Khorrami, Paymon; Zentefis, Alexander K.

Working Paper

Arbitrage and Beliefs

CESifo Working Paper, No. 8490

Provided in Cooperation with:
Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

Suggested Citation: Khorrami, Paymon; Zentefis, Alexander K. (2020) : Arbitrage and Beliefs, CESifo Working Paper, No. 8490, Center for Economic Studies and Ifo Institute (CESifo), Munich

This Version is available at:
https://hdl.handle.net/10419/223562

CESifo WORKING PAPERS

8490
2020

August 2020

Arbitrage and Beliefs

Paymon Khorrami, Alexander K. Zentefis

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.cesifo.org/en/wp
An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.cesifo.org/en/wp

CESifo *Working Paper No. 8490*

Arbitrage and Beliefs

Abstract

We study a segmented-markets setting in which self-fulfilling volatility can arise. The only requirements are (i) asset price movements redistribute wealth across markets (e.g., equities rise as bonds fall) and (ii) some stabilizing force keeps valuation ratios stationary (e.g., cash flow growth rises when valuations rise). We prove that when self-fulfilling volatility exists, arbitrage opportunities must also exist. Conversely, at times when arbitrage profits exist, asset markets are susceptible to self-fulfilling fluctuations. The tight theoretical connection between price volatility and arbitrage is detectable in currency markets by studying deviations from covered interest parity.

JEL-Codes: D840, G110, G120.

Keywords: limits to arbitrage, segmented markets, volatility, self-fulfilling prices, multiple equilibria, covered interest parity.

Paymon Khorrami
Imperial College Business School
London / United Kingdom
p.khorrami@imperial.ac.uk

Alexander K. Zentefis
Yale School of Management
New Haven / CT / USA
alexander.zentefis@yale.edu

July 26, 2020
We are very grateful to Nick Barberis, Jess Benhabib, Harjoat Bhamra, Wenxin Du, Daniel Grosshans, Stefan Nagel, Jung Sakong, Andre Veiga, Pietro Veronesi, Raman Uppal, and especially Stavros Panageas and Dimitri Vayanos for extremely helpful comments. We would also like to thank conference participants at CESifo's Macro/Money/International conference and seminar participants at Imperial College and INSEAD for valuable feedback.

1 Introduction

Arbitrages exist. Empirical research has documented several examples of trades featuring positive profits with zero hold-to-maturity risk. Moving beyond the neoclassical frictionless model of financial markets, a theoretical literature emphasizing *limits to arbitrage* aims to rationalize such trades.[1] More broadly, limits to arbitrage help reconcile high asset returns and extreme price volatility under tame levels of fundamental risk, helping bridge an important gap in financial economics.

In this paper, we argue these models have not gone far enough. More specifically, we prove that a slight twist on a canonical limits-to-arbitrage setting can feature a type of self-fulfilling volatility which has nothing to do with the risk properties of underlying cash flows. In this sense, we can justify return volatilities even higher than the literature has predicted thusfar. While limits to arbitrage are certainly present to some degree, our reliance on equilibrium multiplicity in order to generate volatility immediately raises the question of testability: is this excess volatility really present in the data?

The question of testability is addressed by the second part of the paper. We establish a one-to-one theoretical mapping between the amount of this self-fulfilling volatility and the amount of arbitrage profits on the table. This mapping is a robust prediction, and it is what we later take to currency market data.

To elucidate our theoretical insights, we start with a very stylized setting with some assumed form of market segmentation (one can also think of "investor habitats"). For example, imagine some group of investors (call them A-types) only trades in market A, whereas a different group of investors (call them B-types) only trades in market B. Without some arbitrageur trading in both markets simultaneously (or not trading very actively across markets), what determines asset prices? Usually, we appeal to fundamental valuation: each of these local investors (A or B) knows their risk preferences, so they can look at the underlying cash flows, discount them to the present with a risk adjustment, and market prices should adjust to equal this quantity.

Digging into the details helps here. Fundamental valuation usually works because any other price is associated to a violation of the long-run transversality condition. An asset whose price is above fundamental value features a low dividend yield, and valuations must continuously rise without bound to satisfy investors; the opposite occurs if asset prices are below fundamental value. This type of instability is mechanically why transversality fails and why fundamental value prevails.

But fundamental valuation is complicated by various factors, including the endo-

[1] See related literature section for empirical examples and theoretical antecedents.

geneity of cash flows, the probability distribution to use in forecasting, and of course the discount rate to apply. We show that when cash flows (or beliefs) are endogenous to valuations in a particular way – namely, they are expected to grow faster when valuations are higher – asset prices may not be uniquely determined.

Instead, the endogeneity of perceived growth sets the stage for self-fulfilling expectations of future price changes. Consider an asset that has a high price-dividend ratio today, meaning its cash flows are assumed to also grow faster. Investors will tolerate future valuation declines, because the cash flow growth is enough to satisfy their required returns. Working in reverse, if investors expect future valuation declines, high prices can be justified today. Mechanically, unlike the instability discussed above, these examples involve a *stabilizing force* that brings high valuations back to normal levels (and vice versa), ensuring transversality holds.

Layering any such stabilizing force onto a limits-to-arbitrage model opens the door for self-fulfilling volatility. The basic intuition is that non-fundamental price fluctuations, as long as they disappear "far from steady state," will not push prices outside of the stable region.

There is one additional requirement on beliefs that can be thought of as *redistribution* across markets. Returning to our example with markets A and B, suppose the price of asset A declines for extrinsic (i.e., non-fundamental) reasons. In reality, this could be some random selling putting downward pressure on prices. Having less wealth after the shock, A-type investors will want to cut consumption. To do this, A-types save a bit of the cash flows from asset A in the bond market. By market clearing, B-types must be borrowing this amount, consuming more than the cash flows of asset B. This consumption plan is only optimal if B-types believe their wealth has increased, requiring assets A and B to experience *equal and opposite* extrinsic shocks. In this sense, market clearing dictates the beliefs of A- and B-types must be negatively correlated. This story also suggests that bond market integration is required for our mechanism. This sequence of logic is reflected in Figure 1 below.

The general conditions for self-fulfilling volatility are the essence of Theorem 1 of our paper (Propositions 1-2 consider specific examples). These multiplicity results hold with infinitely-lived agents in a bubble-free economy with dynamically-complete, but imperfectly-integrated, financial markets. It is not required that investors be irrational in forecasting growth, but as we show in an extension, such beliefs expand the set of circumstances that can accommodate multiplicity. Besides the fact that markets are not perfectly integrated, there are no other frictions or constraints – agents are always marginal in their local asset markets.

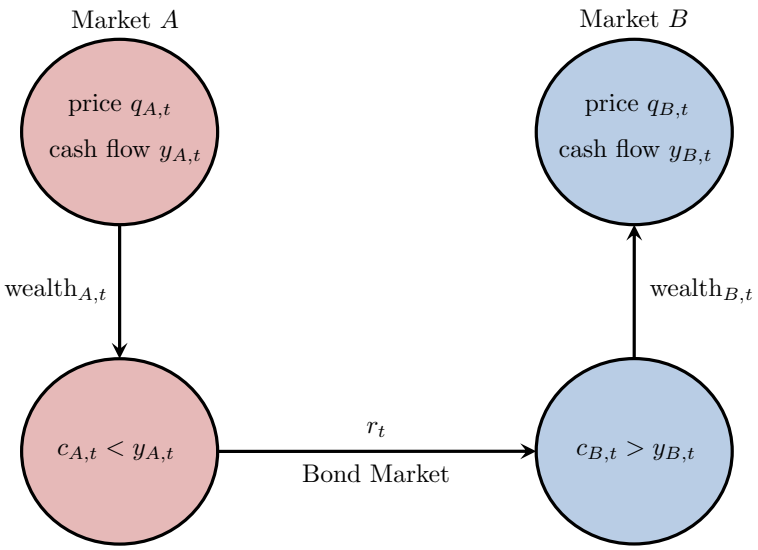

Figure 1: Mechanism of the model. Market A experiences a negative shock to its asset price $q_{A,t}$, without any effect on its cash flow $y_{A,t}$. Optimal consumption $c_{A,t}$ of agent A wants to fall below the cash flow to reflect lower wealth, which requires saving in the bond market at rate r_t. By bond market clearing, consumption $c_{B,t}$ of agent B must be higher than her local asset cash flow $y_{B,t}$. This is only optimal if market B asset prices $q_{B,t}$ rise in a manner that offsets the decline in $q_{A,t}$.

A robust implication that we develop in more detail is the tight connection between volatility and arbitrage (Theorem 2). Going back to Figure 1, if shocks to assets A and B are offsetting, one can construct a riskless portfolio containing both. This portfolio must generate *arbitrage profits*. The reason: A-types demand a risk premium on the extrinsic volatility of asset A, and similarly for B, so a portfolio that buys positive amounts of A and B will earn more than the riskless rate. In fact, a natural and correct conjecture is that the size of arbitrage profits is directly linked to the amount of self-fulfilling volatility (Proposition 3). This theoretical result is what we ultimately test empirically.

Normally, the presence of large arbitrage profits incentivizes trading by relative-value traders. Enter investment fund F. If fund F can trade freely across markets, we return to a neoclassical world without arbitrage and without self-fulfilling volatility. But if fund F encounters the type of frictions articulated by the limits-to-arbitrage literature (e.g., margin requirements, search frictions, myopic performance-based clients), some arbitrage profits will be left on the table. By extension, self-fulfilling volatility will not be fully eliminated. Quantitatively, the magnitude of limits-to-arbitrage frictions disciplines the magnitude of viable self-fulfilling volatility (Proposition 4 and Corollary 5).

Empirically, we examine the hypothesized link between arbitrage profits and volatility in the context of covered interest parity (CIP) deviations. The 3-month CIP deviation serves as our primary measure of available arbitrage profits. The arbitrage strategy bor-

rows in USD and goes long a synthetic US bond built with a foreign bond and currency swaps (or vice versa if the CIP deviation is negative).

We employ several different proxies for volatility. In one exercise, we compute the weighted-average return volatility on the two legs of the CIP (the US bond and the synthetic US bond). We also examine the longer-maturity counterparts of these legs (10-year bonds as opposed to 3-month bonds), as a way to make a duration-adjustment that brings our volatility magnitude closer to what the model calls for (the assets in the model are infinitely-lived). We find very strong positive associations between CIP deviations and each of our volatility proxies: (i) the weighted-average 3-month return volatilities for the CIP components; (ii) the 10-year US bond return volatility; (iii) the 10-year Treasury VIX; and (iv) the 10-year foreign bond return volatility. We obviously cannot determine if volatility is of a self-fulfilling nature or not, which makes our test imperfect, but the strength of the association is striking.

Related literature. Our paper is most closely related to the theoretical literature on limits to arbitrage. Many papers in this literature have focused on developing the implications of specific micro-foundations – such as margin requirements, myopic performanced-based clients, or search frictions.[2] These papers also typically pay detailed attention to the behavior of arbitrageurs. By contrast, we relegate the behavior of arbitrageurs to the background and take a much more reduced-form approach to the actual frictions involved. But as a benefit, we are able to analytically develop all pricing implications, including the new insight of multiplicity, more fully.

Empirically, there are a plethora of documented arbitrage trades: examples include spinoffs (Lamont and Thaler, 2003); "on-the-run / off-the-run" bonds (Krishnamurthy, 2002); covered interest parity (Du, Tepper and Verdelhan, 2018); Treasury spot and future repo rates (Fleckenstein and Longstaff, 2018).

These empirical examples are analyzed for clarity, in the sense that matching cash flows leaves only frictions to explain price differences. Although they may be more difficult to identify empirically, similar frictions may pervade other markets. For instance, Hu, Pan and Wang (2013) suggest that hedge fund capital modulates the closeness of

[2]See Shleifer and Vishny (1997) for a micro-foundation for mis-pricing due to "performance-based arbitrage." See Gromb and Vayanos (2002) for a margin-based analysis of price deviations in multiple identical markets. See Garleanu and Pedersen (2011) for dynamic asset pricing model with margin constraints. See Duffie and Strulovici (2012) for a model in which slow-moving capital arises due to search frictions. See Vayanos and Weill (2008) for a search friction application to the "on-the-run / off-the-run" bond phenomenon. See Duffie (2010) and Gromb and Vayanos (2010) for further reviews of the literature and existing mechanisms. See Biais et al. (2017) for a model where market segmentation arises endogenously due to incentive constraints and preference heterogeneity.

the yield curve to no-arbitrage models. The idea is that there is some market segmentation between Treasuries of different maturities, perhaps because investors have maturity-specific "habitats" in which they like to focus. In such a world, if arbitrageur capital is somewhat limited, prices may deviate from the no-arbitrage benchmark. This may even be true in riskier markets: Ma (2019) suggests that corporate bond and equity markets may be partially segmented, with corporate issuances and buybacks acting as a mechanism to profit from price differences.

Our paper also relates to the literature on self-fulfilling dynamics.[3] Typical self-fulfilling stochastic equilibria build "sunspot shocks" around a locally-stable steady state (essentially lotteries on the multiplicity of deterministic transition paths). Our equilibria share a similar flavor, as the "stabilizing forces" we identify render our deterministic steady state locally-stable. We differ from this literature in some of the assumptions we adopt – we require neither overlapping generations (with the resulting possibility of bubbles)[4] or aggregate increasing returns[5] to induce stability.

Focusing on asset prices, our paper is closer to Hugonnier (2012), Zentefis (2020), and Gârleanu and Panageas (2019). As in those models, our multiplicity arises when there are multiple traded assets and some limits to arbitrage between them. Multiple assets is crucial in the sense that shocks to one asset class must be offset by the others in order to keep aggregate wealth smooth. Where we depart from these three papers is our notion of limits to arbitrage, which drives a distinction in our results and interpretation.

Hugonnier (2012) generates multiplicity from the presence of an aggregate bubble, which arises in stockholder-bondholder economies, and can be sub-divided arbitrarily to redistribute wealth between asset classes.[6] Our markets are segmented cross-sectionally, so we do not require such a bubble to obtain indeterminacy.

Like us, the OLG economy of Gârleanu and Panageas (2019) also has cross-sectionally segmented markets, in the sense that unborns have a disproportionate claim to human capital but cannot trade before birth. Multiplicity arises through wealth redistribution across overlapping generations, as extrinsic stock market shocks are offset by human capital shocks. They interpret this as a volatile aggregate stock market, whereas our equilibrium is better interpreted as self-fulfilling volatility in relative-value trades (e.g.,

[3] Benhabib and Farmer (1999) reviews this class of models in macroeconomics. Farmer (2016) discusses the intellectual history and compares to newer models in which a continuum of steady states arises.

[4] See Azariadis (1981), Cass and Shell (1983), and Farmer and Woodford (1997) [originally published in 1984] for early models with two-period lifetimes.

[5] See Farmer and Benhabib (1994).

[6] That said, the necessity of an aggregate bubble in limited participation economies is fragile, in the sense that *any* amount of entry by non-participants, no matter how tiny, eliminates the bubble (Khorrami, 2018) and would thus eliminate this multi-sector indeterminacy.

basis trades). In fact, their economy features no arbitrage at all times, whereas one of our main contributions is tightly connecting arbitrage profits and volatility.[7]

Zentefis (2020) demonstrates that leverage constraints can generate interesting self-fulfilling price dynamics in multi-asset models. This paper is related in the following sense: our example with endogenous growth rates can be interpreted as the outcome of collateral constraints on investing firms. But we also provide other mechanisms through which multiplicity can arise. By showing that several mechanisms can provide the required "stabilizing force," and by studying a general N-asset economy, our goal is to embark on a more general analysis.

Finally, as we study N "locations" with their own fundamentals, one can naturally take a global perspective. A recent international finance literature assumes cross-sectional segmentation of local equity or sovereign debt markets, with some global intermediary participating in all of them (Gabaix and Maggiori, 2015; Itskhoki and Mukhin, 2017). Our model points out the possibility of self-fulfilling dynamics in such a setting.

The paper is organized as follows. Section 2 presents the model. Section 3 explores when and why the model could have self-fulfilling volatility. Section 4 establishes the central link between the presence of arbitrage opportunities and self-fulfilling volatility. Section 5 studies the key model prediction in the context of currency markets. Section 6 concludes. The Appendix contains the proofs and further analysis.

2 Model

Setup. The model is set in continuous time with $t \geq 0$. The aggregate endowment follows

$$dY_t = gY_t dt.$$

We begin with deterministic endowments for theoretical clarity, but we show in Appendix B.1 that our results extend to a setting with aggregate shocks.

The economy features N locations, which can stand for sectors, industries, countries, or distinct financial markets. The endowment of location n is given by $y_{n,t}$, where

$$dy_{n,t} = y_{n,t} g_{n,t} dt.$$

[7] Our construction of a stochastic equilibrium is also related to theirs, by essentially randomizing over a multiplicity of deterministic transition paths. Indeed, one of our examples of a "stabilizing force" generalizes a reduced-form version of their model. There are two other related papers. Farmer (2018) studies capital asset prices in an OLG economy, with nominal government debt (rather than human capital) as the second asset allowing redistribution. Bacchetta et al. (2012) also has an OLG economy, with the uncleared riskless bond market providing the second asset allowing redistribution (to unmodeled foreigners).

For now, the only restriction on the local growth rate $g_{n,t}$ is that $\sum_{n=1}^{N} y_{n,t} g_{n,t} = g Y_t$ for some exogenously-given constant g. We purposefully leave growth rates otherwise unspecified, because the manner in which they are endogenous to asset prices plays a key role in the type of equilibrium that prevails.

In terms of financial markets, each location offers a single positive-net-supply asset that is a claim to $y_{n,t}$. The equilibrium price of asset n is $q_{n,t} y_{n,t}$, where $q_{n,t}$ is the price-dividend ratio. Let us define the endowment share $\alpha_{n,t} := y_{n,t}/Y_t$ to save notation.

Each location has a different representative agent. This agent can invest only in his or her local asset market and a zero-net-supply short-term bond market that is open to everyone. The equilibrium risk-free rate in the integrated bond market is r_t.

All agents have rational expectations, infinite lives, logarithmic utility, and discount rate $\delta > 0$. Mathematically, their preferences are

$$\mathbb{E}_0\Big[\int_0^\infty e^{-\delta t} \log(c_{n,t}) dt\Big].$$

Clearing of the goods and bond markets is standard: $\sum_{n=1}^{N} c_{n,t} = Y_t$ and $\sum_{n=1}^{N} q_{n,t} y_{n,t} = Q_t Y_t$, where Q_t is the aggregate price-dividend ratio.

Extrinsic Shocks. With market clearing established, we next describe asset prices. In a deterministic economy, any stochastic price changes must inherently originate from agents' self-fulfilling beliefs. To allow for this volatility, we conjecture that the price-dividend ratio of each location's asset follows a stochastic process

$$dq_{n,t} = q_{n,t}\Big[\mu^q_{n,t} dt + \sigma^q_{n,t} d\tilde{Z}_{n,t}\Big], \tag{1}$$

where $\tilde{Z}_{n,t}$ is a one-dimensional Brownian motion. The economy has no intrinsic uncertainty. This shock is therefore *extrinsic*, and it is the source of self-fulfilling asset price volatility, if any exists. Let $\tilde{Z}_t := \big(\tilde{Z}_{n,t}\big)_{n=1}^{N}$ be a vector of all locations' extrinsic shocks.

Economically, the extrinsic $\tilde{Z}$ shocks arise from sources that we do not explicitly model. Investor sentiment or signals that coordinate beliefs might trigger the self-fulfilling fluctuations, in a manner similar to Benhabib et al. (2015). Heterogeneity in opinions between optimists and pessimists akin to Scheinkman and Xiong (2003) can be another source. Correlated institutional demand shocks as described in Koijen and Yogo (2019) can yet be another driver of the price changes.

We allow the extrinsic shocks in the economy to obey an arbitrary correlation structure. A convenient way to represent this structure uses an N-dimensional basis of uncorrelated Brownian motions $Z_t := (Z_{n,t})_{n=1}^{N}$ and an $N \times N$ matrix of constants M that

captures their relation. From these two components, we recast the vector of extrinsic shocks as

$$\tilde{Z}_t = MZ_t. \tag{2}$$

The matrix M is normalized so that $\text{diag}\,[MM'] = (1, \ldots, 1)'$, which preserves $\tilde{Z}_t$ as a collection of Brownian motions. Substituting equation (2) into equation (1) shows that the self-fulfilling shock to asset n at time t is $\sigma^q_{n,t} M_n dZ_t$, where M_n is the n-th row of M.[8]

The matrix M is a crucial parameter of the model. To illustrate its structure, we consider the following examples, which we repeatedly use throughout the text.

Example 1 (Uncorrelated shocks)**.** Suppose M is the identity matrix. This structure implies $\tilde{Z}_t = Z_t$, which renders all extrinsic shocks uncorrelated.

Example 2 (Two-by-two redistribution)**.** Suppose $N = 2$ and let

$$M = \begin{bmatrix} 1 & 0 \\ -1 & 0 \end{bmatrix}.$$

This example presents a setting with two locations and one source of extrinsic uncertainty. The matrix M puts $\tilde{Z}_{1,t} = -\tilde{Z}_{2,t}$, which implies that the self-fulfilling price changes redistribute wealth between the two assets. As one price falls, the other rises.

Example 3 (General redistribution)**.** This example is the N-dimensional counterpart to Example 2. Let $\tilde{M}$ be an $N \times N$ non-singular matrix. Suppose

$$M = \tilde{M} - \frac{1}{N}\mathbf{1}'\tilde{M} \otimes \mathbf{1}.$$

In this structure, each element of the matrix $\tilde{M}$ is reduced by the simple average of its columns. This operation makes the column sums of M equal zero. The key consequence of this design is that $\mathbf{1}'d\tilde{Z}_t = \mathbf{1}'MdZ_t = 0$ almost-surely. Any other linear combination of $d\tilde{Z}_t$ does not equal 0. As a result, $\text{rank}(M) = N - 1$.

[8] Although markets are incomplete in the model, they are dynamically complete. The vector $\tilde{Z}_{n,t} = M_n Z_t$ is generated by N distinct shocks, but it suffices for agent n to only trade $\tilde{Z}_{n,t}$, which is the shock his local asset loads on. Indeed, if we introduce in each market zero-net-supply Arrow securities spanning Z that are traded only in market n, the equilibrium remains unchanged.

3 Self-fulfilling volatility

At first glance, readers may be divided about whether non-fundamental volatility is possible in such a model. On one hand, no "arbitrageur" exists to connect market dynamics across locations, so what disciplines local market prices? On the other hand, the presence of unconstrained fundamental investors having the same preferences and deterministic endowments suggests prices should be assigned a common fundamental value. Here, we shed light on this issue, hopefully clarifying when non-fundamental price dynamics exist and when they do not.

To develop some understanding of necessary conditions for multiplicity, consider the market clearing conditions. Individuals with log utility consume δ fraction of their wealth, so the aggregate wealth-consumption (price-dividend) ratio is $Q_t = \delta^{-1}$. Bond market clearing can then be written as

$$\sum_{n=1}^{N} \alpha_{n,t} q_{n,t} = \delta^{-1}. \tag{3}$$

Because the aggregate wealth-consumption ratio is constant, if there are any extrinsic shocks to $q_{n,t}$, they must be offset by extrinsic shocks to other assets. In this sense, extrinsic shocks (if they affect anything) must be *redistributive* across markets.

This redistribution stems directly from rank(M). By time-differentiating equation (3), we see that the loadings on each of the basis dZ_t shocks must be zero:

$$\sum_{n=1}^{N} \alpha_{n,t} q_{n,t} \sigma^q_{n,t} M_n = 0. \tag{4}$$

Write equation (4) as a matrix equation

$$M' v_t = 0, \tag{5}$$

where $v_t = (\alpha_{1,t} q_{1,t} \sigma^q_{1,t}, \ldots, \alpha_{N,t} q_{N,t} \sigma^q_{N,t})'$ is the column vector of volatilities. If M were full rank, the unique solution to (5) would be $v_t \equiv 0$. However, the singular situation rank$(M) < N$ implies a non-zero time-invariant solution $v_t \equiv v^* \neq 0$ exists. If so, then $\psi_t v^*$ also solves (5) for any scalar process ψ_t. Hence, as long as M is singular (Examples 2-3 but not Example 1), we have a continuum of candidate equilibria.

When are these "candidate equilibria" actual equilibria? It turns out that the only additional requirement is that asset prices be bounded, which ensures the transversality condition holds. Below, we will show how the requirement of boundedness translates

into constraints on model primitives. For now, we summarize this discussion with the following theorem, which collects general necessary conditions that tell us when the model could have self-fulfilling volatility. The proof is in Appendix A.

Theorem 1. *If rank$(M) = N$ (full rank), then equilibrium cannot have self-fulfilling volatility: $(\sigma^q_{1,t}, \ldots, \sigma^q_{N,t}) \equiv 0$ for all t. Conversely, suppose rank$(M) < N$ and let v^* be in the null-space of M'. Given a process $\{\psi_t\}_{t\geq 0}$, an equilibrium can be sustained with volatility $\alpha_{n,t} q_{n,t} \sigma^q_{n,t} = v^*_n \psi_t$ for all n, as long as the resulting $\{(q_{n,t})_{n=1}^N\}_{t\geq 0}$ is a bounded, positive process.*

The remainder of this section sheds light on which model mechanisms help preserve the asset price stationarity required by Theorem 1 and which do not.

Determinacy and instability. The following is a benchmark case in which equilibrium is unique and non-stochastic.

Proposition 1. *Assume constant local growth rates $g_{n,t} = g$. Then, equilibrium cannot have self-fulfilling volatility: $(\sigma^q_{1,t}, \ldots, \sigma^q_{N,t}) \equiv 0$ for all t. All assets have identical constant price-dividend ratios $q_{n,t} = \delta^{-1}$.*

Even though there is no arbitrageur connecting locations, Proposition 1 shows that the presence of our fundamental traders is enough to pin down asset prices. This equilibrium with $q_{n,t} = \delta^{-1}$ always exists, but here it is also unique.

The reason for this strong determinacy is the instability of price-dividend dynamics. To see this, consider the deterministic model with $(\sigma^q_{1,t}, \ldots, \sigma^q_{N,t}) \equiv 0$. In this case, there is no risk compensation, and all assets must earn the riskless rate, i.e.,

$$\underbrace{\dot{q}_{n,t}/q_{n,t} + g}_{\text{capital gain}} + \underbrace{1/q_{n,t}}_{\frac{\text{dividend}}{\text{price}}} = r_t. \tag{6}$$

Furthermore, as the equilibrium involves deterministic individual consumption paths, the interest rate is solely determined by time-discounting and economic growth, i.e., $r_t = \delta + g$. Substituting this into (6), we have $\dot{q}_{n,t} = -1 + \delta q_{n,t}$, a dynamical system that has a single steady state which is unstable. If the price-dividend ratio is below (above) δ^{-1}, then it drifts downwards (upwards) at a pace which accelerates over time, so the required boundedness of Theorem 1 is violated. Adding stochastic shocks does nothing to remedy the core non-stationarity.

Multiplicity and stability. For any self-fulfilling volatility to exist, there must be a *stabilizing force* present that keeps asset prices stationary. Next, we provide an example of what such a stabilizing force might entail.

Proposition 2. *Assume local growth rates satisfy* $g_{n,t} = g + \lambda(q_{n,t} - \delta^{-1})$ *with* $\lambda > \delta^2$. *Then, self-fulfilling volatility is possible: there exists a non-zero process* $\{\psi_t\}_{t\geq 0}$ *such that an equilibrium exists with* $\alpha_{n,t} q_{n,t} \sigma^q_{n,t} = v^*_n \psi_t$ *for all* n, *where* v^* *is in the null-space of* M'. *In terms of the cross-sectional minimums* $\underline{\alpha}_t := \min_n \alpha_{n,t}$, $\underline{x}_t := \min_n x_{n,t}$, *and* $\underline{q}_t := \min_n q_{n,t}$, *the volatility* ψ_t *can be any bounded process that satisfies*

(P1) $\psi_t / \underline{\alpha}_t$ *and* $\psi_t / \underline{x}_t$ *are bounded;*

(P2) ψ_t *vanishes as* $\underline{q}_t$ *approaches* $\delta(\epsilon + \lambda^{-1})$ *from above, for some* $0 < \epsilon < \delta^{-2} - \lambda^{-1}$.

The key condition of Proposition 2 is that local growth is endogenous and increases (sufficiently quickly) with local asset prices. While we do not provide a full micro-foundation for why this might happen, it seems intuitively plausible. For instance, if productive assets are used as collateral, higher valuations fuel greater borrowing and investment (e.g., models like Kiyotaki and Moore (1997)). The exact mathematical condition is relatively modest: for a standard discount rate $\delta = 0.01$, local growth rates must be at least 0.1% higher than average when local valuations are 10% above average.

The dependence of growth on valuations allows for self-fulfilling expectations of future price changes. For example, if an asset has a high price-dividend ratio today, its cash flows also grow faster, which means investors will tolerate future valuation declines. Working in reverse, investor beliefs about future valuation declines can justify high valuations today. A symmetric intuition applies to low-priced assets that must be expected to appreciate.[9]

Mathematically, the endogeneity of growth rates acts as a stabilizing force. To see this clearly, consider again the deterministic model but with these endogenous growth rates. Pricing equation (6) is replaced by (after substituting $g_{n,t}$ for g and using again the fact that $r_t = \delta + g$) the Riccatti equation

$$\dot{q}_{n,t} = -1 + \delta(1 + \lambda/\delta^2) q_{n,t} - \lambda q^2_{n,t}. \tag{7}$$

Supposing $\lambda > \delta^2$, the larger of the two steady-states is the relevant one (with $q_n = \delta^{-1}$) and it is locally stable, in the sense that $\frac{\partial \dot{q}_n}{\partial q_n}\big|_{q_n = \delta^{-1}} = \delta(1 - \lambda/\delta^2) < 0$.[10]

[9] Of course, since $g_{n,t}$ depends on $q_{n,t}$, the growth rate itself follows a (stochastic) process. Had we specified this growth process exogenously, equilibrium would induce a unique valuation ratio $q_{n,t}$, from the present-value formula. In that sense, for each resulting process $g_{n,t}$, there is still a unique valuation ratio $q_{n,t}$ paired to it. The interesting departure here is that multiple pairs (g_n, q_n) solve (i) the growth-valuation link assumed in Proposition 2, coupled with (ii) the present-value formula for asset prices.

[10] In contrast to one-dimensional diffusions, it is often very difficult technically to establish stability of multi-variate SDEs, even moreso if the dynamics are non-Markovian. See Chapters 3.5, 3.7, and 4.4 of

When an economy has such a stabilizing force, some amount of self-fulfilling volatility ψ_t becomes possible. More surprisingly, volatility is essentially arbitrary, with the only restrictions being that it vanish when the economy is "far from steady state" so that the stabilizing force kicks in unabated. This vanishing property is the crux of properties (P1) and (P2). The result provides a strong sense in which lack of cross-market arbitrage can allow prices of similar assets to move distinctly. Later, by introducing some partial cross-market trading, we will place more restrictions on the volatility ψ_t.

How special is the stabilizing force of Proposition 2, and could other settings produce similar multiplicity? In Appendix B.2, we show that similar stability is provided by a type of optimism about local growth, similar in spirit to the extrapolation of Barberis et al. (2015). If agents become more optimistic about growth when asset prices rise, and vice versa, then the equilibrium behaves very similarly to that in Proposition 2. This is true even if all growth rates are $g_{n,t} = g$ in reality. Hence, the stabilizing force provided by endogenous growth rates can be real or merely perceived.

To highlight the wide range of possible of stabilizing forces, we also analyze a substantially different economy in Appendix B.3, using a model with overlapping generations and creative destruction along the lines of Gârleanu and Panageas (2019). As long as incumbent producers (alive cohorts) are sufficiently better able to ward off new producers (newborn cohorts) when asset prices are higher, the model has the right stability properties. It is likely that many other plausible examples of stabilizing forces are out there, and we do not attempt to find them all.

These stabilizing forces and their implications are reminiscent of the extant multiplicity and sunspots literature that builds on seminal papers Azariadis (1981); Cass and Shell (1983); Farmer and Woodford (1997). There, self-fulfilling stochastic equilibria are constructed using "sunspot shocks" around a locally-stable deterministic steady state, which is possible because of the multiplicity of potential transition paths to that steady state. Our construction is similar in spirit, though more general in the type of stochastic processes allowed.

The role of the bond market. Any self-fulfilling equilibria of our model, even going beyond the specific setting of Proposition 2, crucially require the bond market. Without the bond market, agent n only consumes the cash flows from his local asset, $y_{n,t}$. Since

Khasminskii (2011) for some results that apply to multi-variate diffusions (recurrence and stationarity theorems). We are unable to apply these results to our problem because of the properties of $\alpha_{n,t} := y_{n,t}/Y_t$ which has dynamics $d\alpha_{n,t} = \alpha_{n,t}[g_{n,t} - g]dt$ (in particular, it is a transient process with a degenerate diffusion matrix). We are able to sidestep this difficulty here and take a direct approach, because our system has very convenient analytical properties, like the fact that the deterministic version of q_n dynamics are decoupled across locations. See the proof of Proposition 2.

this consumption is deterministic, no asset-price volatility can be justified! A similar result holds even if the local cash flows possess arbitrary fundamental shocks independent of the extrinsic uncertainty, because investors' consumption would be uncorrelated with extrinsic shocks. By contrast, if the bond market is open, agent n can send and receive consumption across locations, with the promise of inter-temporal payback. This opens the door for stochastic individual consumption profiles ($dc_{n,t}$ can load on $d\tilde{Z}_{n,t}$), which then creates a stochastic local pricing kernel (marginal utility loads on $d\tilde{Z}_{n,t}$), and finally justifies price volatility ($dq_{n,t}$ loads on $d\tilde{Z}_{n,t}$).

To see clearly the link between volatility and the pricing kernel, note that any self-fulfilling volatility must be compensated. Agent n holds exposure to the extrinsic shock $\tilde{Z}_{n,t}$ through his exposure to $q_{n,t}$. If we define $\tilde{\pi}_{n,t}$ as the risk price (or Sharpe ratio) associated to this shock, and define the consumption shares $x_{n,t} := c_{n,t}/Y_t$, then

$$\tilde{\pi}_{n,t} = \delta\Big(\frac{\alpha_{n,t}q_{n,t}}{x_{n,t}}\Big)\sigma^q_{n,t}. \tag{8}$$

Intuitively, $y_{n,t}q_{n,t}\sigma^q_{n,t}$ is the total exposure to $\tilde{Z}_{n,t}$ shocks, and $\delta^{-1}x_{n,t}Y_t$ is the wealth of agent n, who bears these shocks. With log utility, the required compensation for Brownian shocks is the per-unit-of-wealth exposure, so dividing these two reveals the risk price. Formula (8) also shows that Sharpe ratios are linked to self-fulfilling volatility: $\sigma^q_{n,t} > 0$ if and only if $\tilde{\pi}_{n,t} > 0$ as well.

Remark 1 ($N = 1$ representative-agent economy)**.** The discussion above also shows that our model with $N = 1$ cannot have self-fulfilling volatility, even if aggregate growth rates are endogenous and linked to valuations as in Proposition 2 (e.g., aggregate growth $g_t = G(Q_t)$ for some increasing function G). The reasoning is that aggregate endowment growth is deterministic over small time-intervals dt, so the representative agent demands no risk premium, and by the risk price formula (8), there is zero asset price volatility.

Remark 2 (Small open economy)**.** Although the multiplicity result requires an open and active bond market, it does not require bond market clearing. Consider a "small open economy" in which the market for claims to $\{y_{n,t}\}_{t\geq 0}$ clears, but the bond market does not. All the results on multiplicity go through. Intuitively, even without bond market clearing, investors can still use the bond to obtain stochastic consumption. Mathematically, the equilibrium interest rate without extrinsic shocks is constant at $r_t = \delta + g$, so it plays no role in stabilizing the steady state of Proposition 2. Given any exogenous constant rate $r \leq \delta + g$, and endogenous local growth rates $g_{n,t} = g + \lambda(q_{n,t} - \delta^{-1})$, the

counterpart to valuation dynamics of equation (7) is

$$\dot{q}_{n,t} = -1 + (r - g + \lambda\delta^{-1})q_{n,t} - \lambda q_{n,t}^2,$$

which has a stable steady state (the larger of the two) if and only if $\lambda > \delta^2(1 + \sqrt{\frac{\delta+g-r}{\delta}})$.

4 Arbitrage Profits

Having seen that self-fulfilling volatility is possible, we next connect it to arbitrage profits. Section 4.1 shows that the presence of this volatility and the existence of arbitrage opportunities are two sides of the same coin. Section 4.2 demonstrates that limits to arbitrage quantitatively discipline the amount of self-fulfilling price changes.

4.1 Volatility implies arbitrages and vice versa

Theorem 1 requires the condition $\text{rank}(M) < N$ in order to have self-fulfilling volatility. To get a sense of what this rank condition means, consider what would happen if a single trader was allowed to participate in all markets. With $\text{rank}(M) < N$, there is some asset that this trader can replicate using the other $N - 1$ assets. But with self-fulfilling volatility, the price of this asset and its replicating portfolio need not move together. In short, this trader would be faced with an *arbitrage opportunity*. In our model, self-fulfilling volatility emerges if and only if arbitrages exist, which provides a more intuitive diagnostic for multiplicity than the rank condition on M. This link goes beyond the specific "stabilizing forces" raised in Section 3 and applies to any conceivable example with self-fulfilling volatility.

Theorem 2. *Self-fulfilling volatility implies an arbitrage. Conversely, if there are arbitrages, then equilibrium must feature self-fulfilling volatility.*

First, consider the second statement of Theorem 2, that arbitrages imply self-fulfilling volatility. If there were no volatility, then all assets earn the riskless rate, so there is no way to combine them into a portfolio that outperforms the riskless rate. This no-arbitrage, no-volatility equilibrium is the only one that can emerge when there is no stabilizing force in the economy.

Conversely, arbitrages exist when self-fulfilling volatility emerges. In the proof in Appendix A, we examine the portfolio that puts $\delta\alpha_{n,t}q_{n,t}$ in each asset $n = 1, \ldots, N$. By equation (4), the portfolio volatility is identically zero. This is where $\text{rank}(M) < N$ is

critical: even if all assets have positive self-fulfilling volatility, we can manufacture a riskless asset from them. The proof shows mathematically why this portfolio pays more than the riskless rate, but the basic intuition comes from the fact that each local market n demands a risk premium on its local asset. A portfolio built as a convex combination of components with risk premia must have a premium itself.

Confirming this intuition, this strategy's excess return over the riskless rate r_t is

$$A_t := \sum_{n=1}^{N} x_{n,t} \tilde{\pi}_{n,t}^2 > 0, \tag{9}$$

where recall $\tilde{\pi}_n$ is the local Sharpe ratio. This can be thought of as a measure of *arbitrage profit* in this model. Note that A_t is already quoted in the standard units used in analysis of arbitrage trades, because the long position $(\delta \alpha_n q_{n,t})_{1 \leq n \leq N}$ is a return (unit-cost portfolio). In addition, the amount of arbitrage profit A_t is exactly the difference between the risk-free rate that prevails without self-fulfilling volatility ($r_t = \delta + g$) and the one with self-fulfilling volatility ($r_t = \delta + g - A_t$). One usually reads this term as a precautionary savings term, but based on this discussion, A_t can also be thought of as the difference between a synthetic bond return and the traded bond return.

The interpretation of the traded and synthetic bond depends on the context. Some common examples are collateralized versus uncollateralized lending (sometimes captured in the TED spread); on-the-run versus off-the-run Treasury bonds; and deviations from covered interest parity (CIP). Measures of this quantity tend to be minimal for much of the time, but can expand to around 3% during financial crisis periods (Fleckenstein and Longstaff, 2018; Du et al., 2018).

The degree of arbitrage profit informs the amount of self-fulfilling volatility. This is because volatility drives location-specific risk prices, which constitute A_t.

Proposition 3. *Let rank$(M) < N$. Then, there exists a non-zero vector v^*, in the null-space of M', such that the self-fulfilling volatility ψ_t of Theorem 1 satisfies*

$$\psi_t = \frac{\delta^{-1}\sqrt{A_t}}{\sqrt{\sum_{n=1}^{N} x_{n,t} (\frac{v_n^*}{x_{n,t}})^2}} \leq \frac{\delta^{-1}\sqrt{A_t}}{\mathbf{1}' v^*}, \tag{10}$$

where A_t is the arbitrage profit given in (9). Consequently, the "average return volatility" $\sigma_t^* := \sum_{n=1}^{N} \frac{\alpha_{n,t} q_{n,t}}{\sum_{i=1}^{N} \alpha_{i,t} q_{i,t}} \sigma_{n,t}^q$ satisfies

$$\sigma_t^* = \delta \psi_t \mathbf{1}' v^* \leq \sqrt{A_t}. \tag{11}$$

The average return volatility σ_t^* defined in Proposition 3 is a scale-free summary statistic for the degree of volatility in our model. The tight link to arbitrage profits, $\sigma_t^* \leq \sqrt{A_t}$, is a bonus. To get a sense of magnitudes, consider arbitrage profits that range from $A_t \in [0, 0.03]$, consistent with the Treasury evidence of Fleckenstein and Longstaff (2018) and the CIP deviations documented in Du et al. (2018). Then, average return volatilities can range from $\sigma_t^* \in [0, 17.3\%]$, a quantitatively-large estimate.

At this point, it should be clear that arbitrages and self-fulfilling volatility are intrinsically linked. A different type of trade that resembles an arbitrage, a so-called *basis trade*, also exists in our model. A basis trade is a long-short strategy designed to capitalize on violations of the law of one price, i.e., price discrepancies between two assets with identical cash flows.

Remark 3 (CIP deviations)**.** Our model can accommodate CIP deviations that we will study empirically. In the real world, one might consider the following basis trade: borrow in Japanese yen, exchange to US dollars in the FX spot market, lend in dollars, and finally convert back into yen via a pre-signed currency futures contract. This trade is a theoretically riskless method to move Japanese yen from today to tomorrow, and it should return the yen risk-free rate. When it does not, CIP fails.

In our model, think of locations as countries, and consider a hypothetical trader that can access all countries' markets. Countries' consumption goods are homogeneous and freely tradable, so the spot and forward exchange rates are always unity.

Next, if one were to construct a "local discount bond" that pays off in the future consumption of the local country, this bond would have price

$$\begin{aligned} b_{n,t\to T} &= e^{-\delta(T-t)}\mathbb{E}_t\left[\frac{c_{n,t}}{c_{n,T}}\right] \\ &= e^{-(\delta+g)(T-t)}\mathbb{E}_t\left[\exp\left(-\int_t^T(-A_u+\frac{1}{2}\tilde{\pi}_{n,u}^2)du-\int_t^T\tilde{\pi}_{n,u}d\tilde{Z}_{n,u}\right)\right] \\ &= e^{-(\delta+g)(T-t)}\tilde{\mathbb{E}}_t^n\left[\exp\left(\int_t^T A_u du\right)\right], \end{aligned}$$

where A_t is given in (9) and $\tilde{\mathbb{E}}^n$ is the local risk-neutral expectation induced by risk prices $\tilde{\pi}_{n,t}$. This expression is an expected discounted sum of arbitrage profits, where discounting is performed by the riskless rate that would prevail without arbitrages. With unitary exchange rates, the CIP deviation (in continuously-compounded units) is given

by the difference in these yields on these discount bonds, i.e.,

$$\Delta_{t\to T}^{\text{CIP}(i,j)} := -\frac{1}{T-t}\Big(\log b_{j,t\to T} - \log b_{i,t\to T}\Big) = \frac{1}{T-t}\log\frac{\tilde{\mathbb{E}}_t^i\big[\exp\big(\int_t^T A_u du\big)\big]}{\tilde{\mathbb{E}}_t^j\big[\exp\big(\int_t^T A_u du\big)\big]}.$$

Note that $\Delta_{t\to T}^{\text{CIP}(i,j)} \neq 0$ if and only if the self-fulfilling equilibrium obtains. Indeed, the risk-neutral measures $\tilde{\mathbb{E}}^i$ and $\tilde{\mathbb{E}}^j$ are different in any equilibrium with self-fulfilling volatility, which is when $A_t > 0$. Conversely, when $A_t = 0$ forever, clearly $\Delta_{t\to T}^{\text{CIP}(i,j)} = 0$.

4.2 Cross-market trading limits volatility

Proposition 3 provided a link between volatility and a measure of arbitrage profits. Given this connection, impediments to capital mobility and cross-market trading, which work to curb arbitrage profits, should curb asset volatility. In this section, we make this argument precise by developing a notion of *limits to arbitrage* and showing how it bounds the degree of volatility.

Motivated by models like Gromb and Vayanos (2002) and Garleanu and Pedersen (2011), we assume that cross-sectional risk prices are linked by some amount of relative-value trading going on in the background. To formalize this notion, we need to examine the location-specific risk prices induced on the basis shocks Z_t. Recall equation (2) connecting $\tilde{Z}_t = MZ_t$. If $\tilde{\pi}_{n,t}$ is the location-n marginal utility response to $d\tilde{Z}_{n,t}$, then

$$\pi_{n,t} := \tilde{\pi}_{n,t} M_n. \tag{12}$$

is the marginal utility response to dZ_t, where M_n is the nth row of M. Note that $\tilde{\pi}_{n,t}$ is a scalar, while $\pi_{n,t}$ is a vector.

We make the following reduced-form assumption about these basis risk prices $\pi_{n,t}$:

$$\|\pi_{j,t} - \pi_{i,t}\| \leq \Pi_t \quad \forall i \neq j. \tag{13}$$

When $\Pi_t > 0$, we say there are *limits to arbitrage*. This terminology is justified by the well-known equivalence between absence of arbitrage and the existence of a stochastic discount factor that prices all assets (hence a single risk price vector across all markets, $\pi_t^* = \pi_{i,t} = \pi_{j,t}$).

In microfounded models, the process for Π_t would be linked to fundamental objects like arbitrageur wealth, preferences, constraints, and trading costs. For example, one can think of Π_t arising due to margin constraints and the limited wealth that arbi-

trageurs can thus deploy in eliminating risk-price differentials. It is only worth trading if risk-price differentials become sufficiently large. Bounds like (13) pervade most models of limits to arbitrage.[11] Here, we take Π_t as given and do not model the behavior of these arbitrageurs, opting instead to characterize equilibrium conditional on partial arbitrage.[12]

Up to now, we have been implicitly assuming $\Pi_t = +\infty$, which is tantamount to infinitely-frictional arbitrage behavior. What happens when there is some partial amount of market segmentation? We have the following link between the degree of market segmentation and the degree of self-fulfilling volatility. The proof is in Appendix A.

Proposition 4. *Let $0 < \Pi_t < +\infty$ and $rank(M) < N$. Then, there exists a non-zero vector v^*, in the null-space of M', such that the self-fulfilling volatility ψ_t of Theorem 1 is bounded by*

$$\psi_t \leq \delta^{-1} L_t^{-1} \Pi_t. \tag{14}$$

where $L_t := \max_{(i,j):i\neq j} \left\| x_{i,t}^{-1} v_i^* M_i - x_{j,t}^{-1} v_j^* M_j \right\|$. The "average return volatility" $\sigma_t^* = \delta \psi_t \mathbf{1}' v^*$ is bounded by

$$\sigma_t^* \leq \mathbf{1}' v^* L_t^{-1} \Pi_t. \tag{15}$$

Intuitively, with large limits-to-arbitrage, there can be large amounts of self-fulfilling volatility, because capital is too slow to correct any such price movements. As limits to arbitrage are relaxed, the amount of self-fulfilling volatility must vanish. Propositions 3 and 4 are thus similar in that they connect volatility to some quantitative measure of arbitrage efficacy (arbitrage profits and limits to arbitrage, respectively). Because of the link between volatility ψ_t and arbitrage profits A_t, limits-to-arbitrage as assumed in (13) also puts clear and intuitive bounds on A_t. This also bounds equilibrium risk prices, like Hansen and Jagannathan (1991), even though our primitive limits-to-arbitrage assumption in (13) is about relative risk prices.

[11] For instance, Proposition 2′ in Appendix B of Garleanu and Pedersen (2011) explicitly shows how margin constraints lead to a range of viable risk premia.

[12] We also do not modify any of the market clearing conditions to account for arbitrageur consumption, which can be justified by the idea that infinite trading would occur if $\|\pi_{j,t} - \pi_{i,t}\| > \Pi_t$ ever occurred, but zero trading is needed otherwise.

Corollary 5. *Under the conditions of Proposition 4, risk prices and arbitrage profits are bounded:*

$$\begin{aligned}\|\pi_{n,t}\| &\leq \frac{v_n^*}{x_{n,t}} L_t^{-1} \Pi_t \\ \sqrt{A_t} &\leq \Big(\sum_{n=1}^{N} x_{n,t} (\frac{v_n^*}{x_{n,t}})^2 \Big)^{1/2} L_t^{-1} \Pi_t,\end{aligned}$$

where $L_t := \max_{(i,j):i\neq j} \left\| x_{i,t}^{-1} v_i^* M_i - x_{j,t}^{-1} v_j^* M_j \right\|$.

To get a quantitative sense of the volatility bounds, we calibrate and simulate our model for an economy stability-inducing endogenous growth rates as in Proposition 2, i.e.,

$$g_{n,t} = g + \lambda(q_{n,t} - \delta^{-1}) \tag{16}$$

To satisfy the stability requirement $\lambda > \delta^2$, we set $\lambda = \delta^2 + 0.01$. We study $N = 10$ locations and set the extrinsic shocks in a similar way as Example 3, which has a zero-sum condition that we referred to as "redistribution":

$$\begin{aligned} M &= \frac{N}{\sqrt{N(N-1)}} \Big[I_N - \frac{1}{N} \mathbf{1} \otimes \mathbf{1}' \Big] \\ &= \frac{1}{\sqrt{N(N-1)}} \begin{bmatrix} N-1 & -1 & -1 & \cdots & -1 & -1 \\ -1 & N-1 & -1 & \cdots & -1 & -1 \\ \vdots & & \ddots & & & \vdots \\ \vdots & & & \ddots & & \vdots \\ -1 & -1 & -1 & \cdots & N-1 & -1 \\ -1 & -1 & -1 & \cdots & -1 & N-1 \end{bmatrix}. \end{aligned} \tag{17}$$

Note that the columns of M sum to zero and have unit norm. It can easily be verified that $v^* = \mathbf{1}$ is the unique element, up to scale, in the null-space of M. To keep things simple, we assume initially equally-sized locations ($\alpha_{n,0} = 1/N$) and initialize the simulation with equally-wealthy locations ($x_{n,0} = 1/N$ for all n). We also set $\delta = g = 0.02$.

In terms of the exogenous arbitrage bounds, we set $\Pi_t = 0.25$ to time-invariant values. The interpretation is that arbitrageurs are only willing to enter and correct Sharpe ratio differentials greater than 0.25. As will be clear shortly, these limits to arbitrage are quantitatively reasonable.

To simulate $\{x_{n,t} : t \geq 0\}$, first note that the analytical dynamics are given by

$$dx_{n,t} = x_{n,t}(1 - x_{n,t})\left[\tilde{\pi}_{n,t}^2 - \sum_{i \neq n} \frac{x_{i,t}}{1 - x_{n,t}} \tilde{\pi}_{i,t}^2\right] dt + x_{n,t}\tilde{\pi}_{n,t} d\tilde{Z}_{n,t}. \tag{18}$$

This is derived by applying Itô's formula to the definition $x_{n,t} := c_{n,t}/Y_t$, where the dynamics of $c_{n,t}$ are given in (27) in the appendix. Because $\tilde{\pi}_{n,t}$ depends on the self-fulfilling volatility, we will assume ψ_t is always at its upper bound, subject to vanishing when needed.[13]

The results for average return volatility σ_t^*, and the associated arbitrage profits A_t, are displayed in Figure 2. The average return volatility σ_t^* fluctuates around 16% (it could be much lower if we assume the economy is not at the upper bound of the volatility bounds). The claim that Π is reasonable can be seen by examining the associated simulated arbitrage profits A_t, which are around 2.5%, the upper range of measured arbitrage profits.

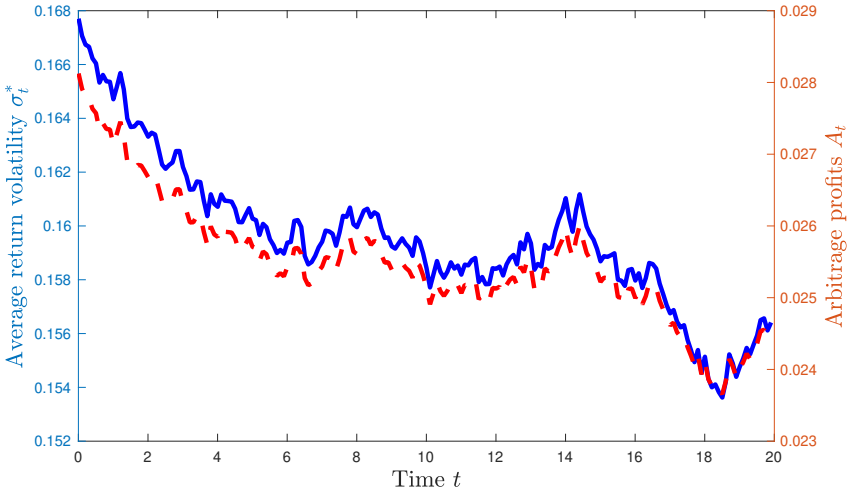

Figure 2: Plotted in solid blue against the left axis are volatility bounds of Proposition 4, from a simulated economy with $N = 10$ equally-sized locations ($\alpha_n = 1/N$) starting with equal initial wealth ($x_{n,0} = 1/N$), with extrinsic shock matrix M given in (17), and with endogenous growth rates $g_{n,t}$ from (16). Plotted in dashed red against the right axes are arbitrage profits A_t from the simulation. The simulation assumes ψ_t is always at the upper bound, except when it needs to vanish, i.e., when $\min_n \alpha_{n,t}$, $\min x_{n,t}$, or $\min_n q_{n,t} - \delta\lambda^{-1}$ become close enough to zero (see Proposition 2). We ensure this by capping the ratio of ψ_t to each of these quantities by 100 in the simulation. Other parameters are described in the text.

[13] In particular, Proposition 2 shows that ψ_t needs to vanish when $\underline{\alpha}_t$, $\underline{x}_t$, or $\underline{q}_t - \delta\lambda^{-1}$ become low enough. We ensure this by capping the ratio of ψ_t to each of these quantities by 100 in the simulation. In our simulation of $T = 20$ years, none of these vanishing conditions are ever binding, but they would in a long enough simulation. In particular, since $d\alpha_{n,t} = \alpha_{n,t}[g_{n,t} - g]dt$, some locations' endowments can eventually shrink relative to aggregate, i.e., $\liminf_{T\to\infty} \alpha_{n,T} = 0$ with positive probability. In such case, self-fulfilling volatility must vanish asymptotically. However, this type of long-run degeneracy is not present if the "stabilizing force" stems from beliefs (Appendix B.2) or creative destruction (Appendix B.3).

These magnitudes can be theoretically verified for the M in (17) by considering the approximation $x_{n,t} = 1/N$ for all n. Then, the bounds simplify to

$$\sigma_t^* = \sqrt{A_t} \leq \sqrt{\frac{N-1}{2N}}\Pi_t.$$

Substituting $N = 10$, we obtain $\sigma_t^* \leq 16.8\%$ and $A_t \leq 2.8\%$, very close to the ranges displayed in Figure 2.

5 Empirical analysis

We now confront the main model prediction with the data. From Proposition 3, recall

$$\text{average return volatility } = \sigma_t^* \leq \sqrt{A_t} = \text{ square-root of arbitrage profits.} \tag{19}$$

Equation (19) predicts that there should be an increasing relationship between the square root of arbitrage profits ($\sqrt{A_t}$) and the value-weighted-average volatilities of assets comprising the arbitrage trade in question (σ_t^*). Building on Du et al. (2018), we investigate deviations from covered interest parity (CIP).

If we take the model very literally, relationship (19) should hold approximately one-for-one: that is, a regression of σ_t^* onto $\sqrt{A_t}$ should recover a regression coefficient of approximately 1.[14] We do not take the prediction so literally for two key reasons, most of which lead one to expect a regression coefficient below 1, in fact. First, in the presence of additional shocks, a substantial fraction of asset-price volatility will not be related to arbitrage profits. Second, the arbitrage of our model is built with long-lived assets, whereas any practical application uses shorter-dated assets that have lower volatility, through short duration alone. We attempt to partially address this below, with a variety of proxies for σ_t^*, but our choices are not without trade-offs.

Data and proxies. As a proxy for arbitrage profits A_t, we measure the 3-month absolute CIP deviations of the G10 currencies against the USD, and take a simple average across these currencies. In particular, for currency i and maturity m, CIP against the US says

$$\frac{1}{p_{US,t}^{(m)}} = \frac{s_{US\to i,t}}{f_{US\to i,t}^{(m)}} \frac{1}{p_{i,t}^{(m)}}, \tag{20}$$

[14] Of course, expression (19) is not an equality, because Jensen's inequality was used to obtain this expression. But as long as the Jensen deviation from equality is either small or uncorrelated with the variation in arbitrage profits, a unit regression coefficient should be expected.

where $p^{(m)}_{US,t}$ and $p^{(m)}_{i,t}$ are the m-maturity zero-coupon bond prices, $s_{US\to i,t}$ denotes the spot exchange rate, and $f^{(m)}_{US\to i,t}$ denotes the forward exchange rate. Exactly as in Du et al. (2018), we define the CIP deviation by the annualized incremental foreign bond yield needed to make (20) hold exactly. We take the absolute value of each G10 currency's CIP deviation to obtain our proxy

$$\hat{A}^{(m)}_{i,t} := \frac{1}{m}\left|\log\left(p^{(m)}_{i,t}\Big/p^{(m)}_{US,t}\right)+\log\left(f^{(m)}_{US\to i,t}\Big/s_{US\to i,t}\right)\right|. \tag{21}$$

For an aggregate time series measure, we take the simple average across currencies:

$$\hat{A}^{(m)}_{t} := \frac{1}{10}\sum_{i=1}^{10}\hat{A}^{(m)}_{i,t}. \tag{22}$$

Using 3-month contracts corresponds to setting $m = 1/4$. Our choice for $\hat{A}^{(m)}_t$ matches a few key features of the arbitrage in the model: it represents a relatively short-term trade that generates positive profits over the riskless rate for sure, if held to maturity. One-month or one-week CIP deviations may match better the fact that A_t represents profits over the infinitesimal time interval $[t, t+dt]$, but these contracts have other idiosyncratic features, as discussed in Du et al. (2018).

An appropriate choice for asset return volatility σ^*_t is more difficult, due to various data limitations and discrepancies between our simple model and the real world. As a baseline, we proxy σ^*_t by the monthly standard deviation of daily log price changes in the 10-year Constant Maturity Treasury note, annualized, i.e.,[15]

$$\hat{\sigma}^*_t := \sqrt{\frac{252}{30}\sum_{u=1}^{30}\left[\log\left(p^{(10)}_{US,t+u}\Big/p^{(10)}_{US,t+u-1}\right)-\frac{1}{30}\sum_{v=1}^{30}\log\left(p^{(10)}_{US,t+v}\Big/p^{(10)}_{US,t+v-1}\right)\right]^2}. \tag{23}$$

This choice for $\hat{\sigma}^*_t$ is simple and transparent but has several drawbacks: (a) it is not a very precise estimate of time-t conditional volatility; (b) it includes no volatility information for the synthetic US bonds, which constitute the other leg of the CIP trade; and (c) unlike the model, in which the assets used to construct A_t correspond to those used to measure σ^*_t, this proxy uses long-maturity bonds instead of the 3-month bonds comprising $\hat{A}^{(1/4)}_t$.

To help address concerns (a)-(c), we also consider three alternative measures of $\hat{\sigma}^*_t$ described briefly below and more extensively in Appendix C.

[15] These are not holding period returns, because we hold the time-to-maturity constant in calculating price at day t and day $t+1$. Without imposing our own model-based interpolation methods, we cannot observe the day-$t+1$ price of a Treasury that was a 10-year bond on day-t.

(a) For a more real-time measure of conditional volatility, we also examine the CBOE's 10-year Treasury VIX (TYVIX), which is the implied 30-day volatility of CBOT futures on 10-year US Treasury Notes. TYVIX applies the CBOE's VIX methodology to options on 10-year US Treasury Note futures.

(b) To include information for the foreign leg, we compute the volatility of foreign 10-year notes (converted to USD via spot exchange rates), analogous to (23). We then take the value-weighted-average of this measure and 10-year US note volatility.

(c) To bring the assets in the volatility construction as close as possible to those used in the arbitrage trade, we examine the value-weighted-average return volatilities of the 3-month US bill and the 3-month synthetic US bill. The prices of these bills are constructed using country-specific IBOR.

Results. Figure 3 summarizes the main empirical finding that arbitrage profits and volatility co-move strongly over time. The average absolute CIP deviation is high when the 10-year US Treasury volatility is high. Figure 4 shows the same result holds disaggregated at the currency level. Appendix C repeats the same figures for our three other proxies of σ_t^* and finds similar results in each case.

To formalize and quantify this link, Table 1, column (1), displays OLS results from a regression of $\hat{\sigma}_t^*$ on $[\hat{A}_{i,t}^{(1/4)}]^{1/2}$ for each currency i. Across all currencies, we document a very strong relationship, both statistically and economically, between 10-year US Treasury note volatility and CIP deviations. Results are strongest for Australia and New Zealand, currencies that play a central role in the carry trade, in practice. Amazingly, despite the caveats outlined at the beginning of this section, the regression coefficients are in the ballpark of 1, in line with the model prediction.

Now, turn to our other proxies for $\hat{\sigma}_t^*$ at the 10-year maturity in columns (2)-(4). Using instead the 10-year Treasury VIX (column 2) or including the 10-year foreign note (column 3) does not change the empirical message. The Treasury VIX regression coefficients are attenuated a bit, whereas the coefficients including the foreign note are magnified. If we use volatility from the 3-month bonds (column 4), our regressions produce slope estimates approximately 40-50 times lower, in line with the relative durations of a 10-year note and a 3-month bill.

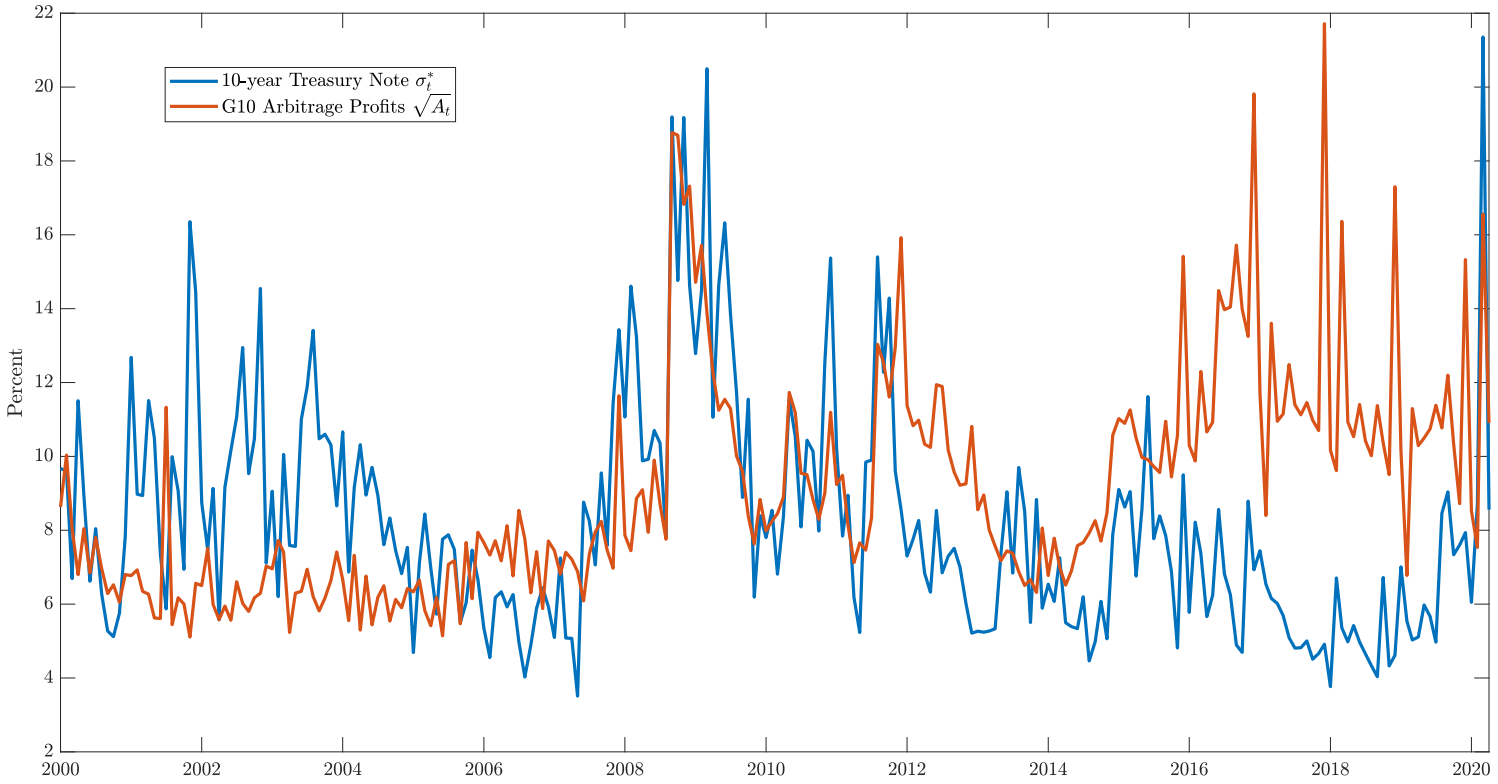

Figure 3: Monthly time-series of proxies for A_t and σ_t^*. Proxy for A_t is $\hat{A}_t^{(1/4)}$ from (22), which is constructed using 3-month absolute CIP deviations (against USD), measured daily, then averaged monthly, then (simple) averaged across the G10 currencies. Proxy for σ_t^* is $\hat{\sigma}_t^*$ from (23), which is constructed using the monthly standard deviation of daily log price changes of a 10-year Constant Maturity US Note. Both measures are annualized. Currency data are from Bloomberg, whereas 10-year Constant Maturity Treasury data are from the Board of Governors of the Federal Reserve System, retrieved from FRED. Data range from Jan. 2000 to Apr. 2020.

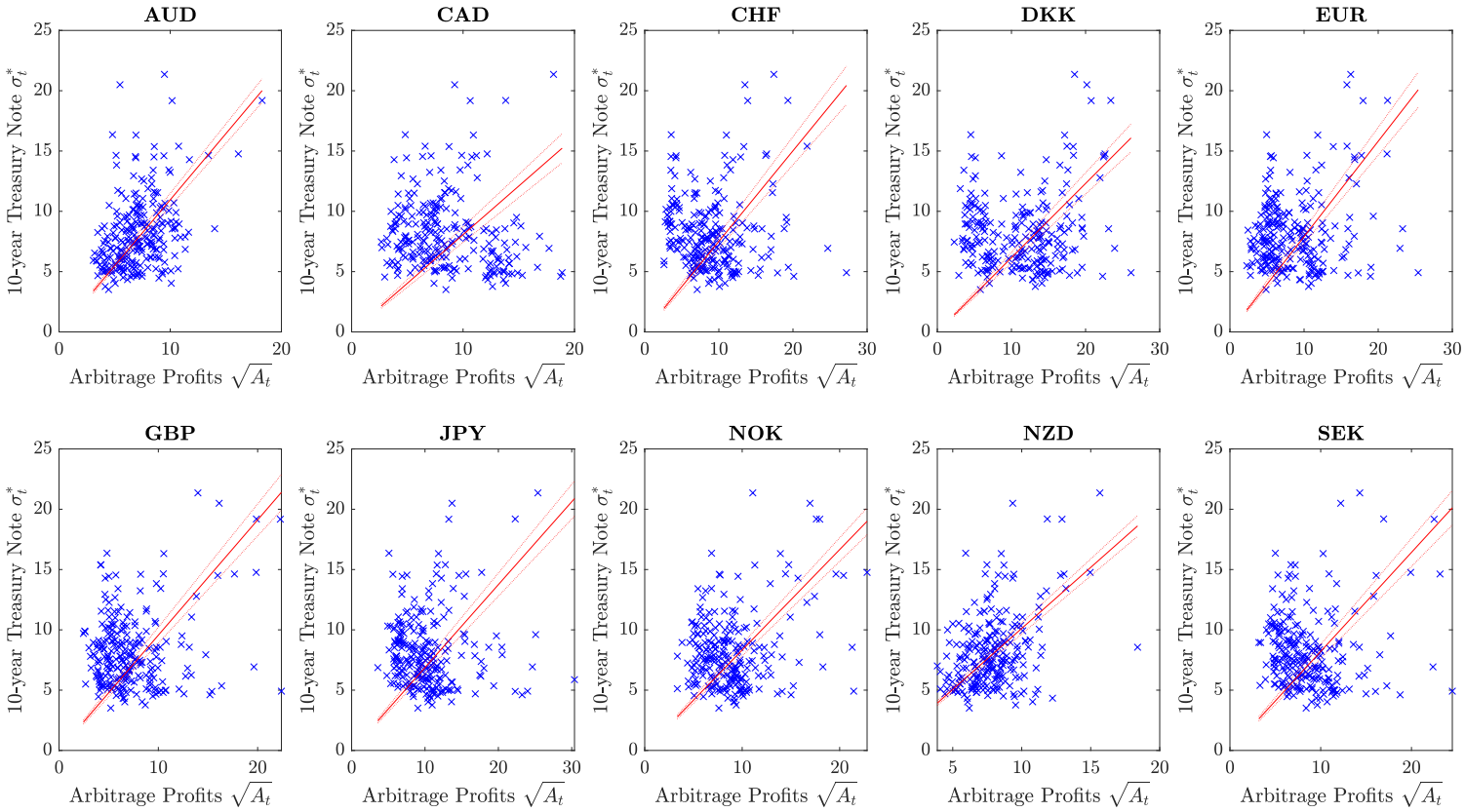

Figure 4: Currency-level OLS regressions of σ_t^* on $\sqrt{A_t}$ (monthly, no intercept). Proxy for A_t is $\hat{A}_t^{(1/4)}$ from (22), which is constructed using 3-month absolute CIP deviations (against USD), measured daily, then averaged monthly, then (simple) averaged across the G10 currencies. Proxy for σ_t^* is $\hat{\sigma}_t^*$ from (23), which is constructed using the monthly standard deviation of daily log price changes of a 10-year Constant Maturity US Note. Both measures are annualized. Regression lines, and 95% confidence intervals (using HAC standard errors), are also displayed. Currency data are from Bloomberg, whereas 10-year Constant Maturity Treasury data are from the Board of Governors of the Federal Reserve System, retrieved from FRED. Data range from Jan. 2000 to Apr. 2020.

	Proxy for $\hat{\sigma}_t^*$			
Currency	(1) 10y US Note	(2) TYVIX	(3) 10y US + foreign	(4) 3m US + foreign
aud	1.095 (0.042)	0.813 (0.035)	1.602 (0.071)	0.033 (0.008)
cad	0.806 (0.126)	0.581 (0.103)	0.871 (0.126)	0.019 (0.005)
chf	0.751 (0.068)	0.538 (0.046)	0.812 (0.070)	0.027 (0.005)
dkk	0.616 (0.047)	0.441 (0.035)	0.629 (0.047)	0.021 (0.005)
eur	0.790 (0.076)	0.561 (0.052)	0.941 (0.075)	0.024 (0.006)
gbp	0.956 (0.083)	0.693 (0.058)	1.153 (0.069)	0.031 (0.009)
jpy	0.687 (0.053)	0.514 (0.048)	0.689 (0.053)	0.025 (0.005)
nok	0.833 (0.052)	0.606 (0.043)	0.815 (0.055)	0.026 (0.006)
nzd	1.012 (0.043)	0.755 (0.047)	1.426 (0.066)	0.031 (0.006)
sek	0.821 (0.070)	0.595 (0.051)	0.789 (0.071)	0.026 (0.006)
Observations	244	208	244*	242

Table 1: OLS regressions of volatility $\hat{\sigma}_t^*$ on arbitrage profits $\sqrt{\hat{A}_t^{(1/4)}}$ across G10 currencies and using four different volatility proxies $\hat{\sigma}_t^*$. Going across the columns, the four volatility proxies are as follows: (1) the monthly volatility of daily log price changes on the 10-year Constant Maturity US Note; (2) the CBOE's 10-year Treasury VIX (TYVIX); (3) the value-weighted average volatilities, computed monthly from daily log price changes, of the 10-year US Treasury Note and the 10-year foreign note, where the latter is adjusted to USD by the spot exchange rate; (4) the value-weighted average volatilities, computed monthly from daily measures of next-two-month holding period returns, on the 3-month US bill and the 3-month synthetic US bill (constructed using a foreign bill, the spot exchange rate, and a forward currency swap). All four measures are annualized. The proxy for arbitrage profits is the currency-specific 3-month absolute CIP deviation as in (21), averaged monthly. Estimated regression coefficients are computed from OLS without an intercept. Standard errors are in parentheses and computed using heteroskedasticity and autocorrelation corrected (HAC) formulas. Currency and foreign note data are from Bloomberg. 10-year Constant Maturity Treasury data are from the Board of Governors of the Federal Reserve System, retrieved from FRED. TYVIX data from the CBOE. Currency data range from Jan. 2000 to Apr. 2020. TYVIX data ranges from Jan 2003 to Apr. 2020. 10-year G10 Note data range from Jan. 2000 to Apr. 2020 for AUD, CAD, DKK, EUR, GBP, NZD, and Jan. 2007 to Apr. 2020 for CHF, NOK, and SEK. *158 (CHF), 239 (DKK), 243 (JPY), 158 (NOK), 160 (SEK).

6 Conclusion

We have demonstrated, in a limits-to-arbitrage framework, the strong connection between the availability of arbitrage profits and the possibility of self-fulfilling volatility. Empirically, we have documented an association between available arbitrage profits in foreign exchange markets and volatility of the underlying instruments in these trades.

Often, the presence of multiple equilibria and self-fulfilling dynamics are viewed as a nuisance for theoretical models. But given that levels of asset-price volatility often far exceed predictions of many theoretical models, our mechanism can help bridge a gap in financial economics.

For example, consider corporate equity and bond markets. Although equity and bond returns are linked, one cannot construct a riskless portfolio from them in a simple way, unlike for covered interest parity.[16] Still, it is entirely possible, and anecdotally true, that equity investors differ from bond investors and that capital is slow moving, whether due to market segmentation or investor habitats. With this in mind, our model suggests some amount of redistributive self-fulfilling volatility should be possible between corporate equity and bond markets. In this sense, our focus on true arbitrages is just for clarity: one can measure accurately the amount of arbitrage profit without having to know investors' pricing kernels. We think that future research could, through a self-fulfilling mechanism, connect frictions such as market segmentation to "volatility puzzles" in other asset markets beyond those with self-evident arbitrage profits.

[16] With dynamic trading, knowledge of the underlying shocks that affect both securities, as well as their sensitivities to those shocks, one could obtain a no-arbitrage relation between equities and bonds.

References

Azariadis, Costas, "Self-fulfilling prophecies," *Journal of Economic Theory*, 1981, *25* (3), 380–396.

Bacchetta, Philippe, Cédric Tille, and Eric Van Wincoop, "Self-fulfilling risk panics," *American Economic Review*, 2012, *102* (7), 3674–3700.

Barberis, Nicholas, Robin Greenwood, Lawrence Jin, and Andrei Shleifer, "X-CAPM: An extrapolative capital asset pricing model," *Journal of Financial Economics*, 2015, *115* (1), 1–24.

Benhabib, Jess and Roger EA Farmer, "Indeterminacy and sunspots in macroeconomics," *Handbook of Macroeconomics*, 1999, *1*, 387–448.

_ , Pengfei Wang, and Yi Wen, "Sentiments and aggregate demand fluctuations," *Econometrica*, 2015, *83* (2), 549–585.

Biais, Bruno, Johan Hombert, and Pierre-Olivier Weill, "Incentive constrained risk sharing, segmentation, and asset pricing," *Working Paper*, 2017.

Blanchard, Olivier J, "Debt, Deficits, and Finite Horizons," *Journal of Political Economy*, 1985, *93* (2), 223–47.

Cass, David and Karl Shell, "Do sunspots matter?," *Journal of Political Economy*, 1983, *91* (2), 193–227.

Cvitanić, Jakša and Ioannis Karatzas, "Convex duality in constrained portfolio optimization," *The Annals of Applied Probability*, 1992, pp. 767–818.

Du, Wenxin, Alexander Tepper, and Adrien Verdelhan, "Deviations from covered interest rate parity," *The Journal of Finance*, 2018, *73* (3), 915–957.

Duffie, Darrell, "Presidential Address: Asset Price Dynamics with Slow-Moving Capital," *The Journal of finance*, 2010, *65* (4), 1237–1267.

_ and Bruno Strulovici, "Capital Mobility and Asset Pricing," *Econometrica*, 2012, *80* (6), 2469–2509.

Farmer, Roger EA, "The evolution of endogenous business cycles," *Macroeconomic Dynamics*, 2016, *20* (2), 544–557.

_ , "Pricing assets in a perpetual youth model," *Review of Economic Dynamics*, 2018, *30*, 106–124.

_ and Jess Benhabib, "Indeterminacy and increasing returns," *Journal of Economic Theory*, 1994, *63*, 19–41.

_ and Michael Woodford, "Self-fulfilling prophecies and the business cycle," *Macroeconomic Dynamics*, 1997, *1* (4), 740–769.

Fleckenstein, Matthias and Francis A Longstaff, "Shadow Funding Costs: Measuring the Cost of Balance Sheet Constraints," *Working Paper*, 2018.

Gabaix, Xavier and Matteo Maggiori, "International liquidity and exchange rate dynamics," *The Quarterly Journal of Economics*, 2015, *130* (3), 1369–1420.

Garleanu, Nicolae and Lasse Heje Pedersen, "Margin-based asset pricing and deviations from the law of one price," *The Review of Financial Studies*, 2011, *24* (6), 1980–2022.

Gârleanu, Nicolae and Stavros Panageas, "What to Expect when Everyone is Expecting: Self-Fulfilling Expectations and Asset-Pricing Puzzles," *Unpublished working paper. University of California, Los Angeles, CA*, 2019.

Gromb, Denis and Dimitri Vayanos, "Equilibrium and Welfare in Markets with Financially Constrained Arbitrageurs," *Journal of Financial Economics*, 2002, *66* (2), 361–407.

__ and __, "Limits of arbitrage," *Annu. Rev. Financ. Econ.*, 2010, *2* (1), 251–275.

Hansen, Lars Peter and Ravi Jagannathan, "Implications of security market data for models of dynamic economies," *Journal of Political Economy*, 1991, *99* (2), 225–262.

Hu, Grace Xing, Jun Pan, and Jiang Wang, "Noise as information for illiquidity," *The Journal of Finance*, 2013, *68* (6), 2341–2382.

Hugonnier, Julien, "Rational Asset Pricing Bubbles and Portfolio Constraints," *Journal of Economic Theory*, 2012, *147* (6), 2260–2302.

Itskhoki, Oleg and Dmitry Mukhin, "Exchange rate disconnect in general equilibrium," 2017.

Khasminskii, Rafail, *Stochastic stability of differential equations*, Vol. 66, Springer Science & Business Media, 2011.

Khorrami, Paymon, "Entry and slow-moving capital: using asset markets to infer the costs of risk concentration," *Available at SSRN 2777747*, 2018.

Kiyotaki, Nobuhiro and John Moore, "Credit cycles," *Journal of Political Economy*, 1997, *105* (2), 211–248.

Koijen, Ralph SJ and Motohiro Yogo, "A demand system approach to asset pricing," *Journal of Political Economy*, 2019, *127* (4), 1475–1515.

Krishnamurthy, Arvind, "The bond/old-bond spread," *Journal of Financial Economics*, 2002, *66* (2-3), 463–506.

Lamont, Owen A and Richard H Thaler, "Can the market add and subtract? Mispricing in tech stock carve-outs," *Journal of Political Economy*, 2003, *111* (2), 227–268.

Ma, Yueran, "Nonfinancial Firms as Cross-Market Arbitrageurs," *The Journal of Finance*, 2019, *74* (6), 3041–3087.

Scheinkman, Jose A and Wei Xiong, "Overconfidence and speculative bubbles," *Journal of Political Economy*, 2003, *111* (6), 1183–1220.

Shleifer, Andrei and Robert W Vishny, "The limits of arbitrage," *The Journal of Finance*, 1997, *52* (1), 35–55.

Vayanos, Dimitri and Pierre-Olivier Weill, "A search-based theory of the on-the-run phenomenon," *The Journal of Finance*, 2008, *63* (3), 1361–1398.

Zentefis, Alexander, "Self-fulfilling asset prices," 2020. Unpublished working paper. Yale University, New Haven, CT.

Appendix

A Proofs for Sections 3 and 4

Proof of Theorem 1. To prove the claim, we need to fill in any details that go beyond the discussion following the statement of Theorem 1. There are four brief steps needed to fill in the details.

Step 1: State prices. Each location has its own risk price $\tilde{\pi}_{n,t}$, which is the marginal utility sensitivity to the $d\tilde{Z}_{n,t}$ shock. The state price density for location n is then given by

$$d\xi_{n,t} = -\xi_{n,t}\Big[r_t dt + \tilde{\pi}_{n,t} d\tilde{Z}_{n,t}\Big]. \tag{24}$$

In these terms, we have the no-arbitrage pricing relation

$$\mu^q_{n,t} + g_{n,t} + \frac{1}{q_{n,t}} - r_t = \sigma^q_{n,t}\tilde{\pi}_{n,t}, \tag{25}$$

which suffices assuming $q_{n,t} > 0$. We can also pose things in terms of the basis shocks. Let $\pi_{n,t}$ be the risk price vector pertaining to dZ_t, which is potentially location-specific because of market segmentation. The link between these two, by substituting equation (2) into (24), is given in equation (12).

Step 2: Optimality. Log agents optimally consume δ fraction of their wealth when there are no bubbles. Investor n wealth is given by $y_{n,t}q_{n,t} + \beta_{n,t}$ where $\beta_{n,t}$ is their risk-free bond market position. Let $\theta_{n,t} := \frac{y_{n,t}q_{n,t}}{y_{n,t}q_{n,t}+\beta_{n,t}}$ be the fraction of wealth this investor puts in the local risky asset. Note that market clearing is imposed automatically in this formula, as the local investor n holds the entirety of the local asset. Given the dynamic conjecture for asset prices, and the consumption-wealth ratio δ, each investor then has consumption dynamics

$$\frac{dc_{n,t}}{c_{n,t}} = \Big[r_t - \delta + \theta_{n,t}\sigma^q_{n,t}\tilde{\pi}_{n,t}\Big]dt + \theta_{n,t}\sigma^q_{n,t}d\tilde{Z}_{n,t}. \tag{26}$$

Under these assumptions, optimal portfolio choices are given by the standard mean-variance formula $\theta_{n,t}\sigma^q_{n,t} = \tilde{\pi}_{n,t}$. Substituting this portfolio choice into (26), equilibrium consumption dynamics are

$$\frac{dc_{n,t}}{c_{n,t}} = \Big[r_t - \delta + \tilde{\pi}^2_{n,t}\Big]dt + \tilde{\pi}_{n,t}d\tilde{Z}_{n,t}. \tag{27}$$

From (24) and (27), we obtain $\xi_{n,t}c_{n,t} = \xi_{n,0}c_{n,0}\exp(-\delta t)$, so that the no-bubble static budget constraint (with wealth defined as $w_{n,t} := y_{n,t}q_{n,t} + \beta_{n,t}$)

$$\mathbb{E}_t\Big[\int_0^\infty \frac{\xi_{n,t+s}}{\xi_{n,t}}c_{n,t+s}ds\Big] = w_{n,t} \tag{28}$$

holds automatically with $c_{n,t} = \delta w_{n,t}$. This confirms that the optimal consumption rule and no bubbles are mutually consistent.

Step 3: Aggregation. Define the consumption shares $x_{n,t} := c_{n,t}/Y_t$ and recall the endowment shares $\alpha_{n,t} := y_{n,t}/Y_t$. Notice that $\theta_{n,t} = \delta\alpha_{n,t}q_{n,t}/x_{n,t}$, which, combined with optimal portfolio

choice, yields equation (8). Time-differentiating the goods market clearing condition $\sum_{n=1}^{N} c_{n,t} = Y_t$ and using (27), we have

$$r_t = \delta + g - \sum_{n=1}^{N} x_{n,t}\tilde{\pi}_{n,t}^2 \tag{29}$$

and

$$0 = \sum_{n=1}^{N} x_{n,t}\tilde{\pi}_{n,t}M_n. \tag{30}$$

Substituting (8) into (30) delivers equation (4). Also, combining the asset-pricing equation (25), which is an equation for μ_n^q, with the risk-free rate equation (29), one can show that (3) holds if and only if $\sum_{n=1}^{N} \alpha_n q_{n,0} = \delta^{-1}$, i.e., if an initial restriction holds for prices. In addition, note that consumption share dynamics are obtained by Itô's formula by equation (18).

Step 4: Transversality. Finally, to ensure no bubbles are present and that free disposal is not violated, we require the transversality condition on prices:

$$\lim_{T\to\infty} \mathbb{E}_t\Big[\xi_{n,T}q_{n,T}\Big] = 0. \tag{31}$$

Note that (31) is violated only if $\mathbb{P}[\limsup_{T\to\infty} q_{n,T} = +\infty \text{ or } \liminf_{T\to\infty} q_{n,T} = -\infty] > 0$, which leads to the sufficiency of boundedness for (31). □

Proof of Proposition 1. Given the transversality condition (31), we have

$$q_{n,t} = \mathbb{E}_t\Big[\int_t^\infty \frac{\xi_{n,s}}{\xi_{n,t}}\frac{y_{n,s}}{y_{n,t}}ds\Big].$$

Using $g_{n,t} = g$ for all (n,t) and $r_t = \delta + g - A_t \leq \delta + g$, where $A_t := \sum_{n=1}^{N} x_{n,t}\tilde{\pi}_{n,t}^2 \geq 0$, we have

$$q_{n,t} = \int_t^\infty e^{-\delta(s-t)}\tilde{\mathbb{E}}_t^n\Big[\exp(\int_t^s A_u du)\Big]ds \geq \int_t^\infty e^{-\delta(s-t)}ds = \delta^{-1},$$

where $\tilde{\mathbb{E}}_t^n$ is the location-n risk-neutral expectation, which is mutually absolutely-continuous with respect to $\mathbb{E}$. Using the bond-market clearing condition (3), we must have $q_{n,t} = \delta^{-1}$ for all (n,t). □

Proof of Proposition 2. Consider $g_{n,t} = g + \lambda(q_{n,t} - \delta^{-1})$ with $\lambda > \delta^2$ and fixed ϵ that satisfies $0 < \epsilon < \delta^{-2} - \lambda^{-1}$. Supposing $\text{rank}(M) < N$, conjecture a stochastic equilibrium exists with $\alpha_{n,t}q_{n,t}\sigma_{n,t}^q = v_n^*\psi_t$ and $\tilde{\pi}_{n,t} = \delta v_n^*\psi_t/x_{n,t}$ for some process ψ_t. Substituting these and all other equilibrium objects into the asset-pricing equation (25), we have

$$dq_{n,t} = \Big[-1 + \Big(\delta + \lambda\delta^{-1} - \delta^2\psi_t^2\sum_{i=1}^{N}\frac{(v_i^*)^2}{x_{i,t}}\Big)q_{n,t} - \lambda q_{n,t}^2 + \delta\frac{(v_n^*\psi_t)^2}{\alpha_{n,t}x_{n,t}}\Big]dt + \frac{v_n^*}{\alpha_{n,t}}\psi_t M_n dZ_t. \tag{32}$$

We show that if properties (P1) and (P2) are satisfied, then $q_{n,t}$ remains bounded for all n. As a preliminary, define

$$D(q) := -1 + (\delta + \lambda\delta^{-1})q - \lambda q^2. \tag{33}$$

When $\psi_t = 0$, all local price-dividend ratios follow $dq_{n,t} = D(q_{n,t})dt$. Note that $D(q) = 0$ is a quadratic equation that has two roots: δ^{-1} and $\delta\lambda^{-1}$. Moreover, $D(q) > 0$ if and only if $q \in (\delta\lambda^{-1}, \delta^{-1})$.

Under property (P2), if $\underline{q}_t = \delta(\epsilon + \lambda^{-1})$, we have $\psi_t = 0$ and so

$$d\underline{q}_t = D\big(\delta(\epsilon + \lambda^{-1})\big)dt > 0.$$

Note that, under property (P1), the drift and diffusion coefficients of $q_{n,t}$ are bounded, so $q_{n,t}$ is almost-surely path-continuous. This proves that the entire path is bounded below: if $q_{n,0} > \delta(\epsilon + \lambda^{-1})$ for all n, then $\{q_{n,t}\}_{t\geq 0} > \delta(\epsilon + \lambda^{-1})$ for each n almost-surely.

On the other hand, bond market clearing (3), plus this lower bound on valuations, implies an upper bound on the maximal valuation:

$$\bar{q}_t := \max_n q_{n,t} < \underbrace{(\alpha_{\bar{n}_t,t}\delta)^{-1} - (1 - \alpha_{\bar{n}_t,t})\delta(\epsilon + \lambda^{-1})}_{:=b_t}, \quad \text{where} \quad \bar{n}_t := \arg\max_n q_{n,t}.$$

It suffices to show that $\mathbb{P}[\sup_t b_t < +\infty] = 1$. As long as $\underline{\alpha}_t > 0$, we always have $b_t < +\infty$. As a result, we need only consider the case $\underline{\alpha}_t = \alpha_{\bar{n}_t,t}$ (i.e., the location with maximal valuation is the location with minimal endowment share) and suppose $\underline{\alpha}_t \searrow 0$. However, since $\bar{q}_t > \delta^{-1}$, we have

$$d\alpha_{\bar{n}_t,t} = \alpha_{\bar{n}_t,t}\lambda[\bar{q}_t - \delta^{-1}]dt > 0,$$

which contradicts $\underline{\alpha}_t \searrow 0$.

In summary, $\{(q_{n,t})_{n=1}^N : t \geq 0\}$ are bounded almost-surely, so the conditions of Theorem 1 are satisfied and $\alpha_{n,t}q_{n,t}\sigma^q_{n,t} = v^*_n\psi_t$ is indeed an equilibrium. This completes the proof. □

Proof of Theorem 2. First, assuming the existence of self-fulfilling volatility, let us find a portfolio that has no risk but pays a positive premium over the riskless rate. Consider a portfolio that goes long $\delta\alpha_{n,t}q_{n,t}$ of each asset $n = 1, \ldots, N$, which costs 1 by equation (3). As stated in equation (25), each asset n has expected excess returns that are given by the product of the location-n risk quantity times the risk price: $\sigma^q_{n,t}\tilde{\pi}_{n,t}$. Using equation (8) to substitute $\tilde{\pi}_{n,t}$, the portfolio excess return is

$$\sum_{n=1}^N x_{n,t}\delta^2\Big(\frac{\alpha_{n,t}q_{n,t}}{x_{n,t}}\Big)^2(\sigma^q_{n,t})^2 \geq 0,$$

which is strictly positive as long as any self-fulfilling volatility obtains. Using the expression for $\tilde{\pi}_{n,t}$, one can easily verify this expression is equivalent to A_t in (9). At the same time, by equation (4), the portfolio volatility is identically zero. This shows that an arbitrage always emerges if there is self-fulfilling volatility.

Next, the claim that absence of self-fulfilling volatility implies no arbitrage follows from (25), whereby all assets return r_t when $\sigma^q_{n,t} = 0$. □

Proof of Proposition 3. Substituting $\alpha_{n,t}q_{n,t}\sigma^q_{n,t} = \psi_t v^*_n$ from Theorem 1 into location-specific risk prices of (8), and substituting the result into (9), we have

$$A_t = \delta^2\psi_t^2\sum_{n=1}^N x_{n,t}\Big(\frac{v^*_n}{x_{n,t}}\Big)^2$$

By inverting this relationship, the amount of self-fulfilling volatility ψ_t can be inferred from A_t, which gives the equality in (10). The upper bound can be obtained by substituting

$$\sum_{n=1}^{N} x_{n,t}\Big(\frac{v_n^*}{x_{n,t}}\Big)^2 \geq \Big(\sum_{n=1}^{N} x_{n,t}\frac{v_n^*}{x_{n,t}}\Big)^2 = (\mathbf{1}'v^*)^2,$$

which holds by Jensen's inequality. To obtain the equality in (11), substitute (3) into the definition of σ_t^* and use the result from Theorem 1 that $\alpha_{n,t}q_{n,t}\sigma_{n,t}^q = \psi_t v_n^*$. To obtain the inequality, use (10). □

Proof of Proposition 4. Substitute equation (12) into equation (8) to get

$$\pi_{n,t} = \delta\Big(\frac{\alpha_{n,t}q_{n,t}}{x_{n,t}}\Big)\sigma_{n,t}^q M_n.$$

Now, use the result of Theorem 1 that $\alpha_{n,t}q_{n,t}\sigma_{n,t}^q = v_n^*\psi_t$. Combining these equations, we have

$$\pi_{n,t} = \delta v_n^*\psi_t\frac{M_n}{x_{n,t}}. \tag{34}$$

Assumption (13) is equivalent to

$$\delta\psi_t \max_{(i,j):i\neq j}\Big\|\frac{v_i^*M_i}{x_{i,t}} - \frac{v_j^*M_j}{x_{j,t}}\Big\| \leq \Pi_t.$$

Solving for ψ_t, we obtain inequality (14). The bounds for σ_t^* are a direct consequence of (14). □

Proof of Corollary 5. To get the both bounds, begin with the volatility bound (14) of Proposition 4 and use

$$\|\pi_{n,t}\| = \delta\psi_t\frac{v_n^*}{x_{n,t}}$$
$$\sqrt{A_t} = \delta\psi_t\sqrt{\sum_{n=1}^{N} x_{n,t}(\frac{v_n^*}{x_{n,t}})^2}.$$

The expression for $\|\pi_{n,t}\|$ comes from taking the norm of equation (34) and using the fact that MM' has ones on its diagonal (this was a normalization). The expression for $\sqrt{A_t}$ comes from expression (10) in Proposition 3. □

B Model extensions

B.1 Aggregate shocks

Here, we allow for aggregate shocks hitting the endowments. Location-specific endowments now follow

$$dy_{n,t} = y_{n,t}[g_{n,t}dt + \nu dB_t],$$

where B_t is an aggregate Brownian shock, independent of the extrinsic shocks Z_t (and by extension $\tilde{Z}_t$). We maintain the restriction $\sum_{n=1}^N y_{n,t}g_{n,t} = Y_t g$. Thus, the aggregate endowment follows

$$dY_t = Y_t[gdt + \nu dB_t].$$

Conjecture that local price-dividend ratios follow

$$dq_{n,t} = q_{n,t}\left[\mu^q_{n,t}dt + \sigma^q_{n,t}d\tilde{Z}_{n,t} + \varsigma^q_{n,t}dB_t\right],$$

where $(\mu^q_{n,t}, \sigma^q_{n,t}, \varsigma^q_{n,t})$ are determined in equilibrium. We will proceed by making one of two possible assumptions on the tradability of this aggregate shock.

Assumption 1. *One of the following holds:*

(a) there are no additional markets open beyond those assumed so far;

(b) there is an integrated market in which agents frictionlessly trade a zero-net-supply Arrow security that has a unit loading on dB_t.

In both cases of Assumption 1, all previous results on self-fulfilling volatility go through. However, we uncover a surprising nuance: equilibrium is consistent with local assets having nearly arbitrary sensitivities to the aggregate shock.

Proposition 6. *With aggregate shocks, the conclusions of Theorem 1 on $(\sigma^q_{n,t})_{n=1}^N$ continue to hold without modification. Regarding $(\varsigma^q_{n,t})_{n=1}^N$, we have the following. Let $(\phi_{n,t})_{n=1}^{N-1}$ be a collection of arbitrary stochastic processes and set $\phi_{N,t} := -\sum_{n=1}^{N-1}\phi_{n,t}$. Then, there exists an equilibrium with $\alpha_{n,t}q_{n,t}\varsigma^q_{n,t} = \phi_{n,t}$ as long as the resulting $\{(q_{n,t})_{n=1}^N\}_{t\geq 0}$ is a bounded, positive process.*

Before giving a formal proof, we provide the basic sketch of the argument. Because our log agents will still consume δ fraction of their wealth in this environment, equilibrium still satisfies equation (3), that $\sum_{n=1}^N \alpha_{n,t}q_{n,t} = \delta^{-1}$. If we time-differentiate this condition as before, matching diffusion terms leads us to

$$\text{(match } dZ_t \text{ terms)} \quad 0 = \sum_{n=1}^N \alpha_{n,t}q_{n,t}\sigma^q_{n,t}M_n \tag{35}$$

$$\text{(match } dB_t \text{ terms)} \quad 0 = \sum_{n=1}^N \alpha_{n,t}q_{n,t}\varsigma^q_{n,t}. \tag{36}$$

Equation (35) is identical to equation (4), which is why the results of Theorem 1 continue to hold. For equation (36), of course it is possible to have $\varsigma^q_{n,t} = 0$ for all n. But we may also set $(\varsigma^q_{n,t})_{n=1}^{N-1}$ arbitrarily, so long as $\varsigma^q_{N,t}$ offsets these sensitivities. Thus, the volatilities have a similar redistributive flavor as before.

This is indeed an equilibrium, as long as the induced dynamics of price-dividend ratios are stationary. To this end, we can easily extend Propositions 1 and 2 to this setting with aggregate shocks. With common growth rates $g_{n,t} = g$, there will be no multiplicity ($\sigma^q_{n,t} = \varsigma^q_{n,t} = 0$), as the only prices consistent with the transversality condition are $q_{n,t} = \delta^{-1}$. With growth rates that increase sufficiently quickly in local valuations, we can generate stochastic multiplicity, because all that is required is to have both $\sigma^q_{n,t}$ and $\varsigma^q_{n,t}$ vanish whenever $\min_n q_{n,t}$ or $\min_n \alpha_{n,t}$ become "too small". We omit the details of these results.[17]

The intuition for self-fulfilling fundamental sensitivities differs depending on whether the shock is hedgable or not. When agents cannot hedge the dB_t shock, the logic is similar to the baseline model: agents adjust their consumption, through the bond market, to their conjecture about how the local asset co-moves with the fundamental shock. When agents trade Arrow securities on dB_t in an integrated market, they do not care whether or not their local asset responds to this shock. Enough hedging and risk-sharing will occur in equilibrium such that individual consumptions all have sensitivity ν to dB_t. Under a particular conjecture about $\varsigma^q_{n,t}$, location-n agents will form a hedging plan in order to undo this exposure. This is self-fulfilling: as long as asset prices move according to the conjecture, the hedging plan was correct.

Proof of Proposition 6. We will nest cases (a) and (b) of Assumption 1 in the following setting. Introduce an Arrow security that pays off $\eta_{n,t}dt + dB_t$ per unit of time, where $(\eta_{n,t})_{n=1}^N$ will be determined endogenously. Thus, agent n faces the state-price density process, modified from (24):

$$d\xi_{n,t} = -\xi_{n,t}\Big[r_t dt + \tilde{\pi}_{n,t} d\tilde{Z}_{n,t} + \eta_{n,t} dB_t\Big]. \tag{37}$$

Let $\theta^{\text{agg}}_{n,t}$ be the fraction of wealth a location-n agent invests in the Arrow security, and let $\theta_{n,t}$ be the fraction of wealth invested in the location-specific capital asset as before. The wealth of agent n has the following dynamics (dynamic budget constraint)

$$\begin{aligned}\frac{dw_{n,t}}{w_{n,t}} = &\Big[r_t - \frac{c_{n,t}}{w_{n,t}} + \theta_{n,t}\sigma^q_{n,t}\tilde{\pi}_{n,t} + \Big(\theta_{n,t}(\nu + \varsigma^q_{n,t}) + \theta^{\text{agg}}_{n,t}\Big)\eta_{n,t}\Big]dt \\ &+ \theta_{n,t}\sigma^q_{n,t} d\tilde{Z}_{n,t} + \Big(\theta_{n,t}(\nu + \varsigma^q_{n,t}) + \theta^{\text{agg}}_{n,t}\Big)dB_t.\end{aligned} \tag{38}$$

To implement (a), where agents are not allowed to trade the Arrow security, we impose a fictitious market clearing condition $\theta^{\text{agg}}_{n,t} = 0$ for all n, which will pin down $\eta_{n,t}$ such that no trading in the Arrow security occurs. From the results of Cvitanić and Karatzas (1992), this implements the same equilibrium as if we never introduced this fictitious market. To implement (b), in which the Arrow market exists and is integrated, we impose $\eta_{n,t} = \eta_t$ for all n and clear the market via $\sum_{n=1}^N x_{n,t}\theta^{\text{agg}}_{n,t} = 0$. In both cases, we have the capital market clearing condition $\theta_{n,t} = y_{n,t}q_{n,t}/w_{n,t}$ as before.

Thus, we may nest cases (a) and (b) by solving unconstrained optimization problems for our investors, augmented with the general state-price density process (37) as long as $\eta_{n,t}$ is chosen appropriately. Given the state-price density, the pricing condition (25) is replaced by

$$\mu^q_{n,t} + g_{n,t} + \frac{1}{q_{n,t}} + \nu\varsigma^q_{n,t} - r_t = \sigma^q_{n,t}\tilde{\pi}_{n,t} + (\nu + \varsigma^q_{n,t})\eta_{n,t}, \tag{39}$$

along with the requirement $q_{n,t} > 0$. Because all agents have log utility and effectively solve unconstrained portfolio problems with homogeneous wealth dynamics (38), they all consume δ fraction of

[17] To prove an analogous result to Proposition 2 formally, it is convenient that all locations have equal exposures ν to the aggregate shock, so that $\alpha_{n,t}$ evolves locally deterministically for all n.

their wealth, i.e., $c_{n,t} = \delta w_{n,t}$. Then, as B_t and $\tilde{Z}_{n,t}$ are independent, optimal consumption dynamics (27) are modified to read

$$\frac{dc_{n,t}}{c_{n,t}} = \left[r_t - \delta + \tilde{\pi}_{n,t}^2 + \eta_{n,t}^2\right]dt + \tilde{\pi}_{n,t}d\tilde{Z}_{n,t} + \eta_{n,t}dB_t.$$

Because $dw_{n,t}/w_{n,t} = dc_{n,t}/c_{n,t}$, we therefore have

$$\tilde{\pi}_{n,t} = \theta_{n,t}\sigma_{n,t}^q = \frac{\delta\alpha_{n,t}q_{n,t}}{x_{n,t}}\sigma_{n,t}^q$$
$$\eta_{n,t} = \theta_{n,t}(\nu + \varsigma_{n,t}^q) + \theta_{n,t}^{\text{agg}} = \frac{\delta\alpha_{n,t}q_{n,t}}{x_{n,t}}(\nu + \varsigma_{n,t}^q) + \theta_{n,t}^{\text{agg}}.$$

The first equation is identical to (8).

Now, we aggregate. First, equation (3) still holds, since agents consume δ fraction of wealth, and since both the bond market and the Arrow markets are in zero net supply. Next, time-differentiate the goods market clearing condition $\sum_{n=1}^N c_{n,t} = Y_t$ and match drift and diffusion terms to obtain

$$r_t = \delta + g - \sum_{n=1}^N x_{n,t}\tilde{\pi}_{n,t}^2 - \sum_{n=1}^N x_{n,t}\eta_{n,t}^2$$
$$0 = \sum_{n=1}^N x_{n,t}\tilde{\pi}_{n,t}M_n$$
$$\nu = \sum_{n=1}^N x_{n,t}\eta_{n,t}.$$

Using the expressions for $\tilde{\pi}_{n,t}$ and $\eta_{n,t}$ above, along with the condition $\sum_{n=1}^N x_{n,t}\theta_{n,t}^{\text{agg}} = 0$ (which holds in cases (a) and (b) both), we obtain equations (35)-(36). Thus, $\sigma_{n,t}^q$ and $\tilde{\pi}_{n,t}$ are solved exactly as in Theorem 1. Letting $(\phi_{n,t})_{n=1}^{N-1}$ be arbitrary processes, and putting $\phi_{N,t} = -\sum_{n=1}^{N-1}\phi_{n,t}$, we may satisfy (36) by setting $\varsigma_{n,t}^q$ by $\phi_{n,t} = \alpha_{n,t}q_{n,t}\varsigma_{n,t}^q$. As before, this is an equilibrium as long as the transversality condition (31) is satisfied, for which it suffices to show that $q_{n,t}$ is almost-surely bounded.

It remains to solve for $(\eta_{n,t})_{n=1}^N$. In case (a), we use $\theta_{n,t}^{\text{agg}} = 0$ in conjunction with the expression for $\eta_{n,t}$ above to get $\eta_{n,t} = \delta\alpha_{n,t}q_{n,t}(\nu + \varsigma_{n,t}^q)/x_{n,t}$. In case (b), we impose $\eta_{n,t} = \eta_t$ for all n, which after substituting into $\nu = \sum_{n=1}^N x_{n,t}\eta_{n,t}$ yields $\eta_t = \nu$. □

B.2 Procyclical perceived growth as a "stabilizing force"

In this section, we show that a particular type of heterogeneous beliefs about growth can lead to the same conclusions about multiplicity as in Proposition 2. In particular, if local investors become sufficiently optimistic about growth when valuations are high (and vice versa), the economy possesses a natural stabilizing force that facilitates adding stochastic extrinsic shocks.

Simplifying assumptions. To simplify this analysis, assume all local growth rates are the same, $g_{n,t} = g$. Furthermore, suppose there are no fundamental shocks (unlike Appendix B.1). Although actual growth rates are the same, assume that local agents irrationally perceive a growth rate $\tilde{g}_{n,t}$ which is potentially decoupled from g.[18]

[18] In this deterministic setting, such divergence constitutes a belief that is mutually singular to the true "probability distribution," but we ignore this issue here.

Solution. In this model, the asset-pricing equation (39) is replaced by

$$\mu^q_{n,t} + \tilde{g}_{n,t} + \frac{1}{q_{n,t}} - r_t = \sigma^q_{n,t}\tilde{\pi}_{n,t}. \tag{40}$$

This is identical to (39), except for the fact that actual growth $g_{n,t}$ is replaced by perceived growth $\tilde{g}_{n,t}$. Consequently, all the stability analysis of Proposition 2 goes through, if $\tilde{g}_{n,t}$ increases fast enough with $q_{n,t}$. Therefore, we state without proof the following proposition. Note that the endowment share $\alpha_{n,t}$ are constant in here, so the counterpart to property (P1) can be relaxed a bit.

Proposition 7. *Assume beliefs about local growth rates satisfy $\tilde{g}_{n,t} = g + \lambda(q_{n,t} - \delta^{-1})$ with $\lambda > \delta^2$. Then, self-fulfilling volatility is possible: there exists a non-zero process $\{\psi_t\}_{t\geq 0}$ such that an equilibrium exists with $\alpha_n q_{n,t}\sigma^q_{n,t} = v^*_n\psi_t$ for all n, where v^* is in the null-space of M'. In terms of the cross-sectional minimums $\underline{x}_t := \min_n x_{n,t}$ and $\underline{q}_t := \min_n q_{n,t}$, the volatility ψ_t can be any bounded process that satisfies*

(P1') *$\psi_t/\underline{x}_t$ are bounded;*

(P2') *ψ_t vanishes as $\underline{q}_t$ approaches $\delta(\epsilon + \lambda^{-1})$ from above, for some $0 < \epsilon < \delta^{-2} - \lambda^{-1}$.*

Extrapolation. As mentioned in the main text, this formulation of beliefs bears some resemblance to extrapolative beliefs modeled in Barberis et al. (2015). To see this, write local agent's perceived expected returns $\tilde{\mu}^R_{n,t}$ in terms of the objective expected returns $\mu^R_{n,t} := \mu^q_{n,t} + g + 1/q_{n,t}$:

$$\tilde{\mu}^R_{n,t} = \mu^R_{n,t} + (\tilde{g}_{n,t} - g).$$

In the equilibrium of Proposition 7, the perceived growth differential $\tilde{g}_{n,t} - g = \lambda(q_{n,t} - \delta^{-1})$ is a stationary, mean-reverting process, and it responds positively to a positive return shock $d\tilde{Z}_{n,t} > 0$. This is very similar to Barberis et al. (2015), in which extrapolators' beliefs are the sum of true expected returns and a stationary, mean-zero process that responds positively to return shocks.

B.3 Creative destruction as a "stabilizing force"

In this section, we consider another model that allows multiplicity. We show how an overlapping generations (OLG) "perpetual youth" economy – built upon Blanchard (1985) – augmented with a particular type of creative destruction – similar to Gârleanu and Panageas (2019) – creates a stabilizing force upon which extrinsic shocks can be layered. In particular, if existing firms can better insulate themselves from creative destruction when asset valuations are high, the economy possesses a natural stabilizing force. The contribution relative to Gârleanu and Panageas (2019) is to show how this is possible with an arbitrary number of assets (corresponding to the N locations) whose markets are in addition not integrated.

Cohorts, Endowments, Markets. In this model, all agents face a constant hazard rate of death $\beta > 0$, with all dying agents replaced by newborns (in the same location), so that population size is constant at 1. To keep matters simple, assume all locations have identical constant endowment growth rates and no shocks. That said, the endowment growth of an individual agent differs from the aggregate growth rate; this is the crucial ingredient in this model.

In particular, we assume some amount of *creative destruction*. The endowments of living agents decay at rate $\kappa_{n,t}$ (obsolescence rate), while newborn agents arrive to the economy with new trees of

total size $\kappa_{n,t} + g$ (or in per capita units, their individual trees are $(\kappa_{n,t} + g)/\beta$ in size). Specifically, the time-t endowment accruing to location-n agents born at time $s \leq t$ is

$$y_{n,t}^{(s)} = y_{n,t}(\kappa_{n,s} + g) \exp\Big[- \int_s^t (\kappa_{n,u} + g) du \Big].$$

To make sure things aggregate, note the aggregate endowment follows

$$dy_{n,t} = d\Big(\int_{-\infty}^t y_{n,t}^{(s)} ds \Big) = y_{n,t}^{(t)} dt + \int_{-\infty}^t dy_{n,t}^{(s)} ds = \underbrace{y_{n,t}(\kappa_{n,t} + g) dt}_{\text{newborn entry}} - \underbrace{y_{n,t}\kappa_{n,t} dt}_{\text{obsolescence}} = y_{n,t} g dt.$$

For now, we leave $\kappa_{n,t}$ unspecified, but note that its formulation will be the determinant of whether multiplicity is possible or not.

Agents can only trade in financial markets while alive. In addition to tradability of claims to local endowments, agents can access a market for annuities that insures their death hazard and provides a stream of $\beta w_{n,t}^{(s)}$ of income per unit of time, where $w_{n,t}^{(s)}$ is the wealth of a location-n agent born at time $s \leq t$ (this is standard in perpetual youth models).

Solution. Under these assumptions, one can show that agents consume $\delta + \beta$ fraction of their wealth, so that bond market clearing condition (3) is replaced by

$$\sum_{n=1}^N \alpha_n q_{n,t} = (\delta + \beta)^{-1},$$

where $q_{n,t}$ is the (aggregated across cohorts) location-n valuation ratio. Let $\xi_{n,t}$ denote the location-n state-price density, which follows

$$d\xi_{n,t} = -\xi_{n,t} \Big[r_t dt + \tilde{\pi}_{n,t} d\tilde{Z}_{n,t} \Big].$$

We will continue to assume a bubble-free equilibrium, so that

$$q_{n,t} = \mathbb{E}_t \Big[\int_t^\infty \frac{\xi_{n,\tau}}{\xi_{n,t}} \frac{y_{n,\tau}^{(s)}}{y_{n,t}^{(s)}} d\tau \Big] \quad \text{(for any birth-date } s \leq t\text{, this yields the same answer).}$$

Critically, this valuation does not incorporate wealth gains due to entry of future newborns. The dynamic counterpart of this valuation equation is, for some diffusion coefficient $\sigma_{n,t}^q$,

$$\frac{dq_{n,t}}{q_{n,t}} = \Big[r_t + \kappa_{n,t} - \frac{1}{q_{n,t}} + \sigma_{n,t}^q \tilde{\pi}_{n,t} \Big] dt + \sigma_{n,t}^q d\tilde{Z}_{n,t}. \tag{41}$$

The equilibrium riskless rate is obtained as follows. The goods market is integrated across locations, so the market clearing condition is given by

$$Y_t = \sum_{n=1}^N y_{n,t} = \sum_{n=1}^N \int_{-\infty}^t \beta e^{-\beta(t-s)} c_{n,t}^{(s)} ds.$$

Optimal consumption dynamics for alive agents are

$$\frac{dc_{n,t}^{(s)}}{c_{n,t}^{(s)}} = \Big[r_t - \delta + \tilde{\pi}_{n,t}^2 \Big] dt + \tilde{\pi}_{n,t} d\tilde{Z}_{n,t},$$

whereas newborn agents consume

$$\beta c_{n,t}^{(t)} = \underbrace{(\delta+\beta)}_{\substack{\text{cons-wealth}\\ \text{ratio}}} \times \underbrace{(\kappa_{n,t}+g)y_{n,t}q_{n,t}}_{\text{newborn wealth}}.$$

Time-differentiating goods market clearing, and using these results, we obtain

$$r_t = \delta + \beta - \sum_{n=1}^{N} x_{n,t}\tilde{\pi}_{n,t}^2 - (\delta+\beta)\sum_{n=1}^{N}\alpha_n q_{n,t}\kappa_{n,t}. \tag{42}$$

Stability. To see how the stabilizing force works, it is instructive to once again study the deterministic equilibrium in which extrinsic shocks have no volatility. Substituting (42) into (41) with $\sigma_{n,t}^q = 0$, we obtain

$$\dot{q}_{n,t} = \underbrace{-1+(\delta+\beta)q_{n,t}}_{\text{unstable component}} - \underbrace{\Big[(\delta+\beta)\sum_{i=1}^{N}\alpha_i q_{i,t}\kappa_{i,t} - \kappa_{n,t}\Big]q_{n,t}}_{\text{stabilizing force}} \quad \text{when} \quad \sigma_{i,t}^q = 0 \quad \forall i. \tag{43}$$

The first piece is the unstable component, propelling valuations further and further away from the "steady state" value $(\delta+\beta)^{-1}$. The second piece, capturing the relative amount of creative destruction in location n, is the stabilizing force.

Based on equation (43), we claim that if $\kappa_{n,t}$ decreases sufficiently rapidly as $q_{n,t}$ increases, then valuation dynamics are stable. To see this in a transparent way, assume similar to Proposition 2 that[19]

$$\kappa_{n,t} = \bar{\kappa} - \lambda\big[q_{n,t} - (\delta+\beta)^{-1}\big]. \tag{44}$$

Then, compute

$$\frac{\partial \dot{q}_n}{\partial q_m}\Big|_{q_i=(\delta+\beta)^{-1}\,\forall\, i} = \begin{cases} \delta+\beta-\lambda(\delta+\beta)^{-1}(1-\alpha_n)-\alpha_n\bar{\kappa}, & \text{if } m=n; \\ \lambda(\delta+\beta)^{-1}\alpha_m - \alpha_m\bar{\kappa}, & \text{if } m\neq n. \end{cases}$$

Construct the steady-state Jacobian matrix as

$$J := \Big[\frac{\partial \dot{q}_n}{\partial q_m}\Big|_{q_i=(\delta+\beta)^{-1}\,\forall\, i}\Big]_{1\le n,m\le N}. \tag{45}$$

Local stability of the steady-state can be determined by the eigenvalues of J. By the Gershgorin circle theorem, all of these eigenvalues will have strictly negative real parts if J has negative diagonal elements and is diagonally dominant. This is easily guaranteed by making $\bar{\kappa}$ and λ large enough, meaning the amount of creative destruction and its sensitivity to prices are both large enough. The result is summarized in the following lemma, with the proof omitted.

Lemma 1. *Assume $\bar{\kappa} > \delta+\beta$ and $\lambda > (\delta+\beta)\bar{\kappa}$. Then, all eigenvalues of J have strictly negative real parts. Consequently, the equilibrium of the creative destruction model is locally stable.*

[19] The fact that $\kappa_{n,t}$ is modeled as a linear function of $q_{n,t}$ implies that it can be negative, which strains the "obsolescence" interpretation. We ignore this issue here, because κ_n can always be linearized near enough to the steady state.

C Other empirical proxies for volatility

Recall our baseline choice for $\hat{\sigma}_t^*$ in (23) uses only the monthly volatility of daily price changes for the 10-year US Treasury Note. This measure is simple and transparent but has several drawbacks: (a) it is not a very precise estimate of time-t conditional volatility; (b) it includes no volatility information for the synthetic US notes, which constitute the other leg of the CIP trade; (c) and unlike the model, in which the assets used to construct A_t correspond to those used to measure σ_t^*, this proxy uses long-maturity notes instead of the 3-month bills comprising $\hat{A}_t^{(1/4)}$.

To help address concerns (a)-(c), we also consider three alternative measures of $\hat{\sigma}_t^*$.

(a) For a more real-time measure of conditional volatility, we also examine the CBOE's 10-year Treasury VIX (TYVIX), which is the implied 30-day volatility of CBOT futures on 10-year US Treasury Notes. The downside of asset-implied volatility is that it corresponds to risk-neutral volatility and may be a biased estimate of actual volatility.

(b) To include information for the foreign leg, we also compute a foreign volatility analogue to (23). In particular, we compute the volatility of the 10-year constant maturity foreign note's daily price changes, measured in USD (i.e., the bond prices are adjusted by the spot exchange rate each day). We then take a value-weighted average of the 10-year Constant Maturity US Treasury Note volatility and this foreign note volatility, which delivers a country-specific volatility measure. The downside of this construction is that it introduces additional variation by using the spot exchange rate to convert future prices to dollars, rather than the forward exchange rate as in the CIP trade.

(c) To bring the assets in the volatility construction as close as possible to those used in the arbitrage trade, we examine the value-weighted average return volatilities of the 3-month US Treasury bill and the 3-month synthetic US bill. The downsides of using this proxy are twofold. First, because we have no way of interpolating the forward exchange rate curve between the 3-month and 1-month forward rates, we construct 2-month holding period returns on both bills. The prices of these bills are constructed using country-specific IBOR. This is very long relative to the maturity of the bill, so our estimate of a conditional volatility is likely to be highly imprecise. Second, mainly due to their short durations, 3-month bill volatilities are mechanically much smaller than those of 10-year notes (approximately 40 times smaller, as their relative durations suggests). Whereas σ_t^* and $\sqrt{A_t}$ are on approximately the same scale in equation (19), which comes from a model with infinitely-lived assets, 3-month bill volatility is not on a scale comparable to the 3-month CIP deviation.

We repeat our analysis with these proxies. Aggregate time series plots associating volatility to arbitrage profits are below in Figures 5, 6, and 7. Disaggregated analysis at the currency level are below in Figures 8, 9, and 10.

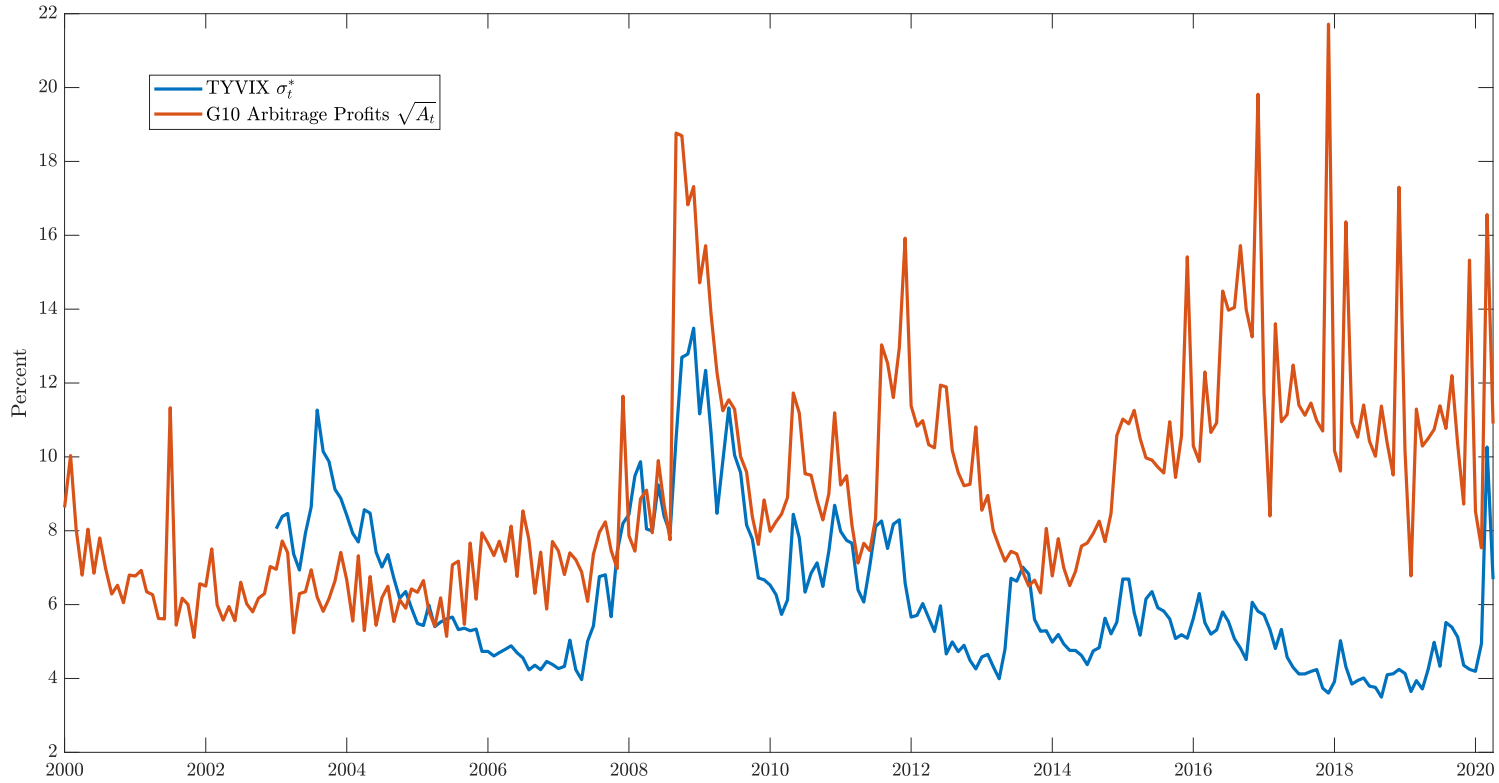

Figure 5: Monthly time-series of proxies for A_t and σ_t^*. Proxy for A_t is $\hat{A}_t^{(1/4)}$ from (22), which is constructed using 3-month absolute CIP deviations (against USD), measured daily, then averaged monthly, then (simple) averaged across the G10 currencies. Proxy for σ_t^* is constructed using the monthly average of TYVIX, which is the implied volatility of CBOE options on 10-year US Treasury Note futures. Both measures are annualized. Currency data are from Bloomberg and range from Jan. 2000 to Apr. 2020. TYVIX data are from the CBOE and range from Jan. 2003 to Apr. 2020.

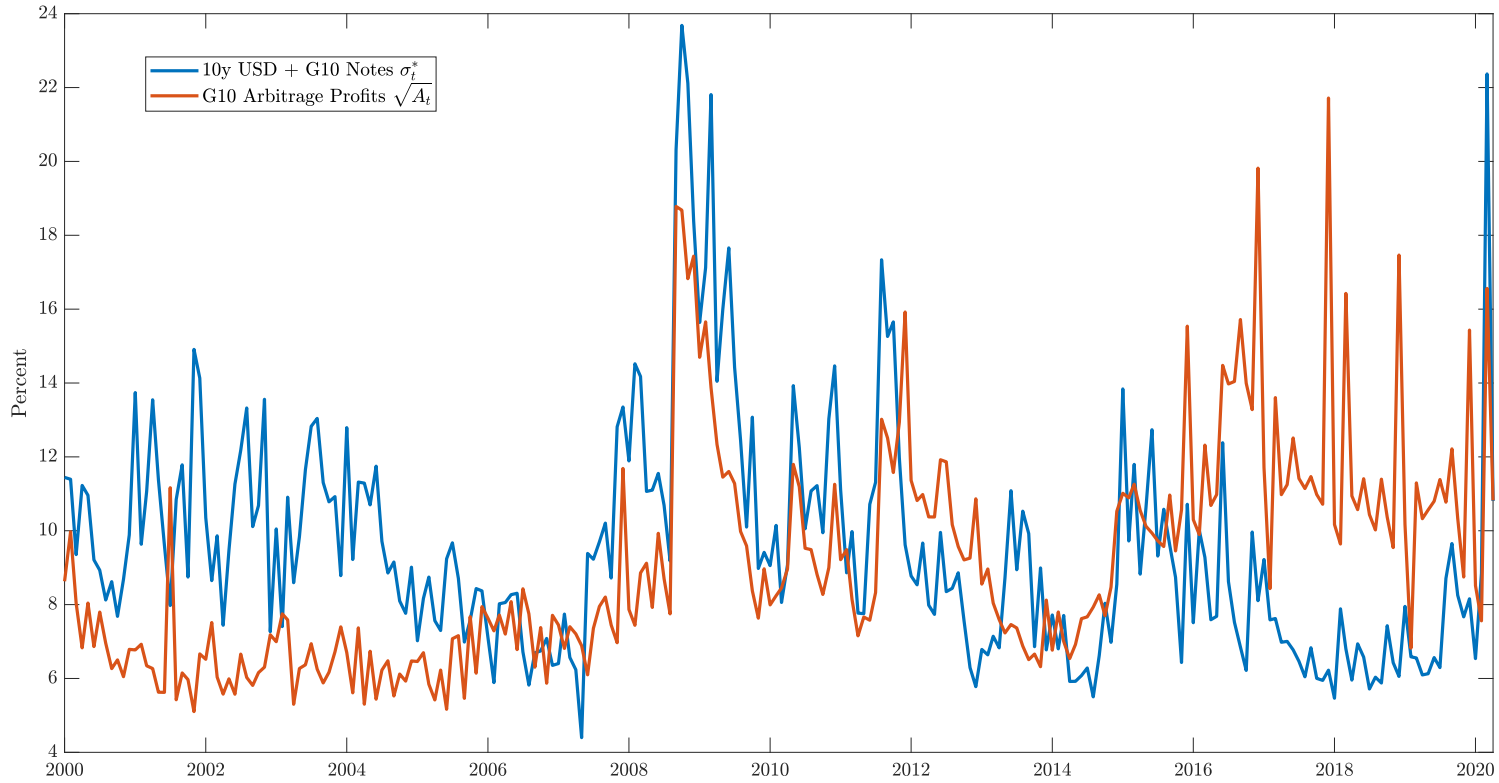

Figure 6: Monthly time-series of proxies for A_t and σ_t^*. Proxy for A_t is $\hat{A}_t^{(1/4)}$ from (22), which is constructed using 3-month absolute CIP deviations (against USD), measured daily, then averaged monthly, then (simple) averaged across the G10 currencies. Proxy for σ_t^* is constructed using the value-weighted average volatilities of daily log price changes on (i) the 10-year Constant Maturity US Treasury Note; and (ii) one of the G10's 10-year constant maturity note, adjusted by the respective spot exchange rate to be measured in USD. Volatility σ_t^* is calculated as a within-month standard deviation of these daily price changes, then (simple) averaged across the G10 currencies. Both measures are annualized. Currency and foreign notes data are from Bloomberg, whereas 10-year US Treasury Note data are from the Board of Governors of the Federal Reserve System, retrieved from FRED. Data range from Jan. 2000 to Apr. 2020.

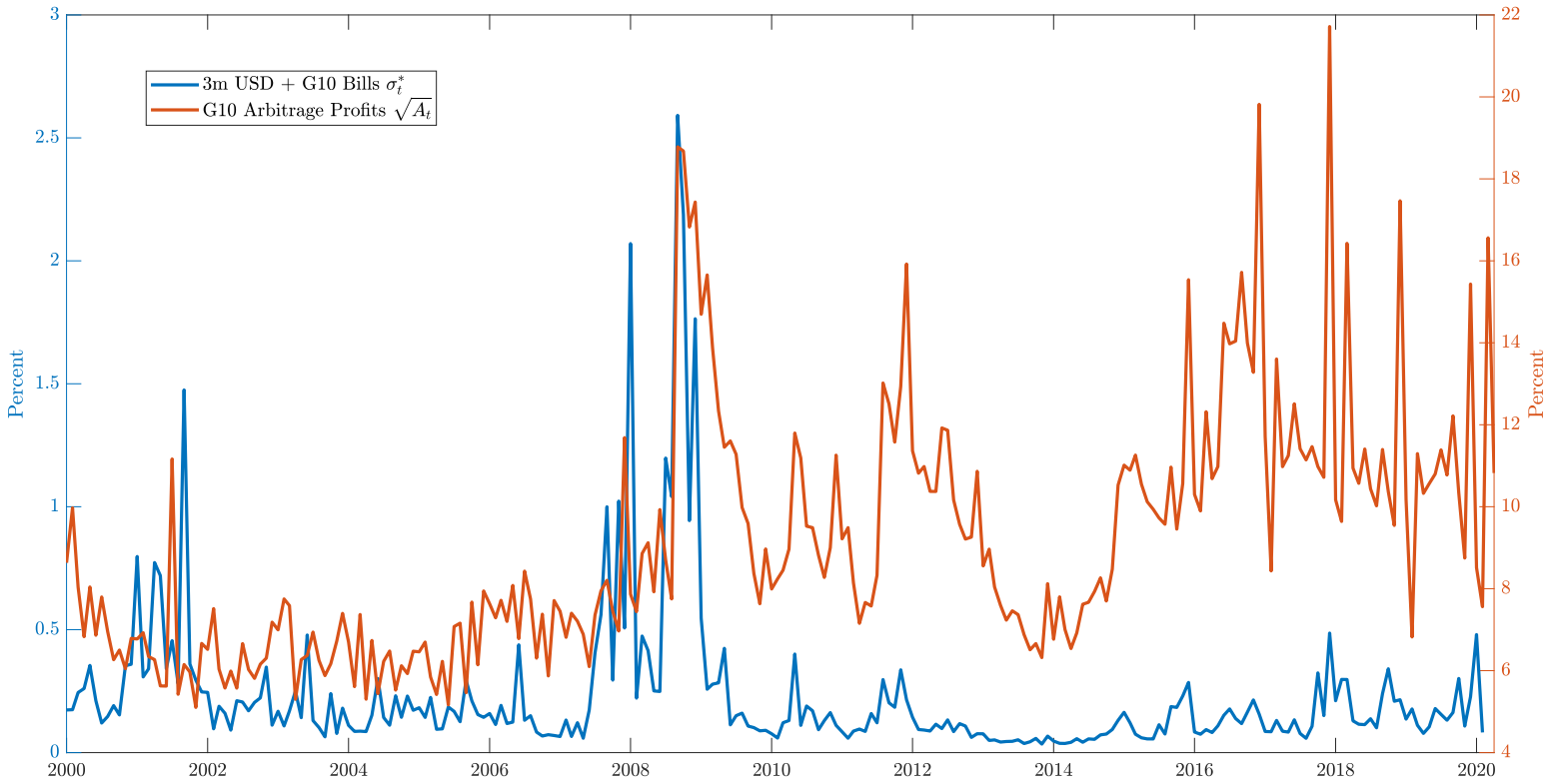

Figure 7: Monthly time-series of proxies for A_t and σ_t^*. Proxy for A_t is $\hat{A}_t^{(1/4)}$ from (22), which is constructed using 3-month absolute CIP deviations (against USD), measured daily, then averaged monthly, then (simple) averaged across the G10 currencies. Proxy for σ_t^* is constructed using the value-weighted average volatilities of next-2-month returns on (i) 3-month US bills; (ii) 3-month synthetic US bills, built using foreign bills, spot and forward exchange rates. The 2-month returns are measured at a daily frequency, then volatility is calculated as a within-month standard deviation of these daily measures, then these measures are (simple) averaged across the G10 currencies. Both measures are annualized. Currency, foreign bills, and US bills data are from Bloomberg. Data range from Jan. 2000 to Apr. 2020.

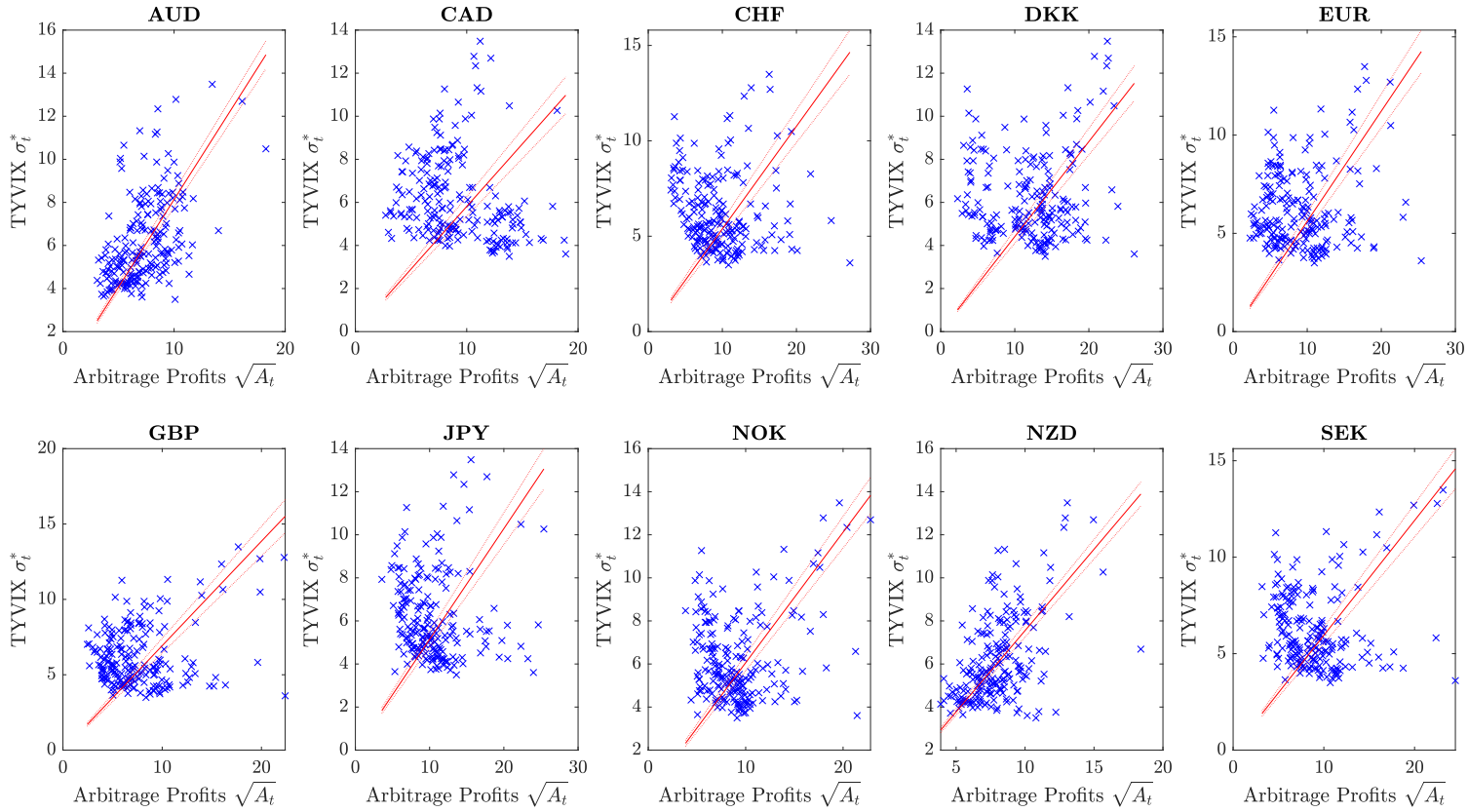

Figure 8: Currency-level OLS regressions of σ_t^* on $\sqrt{A_t}$ (monthly, no intercept). Proxy for A_t is $\hat{A}_t^{(1/4)}$ from (22), which is constructed using 3-month absolute CIP deviations (against USD), measured daily, then averaged monthly, then (simple) averaged across the G10 currencies. Proxy for σ_t^* is constructed using the monthly average of TYVIX, which is the implied volatility of CBOE options on 10-year US Treasury Note futures. Both measures are annualized. Regression lines, and 95% confidence intervals (using HAC standard errors), are also displayed. Currency data are from Bloomberg and TYVIX data are from the CBOE. Data range from Jan. 2003 to Apr. 2020.

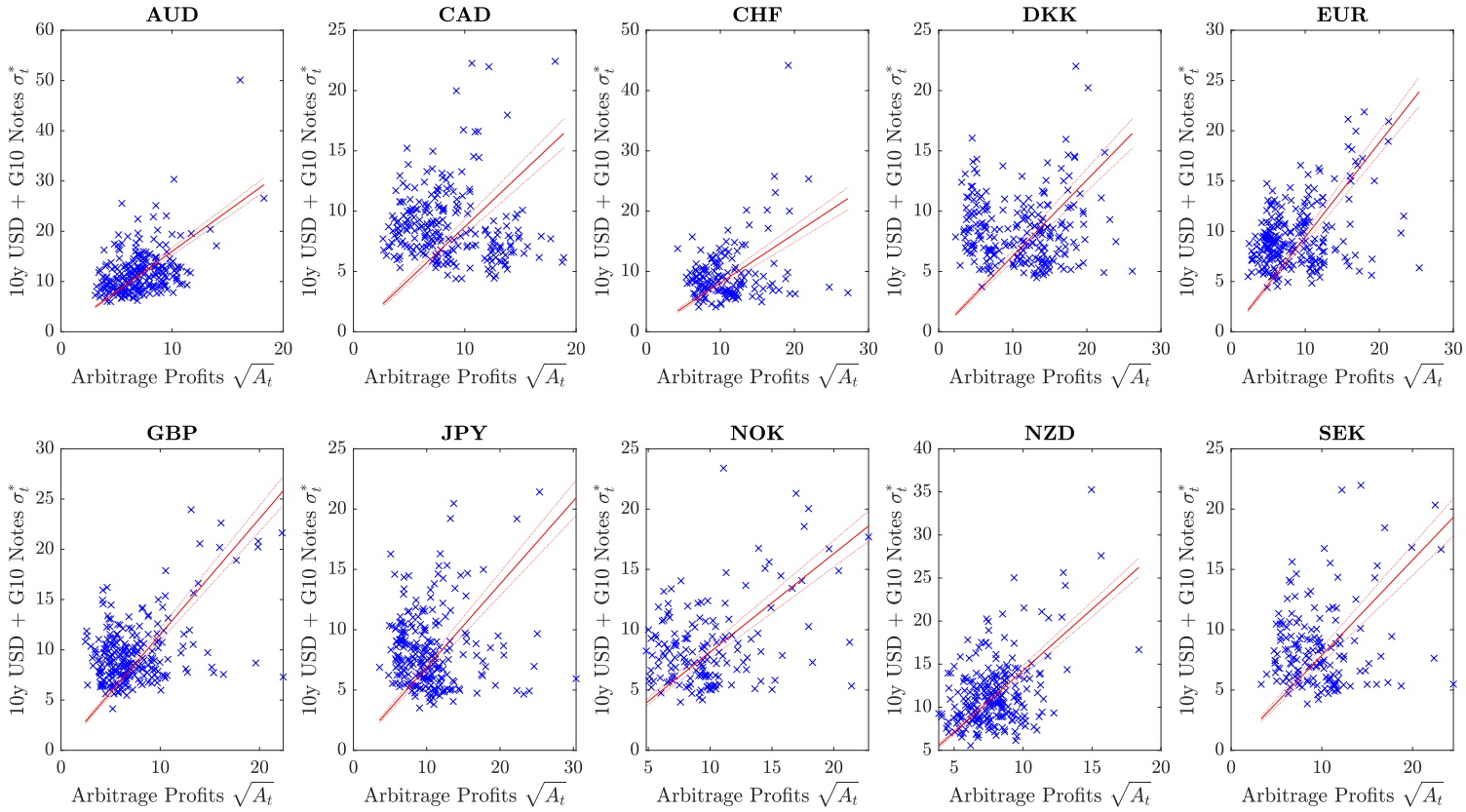

Figure 9: Currency-level OLS regressions of σ_t^* on $\sqrt{A_t}$ (monthly, no intercept). Proxy for A_t is $\hat{A}_t^{(1/4)}$ from (22), which is constructed using 3-month absolute CIP deviations (against USD), measured daily, then averaged monthly, then (simple) averaged across the G10 currencies. Proxy for σ_t^* is constructed using the value-weighted average volatilities of daily log price changes on (i) the 10-year Constant Maturity US Treasury Note; and (ii) one of the G10's 10-year constant maturity note, adjusted by the respective spot exchange rate to be measured in USD. Volatility σ_t^* is calculated as a within-month standard deviation of these daily price changes, then (simple) averaged across the G10 currencies. Both measures are annualized. Regression lines, and 95% confidence intervals (using HAC standard errors), are also displayed. Currency and foreign notes data are from Bloomberg, whereas 10-year US Treasury Note data are from the Board of Governors of the Federal Reserve System, retrieved from FRED. Data range from Jan. 2000 to Apr. 2020.

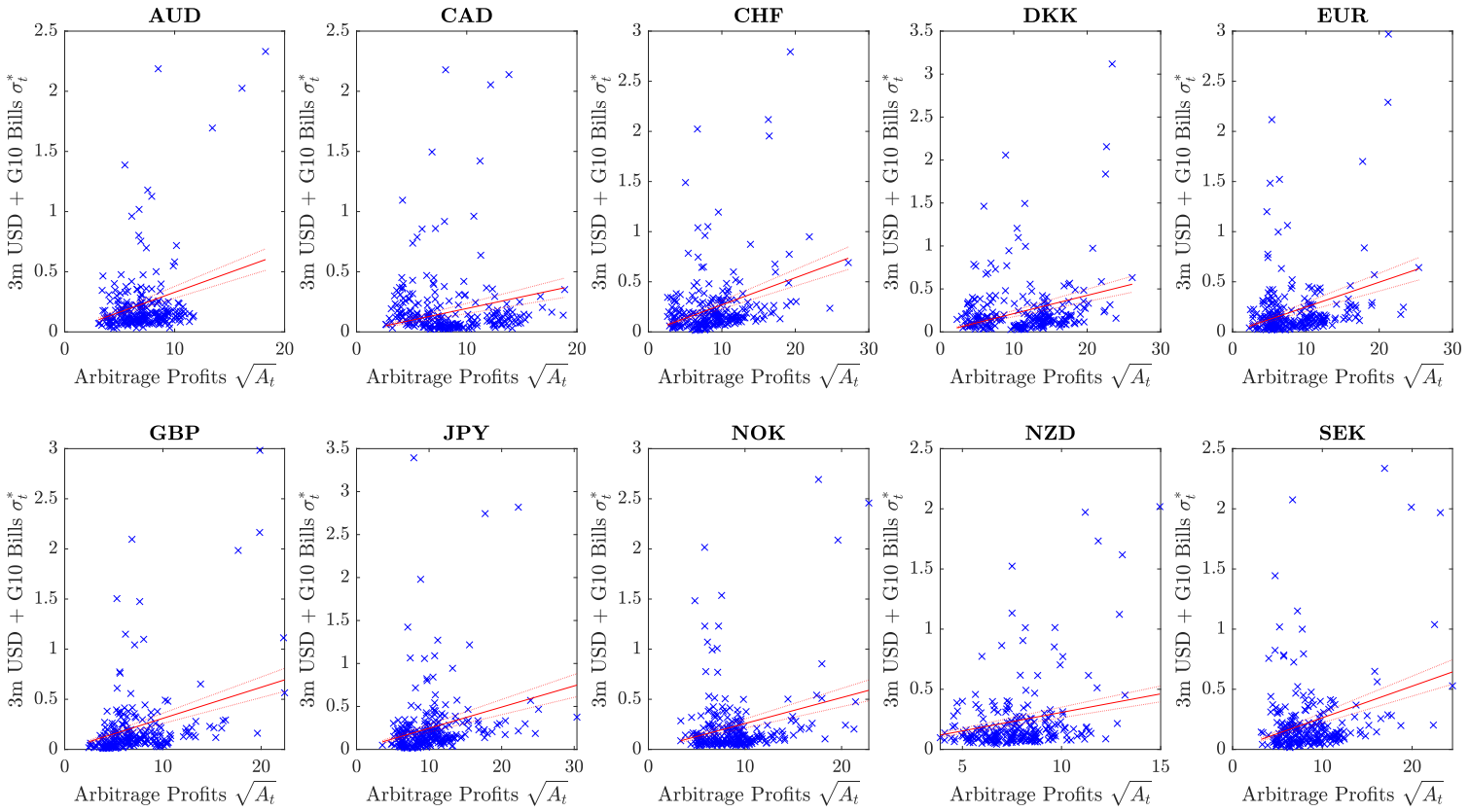

Figure 10: Currency-level OLS regressions of σ_t^* on $\sqrt{A_t}$ (monthly, no intercept). Proxy for A_t is $\hat{A}_t^{(1/4)}$ from (22), which is constructed using 3-month absolute CIP deviations (against USD), measured daily, then averaged monthly, then (simple) averaged across the G10 currencies. Proxy for σ_t^* is constructed using the value-weighted average volatilities of next-2-month returns on (i) 3-month US bills; (ii) 3-month synthetic US bills, built using foreign bills, spot and forward exchange rates. The 2-month returns are measured at a daily frequency, volatility is calculated as a within-month standard deviation of these daily measures, and then these measures are (simple) averaged across the G10 currencies. Both measures are annualized. Regression lines, and 95% confidence intervals (using HAC standard errors), are also displayed. Currency, foreign bills, and US bills data are from Bloomberg. Data range from Jan. 2000 to Apr. 2020.