A Competitive Search Theory of Asset Pricing*

Mahyar Kargar[†] Juan Passadore[‡] Dejanir Silva[§]

March 24, 2021

Abstract

We develop an asset-pricing model with heterogeneous investors and search frictions. Trade is intermediated by risk-neutral dealers subject to capacity constraints. Risk-averse investors can direct their search towards dealers based on price and execution speed. Order flows affect the risk premium, volatility, and equilibrium interest rate. We propose a new solution method to characterize the equilibrium analytically. We assess the quantitative implications of the model in response to a large adverse shock. Consistent with the empirical evidence from the COVID-19 crisis, we find an increase in the risk premium and market illiquidity, and a decline in interest rates.

Keywords: Asset pricing, competitive search, market liquidity, perturbation methods

JEL Classification: G11, G12, D53.

*We thank Adem Atmaz, Nicolas Caramp, Valentin Haddad, Tim Johnson, Moritz Lenel, Francis Longstaff, Stavros Panageas, Avanidhar Subrahmanyam, Pietro Veronesi, Pierre-Olivier Weill, and seminar participants at Einaudi, FGV, Purdue, UCLA Anderson, and the UC Davis Liquidity in Macroeconomics workshop for their helpful comments and suggestions. The views expressed herein are those of the authors and should not be attributed to the IMF, its Executive Board, or its management.

[†]University of Illinois at Urbana-Champaign, Department of Finance. Email: kargar@illinois.edu.

[‡]International Monetary Fund, Research Department. Email: jpassadore@imf.org.

[§]University of Illinois at Urbana-Champaign, Department of Finance. Email: dejanir@illinois.edu, corresponding author.

The Global Financial Crisis of 2007-2009 and the COVID-19 crisis in March 2020 underscore the importance of liquidity frictions in determining asset prices. The onset of these crises triggered an increase in risk premia and market volatility, as well as a flight to safety that depressed short-term interest rates. We also observe large portfolio reallocations, substantial increases in transaction costs and trading volume, and the deterioration of market liquidity. Understanding the joint dynamics of asset prices and liquidity conditions becomes even more relevant as the policy responses to these crises involved measures to reduce market illiquidity and absorb risky assets directly. In particular, the U.S. Federal Reserve's announcements of the unprecedented secondary market corporate bond-buying program in March and April 2020 emphasize supporting market liquidity as their primary goal.[1] The assessment of these policies requires a unified framework that captures how liquidity conditions and risk premia are jointly determined.

To capture the effects of trading frictions on asset prices and risk premia, one needs to depart in important ways from the workhorse models used to study market liquidity. The seminal work of Duffie, Gârleanu, and Pedersen (2005) on the search theory of over-the-counter (OTC) markets restricts investors' ability to adjust their portfolio to keep the problem tractable. Lagos and Rocheteau (2009) relax this restriction by introducing quasi-linear preferences. Despite successfully addressing key aspects of market liquidity, this approach essentially eliminates the effects of search frictions on risk premia. Introducing risk-averse investors to capture such effects makes the problem highly intractable, as this requires keeping track of the entire distribution of investors' asset holdings and wealth.

In this paper, we study a general equilibrium model with risk-averse agents, unrestricted asset holdings, and trading frictions in the secondary market. We make three main contributions to the literature. First, we provide a unified framework to study the joint determination of the risk premium, risk-free rates, and market liquidity in general equilibrium. Second, we propose a new solution method, which we call state-global perturbations, that allows us to analytically characterize the equilibrium even in the case of an infinite-dimensional state space. Third, we apply our model to quantitatively study the financial market response to the COVID-19 shock. We calibrate the model to match key asset pricing and secondary market moments from the corporate bond market. We find that search frictions are quantitatively important in determining the magnitude of the risk premium in the long run. Moreover, we show that in response to a large negative shock, the model can generate a substantial increase in bid-ask spreads, trading volume,

[1]See FAQs for the Secondary Market Corporate Credit Facility from the Federal Reserve Bank of New York: https://www.newyorkfed.org/markets/primary-and-secondary-market-faq/corporate-credit-facility-faq.

and the risk premium, as well as a decline in interest rates to nearly zero. These results are consistent with the empirical evidence from the height of COVID-19 crisis in March 2020.

We consider an endowment economy where risk-averse investors can trade a risk-free bond without frictions and a risky asset in an OTC market with search frictions. Investors differ only in their initial endowments of the risky asset and risk-free bond.[2] Bilateral trades are intermediated by risk-neutral dealers who hold no inventories and have access to a frictionless interdealer market but face capacity constraints. Dealers post contracts specifying the number of shares of the asset and an intermediation fee. Investors choose among these contracts in a competitive search market.[3] The speed at which an investor's order is executed depends on the number of contracts posted by dealers and the mass of investors sending orders for that given contract. Due to the presence of intermediation fees, investors effectively solve a portfolio problem with transaction costs.

Our framework captures three key features. First, unlike the standard portfolio choice problem without transaction costs, in our setting, an investor's risky asset holding is a state variable and not a control variable. Therefore, the joint distribution of investors' wealth and portfolio holdings affects the economy's aggregate behavior, including risk premia, interest rates, and transaction costs. Second, we show that contracts with higher intermediation fees attract more dealers, which leads to a higher trading speed for investors. Thus, we obtain a trade-off between trading speed and transaction costs. Finally, despite the presence of proportional transaction costs, there is no inaction region in the model, and investors trade continuously. This result is a consequence of the endogenous choice of trading speed, as investors may choose cheaper (and slower) trades when the gains from trade are relatively small.

We present two main results on the implications of search frictions for portfolio choice and asset pricing. First, we characterize investors' trading behavior. Whether an agent decides to buy or sell the risky asset depends on the marginal utility of holding an additional unit of the risky asset. We denote this key object by the *marginal value of portfolio rebalancing*. This object is analogous to the marginal utility of the asset in standard search models. We show that this marginal value of rebalancing is the present discounted value of the deviation of investors' actual portfolio share from the frictionless benchmark of

[2] The focus on the heterogeneity of initial endowments is only for expositional purposes, as our results extend to other types of heterogeneity. In particular, we consider a version of the model with heterogeneous risk aversion and Epstein-Zin preferences in our quantitative exercise.

[3] Although our results most likely extend to a case of random search, the competitive search setting has a few advantages. First, it provides seamless integration with standard portfolio theory, where prices are determined by competitive forces instead of bilateral bargaining. Second, it enables us to capture the trade-off between trading speed and trading cost, documented in various OTC markets.

Merton (1971). We thus provide a new microfoundation for the asset valuation function in search models. The marginal value of portfolio rebalancing pins down investors' trading behavior: trading speed and order size increase the further away investors' portfolios are from the target portfolio. Second, we provide a liquidity-adjusted consumption capital asset pricing model (CCAPM). In our model with search frictions, the risk premium depends on how the distribution of the marginal value of portfolio rebalancing covaries with consumption.

The asset pricing results above rely on the relations among endogenous variables. To further characterize the model's implications, we need to obtain an explicit solution. We can obtain analytical results using what we call state-global perturbation techniques. We consider a small-risk approximation of the equilibrium conditions, but in contrast to standard applications of perturbation methods, we do not assume that the economy is near a steady state.[4] Instead, our approach is global in the state space, which is crucial to capture the economy's behavior after large shocks. Moreover, our method handles large state spaces, even infinite-dimensional ones, enabling us to capture the rich investor heterogeneity that typically emerges with search frictions. This method also allows us to obtain closed-form asymptotic expressions for asset prices and trading behavior, despite the presence of trading frictions and time-varying investment opportunities.

Aided by the state-global perturbation techniques, we explore the asset pricing implications of trading frictions and portfolio flows. We show that both the risk premium and the interest rate depend on the asymmetry and dispersion of investors' portfolios. Portfolio dispersion measures how far investors are from their desired portfolio, while portfolio asymmetry captures the relative distance of sellers and buyers to their desired portfolios. We find that portfolio asymmetry amplifies the risk premium relative to a model with no frictions. In particular, if there is a net selling pressure such that sellers' portfolios are further away from the frictionless target than those of buyers', this portfolio asymmetry leads to a higher risk premium than the one in a frictionless benchmark. Moreover, we show that the distribution of investor portfolios matters for the determination of the interest rate. The dispersion in agents' portfolios amplifies the precautionary saving motive, leading to a flight to safety and a decline in the risk-free rate.

We also show how investor portfolios affect market liquidity. We find that higher portfolio dispersion induces a surge in demand for transaction services leading to higher trading volumes, transaction costs,

[4]For a discussion of perturbation methods where the model is linearized around the non-stochastic steady state, see Fernández-Villaverde, Rubio-Ramírez, and Schorfheide (2016).

and dealer profits. Higher dispersion has two impacts on trading speed: (i) investors want to trade faster the further away they are from the target portfolio and (ii) the increased trading costs make them shift to cheaper and slower trades. In equilibrium, due to dealer capacity constraints, the second force dominates, and higher dispersion endogenously leads to an increase in trading delays.

Next, we consider the response of the economy to a large negative aggregate shock. Our model can capture the joint behavior of market liquidity and risk premia observed during the COVID-19 crisis both qualitatively and quantitatively. We show that portfolio asymmetry and dispersion are countercyclical. The increase in asymmetry and dispersion leads to (i) a rise in risk premia and a decline in interest rates, (ii) an increase in trading volume, as investors have a stronger incentives to rebalance their portfolios, and (ii) a deterioration of market liquidity as measured by increased transaction costs and trading delays. All these outcomes are consistent with empirical evidence during the COVID-19 crisis, recently documented in Haddad, Moreira, and Muir (2020), Kargar et al. (2020), O'Hara and Zhou (2020), and others. In particular, Haddad, Moreira, and Muir (2020) emphasize that these patterns are hard to reconcile using existing frictionless models or even those with financial constraints.

To quantitatively assess the impact of large shocks on asset prices and liquidity conditions, we extend the model to incorporate heterogeneous risk aversions. Using secondary market transaction data from TRACE, we calibrate our model to match moments from the corporate bond market before the onslaught of the COVID-19 crisis, as well as standard asset pricing moments. We consider an adverse shock that generates the response of bid-ask spreads consistent with the one observed during the crisis, a 10-fold increase. We find that the model endogenously generates an increase in the risk premium of 25%, interest rates going to zero, and a 20% increase in trading volume. This is roughly consistent with the empirical evidence during this period mentioned above.

Related literature

This paper connects to different strands of the literature. First, our paper is closely related to the literature on search in asset markets. The seminal contribution to this body of work is Duffie, Gârleanu, and Pedersen (2005), DGP hereafter, who introduce search frictions in an asset pricing model where investors and dealers meet randomly over time. In DGP, investors' valuations change exogenously when they receive (uninsurable) idiosyncratic liquidity shocks, and investors can hold at most one unit of the asset. Lagos and Rocheteau (2009) extend this setup by allowing agents to hold arbitrary asset positions. To keep the model tractable, they assume that the utility function is quasi-linear in consumption and that the asset's valuation is exogenous. These assumptions imply, however, that the aggregate risk premium is zero.[5] These papers explain important aspects of market liquidity, such as bid-ask spreads and trading volumes. We contribute to this literature by endogeneizing the asset's valuation and providing a joint theory of the aggregate risk premium and the liquidity risk premium.

Second, our paper is also related to an extensive literature on portfolio choice models with frictions (e.g., Constantinides, 1986; Davis and Norman, 1990; Dumas and Luciano, 1991; Buss and Dumas, 2019). Due to proportional transaction costs, this class of models features an inaction region. Our setup makes two contributions to this literature. First, we show that proportional transaction costs do not lead to a no-trade region in a setup with competitive search because the trading cost depends on the endogenous choice of trading speed. Second, while most of these studies focus on asset prices as primitives of the model, we derive the implications of search frictions for risk premia and interest rates in general equilibrium.

An alternative body of work studies portfolio choice under quadratic transaction costs (e.g., Heaton and Lucas, 1996; Gârleanu and Pedersen, 2013). Our model is related to the latter, who find that the optimal portfolio strategy for an investor with quadratic utility is to aim in front of the Markowitz portfolio target. In our model with search frictions and constant relative risk aversion (CRRA) preferences, we find a similar result where investors should adjust their portfolio towards a target, corresponding to the optimal Merton (1971) portfolio. We provide a liquidity-adjusted CCAPM in the spirit of Acharya and Pedersen (2005). We find that asset returns depend not only on their covariance with the aggregate endowment but also on the marginal value of portfolio rebalancing, which is the key driver of trading behavior in our setup. Our framework also provides a microfoundation for bounded variation and trading delays whose asset pricing implications are studied in Longstaff (2001) and Longstaff (2009), respectively.[6]

Third, our paper contributes to the large literature on market microstructure. These studies rationalize the existence of bid-ask spreads due to risk aversion of dealers (e.g., Stoll, 1978) and adverse selection (e.g., Grossman and Miller, 1988). Kyle (1985) studies the impact of order flow on asset prices. In our paper, the order flow also impacts asset prices. However, we focus on the risk premium response to portfolio flows, while Kyle (1985) abstracts from risk premia effects.

[5]One notable exception is Gârleanu (2009) who studies portfolio choice in an economy with CARA preferences and random search. This approach eliminates wealth effects and the portfolio rebalancing channel that are crucial in our setting.

[6]Brunnermeier and Pedersen (2009) also study the link between asset prices and market liquidity, but focus on dealers' funding constraints instead of search frictions.

Finally, our paper relates to the extensive literature on the asset pricing implications of investor heterogeneity (e.g., Dumas, 1989; Longstaff and Wang, 2012; Gârleanu and Panageas, 2015; Alvarez and Atkeson, 2018). We introduce search frictions in a standard asset pricing model with heterogeneous agents (e.g., Gârleanu and Panageas, 2015). In these studies, the wealth distribution among agents is the main state variable. We show that once we consider secondary market frictions, the distribution of asset holdings is an additional state variable that pins down asset prices. Thus, our paper is also related to studies that emphasize the impact of portfolio flows on the risk premium (e.g., Koijen and Yogo, 2019; Gabaix and Koijen, 2020).

1 Motivating Evidence

In this section, we provide some empirical evidence to motivate our theoretical framework. In particular, we characterize the joint response of asset prices, market liquidity, and portfolio flows during the COVID-19 crisis.

1.1 Asset pricing

As shown in the top two panels of Figure 1, we witnessed a large decline in stocks and corporate bond returns in the U.S. during the most tumultuous period of the crisis in mid-March 2020. During this period, there was also an intense flight-to-safety episode, where the short-term interest rate went to nearly zero. The S&P 500 return and 1-year Treasury rate declined by over 35% and 92%, respectively, while the AAA (high-yield) credit spread increased by a factor of five (three).

1.2 Liquidity

These large movements in asset prices were accompanied by an increase in trading volume and a deterioration of liquidity conditions. Using transaction-level data from the corporate bond market, as shown in the bottom-left panel of Figure 1, we see that the trading volume for customer-to-dealer trades increased by approximately 40% relative to pre-pandemic levels. At the same time, inter-dealer trade volume remained almost unchanged.[7] In the bottom-right panel, we estimate that transaction costs for customer-to-dealer trades increased dramatically in mid-March 2020. Only after unprecedented interventions from the Fed-

[7]For a detailed description of the TRACE data and how we construct our sample, see Appendix OA.2.1.

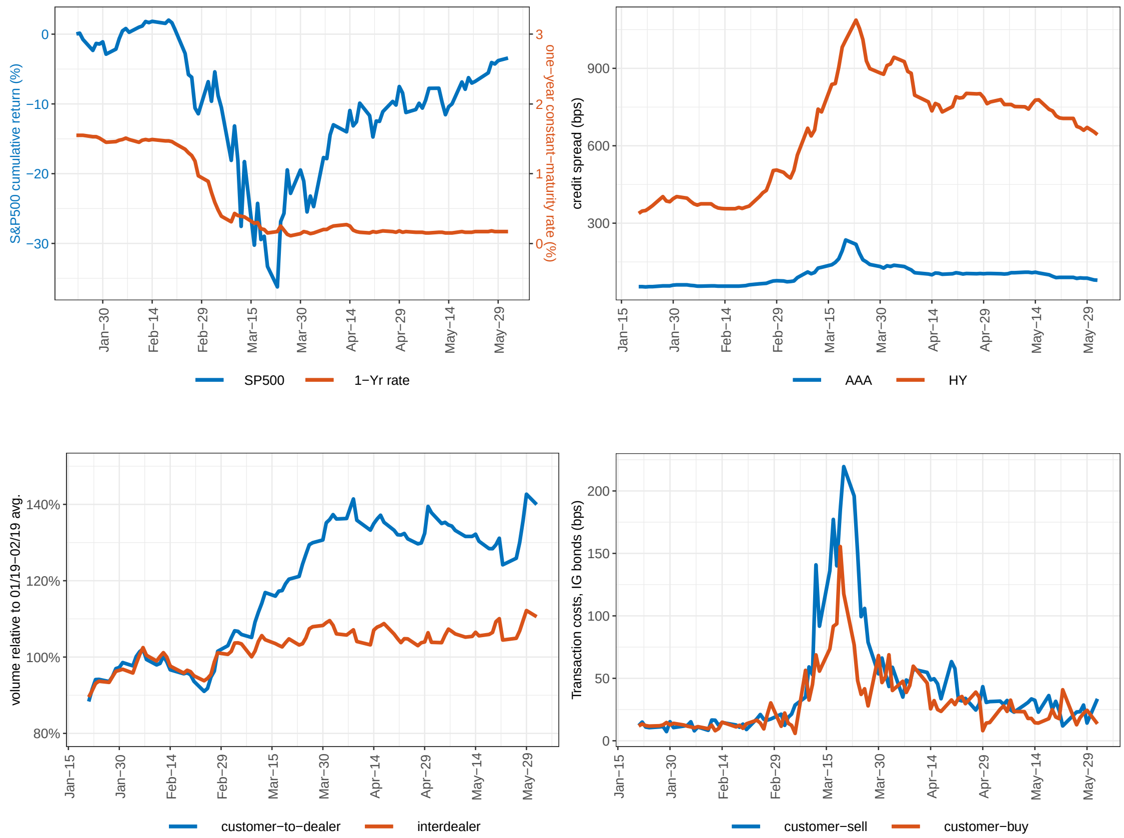

Figure 1. Stock returns, interest rates, credit spreads, trading volume, and transaction costs for corporate bonds during the onset of the COVID-19 pandemic. Source: TRACE, Bloomberg, and FRED.

eral Reserve through the primary and secondary market corporate credit facilities did these patterns start to reverse. However, as of the end of 2020, they have not yet reached their pre-crisis levels. Investors responded to the increase in transaction costs by changing their trading speed, as documented in Kargar et al. (2020). When trading costs for fast trades, where dealers hold corporate bonds in their inventories, rose dramatically in mid-March 2020, customers substituted to slower so-called agency trades where dealers act as matchmakers. Therefore, we witnessed a deterioration of market liquidity along two dimensions, as transaction costs and trading delays both increased in mid-March 2020.

1.3 Portfolio reallocation

There have been massive outflows from fixed income mutual funds and ETFs during the onset of the COVID-19 pandemic. Falato, Goldstein, and Hortaçsu (2020) document that in March 2020, corporate bond funds and ETFs experienced aggregate net outflows of over 4% relative to net asset values, far greater than in previous stress episodes including October 2008. Similarly, as shown in Ma, Xiao, and Zeng (2020), in March 2020, bond mutual funds lost approximately 12% of their assets under management. In addition, as documented in Duffie (2020), in March 2020, foreign investors sold, on net, approximately $300 billion of Treasury bonds and notes, far above normal levels. The intense selling pressure was reflected in the relative transaction costs for buyers and sellers. For corporate bonds, we find that trades where customers sell to dealers have significantly higher transaction costs than customer-buy trades. The bottom-right panel of Figure 1 shows that this difference is approximately 40 bps (95 bps for customer-sell and 55 bps for customer-buy trades) in March 2020 and is much larger during the peak of the crisis in mid-March 2020.

The evidence in Figure 1 indicates substantial movements in asset prices, liquidity conditions, and portfolio flows in mid-March 2020.[8] Building on this evidence, in the next section, we study a model with two key features: competitive search and risk-averse agents. Competitive search allows us to capture the two dimensions of liquidity: transaction costs and trading delays. Having risk-averse agents enables us to study the impact of portfolio flows on risk premia and interest rates.

2 Model

In this section, we present a general equilibrium asset pricing model with search frictions. The model captures the main features of directed search models of OTC markets, as in, for example, Lester, Rocheteau, and Weill (2015), in an environment with risk-averse agents.[9]

2.1 Environment

Time is continuous, $t \in [0, \infty)$. The economy is populated by a continuum of investors and dealers, each with a unit mass. The economy's aggregate endowment follows a geometric Brownian motion:

$$\frac{dY_t}{Y_t} = \mu dt + \sigma dZ_t, \tag{1}$$

where μ and σ are constants and Z_t is a standard Brownian motion defined on a fixed probability space $(\Omega, \mathcal{F}, P)$ and a filtration $\{\mathcal{F}_t, t \geq 0\}$ of sub-σ-algebras of $\mathcal{F}$ satisfying the usual conditions, as defined by

[8]The evidence provided above is not unique to the COVID-19 pandemic. Similar patterns emerged during the 2007-2009 Global Financial Crisis (GFC), as shown in Figure OA.1 in Appendix OA.2.2.

[9]See Wright et al. (2019) for a review of the literature on competitive (or directed) search.

Protter (2004). Investors have access to two assets: a risk-free bond and a risky asset. The risk-free bond market is frictionless so that investors can adjust the amount invested in the riskless asset instantaneously. The risky asset is a claim on the aggregate endowment, traded on a market subject to search frictions.

2.1.1 Dealers and competitive search

We assume that trades on the risky asset are bilateral and intermediated by dealers, that is, investors must buy or sell through dealers. Dealers have continuous access to a frictionless inter-dealer market, where the risky asset trades at price p_t, which evolves according to:

$$\frac{dp_t}{p_t} = \mu_{p,t}dt + \sigma_{p,t}dZ_t,$$

where $\mu_{p,t}$ and $\sigma_{p,t}$ are determined in equilibrium.[10]

Dealers post contracts $\varsigma = (n, \phi) \in \Sigma$ specifying the number of shares $n \in \mathbb{R}$ they sell to investors and the intermediation fee $\phi \in \mathbb{R}_+$ investors must pay to dealers. Dealers hold no inventory. If $n > 0$, the dealer sells to the investor n units of the asset, which are immediately acquired from the inter-dealer market. If $n < 0$, the dealer buys $|n|$ units from the investor, which are immediately sold at the inter-dealer market. The fees determine the effective price faced by investors. Investors ultimately pay $p_t + \phi$ when buying the asset, and they receive the amount $p_t - \phi$ when selling it.

Dealers post a quantity $d_t(n, \phi)$ of the contract $\varsigma = (n, \phi)$. Investors choose which contract to submit an order to. The total mass of investors submitting orders to the contract (n, ϕ) is denoted by $\iota_t(n, \phi)$. We assume that the total number of orders executed at a given moment in time is determined by a constant-returns-to-scale matching function $m(\iota, d)$. This implies that the order of any individual investor is executed at Poisson arrival times with intensity $\alpha(\theta_t(n, \phi)) \equiv m(1, \theta_t(n, \phi))$, where $\theta_t(n, \phi) \equiv d_t(n, \phi)/\iota_t(n, \phi)$ denotes the dealer-to-investor ratio or market tightness. If $\iota_t(n, \phi) = 0$, we assume that $\theta_t(n, \phi) = \infty$. Analogously, a contract (n, ϕ) posted by a dealer is executed at Poisson arrival times with intensity $\alpha(\theta_t(n, \phi))/\theta_t(n, \phi)$. The arrival rate $\alpha(\cdot)$ is given by:

$$\alpha(\theta) = \overline{\alpha}\frac{\theta^\eta}{\eta}, \tag{2}$$

[10] The assumption of a frictionless inter-dealer market follows Duffie, Gârleanu, and Pedersen (2005) and it is standard in the OTC market literature.

where $\overline{\alpha}$ and η control, respectively, the efficiency and concavity of the matching function.

Dealers are risk-neutral and choose $d_t(n, \phi)$ to maximize expected profits,

$$\Pi_{d,t} = \max_{\{d_t(n,\phi)\}_{(n,\phi)\in\Sigma}} \int_{\Sigma} d_t(n, \phi) \frac{\alpha(\theta_t(n, \phi))}{\theta_t(n, \phi)} |n| \phi d\varsigma, \tag{3}$$

subject to a non-negativity constraint $d_t(n, \phi) \geq 0$ and a capacity constraint,

$$\int_{\Sigma} d_t(n, \phi)|n| d\varsigma \leq \overline{d}, \tag{4}$$

where the parameter $\overline{d}$ determines dealers' intermediation capacity. This constraint can be motivated by dealers' costs of posting contracts, where $\overline{d}$ represents the total budget available to dealers to cover such costs. This feature intends to capture the short-run behavior of the supply of intermediation services in the secondary market.

2.1.2 Investors

There is a continuum of investors indexed by $i \in [0, 1]$. Investor i maximizes her utility by choosing consumption $C_{i,t}$ and which contract to send the order $\varsigma_{i,t} = (n_{i,t}, \phi_{i,t})$, given her initial wealth $W_{i,0}$ and the initial number of shares of the risky asset, $S_{i,0}$. Wealth is computed as the value of the riskless bonds held by investor i and the value of the shares evaluated at the inter-dealer price p_t. Investors differ only on their initial endowments of the risky and riskless assets, and this heterogeneity in endowments provides the motivation for trade in this economy.

The investor's problem is given by:

$$V(W_i, S_i, X) = \max_{[C_{i,t}, n_{i,t}, \phi_{i,t}]} \mathbb{E}_0 \left[\int_0^{\infty} e^{-\rho t} \frac{C_{i,t}^{1-\gamma}}{1-\gamma} dt \right], \tag{5}$$

subject to:[11]

$$dW_{i,t} = \left[r_t W_{i,t} + \pi_t p_t S_{i,t} - \frac{1}{2} p_t \chi n_{i,t}^2 - C_{i,t} \right] dt + \sigma_{R,t} p_t S_{i,t} dZ_t - \phi_{i,t} |dS_{i,t}| \tag{6}$$

$$dS_{i,t} = n_{i,t} dN_{i,t}, \tag{7}$$

[11] See Appendix OA.1.1 for a derivation of the investor's budget constraint in the presence of transaction costs.

and a lower bound on wealth to prevent Ponzi schemes, where $N_{i,t}$ is a Poisson process with arrival intensity $\alpha(\theta_t(n_{i,t}, \phi_{i,t}))$.

Investors take as given the process for the interest rate r_t and the risk premium $\pi_t = \mu_{R,t} - r_t$, where the expected return on the risky asset is $\mu_{R,t} = \mu_{p,t} + Y_t/p_t$, and the volatility is $\sigma_{R,t} = \sigma_{p,t}$. Note that returns are computed using the inter-dealer price p_t, so the trading fee $\phi_{i,t}$ is subtracted from investor's wealth when a trade is realized. Investors face a quadratic portfolio adjustment cost $0.5\chi n_{i,t}^2$. This cost captures any cognitive or physical costs investors face in adjusting their portfolio.[12] Note that, in contrast to a standard portfolio problem with transaction costs, the number of shares invested in the risky asset $S_{i,t}$ is a state variable instead of a control variable. The number of shares evolves with the number of orders submitted, $n_{i,t}$, and whether this order is executed, which is determined by the Poisson process, $N_{i,t}$. Therefore, investors face an idiosyncratic order execution risk, as whether their order will be executed depends on the realization of the random variable $N_{i,t}$.

Finally, in a Markov equilibrium, asset prices are a function of the (K-dimensional) aggregate state variable $X_t \in \mathbb{R}^K$, that is, prices satisfy $r_t = r(X_t)$ and $p_t = p(X_t)$. The aggregate state variable follows the stochastic process:

$$dX_t = \mu_{X,t}dt + \sigma_{X,t}dZ_t,$$

where $\mu_{X,t}$ and $\sigma_{X,t}$ are determined in equilibrium.

2.1.3 Market tightness

The market tightness function $\theta_t(n, \phi)$ must be defined for every contract, even for the ones not active in equilibrium. We follow Lester, Rocheteau, and Weill (2015) and impose the following restriction on $\theta_t(n, \phi)$. Let $\theta_t(0, 0) = 0$ and, for $\varsigma \neq (0, 0)$, we assume that

$$\theta_t(\varsigma) = \inf\left\{\theta \geq 0 \; : \; V_t(W_i, S_i|\varsigma, \theta) > V_t(W_i, S_i), \text{ for some investor } i \text{ with state } (W_i, S_i)\right\}, \tag{8}$$

and $\theta_t(\varsigma) = \infty$ if this set is empty, where $V_t(W_i, S_i|\varsigma, \theta)$ is the value for an investor constrained to choose $(n_{i,t}, \phi_{i,t}) = \varsigma$ at date t, given an arrival rate of $\alpha(\theta)$. This assumption captures the idea that if the market tightness is larger than what is given by Equation (8), then investors would send orders to that contract, reducing the market tightness until it coincides with $\theta_t(\varsigma)$ as given above.

[12] See, for example, Heaton and Lucas (1996) and Gârleanu and Pedersen (2013) for models with quadratic transaction costs.

2.1.4 Competitive Search Equilibrium

We provide the definition of equilibrium below.

Definition 1. *A competitive search equilibrium is a list of stochastic processes adapted to the filtration generated by Z_t: the aggregate endowment Y, the price of the claim on aggregate endowment p, the risk-free interest rate r, and the market tightness function $\{\theta_t(n, \phi)\}_{(n,\phi)\in\Sigma}$; the mass of contracts posted by dealers $\{d_t(n, \phi)\}_{(n,\phi)\in\Sigma}$; and a set of stochastic process for each investor $i \in [0, 1]$: wealth W_i, asset holdings S_i, consumption C_i, and a contract (n_i, ϕ_i); such that:*

(i) Aggregate endowment evolves according to (1), given $Y_0 > 0$.

(ii) Given the stochastic process for the market tightness function $\{\theta_t(n, \phi)\}_{(n,\phi)\in\Sigma}$, the mass of contracts $\{d_t(n, \phi)\}_{(n,\phi)\in\Sigma}$ solves the dealer's optimization problem (3).

(iii) Given the stochastic processes for (p_t, r_t) and the market tightness function $\{\theta_t(n, \phi)\}_{(n,\phi)\in\Sigma}$, choices $(C_{i,t}, n_{i,t}, \phi_{i,t})$ solve investor i's optimization problem (5).

(iv) The market tightness function $\{\theta_t(n, \phi)\}_{(n,\phi)\in\Sigma}$ satisfies condition (8).

(v) Markets for consumption goods, risky asset, and risk-free bonds clear

$$\int_0^1 \left(C_{i,t} + 0.5\chi p_t n_{i,t}^2\right) di + \Pi_{d,t} = Y_t, \qquad \int_0^1 S_{i,t} di = 1, \qquad \int_0^1 W_{i,t} di = p_t.$$

2.2 Equilibrium characterization

We next provide a characterization of the equilibrium. We start by considering the dealers' problem and how their behavior leads to a trade-off between execution speed and trading costs. We characterize the investors' marginal valuation of the risky asset and show how this marginal value shapes their trading behavior. We then show how the distribution of investors' marginal valuation pins down equilibrium trading costs, volume, and dealers' compensation.

2.2.1 Dealers' problem

The first-order condition for the dealers' problem is:

$$\frac{\alpha(\theta_t(n, \phi))}{\theta_t(n, \phi)} \frac{\phi}{p_t} \leq v_{d,t}, \tag{9}$$

with equality if $d_t(n, \phi) > 0$, where $p_t v_{d,t}$ is the Lagrange multiplier on the capacity constraint in (4). Notice that the dealers' profits are given by $p_t v_{d,t} \overline{d}$, so $v_{d,t}$ captures the profitability of dealers, which is important when determining intermediation costs in equilibrium.

An important implication of the dealer's optimal choice is that investors will face a trade-off between trading speed and trading costs. In particular, rearranging (9) for an active contract, we obtain:

$$\frac{\phi}{p_t} = \frac{\theta_t(n, \phi)}{\alpha(\theta_t(n, \phi))} v_{d,t}. \tag{10}$$

The intermediation fee as a fraction of the inter-dealer price, ϕ/p_t, is increasing in the market tightness, θ, and the dealers' value $v_{d,t}$. The positive relation between the intermediation fee and market tightness reflects the incentive of dealers to post contracts. Contracts with higher intermediation fees attract more dealers, which increases market tightness $\theta_t(n, \phi)$, making it easier for investors to find a counterparty.[13]

2.2.2 Investors' problem

For the characterization of the investors' problem, it is convenient to have them directly choose the market tightness θ, as the trading fee ϕ is determined by condition (10). Given this change of variables and conditions (6), (7), and (9), the Hamilton-Jacobi-Bellman (HJB) equation for the investor can be written as:

$$\begin{aligned}\rho V = \max_{C,n,\theta} \frac{C^{1-\gamma}}{1-\gamma} + V_W\left[rW + \pi pS - \frac{1}{2}p\chi n^2 - C\right] + V_X \mu_{X,t} + \frac{1}{2}V_{WW}\sigma_R^2(pS)^2 + \frac{1}{2}\sigma_X' V_{XX}\sigma_X \\ + pS\sigma_R V_{WX}\sigma_{X,t} + \left[V\left(W - \frac{\theta v_d}{\alpha(\theta)}p|n|, S+n, X\right) - V(W,S,X)\right]\alpha(\theta). \end{aligned} \tag{11}$$

This HJB equation is reminiscent of the one for portfolio problems with time-varying investment opportunities, where aggregate conditions are described by X_t. The last term captures the impact of search frictions. When an investor meets a dealer, which happens with intensity $\alpha(\theta)$, the number of shares changes by n and wealth is reduced by the trading costs, $\frac{\theta v_d}{\alpha(\theta)}p|n|$.

The problem in (11) encompasses two important benchmarks. In the absence of transaction costs, it corresponds to a standard Merton portfolio problem (see Merton, 1971). In the case of quasi-linear value function $V(W,S,X) = W + v(S)$, the HJB equation would correspond to a version of the competitive

[13] This trade-off between trading speed and trading costs has been empirically documented, for example, in Hendershott and Madhavan (2015) and Li and Schürhoff (2019).

search model proposed by Lester, Rocheteau, and Weill (2015).[14] Importantly, the (indirect) utility that the investor derives from the asset will not be separable from the level of wealth with CRRA preferences, so it will be necessary to keep track of the joint distribution of wealth and risky asset shares to derive the economy's aggregate behavior. Assumption 1 below enables us to focus on the simpler case of only two types.[15] We relax this assumption in Section 5 and show how the case with order execution risk can be analyzed using perturbation methods.

Assumption 1 (Big-family assumption)**.** *Investors belong to two families (types): If $i \leq \nu$, then the investor belongs to family 1 and, if $i > \nu$, the investor belongs to family* 2*. Investors pool their resources inside each family and they perfectly diversify the (idiosyncratic) order execution risk.*

Assumption 1 implies that investors can perfectly diversify their order execution risk, such that a mass $\alpha(\theta)$ of orders are executed every instant. This assumption simplifies the exposition, as it allow us to focus on the two-type case, but it is not necessary for our main results.

The HJB equation under Assumption 1 can be written as:

$$\rho V = \max_{C,n,\theta} \frac{C^{1-\gamma}}{1-\gamma} + V_W \left[rW + \pi pS - \frac{1}{2} p\chi n^2 - C \right] + V_X \mu_X + \frac{1}{2} V_{WW} \sigma_R^2 (pS)^2 + \frac{1}{2} \sigma_X' V_{XX} \sigma_X$$
$$+ pS\sigma_R V_{WX} \sigma_X - V_W \theta v_d p|n| + V_S n\alpha(\theta). \tag{12}$$

Conditions (11) and (12) differ only on the last term, where the difference of value functions in (11) is replaced by the corresponding derivatives in (12). Note that the quadratic adjustment cost plays an important role in the HJB eqution (12), as it guarantees that there is an interior solution to the order size n. In particular, the quadratic adjustment cost plays a role similar to the concavity of the valuation function in standard search models.[16]

The investors' trading behavior depends to a great extent on the marginal value of portfolio rebalancing, Ω, defined as:

$$\Omega(W, S, X) \equiv \frac{V_S(W, S, X)}{V_W(W, S, X)}. \tag{13}$$

[14] The value function would be quasi-linear if, for instance, $\gamma = 0$ and the investor derived some utility flow from holding the asset, similar to money-in-the-utility models.

[15] This assumption was originally introduced by Lucas (1990) in his study of the liquidity effect and it is often adopted in macro-finance models (e.g., Gertler and Kiyotaki, 2010).

[16] For instance, assuming the quasi-linear value function $V(W, S, X) = W - 0.5\chi p \left(S - \overline{S}\right)^2$, the problem with order execution risk would be analogous to problem (12), as the expression $V(W - \phi|n|, S + n, X) - V(W, S, X) = -\phi|n| + V_S n - 0.5\chi pn^2$ would also have a quadratic term in n.

It measures the marginal utility of adjusting the portfolio by one unit, measured in units of wealth. Note that Ω can be positive or negative. In the benchmark case in which there are no frictions, the initial composition of an investors' wealth is not relevant, so $\Omega(W, S, X) = 0$. If the value function is quasi-linear in wealth, $V(W, S, X) = W + v(S)$, then the marginal value of portfolio rebalancing is equal to the marginal utility of holding the asset $v'(S)$. In the case with CRRA preferences, the marginal value of rebalancing depends on both W and S.

The optimatility condition for consumption is given by the standard expression:

$$C^{-\gamma} = V_W. \tag{14}$$

The first-order conditions for the order size and market tightness are given by:

$$n = \frac{1}{\chi}\left[\alpha(\theta)\frac{\Omega(W,S,X)}{p} - \text{sg}\,(n)\, v_d \theta\right], \qquad \alpha'(\theta)\frac{|\Omega(W,S,X)|}{p} = v_d, \tag{15}$$

where $\text{sg}\,(n_{i,t})$ is the sign function. The first equation in (15) corresponds to the optimality condition for the order size. If the benefit of increasing the number of shares, $\alpha(\theta)\frac{\Omega}{p}$, exceeds the expected transaction cost, $v_d\theta$, then the investor will choose a positive order size n. The order size is decreasing in the adjustment cost parameter χ. The second equation in (15) corresponds to the optimality condition for the market tightness. An increase in θ generates an expected gain for investors of $\alpha'(\theta)\frac{|\Omega|}{p}$, which is balanced against the increase in the expected transaction cost v_d.

2.2.3 Inaction region and competitive search

A typical feature of models with (exogeneous) proportional transaction costs is the presence of an inaction region (e.g., Constantinides, 1986; Davis and Norman, 1990). Importantly, this is not the case with endogenous transaction costs, as in search models. If the market tightness is fixed, then it would be optimal to set $n = 0$ whenever $\alpha(\theta)\frac{|\Omega|}{p} < v_d\theta$, as the benefit of trading would not be enough to compensate for the transaction cost. If investors can choose the market tightness, however, they can shift to contracts with smaller transaction costs when the gains from trade are small. Combining the optimality conditions for n and θ, and using Equation (2), we find that it is always optimal to trade a non-zero amount as long as

$\Omega \neq 0$:

$$\theta = \left(\frac{\overline{\alpha}}{v_d} \frac{|\,\Omega\,|}{p} \right)^{\frac{1}{1-\eta}}, \qquad n = \overline{\alpha} \frac{\theta^{\eta}}{\eta} \frac{1-\eta}{\chi} \frac{\Omega}{p} \;. \tag{16}$$

First, investors choose a smaller value for the market tightness, or equivalently higher trading delays, when the gains from trade measured by $|\Omega|$ are small. The trading speed declines in the dealers' profitability v_d. Second, investors buy the asset if $\Omega > 0$ and sell it if $\Omega < 0$. There is no inaction region and investors trade continuously.[17] Therefore, competitive search allows us to accommodate proportional transaction costs maintaining the tractability of continuous trading models.

2.2.4 Marginal value of portfolio rebalancing

The marginal value of rebalancing, $\Omega(W, S, X)$, plays a crucial role in determining the trading behavior of an investor. In Proposition 1, we characterize the marginal value of rebalancing in terms of deviations of the investor's portfolio from a target, which corresponds to the (frictionless) Merton portfolio.

Proposition 1. *The marginal value of rebalancing, denoted by $\Omega_{i,t} \equiv \Omega(W_{i,t}, S_{i,t}, X_t)$, is given by*

$$\Omega_{i,t} = \mathbb{E}_t \left[\int_t^{\infty} \frac{e^{-\rho(s-t)} C_{i,s}^{-\gamma}}{C_{i,t}^{-\gamma}} p_s \gamma_{i,s}^V \sigma_{R,s}^2 \left(Target_{i,s} - \frac{p_s S_{i,s}}{W_{i,s}} \right) ds \right],$$

where $\gamma_{i,t}^V \equiv -\frac{V_{WW,it} W_{i,t}}{V_{W,it}}$ is the value-function risk aversion, and $Target_{i,t}$ is given by:

$$Target_{i,t} = \underbrace{\frac{\mu_{R,t} - r_t}{\gamma_{i,t}^V \sigma_{R,t}^2} + \frac{V_{WX,it}}{\gamma_{i,t}^V V_{W,it}} \frac{\sigma_{X,t}}{\sigma_{R,t}}}_{\text{Merton portfolio share}}. \tag{17}$$

Proof. See Appendix A.1. □

The marginal value of rebalancing is the present discounted value of the deviation of the actual portfolio share, $p_t S_{i,t}/W_{i,t}$, from the frictionless intertemporal portfolio. Note that (17) is the optimal portfolio choice in an economy with time-varying investment opportunities and no frictions, as first characterized by Merton (1971). This optimal portfolio is the sum of the myopic portfolio choice, as in Markowitz's mean-variance analysis, and the intertemporal hedging term, by which investors take into account the evolution over time of the investment opportunity set. Thus, for example, the marginal value of portfolio

[17] Gârleanu and Pedersen (2016) provide evidence indicating that the trading behavior of institutional investors can be well approximated by continuous trading models.

rebalancing is positive if the investor foresees that now and in the future her share of the risky asset is expected to be lower than the one she would choose in a frictionless environment.

Proposition 1 provides a generalization of the results in Gârleanu and Pedersen (2013). In a model with quadratic utility and quadratic transaction costs, they find that investors adjust their portfolio towards the present discounted value of the Markowitz portfolio, which corresponds to the myopic component discussed above. In our setting with CRRA preferences, we obtain as a target the Merton portfolio, which includes intertemporal hedging demands. A second distinction with Gârleanu and Pedersen (2013) is that we derive a partial adjustment towards a target portfolio in the presence of (endogenous) proportional transaction costs. This is a consequence of allowing for endogenous transaction costs, as otherwise investors' behavior would be characterized by an inaction region.

The result in Proposition 1 also provides a new microfoundation for the asset valuation function in search models.[18] In standard search theory, to create a demand for trading assets, studies typically assume that investors' utility flow changes exogenously. In our model, aggregate shocks impact the marginal value of rebalancing through investors' wealth, creating an endogenous trade motive. In particular, the marginal value of holding an additional unit of the asset depends on the expected deviation of the portfolio share from the target. Therefore, it depends on both the wealth and the investment opportunity set.

2.2.5 Spreads and dealers' profitability

Trading fees and the dealers' value can also be characterized as functions of the marginal value of portfolio rebalancing. The intermediation fees that investors must pay to dealers are given by:

$$\phi_{i,t} = \eta|\Omega_{i,t}|. \tag{18}$$

Note that investors who are further away from their desired portfolio, such that they have a high value for $|\Omega_{i,t}|$, pay a higher intermediation fee in equilibrium.

Given our assumption of two families, the bid-ask spread can be defined as follows:

$$\phi_{ba,t} \equiv \frac{\phi_{1,t} + \phi_{2,t}}{p_t},$$

[18] See Weill (2020) for a discussion of alternative microfoundations for the asset valuation function in search models.

where $\phi_{j,t}$ denotes the fee incurred by family $j \in \{1,2\}$. Note that the bid-ask spread is increasing in the dispersion of the marginal value of rebalancing, as a mean-preserving spread of Ω raises the value of $\phi_{ba,t}$. An increase in the dispersion of Ω implies a higher demand for trading, as investors are on average further away from their desired portfolio. This higher demand for trading leads to higher transaction costs in equilibrium.

Combining the expressions in (16) with condition (4), we obtain $v_{d,t}$:

$$v_{d,t} = \overline{\alpha}^{\frac{2}{1+\eta}} \left[\frac{1-\eta}{\eta\chi\overline{d}} \left(\nu \left(\frac{|\Omega_{1,t}|}{p_t} \right)^{\frac{2}{1-\eta}} + (1-\nu) \left(\frac{|\Omega_{2,t}|}{p_t} \right)^{\frac{2}{1-\eta}} \right) \right]^{\frac{1-\eta}{1+\eta}}, \tag{19}$$

where $\Omega_{j,t}$ denotes the marginal value of rebalancing for an investor of type $j \in \{1,2\}$. Analogous to the bid-ask spread, the profitability of dealers increases in the dispersion of the marginal value of portfolio rebalancing. This is in contrast to the models that assume free entry, where the total intermediation capacity adjusts so the dealers' value is pinned down by the entry cost. This endogenous response of the dealer's value plays an important role in determining how liquidity and trading delays behave in periods of crises.

2.2.6 The liquidity-adjusted CCAPM

In Proposition 2, we provide a characterization of the risk premium in this economy.

Proposition 2. *The risk premium is given by:*

$$(\mu_{R,t} - r_t)dt = \gamma \frac{dC_t}{C_t}\frac{dp_t}{p_t} - \sum_{j=1}^{2} \omega_{j,t}^{c} \frac{\mathbb{E}_t[d(e^{-\rho t} C_{j,t}^{-\gamma} \Omega_{j,t})]}{e^{-\rho t} C_{j,t}^{-\gamma} p_t}, \tag{20}$$

where C_t denotes investors' aggregate consumption and $\omega_{j,t}^{c}$ is the consumption share of family $j \in \{1,2\}$.

Proof. See Appendix A.2. □

Expression (20) provides a liquidity-adjusted CCAPM. In the absence of liquidity frictions, Breeden's (1979) Consumption CAPM would hold in this economy, as $\Omega_{j,t} = 0$ for all investors. In the economy with search frictions, the risk premium depends not only on the asset's covariance with aggregate consumption, but also on the distribution of the marginal value of portfolio rebalancing and how it covaries with consumption.

An important implication of expression (20) is that the impact of liquidity frictions on asset prices is not summarized by the bid-ask spread. In particular, this can be seen by noticing that the parameter η, the elasticity of the matching function, enters in the determination of the intermediation fee (see Equation 18), while it does not enter directly into the pricing Equation (20). This suggests that we can make the bid-ask spread as small as possible, and still, liquidity frictions would have a non-trivial effect on asset prices. Therefore, a partial equilibrium analysis that focuses only on the direct impact of bid-ask spreads on returns could severely underestimate the effects of liquidity frictions on asset prices.

2.2.7 Markov Equilibrium

The aggregate state variables is given by $X_t = (Y_t, x_t, s_t)$, where x_t and s_t are defined as:

$$x_t = \frac{\nu W_{1,t}}{\nu W_{1,t} + (1-\nu) W_{2,t}}, \qquad s_t = \frac{\nu S_{1,t}}{\nu S_{1,t} + (1-\nu) S_{2,t}}, \tag{21}$$

and $W_{j,t}$ and $S_{j,t}$ denote the wealth and the number of shares of family $j \in \{1, 2\}$.

Besides the endowment Y_t, the aggregate dynamics is described by two state variables: the share of wealth of type-1 investors and the share of risky assets held be type-1 investors. While the wealth distribution usually appears as an aggregate state variable in heterogeneous-agent asset pricing models, it is the joint distribution of wealth and asset holdings that is relevant in our setting. We restrict our attention to a Markov equilibrium in state variables $X_t = (Y_t, x_t, s_t)$, defined below.

Definition 2. *A Markov equilibrium in state variables $X_t = (Y_t, x_t, s_t)$ is the set of functions: interdealer price $p(X)$, real interest rate $r(X)$, dealers' value $v_d(X)$, individual state functions $W_j(X)$ and $S_j(X)$ for $j \in \{1, 2\}$, value function $V(W, S, X)$, policy functions $\{C(W, S, X), n(W, S, X), \theta(W, S, X)\}$, intermediation fee $\phi(W, S, X)$, and laws of motion for state variables such that:*

(i) The value function satisfies the HJB equation (12), and policy functions C and (n, θ) are given by (14) and (16), respectively.

(ii) The fee $\phi(W, S, X)$ and dealers' value $v_d(X)$ satisfy conditions (10) and (19), respectively.

(iii) Markets for the consumption good, the risk-free bond, and the risky asset clear:

$$\nu\left(C_{1,t} + 0.5\chi p_t n_{1,t}^2\right) + (1-\nu)\left(C_{2,t} + 0.5\chi p_t n_{2,t}^2\right) + p_t v_{d,t}\bar{d} = Y_t, \tag{22}$$

$$\nu W_{1,t} + (1-\nu) W_{2,t} = p_t, \tag{23}$$

$$\nu S_{1,t} + (1-\nu) S_{2,t} = 1, \tag{24}$$

where $C_{j,t} = C\left(W_j(X_t), S_j(X_t), X_t\right)$, $n_{j,t} = n\left(W_j(X_t), S_j(X_t), X_t\right)$, $p_t = p(X_t)$, $v_{d,t} = v_d(X_t)$, $W_{j,t} = W_j(X_t)$, *and* $S_{j,t} = S_j(X_t)$.

(iv) The individual state functions $W_j(X)$ *and* $S_j(X)$ *are determined from* (21), (23), *and* (24).

(v) The state variable Y *evolves according to* (1), *and the laws of motion for* x_t *and* s_t *are given by:*

$$dx_t = \left[(r_t - \mu_{p,t})x_t + \pi_t s_t - \frac{\nu}{2}\chi n_{1,t}^2 - \nu\frac{C_{1,t}}{p_t} + \sigma_{p,t}^2 (x_t - s_t)\right] dt + \sigma_{p,t}(s_t - x_t)\, dZ_t, \tag{25}$$

$$ds_t = \nu n_{1,t}\alpha(\theta_{1,t})dt, \tag{26}$$

where $\theta_{j,t} = \theta\left(W_j(X_t), S_j(X_t), X_t\right)$, *and* $\left(\mu_p, \sigma_p\right)$ *are given by Ito's lemma from* $p(X_t)$.

A complete characterization of the equilibrium in this economy involves determining $V(W, S, Y, x, s)$, which requires solving a system of partial differential equations (PDEs) in five state variables. The presence of individual and aggregate state variables contrasts with standard settings (e.g., Brunnermeier and Sannikov, 2014; DiTella, 2017), in which, using the homotheticity assumption, one can eliminate the dependence of the value function to the individual state variables. Even though an exact solution to this system of PDEs is not available, in the next section, we obtain an explicit analytical characterization using state-global perturbation methods.

3 Asset Pricing Implications of Portfolio Flows

In this section, we show how portfolio flows affect asset prices in our economy with search frictions. We adopt perturbation techniques that allow us to obtain asymptotic closed-form expressions for investors' trading behavior and equilibrium asset prices.

3.1 State-global perturbation

We start by considering a parametric sequence of economies, indexed by $\epsilon > 0$, in which output is given by:

$$\frac{dY_t}{Y_t} = \mu dt + \sigma \sqrt{\epsilon} dZ_t,$$

and the capacity constraint for dealers is given by:

$$\int_{\Sigma} d_t(n, \phi)|n| d\sigma \leq \overline{d}\epsilon.$$

The parameter ϵ simultaneously controls the magnitude of the variance of endowments, $\sigma^2\epsilon$, and the dealers' intermediation capacity $\overline{d}\epsilon$. The special case $\epsilon = 0$ provides a convenient benchmark where equilibrium objects can be easily characterized. We proceed by taking a small-risk approximation, that is, we study how the economy behaves in the neighborhood of $\epsilon = 0$. In particular, we are interested in computing the following perturbation:

$$V(W, S, X; \epsilon) = \underbrace{V^{\star}(W, S, X)}_{\text{benchmark: } \epsilon=0} + \underbrace{\tilde{V}(W, S, X)}_{\text{first-order correction}} \epsilon + \mathcal{O}(\epsilon^2),$$

where we denote $V^{\star}(W, S, X) \equiv V(W, S, X; 0)$ and $\tilde{V}(W, S, X) \equiv V_{\epsilon}(W, S, X; 0)$.

The term $V^{\star}(W, S, X)$ corresponds to the value function in the non-stochastic benchmark, that is, $\epsilon \longrightarrow 0$. The term $\tilde{V}(W, S, X)$ corresponds to the first-order correction, which is the derivative of the value function with respect to ϵ evaluated at $\epsilon = 0$. These first-order corrections are our main objects of interest. We adopt an analogous notation for policy functions and remaining equilibrium objects. For instance, consumption is given by:

$$C(W, S, X; \epsilon) = C^{\star}(W, S, X) + \tilde{C}(W, S, X)\epsilon + \mathcal{O}(\epsilon^2).$$

Note that, in contrast to the common use of perturbation methods in economics, we are not imposing that the solution is linear in the state variables.[19] Instead, we allow the value function and policy functions to depend in a non-linear way on the set of state variables. We refer to this method, which is local in the

[19] Standard applications of perturbation methods involve a linearization (or higher-order perturbations) around a steady state, so the analysis is local in both the state variables and the amount of risk. We provide a detailed discussion of the state-global perturbation techniques and the differences with standard linearization methods in Appendix OA.3.

parameter ϵ, but global in the state space, as a state-global perturbation. The global nature of this method with respect to state variables is critical when considering the economy's response to large shocks.

It is important to note that we scale both the endowment's variance σ^2 and dealers' intermediation capacity $\overline{d}$ by ϵ. This parametrization is necessary to guarantee that the liquidity frictions matter in the case with small risk. As we reduce the endowment's variance, the demand for trading is reduced, as the risky and riskless assets become more similar to each other. By assuming that the intermediation capacity is also reduced with parameter ϵ, we ensure that the supply of liquidity is commensurate with the demand for trading in the economy, so spreads will be positive and trading frictions will be relevant even in the neighborhood of $\epsilon = 0$.

3.1.1 The benchmark economy

In Lemma 1, we characterize the zeroth-order terms in our perturbation: the benchmark (non-stochastic) economy obtained when $\epsilon = 0$.

Lemma 1. *Suppose $\rho + (\gamma - 1)\mu > 0$.*[20] *Then, for the $\epsilon = 0$ economy, the investors' value function and policy functions are given by:*

$$V^\star(W,S,X) = A\frac{W^{1-\gamma}}{1-\gamma}, \qquad C^\star(W,S,X) = (\rho + (\gamma - 1)\mu)\, W, \tag{27}$$

$n^\star(W,S,X) = 0$, and $\theta^\star(W,S,X)$ is indeterminate, where $A^{-\frac{1}{\gamma}} = \rho + (\gamma - 1)\mu$. Dealers' value and orders are given by $v_d^\star = 0$ and $d^\star(n,\phi) = 0$, and equilibrium prices satisfy:

$$\mu_R^\star = r^\star = \rho + \gamma\mu, \qquad q^\star = \frac{1}{\rho + (\gamma - 1)\mu},$$

and $\sigma_R^\star = 0$, where $q_t \equiv p_t/Y_t$ is the price-dividend ratio.

Proof. See Appendix A.3. □

In the economy with $\epsilon = 0$, there is no risk, so the two financial assets are essentially perfect substitutes. The investors then has no incentive to change their initial portfolios, and there is no trade in the risky asset in equilibrium.[21] Therefore, trading frictions play no role in the determination of prices and quantities.

[20] The condition $\rho+(\gamma-1)\mu > 0$ is standard in economies with growth, and it goes back to Brock and Gale (1969). This condition guarantees that the investors' utility is well-defined in equilibrium.

[21] Note, however, that investors may still accumulate or sell riskless assets over time.

As investors do not trade, they are indifferent between any value of the market tightness. Dealers post no contracts and earn no profits. We obtain the standard result that consumption is linear in wealth and the value function is a power function of wealth. Finally, because there is no aggregate risk, the risk premium is equal to zero, and the real interest rate is constant and given by the standard condition.

3.1.2 The small-risk economy

Given the value of $V^\star(W, S, X)$ and the corresponding policy functions, we are able to compute the first-order approximation of the investors' problem.[22] In Proposition 3, we characterize the value function and policy functions in the small-risk economy.

Proposition 3 (Value function and policy functions)**.** *Suppose* $\rho + (\gamma - 1)\mu > 0$.

a. Value function:

$$\tilde{V}(W, S, X) = \frac{AW^{1-\gamma}}{r^\star - \mu}\left[\tilde{r}(X) + \tilde{\pi}(X)\frac{p^\star(X)S}{W} - \frac{\gamma}{2}\tilde{\sigma}_R^2(X)\left(\frac{p^\star(X)S}{W}\right)^2\right], \tag{28}$$

where $\tilde{r}(X)$, $\tilde{\pi}(X)$, *and* $\tilde{\sigma}_R^2(X)$ *denote, respectively, the first-order correction for the interest rate, risk premium, and return variance, and* $p^\star(X_t) = q^\star Y_t$.

b. Marginal value of rebalancing:

$$\tilde{\Omega}(W, S, X) = \frac{\gamma\tilde{\sigma}_R^2(X)}{r^\star - \mu}\left(\frac{\tilde{\pi}(X)}{\gamma\tilde{\sigma}_R^2(X)} - \frac{p^\star(X)S}{W}\right)p^\star(X). \tag{29}$$

c. Policy functions:

$$\tilde{C}(W, S, X) = \left[\frac{\gamma - 1}{\gamma}\tilde{r}(X) + \tilde{\pi}(X)\frac{p^\star(X)S}{W} - \frac{(\gamma + 1)}{2}\tilde{\sigma}^2(X)\left(\frac{p^\star(X)S}{W}\right)^2\right]W \tag{30}$$

$$\theta^\star(W, S, X) = \left(\frac{\overline{\alpha}}{\tilde{v}_d(X)}\frac{|\tilde{\Omega}(W, S, X)|}{p^\star(X)}\right)^{\frac{1}{1-\eta}} \tag{31}$$

$$\tilde{n}(W, S, X) = \alpha\left(\theta^\star(W, S, X)\right)\frac{1 - \eta}{\chi}\frac{\tilde{\Omega}(W, S, X)}{p^\star(X)}. \tag{32}$$

Proof. See Appendix A.4. □

[22] Because the market tightness is not defined at $\epsilon = 0$, we cannot apply the implicit function theorem to compute the first-order corrections. The solution is computed instead using bifurcation methods, as discussed in Judd and Guu (2001).

In Proposition 3, we characterize the value function in terms of the individual state variables (W, S) and equilibrium prices $(\tilde{r}(X), \tilde{\pi}(X), \tilde{\sigma}_R(X), p^\star(X))$. Note that the value function is concave in the portfolio share in our model, whereas in the benchmark model, it does not depend on the portfolio share. Given the value function, we can solve for the marginal value of rebalancing Ω. The value of Ω depends on how the portfolio share compares with the myopic portfolio $\frac{\tilde{\pi}(X)}{\gamma\tilde{\sigma}_R^2(X)}$. For σ sufficiently small, the hedging demand is small compared to the myopic demand and can be ignored up to a first-order approximation. The marginal value of rebalancing is positive when the actual portfolio share is below the target myopic portfolio, while it is negative when the portfolio share is above the target.

Proposition 3 also shows that the consumption function depends on the investors' portfolio position. This result implies that, even though risk-free bonds can be traded in frictionless markets, search frictions affect savings decisions and, ultimately, the economy's interest rate. In particular, the consumption-wealth ratio is a concave function of the portfolio share. Therefore, an increase in portfolio dispersion would depress consumption, everything else constant, ultimately affecting the equilibrium real interest rate.

In Proposition 3, we also characterize the behavior of the order size and market tightness. Note that, even though the market tightness is indeterminate when $\epsilon = 0$, there is a well-defined limit when $\epsilon \longrightarrow 0$, which is given by $\theta^\star(W, S, X)$. As before, the investor is a buyer (seller) when the marginal value of rebalancing is positive (negative). Moreover, the market tightness is higher when the investor is further away from her desired portfolio.

3.2 The aggregate implications of search frictions

Next, we consider the implications of search frictions for the risk premium, risk-free rate, and market liquidity in three propositions. We define two key features of the distribution of investors asset holdings: portfolio dispersion and asymmetry. We then derive various aspects of asset prices and market liquidity as functions of these two features.

3.2.1 Risk premium and order flow

From the market-clearing condition for the risky asset and the law of motion for the holdings of the risky asset by the two types of investors, we obtain the following condition:

$$\nu\alpha(\theta_{1,t})n_{1,t} + (1-\nu)\alpha(\theta_{2,t})n_{2,t} = 0. \tag{33}$$

For concreteness, suppose type-1 investors start with a relatively high portfolio share, $\frac{p_t S_{1,t}}{W_{1,t}} = \frac{s_t}{x_t} > 1$, so they will be sellers in equilibrium. This implies that the portfolio share of type-2 investors satisfies $\frac{p_t S_{2,t}}{W_{2,t}} = \frac{1-s_t}{1-x_t} < 1$, so they will be buyers in equilibrium. Combining (33) with the expressions for market tightness and orders in (31) and (32), we obtain:

$$\underbrace{\nu\left(\frac{s}{x} - \frac{\tilde{\pi}(X)}{\gamma\tilde{\sigma}_R^2(X)}\right)^{\frac{1+\eta}{1-\eta}}}_{\text{Supply of shares: } \mathcal{S}(\tilde{\pi}|X)} = \underbrace{(1-\nu)\left(\frac{\tilde{\pi}(X)}{\gamma\tilde{\sigma}_R^2(X)} - \frac{1-s}{1-x}\right)^{\frac{1+\eta}{1-\eta}}}_{\text{Demand for shares: } \mathcal{D}(\tilde{\pi}|X)}, \tag{34}$$

for $\tilde{\pi}(X)$ satisfying $\frac{1-s}{1-x} \leq \frac{\tilde{\pi}(X)}{\gamma\tilde{\sigma}_R^2(X)} \leq \frac{s}{x}$.[23]

Equation (34) defines the aggregate demand for shares in the economy, $\mathcal{D}(\tilde{\pi} \mid X)$, and the aggregate supply of shares, $\mathcal{S}(\tilde{\pi} \mid X)$. Despite isoelastic preferences, the trading elasticity of buy and sell orders is time-varying and depends on how far an investor is from her target portfolio.[24] Note that $\mathcal{D}(\cdot \mid X)$ is increasing, convex, and it satisfies $\mathcal{D}(0 \mid X) = 0$ and $\mathcal{D}(\tilde{\pi} \mid X) \longrightarrow \infty$ as $\tilde{\pi} \longrightarrow \infty$. The order supply schedule is decreasing, convex, and it satisfies $\mathcal{S}(\tilde{\pi} \mid X) \longrightarrow 0$ as $\tilde{\pi} \longrightarrow \infty$ and $\mathcal{S}(\tilde{\pi} \mid X) > 0$ as $\tilde{\pi} \longrightarrow 0$. Therefore, there exists a unique value of $\tilde{\pi}$ that satisfies the market-clearing condition. The right panel of Figure 2 shows the supply and demand schedules for a given point in the state space, represented by the dot in the left panel. Note that because we are expressing orders as a function of the risk premium, instead of the price of the risky asset, the demand for shares is the upward-sloping curve, and the supply of shares is the downward-sloping curve in Figure 2. In Proposition 4, we give an explicit solution for the risk premium in terms of investors' portfolio positions.

Proposition 4 (Risk premium and investors' portfolios)**.** *Suppose $\rho + (\gamma - 1)\mu > 0$. The risk premium is given by:*

$$\tilde{\pi}(X) = \left[\tilde{\nu}\frac{s}{x} + (1-\tilde{\nu})\frac{1-s}{1-x}\right]\gamma\tilde{\sigma}_R^2(X), \tag{35}$$

where $\tilde{\nu} = \nu^{\frac{1-\eta}{1+\eta}}\left[\nu^{\frac{1-\eta}{1+\eta}} + (1-\nu)^{\frac{1-\eta}{1+\eta}}\right]^{-1}$.

Proof. See Appendix A.5. □

Proposition 4 shows how the risk premium depends not only on the components typically emphasized

[23]Note that $\mathcal{D}(\tilde{\pi} \mid X) = 0$ if $\frac{\tilde{\pi}(X)}{\gamma\tilde{\sigma}_R^2(X)} < \frac{1-s}{1-x}$ and $\mathcal{S}(\tilde{\pi} \mid X) = 0$ if $\frac{\tilde{\pi}(X)}{\gamma\tilde{\sigma}_R^2(X)} > \frac{s}{x}$, so the risk premium must satisfy $\frac{1-s}{1-x} \leq \frac{\tilde{\pi}(X)}{\gamma\tilde{\sigma}_R^2(X)} \leq \frac{s}{x}$ in equilibrium.

[24]In Appendix OA.4, we show that relatively inelastic demand for shares is associated with high risk premium amplification, consistent with Gabaix and Koijen (2020).

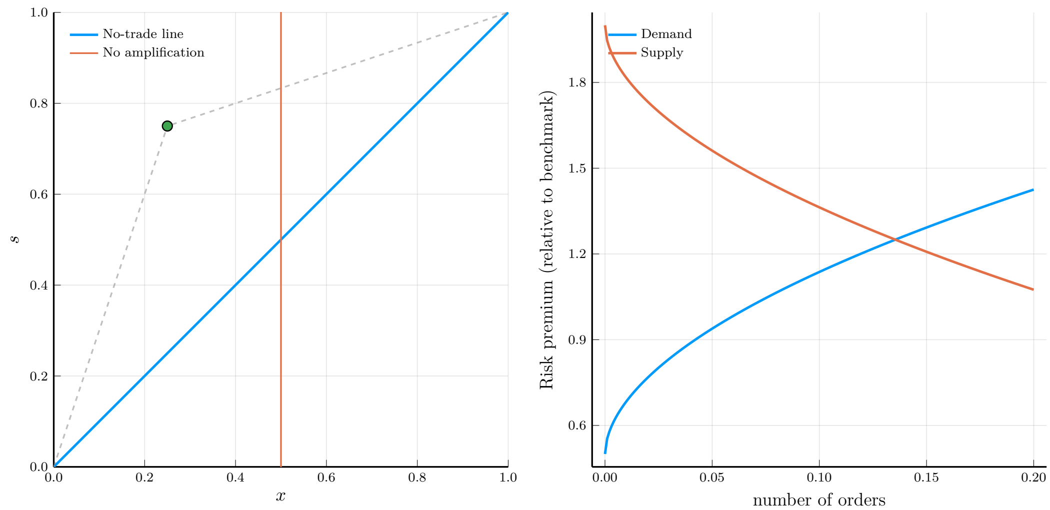

Figure 2. The left panel of this figure depicts the state space. The blue line shows the points where there is no trade and the vertical line shows the points where there is positive trade but no amplification. The right panel shows the determination of the risk premium as the intersection of the supply and demand for shares.

by asset pricing theory, such as risk $\tilde{\sigma}_R^2(X)$ and risk aversion γ, but also on the investors' portfolios. In the absence of search frictions, the distribution of portfolio holdings would not matter for the determination of asset prices, and the risk premium would be given by the standard formula $\gamma\tilde{\sigma}_R^2(X)$.

Relative to the frictionless case, the economy may feature amplification or dampening of the risk premium. There are two benchmark cases where the risk premium coincides with the one in the frictionless economy. First, the case where $s = x$, represented by the 45 degree line in the left panel of Figure 2. The portfolio share is equal to 1 for both investors in this case, and there is no trade. In the absence of frictions, investors would immediately adjust their portfolio to hold 100% of the risky asset, as investors have no incentive to borrow or lend given the assumption of common beliefs and risk aversion. If investors happen to start with a portfolio share of 100%, they have no incentive to trade and the economy with search frictions behaves effectively as a representative-agent economy. Second, the case in which $x = \tilde{\nu}$, represented by the vertical line in the left panel of Figure 2. Even though there is positive trade, there are no effects in the risk premium.

To better understand when positive trade can lead to effects on expected returns, it is useful to consider portfolio dispersion and asymmetry as defined below.

Definition 3. *We define portfolio dispersion, Δ, and (intensive-margin) portfolio asymmetry, a, as follows:*

$$\Delta \equiv \left|\frac{s}{x} - 1\right| + \left|\frac{1-s}{1-x} - 1\right|, \qquad a \equiv \left|\frac{s}{x} - 1\right| - \left|\frac{1-s}{1-x} - 1\right|.$$

Portfolio dispersion measures how far investors are from their desired portfolio and it corresponds to the difference in their portfolio share, $\Delta = \frac{s}{x} - \frac{1-s}{1-x}$ in the case $s > x$. Portfolio asymmetry captures the relative distance of sellers and buyers to their desired portfolio. For instance, if $s/x = 1.5$ and $(1-s)/(1-x) = 0.5$, then there is positive dispersion, $\Delta = 1.0$, but there is no asymmetry, $a = 0$, as buyers and sellers are equidistant to their (long-run) target portfolio.

We can express the risk premium in terms of portfolio dispersion and asymmetry:

$$\tilde{\pi}(X) = \left[1 + \left(\tilde{\nu} - \frac{1}{2}\right)\Delta + \frac{a}{2}\right]\gamma\tilde{\sigma}_R^2(X), \tag{36}$$

assuming $s > x$ for concreteness.

The intuition behind Equation (36) is as follows. Consider the case where $\nu = 0.5$, so there is an equal number of buyers and sellers. In this case, there is no amplification if $a = 0$, that is, buyers and sellers are equally distant from their target portfolio. Buyers and sellers then have an equal influence on the market and prices coincide with those in a representative-agent economy. We obtain amplification if sellers are disproportionately further away from their target portfolio, that is, $a > 0$. This corresponds to the set of points to the left of the vertical line and above the 45 degree line in Figure 2. This situation characterizes net selling pressure in the market. The representative-agent price is not enough to induce buyers to absorb this selling pressure, so the price needs to go down, which increases the risk premium. Therefore, the results suggests that portfolio flows and selling pressure play a key role in determining the risk premium in the presence of search frictions.

Second, consider now the case where $\nu > 0.5$, but $a = 0$, so buyers and sellers are equidistant from the target portfolio. In this case, there is a large number of sellers in the market, which again creates net selling pressure and amplification. This situation captures an asymmetry in orders on the extensive margin, while the case $\nu = 0.5$ and $a > 0$ capture asymmetry on the intensive margin. Risk premium amplification depends then on the asymmetry between investors in the market.

3.2.2 Interest rate

Having determined the level of the risk premium, we next consider the behavior of interest rates. Aggregating the individual consumption decisions and using the market-clearing condition for goods allows us to obtain the level of the interest rate. In Proposition 5, we characterize the risk-free rate in our setting.

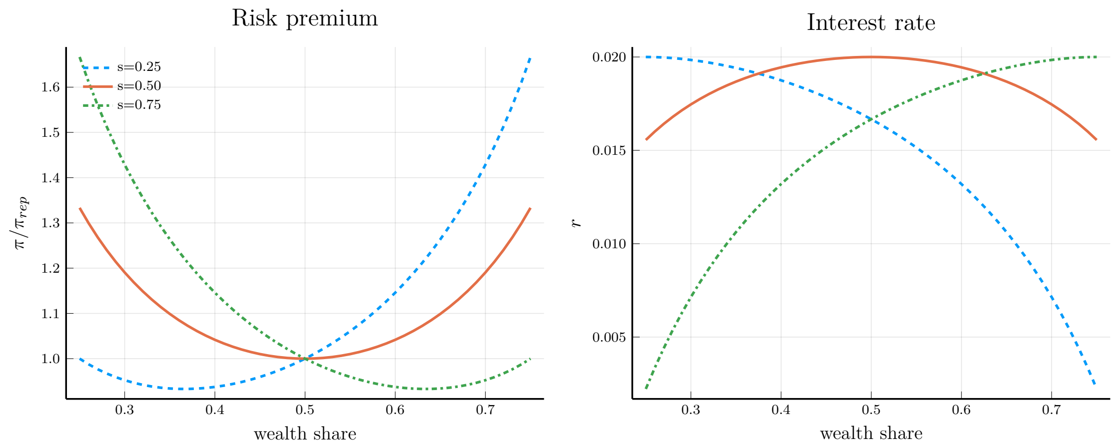

Figure 3. This figure shows the risk premium (left panel) and the interest rate (right panel) as functions of the wealth share x for different values of the asset share s. The risk premium is normalized by the value of the risk premium in a representative agent economy.

Proposition 5 (Interest rate)**.** *The risk-free interest rate is given by:*

$$r(X) = \rho + \gamma\mu - \frac{\gamma(\gamma+1)}{2}\left[x\left(\frac{s}{x}\right)^2 + (1-x)\left(\frac{1-s}{1-x}\right)^2\right]\tilde{\sigma}_R^2(X)\epsilon + \mathcal{O}(\epsilon^2). \tag{37}$$

Proof. See Appendix A.6. □

In the case where investors reached their desired portfolio, $s = x$, the expression in Equation (37) boils down to the standard condition for the interest rate in a frictionless economy. As investors' portfolios deviate from this benchmark, we find that there is a reduction in interest rates relative to the frictionless economy through an amplification of the precautionary savings motive. In particular, we can show that the interest rate is decreasing in the degree of dispersion in investors' portfolios.[25] This result reflects the concavity of the consumption function on the portfolio share. An increase in the dispersion of portfolios leads to a reduction in aggregate consumption in the absence of any price reaction. The interest rate then goes down to restore the equilibrium. The right panel of Figure 3 shows the behavior of the interest rate as a function of x for different values of s. The figure shows that the interest rate is maximized at the point $x = s$, so portfolio shares are equalized across investors.

[25]Note that $x\left(\frac{s}{x}\right)^2 + (1-x)\left(\frac{1-s}{1-x}\right)^2 = Var_x[\frac{S}{W}] + 1$, where $Var_x[\frac{S}{W}]$ is the cross-sectional variance of the portfolio share weighted by the corresponding wealth share. Therefore, an increase in the variance of the portfolio share reduces the interest rate.

3.2.3 Volatility

In Appendix OA.1.3, we derive the expression for return volatility and show that time-varying volatility requires a higher-order correction. Up to the first-order, volatility is constant and given by $\tilde{\sigma}_R(X) = \sigma$. A time-varying endogenous volatility is obtained when considering a second-order approximation.

We also find that the sign of the second-order term for the endogenous volatility is ambiguous. As we can seen from the left panel of Figure 3, the risk premium moves in the opposite direction to the interest rate for most of the state space. The opposing effects of the risk premium and the interest rate are typically present even in frictionless asset pricing models, and they determine the degree of amplification in volatility. If a negative shock increases the risk premium by more than it reduces the interest rate, then the discount rate increases, which amplifies the effect of the negative shock. In frictionless asset pricing models, a high elasticity of intertemporal substitution (EIS) implies that the risk premium effect dominates. In an economy with search frictions, however, the EIS is not sufficient to pin down which effect dominates, as it is possible for the interest effect to dominate even in the high-EIS case.

3.2.4 Market liquidity

In Proposition 6, we characterize the response of volume traded, bid-ask spreads, dealers' value, and trading delays.

Proposition 6 (Market liquidity)**.** *Suppose $\rho + (\gamma - 1)\mu > 0$ and let $\Delta \equiv \left|\frac{s}{x} - \frac{1-s}{1-x}\right|$ denote portfolio dispersion.*

a. Volume traded is determined as follows:

$$\mathbb{V}(X) = \overline{V}\Delta^{\frac{1-\eta}{1+\eta}},$$

where $\overline{V}$ is a positive constant defined in the Appendix.

b. Dealers' value is determined by the following:

$$\tilde{v}_d(X) = \left[\frac{1-\eta}{\chi\eta\overline{d}}\left(\nu\left(\frac{\overline{\alpha}|\tilde{\Omega}_1|}{p^\star(X)}\right)^{\frac{2}{1-\eta}} + (1-\nu)\left(\frac{\overline{\alpha}|\tilde{\Omega}_2|}{p^\star(X)}\right)^{\frac{2}{1-\eta}}\right)\right]^{\frac{1-\eta}{1+\eta}} = \overline{v}_d\Delta^{\frac{2}{1+\eta}}, \tag{38}$$

where $\tilde{\Omega}_j$ is the marginal value of rebalancing for investor of type $j \in \{1,2\}$, and $\overline{v}_d$ is a positive constant defined in the Appendix.

c. Bid-ask spread is determined by:

$$\tilde{\phi}_{ba}(X) = \frac{\eta\gamma\sigma^2}{r^\star - \mu}\Delta.$$

d. Market tightness of investor j, $j \in \{1, 2\}$, is given by:

$$\theta_j^\star(X) = \overline{\theta}_j \Delta^{-\frac{1}{1+\eta}},$$

where $\overline{\theta}_j$ is a positive constant defined in the Appendix.

Proof. See Appendix A.7. □

Volume traded $\mathbb{V}(X)$, defined as the total number of shares traded at a given point in time, is increasing in the portfolio dispersion. The higher the differences in the portfolios, the more distant investors are from their target portfolio, and the higher the demand for trade. We show in Appendix A.7 that parameters affect the volume traded in the expected manner: volume is increasing in the efficiency of the matching function $\overline{\alpha}$ and in the dealers' intermediation capacity $\overline{d}$, and it is decreasing in the investor's adjustment cost parameter χ. Interestingly, volume responds positively to the level of risk and risk aversion, as this leads to an increase in the risk premium and also the benefits of trading the risky asset.

Figure 4 shows the behavior of volume as a function of the wealth share x for different values of the asset share s. The behavior of volume is highly non-linear in the state variables. As we approach the point $x = s$, there is no incentive to trade and the volume goes to zero. As we move away from this point, volume increases quickly, indicating that the volume induced by this rebalancing motive is particularly relevant for large shocks.

The dealer's value $\tilde{v}_d(X)$ is increasing in the dispersion of the marginal value of rebalancing and, ultimately, on the dispersion of portfolios. Similarly, the bid-ask spread $\tilde{\phi}(X)$ is also increasing in the portfolio dispersion. Given the dealers' limited intermediation capacity, an increase in the demand for trading leads to an increase in transaction costs and an increase in the dealers' profits.

Figure 4 also shows the behavior of the dealers' value and of the bid-ask spread. Analogous to trading volume, the dealers' value and the bid-ask spread are both highly non-linear functions of the state variables. Both variables are close to zero in the neighborhood of $x = s$ and they increase sharply as we move away from the point of no-trade.

Portfolio dispersion also has an implication for the second dimension of market liquidity: trading

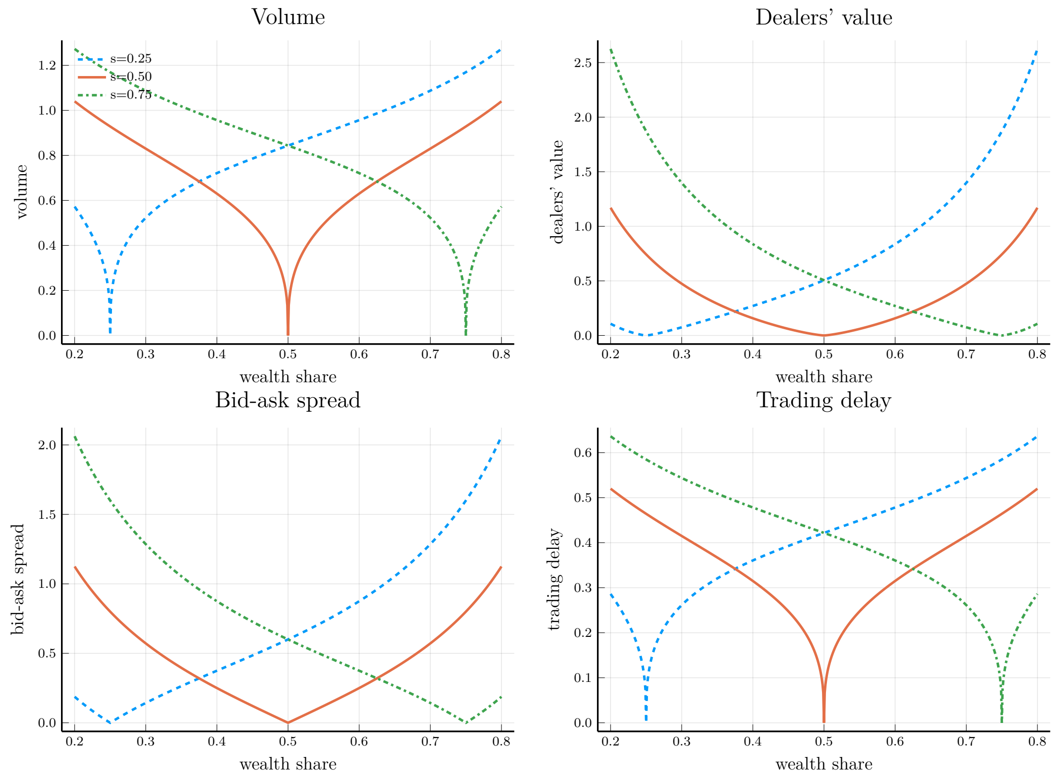

Figure 4. This figure shows the volume traded (top-left panel), dealers' value (top-right panel), the bid-ask spread (bottom-left panel), and trading delay (bottom-right panel) as functions of the wealth share x for different values of the asset share s. Trading delay is defined as the inverse of $\alpha(\theta)$ for the type-1 investor.

delays, which is defined as the inverse of the arrival rate $\alpha(\theta)$. An increase in portfolio dispersion leads to an increase in trading delays. Note that there are opposite forces affecting the investors' choice of market tightness, as can be seen from Equation (31). First, investors choose faster trading when they are further away from their target portfolio. Second, the increase in trading costs induces investors to switch to cheaper and slower trades. In equilibrium, the second force dominates. The reason is that, given the limited intermediation capacity of dealers, it is not possible to simultaneously increase the trading volume and trading speed for all investors. The cost of trading must then increase enough to induce investors to choose instruments with a longer execution time. This is consistent with the deterioration in liquidity conditions and the shift towards slower trades documented in Kargar et al. (2020). Note that we would not obtain this result with a free-entry condition for dealers, as v_d would be pinned down by the entry costs, and only the first force would be at play.

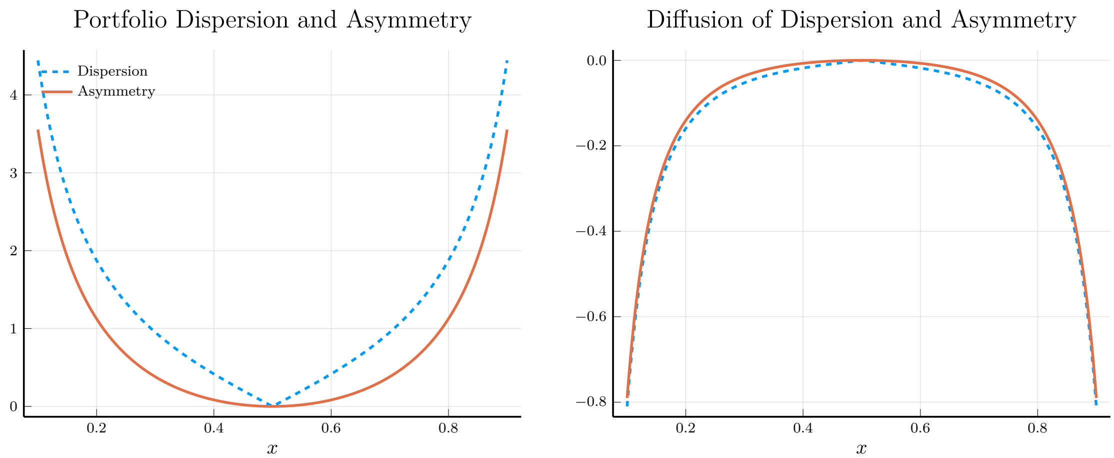

Figure 5. The left panel shows portfolio dispersion and asymmetry as a function of the wealth share x given $s = 0.5$. The right panel shows the diffusion of portfolio dispersion and asymmetry, that is, their exposure to the aggregate risk.

3.3 Countercyclical portfolio asymmetry and dispersion

In this section, we show that portfolio asymmetry and dispersion are countercyclical, so negative shocks will trigger asset prices and liquidity responses in line with the empirical evidence.

The left panel of Figure 5 presents portfolio dispersion and asymmetry, as defined in Definition 3, as a function of x, given $s = 0.5$. We plot the diffusions of these two measures, that is, their exposures to the aggregate shock in the right panel. For all values of x, the diffusion terms are negative, so asymmetry and dispersion increase endogenously in response to a negative shock. For instance, the diffusion term for portfolio dispersion, $\tilde{\sigma}_{\Delta,t}$, for $s > x$ is determined as follows:

$$\tilde{\sigma}_{\Delta,t} = -\left[\frac{s}{x^2} + \frac{1-s}{(1-x)^2}\right](s-x)\sigma < 0,$$

using the fact that the diffusion of x satisfies $\tilde{\sigma}_x = (s - x)\sigma$.

Intuitively, a negative shock increases (decreases) the portfolio share of type-1 (type-2) investors increasing portfolio dispersion. As the negative shock affects the leveraged investor disproportionately more, it also increases asymmetry. As illustrated by Figure 5, these effects are particularly relevant for large shocks. In the neighborhood of $s = x$, asymmetry and dispersion are not very sensitive to changes in the aggregate shock, so small shocks would have a limited impact on aggregate variables. Therefore, trading frictions have only a small impact on asset prices during normal times when asymmetry and dispersion are low, and the market is liquid with low transaction costs. In crisis periods, when asymmetry and dispersion are high, the market liquidity deteriorates, and trading frictions have a large effect on asset prices.

4 Quantitative Implications of the Model

In this section, we study the quantitative implications of the model in the long run and assess its dynamic response to a large adverse shock. First, we extend the model to allow for heterogeneity in investors' risk aversion, which is crucial for search frictions to have a long-run impact on the economy. Next, we calibrate the model to match key asset pricing and secondary market moments from the corporate bond market. Finally, in response to a large negative shock, we show that the model can generate asset pricing and market liquidity dynamics consistent with the evidence from the COVID-19 crisis.

4.1 Heterogeneous target portfolios.

An important feature of the economy we have considered so far is that search frictions have no impact on investors' trading behavior in the long-run. Investors eventually reach their frictionless portfolios (i.e., $s = x$) so both agents invest the same fraction of their wealth in the risky asset. In the long run, portfolios do not react to aggregate shocks and the model behaves as if there is a representative agent. Therefore, liquidity frictions have only a transient effect, as eventually the economy converges to the point of no trade.

To explore the long-run asset pricing implications of trading frictions, we augment the model by introducing heterogeneous risk aversions for investors. We assume that investors have stochastic differential utility, as in Duffie and Epstein (1992), the continuous-time analog of the recursive preferences of Epstein and Zin (1989). We also assume that agents have different risk aversions, $\gamma_1 \leq \gamma_2$, but they have the same elasticity of intertemporal substitution (EIS) ψ. To guarantee that a (non-degenerate) stationary distribution of wealth exists, we assume that investors exit with intensity κ. Therefore, for an investor i who belongs to family $j \in \{1, 2\}$, preferences are given by:

$$V_{i,t} = \mathbb{E}_t \left[\int_t^\infty f_i(C_{i,s}, V_{i,s}) ds \right],$$

where

$$f_i(C,V) = \rho \frac{(1-\gamma_j)V}{1-\psi^{-1}} \left\{ \left(\frac{C}{\left((1-\gamma_j)V\right)^{\frac{1}{1-\gamma_j}}} \right)^{1-\psi^{-1}} - 1 \right\}.$$

In Proposition 7, we extend the characterization of the equilibrium to a case with heterogeneous risk aversions. To avoid repetition, we report the results for only a few selected equilibrium objects and leave a more detailed description of the economy with heterogeneous preferences to Appendix A.9.

Proposition 7 (Heterogeneous risk aversions)**.** *Suppose* $\rho + (\psi^{-1} - 1)\mu > 0$.

a. Marginal value of rebalancing is given by:

$$\tilde{\Omega}_j(W,S,X) = \frac{\gamma_j \tilde{\sigma}_R^2(X)}{r^\star - \mu} \left(\frac{\tilde{\pi}(X)}{\gamma_j \tilde{\sigma}_R^2(X)} - \frac{p^\star(X) S}{W} \right) p^\star(X).$$

b. Risk premium is given by:

$$\tilde{\pi}(X) = \left[\tilde{\nu} \gamma_1 \frac{s}{x} + (1-\tilde{\nu}) \gamma_2 \frac{1-s}{1-x} \right] \sigma^2$$

c. Bid-ask spread is determined using the following:

$$\tilde{\phi}_{ba} = \frac{\eta}{r^\star - \mu} \left| \gamma_1 \frac{s}{x} - \gamma_2 \frac{1-s}{1-x} \right| \sigma^2.$$

Proof. See Appendix A.8. □

The target portfolio now depends on the investors' risk aversion, $\frac{\tilde{\pi}(X)}{\gamma_j \tilde{\sigma}_R^2(X)}$. Investors with low risk aversion have a higher target portfolio and operate with leverage in equilibrium. The determination of the risk premium is analogous to the one in the case with homogeneous preferences, but the effective weights on the portfolio shares are now $\tilde{\nu}\gamma_1$ and $(1-\tilde{\nu})\gamma_2$, which creates another source of asymmetry across investors. Finally, the bid-ask spread is now proportional to the difference of risk-aversion-adjusted portfolio shares, or equivalently, to the difference in the relative distance to the target portfolio.

The law of motion of the wealth share is given by:

$$\tilde{\mu}_x(X) = \left[(x-s)\sigma^2 + \frac{\psi+1}{2} \sigma^2 x(1-x) \left(\gamma_1 \left(\frac{s}{x}\right)^2 - \gamma_2 \left(\frac{1-s}{1-x}\right)^2 \right) \right] - \kappa(x-\nu),$$

where κ denotes the mortality rate and $\tilde{\sigma}_x(X) = (s-x)\sigma$.

We can define a stochastic steady state as the point where $\tilde{\mu}_x(X) = \tilde{\mu}_s(X) = 0$, so equilibrium variables remain constant in the absence of shocks. In the case of homogeneous preferences, this point corresponds to $x = s = \nu$, where investors have reached their desired portfolios, and they have no incentive to trade. Moreover, we have that $\tilde{\sigma}_x(X) = 0$, so the stochastic steady-state is an absorbing state. If the economy ever reaches this point, it would stay there forever. In the case of heterogeneous preferences, the portfolio share of both investors will be equal to their target portfolio only if $\gamma_1 \frac{s}{x} = \gamma_2 \frac{1-s}{1-x}$. This implies that type-1 investors operate with leverage in the stochastic steady state; thus, their wealth share has a positive exposure to risk, $\tilde{\sigma}_x(X) > 0$. Therefore, with heterogeneous risk aversions, the stochastic steady state is no longer an absorbing state, and liquidity frictions impact equilibrium outcomes even in the long run.

4.2 Calibration

Table 1 lists the parameter values used in calibrating the model with heterogeneous risk aversions. Consistent with Gârleanu and Panageas (2015), we choose $\mu = 0.02$ and $\sigma = 0.04$, the drift and diffusion of the aggregate endowment, so that time-integrated data from the model can approximately match the first two moments of annual U.S. consumption growth. We set the elasticity of intertemporal substitution to $\psi = 1.5$, consistent with the asset pricing literature (e.g., Bansal and Yaron, 2004). Following Gârleanu and Panageas (2015), the share of high-type agents in the population is set to $\nu = 0.01$. We set risk aversion coefficients γ_1 and γ_2, the subjective discount rate ρ, and agents' entry and exit rate κ to match the unconditional equity premium (on the levered claim) of approximately 7% (consistent with Barro, 2006), a wealth-weighted average risk aversion of at most 10, the leverage of high-type agents of around 5, and an average real interest rate between 1-1.5%.

We use three moments from the corporate bond market to calibrate the remaining parameters from the model.[26] We target the transaction costs in the secondary market before the COVID crisis from O'Hara and Zhou (2020), which is approximately 40 bps, to identify the concavity of the matching function η. Targeting the annual turnover of corporate-to-dealer trades, which is approximately 0.2 in the TRACE data before the onset of the COVID-19 crisis, we calibrate the portfolio adjustment cost parameter χ. Parameters $\overline{\alpha}$ and $\overline{d}$ that determine the efficiency of the matching function and dealers' intermediation capacity, respectively,

[26]The risky asset in our model is an unlevered claim on the aggregate output, corresponding to a portfolio of (levered) equity and a risky bond. This implies that we will likely underestimate the response of the risk premium to a large negative shock. As shown in Merton (1974), the volatility of the corporate bond responds to a negative shock by more than the volatility of the firm value. By focusing on an unlevered claim on the firm, we obtain a lower bound on the risk premium response of corporate bonds.

Table 1. Parameter values

This table reports the parameter values used in calibrating the model.

	Parameter	Choice
	Preferences & distribution	
ψ	Elasticity of intertemporal substitution	1.5
γ_1	Risk aversion of high-type	1.8
γ_2	Risk aversion of low-type	10.25
ρ	Rate of time preference	0.038
κ	Agents' entry/exit rate	0.32
ν	Share of high-type agents	0.01
	Technology	
μ	Endowment growth rate	0.02
σ	Endowment volatility	0.04
	Trading	
η	Concavity of the matching function	0.89
χ	Portfolio adjustment cost parameter	1.40
$\overline{\alpha}$	Efficiency of the matching function	1.00
$\overline{d}$	Dealers' intermediation capacity	0.64

cannot be separately identified. So, we normalize $\overline{\alpha}$ to one and calibrate $\overline{d}$ to match the contact frequency of customers and dealers in the corporate bond market. Following Hugonnier, Lester, and Weill (2020), we calibrate $\overline{d}$ so that a customer contacts a dealer every five days.

4.3 Quantitative impacts of search frictions

In Table 2, we compare the unconditional moments and targets in the data and the model with and without search frictions. Comparing columns (1) and (2), the baseline model with search frictions closely matches the three target moments from the corporate bond market. The model generates average transaction costs of 38 bps, an annual turnover of 0.2, and a customer-dealer contact frequency of 5.4 days. Moreover, the model with frictions generates an equity premium of 7.1%, a real interest rate of 1.2%, and a leverage of 5.3 for the high-type agents. The baseline model creates a substantial endogenous volatility through amplification of the exogenous risk. In contrast, the frictionless model in column (3) can generate a risk premium of only 2.4%, about three times lower than the one in the baseline model.

The result above shows that search frictions are quantitatively important for the model to generate a level of risk premia consistent with the data. This conclusion contrasts with the findings of Constantinides

Table 2. Unconditional moments and targets from the data and the model

This table presents the unconditional moments from the data and the model with and without search frictions. In column (2), we present the results in our baseline model with search frictions. In column (3), we shut down search frictions and present results in a frictionless model.

Moment	Data/Target	Baseline Model	Frictionless Model
	(1)	(2)	(3)
Average corp. bond trading costs (bps)	40	38	0
Annual turnover for customer-dealer corp. bond trades (%)	0.2	0.2	-
Average customer-dealer contact frequency (days)	5	5.4	-
Average equity premium (%)	7.0	7.1	2.4
Average real interest rate (%)	1.0-1.5	1.2	3.8
Average wealth-weighted risk aversion	≤10	9.5	9.5
Leverage of high-type agents	5	5.3	5.3

(1986) who shows that liquidity frictions have only a limited effect on asset prices in a partial equilibrium model with exogenous transaction costs. In his model, investors substantially reduce their trading activity in the presence of frictions so that they can economize on transaction costs leading to a small liquidity premium.[27] Constantinides's (1986) result, however, only captures the direct effect of transaction costs. In our setting, the reduction in trading activity caused by transaction costs leads to asset misallocation, as investors' portfolio holdings deviate from their target. This asset misallocation by itself has an impact on asset prices. This effect is not present in a partial equilibrium setting, highlighting the importance of adopting a general equilibrium framework to fully capture the implications of liquidity frictions.[28]

4.4 Equilibrium dynamics after a large shock

Next, we consider the quantitative impact of a large negative shock on the economy with heterogeneous risk aversions. We initialize the economy at the stochastic steady state; that is, the state variables are given by $(x, s) = (\overline{x}, \overline{s})$. We consider a shock that increases the bid-ask spread by a factor of 10, roughly the magnitude observed during the COVID-19 crisis. Figure 6 shows the conditional expectation of the risk premium, the interest rate, the bid-ask spread, and trading volume t weeks ahead. The grey area represents one standard deviation bands computed using each variable's conditional distribution t periods ahead.

[27]Note that we also find a substantial reduction in trading activity in our model with search frictions. In particular, we obtain trading strategies of bounded variation while trading in frictionless models has unbounded variation.

[28]The emphasis on the equilibrium implications of trading costs is also shared by Lo, Mamaysky, and Wang (2004), who study the impact of exogenous fixed costs on trading volume and asset prices.

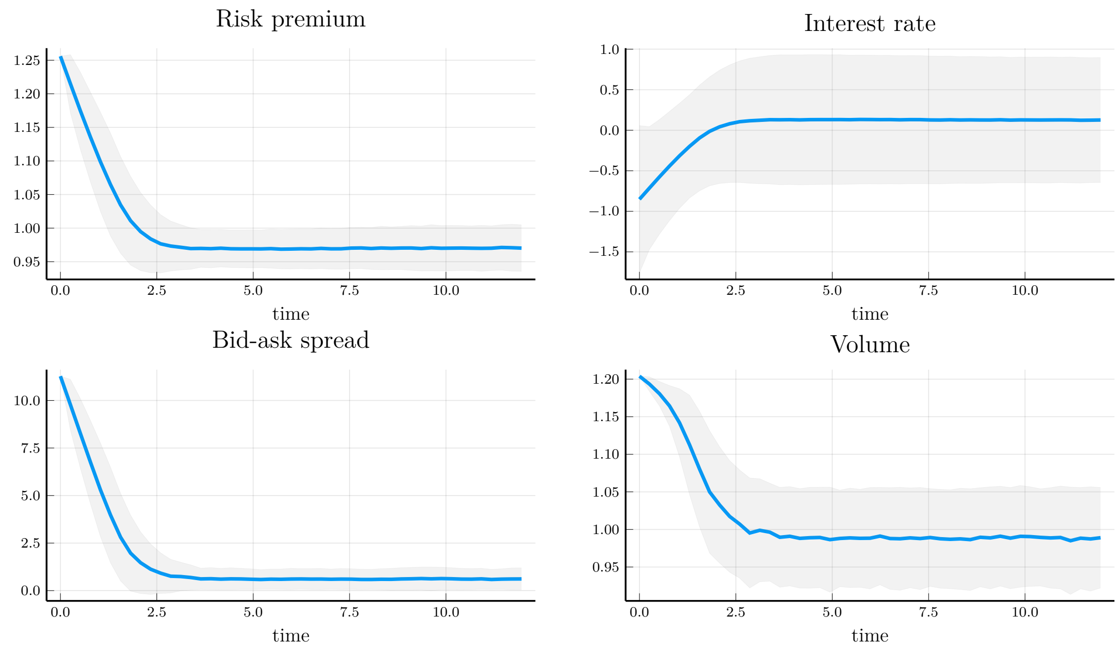

Figure 6. This figure shows impulse-response functions of the risk premium, interest rates, bid-ask spreads, and trading volume in response to a negative aggregate shock. Risk premium, bid-ask spread, and volume are reported as a ratio to their value at the stochastic steady state. Interest rate is reported as the difference to its level in the stochastic steady state. Time is expressed in weeks. The shaded areas indicate one standard deviation bands.

We observe that the model can capture the market reaction during the COVID-19 crisis discussed in Section 1: the risk premium, bid-ask spreads, and trading volume increase, while the interest rate declines. As discussed in Subsection 3.3, following a large adverse shock, portfolio dispersion and asymmetry endogenously increase. The increase in portfolio asymmetry explains the rise in the risk premium, while higher dispersion leads to a reduction of the real interest rate. Investors have stronger incentives to rebalance their portfolio with higher dispersion, leading to increased trading volume. Higher demand for transaction services results in a rise in trading costs given the limited intermediation capacity.

The model can also quantitatively capture the market liquidity dynamics described in Section 1, while generating a substantial movement in asset prices. The bid-ask spread increases by 10-fold, as the volume is about 20% higher than its pre-crisis level, consistent with the evidence in Kargar et al. (2020). The shock leads to a 25% increase in the risk premium, and interest rates go to approximately zero on impact.

5 The Case of an Infinite-dimensional State Space

So far, we have considered the characterization of the economy under the big-family assumption, Assumption 1, so investors can diversify the order execution risk. The main advantage of making this assumption is that the joint distribution of wealth and portfolios can be easily summarized by only two variables: the wealth share x and the share of risky assets s of type-1 investors. In practice, however, investors in OTC markets face uncertainty on whether they can trade immediately. It is important then to consider the case where investors bear the order execution risk in equilibrium. In this case, even when investors start at the same initial conditions, they may end up at different levels of wealth and risky asset shares depending on when they could trade. The aggregate state variables in the economy with order execution risk will be an infinite-dimensional object: the entire joint distribution of wealth and asset shares, which we denote by $G(x, s)$.

Such a problem is typically intractable using standard solution methods. The state-global perturbation techniques discussed above apply in this more general case as well. This enables us to extend our results to an economy with order execution risk. In particular, we show below that the results in Proposition 3 also hold in this more general environment. Given the individual trading behavior, from the market-clearing condition for the risky asset, we obtain the following expression:[29]

$$\underbrace{\int_{\frac{\tilde{\pi}(X)}{\gamma\sigma^2}}^{\infty} \left(\omega - \frac{\tilde{\pi}(X)}{\gamma\sigma^2}\right)^{\frac{1+\eta}{1-\eta}} dH(\omega \mid X)}_{\mathcal{S}(\tilde{\pi}|X)} = \underbrace{\int_0^{\frac{\tilde{\pi}(X)}{\gamma\sigma_R^2(X)}} \left(\frac{\tilde{\pi}(X)}{\gamma\sigma^2} - \omega\right)^{\frac{1+\eta}{1-\eta}} dH(\omega \mid X)}_{\mathcal{D}(\tilde{\pi}|X)},$$

where $H(\omega \mid X)$ is the distribution of portfolio shares in the economy induced by $G(x, s)$.

Note that $\mathcal{D}(\cdot \mid X)$ is increasing, convex, and satisfies $\mathcal{D}(0 \mid X) = 0$ and $\mathcal{D}(\tilde{\pi} \mid X) \to \infty$ as $\tilde{\pi} \to \infty$. Analogously, $\mathcal{S}(\cdot \mid X)$ is decreasing, convex, and satisfies $\mathcal{S}(0 \mid X) > 0$ and $\mathcal{S}(\tilde{\pi} \mid X) \to 0$ as $\tilde{\pi} \to \infty$. Therefore, there exists a unique value $\tilde{\pi}(X)$ that satisfies $\mathcal{D}(\tilde{\pi}(X) \mid X) = \mathcal{S}(\tilde{\pi}(X) \mid X)$. In contrast to the two-type case, changes in the risk premium affect not only the magnitude of buy and sell orders but also the mass of agents who are buyers and sellers. Therefore, the aggregate conditions now affect the magnitude of the selling pressure via the extensive margin. In the two-type case, the selling pressure depends only on the concavity of the matching function and the share of type-1 agents.

In Proposition 8, we extend the results on interest rates and market liquidity for a general distribution $G(x, s)$.

Proposition 8 (Infinite-dimensional state space)**.** *Suppose $\rho + (\gamma - 1)\mu > 0$ and investors are subject to order*

[29]For ease of exposition, we again consider the case of homogeneous preferences. We provide a complete derivation of the case with heterogeneous preferences and order execution risk in Appendix A.9.

execution risk.

a. Interest rate is determined by the following expression:

$$r_t = \rho + \gamma\mu - \frac{\gamma(\gamma+1)}{2} \int \left(\frac{s}{x}\right)^2 x dG(x,s)\tilde{\sigma}_R^2(X)\epsilon + \mathcal{O}(\epsilon^2).$$

b. Intermediation fee is given by the following expression:

$$\tilde{\phi}(x, s \mid X) = \eta\tilde{\Omega}(x, s \mid X),$$

where $\tilde{\Omega}(x, s \mid X) = p^\star \frac{\gamma\sigma^2}{r^\star - \mu}\left[\frac{\pi(X)}{\gamma\sigma^2} - \frac{s}{x}\right]$.

c. Dealers' value is determined by the following expression:

$$\tilde{v}_{d,t} = \left[\frac{1-\eta}{\chi\eta\bar{d}} \int \left(\frac{|\tilde{\Omega}(x,s)|}{p^\star}\right)^{\frac{2}{1-\eta}} dG(x,s)\right]^{\frac{1-\eta}{1+\eta}}.$$

Proof. See Appendix A.9. □

Similar to the results under Assumption 1, the interest rate is decreasing in portfolio dispersion. Formally, a mean-preserving spread of the (wealth-weighted) distribution of portfolio shares leads to a reduction of the interest rates. The intuition is the same as in the previous case: given the concavity of the consumption function on the portfolio share, an increase in portfolio dispersion leads to a decline in interest rates to compensate for the drop in aggregate demand.

Intermediation fees depend on the distance of the investor's portfolio to the (myopic) target. Portfolio dispersion also plays an important role in determining the dealers' profitability. Formally, a mean-preserving spread of the distribution of portfolios leads to an increase in the dealers' value function $\tilde{v}_{d,t}$. An increase in portfolio dispersion implies a higher demand for trading. Given the limited intermediation capacity, this high demand leads to an increase in dealers' profits in equilibrium.

6 Conclusion

The COVID-19 crisis and the unprecedented interventions by the Federal Reserve highlight the need for a unified framework to study the impacts of large adverse shocks on asset prices and market liquidity and to evaluate the implication of the policy aimed to improve market conditions.

This paper presents an asset pricing model with risk-averse investors, unrestricted asset holdings, and (competitive) search frictions that allow us to jointly study the risk premium, risk-free rate, and market liquidity in general equilibrium. To tackle this highly intractable problem, we propose a new methodology, state-global perturbations, that allows us to characterize the equilibrium analytically despite the presence of an infinite-dimensional state space. After a large negative shock, the calibrated model generates asset pricing and trading dynamics consistent with the empirical evidence during the market turmoil in March 2020.

While this paper is an important first step to introduce risk premia in workhorse search models, much work remains to be done. One potentially interesting topic is to examine the implications of balance sheet costs and risk considerations to understand the source of dealers' unwillingness to intermediate during crisis periods. We also believe that our state-global perturbation techniques can be potentially applied to various economic settings in which a high-dimensional state space makes existing solution methods less suitable. Incomplete market models with rich heterogeneity, financial frictions, and aggregate risk is a potential application.

References

Acharya, Viral V., and Lasse Heje Pedersen, 2005, Asset pricing with liquidity risk, *Journal of Financial Economics* 77, 375–410.

Alvarez, Fernando, and Andrew Atkeson, 2018, The risk of becoming risk averse: A model of asset pricing and trade volumes, Working paper, University of Chicago and UCLA.

Bansal, Ravi, and Amir Yaron, 2004, Risks for the long run: A potential resolution of asset pricing puzzles, *Journal of Finance* 59, 1481–1509.

Barro, Robert J., 2006, Rare Disasters and Asset Markets in the Twentieth Century, *The Quarterly Journal of Economics* 121, 823–866.

Breeden, Douglas T., 1979, An intertemporal asset pricing model with stochastic consumption and investment opportunities, *Journal of Financial Economics* 7, 265–296.

Brock, William A, and David Gale, 1969, Optimal growth under factor augmenting progress, *Journal of Economic Theory* 1, 229–243.

Brunnermeier, Markus K., and Lasse Heje Pedersen, 2009, Market Liquidity and Funding Liquidity, *Review of Financial Studies* 22, 2201–2238.

Brunnermeier, Markus K., and Yuliy Sannikov, 2014, A macroeconomic model with a financial sector, *The American Economic Review* 104, 379–421.

Buss, Adrian, and Bernard Dumas, 2019, The dynamic properties of financial-market equilibrium with trading fees, *Journal of Finance* 74, 795–844.

Constantinides, George M., 1986, Capital market equilibrium with transaction costs, *Journal of Political Economy* 94, 842–862.

Davis, Mark H.A., and Andrew R. Norman, 1990, Portfolio selection with transaction costs, *Mathematics of Operations Research* 15, 676–713.

Dick-Nielsen, Jens, 2014, How to clean enhanced TRACE data, Working paper, CBS.

DiTella, Sebastian, 2017, Uncertainty shocks and balance sheet recessions, *Journal of Political Economy* 125, 2038–2081.

Duffie, Darrell, 2020, Still the world's safe haven? Redesigning the US treasury market after the COVID-19 crisis, Hutchins Center Working Paper #62.

Duffie, Darrell, and Larry G. Epstein, 1992, Stochastic Differential Utility, *Econometrica* 60, 353–394.

Duffie, Darrell, Nicolae Gârleanu, and Lasse Heje Pedersen, 2005, Over-the-counter markets, *Econometrica* 73, 1815–1847.

Dumas, Bernard, 1989, Two-person dynamic equilibrium in the capital market, *Review of Financial Studies* 2, 157–188.

Dumas, Bernard, and Elisa Luciano, 1991, An exact solution to a dynamic portfolio choice problem under transactions costs, *Journal of Finance* 46, 577–595.

Epstein, Larry G., and Stanley E. Zin, 1989, Substitution, Risk Aversion, and the Temporal Behavior of Consumption and Asset Returns: A Theoretical Framework, *Econometrica* 57, 937–969.

Falato, Antonio, Itay Goldstein, and Ali Hortaçsu, 2020, Financial fragility in the COVID-19 crisis: The case of investment funds in corporate bond markets, Working paper, FRB, Wharton, and University of Chicago.

Fernández-Villaverde, Jesús, Juan Francisco Rubio-Ramírez, and Frank Schorfheide, 2016, Solution and estimation methods for dsge models, in *Handbook of Macroeconomics*, volume 2, 527–724 (Elsevier).

Gabaix, Xavier, and Ralph S. J. Koijen, 2020, In search of the origins of financial fluctuations: The inelastic markets hypothesis, Working paper, Harvard and University of Chicago.

Gârleanu, Nicolae, 2009, Portfolio choice and pricing in illiquid markets, *Journal of Economic Theory* 144, 532–564.

Gârleanu, Nicolae, and Stavros Panageas, 2015, Young, old, conservative, and bold: The implications of heterogeneity and finite lives for asset pricing, *Journal of Political Economy* 123, 670–685.

Gârleanu, Nicolae, and Lasse Heje Pedersen, 2013, Dynamic trading with predictable returns and transaction costs, *Journal of Finance* 68, 2309–2340.

Gârleanu, Nicolae, and Lasse Heje Pedersen, 2016, Dynamic portfolio choice with frictions, *Journal of Economic Theory* 165, 487–516.

Gertler, Mark, and Nobuhiro Kiyotaki, 2010, Financial intermediation and credit policy in business cycle analysis, in *Handbook of Monetary Economics*, volume 3, 547–599 (Elsevier).

Grossman, Sanford J., and Merton H. Miller, 1988, Liquidity and market structure, *Journal of Finance* 43, 617–633.

Haddad, Valentin, Alan Moreira, and Tyler Muir, 2020, When selling becomes viral: Disruptions in debt markets in the COVID-19 crisis and the Fed's response, *Review of Financial Studies* , Forthcoming.

Heaton, John, and Deborah J. Lucas, 1996, Evaluating the effects of incomplete markets on risk sharing and asset pricing, *Journal of Political Economy* 104, 443–487.

Hendershott, Terrence, and Ananth Madhavan, 2015, Click or call? auction versus search in the over-the-counter market, *Journal of Finance* 70, 419–447.

Hugonnier, Julien, Benjamin Lester, and Pierre-Olivier Weill, 2020, Frictional Intermediation in Over-the-Counter Markets, *The Review of Economic Studies* 87, 1432–1469.

Judd, Kenneth L., and Sy-Ming Guu, 2001, Asymptotic methods for asset market equilibrium analysis, *Economic Theory* 18, 127–157.

Kargar, Mahyar, Benjamin Lester, David Lindsay, Shuo Liu, Pierre-Olivier Weill, and Diego Zúñiga, 2020, Corporate bond liquidity during the COVID-19 crisis, Technical report, National Bureau of Economic Research.

Koijen, Ralph S. J., and Motohiro Yogo, 2019, A demand system approach to asset pricing, *Journal of Political Economy* 127, 1475–1515.

Kyle, Albert S., 1985, Continuous auctions and insider trading, *Econometrica* 1315–1335.

Lagos, Ricardo, and Guillaume Rocheteau, 2009, Liquidity in asset markets with search frictions, *Econometrica* 77, 403–426.

Lester, Benjamin, Guillaume Rocheteau, and Pierre-Olivier Weill, 2015, Competing for order flow in otc markets, *Journal of Money, Credit and Banking* 47, 77–126.

Li, Dan, and Norman Schürhoff, 2019, Dealer networks, *Journal of Finance* 74, 91–144.

Lo, Andrew W., Harry Mamaysky, and Jiang Wang, 2004, Asset prices and trading volume under fixed transactions costs, *Journal of Political Economy* 112, 1054–1090.

Longstaff, Francis A., 2001, Optimal Portfolio Choice and the Valuation of Illiquid Securities, *Review of Financial Studies* 14, 407–431.

Longstaff, Francis A., 2009, Portfolio Claustrophobia: Asset Pricing in Markets with Illiquid Assets, *American Economic Review* 99, 1119–1144.

Longstaff, Francis A., and Jiang Wang, 2012, Asset Pricing and the Credit Market, *Review of Financial Studies* 25, 3169–3215.

Lucas, Robert E., 1990, Liquidity and interest rates, *Journal of Economic Theory* 50, 237–264.

Luenberger, David G, 1997, *Optimization by vector space methods* (John Wiley & Sons).

Ma, Yiming, Kairong Xiao, and Yao Zeng, 2020, Mutual fund liquidity transformation and reverse flight to liquidity, Working paper, Columbia and Wharton.

Merton, Robert C., 1971, Optimum consumption and portfolio rules in a continuous-time model, *Journal of Economic Theory* 373–413.

Merton, Robert C., 1974, On the pricing of corporate debt: The risk structure of interest rates, *Journal of Finance* 29, 449–470.

O'Hara, Maureen, and Xing (Alex) Zhou, 2020, Anatomy of a liquidity crisis: Corporate bonds in the Covid-19 crisis, *Journal of Financial Economics* , Forthcoming.

Protter, Philip E., 2004, *Stochastic Integration and Differential Equations*, second edition (Springer-Verlag).

Schmitt-Grohé, Stephanie, and Martın Uribe, 2004, Solving dynamic general equilibrium models using a second-order approximation to the policy function, *Journal of Economic Dynamics and Control* 28, 755–775.

Stoll, Hans R., 1978, The supply of dealer services in securities markets, *Journal of Finance* 33, 1133–1151.

Weill, Pierre-Olivier, 2020, The search theory of over-the-counter markets, *Annual Review of Economics* 12.

Wright, Randall, Philipp Kircher, Benoît Julien, and Veronica Guerrieri, 2019, Directed search and competitive search: A guided tour, *Journal of Economic Literature* .

Appendix

A Proofs

A.1 Proof of Proposition 1

Proof. ***Step 1: Envelope condition of the HJB with respect to S.*** Differentiating the Hamilton-Jacobi-Bellman equation (12) with respect to S, we obtain:

$$\rho V_S = V_W p(\mu_R - r) + V_{WW}\sigma_R^2 p^2 S + V_{WX}\sigma_X p \sigma_R + \mathcal{D}V_S, \tag{A.1}$$

where $\mathcal{D}$ represents the Dynkin operator, and we omit the investor i and time subscripts.

Step 2: Feymann-Kac. Applying the Feymann-Kac solution to equation (A.1), we obtain:

$$V_{S,t} = \mathbb{E}_t\left[\int_t^\infty e^{-\rho(s-t)} V_{W,s} p_s \sigma_{R,s}^2 \left(\frac{\mu_{R,s} - r_s}{\sigma_{R,s}^2} + \frac{V_{WW,s}}{V_{W,s}} p_s S_s + \frac{V_{WX,s}}{V_{W,s}}\frac{\sigma_{X,s}}{\sigma_{R,s}}\right) ds\right].$$

Dividing the expression above by $V_{W,t}$ and using the condition $V_W = C^{-\gamma}$, we obtain

$$\Omega(W_t, S_t, X_t) = \mathbb{E}_t\left[\int_t^\infty \frac{e^{-\rho(s-t)}C_s^{-\gamma}}{C_t^{-\gamma}} p_s \gamma_s^V \sigma_{R,s}^2 \left(Target_s - \frac{p_s S_s}{W_s}\right) ds\right],$$

where we used the definitions:

$$Target_s \equiv \frac{\mu_{R,s} - r_s}{\gamma_s^V \sigma_{R,s}^2} + \frac{V_{WX,s}}{\gamma_s^V V_{W,s}}\frac{\sigma_{X,s}}{\sigma_{R,s}}, \qquad \gamma_s^V \equiv -\frac{V_{WW,s}W_s}{V_{W,s}}.$$

□

A.2 Proof of Proposition 2

Proof. ***Step 1:*** We can re-write (A.1) as follows:

$$0 = \frac{\mathbb{E}_t[d(e^{-\rho t}V_{S,t})]}{e^{-\rho t}V_{W,t}p_t} + (\mu_{R,t} - r_t)dt + \left[\frac{V_{WW,t}}{V_{W,t}}\sigma_{R,t} pS + \frac{V_{WX,t}}{V_{W,t}}\sigma_{X,t}\right]\sigma_{R,t}dt, \tag{A.2}$$

where we omit the investor subscript i.

Step 2: Applying Ito's lemma to $V_{W,t}$, we obtain:

$$dV_{W,t} - \mathbb{E}_t[dV_{W,t}] = [V_{WW,t}\sigma_{R,t}p_tS_t + V_{WX,t}\sigma_{X,t}]\, dZ_t. \tag{A.3}$$

Step 3: Combining (A.2) and (A.3), and the optimality condition $C_t^{-\gamma} = V_{W,t}$, we obtain:

$$(\mu_{R,t} - r_t)dt = \gamma \frac{dC_{j,t}}{C_{j,t}} \frac{dp_t}{p_t} - \frac{\mathbb{E}_t[d(e^{-\rho t} C_{j,t}^{-\gamma} \Omega_{j,t})]}{e^{-\rho t} C_{j,t}^{-\gamma} p_t}.$$

Step 4: Aggregating across families, we obtain:

$$(\mu_{R,t} - r_t)dt = \gamma \frac{dC_t}{C_t} \frac{dp_t}{p_t} - \sum_{j=1}^{2} \omega_j^c \frac{\mathbb{E}_t[d(e^{-\rho t} C_{j,t}^{-\gamma} \Omega_{j,t})]}{e^{-\rho t} C_{j,t}^{-\gamma} p_t},$$

where C_t denotes investors' aggregate consumption, $C_{j,t}$ is the consumption of family j, and $\omega_{1,t}^c \equiv \nu C_{1,t}/C_t$, $\omega_{2,t} \equiv (1-\nu)C_{2,t}/C_t$ denote the consumption share of families 1 and 2, respectively.

□

A.3 Proof of Lemma 1

Proof. We first consider the investors' problem when $\epsilon = 0$, given prices. Then, we solve the dealers' problem and solve for equilibrium prices.

Step 1: Value Function and Policy Functions. Consider the economy where $\epsilon = 0$. We assume initially that the interest rate $r^\star(X)$ and price-dividend ratio $q^\star(X) = \frac{p(X_t)}{Y_t}$ are constant. Moreover,

$$\mu_X^\star(X) = \sigma_X^\star(X) = \sigma_R^\star(X) = \mu_R^\star(X) - r^\star(X) = 0.$$

We show below that these properties hold in equilibrium. In this case, the investors' problem can be written as:

$$\rho V = \max_{C,n,\theta} \frac{C^{1-\gamma}}{1-\gamma} + V_W[rW - \frac{1}{2}p\chi n^2 - C - \upsilon_d \theta p|n|] + V_S n\alpha(\theta).$$

We then guess-and-verify that the value function is given by $V^\star(W, S, X) = A\frac{W^{1-\gamma}}{1-\gamma}$. Since $V^\star_S = 0$, then it is optimal to have $n^\star = 0$. Plugging $n^\star = 0$ into the expression above, we obtain that the investor is indifferent between any value of θ. Consumption is given by $C^\star(W, S, X) = A^{-\frac{1}{\gamma}} W$. Plugging the value of $n^\star$ and $C^\star$ into the HJB equation, we obtain:

$$A^{-\frac{1}{\gamma}} = \frac{1}{\gamma}\rho + \left(1 - \frac{1}{\gamma}\right) r^\star. \tag{A.4}$$

Step 2: Asset Prices. Using the fact that $n^\star_{i,t} = \Pi^\star_{d,t} = 0$ and the expression for the consumption-wealth ratio, we can write the market-clearing condition for goods as follows:

$$\int_0^1 A^{-\frac{1}{\gamma}} W_{i,t} di = Y_t, \quad \int_0^1 W_{i,t} di = p_t. \tag{A.5}$$

Using the definition of the price dividend ratio, the consumption wealth ratio (A.4), market clearing of goods and risky assets (A.5), and we obtain the price-dividend ratio when $\epsilon = 0$ which is given by:

$$q^\star = \frac{1}{\frac{1}{\gamma}\rho + \left(1 - \frac{1}{\gamma}\right) r^\star}. \tag{A.6}$$

Given that the price-dividend ratio is constant, we know that $\frac{dp_t}{p_t} = \frac{dY_t}{Y_t} = \mu dt$. Using the fact that the risk premium is equal to zero, we obtain the risk free rate as:

$$\frac{Y_t}{p_t} + \frac{1}{p_t}\frac{dp_t}{dt} = \frac{1}{q^\star} + \mu = r^\star \implies r^\star = \rho + \gamma\mu.$$

The consumption-wealth ratio and the dividend yield are then given by:

$$A^{-\frac{1}{\gamma}} = \frac{1}{q^\star} = \rho + (\gamma - 1)\mu.$$

Step 3: Dealers' Problem. Consider a contract $\varsigma = (n, \phi)$, where $n \neq 0$. Given that there is no value of θ which will generate a strict improvement over $V^\star(W, S, X)$, then $\theta_t(\varsigma) = \infty$ and $\frac{\alpha(\theta(\varsigma))}{\theta(\varsigma))} = 0$, according to (8). This implies that the expected profit of the dealer is zero and the capacity constraint is not binding, that is, $v^\star_d = 0$.

Step 4: Aggregate State Variables. The law of motion of x_t, equation (25), satisfies:

$$dx = x\left[r^\star - \mu - A^{-\frac{1}{\gamma}}\right]dt,$$

where we used $\sigma_x = \pi = n_1 = 0$. Using the fact that $r^\star - \mu = A^{-\frac{1}{\gamma}}$, we obtain $\mu_x^\star = 0$. Finally, since $n_1^\star = 0$, we have that $\mu_s^\star = 0$. Therefore, the drift and diffusion terms of (x, s) are equal to zero when $\epsilon = 0$. □

A.4 Proof of Proposition 3

Proof. ***Step 0: Standardizing the HJB.*** The HJB equation is given by:

$$\rho V = \max_{C,n,\theta} \frac{C^{1-\gamma}}{1-\gamma} + V_W\left[rW + \pi pS - \frac{1}{2}p\chi n^2 - C - v_d\theta p|n|\right] + V_S n\alpha(\theta) + V_X\mu_X$$
$$+ \frac{1}{2}V_{WW}\sigma_R^2(pS)^2 + pS\sigma_R V_{WX}\sigma_X + \frac{1}{2}\sigma_X' V_{XX}\sigma_X. \quad \text{(A.7)}$$

We guess and verify that the value function can be written as:

$$V(W_t, S_t, X_t) = Y_t^{1-\gamma}\hat{V}\left(\frac{W_t}{Y_t}, S_t, \hat{X}_t\right), \quad \text{(A.8)}$$

where $\hat{X}_t \equiv (x_t, s_t)$ denote the aggregate state variables besides Y_t. Inserting (A.8) into (A.7), after several steps of algebra, we can write the HJB equation in terms of the scaled variables:

$$\rho^\star\hat{V} = \max_{\hat{C},n,\theta} \frac{\hat{C}^{1-\gamma}}{1-\gamma} + \hat{V}_{\hat{W}}\left[\left(r + \gamma\sigma^2 - \mu\right)\hat{W} + (\pi - \gamma\sigma\sigma_R)qS - \frac{1}{2}q\chi n^2 - \hat{C} - v_d\theta q|n|\right] + \hat{V}_S n\alpha(\theta)$$
$$+ \hat{V}_{\hat{X}}\left(\mu_{\hat{X}} + (1-\gamma)\sigma\sigma_{\hat{X}}\right) + \frac{1}{2}\hat{V}_{\hat{W}\hat{W}}\left(\sigma_R qS - \sigma\hat{W}\right)^2 + (\sigma_R qS - \sigma\hat{W})\hat{V}_{\hat{W}\hat{X}}\sigma_{\hat{X}} + \frac{1}{2}\sigma_{\hat{X}}'\hat{V}_{\hat{X}\hat{X}}\sigma_{\hat{X},t} + \frac{\gamma(\gamma-1)\sigma^2}{2}\hat{V}, \quad \text{(A.9)}$$

where $\rho^\star \equiv \rho + (\gamma - 1)\mu$, $\hat{W}_t \equiv \frac{W_t}{Y_t}$, and $\hat{C}_t \equiv \frac{C_t}{Y_t}$.

Step 1: Write Aggregate Variables in Terms of ϵ. A first-order approximation of the aggregate variables in ϵ gives the expressions:

$$\mu_X(X,\epsilon) = \mu_X^\star(X) + \tilde{\mu}_X(X)\epsilon + o(\epsilon), \quad \mu_q(X,\epsilon) = \mu_q^\star(X) + \tilde{\mu}_q(X)\epsilon + o(\epsilon), \quad \mu_R(X,\epsilon) = \mu_R^\star(X) + \tilde{\mu}_R(X)\epsilon + o(\epsilon)$$
$$r(X,\epsilon) = r^\star(X) + \tilde{r}(X)\epsilon + o(\epsilon), \quad q(X,\epsilon) = q^\star(X) + \tilde{q}(X)\epsilon + o(\epsilon), \quad \pi(X,\epsilon) = \pi^\star(X) + \tilde{\pi}(X)\epsilon + o(\epsilon),$$
$$v_d(X,\epsilon) = v_d^\star(X) + \tilde{v}_d(X)\epsilon + o(\epsilon), \quad \sigma_R(X,\epsilon) = \sigma_R^\star(X) + \tilde{\sigma}_R(X)\sqrt{\epsilon} + \mathcal{O}(\epsilon), \quad \sigma_X(X,\epsilon) = \sigma_X^\star(X) + \tilde{\sigma}_X(X)\sqrt{\epsilon} + \mathcal{O}(\epsilon),$$

where, for instance, $r^*(X) = r(X,0)$ and $\tilde{r}(X) = r_\epsilon(X,0)$. From Lemma 1, recall that we know that:

$$v_d^\star(X) = \sigma_R^\star(X) = \sigma_X^\star(X) = \pi^\star(X) = \mu_q^\star(X) = \mu_X^\star(X) = 0,$$
$$r^\star(X) = \rho + \gamma\mu, \quad \mu_R^\star(X) = \rho + \gamma\mu.$$

Next, plugging the first-order approximation of the aggregate variables into (A.9), we write the HJB equation as follows:

$$\begin{aligned}\rho^\star \hat{V} = {} & \frac{\hat{C}^{1-\gamma}}{1-\gamma} + \hat{V}_{\hat{W}}[(r^\star + \tilde{r}\epsilon + \gamma\sigma^2\epsilon - \mu)\hat{W} + (\tilde{\pi}\epsilon - \gamma\sigma\tilde{\sigma}_R\epsilon)q^\star S - 0.5q^\star\chi n^2 - \hat{C} - \tilde{v}_d\epsilon\theta q^\star|n|] + \hat{V}_S n\alpha(\theta) \\ & + V_{\hat{X}}\left(\tilde{\mu}_{\hat{X}}\epsilon - (\gamma-1)\sigma\tilde{\sigma}_{\hat{X}}\epsilon\right) + \hat{V}_{\hat{W}\hat{W}}\left(\tilde{\sigma}_R^2\frac{(q^\star S)^2}{2}\epsilon - \hat{W}q^\star S\sigma\tilde{\sigma}_R\epsilon + \frac{\sigma^2\epsilon}{2}\hat{W}^2\right) \\ & + V_{\hat{W}\hat{X}}\tilde{\sigma}_{\hat{X}}\left(q^\star S\tilde{\sigma}_{R,t} - \hat{W}\sigma\right)\epsilon + \frac{1}{2}\tilde{\sigma}_{\hat{X}}' V_{\hat{X}\hat{X}}\tilde{\sigma}_{\hat{X}}\epsilon + \frac{\gamma(\gamma-1)}{2}\hat{V}\sigma^2\epsilon + o(\epsilon).\end{aligned}$$

Step 2: Derivative of the Value Function at $\epsilon = 0$. Taking the derivative of the expression above with respect to ϵ, and evaluating the resulting expression at $\epsilon = 0$ we obtain:[1]

$$\begin{aligned}\rho^\star \hat{V}_\epsilon = {} & \hat{V}^\star_{\hat{W}}[(\tilde{r} + \gamma\sigma^2)\hat{W} + (\tilde{\pi} - \gamma\sigma\tilde{\sigma}_R)\, q^\star S] + \hat{V}_{W,\epsilon}[(r^\star - \mu)\hat{W} - A^{-\frac{1}{\gamma}}\hat{W}] \\ & + \frac{1}{2}\hat{V}^\star_{\hat{W}\hat{W}}\left(\tilde{\sigma}_R^2(q^\star S)^2 - 2\hat{W}q^\star S\sigma\tilde{\sigma}_R + \sigma^2\hat{W}^2\right) + \frac{\gamma(\gamma-1)}{2}\hat{V}^\star\sigma^2, \end{aligned} \tag{A.10}$$

where we used the fact that $V_S^\star = V_{\hat{X}}^\star = V_{W\hat{X}}^\star = V_{\hat{X}\hat{X}}^\star = n^\star = 0$. Note that the expression above involves $\hat{V}_\epsilon$ and its derivative in W, $\hat{V}_{W,\epsilon}$. If the term multiplying $\hat{V}_{W,\epsilon}$ were different from zero, then the perturbation would reduce the problem of solving a non-linear PDE into the solution of a linear ordinary differential equation. Given the normalization of the value function, the term multiplying $\hat{V}_{W,\epsilon}$ is actually equal to

[1] Note that the value function, and its derivatives, and the control variables, are also functions of ϵ. After taking the derivative with respect to ϵ, we evaluate the resulting expression at $\epsilon = 0$. This last step yields terms as, for example, $\hat{V}^\star_{\hat{W}}$, which is shorthand notation for $\hat{V}_{\hat{W}}(\hat{W}, S, \hat{X}; 0)$, and was computed in Lemma 1 (without the normalization).

zero, as $A^{-\frac{1}{\gamma}} = r^\star - \mu$. In this case, the problem simplifies further to a linear (algebraic) equation in $\hat{V}_\epsilon$, whose solution is given by:

$$\hat{V}_\epsilon(\hat{W}, S, \hat{X}; 0) = \frac{A\hat{W}^{1-\gamma}}{\rho^\star}\left[\tilde{r}(\hat{X}) + \tilde{\pi}(X)\frac{q^\star S}{\hat{W}} - \frac{\gamma}{2}\tilde{\sigma}_R^2(\hat{X})\left(\frac{q^\star S}{\hat{W}}\right)^2\right].$$

Converting back to the expression in levels (recall that we standardized: by Y_t), we have that:

$$\tilde{V}(W, S, X) = \frac{AW^{1-\gamma}}{\rho^\star}\left[\tilde{r}(X) + \tilde{\pi}(X)\frac{p^\star(X)S}{W} - \frac{\gamma}{2}\tilde{\sigma}_R^2(X)\left(\frac{p^\star(X)S}{W}\right)^2\right]. \tag{A.11}$$

Step 3: Consumption Policy Function. Given the expression for the value function, we can solve for the policy functions. Consumption is given by:

$$\begin{aligned} C(W, S, X; \epsilon) &= V_W(W, S, X; \epsilon)^{-\frac{1}{\gamma}} \\ &= V_W^\star(W, S, X)^{-\frac{1}{\gamma}} - \frac{1}{\gamma}V_W^\star(W, S, X)^{-\frac{1+\gamma}{\gamma}}\tilde{V}_W(W, S, X)\epsilon + \mathcal{O}(\epsilon^2), \end{aligned} \tag{A.12}$$

where the first equality follows from the first order condition for consumption, and the second equality is a first order expansion of $V_W(W, S, X; \epsilon)^{-\frac{1}{\gamma}}$ in ϵ. The first-order correction for consumption, which is the second term in the right hand side of (A.12), is then given by:

$$\tilde{C}(W, S, X) = \frac{A^{-\frac{1}{\gamma}}}{\rho^\star}\left[\frac{\gamma - 1}{\gamma}\tilde{r}(X) + \tilde{\pi}(X)\frac{p^\star(X)S}{W} - \frac{(\gamma + 1)}{2}\tilde{\sigma}_R^2(X)\left(\frac{p^\star(X)S}{W}\right)^2\right]W,$$

where using the fact that $A^{-\frac{1}{\gamma}} = \rho^\star$, we obtain expression (30).

Step 4: Market Tightness. From equation (16), note that the market tightness is given by:

$$\theta^{1-\eta} = \overline{\alpha}\frac{V_S(W, S, X)\, n}{V_W(W, S, X)\, v_d p|n|}.$$

Rearranging the expression above, we obtain that in the limit when $\epsilon \longrightarrow 0$,[2] market tightness is well

[2]Note that the limit is indeterminate. However, note that:

$$\lim_{\epsilon\to 0}\frac{V_S}{\epsilon}\left(\frac{v_d}{\epsilon}\right)^{-1} = \lim_{\epsilon\to 0}\left(\frac{V_S^\star}{\epsilon} + \frac{\tilde{V}_S\epsilon}{\epsilon} + \mathcal{O}(\epsilon^2)\right)\left(\frac{v_d^\star}{\epsilon} + \frac{\tilde{v}_d\epsilon}{\epsilon} + \mathcal{O}(\epsilon^2)\right)^{-1} = \frac{\tilde{V}_S}{\tilde{v}_d}.$$

A similar reasoning applies to $\frac{n}{|n|}$.

defined and given by:

$$\theta^\star(W,S,X) = \left[\frac{\overline{\alpha}}{\tilde{v}_d(X)}\frac{\tilde{\Omega}(W,S,X)}{p^\star(X)}\frac{\tilde{n}}{|\tilde{n}|}\right]^{\frac{1}{1-\eta}},$$

where:

$$\begin{aligned}\tilde{\Omega}(W,S,X) &= \frac{\tilde{V}_S(W,S,X)}{V^\star_W(W,S,X)},\\ &= \frac{\gamma\tilde{\sigma}^2_R(X)}{\rho^\star}\left(\frac{\tilde{\pi}(X)}{\gamma\tilde{\sigma}_R(X)^2}-\frac{p^\star(X)S}{W}\right)p^\star(X).\end{aligned}$$

The market tightness is then given by:

$$\theta^\star(W,S,X) = \left[\frac{\overline{\alpha}\gamma\tilde{\sigma}^2_R(X)}{\rho^\star\,\tilde{v}_d(X)}\left|\frac{\tilde{\pi}(X)}{\gamma\tilde{\sigma}_R(X)^2}-\frac{p^\star(X)S}{W}\right|\right]^{\frac{1}{1-\eta}}.$$

Step 5: Orders. The order size can be written as:

$$\tilde{n}(W,S,X) = \alpha(\theta^\star(W,S,X))\frac{1-\eta}{\chi}\frac{\gamma\tilde{\sigma}^2_R(X)}{\rho^\star}\left(\frac{\tilde{\pi}(X)}{\gamma\tilde{\sigma}^2_R(X)}-\frac{p^\star(X)S}{W}\right).$$

Step 6: Intermediation Fees. For further reference, note that the fees satisfy the condition:

$$\phi(W,S,X;\epsilon) = \frac{\theta(W,S,X;\epsilon)}{\alpha(\theta(W,S,X;\epsilon))}p(X;\epsilon)v_d(X;\epsilon).$$

Using the fact that $v_d = \mathcal{O}(\epsilon)$, we obtain that $\phi(W,S,X;0)=0$. The first-order term is given by:

$$\tilde{\phi}(W,S,X) = \frac{\theta^\star(W,S,X)}{\alpha(\theta^\star(W,S,X))}p^\star(X)\tilde{v}_d(X) = \eta|\tilde{\Omega}(W,S,X)|.$$

□

A.5 Proof of Proposition 4

Proof. Rearranging the market-clearing condition, we obtain the expression:

$$\nu\left|\frac{s}{x}-\frac{\tilde{\pi}(X)}{\gamma\tilde{\sigma}^2_R(X)}\right|^{\frac{1+\eta}{1-\eta}} = (1-\nu)\left|\frac{\tilde{\pi}(X)}{\gamma\tilde{\sigma}^2_R(X)}-\frac{1-s}{1-x}\right|^{\frac{1+\eta}{1-\eta}}.$$

Note that the expression above holds whether or not $s > x$, the case considered in the main text. Rearranging the expression above, we obtain:

$$\tilde{\pi}(X) = \left[\frac{\nu^{\frac{1-\eta}{1+\eta}}}{\nu^{\frac{1-\eta}{1+\eta}} + (1-\nu)^{\frac{1-\eta}{1+\eta}}} \frac{s}{x} + \frac{(1-\nu)^{\frac{1-\eta}{1+\eta}}}{\nu^{\frac{1-\eta}{1+\eta}} + (1-\nu)^{\frac{1-\eta}{1+\eta}}} \frac{1-s}{1-x} \right] \gamma \tilde{\sigma}_R^2(X).$$

Defining $\tilde{\nu} \equiv \nu^{\frac{1-\eta}{1+\eta}} \left[\nu^{\frac{1-\eta}{1+\eta}} + (1-\nu)^{\frac{1-\eta}{1+\eta}} \right]^{-1}$, we obtain expression in Equation (35). □

A.6 Proof of Proposition 5

Proof. We can write the market clearing condition for goods (22) as follows:

$$x_t \frac{\hat{C}(\hat{W}_{1,t}, S_{1,t}, \hat{X}_t; \epsilon)}{\hat{W}_{1,t}} + \nu\chi \frac{n^2(\hat{W}_{1,t}, S_{1,t}, \hat{X}_t; \epsilon)}{2} + (1-x_t) \frac{\hat{C}(\hat{W}_{2,t}, S_{2,t}, \hat{X}_t; \epsilon)}{\hat{W}_{2,t}} + (1-\nu)\chi \frac{n^2(\hat{W}_{2,t}, S_{2,t}, \hat{X}_t; \epsilon)}{2}$$
$$+ v_d(X_t; \epsilon)\overline{d}\epsilon = \frac{1}{q_t}, \tag{A.13}$$

The first-order approximation of the expression above gives:

$$x \left[\frac{\gamma-1}{\gamma} \tilde{r}(X) + \tilde{\pi}(X) \frac{s}{x} - \frac{(\gamma+1)}{2} \tilde{\sigma}_R^2(X) \left(\frac{s}{x} \right)^2 \right] + (1-x) \left[\frac{\gamma-1}{\gamma} \tilde{r}(X) + \tilde{\pi}(X) \frac{1-s}{1-x} - \frac{(\gamma+1)}{2} \tilde{\sigma}_R^2(X) \left(\frac{1-s}{1-x} \right)^2 \right] \tag{A.14}$$
$$= -\frac{\tilde{q}(X)}{q^\star(X)}. \tag{A.15}$$

where we used that n^2 and $v_d \overline{d}\epsilon$ are of order ϵ^2. From the expression for expected returns in the risky asset, we have that:

$$\frac{1}{q_t} + \mu + \mu_{q,t} + \sigma\sqrt{\epsilon}\sigma_{q,t} = r_t + \pi_t.$$

Computing the first-order approximation of the expression above, we obtain:

$$-\frac{\tilde{q}(X)}{q^\star(X)^2} = \tilde{r}(X) + \tilde{\pi}(X), \tag{A.16}$$

where we used the fact that $\mu_{q,t} = \mathcal{O}(\epsilon^2)$ and $\sqrt{\epsilon}\sigma_{q,t} = \mathcal{O}(\epsilon^2)$. From (A.14) and (A.6), we obtain:

$$\frac{\gamma-1}{\gamma} \tilde{r}(X) + \tilde{\pi}(X) - \frac{\gamma+1}{2} \left[x \left(\frac{s}{x} \right)^2 + (1-x) \left(\frac{1-s}{1-x} \right)^2 \right] \tilde{\sigma}_R^2(X) = \tilde{r}(X) + \tilde{\pi}(X).$$

Rearranging the expression above, we obtain:

$$\tilde{r}(X) = -\frac{\gamma(\gamma+1)}{2} \left[x \left(\frac{s}{x} \right)^2 + (1-x) \left(\frac{1-s}{1-x} \right)^2 \right] \tilde{\sigma}_R^2(X)$$

Using the fact that $r(X, \epsilon) = r^\star(X) + \tilde{r}(X)\epsilon + \mathcal{O}(\epsilon)$, we obtain the expression in Equation (37). □

A.7 Proof of Proposition 6

Proof. ***Step 1: Dealers' Value.*** Plugging (31) and (32), which are the expressions for $\theta^\star$ and $\tilde{n}$, into the dealers' capacity constraint (with equality), condition (4), and using the definition of the market tightness to solve for mass of contracts posted by dealers, we obtain:

$$\tilde{v}_d(X) = \left[\frac{1-\eta}{\chi\eta\overline{d}}\left(\nu\left(\frac{\overline{\alpha}|\tilde{\Omega}_1|}{p^\star(X)}\right)^{\frac{2}{1-\eta}} + (1-\nu)\left(\frac{\overline{\alpha}|\tilde{\Omega}_2|}{p^\star(X)}\right)^{\frac{2}{1-\eta}}\right)\right]^{\frac{1-\eta}{1+\eta}}.$$

Using the expression for the marginal value of portfolio rebalancing (29) and the expression for the risk premium (35), we obtain:

$$\frac{|\Omega_1(x,s)|}{p} = \frac{\gamma\sigma^2}{r^\star - \mu}(1-\tilde{\nu})\left|\frac{s}{x} - \frac{1-s}{1-x}\right|, \qquad \frac{|\Omega_2(x,s)|}{p} = \frac{\gamma\sigma^2}{r^\star - \mu}\tilde{\nu}\left|\frac{s}{x} - \frac{1-s}{1-x}\right|. \tag{A.17}$$

Combining the expressions above, we get:

$$\tilde{v}_d(X) = \overline{v}_d\left|\frac{s}{x} - \frac{1-s}{1-x}\right|^{\frac{2}{1+\eta}},$$

where:

$$\overline{v}_d \equiv \left(\overline{\alpha}\frac{\gamma\sigma^2}{r^\star - \mu}\right)^{\frac{2}{1+\eta}}\left[\frac{1-\eta}{\chi\eta\overline{d}}\left(\nu(1-\tilde{\nu})^{\frac{2}{1-\eta}} + (1-\nu)\tilde{\nu}^{\frac{2}{1-\eta}}\right)\right]^{\frac{1-\eta}{1+\eta}}.$$

Note that we can write $\overline{v}_d$ as follows:

$$\overline{v}_d = \left(\overline{\alpha}\frac{\gamma\sigma^2}{r^\star - \mu}\right)^{\frac{2}{1+\eta}}\left(\frac{1-\eta}{\chi\eta\overline{d}}\right)^{\frac{1-\eta}{1+\eta}}\frac{\nu^{\frac{1-\eta}{1+\eta}}(1-\nu)^{\frac{1-\eta}{1+\eta}}}{\nu^{\frac{1-\eta}{1+\eta}} + (1-\nu)^{\frac{1-\eta}{1+\eta}}}.$$

Step 2: Market Tightness. Combining (31) with the expressions for the marginal value of portfolio rebalancing and the dealer's value derived above, we obtain the market tightness for investor j:

$$\theta_j^\star(x,s) = \left(\frac{\overline{\alpha}(1-\tilde{\nu}_j)}{\tilde{v}_d(x,s)}\left|\frac{1-s}{1-x} - \frac{s}{x}\right|\frac{\gamma\sigma^2}{\rho^\star}\right)^{\frac{1}{1-\eta}},$$

where $\tilde{\nu}_j = \mathbf{1}_{j=1}\tilde{\nu} + \mathbf{1}_{j=2}(1-\tilde{\nu})$. Plugging the expression for the dealer's value into the expression above, we obtain:

$$\theta_j^\star(x,s) = \left[\nu_j\overline{\alpha}\frac{\gamma\sigma^2}{r^\star-\mu}\frac{1-\eta}{\chi\eta\overline{d}}\right]^{-\frac{1}{1+\eta}}\left|\frac{1-s}{1-x}-\frac{s}{x}\right|^{-\frac{1}{1+\eta}}$$
$$= \overline{\theta}_j\Delta^{-\frac{1}{1+\eta}}, \tag{A.18}$$

where $\overline{\theta}_j \equiv \left[\nu_j\overline{\alpha}\frac{\gamma\sigma^2}{r^\star-\mu}\frac{1-\eta}{\chi\eta\overline{d}}\right]^{-\frac{1}{1+\eta}}$ and Δ is the portfolio dispersion defined in Definition 3.

Step 3: Volume. Volume traded can be written as:

$$\mathbb{V}(x,s) = \nu|n_1(x,s)|\alpha(\theta_1(x,s)),$$

where $n_1(x,s)$ and $\theta_1(x,s)$ denote the number of shares and the market tightness for investor 1. By market clearing of the risky asset, we would obtain the same result if we used the order size and market tightness for investor 2. Plugging (A.17) and (A.18), which are the expressions for the marginal value of portfolio rebalancing and the market tightness, into (31) we obtain the expression for the trading volume as a function of the state variables:

$$\mathbb{V}(x,s) = \nu(1-\tilde{\nu})\frac{\gamma\sigma^2}{\rho^\star}\left(\frac{\overline{\alpha}}{\eta}\right)^2\frac{1-\eta}{\chi}\overline{\theta}_1^{2\eta}\left|\frac{1-s}{1-x}-\frac{s}{x}\right|^{\frac{1-\eta}{1+\eta}}\epsilon + o(\epsilon),$$

We can write the expression above as follows:

$$\mathbb{V}(x,s) = \overline{V}\Delta^{\frac{1-\eta}{1+\eta}}\epsilon + o(\epsilon),$$

where:

$$\overline{V} = \frac{\nu^{\frac{1-\eta}{1+\eta}}(1-\nu)^{\frac{1-\eta}{1+\eta}}}{\nu^{\frac{1-\eta}{1+\eta}}+(1-\nu)^{\frac{1-\eta}{1+\eta}}}\left[\frac{1-\eta}{\chi\rho^\star}\left(\frac{\overline{\alpha}\overline{d}^\eta}{\eta}\right)^{\frac{2}{1-\eta}}\gamma\sigma^2\right]^{\frac{1-\eta}{1+\eta}}.$$

Note how volume traded is increasing in the efficiency of the matching function $\overline{\alpha}$ and the dealers' intermediation capacity $\overline{d}$. Volume is decreasing in the adjustment cost parameter χ. Volume is also increasing in the amount of risk σ^2 and the risk aversion coefficient γ.

Step 4: Bid-ask Spread. Combining (18) and (A.17), we obtain:

$$\phi_{ba}(x,s) = \frac{\eta}{r^\star-\mu}\left|\frac{s}{x}-\frac{1-s}{1-x}\right|\gamma\sigma^2\epsilon = \frac{\eta\gamma\sigma^2}{r^\star-\mu}\Delta\epsilon.$$

□

A.8 Proof of Proposition 7

Proof. This proof follows closely the steps of Propositions 3, 4, and 6.

Step 1: Marginal Value of Portfolio Rebalancing. Note that in the derivation of the results in Proposition 3, we did not use the fact that investors have homogeneous preferences, as it characterizes the solution of the investors' problem for arbitrary values of the interest rate, risk premium, and volatility. Therefore, it is immediate that the marginal value of rebalancing satisfies:

$$\tilde{\Omega}_k(W,S,X) = \frac{\gamma_k \tilde{\sigma}_R^2(X)}{r^\star - \mu}\left(\frac{\tilde{\pi}(X)}{\gamma_k \tilde{\sigma}_R^2(X)} - \frac{p^\star(X)S}{W}\right)p^\star(X),$$

which proves part (a).

Step 2: Risk Premium. Following similar steps to those of Proposition 4, we obtain:

$$\nu\left|\gamma_1 \tilde{\sigma}_R^2(X)\frac{s}{x} - \tilde{\pi}(X)\right|^{\frac{1+\eta}{1-\eta}} = (1-\nu)\left|\tilde{\pi}(X) - \gamma_2 \tilde{\sigma}_R^2(X)\frac{1-s}{1-x}\right|^{\frac{1+\eta}{1-\eta}}.$$

Rearranging the expression above yields part (b).

Step 3: Bid-ask Spread. Following similar steps to Proposition 6, combining (18) with the marginal value of portfolio rebalancing, we obtain part (c). □

A.9 Proof of Proposition 8

Proof. We prove Proposition 8 for the more general case with heterogeneous risk aversion, Epstein-Zin preferences, and order execution risk. We focus on the case of two types of investors: $\gamma_i = \gamma_1$ if $i \leq \nu$ and $\gamma_i = \gamma_2$ if $i > \nu$. Investors have continuous-time Epstein-Zin (recursive) preferences:

$$V_{i,t} = \mathbb{E}_t\left[\int_t^\infty f_i(C_{i,s}, V_{i,s})ds\right],$$

where:

$$f_i(C,V) = \rho\frac{(1-\gamma_i)V}{1-\psi^{-1}}\left[\left(\frac{C}{((1-\gamma_i)V)^{\frac{1}{1-\gamma_i}}}\right)^{1-\psi^{-1}} - 1\right].$$

Step 1: Deriving the detrended HJB equation.

Step 1.1: HJB equation. The HJB equation is given by:

$$0 = \max_{C,n,\theta} f_i(C, V_{i,t}) + \mathbb{E}[dV_{i,t}].$$

Applying Ito's lemma to the expression above and dropping the i subscripts, we obtain:

$$\rho \frac{(1-\gamma)V}{1-\psi^{-1}} = \max_{C,n,\theta} \rho \frac{(1-\gamma)V}{1-\psi^{-1}} \left[\frac{C}{((1-\gamma)V)^{\frac{1}{1-\gamma}}} \right]^{1-\psi^{-1}} + V_W \left[rW + \pi pS - \frac{\chi}{2} pn^2 - C \right] + V_X \mu_X$$
$$+ \frac{1}{2} V_{WW} \sigma_R^2 (pS)^2 + pS\sigma_R V_{WX} \sigma_X + \frac{1}{2} \sigma_X' V_{XX} \sigma_X + \left[V\left(W - \frac{\theta v_d}{\alpha(\theta)} p|n|, S+n, X \right) - V(W,S,X) \right] \alpha(\theta).$$

Step 1.2: De-trending. As in the CRRA case, it is convenient to work with detrended variables. We write the normalized consumption, wealth, and value function as $\hat{C}_{i,t} = \frac{C_{i,t}}{Y_t}$, $\hat{W}_{i,t} = \frac{W_{i,t}}{Y_t}$ and $\hat{V}_{i,t} = \frac{V_{i,t}}{Y_t^{1-\gamma_i}}$, respectively. The normalized value function can be written as $\hat{V}_{i,t} = \hat{V}(\hat{W}_{i,t}, S_{i,t}, \hat{X}_t; \epsilon)$, where $X = (Y, \hat{X})$, and $\hat{X}$ is the relevant state variable for the detrended economy.

Step 1.3: Detrended HJB Equation. Using the definition of the detrended variables into the HJB equation, after several algebraic steps, we obtain:

$$\rho \frac{(1-\gamma)\hat{V}}{1-\psi^{-1}} = \max_{\hat{C},n,\theta} \rho \frac{(1-\gamma)\hat{V}}{1-\psi^{-1}} \left[\frac{\hat{C}}{((1-\gamma)\hat{V})^{\frac{1}{1-\gamma}}} \right]^{1-\psi^{-1}} + \hat{V}_{\hat{W}} \left[r\hat{W} + \pi qS - \frac{\chi}{2} qn^2 - \hat{C} \right] + \left[(1-\gamma)\hat{V} - \hat{V}_{\hat{W}} \hat{W} \right] \mu + \hat{V}_{\hat{X}} \mu_{\hat{X}}$$
$$+ \frac{1}{2} \hat{V}_{\hat{W}\hat{W}} \sigma_R^2 (qS)^2 + qS\sigma_R \left[-\gamma \hat{V}_{\hat{W}} - \hat{V}_{\hat{W}\hat{W}} \hat{W} \right] \sigma + qS\sigma_R \hat{V}_{\hat{W}\hat{X}} \sigma_{\hat{X}} + \frac{1}{2} \left[\gamma(\gamma-1)\hat{V} + 2\gamma \hat{V}_{\hat{W}} \hat{W} + \hat{V}_{\hat{W}\hat{W}} \hat{W}^2 \right] \sigma^2$$
$$+ \sigma \left[(1-\gamma)\hat{V}_{\hat{X}} - \hat{V}_{\hat{W}\hat{X}} \hat{W} \right] \sigma_{\hat{X}} + \frac{1}{2} \sigma_{\hat{X}}' V_{\hat{X}\hat{X}} \sigma_{\hat{X}} + \left[\hat{V}\left(\hat{W} - \frac{\theta v_d}{\alpha(\theta)} q|n|, S+n, \hat{X} \right) - \hat{V}(\hat{W}, S, X) \right] \alpha(\theta),$$

where we used the fact that $V_X \mu_X = V_Y \mu Y + V_{\hat{X}} \mu_{\hat{X}}$, $\sigma_X' V_{XX} \sigma_X = V_{YY} \sigma^2 Y^2 + 2V_{\hat{X}Y} \sigma_{\hat{X}} \sigma Y + \sigma_{\hat{X}}' V_{\hat{X}\hat{X}} \sigma_{\hat{X}}$, and $V_{WX} \sigma_X = V_{WY} \sigma Y + V_{W\hat{X}} \sigma_{\hat{X}}$.[3]

Combining common terms, we can rewrite the expression above as follows:

$$\hat{\rho} \frac{(1-\gamma)\hat{V}}{1-\psi^{-1}} = \max_{\hat{C},n,\theta} \rho \frac{(1-\gamma)\hat{V}}{1-\psi^{-1}} \left[\frac{\hat{C}}{((1-\gamma)\hat{V})^{\frac{1}{1-\gamma}}} \right]^{1-\psi^{-1}} + \hat{V}_{\hat{W}} \left[(r - \mu + \gamma\sigma^2)\hat{W} + (\pi - \gamma\sigma\sigma_R) qS - \frac{\chi}{2} qn^2 - \hat{C} \right]$$
$$+ \hat{V}_{\hat{X}} (\mu_{\hat{X}} + (1-\gamma)\sigma\sigma_{\hat{X}}) + \frac{1}{2} \hat{V}_{\hat{W}\hat{W}} \left[qS\sigma_R - \sigma\hat{W} \right]^2 + (qS\sigma_R - \hat{W}\sigma) \hat{V}_{W\hat{X}} \sigma_{\hat{X}} + \frac{1}{2} \sigma_{\hat{X}}' V_{\hat{X}\hat{X}} \sigma_{\hat{X}}$$
$$+ \left[\hat{V}\left(\hat{W} - \frac{\theta v_d}{\alpha(\theta)} q|n|, S+n, \hat{X} \right) - \hat{V}(\hat{W}, S, X) \right] \alpha(\theta),$$

[3]Note that using $V(W,S,X) = Y^{1-\gamma} \hat{V}\left(\frac{W}{Y}, S, \hat{X} \right)$ we can express the derivatives of $V(\cdot)$ in terms of the detrended value function $\hat{V}(\cdot)$.

where we have defined $\hat{\rho} \equiv \rho - (1-\psi^{-1})\left(\mu - \frac{\gamma\sigma^2}{2}\right)$.

Step 2: The benchmark economy. We will consider next the benchmark economy where $\sigma = 0$. *Step 2.1: Solution to the investor's problem.* We initially assume that in this economy we have:

$$\pi^\star = \sigma_R^\star = v_d^\star = \mu_{\hat{X}}^\star = \sigma_{\hat{X}}^\star = 0,$$

where the superscript $\star$ is used to denote variables in the benchmark economy. The HJB equation is given by:

$$(\rho - (1-\psi^{-1})\mu)\frac{(1-\gamma)\hat{V}^\star}{1-\psi^{-1}} = \max_{\hat{C},n,\theta} \rho \frac{(1-\gamma)\hat{V}^\star}{1-\psi^{-1}}\left[\frac{\hat{C}}{((1-\gamma)\hat{V}^\star)^{\frac{1}{1-\gamma}}}\right]^{1-\psi^{-1}} + \hat{V}_{\hat{W}}^\star\left[(r^\star - \mu)\hat{W} - \frac{\chi}{2}q^\star n^2 - \hat{C}\right]$$
$$+\left[\hat{V}^\star\left(\hat{W}, S+n, \hat{X}\right) - \hat{V}^\star(\hat{W}, S, X)\right]\alpha(\theta).$$

We guess-and-verify that the value function takes the form:

$$\hat{V}^\star(\hat{W}, S, \hat{X}) = A^{\frac{1-\gamma}{1-\psi^{-1}}}\frac{W^{1-\gamma}}{1-\gamma}. \tag{A.19}$$

The HJB equation can then be written as:

$$\frac{\rho^\star}{1-\psi^{-1}} = \max_{\hat{C},n,\theta}\frac{\rho}{1-\psi^{-1}}\left[\frac{\hat{C}}{A^{\frac{1}{1-\psi^{-1}}}\hat{W}}\right]^{1-\psi^{-1}} + (r^\star - \mu) - \frac{\chi}{2}\frac{q^\star n^2}{\hat{W}} - \frac{\hat{C}}{\hat{W}}.$$

The policy functions are given by:

$$\hat{C}^\star(\hat{W}, S, \hat{X}) = \rho^\psi A^{-\psi}, \qquad n^\star(\hat{W}, S, \hat{X}) = 0, \tag{A.20}$$

where $\theta^\star(\hat{W}, S, \hat{X})$ is indeterminate. Plugging the policy functions back into the HJB equation, we obtain:

$$\rho^\psi A^{-\psi} = \psi\rho + (1-\psi)r^\star.$$

Step 2.2: Determining asset prices. From the market clearing condition for goods, we have that:

$$\int_0^1 \hat{C}_{i,t}di = 1 \Rightarrow \rho^\psi A^{-\psi}\int_0^1 \hat{W}_{i,t}di = 1 \Rightarrow \rho^\psi A^{-\psi} = \frac{1}{q^\star},$$

where we used the market clearing condition for the risky asset $\int_0^1 \hat{W}_{i,t} di = q_t^\star$.

Given that the risk premium is zero, then $\mu_R^\star = \frac{1}{q^\star} + \mu = r^\star$. Using the expression for the consumption-wealth ratio derived above, we obtain:

$$\psi\rho + (1-\psi)r^\star = r^\star - \mu \Longrightarrow r^\star = \rho + \psi^{-1}\mu,$$

and $\rho^\psi A^{-\psi} = (q^\star)^{-1} = \rho - (1-\psi^{-1})\mu$, a quantity that is assumed to be positive.

Step 2.3: Joint distribution of wealth and asset holdings. Define the wealth share of investor i as $x_{i,t} \equiv \frac{\hat{W}_{i,t}}{\int_0^1 \hat{W}_{i,t} di}$ and the asset share of investor i as $s_{i,t} \equiv \frac{S_{i,t}}{\int_0^1 S_{i,t} di}$. Let $G_{1,t}(x,s) = \frac{1}{\nu}\int_0^\nu \mathbf{1}_{\{x_{i,t}\leq x, s_{i,t}\leq s\}} di$ and $G_{2,t}(x,s) = \frac{1}{1-\nu}\int_\nu^1 \mathbf{1}_{\{x_{i,t}\leq x, s_{i,t}\leq s\}} di$ denote the joint distribution of $x_{i,t}$ and $s_{i,t}$ conditional on being of type 1 and type 2, respectively. This pair of distributions corresponds to the aggregate state variable in this economy, that is, $\hat{X}_t = (G_{1,t}(\cdot), G_{2,t}(\cdot))$.[4] From the market clearing condition for the risky asset, we have that $x_{i,t} = \frac{\hat{W}_{i,t}}{q_t}$. At the benchmark economy, the price-dividend ratio $q^\star$ is constant. Using the fact that $\rho^\psi A^{-\psi} = r^\star - \mu$, we have that normalized wealth is constant, so $x_{i,t}$ is constant for all $i \in [0,1]$. Given that $n^\star = 0$, then $s_{i,t}$ is also constant for all $i \in [0,1]$. Therefore, $G_{j,t}(x,s)$ is constant in the benchmark economy, so $\mu_{\hat{X}}^\star = \sigma_{\hat{X}}^\star = 0$.

Step 2.4: Dealers' problem. Consider a contract $\varsigma = (n,\phi)$, where $n \neq 0$. Given that there is no value of θ which will generate a strict improvement over $V^\star(W,S,X)$, then $\theta_t(\varsigma) = \infty$ and $\frac{\alpha(\theta(\varsigma))}{\theta(\varsigma))} = 0$, according to (8). This implies that the expected profit of the dealer is zero and the capacity constraint is not binding, that is, $v_d^\star = 0$.

Step 3: The first-order approximation of the investor's problem.

Step 3.1: The detrended value function. We consider next the first-order correction terms. We start by taking a first-order approximation of the aggregate variables with respect to ϵ:

$$r(X,\epsilon) = r^\star(X) + \tilde{r}(X)\epsilon + \mathcal{O}(\epsilon^2),$$

and analogously for the remaining aggregate variables. The diffusion terms can be written as follows:

$$\sigma_R(X,\epsilon) = \sigma_R^\star(X) + \tilde{\sigma}_R(X)\sqrt{\epsilon} + \mathcal{O}(\epsilon), \qquad \sigma_X(X,\epsilon) = \sigma_X^\star(X) + \tilde{\sigma}_X(X)\sqrt{\epsilon} + \mathcal{O}(\epsilon).$$

[4] This implies that the state space is infinite-dimensional in this case. Therefore, the derivatives of the value function with respect to $\hat{X}$ in the HJB equation should be interpreted as Fréchet derivatives, as discussed e.g. in Luenberger (1997).

Plugging these expressions into the detrended value function, we obtain:

$$\rho^\star \frac{(1-\gamma)\hat{V}}{1-\psi^{-1}} = \max_{\hat{C},n,\theta} \rho \frac{(1-\gamma)\hat{V}}{1-\psi^{-1}} \left[\frac{\hat{C}}{((1-\gamma)\hat{V})^{\frac{1}{1-\gamma}}}\right]^{1-\psi^{-1}} + \hat{V}_{\hat{W}} \left[(r^\star + \tilde{r}\epsilon - \mu + \gamma\sigma^2\epsilon)\hat{W} + (\tilde{\pi} - \gamma\sigma\tilde{\sigma}_R)\epsilon q^\star S - \frac{\chi}{2} q^\star n^2 - \hat{C}\right]$$
$$+ \hat{V}_{\hat{X}}(\tilde{\mu}_{\hat{X}} + (1-\gamma)\sigma\sigma_{\hat{X}})\epsilon + \frac{1}{2}\hat{V}_{\hat{W}\hat{W}}\left[q^\star S\tilde{\sigma}_R - \sigma\hat{W}\right]^2 \epsilon + (q^\star S\tilde{\sigma}_R - \hat{W}\sigma)\hat{V}_{W\hat{X}}\tilde{\sigma}_{\hat{X}}\epsilon + \frac{1}{2}\tilde{\sigma}'_{\hat{X}} V_{\hat{X}\hat{X}}\tilde{\sigma}_{\hat{X}}\epsilon$$
$$+ \left[\hat{V}\left(\hat{W} - \frac{\theta\tilde{v}_d\epsilon}{\alpha(\theta)} q^\star|n|, S+n, \hat{X}\right) - \hat{V}(\hat{W}, S, X)\right]\alpha(\theta) - (1-\gamma)\hat{V}\frac{\gamma\sigma^2}{2}\epsilon + \mathcal{O}(\epsilon^2).$$

Taking the derivative of the expression above with respect to ϵ and evaluating at $\epsilon = 0$, we obtain:

$$\rho^\star \frac{(1-\gamma)\hat{V}_\epsilon}{1-\psi^{-1}} = \frac{\psi^{-1}-\gamma}{1-\gamma}\frac{\rho(1-\gamma)\hat{V}_\epsilon}{1-\psi^{-1}}\left[\frac{\hat{C}^\star}{((1-\gamma)\hat{V}^\star)^{\frac{1}{1-\gamma}}}\right]^{1-\psi^{-1}} + \hat{V}_{\hat{W},\epsilon}\left[(r^\star - \mu)\hat{W} - \hat{C}^\star\right]$$
$$+ \hat{V}^\star_{\hat{W}}\left[(\tilde{r} + \gamma\sigma^2)\hat{W} + (\tilde{\pi} - \gamma\sigma\tilde{\sigma}_R)q^\star S\right] + \frac{1}{2}\hat{V}^\star_{\hat{W}\hat{W}}\left[q^\star S\tilde{\sigma}_R - \sigma\hat{W}\right]^2 - (1-\gamma)\hat{V}^\star\frac{\gamma\sigma^2}{2},$$

where we used the fact that $V^\star_{\hat{X}} = V^\star_{W\hat{X}} = V^\star_{\hat{X}\hat{X}} = n^\star = 0$.

Using (A.19), which is the expression for $\hat{V}^\star$, and the fact that $\rho^\psi A^{-\psi} = r^\star - \mu$, we obtain:

$$\rho^\star \frac{(1-\gamma)\hat{V}_\epsilon}{1-\psi^{-1}} = \frac{\psi^{-1}-\gamma}{1-\gamma}\frac{\rho^\star(1-\gamma)\hat{V}_\epsilon}{1-\psi^{-1}} + A^{\frac{1-\gamma}{1-\psi^{-1}}} W^{1-\gamma}\left[\tilde{r} + \tilde{\pi}\frac{q^\star S}{\hat{W}} - \frac{\gamma}{2}\tilde{\sigma}_R^2\left(\frac{q^\star S}{\hat{W}}\right)^2\right].$$

After some rearrangement, we obtain the expression for $\hat{V}_\epsilon$:

$$\hat{V}_\epsilon(\hat{W}, S, \hat{X}) = \frac{A^{\frac{1-\gamma}{1-\psi^{-1}}} W^{1-\gamma}}{\rho^\star}\left[\tilde{r}(\hat{X}) + \tilde{\pi}(\hat{X})\frac{q^\star(\hat{X})S}{\hat{W}} - \frac{\gamma}{2}\tilde{\sigma}_R^2(\hat{X})\left(\frac{q^\star(\hat{X})S}{\hat{W}}\right)^2\right].$$

Step 3.2: Consumption. The first-order condition for consumption is given by:

$$\rho\left((1-\gamma)\hat{V}\right)^{1-\frac{1-\psi^{-1}}{1-\gamma}} \hat{C}^{-\psi^{-1}} = \hat{V}_W.$$

After some rearrangement, it gives the policy function for consumption:

$$\hat{C}(\hat{W}, S, \hat{X}; \epsilon) = \rho^\psi \left((1-\gamma)\hat{V}(\hat{W}, S, \hat{X}; \epsilon)\right)^{\frac{1-\gamma\psi}{1-\gamma}} \hat{V}_{\hat{W}}^{-\psi}(\hat{W}, S, \hat{X}; \epsilon).$$

The derivative of the expression above with respect to ϵ is given by:

$$\hat{C}_\epsilon(\hat{W}, S, \hat{X}; 0) = \frac{1-\gamma\psi}{1-\gamma}\rho^\psi \left(A^{\frac{1-\gamma}{1-\psi^{-1}}} W^{1-\gamma}\right)^{\gamma\frac{1-\psi}{1-\gamma}} (A^{\frac{1-\gamma}{1-\psi^{-1}}} W^{-\gamma})^{-\psi}(1-\gamma)\hat{V}_\epsilon(\hat{W}, S, \hat{X}; 0)$$
$$- \psi\rho^\psi \left(A^{\frac{1-\gamma}{1-\psi^{-1}}} W^{1-\gamma}\right)^{\frac{1-\gamma\psi}{1-\gamma}} \left(A^{\frac{1-\gamma}{1-\psi^{-1}}} W^{-\gamma}\right)^{-\psi-1} \hat{V}_{\hat{W},\epsilon}(\hat{W}, S, \hat{X}; 0).$$

Plugging in the expressions for $\hat{V}_\epsilon(\hat{W}, S, \hat{X}; 0)$ and $\hat{V}_{\hat{W},\epsilon}(\hat{W}, S, \hat{X}; 0)$, we get:

$$\hat{C}_\epsilon(\hat{W}, S, \hat{X}; 0) = (1-\gamma\psi)\frac{\rho^\psi A^{-\psi}}{\rho^\star} W \left[\tilde{r}(\hat{X}) + \tilde{\pi}(\hat{X})\frac{q^\star(\hat{X})S}{\hat{W}} - \frac{\gamma}{2}\tilde{\sigma}_R^2(\hat{X})\left(\frac{q^\star(\hat{X})S}{\hat{W}}\right)^2\right]$$
$$- \psi\frac{\rho^\psi A^{-\psi}}{\rho^\star} W \left[(1-\gamma)\tilde{r}(\hat{X}) - \gamma\tilde{\pi}(\hat{X})\frac{q^\star(\hat{X})S}{\hat{W}} + (\gamma+1)\frac{\gamma}{2}\tilde{\sigma}_R^2(\hat{X})\left(\frac{q^\star(\hat{X})S}{\hat{W}}\right)^2\right]. \quad \text{(A.21)}$$

From (A.20) and (A.21), the first-order expansion for consumption is given by:

$$\hat{C}(\hat{W}, S, \hat{X}; 0) = \rho^\psi A^{-\psi} W + W\left[(1-\psi)\tilde{r}(\hat{X}) + \tilde{\pi}(\hat{X})\frac{q^\star(\hat{X})S}{\hat{W}} - (1+\psi)\frac{\gamma}{2}\tilde{\sigma}_R^2(\hat{X})\left(\frac{q^\star(\hat{X})S}{\hat{W}}\right)^2\right]\epsilon + \mathcal{O}(\epsilon^2).$$

Step 3.3: Market tightness. The first-order condition for the market tightness is given by:

$$\left[\hat{V}\left(\hat{W} - \frac{\theta v_d}{\alpha(\theta)}q|n|, S+n, \hat{X}\right) - \hat{V}(\hat{W}, S, X)\right]\alpha'(\theta) = (1-\eta)\hat{V}_{\hat{W}}\left(\hat{W} - \frac{\theta v_d}{\alpha(\theta)}q|n|, S+n, \hat{X}\right) v_d q|n|. \quad \text{(A.22)}$$

Note that the right-hand side of the expression above is of order $\mathcal{O}(\epsilon^2)$, as $n = \mathcal{O}(\epsilon)$ and $v_d = \mathcal{O}(\epsilon)$. We will show next that the left-hand side is also of order $\mathcal{O}(\epsilon^2)$. The parameter ϵ affects the left-hand side through several terms, including the choice variables θ and n, the aggregate variables v_d and q, and the value function itself. Given the presence of the second-order term $v_d|n|$, we can ignore the effects through θ and n and fix them at their zeroth-order terms. Let $\mathcal{N}(\epsilon)$ denote the term in brackets above (up to second-order in ϵ), as a function of ϵ:

$$\mathcal{N}(\epsilon) = \hat{V}\left(\hat{W} - \frac{\theta^\star\tilde{v}_d}{\alpha(\theta^\star)}q^\star|\tilde{n}|\epsilon^2, S+\tilde{n}\epsilon, \hat{X}; \epsilon\right) - \hat{V}(\hat{W}, S, X; \epsilon).$$

Note that $\mathcal{N}(0) = 0$ and that the derivative of $\mathcal{N}(\epsilon)$ is given by:

$$\mathcal{N}'(\epsilon) = -\left.\hat{V}_{\hat{W}}\right|_{\left(\hat{W}-\frac{\theta^\star \tilde{v}_d}{\alpha(\theta^\star)}q^\star|\tilde{n}|\epsilon^2, S+\tilde{n}\epsilon, \hat{X};\epsilon\right)} 2\frac{\theta^\star \tilde{v}_d}{\alpha(\theta^\star)} q^\star |\tilde{n}|\epsilon + \left.\hat{V}_S\right|_{\left(\hat{W}-\frac{\theta^\star \tilde{v}_d}{\alpha(\theta^\star)}q^\star|\tilde{n}|\epsilon^2, S+\tilde{n}\epsilon, \hat{X};\epsilon\right)} \tilde{n} + \\ + \left.\hat{V}_\epsilon\right|_{\left(\hat{W}-\frac{\theta^\star \tilde{v}_d}{\alpha(\theta^\star)}q^\star|\tilde{n}|\epsilon^2, S+\tilde{n}\epsilon, \hat{X};\epsilon\right)} - \left.\hat{V}_\epsilon\right|_{\left(\hat{W}, S, \hat{X};\epsilon\right)}.$$

Using the fact that $\hat{V}_S^\star = 0$, we have that $\mathcal{N}'(0) = 0$. The second derivative of $\mathcal{N}(\epsilon)$ evaluated at $\epsilon = 0$ is given by:

$$\mathcal{N}''(0) = -\hat{V}_{\hat{W}}(\hat{W}, S, \hat{X}; 0) 2\frac{\theta^\star \tilde{v}_d}{\alpha(\theta^\star)} q^\star |\tilde{n}| + 2\hat{V}_{S,\epsilon}(\hat{W}, S, \hat{X}; 0)\tilde{n}.$$

Taking a second-order expansion of the left-hand side and right-hand side of (A.22) with respect to ϵ, then gives:

$$\left[-\hat{V}^\star_{\hat{W}} \frac{\theta^\star \tilde{v}_d}{\alpha(\theta^\star)} q^\star |\tilde{n}| + \hat{V}_{S,\epsilon}\tilde{n}\right] \overline{\alpha}(\theta^\star)^{\eta-1} = (1-\eta)\hat{V}^\star_{\hat{W}} \tilde{v}_d q^\star |\tilde{n}|,$$

where we used the fact that $\mathcal{N}(\epsilon) = \mathcal{N}(0) + \mathcal{N}'(0)\epsilon + \frac{1}{2}\mathcal{N}''(0)\epsilon^2 + \mathcal{O}(\epsilon^3)$.

Rearranging the expression above, we obtain:

$$\theta^\star(\hat{W}, S, \hat{X}) = \left[\frac{\overline{\alpha}}{\tilde{v}_d(\hat{X})} \frac{\tilde{\Omega}(\hat{W}, S, \hat{X})}{q^\star(\hat{X})} \frac{\tilde{n}}{|\tilde{n}|}\right]^{\frac{1}{1-\eta}}, \tag{A.23}$$

where $\tilde{\Omega}(\hat{W}, S, \hat{X})$ is the first-order correction for the marginal value of rebalancing, given by:

$$\tilde{\Omega}(\hat{W}, S, \hat{X}) = \frac{\hat{V}_{S,\epsilon}(\hat{W}, S, \hat{X}; 0)}{\hat{V}^\star_{\hat{W}}(\hat{W}, S, \hat{X})} = \left[\tilde{\pi}(\hat{X}) - \gamma\tilde{\sigma}^2_R(\hat{X}) \frac{q^\star(\hat{X})S}{\hat{W}}\right] q^\star(\hat{X}).$$

Step 3.4: Order size. The first-order condition for the order size is given by:

$$\hat{V}_S\left(\hat{W} - \frac{\theta v_d}{\alpha(\theta)} q|n|, S+n, \hat{X}\right)\alpha(\theta) = \hat{V}_{\hat{W}}\left(\hat{W} - \frac{\theta v_d}{\alpha(\theta)} q|n|, S+n, \hat{X}\right)\theta v_d q sg(n) + \chi \hat{V}_{\hat{W}}(\hat{W}, S, \hat{X}) qn.$$

Dividing the expression above by ϵ and taking the limit as ϵ goes to zero, we obtain:

$$\hat{V}_{S,\epsilon}\left(\hat{W}, S, \hat{X}\right)\alpha(\theta^\star) = \hat{V}^\star_{\hat{W}}\left(\hat{W}, S, \hat{X}\right)\theta\tilde{v}_d q^\star sg(\tilde{n}) + \chi\hat{V}^\star_{\hat{W}}(\hat{W}, S, \hat{X}) q^\star \tilde{n}.$$

Rearranging the expression above, we obtain:

$$\tilde{n}(\hat{W},S,\hat{X}) = \frac{1}{\chi}\left[\alpha(\theta^{\star}(\hat{W},S,\hat{X}))\frac{\tilde{\Omega}(\hat{W},S,\hat{X})}{q^{*}(\hat{X})} - sg(\tilde{n})\theta^{\star}(\hat{W},S,\hat{X})\tilde{v}_d(\hat{X})\right].$$

From Equation (A.23), we have that:

$$\theta^{\star}\,\tilde{v}_d = \eta\alpha(\theta^{\star})\frac{\tilde{\Omega}(\hat{W},S,\hat{X})}{q^{\star}(\hat{X})}\frac{\tilde{n}}{|\tilde{n}|}.$$

Therefore, we can write the expression for $\tilde{n}$ as follows:

$$\tilde{n}(\hat{W},S,\hat{X}) = \alpha(\theta^{\star}(\hat{W},S,\hat{X}))\frac{1-\eta}{\chi}\frac{\tilde{\Omega}(\hat{W},S,\hat{X})}{q^{*}(\hat{X})}. \tag{A.24}$$

Step 4: Dealers' value and intermediation fee. *Step 4.1: Dealers' value.* Using the fact that $\theta_t(n,\phi) = d_t(n,\phi)/\iota_t(n,\phi)$, the capacity constraint for dealers (4) with equality can be written as follows:

$$\int_{\Sigma}\iota_t(n,\phi)\theta_t(n,\phi)|n|d\varsigma = \overline{d}\epsilon,$$

where $\iota_t(n,\phi)$ is the mass of investors sending an order to contract (n,ϕ) at period t. We can write the expression above in terms of the joint distribution of wealth and asset holdings as follows:[5]

$$\sum_{j=1}^{2}\nu_j\int\theta_j(xq_t,s,\hat{X})|n_j(xq_t,s,\hat{X})|dG_{j,t}(x,s) = \overline{d}\epsilon,$$

where we used the fact that $\hat{W}_{i,t} = x_{i,t}q_t$ and $S_{i,t} = s_{i,t}$. Dividing the expression above by ϵ and taking the limit as $\epsilon \longrightarrow 0$, we obtain:

$$\sum_{j=1}^{2}\nu_j\int\theta_j^{\star}(xq^{\star},s,\hat{X})|\tilde{n}_j(xq^{\star},s,\hat{X})|dG_{j,t}(x,s) = \overline{d},$$

[5]To ease the notation, we performed the derivation of the value function and policy functions omitting the dependence on the investor's type. For the aggregation results, we need to take this dependence into account explicitly. Therefore, we use the notation $\tilde{n}_j(\hat{W},S,\hat{X})$ for the order size of an investor of type $j \in \{1,2\}$ and similarly for the remaining policy functions and value function.

Plugging in the expressions for the market tightness (A.23) and order size (A.24), we obtain:

$$\sum_{j=1}^{2} \nu_j \int \frac{\overline{\alpha}}{\chi}\frac{1-\eta}{\eta}\left[\frac{\overline{\alpha}}{\tilde{v}_d(\hat{X})}\right]^{\frac{1+\eta}{1-\eta}}\left[\frac{\tilde{\Omega}_j(xq^\star, s, \hat{X})}{q^\star(\hat{X})}\right]^{\frac{2}{1-\eta}} dG_{j,t}(x,s) = \overline{d},$$

Rearranging the expression above, we can solve for $\tilde{v}_d(\hat{X})$ as:

$$\tilde{v}_d(\hat{X}) = \left[\sum_{j=1}^{2} \nu_j \int \frac{1-\eta}{\eta\chi\overline{d}}\left[\overline{\alpha}\frac{\tilde{\Omega}_j(xq^\star, s, \hat{X})}{q^\star(\hat{X})}\right]^{\frac{2}{1-\eta}} dG_{j,t}(x,s)\right]^{\frac{1-\eta}{1+\eta}}.$$

Step 4.2: Intermediation fee. Using expression (10), we obtain the first-order correction for the intermediation fee (normalized by Y_t):

$$\tilde{\phi}_j(\hat{W}, S, \hat{X}) = \frac{\theta_j^\star(\hat{W}, S, \hat{X})}{\alpha(\theta_j^\star(\hat{W}, S, \hat{X}))}\tilde{v}_d(\hat{X})\tilde{q}^\star(\hat{X}).$$

From (A.23), we have that $\frac{\theta_j^\star \tilde{v}_d}{\alpha(\theta_j^\star)} = \eta\frac{|\tilde{\Omega}(\hat{W},S,\hat{X})|}{q^\star(\hat{X})}$, so we can write the expression above as follows:

$$\tilde{\phi}_j(\hat{W}, S, \hat{X}) = \eta|\tilde{\Omega}_j(\hat{W}, S, \hat{X})|.$$

Step 5: Interest rate. The market clearing condition for goods can be written as:

$$\sum_{j=1}^{2} \nu_j \int \left(x\frac{\hat{C}_j(x\hat{q}(\hat{X}), s, \hat{X})}{x\hat{q}(\hat{X})} + 0.5\chi n_j^2(x\hat{q}(\hat{X}), s, \hat{X})\right) dG_j(x,s) + v_{d,t}\overline{d}\epsilon = \frac{1}{q(\hat{X})}.$$

Taking a first-order approximation in ϵ of the expression above, we obtain:

$$\sum_{j=1}^{2} \nu_j \int x\left[(1-\psi)\tilde{r}(\hat{X}) + \tilde{\pi}(\hat{X})\frac{s}{x} - (1+\psi)\frac{\gamma_j}{2}\tilde{\sigma}_R^2(\hat{X})\left(\frac{s}{x}\right)^2\right] dG_j(x,s) = -\frac{\tilde{q}(\hat{X})}{q^\star(\hat{X})^2}. \tag{A.25}$$

The first-order approximation of $1/q_t$ can be expressed in terms of the interest rate and the risk premium:

$$-\frac{\tilde{q}(X)}{q^\star(X)^2} = \tilde{r}(X) + \tilde{\pi}(X), \tag{A.26}$$

using the fact that $\frac{1}{q_t} + \mu + \mu_{q,t} + \sigma\sqrt{\epsilon}\sigma_{q,t} = r_t + \pi_t$, $\mu_{q,t} = \mathcal{O}(\epsilon^2)$ and $\sqrt{\epsilon}\sigma_{q,t} = \mathcal{O}(\epsilon^2)$. From (A.25) and (A.26), we obtain:

$$\tilde{r}(\hat{X}) = -\frac{1+\psi^{-1}}{2}\sum_{j=1}^{2} \nu_j \int x\gamma_j \tilde{\sigma}_R^2(\hat{X}) \left(\frac{s}{x}\right)^2 dG_j(x,s).$$

□

Online Appendix

OA.1 Derivations

OA.1.1 Investors' flow budget constraints

Let $B_{i,t}$ denote the total amount invested in the risk-free asset at time t for investor i. Then, $B_{i,t}$ evolves according to:

$$dB_{i,t} = \left[r_t B_{i,t} + S_{i,t} Y_t - \frac{1}{2} p_t \chi n_{i,t}^2 - C_{i,t}\right] dt - p_t dS_{i,t} - \phi_{i,t} |dS_{i,t}|.$$

Let $W_{i,t} \equiv B_{i,t} + p_t S_{i,t}$ denote investor i's wealth, assessed at the inter-dealer price p_t. Investor's wealth evolves according to:

$$\begin{aligned} dW_{i,t} &= dB_{i,t} + dp_t S_{i,t} + p_t dS_{i,t} \\ &= \left[r_t B_{i,t} - \frac{1}{2} p_t \chi n_{i,t}^2 - C_{i,t}\right] dt + S_{i,t}(Y_t dt + dp_t) - \phi_{i,t} |dS_{i,t}| \\ &= \left[r_t W_{i,t} + p_t S_{i,t}(\mu_{R,t} - r_t) - \frac{1}{2} p_t \chi n_{i,t}^2 - C_{i,t}\right] dt + p_t S_{i,t} \sigma_{R,t} dZ_t - \phi_{i,t} |dS_{i,t}|, \end{aligned}$$

where $\mu_{R,t} = \dfrac{Y_t}{p_t} + \mu_{p,t}$ and $\sigma_{R,t} = \sigma_{p,t}$.

OA.1.2 Investors' flow budget constraints: No order execution risk

The evolution of wealth and stocks is given by:

$$\begin{aligned} dW_{i,t} &= \left[r_t W_{i,t} + (\mu_{R,t} - r_t) p_t S_{i,t} - \frac{1}{2} p_t \chi n_{i,t}^2 - C_{i,t}\right] dt + \sigma_{R,t} p_t S_{i,t} dZ_t - \phi_{i,t} |dS_{i,t}| \\ dS_{i,t} &= n_{i,t} dN_{i,t}. \end{aligned}$$

When there is not execution risk, it holds that $dN_{i,t} = n_{i,t} \alpha(\theta_{i,t}) dt$. Furthermore $\phi_{i,t} |dS_{i,t}| = \phi_{i,t} |n_{i,t} \mid \alpha(\theta_{i,t}) dt$. Thus, the budget constraint is now given by:

$$\begin{aligned} dW_{i,t} &= \left[r_t W_{i,t} + (\mu_{R,t} - r_t) p_t S_{i,t} - \frac{1}{2} p_t \chi n_{i,t}^2 - C_{i,t}\right] dt + \sigma_{R,t} p_t S_{i,t} dZ_t - \phi_{i,t} |n_{i,t} \mid \alpha(\theta_{i,t}) dt \\ dS_{i,t} &= n_{i,t} \alpha(\theta_{i,t}) dt. \end{aligned}$$

OA.1.3 Return volatility

The variance of returns is given by the following expression:

$$
\begin{aligned}
\sigma_R^2(X) &= \sigma^2\epsilon + 2\sigma\sigma_q\sqrt{\epsilon} + \sigma_q^2 \\
&= \sigma^2\epsilon + 2\sigma^2\frac{\tilde{q}_x}{q^\star}(s-x)\,\epsilon^2 + o(\epsilon^2),
\end{aligned}
$$

where we used that $\sigma_x(x,s) = (s-x)\sigma\sqrt{\epsilon} + \mathcal{O}(\epsilon)$ and $\sigma_q = \frac{q_x}{q}\sigma_x = \frac{\tilde{q}_x}{q^\star}(s-x)\sigma\epsilon\sqrt{\epsilon} + o(\epsilon\sqrt{\epsilon})$. Recall that the first-order approximation of the expression for the expected return on the risky asset is given by:

$$
\tilde{q}(x,s) = -q^\star(x,s)^2(\tilde{r}(x,s) + \tilde{\pi}(x,s)).
$$

The derivative of the expression above with respect to x is given by:

$$
\tilde{q}_x(x,s) = -q^\star(x,s)^2\left[\tilde{r}_x(x,s) + \tilde{\pi}_x(x,s)\right], \tag{OA.1.1}
$$

where

$$
\begin{aligned}
\tilde{r}_x(x,s) &= \frac{(\gamma+1)}{2}\left[\left(\frac{s}{x}\right)^2 - \left(\frac{1-s}{1-x}\right)^2\right]\gamma\sigma^2 \\
\tilde{\pi}_x(x,s) &= -\left[\tilde{\nu}\frac{s}{x^2} - (1-\tilde{\nu})\frac{1-s}{(1-x)^2}\right]\gamma\sigma^2.
\end{aligned}
$$

The variance of returns can be written as:

$$
\sigma_R^2(X) = \sigma^2\epsilon - 2\sigma^2 q^\star\left[\tilde{r}_x(x,s) + \tilde{\pi}_x(x,s)\right](s-x)\epsilon^2 + o(\epsilon^2).
$$

Note that $\left[\tilde{r}_x(x,s) + \tilde{\pi}_x(x,s)\right](s-x)\sigma\sqrt{\epsilon}$ is the sensibility of the discount rate to the aggregate shock. If this term is negative, so a negative shock increases the discount rate, then the volatility of returns will be larger than the volatility of the dividends σ. The reason is that a countercyclical discount rate amplifies the effect of shocks on dividends. Expected returns will be countercyclical if, for instance, the risk premium increases by more than interest rates fall in response to a negative shock.

OA.2 Data and Additional Motivating Evidence

OA.2.1 Corporate bond transactions data

We use data on secondary market transactions in the corporate bond market from the Trade Reporting and Compliance Engine (TRACE), made available by the Financial Industry Regulatory Authority (FINRA). The TRACE data provide detailed information on all secondary market transactions self-reported by FINRA member dealers. These include a bond's CUSIP, trade execution time and date, transaction price ($100 = par), the volume traded (in dollars of par), a buy/sell indicator, and flags for dealer-to-customer and inter-dealer trades.

We first filter the report data for trade corrections and cancellations following the procedure described in Dick-Nielsen (2014). We then merge the filtered data set with the TRACE master file, which contains bond grade information, and also with the Mergent Fixed Income Securities Database (FISD) to obtain bond fundamentals, such as issuing date, issuing amount, amount outstanding, etc. Following the literature, we exclude bonds with special characteristics, such as variable coupon, convertible, exchangeable, and puttable, as well as asset-backed securities and privately placed instruments. We follow Kargar et al. (2020) to calculate transaction costs for the so-called risky-principal trades where dealers hold bonds in inventories.

OA.2.2 Evidence form the GFC

During the 2007-2009 Global Financial Crisis (GFC), patterns similar to the COVID-19 crisis episode emerged for asset prices, credit spreads, and trading costs, and turnover appeared, as shown in Figure OA.1. In the corporate bond market, trading costs and turnover increased, while bond prices declined simultaneously, albeit more gradually than for the COVID-19 episode.

OA.3 State-global Perturbations

In this section, we present the state-global perturbation method in more detail and show how they differ from standard linearization methods. We start by discussing the standard linearization procedure and its limitations, and then study the state-global perturbation techniques and show how they can overcome these limitations.

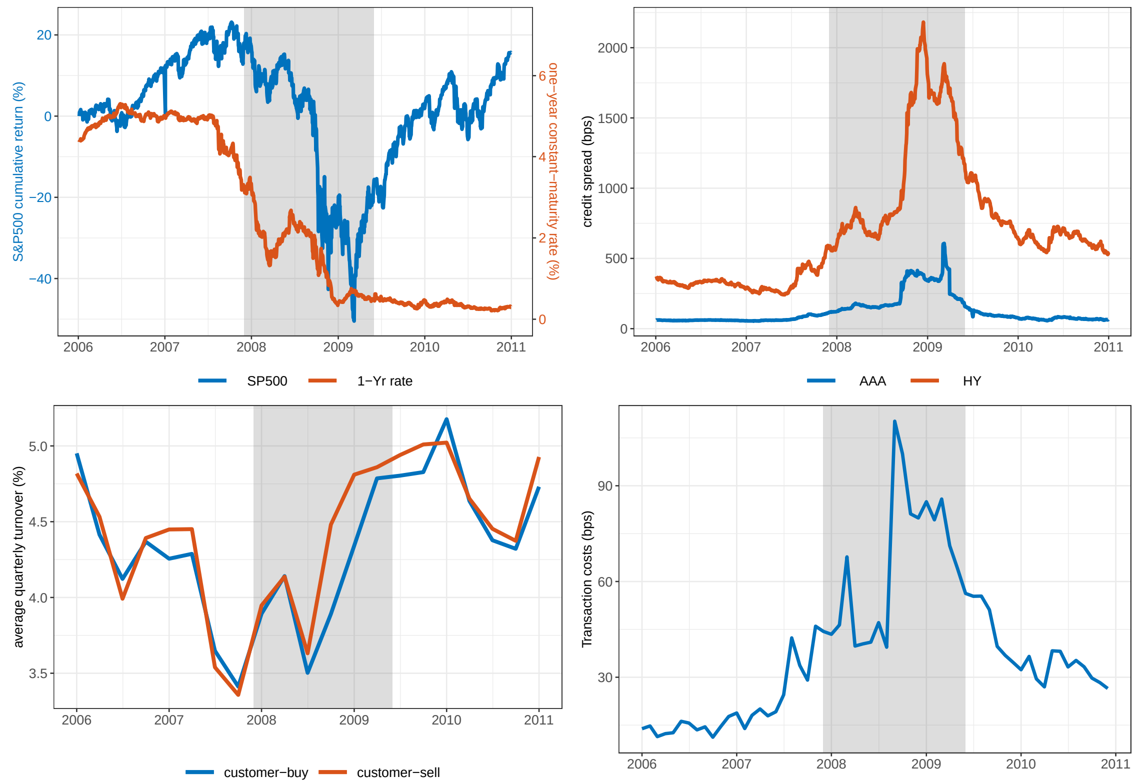

Figure OA.1. Stock returns, interest rates, credit spreads, turnover, and transaction costs for corporate bonds during the Great Recession. The vertical shaded bars indicate NBER recessions. Source: TRACE, Bloomberg, and FRED.

OA.3.1 Linearization

Discrete-time dynamic stochastic general equilibrium (DSGE) models can typically be written as:

$$\mathbb{E}_t \mathcal{H}(y, y', x, x'; \epsilon) = 0, \tag{OA.3.1}$$

where y is a $n_y \times 1$ vector of controls, x is a $n_x \times 1$ vector of states, and $n = n_x + n_y$. The function $\mathcal{H}(\cdot)$ returns a $n \times 1$ vector. The parameter ϵ controls the volatility of innovations and it is a perturbation parameter.

The solution of the model consists of policy functions and the laws of motion for the state variables:

$$y = g(x; \epsilon) \tag{OA.3.2}$$

$$x' = h(x; \epsilon) + \epsilon \Sigma u, \tag{OA.3.3}$$

where g maps $\mathbb{R}^{n_x} \times \mathbb{R}_+$ into $\mathbb{R}^{n_y}$, h maps $\mathbb{R}^{n_x} \times \mathbb{R}_+$ into $\mathbb{R}^{n_x}$, Σ is a $n_x \times n_u$ matrix, and u is a $n_u \times 1$ vector of white noise disturbances.

The case where $\epsilon = 0$ corresponds to a perfect foresight economy in which there is no uncertainty. We assume that the policy functions are known at the non-stochastic steady state, that is, at $x = \overline{x}$ we have:

$$\mathcal{H}(\overline{y}, \overline{y}, \overline{x}, \overline{x}, 0) = 0, \tag{OA.3.4}$$

where $\overline{y} = g(\overline{x}, 0)$ and $\overline{x} = h(\overline{x}, 0)$.

The functions g and h are usually characterized by using a first-order perturbation method, also known as linearization. In particular, we take a Taylor expansion of g and h around $x = \overline{x}$ and $\epsilon = 0$:[1]

$$g(x; \epsilon) = g(\overline{x}; 0) + g_x(\overline{x}; 0)(x - \overline{x}) + g_\epsilon(\overline{x}; 0)\epsilon + \mathcal{O}(||x - \overline{x}, \epsilon||^2) \tag{OA.3.5}$$

$$h(x; \epsilon) = h(\overline{x}; 0) + h_x(\overline{x}; 0)(x - \overline{x}) + h_\epsilon(\overline{x}; 0)\epsilon + \mathcal{O}(||x - \overline{x}, \epsilon||^2). \tag{OA.3.6}$$

OA.3.1.1 Key features of the linearization solution

Three aspects of the solution are important to emphasize. First, the solution is accurate to the extent that the perturbation parameter ϵ and the distance to the steady state $||x - \overline{x}||$ are small. The analysis is, therefore, valid only in the neighborhood of the steady state and cannot inform the economy's response to large shocks, when the state variable x can deviate significantly from $\overline{x}$. Importantly, this local approximation requires only local information; that is, one needs to know only $\overline{y}$ and $\overline{x}$ to compute the solution. Second, the solution is linear in the state variable x. This implies, in particular, that the effect of shocks u on the policy functions y is constant, so the first-order approximation is not able to capture any state-dependency of the effects of shocks. Third, as shown by Schmitt-Grohé and Uribe (2004), the coefficients on the parameter ϵ are equal to zero, that is, $g_\epsilon(\overline{x}, 0) = h_\epsilon(\overline{x}, 0) = 0$. As a result, the solution satisfies the certainty-equivalence property, where the policy functions coincide with the one in a non-stochastic economy. Risk does not affect the economic outcomes up to a first-order approximation. In particular, the risk premium would be equal to zero and portfolio choice would be indeterminate. A second-order approximation would be required to obtain a non-zero risk premium, and a third-order approximation would be required to capture time variation in the risk premium.

[1] The coefficients of the approximation can be computed as follows. First, define the function $F(x, \epsilon) \equiv \mathbb{E}_t \mathcal{H}(g(x; \epsilon), g(h(x; \epsilon) + \epsilon\Sigma u; \epsilon), x, h(x; \epsilon) + \epsilon\Sigma u; \epsilon) = 0$, which is identically equal to zero. The derivatives of $F(x, \epsilon)$ are then also equal to zero. In the regular case, the coefficients on $(x - \overline{x})$ and ϵ can be computed by solving the system of equations $F_x(\overline{x}, 0) = 0$ and $F_\epsilon(\overline{x}, 0) = 0$.

OA.3.2 State-global perturbation

We next consider a continuous-time problem, which can be written in general form as follows:

$$\mathcal{H}(y, y_x, y_{xx}, x; \epsilon) = 0, \tag{OA.3.7}$$

where y is a $n_y \times 1$ vector of controls, x is a $n_x \times 1$ vector of states, and ϵ is a perturbation parameter. The function $\mathcal{H}(\cdot)$ returns a $n \times 1$ vector.

The solution of the model consists of policy functions and the law of motion of state variables:

$$y = g(x; \epsilon) \tag{OA.3.8}$$

$$dx_t = \mu^x(y, y_x, y_{xx}, x; \epsilon)dt + \sigma^x(y, y_x, y_{xx}, x; \epsilon)dZ_t. \tag{OA.3.9}$$

We assume that the risk exposure of the state variable, $\sigma^x(\cdot)$, is equal to zero when $\epsilon = 0$, so we focus on a small risk approximation. Notice that ϵ can also directly affect other equations in the system, which would include the case discussed in Section 3 where the intermediation capacity parameter was also a function of ϵ.

In contrast to the standard linearization procedure, we assume that the policy is known for all values of x at $\epsilon = 0$, that is, the function $g(x, 0)$ is known, as well as the drift and diffusion of x at $\epsilon = 0$, for all x. Therefore, the method requires global information on the solution, not only on the value of the solution at the steady state. In the context of our model, We provide this global solution in Lemma 1, which takes a relatively simple form in our case.

We next consider a first-order perturbation in ϵ, instead of an approximation in x and ϵ, which assumes the following form:[2]

$$g(x; \epsilon) = g(x; 0) + g_\epsilon(x; 0)\epsilon + \mathcal{O}(\epsilon^2), \tag{OA.3.10}$$

and analogous expressions hold for $\mu^x(\cdot)$ and $\sigma^x(\cdot)$.

[2]The coefficients of the approximation can be computed as follows. First, define the function $F(x, \epsilon) \equiv \mathcal{H}(g(x; \epsilon), g_x(x; \epsilon), g_{xx}(x; \epsilon), x; \epsilon) = 0$, which is identically equal to zero. The derivatives of $F(x, \epsilon)$ are then also equal to zero. In the regular case, the coefficients on ϵ can be computed by solving the system of equations $F_\epsilon(x, 0) = 0$ for all x.

OA.3.2.1 Key features of the state-global solution

Three aspects of the solution are important to emphasize. First, the solution is accurate to the extent that the perturbation parameter ϵ is small. Importantly, there is no requirement that the state variable x needs to be close to the non-stochastic steady state value $\bar{x}$, allowing us to study the behavior of the economy far from the steady state, which is crucial when considering large shocks. Second, the solution is potentially non-linear in the state variable x. This enables us to capture the state-dependent effects of shocks, where the impact of shocks depend on its magnitude and initial condition. Third, the coefficient on ϵ is, in general, non-zero and it is a function of the state variable x. This implies that the certainty-equivalence property is not satisfied, so risk has a first-order impact on the policy functions, and we are able to obtain a time-varying risk premia without having to resort to higher-order approximations.

OA.3.2.2 The role of bifurcation theory

In the regular case where the elements of $g_\epsilon(x;0)$ can all be determined, the coefficients $g_\epsilon(x;0)$ can be obtained by applying the implicit function theorem. However, as is often the case in the small-risk approximations of portfolio problems, some elements of the solution are indeterminate in the non-stochastic economy. For instance, in a frictionless portfolio problem, the portfolio share on the "risky" asset is indeterminate when there is no risk. In our case, due to the presence of transaction costs, the order size is determined in the limit with no risk, as it is optimal to not trade in that case. In contrast, the market tightness is indeterminate, as the trading speed is not relevant when investors have no incentive to trade. This indeterminacy leads to a violation of the regularity condition necessary to apply the implicit function theory, which implies that a different method is necessary to compute the coefficients $g_\epsilon(x;0)$.

Judd and Guu (2001) show that bifurcation theory can be used to compute approximations in situations where the implicit function theorem does not hold. Intuitively, the indeterminacy is resolved by considering the limit of the solution as $\epsilon \longrightarrow 0$, which selects a unique value for the solution at $\epsilon = 0$. We apply the methods of Judd and Guu (2001) to compute the value of the market tightness in Proposition 3.

OA.4 The Role of the Trading Elasticity

The degree of amplification of the risk premium is related to the relative trading elasticity. The elasticity of buy and sell orders is not constant in this economy, despite the iso-elastic preferences. The demand elasticity depends on the matching function parameter η and on how far an investor is from the target portfolio:

$$\frac{\partial \log \mathcal{D}}{\partial \log \pi} = \frac{1+\eta}{1-\eta} \frac{\tilde{\pi}(X)}{\tilde{\pi}(X) - \gamma \sigma_R^2(X)\omega_b}, \tag{OA.4.1}$$

where ω_b is the portfolio share of buyers. It can be shown that the ratio of the trading elasticity of buyers to sellers is given by $(1 - \tilde{\nu})/\tilde{\nu}$, when type-1 investors are sellers. Therefore, when $\tilde{\nu}$ is large, it means that demand for shares is relatively inelastic and this coincides with the region where there is amplification of the risk premium. The right panel of Figure OA.2 shows how the risk premium responds as we vary the ratio of demand to supply trading elasticity. We find that there is amplification in the inelastic demand region, effectively relating the amount of selling pressure (on the extensive margin) to the market elasticity. This result is in line with the role of the market demand elasticity discussed by Gabaix and Koijen (2020).

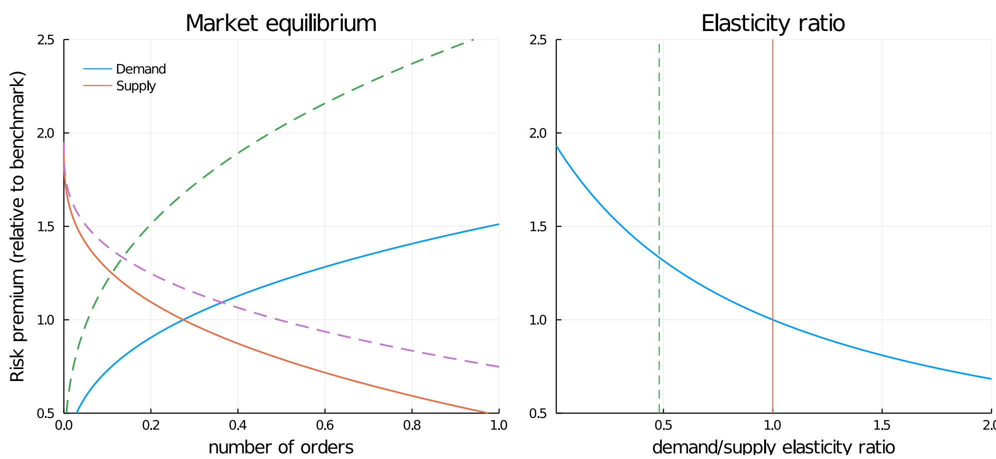

Figure OA.2. The left panel of this figure depicts the determination of orders and the risk premium. The right panel shows the relation between the ratio of demand and supply trading elasticities and risk premium. The solid lines represent the case $\nu = 0.5$ and the dashed lines represent the case $\nu > 0.5$. To construct this figure, we use the calibration discussed in Table 1 of Section 4 specialized to the case of no heterogeneity in risk aversion.