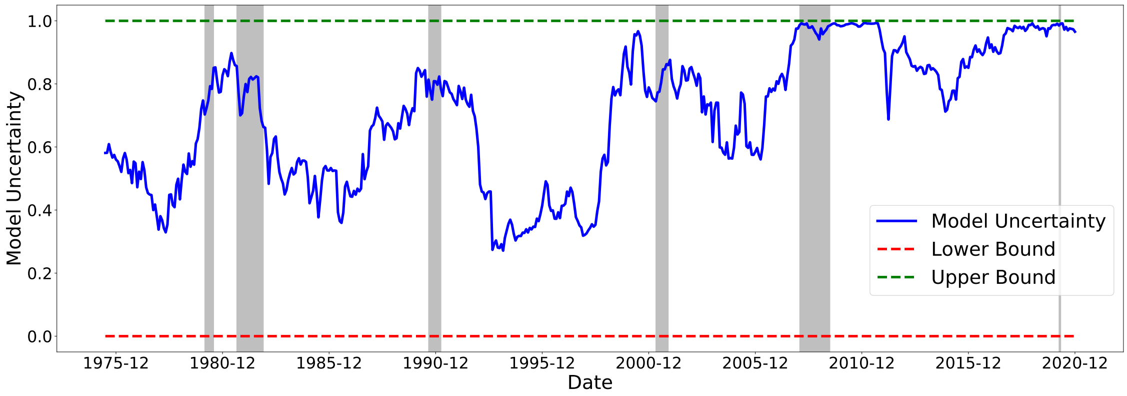


Figure 3.1: Time-Series of Model Uncertainty (3-Year Rolling Window)

The figure plots the time-series of model uncertainty about the linear stochastic discount factor (SDF). We consider 14 prominent factors from the past literature (see Section 3.2 for details). At the end of each month, we compute the posterior model probabilities using the daily factor returns in the past three years. We use the entropy of model probabilities to quantify model uncertainty in the cross-section. The sample ranges from July 1972 to December 2020. Since we use a three-year rolling window, the model uncertainty index starts from June 1975. The red line and green lines show the lower (0) and upper bounds (1) of model uncertainty. Shaded areas are NBER-based recession periods for the US.

We repeat the exercise in European and Asian Pacific stock markets. While the time-series pattern in Europe is roughly the same as the US stock market, the Asian Pacific equity market displays certain unique behaviours. For example, model uncertainty in this market is exceptionally high during the 1997 Asian financial crisis.

Second, we show the time-varying importance of Bayesian model averaging (BMA) in portfolio choice. Following past literature (e.g., Barillas and Shanken (2018)), we use as the criterion the out-of-sample (OOS) Sharpe ratio implied by factor models. We split the full sample into three equal subsamples based on model uncertainty and denoted them as low, middle, and high model uncertainty dates. In particular, we compare BMA with the top one model ranked by posterior model probabilities. The critical observation is that BMA outperforms the top model only in high model uncertainty dates, whereas they have almost identical performance in other periods. Therefore, when model uncertainty is relatively high, investors are better off if they aggregate the information over the space of all models instead of selecting a specific high probability model. Third, model uncertainty is a crucial determinant of mutual fund flows, regardless of being an exogenous cause or a merely propagating mechanism. We adopt the canonical Vector Autoregression (VAR) model to study the dynamic responses of fund flows to uncertainty shocks. Most strikingly, model uncertainty innovations induce sharp outflows from the US equity funds and inflows to US government bond funds, with effects persisting for around three years. These outflows mainly come from small-cap and style funds but not large-cap or sector funds. In addition, we do not observe significant inflows to money market funds, so there is little evidence of "flight-to-liquidity" following high model uncertainty. Hence, investors' asset allocation decisions tend to respond to our uncertainty measure consistent with the conventional wisdom of "flight-to-safety": Facing high cross-sectional uncertainty, they reduce risky asset positions, especially in small-cap stocks and actively-managed (style) funds, and reallocate proceeds into safe assets such as government bonds.

It is also worth noting that similar fund flows patterns do not emerge when using volatility-driven uncertainty indices such as VXO and financial uncertainty. We document some evidence that VXO and financial uncertainty innovations relate to future inflows to money market funds, consistent with "flight-to-liquidity." However, dynamic responses of fund flows to these two uncertainty measures tend to be transitory and sensitive to identification assumptions, while those to model uncertainty shocks are very persistent and robust.

Fourth, we find that high cross-sectional model uncertainty is associated with investors' expectations and confidence about the stock market. We quantify investors' expectations using surveys from the American Association of Individual Investors (AAII) and their confidence levels using the Investor Behavior Project at Yale University. When our uncertainty measure goes up, both individual and institutional investors become more pessimistic about the stock market. More intriguingly, individual investors tend to "react" more aggressively (in terms of pessimism) to our cross-sectional uncertainty measure.

Literature. Our paper belongs to three strands of literature. The first is on new methodologies for factor models and risk premia, with a particular focus on adapting and refining methods from the machine learning and high-dimensional statistics literature (see Giglio, Kelly, and Xiu (2021) for a review of recent advancement). The most related paper to ours is Kozak, Nagel, and Santosh (2020), who carefully examines sparsity under the linear SDF framework. They conclude that a sparse linear SDF model constructed from characteristics-sorted factors cannot explain the cross-sectional variation out-of-the-sample. Complementing their findings, our model uncertainty index further rules out the possibility of any dominant model (even "dense" ones with many factors). However, we also find that a smaller number of parsimonious models do perform well in explaining the cross section of expected returns within specific sample periods.

There is an increasing interest in developing uncertainty measures for both real (e.g., Bloom (2009); Baker et al. (2016); Jurado et al. (2015)) and financial activities (e.g., Manela and Moreira (2017)). Dew-Becker and Giglio (2021) propose a cross-sectional uncertainty measure using long history of option prices, which can be interpreted as uncertainties about general firm outcomes. Our contribution to the literature is to propose a conceptually new index capturing equity investors' uncertainty about factor models describing the cross section of returns.

Our paper also contributes to the literature on Bayesian inferences about factor models and Bayesian portfolio choices (Kandel and Stambaugh, 1996; Barberis, 2000; Pástor, 2000; Avramov, 2002; Barillas and Shanken, 2018; Chib et al., 2020). The use of Zellner's g-priors for parameter uncertainties in time-series return regressions first appears in Kandel and Stambaugh (1996). Avramov (2002) extends the their framework to account for both parameter and model uncertainties. In comparison, we assign the g-prior to factor loadings of linear SDF models. We develop an approach for hyper-prior tuning, addressing the issue of artificially favoring sparse models that may appear in existing Bayesian methods. Our posteriors are analytically tractable, which depend only on in-sample maximal Sharpe ratios and model dimensions, recasting the classical GRS test intuition (Gibbons, Ross, and Shanken, 1989) in light of the factor zoo. Empirically, the existing Bayesian factor model and portfolio choice literature focuses on normative questions: what should a Bayesian investor do (in terms of choosing factors and building portfolios)? We, on the other hand, take a positive perspective and examine implications of the Bayesian view (on model uncertainties) to aggregate flow of funds.

3.1 Theory and method

Throughout our analysis, we focus on the cross section of *excess* returns and their risk premia. Denote by $\boldsymbol{R}$, a random vector of dimension N, the excess returns under consideration.[4] Out of these excess returns, some would be regarded as asset pricing factors that drive the whole cross section of $\mathbb{E}[\boldsymbol{R}]$. Common examples of these factors include the market excess return in the CAPM and long-short portfolios in multi-factor asset pricing models. We use the notation $\boldsymbol{f}$, a subset of p excess returns from $\boldsymbol{R}$ ($p \leq N$), to represent these factors.[5] A linear factor model for excess returns in the stochastic discount factor (SDF) form can be written as (see Chapter 13 of Cochrane (2005) for a detailed exposition):

$$m = 1 - (\boldsymbol{f} - \mathbb{E}[\boldsymbol{f}])^\top \boldsymbol{b}, \tag{3.1}$$

$$\mathbb{E}[\boldsymbol{R} \times m] = \boldsymbol{0}, \tag{3.2}$$

or equivalently,

$$\mathbb{E}[\boldsymbol{R}] = \mathrm{cov}[\boldsymbol{R}, \boldsymbol{f}]\boldsymbol{b}, \tag{3.3}$$

where m is an SDF that prices assets, i.e., it is such that the prices of excess returns all equal zero. Since the pricing equation (3.2) is scale-invariant, we normalize the constant term in the SDF to one. The covariance term, $\mathrm{cov}[\boldsymbol{R}, \boldsymbol{f}]$, is an $N \times p$ matrix.[6] Its entry in the ith row and jth column is the covariance between excess return R_i ($i = 1, \dots, N$)

[4] Excess returns in our context can be returns on risky assets less the risk-free rate, and more generally, returns on long-short portfolio positions with zero initial costs.

[5] We intentionally let the factors $\boldsymbol{f}$ be a subset of excess returns $\boldsymbol{R}$ to enforce that factors themselves are correctly priced, that is, their price being zero, by the factor models we write down next.

[6] The simplified equation (3.3) is due to $\mathrm{cov}[\boldsymbol{R}, \boldsymbol{f}] = \mathbb{E}\left[(\boldsymbol{R} - \mathbb{E}[\boldsymbol{R}])(\boldsymbol{f} - \mathbb{E}[\boldsymbol{f}])^\top\right] = \mathbb{E}\left[\boldsymbol{R}(\boldsymbol{f} - \mathbb{E}[\boldsymbol{f}])^\top\right]$.

and factor f_j $(j = 1 \ldots, p)$.

Remark. Linear factor characterization of SDFs relates to the results of Hansen and Jagannathan (1991): Assuming no arbitrage, an SDF within the space spanned by all tradable excess returns $\boldsymbol{R}$ can be written as

$$m = 1 - (\boldsymbol{R} - \mathbb{E}[\boldsymbol{R}])^\top (\text{var}[\boldsymbol{R}])^{-1} \mathbb{E}[\boldsymbol{R}].$$

Clearly, $\mathbb{E}[m \times \boldsymbol{R}] \equiv 0$ under the specification of m above. This SDF treats all excess returns as asset pricing factors, i.e., $\boldsymbol{f} = \boldsymbol{R}$ and $\boldsymbol{b} = (\text{var}[\boldsymbol{R}])^{-1} \mathbb{E}[\boldsymbol{R}]$ under the specification of equation (3.1).

3.1.1 A simple framework for incorporating model uncertainty

Now we would like to formalize the concept of model uncertainty. A priori, we do not know factors that enter the SDF. As a result, for a given set of p factors $\boldsymbol{f} = [f_1, \ldots, f_p]^\top$, a total number of 2^p models for the linear SDF are possible candidates. To capture uncertainty regarding this pool of models, we index the whole set of 2^p models using a p-dimensional vector of indicator variables $\boldsymbol{\gamma} = [\gamma_1, \ldots, \gamma_p]^\top$, with $\gamma_j = 1$ representing that factor f_j is included into the linear SDF, while with $\gamma_j = 0$ meaning that f_j is excluded. This vector $\boldsymbol{\gamma}$ uniquely defines a model for the SDF,[7] denoted by $\mathcal{M}_{\boldsymbol{\gamma}}$: Under $\mathcal{M}_{\boldsymbol{\gamma}}$, the linear SDF is

$$m_{\boldsymbol{\gamma}} = 1 - (\boldsymbol{f}_{\boldsymbol{\gamma}} - \mathbb{E}[\boldsymbol{f}_{\boldsymbol{\gamma}}])^\top \boldsymbol{b}_{\boldsymbol{\gamma}} \tag{3.4}$$

and the expected excess returns are such that

$$\mathbb{E}[\boldsymbol{R}] = \text{cov}[\boldsymbol{R}, \boldsymbol{f}_{\boldsymbol{\gamma}}]\boldsymbol{b}_{\boldsymbol{\gamma}} \tag{3.5}$$

where $\boldsymbol{f}_{\boldsymbol{\gamma}}$ is a $p_{\boldsymbol{\gamma}}$-dimensional vector that contains all the factors that are included under the current model;[8] $\boldsymbol{b}_{\boldsymbol{\gamma}}$ is a $p_{\boldsymbol{\gamma}}$-dimensional vector, the elements of which are market prices of risk; $\text{cov}[\boldsymbol{R}, \boldsymbol{f}_{\boldsymbol{\gamma}}]$ is now an $N \times p_{\boldsymbol{\gamma}}$ covariance matrix. The two equations above are counterparts of (3.1) and (3.3) after incorporating model uncertainty.

We choose to study model uncertainty under the linear SDF specification for two reasons. First, linear SDF models enable us to focus on the cross section of expected returns. Adding in the time-series dimension, model uncertainty has been introduced to panel regressions of returns on characteristics (e.g., Avramov (2002)) or asset pricing factors (e.g., Barillas and Shanken (2018) and Chib et al. (2020)). Model uncertainties under these settings blend in information regarding time-series predictability.

Most importantly, linear factor models in the SDF form enable us to ask the following question: Does one set of factors drive out another? As in Cochrane (2005, p. 261), the elements of $\boldsymbol{b}$ address the question of "should I include factor j given the other factors?"

[7] For notation simplicity, we use "$-\boldsymbol{\gamma}$" to denote the set of factors that are excluded from now on. That is, it is always the case that elements in vector $\boldsymbol{f}$ are unions of elements in $\boldsymbol{f}_{\boldsymbol{\gamma}}$ and $\boldsymbol{f}_{-\boldsymbol{\gamma}}$, and the intercept of elements in $\boldsymbol{f}_{\boldsymbol{\gamma}}$ and $\boldsymbol{f}_{-\boldsymbol{\gamma}}$ is empty.

[8] $p_{\boldsymbol{\gamma}} = \sum_{j=1}^{p} I[\gamma_j = 1]$ is the total number of factors that are included under model $\mathcal{M}_{\boldsymbol{\gamma}}$.

For js such that $b_j = 0$, the answer would be "no", which maps directly into our model uncertainty framework.

Another object of interests is the factor risk premia $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_p]^\top$. Under model $\mathcal{M}_{\boldsymbol{\gamma}}$, $\boldsymbol{\lambda} = \mathrm{cov}[\boldsymbol{f},\, \boldsymbol{f}_{\boldsymbol{\gamma}}]\boldsymbol{b}_{\boldsymbol{\gamma}}$. Clearly, for factors that *do not* enter the SDF (their market prices of risk $\boldsymbol{b}_{-\boldsymbol{\gamma}} = \boldsymbol{0}$), their risk premia $\boldsymbol{\lambda}_{-\boldsymbol{\gamma}}$ are not necessarily zero. Thus, there is no clear theory guidance to introducing the latent variable $\boldsymbol{\gamma}$ for the risk premia. Knowing whether factors' risk premia equal zero or not *does not* help distinguish factor models. Then, of course, to capture the uncertainty regarding factor models, we need to account for the uncertainty regarding whether the elements in $\boldsymbol{b}$ are zero or not.

3.1.2 Prior specification and empirical Bayes inference

We now present a Bayesian framework to understand and quantify model uncertainty in the cross-section of expected stock returns, under the linear SDF setting. With observed data for excess returns, denoted by $\mathcal{D} = \{\boldsymbol{R}_t\}_{t=1}^T$, our primary goal is to evaluate the probability of each model $\mathcal{M}_{\boldsymbol{\gamma}}$ given the observed data $p[\mathcal{M}_{\boldsymbol{\gamma}} \mid \mathcal{D}]$. Bayesian inference offers a natural way of computing these posterior model probabilities.

We (as have many others) assume that the observed excess returns are generated from a multivariate Gaussian distribution:

$$\boldsymbol{R}_1, \ldots, \boldsymbol{R}_T \overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{\mu},\, \boldsymbol{\Sigma}). \tag{3.6}$$

The linear SDF model $\mathcal{M}_{\boldsymbol{\gamma}}$ then sets a restriction on this distribution through the following moment condition:

$$\boldsymbol{\mu} = \boldsymbol{C}_{\boldsymbol{\gamma}}\boldsymbol{b}_{\boldsymbol{\gamma}}, \tag{3.7}$$

where $\boldsymbol{C}_{\boldsymbol{\gamma}} = \mathrm{cov}[\boldsymbol{R}, \boldsymbol{f}_{\boldsymbol{\gamma}}]$ consists of a subset of columns in $\boldsymbol{\Sigma}$. We adopt an empirical Bayes strategy by treating the variance-covariance matrix $\boldsymbol{\Sigma}$ as known initially to derive the posterior model probability $p[\mathcal{M}_{\boldsymbol{\gamma}} \mid \mathcal{D}]$, and then substituting this matrix with a moment estimator[9].

Now we proceed to assign priors for $\boldsymbol{b}_{\boldsymbol{\gamma}}$. Our prior specification is motivated by the g-prior proposed by Arnold Zellner (see Zellner (1986)). We assume that *conditional* on choosing model $\mathcal{M}_{\boldsymbol{\gamma}}$,

$$\boldsymbol{b}_{\boldsymbol{\gamma}} \mid \mathcal{M}_{\boldsymbol{\gamma}} \sim \mathcal{N}\left(\boldsymbol{0},\, \frac{g}{T}\left(\boldsymbol{C}_{\boldsymbol{\gamma}}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{C}_{\boldsymbol{\gamma}}\right)^{-1}\right), \quad g > 0 \tag{3.8}$$

where T is the sample size for the observed excess returns. The parameter g is related to the effective sample size or level of uncertainty for an "conceptual or imaginary sample" according to Zellner (1986).

[9] Empirical Bayes approaches use data to facilitate prior assignments. Here although the matrix $\boldsymbol{\Sigma}$ is a likelihood parameter, it also enters the prior for $\boldsymbol{b}_{\boldsymbol{\gamma}}$, as will become clear next when we introduce our prior specification. Thus we are still using data to pin down (hyper)parameters in the priors. The use of moment estimators to replace parameters in the prior distributions dates all the way back to the seminal James-Stein estimator (James and Stein (1961)). For a monograph on modern empirical Bayes methods, see Efron (2012).

Following the reasoning of Zellner (1986), we generalize the original g-prior and adapt it to our specific setting. Before making inference about different linear SDF models using the observed excess return data $\mathcal{D}$, we consider an "imaginary" sample of size T', denoted by $\mathcal{D}' = \{\boldsymbol{R}'_t\}_{t=1}^{T'}$, where the sample size is allowed to be different from T by a scalar g such that $T' = T/g$. This parameter g also governs level of uncertainty about our imaginary sample relative to the data sample we have [10]. Under model $\mathcal{M}_{\boldsymbol{\gamma}}$, excess returns observed in this sample are distributed as follows: $\boldsymbol{R}'_1, \ldots, \boldsymbol{R}'_{T'} \overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{C}_{\boldsymbol{\gamma}}\boldsymbol{b}_{\boldsymbol{\gamma}}, \boldsymbol{\Sigma})$. Assigning a non-informative prior on $\boldsymbol{b}_{\boldsymbol{\gamma}}$, which is flat everywhere[11], we can derive the "posterior" of $\boldsymbol{b}_{\boldsymbol{\gamma}}$ given this conceptual data sample as $[\boldsymbol{b}_{\boldsymbol{\gamma}} \mid \mathcal{M}_{\boldsymbol{\gamma}}, \mathcal{D}'] \sim \mathcal{N}\left(\boldsymbol{b}'_{\boldsymbol{\gamma}}, g/T \times \left(\boldsymbol{C}_{\boldsymbol{\gamma}}^{\top}\boldsymbol{\Sigma}^{-1}\boldsymbol{C}_{\boldsymbol{\gamma}}\right)^{-1}\right)$, where the posterior mean $\boldsymbol{b}'_{\boldsymbol{\gamma}}$ is related to the particular hypothetical data set $\mathcal{D}'$ in mind, while the posterior variance is not (a celebrated result for conditional normal distributions). This leaves the posterior mean $\boldsymbol{b}'_{\boldsymbol{\gamma}}$ largely undetermined for we can have infinite degrees of freedom "imagining" the data set $\mathcal{D}'$. If we would like to use this posterior as our prior for $\boldsymbol{b}_{\boldsymbol{\gamma}}$, resorting to the Bayesian philosophy that "today's posterior is tomorrow's prior" (Lindley, 2000), we at lease need to find a way of determining $\boldsymbol{b}'_{\boldsymbol{\gamma}}$, the current posterior mean.

Zellner (1986) relies on the rational expectation hypothesis to pin down $\boldsymbol{b}'_{\boldsymbol{\gamma}}$. Suppose that we have an anticipatory value for $\boldsymbol{b}_{\boldsymbol{\gamma}}$, denoted by $\boldsymbol{b}^a_{\boldsymbol{\gamma}}$, in addition to the imaginary sample $\mathcal{D}'$ (as well as the initial diffuse prior for $\boldsymbol{b}_{\boldsymbol{\gamma}}$). The rational expectation hypothesis says that $\boldsymbol{b}^a_{\boldsymbol{\gamma}} = \mathbb{E}[\boldsymbol{b}_{\boldsymbol{\gamma}} \mid \mathcal{M}_{\boldsymbol{\gamma}}, \mathcal{D}'] = \boldsymbol{b}'_{\boldsymbol{\gamma}}$. Now we have a reference informative prior distribution that does not depend on the hypothetical sample, which is

$$\boldsymbol{b}_{\boldsymbol{\gamma}} \mid \mathcal{M}_{\boldsymbol{\gamma}} \sim \mathcal{N}\left(\boldsymbol{b}^a_{\boldsymbol{\gamma}}, \frac{g}{T}\left(\boldsymbol{C}_{\boldsymbol{\gamma}}^{\top}\boldsymbol{\Sigma}^{-1}\boldsymbol{C}_{\boldsymbol{\gamma}}\right)^{-1}\right).$$

To determine whether a model $\mathcal{M}_{\boldsymbol{\gamma}}$ is sensible or not, we are basically testing $H_0 : \boldsymbol{b}_{\boldsymbol{\gamma}} = \boldsymbol{0}$ versus $H_1 : \boldsymbol{b}_{\boldsymbol{\gamma}} \in \mathbb{R}^{p_{\boldsymbol{\gamma}}}$. These tests help us distinguish between different models as model $\mathcal{M}_{\boldsymbol{\gamma}}$ already impose the condition that $\boldsymbol{b}_{-\boldsymbol{\gamma}} = \boldsymbol{0}$. Following the suggestion of Zellner (1986), we set $\boldsymbol{b}^a_{\boldsymbol{\gamma}} = \boldsymbol{0}$, that is, the anticipatory expectations are the values under the null. This finally gives us the prior specification in (3.8).

Remark. One might attempt to assign an objective prior, such as the Jeffreys prior,

[10] In Zellner (1986), the scalar g is used to capture the fact that the variance of the hypothetical sample can be different from the variance of the sample under study. These two arguments (effective sample size v.s. variance of the hypothetical data set) are isomorphic because they will lead to the same g-prior specification. Our sample-size based arguments echo the ideas of factional and intrinsic Bayes factor in the mid 90's (see O'Hagan (1995) and Berger and Pericchi (1996)), which aim to "transform" improper priors to proper ones. Similar ideas for specifying priors are adopted in the paper by Shmuel Kandel and Robert F. Stambaugh in the finance literature to discipline the specification of informative priors Kandel and Stambaugh (1996).

[11] This flat prior is non-informative in the sense that it is a Jeffreys prior, a common notion of prior objectiveness or non-informativeness in Bayesian analysis Jeffreys (1946). Under our setting, we treat $\boldsymbol{\Sigma}$ as known. As a result, Jeffreys prior for $\boldsymbol{b}_{\boldsymbol{\gamma}}$ is proportional to a constant, i.e., it is flat. Of remark, this flatness outcome is not true if the covariance matrix is unknown, under which the Jeffreys prior would specify that the joint density of $\pi(\boldsymbol{b}_{\boldsymbol{\gamma}}, \boldsymbol{\Sigma})$ is proportional to $\boldsymbol{\Sigma}^{-\frac{N+2}{2}}$. Some existing work (e.g. Barillas and Shanken (2018)) specifies a prior such that $\pi(\boldsymbol{b}_{\boldsymbol{\gamma}}, \boldsymbol{\Sigma}) \propto \boldsymbol{\Sigma}^{-\frac{N+1}{2}}$, which is the so-called independence Jeffreys prior (*not* the original Jeffreys-rule prior) imposing the assumption that $\boldsymbol{b}_{\boldsymbol{\gamma}}$ and $\boldsymbol{\Sigma}$ are independent at the prior level.

to $\boldsymbol{b}_{\boldsymbol{\gamma}}$. In this case, it is an improper flat prior as we have discussed early on. This would be desirable without model uncertainty, for it will lead to proper posterior distributions. However, with model uncertainty, improper priors can only be assigned to *common* parameters across models, which is clearly not the case for $\boldsymbol{b}_{\boldsymbol{\gamma}}$. Otherwise, posterior model probability would be indeterminate. This is a well-known result in Bayesian statistics and has also been pointed out in the finance literature (e.g., Cremers (2002)).

Our g-prior specification in (3.8) leads to a surprisingly simple expression for the variance of the SDF, which is summarized in Proposition 12.

Proposition 12. *Under model $\mathcal{M}_{\boldsymbol{\gamma}}$, in which $m_{\boldsymbol{\gamma}} = 1 - \left(\boldsymbol{f}_{\boldsymbol{\gamma}} - \mathbb{E}[\boldsymbol{f}_{\boldsymbol{\gamma}}]\right)^{\top} \boldsymbol{b}_{\boldsymbol{\gamma}}$, the g-prior specification for $\boldsymbol{b}_{\boldsymbol{\gamma}}$ implies that*

$$\text{var}[m_{\boldsymbol{\gamma}} \mid g] = \frac{g p_{\boldsymbol{\gamma}}}{T}.$$

According to Proposition 12, volatility of the SDF ($= \sqrt{g p_{\boldsymbol{\gamma}}/T}$) under a certain model is determined by the conditionality of that model, at least at the prior level. The renowned Hansen-Jagannathan bound states that this volatility (times the gross risk-free rate) sets an upper bounds on any achievable Shape ratios in the economy Hansen and Jagannathan (1991); Cochrane and Saá-Requejo (2000) regards portfolio positions with high Sharpe ratios as deals that are too good to be realized in the market. These arguments imply that models with too many factors are not likely to be realistic *a priori*.

The g-prior offers us an analytically tractable framework to make posterior inference. Under the g-prior, we can integrate out $\boldsymbol{b}_{\boldsymbol{\gamma}}$ and calculate the marginal likelihood of observing the excess return data $\mathcal{D}$ based on each model. All these marginal likelihoods are available in closed form and results are collected in Proposition 13.

Proposition 13. *The marginal likelihood of observing excess return data $\mathcal{D}$ under model $\mathcal{M}_{\boldsymbol{\gamma}}$ is*

$$\mathbb{P}[\mathcal{D} \mid \mathcal{M}_{\boldsymbol{\gamma}}, g] = \exp\left\{-\frac{T-1}{2}\text{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}\right) - \frac{T}{2}\left(\text{SR}^2_{\text{max}} - \frac{g}{1+g}\text{SR}^2_{\boldsymbol{\gamma}}\right)\right\} \frac{(1+g)^{-\frac{p_{\boldsymbol{\gamma}}}{2}}}{(2\pi)^{\frac{NT}{2}}|\boldsymbol{\Sigma}|^{\frac{T}{2}}},$$

where

$$\boldsymbol{S} = \frac{1}{T-1}\sum_{t=1}^{T}(\boldsymbol{R} - \overline{\boldsymbol{R}})(\boldsymbol{R} - \overline{\boldsymbol{R}})^{\top},$$

is the in-sample variance-covariance matrix for the excess returns; SR^2_{max} *is the maximal squared Sharpe ratio achievable from forming portfolios using all excess returns under consideration;* $\text{SR}^2_{\boldsymbol{\gamma}}$ *is the maximal squared Sharpe ratio from combining all factors under model $\mathcal{M}_{\boldsymbol{\gamma}}$. These two Sharpe ratios are both in-sample values and it is always the case that* $\text{SR}^2_{\boldsymbol{\gamma}} \leq \text{SR}^2_{\text{max}}$ *for all $\boldsymbol{\gamma}$.*

Proposition 13 has a couple of implications. To begin with, we can calculate the marginal likelihood for a very special model, the null model, in which $\boldsymbol{\gamma} = \boldsymbol{0}$. SDF $m_{\boldsymbol{\gamma}}$ in this case is a constant, characterizing a risk-neutral market. Under this setup, $p_{\boldsymbol{\gamma}}$

equals zero because no factors are included, and the maximal squared Sharpe ratio $\mathrm{SR}^2_{\boldsymbol{\gamma}}$ is also zero. Plugging these two quantities into the expression in Proposition 13, we have $p[\mathcal{D} \mid \mathcal{M}_{\mathbf{0}}, g] \equiv p[\mathcal{D} \mid \mathcal{M}_{\mathbf{0}}]$, because the posterior marginal likelihood under the null model does not depend on the scalar g. The Bayes factor, which is the ratio between marginal likelihoods under two different models, that compares model $\mathcal{M}_{\boldsymbol{\gamma}}$ with the null model $\mathcal{M}_{\mathbf{0}}$ is

$$\begin{aligned} \mathrm{BF}_{\boldsymbol{\gamma}}(g) &= \frac{p[\mathcal{D} \mid \mathcal{M}_{\boldsymbol{\gamma}}, g]}{p[\mathcal{D} \mid \mathcal{M}_{\mathbf{0}}]} \\ &= \exp\left\{ \frac{Tg}{2(1+g)} \mathrm{SR}^2_{\boldsymbol{\gamma}} - \frac{p_{\boldsymbol{\gamma}}}{2} \log(1+g) \right\}. \end{aligned} \tag{3.9}$$

This Bayes factor can be regarded as evidence of model $\mathcal{M}_{\boldsymbol{\gamma}}$ against the null model. To further compare two arbitrary models $\mathcal{M}_{\boldsymbol{\gamma}}$ and $\mathcal{M}_{\boldsymbol{\gamma}'}$, we can calculate the Bayes factor

$$\begin{aligned} \mathrm{BF}_{\boldsymbol{\gamma},\boldsymbol{\gamma}'}(g) &= \frac{\mathrm{BF}_{\boldsymbol{\gamma}}(g)}{\mathrm{BF}_{\boldsymbol{\gamma}'}(g)} \\ &= \exp\left\{ \frac{Tg}{2(1+g)} \left(\mathrm{SR}^2_{\boldsymbol{\gamma}} - \mathrm{SR}^2_{\boldsymbol{\gamma}'}\right) - \frac{p_{\boldsymbol{\gamma}} - p_{\boldsymbol{\gamma}'}}{2} \log(1+g) \right\}, \end{aligned} \tag{3.10}$$

which is, by definition, the (marginal) likelihood ratio $p[\mathcal{D} \mid \mathcal{M}_{\boldsymbol{\gamma}}, g]/p[\mathcal{D} \mid \mathcal{M}_{\boldsymbol{\gamma}'}, g]$. A large Bayes factor $\mathrm{BF}_{\boldsymbol{\gamma},\boldsymbol{\gamma}'}(g)$ lends evidence to favor model $\mathcal{M}_{\boldsymbol{\gamma}}$ against model $\mathcal{M}_{\boldsymbol{\gamma}'}$.

A first observation based on equation (3.10) is that although the marginal likelihood in Proposition 13 depends on the test assets (the pre-specified set of excess returns that define $\boldsymbol{R}$), the Bayes factors do not. The Bayes factors are only determined by the in-sample time series of the factors that enter the linear SDF, through the maximal Sharpe ratios and the number of factors. A key assumption driving this outcome is that factors are a subset of the testing assets. In other words, the linear factor SDF model must price the factors themselves correctly. This finding is reminiscent of the observation that, when estimating factor risk premia in linear factor models, the efficient GMM objective function assigns zero weights to the testing assets except for the factors entering the SDF (See for example, (Cochrane, 2005, Page 244-245)).

The Bayes factor above illustrates a clear trade-off when comparing models. With the number of factors fixed, models in which factors can generate larger in-sample Sharpe ratios are always preferred. This echoes the intuitions behind the GRS tests in Gibbons et al. (1989), which show the link between time-series tests of the factor models and the mean-variance efficiency of factor portfolios. Under our setting, when the factor portfolios deliver large maximal Sharpe ratios, it is evidence that they are more likely to span the excess return space, thus favoring the linear SDF constructed from these factors. On the other hand, it is a simple mechanical phenomenon that maximal Sharpe ratio $\mathrm{SR}_{\boldsymbol{\gamma}}$ increases as additional assets are added into the factor portfolio. Thus the penalty term on model dimensionality $p_{\boldsymbol{\gamma}}$ imposed by the g-prior plays an key role in preventing the Bayes factor to favor large models blindly. In order to properly penalize large models, g cannot be too small, as $\mathrm{SR}_{\boldsymbol{\gamma}}$ always increases after one augments the linear SDF.

Perhaps the most desirable feature of our Bayes factor calculation in equation (3.10) is that it helps us understand the aforementioned trade-off quantitatively. When model dimension is increased by one ($p_{\boldsymbol{\gamma}} - p_{\boldsymbol{\gamma}'} = 1$), the maximal squared Sharpe ratio (times the sample size T) of the factor portfolio has to increase by at least $(1+g)/g \times \log(1+g)$ to lend support to the augmented model, that is,

$$T\left(\mathrm{SR}^2_{\boldsymbol{\gamma}} - \mathrm{SR}^2_{\boldsymbol{\gamma}'}\right) > \frac{1+g}{g}\log(1+g).$$

However, it is always the case that $T\left(\mathrm{SR}^2_{\boldsymbol{\gamma}} - \mathrm{SR}^2_{\boldsymbol{\gamma}'}\right) \leq T\mathrm{SR}^2_{\max}$. Then for g large enough, the inequality above will always be violated, as the function $(1+g)/g \times \log(1+g)$ is monotonically increasing and unbounded. As a result, smaller models will always be supported by the Bayes factor. Under the extreme case that $g \to \infty$, from equation (3.9), $\mathrm{BF}_{\boldsymbol{\gamma}}(g) \to 0$. Paradoxically, the most favorable model will always be the null model. The case under which $g \to \infty$ corresponds to the conventional diffuse priors; and the fact that, with model uncertainty, diffuse priors always support the null model is sometimes called the Bartlett's paradox (Bartlett (1957)). Of note, this paradox poses another refutation to the use of improper diffuse priors under model uncertainty, in addition to posterior indeterminacy that has been pointed out earlier.

3.1.3 A prior for the parameter g

Discussions above point to the subtlety of choosing the parameter g. Instead of plugging in particular numbers for g, a natural way under our Bayesian framework is to integrate out g with a proper prior for it. A prior on g, namely $\pi[g]$, is equivalent to assigning a scale-mixture of g priors for $\boldsymbol{b}_{\boldsymbol{\gamma}}$. This idea is adapted from Liang et al. (2008), who argues that this type of mixture priors provides more robust posterior inference. As a result, our g prior specification will be modified to

$$\pi[\boldsymbol{b}_{\boldsymbol{\gamma}} \mid \mathcal{M}_{\boldsymbol{\gamma}}] \propto \int_0^\infty \mathcal{N}\left(\boldsymbol{b}_{\boldsymbol{\gamma}} \,\middle|\, \boldsymbol{0},\, \frac{g}{T}\left(\boldsymbol{C}_{\boldsymbol{\gamma}}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{C}_{\boldsymbol{\gamma}}\right)^{-1}\right) \pi[g]\, \mathrm{d}g, \tag{3.11}$$

where the prior for g is such that

$$\pi[g] = \frac{a-2}{2}(1+g)^{-\frac{a}{2}}, \quad g > 0.$$

This prior $\pi[g]$ is improper when $a \leq 2$. A special case when $a = 2$ corresponds to the Jeffreys prior in Liang et al. (2008). Because the marginal likelihood of the null model does not depend on g (recall that $p[\mathcal{D} \mid \mathcal{M}_{\boldsymbol{0}},\, g] \equiv p[\mathcal{D} \mid \mathcal{M}_{\boldsymbol{0}}]$), improper priors will lead to indeterminacy in the ratio

$$\begin{aligned}\mathrm{BF}_{\boldsymbol{\gamma}} &= \frac{\int_0^\infty p[\mathcal{D} \mid \mathcal{M}_{\boldsymbol{\gamma}},\, g]\pi[g]\,\mathrm{d}g}{\int_0^\infty p[\mathcal{D} \mid \mathcal{M}_{\boldsymbol{0}}]\pi[g]\,\mathrm{d}g} \\ &= \int_0^\infty \frac{p[\mathcal{D} \mid \mathcal{M}_{\boldsymbol{\gamma}},\, g]}{p[\mathcal{D} \mid \mathcal{M}_{\boldsymbol{0}}]}\pi[g]\,\mathrm{d}g \qquad (3.12)\end{aligned}$$

up to an arbitrary constant, which is the Bayes factor under the new mixture of g prior specification. Thus we force $a > 2$.

This additional prior on g also leads to refinements on the volatility of the SDF. Based on the result from Proposition 12, the unconditional volatility of the SDF for model $\mathcal{M}_{\boldsymbol{\gamma}}$ must satisfy

$$\text{var}[m_{\boldsymbol{\gamma}}] \geq \mathbb{E}[\text{var}[m_{\boldsymbol{\gamma}} \mid g]] = \frac{p_{\boldsymbol{\gamma}}}{T}\mathbb{E}[g].$$

The prior $\pi[g]$ is such that $\mathbb{E}[g] = \infty$ if $a \leq 4$, and that $\mathbb{E}[g] = 2/(a-4)$ if $a > 4$. To make sure that the variance of the SDF does not explode, we need $a > 4$. And if we follows the argument of Cochrane and Saá-Requejo (2000) to set an upper limit on the maximal achievable Sharpe ratio in the economy[12], denoted by SR_∞, then

$$R_f^2\text{SR}_\infty^2 = \text{var}[m_{\boldsymbol{\gamma}}] \geq \mathbb{E}[\text{var}[m_{\boldsymbol{\gamma}} \mid g]] = \frac{2p_{\boldsymbol{\gamma}}}{T(a-4)},$$

where R_f represents the risk-free rate. For the investor in the economy to be not risk-neutral, the SDF must include at least one factor, that is, $p_{\boldsymbol{\gamma}} \geq 1$ (for example, under the CAPM world). As a result, we will require that

$$a \geq 4 + \frac{2}{TR_f^2\text{SR}_\infty^2}.$$

Another way of looking at our prior for g is that it is equivalent to

$$\frac{g}{1+g} \sim \text{Beta}\left(1,\, \frac{a}{2} - 1\right).$$

This ratio is crucial in that it determines the contribution of data evidence when making posterior inferences. It is sometimes referred to as the “shrinkage factor”. To see this more clearly, we can calculate the posterior of the cross-sectional expected return $\boldsymbol{\mu} = \boldsymbol{C}_{\boldsymbol{\gamma}}\boldsymbol{b}_{\boldsymbol{\gamma}}$, which is given as follows

$$\mathbb{E}[\boldsymbol{\mu} \mid \mathcal{M}_{\boldsymbol{\gamma}},\, g,\, \mathcal{D}] = \frac{g}{1+g}\boldsymbol{C}_{\boldsymbol{\gamma}}\left\{\text{var}[\boldsymbol{f}_{\boldsymbol{\gamma}}]\right\}^{-1}\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{f}_{\boldsymbol{\gamma},t}\right).$$

Under all models, the posterior mean of expected returns are scaled by a fixed factor $g/(1+g) \in (0,1)$. Our prior specification is equivalent to a Beta distribution for this shrinkage factor, and the prior mean for it is

$$\mathbb{E}\left[\frac{g}{1+g}\right] = \frac{2}{a} \leq \frac{1}{2 + \left(TR_f^2\text{SR}_\infty^2\right)^{-1}}.$$

In order to give enough credit to the data-driven estimates and avoid over-shrinkage, we choose the smallest possible a such that $\mathbb{E}\left[g/(1+g)\right]$ is as large as possible *a priori*, that is, we pick $a = 4 + 2/(TR_f^2\text{SR}_\infty^2)$. Under this choice, the prior expectation for the shrinkage

[12] Note that this must be larger than the maximal in-sample Sharpe ratio of portfolios formed using excess returns under our consideration, denoted by $\text{SR}_{\max}$ in Proposition 13.

factor is still strictly smaller than one half, but can be very close (the ratio $2/(TR_f^2\mathrm{SR}_\infty^2)$ is usually very small).

3.1.4 Posterior probability of models

We can integrate out the parameter g according to equation (3.12) to find the Bayes factors under the mixture of g-priors. Proposition 14 presents the results.

Proposition 14. *The Bayes factor for comparing model $\mathcal{M}_{\boldsymbol{\gamma}}$ with the null model $\mathcal{M}_{\mathbf{0}}$ is*

$$\mathrm{BF}_{\boldsymbol{\gamma}} = \left(\frac{a-2}{2}\right)\exp\left(\frac{T}{2}\mathrm{SR}_{\boldsymbol{\gamma}}^2\right)\left(\frac{T}{2}\mathrm{SR}_{\boldsymbol{\gamma}}^2\right)^{-s_{\boldsymbol{\gamma}}}\underline{\Gamma}\left(s_{\boldsymbol{\gamma}},\ \frac{T}{2}\mathrm{SR}_{\boldsymbol{\gamma}}^2\right),$$

where

$$\underline{\Gamma}(s,x) = \int_0^x t^{s-1}e^{-t}\,\mathrm{d}t$$

is the lower incomplete Gamma function (Abramowitz and Stegun, 1965, Page 263); the scalar $s_{\boldsymbol{\gamma}}$ is defined as

$$s_{\boldsymbol{\gamma}} = \frac{p_{\boldsymbol{\gamma}}+a}{2} - 1.$$

This Bayes factor is always increasing in $\mathrm{SR}_{\boldsymbol{\gamma}}^2$ always decreasing in $p_{\boldsymbol{\gamma}}$.

The Bayes factor that compares any two models can be computed as

$$\mathrm{BF}_{\boldsymbol{\gamma},\boldsymbol{\gamma}'} = \frac{\mathrm{BF}_{\boldsymbol{\gamma}}}{\mathrm{BF}_{\boldsymbol{\gamma}'}},$$

which is the same as what we have done earlier. Bayes factors decide the posterior odds of one model against another:

$$\frac{\mathbb{P}[\mathcal{M}_{\boldsymbol{\gamma}} \mid \mathcal{D}]}{\mathbb{P}[\mathcal{M}_{\boldsymbol{\gamma}'} \mid \mathcal{D}]} = \frac{\pi[\mathcal{M}_{\boldsymbol{\gamma}}]}{\pi[\mathcal{M}_{\boldsymbol{\gamma}'}]} \times \mathrm{BF}_{\boldsymbol{\gamma},\boldsymbol{\gamma}'}.$$

Equivalently, the posterior odds give us the posterior model probabilities: for model $\mathcal{M}_{\boldsymbol{\gamma}}$, its posterior probability given the excess return data is

$$\mathbb{P}[\mathcal{M}_{\boldsymbol{\gamma}} \mid \mathcal{D}] = \frac{\mathrm{BF}_{\boldsymbol{\gamma}}\pi[\mathcal{M}_{\boldsymbol{\gamma}}]}{\sum_{\boldsymbol{\gamma}}\mathrm{BF}_{\boldsymbol{\gamma}}\pi[\mathcal{M}_{\boldsymbol{\gamma}}]},$$

which is a direct outcome of the Bayes' rule. We can then define a model uncertainty measure as the entropy of the posterior model probabilities:

$$\mathrm{entropy}[\mathcal{M}_{\boldsymbol{\gamma}} \mid \mathcal{D}] = \sum_{\boldsymbol{\gamma}} \log(p[\mathcal{M}_{\boldsymbol{\gamma}} \mid \mathcal{D}])p[\mathcal{M}_{\boldsymbol{\gamma}} \mid \mathcal{D}]. \tag{3.13}$$

Roughly speaking, larger entropy corresponds to higher model uncertainty. For example, suppose that we have only two candidate models. If one of them has a posterior model probability of 99%, we should be confident about this high-probability model. Actually, the model uncertainty is almost zero in this scenario. However, if the posterior probability of each model is around 50%, then choosing the true model is equivalent to flipping a fair coin. In this case, the model uncertainty in equation (3.13) is maximized.

3.2 Data

In our primary empirical implementation, we combine 14 prominent factors from the past literature and measure model uncertainty in this small zoo of factors. First, we include notable Fama-French five factors (Fama and French (2016)) plus the momentum factor (Jegadeesh and Titman (1993)). In addition, we consider the q-factor model from Hou et al. (2015) and include their size, investment, and profitability factors. The factor models mentioned earlier are based on rational asset pricing theory. Taking the insights from behavioural models, Daniel et al. (2020) propose a three-factor model consisting of the market factor, the short-term behavioural factor (PEAD), and the long-term behavioural factor (FIN). Finally, we include the HML devil, the quality-minus-junk factor, and the betting-against-beta factor from the AQR library. Appendix 3.9.1 presents the detailed description of these factors.

Table 3.7 reports the annualised mean returns and Sharpe ratios of 14 factors. First, most of them (except for two size factors) have enormous Sharpe ratios in the full sample from July 1972 to December 2020. In particular, the short-term behavioural factor (PEAD) seems to be the most profitable historically. Furthermore, I split the entire sample into two equal subsamples. Consistent with past literature (e.g., McLean and Pontiff (2016)), the performance of many factor strategies decline significantly from subsample one to two. Most strikingly, the annualised Sharpe ratio of the value factors has plunged from above 0.9 to nearly zero in the second subsample. This observation suggests that we should focus on the out-of-sample instead of the in-sample Sharpe ratio in evaluating factor models.

With the estimate of model uncertainty, we next compare it with other uncertainty measures and economic variables. Bloom (2009) uses the jumps in VXO/VIX indices as the stock market uncertainty shock. We download the time-series of VXO/VIX indices from Wharton Research Data Services (WRDS). Baker et al. (2016) develop indices of economic policy uncertainty (EPU), which can be downloaded from Nick Bloom's website. Other uncertainty measures that we use include the macro, real and financial uncertainty measures in Ludvigson et al. (2021) and Jurado et al. (2015). We download them from the authors' websites. In addition, we compare our model uncertainty with the intermediary factor from He et al. (2017), the term yield spread (the yield on ten-year government bonds minus the yield on three-month treasury bills), and the credit spread (the yield on BAA corporate bonds minus the yield on AAA corporate bonds). We download the intermediary factor from the authors' websites and the bond yields from the Federal Reserve Bank of St. Louis.

Moreover, we obtain mutual fund data from the Center for Research in Security Prices

(CRSP) survivorship-bias-free mutual fund database.[13] In particular, we are interested in monthly mutual fund flows, so we download the monthly total net assets, monthly fund returns, and the codes of fund investment objectives. To normalise the aggregate fund flows, we divide the equity (fixed-income) fund flows across all funds within a particular investment objective by the total market capitalisation of all listed companies in CRSP (US GDP). In addition, we download the total market value of all US-listed stocks from CRSP.[14]

Finally, we study the relationship between our model uncertainty measure and investors' expectations about future stock market performance. In our paper, we use the survey data from the American Association of Individual Investors (AAII) survey and Shiller's survey conducted by the International Center for Finance at the University of Yale. We download the related data from their official websites.

3.3 Measuring model uncertainty

We now adopt the perspective of Bayesian investors and construct the time series of model uncertainty. At the end of each month, we use all daily factor returns in the past three years to estimate the posterior model probabilities, $p[\mathcal{M}_{\boldsymbol{\gamma}} \mid \mathcal{D}]$, and compute the entropy as in equation (3.13). We choose the hyper-parameter a to be four in the benchmark case. We also present the results obtained from alternative rolling windows and other choices of a in robustness checks (see Section 3.7).

The behavioural factors in Daniel et al. (2020) are available only from July 1972, and we use 36-month data in the estimation, so the model uncertainty measure starts from June 1975. Since some factors are highly correlated, we consider models that contain at most one version of the factors in each of the following categories: (a) size (SMB or ME); (b) profitability (RMW or ROE); (c) value (HML or HML Devil); (d) investment (CMA or IA). We refer to size, profitability, value, and investment as categorical factors. Therefore, there are ten effective factors, including market, size, profitability, value, investment, short-term and long-term behavioural factors, momentum, QMJ, and BAB.

The blue line in Figure 3.1 plots the time series of model uncertainty of linear SDFs, and the sample period spans from June 1975 to December 2020. The red and green dotted lines show the lower and upper bounds of model uncertainty, respectively. The lower entropy bound is always zero, i.e., when there is one dominant model with the posterior model probability of 100%. On the contrary, uncertainty is maximized when the posterior model probabilities are equalized across all models. Because we have 14 factors, and only one of the categorical factors could be selected into the true model, there are 5,184 different candidate models.[15] The upper bound of model uncertainty is around

[13] All variables are calculated (or derived) based on data from database name ©CRSP survivor-bias-free Mutual Funds, Center for Research in Security Prices (CRSP®), The University of Chicago Booth School of Business.

[14] All variables are calculated (or derived) based on data from database name ©CRSP Monthly Stock, Center for Research in Security Prices (CRSP®), The University of Chicago Booth School of Business.

[15] The model in our framework is indexed by $\boldsymbol{\gamma}$: $\gamma_j \in \{0, 1\}$ and $\gamma_j = 1$ implies that the factor j should

8.55.[16] To normalize the model uncertainty index, we divide it by 8.55. Hence, the upper bound is one in Figure 3.1.

The model uncertainty index has several interesting features that could shed light on the nature of uncertainty about the linear SDF. First, we observe a surprisingly high level of model uncertainty. Specifically, the average (median) model uncertainty is around 0.70 (0.75), with the first and third quartiles equal to 0.53 and 0.87, respectively. Hence, most of the time, Bayesian investors are not confident about the true SDF model. Second, model uncertainty fluctuates significantly over time. In particular, the index varies from the lowest value of 0.27 to the highest 0.99, representing economic states in which Bayesian investors find it almost unlikely to determine the true SDF model. The standard deviation of the index is 0.21. Overall, model uncertainty is a dynamic phenomenon. Finally, model uncertainty is persistent by construction since we use a rolling window of 36 months in the estimation. The first-order autocorrelation is 0.98, and the autocorrelation coefficients strictly decrease in time lags, with insignificant autocorrelations after 30 lags.

Figure 3.1 also suggests the countercyclical nature of model uncertainty. In particular, the 1990s was a remarkable period: it was remembered as a period of strong economic growth, low inflation and unemployment rate, and high stock returns. During the 1990s, model uncertainty is the lowest across our sample. As the orange dots in Figure 3.2 suggest, posterior probabilities of the top two models are significantly larger than others. Hence, investors are relatively confident about the true SDF model.

In addition, peaks in model uncertainty tend to coincide with major events in the US stock markets and economy. Important examples include the dot-com crash in 2000 and the global financial crisis in 2008 when model uncertainty almost touches its upper bound. Specifically, the blue dots in Figure 3.2 show that posterior probabilities of the top 50 models, in December 2007, are almost equalized. In other words, it is virtually infeasible to distinguish models based on the observed data. The 2008 crisis is noteworthy because model uncertainty stays at a high level for a prolonged period. In contrast, it declines shortly after other crises/recessions. In the recent five years, model uncertainty has slowly increased from 0.7 to 1 at the end of 2020.

Interestingly, we do not observe a spike in model uncertainty during the 1987 flash crash. The potential reason is that the 1987 market crash was not long-lasting. Even though S&P 500 index declined by more than 20% in one day, the crisis was not caused by any economic recession, and the market recovered rapidly. Instead, the leading cause was synchronous program trading, illiquidity in the market, and the subsequent market panic. Since our uncertainty measure is based on past-three-year daily data, the impact of short-term market chaos is averaged out.

In conclusion, our model uncertainty measure displays considerable time-series variation: it is particularly sizable in bad economic states. The stock market crash that lasts

be included into true SDF. We do not have restrictions on the market, short-term reversal, long-term reversal, momentum, QMJ, and BAB, so the number of models for these 6 factors is 2^6. For SMB and ME, we only allow three cases: (0,0), (1,0) or (0,1). Therefore, each categorical factor has 3 (instead of 4) possibilities. The total number of candidate models equals $2^6 \times 3^4 = 5184$.

[16] upper bound $= -\sum_{\gamma} \frac{1}{5184} \times log(\frac{1}{5184}) = log(5184) \simeq 8.55$..

only for a short period, such as the 1987 flash crash, is not captured by our model uncertainty measure. Furthermore, the cyclical behaviours of model uncertainty imply another layer of investment risk: when investors experience bear stock markets, they are also the most uncertain about the true model in the cross-section, or equivalently, which portfolio of factor strategies they should hold. This further motivates us to study how model uncertainty relates to investors' portfolio choices and expectations. We investigate these topics in section 3.4 and 3.5.

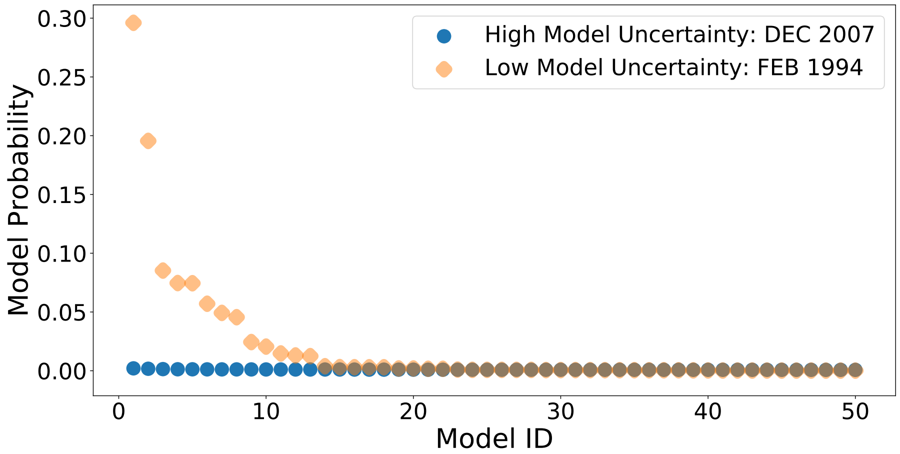


Figure 3.2: Posterior Probabilities of Top 50 models: High vs. Low Model Uncertainty

The figure plots the posterior probabilities of the top 50 models ranked by their posterior probabilities. At the end of each month, we compute the posterior model probabilities using the daily factor returns in the past three years. We use the entropy of model probabilities to quantify model uncertainty in the cross-section. We observe low model uncertainty in February 1994 (orange diamonds) but high model uncertainty in December 2007 (blue dots).

3.3.1 Does model uncertainty matter?

Should investors take into account model uncertainty in the cross-section? A natural hypothesis is that model uncertainty plays a more critical role when it is more sizable. The logic is as follow. When model uncertainty is relatively low, the factor model with the highest model probability dominates others, such as the orange diamonds in Figure 3.2. Hence, investors are more willing to trust the top model ranked by the Bayesian posterior probabilities. In contrast, the top model is not informative if model uncertainty is relatively high, such as during market crashes. In this case, they may prefer to aggregate the information over the space of all models.

The Bayesian model averaging (BMA) is one common approach to aggregating models. It enables us to flexibly model investors' uncertainty about potentially relevant factors. In the SDF model, we are interested in the risk prices, $\boldsymbol{b}$. The BMA of $\boldsymbol{b}$ is defined as

$$\boldsymbol{b}_{bma} := \mathbb{E}[\boldsymbol{b} \mid \mathcal{D}] = \sum_{\gamma} \mathbb{E}[\boldsymbol{b} \mid \mathcal{M}_{\gamma}, \mathcal{D}] \times P(\mathcal{M}_{\gamma} \mid \mathcal{D}). \tag{3.14}$$

Rather than considering the expectation of $\boldsymbol{b}$ conditional on a specific model, we take the weighted average of the model-implied expectations, where the weights are posterior model probabilities. Intuitively, models with high probabilities are more influential in BMA.

BMA deviates sharply from the traditional model selection, in which researchers always use a particular criterion (e.g., adjusted R2, model probabilities, etc.) to select a single model and presume that the selected model is correct. Past literature also shows the importance of model averaging in asset pricing (e.g., Avramov (2002), Bryzgalova et al. (2021), Avramov et al. (2021)).

We now compare the performance of BMA with the top Bayesian model. The performance metric that we use is the out-of-sample (OOS) Sharpe ratio of factor models. We also compare our Bayesian procedure with several candidate models: (1) All 14 factors (All), (2) Carhart (1997) four-factor model (Carhart4), (3) Fama and French (2016) five-factor model (FF5), (4) Hou et al. (2015) q-factor model (HXZ4), and (5) Daniel et al. (2020) behavioural factor model (DHS3).

For each factor model $\boldsymbol{\gamma}$ in month t, we estimate the risk prices of $\boldsymbol{f_\gamma}$ via the standard GMM estimation: $\hat{\boldsymbol{b}}_{\boldsymbol{\gamma}} = (\text{var}[\boldsymbol{f_\gamma}])^{-1}(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{f_{\gamma t}})$, where the covariance matrix and mean returns of $\boldsymbol{f_\gamma}$ are estimated using the data from month $t-35$ to month t, consistent with Figure 3.1. The tangency portfolio conditional on model $\boldsymbol{\gamma}$ is $\hat{\boldsymbol{b}}_{\boldsymbol{\gamma}}^{\top}\boldsymbol{f_{\gamma,t+1}}$, and the BMA tangency portfolio is $\boldsymbol{b}_{bma}^{\top}\boldsymbol{f_{t+1}}$.[17] We update the tangency portfolio each month.[18]

We also test the null hypothesis that BMA and the model $\boldsymbol{\gamma}$ have an identical Sharpe ratio, i.e., H_0 : $\text{SR}^2_{bma} = \text{SR}^2_{\gamma}$, using the non-parametric Bootstrap. Under H_0, the expected return of the tangency portfolio implied by the model $\boldsymbol{\gamma}$ is linear in that of BMA: $\mathbb{E}[R^{\gamma}_t] = \mathbb{E}[R^{bma}_t]\sigma(R^{\gamma}_t)/\sigma(R^{bma}_t)$. We adjust the average return of R^{γ}_t using the previous equality and draw 100,000 sample paths of $\{R^{\gamma}_{t^\star}, R^{bma}_{t^\star}\}^{T}_{t^\star=1}$ with replacement, where T is the sample size in the observed dataset. If the difference in Sharpe ratios between BMA and model $\boldsymbol{\gamma}$ in the observed dataset is larger than 90% (95%, 99%) of those in simulated datasets, we claim that H_0 is rejected by the data at 10% (5%, 1%) significance level.

We start with describing the full-sample performance, as shown in the first row of Table 3.1. First, our Bayesian procedure successfully selects the model that outperforms traditional factor models in the out-of-sample. The top Bayesian model (see column (2)) has an OOS Sharpe ratio of 1.75, which is virtually comparable to the model composed of all 14 factors (see column (3)). Second, BMA beats the top Bayesian model. The performance gain is statistically significant, thought its economic magnitude being relatively small.

One may be concerned that these 14 factors are data-mined, so choosing the top model only reflects data snooping rather than the outperformance of our Bayesian procedure. We further split the whole sample into three equal subsamples to tackle this concern. Consistent with past literature, the performance of factor models tends to decline over time, and

[17] For model $\boldsymbol{\gamma}$, we scale the tangency weights $\hat{\boldsymbol{b}}_{\boldsymbol{\gamma}}$ each month such that the target monthly portfolio volatility is 1% based on historical data from month $t-35$ to month t.

[18] Moreover, the top Bayesian model (with the highest model probability) is time-varying.

Table 3.1: Out-of-Sample Model Performance

	(1) BMA	(2) Top 1	(3) All	(4) Carhart4	(5) FF5	(6) HXZ4	(7) DHS3
Full Sample: 07/1975 - 12/2020	1.818	1.750	1.772	0.736	0.938	1.135	1.639
	-	**	-	***	***	***	-
Subsample I: 07/1975 - 08/1990	2.327	2.226	2.293	1.014	1.589	1.853	2.142
	-	**	-	***	***	*	-
Subsample II: 09/1990 - 10/2005	2.094	2.145	2.095	0.927	0.916	1.222	2.072
	-	-	-	***	***	***	-
Subsample III: 11/2005 - 12/2020	1.106	0.940	0.986	0.317	0.452	0.517	0.795
	-	**	-	***	***	**	*
Low Model Uncertainty	2.572	2.565	2.568	1.288	1.624	1.829	2.282
	-	-	-	***	***	***	-
Middle Model Uncertainty	1.717	1.653	1.771	0.450	0.677	1.232	1.818
	-	-	-	***	***	**	-
High Model Uncertainty	1.251	1.125	1.106	0.564	0.584	0.552	0.897
	-	*	*	***	***	***	**

This table reports the out-of-sample (annualised) Sharpe ratio of (1) BMA: the Bayesian model averaging of factor models, (2) Top 1: the top Bayesian model ranked by posterior model probabilities, (3) All: include all 14 factors, (4) Carhart4: Carhart (1997) four-factor model, (5) FF5: Fama and French (2016) five-factor model, (6) HXZ4: Hou et al. (2015) q-factor model, and (7) DHS3: the market factor plus two behavioural factors in Daniel et al. (2020). We also report the results on testing the null hypothesis that the Sharpe ratio of BMA is equal to the model $\boldsymbol{\gamma}$, i.e., $H_0 : \mathrm{SR}^2_{bma} = \mathrm{SR}^2_{\boldsymbol{\gamma}}$. We use the non-parametric Bootstrap to test the null hypothesis. *, ** and *** denote significance at the 90%, 95%, and 99% level, respectively.

the drops in Sharpe ratios are particularly enormous from subsample II (September 1990 - October 2005) to subsample III (November 2005 - December 2020). In addition, BMA is more valuable in the third subsample: its Sharpe ratio (1.106) is significantly higher than other models except for the one composed of all 14 factors.

Whether the performance of factor models is related to model uncertainty? The short answer is yes. On average, the performance of factor models declines as model uncertainty increases. Specifically, when model uncertainty is low, both the top model and BMA have similar Sharpe ratios of around 2.57, which are exceptionally high. In other words, investors should be confident about the top model chosen by our Bayesian procedure in low uncertainty states. On the contrary, it is particularly beneficial to incorporate model uncertainty into portfolio choice when model uncertainty is high. As the last row suggests, BMA has an OOS Sharpe ratio of 1.25, significantly larger than any other specifications.

In summary, there are two takeaways from Table 3.1. First, our Bayesian procedure is competent to pick the model that has satisfactory OOS performance. Second, model uncertainty matters and is particularly noteworthy when it is relatively high. In this scenario, BMA, which aggregates the information across all models, is salient for real-time portfolio choice.

3.3.2 Decomposing model uncertainty

The posterior model probabilities (see Proposition 14) are closely related to the model-implied squared Sharpe ratio, SR_{γ}^2. As we include more factors, the in-sample SR_{γ}^2 always rises. Only when a few factor models dominate others can we be confident about the true model. In other words, when the distances in SR_{γ}^2 are sizable across different factor models, we can easily differentiate them and observe low model uncertainty. In contrast, when factor models have similar SR_{γ}^2, model uncertainty tends to be high.

Figure 3.3 plots the time-series of distances in SR_{γ}^2. More precisely, we show the difference between the maximal SR_{γ}^2 and the $90th$-quantile of SR_{γ}^2, as well as the difference between its maximum and median. T4he difference in SR_{γ}^2 decreases obviously before the stock market crashes and remains at a low level during the bear markets. For example, the distance between the highest and medium in-sample SR_{γ}^2 is close to 0.2 (daily) between 1997 and 1998, but it plunges to almost 0 from 1998 to 2000. After the tech bubble, factor models have been becoming more similar in terms of in-sample SR_{γ}^2.

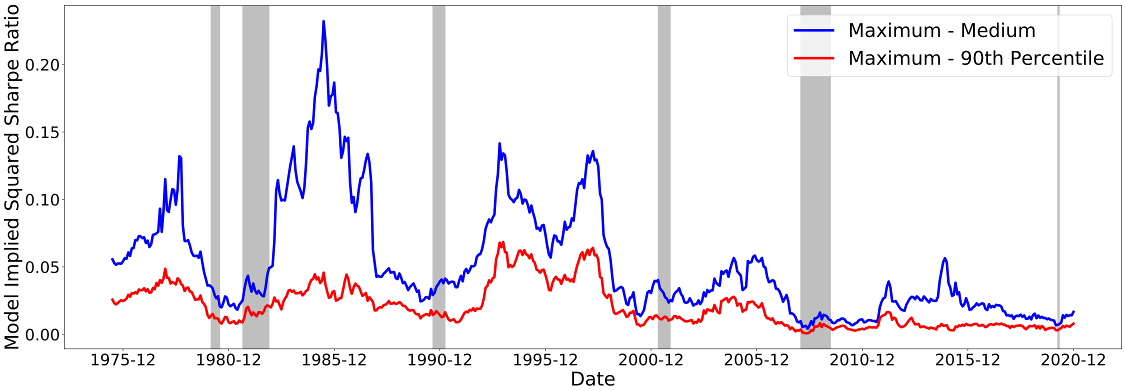
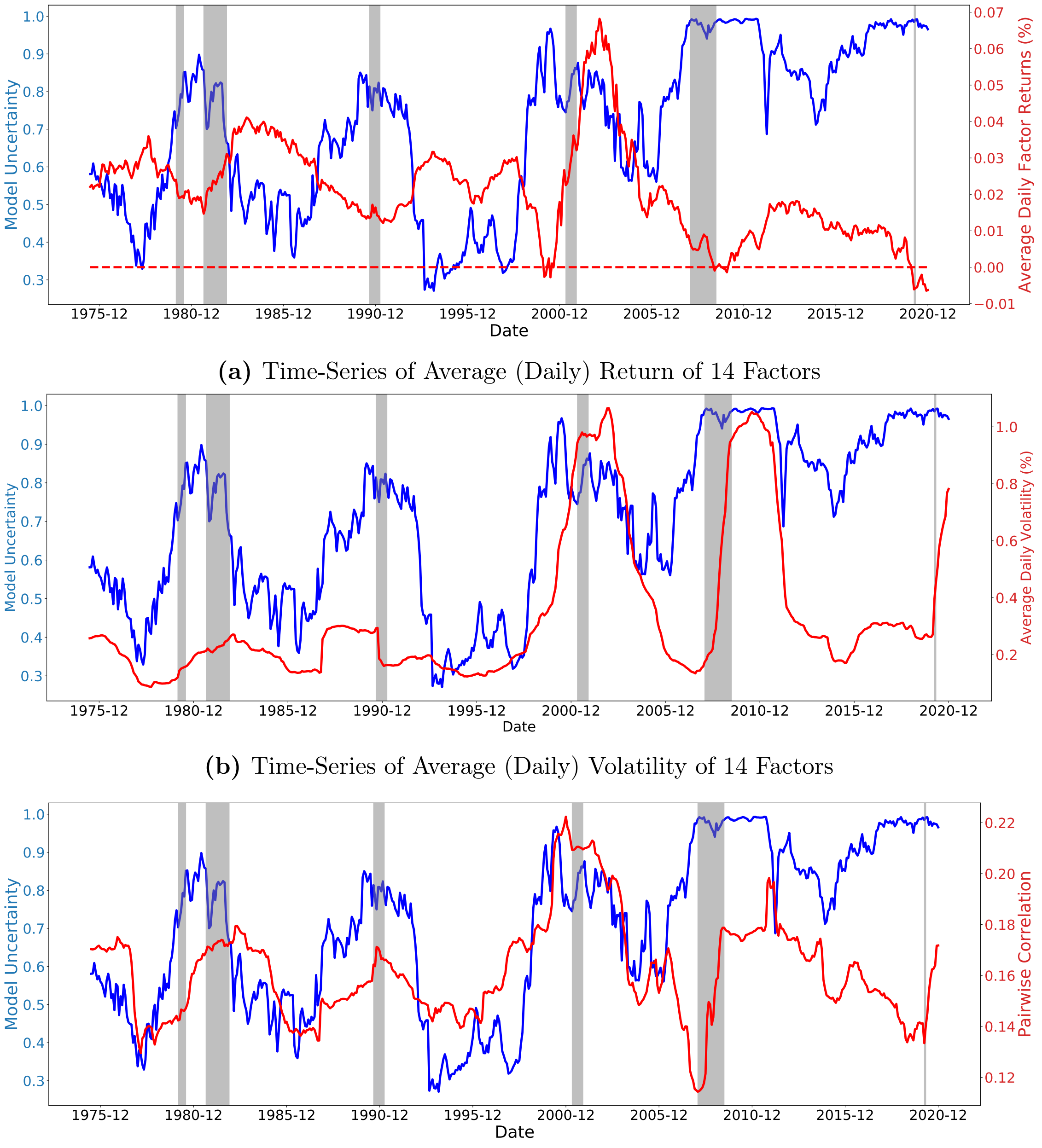


Figure 3.3: Time-Series of Model-Implied Squared Sharpe Ratio (3-Year Rolling Window)

The figure plots the time series of distances in SR_{γ}^2 from June 1975 to December 2020. We present the difference between the highest SR_{γ}^2 and the 90th-quantile of SR_{γ}^2, as well as the difference between the highest SR_{γ}^2 and medium SR_{γ}^2. SR_{γ}^2 is the model-implied squared Sharpe ratio, $\mathbb{E}_T[\boldsymbol{f}_{\gamma}]^{\intercal}\boldsymbol{V}_{\gamma}^{-1}\mathbb{E}_T[\boldsymbol{f}_{\gamma}]$. $\mathbb{E}_T[\boldsymbol{f}_{\gamma}]$ and $\boldsymbol{V}_{\gamma}$ are estimated using the daily factor returns in the past 36 months.

Theoretically, SR_{γ}^2 is determined by mean returns of factors and their covariance matrix. We further analyze SR_{γ}^2 by dipping into three parts: (a) average daily factors returns in the past three years; (b) average daily factor volatility in the past three years; (c) average pairwise correlation among daily factor returns in the past three years. Figure 3.4 plots these time series.

In Figure 3.4a, we show that the average daily return of all 14 factors is incredibly volatile. The average daily return also exhibits cyclical patterns. Specifically, it declines during the run-ups of stock markets. However, it plummets to the bottom during the market crash and recovers gradually after the bear markets. In the recent three most influential market crashes (dot-com bubble, 2008 global financial crisis, and the Covid-19), the average factor returns decline to near zeros. In the past decade, the profitability of these 14 factors is no longer comparable to their historical performance. One potential reason is that more investors implement the same investment strategies after the publication of these factors (see McLean and Pontiff (2016)).



(a) Time-Series of Average (Daily) Return of 14 Factors

(b) Time-Series of Average (Daily) Volatility of 14 Factors

(c) Time-Series of Average Pairwise Correlation of 14 Factors

Figure 3.4: Decomposing the Model Uncertainty

The figures plot the time-series of (a) average daily returns of factors, (b) average daily factor volatility, and (c) average pairwise (absolute) correlation among daily factor returns in the past three years, and these statistics are estimated using the daily factor returns in the past 36 months.

Figure 3.4b plots the average volatility of 14 factors. Even though the average factor volatility increases in the bear markets, the factor returns before the dot-com bubble are not as volatile as after 2000. Typically, the average standard deviation of 14 factors is between 0.2% and 0.4%. During the dot-com bubble and recent global financial crisis, it surges to higher than 1% daily. However, it is evident from figure 3.4b that model uncertainty does not have the same time-series pattern as the average factor volatility.

During market crashes, it is highly likely that arbitrageurs who invest in these factor strategies will exit the market simultaneously, thus driving up comovements among factors. Since the correlation matrix of factors determines the extent to which investors can diversify their investment, it could potentially influence the distances in SR^2_γ. To illustrate this point, we plot the time series of the average pairwise correlation of 14 factors.[19] The average correlation exhibits a similar cyclical pattern as model uncertainty. However, there are two key differences: (a) the average correlation decreases before the 2008 crisis while our model uncertainty starts to climb up from 2006, and (b) model uncertainty increases from 2015 to 2019, while the average correlation among factors declines during the same period.

To sum up, model uncertainty is high when the distances in SR^2_γ among different factor models are low. Since the in-sample SR^2_γ always increases with more factors included, we are uncertain about whether to include an additional factor if the benefit of including it is only marginal. Furthermore, model uncertainty about linear SDFs increases dramatically during the run-ups and stands at the peak during bear markets because different factor models are highly analogous.

3.3.3 Correlation with other economic variables

Figure 3.1 indicates that model uncertainty increases during times of extreme uncertainty in the financial markets and economy. A natural question is how our model uncertainty index correlates with a number of key financial and macroeconomic variables known as capturing critical financial and economic fluctuations.

There are several notable uncertainty measures in the literature. The first measure is VXO/VIX index[20] (used in Bloom (2009)), which quantifies forward-looking market volatility. Subsequent to Bloom (2009), Ludvigson et al. (2021) and Jurado et al. (2015) develop the real, macro and financial uncertainty measures by exploiting a large set of macro and financial variables.[21] Baker et al. (2016) use the coverage of economic or policy-related keywords in the media as proxies for economic policy uncertainty.

[19] At the end of each month t, we use daily factor returns from month $t-35$ to month t to compute the pairwise correlation between any two factors, denoted as ρ_{ij}. The average is computed as $\frac{1}{N\times(N-1)}\sum_{i\neq j}|\rho_{ij}|$.

[20] VIX and VXO index are essentially the same: the correlation between them is higher than 0.98.

[21] They quantify the h-period ahead uncertainty by the extent to which a particular set of economic variables (either real, macro, or financial) become more or less predictable from the perspective of economic agents. Suppose there is a set of economic indicators, $\boldsymbol{Y_t} = (y_{1t}, ..., y_{Lt})^\intercal$. For each variable, they find the conditional volatility of the prediction errors: $u_{jt}(h) = \sqrt{E[(y_{j,t+h} - E[y_{j,t+h}|I_t])^2|I_t]}$. The aggregate uncertainty is quantified by the average conditional volatility of the prediction error of each economic indicator: $u_t(h) = \sum_{j=1}^{L}\omega_j u_{jt}(h)$, where ω_j is the weight on the j-th economic indicator. The detailed econometric framework could be found in the original papers. Our paper considers their one-period ahead uncertainty measures.

In addition to uncertainty measures, we compare model uncertainty with the intermediary factor from He et al. (2017), the term yield spread (the yield on ten-year government bonds minus the yield on three-month treasury bills), and the credit spread (the yield on BAA corporate bonds minus the yield on AAA corporate bonds).

We report in Table 3.2 the results from the regression of model uncertainty on its one-period lag and some contemporaneous economic variables. By running these regressions, we do not intend to study the causal relationship between model uncertainty and other economic variables. Instead, our objective is to describe the contemporaneous relation between them. We also want to point out that model uncertainty is persistent[22] since it is constructed in a rolling window of 36 months. Therefore, we need to be careful in statistical inference. In all following tables, we use Newey-West standard errors (see Newey and West (1987)) with 36 lags in the regressions involving model uncertainty.

As Table 3.2 shows, a number of economic variables are significantly related to model uncertainty, even after we control one-period lagged entropy in the regressions. For example, model uncertainty is positively correlated with financial uncertainty and the VXO index but almost orthogonal to real, macro, and two economic policy uncertainty measures. This finding is intuitive since model uncertainty mainly refines information in financial markets. In addition, the intermediary factor and term yield spread negatively relate to model uncertainty. In column (10), we run horse racing among the VXO index, the intermediary factor, and term yield spread: While the coefficient estimates of the VXO index and term yield spread still remain significant, the intermediary factor becomes inconsequential.

Comments. Conceptually, our model uncertainty index quantifies a different layer of uncertainty from other measures. The stock market volatility, proxied by the VXO index, measures the second-moment investment risk. Three uncertainty measures in Ludvigson et al. (2021) and Jurado et al. (2015) are essentially volatilities of prediction errors. In other words, they measure the dispersion of unexpected changes in economic indicators. Two economic policy uncertainty indices in Baker et al. (2016) are to quantify public attention to economic policy. In contrast, our paper quantifies model uncertainty about linear SDFs. Since we know the lower and upper bounds of entropy, we can easily detect the degree of model uncertainty in the cross-section. For example, model uncertainty reaches its upper bound in some periods, implying that different models' posterior probabilities are almost identical. In short, our model uncertainty index is complementary to other uncertainty measures developed in the past literature. More importantly, ours provides a new angle of analyzing and understanding investment uncertainty.

[22] Strong persistence of the time-series process is ubiquitous in other uncertainty measures. Table 3.8 shows the AR(1) coefficients of the other six uncertainty sequences, and we find that the real, macro and financial uncertainty measures also have AR(1) coefficients less than but close to 1. It is well-known that the volatility of asset returns tends to cluster. When we run the AR(1) for the VXO index, the coefficient estimate of ρ is 0.812. Only the second economic policy uncertainty measure (EPU_2) suffers less from massive autocorrelations.

Table 3.2: Regressions of Model Uncertainty on Contemporaneous Variables

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Lagged Entropy	0.979***	0.982***	0.983***	0.985***	0.983***	0.983***	0.986***	0.985***	0.986***	0.983***
	(128.85)	(142.76)	(146.97)	(106.55)	(105.75)	(129.25)	(161.37)	(150.06)	(154.04)	(131.19)
Financial Uncertainty	0.212*									
	(1.95)									
Macro Uncertainty		0.174								
		(1.53)								
Real Uncertainty			0.140							
			(1.20)							
EPU I				0.000						
				(0.33)						
EPU II					0.000					
					(1.07)					
VIX/VXO						0.005**				0.004**
						(2.20)				(2.34)
Intermediary Factor							-0.503**			-0.196
							(-2.01)			(-0.71)
Term Spread								-0.034***		-0.033**
								(-3.44)		(-2.44)
Default Spread									-0.003	
									(-0.09)	
Sample size	546	546	546	432	432	420	546	546	546	420

The table reports the results from the regression of model uncertainty on its one-period lag and some contemporaneous economic variables (X_{t+1}):

$$Entropy_{t+1} = \beta_0 + \beta_1 Entropy_t + \rho X_{t+1} + \epsilon_{t+1}.$$

X_{t+1} include a) financial, macro, and real uncertainty measures from Ludvigson et al. (2021) and Jurado et al. (2015) in columns (1) - (3), b) two economic policy uncertainty (EPU) indices from Baker et al. (2016) in columns (4) and (5), c) VXO index in column (6), d) the intermediary factor from He et al. (2017) in column (7), e) term spread in column (8), f) default spread in column (9), and g) VXO index, the intermediary factor, and the term spread in column (10). The t-statistics are computed using Newey-West standard errors with 36 lags. *, ** and *** denote significance at the 90%, 95%, and 99% level, respectively.

3.4 Mutual fund flows

If investors consider model uncertainty a crucial source of investment risk, a natural prediction is that their portfolio choice decisions are related to our model uncertainty measure. The difficulty in empirical tests arises due to the lack of observations in their complete portfolio choice. To tackle this issue, we rely on mutual fund flows, which have been studied extensively by the past literature due to their availability. Also, the mutual fund sector is one of the largest financial intermediaries through which individual investors participate in the US stock markets. Hence, we use mutual fund flows as proxies for investors' portfolio rebalancing and study how mutual fund investors react to model uncertainty shocks.

The data is available on CRSP survivor-bias-free US mutual fund database. The database includes investment style or objective codes from three different sources over the whole life of the database.[23] The CRSP style code consists of up to four letters. For example, a fund with the style "EDYG" means that i) this fund mainly invests in domestic equity markets (E = Equity, D = Domestic), and ii) it has a specific investment

[23] From 1962 to 1993, Wiesenberger objective codes are used. Strategic insight objective codes are populated between 1993 and 1998. Lipper objective codes start in 1998. Instead of using the three measures mentioned above directly, CRSP builds its objective codes based on them.

style "Growth" (Y = Style, G = Growth).[24] The quality of data before 1991 is low because the CRSP investment objective code is incomplete. For example, only domestic equity "style" funds and mixed fixed income and equity funds are recorded before 1991. Also, the market values of institutional holdings proportional to the total market value of all stocks (in CRSP) were tiny. Therefore, we focus on the sample from January 1991 to December 2020.

To begin with, we define the aggregate mutual fund flows. Following the literature (see Lou (2012)), we calculate the net fund flows to each fund i in period t as

$$Flow_{i,t} = TNA_{i,t} - TNA_{i,t-1} \times (1 + RET_{i,t}) \tag{3.15}$$

where $TNA_{i,t}$ and $RET_{i,t}$ are total net assets and gross returns of fund i in period t. Next, we aggregate individual fund flows in each period across all funds in a specific group (e.g. all large-cap funds) and scale the aggregate flows by the lagged total market capitalization of all stocks in CRSP:

$$Flows_t^Y = \frac{\sum_{i \in Y} Flow_{i,t}}{\text{CRSP-Market-Cap}_{t-1}}, \tag{3.16}$$

where Y specifies a certain investment objective, such as small-cap funds.

We use the canonical Vector Autoregression (VAR) model to study the dynamic responses of fund flows to model uncertainty shocks. Specifically, we consider the following reduced-form VAR(l) model:

$$\boldsymbol{Y}_t = \boldsymbol{B}_0 + \boldsymbol{B}_1 \boldsymbol{Y}_{t-1} + \cdots + \boldsymbol{B}_l \boldsymbol{Y}_{t-l} + \boldsymbol{u}_t, \tag{3.17}$$

where l denotes the lag order, $\boldsymbol{Y}_t$ is a $k \times 1$ vector of economic variables, $\boldsymbol{u}_t$ is a $k \times 1$ vector of reduced-form innovations with the covariance matrix $\boldsymbol{\Sigma}_u$, and $(\boldsymbol{B}_0, \boldsymbol{B}_1, \ldots, \boldsymbol{B}_l)$ are the coefficient matrices.

Past literature often relates reduced-form innovations to structural shocks, i.e., $\boldsymbol{u_t} = \boldsymbol{S\epsilon_t}$, where $\boldsymbol{S}$ is a $k \times k$ non-singular matrix, and $\boldsymbol{\epsilon_t}$ is a $k \times 1$ vector of structural shocks, which are orthogonal to each other by definition. We use the Cholesky decomposition to identify the dynamic responses to uncertainty shocks, so the ordering of economic variables in $\boldsymbol{Y_t}$ is equivalent to different identification assumptions, which are specified below.

3.4.1 Aggregate equity vs fixed-income funds

Since our model uncertainty measure is based on factors in the US, we delete all foreign mutual funds. In the baseline analysis, we consider the aggregate mutual fund flows to the entire equity and fixed-income markets. That is, we study the VAR regression in equation (3.17), where $\boldsymbol{Y_t}^\top = (Entropy_t, Flows_t^{FI}, Flows_t^{Equity})$. We next use impulse response functions (IRFs) to better understand the dynamic effects and propagating mechanisms of uncertainty shocks.

IRFs greatly depend on the identification assumption, i.e., whether model uncertainty

[24] More details are in the handbook of CRSP survivor-bias-free US mutual fund database.

is an exogenous source of fluctuations in fund flows or an endogenous response. In the first case, model uncertainty is a cause of fund flows, while it acts as a propagating mechanism in the latter case. Without taking a strong stance on the identification assumption, we aim to investigate the dynamic relationship between fund flows and several uncertainty measures, either as a cause or propagating mechanism. To make as few assumptions as possible, we focus only on the dynamic responses to uncertainty shocks and are silent on how innovations in fund flows affect model uncertainty. This simplification allows us to ignore the ordering of other economic variables beyond model uncertainty.

In the benchmark case, we place model uncertainty first in the VAR. Hence, the implicit identification assumption is that fund flows react to the contemporaneous uncertainty shocks, while model uncertainty does not respond to the shocks to mutual funds in the current period. We consider a different identification assumption in robustness checks in Section 3.7; that is, we put model uncertainty as the last element in $\boldsymbol{Y_t}$. As shown below, the IRFs to model uncertainty shocks are essentially robust to the alternative identification strategy, whereas the IRFs to other uncertainty measures are not.

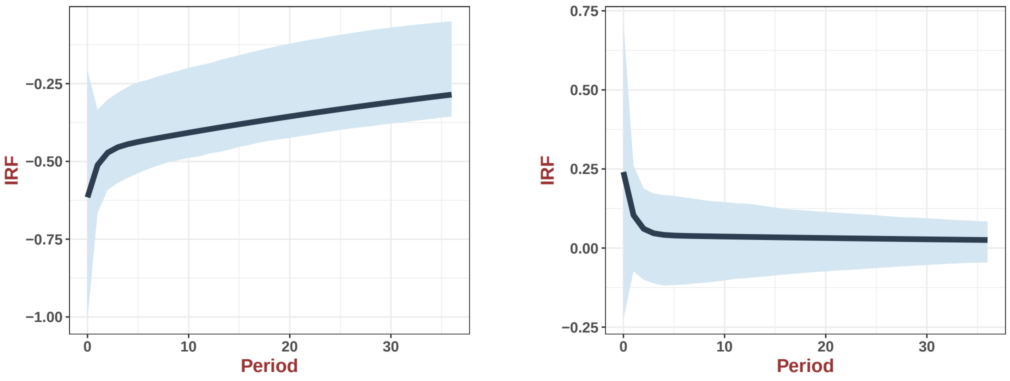
Table 3.3: VAR Estimation of Monthly Entropy, Flows to Domestic Equity Funds, and Flows to Domestic Fixed-Income Funds

	$Entropy_{t+1}$		$Flows^{FI}_{t+1}$		$Flows^{Equity}_{t+1}$	
	Coefficient	t-statistic	Coefficient	t-statistic	Coefficient	t-statistic
Intercept	0.042	1.266	-0.064	-0.259	1.615***	8.197
$Entropy_t$	0.985***	140.610	-0.012	-0.178	-0.344***	-8.106
$Flows^{FI}_t$	0.009	1.198	0.247***	3.856	-0.081	-1.329
$Flows^{Equity}_t$	-0.003	-0.331	-0.093	-1.500	0.240***	4.044
MKT_t	-0.008	-0.980	-0.054	-0.642	0.062	0.970
VXO_t	0.006	0.483	0.170**	2.115	-0.010	-0.251

This table reports the results from the VAR estimation in equation (3.17), where $\boldsymbol{Y_t}^\top = (Entropy_t, Flows^{FI}_t, Flows^{Equity}_t)$. $Entropy_t$ is the model uncertainty measure, and $Flows^{FI}_t$ ($Flows^{Equity}_t$) is the aggregate flows to the domestic fixed-income (equity) mutual funds, normalized by the lagged total market capitalization of all stocks in CRSP (see equation (3.16)). The lag is chosen by BIC and equals one. In addition, we standardize all economic variables such that they have unit variances. We also control for the lagged market return (MKT_t) and VXO index (VXO_t) in each regression. The sample spans from January 1991 to December 2020. We report both coefficient estimates and t-statistics, calculated using Newey-West standard errors with 36 lags. *, ** and *** denote significance at the 90%, 95%, and 99% level, respectively.

Table 3.3 reports the results from the VAR estimation. The sample ranges from January 1991 to December 2020. The lag is chosen by BIC and equals one. In addition, we standardize all economic variables such that they have unit variances. We also include the lagged market return and VXO index as control variables in each regression. The reported t-statistics are based on the Newey-West estimate of the covariance matrix with 36 lags. First, model uncertainty only relates to its lag. Second, the VXO index positively predicts the aggregate flows to fixed-income funds: one standard deviation increase in VXO predicts 0.17 standard deviation inflows to fixed-income funds. Third, model uncertainty negatively forecasts equity fund flows, and the coefficient estimate is sizable in both economic and statistical senses. In particular, one standard deviation increase in model uncertainty implies 0.34 standard deviation equity fund outflows. Although we cannot interpret the regression results as causal, we still find that investors in domestic equity mutual funds tend to decrease their exposures when model uncertainty increases.

Figure 3.5 shows the dynamic responses of fund flows to model uncertainty shocks in VAR-1. Most strikingly, model uncertainty innovations sharply induce fund outflows from the US equity market, with the effects persisting even after 36 months, as depicted in Panel (a). The impulse response functions (IRFs) start from around -0.6 in period zero and slowly decline to -0.35 in period 36, significantly negative based on the 90% standard error bands. In contrast, model uncertainty has negligible effects on fixed-income fund flows (see Panel (b)).



(a) Equity Fund Flows to Entropy **(b)** Fixed-Income Fund Flows to Entropy

Figure 3.5: Impulse Responses of Equity and Fixed-Income Mutual Fund Flows using Entropy as Uncertainty

This figure shows the dynamic impulse response functions (IRFs) of fund flows to model uncertainty shocks in VAR-1. The shaded area denotes the 90 percent standard error bands. We consider mutual fund flows to aggregate equity and fixed-income markets in the US. We normalize the IRFs such that the model uncertainty shock increases one standard deviation model uncertainty. We place model uncertainty first in the VAR. Hence, the implicit identification assumption is that fund flows react to the contemporaneous uncertainty shocks, while model uncertainty does not respond to the shocks to mutual funds in the current period. The data are monthly and span the period 1991:01 - 2020:12.

3.4.2 Different equity mutual funds

We further study the heterogeneous responses of different equity mutual funds to model uncertainty shocks. In particular, we split equity mutual funds into four categories: (a) style funds that specialize in factor investing, (b) sector funds that invest in specific industries (e.g., gold, oil, etc.), (c) small-cap funds that invest in relatively small stocks,[25] and (d) large-cap funds that invest in large stocks.

Table 3.4 reports the results from the VAR estimation in equation (3.17), where $\boldsymbol{Y_t}^\top = (Entropy_t, Flows_t^{style}, Flows_t^{sector}, Flows_t^{small}, Flows_t^{large})$. The lag of VAR is chosen by BIC and equals one. Since the cap-based investment objective code is available after 1997, the sample begins in January 1998. First, after controlling its lag, model uncertainty is

[25] When we mention small funds, we refer to the funds with the CRSP investment objective codes equal “EDCM”, “EDCS”, and “EDCI”.

Table 3.4: VAR Estimation of Monthly Entropy and Flows to Domestic Equity Funds with Different Investment Objectives

	$Entropy_{t+1}$		$Flows_{t+1}^{style}$		$Flows_{t+1}^{sector}$		$Flows_{t+1}^{small}$		$Flows_{t+1}^{large}$	
	Coefficient	t-statistic	Coefficient	t-statistic	Coefficient	t-statistic	Coefficient	t-statistic	Coefficient	t-statistic
Intercept	0.268***	3.270	1.582***	5.195	0.180	0.974	0.533	1.525	0.347	0.746
$Entropy_t$	0.952***	67.064	-0.261***	-5.672	-0.021	-0.525	-0.121**	-1.967	-0.014	-0.209
$Flows_t^{style}$	-0.012	-1.048	0.211***	2.936	-0.056	-1.034	-0.003	-0.054	0.003	0.034
$Flows_t^{sector}$	0.031	1.553	-0.056	-1.089	0.254*	1.686	-0.059	-0.664	-0.123**	-2.266
$Flows_t^{small}$	-0.001	-0.035	0.010	0.169	0.039	0.541	0.424***	6.081	0.089	1.225
$Flows_t^{large}$	0.019*	1.682	0.062	1.181	-0.043	-0.661	-0.107*	-1.731	0.092	1.164
R_t^{style}	0.191	0.987	0.627	0.682	0.401	0.944	-0.192	-0.215	-1.891*	-1.652
R_t^{sector}	0.043	0.900	0.121	0.956	0.367	0.957	-0.210	-1.115	0.010	0.056
R_t^{small}	-0.165**	-2.566	-0.212	-0.983	-0.126	-0.363	0.535**	1.967	0.383	1.527
R_t^{large}	-0.099	-0.741	-0.437	-0.576	-0.605	-1.610	-0.022	-0.034	1.468*	1.653
VXO_t	0.006	0.510	-0.027	-0.467	0.067	1.022	0.093*	1.957	-0.013	-0.151

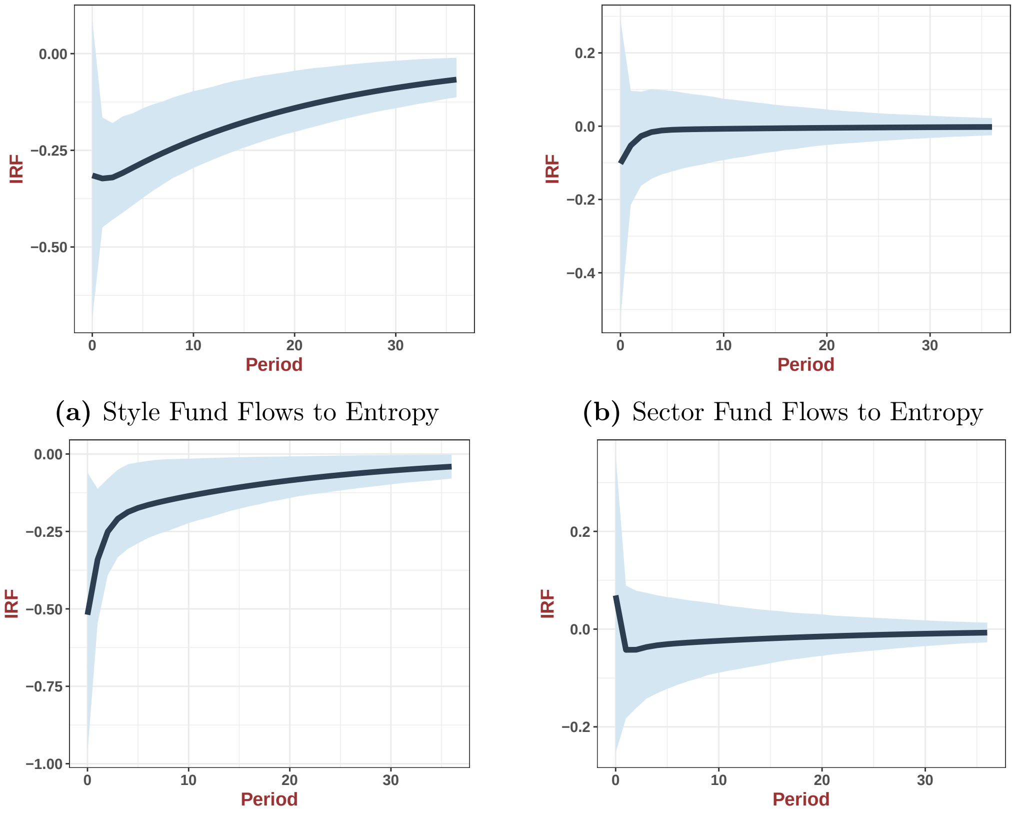
This table reports the results from the VAR estimation in equation (3.17), where $\boldsymbol{Y_t}^\top = (Entropy_t, Flows_t^{style}, Flows_t^{sector}, Flows_t^{small}, Flows_t^{large})$. $Entropy_t$ is the model uncertainty measure, and $Flows_t^{style}$ ($Flows_t^{sector}$, $Flows_t^{small}$, $Flows_t^{large}$) is the aggregate flows to the domestic style (sector, small-cap, large-cap) mutual funds, normalized by the lagged total market capitalization of all stocks in CRSP (see equation (3.16)). The lag is chosen by BIC and equals one. In addition, we standardize all economic variables such that they have unit variances. We also control for the lagged fund returns of each type and VXO index in each regression. The sample spans from January 1998 to December 2020. We report both coefficient estimates and t-statistics, calculated using Newey-West standard errors with 36 lags. *, ** and *** denote significance at the 90%, 95%, and 99% level, respectively.

negatively predicted by large-cap fund flows and small-cap fund returns. Second, model uncertainty negatively forecasts style and small-cap fund flows, and the coefficients are sizable. Specifically, if model uncertainty rises by one standard deviation, style (small-cap) fund flows tend to drop by 0.26 (0.12) standard deviation over the next period. On the contrary, we do not discover a significant relationship between model uncertainty and sector (large-cap) fund flows.

Different from model uncertainty, the traditional volatility-based uncertainty measure (VXO) plays a limited role in the VAR regression. It can marginally predict small-cap fund flows, but the sign of coefficient estimate is counter-intuitive: when uncertainty goes up, investors tend to invest more in small-cap funds. Instead, we observe a negative response of small-cap funds when using entropy as the uncertainty measure. Therefore, we argue that our model uncertainty index captures an essential source of investment risk for equity investors, which is omitted by the traditional VXO index.

Figure 3.6 shows the dynamic responses of four different types of equity fund flows to model uncertainty shocks in VAR-1. Consistent with Table 3.4, model uncertainty shocks reduce future style fund flows, and the effects are long-lasting (see Panel (a)). This observation is intuitive. Style funds refer to the growth, income, growth & income and "hedged" funds, so they are more likely to rely on the factor strategies used in constructing model uncertainty. Therefore, the outflows from style equity funds are remarkably enormous when the model uncertainty is high.

Moreover, we observe significantly negative IRFs of small-cap funds (see Panel (c)), although the effects are not as persistent as in style funds. This observation is reasonable since we include two size factors in model uncertainty. On the contrary, sector and large-cap funds almost do not respond to model uncertainty shocks. One potential explanation is that these two types of funds are primarily passive-investing funds, but model uncertainty mainly affects actively-managed funds.



(a) Style Fund Flows to Entropy

(b) Sector Fund Flows to Entropy

(c) Small-Cap Fund Flows to Entropy

(d) Large-Cap Fund Flows to Entropy

Figure 3.6: Impulse Responses of Equity Fund Flows with Different Investment Objective Codes using Entropy as Uncertainty

This figure shows the dynamic impulse response functions (IRFs) of fund flows to model uncertainty shocks in VAR-1. The shaded area denotes the 90 percent standard error bands. We consider equity fund flows with different investment objective codes (style, sector, small-cap, and large-cap). We normalize the IRFs such that the model uncertainty shock increases one standard deviation model uncertainty. We place model uncertainty first in the VAR. Hence, the implicit identification assumption is that fund flows react to the contemporaneous uncertainty shocks, while model uncertainty does not respond to the shocks to mutual funds in the current period. The data are monthly and span the period 1998:01 - 2020:12.

3.4.3 Different fixed-income funds

Similar to the previous section, we divide all fixed-income mutual funds into four categories: (a) government bond funds, (b) money market funds, (c) corporate bond funds, and (d) municipal bond funds. This subsection repeats a similar VAR estimation and investigates the dynamic responses of fixed-income fund flows to model uncertainty shocks.

Table 3.5 shows the results from the VAR-1 regression. According to columns (3) and (4), model uncertainty positively predicts the aggregate fund flows in US government bonds. US government bonds are notable for their superior safety over other asset classes. Hence, investors tend to allocate more wealth to safe assets when model uncertainty is more substantial. In contrast, model uncertainty negatively forecasts corporate fund flows, so mutual fund investors reduce their exposure to corporate bonds following high model uncertainty.

Table 3.5: VAR Estimation of Monthly Entropy and Flows to Domestic Fixed-Income Funds with Different Investment Objectives

	$Entropy_{t+1}$		$Flows_{t+1}^{gov}$		$Flows_{t+1}^{money}$		$Flows_{t+1}^{corp}$		$Flows_{t+1}^{muni}$	
	Coefficient	t-statistic	Coefficient	t-statistic	Coefficient	t-statistic	Coefficient	t-statistic	Coefficient	t-statistic
Intercept	0.381***	2.812	-0.574***	-3.305	-0.191	-0.724	1.033***	4.198	-0.132	-0.576
$Entropy_t$	0.983***	97.964	0.182**	2.535	-0.049	-0.816	-0.189***	-2.712	0.093	1.621
$Flows_t^{gov}$	0.016**	2.407	0.341***	4.580	0.090	1.325	0.094	1.448	0.136**	1.991
$Flows_t^{money}$	0.011	1.528	-0.017	-0.369	0.252***	3.125	-0.064	-1.072	0.016	0.299
$Flows_t^{corp}$	-0.012	-0.990	0.008	0.235	0.029	0.709	0.161**	2.264	0.166***	2.591
$Flows_t^{muni}$	-0.022**	-2.067	0.131**	2.064	-0.095	-1.380	0.200	1.455	0.193	1.336
VXO_t	0.006	0.590	0.044	0.623	0.191**	2.298	-0.007	-0.084	-0.094	-1.422

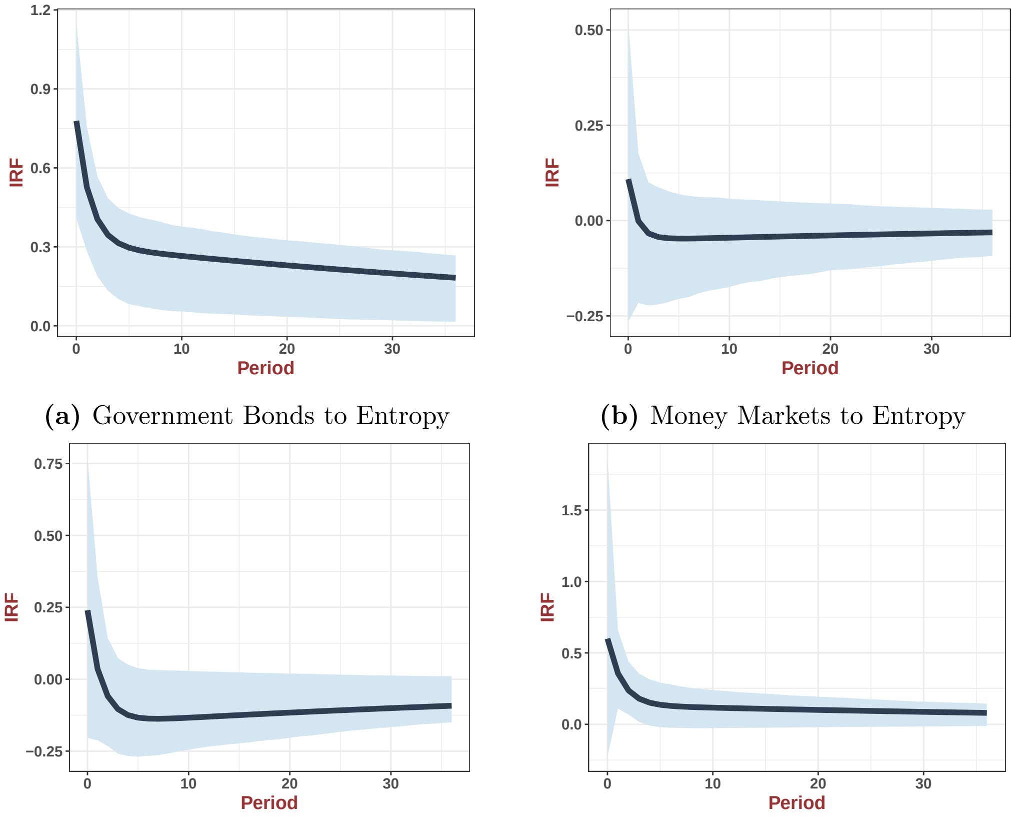
This table reports the results from the VAR estimation in equation (3.17), where $\boldsymbol{Y_t}^\top = (Entropy_t, Flows_t^{gov}, Flows_t^{money}, Flows_t^{corp}, Flows_t^{muni})$. $Entropy_t$ is the model uncertainty measure, and $Flows_t^{gov}$ ($Flows_t^{money}$, $Flows_t^{corp}$, $Flows_t^{muni}$) is the aggregate flows to the domestic government bond (money market, corporate bond, and municipal bond) mutual funds, normalized by the lagged total market capitalization of all stocks in CRSP (see equation (3.16)). The lag is chosen by BIC and equals one. In addition, we standardize all economic variables such that they have unit variances. We also control for the VXO index in each regression. The sample spans from January 1998 to December 2020. We report both coefficient estimates and t-statistics, calculated using Newey-West standard errors with 36 lags. *, ** and *** denote significance at the 90%, 95%, and 99% level, respectively.

Next, we report the IRFs of different fixed-income funds to entropy shocks in Figure 3.7. Not surprisingly, we document sharp dynamic inflows to government bond funds. As Panel (a) suggests, one standard deviation increase in model uncertainty corresponds to more than 0.7 standard deviation increase in government bond fund inflows at time zero, and the dynamic response persists for more than 36 periods. On the contrary, the IRFs of other fixed-income fund flows are not significant.

In addition, it is worth noting that we do not observe a significant relationship between model uncertainty and money market funds. The difference between money market and government bond funds is that the first type has a smaller duration and more liquid, while the latter consists of government bonds of different maturities. Unlike model uncertainty, the VXO index significantly predicts positive inflows to money market funds. We interpret these facts as evidence that high model uncertainty induces “flight to safety”, whereas high VXO implies “flight to liquidity”. Combined with the previous analyses, we conclude that mutual fund investors transfer their wealth from style and small-cap equity funds to government bonds, which are famous for their superior safety.

3.4.4 Comparison with other uncertainty measures

One major concern about the previous analyses is that model uncertainty is correlated with other uncertainty indicators, so the dynamic responses of mutual fund flows to model uncertainty shocks are confounded by them. Hence, we study how other uncertainty measures affect mutual fund flows in this section and compare their dynamic responses with the previous results. We consider the VXO index and financial uncertainty in Jurado et al. (2015) since these two measures are significantly associated with our model uncertainty measure, as we show in Table 3.2.



(a) Government Bonds to Entropy

(b) Money Markets to Entropy

(c) Corporate Bonds to Entropy

(d) Municipal Bonds to Entropy

Figure 3.7: Impulse Responses of Fixed-Income Fund Flows with Different Investment Objective Codes using Entropy as Uncertainty

This figure shows the dynamic impulse response functions (IRFs) of fund flows to model uncertainty shocks in VAR-1. The shaded area denotes the 90 percent standard error bands. We consider fixed-income fund flows with different investment objective codes (government bonds, money market, corporate bonds, and municipal bonds). We normalize the IRFs such that the model uncertainty shock increases one standard deviation model uncertainty. We place model uncertainty first in the VAR. Hence, the implicit identification assumption is that fund flows react to the contemporaneous uncertainty shocks, while model uncertainty does not respond to the shocks to mutual funds in the current period. The data are monthly and span the period 1991:01 - 2020:12.

Figure 3.8 plots the dynamic responses of four different types of equity fund flows to VXO or financial uncertainty shocks in VAR-1. Consistent with the previous identification assumption, We place VXO or financial uncertainty first in the VAR. We also control the lagged model uncertainty in each regression. First, as Panels (a) and (b) indicate, style funds experience massive outflows when VXO or financial uncertainty increases. However, these effects are temporary; that is, the IRFs of fund flows reverse back to zeros immediately after time zero. On the contrary, model uncertainty shocks are followed by persistent outflows from style funds even beyond 36 periods. Similarly, the dynamic responses of fund flows to sector/small-cap/large-cap funds are also transitory and not significant (except for Panel (c) at period zero).

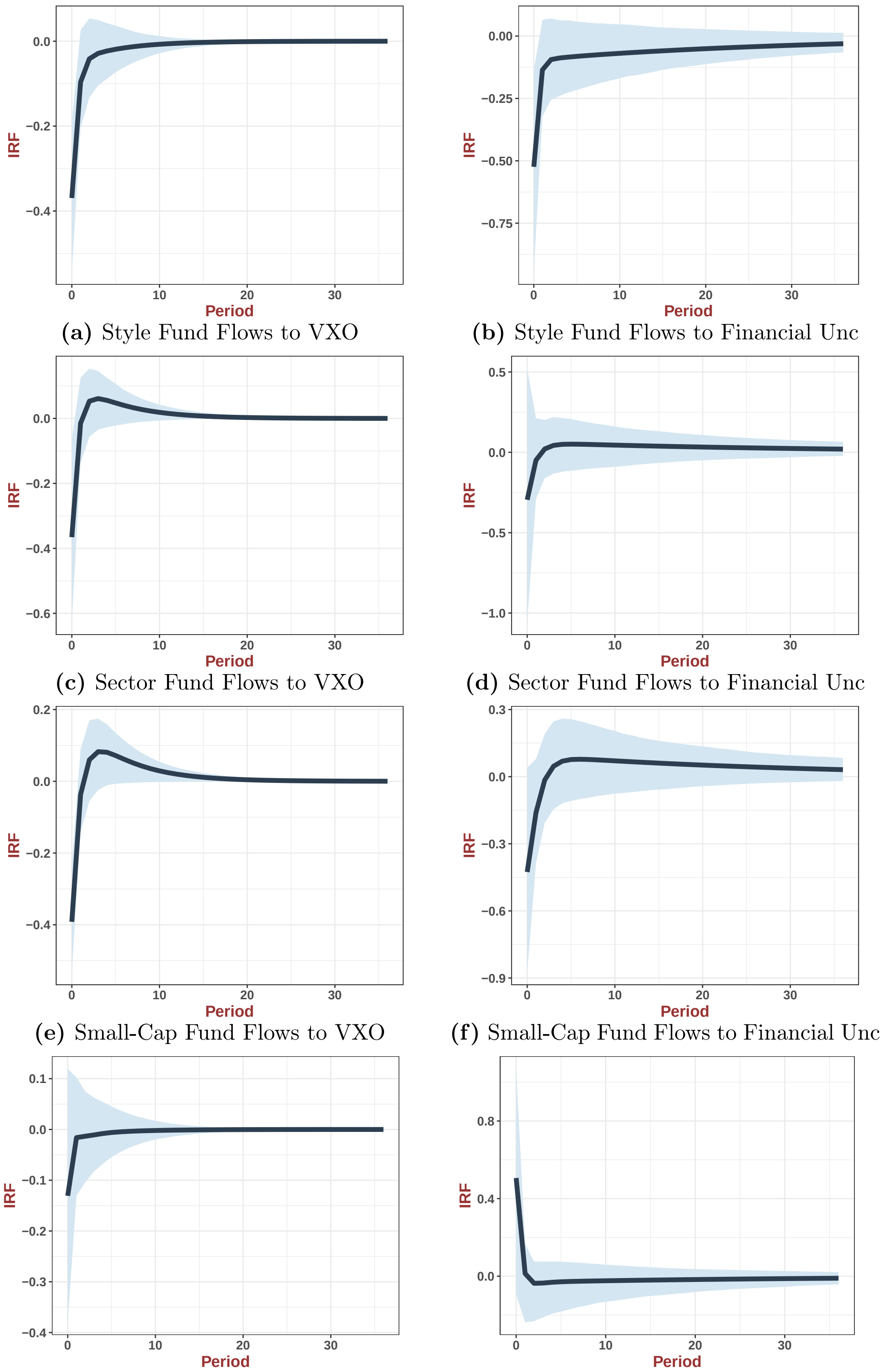
We further consider the dynamic responses of fixed-income funds in Figure 3.9. When VXO or financial uncertainty goes up, government bond funds tend to experience massive inflows, although these effects are less than 50% of those following model uncertainty shocks (see Figure 3.7(a)). Most strikingly, we document massive inflows to money market funds after positive VXO and financial uncertainty shocks. In contrast, model uncertainty does not play a part in money market funds. In other words, model uncertainty shocks primarily induce “flight to safety", while other volatility-based uncertainty measures are mainly related to “flight to liquidity".

In summary, our model uncertainty measure captures some unique dynamic responses of fund flows, and notably, they are different from traditional volatility-based measures, such as VXO and financial uncertainty. In particular, we observe significant fund inflows to government bond funds and outflows from style and small-cap equity funds. In contrast, VXO and financial uncertainty shocks fail to generate similar dynamic responses. Finally, as we will show in Section 3.7, the IRFs of fund flows to model uncertainty shocks are virtually robust to an alternative identification assumption, whereas the effects of VXO or financial uncertainty shocks tend to be fairly sensitive.

3.5 Investors’ expectations

This section investigates whether our model uncertainty measure correlates with investors’ expectations of the stock markets. The first measure is from the American Association of Individual Investors (AAII). The survey is completed weekly by registered members of AAII, and it asks the investors whether they are bearish, neutral, or bullish on the stock market for the next six months. Since our model uncertainty measure is of monthly frequency, we use the expectation measures in the last week of each month.

We also consider Robert Shiller’s stock market confidence indices from the survey conducted by the International Center for Finance at the University of Yale. Our paper focuses on the US one-year confidence index and US crash confidence index. Specifically, the one-year confidence index is the percentage of the individual or institutional investors expecting an increase in the Dow in a year. In contrast, the crash confidence index is the percentage of individual or institutional investors who believe the probability of a



(a) Style Fund Flows to VXO

(b) Style Fund Flows to Financial Unc

(c) Sector Fund Flows to VXO

(d) Sector Fund Flows to Financial Unc

(e) Small-Cap Fund Flows to VXO

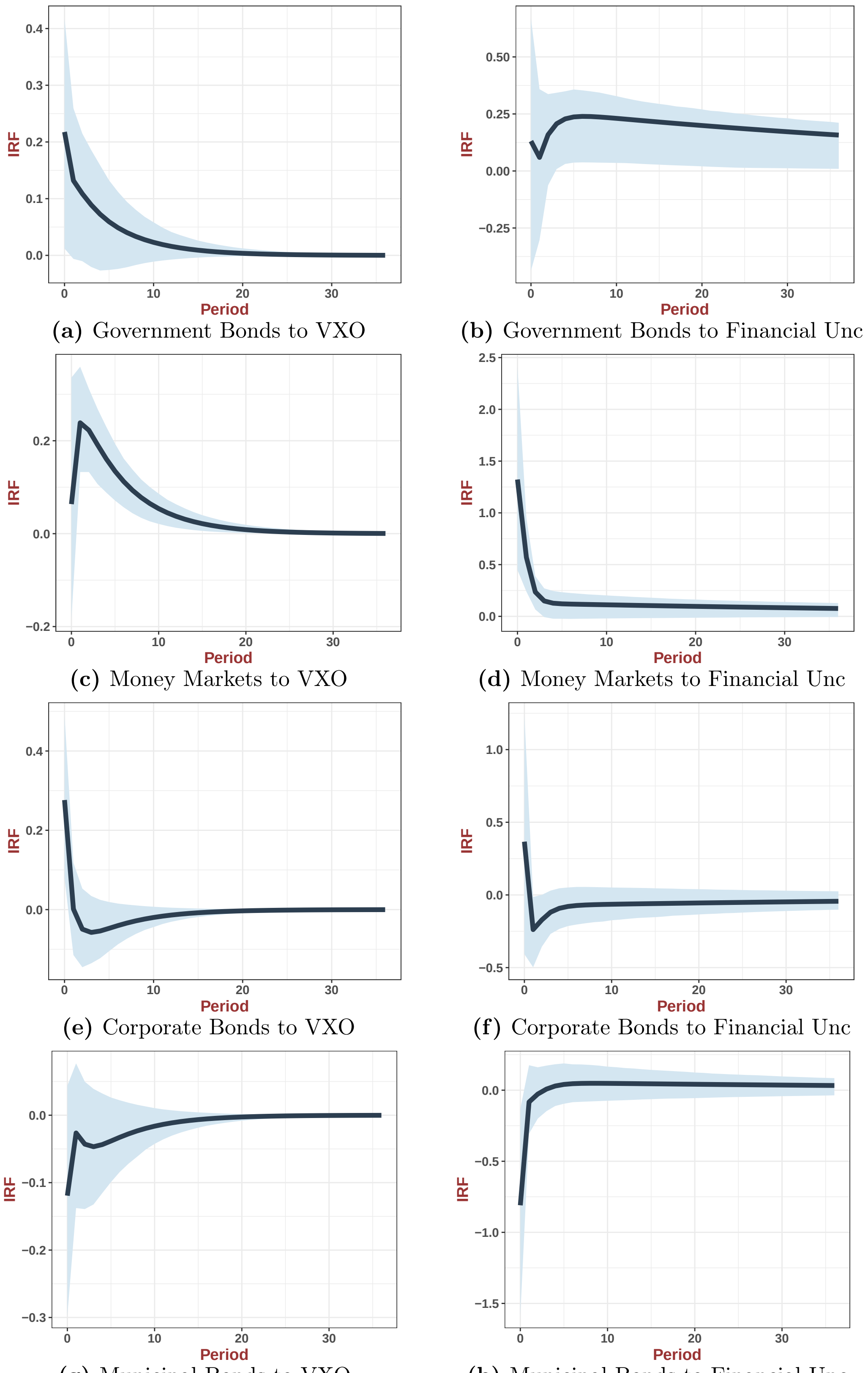
(f) Small-Cap Fund Flows to Financial Unc

(g) Large-Cap Fund Flows to VXO

(h) Large-Cap Fund Flows to Financial Unc

Figure 3.8: Impulse Responses of Equity Fund Flows with Different Investment Objective Codes using VXO and Financial Uncertainty as Uncertainty Measures

This figure shows the dynamic impulse response functions (IRFs) of equity fund flows to VXO and financial uncertainty shocks in VAR-1. Other details can be found in the footnote of Figure 3.6.



(a) Government Bonds to VXO

(b) Government Bonds to Financial Unc

(c) Money Markets to VXO

(d) Money Markets to Financial Unc

(e) Corporate Bonds to VXO

(f) Corporate Bonds to Financial Unc

(g) Municipal Bonds to VXO

(h) Municipal Bonds to Financial Unc

Figure 3.9: Impulse Responses of Fixed-Income Fund Flows with Different Investment Objective Codes using VXO and Financial Uncertainty as Uncertainty Measures

This figure shows the dynamic impulse response functions (IRFs) of fixed-income fund flows to VXO and financial uncertainty shocks in VAR-1. Other details can be found in the footnote of Figure 3.7.

catastrophic stock market crash in the next six months is lower than 10%. Roughly speaking, the higher the indices are, the more confident individual or institutional investors are about the stock market.

We consider the following time-series regression:

$$Exp_{t+1} = \beta_0 + \gamma Entropy_t + \psi X_t + \epsilon_{t+1} \tag{3.18}$$

where Exp_{t+1} is the one-period ahead expectation measure, $Entropy_t$ is the model uncertainty measure in period t, and X_t includes other control variables up to time t, such as lagged expectation indices, VXO and etc. Since all expectation indices are autocorrelated, we control their one and two-period lags in all regressions.[26] We further control lagged market returns (S&P 500 index) in the regression for investors' expectations on the market are extrapolative (see Greenwood and Shleifer (2014)).

In table 3.6(a), we regress AAII sentiment indices on model uncertainty to explore how individual investors change their attitudes towards the stock market in response to variation in model uncertainty. To increase the interpretability of our results, we standard model uncertainty to have unit variance, so coefficient estimates of $Entropy_t$ are interpreted as the increases in the percentages of bullish/neutral/bearish investors when model uncertainty grows by one standard deviation.

In columns (1) and (2), $Entropy_t$ cannot predict the next-period percentage of bullish investors. Specifically, the average investors become less bullish if model uncertainty in the cross-section goes up, but this prediction is not sharp. Columns (3) and (4) regress the percentage of neutral investors on lagged model uncertainty: If model uncertainty increases by one standard deviation, the fraction of neutral investors declines by 0.605% or 0.434%, depending on the regression setup.

The next question is, in which direction do bullish investors change their attitudes? Columns (5) and (6) indicate that investors are more likely to be bearish following an increase in model uncertainty. Our interpretation is that some neutral investors become bearish after observing a higher level of model uncertainty. Finally, we regress the difference between fractions of bullish and bearish investors on entropy. The coefficient estimate of entropy is negative and significant at the 10% level. Overall, when model uncertainty goes up, market participants tend to be more pessimistic about the future stock market performance.

Table 3.6(b) regresses Shiller's confidence indices on entropy. Unlike the AAII sentiment index, we also observe the expectations of institutional investors. The results are generally similar to table 3.6: Investors tend to be more pessimistic about the stock market when model uncertainty increases. They also believe that a market crash is more likely to occur following higher model uncertainty. One interesting empirical fact is that the coefficient estimates of $Entropy_t$ in the regressions of individual investors' confidence indices are always more negative than institutional investors. Hence, individual investors

[26] The coefficient estimate of 3-period lagged variable is close to zero and insignificant, so we include only the first two lags.

Table 3.6: Investors' Expectations, Confidence Indices, and Model Uncertainty

Panel (a). AAII Sentiment Index

Exp_{t+1} =	Bullish		Neutral		Bearish		Bullish - Bearish	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$Entropy_t$	-0.280	-0.374	-0.605**	-0.434**	1.043**	1.036***	-1.511*	-1.574**
	(-0.683)	(-1.122)	(-2.121)	(-2.102)	(2.499)	(2.656)	(-1.826)	(-2.127)
VXO_t	0.022	0.079	0.016	-0.009	-0.008	-0.034	0.016	0.118
	(0.311)	(1.500)	(0.211)	(-0.161)	(-0.169)	(-0.500)	(0.157)	(1.060)
Exp_t	0.418***	0.373***	0.487***	0.452***	0.367***	0.335***	0.373***	0.331***
	(8.593)	(6.954)	(9.709)	(10.249)	(9.238)	(7.155)	(7.325)	(5.983)
Exp_{t-1}	0.098**	0.158***	0.213***	0.253***	0.182***	0.208***	0.118***	0.160***
	(2.434)	(3.531)	(6.103)	(6.209)	(5.676)	(5.850)	(3.151)	(3.623)
Lagged Market Returns	NO	YES	NO	YES	NO	YES	NO	YES
Sample Size	400	396	400	396	400	396	400	396
R^2_{adj}	21.76%	22.53%	43.11%	44.98%	27.24%	26.79%	20.92%	21.01%

Panel (b). Shiller's Confidence Indices

Exp_{t+1} =	1-Year Confidence Index - Institution		1-Year Confidence Index - Individual		Crash Confidence Index - Institution		Crash Confidence Index - Individual	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$Entropy_t$	-0.365***	-0.379***	-0.546***	-0.682***	-0.562***	-0.635***	-0.754***	-0.754***
	(-2.727)	(-2.952)	(-2.733)	(-5.405)	(-3.265)	(-3.335)	(-5.790)	(-6.048)
VXO_t	0.025*	0.030	0.044*	0.080***	-0.058**	-0.034	-0.046***	-0.001
	(1.767)	(0.829)	(1.705)	(4.204)	(-2.153)	(-1.066)	(-2.712)	(-0.047)
Exp_t	1.133***	1.165***	0.931***	0.949***	1.068***	1.065***	1.086***	1.071***
	(16.820)	(18.984)	(11.730)	(15.898)	(19.165)	(21.126)	(16.459)	(13.272)
Exp_{t-1}	-0.270***	-0.304***	-0.015	-0.045	-0.217***	-0.219***	-0.241***	-0.208***
	(-3.603)	(-4.449)	(-0.212)	(-0.823)	(-3.540)	(-4.000)	(-4.268)	(-3.078)
Lagged Market Returns	NO	YES	NO	YES	NO	YES	NO	YES
Sample Size	232	228	232	228	232	228	232	228
R^2_{adj}	82.70%	83.38%	93.17%	93.25%	87.44%	87.01%	92.24%	92.82%

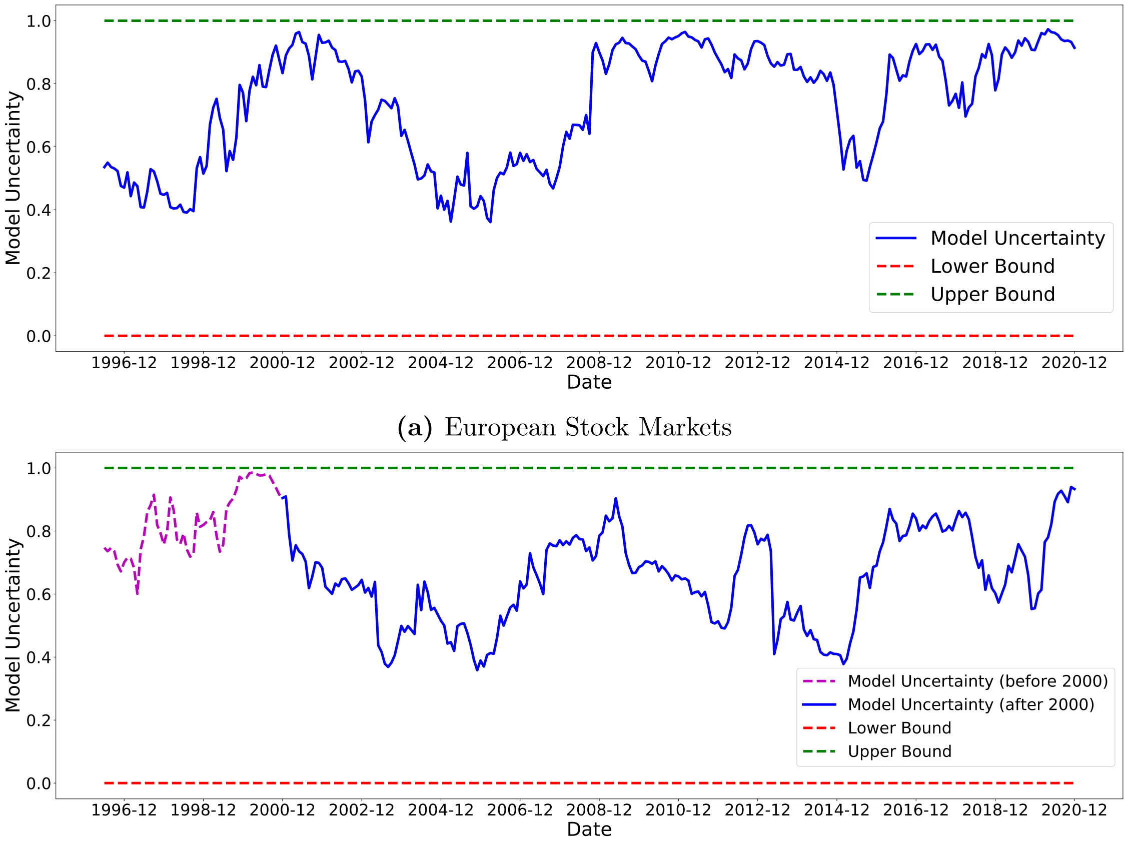
The table reports empirical results in regression: $Exp_{t+1} = \beta_0 + \gamma Entropy_t + \psi X_t + \epsilon_{t+1}$, where Exp_{t+1} is the one-period ahead expectation/confidence index, $Entropy_t$ is the model uncertainty measure in period t, and X_t includes other control variables up to time t, such as lagged expectation/confidence indices, VXO and etc. Since all expectation/confidence indices are autocorrelated, we control their one and two-period lags (Exp_t and Exp_{t-1}) in all regressions. We further control lagged market returns in the regression (we include six lags). In Panel (a), expectation indices come from the survey conducted by the American Association of Individual Investors (AAII). The survey is completed weekly by registered members of AAII, and it asks the investors whether they are bearish, neutral or bullish on the stock market for the next six months. Therefore, we have the data regarding the percentages of bearish, neutral or bullish respondents each week. Since our model uncertainty measure is monthly, we use the expectation index in the final week of each month. In Panel (b), confidence indices come from Shiller's survey. We focus on the US one-year confidence index and US crash confidence index. The one-year confidence index is the percentage of the individual or institutional investors expecting an increase in the Dow in a year. In contrast, the crash confidence index is the percentage of individual or institutional investors who think that the probability of a catastrophic stock market crash in the next six months is lower than 10%. The t-statistics are computed using Newey-West standard errors with 36 lags. *, ** and *** denote significance at the 90%, 95%, and 99% level.

react more dramatically to the changes in model uncertainty than institutional ones.

In short, we conclude that higher model uncertainty generally predicts that investors in the survey, be it individual or institutional, will become more pessimistic about the future stock market performance.

3.6 Evidence in European and Asia-pacific markets

This section presents the time series of model uncertainty in European and Asian Pacific stock markets. Instead of using all 14 factors in the US stock market, we include only nine of them because of the limited data availability. Specifically, short-term and long-term behavioural factors are excluded because they are unavailable in international markets. For the same reason, we ignore the size (ME), profitability (ROE), and investment (IA) in Hou et al. (2015), and we believe that the Fama-French five factors capture similar systematic risks. Finally, we end up with nine candidates: MKT, SMB, HML, RMW, CMA, MOM, QMJ, BAB, and HML devil. Either HML or HML devil can enter the true SDF. Since the AQR library only provides the QMJ factor from July 1993, and we use a three-year rolling window, our model uncertainty measure starts from June 1996.



(a) European Stock Markets

(b) Asian Pacific Stock Markets

Figure 3.10: Model Uncertainty in European and Asian Pacific Markets

The figure plots the time series of model uncertainty about the linear stochastic discount factor (SDF) in European and Asian Stock Markets. The construction of model uncertainty is the same as in figure 3.1 except that we use only nine factors to calculate the posterior model probabilities. Details about used factors could be found in section 3.6. The sample ranges from July 1993 to December 2020. Since we use 3-year rolling window, the model uncertainty index starts from June 1996. The red line and green lines in the figure show the lower (0) and upper bounds (1) of model uncertainty.

Figure 3.10a plots the time series of model uncertainty in the European stock market from June 1996 to December 2020. Several results stand out. The time-series patterns in European markets[27] are remarkably similar to the US stock market. In particular, model

[27] European markets include the following countries: Austria, Belgium, Switzerland, Germany, Den-

uncertainty increases from 1999 and reaches its first peak between 2000 and 2001 because of the dot-com bubble burst. During these periods, model uncertainty almost touches its upper bound. After 2002, model uncertainty declines gradually and remains relatively low until the start of the 2008 global financial crisis. During this long-lasting economic and stock market crisis, model uncertainty stays close to the upper bound from 2008 to 2012 and only declines gradually after 2012. Finally, the uncertainty index shoots up again after 2015, similar to what we observe in the US market.

We next turn to discuss the findings in Asian Pacific markets.[28] It is worth noting that we observe some unique time-series variation in Asian stock markets. According to figure 3.10b, model uncertainty is high starting from 1997 due to the profound 1997 Asian financial crisis. Asian stock markets were over-heated, and market crashes appeared in almost every Asian country. The dot-com bubble in 2000 led to another peak in model uncertainty, which almost reaches the upper bound. However, the Asian markets recovered quickly after 2000, so the model uncertainty index declines afterwards. Another steady increase in model uncertainty appears before and during the 2008 crisis, but the entropy is not as high as in the late 1990s and drops immediately from 2009. This particular pattern is unlike the US and European markets, in which we observe higher model uncertainty of the 2008 crisis than the dot-com bubble.

Another steady increase in model uncertainty appears before and during the 2008 crisis, but the entropy is not as high as in the late 1990s and drops immediately from 2009. This particular pattern is unlike the US and European markets, in which we observe higher model uncertainty of the 2008 crisis than the dot-com bubble. One potential explanation is that the 1997 Asian financial crisis, combined with the burst of the dot-com bubble in 2000, was more destructive than the 2008 financial crisis. There is a short-term upward jump in model uncertainty between 2011 and 2012 when the US government bonds were downgraded. Similar to US and European markets, model uncertainty surges from the beginning of 2015.

In short, the international market evidence in this section lends further support to the time-varying nature of model uncertainty. First, model uncertainty is high in many periods, way above its lower bound. Second, it fluctuates significantly over time and coincides with major events in corresponding asset markets. However, model uncertainty is not all alike. For example, Asian markets display unique behaviours that distinguish them from the US and European markets.

3.7 Robustness

This section considers several robust checks, including alternative hyper-parameter a in estimating factor models, alternative rolling windows in constructing the time series of model uncertainty, and a different identification assumption under which we re-estimate

mark, Spain, Finland, France, UK, Greece, Ireland, Italy, Netherlands, Norway, Portugal, and Sweden.

[28] By saying the Asian Pacific market, we refer to the stock markets in Australia, Hong Kong, New Zealand, and Singapore.

the dynamic responses of fund flows to uncertainty shocks.

3.7.1 Alternative hyper-parameter a

One important choice in our Bayesian inference is the value of hyper-parameter a. In the benchmark case, we assign a to be 4. Just as Section 3.1 shows, a higher a implies a stronger shrinkage for factors' risk prices, $\boldsymbol{b}$.

Figure 3.12 plots the time series of model uncertainty using different values of a, including 3, 8, 16. Several findings stand out. First, we find that the time-series patterns in model uncertainty are not sensitive to the choice of a. In fact, the sequences under different values of a are virtually identical. Second, model uncertainty is increasing in a. This observation is not surprising since a larger a mechanically shrinks all candidate models to the null model, rendering factor models to become more similar and driving up model uncertainty.

3.7.2 Alternative rolling windows

There is a trade-off in choosing the length of the rolling window. On the one hand, we prefer a larger time-series sample to achieve higher precision in estimating model parameters. The one-year or two-year daily sample is insufficient since estimating factors' expected returns and their covariance matrix is challenging. On the other hand, larger sample size is not always desirable since it implicitly assumes that factor models remain constant and robust over a long period. As many research (e.g. McLean and Pontiff (2016)) suggest, factors' performances deteriorate post-publication. Moreover, a long estimation period of 10 or 20 years will average valuable information concerning factors' cyclical behaviours.

Motivated by the above discussion, we consider four-year and five-year rolling windows in Figure 3.13. There is one tiny difference: Model uncertainty tends to be smoother in longer rolling windows, especially the five-year window. Beyond that, the time-series properties are similar to those found in a three-year rolling window.

3.7.3 Alternative VAR identification assumption

Another robustness check concerns the identification assumption in our VAR analysis. In Section 3.4, we put uncertainty measures first in $\boldsymbol{Y_t}$. We now consider an alternative setup, in which uncertainty measures are the last variables in $\boldsymbol{Y_t}$. In other words, we allow uncertainty measures to correlate with contemporaneous shocks to mutual fund flows, but uncertainty shocks do not affect mutual fund flows simultaneously. Although model uncertainty is an endogenous response to innovations in fund flows under this assumption, it is still worth investigating whether model uncertainty is a key player to propagate those exogenous shocks over a long-lasting period.

Figures 3.15 and 3.16 plot the IRFs of fund flows to three uncertainty measures. Under the current assumption, the IRFs are zeros at period zero by construction. The first column shows the dynamic responses to model uncertainty shocks. Similar to the observations in

Figures 3.6 and 3.7, an increase in model uncertainty relates to persistent outflows from style and small-cap funds but sharp inflows to government bond funds. The dynamic effects are bounded well away from zero even beyond 36 months, although they decline slowly over time. Hence, the main results in Figures 3.6 and 3.7 are largely robust.

The second and third columns show the IRFs of fund flows using VXO and financial uncertainty. Surprisingly, VXO shocks imply positive inflows to small-cap funds. On average, one standard deviation increase in the VXO index corresponds to more than 0.1 standard deviation fund inflows, and these positive dynamic responses last for around 20 months. However, the 90% confidence interval of IRFs covers zero effects, so they are on the edge of being consequential. Beyond that, the IRFs in other panels are virtually zeros, so there is little evidence that mutual fund investors react to VXO or financial uncertainty shocks.

Finally, we observe significant inflows to money market funds following positive VXO shocks, and the dynamic responses have similar economic sizes to those in Figure 3.9. The key difference under the new identification assumption is that the IRFs of money market funds to financial uncertainty shocks are no longer significant. In other words, the dynamic responses to financial uncertainty shocks in Figure 3.9 are driven mainly by the identification assumption.

To conclude, model uncertainty has robust and persistent effects on mutual fund flows, particularly the style, small-cap, and government bond funds. We argue that model uncertainty is a crucial determinant of mutual fund flows, regardless of being an exogenous cause or a merely propagating mechanism. On the contrary, the dynamic responses of fund flows to volatility-based measures, be it VXO or financial uncertainty, are more or less sensitive to different identification assumptions. In fact, there is little evidence that equity mutual fund investors respond to VXO or uncertainty shocks.

3.8 Conclusions

We develop a new measure of model uncertainty in the cross-sectional asset pricing under the linear SDF specification. Roughly speaking, the measure is based on the entropy of Bayesian posterior probabilities for all possible factor models. The critical observation is that model uncertainty is countercyclical: it begins to climb up right before the stock market crashes and remains at its peaks during bear markets. Since we can calculate the lower and upper bound of entropy, we can easily discern when model uncertainty is abnormally high or low. In contrast, other uncertainty measures in past literature do not have this satisfactory property. We find that model uncertainty almost touches its upper bounds in the burst of the dot-com bubble and the 2008 financial crisis.

If investors consider model uncertainty as another source of investment risk, their portfolio choice and expectations of the stock market should be naturally related to model uncertainty. Our second key observation is that model uncertainty can predict the next-period mutual fund flows, even after controlling past fund flows, VXO, and the past performance of mutual funds. In particular, investors seem to reduce their investment in style and small-cap mutual funds but allocate more of their wealth to safer US government bond funds. Model uncertainty is also closely related to investors' expectations and confidence. We document that investors in the survey, no matter individual or institutional investors, are more pessimistic about the stock market when confronted with higher model uncertainty. We find similar countercyclical behaviours of model uncertainty in European and Asian Pacific stock markets.

As model uncertainty in the cross-section is an important source of investment risk, future theoretical research on portfolio choice should incorporate it into the model. Even though a few partial equilibrium models have considered model uncertainty of mean-variance portfolios, no such a general equilibrium model exists, at least according to our knowledge. Future research could attempt to endogenize model uncertainty in the general equilibrium model and explain its countercyclical behaviours.

3.9 Appendices

3.9.1 Description of factors

CAPM. The CAPM in Sharpe (1964) and Lintner (1965) is the pioneer of linear factor models. The only factor in CAPM is the excess return on the market portfolio (MKT). The data comes from Ken French's website.

Fama-French Five-factor model. Fama and French (1992) extend CAPM by introducing SMB and HML, where SMB is the return difference between portfolios of small and large stocks, and HML is the return difference between portfolios of stocks with high and low book-to-market ratios. Fama and French (2016) further include a profitability factor (RMW) and one investment factor (CMA). Again, the data comes from Ken French's website.

Momentumn. Jegadeesh and Titman (1993) find that stocks that perform well or poorly in the past three to 12 months continue their performance in the next three to 12 months. Therefore, investors can outperform the market by buying past winners and selling past losers. We download the momentum (MOM) factor from Ken French's data library.

q-factor model. Hou et al. (2015) introduce a four-factor model that includes market excess return (MKT), a new size factor (ME), an investment factor (IA), and finally, the profitability factor (ROE).[29]

Behavioral Factors. Daniel et al. (2020) propose a three-factor model consisting of the market factor and two theory-based behavioural factors. The short-term behavioural factor is based on the post-earnings announcement drift (PEAD) and captures the underreaction to quarterly earnings announcements in the short horizon. Instead, the long-term behavioural factor (FIN) is based on the one-year net and five-year composite share issuance.

Quality-minus-junk. Asness et al. (2019) groups the listed companies into the quality and junk stocks. They find that a quality-minus-junk (QMJ) strategy generate high positive abnormal returns. We download the QMJ factor from the AQR data library.

Betting-against-beta. One of the most prominent failures of CAPM is that the security market line is too flat, so the risk premia of high-beta stocks are not as substantial as CAPM suggests. Frazzini and Pedersen (2014) constructs market-neutral betting-against-beta (BAB) factor that longs the low-beta stocks and shorts high-beta assets. We download the BAB factor from the AQR data library.

HML Devil. Asness and Frazzini (2013) propose an alternative way to construct the value factor, which relies on more timely market value information. We download the HML Devil factor from the AQR data library.

3.9.2 Additional tables and figures

[29] We are grateful to the authors for sharing the data with us.

Table 3.7: Summary Statistics of 14 Factors

	Full Sample		Subsample I		Subsample II	
	Mean (%)	SR	Mean (%)	SR	Mean (%)	SR
MKT	7.36	0.43	5.54	0.40	9.18	0.47
ME	1.97	0.22	1.79	0.23	2.16	0.21
IA	3.92	0.66	6.36	1.38	1.48	0.21
ROE	6.21	0.91	8.50	1.72	3.92	0.47
SMB	1.24	0.14	0.89	0.12	1.58	0.16
HML	3.39	0.37	6.30	1.03	0.48	0.04
RMW	3.26	0.52	2.77	0.73	3.74	0.47
CMA	3.42	0.59	4.76	1.05	2.07	0.30
MOM	6.89	0.55	8.94	1.22	4.85	0.30
QMJ	4.31	0.63	3.76	0.94	4.85	0.55
BAB	10.10	1.00	11.99	1.81	8.21	0.65
HML_devil	3.03	0.30	5.80	0.90	0.27	0.02
FIN	8.47	0.73	11.67	1.36	5.28	0.38
PEAD	7.57	1.30	9.34	2.00	5.80	0.85

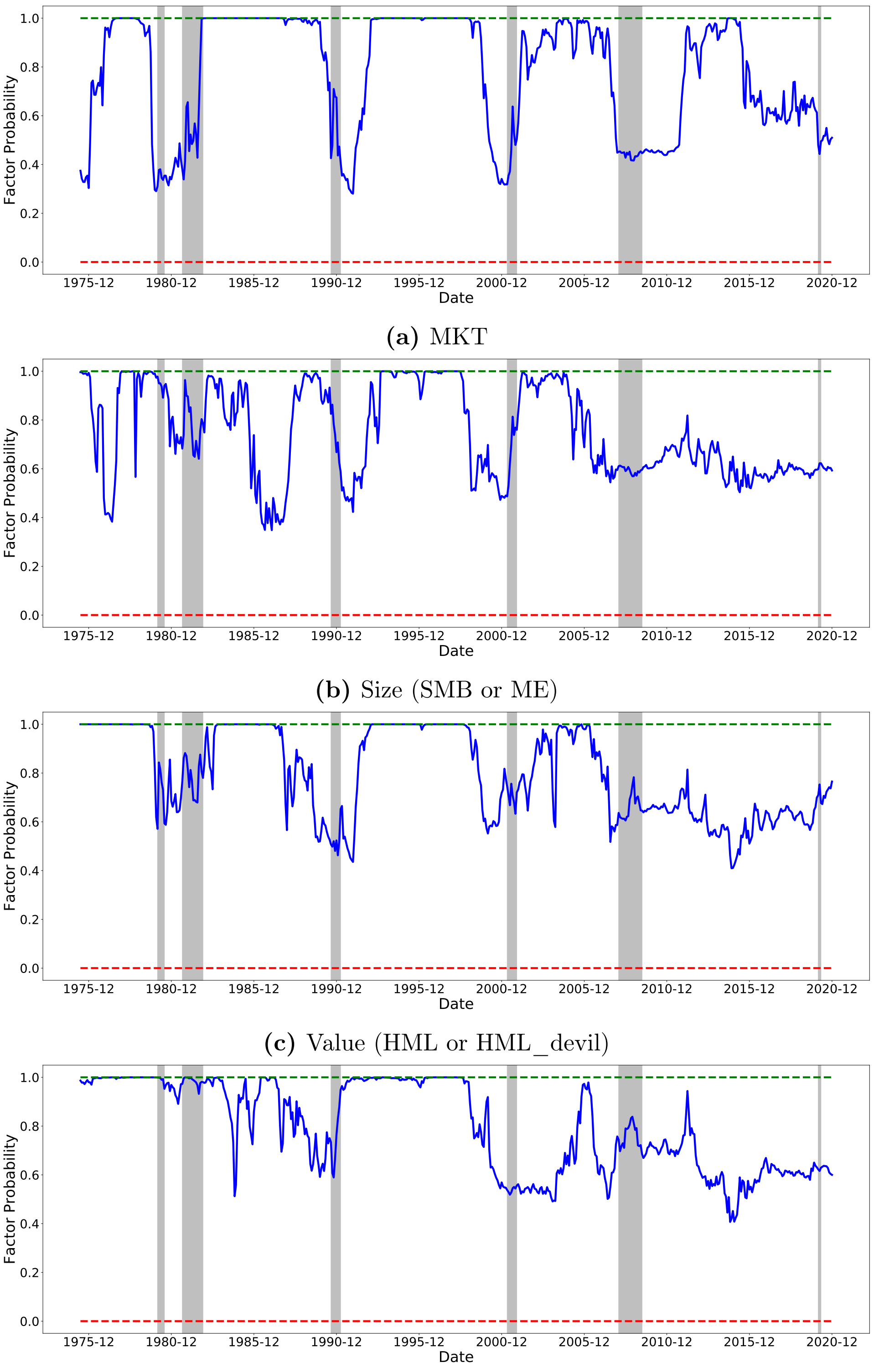
This table reports the annualised mean returns and annualised Sharpe ratios of 14 factors listed in Appendix 3.9.1. The full sample starts from July 1972 to December 2020. We further split the entire sample into two equal subsamples.

Table 3.8: Summary of First-Order Autoregression

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Entropy	Financial	Macro	Real	EPU_1	EPU_2	VXO
AR(1)	0.986***	0.977***	0.985***	0.984***	0.844***	0.700***	0.812***
	(158.08)	(98.78)	(73.92)	(46.84)	(24.64)	(14.30)	(23.40)
Sample size	546	546	546	546	431	431	419
R^2	0.9697	0.9523	0.9667	0.9514	0.6929	0.5945	0.6586

t statistics in parentheses: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

The table reports empirical results in the first-order autoregression of seven uncertainty measures: $y_{t+1} = \alpha + \rho y_t + \epsilon_{t+1}$. Entropy is our model uncertainty measure. Financial, macro and real uncertainty measures come from Ludvigson et al. (2021) and Jurado et al. (2015). EPU_1 and EPU_2 are two economic policy uncertainty sequences from Baker et al. (2016). VXO is the forward-looking market volatility traded in CME. The t-statistics are computed using Newey-West standard errors with 36 lags. *, ** and *** denote significance at the 90%, 95%, and 99% level, respectively.



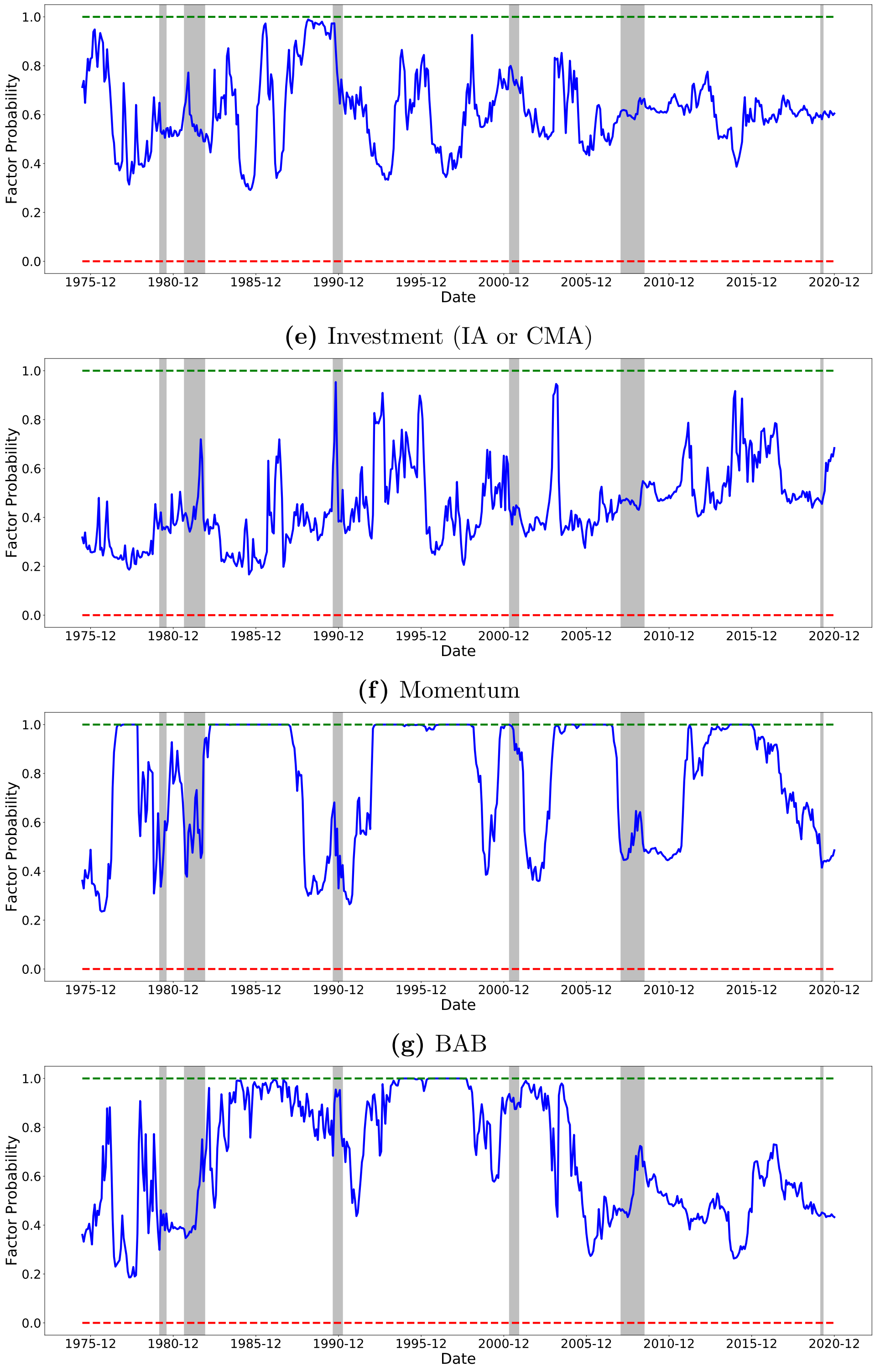
(a) MKT

(b) Size (SMB or ME)

(c) Value (HML or HML_devil)

(d) Profitability (ROE or RMW)

Figure 3.11: Time Series of Posterior Factor Probabilities



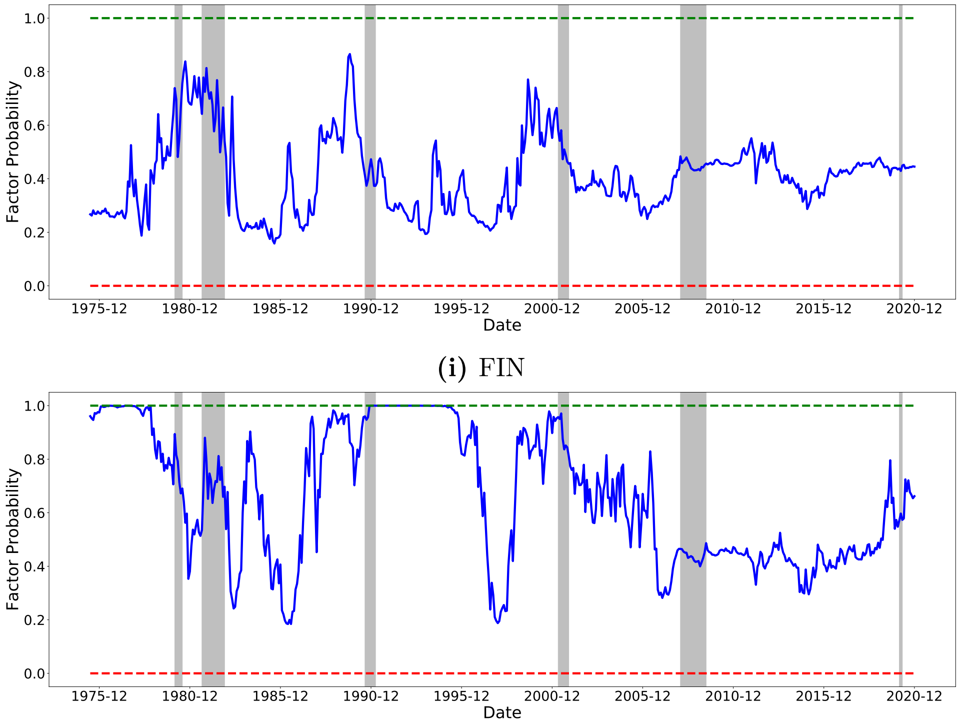
(e) Investment (IA or CMA)

(f) Momentum

(g) BAB

(h) QMJ

Figure 3.11: Time Series of Posterior Factor Probabilities (Continued)



(i) FIN

(j) PEAD

Figure 3.11: Time Series of Posterior Factor Probabilities (Continued)

The figures plot the time series of posterior marginal probabilities of 14 factors. At the end of each month, we estimate models using the daily factor returns in the past three years. The sample ranges from July 1972 to December 2020. Since we use a three-year rolling window, the time series of factor probabilities start from June 1975. Shaded areas are NBER-based recession periods for the US.

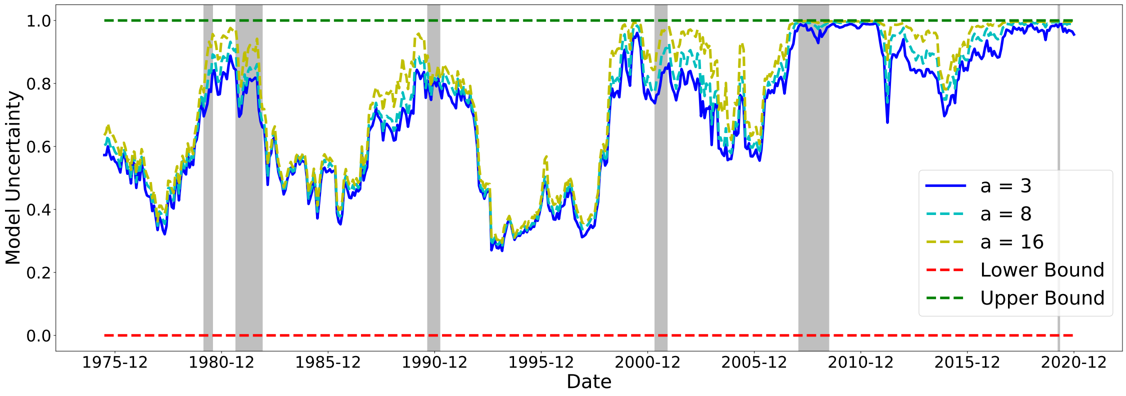
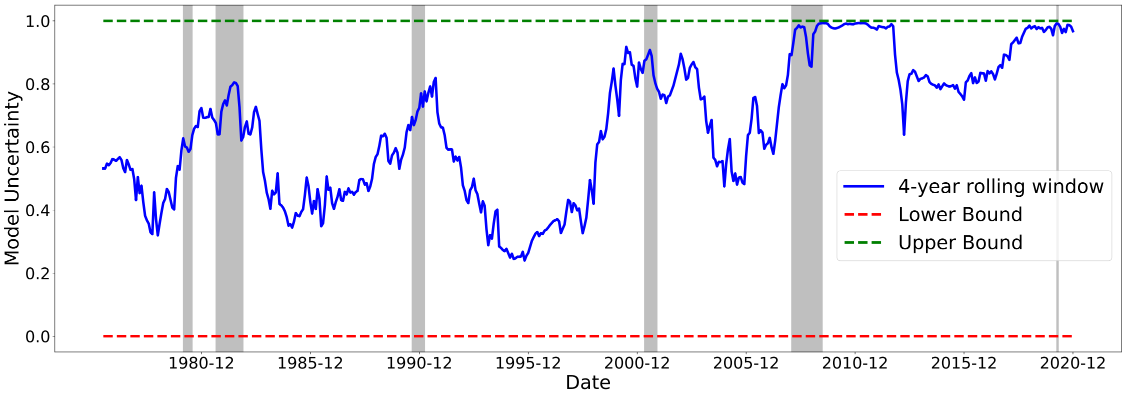
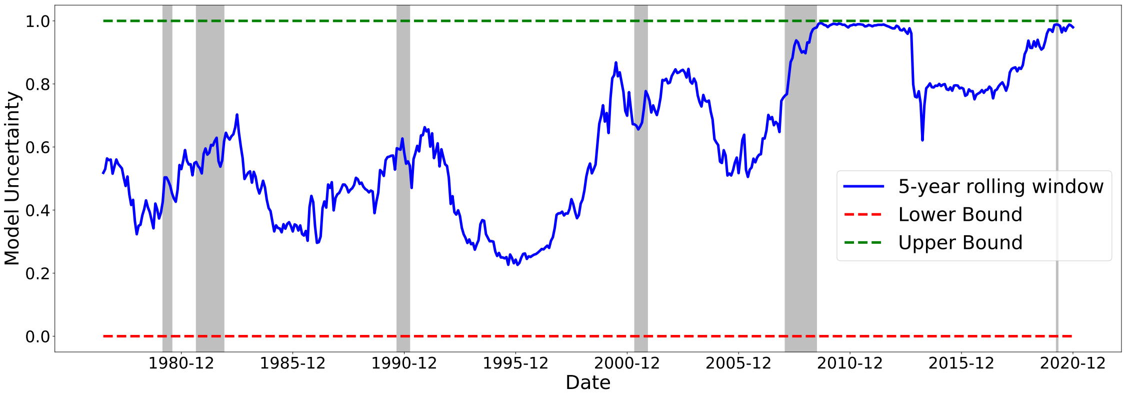


Figure 3.12: Time-Series of Model Uncertainty (3-Year Rolling Window) using different values of the hyper-parameter, $a \in \{3, 8, 16\}$

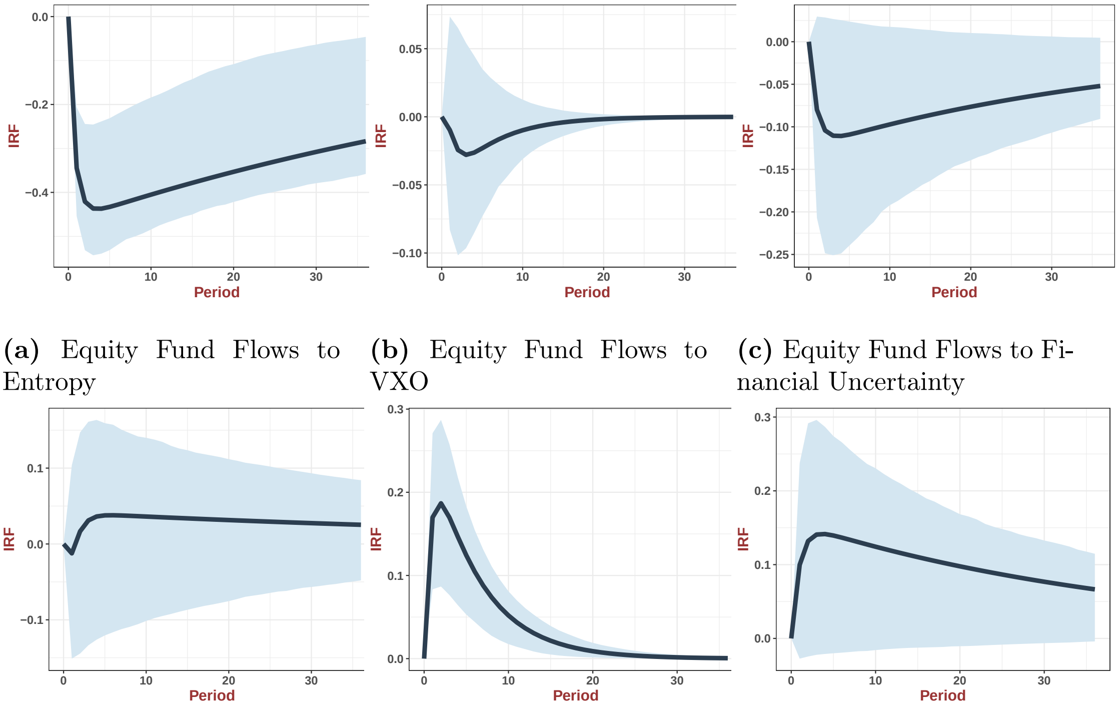


(a) Model Uncertainty in 4-Year Rolling Window



(b) Model Uncertainty in 5-Year Rolling Window

Figure 3.13: Alternative Rolling Windows



(a) Equity Fund Flows to Entropy

(b) Equity Fund Flows to VXO

(c) Equity Fund Flows to Financial Uncertainty

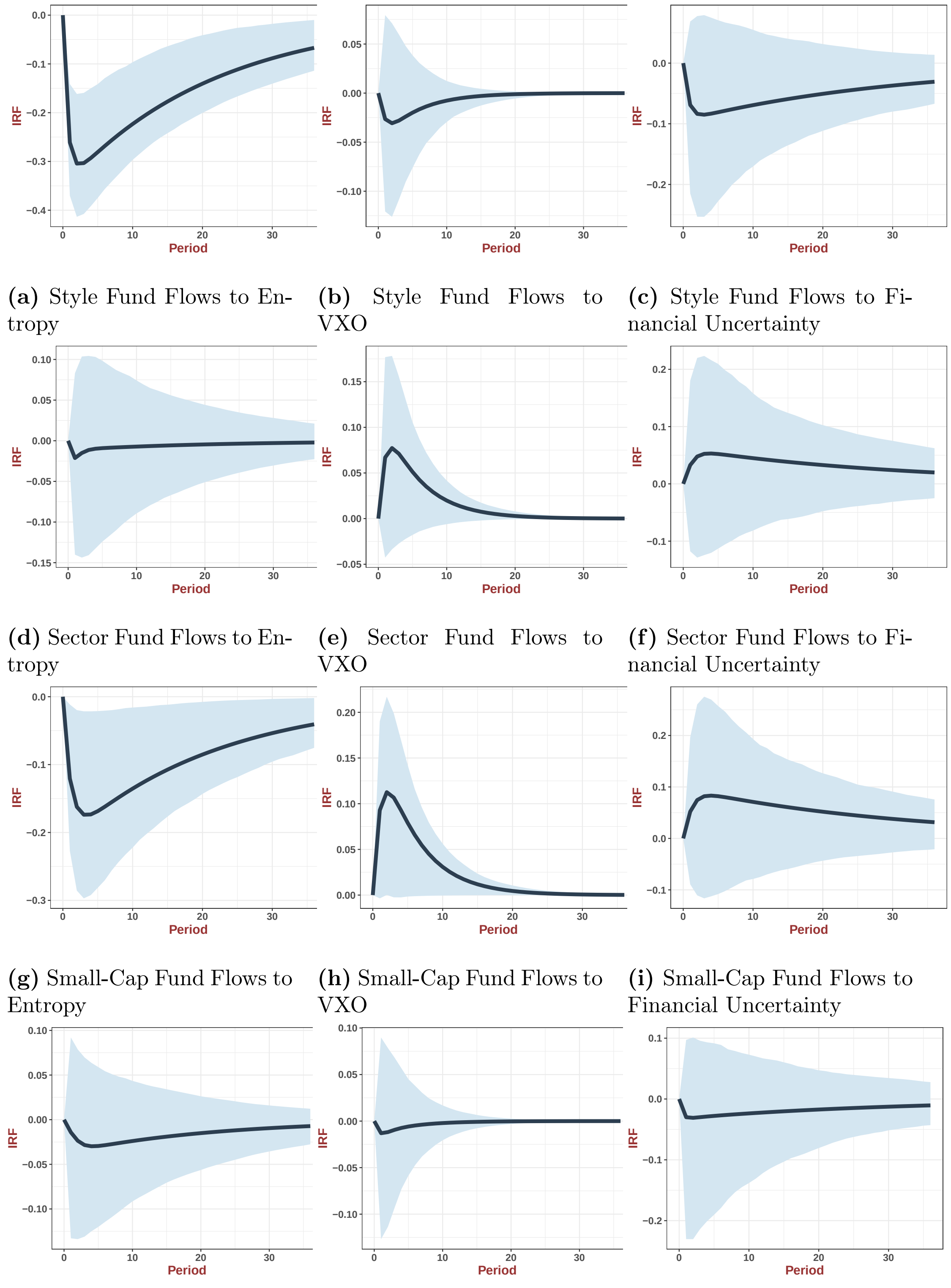
(d) Fixed-Income Fund Flows to Entropy

(e) Fixed-Income Fund Flows to VXO

(f) Fixed-Income Fund Flows to Financial Uncertainty

Figure 3.14: Robustness Check: Impulse Responses of Equity and Fixed-Income Fund Flows under Alternative Identification Assumption

This figure shows the dynamic impulse response functions (IRFs) of equity and fixed-income fund flows to uncertainty shocks in VAR-1. We identity the IRFs by putting uncertainty last in VAR.



(a) Style Fund Flows to Entropy

(b) Style Fund Flows to VXO

(c) Style Fund Flows to Financial Uncertainty

(d) Sector Fund Flows to Entropy

(e) Sector Fund Flows to VXO

(f) Sector Fund Flows to Financial Uncertainty

(g) Small-Cap Fund Flows to Entropy

(h) Small-Cap Fund Flows to VXO

(i) Small-Cap Fund Flows to Financial Uncertainty

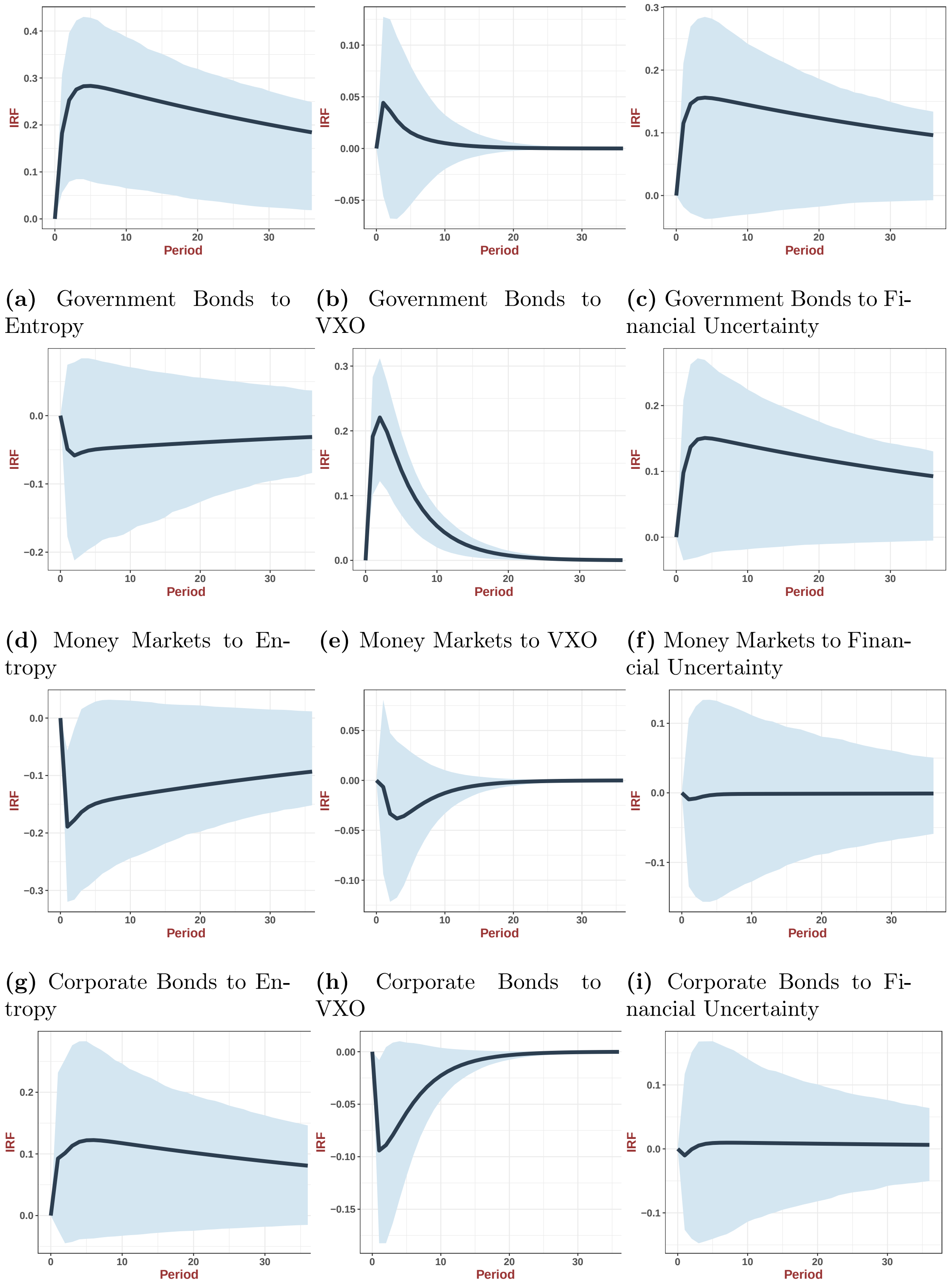
(j) Large-Cap Fund Flows to Entropy

(k) Large-Cap Fund Flows to VXO

(l) Large-Cap Fund Flows to Financial Uncertainty

Figure 3.15: Robustness Check: Impulse Responses of Equity Fund Flows with Different Investment Objective Codes under Alternative Identification Assumption

This figure shows the dynamic impulse response functions (IRFs) of equity fund flows to uncertainty shocks in VAR-1. We identity the IRFs by putting uncertainty last in VAR.



(a) Government Bonds to Entropy

(b) Government Bonds to VXO

(c) Government Bonds to Financial Uncertainty

(d) Money Markets to Entropy

(e) Money Markets to VXO

(f) Money Markets to Financial Uncertainty

(g) Corporate Bonds to Entropy

(h) Corporate Bonds to VXO

(i) Corporate Bonds to Financial Uncertainty

(j) Municipal Bonds to Entropy

(k) Municipal Bonds to VXO

(l) Municipal Bonds to Financial Uncertainty

Figure 3.16: Robustness Check: Impulse Responses of Fixed-Income Fund Flows with Different Investment Objective Codes under Alternative Identification Assumption

This figure shows the dynamic impulse response functions (IRFs) of fixed-income fund flows to uncertainty shocks in VAR-1. We identity the IRFs by putting uncertainty last in VAR.

3.9.3 Proofs

Proof of Proposition 12

Proof. As in section 3.1.2, we assign g-prior for $\boldsymbol{b}_\gamma$: $\boldsymbol{b}_\gamma \mid \mathcal{M}_\gamma,\, g \sim \mathcal{N}\left(\mathbf{0},\, \frac{g}{T}\left(\boldsymbol{C}_\gamma^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{C}_\gamma\right)^{-1}\right)$. From lemma 6, $\boldsymbol{C}_\gamma^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{C}_\gamma = \boldsymbol{V}_\gamma$, so the prior distribution for $\boldsymbol{b}_\gamma$ is simplified as $\mathcal{N}\left(\mathbf{0},\, \frac{g}{T}\boldsymbol{V}_\gamma^{-1}\right)$. Thus, the variance of linear SDF m_γ, conditioned that g and $\boldsymbol{V}_\gamma$ are known, is

$$\begin{aligned}
\operatorname{var}[m_\gamma] &= \mathbb{E}\left[\operatorname{var}\left[(\boldsymbol{f}_\gamma - \mathbb{E}[\boldsymbol{f}_\gamma])^\top \boldsymbol{b}_\gamma \mid \boldsymbol{b}_\gamma\right]\right] + \operatorname{var}\left[\mathbb{E}\left[1 - (\boldsymbol{f}_\gamma - \mathbb{E}[\boldsymbol{f}_\gamma])^\top \boldsymbol{b}_\gamma \mid \boldsymbol{b}_\gamma\right]\right] \\
&= \mathbb{E}\left[\operatorname{tr}\left(\boldsymbol{b}_\gamma^\top \boldsymbol{V}_\gamma \boldsymbol{b}_\gamma\right)\right] + \operatorname{var}\left[1 - \mathbf{0}^\top \boldsymbol{b}_\gamma\right] \\
&= \operatorname{tr}\left(\boldsymbol{V}_\gamma \mathbb{E}\left[\boldsymbol{b}_\gamma \boldsymbol{b}_\gamma^\top\right]\right) + 0 \\
&= \operatorname{tr}\left(\boldsymbol{V}_\gamma \frac{g}{T} \boldsymbol{V}_\gamma^{-1}\right) \\
&= \frac{g p_\gamma}{T}
\end{aligned}$$

This completes the proof of Proposition 12. □

Proof of Proposition 13

We begin the proof of Proposition 13 with the following lemma.

Lemma 6. *Define* $\boldsymbol{V}_\gamma = \operatorname{var}[\boldsymbol{f}_\gamma]$, $\boldsymbol{C}_\gamma = \operatorname{cov}[\boldsymbol{R}, \boldsymbol{f}_\gamma]$, *and* $\boldsymbol{\Sigma} = \operatorname{var}[\boldsymbol{R}]$, *then*

$$\boldsymbol{\Sigma}^{-1}\boldsymbol{C}_\gamma = \begin{pmatrix} \boldsymbol{I}_{p_\gamma} \\ \mathbf{0}_{(N-p_\gamma)} \end{pmatrix}, \qquad \boldsymbol{R}\boldsymbol{\Sigma}^{-1}\boldsymbol{C}_\gamma = \boldsymbol{f}_\gamma, \qquad \boldsymbol{C}_\gamma^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{C}_\gamma = \boldsymbol{V}_\gamma.$$

Proof. Recall that under our specification, it is always that $\boldsymbol{f}_\gamma \subseteq \boldsymbol{f} \subseteq \boldsymbol{R}$. Without loss of generality, the vector $\boldsymbol{R}$ can be arranged as

$$\boldsymbol{R} = \begin{pmatrix} \boldsymbol{f}_\gamma \\ \boldsymbol{f}_{-\gamma} \\ \boldsymbol{r}^e \end{pmatrix}$$

where $\boldsymbol{r}_t^e$ is a vector of test assets that are excess returns themselves but are excluded from factors under consideration (i.e., $\boldsymbol{f}$). Then

$$\boldsymbol{\Sigma} = \operatorname{var}[\boldsymbol{R}] = \begin{pmatrix} \boldsymbol{V}_\gamma & \boldsymbol{U}_\gamma^\top \\ \boldsymbol{U}_\gamma & \boldsymbol{V}_{-\gamma} \end{pmatrix}, \quad \boldsymbol{C}_\gamma = \operatorname{cov}[\boldsymbol{R}, \boldsymbol{f}_\gamma] = \begin{pmatrix} \boldsymbol{V}_\gamma \\ \boldsymbol{U}_\gamma \end{pmatrix},$$

where

$$\boldsymbol{V}_\gamma = \operatorname{var}[\boldsymbol{f}_\gamma], \quad \boldsymbol{V}_{-\gamma} = \operatorname{var}\left[\begin{pmatrix} \boldsymbol{f}_{-\gamma} \\ \boldsymbol{r}^e \end{pmatrix}\right], \quad \boldsymbol{U}_\gamma = \operatorname{cov}\left[\begin{pmatrix} \boldsymbol{f}_{-\gamma} \\ \boldsymbol{r}^e \end{pmatrix}, \boldsymbol{f}_\gamma\right].$$

Inverting $\boldsymbol{\Sigma}$ blockwise, we have

$$\boldsymbol{\Sigma}^{-1} = \begin{pmatrix} (\boldsymbol{V}_\gamma - \boldsymbol{U}_\gamma^\top \boldsymbol{V}_{-\gamma}^{-1}\boldsymbol{U}_\gamma)^{-1} & -\boldsymbol{V}_\gamma^{-1}\boldsymbol{U}_\gamma^\top(\boldsymbol{V}_{-\gamma} - \boldsymbol{U}_\gamma \boldsymbol{V}_{-\gamma}^{-1}\boldsymbol{U}_\gamma^\top)^{-1} \\ -\boldsymbol{V}_{-\gamma}^{-1}\boldsymbol{U}_\gamma(\boldsymbol{V}_\gamma - \boldsymbol{U}_\gamma^\top \boldsymbol{V}_\gamma^{-1}\boldsymbol{U}_\gamma)^{-1} & (\boldsymbol{V}_{-\gamma} - \boldsymbol{U}_\gamma \boldsymbol{V}_\gamma^{-1}\boldsymbol{U}_\gamma^\top)^{-1} \end{pmatrix}.$$

or exchanging the two off-diagonal blocks and taking transposes,

$$\boldsymbol{\Sigma}^{-1} = \begin{pmatrix} (\boldsymbol{V}_\gamma - \boldsymbol{U}_\gamma^\top \boldsymbol{V}_{-\gamma}^{-1}\boldsymbol{U}_\gamma)^{-1} & -(\boldsymbol{V}_\gamma - \boldsymbol{U}_\gamma^\top \boldsymbol{V}_\gamma^{-1}\boldsymbol{U}_\gamma)^{-1}\boldsymbol{U}_\gamma^\top \boldsymbol{V}_{-\gamma}^{-1} \\ -(\boldsymbol{V}_{-\gamma} - \boldsymbol{U}_\gamma \boldsymbol{V}_{-\gamma}^{-1}\boldsymbol{U}_\gamma^\top)^{-1}\boldsymbol{U}_\gamma \boldsymbol{V}_\gamma^{-1} & (\boldsymbol{V}_{-\gamma} - \boldsymbol{U}_\gamma \boldsymbol{V}_\gamma^{-1}\boldsymbol{U}_\gamma^\top)^{-1} \end{pmatrix}.$$

Thus

$$\begin{aligned} \boldsymbol{\Sigma}^{-1}\boldsymbol{C}_\gamma &= \begin{pmatrix} (\boldsymbol{V}_\gamma - \boldsymbol{U}_\gamma^\top \boldsymbol{V}_{-\gamma}^{-1}\boldsymbol{U}_\gamma)^{-1} & -(\boldsymbol{V}_\gamma - \boldsymbol{U}_\gamma^\top \boldsymbol{V}_\gamma^{-1}\boldsymbol{U}_\gamma)^{-1}\boldsymbol{U}_\gamma^\top \boldsymbol{V}_{-\gamma}^{-1} \\ -(\boldsymbol{V}_{-\gamma} - \boldsymbol{U}_\gamma \boldsymbol{V}_{-\gamma}^{-1}\boldsymbol{U}_\gamma^\top)^{-1}\boldsymbol{U}_\gamma \boldsymbol{V}_\gamma^{-1} & (\boldsymbol{V}_{-\gamma} - \boldsymbol{U}_\gamma \boldsymbol{V}_\gamma^{-1}\boldsymbol{U}_\gamma^\top)^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{V}_\gamma \\ \boldsymbol{U}_\gamma \end{pmatrix} \\ &= \begin{pmatrix} (\boldsymbol{V}_\gamma - \boldsymbol{U}_\gamma^\top \boldsymbol{V}_{-\gamma}^{-1}\boldsymbol{U}_\gamma)^{-1}\boldsymbol{V}_\gamma - (\boldsymbol{V}_\gamma - \boldsymbol{U}_\gamma^\top \boldsymbol{V}_\gamma^{-1}\boldsymbol{U}_\gamma)^{-1}\boldsymbol{U}_\gamma^\top \boldsymbol{V}_{-\gamma}^{-1}\boldsymbol{U}_\gamma \\ -(\boldsymbol{V}_{-\gamma} - \boldsymbol{U}_\gamma \boldsymbol{V}_{-\gamma}^{-1}\boldsymbol{U}_\gamma^\top)^{-1}\boldsymbol{U}_\gamma + (\boldsymbol{V}_{-\gamma} - \boldsymbol{U}_\gamma \boldsymbol{V}_\gamma^{-1}\boldsymbol{U}_\gamma^\top)^{-1}\boldsymbol{U}_\gamma \end{pmatrix} \\ &= \begin{pmatrix} \boldsymbol{I}_{p_\gamma} \\ \boldsymbol{0}_{(N-p_\gamma)} \end{pmatrix}, \end{aligned}$$

which directly implies that $\boldsymbol{R}\boldsymbol{\Sigma}^{-1}\boldsymbol{C}_\gamma = \boldsymbol{f}_\gamma$ and that $\boldsymbol{C}_\gamma^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{C}_\gamma = \boldsymbol{V}_\gamma$. □

Under this lemma, we prove Proposition 13 as follows.

Proof. Since

$$[\boldsymbol{R}_t \mid \boldsymbol{b}_\gamma, \mathcal{M}_\gamma] \overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{C}_\gamma \boldsymbol{b}_\gamma, \boldsymbol{\Sigma}), \quad t = 1, \ldots, T,$$

under our distributional assumption and

$$[\boldsymbol{b}_\gamma \mid \mathcal{M}_\gamma, g] \sim \mathcal{N}\left(\boldsymbol{0}, \frac{g}{T}\left(\boldsymbol{C}_\gamma^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{C}_\gamma\right)^{-1}\right),$$

under our g-prior specification, we can integrate out $\boldsymbol{b}_\gamma$ and reach the following distributional results for the observed dataset $\mathcal{D} = \{\boldsymbol{R}_1, \ldots, \boldsymbol{R}_T\}$:

$$[\boldsymbol{R}_1^\top, \ldots, \boldsymbol{R}_T^\top]^\top \triangleq \boldsymbol{R}_{[1:T]} \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{I}_T \otimes \boldsymbol{\Sigma} + \frac{g}{T}(\boldsymbol{1}_T \otimes \boldsymbol{C}_\gamma)\left(\boldsymbol{C}_\gamma^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{C}_\gamma\right)^{-1}(\boldsymbol{1}_T \otimes \boldsymbol{C}_\gamma)^\top\right),$$

where ⊗ performs the matrix Kronecker product. As a result,

$$\begin{aligned} &\mathbb{P}[\mathcal{D} \mid \mathcal{M}_\gamma, g] \\ =& \exp\left\{-\frac{1}{2}\boldsymbol{R}_{[1:T]}^\top \left[\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}^{-1} + \frac{g}{T}(\boldsymbol{1}_T \otimes \boldsymbol{C}_\gamma)\left(\boldsymbol{C}_\gamma^\top \boldsymbol{\Sigma}\boldsymbol{C}_\gamma\right)^{-1}(\boldsymbol{1}_T \otimes \boldsymbol{C}_\gamma)^\top\right]^{-1} \boldsymbol{R}_{[1:T]}\right\} \\ &\times \left|\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}^{-1} + \frac{g}{T}(\boldsymbol{1}_T \otimes \boldsymbol{C}_\gamma)\left(\boldsymbol{C}_\gamma^\top \boldsymbol{\Sigma}\boldsymbol{C}_\gamma\right)^{-1}(\boldsymbol{1}_T \otimes \boldsymbol{C}_\gamma)^\top\right|^{-\frac{1}{2}} (2\pi)^{-\frac{NT}{2}}. \end{aligned}$$

By the Sherman-Morrison-Woodbury formula,[30]

$$
\begin{aligned}
&\left[\boldsymbol{I}_T \otimes \boldsymbol{\Sigma} + \frac{g}{T}(\mathbf{1}_T \otimes \boldsymbol{C}_{\gamma})\left(\boldsymbol{C}_{\gamma}^{\top}\boldsymbol{\Sigma}^{-1}\boldsymbol{C}_{\gamma}\right)^{-1}(\mathbf{1}_T \otimes \boldsymbol{C}_{\gamma})^{\top}\right]^{-1} \\
=&\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}^{-1} - \\
&\quad[\mathbf{1}_T \otimes (\boldsymbol{\Sigma}^{-1}\boldsymbol{C}_{\gamma})]\left(\frac{T}{g}\boldsymbol{C}_{\gamma}^{\top}\boldsymbol{\Sigma}^{-1}\boldsymbol{C}_{\gamma} + (\mathbf{1}_T \otimes \boldsymbol{C}_{\gamma})^{\top}(\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}^{-1})(\mathbf{1}_T \otimes \boldsymbol{C}_{\gamma})\right)^{-1}[\mathbf{1}_T^{\top} \otimes (\boldsymbol{C}_{\gamma}^{\top}\boldsymbol{\Sigma}^{-1})] \\
=&\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}^{-1} - \frac{g}{(1+g)T}[\mathbf{1}_T \otimes (\boldsymbol{\Sigma}^{-1}\boldsymbol{C}_{\gamma})]\left(\boldsymbol{C}_{\gamma}^{\top}\boldsymbol{\Sigma}^{-1}\boldsymbol{C}_{\gamma}\right)^{-1}[\mathbf{1}_T^{\top} \otimes (\boldsymbol{C}_{\gamma}^{\top}\boldsymbol{\Sigma}^{-1})].
\end{aligned}
$$

By the generalized Sylvester's theorem for determinants,[31]

$$
\begin{aligned}
&\left|\boldsymbol{I}_T \otimes \boldsymbol{\Sigma} + \frac{g}{T}(\mathbf{1}_T \otimes \boldsymbol{C}_{\gamma})\left(\boldsymbol{C}_{\gamma}^{\top}\boldsymbol{\Sigma}^{-1}\boldsymbol{C}_{\gamma}\right)^{-1}(\mathbf{1}_T \otimes \boldsymbol{C}_{\gamma})^{\top}\right| \\
=&\frac{|Tg^{-1}\boldsymbol{C}_{\gamma}^{\top}\boldsymbol{\Sigma}^{-1}\boldsymbol{C}_{\gamma} + (\mathbf{1}_T \otimes \boldsymbol{C}_{\gamma})^{\top}(\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}^{-1})(\mathbf{1}_T \otimes \boldsymbol{C}_{\gamma})|}{|Tg^{-1}\boldsymbol{C}_{\gamma}^{\top}\boldsymbol{\Sigma}^{-1}\boldsymbol{C}_{\gamma}| \times |\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}^{-1}|} \\
=&\frac{|(g^{-1}+1)\boldsymbol{C}_{\gamma}^{\top}\boldsymbol{\Sigma}^{-1}\boldsymbol{C}_{\gamma}|}{|g^{-1}\boldsymbol{C}_{\gamma}^{\top}\boldsymbol{\Sigma}^{-1}\boldsymbol{C}_{\gamma}| \times |\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}^{-1}|} \\
=&\frac{(1+g)^{p_{\gamma}}}{|\boldsymbol{\Sigma}^{-1}|^{T}},
\end{aligned}
$$

the last equation of which is due to the fact that $\boldsymbol{C}_{\gamma}^{\top}\boldsymbol{\Sigma}^{-1}\boldsymbol{C}_{\gamma} = \boldsymbol{V}_{\gamma}$ according to the lemma.

Plugging the two results above back to our original formula of $\mathbb{P}[\mathcal{D} \mid \mathcal{M}_{\gamma},\, g]$, we get

$$
\begin{aligned}
&\mathbb{P}[\mathcal{D} \mid \mathcal{M}_{\gamma},\, g] \\
=&\exp\left\{-\frac{1}{2}\boldsymbol{R}_{[1:T]}^{\top}\left[\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}^{-1} - \frac{g}{(1+g)T}[\mathbf{1}_T \otimes (\boldsymbol{\Sigma}^{-1}\boldsymbol{C}_{\gamma})]\left(\boldsymbol{C}_{\gamma}^{\top}\boldsymbol{\Sigma}^{-1}\boldsymbol{C}_{\gamma}\right)^{-1}[\mathbf{1}_T^{\top} \otimes (\boldsymbol{C}_{\gamma}^{\top}\boldsymbol{\Sigma}^{-1})]\right]\boldsymbol{R}_{[1:T]}\right\} \\
&\quad\times\frac{|\boldsymbol{\Sigma}^{-1}|^{\frac{T}{2}}}{(1+g)^{\frac{p_{\gamma}}{2}}(2\pi)^{\frac{NT}{2}}} \\
=&\exp\left\{-\frac{1}{2}\sum_{t=1}^{T}\boldsymbol{R}_t^{\top}\boldsymbol{\Sigma}^{-1}\boldsymbol{R}_t + \frac{g}{1+g}\frac{T}{2}\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{f}_{\gamma,t}\right)^{\top}\boldsymbol{V}_{\gamma}^{-1}\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{f}_{\gamma,t}\right)\right\}\frac{(1+g)^{-\frac{p_{\gamma}}{2}}}{(2\pi)^{\frac{NT}{2}}|\boldsymbol{\Sigma}|^{\frac{T}{2}}} \\
=&\exp\left\{-\frac{T-1}{2}\mathrm{tr}\left(\boldsymbol{S}\boldsymbol{\Sigma}^{-1}\right) - \frac{T}{2}\Big(\underbrace{\overline{\boldsymbol{R}}^{\top}\boldsymbol{\Sigma}^{-1}\overline{\boldsymbol{R}}}_{\mathrm{SR}^2_{\max}} - \frac{g}{1+g}\underbrace{\overline{\boldsymbol{f}}_{\gamma}^{\top}\boldsymbol{V}_{\gamma}^{-1}\overline{\boldsymbol{f}}_{\gamma}}_{\mathrm{SR}^2_{\gamma}}\Big)\right\}\frac{(1+g)^{-\frac{p_{\gamma}}{2}}}{(2\pi)^{\frac{NT}{2}}|\boldsymbol{\Sigma}|^{\frac{T}{2}}}
\end{aligned}
$$

where $\overline{\boldsymbol{R}} = \left(\sum_{t=1}^{T}\boldsymbol{R}_t\right)/T$, $\overline{\boldsymbol{f}}_{\gamma} = \left(\overline{\boldsymbol{f}}_{\gamma,t}\right)/T$; the second equation is due to the fact that $\boldsymbol{R}\boldsymbol{\Sigma}^{-1}\boldsymbol{C}_{\gamma} = \boldsymbol{f}_{\gamma}$ and that $\boldsymbol{C}_{\gamma}^{\top}\boldsymbol{\Sigma}^{-1}\boldsymbol{C}_{\gamma} = \boldsymbol{V}_{\gamma}$ as demonstrated in the lemma. □

[30] $(A+UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1}+VA^{-1}U)^{-1}VA^{-1}$ for "well-behaved" matrices A, U, C, V.

[31] $|X+ACB| = |X| \times |C| \times |C^{-1}+BX^{-1}A|$ for "well-behaved" matrices A, B, C, X.

Proof of Proposition 14

Proof. When comparing different models, the common factor unrelated to $(\mathcal{M}_\gamma,\, g)$ can be ignored, so we simplify the marginal likelihood of data as following:

$$p[\mathcal{D} \mid \mathcal{M}_\gamma,\, g] \propto (1+g)^{-\frac{p_\gamma}{2}} \exp\left\{\frac{gT}{2(1+g)}\mathrm{SR}^2_\gamma\right\} \tag{3.19}$$

An equivalent way to think about equation (3.19) is to treat it as the Bayes factor of model $\mathcal{M}_\gamma$ relative to $\mathcal{M}_0$. One amazing fact is that $p[\mathcal{D} \mid \mathcal{M}_0,\, g]$ does not depend on g[32]. Therefore, the Bayes factor could be defined as

$$\mathrm{BF}_\gamma(g) = \frac{p[\mathcal{D} \mid \mathcal{M}_\gamma,\, g]}{p[\mathcal{D} \mid \mathcal{M}_0,\, g]} = (1+g)^{-\frac{p_\gamma}{2}} \exp\left\{\frac{gT}{2(1+g)}\mathrm{SR}^2_\gamma\right\}$$

The prior for g is such that $\pi[g] = \frac{a-2}{2}(1+g)^{-\frac{a}{2}}$. We calculate the marginal likelihood of data only conditional on model $\mathcal{M}_\gamma$ by integrating out g in equation (3.19).

$$\begin{aligned}
p[\mathcal{D} \mid \mathcal{M}_\gamma] &\propto \frac{a-2}{2}\int_0^\infty (1+g)^{-\frac{p_\gamma+a}{2}} \exp\left\{\frac{g}{1+g}\left[\frac{T}{2}\mathrm{SR}^2_\gamma\right]\right\} \mathrm{d}g \\
&= \frac{a-2}{2}\exp\left\{\frac{T}{2}\mathrm{SR}^2_\gamma\right\}\int_0^\infty (1+g)^{-\frac{p_\gamma+a}{2}} \exp\left\{-\frac{1}{1+g}\left[\frac{T}{2}\mathrm{SR}^2_\gamma\right]\right\} \mathrm{d}g \\
&= \frac{a-2}{2}\exp\left\{\frac{T}{2}\mathrm{SR}^2_\gamma\right\}\int_0^1 k^{\frac{p_\gamma+a}{2}-2} \exp\left\{-k\left[\frac{T}{2}\mathrm{SR}^2_\gamma\right]\right\} \mathrm{d}k \\
&= \frac{a-2}{2}\exp\left\{\frac{T}{2}\mathrm{SR}^2_\gamma\right\}\left(\frac{T}{2}\mathrm{SR}^2_\gamma\right)^{1-\frac{p_\gamma+a}{2}}\int_0^{\frac{T}{2}\mathrm{SR}^2_\gamma} t^{\frac{p_\gamma+a}{2}-2} e^{-t}\,\mathrm{d}t \\
&= \frac{a-2}{2}\exp\left\{\frac{T}{2}\mathrm{SR}^2_\gamma\right\}\left(\frac{T}{2}\mathrm{SR}^2_\gamma\right)^{-s_\gamma}\underline{\Gamma}\left(s_\gamma,\, \frac{T}{2}\mathrm{SR}^2_\gamma\right)
\end{aligned}$$

where $\underline{\Gamma}(s,x) = \int_0^x t^{s-1}e^{-t}\,\mathrm{d}t$ is the lower incomplete Gamma function; the scalar s_γ is defined as $s_\gamma = \frac{p_\gamma+a}{2} - 1$. We have proved the formula of Bayes factor BF_γ in Proposition 3. To prove that the Bayes factor is always increasing in SR^2_γ always decreasing in p_γ, we use the original representation of Bayes Factor, that is,

$$\mathrm{BF}_\gamma = \frac{a-2}{2}\int_0^\infty (1+g)^{-\frac{p_\gamma+a}{2}} \exp\left\{\frac{gT}{2(1+g)}\mathrm{SR}^2_\gamma\right\} \mathrm{d}g$$

Take the first-order derivative with respect to SR^2_γ and p_γ:

$$\frac{\partial \mathrm{BF}_\gamma}{\partial \mathrm{SR}^2_\gamma} = \frac{a-2}{2}\int_0^\infty \frac{gT}{2(1+g)}(1+g)^{-\frac{p_\gamma+a}{2}} \exp\left\{\frac{gT}{2(1+g)}\mathrm{SR}^2_\gamma\right\} \mathrm{d}g > 0,$$

$$\frac{\partial \mathrm{BF}_\gamma}{\partial p_\gamma} = \frac{a-2}{2}\int_0^\infty -\frac{\log(1+g)}{2}(1+g)^{-\frac{p_\gamma+a}{2}} \exp\left\{\frac{gT}{2(1+g)}\mathrm{SR}^2_\gamma\right\} \mathrm{d}g < 0,$$

□

[32] $p[\mathcal{D} \mid \mathcal{M}_0,\, g] = (2\pi)^{-\frac{NT}{2}}|\boldsymbol{\Sigma}|^{-\frac{T}{2}} \exp\left\{-\frac{T-1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}\right) - \frac{T}{2}\mathrm{SR}^2_{\max}\right\}$.

Bibliography

Abel, A. B. (1983). Optimal investment under uncertainty. *American Economic Review 73*(1), 228–233.

Abramowitz, M. and I. A. Stegun (1965). *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables*, Volume 55. Courier Corporation.

Acharya, V. V., L. A. Lochstoer, and T. Ramadorai (2013). Limits to arbitrage and hedging: Evidence from commodity markets. *Journal of Financial Economics 109*(2), 441–465.

Adrian, T. and H. S. Shin (2014). Procyclical leverage and value-at-risk. *Review of Financial Studies 27*(2), 373–403.

Aït-Sahalia, Y. and A. W. Lo (1998). Nonparametric estimation of state-price densities implicit in financial asset prices. *Journal of Finance 53*(2), 499–547.

Akram, Q. F., D. Rime, and L. Sarno (2008). Arbitrage in the foreign exchange market: Turning on the microscope. *Journal of International Economics 76*(2), 237–253.

Akram, Q. F., D. Rime, and L. Sarno (2009). Does the law of one price hold in international financial markets? Evidence from tick data. *Journal of Banking and Finance 33*(10), 1741–1754.

Amiti, M., P. McGuire, and D. E. Weinstein (2019). International bank flows and the global financial cycle. *IMF Economic Review 67*(1), 61–108.

Andersen, L., D. Duffie, and Y. Song (2019). Funding value adjustments. *Journal of Finance 74*(1), 145–192.

Andersen, T. G., L. Benzoni, and J. Lund (2002). An empirical investigation of continuous-time equity return models. *Journal of Finance 57*(3), 1239–1284.

Andersen, T. G., T. Bollerslev, F. X. Diebold, and H. Ebens (2001). The distribution of realized stock return volatility. *Journal of Financial Economics 61*(1), 43–76.

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

Asness, C. and A. Frazzini (2013). The devil in hml's details. *The Journal of Portfolio Management 39*(4), 49–68.

Asness, C. S., A. Frazzini, and L. H. Pedersen (2019). Quality minus junk. *Review of Accounting Studies 24*(1), 34–112.

Atkeson, A. G., A. d'Avernas, A. L. Eisfeldt, and P.-O. Weill (2019). Government guarantees and the valuation of American banks. *NBER Macroeconomics Annual 33*(1), 81–145.

Augustin, P., M. Chernov, L. Schmid, and D. Song (2020). The term structure of CIP violations. Technical report, National Bureau of Economic Research.

Avdjiev, S., W. Du, C. Koch, and H. S. Shin (2019). The dollar, bank leverage, and deviations from covered interest parity. *American Economic Review: Insights 1*(2), 193–208.

Avramov, D. (2002). Stock return predictability and model uncertainty. *Journal of Financial Economics 64*(3), 423–458.

Avramov, D., S. Cheng, L. Metzker, and S. Voigt (2021). Integrating factor models.

Baba, N. and F. Packer (2009). Interpreting deviations from covered interest parity during the financial market turmoil of 2007–08. *Journal of Banking and Finance 33*(11), 1953–1962.

Baker, S. R., N. Bloom, and S. J. Davis (2016). Measuring economic policy uncertainty. *The Quarterly Journal of Economics 131*(4), 1593–1636.

Barberis, N. (2000). Investing for the long run when returns are predictable. *Journal of Finance 55*(1), 225–264.

Barillas, F. and J. Shanken (2018). Comparing asset pricing models. *Journal of Finance 73*(2), 715–754.

Baron, M., E. Verner, and W. Xiong (2020, 10). Banking Crises Without Panics*. *The Quarterly Journal of Economics 136*(1), 51–113.

Bartlett, M. (1957). A comment on d.v. lindley's statistical paradox. *Biometrika 44*(3-4), 533–533.

Bates, D. S. (1991). The crash of '87: Was it expected? the evidence from options markets. *Journal of Finance 46*(3), 1009–1044.

Berger, J. O. and L. R. Pericchi (1996). The intrinsic bayes factor for model selection and prediction. *Journal of the American Statistical Association 91*(433), 109–122.

Bernardo, A. E. and O. Ledoit (2000). Gain, loss, and asset pricing. *Journal of Political Economy 108*(1), 144–172.

Berry, S., J. Levinsohn, and A. Pakes (1995). Automobile prices in market equilibrium. *Econometrica 63*(4), 841–890.

Bertsekas, D. P. (2009). *Convex Optimization Theory*. Athena Scientific Belmont.

Bloom, N. (2009). The impact of uncertainty shocks. *Econometrica 77*(3), 623–685.

Borio, C. E., R. N. McCauley, P. McGuire, and V. Sushko (2016). Covered interest parity lost: understanding the cross-currency basis. *BIS Quarterly Review*.

Boyer, B., T. Mitton, and K. Vorkink (2009). Expected idiosyncratic skewness. *Review of Financial Studies 23*(1), 169–202.

Breeden, D. T. and R. H. Litzenberger (1978). Prices of state-contingent claims implicit in option prices. *Journal of Business 51*(4), 621–651.

Breiman, L. (1960). Investment policies for expanding businesses optimal in a long-run sense. *Naval Research Logistics 7*(4), 647–651.

Brunnermeier, M. K. and Y. Sannikov (2014). A macroeconomic model with a financial sector. *American Economic Review 104*(2), 379–421.

Bryzgalova, S., J. Huang, and C. Julliard (2021). Bayesian solutions for the factor zoo: We just ran two quadrillion models.

Burkard, R., M. Dell'Amico, and S. Martello (2012). *Assignment Problems: Revised Reprint*. SIAM.

Caballero, R. J. and A. Krishnamurthy (2008). Collective risk management in a flight to quality episode. *Journal of Finance 63*(5), 2195–2230.

Calomiris, C. W. and D. Nissim (2014). Crisis-related shifts in the market valuation of banking activities. *Journal of Financial Intermediation 23*(3), 400–435.

Carhart, M. M. (1997). On persistence in mutual fund performance. *Journal of Finance 52*(1), 57–82.

Carr, P. and D. Madan (2001). *Towards a Theory of Volatility Trading*, pp. 458–476. Cambridge University Press.

Carr, P. and L. Wu (2009). Variance risk premiums. *Review of Financial Studies 22*(3), 1311–1341.

Chen, J., H. Hong, and J. C. Stein (2001). Forecasting crashes: Trading volume, past returns, and conditional skewness in stock prices. *Journal of financial Economics 61*(3), 345–381.

Chib, S., X. Zeng, and L. Zhao (2020). On comparing asset pricing models. *Journal of Finance 75*(1), 551–577.

Cochrane, J. H. (2005). *Asset pricing: Revised edition.* Princeton University Press.

Cochrane, J. H. and J. Saá-Requejo (2000). Beyond arbitrage: Good-deal asset price bounds in incomplete markets. *Journal of Political Economy 108*(1), 79–119.

Constantinides, G. M., J. C. Jackwerth, and S. Perrakis (2008). Mispricing of S&P 500 index options. *Review of Financial Studies 22*(3), 1247–1277.

Coval, J. and E. Stafford (2007). Asset fire sales (and purchases) in equity markets. *Journal of Financial Economics 86*(2), 479–512.

Cremers, M. (2002). Stock return predictability: A bayesian model selection perspective. *Review of Financial Studies 15*(4), 1223–1249.

Cvitanić, J. and I. Karatzas (1992). Convex duality in constrained portfolio optimization. *Annals of Applied Probability 2*(4), 767–818.

Daniel, K., D. Hirshleifer, and L. Sun (2020). Short-and long-horizon behavioral factors. *The Review of Financial Studies 33*(4), 1673–1736.

Dew-Becker, I. and S. Giglio (2021+). Cross-sectional uncertainty and the business cycle: evidence from 40 years of options data. *American Economic Jounral: Macroeconomics forthcoming.*

Driscoll, J. C. and A. C. Kraay (1998). Consistent covariance matrix estimation with spatially dependent panel data. *Review of Economics and Statistics 80*(4), 549–560.

Du, W., A. Tepper, and A. Verdelhan (2018). Deviations from covered interest rate parity. *Journal of Finance 73*(3), 915–957.

Efron, B. (2012). *Large-scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*, Volume 1. Cambridge University Press.

Eilers, P. H. and B. D. Marx (1996). Flexible smoothing with b-splines and penalties. *Statistical Science 11*(2), 89–121.

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

Fama, E. F. and K. R. French (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics 33*(1), 3–56.

Fama, E. F. and K. R. French (2016). Dissecting anomalies with a five-factor model. *The Review of Financial Studies 29*(1), 69–103.

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

Frenkel, J. A. and R. M. Levich (1975). Covered interest arbitrage: Unexploited profits? *Journal of Political Economy 83*(2), 325–338.

Frenkel, J. A. and R. M. Levich (1977). Transaction costs and interest arbitrage: Tranquil versus turbulent periods. *Journal of Political Economy 85*(6), 1209–1226.

Gabaix, X. and M. Maggiori (2015). International liquidity and exchange rate dynamics. *Quarterly Journal of Economics 130*(3), 1369–1420.

Gârleanu, N. and L. H. Pedersen (2011). Margin-based asset pricing and deviations from the law of one price. *Review of Financial Studies 24*(6), 1980–2022.

Gârrleanu, N., L. H. Pedersen, and A. M. Poteshman (2008). Demand-based option pricing. *Review of Financial Studies 22*(10), 4259–4299.

Gibbons, M. R., S. A. Ross, and J. Shanken (1989). A test of the efficiency of a given portfolio. *Econometrica 57*(5), 1121–1152.

Giglio, S., B. Kelly, and D. Xiu (2021+). Factor models, machine learning, and asset pricing. *Annual Review of Financial Economics, Forthcoming forthcoming.*

Greenwood, R. and A. Shleifer (2014). Expectations of returns and expected returns. *The Review of Financial Studies 27*(3), 714–746.

Greenwood, R., A. Shleifer, and Y. You (2019). Bubbles for Fama. *Journal of Financial Economics 131*(1), 20–43.

Greenwood, R. and D. Vayanos (2010). Price pressure in the government bond market. *American Economic Review 100*(2), 585–90.

Gromb, D. and D. Vayanos (2002). Equilibrium and welfare in markets with financially constrained arbitrageurs. *Journal of Financial Economics 66*(2), 361–407.

Gromb, D. and D. Vayanos (2018). The dynamics of financially constrained arbitrage. *Journal of Finance 73*(4), 1713–1750.

Guerrieri, V. and R. Shimer (2014). Dynamic adverse selection: A theory of illiquidity, fire sales, and flight to quality. *American Economic Review 104*(7), 1875–1908.

Hansen, L. P. and R. Jagannathan (1991). Implications of security market data for models of dynamic economies. *Journal of political economy 99*(2), 225–262.

Hayashi, F. (1982). Tobin's marginal q and average q: A neoclassical interpretation. *Econometrica 50*(1), 213–224.

He, Z., B. Kelly, and A. Manela (2017). Intermediary asset pricing: New evidence from many asset classes. *Journal of Financial Economics 126*(1), 1–35.

He, Z. and A. Krishnamurthy (2013). Intermediary asset pricing. *American Economic Review 103*(2), 732–70.

Hodges, H. M. (1996). Arbitrage bounds of the implied volatility strike and term structures of european-style options. *Journal of Derivatives 3*(4), 23–35.

Hofer, M. and M. R. Iacò (2014). Optimal bounds for integrals with respect to copulas and applications. *Journal of Optimization Theory and Applications 3*, 999–1011.

Hou, K., C. Xue, and L. Zhang (2015). Digesting anomalies: An investment approach. *The Review of Financial Studies 28*(3), 650–705.

Ivashina, V., D. S. Scharfstein, and J. C. Stein (2015). Dollar funding and the lending behavior of global banks. *Quarterly Journal of Economics 130*(3), 1241–1281.

Jackwerth, J. C. and M. Rubinstein (1996). Recovering probability distributions from option prices. *Journal of Finance 51*(5), 1611–1631.

James, W. and C. Stein (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, Berkeley, Calif., pp. 361–379. University of California Press.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences 186*(1007), 453–461.

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

Jermann, U. (2020). Negative swap spreads and limited arbitrage. *Review of Financial Studies 33*(1), 212–238.

Jurado, K., S. C. Ludvigson, and S. Ng (2015). Measuring uncertainty. *American Economic Review 105*(3), 1177–1216.

Kandel, S. and R. F. Stambaugh (1996). On the predictability of stock returns: an asset-allocation perspective. *Journal of Finance 51*(2), 385–424.

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

Klingler, S. and S. Sundaresan (2019). An explanation of negative swap spreads: Demand for duration from underfunded pension plans. *Journal of Finance 74*(2), 675–710.

Koijen, R. S. and M. Yogo (2019). A demand system approach to asset pricing. *Journal of Political Economy 127*(4), 1475–1515.

Kondor, P. (2009). Risk in dynamic arbitrage: the price effects of convergence trading. *Journal of Finance 64*(2), 631–655.

Kondor, P. and D. Vayanos (2019). Liquidity risk and the dynamics of arbitrage capital. *Journal of Finance 74*(3), 1139–1173.

Kozak, S., S. Nagel, and S. Santosh (2018). Interpreting factor models. *The Journal of Finance 73*(3), 1183–1223.

Kozak, S., S. Nagel, and S. Santosh (2020). Shrinking the cross-section. *Journal of Financial Economics 135*(2), 271–292.

Kuhn, H. W. (1955). The Hungarian method for the assignment problem. *Naval Research Logistics 2*(1-2), 83–97.

Kyle, A. S. and W. Xiong (2001). Contagion as a wealth effect. *Journal of Finance 56*(4), 1401–1440.

Latane, H. A. (1959). Criteria for choice among risky ventures. *Journal of Political Economy 67*(2), 144–155.

Liang, F., R. Paulo, G. Molina, M. A. Clyde, and J. O. Berger (2008). Mixtures of g priors for bayesian variable selection. *Journal of the American Statistical Association 103*(481), 410–423.

Lindley, D. V. (2000). The philosophy of statistics. *Journal of the Royal Statistical Society: Series D (The Statistician) 49*(3), 293–337.

Lintner, J. (1965, December). Security prices, risk, and maximal gains from diversification. *Journal of Finance 20*, 587–615.

Longstaff, F. (2004). The flight-to-liquidity premium in u.s. treasury bond prices. *Journal of Business 77*(3), 511–526.

Lou, D. (2012). A flow-based explanation for return predictability. *Review of Financial Studies 25*(12), 3457–3489.

Lucas, R. E. (1967). Adjustment costs and the theory of supply. *Journal of Political Economy 75*(4), 321–334.

Ludvigson, S. C., S. Ma, and S. Ng (2021). Uncertainty and business cycles: exogenous impulse or endogenous response? *American Economic Journal: Macroeconomics 13*(4), 369–410.

Manela, A. and A. Moreira (2017). News implied volatility and disaster concerns. *Journal of Financial Economics 123*(1), 137–162.

Martin, I. (2017). What is the expected return on the market? *Quarterly Journal of Economics 132*(1), 367–433.

Martin, I. (2018). Options and the gamma knife. *Journal of Portfolio Management 44*(6), 47–55.

Martin, I. and C. Wagner (2019). What is the expected return on a stock? *Journal of Finance 74*(4), 1887–1929.

Mas-Colell, A., M. D. Whinston, and J. R. Green (1995). *Microeconomic Theory*. Oxford University Press.

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

Merton, R. C. (1973). An intertemporal capital asset pricing model. *Econometrica 41*(5), 867–887.

Molchanov, I. and F. Molinari (2018). *Random Sets in Econometrics*, Volume 60. Cambridge University Press.

Nelsen, R. B. (2007). *An Introduction to Copulas*. Springer Science & Business Media.

Newey, W. K. and K. D. West (1987). A simple, positive semidefinite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica 55*, 703–08.

O'Hagan, A. (1995). Fractional bayes factors for model comparison. *Journal of the Royal Statistical Society: Series B (Methodological) 57*(1), 99–118.

Pástor, L. (2000). Portfolio selection and asset pricing models. *Journal of Finance 55*(1), 179–223.

Pástor, L. and P. Veronesi (2006). Was there a nasdaq bubble in the late 1990s? *Journal of Financial Economics 81*(1), 61–100.

Pástor, L. and P. Veronesi (2009). Technological revolutions and stock prices. *American Economic Review 99*(4), 1451–83.

Petersen, M. A. (2009). Estimating standard errors in finance panel data sets: Comparing approaches. *Review of Financial Studies 22*(1), 435–480.

Puriya, A. and F. Bräuning (2021). Demand effects in the FX forward market: Micro evidence from banks' dollar hedging. *Review of Financial Studies 34*(9), 4177–4215.

Pya, N. and S. N. Wood (2015). Shape constrained additive models. *Statistics and Computing 25*(3), 543–559.

Rogers, L. and M. Tehranchi (2010). Can the implied volatility surface move by parallel shifts? *Finance and Stochastics 14*(2), 235–248.

Ross, S. A. (1976a). The arbitrage theory of capital asset pricing. *Journal of Economic Theory 13*(3), 341–360.

Ross, S. A. (1976b). Options and efficiency. *Quarterly Journal of Economics 90*(1), 75–89.

Rubinstein, M. (1994). Implied binomial trees. *Journal of Finance 49*(3), 771–818.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance 19*(3), 425–42.

Shorrocks, A. F. (2013). Decomposition procedures for distributional analysis: a unified framework based on the shapley value. *Journal of Economic Inequality 11*(1), 99.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris 8*, 229–231.

Stock, J. H. and M. Yogo (2005). Testing for weak instruments in linear iv regression. In D. W. K. Andrews and J. H. Stock (Eds.), *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, pp. 80–108. Cambridge University Press.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Annals of Statistics 10*(4), 1040–1053.

Taylor, M. P. (1987). Covered interest parity: a high-frequency, high-quality data study. *Economica 45*(216), 429–438.

Tchen, A. H. (1980). Inequalities for distributions with given marginals. *Annals of Probability 8*(4), 814–827.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological) 58*(1), 267–288.

Vayanos, D. (2004). Flight to quality, flight to liquidity, and the pricing of risk.

Vayanos, D. and J.-L. Vila (2021). A preferred-habitat model of the term structure of interest rates. *Econometrica 89*(1), 77–112.

Xu, C. (2020). Reshaping global trade: the immediate and long-run effects of bank failures. Technical report, Available at SSRN 3710455.

Zellner, A. (1986). On assessing prior distributions and bayesian regression analysis with g-prior distributions. In P. K. Goel and A. Zellner (Eds.), *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, Chapter 29, pp. 233–243. Amsterdam: North-Holland/Elsevier.