Which Intermediary Costs Matter for Asset Prices?*

Manav Chaudhary† Julie Zhiyu Fu‡ Jian Li§

July 7, 2026

[Click here for the latest version]

Abstract

When intermediaries such as dealers lack the balance-sheet capacity to absorb investor demand, asset prices deviate from fundamentals. Arbitrage spreads (e.g., Treasury-OIS spreads) are widely used to diagnose such distortions, yet it is the price level (e.g., Treasury yields) that ultimately governs borrowing costs and monetary policy transmission. We study to what extent spreads reflect distortions that matter for price levels. We develop a model where intermediaries face two costs: one proportional to portfolio risk, another tied to gross position size. While gross position costs predominantly drive spread distortions, risk costs primarily drive price-level distortions. The theory delivers a sufficient statistic: differences in the rate at which price levels and spreads revert after a demand shock separately identify these costs. We apply this framework to U.S. Treasury and OIS markets using high-frequency demand shocks identified from Treasury auctions. Risk costs dominate on average: demand shocks move yields substantially while leaving spreads largely unchanged. A calibrated model reinforces the disconnect: relaxing position costs such as the supplementary leverage ratio sharply reduces spread volatility with little effect on yield volatility, while easing risk-based costs does the reverse. Overall, spreads alone miss the dominant source of price-level distortions.

*We thank Ashwini Agarwal, Daniel Barth, Kent Daniel, Greg Duffee, Peter Kondor, Ian Martin, Daniel Neuhann, Tomasz Piskorski, Tano Santos, Dimitri Vayanos, Stijn Van Nieuwerburgh, Laura Veldkamp, Kathy Yuan and seminar participants at the Bank of England, Bayes Business School, Booth Banking Workshop, Bridging Theory and Empirical Research in Finance Conference, Columbia Business School, David Backus Memorial Conference on Macro-Finance, Demand in Asset Markets workshop, HEC Paris, Imperial College, Frankfurt School, London Junior Conference, London School of Economics, Goethe University, Olin Business School, University College London, and WAPFIN for helpful comments.

†London School of Economics. Email: m.chaudhary6@lse.ac.uk

‡Washington University in St. Louis, Olin Business School. Email: z.fu@wustl.edu

§Columbia Business School. Email: jl5964@columbia.edu

1 Introduction

Arbitrage spreads—the price difference between two assets that deliver identical payoffs—are a standard way to assess whether intermediaries lack the balance-sheet capacity to keep asset prices aligned with fundamental values. The Treasury-OIS spread, the cash-futures basis, and the Covered Interest Parity (CIP) spread are prominent examples. Absent frictions, intermediaries such as broker-dealers and hedge funds should eliminate these spreads by trading against them. When an arbitrage relationship "breaks," this signals that intermediation capacity is scarce. Central banks treat these spreads as barometers of market functioning,[1] and a large empirical literature uses them to track the effects of intermediary constraints. But the real economic consequences of limited intermediation capacity—borrowing costs, discount rates, the transmission of monetary policy—depend on how far the *level* of asset prices, such as Treasury yields, is pushed from fundamentals, not on spreads.

Do arbitrage spreads provide an accurate picture of these price-level distortions? In canonical intermediary asset pricing models, intermediaries face risk-based costs that drive asset price levels (He and Krishnamurthy, 2013; Brunnermeier and Sannikov, 2014; Vayanos and Vila, 2021). But these risk costs do not penalize riskless arbitrage trades, so spreads are eliminated even when price levels are substantially distorted. Models of arbitrage violations show that spreads arise from costs tied to gross positions (Gromb and Vayanos, 2002; Gârleanu and Pedersen, 2011; Du, Tepper, and Verdelhan, 2018), but do not address what these costs imply for price levels. Answering this question requires a framework that jointly models how different intermediation costs affect both spreads and price levels, and empirical tools to tell them apart.

We build a model with slow-moving capital in which intermediaries arbitrage across a cash and synthetic asset with identical payoffs. Intermediaries face two broad costs: a risk-based cost penalizing portfolio risk and a gross position cost tied to the size of positions. These map into the two ways constraints limit balance-sheet capacity—on composition (e.g. Stress Capital Buffer or Value at Risk) and, separately, on size (e.g. the Supplementary Leverage Ratio or margin requirements). The model delivers a sharp distinction between them: gross position costs matter primarily for arbitrage spreads, while risk-based costs are the dominant force behind movements in price levels. Indeed, absent gross position costs, the model admits an equilibrium where two identical assets are perfectly integrated and the

[1] The Federal Reserve's expansion of Treasury purchases in March 2020 was prompted in part by the breakdown of the Treasury cash-futures basis, and the Bank of England's temporary gilt purchase programme in September 2022 was triggered by sharp dislocations in gilt spreads. In both cases, subsequent spread compression was cited as evidence that the interventions had restored market functioning. See Board of Governors of the Federal Reserve System (2020) and Bank of England (2025) for more details.

spread is always zero. While spreads are informative of the gross position cost, they are not informative of the risk-based cost—the dominant driver of price-level distortions.

Given the different implications for spreads and price level, it is important to quantify the relative magnitude of the two costs. Our model delivers a sufficient statistic to separately identify the two costs. An exogenous demand shock generates a price impact whose initial rate of decay is pinned down by intermediary costs. Because the two costs affect yields and spreads differently, their distinct decay patterns allow separate identification. We take this approach to the U.S. Treasury and overnight index swap (OIS) markets, using high-frequency demand shocks around Treasury auction releases (Ray, Droste, and Gorodnichenko, 2024). We find that risk-based costs account for the dominant share of the cost intermediaries face when trading against price-level distortions. Zooming into the GFC and COVID, we find gross position costs appear to play a larger role, though risk-based costs remain important. The shift is particularly pronounced during COVID, consistent with the SLR binding. To assess what spreads actually revealed about price-level distortions in this episode, we calibrate the model. While spreads directionally reflected increased distortions in Treasury yields, they substantially overstated their magnitude. Finally, we use the calibrated model to evaluate the impact relaxing dealer regulation may have on Treasury market functioning. There is a sharp asymmetry: reducing gross position costs compresses spread responses substantially while leaving yield dynamics largely unchanged, whereas reducing risk costs does the reverse. Overall, spreads provide a limited window into price-level distortions—they primarily reflect gross position costs, which are not the dominant driver of price levels.

We describe the model in more detail. We model two assets with identical fundamental values, traded by three types of agents: institutional investors such as pension funds and asset managers, intermediaries such as dealers and hedge funds, and noise traders. Institutional investors are long-term investors with downward-sloping demand that depends on the assets' prices and their fundamental value. To capture the slow adjustments in their positions, we assume they can only partially adjust their position towards the target demand each period. Intermediaries act as arbitrageurs between the two markets, facing two types of costs: a risk-based cost for bearing portfolio risk, and a gross position cost that captures constraints tied to the size of asset positions. The two are distinct: the risk-based cost reflects intermediaries' concern for the overall risk of their portfolio, while the gross position cost makes balance sheet space costly even when a position adds little to portfolio risk. Finally, noise traders trade for exogenous reasons and generate demand shocks.

We use the model to study how different types of intermediation costs shape prices and spreads differentially. Risk-based costs *integrate* the two markets, while gross position costs

segment them. The former encourages intermediaries to take offsetting positions that hedge risk, linking the two markets through active arbitrage and raising return correlation; the latter penalizes the gross positions required to do so, driving prices apart. As a result, gross position costs disproportionately affect spread volatility, while risk-based costs have a much larger impact on price volatility. This distinction determines when spreads are—and are not—informative about distortions in price levels.

Given their different implications, we seek to learn about the relative importance of these two costs. Our approach relies on a sufficient statistic: the speed at which prices and spreads revert following an exogenous demand shock. The logic is as follows. Intermediaries act as short-horizon liquidity providers—they absorb a demand shock on impact and gradually unwind their inventories as slow-moving institutional investors take over the position. The impulse response of prices therefore has two components: a long-run response pinned down by institutional demand, and a transitory amplification reflecting the compensation intermediaries require to hold inventory while this adjustment takes place.

We cannot identify intermediation costs from the price impact directly, since the observed response at any horizon reflects both end-investor demand and intermediary costs. But we can identify them from the *rate of reversion*. In a competitive market, the return an intermediary earns on its inventory—the rate at which the transitory price impact decays—equals its marginal cost of intermediation. The initial slopes of the price reversion and spread reversion depend on the risk costs and the gross position costs differentially. Hence, using estimates of the two slopes, we can separately identify the two costs. This identification strategy applies broadly across a wide set of models. The key assumption is that intermediaries are the marginal liquidity providers on impact, a common feature in many models of intermediation that is consistent with the empirical evidence.

We apply this framework to study the U.S. Treasury cash market and the overnight index swap (OIS) market. Following Ray, Droste, and Gorodnichenko (2024), we identify Treasury demand shocks using the high-frequency price impact of Treasury auction result releases. In U.S. Treasury auctions, the amount to be issued is preannounced, but the investor demand is only revealed when the auction closes and results are published. As a result, yield movements in a tight window around the result release capture unanticipated variation in demand and provide a high-frequency proxy for exogenous demand shocks to the Treasury market. Treasury auctions are a particularly attractive source of demand shocks: they are the dominant driver of primary-dealer inventory changes (Fleming, Nguyen, and Rosenberg, 2024) and are accompanied by elevated secondary market trading activity (Somogyi, Wallen, and Xu, 2025).

We estimate impulse responses of yields and spreads to auction shocks using local projections and recover initial decay rates from the estimated impulse responses. The theory framework then allows us to map these initial decay rates to the underlying intermediation costs. Because we do not have a direct measure of the size of the demand shocks, we can only identify the relative importance of the two types of intermediation costs, not their absolute magnitudes.

In the full sample (2008–2022), we find that risk-based costs are, on average, the dominant driver of price-level dynamics. In response to auction shocks, Treasury yields and OIS rates move closely together, while the Treasury–OIS spread exhibits little systematic response, implying that risk-based costs dominate intermediation costs on average. Hence, focusing solely on spreads would tend to understate the extent of illiquidity in the market, as spreads miss much of the relevant price-level variation in the Treasury market.

The crisis periods tell a different story. In both the GFC and COVID episodes, gross position costs appear to play a larger role than in normal times, consistent with the widening of no-arbitrage spreads during these periods. As the risk contribution measure declines, we find that the spread response to demand shocks is larger and the spread variance increases more relative to level variance— the pattern our model predicts when gross-position costs are elevated. These effects are especially pronounced during COVID, consistent with the Supplementary Leverage Ratio (SLR) binding after Dodd–Frank. This suggests that in a post-SLR regulatory regime, spreads may become more informative about price-level distortions during stress, but quantifying their implications for price dynamics requires a calibrated model.

Our model suggests that the short-run variance of Treasury yields and OIS spreads relative to the long-run variance captures the amplification effect of intermediation costs. Hence we calibrate the model by targeting the unconditional term structures of variance of Treasury yields and the Treasury-OIS spread. We then conduct two exercises: First, we apply the calibrated model to the COVID-19 episode to quantify the role of elevated gross position costs. The elevated gross position cost increased spread variance roughly ten times more than yield variance. So while spreads directionally reflected increased distortions in the Treasury yield during COVID-19, they overstated the magnitude of the increase in yield distortions.

Second, given the current debate on relaxing banking regulations, we examine how hypothetical reductions in each cost component affect Treasury market functioning. A sharp asymmetry emerges: a reduction in the gross position costs (as would arise from relaxing the SLR) compresses the spread response to demand shocks and spread volatility substantially,

while leaving yield dynamics relatively unchanged. A reduction in the risk cost (as would arise from relaxing the Stress Capital Buffer) does the opposite, materially reducing yield price impact with little effect on the spread. These counterfactuals illustrate how different regulatory tools affect Treasury market functioning through distinct channels.

The rest of the paper proceeds as follows. We review the relevant literature in the remainder of this section. Section 2 describes the model and our theoretical results. Section 3 discusses our identification strategy and describes the estimation procedure. Section 4 presents the empirical results. Section 5 calibrates the model and presents counterfactual exercises. Section 6 concludes.

1.1 Literature

We contribute to the intermediary asset pricing literature by addressing two related gaps that have received limited systematic attention: the absence of a unified framework that jointly models how different types of intermediation costs affect prices and spreads, and the disconnect between empirical evidence from arbitrage spreads and implications for asset price levels (Haddad and Muir, 2025). We tackle both challenges with a theoretical framework and an empirical identification strategy that the framework motivates, quantifying the relative importance of different types of intermediation costs.

The theoretical literature models intermediaries as marginal investors and studies how their constraints and frictions affect asset prices. Both risk-based constraints, such as capital adequacy tests and value-at-risk limits, and gross position costs, such as the supplementary leverage ratio and margin requirements, feature prominently in theoretical work (e.g., Gromb and Vayanos, 2002; Adrian and Shin, 2014; Gârleanu and Pedersen, 2011; Du, Hébert, and Huber, 2023). In particular, recent work has highlighted the role of position costs in Treasury market dysfunction during the Covid crisis (He, Nagel, and Song, 2022; Kashyap et al., 2025).[2] Much of the existing work, however, either focuses on a single asset without separately identifying risk-based and gross position costs (e.g., He and Krishnamurthy, 2013; Samuel G. Hanson, Malkhozov, and Venter, 2024), or studies one type of constraint in isolation (e.g., Kondor, 2009; Du, Hébert, and W. Li, 2023). We provide a two-asset framework that jointly studies the effects of both types of costs on prices and spreads.[3] The model delivers a sharp distinction: risk-based costs disproportionately affect price levels, while gross

[2]Given the breadth of this literature, we do not attempt a comprehensive review here. For recent surveys, see He and Krishnamurthy (2018) and Haddad and Muir (2025).

[3]Hazelkorn, Moskowitz, and Vasudevan (2023) construct a model in which arbitrageurs face balance sheet cost and liquidity providers are risk averse. We highlight the difference in how the two types of costs affect spread and price level differentially, and quantify their relative importance.

position costs disproportionately affect spreads.

The empirical literature shows that intermediary leverage comoves with returns and prices risk across asset classes (e.g., Adrian, Etula, and Muir, 2014; He, Kelly, and Manela, 2017; Haddad and Muir, 2021), supporting intermediary asset pricing models. Such evidence, however, is subject to critiques regarding the endogeneity of intermediary leverage (Santos and Veronesi, 2022). The existence of arbitrage spreads in heavily intermediated markets, such as CIP deviations, and their widening when intermediary constraints tighten provide more direct evidence in favor of intermediary frictions (see, for example, Du, Tepper, and Verdelhan, 2018; Fleckenstein and Longstaff, 2020; Boyarchenko et al., 2018; Duffie et al., 2023; Barth and Kahn, 2025).[4] The remaining question is quantitative: how informative are spread movements about distortions in price levels? We contribute by separately identifying the two types of costs and showing that risk-based costs are the dominant driver of price-level dynamics, while spreads alone are a limited diagnostic for the extent of intermediation costs. Interestingly, Klingler and Sundaresan (2023) find that in the Treasury bill market, changes in Treasury yields are the predominant driver of the spread changes, highlighting the role of balance sheet cost in the short-term market.

Our methodology draws on the recent literature focusing on the role of demand shocks in asset pricing. Our model builds on the workhorse framework of Vayanos and Vila (2021), which studies how arbitrageurs absorb demand shocks. Our empirical strategy builds on the key insight from demand-system asset pricing that the price response to a demand shock reveals the underlying frictions faced by the arbitrageur (Koijen and Yogo, 2019; Gabaix and Koijen, 2021). The existing literature typically estimates price impact or demand elasticities in reduced form. We take a step further by interpreting the impulse responses of prices and spreads to a demand shock through the lens of our model, thereby identifying the structural parameters governing risk-based and gross position costs. Existing work in demand-system asset pricing highlights the importance of accounting for heterogeneous substitution in fixed income markets (Chaudhary, Fu, and J. Li, 2025); we study what determines the intensity of substitution between similar assets and limits to arbitrage more broadly.

2 Model

To study the implication of intermediary frictions on asset prices and spreads, we introduce a preferred-habitat style model with two assets (Vayanos and Vila, 2021). We first describe

[4] D. Li, Petrasek, and Tian (2025) show that during COVID, dealer risk limits also played a role in the Treasury market.

the model setup in detail and analyze the equilibrium. We then discuss how different types of intermediary costs matter for asset prices and their relative spreads, and characterize the price dynamics following demand shocks that form the basis for our identification strategy.

2.1 Setup

Time is continuous and there are two assets with identical fundamental values. Our model features three types of agents: institutional investors that demand both assets but adjust their capital only sluggishly, an intermediary sector that arbitrages across markets and noise traders in each market.[5]

Assets. There are two assets in the economy: a cash asset denoted by C, such as the Treasury bonds, and a synthetic asset denoted by S, which is a derivative contract that replicates the payoff of the cash asset, such as the OIS swap. The two assets have the same fundamental value V_t, which follows a Brownian motion. Absent any frictions, the prices of the two assets should be identical and equal to V_t, i.e., $P_{C,t} = P_{S,t} = V_t$. We interpret V_t as the long-term fundamental value of the two assets. We are interested in how different types of intermediation costs affect the level of asset prices and the relative spread between the two assets, defined as $P_{S,t} - P_{C,t}$, differently. We use $\mathbf{P}_t$ to denote the 2 by 1 price vector of the two assets.

Institutional Investors. The main focus of this paper is on the intermediary sector; to this end, we model institutional investors in a stylized fashion, as slow-moving capital with reduced-form demand. These are long-term investors who care about the fundamental payoff of the assets. They have downward sloping demand curves and each period, their positions adjust only partially toward their target allocation (Duffie, 2010). These could be mutual funds, pension funds and insurance companies — investors that either face tight investment mandates or lack the sophistication (or incentives) to arbitrage actively across markets. In other words, the two assets are not perfectly substitutable for the end investors. They are also often slow to respond when prices deviate from fundamentals, hence we model them as slow-moving capital, similar to Greenwood, Samuel G Hanson, and Liao (2018).

We denote their holdings in the cash and synthetic asset markets by $y_{C,t}$ and $y_{S,t}$, respectively. Each instantaneous period, a fraction kdt of the investors in each market re-optimize their holdings to their target demand $Z_{i,t}$ ($i \in \{C, S\}$), while the rest keep their previous

[5] We separate noise traders from institutional investors for expositional clarity, but their shocks can equally be thought of as arising from institutional investors.

holdings. The target demand curve of the institutional investors who re-optimize in each period is given by

$$\begin{pmatrix} Z_{C,t} \\ Z_{S,t} \end{pmatrix} = -\boldsymbol{\zeta} \begin{pmatrix} P_{C,t} - V_t \\ P_{S,t} - V_t \end{pmatrix} + \boldsymbol{\theta}, \tag{1}$$

where $\boldsymbol{\zeta}$ is the demand elasticity matrix. The target demand of institutional investors is decreasing in the prices and increasing in the fundamental value. The constant term $\boldsymbol{\theta}$ captures an exogenous component of the demand, which captures other unmodelled components of demand such as institutional investors' hedging incentives and flows.

The dynamics of institutional investors' aggregate holdings are hence given by

$$\begin{pmatrix} dy_{C,t} \\ dy_{S,t} \end{pmatrix} = k \begin{pmatrix} Z_{C,t} - y_{C,t} \\ Z_{S,t} - y_{S,t} \end{pmatrix} dt. \tag{2}$$

The institutional investors' holdings as a sector adjust slowly to their targeted demand.

Intermediaries. The intermediary sector acts as the arbitrageur between the two markets. An important feature of intermediaries is that they provide liquidity to the markets when there are demand shocks. They are risk-averse and choose their positions in the two asset markets, $x_{i,t}$ ($i \in \{C, S\}$) to maximize their instantaneous expected utility, subject to two types of intermediation costs: risk-based cost and an additional gross position cost in each asset. More specifically, the intermediary's optimization problem is given by

$$\max_{x_{C,t}, x_{S,t}} \mathbb{E}_t \left[dW_t \right] - \frac{\gamma}{2} \operatorname{Var}_t(dW_t) - \left(\frac{\psi_C}{2} x_{C,t}^2 + \frac{\psi_S}{2} x_{S,t}^2 \right) dt \tag{3}$$

where

$$dW_t = x_{C,t} dP_{C,t} + x_{S,t} dP_{S,t} + \left(W_t - \sum_{i \in \{C,S\}} x_{i,t} P_{i,t} \right) r dt. \tag{4}$$

The first term captures the expected return of the intermediaries' portfolio dW_t, where r is the risk-free rate. Since intermediaries are short-term investors, they do not care about the fundamental value of the assets and only care about the instantaneous return of their portfolio. Given the frequency of our empirical analysis is at the daily level, the risk-free return is close to zero. Hence in the theoretical analysis we set it to zero for simplicity. All results are robust to the general case.

The next two terms represent two types of costs faced by intermediaries. The term $\frac{\gamma}{2}\operatorname{Var}_t(dW_t)$ captures the intermediaries' risk-based cost, and γ is the risk-aversion coefficient. This cost is proportional to the riskiness of their entire portfolio, reflecting intermediary costs such as value-at-risk limits and other variance-based constraints. The last term $\left(\frac{\psi_C}{2}x_{C,t}^2 + \frac{\psi_S}{2}x_{S,t}^2\right)dt$ captures the additional cost related to each asset position, reflecting balance sheet costs that are increasing in position size but do not depend on the rest of the portfolio, such as the Supplementary Leverage Ratio (SLR) constraint. We refer to ψ_C and ψ_S as the gross position cost parameters.

The intermediaries' optimization problem leads to the following first-order condition:

$$\mathbb{E}_t\left[\begin{pmatrix} dP_{C,t} \\ dP_{S,t} \end{pmatrix}\right] = \underbrace{(\gamma\boldsymbol{\Sigma}_P + \boldsymbol{\Psi})}_{\mathbf{C}}\begin{pmatrix} x_{C,t} \\ x_{S,t} \end{pmatrix} dt \tag{5}$$

where $\boldsymbol{\Sigma}_P \equiv \operatorname{Cov}(d\mathbf{P}_t)/dt$ is the instantaneous covariance matrix of asset price changes, $\boldsymbol{\Psi} \equiv \operatorname{diag}(\psi_C, \psi_S)$, and $\mathbf{C}$ represents the total instantaneous marginal cost matrix.

Intuitively, the intermediary's expected return must equal the marginal cost of holding a given inventory position. The cost matrix $\mathbf{C}$ reflects two sources of friction: the risk-based cost, which scales with the intermediary's risk aversion and the instantaneous variance of the portfolio, and the position cost, captured by the diagonal matrix with entries ψ_C and ψ_S, penalizing large asset positions regardless of portfolio composition.

We do not attach a literal interpretation to either cost term. Instead, the two terms summarize the broad classes of intermediation costs emphasized in the literature and in policy discussions. The risk-based term captures forces that make intermediaries care about overall portfolio risk, including risk aversion and regulations tied to measured risk. The position-cost term captures constraints, such as the SLR, that make balance sheet space costly even when a position adds little portfolio risk. We keep the position cost deliberately simple—the marginal cost is linear in the position $\boldsymbol{\Psi}\mathbf{x}_t$, as the goal is not to model a specific regulation, but to summarize the marginal cost of using balance sheet capacity in a form that can be taken to the data. With this interpretation, the marginal cost in the first-order condition (5) can be seen as a local approximation to a broader set of constraints and costs. In Appendix A.7, we discuss how different constraints and costs fit into this framework and how they affect the interpretation of our results.[6]

Because the intermediaries maximize their instantaneous expected utility, the dynamics only matter through future prices. In Appendix A.8, we consider a fully dynamic model

[6]Related, if the intermediary is long-lived with hedging demand, we can also represent the intermediary's first order condition in a similar form.

where the intermediary is long-lived and has hedging demand. We discuss the mapping between the two models and how it affects the interpretation of our results on intermediary costs.

Noise Traders. In each market, the noise traders hold $\beta_{i,t}$ ($i \in \{C, S\}$) units of the asset. We use $\boldsymbol{\beta}_t$ to denote the vector form. Then $d\beta_{C,t}$ and $d\beta_{S,t}$ captures the demand shocks that originate from the noise traders. $(\beta_{C,t}, \beta_{S,t}, V_t)^\top$ follows an Ornstein–Uhlenbeck process with mean reversion coefficient $\boldsymbol{\eta} = \text{diag}(\eta_{\beta_C}, \eta_{\beta_S}, \eta_v)$ and potentially correlated shocks.

$$\begin{pmatrix} d\beta_{C,t} \\ d\beta_{S,t} \\ dV_t \end{pmatrix} = \boldsymbol{\eta} \begin{pmatrix} \bar{\beta_C} - \beta_{C,t} \\ \bar{\beta_S} - \beta_{S,t} \\ \bar{V} - V_t \end{pmatrix} dt + \hat{\boldsymbol{\Sigma}}^{\frac{1}{2}} d\mathbf{B}_t \tag{6}$$

where

$$\hat{\boldsymbol{\Sigma}} = \begin{pmatrix} \sigma_C^2 & \sigma_{CS} & \sigma_{CV} \\ \sigma_{CS} & \sigma_S^2 & \sigma_{SV} \\ \sigma_{CV} & \sigma_{SV} & \sigma_V^2 \end{pmatrix}. \tag{7}$$

Our general framework allows for flexible mean reversion. However, to have the sharpest characterization of the price dynamics, in the following analysis we consider the asymptotic case where the mean reversion $\boldsymbol{\eta}$ goes to $\mathbf{0}$ so that the noise trader demand and the fundamental value behave like martingales.

Finally, we assume the asset supply is fixed at $\bar{S}_i$ for $i \in \{C, S\}$. The market clearing condition implies

$$x_{i,t} + y_{i,t} + \beta_{i,t} = \bar{S}_i, \quad i \in \{C, S\}. \tag{8}$$

Due to the quadratic costs in the intermediaries' optimization problem, the equilibrium price is linear in the state variables. We characterize the equilibrium in closed form in the next section.

2.2 Equilibrium Analysis

We characterize the equilibrium price and quantity dynamics in closed form in this section. The proof is in Appendix A. Similar to the Vayanos-Vila framework, we conjecture the equilibrium price of the two assets can be expressed as a linear function of the state variables,

i.e.,

$$\mathbf{P}_t = \underbrace{\mathbf{p}}_{2\times 5}\mathbf{s}_t + \underbrace{\bar{\mathbf{P}}}_{2\times 1} \tag{9}$$

where the state variables are the position of intermediaries in each market, noise trader position in each market, and the fundamental value of the asset, i.e., $\mathbf{s}_t = (x_{C,t}, x_{S,t}, \beta_{C,t}, \beta_{S,t}, V_t)^\top$. Furthermore, we denote $\mathbf{p} = (\boldsymbol{\lambda}_x, \boldsymbol{\lambda}_\beta, \boldsymbol{\lambda}_V)$ where $\boldsymbol{\lambda}_x$ (2×2), $\boldsymbol{\lambda}_\beta$ (2×2) and $\boldsymbol{\lambda}_V$ (2×1) are the price loadings on the state variables respectively.

Let the state variable dynamics be

$$d\mathbf{s}_t = -\boldsymbol{\Gamma}(\mathbf{s}_t - \bar{\mathbf{s}})dt + \boldsymbol{\Sigma}^{\frac{1}{2}} d\mathbf{B}_t \tag{10}$$

where $\boldsymbol{\Sigma}$ is the instantaneous innovation covariance matrix, i.e., $\boldsymbol{\Sigma} = \mathrm{Cov}(d\mathbf{s}_t)/dt$. The explicit expression of $\boldsymbol{\Sigma}$ is given by (51) in Appendix A.

Define

$$\boldsymbol{\Lambda} \equiv k(\mathbf{I} - \boldsymbol{\zeta}\boldsymbol{\lambda}_x) \tag{11}$$

then the mean reversion matrix $\boldsymbol{\Gamma}$ is given by

$$\boldsymbol{\Gamma} = \begin{pmatrix} \boldsymbol{\Lambda} & k(\mathbf{I} - \boldsymbol{\zeta}\boldsymbol{\lambda}_\beta) - \boldsymbol{\eta}_\beta & -k\boldsymbol{\zeta}(\boldsymbol{\lambda}_V - \mathbf{1}) \\ \mathbf{0}_{2\times 2} & \boldsymbol{\eta}_\beta & \mathbf{0}_{2\times 1} \\ \mathbf{0}_{1\times 2} & \mathbf{0}_{1\times 2} & \eta_v \end{pmatrix} \tag{12}$$

Throughout the analysis, we consider the equilibrium where $\boldsymbol{\Lambda}$ has positive eigenvalues to ensure stability.

Proposition 1 characterizes the equilibrium price in closed form. The long-run price dynamics are governed by the demand of institutional investors: the steady-state price vector $\bar{\mathbf{P}}$ is determined by the natural demand of the institutional investors relative to the asset supply, scaled by the demand elasticities of the institutional investors. The price of each asset is increasing in the fundamental value V_t with a loading of 1. The price loading on the noise trader position $\boldsymbol{\lambda}_\beta$ is given by the inverse of the demand elasticity matrix $\boldsymbol{\zeta}$. This is because ultimately the institutional investors take the offsetting positions of the noise traders, and hence their demand elasticity determines the price loading, holding the intermediary sector's holdings constant.

The key object that determines the short-run price dynamics is the price loading on

the intermediary position $\boldsymbol{\lambda}_x$, which captures how much prices must adjust to compensate intermediaries for bearing extra inventory. It is pinned down by the cost matrix $\mathbf{C}$ through a matrix quadratic equation.

Proposition 1. *The equilibrium price is given by*

$$\mathbf{P}_t = \underbrace{\bar{\mathbf{P}}}_{\text{Steady state}} + \mathbf{1} \underbrace{V_t}_{\text{Fundamentals}} + \boldsymbol{\lambda}_x \underbrace{\begin{pmatrix} x_{C,t} \\ x_{S,t} \end{pmatrix}}_{\text{Intermediary holdings}} + \boldsymbol{\lambda}_\beta \underbrace{\begin{pmatrix} \beta_{C,t} \\ \beta_{S,t} \end{pmatrix}}_{\text{Noise trader holdings}} \tag{13}$$

where $\boldsymbol{\lambda}_x$ is the solution to the following matrix quadratic equation:

$$-\boldsymbol{\lambda}_x \boldsymbol{\Lambda} = \mathbf{C} \equiv \begin{pmatrix} \psi_C & 0 \\ 0 & \psi_S \end{pmatrix} + \gamma \begin{pmatrix} \boldsymbol{\lambda}_x & \boldsymbol{\zeta}^{-1} & \mathbf{1} \end{pmatrix} \boldsymbol{\Sigma} \begin{pmatrix} \boldsymbol{\lambda}_x \\ \boldsymbol{\zeta}^{-1} \\ \mathbf{1} \end{pmatrix} \tag{14}$$

$\boldsymbol{\lambda}_\beta$ and $\bar{\mathbf{P}}$ are given by

$$\boldsymbol{\lambda}_\beta = \boldsymbol{\zeta}^{-1} \tag{15}$$

$$\bar{\mathbf{P}} = \boldsymbol{\zeta}^{-1} \left(\boldsymbol{\theta} - \bar{\mathbf{S}} \right) \tag{16}$$

2.3 Differential Roles of Intermediary Costs

The two types of intermediary costs play distinct roles in shaping prices and spreads: risk-based costs integrate the two markets, while gross position costs segment them. To build this intuition, we first consider the case without position costs, i.e., $\psi_C = \psi_S = 0$. As Corollary 1 shows, the model then admits an equilibrium in which the two asset markets are perfectly integrated: prices are identical and the spread is always zero.

Corollary 1. *When $\psi_C = \psi_S = 0$, under the regularity condition (64) in Appendix A.2, there exists a solution in which the two asset markets are perfectly integrated and the equilibrium prices of the two assets are identical. The equilibrium price is given by*

$$\mathbf{P}_t = \left(\bar{P} + V_t + \lambda_x \mathbf{1}^\top \mathbf{x}_t + \lambda_\beta \mathbf{1}^\top \boldsymbol{\beta}_t \right) \mathbf{1}. \tag{17}$$

where λ_x is a scalar characterized by the quadratic equation (77) in Appendix A.2 and $\lambda_\beta = \frac{1}{\mathbf{1}^\top \boldsymbol{\zeta} \mathbf{1}}$.

Notice that this perfectly integrated equilibrium exists *despite* the imperfect substitutability of the two assets for institutional investors. Indeed, in this equilibrium, how institutional investors substitute across the two markets becomes irrelevant. Their elasticity matrix enters the equilibrium only through $\mathbf{1}^\top \boldsymbol{\zeta} \mathbf{1}$ — the aggregate demand elasticity across the two markets. To see why, note that without position costs, intermediaries can hold long-short positions indefinitely at no cost, as long as they are perfectly hedged.[7] They can therefore adjust the relative supply of the two assets to accommodate any institutional demand composition, leaving no effect on the prices.

We now turn to the case where both types of costs are present, and show they have distinct effects on prices and spreads. To obtain explicit analytical results, we focus on a symmetric special case with $\psi_C = \psi_S = \psi$, $\boldsymbol{\zeta}_{11} = \boldsymbol{\zeta}_{22}$, and mutually independent demand and fundamental shocks. When the two markets are symmetric, we can define

$$\begin{pmatrix} M_{l,\infty} & 0 \\ 0 & M_{s,\infty} \end{pmatrix} \equiv \frac{1}{2} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \boldsymbol{\zeta}^{-1} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}^\top \tag{18}$$

where $M_{l,\infty}$ and $M_{s,\infty}$ are the long-run price impacts for the average price level and the relative spread, respectively.

We define $P_{l,t} = \frac{P_{C,t}+P_{S,t}}{2}$ as the average price of the two markets, and $P_{s,t} = \frac{P_{S,t}-P_{C,t}}{2}$ as the (half) spread between the two markets.[8] Furthermore, define $V_{l,0}$ and $V_{s,0}$ as the instantaneous variances of $P_{l,t}$ and $P_{s,t}$, respectively.

$$V_{l,0} \equiv \mathrm{Var}(dP_{l,t}) = \frac{1}{4}\mathrm{Var}(dP_{C,t} + dP_{S,t}), \quad V_{s,0} \equiv \mathrm{Var}(dP_{s,t}) = \frac{1}{4}\mathrm{Var}(dP_{S,t} - dP_{C,t}) \tag{19}$$

We derive the following comparative statics results in Proposition 2. Both risk-based cost and gross position cost increase the variance of the average price and the variance of the spread. Intuitively, higher intermediation costs make it more costly for intermediaries to absorb shocks, hence they demand higher compensation and larger price fluctuations.

Proposition 2. *Under the symmetric model, the instantaneous variance of the average price $V_{l,0}$ and the instantaneous variance of the spread $V_{s,0}$ are increasing in both the risk-aversion*

[7] Formally, we show in Appendix A.2 that in this equilibrium intermediaries can permanently absorb offsetting demand shocks, i.e., holding $\beta_{C,t} + \beta_{S,t}$ fixed, $\frac{\partial \mathbb{E}_t[x_{C,t+\tau} - x_{S,t+\tau}]}{\partial(\beta_{C,t} - \beta_{S,t})} = -1$ for all $\tau \geq 0$.

[8] For the theoretical derivation it is often more convenient to work with the half-spread. For the ease of exposition we simply refer to it as the spread when the difference is immaterial.

parameter γ and the position cost ψ, i.e.,

$$\frac{\partial V_{l,0}}{\partial \gamma} > 0, \quad \frac{\partial V_{l,0}}{\partial \psi} > 0, \quad \frac{\partial V_{s,0}}{\partial \gamma} > 0, \quad \frac{\partial V_{s,0}}{\partial \psi} > 0 \tag{20}$$

However, even though both types of costs increase the variance of price and spread, the relative magnitude is different. In Proposition 3, we show that the gross position cost increases the variance of the spread more relative to the variance of the price level, while the risk-based cost matters more for the variance of the price level than that of the spread.

Proposition 3. *Under the symmetric model, the ratio between the instantaneous variance of the spread and the instantaneous variance of the average price, i.e., $V_{s,0}/V_{l,0}$, is increasing in the position cost ψ, i.e.,*

$$\frac{\partial (V_{s,0}/V_{l,0})}{\partial \psi} > 0 \tag{21}$$

Furthermore, when $\frac{2M_{s,\infty}}{1-\gamma\sigma_\beta^2 M_{s,\infty}/k} < \frac{M_{l,\infty}}{1-\gamma\sigma_\beta^2 M_{l,\infty}/k}$, the ratio of the instantaneous variance of the spread and the instantaneous variance of the average price is decreasing in the risk-aversion parameter γ, i.e.,

$$\frac{\partial (V_{s,0}/V_{l,0})}{\partial \gamma} < 0 \tag{22}$$

Intuitively, to arbitrage the spread away between the two assets, the intermediaries need to take a hedged position $x_{C,t} - x_{S,t}$. This position nets out most of the risk from price movements but still requires the intermediaries to hold a large gross position. Hence the position cost has a large impact on this trading strategy, and therefore affects the spread directly. In contrast, to arbitrage the price level (when it deviates from the fundamental value), the intermediaries need to take a net long position in the two assets, i.e., a market position $(x_{C,t} + x_{S,t})$, and bear risks from their price movements. As a result, such trading strategy is more affected by the risk-based costs, and hence the risk aversion coefficient has a larger impact on the level of the prices.

An equivalent way to state the above result is through the instantaneous return correlation between the two assets, since the variance ratio maps one-to-one into the correlation.

Corollary 2. *The instantaneous return correlation $\rho \equiv \mathrm{Corr}(dP_{C,t}, dP_{S,t})$ is decreasing in the position cost ψ and, under the same sufficient condition as in Proposition 3, increasing*

in the risk-aversion parameter γ:

$$\frac{\partial \rho}{\partial \psi} < 0, \qquad \frac{\partial \rho}{\partial \gamma} > 0 \tag{23}$$

Intuitively, risk-based costs *integrate* the two markets: they encourage arbitrageurs to take offsetting positions that hedge risk, linking the two markets through active arbitrage and raising return correlation. Gross position costs, by contrast, *segment* the markets by penalizing the gross positions required to do so, driving prices apart.

These results highlight the importance of distinguishing between the two types of intermediary costs. While gross position costs have a substantial impact on the spread, risk-based costs are more influential for the price level. As risk-based costs rise, the wedge between spread variance and price-level variance widens, implying that the spread can miss much of the movements in the price level. By contrast, when position costs dominate, spread variance tracks price-level variance more closely, making the spread a more informative proxy for price-level effects. To separately identify the two cost components, we next characterize the dynamics of prices and spreads following demand shocks. As we show, the speed at which prices revert is pinned down by intermediary costs alone, providing the basis for our identification strategy.

2.4 Price Dynamics

Proposition 4 characterizes the price and position dynamics following a demand shock, which are determined by the eigenvalues of the matrix $\mathbf{\Lambda}$, as defined in (11). To ensure stability, we need $\mathbf{\Lambda}$ to have positive eigenvalues. The magnitudes of the eigenvalues control the speed of convergence to long-run behavior.

Proposition 4. *The price and position dynamics following a demand shock are given by*

$$\frac{\partial \mathbb{E}_t[\mathbf{x}_{t+\tau}]}{\partial \boldsymbol{\beta}_t^\top} = -e^{-\mathbf{\Lambda}\tau} \tag{24}$$

$$\mathbf{M}_\tau \equiv \frac{\partial \mathbb{E}_t[\mathbf{P}_{t+\tau}]}{\partial \boldsymbol{\beta}_t^\top} = -\boldsymbol{\lambda}_x e^{-\mathbf{\Lambda}\tau} + \boldsymbol{\zeta}^{-1} \tag{25}$$

where

$$e^{-\mathbf{\Lambda}\tau} = \frac{e^{-\nu_1\tau}}{\nu_2 - \nu_1}(\nu_2 \mathbf{I} - \mathbf{\Lambda}) + \frac{e^{-\nu_2\tau}}{\nu_1 - \nu_2}(\nu_1 \mathbf{I} - \mathbf{\Lambda}), \tag{26}$$

and ν_1 and ν_2 are the two eigenvalues of $\mathbf{\Lambda}$.

Note that, on impact, the intermediary absorbs the entire shock, hence the intermediary position immediately after the shock is given by

$$\frac{\partial \mathbb{E}_t[\mathbf{x}_{t+\tau}]}{\partial \boldsymbol{\beta}_t^\top}\Big|_{\tau=0} = -\mathbf{I} \tag{27}$$

In the long run, the intermediary offloads the entire demand shock to the institutional investors, and the intermediary position reverts back to its steady state

$$\lim_{\tau\to\infty} \frac{\partial \mathbb{E}_t[\mathbf{x}_{t+\tau}]}{\partial \boldsymbol{\beta}_t^\top} = \mathbf{0} \tag{28}$$

Figure 1 provides a numerical illustration of the price and position dynamics following a unit demand shock in the cash market ($\Delta\beta_{C,t} = 1$). Panel (a) shows the cash market quantity dynamics: on impact, the arbitrageur absorbs the entire shock ($\Delta x_C = -1$), while institutional investors do not respond instantaneously. Over time, as prices adjust, institutional investors gradually increase their cash holdings and the arbitrageur unwinds its position, both converging to their long-run values. Panel (b) shows the cross-market spillover to the synthetic asset: even though the shock originates in the cash market, the arbitrageur takes an offsetting position in the synthetic market, which is then gradually absorbed by institutional investors.

In terms of price dynamics, in the long run, the price response is only determined by the demand elasticity of the institutional investors, reflecting the fact that intermediaries are short-term investors in the market. That is,

$$\mathbf{M}_\infty \equiv \lim_{\tau\to\infty} \mathbf{M}_\tau = \boldsymbol{\zeta}^{-1}. \tag{29}$$

On impact, the price impact is

$$\mathbf{M}_0 \equiv \mathbf{M}_\tau\big|_{\tau=0} = -\boldsymbol{\lambda}_x + \boldsymbol{\zeta}^{-1}. \tag{30}$$

That said, along the transition path, the price response does depend on the intermediary position. In fact, we can rewrite the price impact as

$$\mathbf{M}_\tau = \overbrace{\underbrace{\boldsymbol{\zeta}^{-1}}_{\text{Institution elasticity}}}^{\text{Long-run impact}} + \overbrace{\underbrace{\partial \mathbf{x}_{t+\tau}/\partial \boldsymbol{\beta}_t^\top}_{\text{Intermediation share}} \times \underbrace{\boldsymbol{\lambda}_x}_{\text{Intermediary required comp.}}}^{\text{Transitory impact driven inventory dynamics}} \tag{31}$$

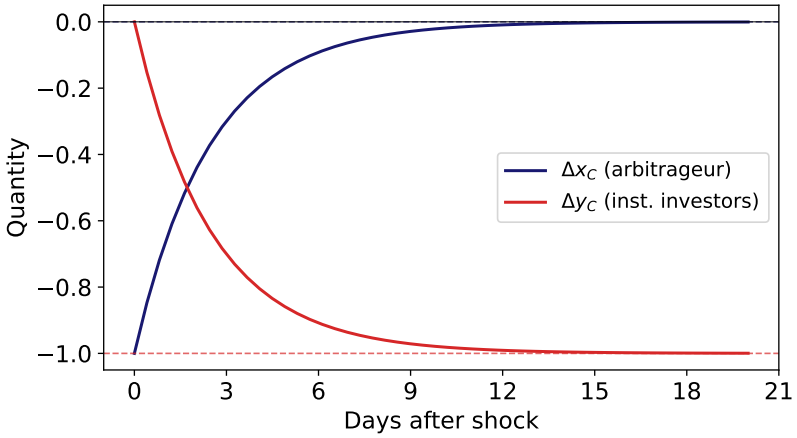
The price response reflects both a long-run component, which is determined by the de-

mand elasticity of the preferred-habitat investors, and a transitory component, which is driven by the inventory dynamics of the intermediary sector. The transitory component captures the fact that intermediaries are short-term investors in the market, and they require compensation for bearing inventory in the short run.

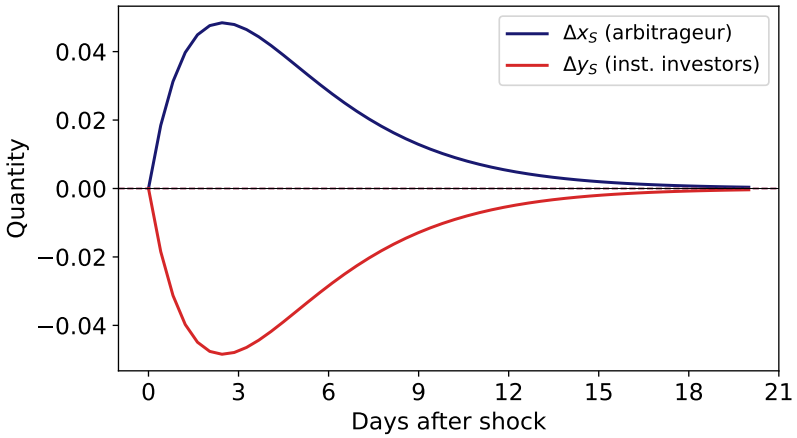
Panel (c) of Figure 1 illustrates the price responses in the numerical example. Both the cash price and the synthetic price jump on impact and subsequently decay toward their long-run responses, governed by $\mathbf{M}_\infty = \boldsymbol{\zeta}^{-1}$. The spread $P_{s,t} = (P_{S,t} - P_{C,t})/2$ moves away from zero on impact and converges to its long-run value, governed by $-M_{s,\infty}/2$.

The decomposition in equation (31) shows that the price impact at any horizon reflects both institutional demand elasticities and intermediary costs. However, the rate at which prices revert has a sharper interpretation: it equals the intermediaries' expected return on their inventory, which by their first-order condition (5) is pinned down by the marginal cost of intermediation. We formalize this identification strategy in the next section.

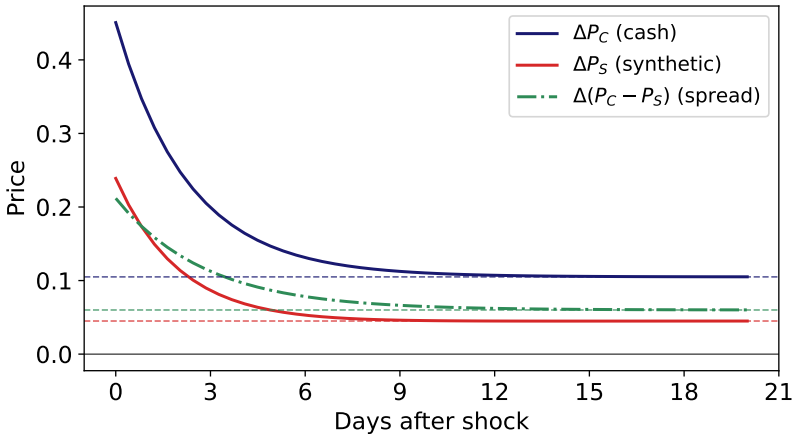
Figure 1: Impulse responses to a unit cash demand shock



(a) Cash market quantities



(b) Synthetic market quantities



(c) Price impulse responses

This figure illustrates the price and quantity dynamics in response to a demand shock $\Delta\beta_C = 1$ and $\Delta\beta_S = 0$. Dashed lines indicate long-run values. Parameter values: $k = 0.1$, $\gamma = 2$, $\psi = 0.05$, $\sigma_\beta = 0.2$, $\sigma_V = 0.2$, $M_{s,\infty} = 0.06$, $M_{l,\infty} = 0.15$.

3 Identification and Estimation

We develop a new method for identifying intermediary costs from observed price dynamics following demand shocks. We show that the intermediaries' first-order condition links the initial rate of price reversion to marginal intermediation costs, providing a direct way to identify risk-based and gross position costs from observed price dynamics. We formalize this identification strategy below and then take it to the U.S. Treasury cash and OIS markets, using high-frequency demand shocks from Treasury auction result releases.

3.1 Identification from the Slope of Price Impact

Building on the price impact decomposition in (31), we show that the cost parameters can be identified from the slope of the price impact curve following a demand shock.

Formally, the rate of price reversion is given by

$$-\frac{\partial \mathbf{M}_\tau}{\partial \tau} \;=\; -\mathbb{E}\left[\frac{\partial}{\partial \tau}\left(\frac{\partial \mathbf{P}_{t+\tau}}{\partial \boldsymbol{\beta}_t^\top}\right)\right] = \underbrace{-\partial \mathbf{x}_{t+\tau}/\partial \boldsymbol{\beta}_t^\top}_{\text{Intermediation share}} \times \underbrace{\mathbf{C}}_{\text{marginal required comp.}} \tag{32}$$

and is determined by the intermediary's cost parameters alone. Intuitively, the rate of price reversion is the return that intermediaries get from bearing the inventory at a given instant, which should equal their inventory share times the marginal cost of bearing the inventory. The logic of this identification strategy follows directly from the intermediary's first order condition.

With detailed quantity data on intermediary holdings, we can observe $-\partial \mathbf{x}_{t+\tau}/\partial \boldsymbol{\beta}_t^\top$ along the whole transition path and back out the cost parameters from the changes in the price impact of the cash and synthetic assets. Absent such quantity data, we can still identify the cost parameters by leveraging the fact that the institutional investors have slow moving capital and the entire demand shock is absorbed by the intermediary on impact, i.e.,

$$-\frac{\partial \mathbf{M}_\tau}{\partial \tau}\Big|_{\tau=0} = -\mathbb{E}\left[\frac{\partial}{\partial \tau}\left(\frac{\partial \mathbf{P}_{t+\tau}}{\partial \boldsymbol{\beta}_t^\top}\right)\right]\Big|_{\tau=0} = \underbrace{-\partial \mathbf{x}_{t+\tau}/\partial \boldsymbol{\beta}_t^\top\Big|_{\tau=0}}_{\text{Intermediation share=100\%}} \times \underbrace{\mathbf{C}}_{\text{marginal required comp.}} = \mathbf{C} \tag{33}$$

This means that even without detailed holdings data, we can still identify the cost parameters from the initial rate at which prices revert.

Figure 2 illustrates the intuition behind this identification strategy. The average price impact of a demand shock decomposes into two regions. The long-run level reflects insti-

tutional investors' demand elasticity, since they eventually absorb the entire demand shock. The additional transitory component above the long-run level reflects intermediary compensation: intermediaries absorb the shock on impact and require return compensation to hold the extra inventory while institutional investors gradually step in. The price impact therefore starts high on impact and decays toward the long-run level as intermediaries offload their positions. At each instant, the rate of price reversion equals the intermediaries' expected return on their inventory—their marginal revenue from providing liquidity. Under competitive intermediation, marginal revenue equals marginal cost in equilibrium, so the reversion rate equals the intermediation share times the marginal cost of intermediation $\mathbf{C}$. On impact at $\tau = 0$, the intermediation share is 100%, so the initial reversion rate directly reveals the marginal cost of intermediation.

This gives us a direct way to identify the intermediary cost matrix $\mathbf{C}$ without estimating the full economy. The identification comes directly from the intermediary's first order condition and the main assumption we need is that the entire demand shock is absorbed by the intermediary on impact.[9] This identification approach can be applied to a wide range of models where the intermediary sector is the supplier of liquidity. By construction, the approach identifies the cost faced by fast-moving liquidity providers. In the Treasury and OIS markets, the natural candidates for this role are dealers and hedge funds.

In our specific model, the intermediary cost matrix $\mathbf{C}$ has a particular structure, i.e.,

$$\mathbf{C} = \begin{pmatrix} \psi_C + \gamma\sigma_C^2 & \gamma\sigma_{CS} \\ \gamma\sigma_{CS} & \psi_S + \gamma\sigma_S^2 \end{pmatrix} \tag{34}$$

where σ_C, σ_S, and σ_{CS} are all empirical observables. This allows us to back out the gross position costs ψ_i and the risk-aversion parameter γ from the reversion rate of the price and spread response to demand shocks. Similar to the intuition on relative volatility ratio, the hedged position reflects the gross position cost more than risk-based costs, hence the reversion rate of the spread in response to the demand shock is more informative about ψ_i. On the other hand, the market position is more informative about the risk-based cost and hence the level of price response is more informative about γ. Combining the two allows us to identify the two types of costs separately.

[9] Note that we do not need to assume symmetry of the Treasury and the OIS market for the identification to hold.

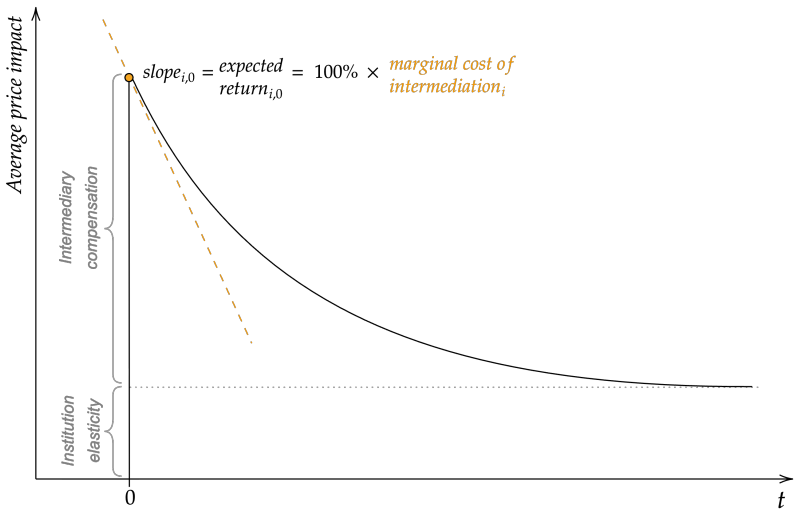


Figure 2: Identification from the initial decay rate. The figure illustrates the price impact dynamics following a positive permanent demand shock. Initially, the entire shock is accommodated by intermediaries—who sell (short) Treasuries to the noise traders—so the intermediation share is 100%. The initial reversion rate of the price impact curve therefore directly reveals the marginal cost matrix $\mathbf{C}$.

3.2 Treasury Yields and OIS Rates

We measure the price level response using the 5-year constant-maturity Treasury yield, sourced from the U.S. Department of the Treasury's daily yield curve data. For the synthetic price level, we use the 5-year overnight index swap (OIS) rate from Bloomberg, where the underlying reference rate is the effective federal funds rate. The Treasury-OIS spread—the difference between the Treasury yield and the OIS rate—serves as our measure of the no-arbitrage spread.[10] The OIS rate provides a particularly clean synthetic price level benchmark because OIS contracts are derivatives settled in cash and do not require dealers to hold physical inventory on their balance sheets.

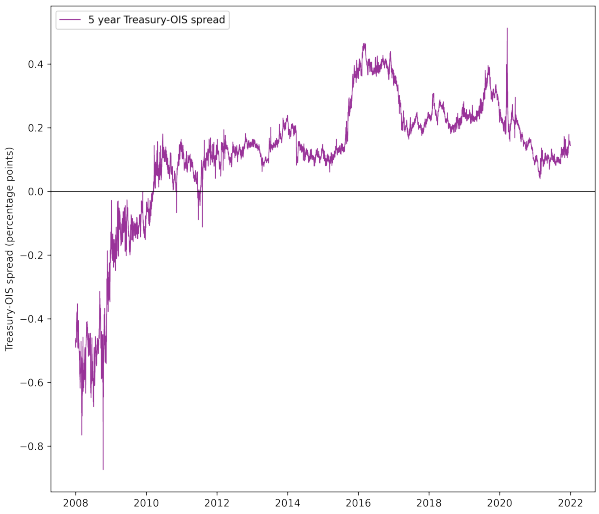
[10]The Treasury-OIS spread partly reflects the "inconvenience yield" of intermediating Treasuries: He, Nagel, and Song (2022) show that this spread captures the balance sheet costs borne by dealer-intermediaries when absorbing demand and supply shocks in the Treasury market.

Figure 3: 5-year Treasury yield and Treasury-OIS spread

(a) Treasury yield and OIS rate



(b) Treasury–OIS spread



The figure plots the 5-year constant-maturity Treasury yield, 5-year OIS rate, and the 5-year Treasury-OIS spread over the sample period. Panel (a) shows the Treasury yield and the OIS rate. Panel (b) shows the Treasury-OIS spread. The Treasury yield is from the U.S. Department of the Treasury's daily yield curve data. The OIS rate underlying the spread is the 5-year overnight index swap rate from Bloomberg, referenced to the effective federal funds rate.

Table 1: Summary statistics: Treasury yield and OIS rate

	Mean	Std. Dev.	Min	Median	Max	N
Treasury yield	1.66	0.74	0.20	1.64	3.73	3,621
OIS rate	1.54	0.84	0.04	1.47	4.33	3,621
Treasury-OIS spread	0.11	0.21	-0.87	0.14	0.51	3,621

Summary statistics for the 5-year constant-maturity Treasury yield, the 5-year OIS rate, and the Treasury-OIS spread. The sample spans January 2008 to January 2022. Treasury yield and OIS rate are in percent; the Treasury-OIS spread is in percentage points. The correlation between daily changes in the Treasury yield and the OIS rate is 0.92.

Figure 3 plots the two series and the resulting spread over the sample period, and Table 1 reports summary statistics. The 5-year Treasury yield averages 1.66% over the sample, closely tracking the OIS rate (mean 1.54%), with a correlation of 0.92 between their daily

changes. The Treasury-OIS spread averages 11 basis points but exhibits substantial variation, ranging from −87 to 51 basis points, with notable spikes during the Global Financial Crisis and the COVID-19 episode visible in Panel (b).

3.3 Demand Shock Construction

We identify demand shocks to the Treasury market using the high-frequency price impact of Treasury auction result releases, following Ray, Droste, and Gorodnichenko (2024). In U.S. Treasury auctions, the total quantity to be issued is announced days in advance, but the composition of investor demand is not revealed until the auction closes and results are released. The yield change in a narrow window around the release of auction results therefore reflects unexpected shifts in investor demand, providing a measure of exogenous demand shocks to the Treasury market.

Figure 4 illustrates the construction. Let $y_{t,\text{pre}}$ denote the yield of the auctioned security measured at 12:50pm, approximately 10 minutes before the competitive bidding closes, and $y_{t,\text{post}}$ the yield at 1:10pm, after the results have been released. For reopenings, yields are measured on the outstanding (on-the-run) security; for new issues, yields are measured on the when-issued forward contract.[11] The demand shock on auction date t is defined as

$$\beta_{\text{Treasury},t} \propto u_{\text{Treasury},t} = y_{t,\text{post}} - y_{t,\text{pre}}. \tag{35}$$

The intraday yield data are constructed from tick-level quotes sourced from GovPX, which provides real-time prices for on-the-run and when-issued Treasury securities. On non-auction days, the demand shock is set to zero.

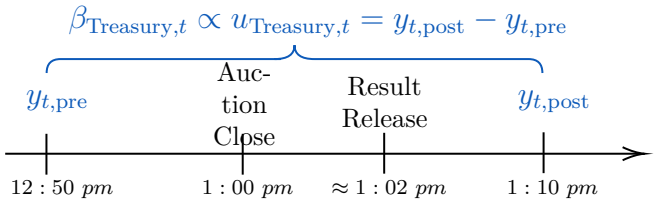
An important caveat is that we do not directly observe the underlying demand shock $\beta_{\text{Treasury},t}$, but rather its price impact $u_{\text{Treasury},t} = y_{t,\text{post}} - y_{t,\text{pre}}$. As we discuss below, assuming the price impact is proportional to the demand shock, we can use it to identify the relative importance of risk-based versus gross position costs.

Conversely, weaker-than-expected demand raises yields. Figure 5 plots the demand shocks over the sample period, and Table 2 summarizes their distribution. Of the 168 five-year Treasury note auctions in our sample, 165 have GovPX intraday price data available in the event window.[12] These produce shocks that are approximately mean-zero (0.17 bps) with a standard deviation of 1.77 basis points, ranging from −5.55 to 8.50 basis points. The

[11] When-issued contracts are forward contracts that settle on the issue date, providing a pre-issuance price for securities that do not yet trade in the secondary market.

[12] Three auction dates lack GovPX trades in the narrow pre- or post-window: February 28, 2008; August 27, 2014; and September 25, 2018.

Figure 4: Timeline of Treasury auction shock construction

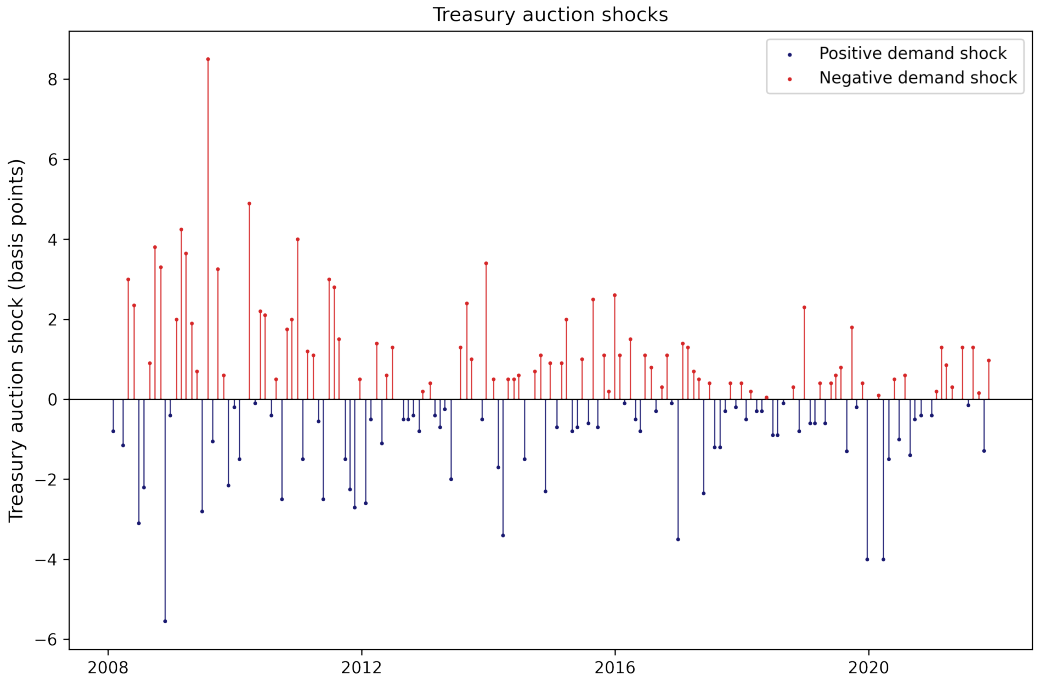


The figure illustrates the construction of Treasury auction demand shocks following Ray, Droste, and Gorodnichenko (2024). The pre-auction yield $y_{t,\text{pre}}$ is measured at 12:50pm, before the competitive bidding closes at 1:00pm. The post-auction yield $y_{t,\text{post}}$ is measured at 1:10pm, after the results are released at approximately 1:02pm. The demand shock is $\beta_{\text{Treasury},t} \propto u_{\text{Treasury},t} = y_{t,\text{post}} - y_{t,\text{pre}}$.

largest shocks are during the Global Financial Crisis and the Covid period, consistent with heightened uncertainty during these episodes. Note that a positive demand shock—more demand for Treasuries than expected—pushes bond prices up and hence yields down, so that $u_{\text{Treasury},t} < 0$.[13]

[13]Somogyi, Wallen, and Xu, 2025 estimate Treasury demand elasticity from auction data and find that it has become less elastic after 2010, though it remains relatively stable during our sample period.

Figure 5: Treasury auction demand shocks



The figure plots Treasury auction demand shocks for 5-year Treasury note auctions over the sample period (2008–2022). Each point represents the yield change $u_{\text{Treasury},t} = y_{t,\text{post}} - y_{t,\text{pre}}$ around a single auction result release, measured in basis points.

Table 2: Summary statistics: Treasury auction demand shocks

	Mean	Std. Dev.	Min	Median	Max	N
5Y auction shock (bps)	0.17	1.77	-5.55	0.05	8.50	165

Summary statistics for 5-year Treasury note auction demand shocks. The shock is the yield change from 12:50pm to 1:10pm around the release of auction results, measured in basis points. The sample spans January 2008 to January 2022.

3.3.1 Demand Shock and Identification

To identify intermediary costs, we need an exogenous shock to intermediaries' holdings in one market. As intermediaries adjust their positions in response, the subsequent price dynamics across the two markets reveal the magnitudes of the two types of costs. Our demand-

shock measure meets this requirement as long as yield changes around auction-result releases reflect shifts in investor demand for Treasury securities that are both: (i) exogenous to macroeconomic fundamentals, and (ii) specific to the Treasury market rather than the OIS market. The key distinction between fundamental shocks and demand shocks is that prices in both markets load directly on fundamental shocks, whereas demand shocks affect the other market only through intermediaries' arbitrage activity. The extent of the spillover therefore reflects intermediary constraints. For the same reason, correlated demand shocks pose an identification challenge, since they move both markets simultaneously.

The high-frequency nature of our identification strategy makes it particularly plausible that our Treasury market demand shocks are orthogonal to both the fundamental shocks as well as demand shocks in the OIS market. First, the revelation of fundamental information is unlikely to contaminate our shock measure. The narrow window of a few minutes around the auction result release rules out the arrival of macroeconomic news during the event window. A subtler issue is that auction participants may possess private information about fundamentals, so that their demand reveals fundamentals rather than reflecting an exogenous shift. This channel also appears unlikely: Ray, Droste, and Gorodnichenko (2024) show that surprise movements in auction demand are driven by relatively unsophisticated institutional investors—foreign monetary authorities, investment funds, and insurance companies—rather than hedge funds or broker-dealers who are more likely to possess private information.[14]

Second, the high-frequency window implies that the demand shocks we capture originate from investors who face frictions in accessing the secondary Treasury market or OIS markets. If these investors could trade freely, they would have already expressed their demand in the secondary market or through OIS positions outside the auction window. The fact that their demand is revealed only at auction indicates they are restricted to the primary market, making it unlikely that they simultaneously generate demand shocks in OIS.

By definition, the new information revealed in the tight window around the auction result release must be revealed through current demand. Consistent with our model, demand shocks can be persistent and current demand can contain information about future demand. As we discuss below, the exact magnitude of the total demand shock is irrelevant for calculating the relative importance of risk-based versus gross position costs, as long as the price impact is proportional to the underlying demand shock.

[14]Table 3 confirms that our results are robust to excluding the largest demand shocks and crisis periods, where information revelation is most plausible.

3.4 Mapping Yield and Spread Decay Rates to Intermediation Costs

We can test our model predictions using any two out of the three series—Treasury yields, OIS rates, or the Treasury-OIS spread—since any two can be used to construct the third. We focus on the yield level and the Treasury-OIS spread. Two considerations motivate this choice. First, the academic literature and policy makers frequently focus on the Treasury-OIS spread as an indicator of intermediation capacity in the Treasury market (e.g., He, Nagel, and Song, 2022), facilitating comparison with existing work. Second, our theory suggests that spreads provide a more direct measure of capacity costs, helping with statistical inference.

As derived in Section 3.1, intermediaries' first-order condition implies that the rate at which expected price levels revert following a demand shock is proportional to the cost matrix $\mathbf{C}$. As we only have demand shocks to the Treasury market ($\beta_{C,t} \equiv \boldsymbol{\beta}\mathbf{e}_1$), our identification strategy identifies the first column of the cost matrix $\mathbf{C}$, and hence the gross position cost ψ_C associated with holding cash Treasuries. We restate the first-order condition in terms of yields and spreads below.

Let $y_{C,t}$ denote the Treasury yield and $y_{S,t}$ the OIS rate. Denote the *instantaneous decay rates* of the yield level and the Treasury-OIS spread as

$$m_l := \frac{\partial}{\partial \tau} \frac{\partial \mathbb{E}_t[y_{C,t+\tau}]}{\partial \beta_{C,t}}\bigg|_{\tau=0}, \quad m_s := \frac{\partial}{\partial \tau} \frac{\partial \mathbb{E}_t[y_{C,t+\tau} - y_{S,t+\tau}]}{\partial \beta_{C,t}}\bigg|_{\tau=0}, \tag{36}$$

where $\beta_{C,t}$ is the demand shock to the Treasury market. Intuitively, m_l measures how fast the yield impact reverts (a more negative m_l indicates faster mean reversion), while m_s measures the same for the no-arbitrage spread. [15]

From the first-order condition, these observable decay rates satisfy

$$\begin{pmatrix} m_l \\ m_s \end{pmatrix} = -\tilde{\lambda} \begin{pmatrix} \psi_C + \gamma\sigma_C^2 \\ \psi_C + \gamma\sigma_{C,s} \end{pmatrix}, \tag{37}$$

where $\sigma_C^2 := \mathrm{Var}(dy_{C,t})/dt$ is the instantaneous variance of Treasury yields, $\sigma_{C,s} := \mathrm{Cov}(dy_{C,t}, d(y_{C,t} - y_{S,t}))/dt$ is the covariance of yields and spreads, and $\tilde{\lambda} > 0$ is a positive scaling constant reflecting the fact that we do not observe the size of the demand shocks, only their price impact, which we assume to be proportional.[16]

[15] These are the yield-space analogues of the model derivatives $\partial_\tau M_{l,\tau}|_{\tau=0}$ and $\partial_\tau M_{s,\tau}|_{\tau=0}$. Up to first order, $y_{C,t} - y_{S,t} \approx 2P_{s,t}/(DP)$, so the mapping preserves the sign and differs only by a scale factor.

[16] In addition, to map from prices to yields we use a first order approximation around the bonds steady

Once these decay rates are estimated, we can quantify the relative importance of the risk and gross position costs. The two-equation system (37) can be inverted to express the structural parameters (up to the common scale $\tilde{\lambda}$) as functions of the observables m_l, m_s, and the yield covariance structure. Hence we can construct summary statistics that measure the fraction of marginal intermediation cost attributable to risk-based costs. The risk contribution to the intermediation costs associated with arbitraging the Treasury yield and the Treasury-OIS spread are,

$$\hat{C}_{r,l} \equiv \frac{\gamma\sigma_C^2}{\psi_C + \gamma\sigma_C^2} = \frac{\sigma_C^2}{\sigma_C^2 - \sigma_{C,s}}\left(1 - \frac{m_s}{m_l}\right), \tag{38}$$

$$\hat{C}_{r,s} \equiv \frac{\gamma\sigma_{C,s}}{\psi_C + \gamma\sigma_{C,s}} = \frac{\sigma_{C,s}}{\sigma_C^2 - \sigma_{C,s}}\left(\frac{m_l}{m_s} - 1\right). \tag{39}$$

The contribution of the gross position cost is one minus the contribution of the risk-based cost.[17]

To build intuition, consider the following two cases. If Treasury yields and OIS rates move in lockstep after a demand shock—so that the spread does not respond ($m_s = 0$)—then the entire yield reversion is attributable to risk costs ($\hat{C}_{r,l} = 100\%$). This is because arbitraging the spread is nearly risk-free, so if the spread shows no movement, it must be that risk costs, not gross position costs, are dominating intermediation. Conversely, if the spread also reverts ($m_s \neq 0$), then gross position costs contribute to the yield response, and $\hat{C}_{r,l} < 100\%$. Hence, testing whether $m_s = 0$ is a sufficient condition to see if gross position costs matter.

3.5 Estimation Procedure

Local projections. We estimate the impulse response of yields and spreads to Treasury auction demand shocks using the local projection framework of Jordà (2005). To address the noise inherent in local projection estimation, we work at weekly frequency, measuring the cumulative impact at $\tau = 0, 1, 2, \ldots, H$ weeks after the auction.[18][19] For each outcome variable $i \in \{l, s\}$ (yield level and Treasury-OIS spread, respectively) and weekly horizon τ,

state duration D and price level P which are also absorbed in $\tilde{\lambda}$.

[17] Because the risk-based cost contributions depend only on the ratio m_s/m_l and the yield covariance structure, they are identified without knowledge of the demand shock magnitude; the scale of the shock cancels in the ratio.

[18] We use multiples of five trading days to define a week in order to ensure the initial impact is always exactly 5 days after the shock, and the subsequent impacts are always a multiple of 5 days after the previous impact.

[19] In Appendix B.1 we show that the results are robust to estimating at daily frequency.

we estimate

$$\Delta_\tau y_{i,t} = a_{i,\tau} + \theta_{i,\tau}\, u_t + c_{i,\tau}\, \Delta_\tau y_{i,t-\tau-1} + \varepsilon_{i,t+\tau}, \tag{40}$$

where $\Delta_\tau y_{i,t} := y_{i,t+\tau} - y_{i,t-1}$ is the cumulative change from the day before auction day t to τ weeks later and u_t is the Treasury auction demand shock.[20] The coefficient $\theta_{i,\tau}$ traces out the impulse response function (IRF) at horizon τ.

Decay estimation. We parameterize the IRF as $\theta_i(\tau) = b_{i,0} + b_{i,1}\,\tau$ and estimate the coefficients jointly via the panel regression

$$\Delta_\tau y_{i,t} = a_{i,\tau} + (b_{i,0} + b_{i,1}\,\tau)\, u_t + c_{i,\tau}\, \Delta_\tau y_{i,t-\tau-1} + \varepsilon_{i,t+\tau}, \quad \tau = 0, \ldots, H;\ i = l, s. \tag{41}$$

The initial decay rate is then $\hat{m}_i = \hat{b}_{i,1}$. We estimate the level and spread equations jointly, which directly provides the covariance between $\hat{m}_l$ and $\hat{m}_s$ needed for inference on the risk contributions. Standard errors for the decay rates $\hat{m}_l$ and $\hat{m}_s$ are computed using date-clustered covariance matrices, which account for the correlation across horizons within each auction date.

Choosing a linear specification over higher-order polynomials reflects a natural bias-variance trade-off. Identifying the decay rate near $\tau = 0$ requires a short estimation window, leaving just one to three weekly horizons—too few for higher-order polynomials, which would overfit. The impulse responses in Figure 6 confirm that decay is approximately linear over this range, so the restriction costs little in practice.

Mapping decay rates to contribution. Contribution of risk to yields and spreads must lie between 0% and 100%. We impose these bounds when estimating the risk contributions using constrained optimization and use a date-cluster bootstrap for inference; the details are described in Appendix B.2.[21]

[20] The lagged dependent variable $\Delta_\tau y_{i,t-\tau-1}$ addresses autocorrelation in the errors, following Montiel Olea and Plagborg-Møller (2021).

[21] Overall, we find constraints have negligible impact on local projection fit and slope estimates, but mainly address the issue of near-zero decay rates inducing noise in contribution estimation. See Appendix B.2 for details.

4 Estimation Results

We present the main findings in this section. Section 4.1 reports baseline results for the full sample (2008–2022). Although yields move considerably in response to Treasury demand shocks, no-arbitrage spreads respond little, suggesting that risk-based costs are on average the dominant cost. Section 4.2 then examines two crisis episodes. In both, gross position costs appear to play a larger role than in normal times—consistent with the widening of no-arbitrage spreads during these periods. The magnitude differs across the two crises, however: during the GFC, risk-based costs remain dominant, while during COVID gross position costs account for a substantially larger share, consistent with post-Dodd-Frank Supplementary Leverage Ratio (SLR) constraints becoming binding.

4.1 Baseline Results

Figure 6 shows the impulse response functions to Treasury auction demand shocks for the full sample (2008–2022). Panel (a) shows that Treasury yields and the OIS rate move in lockstep: both initially increase and then slowly decay toward zero. This is consistent with slow-moving investors entering the market over time and hence increasing its absorptive capacity. In line with the lockstep movement, panel (b) shows that the Treasury-OIS spread does not respond to the shock. This pattern indicates that risk-based costs, rather than gross position costs, are the main driver of intermediation capacity.[22]

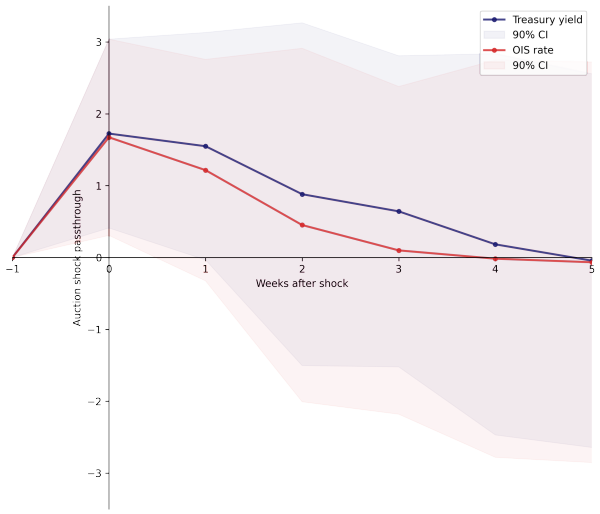
The intuition is as follows: arbitraging the Treasury-OIS spread is nearly risk-free, since going long one asset and short the other hedges away risk. Risk-based costs therefore do not inhibit this arbitrage trade. Gross position costs, by contrast, penalize long/short positions regardless of risk exposure. The absence of a spread response thus indicates active arbitrage and suggests that gross position costs play a limited role on average—that is, in normal times, which constitute most of the sample. We turn next to quantitatively confirming this interpretation and then show that it does not hold during periods of recent market stress.

Table 3 quantifies the risk contribution—the share of the yield's initial decay rate attributable to risk-based costs, as defined in Section 3.1—for the 5-year note auction demand shock. The gross position cost contribution is one minus the risk contribution. The baseline specification uses a linear decay function estimated over a three-week window ($\tau = 0$ to $\tau = 3$ weeks). The impulse responses in Figure 6 are approximately linear over this range,

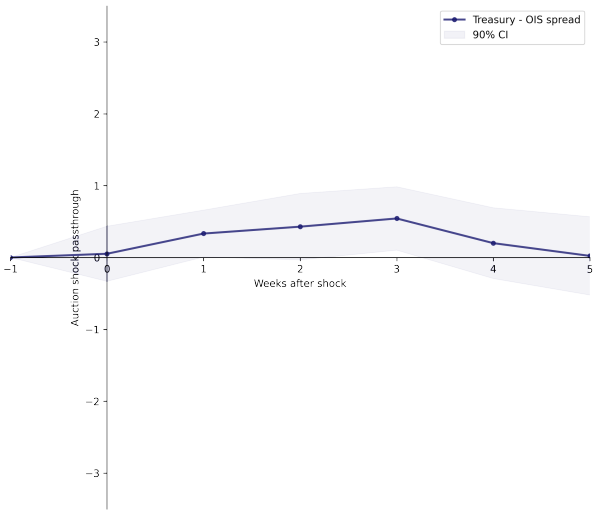
[22] In Appendix A.7, we show that if the intermediary face hard capacity constraint, under reasonable assumptions, our estimation approach underestimates the contribution of the risk-based costs. Similarly, hedging demand also implies an underestimation of the risk-based cost contribution. Hence, the finding that risk-based costs are dominant is robust to these potential biases.

Figure 6: Impulse response to Treasury auction demand shocks: Full sample (2008–2022)

(a) Treasury yield and OIS rate



(b) Treasury–OIS spread



The figures plot impulse response functions of 5-year Treasury yields, the OIS rate, and the Treasury–OIS spread to Treasury auction demand shocks over the full sample (2008–2022), estimated at weekly frequency. The demand shock is constructed from 5-year Treasury note auctions. Panel (a) shows the response of Treasury yields (blue) and the OIS rate (red), with 90% confidence bands. Panel (b) shows the response of the Treasury–OIS spread. The sample spans 165 auction dates.

making this a natural choice.[23]

Under this baseline, the 90% confidence interval for the risk contribution is 69% to 100%, meaning that risk-based costs account for the dominant share of intermediation costs. The point estimate is 100%. This is consistent with the visual evidence in Figure 6: the yield and OIS rate move in lockstep while the spread remains flat.

Table 3 confirms that the result is robust to alternative estimation choices. Shortening the estimation window to two weeks or one week yields confidence intervals of 44%–100% and 23%–100%, respectively. The intervals widen as expected—shorter windows reduce the effective number of observations—but the point estimate remains at 100% in every specification. Because the contribution is bounded between $[0, 100]$, the parameter estimates are highly non-normal; the confidence intervals reflect sampling noise that pulls the lower bound away from the boundary.[24]

[23] In line with Fleming, Nguyen, and Rosenberg (2024) who show that primary-dealer Treasuries positions, especially for coupon-paying bonds like 5-year notes, are slow to mean-revert. Positions persist for several weeks to months.

[24] Appendix B.3 shows the direction of the price response and its subsequent decay align with the sign of

A potential concern is that auction demand may reveal private information about fundamentals rather than reflect an exogenous demand shift. As discussed, this seems unlikely since the demand shock is predominantly driven by long-term investors rather than dealers and hedge funds, who are unlikely to possess private information (Ray, Droste, and Gorodnichenko, 2024). Nonetheless, private information revelation is most likely during episodes of extreme market stress or on dates with unusually large demand surprises. Table 3 examines this by dropping the ten auction dates with the largest demand shocks by absolute value and by excluding the GFC and COVID periods entirely. In both cases, risk-based costs remain dominant, suggesting our result is not driven by information revelation. The sixth row broadens the shock universe to 2-, 3-, 5-, and 7-year note auctions (`term_group=[2,7]`), holding the dependent variables fixed at the 5-year Treasury and 5-year Treasury–OIS spread. The point estimate of the risk contribution is unchanged at 100% (90% CI: [74, 100]) and is based on 614 auction-shock dates over 2008–2022, roughly four times the number of 5-year-only shocks. This addresses the concern that the 5-year-only specification may have insufficient effective observations.

the demand shock, ruling out resolution of uncertainty due to gradually learning results of the auction.

Table 3: Estimation results: Full sample (2008–2022)

	Risk Contribution	Window
Baseline	100.0%***	3 weeks
	[67,100]	
2-week window	100.0%***	2 weeks
	[43,100]	
1-week window	100.0%**	1 week
	[23,100]	
Drop 10 largest	100.0%***	3 weeks
	[52,100]	
Excl. GFC & COVID	100.0%***	3 weeks
	[64,100]	
Multiple tenor auctions	100.0%***	3 weeks
	[74,100]	

The table reports 90% confidence intervals for the risk contribution of the yield response to Treasury auction demand shocks across different specifications. The risk contribution measures the share of the yield's initial decay rate attributable to risk-based costs. The sample spans January 2008 to January 2022, with 165 auction dates. The demand shock is constructed from 5-year Treasury note auctions. Confidence intervals are 5th and 95th percentiles of the date-cluster bootstrap distribution (1,000 replications). Significance: *** p<0.01, ** p<0.05, * p<0.1.

4.2 Crisis Episodes: A Case Study

We now examine two crisis episodes—the Global Financial Crisis (GFC) and the COVID-19 crisis—to assess whether the relative importance of the two costs shifts under stress. Crisis windows are short, leaving few auction events to estimate the risk contribution. We therefore treat this section as a case study: indicative evidence that crises differ from normal times, not a precise decomposition.

To accompany the risk contribution estimate, we report two model-motivated metrics that point to the role of each cost without identifying it directly. The first is the on-impact spread response to a demand shock, estimated from the local projection in (40). By the decomposition in (31), this response reflects both intermediary costs and slow-moving investors' demand elasticity. Holding these elasticities fixed within a window, a larger spread

response signals a greater role for ψ. The second is the ratio of price level variance to spread variance, V_l/V_s. Proposition 3 shows it falls as ψ rises relative to γ: gross position costs amplify spread variance more than level variance. A lower ratio therefore points to a larger role for gross position costs.

Table 4 reports all three metrics across four regimes: the full-sample baseline, the GFC, and two COVID sub-episodes separated by the Supplementary Leverage Ratio (SLR) exemption, which took effect on April 1, 2020. We label the period before the exemption (February–March 2020) as *Covid SLR* and the period after (April 2020–March 2021) as *Covid post-SLR*. All three metrics move in the direction of gross position costs becoming more important during crisis periods relative to the baseline: the risk contribution falls, the spread impact rises, and the variance ratio declines. The pattern is most pronounced during the Covid SLR window, where the spread impact is 2.46 percentage points compared to near zero in all other periods, and the variance ratio falls from 22.9 in the baseline to 13.8. The confidence intervals for the risk contribution, however, are wide in every crisis column, reflecting the limited number of auction events in each subperiod.

Across the three metrics, gross position costs appear to play a larger role during the Covid SLR period than during the GFC. This is consistent with post-GFC regulatory changes. The SLR, introduced under Basel III, penalizes the size of the balance sheet regardless of its risk exposure—mapping to a rise in ψ in our model.

The pre/post-SLR split sharpens the interpretation. After the exemption was granted, all three metrics reverse: the spread impact turns negative and insignificant, the variance ratio rises back toward its baseline level, and the risk contribution point estimate returns to 100%. This before/after pattern is difficult to explain absent a regulatory mechanism, and is consistent with He, Nagel, and Song (2022), who document that the SLR was a binding constraint on Treasury market intermediation during this period.

Table 4: Crisis episodes: A case study

	Baseline	GFC	Covid SLR	Covid post-SLR
Risk contribution	100.0%	83.6%	0.0%	100.0%
	[67, 100]	[0, 100]	[0, 100]	[0, 100]
Spread impact (pp)	0.05	-0.56	2.46***	-0.75
Variance Ratio (V_l/V_s)	22.9	16.5	13.8	17.8

The table reports three metrics across four regimes. The baseline uses the full sample (January 2008 to January 2022). The GFC covers December 2007 to June 2009 (NBER recession dates). Covid SLR covers February 2020 to March 2020, the period before the Federal Reserve's SLR exemption took effect on April 1, 2020. Covid post-SLR covers April 2020 to March 2021. The risk contribution measures the share of the yield's initial decay rate attributable to risk-based costs; confidence intervals are 5th and 95th percentiles of the date-cluster bootstrap distribution (1,000 replications). The spread impact is the estimated coefficient at $h = 5$ (one week) from the local projection (40). The variance ratio V_l/V_s is the ratio of the unconditional variance of daily yield changes to daily spread changes within each window. Significance: *** p<0.01, ** p<0.05, * p<0.1.

5 Calibration

We complement the reduced-form estimation in Section 4 with a structural calibration using unconditional moments, in particular, leveraging the variance term structures of the average yield level and the Treasury-OIS spread. We show that the rate at which variance decays across horizons in each market reveals the relative compensation for taking level versus spread positions.

We further conduct two counterfactual exercises using the calibrated model. The first applies the model to the COVID-19 episode, where our reduced-form results indicate that gross position costs played a larger role in constraining intermediation. The second examines the broader implications of relaxing different regulatory tools—such as the SLR versus the Stress Capital Buffer—for Treasury market functioning.

5.1 Identification of Parameters

Variance Term Structures. The key empirical moments for identification are the yield-space variance term structures of the average yield and the Treasury-OIS (half) spread,

defined as the per-period variance of H-day changes as a function of the horizon H.[25] These moments follow exponential decay

$$V^y_{i,H} = V^y_{i,\infty} + (V^y_{i,0} - V^y_{i,\infty})\frac{1 - e^{-\nu_i H}}{\nu_i H}, \quad i \in \{l, s\}, \tag{42}$$

where $V^y_{i,0}$ is the instantaneous variance in the yield-space, $V^y_{i,\infty}$ is the long-run variance, and ν_i is the decay rate of the variance—it is the eigenvalue of $\mathbf{\Lambda}$ defined in Proposition 4. After converting the fitted moments to price-return units via modified duration, the model implies

$$V_{l,0} = \tfrac{1}{2}\sigma^2_\beta M^2_{l,0} + \sigma^2_V, \qquad V_{l,\infty} = \tfrac{1}{2}\sigma^2_\beta M^2_{l,\infty} + \sigma^2_V, \tag{43}$$

$$V_{s,0} = \tfrac{1}{2}\sigma^2_\beta M^2_{s,0}, \qquad V_{s,\infty} = \tfrac{1}{2}\sigma^2_\beta M^2_{s,\infty}. \tag{44}$$

The level variance loads on both the fundamental shock variance σ^2_V and the demand shock variance σ^2_β through the short-run and long-run level impact multipliers, while the spread variance loads only on demand shocks.

A key insight of the model is that the ratio of the short-run price impact $M_{i,0}$ to the long-run price impact $M_{i,\infty}$, which measures the amplification effect of intermediaries, equals the ratio of the price mean reversion rate ν_i to the institutional capital adjustment speed k:

$$\frac{M_{i,0}}{M_{i,\infty}} = \frac{\nu_i}{k}, \quad i \in \{l, s\}.$$

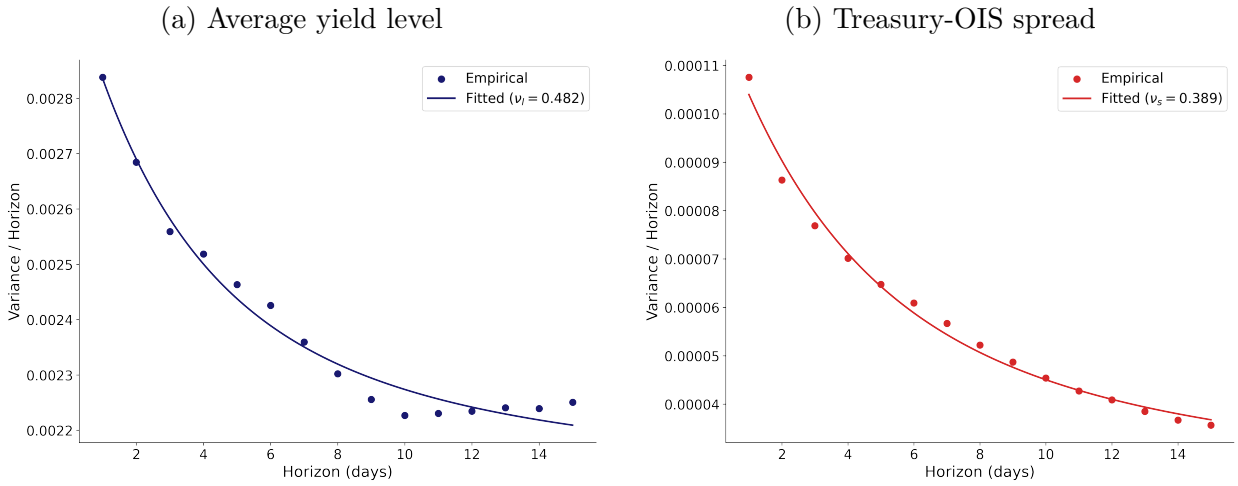
This relation follows directly from (11). The intuition mirrors our reduced-form identification strategy in Section 3.1: the cost of holding inventory is compensated through price mean reversion, so the larger the cost, the faster prices must revert to compensate intermediaries per unit of time, i.e., a larger ν_i. Correspondingly, the amplification due to intermediary frictions $\frac{M_{i,0}}{M_{i,\infty}}$ is high as well. On the other hand, when institutional investors enter at a faster rate, the amplification effect is smaller. Just as the IRF slope ratio m_s/m_l reveals the relative cost of the spread versus level positions, the variance decay rates ν_l and ν_s reveal the relative amplification of level and spread price impacts.

Figure 7 displays the empirical variance term structures for the level factor and spread together with the fitted curves, and Table 5 reports the estimated variance term structure parameters in yield units. Details of the estimation and calibration procedure are discussed in Appendix C. The level variance decays faster than the spread variance, with $\nu_l = 0.48$

[25] We map the empirical moments to the symmetric $r = 0$ model to reduce the number of parameters to be identified. As the two markets closely track each other, this simplification serves as a reasonable approximation. We also assume the shocks are permanent and uncorrelated.

versus $\nu_s = 0.39$, indicating that intermediaries require greater compensation for holding level positions than spread positions.

Figure 7: Variance term structures of the average yield level and Treasury-OIS spread



Each panel plots the empirical per-period yield-space variance $\widehat{V}^y_{i,H} = \mathrm{Var}(\Delta_H y_i)/H$ (dots) and the fitted parametric curve (42) (curve) as a function of horizon H in days. The left panel uses the residualized average of the 5-year Treasury yield and the 5-year OIS rate, $(y^T_{5y,t} + y^{OIS}_{5y,t})/2 - \widehat{\beta} y_{3m,t}$. The right panel uses the 5-year Treasury-OIS spread, $(y^T_{5y,t} - y^{OIS}_{5y,t})/2$. The sample spans January 2008 to January 2022.

Calibrated parameters. To map the decay rates to structural parameters, we first convert the fitted moments in Table 5 from yield units to price-return units using modified duration and then leverage the equilibrium conditions to recover $(\gamma, \psi, k, \sigma_\beta, \sigma_V, M_{s,\infty})$.[26] The details are explained in Appendix C.

Table 6 reports the calibrated structural parameters. Under the calibrated parameter values, the risk contribution to the intermediation cost for the level portfolio is $\frac{2\gamma V_{l,0}}{\psi + 2\gamma V_{l,0}} \approx 80\%$, consistent with the reduced-form evidence in Section 4.[27] The institutional capital adjustment speed is $k = 0.16$, implying that it takes about 6 days on average for institutional capital to adjust to shocks.

[26] We externally calibrate the long-run level multiplier $M_{l,\infty} = 0.15$ from Chaudhary, Fu, and Zhou (2025). This choice is innocuous as the model is scale-invariant to the level price impact $M_{l,\infty}$.

[27] The risk contribution from the calibration is smaller than the point estimate in the auction-shock estimates. However, it is well within the range of the confidence interval.

Table 5: Estimated variance term structure moments

	Spread	Level
$V^y_{i,0}$ (instantaneous)	1.2	30
$V^y_{i,\infty}$ (long-run)	0.19	21
ν_i (decay rate)	0.39	0.48

Estimated yield-space parameters of the variance term structure (42) for the average yield level ($i = l$) and Treasury-OIS spread ($i = s$). $V^y_{i,0}$ is the short-run instantaneous variance, $V^y_{i,\infty}$ is the long-run permanent component, and ν_i is the mean-reversion rate. Variance units are in squared basis points per day; decay rates are in day^{-1}. Appendix C converts $(V^y_{i,0}, V^y_{i,\infty})$ into price-return units before recovering the structural parameters. The sample spans January 2008 to January 2022.

Table 6: Calibrated structural parameters

Parameter	Description	Value
γ	Risk cost	0.87
ψ	Gross position cost	0.03
k	Slow-moving capital speed	0.16
σ_β	Demand shock volatility	0.47
σ_V	Fundamental shock volatility	0.21
$M_{s,\infty}$	Spread demand multiplier	0.06
$M_{l,\infty}$	Level demand multiplier (external)	0.15

Calibrated structural parameters recovered from matching the estimated variance term structure moments in Table 5. The long-run level multiplier $M_{l,\infty} = 0.15$ is set externally following Chaudhary, Fu, and Zhou (2025).

5.2 Counterfactual Analysis

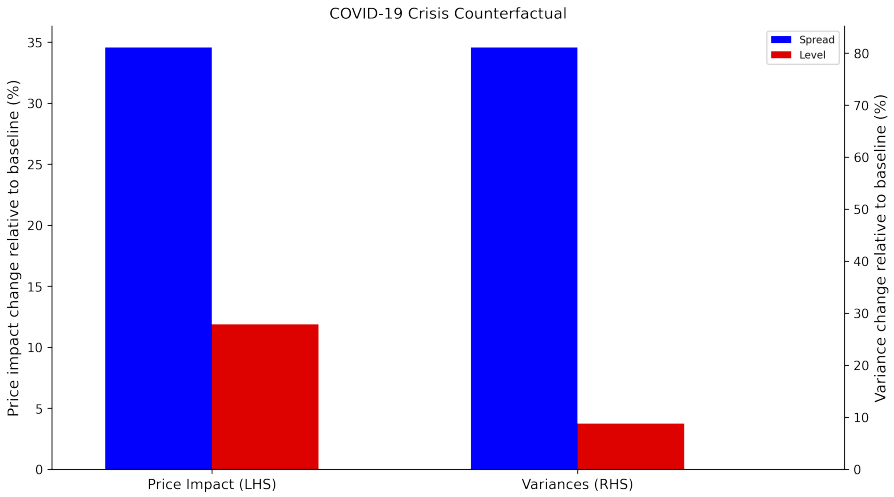
With the structural parameters in hand, we conduct two counterfactual experiments that illustrate how changes in the two cost components propagate asymmetrically into yield volatility and spread volatility, as well as their responses to demand shocks. The first quantifies the COVID-19 episode, where our reduced-form results indicate that gross position costs played a more important role. The second examines the broader implications of regulatory changes for Treasury market functioning.

5.2.1 COVID-19 Crisis

Our first counterfactual zooms in on the COVID-19 episode of 2020, during which Treasury market functioning deteriorated sharply and intermediary balance sheets came under severe pressure. The stress was alleviated quickly due to the Federal Reserve's aggressive interventions. We are interested in learning about the market outcomes if the gross position costs had remained elevated for longer periods of time. To this end, we adjust the gross position cost of the model so that it reproduces the compression in the level-to-spread variance ratio V_l/V_s observed during the Covid SLR window (Table 4). Proposition 3 shows this ratio falls as gross position costs rise, which pins down the implied ψ. We then compute the implied effects on price impact and short-run variances.

Our estimation result in Section 4 suggests that the gross-position cost relative to the risk cost was substantially higher during the COVID crisis than over the full sample, reflecting the severe balance-sheet pressure dealers faced during the dash-for-cash episode in March 2020. As we increase the gross position cost to match the COVID-period variance ratio, our model implies a larger increase in the spread response to demand shocks than the yield response. Figure 8 displays the results. The elevated gross position cost increases spread price impact by about 35%, compared with a 12% increase in yield price impact. This is consistent with the empirical observation that the Treasury-OIS spread widened dramatically in March 2020, as the selling pressure built up in the Treasury market. The changes in short-run variance are more dramatic: the elevated gross position cost increases spread variance by about 81%, while increasing yield variance by only 9%. These results reflect the fact that the gross position cost has a disproportionate effect on spread relative to yield. So while spreads directionally reflected increased distortions in the Treasury yield during COVID-19, they overstated the magnitude of yield distortions.

Figure 8: COVID-19 counterfactual: implied effects of elevated gross position costs



The figure displays counterfactual price impact changes and short-run variance changes for the yield level and Treasury-OIS spread under a COVID-consistent parameterization. The gross position cost ψ is recalibrated to match the proportional decline in the level-to-spread variance ratio V_l/V_s during the Covid SLR window (Table 4); all other parameters are held at their full-sample calibrated values from Table 6.

5.2.2 Banking Deregulation

We now ask how Treasury market outcomes would change under a relaxation of intermediary constraints. Specifically, we compute model-implied outcomes under two scenarios: a 50% reduction in ψ, the gross position cost (as would arise from relaxing the SLR), holding γ fixed; and a 50% reduction in γ, the risk cost (as would arise from reducing the Stress Capital Buffer), holding ψ fixed. For each counterfactual, we resolve the model to obtain the implied short-run amplification factors due to intermediary frictions, and compute the resulting percentage changes in price impact of demand shocks and short-run variance.

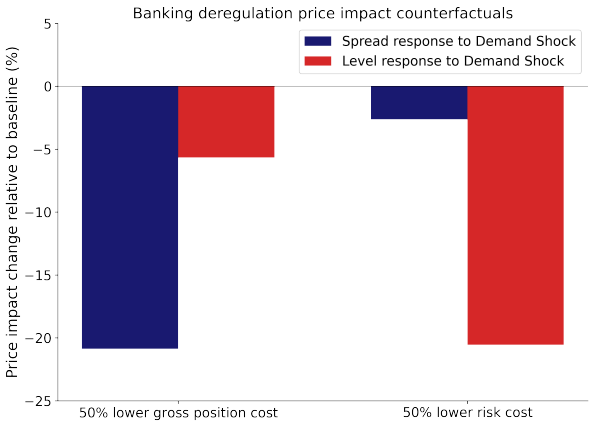
Figure 9 displays the results. The left panel shows percentage changes in price impact, and the right panel shows percentage changes in short-run variance. A sharp asymmetry emerges: reducing the gross position cost ψ compresses the spread response to demand shocks substantially, by about 21%, while having little effect on yield response—reducing it by only 6%. On the other hand, reducing the risk cost γ reduces yield price impact considerably by about 21%, while leaving spread response nearly unaffected (3%). The changes in short-run variance paint a consistent picture: reduction in ψ reduces spread variance by about 37%, yet has very little impact on yield variance (4%). These findings reflect our theoretical findings in Proposition 3 that the gross position cost disproportionately affects spread variance while

the risk cost disproportionately affects yield variance.

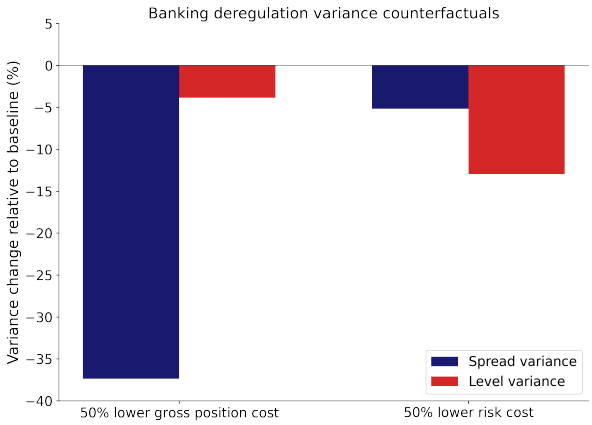
These results have direct regulatory implications. Relaxing gross position costs operates mainly through ψ and therefore has a large effect on the Treasury-OIS spread but a smaller effect on outright yield volatility and Treasury market liquidity. Relaxing risk-based constraints operates through γ and mainly affects yield levels.

Figure 9: Counterfactual effects of reducing intermediary costs by 50%

(a) Change in price impact



(b) Change in short-run variance



Each panel shows the percentage change in price impact amplification (left) and short-run variance (right) for the yield level and Treasury-OIS spread under two counterfactual scenarios: a 50% reduction in the gross position cost ψ and a 50% reduction in the risk cost γ. All other parameters are held at their calibrated values from Table 6. We resolve the model in each case and generate the implied effects on the Treasury markets.

6 Conclusion

Arbitrage spreads are widely used to diagnose market dysfunction, yet they primarily reflect gross position costs—not the risk-based costs that our estimates suggest are the dominant driver of price-level distortions. This paper develops a framework that jointly models how different intermediation costs shape asset price levels and spreads, and introduces a sufficient statistic approach to separately identify them from the initial decay rates of prices following exogenous demand shocks. Applied to Treasury auction shocks, we find that risk-based costs account for the dominant share of intermediation costs on average. During both the GFC and COVID crises, gross position costs appear to have played a larger role—particularly during COVID, where they significantly account for intermediation costs, consistent with Supplementary Leverage Ratio binding.

These findings carry several broader implications. First, spreads can be a misleading barometer of market dysfunction in both directions. In normal times, tight spreads may mask substantial price-level distortions driven by risk costs. In crisis episodes, elevated spreads may overstate the magnitude of yield distortions. Second, our counterfactual analysis suggests that regulatory tools are not interchangeable: relaxing the SLR compresses spreads substantially while having limited effect on yields, whereas reducing risk-based constraints such as the Stress Capital Buffer does the reverse. Policy discussions about Treasury market resilience therefore need to specify which dimension of market functioning is the target.

More broadly, the sufficient statistic approach developed here identifies intermediary cost parameters from price dynamics alone, without requiring data on dealer positions or identical assets—only that intermediaries are the marginal liquidity providers on impact. This makes the framework a misspecification-robust approach to estimate the cost parameters that enter intermediary first-order conditions across a broad class of models.

Finally, our results point to a gap in market monitoring: the standard real-time indicators of intermediation capacity—arbitrage spreads—primarily track gross position costs and may miss the risk-based frictions that appear to dominate price-level dynamics. Developing high-frequency diagnostics that more directly reflect risk-based costs is an important open problem. More broadly, extending the level-versus-spread decomposition to other asset classes and incorporating time-varying costs would deepen our understanding of how intermediation frictions shape asset prices across markets and over time.

References

Adrian, Tobias, Erkko Etula, and Tyler Muir (Dec. 2014). "Financial Intermediaries and the Cross-Section of Asset Returns". In: *The Journal of Finance* 69.6, pp. 2557–2596. ISSN: 0022-1082, 1540-6261. DOI: 10.1111/jofi.12189.

Adrian, Tobias and Hyun Song Shin (Feb. 1, 2014). "Procyclical Leverage and Value-at-Risk". In: *The Review of Financial Studies* 27.2, pp. 373–403. ISSN: 0893-9454. DOI: 10.1093/rfs/hht068.

Bank of England (2025). *Enhancing the Resilience of the Gilt Repo Market*. Discussion Paper. Bank of England. URL: https://www.bankofengland.co.uk/paper/2025/discussion-paper/enhancing-the-resilience-of-the-gilt-repo-market.

Barth, Daniel and R. Jay Kahn (Oct. 1, 2025). "Hedge Funds and the Treasury Cash-Futures Basis Trade". In: *Journal of Monetary Economics* 155, p. 103823. ISSN: 0304-3932. DOI: 10.1016/j.jmoneco.2025.103823.

Board of Governors of the Federal Reserve System (May 15, 2020). *Financial Stability Report: May 2020*. Board of Governors of the Federal Reserve System.

Boyarchenko, Nina et al. (July 1, 2018). *Bank-Intermediated Arbitrage*. DOI: 10.2139/ssrn.3200041. Social Science Research Network: 3200041. Pre-published.

Brunnermeier, Markus K. and Yuliy Sannikov (Feb. 2014). "A Macroeconomic Model with a Financial Sector". In: *American Economic Review* 104.2, pp. 379–421. DOI: 10.1257/aer.104.2.379.

Chaudhary, Manav, Julie Zhiyu Fu, and Jian Li (2025). "Corporate Bond Multipliers: Substitutes Matter". Working paper.

Chaudhary, Manav, Julie Zhiyu Fu, and Haonan Zhou (Oct. 2025). "Anatomy of the Treasury Market: Who Moves Yields?" Working paper. URL: https://fuzhiyu.me/TreasuryGIVPaper/Treasury_GIV_draft.pdf.

Du, Wenxin, Benjamin Hébert, and Amy Wang Huber (Apr. 1, 2023). "Are Intermediary Constraints Priced?" In: *The Review of Financial Studies* 36.4, pp. 1464–1507. ISSN: 0893-9454. DOI: 10.1093/rfs/hhac050.

Du, Wenxin, Benjamin Hébert, and Wenhao Li (Dec. 1, 2023). "Intermediary Balance Sheets and the Treasury Yield Curve". In: *Journal of Financial Economics* 150.3, p. 103722. ISSN: 0304-405X. DOI: 10.1016/j.jfineco.2023.103722.

Du, Wenxin, Alexander Tepper, and Adrien Verdelhan (2018). "Deviations from Covered Interest Rate Parity". In: *The Journal of Finance* 73.3, pp. 915–957. ISSN: 1540-6261. DOI: 10.1111/jofi.12620.

Duffie, Darrell (2010). "Presidential Address: Asset Price Dynamics with Slow-Moving Capital". In: *The Journal of Finance* 65.4, pp. 1237–1267. ISSN: 1540-6261. DOI: 10.1111/j.1540-6261.2010.01569.x.

Duffie, Darrell et al. (Aug. 2023). *Dealer Capacity and U.S. Treasury Market Functionality*. Staff Reports (Federal Reserve Bank of New York). Federal Reserve Bank of New York. DOI: 10.59576/sr.1070.

Fleckenstein, Matthias and Francis A Longstaff (Nov. 1, 2020). "Renting Balance Sheet Space: Intermediary Balance Sheet Rental Costs and the Valuation of Derivatives". In: *The Review of Financial Studies* 33.11, pp. 5051–5091. ISSN: 0893-9454. DOI: 10.1093/rfs/hhaa033.

Fleming, Michael, Giang Nguyen, and Joshua Rosenberg (Aug. 1, 2024). "How Do Treasury Dealers Manage Their Positions?" In: *Journal of Financial Economics* 158, p. 103885. DOI: 10.1016/j.jfineco.2024.103885.

Gabaix, Xavier and Ralph S. J. Koijen (June 28, 2021). *In Search of the Origins of Financial Fluctuations: The Inelastic Markets Hypothesis*. w28967. National Bureau of Economic Research. DOI: 10.3386/w28967.

Gârleanu, Nicolae and Lasse Heje Pedersen (June 1, 2011). "Margin-Based Asset Pricing and Deviations from the Law of One Price". In: *The Review of Financial Studies* 24.6, pp. 1980–2022. ISSN: 0893-9454. DOI: 10.1093/rfs/hhr027.

Greenwood, Robin, Samuel G Hanson, and Gordon Y Liao (Apr. 2018). "Asset Price Dynamics in Partially Segmented Markets". In: *The Review of Financial Studies* 31.9, pp. 3307–3343. ISSN: 0893-9454. DOI: 10.1093/rfs/hhy048. eprint: https://academic.oup.com/rfs/article-pdf/31/9/3307/25521746/hhy048.pdf.

Gromb, Denis and Dimitri Vayanos (Nov. 1, 2002). "Equilibrium and Welfare in Markets with Financially Constrained Arbitrageurs". In: *Journal of Financial Economics*. Limits on Arbitrage 66.2, pp. 361–407. ISSN: 0304-405X. DOI: 10.1016/S0304-405X(02)00228-3.

Haddad, Valentin and Tyler Muir (Dec. 2021). "Do Intermediaries Matter for Aggregate Asset Prices?" In: *The Journal of Finance* 76.6, pp. 2719–2761. ISSN: 0022-1082, 1540-6261. DOI: 10.1111/jofi.13086.

— (2025). "Intermediaries and Asset Prices". In.

Hanson, Samuel G., Aytek Malkhozov, and Gyuri Venter (Apr. 1, 2024). "Demand-and-Supply Imbalance Risk and Long-Term Swap Spreads". In: *Journal of Financial Economics* 154, p. 103814. ISSN: 0304-405X. DOI: 10.1016/j.jfineco.2024.103814.

Hazelkorn, Todd M., Tobias J. Moskowitz, and Kaushik Vasudevan (2023). "Beyond Basis Basics: Liquidity Demand and Deviations from the Law of One Price". In: *The Journal of Finance* 78.1, pp. 301–345. DOI: https://doi.org/10.1111/jofi.13198. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jofi.13198. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/jofi.13198.

He, Zhiguo, Bryan Kelly, and Asaf Manela (Oct. 1, 2017). "Intermediary Asset Pricing: New Evidence from Many Asset Classes". In: *Journal of Financial Economics* 126.1, pp. 1–35. ISSN: 0304-405X. DOI: 10.1016/j.jfineco.2017.08.002.

He, Zhiguo and Arvind Krishnamurthy (Apr. 2013). "Intermediary Asset Pricing". In: *American Economic Review* 103.2, pp. 732–70. DOI: 10.1257/aer.103.2.732.

— (Nov. 1, 2018). "Intermediary Asset Pricing and the Financial Crisis". In: *Annual Review of Financial Economics* 10 (Volume 10, 2018), pp. 173–197. ISSN: 1941-1367, 1941-1375. DOI: 10.1146/annurev-financial-110217-022636.

He, Zhiguo, Stefan Nagel, and Zhaogang Song (Jan. 1, 2022). "Treasury Inconvenience Yields during the COVID-19 Crisis". In: *Journal of Financial Economics* 143.1, pp. 57–79. ISSN: 0304-405X. DOI: 10.1016/j.jfineco.2021.05.044.

Jordà, Òscar (2005). "Estimation and Inference of Impulse Responses by Local Projections". In: *American Economic Review* 95.1, pp. 161–182. ISSN: 0002-8282. DOI: 10.1257/0002828053828518.

Kashyap, Anil K et al. (2025). "Treasury Market Dysfunction and the Role of the Central Bank". In: *Brookings Papers on Economic Activity*. Spring 2025.

Klingler, Sven and Suresh Sundaresan (2023). "Diminishing treasury convenience premiums: Effects of dealers' excess demand and balance sheet constraints". In: *Journal of Monetary Economics* 135, pp. 55–69. ISSN: 0304-3932. DOI: https://doi.org/10.1016/j.jmoneco.2023.01.002. URL: https://www.sciencedirect.com/science/article/pii/S0304393223000089.

Koijen, Ralph S. J. and Motohiro Yogo (2019). "A Demand System Approach to Asset Pricing". In: *Journal of Political Economy* 127.4, pp. 1475–1515. ISSN: 0022-3808. DOI: 10.1086/701683.

Kondor, Peter (2009). "Risk in Dynamic Arbitrage: The Price Effects of Convergence Trading". In: *The Journal of Finance* 64.2, pp. 631–655. DOI: https://doi.org/10.1111/j.1540-6261.2009.01445.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-6261.2009.01445.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.2009.01445.x.

Li, Dan, Lubomir Petrasek, and Mary Tian (2025). "Risk-averse Dealers in a Risk-free Market - The Role of Trading Desk Risk Limits". Working paper.

Montiel Olea, José Luis and Mikkel Plagborg-Møller (2021). "Local Projection Inference Is Simpler and More Robust than You Think". In: *Econometrica : journal of the Econometric Society* 89.4, pp. 1789–1823. ISSN: 1468-0262. DOI: 10.3982/ECTA18756.

Ray, Walker, Michael Droste, and Yuriy Gorodnichenko (Sept. 1, 2024). "Unbundling Quantitative Easing: Taking a Cue from Treasury Auctions". In: *Journal of Political Economy*. DOI: 10.1086/729581.

Santos, Tano and Pietro Veronesi (Aug. 1, 2022). "Leverage". In: *Journal of Financial Economics* 145 (2, Part B), pp. 362–386. ISSN: 0304-405X. DOI: 10.1016/j.jfineco.2021.09.001.

Somogyi, Fabricius, Jonathan Wallen, and Lingdi Xu (2025). "What Treasury Auctions Reveal About Investor Demand". Working Paper, Harvard Business School.

Vayanos, Dimitri and Jean-Luc Vila (2021). "A Preferred-Habitat Model of the Term Structure of Interest Rates". In: *Econometrica : journal of the Econometric Society* 89.1, pp. 77–112. DOI: 10.3982/ECTA17440. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA17440.

A Derivations and Proofs

Law of motion. From market clearing condition we have

$$\mathbf{x}_t + \mathbf{y}_t + \boldsymbol{\beta}_t = \bar{\mathbf{S}} \tag{45}$$

The law of motion for $dy_{i,t}$ is given by

$$dy_{i,t} = k(Z_{i,t} - y_{i,t})dt \tag{46}$$

Plugging in the expression for $Z_{i,t}$ and writing in vector form, we have

$$d\mathbf{y}_t = k(-\boldsymbol{\zeta}\mathbf{P}_t + V_i\boldsymbol{\zeta}\mathbf{1} + \boldsymbol{\theta} - \mathbf{y}_t)dt \tag{47}$$

Substituting the market clearing condition, we have

$$d\mathbf{x}_t = -k\Big(\mathbf{x}_t - \boldsymbol{\zeta}\mathbf{P}_t + V_t\boldsymbol{\zeta}\mathbf{1} + \boldsymbol{\theta} - \bar{\mathbf{S}} + \boldsymbol{\beta}_t\Big)dt - d\boldsymbol{\beta}_t. \tag{48}$$

Next, substituting (9) in (48), we get

$$d\mathbf{x}_t = -k\Big(\mathbf{x}_t - \boldsymbol{\zeta}\mathbf{p}\mathbf{s}_t - \boldsymbol{\zeta}\bar{\mathbf{P}} + V_t\boldsymbol{\zeta}\mathbf{1} + \boldsymbol{\theta} - \bar{\mathbf{S}} + \boldsymbol{\beta}_t\Big)dt - d\boldsymbol{\beta}_t. \tag{49}$$

$$= -k(\mathbf{I} - \boldsymbol{\zeta}\boldsymbol{\lambda}_x)\mathbf{x}_t dt - k(\mathbf{I} - \boldsymbol{\zeta}\boldsymbol{\lambda}_\beta)\boldsymbol{\beta}_t dt + k\boldsymbol{\zeta}(\boldsymbol{\lambda}_V - \mathbf{1})V_t dt + \text{constant}dt - d\boldsymbol{\beta}_t. \tag{50}$$

This implies the instantaneous innovation covariance matrix is given by

$$\boldsymbol{\Sigma} \equiv \frac{\text{Cov}(d\mathbf{s}_t)}{dt} = \begin{pmatrix} \sigma_C^2 & \sigma_{CS} & -\sigma_C^2 & -\sigma_{CS} & -\sigma_{CV} \\ \sigma_{CS} & \sigma_S^2 & -\sigma_{CS} & -\sigma_S^2 & -\sigma_{SV} \\ -\sigma_C^2 & -\sigma_{CS} & \sigma_C^2 & \sigma_{CS} & \sigma_{CV} \\ -\sigma_{CS} & -\sigma_S^2 & \sigma_{CS} & \sigma_S^2 & \sigma_{SV} \\ -\sigma_{CV} & -\sigma_{SV} & \sigma_{CV} & \sigma_{SV} & \sigma_V^2 \end{pmatrix} \tag{51}$$

and the mean reversion matrix $\boldsymbol{\Gamma}$ is given by

$$\boldsymbol{\Gamma} = \begin{pmatrix} \boldsymbol{\Lambda} & k(\mathbf{I} - \boldsymbol{\zeta}\boldsymbol{\lambda}_\beta) - \boldsymbol{\eta}_\beta & -k\boldsymbol{\zeta}(\boldsymbol{\lambda}_V - \mathbf{1}) \\ \mathbf{0}_{2\times 2} & \boldsymbol{\eta}_\beta & \mathbf{0}_{2\times 1} \\ \mathbf{0}_{1\times 2} & \mathbf{0}_{1\times 2} & \eta_v \end{pmatrix} \tag{52}$$

where $\mathbf{\Lambda} = k(\mathbf{I} - \boldsymbol{\zeta}\boldsymbol{\lambda}_x)$, and $\boldsymbol{\eta}_\beta = \begin{pmatrix} \eta_{\beta_C} & 0 \\ 0 & \eta_{\beta_S} \end{pmatrix}$.

A.1 Proof of Proposition 1

To solve for the equilibrium price, we plug the conjectured price function into the first order condition (5). We have

$$-\mathbf{p}\mathbf{\Gamma}(\mathbf{s}_t - \bar{\mathbf{s}}) = \mathbf{C}\mathbf{x}_t \tag{53}$$

$$-\begin{pmatrix} \boldsymbol{\lambda}_x & \boldsymbol{\lambda}_\beta & \boldsymbol{\lambda}_V \end{pmatrix} \begin{pmatrix} \mathbf{\Lambda} & k(\mathbf{I} - \boldsymbol{\zeta}\boldsymbol{\lambda}_\beta) & -k\boldsymbol{\zeta}(\boldsymbol{\lambda}_V - \mathbf{1}) \\ \mathbf{0}_{2\times 2} & \boldsymbol{\eta}_\beta & \mathbf{0}_{2\times 1} \\ \mathbf{0}_{1\times 2} & \mathbf{0}_{1\times 2} & \eta_v \end{pmatrix} \begin{pmatrix} \mathbf{x}_t - \bar{\mathbf{x}} \\ \boldsymbol{\beta}_t - \bar{\boldsymbol{\beta}} \\ V_t - \bar{V} \end{pmatrix} = \mathbf{C}\mathbf{x}_t \tag{54}$$

Expanding the equation we get

$$-\boldsymbol{\lambda}_x\mathbf{\Lambda}(\mathbf{x}_t - \bar{\mathbf{x}}) - \boldsymbol{\lambda}_\beta\boldsymbol{\eta}_\beta(\boldsymbol{\beta}_t - \bar{\boldsymbol{\beta}}) - \boldsymbol{\lambda}_V\eta_v(V_t - \bar{V}) - k\boldsymbol{\lambda}_x(\mathbf{I} - \boldsymbol{\zeta}\boldsymbol{\lambda}_\beta)(\boldsymbol{\beta}_t - \bar{\boldsymbol{\beta}}) \tag{55}$$

$$+ k\boldsymbol{\lambda}_x\boldsymbol{\zeta}(\boldsymbol{\lambda}_V - \mathbf{1})(V_t - \bar{V}) = \mathbf{C}\mathbf{x}_t \tag{56}$$

Matching the coefficients we have

$$[\mathbf{x}_t]: \quad -\boldsymbol{\lambda}_x\mathbf{\Lambda} = \mathbf{C} \tag{57}$$

$$[\boldsymbol{\beta}_t]: \quad -\boldsymbol{\lambda}_\beta\boldsymbol{\eta}_\beta - k\boldsymbol{\lambda}_x(\mathbf{I} - \boldsymbol{\zeta}\boldsymbol{\lambda}_\beta) = \mathbf{0} \tag{58}$$

$$[V_t]: \quad -\boldsymbol{\lambda}_V\eta_v + k\boldsymbol{\lambda}_x\boldsymbol{\zeta}(\boldsymbol{\lambda}_V - \mathbf{1}) = \mathbf{0} \tag{59}$$

Consider the limiting case where $\eta_{\beta_i} \to 0$ and $\eta_v \to 0$. When $\boldsymbol{\lambda}_x$ has full rank, from (58) and (59) we must have

$$\boldsymbol{\lambda}_\beta = \boldsymbol{\zeta}^{-1} \tag{60}$$

$$\boldsymbol{\lambda}_V = \mathbf{1} \tag{61}$$

To solve for $\boldsymbol{\lambda}_x$, we can plug the expression for $\boldsymbol{\lambda}_\beta$ and $\boldsymbol{\lambda}_V$ back into $\mathbf{C}$ in (57)

$$-\boldsymbol{\lambda}_x\mathbf{\Lambda} = \begin{pmatrix} \psi_C & 0 \\ 0 & \psi_S \end{pmatrix} + \gamma \begin{pmatrix} \boldsymbol{\lambda}_x & \boldsymbol{\zeta}^{-1} & \mathbf{1} \end{pmatrix} \mathbf{\Sigma} \begin{pmatrix} \boldsymbol{\lambda}_x \\ \boldsymbol{\zeta}^{-1} \\ \mathbf{1} \end{pmatrix} \tag{62}$$

in which the only unknown is $\boldsymbol{\lambda}_x$. This is a quadratic matrix equation in $\boldsymbol{\lambda}_x$ and is only well-defined when $\psi_C > 0$ and $\psi_S > 0$.

Finally, matching the coefficients on the constant term we have $\bar{\mathbf{x}} = \mathbf{0}$, which implies

$$\bar{\mathbf{P}} = \boldsymbol{\zeta}^{-1}\left(\boldsymbol{\theta} - \bar{\mathbf{S}}\right) \tag{63}$$

A.2 Proof of Corollary 1: Perfectly-integrated equilibrium

When $\psi_C = \psi_S = 0$, Proposition 1 does not imply uniqueness because (57) may admit rank-one solutions for $\boldsymbol{\lambda}_x$.

We show that under the regularity condition,

$$k\mathbf{1}^\top\boldsymbol{\zeta}\mathbf{1} > \gamma\tilde{\sigma}_\beta^2, \tag{64}$$

where $\tilde{\sigma}_\beta^2 \equiv \sigma_C^2 + 2\sigma_{CS} + \sigma_S^2$, there exists another comoving equilibrium. Here we show the existence by guess and verify.

Conjecture that both prices have identical loadings:

$$\boldsymbol{\lambda}_x = \lambda_x\mathbf{1}\mathbf{1}^\top, \qquad \boldsymbol{\lambda}_\beta = \lambda_\beta\mathbf{1}\mathbf{1}^\top, \qquad \boldsymbol{\lambda}_V = \lambda_V\mathbf{1}. \tag{65}$$

This implies $P_{C,t} = P_{S,t}$ for all states.

Under the same limiting case as in Proposition 1, $\eta_{\beta_C} = \eta_{\beta_S} = \eta_v = 0$, equations (58) and (59) become

$$\boldsymbol{\lambda}_x(\mathbf{I} - \boldsymbol{\zeta}\boldsymbol{\lambda}_\beta) = \mathbf{0}, \qquad \boldsymbol{\lambda}_x\boldsymbol{\zeta}(\boldsymbol{\lambda}_V - \mathbf{1}) = \mathbf{0}. \tag{66}$$

Using our conjecture, the above equations reduce to

$$\lambda_x\left(1 - \lambda_\beta\mathbf{1}^\top\boldsymbol{\zeta}\mathbf{1}\right)\mathbf{1}\mathbf{1}^\top = \mathbf{0}, \qquad \lambda_x\left(\mathbf{1}^\top\boldsymbol{\zeta}\mathbf{1}\right)(\lambda_V - 1)\mathbf{1} = \mathbf{0}. \tag{67}$$

Hence

$$\lambda_\beta = \frac{1}{\mathbf{1}^\top\boldsymbol{\zeta}\mathbf{1}}, \qquad \lambda_V = 1, \tag{68}$$

That is, λ_β is the reciprocal of the aggregate elasticity.

We solve λ_x from (57). Using the conjecture above, we have:

$$\mathbf{\Lambda} = k(\mathbf{I} - \boldsymbol{\zeta}\boldsymbol{\lambda}_x) = k(\mathbf{I} - \lambda_x \boldsymbol{\zeta}\mathbf{1}\mathbf{1}^\top), \tag{69}$$

so

$$-\boldsymbol{\lambda}_x\mathbf{\Lambda} = -k\lambda_x \left(1 - \frac{\lambda_x}{\lambda_\beta}\right)\mathbf{1}\mathbf{1}^\top. \tag{70}$$

On the right-hand side of (57), as $\psi_C = \psi_S = 0$, $\mathbf{C} = \gamma\mathbf{p}\mathbf{\Sigma}\mathbf{p}^\top$ under the comoving equilibrium. Define

$$\delta \equiv 1 - \frac{\lambda_x}{\lambda_\beta}, \qquad \tilde{\sigma}_{\beta V} \equiv \sigma_{CV} + \sigma_{SV}. \tag{71}$$

Then the diffusion term (denoted with superscript diff) of the price is given by

$$\begin{aligned} dP_{i,t}^{\text{diff}} &= \lambda_x(dx_{C,t}^{\text{diff}} + dx_{S,t}^{\text{diff}}) + \lambda_\beta(d\beta_{C,t} + d\beta_{S,t}) + dV_t & (72)\\ &= \lambda_\beta\delta(d\beta_{C,t} + d\beta_{S,t}) + dV_t, \qquad i \in \{C, S\}, & (73) \end{aligned}$$

where we used $d\mathbf{x}_t^{\text{diff}} = -d\boldsymbol{\beta}_t$ from (50). Therefore

$$\text{Var}\left(dP_{i,t}^{\text{diff}}\right) = \tilde{\sigma}_\beta^2\lambda_\beta^2\delta^2 + 2\tilde{\sigma}_{\beta V}\lambda_\beta\delta + \sigma_V^2, \tag{74}$$

and hence

$$\begin{aligned} \mathbf{p}\mathbf{\Sigma}\mathbf{p}^\top &= \left(\tilde{\sigma}_\beta^2\lambda_\beta^2\delta^2 + 2\tilde{\sigma}_{\beta V}\lambda_\beta\delta + \sigma_V^2\right)\mathbf{1}\mathbf{1}^\top, & (75)\\ \mathbf{C} = \gamma\mathbf{p}\mathbf{\Sigma}\mathbf{p}^\top &= \gamma\left(\tilde{\sigma}_\beta^2\lambda_\beta^2\delta^2 + 2\tilde{\sigma}_{\beta V}\lambda_\beta\delta + \sigma_V^2\right)\mathbf{1}\mathbf{1}^\top. & (76) \end{aligned}$$

Then (57) reduces to the scalar condition

$$-k\lambda_x\left(1 - \frac{\lambda_x}{\lambda_\beta}\right) = \gamma\left(\tilde{\sigma}_\beta^2\lambda_\beta^2\delta^2 + 2\tilde{\sigma}_{\beta V}\lambda_\beta\delta + \sigma_V^2\right). \tag{77}$$

Rearranging (77), we obtain

$$\left(k - \gamma\lambda_\beta\tilde{\sigma}_\beta^2\right)\delta^2 + (-k - 2\gamma\tilde{\sigma}_{\beta V})\delta - \frac{\gamma\sigma_V^2}{\lambda_\beta} = 0. \tag{78}$$

Under the regularity condition (64), (78) has one positive root and one negative root.

Moreover,

$$\mathbf{\Lambda} = k(\mathbf{I} - \lambda_x \boldsymbol{\zeta}\mathbf{1}\mathbf{1}^\top). \tag{79}$$

For any vector u such that $\mathbf{1}^\top u = 0$, we have $\boldsymbol{\zeta}\mathbf{1}\mathbf{1}^\top u = \mathbf{0}$, so

$$\mathbf{\Lambda} u = ku. \tag{80}$$

Thus one eigenvalue is $\nu_s = k$. Also,

$$\mathbf{\Lambda}(\boldsymbol{\zeta}\mathbf{1}) = k\left(\boldsymbol{\zeta}\mathbf{1} - \lambda_x \boldsymbol{\zeta}\mathbf{1}\mathbf{1}^\top \boldsymbol{\zeta}\mathbf{1}\right) \tag{81}$$

$$= k\left(1 - \frac{\lambda_x}{\lambda_\beta}\right)\boldsymbol{\zeta}\mathbf{1} = k\delta\,\boldsymbol{\zeta}\mathbf{1}. \tag{82}$$

So the other eigenvalue is $\nu_l = k\delta$. Choosing the positive root of (78) yields $\nu_l > 0$, hence both eigenvalues of $\mathbf{\Lambda}$ are positive.

Dynamics under the perfectly integrated equilibrium. To see how this equilibrium is maintained, it is instructive to analyze the dynamics of the intermediary inventory. Define

$$x_{l,t} \equiv x_{C,t} + x_{S,t}, \qquad \beta_{l,t} \equiv \beta_{C,t} + \beta_{S,t}, \qquad x_{s,t} \equiv x_{C,t} - x_{S,t}, \qquad \beta_{s,t} \equiv \beta_{C,t} - \beta_{S,t}. \tag{83}$$

Since $P_{C,t} = P_{S,t}$ in the comoving equilibrium, the common price can be written as

$$\mathbf{P}_t = \left(\bar{P} + V_t + \lambda_x x_{l,t} + \lambda_\beta \beta_{l,t}\right)\mathbf{1}. \tag{84}$$

Substituting the comoving coefficients into (50) gives

$$d\mathbf{x}_t = -k(\mathbf{I} - \lambda_x \boldsymbol{\zeta}\mathbf{1}\mathbf{1}^\top)\mathbf{x}_t dt - k(\mathbf{I} - \lambda_\beta \boldsymbol{\zeta}\mathbf{1}\mathbf{1}^\top)\boldsymbol{\beta}_t dt - d\boldsymbol{\beta}_t + \text{constant}dt. \tag{85}$$

Because $\lambda_\beta = (\mathbf{1}^\top \boldsymbol{\zeta}\mathbf{1})^{-1}$, left-multiplying by $\mathbf{1}^\top$ yields

$$dx_{l,t} = -\nu_l x_{l,t} dt - d\beta_{l,t} + \text{constant}dt, \qquad \nu_l \equiv k\left(1 - \frac{\lambda_x}{\lambda_\beta}\right). \tag{86}$$

Hence, in the martingale limit $\eta_{\beta_C} = \eta_{\beta_S} = 0$, a pure relative-demand shock holding $\beta_{C,t}+\beta_{S,t}$ fixed leaves the aggregate state unchanged. By (84), this implies

$$\frac{\partial \mathbb{E}_t[P_{i,t+\tau}]}{\partial(\beta_{C,t} - \beta_{S,t})} = 0, \qquad i \in \{C, S\}. \tag{87}$$

To characterize quantities, left-multiply (85) by $(1, -1)$ to obtain

$$dx_{s,t} = -k(x_{s,t} + \beta_{s,t})dt + k\left((1,-1)\boldsymbol{\zeta}\mathbf{1}\right)\left(\lambda_x x_{l,t} + \lambda_\beta \beta_{l,t}\right)dt - d\beta_{s,t} + \text{constant}dt. \tag{88}$$

The second term depends only on the level states $x_{l,t}$ and $\beta_{l,t}$, so under a pure relative-demand shock it is unchanged because $x_{l,t}$ and $\beta_{l,t}$ are unchanged. Therefore

$$\frac{\partial \mathbb{E}_t[x_{s,t+\tau}]}{\partial \beta_{s,t}} = -1, \tag{89}$$

Thus the relative demand imbalance is fully warehoused by arbitrageurs.

Non-comoving equilibrium. We note that when $\psi_C = \psi_S = 0$, the model yields multiple equilibria. There may exist another non-comoving equilibrium governed by the equilibrium conditions characterized by Proposition 1. To see that, we consider the model under symmetric case with $\eta_{\beta_C} = \eta_{\beta_S} = \eta_v = 0$, $r = 0$, $\psi_C = \psi_S = 0$, and

$$\boldsymbol{\zeta} = \begin{pmatrix} \zeta_d & \zeta_o \\ \zeta_o & \zeta_d \end{pmatrix}, \qquad \boldsymbol{\lambda}_x = \begin{pmatrix} \frac{l_x+s_x}{2} & \frac{l_x-s_x}{2} \\ \frac{l_x-s_x}{2} & \frac{l_x+s_x}{2} \end{pmatrix}. \tag{90}$$

From (58) and (59) (with full-rank $\boldsymbol{\lambda}_x$), we have $\boldsymbol{\lambda}_\beta = \boldsymbol{\zeta}^{-1}$ and $\boldsymbol{\lambda}_V = \mathbf{1}$. Similar to the analysis in the main text, we define the level and spread components of $\boldsymbol{\lambda}_x$ and $\boldsymbol{\zeta}^{-1}$ as

$$M_{l,\infty} \equiv \frac{1}{\zeta_d + \zeta_o}, \quad M_{s,\infty} \equiv \frac{1}{\zeta_d - \zeta_o}, \tag{91}$$

where $(M_{l,\infty}, M_{s,\infty})$ are exactly the rotated long-run impacts in (18), and the associated level/spread price and variance objects $(P_{l,t}, P_{s,t}, V_{l,0}, V_{s,0})$ are defined in (19). Under the same symmetric-shock assumptions as in Proposition 2, (57) decouples into

$$k\frac{l_x}{M_{l,\infty}}(l_x - M_{l,\infty}) = \gamma\left(\sigma_\beta^2(l_x - M_{l,\infty})^2 + 2\sigma_V^2\right), \tag{92}$$

$$k\frac{s_x}{M_{s,\infty}}(s_x - M_{s,\infty}) = \gamma\sigma_\beta^2(s_x - M_{s,\infty})^2. \tag{93}$$

Define $\phi_l \equiv 1 - \gamma\sigma_\beta^2 M_{l,\infty}/k$ and $\phi_s \equiv 1 - \gamma\sigma_\beta^2 M_{s,\infty}/k$. Then (92) gives

$$M_{l,\infty} - l_x = \frac{M_{l,\infty}}{2\phi_l}\left(1 + \sqrt{1 + \frac{8\phi_l\gamma\sigma_V^2}{kM_{l,\infty}}}\right), \tag{94}$$

and (93) gives two branches:

$$M_{s,\infty} - s_x = 0 \qquad \text{or} \qquad M_{s,\infty} - s_x = \frac{M_{s,\infty}}{\phi_s}. \tag{95}$$

Now, under $\phi_l > 0$ and $\phi_s > 0$, the eigenvalues of

$$\mathbf{\Lambda} = k(\mathbf{I} - \boldsymbol{\zeta}\boldsymbol{\lambda}_x) \tag{96}$$

are

$$\nu_l = k\left(1 - \frac{l_x}{M_{l,\infty}}\right), \qquad \nu_s = k\left(1 - \frac{s_x}{M_{s,\infty}}\right). \tag{97}$$

For the branch $M_{s,\infty} - s_x = \frac{M_{s,\infty}}{\phi_s}$, we obtain

$$\nu_l = \frac{k}{2\phi_l}\left(1 + \sqrt{1 + \frac{8\phi_l\gamma\sigma_V^2}{kM_{l,\infty}}}\right) > 0, \qquad \nu_s = \frac{k}{\phi_s} > 0. \tag{98}$$

Moreover,

$$s_x = M_{s,\infty} - \frac{M_{s,\infty}}{\phi_s} = -\frac{\gamma\sigma_\beta^2 M_{s,\infty}^2}{k - \gamma\sigma_\beta^2 M_{s,\infty}} \neq 0 \tag{99}$$

when $\gamma > 0$, so the solution is non-comoving. Therefore, besides the comoving solution in Corollary 1, there exists an analytically characterized non-comoving equilibrium with both eigenvalues of $\mathbf{\Lambda}$ strictly positive.

A.3 Proof of Proposition 4

The proof below applies to the full-rank equilibrium characterized by Proposition 1. The comoving rank-one equilibrium in Corollary 1 is characterized separately above.

In the case when $\boldsymbol{\psi} \neq \mathbf{0}$, we can plug in the solution for prices into the law of motion for

$\mathbf{x}_t$ in (50) to get

$$d\mathbf{x}_t = -\mathbf{\Lambda}\mathbf{x}_t dt - d\boldsymbol{\beta}_t + \text{constant}dt \tag{100}$$

For stationarity, we need the eigenvalues of $\mathbf{\Lambda}$ to be positive, which ensures $e^{-\mathbf{\Lambda}T} \to 0$ as $T \to \infty$. Hence the price impact following a demand shock decays exponentially at rate $\mathbf{\Lambda}$.

$$\frac{\partial \mathbb{E}_t[\mathbf{x}_{t+\tau}]}{\partial \boldsymbol{\beta}_t^\top} = -e^{-\mathbf{\Lambda}\tau} \tag{101}$$

where

$$e^{-\mathbf{\Lambda}\tau} = \frac{e^{-\nu_1\tau}}{\nu_2 - \nu_1}(\nu_2\mathbf{I} - \mathbf{\Lambda}) + \frac{e^{-\nu_2\tau}}{\nu_1 - \nu_2}(\nu_1\mathbf{I} - \mathbf{\Lambda}) \tag{102}$$

Here ν_1 and ν_2 are the two eigenvalues of $\mathbf{\Lambda}$, and they control the speed of convergence to long-run behavior.

Next, we analyze the price dynamics following a demand shock. Using

$$\mathbf{P}_t = \boldsymbol{\lambda}_x \mathbf{x}_t + \boldsymbol{\lambda}_\beta \boldsymbol{\beta}_t + \boldsymbol{\lambda}_V V_t + \bar{\mathbf{P}} \tag{103}$$

$$\frac{\partial \mathbb{E}_t[\mathbf{P}_{t+\tau}]}{\partial \boldsymbol{\beta}_t^\top} = \boldsymbol{\lambda}_x \frac{\partial \mathbb{E}_t[\mathbf{x}_{t+\tau}]}{\partial \boldsymbol{\beta}_t^\top} + \boldsymbol{\lambda}_\beta \tag{104}$$

Plug in $\frac{\partial \mathbb{E}_t[\mathbf{x}_{t+\tau}]}{\partial \boldsymbol{\beta}_t^\top}$ from (101) and $\boldsymbol{\lambda}_\beta$ from (60), we have

$$\frac{\partial \mathbb{E}_t[\mathbf{P}_{t+\tau}]}{\partial \boldsymbol{\beta}_t^\top} = -\boldsymbol{\lambda}_x e^{-\mathbf{\Lambda}\tau} + \boldsymbol{\zeta}^{-1} \tag{105}$$

A.4 Proof of Proposition 2

Given the definition of $P_{l,t}$ and $P_{s,t}$,

$$\text{Var}(dP_{l,t})/dt = \frac{1}{4}(1,1)\mathbf{\Sigma}_P \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \frac{1}{4}(1,1)\mathbf{p}\mathbf{\Sigma}\mathbf{p}^\top \begin{pmatrix} 1 \\ 1 \end{pmatrix} \tag{106}$$

$$\text{Var}(dP_{s,t})/dt = \frac{1}{4}(-1,1)\mathbf{\Sigma}_P \begin{pmatrix} -1 \\ 1 \end{pmatrix} = \frac{1}{4}(-1,1)\mathbf{p}\mathbf{\Sigma}\mathbf{p}^\top \begin{pmatrix} -1 \\ 1 \end{pmatrix} \tag{107}$$

Since the two markets are symmetric, we denote $\sigma_C = \sigma_S = \sigma_\beta$. Given the shock correlations are zero, we have

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_\beta^2 & 0 & -\sigma_\beta^2 & 0 & 0 \\ 0 & \sigma_\beta^2 & 0 & -\sigma_\beta^2 & 0 \\ -\sigma_\beta^2 & 0 & \sigma_\beta^2 & 0 & 0 \\ 0 & -\sigma_\beta^2 & 0 & \sigma_\beta^2 & 0 \\ 0 & 0 & 0 & 0 & \sigma_V^2 \end{pmatrix} \tag{108}$$

For convenience, we define $\mathbf{E} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}$. Given the symmetry of the two markets,

$$\lambda_{C,x_C} = \lambda_{S,x_S} \quad \lambda_{C,x_S} = \lambda_{S,x_C} \tag{109}$$

$$\lambda_{C,\beta_C} = \lambda_{S,\beta_S} \quad \lambda_{C,\beta_S} = \lambda_{S,\beta_C} \tag{110}$$

$$\lambda_{C,V} = \lambda_{S,V} \tag{111}$$

Hence $\mathbf{E}^\top \boldsymbol{\lambda}_x \mathbf{E}$ and $\mathbf{E}^\top \boldsymbol{\lambda}_\beta \mathbf{E}$ are diagonal matrices, we define

$$\begin{pmatrix} l_x & 0 \\ 0 & s_x \end{pmatrix} \equiv \mathbf{E}^\top \boldsymbol{\lambda}_x \mathbf{E}, \qquad \begin{pmatrix} M_{l,\infty} & 0 \\ 0 & M_{s,\infty} \end{pmatrix} \equiv \mathbf{E}^\top \boldsymbol{\lambda}_\beta \mathbf{E} \tag{112}$$

Recall that Proposition 1 gives $\boldsymbol{\lambda}_\beta = \boldsymbol{\zeta}^{-1}$ and symmetry implies that $\boldsymbol{\zeta}^{-1}$ is diagonalized by the level/spread basis $\mathbf{E}$, so the entries $M_{l,\infty}$ and $M_{s,\infty}$ are exactly the long-run level and spread price impacts. Define the corresponding on-impact price impacts by

$$M_{l,0} \equiv M_{l,\infty} - l_x, \qquad M_{s,0} \equiv M_{s,\infty} - s_x. \tag{113}$$

Hence we can write

$$d \begin{pmatrix} P_{l,t} \\ P_{s,t} \end{pmatrix} = \frac{1}{\sqrt{2}} \mathbf{E}^\top \mathbf{p} \, d\mathbf{s}_t = \frac{1}{2} \begin{pmatrix} l_x & l_x \\ -s_x & s_x \end{pmatrix} d\mathbf{x}_t + \frac{1}{2} \begin{pmatrix} M_{l,\infty} & M_{l,\infty} \\ -M_{s,\infty} & M_{s,\infty} \end{pmatrix} d\boldsymbol{\beta}_t + \frac{1}{2} \begin{pmatrix} 2 \\ 0 \end{pmatrix} dV_t \tag{114}$$

$$\mathrm{Var}(d \begin{pmatrix} P_{l,t} \\ P_{s,t} \end{pmatrix})/dt = \frac{1}{4} \begin{pmatrix} l_x & l_x & M_{l,\infty} & M_{l,\infty} & 2 \\ -s_x & s_x & -M_{s,\infty} & M_{s,\infty} & 0 \end{pmatrix} \mathbf{\Sigma} \begin{pmatrix} l_x & -s_x \\ l_x & s_x \\ M_{l,\infty} & -M_{s,\infty} \\ M_{l,\infty} & M_{s,\infty} \\ 2 & 0 \end{pmatrix} \tag{115}$$

To expand, first compute $\mathbf{\Sigma}$ times each column of the right matrix. For column 1:

$$\mathbf{\Sigma}\begin{pmatrix} l_x \\ l_x \\ M_{l,\infty} \\ M_{l,\infty} \\ 2 \end{pmatrix} = -\sigma_\beta^2 M_{l,0}\begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 2\sigma_V^2 \end{pmatrix} \tag{116}$$

Dotting with each row of the left matrix:

$$\begin{aligned} \text{Row 1:} \quad & -2l_x\sigma_\beta^2 M_{l,0} + 2M_{l,\infty}\sigma_\beta^2 M_{l,0} + 4\sigma_V^2 = 2\sigma_\beta^2 M_{l,0}^2 + 4\sigma_V^2 \\ \text{Row 2:} \quad & s_x\sigma_\beta^2 M_{l,0} - s_x\sigma_\beta^2 M_{l,0} + M_{s,\infty}\sigma_\beta^2 M_{l,0} - M_{s,\infty}\sigma_\beta^2 M_{l,0} = 0 \end{aligned}$$

For column 2:

$$\mathbf{\Sigma}\begin{pmatrix} -s_x \\ s_x \\ -M_{s,\infty} \\ M_{s,\infty} \\ 0 \end{pmatrix} = \sigma_\beta^2 M_{s,0}\begin{pmatrix} 1 \\ -1 \\ -1 \\ 1 \\ 0 \end{pmatrix} \tag{117}$$

Dotting with each row:

$$\begin{aligned} \text{Row 1:} \quad & l_x\sigma_\beta^2 M_{s,0} - l_x\sigma_\beta^2 M_{s,0} - M_{l,\infty}\sigma_\beta^2 M_{s,0} + M_{l,\infty}\sigma_\beta^2 M_{s,0} = 0 \\ \text{Row 2:} \quad & -2s_x\sigma_\beta^2 M_{s,0} + 2M_{s,\infty}\sigma_\beta^2 M_{s,0} = 2\sigma_\beta^2 M_{s,0}^2 \end{aligned}$$

Combining with the $\frac{1}{4}$ prefactor, the covariance matrix is diagonal:

$$\text{Var}\left(d\begin{pmatrix} P_{l,t} \\ P_{s,t} \end{pmatrix}\right)/dt = \frac{1}{2}\begin{pmatrix} \sigma_\beta^2 M_{l,0}^2 + 2\sigma_V^2 & 0 \\ 0 & \sigma_\beta^2 M_{s,0}^2 \end{pmatrix} \tag{118}$$

Level and spread returns are instantaneously uncorrelated. The variances are

$$\text{Var}(dP_{l,t})/dt = \frac{1}{2}\sigma_\beta^2 M_{l,0}^2 + \sigma_V^2 \tag{119}$$

$$\text{Var}(dP_{s,t})/dt = \frac{1}{2}\sigma_\beta^2 M_{s,0}^2 \tag{120}$$

These variance formulas correspond to the average-price and half-spread objects $(P_{l,t}, P_{s,t})$. To solve for (l_x, s_x), however, it is more convenient to rotate (57) by the orthonormal basis $\mathbf{E}$ itself. That rotated fixed-point system is written for the level and spread eigen-components, so no additional $\frac{1}{2}$ enters the Riccati equations below. To solve for l_x and s_x, we multiply both sides of (57) by $\mathbf{E}$

$$-\begin{pmatrix} l_x & 0 \\ 0 & s_x \end{pmatrix} k(\mathbf{I} - \tilde{\boldsymbol{\zeta}} \begin{pmatrix} l_x & 0 \\ 0 & s_x \end{pmatrix}) = \begin{pmatrix} \psi & 0 \\ 0 & \psi \end{pmatrix} + \gamma \begin{pmatrix} l_x & l_x & M_{l,\infty} & M_{l,\infty} & 2 \\ -s_x & s_x & -M_{s,\infty} & M_{s,\infty} & 0 \end{pmatrix} \boldsymbol{\Sigma} \begin{pmatrix} l_x & -s_x \\ l_x & s_x \\ M_{l,\infty} & -M_{s,\infty} \\ M_{l,\infty} & M_{s,\infty} \\ 2 & 0 \end{pmatrix} \tag{121}$$

$$-\begin{pmatrix} l_x & 0 \\ 0 & s_x \end{pmatrix} k(\mathbf{I} - \tilde{\boldsymbol{\zeta}} \begin{pmatrix} l_x & 0 \\ 0 & s_x \end{pmatrix}) = \begin{pmatrix} \psi & 0 \\ 0 & \psi \end{pmatrix} + \gamma \begin{pmatrix} \sigma_\beta^2 M_{l,0}^2 + 2\sigma_V^2 & 0 \\ 0 & \sigma_\beta^2 M_{s,0}^2 \end{pmatrix} \tag{122}$$

where $\tilde{\boldsymbol{\zeta}}$ is also diagonal and is equal to

$$\tilde{\boldsymbol{\zeta}} = \begin{pmatrix} \frac{1}{M_{l,\infty}} & 0 \\ 0 & \frac{1}{M_{s,\infty}} \end{pmatrix} \tag{123}$$

Hence l_x and s_x are decoupled and are defined by the following equations explicitly,

$$k \frac{l_x}{M_{l,\infty}} (l_x - M_{l,\infty}) = \psi + \gamma \left(\sigma_\beta^2 M_{l,0}^2 + 2\sigma_V^2 \right) \tag{124}$$

$$k \frac{s_x}{M_{s,\infty}} (s_x - M_{s,\infty}) = \psi + \gamma \sigma_\beta^2 M_{s,0}^2 \tag{125}$$

define the risk adjusted factors as

$$\phi_l = 1 - \frac{\gamma \sigma_\beta^2 M_{l,\infty}}{k} \qquad \phi_s = 1 - \frac{\gamma \sigma_\beta^2 M_{s,\infty}}{k} \tag{126}$$

The two equations above can be written as

$$\phi_l M_{l,0}^2 - M_{l,\infty} M_{l,0} - (\psi + 2\gamma\sigma_V^2) \frac{M_{l,\infty}}{k} = 0 \tag{127}$$

$$\phi_s M_{s,0}^2 - M_{s,\infty} M_{s,0} - \psi \frac{M_{s,\infty}}{k} = 0 \tag{128}$$

Hence

$$M_{l,0} = \frac{M_{l,\infty}(1 + \sqrt{1 + \frac{4\phi_l}{kM_{l,\infty}}(\psi + 2\gamma\sigma_V^2)})}{2\phi_l} \tag{129}$$

$$M_{s,0} = \frac{M_{s,\infty}(1 + \sqrt{1 + \frac{4\phi_s}{kM_{s,\infty}}\psi})}{2\phi_s} \tag{130}$$

To see how ψ affects the volatilities, define the instantaneous level and spread variances

$$V_{s,0} \equiv \frac{\mathrm{Var}(dP_{s,t})}{dt} = \frac{1}{2}\sigma_\beta^2 M_{s,0}^2, \tag{131}$$

$$V_{l,0} \equiv \frac{\mathrm{Var}(dP_{l,t})}{dt} = \frac{1}{2}\sigma_\beta^2 M_{l,0}^2 + \sigma_V^2. \tag{132}$$

the ψ-derivatives are

$$\frac{\partial V_{s,0}}{\partial \psi} = \sigma_\beta^2 M_{s,0} \frac{\partial M_{s,0}}{\partial \psi} \tag{133}$$

$$\frac{\partial V_{l,0}}{\partial \psi} = \sigma_\beta^2 M_{l,0} \frac{\partial M_{l,0}}{\partial \psi} \tag{134}$$

From (129) and (130), it is straightforward to see that

$$\frac{\partial M_{s,0}}{\partial \psi} > 0 \qquad \frac{\partial M_{l,0}}{\partial \psi} > 0 \tag{135}$$

Higher capacity cost ψ therefore raises both level and spread variance.

Similarly

$$\frac{\partial V_{s,0}}{\partial \gamma} = \sigma_\beta^2 M_{s,0} \frac{\partial M_{s,0}}{\partial \gamma} \tag{136}$$

$$\frac{\partial V_{l,0}}{\partial \gamma} = \sigma_\beta^2 M_{l,0} \frac{\partial M_{l,0}}{\partial \gamma} \tag{137}$$

Plug in $\frac{\partial M_{s,0}}{\partial \gamma}$ and $\frac{\partial M_{l,0}}{\partial \gamma}$ we have

$$\frac{\partial V_{s,0}}{\partial \gamma} = \frac{\sigma_\beta^4 M_{s,0}^3}{k\sqrt{1 + \frac{4\phi_s}{kM_{s,\infty}}\psi}} > 0 \tag{138}$$

$$\frac{\partial V_{l,0}}{\partial \gamma} = \frac{\sigma_\beta^2 M_{l,0}(\sigma_\beta^2 M_{l,0}^2 + 2\sigma_V^2)}{k\sqrt{1 + \frac{4\phi_l}{kM_{l,\infty}}(\psi + 2\gamma\sigma_V^2)}} > 0 \tag{139}$$

Hence higher risk aversion γ also raises both level and spread variance.

A.5 Proof of Proposition 3

The variance ratio of spread to level is given by

$$R \equiv \frac{V_{s,0}}{V_{l,0}} = \frac{\sigma_\beta^2 M_{s,0}^2}{\sigma_\beta^2 M_{l,0}^2 + 2\sigma_V^2} \tag{140}$$

Define $M_{a,0}(a, \tilde{\psi})$ as the positive root to $\phi(a)M_{a,0}^2 - aM_{a,0} - \tilde{\psi}a/k = 0$, so that $M_{l,0} = M_{l,0}(M_{l,\infty}, \psi + 2\gamma\sigma_V^2)$ and $M_{s,0} = M_{s,0}(M_{s,\infty}, \psi)$. Implicit differentiation gives

$$\frac{\partial M_{a,0}}{\partial \tilde{\psi}} = \frac{1}{kD(a,\tilde{\psi})} > 0, \qquad \frac{\partial M_{a,0}}{\partial a} = -\frac{\phi'(a)M_{a,0}^2 - M_{a,0} - \tilde{\psi}/k}{2\phi(a)M_{a,0} - a} > 0, \tag{141}$$

where $D(a,\tilde{\psi}) \equiv \sqrt{1 + \frac{4\phi(a)}{ka}\tilde{\psi}}$, $\phi(a) \equiv 1 - \frac{\gamma\sigma_\beta^2 a}{k}$, and $\phi'(a) = -\frac{\gamma\sigma_\beta^2}{k} < 0$. In the symmetric case with substitutable assets ($\boldsymbol{\zeta}_{12} < 0$), $M_{l,\infty} > M_{s,\infty}$ and the effective cost in the level equation, $\psi + 2\gamma\sigma_V^2$, strictly exceeds ψ. Monotonicity in both arguments therefore implies $M_{l,0} > M_{s,0}$. Combining with (131)–(132),

$$R = \frac{\sigma_\beta^2 M_{s,0}^2}{\sigma_\beta^2 M_{l,0}^2 + 2\sigma_V^2} < \frac{\sigma_\beta^2 M_{l,0}^2}{\sigma_\beta^2 M_{l,0}^2 + 2\sigma_V^2} < 1, \tag{142}$$

with strict inequality whenever $\sigma_V^2 > 0$. Hence $R < 1$.

Define $Q \equiv \frac{M_{s,0}}{M_{l,0}}$, we can write

$$R = \frac{Q^2}{1 + \frac{2\sigma_V^2}{\sigma_\beta^2 M_{l,0}^2}} \tag{143}$$

Taking log and derivative with respect to ψ we have

$$\frac{\partial \log R}{\partial \psi} = 2\frac{\partial \log Q}{\partial \psi} + \frac{4\sigma_V^2}{\sigma_\beta^2 M_{l,0}^2 + 2\sigma_V^2} \underbrace{\frac{\partial \log M_{l,0}}{\partial \psi}}_{>0} \tag{144}$$

Furthermore,

$$\frac{\partial \log Q}{\partial \psi} = \frac{1}{M_{s,0}}\frac{\partial M_{s,0}}{\partial \psi} - \frac{1}{M_{l,0}}\frac{\partial M_{l,0}}{\partial \psi} \tag{145}$$

where

$$\frac{\partial M_{s,0}}{\partial \psi} = \frac{1}{k\sqrt{1+\frac{4\phi_s}{kM_{s,\infty}}\psi}} \tag{146}$$

$$\frac{\partial M_{l,0}}{\partial \psi} = \frac{1}{k\sqrt{1+\frac{4\phi_l}{kM_{l,\infty}}(\psi+2\gamma\sigma_V^2)}} \tag{147}$$

$$\frac{1}{M_{s,0}}\frac{\partial M_{s,0}}{\partial \psi} = \frac{1}{k\sqrt{1+\frac{4\phi_s}{kM_{s,\infty}}\psi}}\frac{2\phi_s}{M_{s,\infty}(1+\sqrt{1+\frac{4\phi_s}{kM_{s,\infty}}\psi})} \tag{148}$$

$$\frac{1}{M_{l,0}}\frac{\partial M_{l,0}}{\partial \psi} = \frac{1}{k\sqrt{1+\frac{4\phi_l}{kM_{l,\infty}}(\psi+2\gamma\sigma_V^2)}}\frac{2\phi_l}{M_{l,\infty}(1+\sqrt{1+\frac{4\phi_l}{kM_{l,\infty}}(\psi+2\gamma\sigma_V^2)})} \tag{149}$$

Define

$$g(a,\tilde{\psi}) \equiv \frac{\phi(a)}{aD(a,\tilde{\psi})(1+D(a,\tilde{\psi}))}, \qquad D(a,\tilde{\psi}) \equiv \sqrt{1+\frac{4\phi(a)}{ka}\tilde{\psi}} \quad \phi(a) \equiv 1-\frac{\gamma\sigma_\beta^2 a}{k} \tag{150}$$

Given $D(a,\tilde{\psi})$ is increasing in $\tilde{\psi}$, we have $g(a,\tilde{\psi})$ is decreasing in $\tilde{\psi}$. Furthermore, $g(a,\tilde{\psi})$ is decreasing in a. Since we have $M_{l,\infty} > M_{s,\infty}$, we have $g(M_{l,\infty}, \psi+2\gamma\sigma_V^2) < g(M_{s,\infty},\psi)$, which implies $\frac{1}{M_{s,0}}\frac{\partial M_{s,0}}{\partial \psi} > \frac{1}{M_{l,0}}\frac{\partial M_{l,0}}{\partial \psi}$. Hence $\frac{\partial \log Q}{\partial \psi} > 0$.

Similarly, taking derivative with respect to γ we have

$$\frac{\partial \log R}{\partial \gamma} = \frac{\partial \log V_{s,0}}{\partial \gamma} - \frac{\partial \log V_{l,0}}{\partial \gamma} \tag{151}$$

where

$$\frac{\partial \log V_{s,0}}{\partial \gamma} = \frac{1}{V_{s,0}}\frac{\partial V_{s,0}}{\partial \gamma} = \frac{2\sigma_\beta^2 M_{s,0}}{k\sqrt{1+\frac{4\phi_s}{kM_{s,\infty}}\psi}} \tag{152}$$

$$\frac{\partial \log V_{l,0}}{\partial \gamma} = \frac{1}{V_{l,0}}\frac{\partial V_{l,0}}{\partial \gamma} = \frac{2\sigma_\beta^2 M_{l,0}}{k\sqrt{1+\frac{4\phi_l}{kM_{l,\infty}}(\psi+2\gamma\sigma_V^2)}} \tag{153}$$

Hence

$$\frac{\partial \log R}{\partial \gamma} = \frac{2\sigma_\beta^2}{k}\left(\frac{M_{s,0}}{\sqrt{1+\frac{4\phi_s}{kM_{s,\infty}}\psi}} - \frac{M_{l,0}}{\sqrt{1+\frac{4\phi_l}{kM_{l,\infty}}(\psi+2\gamma\sigma_V^2)}}\right) \tag{154}$$

$$= \frac{2\sigma_\beta^2}{k}\left(\frac{M_{s,\infty}}{2\phi_s}\left(1+\frac{1}{\sqrt{1+\frac{4\phi_s}{kM_{s,\infty}}\psi}}\right) - \frac{M_{l,\infty}}{2\phi_l}\left(1+\frac{1}{\sqrt{1+\frac{4\phi_l}{kM_{l,\infty}}(\psi+2\gamma\sigma_V^2)}}\right)\right) \tag{155}$$

we have

$$\frac{M_{s,\infty}}{2\phi_s}\left(1+\frac{1}{\sqrt{1+\frac{4\phi_s}{kM_{s,\infty}}\psi}}\right) < \frac{M_{s,\infty}}{\phi_s} \tag{156}$$

and

$$\frac{M_{l,\infty}}{2\phi_l}\left(1+\frac{1}{\sqrt{1+\frac{4\phi_l}{kM_{l,\infty}}(\psi+2\gamma\sigma_V^2)}}\right) > \frac{M_{l,\infty}}{2\phi_l} \tag{157}$$

Under the condition that $\frac{2M_{s,\infty}}{\phi_s} < \frac{M_{l,\infty}}{\phi_l}$, we have $\frac{\partial \log R}{\partial \gamma} < 0$.

A.6 Proof of Corollary 2

By symmetry, $\mathrm{Var}(dP_{C,t}) = \mathrm{Var}(dP_{S,t}) \equiv \sigma_P^2$. The level and spread variances are

$$V_{l,0} = \frac{1}{4}\mathrm{Var}(dP_{C,t} + dP_{S,t}) = \frac{\sigma_P^2(1+\rho)}{2}, \quad V_{s,0} = \frac{1}{4}\mathrm{Var}(dP_{S,t} - dP_{C,t}) = \frac{\sigma_P^2(1-\rho)}{2} \tag{158}$$

where $\rho = \mathrm{Corr}(dP_{C,t}, dP_{S,t})$. Taking the ratio gives

$$R \equiv \frac{V_{s,0}}{V_{l,0}} = \frac{1-\rho}{1+\rho} \quad \Longleftrightarrow \quad \rho = \frac{1-R}{1+R} \tag{159}$$

Since ρ is strictly decreasing in R,

$$\frac{\partial \rho}{\partial \theta} = -\frac{2}{(1+R)^2}\frac{\partial R}{\partial \theta} \tag{160}$$

for any parameter θ. From Proposition 3, $\partial R/\partial \psi > 0$ implies $\partial \rho/\partial \psi < 0$, and $\partial R/\partial \gamma < 0$ (under the sufficient condition) implies $\partial \rho/\partial \gamma > 0$.

A.7 Microfoundation of Gross Position Costs

In this subsection, we show that different forms of balance-sheet costs and constraints lead to an intermediary first-order condition of the same form as in our model,

$$\mathbb{E}_t[d\mathbf{P}_t]/dt = \bar{\boldsymbol{\Psi}} + (\gamma\boldsymbol{\Sigma}_P + \boldsymbol{\Psi})\,\mathbf{x}_t, \tag{161}$$

where $\bar{\boldsymbol{\Psi}}$ is a constant wedge and $\boldsymbol{\Psi}$ encodes the slope of the marginal position cost. Throughout we write $\boldsymbol{\mu}_t \equiv \mathbb{E}_t[d\mathbf{P}_t]/dt$. Since our estimation identifies the response of $\boldsymbol{\mu}_t$ to demand-induced changes in $\mathbf{x}_t$, the constant wedge is differenced out as it does not contribute to the variations in prices, and the slope $\boldsymbol{\Psi}$ is the object that matters for our estimates. We characterize $\boldsymbol{\Psi}$ for each specification and discuss what it implies for the interpretation of our estimated risk contribution.

In the main text, the gross position cost leads to a diagonal $\boldsymbol{\Psi}$, but in general it need not be. Its off-diagonal entry is the *cross-position cost*: a positive entry means that holding more positive amount of one asset increases the marginal cost of holding positive amount of the other asset. In other words it penalizes same-direction positions, behaving like the risk cost. On the other hand, a negative entry penalizes offsetting positions. When intermediaries hold long-short positions in the two assets, as they do empirically (Samuel G. Hanson, Malkhozov, and Venter, 2024), standard balance-sheet constraints and costs typically make this cross-position cost negative, as the examples below illustrate. Intuitively, when an intermediary is constrained in its total gross position, or faces a convex gross-position cost, a larger position in the long leg makes the short leg more costly and pushes the intermediary to cut the short position.

The sign of the cross-position cost highlights the distinction between the gross-position cost and the risk cost: the risk cost makes the two assets substitutes and so encourages long-short positions, whereas a negative cross-position cost makes them complements. When the true model features negative cross-position costs while we decompose costs under the diagonal $\boldsymbol{\Psi}$ as in the main text, our estimate of the risk cost's contribution is a lower bound.

Below, we first show how the bias depends on the sign of the cross-position cost, and then show that the signs for several commonly used constraints and costs implies our estimation of the risk contribution is a lower bound.

Bias from imposing a diagonal non-risk slope. The estimation procedure of Section 3.1 identifies the first column of the marginal-cost matrix $\mathbf{C} = \gamma\boldsymbol{\Sigma}_P + \boldsymbol{\Psi}$ from the demand shock to the cash market, and reports the risk contribution $\widehat{C}_{r,l} = \gamma\sigma_C^2/(\psi_C + \gamma\sigma_C^2)$

as the share of the marginal cost attributable to risk rather than to the gross-position term ψ_C. This procedure maintains the diagonal $\mathbf{\Psi}$ specified in the model. We now apply it when the true non-risk slope carries a nonzero cross-position cost, and trace the resulting bias.

To trace the bias, write the true first-column decomposition as

$$\mathbf{C}_{\text{true}}\mathbf{e}_C = \gamma\mathbf{\Sigma}_P\mathbf{e}_C + \begin{pmatrix}\psi_C\\ \psi_{SC}\end{pmatrix}, \tag{162}$$

where $\mathbf{e}_C = (1,0)^\top$, and $\psi_C > 0$ and $\psi_{SC} < 0$ are the own and cross entries of the non-risk slope in the long-short region. The two identified entries are, up to a scaling factor,

$$C_{CC} = \gamma\sigma_C^2 + \psi_C, \qquad C_{SC} = \gamma\sigma_{SC} + \psi_{SC}, \tag{163}$$

with σ_C^2 and σ_{SC} the own and cross entries of $\mathbf{\Sigma}_P$ in the first column. The diagonal procedure attributes the entire cross entry of $\mathbf{C}$ to risk, $C_{SC} = \widehat{\gamma}\sigma_{SC}$, so it infers

$$\widehat{\gamma} = \gamma + \frac{\psi_{SC}}{\sigma_{SC}}. \tag{164}$$

The price covariance of the two assets is positive ($\sigma_{SC} > 0$), so the bias $\widehat{\gamma} - \gamma$ inherits the sign of ψ_{SC}. When $\psi_{SC} < 0$, the procedure understates the risk scale. Correspondingly, the estimation procedure infers the gross-position cost from $\widehat{\psi}_C = C_{CC} - \widehat{\gamma}\sigma_C^2$, overstating the gross-position cost:

$$\widehat{\psi}_C = \psi_C - \frac{\psi_{SC}}{\sigma_{SC}}\sigma_C^2 > \psi_C. \tag{165}$$

Next, we derive $\mathbf{\Psi}$, and in particular the sign of its cross-position entry, for several standard gross-position costs and constraints.

Linear absolute-value gross-position costs. The simplest gross-position friction is a cost proportional to the total gross position,

$$G_{\text{abs}}(\mathbf{x}) = \kappa\left(\widetilde{\psi}_C|x_C| + \widetilde{\psi}_S|x_S|\right), \tag{166}$$

with cost intensity $\kappa > 0$ and positive weights $\widetilde{\psi}_C$ and $\widetilde{\psi}_S$. This cost captures SLR-style regulation, or risk-based capital (RBC) charges when the weights depend on asset risk.

The intermediary's instantaneous problem is

$$\max_{\mathbf{x}\in\mathbb{R}^2} \boldsymbol{\mu}_t^\top \mathbf{x} - \frac{\gamma}{2}\mathbf{x}^\top \boldsymbol{\Sigma}_P \mathbf{x} - G_{\text{abs}}(\mathbf{x}), \tag{167}$$

with first-order condition

$$\boldsymbol{\mu}_t = \gamma \boldsymbol{\Sigma}_P \mathbf{x} + \partial_{\mathbf{x}} G_{\text{abs}}(\mathbf{x}). \tag{168}$$

Within a fixed sign branch, the absolute-value terms are affine. Let $s_i = \text{sgn}(x_i)$; then the first-order condition takes the form (161) with

$$\bar{\boldsymbol{\Psi}} = \partial_{\mathbf{x}} G_{\text{abs}}(\mathbf{x}) = \kappa \begin{pmatrix} s_C \widetilde{\psi}_C \\ s_S \widetilde{\psi}_S \end{pmatrix}, \qquad \boldsymbol{\Psi} = \partial^2_{\mathbf{x}} G_{\text{abs}}(\mathbf{x}) = \mathbf{0}. \tag{169}$$

The linear absolute-value cost adds a constant wedge between the expected return and the risk compensation. Because the cost is linear, the wedge does not grow with the size of the position: it shifts the intercept of the first-order condition but contributes nothing to $\boldsymbol{\Psi}$ and so does not affect our estimates.

Quadratic costs on total gross positions. By contrast, consider a quadratic cost in the same total gross position,

$$G_{\text{quad}}(\mathbf{x}) = \frac{\kappa}{2}\left(\widetilde{\psi}_C |x_C| + \widetilde{\psi}_S |x_S|\right)^2. \tag{170}$$

This cost differs from the gross-position cost in the main model: there each position carries its own marginal cost, whereas here the two positions share a common marginal cost through the total gross position. As we show below, the shared marginal cost discourages long-short positions even further.

Replacing G_{abs} with G_{quad} in (167), the first-order condition within a fixed sign branch becomes

$$\boldsymbol{\mu}_t = \gamma \boldsymbol{\Sigma}_P \mathbf{x} + \kappa \left(\widetilde{\psi}_C |x_C| + \widetilde{\psi}_S |x_S|\right) \begin{pmatrix} s_C \widetilde{\psi}_C \\ s_S \widetilde{\psi}_S \end{pmatrix}, \tag{171}$$

so the marginal cost of each position scales with the total gross position. Differentiating the

marginal cost with respect to $\mathbf{x}$ gives the slope

$$\boldsymbol{\Psi} = \partial^2_{\mathbf{x}} G_{\text{quad}}(\mathbf{x}) = \kappa \begin{pmatrix} \widetilde{\psi}_C^2 & s_C s_S \widetilde{\psi}_C \widetilde{\psi}_S \\ s_C s_S \widetilde{\psi}_C \widetilde{\psi}_S & \widetilde{\psi}_S^2 \end{pmatrix}. \tag{172}$$

When the intermediary takes long-short positions, which is the empirically relevant case, the cross-position cost $\kappa s_C s_S \widetilde{\psi}_C \widetilde{\psi}_S$ is negative: increasing the long leg raises the total gross position and with it the marginal cost of expanding the short leg. In this case, our risk contribution measure is a lower bound on the true risk contribution.

Hard capacity constraints. The examples above are formulated as costs; a hard capacity constraint delivers the same economics as the limit of a convex capacity cost as its curvature becomes large. In practice, asset prices enter the constraint and adjust endogenously to clear the market. For example, margin and leverage constraints (Gârleanu and Pedersen, 2011; He and Krishnamurthy, 2013) restrict the market value of positions rather than the number of units. Hence, we consider a gross-dollar-exposure constraint of the form

$$K_{\text{abs}}(\mathbf{x}; \mathbf{P}) = \widetilde{\psi}_C |P_C x_C| + \widetilde{\psi}_S |P_S x_S| - \bar{K} \leq 0, \tag{173}$$

with positive per-dollar weights and capacity $\bar{K}$. The intermediary solves

$$\max_{\mathbf{x} \in \mathbb{R}^2} \boldsymbol{\mu}_t^\top \mathbf{x} - \frac{\gamma}{2} \mathbf{x}^\top \boldsymbol{\Sigma}_P \mathbf{x} \quad \text{subject to} \quad K_{\text{abs}}(\mathbf{x}; \mathbf{P}) \leq 0, \tag{174}$$

with first-order condition

$$\boldsymbol{\mu}_t = \gamma \boldsymbol{\Sigma}_P \mathbf{x} + \lambda \partial_{\mathbf{x}} K_{\text{abs}}, \tag{175}$$

where $\lambda \geq 0$ is the multiplier on the capacity constraint, equal to zero when the constraint is slack. When the constraint is binding, the non-risk wedge is $\lambda \partial_{\mathbf{x}} K_{\text{abs}}$.

In the same spirit as our empirical strategy, we linearize the first-order condition along the equilibrium path of a demand shock, which is the response our estimation identifies. Along this path, the intermediary absorbs the entire shock on impact and prices respond through the on-impact price-impact matrix $\mathbf{M}_0$:

$$d\mathbf{x} = -d\boldsymbol{\beta}, \qquad d\mathbf{P} = \mathbf{M}_0 d\boldsymbol{\beta}. \tag{176}$$

The constraint stays binding along the path,

$$\partial_{\mathbf{x}} K_{\text{abs}}^{\top} d\mathbf{x} + \partial_{\mathbf{P}} K_{\text{abs}}^{\top} d\mathbf{P} = 0, \tag{177}$$

so the price response frees exactly the dollar capacity that the forced position change consumes—the formal sense in which the price is the margin that adjusts. Totally differentiating the first-order condition (175) along the path, and using $\partial_{\mathbf{x}}^2 K_{\text{abs}} = \mathbf{0}$ within a fixed sign branch, gives the response of $\boldsymbol{\mu}_t$ to the absorbed position change:

$$d\boldsymbol{\mu}_t = (\gamma \boldsymbol{\Sigma}_P + \boldsymbol{\Psi})\, d\mathbf{x}, \qquad \boldsymbol{\Psi} = \partial_{\mathbf{x}} K_{\text{abs}} \frac{d\lambda}{d(-\boldsymbol{\beta})^{\top}} - \lambda \partial_{\mathbf{xP}}^2 K_{\text{abs}} \mathbf{M}_0. \tag{178}$$

The first term is the shadow-price channel: the multiplier changes as the intermediary absorbs demand pressure. The second term is the price-base channel: prices revalue how much capacity each unit position uses.

Fully characterizing the entries of $\boldsymbol{\Psi}$ requires solving the equilibrium model under this constraint, which goes beyond this illustrative example. Signing the cross-position cost, however, requires only three weak assumptions in equilibrium: (i) positive own price impact, $M_{0,CC} > 0$, i.e., a demand shock moves the price of the market it hits in its own direction; (ii) a nonnegative own-market wedge response, $\psi_C \geq 0$, i.e., absorbing demand pressure in the cash market does not lower the constraint's marginal wedge $\lambda(\partial_{\mathbf{x}} K_{\text{abs}})_C$ on the cash position—equivalently, the own-market cost response exceeds the risk-based compensation, $C_{CC} \geq \gamma \sigma_C^2$, consistent with our estimate of a positive own gross-position cost; and (iii) return-impact diagonal dominance, $M_{0,CC}/P_C \geq M_{0,SC}/P_S$, i.e., demand pressure in the cash market moves the cash price at least as much as it moves the synthetic price.

For concreteness, consider a negative demand shock to the cash asset, $d\beta_C < 0$, which the intermediary absorbs by taking on the position $dx_C = -d\beta_C > 0$. We work in the empirically relevant binding branch where the intermediary is long the cash asset and short the synthetic asset, $x_C > 0$ and $x_S < 0$, so that for positive prices $\partial_{\mathbf{x}} K_{\text{abs}} = (\widetilde{\psi}_C P_C, -\widetilde{\psi}_S P_S)^{\top}$ and $\partial_{\mathbf{xP}}^2 K_{\text{abs}} = \text{diag}(\widetilde{\psi}_C, -\widetilde{\psi}_S)$. The cash-shock column of (178) then has entries

$$\psi_C = \widetilde{\psi}_C \left(P_C \frac{d\lambda}{d(-\beta_C)} - \lambda M_{0,CC} \right), \qquad \psi_{SC} = \widetilde{\psi}_S \left(\lambda M_{0,SC} - P_S \frac{d\lambda}{d(-\beta_C)} \right). \tag{179}$$

The own entry ties the multiplier response to observables: rearranging it gives

$$\frac{d\lambda}{d(-\beta_C)} = \frac{\psi_C}{\widetilde{\psi}_C P_C} + \frac{\lambda M_{0,CC}}{P_C} \geq \frac{\lambda M_{0,CC}}{P_C} > 0, \tag{180}$$

where the first inequality uses assumption (ii) and the second uses assumption (i) with $\lambda > 0$ on the binding branch. Absorbing demand pressure therefore raises the shadow cost of balance-sheet capacity. Substituting this lower bound into the cross entry of (179),

$$\psi_{SC} \leq \lambda \widetilde{\psi}_S \left(M_{0,SC} - M_{0,CC} \frac{P_S}{P_C} \right) \leq 0, \tag{181}$$

where the second inequality uses the diagonal dominance in assumption (iii). Under these assumptions, the hard market-value constraint generates a cross-position cost when the intermediary holds long-short positions.

A.8 Intermediaries with Hedging Demand

This subsection extends the baseline intermediary problem to allow hedging demand. We do not characterize the full equilibrium with exact hedging demand: as the derivation below shows, exact hedging demand breaks the affine state-variable structure of the model and makes the global equilibrium intractable. Instead, the goal is to understand how hedging demand fits into our existing framework and how it affects the interpretation of risk costs versus gross-position costs.

To introduce intertemporal hedging demand while preserving the rest of the model, let intermediaries be perpetual-youth CARA cohorts that die at Poisson intensity $\delta > 0$. The death rate δ serves as the effective discount rate and controls how forward-looking intermediaries are.

A living cohort chooses its own position $\tilde{\mathbf{x}}_t = (\tilde{x}_{C,t}, \tilde{x}_{S,t})^\top$ while taking the aggregate state

$$\mathbf{s}_t = (x_{C,t}, x_{S,t}, \beta_{C,t}, \beta_{S,t}, V_t)^\top \tag{182}$$

as given. The cohort has terminal payoff $-\exp(-\gamma W)$ at death and no intermediate consumption. With $r = 0$, wealth evolves as

$$dW_t = \tilde{\mathbf{x}}_t^\top d\mathbf{P}_t - \frac{1}{2} \tilde{\mathbf{x}}_t^\top \boldsymbol{\Psi} \tilde{\mathbf{x}}_t dt, \qquad \boldsymbol{\Psi} \equiv \operatorname{diag}(\psi_C, \psi_S). \tag{183}$$

Let $\boldsymbol{\mu}(\mathbf{s}) = \mathbb{E}_t[d\mathbf{P}_t]/dt$ and let $\boldsymbol{\mu}_s(\mathbf{s}) = \mathbb{E}_t[d\mathbf{s}_t]/dt$. The stationary HJB is

$$0 = \max_{\tilde{\mathbf{x}}} \left\{ \left(\tilde{\mathbf{x}}^\top \boldsymbol{\mu}(\mathbf{s}) - \frac{1}{2} \tilde{\mathbf{x}}^\top \boldsymbol{\Psi} \tilde{\mathbf{x}} \right) \partial_W J + \boldsymbol{\mu}_s(\mathbf{s})^\top \partial_{\mathbf{s}} J \right.$$
$$\left. + \frac{1}{2} \tilde{\mathbf{x}}^\top \boldsymbol{\Sigma}_P \tilde{\mathbf{x}} \partial_{WW} J + \tilde{\mathbf{x}}^\top \boldsymbol{\Sigma}_{Ps} \partial_{W\mathbf{s}} J + \frac{1}{2} \operatorname{tr}\left(\boldsymbol{\Sigma} \partial^2_{\mathbf{ss}} J \right) + \delta \left[-\exp(-\gamma W) - J(W, \mathbf{s}) \right] \right\}. \tag{184}$$

Guess that the intermediary's value function takes the form

$$J(W, \mathbf{s}) = -\exp(-\gamma W - g(\mathbf{s})). \tag{185}$$

The function $g(\mathbf{s})$ captures how the current aggregate state scales continuation value. At death, the payoff sets the continuation component to zero; before death, higher $g(\mathbf{s})$ raises value and therefore represents better investment opportunities.

The two derivatives that enter the portfolio choice are

$$-\frac{\partial_{WW} J}{\partial_W J} = \gamma, \qquad -\frac{\partial_{W\mathbf{s}} J}{\partial_W J} = \partial_{\mathbf{s}} g(\mathbf{s}). \tag{186}$$

Thus the coefficient on instantaneous return risk remains exactly γ; the finite lifetime affects demand only through the state-gradient term.

Define the price covariance and price-state covariance as

$$\boldsymbol{\Sigma}_P \equiv \frac{\operatorname{Cov}_t(d\mathbf{P}_t)}{dt} = \mathbf{p} \boldsymbol{\Sigma} \mathbf{p}^\top, \qquad \boldsymbol{\Sigma}_{Ps} \equiv \frac{\operatorname{Cov}_t(d\mathbf{P}_t, d\mathbf{s}_t)}{dt} = \mathbf{p} \boldsymbol{\Sigma}. \tag{187}$$

The first-order condition with respect to $\tilde{\mathbf{x}}$ yields

$$\frac{\mathbb{E}_t[d\mathbf{P}_t]}{dt} = (\gamma \boldsymbol{\Sigma}_P + \boldsymbol{\Psi})\, \tilde{\mathbf{x}}_t + \mathbf{h}(\mathbf{s}_t), \tag{188}$$

where $\mathbf{h}(\mathbf{s})$ captures hedging demand:

$$\mathbf{h}(\mathbf{s}) \equiv \boldsymbol{\Sigma}_{Ps} \partial_{\mathbf{s}} g(\mathbf{s}). \tag{189}$$

Equation (188) is the main model's intermediary FOC plus an intertemporal hedge wedge. This wedge compensates intermediaries for holding assets whose price innovations covary with good states.

To characterize g, substitute the optimal cohort policy into the HJB. With

$$\mathbf{C} \equiv \mathbf{\Psi} + \gamma \mathbf{\Sigma}_P, \tag{190}$$

the exponent g solves

$$\begin{aligned} 0 =& \boldsymbol{\mu}_s(\mathbf{s})^\top \partial_{\mathbf{s}} g - \frac{1}{2} \partial_{\mathbf{s}} g^\top \mathbf{\Sigma} \partial_{\mathbf{s}} g + \frac{1}{2} \operatorname{tr}\left(\mathbf{\Sigma} \partial^2_{\mathbf{ss}} g\right) + \delta(1 - \exp(g)) \\ &+ \gamma \left[(\tilde{\mathbf{x}}^\star)^\top \left(\boldsymbol{\mu} - \mathbf{\Sigma}_{Ps} \partial_{\mathbf{s}} g \right) - \frac{1}{2} (\tilde{\mathbf{x}}^\star)^\top \mathbf{C} \tilde{\mathbf{x}}^\star \right]. \end{aligned} \tag{191}$$

The bracketed term is the optimized net flow value: expected return compensation, net of the hedge covariance term, minus the instantaneous risk and balance-sheet cost.

Since CARA preferences and $r = 0$ make positions independent of wealth, all living cohorts choose the same $\tilde{\mathbf{x}}^\star(\mathbf{s})$; with survivor mass normalized to one, market clearing imposes $\tilde{\mathbf{x}}^\star_t = \mathbf{x}_t$. Hence in equilibrium,

$$\mathbf{x} = \tilde{\mathbf{x}}^\star = \mathbf{C}^{-1} \left(\boldsymbol{\mu} - \mathbf{\Sigma}_{Ps} \partial_{\mathbf{s}} g \right), \tag{192}$$

and the optimized net flow value in (191) collapses to

$$\gamma \left[(\tilde{\mathbf{x}}^\star)^\top \left(\boldsymbol{\mu} - \mathbf{\Sigma}_{Ps} \partial_{\mathbf{s}} g \right) - \frac{1}{2} (\tilde{\mathbf{x}}^\star)^\top \mathbf{C} \tilde{\mathbf{x}}^\star \right] = \frac{\gamma}{2} \mathbf{x}^\top \mathbf{C} \mathbf{x}. \tag{193}$$

That is, the optimized flow value is exactly $\gamma/2$ times the instantaneous holding cost. Hence (191) yields a linear equation in $\exp(-g(\mathbf{s}))$:

$$\begin{aligned} &\left(\delta + \frac{\gamma}{2} \mathbf{x}^\top \mathbf{C} \mathbf{x} \right) \exp(-g(\mathbf{s})) - \boldsymbol{\mu}_s(\mathbf{s})^\top \partial_{\mathbf{s}} \exp(-g(\mathbf{s})) \\ &\quad - \frac{1}{2} \operatorname{tr}\left(\mathbf{\Sigma} \partial^2_{\mathbf{ss}} \exp(-g(\mathbf{s})) \right) = \delta. \end{aligned} \tag{194}$$

By Feynman-Kac,

$$\exp(-g(\mathbf{s})) = \delta \int_0^\infty e^{-\delta u} \mathbb{E}_{\mathbf{s}} \left[\exp\left(- \int_0^u \frac{\gamma}{2} \mathbf{x}_v^\top \mathbf{C} \mathbf{x}_v dv \right) \right] du. \tag{195}$$

Equation (195) has an intuitive interpretation. The weight $\delta e^{-\delta u}$ is the density of the cohort death time. Thus the state scaling factor $\exp(-g(\mathbf{s}))$ averages the exponentially cumulated future flow value $\frac{\gamma}{2} \mathbf{x}_v^\top \mathbf{C} \mathbf{x}_v$ over the cohort's lifetime.

Equation (195), however, generally breaks the global affine state-space structure. Because

g is the negative log of a lifetime mixture of exponential-quadratic objects, $\partial_{\mathbf{s}} g(\mathbf{s})$ need not be affine in the state. The global equilibrium condition

$$\mathbf{p}\boldsymbol{\mu}_s(\mathbf{s}) = \mathbf{C}\mathbf{x} + \boldsymbol{\Sigma}_{Ps}\partial_{\mathbf{s}} g(\mathbf{s}) \tag{196}$$

therefore does not generally preserve the draft's affine-price OU class.

We therefore analyze hedging demand locally around the myopic benchmark of large δ, expanding (195) directly. Around $u = 0$, the inner expectation is

$$\mathbb{E}_{\mathbf{s}}\left[\exp\left(-\int_0^u \frac{\gamma}{2}\mathbf{x}_v^\top \mathbf{C}\mathbf{x}_v dv\right)\right] = 1 - \frac{\gamma}{2}\mathbf{x}^\top \mathbf{C}\mathbf{x}u + O(u^2). \tag{197}$$

Integrating against the death-time density, with $\delta\int_0^\infty e^{-\delta u}u du = \delta^{-1}$, gives $\exp(-g(\mathbf{s})) = 1 - \frac{\gamma}{2\delta}\mathbf{x}^\top\mathbf{C}\mathbf{x} + O(\delta^{-2})$, i.e.,

$$g(\mathbf{s}) = \frac{\gamma}{2\delta}\mathbf{x}^\top \mathbf{C}\mathbf{x} + O(\delta^{-2}). \tag{198}$$

To leading order, the value scaling factor equals the instantaneous flow value $\frac{\gamma}{2}\mathbf{x}^\top\mathbf{C}\mathbf{x}$ accrued over the cohort's expected remaining lifetime δ^{-1}. Consequently,

$$\partial_{\mathbf{s}} g(\mathbf{s}) = \frac{\gamma}{\delta}\begin{pmatrix}\mathbf{C}\mathbf{x}\\ \mathbf{0}_{3\times 1}\end{pmatrix} + O(\delta^{-2}). \tag{199}$$

Let $\boldsymbol{\Sigma}_{Px}$ denote the first two columns of $\boldsymbol{\Sigma}_{Ps}$, so that it is the covariance between price innovations and the inventory-state innovations. Then the leading hedge wedge is

$$\mathbf{h}(\mathbf{s}) = \frac{\gamma}{\delta}\boldsymbol{\Sigma}_{Px}\mathbf{C}\mathbf{x} + O(\delta^{-2}). \tag{200}$$

The intermediary FOC with hedging demand can therefore be written, to leading order, as

$$\boldsymbol{\mu}(\mathbf{s}) = \underbrace{\left(\mathbf{I} + \frac{\gamma}{\delta}\boldsymbol{\Sigma}_{Px}\right)\mathbf{C}}_{\mathbf{C}^{\text{dyn.}}}\mathbf{x} + O(\delta^{-2}). \tag{201}$$

That is, to leading order intermediaries trade as if they faced a dynamic cost matrix $\mathbf{C}^{\text{dyn.}}$: the instantaneous cost matrix adjusted for the comovement of each asset's price with the inventory state.

Hedging demand in a symmetric model. To gain intuition into how hedging demand augments the instantaneous cost matrix, we consider the symmetric benchmark of Proposition 2, generalized so that the demand shocks to assets C and S have correlation ρ_β while the fundamental shock remains orthogonal to demand shocks. We rotate the two assets into the market portfolio l and the spread portfolio s, with portfolio weights $\frac{1}{\sqrt{2}}(1,1)^\top$ and $\frac{1}{\sqrt{2}}(1,-1)^\top$—the level/spread decomposition in (19) rescaled to make the rotation orthonormal. Let σ_l^2 and σ_s^2 denote the price variances of the two portfolios, and let $M_{l,0}$ and $M_{s,0}$ denote the on-impact price impacts of the level and spread components, as defined in Appendix A.4. The portfolio variances are

$$\sigma_l^2 = (1+\rho_\beta)\sigma_\beta^2 M_{l,0}^2 + 2\sigma_V^2, \qquad \sigma_s^2 = (1-\rho_\beta)\sigma_\beta^2 M_{s,0}^2. \tag{202}$$

We use tildes to denote objects in this rotated basis; in particular, $\tilde{\mathbf{C}}$ and $\tilde{\mathbf{\Sigma}}_{Px}$ are the level/spread representations of $\mathbf{C}$ and $\mathbf{\Sigma}_{Px}$. Because the rotation is orthonormal and $\mathbf{\Psi} = \psi\mathbf{I}$ in the symmetric model, rotating $\mathbf{C} = \mathbf{\Psi} + \gamma\mathbf{\Sigma}_P$ gives

$$\tilde{\mathbf{C}} = \begin{pmatrix} \psi + \gamma\sigma_l^2 & 0 \\ 0 & \psi + \gamma\sigma_s^2 \end{pmatrix}, \tag{203}$$

which reduces to the effective costs in the proof of Proposition 2 when $\rho_\beta = 0$.

The hedge adjustment works through the covariance between price innovations and inventory innovations. Since the diffusion part of $d\mathbf{x}_t$ is $-d\boldsymbol{\beta}_t$, this covariance is negative, and symmetry makes it diagonal in the level/spread basis:

$$\tilde{\mathbf{\Sigma}}_{Px} = -\sigma_\beta^2 \begin{pmatrix} (1+\rho_\beta)M_{l,0} & 0 \\ 0 & (1-\rho_\beta)M_{s,0} \end{pmatrix}. \tag{204}$$

Combining (204) with (203), the dynamic cost matrix is

$$\begin{aligned} \tilde{\mathbf{C}}^{\text{dyn.}} &\equiv \left(\mathbf{I} + \frac{\gamma}{\delta}\tilde{\mathbf{\Sigma}}_{Px}\right)\tilde{\mathbf{C}} \\ &= \begin{pmatrix} 1 - \frac{\gamma}{\delta}\sigma_\beta^2(1+\rho_\beta)M_{l,0} & 0 \\ 0 & 1 - \frac{\gamma}{\delta}\sigma_\beta^2(1-\rho_\beta)M_{s,0} \end{pmatrix} \begin{pmatrix} \psi + \gamma\sigma_l^2 & 0 \\ 0 & \psi + \gamma\sigma_s^2 \end{pmatrix} + O(\delta^{-2}). \end{aligned} \tag{205}$$

Both portfolios provide a hedging benefit: their prices fall precisely when intermediaries take on larger positions, and the depressed prices predict higher future returns and hence better investment opportunities. Hedging demand therefore lowers the contemporaneous

compensation intermediaries require, dampening each portfolio's effective cost. The dampening is stronger the larger the price–inventory covariance: $\sigma_\beta^2(1+\rho_\beta)M_{l,0}$ for the market portfolio and $\sigma_\beta^2(1-\rho_\beta)M_{s,0}$ for the spread portfolio. Under the condition

$$(1+\rho_\beta)M_{l,0} > (1-\rho_\beta)M_{s,0}, \tag{206}$$

the market portfolio offers the larger hedging benefit, and its expected return captures a smaller share of its holding cost.

Let m_l and m_s denote the observed instantaneous decay rates of the market-portfolio and spread price responses—the symmetric-benchmark analogues of (36). Suppose an econometrician infers the risk contribution to the market portfolio's intermediation cost from these decay rates assuming the myopic model, in which the decay rates are proportional to the instantaneous costs. Following (38), she computes

$$\hat{C}_{r,l} = \frac{\sigma_l^2}{\sigma_l^2 - \sigma_s^2}\left(1 - \frac{m_s}{m_l}\right). \tag{207}$$

Without hedging demand ($\delta \to \infty$), the decay rates are proportional to the diagonal entries of $\tilde{\mathbf{C}}$ in (203), so $m_s/m_l = (\psi+\gamma\sigma_s^2)/(\psi+\gamma\sigma_l^2)$ and the estimator recovers the true risk contribution:

$$\hat{C}_{r,l} = \frac{\sigma_l^2}{\sigma_l^2 - \sigma_s^2} \cdot \frac{\gamma(\sigma_l^2 - \sigma_s^2)}{\psi + \gamma\sigma_l^2} = \frac{\gamma\sigma_l^2}{\psi + \gamma\sigma_l^2} \equiv C_{r,l}. \tag{208}$$

With hedging demand, however, the decay rates are proportional to the diagonal entries of the dynamic cost matrix $\tilde{\mathbf{C}}^{\text{dyn.}}$ in (205), so the observed ratio becomes

$$\frac{m_s}{m_l} = \frac{1 - \frac{\gamma}{\delta}\sigma_\beta^2(1-\rho_\beta)M_{s,0}}{1 - \frac{\gamma}{\delta}\sigma_\beta^2(1+\rho_\beta)M_{l,0}} \cdot \frac{\psi + \gamma\sigma_s^2}{\psi + \gamma\sigma_l^2} + O(\delta^{-2}). \tag{209}$$

Under (206), the first factor exceeds one: the market portfolio's cost is dampened more, raising the observed ratio m_s/m_l above its myopic counterpart. Substituting (209) into (207) then gives $\hat{C}_{r,l} < C_{r,l}$. Ignoring hedging demand therefore makes the estimated risk contribution a lower bound on the true one.

B Additional Estimation Results

This appendix collects additional estimation results and robustness exercises for the analysis in Section 3.

B.1 Robustness: Daily Frequency Estimation

We also estimate the model at daily frequency, fitting the decay over three windows that all begin at impact: days 0–25, 0–20, and 0–15. Table 7 reports the risk contribution of the yield response in each case. The risk channel dominates across all three windows.

Table 7: Robustness: Daily frequency estimation

	Risk Contribution	Window
Daily, h=0 to 25	100.0%***	25 days
	[77,100]	
Daily, h=0 to 20	100.0%***	20 days
	[56,100]	
Daily, h=0 to 15	100.0%**	15 days
	[27,100]	

The table reports 90% confidence intervals for the risk contribution of the yield response to Treasury auction demand shocks estimated at daily frequency. The sample spans January 2008 to January 2022, with 165 auction dates. Confidence intervals are 5th and 95th percentiles of the date-cluster bootstrap distribution (1,000 replications). Significance: *** p<0.01, ** p<0.05, * p<0.1.

B.2 Constrained Estimation and Bootstrap Inference

The requirement that risk contributions lie in $[0, 1]$ imposes bounds on the ratio of decay rates:

$$\frac{m_s}{m_l} \in \left[\frac{\sigma_{C,s}}{\sigma_C^2}, 1 \right]. \tag{210}$$

The lower bound equals $1 - \beta_{C,S}$, where $\beta_{C,S}$ is the coefficient from regressing Treasury yield changes on OIS rate changes. Under exact no-arbitrage ($\beta_{C,S} = 1$), the lower bound is zero—the spread does not respond at all. The upper bound of one requires that the spread decays no faster than the level.

When computing the risk contribution estimates $\hat{C}_{r,l}$ and $\hat{C}_{r,s}$ from equations (38)–(39), we impose the theoretical bounds (210) to ensure that the estimated contributions are economically meaningful ($\hat{C}_{r,l}, \hat{C}_{r,s} \in [0,1]$). This requires that the ratio m_s/m_l lie in $[\sigma_{C,s}/\sigma_C^2, 1]$.

Constrained optimization. We first project out control variables (intercepts and lagged dependent variables) from both the dependent variable and shock regressors using OLS, following the FWL theorem. Working with the projected residuals, we solve the constrained least squares problem

$$\min_{\mathbf{b}} \frac{1}{2}\mathbf{b}^\top(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})\mathbf{b} - (\tilde{\mathbf{X}}^\top\tilde{\mathbf{y}})^\top\mathbf{b} \quad \text{subject to} \quad \frac{m_s}{m_l} \in \left[\frac{\sigma_{C,s}}{\sigma_C^2}, 1\right], \tag{211}$$

where $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{y}}$ are the FWL-projected shock regressors and dependent variable, $\mathbf{b}$ is the vector of polynomial coefficients, and $m_l = b_{l,1}$, $m_s = b_{s,1}$ are the linear coefficients. The constraints are implemented as inequality constraints on the elements of $\mathbf{b}$ and solved via sequential quadratic programming (SLSQP).

Bootstrap inference. We use a date-cluster bootstrap for inference: in each of 1,000 replications, we resample trading dates with replacement (preserving the within-date correlation structure), re-estimate the constrained problem on the bootstrap sample, and record the resulting estimates. The standard errors for $\hat{m}_l$ and $\hat{m}_s$ are computed as the standard deviation of the bootstrap distribution. Confidence intervals for the risk contributions $\hat{C}_{r,l}$ and $\hat{C}_{r,s}$ are obtained directly from the bootstrap distribution: in each replication, we compute the contribution from the bootstrapped slope estimates and report the 5th and 95th percentiles.

R^2 loss diagnostic. To assess the cost of imposing the theoretical bounds, we report the R^2 loss: the difference in R^2 between the unconstrained and constrained fits. A small R^2 loss indicates that the constraint is nearly satisfied by the data and the constrained estimates are close to the unrestricted optimum.

Table 8 reports this diagnostic for the three baseline specifications alongside the constrained risk contribution estimates. For the full sample, the constraint binds with an R^2 loss of 2.21% of the unconstrained R^2. For the GFC and COVID subsamples, the constraint does not bind. The small R^2 loss for the full sample reflects the fact that the theoretical bounds depend on the ratio of the spread and level decay slopes. When the spread slope is near zero—as in the full sample, where the spread shows no statistically signifi-

cant response—small estimation noise can push the unconstrained ratio slightly outside the admissible region, but the correction required to restore feasibility is minimal.

Table 8: R^2 loss from imposing theoretical bounds

	Unconstrained		Constrained		ΔR^2 (% of unconstrained)
	Yield	Spread	Yield	Spread	
Full sample	-0.078	0.031	-0.075	-0.008	2.82
	(0.063)	(0.022)	(0.057)	(0.010)	
GFC	-0.112	-0.029	—	—	—
	(0.111)	(0.026)			
COVID	-0.072	-0.049	—	—	—
	(0.116)	(0.111)			

The table reports initial decay slope estimates with and without imposing the theoretical bounds on the ratio of decay rates (equation (210)). The ΔR^2 column reports the change in R^2 as a percentage of the unconstrained R^2. All specifications use a linear decay function. The full sample uses a 3-week estimation window; crisis episodes use a 5-week window. Standard errors use date-cluster bootstrap (1,000 replications). Significance: *** p<0.01, ** p<0.05, * p<0.1.

B.3 Robustness: Assessing Resolution of Uncertainty Channel

An alternative explanation for the observed yield decay following demand shocks is the gradual resolution of uncertainty.[28] After the auction closes, market participants' uncertainty about demand declines as they gradually learn the auction outcome, generating a gradual decline in yields. Under this interpretation, the yield decay would not reflect intermediary costs but rather uncertainty gradually reducing. Crucially, this channel predicts that yields should decline following the auction close *regardless* of the sign of the demand shock. This would mean that we should get different signs for the decay rates following positive and negative demand shocks (negative and positive, respectively).

To assess the plausibility of this channel, we estimate the decay rates separately for positive and negative demand shocks using interaction terms in our baseline specification. Table 9 reports the results. Even though power is reduced because of effectively halving

[28]Resolution of uncertainty must be gradual for it to pose a concern for our identification. In a frictionless market, the resolution of uncertainty would be incorporated instantaneously into prices, affecting the price impact but not the decay rate that we use for identification.

the sample of shocks, the yield decay rates are negative for both positive and negative shocks (-0.101 and -0.081 pp/week, respectively), indicating that the yield impact reverts regardless of the shock's sign. This pattern is inconsistent with the resolution-of-uncertainty channel, but consistent with the intermediary cost interpretation—required compensation for bearing inventory risk generates mean reversion.

Table 9: Robustness: Shock sign interaction

	Initial decay rate (pp/week)	Decay estimation
	Yield	Window
Baseline	-0.075	3 weeks
	(0.057)	
Positive shocks	-0.080	3 weeks
	(0.086)	
Negative shocks	-0.074	3 weeks
	(0.100)	

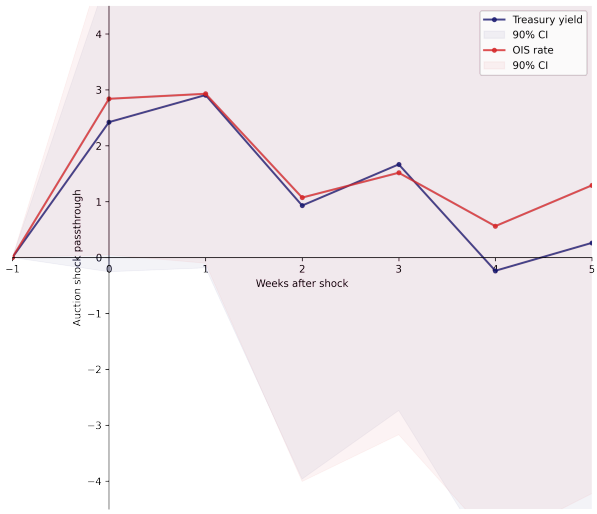
The table reports initial decay rates for positive and negative demand shocks. The specification uses 5-year Treasury note auction demand shocks over the full sample (2008–2022) with a linear decay function and 3-week estimation window. The coefficients are scaled at the daily frequency. Standard errors use date-cluster bootstrap (1,000 replications). Significance: *** p<0.01, ** p<0.05, * p<0.1.

B.4 Crisis Episode Impulse Responses

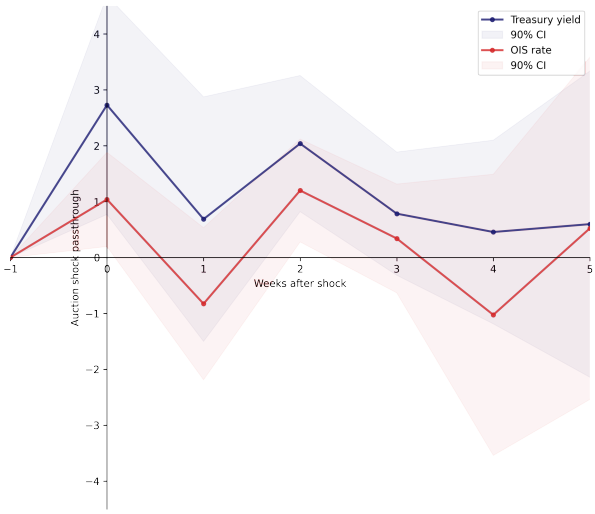
Figure 10 shows the impulse responses of Treasury yields and the OIS rate to 5-year Treasury note auction demand shocks during the two crisis episodes. In both episodes, yields and OIS rates co-move closely. The spread impulse responses are reported in the main text (Figure **??**).

Figure 10: Impulse response of Treasury yields and OIS rates: Crisis episodes

(a) GFC (2008–2010)



(b) COVID (2020–2021)



The figures plot impulse response functions of Treasury yields (blue) and the OIS rate (red) to 5-year Treasury note auction demand shocks for two crisis episodes, estimated at weekly frequency, with 90% confidence bands. Panel (a) shows the GFC episode (January 2008 to January 2010). Panel (b) shows the COVID episode (February 2020 to February 2021).

C Calibration Details

This appendix summarizes how we estimate the yield-space variance term structures and map them to the structural parameters in Section 5.

C.1 Estimating variance term structures

For each series $i \in \{l, s\}$ and horizon $H = 1, \ldots, 15$, we estimate the empirical per-period variance

$$\widehat{V}^{y}_{i,H} = \mathrm{Var}(\Delta_H y_i)/H$$

using overlapping H-day changes. For the spread series, we use the 5-year Treasury-OIS half-spread,

$$y_{s,t} = \frac{y^T_{5y,t} - y^{OIS}_{5y,t}}{2}.$$

For the level series, we proxy the model's average yield factor with the residualized average level

$$y^{\text{res}}_{l,t} = \frac{y^T_{5y,t} + y^{OIS}_{5y,t}}{2} - \widehat{\beta} y_{3m,t},$$

where $\widehat{\beta}$ is estimated from full-sample daily changes. The 3-month Treasury yield serves as a proxy for the short-rate factor. Residualizing the average level on this factor removes autocorrelation in the fundamentals, and hence provides a parsimonious empirical analogue of the model's $\eta_v = 0$ specialization. We then fit equation (42) separately to the level and spread profiles by nonlinear least squares, recovering $(V^y_{i,0}, V^y_{i,\infty}, \nu_i)$. In the calibration, we use the fitted short-horizon intercept $V^y_{i,0}$ rather than the raw one-day variance.

The main-text table reports these fitted moments in yield units. The model is written in price-return units, so using the sample-average modified duration D of the 5-year Treasury, we convert the fitted moments via

$$V_{i,0} = D^2 V^y_{i,0}, \qquad V_{i,\infty} = D^2 V^y_{i,\infty}, \qquad i \in \{l, s\},$$

while the decay rates ν_i are unchanged.

C.2 Recovering structural parameters

We calibrate the symmetric $r = 0$ case used in the main text and impose $\boldsymbol{\eta}_\beta = \mathbf{0}$, $\eta_v = 0$, $\text{Var}(d\beta_C) = \text{Var}(d\beta_S) = \sigma^2_\beta dt$, and zero covariance between the two demand shocks and the fundamental shock. Because the variance moments identify the model only up to one common scale, we fix $M_{l,\infty} = 0.15$ following Chaudhary, Fu, and Zhou (2025). Under these maintained restrictions, the remaining unknowns are $(k, \sigma_\beta, \sigma_V, M_{s,\infty}, \gamma, \psi)$.

The reason this normalization is needed is that, for any $c > 0$, the transformation

$$M'_{l,\infty} = cM_{l,\infty}, \qquad M'_{s,\infty} = cM_{s,\infty}, \qquad \sigma'_\beta = \sigma_\beta / c, \qquad \gamma' = c\gamma, \qquad \psi' = c\psi$$

leaves the fitted variance term structures unchanged, while k and σ_V are unaffected. Under

this rescaling, the on-impact multipliers $M_{l,0}$ and $M_{s,0}$ also scale by c, but the decay rates ν_l and ν_s, the variance moments, and the implied slope ratio all remain the same. Hence the data identify the relative amplification of level and spread price impacts, but not the absolute scale of $M_{l,\infty}$ without an external normalization.

We recover them sequentially. First, using

$$V_{s,0} = \tfrac{1}{2}\sigma_\beta^2 M_{s,0}^2, \qquad V_{s,\infty} = \tfrac{1}{2}\sigma_\beta^2 M_{s,\infty}^2, \qquad \frac{M_{s,0}}{M_{s,\infty}} = \frac{\nu_s}{k},$$

we obtain

$$k = \nu_s \sqrt{\frac{V_{s,\infty}}{V_{s,0}}}.$$

Second, the level moments imply

$$\sigma_\beta^2 = \frac{2(V_{l,0} - V_{l,\infty})}{M_{l,\infty}^2 \left[\left(\frac{\nu_l}{k}\right)^2 - 1 \right]}, \qquad \sigma_V^2 = V_{l,\infty} - \tfrac{1}{2}\sigma_\beta^2 M_{l,\infty}^2.$$

Third, long-run spread variance pins down

$$M_{s,\infty} = \sqrt{\frac{2V_{s,\infty}}{\sigma_\beta^2}}, \qquad M_{l,0} = \frac{\nu_l}{k} M_{l,\infty}, \qquad M_{s,0} = \frac{\nu_s}{k} M_{s,\infty}.$$

Finally, the intermediary first-order condition implies

$$\begin{aligned} A_s &\equiv \frac{k}{M_{s,\infty}} M_{s,0}^2 - k M_{s,0} = \psi + 2\gamma V_{s,0}, \\ A_l &\equiv \frac{k}{M_{l,\infty}} M_{l,0}^2 - k M_{l,0} = \psi + 2\gamma V_{l,0}. \end{aligned}$$

Solving these two equations yields

$$\gamma = \frac{A_l - A_s}{2(V_{l,0} - V_{s,0})}, \qquad \psi = \frac{V_{l,0} A_s - V_{s,0} A_l}{V_{l,0} - V_{s,0}}.$$