ISSN 0956-8549

Causal Inference for Asset Pricing

By

Valentin Haddad

Zhiguo He

Paul Huebner

Péter Kondor

Erik Loualiche

FINANCIAL MARKETS GROUP DISCUSSION PAPER NO. 977

August 2026

Any opinions expressed here are those of the authors and not necessarily those of the FMG. The research findings reported in this paper are the result of the independent research of the authors and do not necessarily reflect the views of the LSE.

Causal Inference for Asset Pricing*

Valentin Haddad Zhiguo He Paul Huebner
Péter Kondor Erik Loualiche

June 2026

Abstract

Portfolio choice involves substituting across many assets at once, complicating inference about asset demand. An elementary condition often captures this behavior in theory and practice: homogeneous substitution conditional on observables (e.g., factor loadings, maturity, credit ratings). We characterize natural experiments identifying demand elasticity and price impact under this condition. Cross-sectional IV and difference-in-differences identify relative elasticity, own- minus cross-price elasticity for assets sharing observables. But a missing-coefficient problem leaves substitution unidentified: the coefficients on observables mechanically absorb it. Identifying substitution requires time-series regressions on portfolios sorted on observables. We apply the framework to corporate bonds, comparing alternative Fed asset-purchase programs.

*Haddad: Anderson School of Management, UCLA and NBER, valentin.haddad@anderson.ucla.edu; He: Graduate School of Business, Stanford University and NBER, hezhg@stanford.edu; Huebner: Stockholm School of Economics, paul.huebner@hhs.se; Kondor: London School of Economics and CEPR, p.kondor@lse.ac.uk; and Loualiche: Carlson School of Management, University of Minnesota and NBER, eloualic@umn.edu. We thank Steven Berry, Mikhail Chernov, Carter Davis, Zhiyu Fu, William Fuchs, Satoshi Fukuda, Xavier Gabaix, Sergei Glebkin, Paul Goldsmith-Pinkham, Elise Gourier, Lars Peter Hansen, Kristy Jansen, Ralph Koijen, Jane Li, Lira Mota, Daniel Neuhann, Anna Pavlova, Aleksandra Rzeźnik, Milena Wittwer, Hanbin Yang, Motohiro Yogo, and seminar and conference participants at Minnesota Carlson, Chicago Booth Alumni Insights, UCLA Anderson, Stockholm School of Economics, Demand in Asset Markets Working Group, USC Macro-Finance Workshop, Adam Smith Workshop, Spring Finance Workshop, SFS Cavalcade, FIRS, CFF Asset Pricing and Machine Learning, LBS Summer Symposium, SoFiE Pre-Conference, Gerzensee, NBER Summer Institute, USC, SAFE Asset Pricing Workshop, Chicago Booth, Harvard, Baruch, Yale, University of Zurich, Duke, CEPR Paris Symposium, Imperial College, AFA Annual Meeting, MIT, Toronto Rotman, HEC-McGill Winter Finance, University of Bonn, Boston College, and Princeton Conference on Asset Demand for helpful comments and suggestions.

Introduction

Causal inference methods that leverage plausibly exogenous sources of variation have become essential tools in empirical economics (Angrist and Pischke, 2009). Recently, these methods have gained traction in asset pricing to better understand the demand for financial assets, through both specific experiments like index inclusions or central banks' asset purchases (Shleifer, 1986; Chang et al., 2014; Krishnamurthy and Vissing-Jorgensen, 2011), and as building blocks for demand systems (Koijen and Yogo, 2019; Haddad et al., 2025b). However, these approaches differ sharply from traditional empirical methods in asset pricing (see, e.g., Cochrane, 2005; Campbell, 2017), which instead prioritize tests of equilibrium relationships such as Euler equations or the CAPM.

We provide a framework for using causal inference in the asset pricing context. We put forward elementary conditions that allow using the standard toolbox of causal inference while capturing the rich substitution patterns of portfolio choice. Under these conditions, we fully characterize what sources of variation and estimation procedures identify portfolio demand and its equilibrium impact. Our characterization reveals a missing-coefficient problem: cross-sectional evidence identifies how investors pick among similar assets but not how they rotate across different ones, so the cross-section alone cannot answer many cross-sectional questions.

More broadly, our focus is on how portfolio choice responds to asset prices, or conversely, how asset prices respond to an exogenous shift in demand. Natural experiments offer a window into these questions. If the Fed purchases quasi-randomly some bonds but not others, prices respond, and one can look at how various investors adjust their portfolios. One can also look at the surprise price change following the announcement of a broad purchase program by the Fed, or an intervention such as Operation Twist, buying more long-term than short-term bonds.[1] But what exactly is learned about investor demand from analyzing these different experiments? Alternatively, one can start from a counterfactual question, such as "what would be the effect of an increase in investor preference for firms with good ESG performance?" and ask which evidence is necessary to inform this counterfactual.

In all generality, all these aspects of demands are necessarily intertwined. Portfolio choice (Markowitz, 1952), and really basic theory of demand (Debreu, 1959), imply that the demand for each asset responds not only to the price of one asset, but to the price of every asset. Formally, with N assets, the slope of the demand curve, the demand elasticity, is an $N \times N$ matrix: $\boldsymbol{\mathcal{E}} = \partial D / \partial P$. Furthermore, prices are connected in equilibrium.[2] In the example

[1] Selgrad (2023) analyzes variation in which specific bond the Fed purchases in the implementation of quantitative easing programs. Krishnamurthy and Vissing-Jorgensen (2011) focus on the response to the announcement of broad QE programs.

[2] Fuchs et al. (2025a) brought attention to the role of equilibrium spillovers in demand estimation.

above, even if the Fed does not buy a bond, the price of this bond might respond to purchases of its substitutes. Without any restrictions, the only way to make progress is to have N different exogenous sources of variation in prices, a tall order.

Imposing further economic restrictions on the elasticity matrix is the natural way forward, but the researcher must be careful not to oversimplify. The most extreme simplification assumes that demand for an asset depends only on its own price. This makes the econometrics easy (standard SUTVA), but it ignores that investors manage portfolios. In classic finance, each asset is a bundle of exposures to common risk factors, and two assets are substitutes when they are exposed to similar risks. This implies that demand elasticity has two distinct components that must be modeled simultaneously. There is the "stock picker" logic, where if Ford gets expensive relative to GM, you swap them because they have the same risk profile. But there is also the "factor management" logic, where if the value premium decreases and stocks with a high book-to-market ratio get expensive, you might rotate your portfolio out of them and into other factors. We show that leading structural models like Koijen and Yogo (2019) and Gabaix and Koijen (2021) prioritize tractability in ways that limit their ability to capture the latter dimension. They capture the stock-picking behavior well, but implicitly tether how investors can rotate between broad risk factors. We provide a concrete set of examples illustrating this limitation: these models cannot explain standard portfolio adjustments in response to price changes in high-beta versus low-beta assets. Instead, we enrich the family of possible foundations for portfolio choice while maintaining tractability in estimation.

Our framework relies on a transparent restriction: *homogeneous substitution conditional on observables*. This condition specifies that investors' demand for any two assets with identical observables responds identically to a price change in any other asset. For instance, if Ford and GM carry the same risk and capital charge, an investor's demand for them must respond identically to a price shock in First Solar. This condition holds for many foundations: whether investors are managing risk (using factor loadings), navigating balance sheet constraints (using capital charges), or targeting green assets (using ESG scores). This flexibility allows us to capture the diverse motives driving asset demand while keeping the econometrics tractable. Indeed, the restriction yields as many degrees of freedom in substitution as the mixed logit familiar from discrete choice (Berry et al., 1995), but delivers that flexibility through a structure native to portfolio choice, and linearly. Beyond theoretical justifications, we provide a variety of empirical approaches to assess the plausibility of this assumption, and tests to falsify it.

Our assumption simplifies the daunting $N \times N$ elasticity matrix into two naturally interpretable components: *relative elasticity* and *substitution*. The relative elasticity captures the

micro response of the stock picker logic: if Ford becomes expensive relative to GM, how much do you swap them? In our baseline, the relative elasticity is a single number; we also show how to make it heterogeneous across assets. The substitution matrix captures reallocation across stocks driven by observables: if assets with a high beta and low capital charge become expensive, how much do you rotate into other assets? This matrix is driven by $K \times K$ parameters, where K is the number of observables. Identifying these two components requires two distinct empirical strategies.

We formally show that standard cross-sectional identification methods, like instrumental variables or difference-in-differences, successfully recover the relative elasticity. By comparing a treated asset (Ford) to a control asset (GM) with the same observables, the shared substitution effects — e.g., that they both respond to a shock in Tech stocks — wash out. More generally, we prove that controlling for observables in cross-sectional regressions of changes in demand on changes in prices is sufficient to absorb these substitution effects. The researcher, nevertheless, still needs an instrument, since relative prices remain endogenous to demand shifts (e.g., a shock to preference for Ford vs. GM). The resulting regression coefficient consistently estimates the relative elasticity, the difference between own- and cross-price elasticities. Crucially, this means that the seemingly naive cross-sectional regression that appears to ignore substitution works, but it only answers a specific question: how demand responds to *relative* price changes, not absolute ones nor those affecting multiple assets.[3]

However, the cross-section is blind to substitution. Substitution is a rotation across observables — selling low-beta stocks to buy high-beta ones. Cross-sectional regressions work precisely by controlling for these observables. They mechanically absorb these rotations into the coefficients on the observables, creating a "missing coefficient" problem: these coefficients cannot distinguish between buying high-beta stocks because their price dropped relative to low-beta stocks (substitution, the object of interest) and buying them because you like beta more (a demand shift). More broadly, a single cross-section cannot tell substitution apart from a demand shift whenever both act through the same characteristics. This issue echoes the "missing intercept" problem in macroeconomics: general equilibrium effects are absorbed by time fixed effects. But the trap here is subtler: substitution hides in the slopes, not just the intercept. To isolate substitution, the researcher must look to the time series along a few dimensions. Practically, the researcher aggregates data into portfolios along the observables that drive substitution, such as sorting stocks based on their beta. Then, they need to identify exogenous time-series shocks that move the price of these portfolios. By observing how investors tilt their portfolios across these characteristics — swapping high-beta for low-

[3]The same reasoning rationalizes the demand regression introduced in Koijen and Yogo (2019). However, unlike our partial identification result, parametric restrictions constrain substitution so much in their model that the researcher recovers the full elasticity matrix from the cross-sectional regression alone.

beta stocks — in response to these shocks, the researcher recovers the substitution.

Our framework thus unifies two distinct empirical traditions: cross-sectional instruments (like index inclusions) to measure relative elasticity, and time-series shocks (like aggregate Fed purchases) to measure substitution. By combining these two approaches, we can fully characterize the elasticity matrix and answer counterfactuals that require both logics simultaneously, such as quantifying the equilibrium impact of Operation Twist or a broad ESG mandate. Notably, there is no neat separation of micro and macro. Instead, substitution driven by observables constitutes a "meso" layer: this force is essential to answer questions about the cross-section, but requires identification from broad time series.

Our framework applies equally to the measurement of price impacts, or multipliers — the effect of exogenous supply or demand shocks on asset prices. Questions like "how do Fed asset purchases affect bond prices?" reverse the variables of interest: instead of asking how prices move quantities, we ask how quantities move prices. We show that because the multiplier matrix is the inverse of the elasticity matrix of aggregate demand, our assumptions of homogeneous substitution and constant relative elasticity translate directly to this setting. Consequently, the same decomposition holds: cross-sectional regressions identify the relative price impact, while time-series regressions on portfolios recover the aggregate and meso-level multipliers.

We assess the robustness of our approach to a number of potential concerns. Minor deviations from our assumption have a small effect on estimates. We outline how to entertain richer heterogeneity at the cost of stronger assumptions and additional sources of variation. We confirm that our approach is compatible with no-arbitrage conditions and show how to assess empirically whether they play a role. Naturally, settings characterized by deviations from no-arbitrage relations also lend themselves particularly well to our framework.[4]

In the final part of the paper, we illustrate practically how to apply our framework by studying price impact in the corporate bond market. Consistent with Chaudhary et al. (2022), we find a relative price impact indistinguishable from zero. Going beyond this micro estimate, we recover the spillover structure and show that it is well captured by substitution linear in two observables, duration and credit rating. We use these estimates to evaluate counterfactual designs of central bank asset purchase programs and find that program design — broad versus targeted, as in the SMCCF — materially changes both aggregate price impact and which bonds benefit, including through spillovers to non-purchased bonds.

Taken as a whole, this paper arms the researcher with a practical guide: which assump-

[4]Examples include covered-interest-parity deviations (Du et al., 2018), the CDS-bond basis (Bai and Collin-Dufresne, 2019), violations of put-call parity (Van Binsbergen et al., 2022), or the pledgeability premium of corporate bonds (Chen et al., 2023); these deviations reveal that investors do not treat these assets as redundant.

tions justify causal inference in asset markets, how to assess them, which design answers which question, and how to read the resulting estimates.

Related Literature. A long tradition in finance uses plausibly exogenous sources of variation to understand portfolio decisions and the price impact of shifts in demand. Prominent examples include the effect of index inclusion (Shleifer, 1986; Harris and Gurel, 1986; Chang et al., 2014; Pavlova and Sikorskaya, 2022; Greenwood and Sammon, 2024), institutional ownership and fund flows (Gompers and Metrick, 2001; Coval and Stafford, 2007; Lou, 2012; Ben-David et al., 2022; Hartzmark and Solomon, 2022), central bank asset purchases (Krishnamurthy and Vissing-Jorgensen, 2011; Selgrad, 2023; Haddad et al., 2021, 2025a), financial constraints (Du et al., 2018; Greenwood and Vissing-Jorgensen, 2018; Haddad and Muir, 2021; Chen et al., 2023), or exchange rates (Evans and Lyons, 2002; Froot and Ramadorai, 2005). This work often incorporates thorough analysis of exogeneity, in particular in the wake of the "credibility revolution" (e.g., Angrist and Pischke, 2009). However, this literature is often more scant in considering a central feature of asset pricing theory, substitution across assets, and whether it affects the validity of inference and the interpretation of estimates. Our framework provides a simple bridge between classical discussions of causal inference and the role of substitution.

Another approach fully specifies and estimates models of portfolio demand and their equilibrium implications. In a seminal article, Koijen and Yogo (2019) derive and estimate a logit model of portfolio choice in the stock market. Subsequent work uses this model either structurally, or as a semi-structural simplification to introduce other mechanisms (e.g. Haddad et al. (2025b)). Applications include quantifying the impact of the rise of passive investing, preferences for sustainable assets (Koijen et al., 2023; Van der Beck, 2021), or the transmission of monetary policy (Lu and Wu, 2023), and have found echo in other settings: the stock market overall (Gabaix and Koijen, 2021), corporate bonds (Bretscher et al., 2022), treasuries (Jansen et al., 2024; Fang, 2023; Fang and Xiao, 2024), or exchange rates (Koijen and Yogo, 2024; Jiang et al., 2024).

We build on some of the insights from estimation inside of these models. Some important ideas are controlling for common exposures (Koijen and Yogo, 2019), the distinction between micro and macro elasticity (Gabaix and Koijen, 2021; Li and Lin, 2022), heterogeneous substitution (Chaudhary et al., 2022; Aghaee, 2024), substitution along factors (An et al., 2024; An and Huber, 2025; Peng and Wang, 2023), and accounting flexibly for spillovers (Fuchs et al., 2025a). Naturally, our simple conditions cannot cover every model; we leave aside considerations of strategic responses (Haddad et al., 2025b), dynamics (Greenwood et al., 2018; Gabaix and Koijen, 2021; Huebner, 2024; He et al., 2025), state-contingent demand

shocks (Haddad et al., 2025a), intermediary distress (He et al., 2022), or bidding in auctions (Allen et al., 2018). In this context, the contribution of our framework is twofold: it not only provides a unifying formalism to discuss identification across models but also allows discussion of what can be learned from the data before espousing a specific model.

Finally, the role of spillovers is not limited to asset pricing and has been recognized in many other contexts. The industrial organization literature often relaxes the assumption of independence of irrelevant alternatives (IIA) and includes heterogeneous substitution in discrete choice models, such as Berry et al. (1995). While without observables, our assumptions would be closely related to IIA, including the observables entertains heterogeneous substitution. Our setting of portfolio choice in line with finance theory does so without introducing nonlinearities, lending itself to using linear regressions. Berg et al. (2021) discuss spillovers in corporate finance. In macroeconomics, a key concern is the missing intercept problem due to general equilibrium effects, with some recent contributions such as Chodorow-Reich et al. (2021), Guren et al. (2021), Huber (2023), and Wolf (2023).

1 The Challenge of Causal Inference in Asset Pricing

We introduce our estimation target, the demand elasticity matrix, and its central role in answering questions about asset prices. Estimation faces two hurdles. First, rich substitution patterns and price endogeneity make estimation impossible without imposing restrictions. Second, restrictions must be flexible enough to capture well-known asset pricing patterns, something leading structural models fail to do. Our framework in Section 2 solves both problems by relying on simple and flexible assumptions.

1.1 Demand and Demand Elasticity

We focus on a generic setting for identifying the demand for financial assets of an investor, with the goal to understand their demand elasticity — that is, how the investor adjusts their portfolio in response to prices.

Demand elasticity. When an investor maximizes their utility given prices—or simply responds to prices through whatever decision process they follow—this choice characterizes demand as a function of prices and other aspects of the decision problem.[5] The demand

[5]Becker (1962) explains the generality of demand as a function of prices.

elasticity is the partial derivative of this function with respect to prices:

$$\boldsymbol{\mathcal{E}} = \frac{\partial D}{\partial P}, \tag{1}$$

where, in a slight abuse of language, we refer to $\boldsymbol{\mathcal{E}}$ as an elasticity irrespective of logs or levels. Importantly, $\boldsymbol{\mathcal{E}}$ is a matrix of size $N \times N$, where N is the number of assets. The diagonal elements $\partial D_i/\partial P_i$ measure the own-price elasticities: how demand for Apple responds to the price of Apple. The off-diagonal elements $\mathcal{E}_{ij} = \partial D_i/\partial P_j$ capture cross-price elasticities: how demand for Apple responds to the price of Nvidia.

In the finance context, these cross-price elasticities are essential to a key insight dating back to Markowitz (1952): assets are not distinct goods but instead alternative means of saving with different risk and reward. Investors choose portfolios optimally combining these assets. The most standard example of this approach is mean-variance optimization: an investor chooses their portfolio to maximize $\mathbf{E}(W) - \frac{\gamma}{2}\operatorname{var}(W)$ where W is their future wealth, and γ measures their absolute risk aversion. Denoting $\bar{\Pi}$ the vector of investor's expected payoffs and Σ their covariance matrix, the vector of demand (in number of shares) is

$$D = \frac{1}{\gamma}\Sigma^{-1}(\bar{\Pi} - P). \tag{2}$$

The matrix of elasticity $\boldsymbol{\mathcal{E}}$ is determined by risk aversion and the covariance between assets: $\boldsymbol{\mathcal{E}} = -\gamma^{-1}\Sigma^{-1}$.[6] When assets are correlated with each other, they become close substitutes, and their demands respond to each other's prices.

Modern finance research acknowledges many deviations from this simple setting: investors have different beliefs and various cognitive limitations; institutions face many regulations and constraints that influence their portfolio decisions. Still the basic idea of portfolio choice as opposed to asset choice remains. Appendix B explores the elasticity across many models. Note also that two investors sharing the same elasticity respond the same way to changes in prices irrespective of their underlying trading motives.

The demand elasticity of investors is crucial to understand how the equilibrium changes in response to changes in market macrostructure, the broad organization of financial markets (Haddad and Muir, 2025). For example, the impact of a counterfactual portfolio liquidation on prices is driven by the inverse of the aggregate demand elasticity. That is, if a specific investor wakes up and must sell quantities ΔQ, the vector of prices moves by $\Delta P = -\boldsymbol{\mathcal{E}}_{agg}^{-1}\Delta Q$.[7]

[6] If $\overline{\Pi}$ and Σ respond to changes in prices, the elasticity formula would be different. Koijen and Yogo (2019) characterize demand curves in such a setting. Koijen et al. (2023) add hedging demands.

[7] The equilibrium prices equate the aggregate demand D_{agg} (the sum of demand across investors) to the aggregate supply S, in that $D_{agg}(P) = S$. If demand curves shift by an amount ΔQ, the new equilibrium price $P + \Delta P$ satisfies $D_{agg}(P + \Delta P) + \Delta Q = S$. With the implicit function theorem, the price impact or

More broadly, armed with investors' demand elasticities, the researcher can answer questions about altering a subset of investors: for example bank regulation, quantitative easing, etc. Appendix B.1 characterizes the set of counterfactual questions that can be answered with demand elasticities in standard competitive equilibrium settings. Appendix B.6 further shows how demand elasticities remain informative in models of asymmetric information or with strategic motives (see also Haddad et al., 2025b).

Structural demand equation. The objective of this paper is to estimate the demand elasticity $\boldsymbol{\mathcal{E}}$ using data on prices and portfolio positions. To do so, one must specify a structural demand equation. Portfolio positions change because of prices (movements along the demand curve) and other reasons (shifts in the demand curve). We study a linear demand in which the elasticity is constant across the experiment and these other movements are collected in the component ϵ:

$$\Delta D_t = \boldsymbol{\mathcal{E}} \Delta P_t + \epsilon_t \quad \Longleftrightarrow \quad \Delta D_{i,t} = \sum_j \mathcal{E}_{ij} \Delta P_{j,t} + \epsilon_{i,t}, \tag{3}$$

where i and j index assets, and t indexes time periods; subsequently, we often drop the latter when focusing on a single time period. In the mean-variance example, the investor's beliefs about expected payoffs could move over time. Then, equation (3) holds exactly with $\epsilon = \frac{1}{\gamma}\Sigma^{-1}\Delta\bar{\Pi}$, where $\Delta\bar{\Pi}$ is the vector of change in beliefs. Throughout the paper, we will discuss other forces that can drive portfolio decisions given prices and contribute to ϵ. The key common property, echoing our earlier discussion of counterfactuals, is that the ϵ remains stable through the experiments of interest.

One can specify equation (3) in levels or logs depending on the model of demand. For example, models like CARA preferences are better behaved in levels, while logit demand aligns with logs. In practice, the choice of units should be driven by regularity in the data and the type of model the researcher believes best matches this regularity.[8] Also, while we focus on writing specifications in changes to match the standard difference-in-difference framework, similar arguments apply without changes. Regardless of units, if the model is not linear (or log-linear) in prices, we focus on a first order approximation of demand.[9] Appendix

multiplier matrix is the inverse of the elasticity matrix of aggregate demand: $\boldsymbol{\mathcal{M}} = -\left(\frac{\partial D_{agg}}{\partial P}\right)^{-1} = -\boldsymbol{\mathcal{E}}_{agg}^{-1}$. Section 4 considers empirical approaches to estimate $\boldsymbol{\mathcal{M}}$ directly as opposed to the demand elasticity of each investor separately.

[8] Appendix B.3 reviews the appropriate units for standard models. See Appendix B.2 for a discussion of the link between elasticity and the multiplier in different units.

[9] For example, if demand depends on the price vector and various observables X and unobservable quantities ξ, a demand function $D(P, X, \xi)$ can be linearized around a specific value with $\Delta D = \boldsymbol{\mathcal{E}}\Delta P + (\partial D/\partial X)\Delta X + (\partial D/\partial \xi)\Delta\xi$. By definition, the elasticity, $\boldsymbol{\mathcal{E}} = \partial D/\partial P$, is fixed in this linearization, and

D discusses the nonlinear case.

1.2 Restrictions are Necessary to Estimate Demand Elasticity

The combination of rich substitution and price endogeneity poses a challenge to the estimation of demand elasticity.

Endogeneity of prices. Equation (3) suggests regressing demand on prices to estimate the demand elasticity $\boldsymbol{\mathcal{E}}$. This approach is biased when shifts in the demand curve ϵ are correlated with prices. Such a correlation is very natural. If positive news comes out about a company, the investor would be inclined to buy more of its shares (a positive ϵ) if the price remained unchanged. However, the stock price is also likely to rise in response to the news. Similarly, periods when many investors face financial distress and seek to liquidate their portfolios (a negative ϵ) tend to coincide with depressed asset prices.

A natural solution for this challenge, in line with the causal inference literature, is to use natural experiments: situations where variation in prices is unrelated to shifts in the demand curve of the investor. For example, one could focus on the inclusion of a stock in an index as an event that increases its price (Shleifer, 1986). Then, measuring the change in portfolio of an investor for whom the index status of the stock is irrelevant reveals their demand elasticity.

Substitution. However, it is not that simple: equation (3) highlights that all prices matter for all demands due to substitution between assets. Hence, experiments for a single asset at a time can generally not be used to recover the matrix of elasticity $\boldsymbol{\mathcal{E}}$, or even the own-price elasticity for this asset.[10] In the language of causal inference, the stable unit treatment value assumption (SUTVA) is violated: treatment on one unit (for us, an asset) affects other units, because cross-price elasticities are not zero. A way out would be to include all prices in the demand estimation regression. This is not possible in practice because it requires exogenous

$\epsilon = (\partial D/\partial X)\Delta X + (\partial D/\partial \xi)\Delta\xi$. Linearity implies that one need not separate the components of ϵ to identify $\boldsymbol{\mathcal{E}}$, which is what matters for our counterfactuals of interest. Separating these components would be necessary for other counterfactual exercises such as the effect of changes in beliefs about a specific asset, leading to other estimation strategies (see, e.g., Berry and Haile, 2021).

[10] Consider for simplicity a setting with three assets and no residual shift in demand. In addition, assume that there is no cross-elasticity between assets 1 and 2 and that they have the same own-price elasticity $\mathcal{E}_0$. One might want to use an experiment that affects asset 1 but not asset 2 and use it as an instrument to estimate this elasticity. The corresponding instrumental variable would have $Z_1 = 1$ and $Z_2 = 0$. In a regression with assets 1 and 2, this leads to an IV estimate of the elasticity $\hat{\mathcal{E}} = (\Delta D_1 - \Delta D_2)/(\Delta P_1 - \Delta P_2) = \mathcal{E}_0 + (\mathcal{E}_{13} - \mathcal{E}_{23})\Delta P_3/(\Delta P_1 - \Delta P_2)$, which is a biased estimator of the own-price elasticity $\mathcal{E}_0$.

sources of variation for each one of the (thousands of) individual prices.[11]

Faced with this challenge, one can deem causal inference hopeless for asset pricing and throw their hands in the air. However, there is a more constructive approach: acknowledge that additional assumptions about the nature of substitution are necessary. After all, this is the second part of Markowitz' argument: basic economics can inform us about the structure of substitution across assets. In the rest of the paper, we follow this path and put forward simple, flexible conditions guided by these economic principles. Our approach is in the spirit of the causal inference literature: our conditions allow to estimate the elasticity matrix without fully specifying how the equilibrium emerges by using standard econometric methods (Angrist and Pischke, 2009).

These restrictions are flexible enough to model rich substitution patterns in financial markets. Next, we highlight why this flexibility is important.

1.3 Restrictions Should Accommodate Standard Finance Logic

Empirical finance research highlights that price movements are often driven by differential exposure to common factors. In standard finance theories, investors choose their portfolio to manage exposure to common factors. These insights have implications for the structure of the elasticity matrix — specifically cross-asset substitution — that an estimation framework must accommodate. We also show that leading structural models of demand do not.

A workhorse finance model. In standard models, investors focus on factor exposure as a way to manage their risk. Consider the Markowitz model introduced in Section 1.1 with power utility with relative risk aversion γ.[12] We assume that the covariance matrix has a factor structure such that log excess returns admit the representation $R^e_{it} = \alpha_i + \beta_i F_t + v_{it}$, where $v_{it} \sim \mathcal{N}(0, \sigma_v^2)$ is i.i.d across assets and orthogonal to the factor F_t (as in Campbell and Viceira (2002) or Koijen and Yogo (2019)). We focus on a single factor structure for illustrative purposes; the matrix versions of all the expressions hold with multiple factors. Each asset has a factor loading β_i and α_i is the component of expected return not explained by these factor loadings ($\beta'\alpha = 0$ by construction).

In this model, the relation between the portfolio share in each asset (the natural unit of demand) and the log price satisfies the structural form of equation (3); it can be rearranged

[11] This issue is the well-known challenge of needing N instruments for estimating a flexible linear demand function as in Deaton and Muellbauer (1980). In non-parametric settings, Berry and Haile (2014) show that $2N$ instruments are required for full identification of a demand function. However, N instruments are enough if one only wants to characterize price counterfactuals (Borusyak et al., 2026), as we do.

[12] The agent maximizes expected utility $\mathbf{E}[W^{1-\gamma}/(1-\gamma)]$ with risk-free rate r_f and log-normally distributed payoffs $\log(\Pi) \sim \mathcal{N}(\bar{\Pi}, \Sigma)$. Then log returns are $R_i = \log(\Pi_i/P_i)$ which we log-linearize following Campbell and Viceira (2002). The appropriate notion of excess returns is $R^e_i = R_i - r_f + \frac{1}{2}\sigma^2_{R_i}$.

in an intuitive form with the factor representation as:

$$D_i = \text{portfolio share}_i = \frac{1}{\gamma}\left[\Sigma^{-1}\left(\bar{\Pi} - \log P - r_f + \frac{1}{2}\operatorname{diag}(\Sigma)\right)\right]_i \tag{4}$$

$$= \beta_i \Upsilon \mathbf{E}(F) + \frac{1}{\gamma\sigma_v^2}\alpha_i, \tag{5}$$

where the scalar Υ depends on risk aversion and the covariance of returns but not prices; instead prices enter through the factor premium $\mathbf{E}(F)$ and the alphas. Two forces shape portfolio decisions. First, investors engage in factor management: they invest in a portfolio with weights proportional to the factor loadings β. They buy more or less of this portfolio depending on the risk premium of the factor. Second, when the factor model does not completely explain returns and the alphas are different from zero, investors engage in "arbitrage" by tilting their investment towards assets with positive alphas.[13]

When the vector of prices changes, holding payoffs constant, expected returns change as well — both the alphas and the factor premium. Hence the same two forces also determine the elasticity matrix $\boldsymbol{\mathcal{E}} = (\gamma\Sigma)^{-1}$, the partial derivative of the portfolio share with respect to the log price:[14]

$$\boldsymbol{\mathcal{E}} = \beta\Psi\beta' - \frac{1}{\gamma\sigma_v^2}\mathbf{I}, \tag{6}$$

where β is the $N \times 1$ vector of factor loadings and Ψ is a counterpart of Υ in (5). Cross-asset substitution only depends on the first term. The substitution between asset i and j is determined by their factor loadings β_i and β_j; we later show that similar substitution patterns can arise for other motives than risk, in which case β_i is replaced by another asset characteristic.[15] Concretely, if the price of the beta-sorted portfolio increases, this implies a decrease in factor premium, and will alter demand for individual assets more or less depending on their own beta. The second component, driven by "arbitrage," is diagonal and hence only affects own-price elasticities: if an individual asset is cheaper, the investor increases their position in this specific asset.

Empirical evidence. In the data, we do not know a priori the motives of investors. Yet, the properties of returns are often highly suggestive that motives in the style of factor management are relevant. Treasury bond returns are well explained by two or three common

[13] While often called arbitrage, these trades are not the textbook's risk-free profit opportunities.

[14] See derivations in Appendix B.3. The covariance matrix is $\Sigma = \beta\Sigma_F\beta' + \sigma_v^2\mathbf{I}$. We have $\Psi = \frac{1}{\gamma\sigma_v^2}\left(\sigma_v^2\Sigma_F^{-1} + \beta'\beta\right)^{-1}$, and $\Upsilon = \frac{1}{\gamma\sigma_v^2}\mathbf{I} - \Psi\beta'\beta$, both being scalars in our single-factor structure.

[15] Thus, our framework does not require taking a stand on the age-old debate between factors and characteristics (Daniel and Titman, 1997; Kelly et al., 2019).

components (Litterman and Scheinkman, 1991). Much work focuses on common factors in stock returns (Fama and French, 1993), although with more debate on whether a low-dimensional representation explains the data. Asset comovement alone does not necessarily imply that substitution occurs along those dimensions.

Evidence from natural experiments where quantity shocks are known is more suggestive of the importance of this type of substitution. For example, Greenwood and Vayanos (2014) show that when the overall supply of treasuries changes, the response of bond yields lines up with maturity, even though the composition of the supply changes does not. Krishnamurthy and Vissing-Jorgensen (2011) or Haddad et al. (2024a) document a similar pattern in response to the Fed's announcement of bond purchases for quantitative easing operations. In the European context, the response across countries of bond yields to asset purchases by central banks lines up well with the riskiness of each country, even though purchases are mostly proportional to country size (e.g., Haddad et al., 2024b). For example, in some interventions, the ECB does not purchase Greek debt, yet its spread decreases more than that of any other country. Finally, Section 5 provides evidence of such substitution across different maturities and quantities of credit risk in corporate bonds.

Limits of leading structural models. To illustrate how existing models cannot capture these forces, we consider the following experiment. In our economy, there is an investor with demand function given by equation (4), as in Campbell and Viceira (2002), with a one-factor structure and varying asset supply. The researcher observes realizations of equilibrium prices and portfolio positions, with all demand parameters held constant (risk-aversion, covariance matrix, and expected payoffs). She estimates her favorite model of portfolio choice from the data; estimation does not face identification challenges because there are no demand shocks in the simulated data.

Then the researcher uses her estimated model to infer the counterfactual change in supply necessary to generate a specific change in prices: a larger observed increase in the price of high-beta assets than that of low-beta assets as in Figure 1, panel A.[16] Concretely, if a researcher had estimated a model of bond demand and then observed the response of prices to the announcement of QE, what pattern of Fed purchases would they predict?

Panel B represents the answer under different models of demand. In our economy, the actual supply shock is an equal decrease in the amount of all assets, the dashed black line. This lowers the amount of factor risk in the economy, which yields a smaller factor premium in equilibrium, and disproportionally affects the price of high-beta assets. Appendix E discusses

[16] One can also interpret this experiment without any reference to an equilibrium. The researcher estimates a model of demand from portfolio responses to exogenous changes in prices. Then they predict the response to a change in price as in panel A of Figure 1.

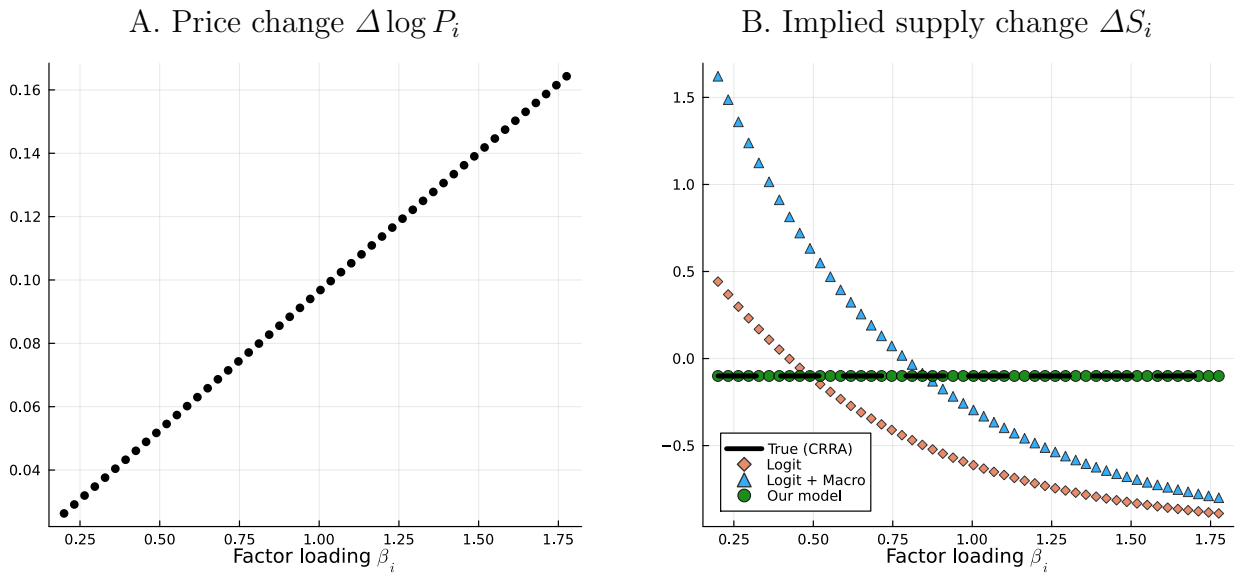


Figure 1: **Supply inference under alternative demand models.** Panel A plots the equilibrium log price change $\Delta \log P_i$ against asset beta β_i when the supply of all assets is reduced uniformly by 10%. High-beta assets experience larger price increases because the reduction in aggregate factor risk lowers the factor premium. Panel B shows the supply change each demand model infers as necessary to generate the price pattern in Panel A. The true shock is a uniform decrease across all assets (dashed black line). Demand models are logit (red), logit enriched with a macro elasticity (blue), and our model (green) introduced in Section 2. Parameters are calibrated to U.S. stock market data (1980–2021); see Appendix E and Table A2 for details on the log-linearized CRRA economy, estimation, and robustness across alternative specifications.

details of the experiment, variation in parameter choices, and alternative counterfactuals. If researchers instead favor a logit model of demand with beta as a characteristic, as in Koijen and Yogo (2019), they would predict a change in supply that is decreasing with beta, the red line.[17] Because the logit model does not entertain substitution of the form highlighted in (6), it attributes the differential change in price along betas to a differential change in supply.[18] One might worry that the logit model does not capture aggregate forces, specifically the macro elasticity emphasized by Gabaix and Koijen (2024). We enrich logit with a macro elasticity parameter (as suggested in their appendix G.4) and find no qualitative change in the required supply: on the blue line, the model still needs a differential change in supply to explain the change in prices; we explain precisely why in Section 3.2.

Anticipating the next section, we also fit the model of demand that we develop in this

[17] While Koijen and Yogo (2019) start with a framework similar to the theory of this section, they include additional restrictions that specialize the demand function to logit.

[18] Appendix E shows that allowing the logit coefficient to depend on beta does not match the true values quantitatively. We also present alternative experiments where the failure is qualitative because some assets not purchased experience the largest price change.

paper. Because the model accounts for asset substitution along observable characteristics such as the betas, it is able to infer the actual change in supply from the differential change in price, the green line.

2 A Model of Asset Demand

In this section, we state our model of the elasticity matrix: we give an assumption that is sufficient to make estimation tractable using standard causal inference methods, but flexible enough to entertain many motives for investment. Specifically, we show that it captures a wide range of theoretical foundations, including not only risk-based incentives as in Section 1.3 but also other motives such as regulatory constraints or non-pecuniary objectives.

2.1 Framework

Investors may substitute across assets in diverse ways, as long as these differences are observable.

Assumption A1 (Homogeneous substitution conditional on observables) *Any pair of assets in the estimation sample $\mathcal{S}$ with the same observables shares the same cross-price elasticity with respect to each third asset, within or outside of the estimation sample:*

$$\mathcal{E}_{il} = \mathcal{E}_{jl}, \quad \text{for all } i, j \in \mathcal{S} \text{ such that } X_i = X_j, \text{ and } l \neq i, j, \tag{7}$$

where X_i is the $K \times 1$ vector of observables for asset i. These cross-elasticities are parametrized by a bilinear form $\mathcal{E}_{cross}$:

$$\mathcal{E}_{il} = \mathcal{E}_{cross}(X_i, X_l) = X_i' \boldsymbol{\mathcal{E}}_X X_l, \tag{8}$$

where $\boldsymbol{\mathcal{E}}_X$ is a $K \times K$ matrix (which may not be symmetric).

Assumption A1 states that for two assets that are comparable along observables, if the price of any third asset—either within or outside the estimation sample—moves, then substitution between the third asset and the two comparable assets is the same. Specifically, take two firms that share the same characteristics, say Ford and General Motors. When the price of Netflix moves, the investor changes their position as much for Ford assets as for General Motors assets. In contrast, the investor might do something different for a firm with different characteristics, such as Target.

Importantly, assuming that the investor substitutes in the same way with Ford and General Motors is not the same as assuming that the two assets are identical. Features specific to each asset are still allowed to affect how much the investor demands of them, as materialized by the residual ϵ_i in equation (3). For example, in risk-based models, two assets with the same observables can have the same comovement with other assets, which drives substitution, but each of them has a distinct idiosyncratic component to their returns; see Section 2.2.1 below. In practice, the set of observables can include any variable measured prior to the price and quantity changes, including variables derived from past prices. We discuss principles to guide this selection throughout the remainder of the paper.

Analytically, equation (7) implies that cross-price elasticities are a function of the observables, which we write as $\mathcal{E}_{cross}(X_i, X_l)$. To make the model tractable, equation (8) parametrizes this function as a bilinear form in observables. This simple specification encompasses a large space of potential substitution patterns because observables could already include nonlinear transformations of more primitive variables or dummies for their levels.[19] We demonstrate this versatility and practicality in Section 2.2.

Remarkably, this intuitive assumption yields a decomposition of the elasticity matrix into two naturally interpretable components:[20]

$$\boldsymbol{\mathcal{E}} = \text{relative elasticity} + \text{substitution} \tag{9}$$

$$= \text{diagonal}(\boldsymbol{\mathcal{E}}_{relative}) + X\boldsymbol{\mathcal{E}}_X X', \tag{10}$$

where $X = [X_1'; X_2'; \cdots; X_N']$ is the $N \times K$ matrix of observables. In this framework, the elasticity matrix is entirely characterized by substitution, encoded by the $K \times K$ matrix $\boldsymbol{\mathcal{E}}_X$, and the vector of relative elasticity $\boldsymbol{\mathcal{E}}_{relative}$, an asset specific component. These two components generalize the structure of "arbitrage" for the relative elasticity, and "factor management" for the substitution matrix that we observed in equation (6). The substitution matrix encodes cross-price elasticities that depend on observables; we provide an interpretation of the relative elasticity in the next section.

To facilitate exposition, we regularize the problem further by assuming that the relative elasticity is constant across assets. In the language of causal inference methods, this corresponds to assuming a homogeneous treatment effect. There are many standard ways to relax this constraint. Section 3.1.2 extends the framework to consider situations where the relative elasticity is not constant and depends on either observable or unobservable sources

[19] Furthermore, it is most often natural to include a constant 1 as part of the observables. If the elasticity satisfies the assumptions for a set of observables X, it also does so for a linear transformation of these observables. For example, one can demean or standardize the observables without loss of generality.

[20] See proof in Appendix A.2.

of variation.

Assumption A2 (Constant relative elasticity) *Assets in the estimation sample have the same value of relative elasticity $\mathcal{E}_{relative}$ with respect to other assets with the same characteristics:*

$$\mathcal{E}_{ii} - \mathcal{E}_{ji} = \mathcal{E}_{relative}, \quad \text{for all } i, j \in \mathcal{S} \text{ such that } X_i = X_j. \tag{11}$$

Assumption A2 ensures a form of symmetry in how the investor responds to the price of assets with the same observables in the sample. It focuses on a specific dimension: the difference between the own-price and cross-price elasticity. We call this difference the relative elasticity. It represents how the demand for one asset relative to another shifts when the price of the asset changes relative to the other, when both assets have the same observables. The next section explains why this quantity is the natural target of cross-sectional regressions.

In some cases, an asset i does not have a "twin" j with identical observables. This often occurs when the observables are continuous variables, such as the sales of a firm. In this situation, we replace Assumption A2 with its natural extension: $\mathcal{E}_{ii} - \mathcal{E}_{cross}(X_i, X_i) = \mathcal{E}_{relative}$. Combining equation (10) and Assumption A2 simplifies the decomposition of elasticity:

$$\boldsymbol{\mathcal{E}} = \mathcal{E}_{relative}\mathbf{I} + X\boldsymbol{\mathcal{E}}_X X'. \tag{12}$$

In Section 3, we show how to use standard causal inference tools to estimate the two components in this setting: the relative elasticity $\mathcal{E}_{relative}$ (Section 3.1) and the substitution matrix $\boldsymbol{\mathcal{E}}_X$ (Section 3.3).[21] Of course, this tractability is only valuable if the assumptions are plausible in practice; we now turn to this question.

2.2 Applying the Assumptions

The applied researcher who wants to use our framework must motivate Assumption A1 in their setting using three complementary strategies. First, they can take a stand on the investor's motives for investment, that is, the underlying theory of portfolio choice. Second, they should choose appropriately which observables to include in X. Finally, they can restrict attention to a specific sample of assets.

The point here is not to argue that our assumption is always true; it is not. Still, we highlight that our assumption holds in a range of situations with choices that are intuitive, close

[21] These parameters yield a similar number of degrees of freedom as in the mixed logit model: $\mathcal{E}_{relative}$ is the counterpart to the price coefficient, and $\mathcal{E}_X$ corresponds to the covariance matrix of the random coefficients. Our specification folds the "linear coefficients" on characteristics inside of the shifter ϵ which is allowed to depend on X in equation (3).

to common empirical practice, and line up with standard finance theory. In Section 3.3.2, we discuss diagnostics that allow the researcher to detect when the assumed structure of substitution misses relevant dimensions.

2.2.1 Risk-based motives

The classic theory of portfolio choice relies on the tradeoff between risk and return. Then, the elasticity matrix as in equation (12) is driven by the covariance of asset returns.[22] When this covariance is driven by exposures to a common factor like in Section 1.3, Assumption A1 holds. Each asset's exposure β_i is the observable X_i driving substitution, as can be immediately observed in equation (6).[23] The exact same equation holds with multiple factors, with $\beta_i = X_i$ being the $K \times 1$ vector of factor exposures. Alternatively, under the assumption that the betas are a function of a set of asset characteristics, it is enough to use these characteristics as the observables (Koijen and Yogo, 2019). This corresponds to the instrumented principal components analysis model of Kelly et al. (2019, 2020).

The use of factor models is ubiquitous in empirical finance. Here, the specific notion of factor is for explaining the covariance matrix, not expected returns. For Treasuries, such a covariance is a strong feature of the data. For stocks, factor analysis typically reveals an important "market" factor, followed by a long succession of factors of relatively similar magnitude (Kelly et al., 2019; Kozak et al., 2020; Lopez-Lira and Roussanov, 2020), making it more challenging to capture substitution with a few observables. However, it is also important to note that actual investors might not use a fully sophisticated or rational estimate of covariance matrix for making portfolio decisions. For example, many financial institutions use factor structures such as the Barra model to assess and manage the risk of their portfolios; in this case, these factor loadings are the relevant observables for substitution.

2.2.2 Non-risk drivers of substitution

In practice, portfolio decisions respond to many other dimensions than risk and return. Some investors care about non-pecuniary aspects of the stocks they hold, such as their carbon emissions or ESG characteristics. Mutual funds, pension funds, and endowments often operate under mandates that require a specific mix of assets, while others are guided by broader objectives outlined in their prospectus. When hedge funds take on leveraged positions, haircuts apply and they have to post margins. Banks and insurance companies

[22] Beyond the static setting, He et al. (2025) shows that the same property holds in some dynamic economies.

[23] Furthermore, Assumption A2 holds if idiosyncratic volatility is constant across assets. In practice, one might be reluctant to assume constant idiosyncratic volatility; Section 3.1.2 shows how to relax this condition.

must ensure that their portfolios satisfy various regulatory targets such as capital adequacy ratios, leverage requirements, or liquidity requirements.

All these dimensions affect which assets these investors choose in the first place, but also how they rebalance their portfolio when prices move. For example, if one of your more environmentally friendly stocks appears overpriced, you might shed this position and replace it with another similarly green position. Hence, these motives can play an important role in the elasticity matrix. We show formally how this is consistent with our assumptions. Consider a generic representation of such a motive, by adding a quadratic cost and a linear constraint to the mean-variance optimization problem:

$$\max_{D} \quad D'(M-P) - \frac{\gamma}{2} D'\Sigma D - \frac{\kappa}{2}\left(D'X^{(1)}\right)^2 \tag{13}$$

$$\text{such that} \quad D'X^{(2)} \leqslant \Theta. \tag{14}$$

The quadratic cost $\kappa/2\left(D'X^{(1)}\right)^2$ captures smooth investment priorities: the more carbon-emitting stocks an investor holds, the less willing she is to hold additional carbon-emitting stocks. The variable $X^{(1)}$ measures the relative contribution of each asset to this total cost — e.g. its carbon emissions — while κ measures the overall willingness to hold carbon emitting stocks. The linear constraint represents hard targets, such as the liquidity ratio that a bank must hold. There, $X^{(2)}$ measures the contribution of each position to the constraint — e.g. its liquidity weight — and Θ is the maximum value capturing the regulatory requirement.

When prices move, the investor balances risk-return considerations with these other non-risk objectives. All these dimensions will shape substitution patterns. We show in Appendix B.4 that the elasticity matrix for this investor satisfies assumptions A1 and A2 with respect to the stacked set of observables $X = [X^{(1)}, X^{(2)}, X^{(3)}]$, where $X^{(3)}$ are the observables necessary to capture the covariance matrix.[24]

Concretely, this implies that the framework accommodates non-risk motives. While the empirical researcher need not know every detail about the origin of these motives, they must take a stand on which observables capture how each asset contributes to them; for example, a firm's carbon emissions for a fund with an ESG mandate.

2.2.3 Experimental design

Instead of imposing structure on all portfolio decisions, researchers can restrict their attention to a subset of assets for which the assumptions are plausible, and for which, as we will see

[24]Clearly, to be able to use our results, the covariance matrix Σ also has to satisfy the assumptions with respect to a set of observables $X^{(3)}$. Specifically, Σ has to be such that the elasticity matrix of the mean-variance problem without constraint and cost function satisfies assumptions A1 and A2.

later on, they have appropriate sources of exogenous variation. This corresponds to choosing judiciously the estimation sample $\mathcal{S}$ in Assumptions A1 and A2. For example, one might opt to focus on a narrow set of highly comparable assets, making the assumptions plausible. Of course, this path reduces how much can be identified about the overall elasticity matrix.

Return spreads or homogeneous estimation sample. A widespread practice in asset pricing to isolate local mechanisms is to focus on the behavior of the spread between the price of two comparable assets. In a demand estimation context, Drechsler et al. (2024) assess the effect of monetary policy on mortgage markets by studying the spread between the mortgage rate and the rate of a Treasury with the same duration. Another example could be multiple corporate bonds from the same issuer with similar maturity (see Coppola (2025)).

Our framework formalizes the implicit assumptions behind this procedure. This corresponds to imposing the conditions on a subsample $\mathcal{S}$ without observables. In this case, assumption A1 implies that the cross-price elasticity is the same for all assets in the estimation sample, while assumption A2 additionally implies that the own-price elasticity is the same for all assets in the estimation sample:

$$\mathcal{E}_{ii} = \mathcal{E}_{own}, \quad \text{for all } i \in \mathcal{S}, \quad \text{and} \quad \mathcal{E}_{ij} = \mathcal{E}_{cross}, \quad \text{for all } i \neq j \in \mathcal{S}. \tag{15}$$

In the simple case of a spread this assumes symmetry in the two legs of the spread, a treated and a control.

Moving on to outside assets, which could be a vast set, the substitution between them and the assets in the estimation sample is generally not constant. Still, assumption A1 also implies that, for each outside asset, all assets in the estimation sample have the same cross-elasticity. In other words, the demand for any asset in the sample responds in the same way to a change in the prices of each outside asset. Figure 2 illustrates such an elasticity matrix. This setting corresponds to a situation in which observables are constant within the estimation sample, while they can vary arbitrarily across assets outside the estimation sample.

In simple risk-based models like in Section 1.3, in which elasticities are proportional to the inverse of the covariance matrix, this means that all assets in the sample, say i and j, have the same variance and covariance with each other; and for any outside asset k, the covariance of its return with that of any asset in the sample is constant: $\text{cov}(R_i, R_k) = \text{cov}(R_j, R_k)$. In practice, outside assets are plentiful and this latter condition cannot be fully assessed. Still, just as one would assess balance on observables (Athey and Imbens, 2017), researchers should present some corroborating evidence that reports average covariances with a set of broad portfolios for treated and control assets (see Section 5.1 for implementation).

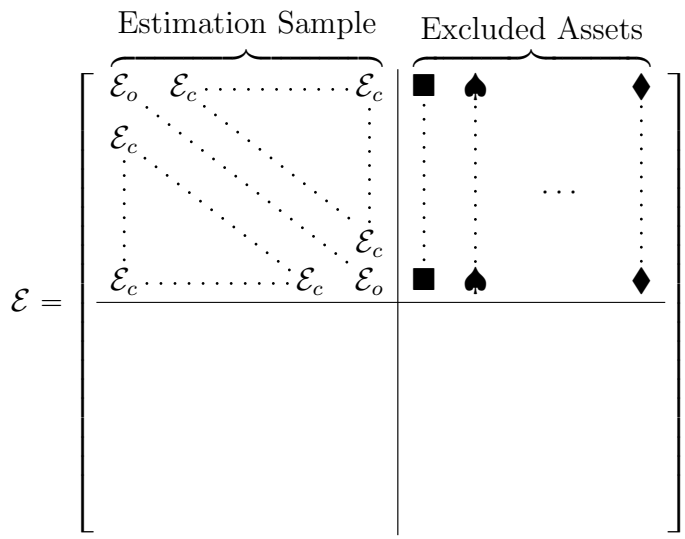


Figure 2: Elasticity matrix satisfying assumptions A1 and A2 for a local experiment.

Groups of assets. In some settings, the researcher can delineate many such subsamples. Even though homogeneous substitution is not plausible across the whole sample, it is plausible within each of the subsamples. For example, homogeneity might hold for a set of firms in a narrow industry but not across these industries. Practically this corresponds to including group dummies as observables. Chaudhary et al. (2022) explain how omitting group effects in such a situation leads to biased inference. They document the relevance of this bias when measuring the effect of fund flows on corporate bond prices.

Mimicking portfolios as synthetic controls. A variation of this approach particularly well-suited for event-study settings is to construct synthetic controls (Abadie et al., 2010) in the style of hedging portfolios. With a set of treated assets, one can construct portfolios of other assets as the control group for a difference-in-difference study. There are two requirements for this approach to be valid. First, the observables or the factor exposures of the control portfolio must be the same as that of the treated asset. Second, each asset in the control portfolio (as opposed to the combined portfolio returns) must have similar residual volatility as the treated asset.

3 Estimation

We show how to estimate the elasticity matrix under the two previous assumptions, with the help of exogenous sources of variation. As a reminder, demand is determined by equation (3)

where the elasticity matrix takes the form highlighted in Section 2.1:

$$\Delta D = \boldsymbol{\mathcal{E}} \Delta P + \epsilon, \quad \text{with } \boldsymbol{\mathcal{E}} = \mathcal{E}_{relative} \mathbf{I} + X \boldsymbol{\mathcal{E}}_X X'. \tag{16}$$

We show separately how to estimate the two components.

3.1 Estimating Relative Elasticity

We show that a simple causal inference regression estimates the relative elasticity consistently. Then, we discuss robustness and extensions of the framework.

3.1.1 Identification using standard cross-sectional approaches

Pretend for a moment that demand for asset i responds to the price of asset i only, not to the price of other assets, $\Delta D_i = \mathcal{E} \Delta P_i + \epsilon_i$, where $\mathcal{E}$ is now a scalar. This is the canonical setting of causal inference, and running an instrumental variable estimation is natural in this situation. In this model, one regresses the change in demand for each asset ΔD_i on the change in the price of this asset ΔP_i, using Z_i as an instrument for the price change. This corresponds to the two-stage least squares specification:

$$\Delta D_i = \widehat{\mathcal{E}} \Delta P_i + \theta' X_i + e_i, \tag{17}$$

$$\Delta P_i = \lambda Z_i + \eta' X_i + u_i, \tag{18}$$

where X_i is the set of observables for each asset.[25] A common case is when the instrument is binary (treatment and control) in which case the estimation is an instrumented difference-in-difference.

The two standard conditions for identification are the relevance and exclusion restrictions. The exclusion restriction captures the idea that the instrument does not affect demand through other channels than the price: $Z_i \perp \epsilon_i | X_i$.[26] In other words, the instrument is not correlated with unobservable shifts in the demand curve in the cross-section of assets. For example, even if the experiment leads to general equilibrium effects such as changing the risk-free rate, the exclusion restriction can still be satisfied if the impact of these effects across assets does not correlate with which asset is treated. Relevance is the idea that the

[25] Throughout the paper, we assume that the matrix X has full column rank, that is, redundant observables have been removed.

[26] The literature sometimes uses variations of this condition. Koijen and Yogo (2019) focus on $E(\exp(\epsilon_i)|X_i, Z_i) = 1$, leading to non-linear estimation. Graves (2025) imposes zero conditional median as opposed to mean, which leads to sequential censored quantile regressions that account for censoring of positions at zero. We favor the linear case for simplicity of exposition and to reflect the most common approach in the empirical literature.

instrument Z_i creates variation in prices: $\lambda \neq 0$. In practice, it is not enough for the first stage to be significant at standard confidence levels; it must be strong to avoid issues related to the weak-instrument problem (Stock and Yogo, 2005; Olea and Pflueger, 2013).

Going back to the realistic case with cross-asset substitution, this estimation method seems to ignore the effect of other assets. The next proposition shows that the naive estimator correctly estimates the relative elasticity under our assumptions.

Proposition 1 *Under assumptions A1 and A2, as well as the standard relevance and exclusion restrictions, the two-stage least squares estimation of equations* (17) *and* (18) *identifies the relative elasticity:*

$$\widehat{\mathcal{E}} = \mathcal{E}_{relative}. \tag{19}$$

When the IV estimation is well specified, it identifies the relative elasticity: the difference between the own-price elasticity and the cross-price elasticity for two assets with the same observables. While this result stands in contrast to the intuition of measuring "how demand for each asset responds to its own price," it is natural. A cross-sectional regression is a comparison across assets in the sample. Even if only the price of the treated asset is shocked, the regression coefficient will still be driven by the response of demand for this asset relative to that for the comparable control asset—hence the relative intensity of the own- and cross-elasticity conditional on observables. In other words, $\widehat{\mathcal{E}}$ answers the question: how does the demand for one asset relative to another comparable asset respond to the relative price of these assets?

Proof for the simple case. Appendix A.1 proves Proposition 1. To understand the mechanics of this result, consider a simple case with two assets with the same observables ($X_1 = X_2$) and no demand shifts ϵ. The changes in demands are:

$$\Delta D_1 = \mathcal{E}_{11}\Delta P_1 + \mathcal{E}_{12}\Delta P_2 + \sum_{k>2} \mathcal{E}_{1k}\Delta P_k; \tag{20}$$

$$\Delta D_2 = \mathcal{E}_{22}\Delta P_2 + \mathcal{E}_{21}\Delta P_1 + \sum_{k>2} \mathcal{E}_{2k}\Delta P_k. \tag{21}$$

Assumption A1 implies that the cross-elasticities with respect to other assets ($k > 2$) are identical:

$$\sum_{k>2} \mathcal{E}_{1k}\Delta P_k = \sum_{k>2} \mathcal{E}_{2k}\Delta P_k. \tag{22}$$

When computing the difference $\Delta D_1 - \Delta D_2$, this response to other prices disappears, effectively removing the omitted variable problem due to other assets:

$$\Delta D_1 - \Delta D_2 = (\mathcal{E}_{11} - \mathcal{E}_{21})\Delta P_1 - (\mathcal{E}_{22} - \mathcal{E}_{12})\Delta P_2. \tag{23}$$

Assumption A2 implies that the coefficients on each of the prices are the relative elasticity:

$$\mathcal{E}_{11} - \mathcal{E}_{21} = \mathcal{E}_{22} - \mathcal{E}_{12} = \mathcal{E}_{relative}. \tag{24}$$

Both the response of demand to the own price (measured by $\mathcal{E}_{11}$) and the response to the price of the other asset (measured by $\mathcal{E}_{21}$) shape this comparison. Hence, the regression coefficient is the relative elasticity:

$$\widehat{\mathcal{E}} = \frac{\Delta D_1 - \Delta D_2}{\Delta P_1 - \Delta P_2} = \mathcal{E}_{relative}. \tag{25}$$

The role of observables. In the richer case where observables differ across assets, it becomes important to control for observables. Two different assets respond differently to the price of other assets. However, assumption A1 ensures that these responses only depend on each asset's observables X_i:

$$\begin{aligned}
\Delta D_i &= \mathcal{E}_{ii}\Delta P_i + \sum_{j\neq i} X_i' \boldsymbol{\mathcal{E}}_X X_j \Delta P_j + \epsilon_i && (26)\\
&= \left(\mathcal{E}_{ii} - X_i'\boldsymbol{\mathcal{E}}_X X_i\right)\Delta P_i + \sum_{j} X_i'\boldsymbol{\mathcal{E}}_X X_j \Delta P_j + \epsilon_i && (27)\\
&= \underbrace{\left(\mathcal{E}_{ii} - X_i'\boldsymbol{\mathcal{E}}_X X_i\right)}_{\mathcal{E}_{relative}}\Delta P_i + X_i' \underbrace{\sum_{j} \boldsymbol{\mathcal{E}}_X X_j \Delta P_j}_{\text{constant across assets}} + \epsilon_i. && (28)
\end{aligned}$$

The second term in (28) highlights that substitution, while depending on all other prices, is proportional to X_i. As a result, controlling for X_i in a cross-sectional regression — that is, including a coefficient θ on observables — absorbs the effects of substitution.[27] Furthermore, the first term in (28) shows that the regression (after controlling for X_i) is equivalent to making pairwise comparisons of assets that have the same observables. Hence, following the same reasoning as in the simple case, the estimate $\widehat{\mathcal{E}}$ recovers the relative elasticity.

[27] In fact, this reasoning shows that one can extend the result of Proposition 1 with a weaker form of assumption A1: $\mathcal{E}_{il} = \mathcal{E}_{cross,l}(X_i) = X_i'Y_l$ for arbitrary vectors Y_l. In other words, the dependence to other assets for a given X_i can be arbitrary and does not need to be parametrized by observable characteristics X_l.

What about equilibrium? One might worry that because prices are an equilibrium outcome, a natural experiment used to create an instrument will generate spillovers across assets, and that this threatens estimation. Say for example that the instrument comes from the Fed quasi-randomly purchasing some bonds but not others, as in Selgrad (2023). Even if the Fed does not buy a specific bond, its price might still respond to purchases of its substitutes. This type of spillovers in the first stage in (18) does not affect our identification result. Instrument exogeneity only requires that treatment status (whether the Fed bought the bond or not) is unrelated to shifts in the investor's demand curve such as changes in their preferences or their views about the assets. That is, the instrument Z_i — which, in the context of the structural equation (3), may affect the equilibrium prices of other assets ΔP_j — is orthogonal to the demand shock ϵ_i, conditional on observables X_i. By modeling the full demand elasticity matrix as in equation (16) rather than the elasticity of a single asset in isolation, our approach directly addresses the identification challenge recently raised by Fuchs et al. (2025b).

Under our assumptions, an economically meaningful situation that threatens identification is when assets are redundant and markets are frictionless, such that a no-arbitrage relation perfectly ties their prices together.[28] Fuchs et al. (2025a) also discuss how this case creates challenges for demand estimation. In such a setting, the relevance condition cannot hold because there is no change in the relative price of treated and control, and identification fails. Of course, the relevance condition can and should be assessed empirically in the first stage. In finite sample, it is not enough for the first stage to be significant at standard confidence levels; it must be strong to avoid bias with weak instruments.[29]

The potential for this issue in the first stage highlights a key consideration when selecting an appropriate control group for a given treated asset. While a control asset should be similar enough to the treated to satisfy Assumptions A1 and A2, it should not be identical. Because distinct assets naturally contain idiosyncratic risk, their relative prices can fluctuate without triggering arbitrage opportunities.[30] More broadly, non-risk factors such as regulatory constraints can introduce equilibrium wedges that allow similar assets to be priced differently.

Partial identification. It might be tempting to conclude that the N-dimensional demand curve of equation (3) in which all prices matter for all demands is equivalent to a uni-dimensional demand curve that only depends on the own price and characteristics, as

[28] Here, what is important is that these assets are perfect substitutes at the aggregate level (so that their equilibrium prices are exactly the same), not so much that the investor whose demand is estimated treats them as such.

[29] This can be assessed using the tests of Stock and Yogo (2005) and Olea and Pflueger (2013) for example.

[30] For example, the monthly returns on Ford and GM stocks exhibit a correlation of 0.53 between 2002 and 2025, and their bonds of maturity 3–7 years have a correlation of 0.71, both numbers far from 1.

estimated in the regression equation (17). This would be incorrect: the equivalent representation only holds when fixing a specific vector of prices; in other words, while equilibrium quantities demanded satisfy equation (17), the N-dimensional demand curve does not. To put it more formally, Proposition 1 offers a partial identification result, where the naive estimator recovers the relative elasticity but not substitution. As we showed in Section 1.3, models like logit that have the same structure as (17) cannot capture the substitution of traditional finance models. More generally, Section 3.2 shows that you cannot recover the substitution matrix $\boldsymbol{\mathcal{E}}_X$ from the cross-section alone.

3.1.2 Robustness and extensions

Robustness to deviations from the assumptions. In practice, assumptions A1 and A2 are approximations of reality. In Appendix C, we assess whether the result of Proposition 1 is robust to small deviations. We show that, as long as the first stage is economically large — that is, the instrument induces a substantial cross-sectional spread in prices — the two-stage least squares estimator recovers the relative elasticity up to a bias that is proportional to the distance to the assumptions; small deviations, small bias. For example, a small amount of measurement error in factor loadings is not consequential. Again, the only exception to this rule is with a weak or zero first stage: weak instruments lead to an extreme sensitivity to misspecification (Bound et al., 1995). Complementary to the finite-sample weak-IV tests discussed in the previous section, the relevant diagnostic is the magnitude of the price spread the instrument induces, measured in economic units (Conley et al., 2012; Van Kippersluis and Rietveld, 2018; Andrews et al., 2019). The sensitivity to misspecification with near-redundant assets in frictionless markets highlighted in Fuchs et al. (2025a) is a special case of a weak first stage.

Observed heterogeneity in relative elasticity. We can also entertain deviations from the assumption of constant relative elasticity. Just like Assumption A1 allows cross-elasticities to depend on observables, one can relax Assumption A2 to let the relative elasticity depend on observables. This corresponds to replacing the condition (11) by:

$$\mathcal{E}_{ii} - \mathcal{E}_{cross}(X_i, X_i) = \mathcal{E}_{relative}(X_i) = \mathcal{E}_r' X_i, \tag{29}$$

where $\mathcal{E}_r$ is a vector of dimension K that maps observables to the relative elasticity for asset i. For example, if the observable captures the size of a company, this relation allows big stocks to have a different relative elasticity than small stocks, an approach taken, for example, in Haddad et al. (2025b). Another useful application is in the context of the factor models of

Section 2.2.1. There, we have seen that idiosyncratic volatility controls the relative elasticity. Therefore, one could include the idiosyncratic volatility of each asset as an observable.

Estimating $\mathcal{E}_r$ simply requires regressing relative demand on the price change and its interactions with observable characteristics, instrumenting appropriately. Appendix Section A.5 proves this result and provides all details on implementation.

Unobserved heterogeneity. A more complicated situation arises when there is unobserved variation in relative elasticity across assets, so that each asset has its own $\mathcal{E}_{relative,i}$. Nothing comes for free: accommodating this more flexible elasticity matrix requires stronger assumptions on the instruments to maintain meaningful identification. Specifically, the assignment of instruments across assets should not be systematically related to the heterogeneity in relative elasticities. An example of this is index inclusions: around the inclusion threshold, included and excluded assets are closely related in size but might randomly differ in industries or other characteristics.

Appendix A.6 formalizes conditions to estimate relative elasticities in this case where Assumption A1 holds but Assumption A2 does not. The two-stage least squares estimation identifies a weighted average of the relative elasticities:[31]

$$\widehat{\mathcal{E}} = \frac{\mathbf{E}_i[\lambda_i \mathcal{E}_{relative,i}]}{\mathbf{E}_i[\lambda_i]}, \tag{30}$$

where λ_i is the residualized first stage effect of the instrument on asset i's price. The relative elasticities of assets with a high λ_i are given a greater weight. If more illiquid assets have both a higher impact of the instrument λ_i and investors trade them more inelastically (smaller $\mathcal{E}_{relative,i}$), estimates of relative elasticity will be lower than the unweighted average relative elasticity, and overstate how inelastic the typical asset is.

3.2 The Missing-Coefficient Problem

Standard cross-sectional causal inference can estimate the relative elasticity, a useful moment for answering micro questions comparing individual assets. There are many other interesting questions concerning the effect of changes in asset prices. For instance, how do investors rebalance when the price of small stocks changes relative to big stocks, or when the price of long-duration bonds changes relative to short-duration ones? At the most aggregated level, how do investors respond when all stocks become more expensive?

[31] Formally, the estimator yields a local average treatment effect with an added monotonicity condition that the relative impact of the instrument always affects prices in the same direction — all λ_i sharing the same sign. See discussion in Appendix A.6.

Addressing these questions hinges on estimating cross-price elasticities separately from relative (or own-price) elasticities. Under assumptions A1 and A2, this corresponds to estimating substitution $\boldsymbol{\mathcal{E}}_X$. We show that, unlike for the relative elasticity, observing a single cross-section does not allow identification of the substitution matrix. This identification challenge sheds light on how leading structural demand frameworks achieve tractability by implicitly assuming away this missing coefficient problem.

3.2.1 Substitution cannot be estimated using the cross-section

The source of this limitation is the classical challenge of demand estimation: substitution is a response to change in prices along the observables, but demand might also shift along the observables. Formally, the demand shifters ϵ can in general be correlated with the observable X: the investor's demand for assets with high values of X relative to low values of X can change, even holding prices constant. For example, the investors might become more optimistic about the relative prospects of tech firms relative to industrial firms.

Denote $\epsilon_X = (X'X)^{-1}X'\epsilon$ the projection of demand shifts on the observables and $\epsilon_{idio,i} = \epsilon_i - X_i'\epsilon_X$ the residual from this projection. We then can rewrite equation (28) as:

$$\Delta D_i = \mathcal{E}_{relative}\Delta P_i + \epsilon_{idio,i} + X_i' \underbrace{(\boldsymbol{\mathcal{E}}_X X'\Delta P + \epsilon_X)}_{\text{coefficient } \theta \ (K\times 1)}. \tag{31}$$

Notice that the effect of substitution $\boldsymbol{\mathcal{E}}_X X'\Delta P$ always appears together with the demand shifter ϵ_X. Furthermore, in a single cross-section, irrespective of the number of assets, there is only a single vector θ for the combination of these two quantities, and hence they cannot be separated using shocks to one but not the other. Appendix Section A.3 proves this impossibility result:

Proposition 2 *Under assumptions A1 and A2, no free parameter of the matrix $\boldsymbol{\mathcal{E}}_X$ can be identified from a single cross section, even under the restriction that some of the coefficients of $\boldsymbol{\mathcal{E}}_X$ are 0.*

This observation is the flip side of the result that facilitated the estimation of relative elasticity: the effect of substitution is always proportional to the observables and is absorbed by the coefficients θ in a cross-sectional regression, as shown by the second term in (28). Demand shifts proportional to the observables also shape the coefficient θ. When the observable X is a constant, this issue is the classic missing intercept problem: substitution is absorbed in the intercept of the cross-sectional regression, but so are aggregate shifts in demand. When substitution depends on other variables, this is a "missing coefficient" problem.

While the well-known missing intercept problem implies that cross-sectional regressions cannot identify aggregate effects (e.g., Guren et al., 2021; Wolf, 2023; Gabaix and Koijen, 2024), our result reveals a much more pervasive blindness. It demonstrates that even the relative forces driving substitution across assets are absorbed by the slope coefficients, rendering them indistinguishable from demand shifts in standard cross-sectional estimation. In other words, the cross-section is not enough to answer questions about the cross-section.

3.2.2 Leading structural models of portfolio decisions and the missing coefficient problem

Leading structural portfolio models such as the logit demand used in Koijen and Yogo (2019) circumvent the missing coefficient problem. While this feature is an attractive property for tractability, we show it precludes these models from capturing the standard substitution based on risk we explored in Section 1.3. Denote by ω the vector of investor's portfolio shares; in units of log portfolio share on log price, a logit demand system implies an elasticity matrix of the form

$$\boldsymbol{\mathcal{E}}^{logit} = -\alpha \mathbf{I}_N + \alpha \mathbf{1}\omega'. \tag{32}$$

This expression implies that logit satisfies Assumptions A1 and A2, where the observables are initial portfolio weights and a constant.[32] Proposition 1 applies, and the cross-sectional estimation identifies the relative elasticity $\mathcal{E}_{relative} = -\alpha$. Because α is the only parameter governing substitution, this cross-sectional estimate pins down the entire elasticity matrix under the logit specification (see equation (32)); there is no missing coefficient problem.

However, comparing this elasticity matrix to the one in the standard model of Section 1.3 (equation (6)) reveals that the logit's restriction on substitution prevents the model from generating the behavior of investors in that workhorse finance theory. As our example demonstrated, a drop in total supply decreases the risk premium, and the required return on riskier assets decreases more than that of less risky assets; otherwise, investors with standard utility functions would not continue to hold the supply of assets. This implies that investors' substitution matrix is sensitive to risk, i.e., depends on β in our example. The logit model does not offer a degree of freedom to accommodate this dependence (even with an extension including an additional parameter for the macro elasticity as in Gabaix and Koijen (2021)).[33]

[32] Formally, the observables for asset i are a constant, $X_i^{[1]} = 1$, and the investor's initial portfolio weight in the asset, $X_i^{[2]} = \omega_i$. Then the substitution matrix is $\boldsymbol{\mathcal{E}}_X = \begin{pmatrix} 0 & \alpha \\ 0 & 0 \end{pmatrix}$.

[33] In a logit asset demand system, substitution operates via the outside asset (the risk-free asset), and it is proportional to investors' pre-determined portfolio shares $\{\omega_i\}$. More precisely, given some counterfactual

For a consistent estimation of the demand elasticity, the extent of this sensitivity still has to be estimated, and the missing coefficient problem arises.

3.2.3 The missing coefficient problem in nonlinear models

The missing coefficient problem is a property of substitution rather than of our linear specification. Whenever substitution depends on observables, variations in the cross-section cannot tell it apart from demand shifts that depend on those same characteristics. For example, even if the cross-elasticity function $\mathcal{E}_{cross}(X_i, X_j)$ is nonlinear, the total impact of substitution is a fixed function $f(X_i)$ in a single cross-section, $f(\cdot) = \sum_j \mathcal{E}_{cross}(\cdot, X_j)$. Similarly demand shifts can be projected on X_i: $\epsilon_i = g(X_i) + \xi_i$ with $g(X_i) = E(\epsilon_i|X_i)$. If the function g is unknown, substitution f cannot be separated from the impact of demand shifts g.

This issue also arises in models of market-level demand for differentiated products like the mixed logit of Berry et al. (1995), where a single time period corresponds to a single market. To relate to our framework, start from the key idea of Berry (1994) inversion to rearrange the model in terms of unobserved utility. Hence, estimation involves separating shifts in demand (mean utility) from cross-product effects (encoded by the inverse share function) mediated by characteristics.[34] Our results highlight that estimation in a single market (Berry et al., 2004) with variation in characteristics depends crucially on functional form restrictions on mean utility: assuming a linear dependence to characteristics while substitution stays nonlinear. This is visible even in the instruments designed for that framework: linear BLP instruments as well as the differentiation instruments of Gandhi and Houde (2019) reduce within a market to functions of a product's own characteristics, and only those with structure different from shifts in mean utility help identification.[35]

3.3 Estimating Substitution

In light of the missing-coefficient problem, we turn to the time series to estimate substitution. We show how to estimate the substitution matrix $\boldsymbol{\mathcal{E}}_X$ with $K + 1$ sources of variation in the

price vector $\mathbf{P}$, the resulting demand in terms of portfolio share satisfies $\log \omega_i(\mathbf{P}) = \log \omega_0(\mathbf{P}) + \hat{\mathcal{E}} \ln P_i + \theta' X_i$, where X_i potentially includes the risk exposure β_i. Therefore, under a logit asset demand system the movement of risk premium (reflected by a change in the vector $\mathbf{P}$) only affects the average of portfolio demand (the intercept) but not how it depends on the risk exposure (the coefficient θ). Capturing changes in portfolio dependent on β requires allowing a dependency $\theta(\mathbf{P})$.

[34] The generic structure of these models (Berry and Haile, 2021) is a product-specific mean utility $\delta_j = g(X_j, p_j) + \xi_j$, and an invertible share function determining market shares $s = \sigma(\delta, X)$. Then estimation relies on writing $\xi_j = -g(X_j, p_j) + \sigma_j^{-1}(s, X)$. See Gandhi and Nevo (2021) for an intuitive explanation of cross-sectional identification in these models.

[35] The central role of functional form assumptions to go around the missing coefficient problem also helps understand the importance of using optimal instruments (Reynaert and Verboven, 2014).

time series: one source for each observable driving substitution, plus one for the overall aggregate.

3.3.1 Decomposing substitution

To develop an identification strategy for substitution, we derive a convenient decomposition. In standard finance models like APT (Ross, 1976) or ICAPM (Merton, 1973), a common insight is that one does not need to look at every asset independently but can focus on a few curated portfolios. While we entertain a much richer class of demand functions, a similar property arises in our framework. A few portfolios are enough to capture the effect of substitution.

Proposition 3 *Take an elasticity matrix $\boldsymbol{\mathcal{E}}$ satisfying assumptions A1 and A2, and consider generic changes in demand and price connected by this matrix $\Delta D = \boldsymbol{\mathcal{E}} \Delta P$. Define the change in demand and price aggregated along observables (vectors of size K) and the idiosyncratic counterparts (scalars) as:*

$$\Delta D_X = (X'X)^{-1} X' \Delta D, \qquad \Delta P_X = (X'X)^{-1} X' \Delta P, \tag{33}$$

$$\Delta D_{idio,i} = \Delta D_i - X_i' \Delta D_X, \qquad \Delta P_{idio,i} = \Delta P_i - X_i' \Delta P_X, \tag{34}$$

such that $\Delta D = \Delta D_{idio} + X \Delta D_X$ and $\Delta P = \Delta P_{idio} + X \Delta P_X$. The response of changes in demand to a change in prices can be decomposed into two sets of components:

$$\text{Micro:} \qquad \Delta D_{idio,i} = \mathcal{E}_{relative} \Delta P_{idio,i}, \tag{35}$$

$$\text{Meso-Macro:} \qquad \Delta D_X = \breve{\boldsymbol{\mathcal{E}}} \Delta P_X, \tag{36}$$

where the $K \times K$ matrix $\breve{\boldsymbol{\mathcal{E}}} = \mathcal{E}_{relative} \mathbf{I}_K + \boldsymbol{\mathcal{E}}_X X'X$.

The response of demand to prices can be decomposed into two components. The idiosyncratic part of the response of demand to prices is driven by the relative elasticity discussed above in Section 3.1. The other component of the response corresponds to substitution effects and is summarized by the response of a few broad portfolios. The vectors ΔP_X and ΔD_X represent the return and the demand response for long-short portfolios along the observables X. Specifically, these are the coefficients from a regression of ΔD or ΔP on X. Why can we interpret these coefficients as coming from portfolios? This insight goes back to Fama and MacBeth (1973), who point out that the $(X'X)^{-1}X'$ can be interpreted as a set of portfolio weights based on the variables in X. Just as in this classic method, the decomposition holds irrespective of whether the observables X vary over time or not.

Two examples help gain intuition on how the decomposition works. First, consider a setting where there is no variation in observables, that is when the only variable in X is a constant as in Gabaix and Koijen (2021). Then, ΔD_X and ΔP_X are of dimension 1 and represent the average demand and return across assets: $\Delta D_X = N^{-1}\sum_i \Delta D_i$ and likewise for ΔP_X. Then the scalar $\breve{\mathcal{E}}$ in equation (36) is the macro elasticity, how aggregate demand for assets responds to the aggregate return.

Now, in addition to the constant $X^{[1]} = 1$, add heterogeneous substitution along one variable $X^{[2]}$ assumed to have mean zero and unit standard deviation. This adds an additional component to $\Delta D_{X^{[2]}}$ and $\Delta P_{X^{[2]}}$. These quantities measure how the price and the demand for assets with larger values of X change relative to those with lower values. Specifically, $\Delta P_{X^{[2]}} = N^{-1}\sum_i X_i^{[2]}\Delta P_i$ is the change in price of a long-short portfolio sorted on this characteristic; $\Delta D_{X^{[2]}} = N^{-1}\sum_i X_i^{[2]}\Delta D_i$ is the change in portfolio tilt along the characteristic (see Appendix A.4). For example, if X measures the duration of bonds, $\Delta P_{X^{[2]}}$ is the return of a portfolio that goes long high-duration bonds and short low-duration bonds; likewise, $\Delta D_{X^{[2]}}$ measures the portfolio tilt relative to the average duration. This additional component highlights again that heterogeneous substitution cannot be captured with only relative (or micro) and macro elasticity but instead adds an additional intermediate "meso" layer to how demand responds to prices. The formulas in equations (33)-(36) generalize this example to multiple observables.

3.3.2 Using the time series to estimate substitution

The decomposition shows that the effect of substitution is characterized through the behavior of a few time series (as many as the number of observables K), which are the meso and macro aggregators in ΔD_X and ΔP_X. However, to estimate substitution, the researcher must still contend with the presence of shifts in demand that are correlated with the price:

$$\Delta D_{X,t} = \breve{\boldsymbol{\mathcal{E}}}\Delta P_{X,t} + \epsilon_{X,t}. \tag{37}$$

The next proposition shows how to identify $\breve{\boldsymbol{\mathcal{E}}}$ with a set of time series instruments.

Proposition 4 *Consider a set of K time series instruments $Z_{X,t}$ for $\Delta P_{X,t}$, such that the exclusion restriction, $Z_{X,t} \perp \epsilon_{X,t}$, and the relevance condition,* $\operatorname{rank}(\operatorname{cov}(Z_X, \Delta P_X)) = K$,

both hold. Under assumptions A1 and A2, the set of K two-stage least squares regressions:

$$\Delta D_{X^{[k]},t} = \sum_l \breve{\boldsymbol{\mathcal{E}}}_{k,l} \Delta P_{X^{[l]},t} + e^{[k]}_{X,t}, \tag{38}$$

$$\Delta P_{X^{[k]},t} = \sum_l \boldsymbol{\Lambda}_{k,l} Z_{X^{[l]},t} + u^{[k]}_{X,t}, \tag{39}$$

correctly identifies the matrix $\breve{\boldsymbol{\mathcal{E}}}$.[36] *Combined with an estimate $\widehat{\mathcal{E}}$ of $\mathcal{E}_{relative}$, this provides an estimate of the substitution matrix $\boldsymbol{\mathcal{E}}_X = (\breve{\boldsymbol{\mathcal{E}}} - \widehat{\mathcal{E}}\mathbf{I}_K)(X'X)^{-1}$.*

Proposition 4 explains that substitution is identified from regressing the portfolio tilt along each observable ($\Delta D_{X^{[k]}}$) on the change in price of all portfolios sorted on observables (ΔP_X), instrumented by the variables (Z_X). In this regression, the notion of instrument and exogenous variation comes from the time series. That is, the researcher needs a set of variables that move prices but are orthogonal with demand shifts. The use of such instruments is the solution to the missing coefficients problem introduced in Section 3.2, because they separate the effects of substitution on demand from shifts in demand for the observables.

Take the example of bond demand where the researcher focuses on a single observable: duration. Estimating substitution requires running two distinct regressions: one measuring the sensitivity of aggregate demand to the returns of both an aggregate and a duration-tilted portfolio, and another measuring the sensitivity of duration demand to these same returns. Both price variables are required in each regression to capture heterogeneous substitution. For example, if all bonds become more expensive, the investor might be less inclined to invest in a duration-tilted portfolio. To consistently estimate these coefficients, the researcher must employ two separate instruments, one for each endogenous price variable. One could use surprises to government debt issuance or the central bank's overall asset purchases, both in overall quantity (e.g., Greenwood and Vayanos, 2014; Krishnamurthy and Vissing-Jorgensen, 2011) or changes in the duration of the purchases (e.g., Hubert de Fraisse, 2022). If one is only interested in the macro multiplier, it is tempting to simplify the analysis by using a single instrument for the aggregate price and controlling for the price of duration. However, our analysis highlights that this is a case of a bad control situation which introduces endogeneity (Angrist and Pischke, 2009).

This result highlights a fundamental tension between estimating relative elasticity and

[36] The estimation equations also take the following matrix form:

$$\Delta D_{X,t} = \breve{\boldsymbol{\mathcal{E}}} \Delta P_{X,t} + e_{X,t} \tag{40}$$

$$\Delta P_{X,t} = \boldsymbol{\Lambda} Z_{X,t} + u_{X,t}. \tag{41}$$

substitution as the set of observables expands. As discussed in Section 3.1, incorporating a rich set of observables can strengthen causal inference for relative elasticities by improving robustness. By contrast, estimating substitution becomes increasingly demanding as dimensionality grows, since each additional observable requires its own source of exogenous variation. In asset classes with a low-dimensional factor structure, such as Treasuries, this tension is not an issue. In richer environments like equities, however, the researcher must impose structure to keep the problem tractable. A useful guiding principle is to allow heterogeneity along the dimensions that are central to the question at hand and those that feature prominently in the broader literature. For example, when studying the impact of ESG preferences, it is important to distinguish investors' willingness to adjust their green tilt in response to changes in the green-brown premium — an economically natural margin that is typically abstracted from in existing estimates.[37] An alternative strategy explored in the literature is to rely on automatically generated instruments, such as granular instrumental variables (Gabaix and Koijen, 2024), which offer a different way of addressing the challenge of dimensionality.

Falsifying the assumptions. Beyond arguing ex ante for Assumptions A1 and A2, the researcher can subject the framework to falsification tests after estimation. We describe three diagnostics, each flagging a different way the specification might miss a relevant dimension of substitution. A first family of tests, in the spirit of Hausman (1978) and Mundlak (1978), checks whether the estimated relative elasticity is invariant to the source of variation used to identify it. For instance, one can repeat estimation restricting attention to within-issuer variation rather than exploiting the full cross-section of corporate bonds. If the substitution structure is correctly specified, alternative estimators should agree; a meaningful divergence indicates that unmodeled drivers of substitution contaminate the regression. A second test targets a specific candidate: choose an additional characteristic that might drive substitution, add it to X, and re-estimate. If this inclusion alters estimates, that is direct evidence that the candidate characteristic does shape substitution and should be modeled explicitly. A third test addresses the structure of substitution itself: the linear-in-observables form $X\boldsymbol{\mathcal{E}}_X X'$ delivers overidentifying restrictions. Estimating meso responses non-parametrically for separate buckets of X should reproduce the linear pattern implied by the framework, and departures reveal non-linearities in substitution that the specification as written cannot capture. We implement all three diagnostics for the corporate bond market in Section 5.

[37] Neither existing demand system estimates (Van der Beck, 2021; Koijen et al., 2023) nor estimates based on classic finance models allow for this force (Berk and Van Binsbergen, 2025).

An alternative approach: assuming symmetric groups. As discussed at the end of Section 2.2.3, one way to ensure assets are comparable is to bin them into groups using dummy variables. Without additional assumptions, however, estimating substitution within this framework would require as many instruments as there are groups. A common alternative to our linear-in-observables approach is to assume symmetry across the groups: the substitution matrix $\boldsymbol{\mathcal{E}}_X$ has constant diagonal and off-diagonal components. In other words, the matrix $\boldsymbol{\mathcal{E}}_X$ (as opposed to $\boldsymbol{\mathcal{E}}$) satisfies Assumptions A1 and A2 with only a constant as the observable; there is homogeneous substitution *across* group-level portfolios. Chaudhary et al. (2022) estimate such a two-tier structure, and the nested logit model is another example.[38] In such a setting, a single source of exogenous variation in the cross-section of groups permits estimation of the relative elasticity across groups.

While this symmetry assumption improves tractability, it sacrifices the heterogeneous substitution essential to finance, a feature our model is designed to capture. Consider, for instance, a setting where investors manage portfolio duration, meaning duration itself drives substitution across bonds. In this context, if one groups bonds by duration, Assumptions A1 and A2 are plausible at the asset level (using group dummies as observables), but not at the group level: substitution between 1–5 year bonds and 5–10 year bonds differs from substitution between 1–5 year bonds and 15–20 year bonds.

4 Price Impact

In some contexts, the researcher is not interested separately in the demand of each investor, but instead finds it sufficient to understand the aggregate demand for assets. The price impact of a shift in demand is the inverse of the slope of the aggregate demand curve, $\boldsymbol{\mathcal{M}} = -\boldsymbol{\mathcal{E}}_{agg}^{-1}$ (see Section 1.1).

In this section, we first explain how the direct estimation of price impact is different from that of demand elasticity through the simple case of one asset. Then, we show how to estimate the multiplier matrix $\boldsymbol{\mathcal{M}}$, in particular how to do so accounting for spillovers to other assets (the off-diagonal elements of $\boldsymbol{\mathcal{M}}$) by relying on assumptions A1 and A2.

Simple causal inference of price impact. Price impact measures how much prices change in response to an exogenous shift in demand curve. Because in equilibrium aggregate demand does not change if assets are in fixed supply, the empirical setup differs from that of demand estimation. To understand the basic intuition, start with one asset to put aside issues of substitution. While equilibrium demand is fixed, it is possible for demand curves to

[38] See Fang (2023) and Koijen and Yogo (2024) for variations of the nested logit structure.

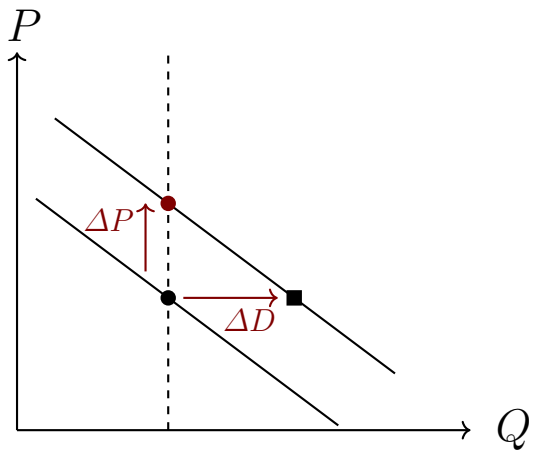


Figure 3: Equilibrium Effect of an Upward Shift in Demand Curve

shift; and we are interested in measuring the impact of such a shift. An idealized example would be an investor waking up in the morning and deciding to buy one share of Apple for no specific reason. Then, the aggregate demand curve for the asset shifts to the right by one unit. In equilibrium, the price must adjust upwards to satisfy market clearing. Figure 3 illustrates this process. Similarly supply shocks can be viewed as the negative of a demand shock and be treated likewise.

In practice, the econometrician starts from a shift of the demand curve Z_i (measured as the actual amount); examples of such shocks from the literature include asset purchases by central banks or rebalancing due to flows in and out of mutual funds (Lou, 2012). Armed with this shift in demand, we run the regression:

$$\Delta P_i = \widehat{\mathcal{M}} Z_i + \epsilon_i, \qquad Z_i \perp \epsilon_i. \tag{42}$$

The exclusion restriction $Z_i \perp \epsilon_i$ requires that the change in demand under consideration must be orthogonal to any other demand shifts in the economy. For example, if a group of investors systematically mimics the Fed's asset purchases, exogeneity is violated and the regression will be biased; the measured shock undercounts the actual change in demand, and overestimates the price impact.

There is no first stage because Z_i directly measures the magnitude of the shift in the demand curve. This shift does not materialize in equilibrium quantity demanded: prices adjust so that the total quantity demanded stays equal to the fixed supply. Equivalently, this identification condition corresponds to assuming that, if one could measure quantities before prices adjust — the out-of-equilibrium square in Figure 3 — the first-stage coefficient

would be one.[39]

Handling substitution. Stepping away from the simplistic case of one asset and no spillover, the same issue as for the estimation of demand elasticity arises: all prices are determined together in equilibrium. There is no such thing as "the multiplier" but instead a matrix $\boldsymbol{\mathcal{M}}$ of own-demand and cross-demand multipliers such that $\Delta P = \boldsymbol{\mathcal{M}}\Delta D$, with $\boldsymbol{\mathcal{M}} = -\boldsymbol{\mathcal{E}}_{agg}^{-1}$ (this is a matrix inverse, not element-by-element, hence own-price multiplier depends on both the own-price and cross-price elasticity). Notice that in a slight abuse of notation, ΔD now represents the shift of the aggregate demand curve.[40] This implies that in response to a vector of shocks Z, price changes will be:

$$\Delta P = \boldsymbol{\mathcal{M}} Z + \epsilon, \tag{43}$$

where ϵ captures the impact of all other demand shocks. The following lemma suggests that the assumptions we introduced for elasticity also help with the estimation of the multiplier.

Lemma 1 *The aggregate elasticity $\boldsymbol{\mathcal{E}}_{agg}$ satisfies assumptions A1 and A2 if and only if the multiplier matrix $\boldsymbol{\mathcal{M}}$ does.*

Thanks to this equivalence, we obtain the mirror image of Propositions 1 and 4 for the multiplier matrix: relative multiplier and spillovers can be estimated using classic cross-sectional shocks, and a set of portfolio-level time-series shocks, respectively.

Proposition 5 *Assume that the multiplier matrix $\boldsymbol{\mathcal{M}}$ satisfies assumptions A1 and A2, and hence can be written as $\boldsymbol{\mathcal{M}} = \mathcal{M}_{relative}\mathbf{I} + X\boldsymbol{\mathcal{M}}_X X'$.*

- Relative multiplier: *If the cross-sectional demand shocks $\{Z_i\}_{i\in\{1,\dots,N\}}$ satisfy the exclusion restriction $Z_i \perp \epsilon_i | X_i$, the cross-sectional regression*

$$\Delta P_i = \widehat{\mathcal{M}} Z_i + \theta' X_i + u_i \tag{44}$$

identifies the relative multiplier $\widehat{\mathcal{M}} = \mathcal{M}_{relative}$, and $\mathcal{M}_{relative} = -1/\mathcal{E}_{agg,relative}$.

[39] If the researcher is willing to make stronger assumptions on which group of investors are affected by the demand shock, they can run a first-stage converting an instrument into a demand shock for that group. However, the stronger assumptions correspond to assuming that the demand shock for the subgroup coincides with the aggregate demand shock, what we refer to as assuming a first-stage coefficient of 1.

[40] That is, ΔD in this section corresponds to the demand shifts aggregated across investors $\epsilon_{agg} = \sum_j \epsilon_j$ in the notation of the previous sections.

- Spillovers: *If the time series demand shocks for each observable* $Z_{X,t} = \{Z_{k,t}\}_{k\in\{1,\ldots,K\},t\in\{1,\ldots,T\}}$ *satisfy the exclusion restriction* $Z_{X,t} \perp \epsilon_{X,t}$*, the set of time-series regressions*

$$\Delta P_{X,t} = \widetilde{\boldsymbol{\mathcal{M}}} Z_{X,t} + u_{X,t} \tag{45}$$

identifies the spillover matrix, with $\boldsymbol{\mathcal{M}}_X = (\widetilde{\boldsymbol{\mathcal{M}}} - \widehat{\mathcal{M}}\mathbf{I}_K)(X'X)^{-1}$.

The relative multiplier measures how the price of one asset relative to another comparable one changes if the relative demand for these assets shifts—how much does the price of Ford change relative to the price of General Motors if the demand for Ford changes relative to the demand for General Motors?[41] Interestingly, in this case the relative multiplier coincides with the inverse of the relative elasticity, $\mathcal{M}_{relative} = -1/\mathcal{E}_{agg,relative}$.[42] In contrast, the cross multipliers captured by the matrix $\boldsymbol{\mathcal{M}}_X$ measure spillovers — how much an increase in demand for one asset affects the price of another asset.

Of course, the researcher should present arguments for the validity of assumptions A1 and A2. The approaches to make this case that we discuss in Section 2 are also relevant here. The next section provides a practical example for the corporate bond market.

5 Empirical Example: Corporate Bond Multipliers

We apply our framework to corporate bonds and estimate the multiplier matrix for that market. First, we choose a source of variation in demand and observables for assumptions A1 and A2, discuss their validity, and conduct pre-estimation diagnostics. Second, we estimate the two components of price impact: relative multiplier and substitution. We conduct a series of specification tests to further evaluate the validity of our setup. Last, we use the estimates to inform counterfactual questions about various designs of central bank interventions in corporate bond markets.

5.1 Identification Strategy and Its Plausibility

Identification of the multiplier relies on combining a source of shifts in demand with assumptions about the structure of the multiplier matrix (A1 and A2). We describe these choices, and discuss their plausibility.

[41] A benefit of focusing on multipliers: the estimation does not become degenerate when the prices of two assets with identical payoffs remain equal. In this setting, the relative multiplier is 0 and the regression can still identify it.

[42] This conjunction occurs despite neither own-price nor the cross-price elasticities being stable by inversion ($\mathcal{M}_{ij} \neq -1/\mathcal{E}_{ij}$); inverting a matrix is different from inverting each of its elements. Gabaix and Koijen (2021) derive this inversion result for the special case without observables shaping substitution.

Data and source of shifts in demand. We study the market for corporate bonds, as in Chaudhary et al. (2022). We obtain data on individual corporate bond returns from the WRDS Bond Returns database between 2010 and 2024 and mutual fund bond holdings and flows from the CRSP Survivor-Bias-Free US Mutual Fund Database. As a source of shifts in demand, we use a refinement of flow-induced trading by mutual funds introduced by Coval and Stafford (2007) and Lou (2012). Flow-induced trading aims to capture a mechanical component of how mutual funds adjust their positions in response to flows. The idea is that funds tend to allocate new inflows towards their larger existing positions, independent of new arrival of information about these positions. The instrument aggregates these shocks across funds:

$$Z_{it} = \sum_k \frac{A_{k,t-1}\widetilde{w}_{i,k,t-1}}{P_{i,t-1}S_{i,t-1}} f_{kt}, \tag{46}$$

where $A_{k,t-1}$ is fund k's assets under management, $\widetilde{w}_{i,k,t-1}$ is the portfolio weight of k on asset i residualized by removing seven common factors from raw weights following Chaudhry (2025), f_{kt} is the flow into fund k at date t, and $P_{i,t-1}S_{i,t-1}$ is the total supply of bond i at $t-1$ (the product of its price and quantity outstanding).

Exogeneity assumption. Exogeneity requires the changes in demand measured by the instrument to be unrelated to any unobserved source of demand — equation (42). Overall mutual fund flows generically do not satisfy this assumption, but the instrument can. Specifically, equation (46) has the structure of a Bartik shift-share instrument: fund flows f_{kt} are the "shifts" and the residualized portfolio weights $\widetilde{w}_{i,k,t-1}$ are the "shares." Goldsmith-Pinkham et al. (2020) show that exogeneity of the shares is sufficient for identification in this setting, without requiring exogenous shifts. That is, exogeneity is satisfied if the weights $\widetilde{w}_{i,k,t-1}$ are uncorrelated with unobserved demand shocks for bond i at time t.

The residualization of Chaudhry (2025) is crucial to argue that the weights are exogenous. If instead, one used actual portfolio shares, the condition could be violated for a couple of reasons. First, households might substitute from holding bonds directly to holding them through mutual funds. Bonds with high mutual fund shares would then mechanically be bonds experiencing offsetting decreases in direct household demand, so that shares correlate with an unobserved demand shift. Second, bonds with high mutual fund ownership might also attract demand from other institutional investors, such as insurance companies, pursuing similar strategies. Then portfolio shares would correlate with unobserved demand from these other investors, biasing the estimated price impact.[43] Empirically, raw portfolio weights

[43]Because the shares are lagged, violation requires persistent correlation between portfolio composition

load on common fund characteristics such as growth, size, and income style, suggesting that fund strategies are more likely to overlap with those of other institutions. Removing many common factors isolates idiosyncratic variation in portfolio composition that is less likely to be shared with other investor types, making exogeneity more plausible. For the estimation of spillovers, exogeneity must also hold in the time series after the aggregation along the observables we introduce next.[44]

Structure of substitution. Our estimation strategy also relies on homogeneous substitution conditional on a set of observables (assumptions A1 and A2).[45] The simplest approach would be to assume unconditional homogeneous substitution. This is unappealing because, if the demand for long-term bonds rises, this will affect the price of other long-term bonds differently from the price of short-term bonds. As discussed in Section 2.2.3, a simple diagnostic for the plausibility of this concern is a test of balance on covariances: do the treated bonds comove in the same way with broad portfolios as control bonds? If they do not, investors would likely substitute these bonds differently.

Figure 4 suggests that they do not. For a given date, we form a long-short portfolio based on whether the instrument Z_{it} is above or below median on that date and compute the beta of this portfolio on a series of broad indices in a two-year range around the date of the sort — the blue dot for that date.[46] Each panel corresponds to a different index: a broad bond index, long-short portfolios based on credit ratings and maturity, and a broad stock index. Bonds with a high instrumented inflow comove more strongly with the credit-rating-sorted portfolio and more weakly with the broad bond index and the duration-sorted portfolio: the blue dots in these panels are systematically away from zero.

It is more plausible to assume homogeneous substitution conditional on duration and credit rating, the two variables that are the most salient characteristics of a corporate bond. Hereafter, we standardize these variables to be mean zero and variance one for each date.[47] We revisit our diagnostic for this case: we form portfolios sorted on $Z_{idio,it}$, the residualized instrument with respect to both duration and credit rating. This corresponds to the orange dots in Figure 4, which are much closer to zero, bolstering the plausibility of our assumptions. After the estimation, we provide tests of overidentifying restrictions to assess further whether

and unobserved demand—less likely than a contemporaneous one, but not ruled out when allocations and demand patterns are slow-moving (e.g., Huebner, 2024).

[44]Exogeneity after aggregation is inherited from the exogeneity at the individual level under a variety of conditions: if weights are not time-varying, if shocks have a factor structure, or if the instruments are centered (Borusyak and Hull, 2023).

[45]The exclusion restriction is also conditional on observables.

[46]Appendix Figure A6 reports exposures for the long and short legs separately.

[47]We use credit ratings as a continuous variable by assigning the historical average 10-year default probabilities to each rating category, from Table 65, column D, of Vazza and Kraemer (2013).

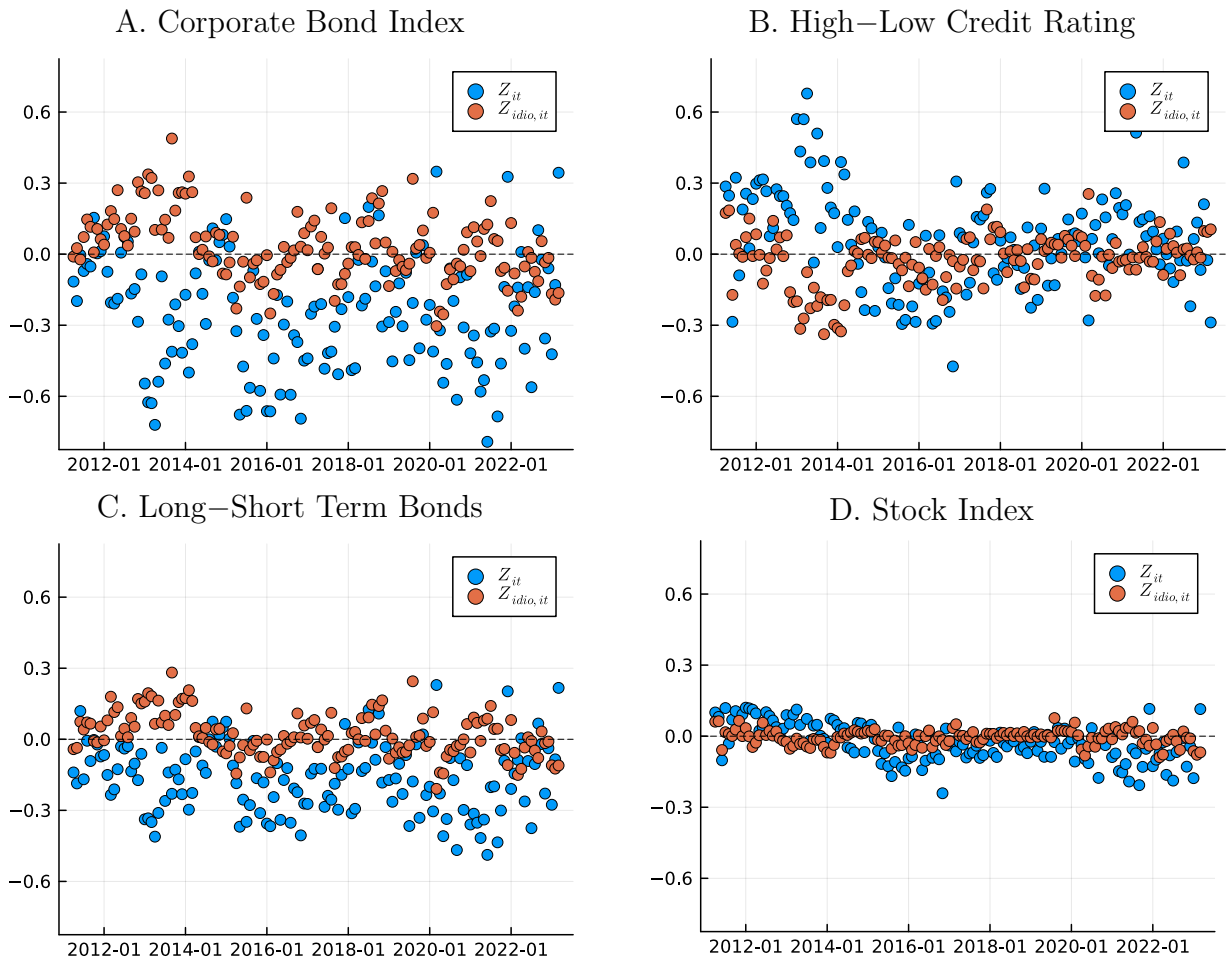


Figure 4: **Balance on covariances: exposure of portfolios sorted on the instrument.** Each dot represents the factor exposure of a portfolio sorted on the values of the instrument Z_{it} (blue) or the residualized instrument $Z_{idio,it}$ at the corresponding month between 2011:04 and 2023:03. Equation (46) defines Z_{it}, while $Z_{idio,it}$ is obtained by cross-sectionally orthogonalizing the instrument at each date to duration and credit risk (long-run average default probabilities based on S&P ratings). At each date, we form long–short equal-weighted portfolios based on whether the value of the instrument for that bond is above or below the median. We compute the returns of these portfolios over two years centered around t, excluding t, and regress separately these returns on four aggregate factors to obtain the factor loadings represented on the figure. Panel A reports exposures to an aggregate investment-grade corporate bond factor, the total return of the ICE BofA US Corporate Index. Panel B considers a credit-sorted factor, the difference in total return between the ICE BofA US High Yield Index and the ICE BofA US Corporate Index. Panel C uses a duration-sorted factor, the difference in total return between the ICE BofA 15+ Year US Corporate Index and the ICE BofA 1-3 Year US Corporate Index. All bond indices are obtained from FRED. Panel D uses the Fama and French (1993) excess stock market return, obtained from Kenneth French's data library.

our framework is misspecified.

Finally, regarding Assumption A2, Appendix Figure A7 shows a similar picture for idiosyncratic volatilities of bonds with high and low values of the instrument. Absent controls for duration and credit rating, these bonds have different idiosyncratic volatilities; after controlling, the volatilities are close to identical.

5.2 Estimates

With these assumptions, we can identify the multiplier matrix by following Proposition 5. For units of change in prices, we use total returns R_{it}, the standard measure in asset pricing with well-behaved statistical properties.[48] We estimate the relative multiplier $\widehat{\mathcal{M}}$ with the cross-sectional estimator:

$$R_{it} = \widehat{\mathcal{M}}\, Z_{it} + \theta_{0,t} + \theta_{DUR,t} DUR_{it} + \theta_{PD,t} PD_{it} + u_{it}. \tag{47}$$

The spillover matrix $\widetilde{\boldsymbol{\mathcal{M}}}$ is obtained from the time-series regressions:

$$R_{agg,t} = \widetilde{\boldsymbol{\mathcal{M}}}_{1,1} Z_{agg,t} + \widetilde{\boldsymbol{\mathcal{M}}}_{1,2} Z_{DUR,t} + \widetilde{\boldsymbol{\mathcal{M}}}_{1,3} Z_{PD,t} + u_{agg,t} \tag{48}$$

$$R_{DUR,t} = \widetilde{\boldsymbol{\mathcal{M}}}_{2,1} Z_{agg,t} + \widetilde{\boldsymbol{\mathcal{M}}}_{2,2} Z_{DUR,t} + \widetilde{\boldsymbol{\mathcal{M}}}_{2,3} Z_{PD,t} + u_{DUR,t} \tag{49}$$

$$R_{PD,t} = \widetilde{\boldsymbol{\mathcal{M}}}_{3,1} Z_{agg,t} + \widetilde{\boldsymbol{\mathcal{M}}}_{3,2} Z_{DUR,t} + \widetilde{\boldsymbol{\mathcal{M}}}_{3,3} Z_{PD,t} + u_{PD,t}, \tag{50}$$

where the subscripts agg, DUR, and PD represent aggregation over all bonds, duration, and credit risk respectively, as in equation (33) (each period, construct $Z_{X,t} = (X_t'X_t)^{-1}X_t'Z_t$).

Relative multiplier. Table 1 estimates the relative multiplier. Column 1 implements the regression of equation (47) to find a relative multiplier $\widehat{\mathcal{M}} \approx 0$, a tightly estimated zero. The estimate suggests that if the demand for one bond relative to another one with the same credit rating and duration increases by 1%, their relative bond price does not change much (the insignificant point estimate corresponds to a response of 5.5bps), as in Chaudhary et al. (2022).

Columns 2 and 3 regress bond returns directly on the residualized instrument $Z_{idio,it}$, with and without controls for the characteristics. This leads to the exact same estimate of the relative multiplier, a mathematical property independent of the specific dataset. This observation highlights that the source of variation for the estimates is variation in the residual component of the instrument $Z_{idio,it}$.

Column 4 estimates the model without controlling for duration and credit risk. This leads to a sizable departure from our baseline estimates, suggesting it is important to account for substitution. Column 5 finds further departure when removing the date fixed effects.

We can also test whether our specification misses some dimensions of substitution. If

[48] It might be tempting to "divide" prices by the observables to remove the role of these variables, e.g. working with changes in yields as opposed to returns. When multiple dimensions drive substitution, no single change of units can make substitution homogeneous. For example, Appendix F.2 reports counterparts of our estimates for yields, and finds that heterogeneity with respect to duration is mitigated in these units but credit ratings remain relevant.

Table 1: **Relative multiplier $\widehat{\mathcal{M}}$ in corporate bonds**

	Return R_{it}				
	(1)	(2)	(3)	(4)	(5)
Demand shock:					
Z_{it}	0.055			0.492***	2.080***
	(0.084)			(0.128)	(0.387)
$Z_{idio,it}$		0.055	0.055		
		(0.084)	(0.087)		
Date Fixed Effects	Yes	Yes	Yes	Yes	
Duration × Date Fixed Effects	Yes	Yes			
Credit Risk × Date Fixed Effects	Yes	Yes			
N	1,041,985	1,041,985	1,041,985	1,041,985	1,041,985
R^2	0.464	0.464	0.293	0.293	0.018

Table 1 reports the results of relative multiplier regressions of bond returns R_{it} on demand shocks Z_{it} and $Z_{idio,it}$ for U.S. non-defaulted corporate bonds. Specifications (1) and (4)–(5) use the flow-induced trading demand shock Z_{it} defined in Equation (46). Specification (1) includes a time fixed effect and controls for a continuous duration variable and a continuous credit risk variable based on average historical default probabilities for each S&P credit rating category, while specification (4) uses only date fixed effects and specification (5) only a common intercept. Specifications (2)–(3) use the demand shock $Z_{idio,it}$ orthogonalized to duration and credit risk each period, with and without controlling for duration and credit risk in the regression. The regressions weigh each date equally. The sample period is 2010:04 to 2024:03. Standard errors are clustered by date and bond issuer.

our model is well specified, including controls for additional observables or changing the level of variation of the data (Hausman, 1978; Mundlak, 1978) would not alter the estimated relative multiplier. Appendix F.1 confirms that controlling for bond liquidity or focusing on within-issuer variation does not.

Spillover matrix. Table 2 presents estimates of the matrix $\widetilde{\boldsymbol{\mathcal{M}}}$ which encodes the meso- and macro multipliers. Columns 1–3 implement equations (48)–(50). Combining these estimates yields

$$\widetilde{\boldsymbol{\mathcal{M}}} = \begin{pmatrix} 14.430 & 0.425 & 1.127 \\ 7.773 & 4.531 & -0.868 \\ 3.764 & -5.789 & 3.383 \end{pmatrix}, \tag{51}$$

and the implied spillover matrix is $\boldsymbol{\mathcal{M}}_X = \left(\widetilde{\boldsymbol{\mathcal{M}}} - \widehat{\mathcal{M}}\mathbf{I}_K\right)(X'X)^{-1}$.

Table 2: **Spillover matrix $\widetilde{\mathcal{M}}$ in corporate bonds**

	Return				
	$R_{agg,t}$	$R_{DUR,t}$	$R_{PD,t}$	R_{it}	$R_{agg,t}$
	(1)	(2)	(3)	(4)	(5)
$Z_{agg,t}$	14.430***	7.773***	3.764**	14.430***	15.032***
	(2.707)	(1.803)	(1.392)	(2.689)	(2.393)
$Z_{DUR,t}$	0.425	4.531	-5.789	0.425	
	(6.713)	(4.293)	(3.948)	(6.646)	
$Z_{PD,t}$	1.127	-0.868	3.383**	1.127	
	(1.866)	(1.404)	(1.025)	(1.854)	
$Z_{idio,it}$				0.055	
				(0.087)	
N	168	168	168	1,041,985	168
R^2	0.343	0.173	0.384	0.100	0.342

Table 2 reports the results of spillover matrix regressions of bond returns on demand shocks for non-defaulted U.S. corporate bonds. Specifications (1)–(3) jointly estimate the matrix $\widetilde{\mathcal{M}}$ from Proposition 5, which together with the relative multiplier $\widehat{\mathcal{M}}$ determines the spillover matrix between observables, $\boldsymbol{\mathcal{M}}_X$. Specification (4) estimates multipliers mechanically identical to specification (1) using disaggregated, repeated cross-sectional regressions, while adding the relative multiplier $\widehat{\mathcal{M}}$. Specification (5) estimates the macro multiplier in isolation by regressing aggregate bond returns $R_{agg,t}$ on the aggregated instrument $Z_{agg,t}$ in the time series. The $K = 3$ observables are a vector of ones, standardized duration, and standardized credit risk. The sample period is 2010:04 to 2024:03. Robust standard errors are used for specifications (1) to (3) and (5). For specification (4), standard errors are clustered by date and issuer, and regressions are weighted such that each date receives equal weight.

Column 4 shows that when using the full panel, rather than time-series regressions, one identifies the exact same meso- and macro-multipliers as well as the relative multiplier (coefficient on $Z_{idio,it}$).[49] This confirms that the panel carries no additional information on spillovers, and underscores the separation result inherent to our structure, whereby the cross-section identifies the relative multiplier, while the time-series identifies spillovers.

The first term in the estimated matrix, $(\widetilde{\boldsymbol{\mathcal{M}}})_{11} = 14.430$, is the aggregate multiplier: a uniform 1% increase in the demand for all bonds leads to a 14% increase in bond prices.[50] The other terms in the first row characterize the average impact of a shock to the relative demand of long-term vs. short-term bonds, or risky vs. safe bonds. For example, a 1%

[49]This property is not specific to the aggregate returns in column 1, but applies more generally to observable-based portfolios in columns 2 and 3. Appendix Table A7 reports this extended specification.

[50]The current literature does not provide estimates at this level of aggregation; existing estimates of portfolio-level multipliers find values between 2.3 and 6 (e.g., Bouveret and Yu, 2021; Chaudhary et al., 2022; Darmouni et al., 2023); see Haddad et al. (2025a) for a review.

increase in the relative demand for risky bonds leads to an insignificant 1.13% increase in bond prices. The second and third rows of $\widetilde{\boldsymbol{\mathcal{M}}}$ capture heterogeneous impacts of demand shocks in the cross section.

Figure 5 builds intuition around these numbers. Panel A reports the impact of an aggregate demand shock on the returns of bonds of various durations. The blue line and shaded area represent estimates and standard errors from an estimate constrained to have constant effects. The orange line is our main estimate: a line combining the level $(\widetilde{\boldsymbol{\mathcal{M}}})_{11} = 14.430$ with the slope $(\widetilde{\boldsymbol{\mathcal{M}}})_{21} = 7.773$. Long-duration bonds respond more to an aggregate demand shock than short-duration bonds, with multipliers ranging from 5 to 25. This heterogeneity is statistically significant; the orange line is different from the blue line. Panel B shows a similar pattern for the response of bonds with different credit risk. The slope of the orange line is now $(\widetilde{\boldsymbol{\mathcal{M}}})_{31} = 3.764$, again statistically and economically significant.

Panels C and D report how demand shocks to long-short portfolios formed on duration and credit risk affect the corresponding portfolio returns, respectively. For duration (panel C), there is a clear monotonic pattern whereby riskier bonds are affected more, but differences are statistically insignificant. For credit risk (panel D), the pattern is similar, and the results are statistically significant.[51]

Finally, we can also assess the validity of our linear specification for spillovers. We estimate the impact of the three demand shocks separately for bonds in different duration or credit risk buckets, the green dots in Figure 5.[52] Under the null of our model, the estimates should line up with the orange line. Linearity holds well across all four panels, suggesting that these overidentifying restrictions stand.

In Section 3.3.2, we pointed out the potential bias when estimating the macro multiplier without accounting for demand shocks along other dimensions. Excluding the instruments Z_{DUR} and Z_{PD} can result in omitted variable bias, if they correlate with Z_{agg}. Empirically, aggregate flows from mutual funds are not uniform across bonds; they tilt towards short-term bonds (Z_{agg} and Z_{DUR} have a correlation of -0.41). Column 5 implements a regression of aggregate returns on aggregate demand only, similar to Gabaix and Koijen (2021). We find an estimate close to our baseline, suggesting the bias is not a concern in this setting.

5.3 Counterfactual Analysis

The estimated multiplier matrix provides answers to many interesting counterfactual questions, with different parts of $\boldsymbol{\mathcal{M}}$ informing outcomes of different policies. For illustration,

[51] The last two dimensions, the impact of duration shocks on prices of credit and vice versa, are in Figure A8.

[52] The exclusion restriction underlying these more granular estimates is more stringent, applying to each bucket individually.

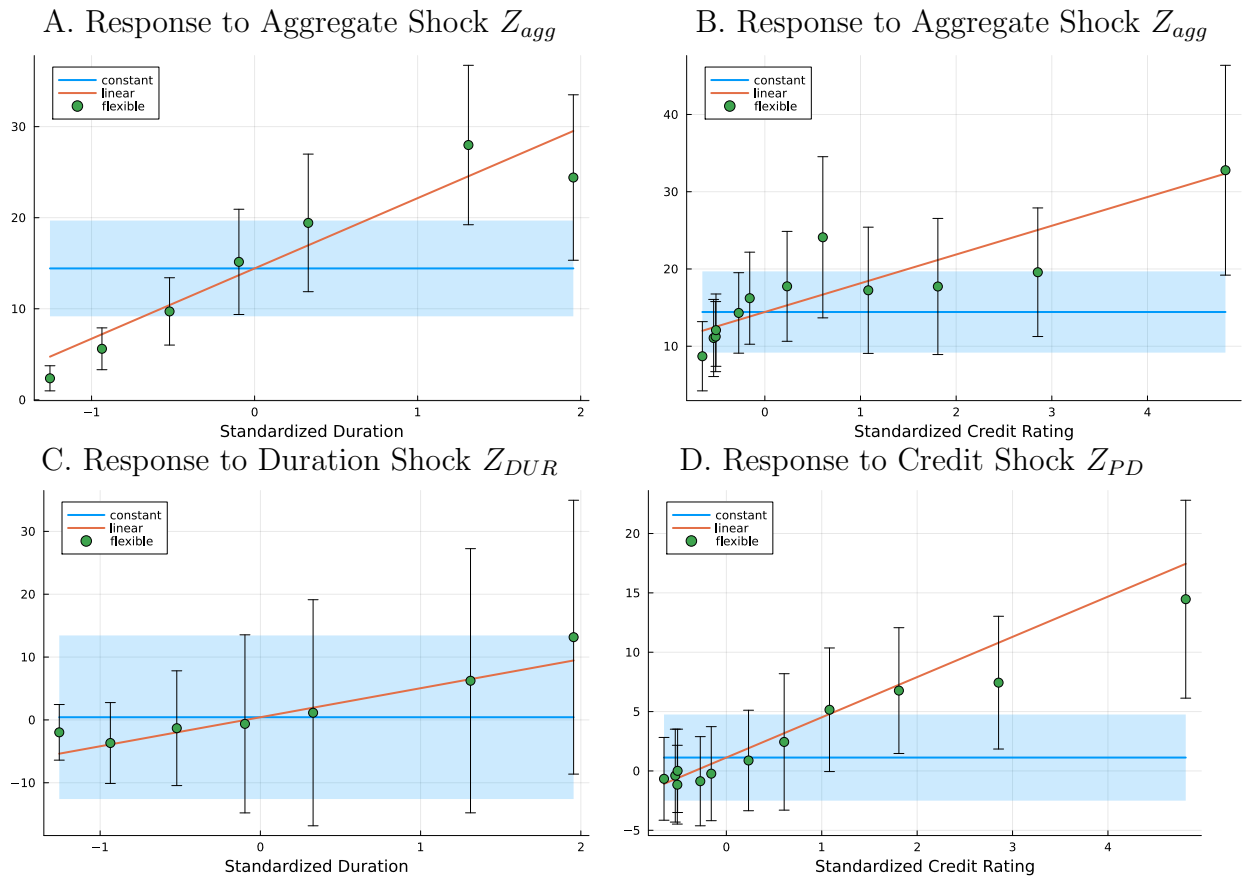


Figure 5: **Meso and macro multipliers in the cross-section.**
We report the response of portfolios of corporate bonds to aggregate demand shocks Z_{agg} (Panels A and B, in the cross-sections of duration and credit rating) and shocks along duration Z_{DUR} (Panel C) and Z_{PD} (Panel D). Bonds are grouped in seven buckets based on duration: <1 year, 1–3 years, 3–5 years, 5–7 years, 7–10 years, 10–15 years, and 15+ years. Bonds are also grouped by S&P credit rating, with individual notches from A+ through B-, and with AAA/AA and CCC/C ratings pooled at the extremes. The blue lines correspond to the estimates from column 1 of Table 2, which assume identical responses. The orange lines incorporate all results from columns 1 to 3, which include linear interaction terms with either duration or credit risk, neutralizing the other. The green dots estimate multipliers separately by duration or credit rating bucket in a pooled panel regression. The sample period is 2010:04 to 2024:03.

consider different implementations of a hypothetical Federal Reserve corporate bond purchase program. From the point estimates of the previous section, the full multiplier matrix can be constructed as $\boldsymbol{\mathcal{M}} = \widehat{\mathcal{M}}\mathbf{I} + X\boldsymbol{\mathcal{M}}_X X'$. Applying this matrix to a vector of interventions yields the full cross-section of price responses.

Relative counterfactuals. Imagine the Federal Reserve conducts an asset purchase program in corporate bonds. Instead of buying all bonds, it randomly selects a subset. This mirrors aspects of the Fed's implementation of QE in Treasury securities (Selgrad, 2023). How will these purchases affect the prices of purchased bonds compared to control bonds with the same duration? The answer depends only on the relative multiplier $\widehat{\mathcal{M}}$. With

$\widehat{\mathcal{M}} \approx 0$, the relative price of the two bonds would not change: substitutability between bonds with the same duration is so strong that buying one bond but not the other moves their prices by the same amount. However, this micro-level comparison does not reveal the total impact of a broad purchase program.

Macro counterfactuals. Consider what happens if the Federal Reserve decides to purchase all corporate bonds. How much would prices change? Unlike in Gabaix and Koijen (2021), the answer is not simply characterized by one macro multiplier; it depends on bond duration and credit rating. For example, consistent with a factor structure based on duration (e.g., Vayanos and Vila, 2021), long-duration bond prices would move more. If the Fed buys 1% of the supply of all corporate bonds, the price of bond i with observables X_i increases by $X_i'\widetilde{\boldsymbol{\mathcal{M}}}(1,0,0)'$ percent.[53] The price of a BBB rated corporate bond with duration of 1 year increases by 5.3%, while another bond with the same credit rating and a duration of 10 years increases by 18.7%. Effects also increase with credit risk; the price of a 3-year AA- bond increases by 6.4%, while that of a B bond increases by 21.8%.[54]

Meso-macro counterfactuals. The Secondary Market Corporate Credit Facility (SMCCF), announced in March 2020, restricted purchases to investment-grade bonds maturing within five years — 36% of bonds were eligible. Suppose the Fed buys 1% of the supply of each eligible bond. This intervention raises prices by 2.92% on 1-year BBB bonds and 5.26% on 10-year BBB bonds; for 3-year AA- bonds the impact is 2.08% and for 3-year B bonds 13.27%.[55]

These numbers are mechanically smaller than in the previous counterfactual because the Fed buys only 36% of the bonds; the cross-sectional patterns also differ. Interestingly, ineligible bonds with high credit-risk exposure see outsized impact: 3-year B bonds reach over 60% of their macro response without being purchased by the Fed. The Fed's concentration on short-term bonds implicitly tilts demand against long duration, which through cross-asset substitution transmits along the credit-risk dimension. Our substitution structure captures this pattern through the point estimate for the off-diagonal entry $\widetilde{\boldsymbol{\mathcal{M}}}_{32} = -5.79$. The duration-credit risk spillover channel of Li (2025) offers a foundation for it: when demand

[53] The calculation simplifies in the case where the demand shock is proportional to the observables, $\Delta D = Xd$, then $\boldsymbol{\mathcal{M}}\Delta D = \boldsymbol{\mathcal{M}}Xd = \left(\widehat{\mathcal{M}}X + X\boldsymbol{\mathcal{M}}_X X'X\right)d = \left(\widehat{\mathcal{M}}X + X\left(\widetilde{\boldsymbol{\mathcal{M}}} - \widehat{\mathcal{M}}\mathbf{I}_K\right)\right)d = X\widetilde{\boldsymbol{\mathcal{M}}}d$.

[54] In the March 2020 cross-section, a 1-year BBB bond has a standardized credit rating of -0.141 and standardized duration of -1.105. A 1% aggregate demand shock moves its price by $14.430 + (-1.105) \times 7.773 + (-0.141) \times 3.764 \simeq 5.3\%$.

[55] Instead of using the full multiplier matrix $\boldsymbol{\mathcal{M}}$, an intuitive approximation projects the demand shock on the observables: a positive intercept of 0.36 equal to the share of bonds purchased, a negative duration tilt of -0.34, and a negative credit-risk tilt of -0.18. The impact is then given by $X_i'\widetilde{\boldsymbol{\mathcal{M}}}(0.36, -0.34, -0.18)'$ which we find to be within 2bps of the exact counterfactual.

tilts away from long-term bonds the term premium increases, which, through life insurers' negative post-GFC duration gap, boosts their risk-bearing capacity and their demand for risky corporate bonds, compressing credit spreads.

6 Concluding Remarks

This paper provides a framework for causal inference with asset prices and quantities. We give conditions for estimating the elasticity matrix of asset demand accounting for the natural spillovers that exist between assets when making portfolio choices. Our main assumption is homogeneous substitution conditional on observables: two assets with the same observables are comparable if the demand for them responds in the same way to the price of every other asset. We show that this condition maps naturally to restrictions often imposed in standard asset pricing models, and also provide guidelines to design experiments satisfying this condition and assess its plausibility in the data.

When our conditions hold, the standard cross-sectional difference-in-differences or instrumental variable approach identifies the relative elasticity between comparable assets—that is, the difference between their own-price and cross-price elasticity. Substitution is identified using a set of time series regressions on portfolios sorted on the observables. Identifying both relative elasticity and substitution is crucial to answer many counterfactual questions such as the response to shocks across broad categories of assets. These tools and principles offer a straightforward package for researchers wanting to use natural experiments to better understand investment decisions and their equilibrium impact.

Because our conditions are flexible, they can guide empirical design without having to take a strong stance on a specific structural model. Still, these causal estimates should only be a first step towards a deeper understanding of how investors and institutions make portfolio decisions, and how those decisions shape equilibrium prices. In richer models, it becomes interesting to understand how demand elasticities change and respond to outside forces such as an increase in passive investing (Haddad et al., 2025b), the Fed's support programs (Haddad et al., 2025a; Jansen et al., 2024), persistent selling pressures (He et al., 2025), and the broader market macrostructure or other macroeconomic conditions.

References

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller, "Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program," *Journal of the American statistical Association*, 2010, *105* (490), 493–505.

Admati, Anat R., "A Noisy Rational Expectations Equilibrium for Multi-Asset Securities Markets," *Econometrica*, 1985, *53* (3), 629–657.

Aghaee, Alireza, "The Flattening Demand Curves," *Working paper*, 2024.

Allen, Jason, Jakub Kastl, and Milena Wittwer, "Estimating demand systems with bidding data," *Available at SSRN 5171755*, 2018.

An, Yu and Amy W. Huber, "Demand Propagation Through Traded Risk Factors," Technical Report 2025.

___ , Yinan Su, and Chen Wang, "Quantity, Risk, and Return," *Working Paper*, 2024.

Anderson, Simon P, André De Palma, and J-F Thisse, "A representative consumer theory of the logit model," *International Economic Review*, 1988, pp. 461–466.

Andrews, Isaiah, James H Stock, and Liyang Sun, "Weak instruments in instrumental variables regression: Theory and practice," *Annual Review of Economics*, 2019, *11* (1), 727–753.

Angrist, Joshua D., "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants," *Econometrica*, 1998, *66* (2), 249–288.

___ and Jörn-Steffen Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press, 2009.

Angrist, Joshua D, Guido W Imbens, and Donald B Rubin, "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 1996, *91* (434), 444–455.

Athey, Susan and Guido W Imbens, "The state of applied econometrics: Causality and policy evaluation," *Journal of Economic perspectives*, 2017, *31* (2), 3–32.

Bai, Jennie and Pierre Collin-Dufresne, "The CDS-bond basis," *Financial Management*, 2019, *48* (2), 417–439.

Becker, Gary S., "Irrational Behavior and Economic Theory," *Journal of Political Economy*, 1962, *70* (1), 1–13.

Ben-David, Itzhak, Jiacui Li, Andrea Rossi, and Yang Song, "Ratings-driven demand and systematic price fluctuations," *The Review of Financial Studies*, 2022, *35* (6), 2790–2838.

Berg, Tobias, Markus Reisinger, and Daniel Streitz, "Spillover effects in empirical corporate finance," *Journal of Financial Economics*, 2021, *142* (3), 1109–1127.

Berk, Jonathan B and Jules H Van Binsbergen, "The impact of impact investing," *Journal of Financial Economics*, 2025, *164*, 103972.

Berry, Steve, Oliver B Linton, and Ariel Pakes, "Limit theorems for estimating the parameters of differentiated product demand systems," *The Review of Economic Studies*, 2004, *71* (3), 613–654.

Berry, Steven, James Levinsohn, and Ariel Pakes, "Automobile Prices in Market Equilibrium," *Econometrica*, 1995, *63* (4), 841–890.

Berry, Steven T., "Estimating Discrete-Choice Models of Product Differentiation," *The RAND Journal of Economics*, 1994, *25* (2), 242–262.

___ and Philip A. Haile, "Identification in differentiated products markets using market level data," *Econometrica*, 2014, *82* (5), 1749–1797.

___ and ___ , "Foundations of demand estimation," in Kate Ho, Ali Hortaçsu, and Alessandro Lizzeri, eds., *Handbook of Industrial Organization*, Vol. 4 (1) of *Handbook of Industrial Organization*, Elsevier, 2021, pp. 1–62.

Binsbergen, Jules H Van, William F Diamond, and Marco Grotteria, "Risk-free interest rates," *Journal of Financial Economics*, 2022, *143* (1), 1–29.

Borusyak, Kirill and Peter Hull, "Nonrandom exposure to exogenous shocks," *Econometrica*, 2023, *91* (6), 2155–2185.

___ , Jiafeng Chen, Peter Hull, and Lihua Lei, "Nonparametric identification of demand without exogenous product characteristics," Technical Report, National Bureau of Economic Research 2026.

Bound, John, David A Jaeger, and Regina M Baker, "Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak," *Journal of the American statistical association*, 1995, *90* (430), 443–450.

Bouveret, Antoine and Jie Yu, "Risks and vulnerabilities in the US bond mutual fund industry," Technical Report 2021.

Bretscher, Lorenzo, Lukas Schmid, Ishita Sen, and Varun Sharma, "Institutional corporate bond pricing," *Swiss Finance Institute Research Paper*, 2022, *21-07*.

Campbell, John Y., *Financial Decisions and Markets: A Course in Asset Pricing*, Princeton University Press, 2017.

___ **and Luis Viceira**, *Strategic Asset Allocation: Portfolio Choice for Long-Term Investors*, first ed., Oxford, UK: Oxford University Press, 2002.

Chang, Yen-Cheng, Harrison Hong, and Inessa Liskovich, "Regression Discontinuity and the Price Effects of Stock Market Indexing," *The Review of Financial Studies*, 07 2014, *28* (1), 212–246.

Chaudhary, Manav, Zhiyu Fu, and Jian Li, "Corporate bond multipliers: Substitutes matter," Technical Report 2022.

Chaudhry, Aditya, "The Impact of Prices on Analyst Cash Flow Expectations: Reconciling Subjective Beliefs Data with Rational Discount Rate Variation," *Journal of Financial Economics*, 2025, *forthcoming*.

Chen, Hui, Zhuo Chen, Zhiguo He, Jinyu Liu, and Rengming Xie, "Pledgeability and asset prices: Evidence from the Chinese corporate bond markets," *The Journal of Finance*, 2023, *78* (5), 2563–2620.

Chodorow-Reich, Gabriel, Plamen T Nenov, and Alp Simsek, "Stock market wealth and the real economy: A local labor market approach," *American Economic Review*, 2021, *111* (5), 1613–1657.

Cochrane, John H, *Asset pricing*, Princeton University Press, 2005.

Conley, Timothy G, Christian B Hansen, and Peter E Rossi, "Plausibly exogenous," *Review of Economics and Statistics*, 2012, *94* (1), 260–272.

Coppola, Antonio, "In safe hands: The financial and real impact of investor composition over the credit cycle," *The Review of Financial Studies*, 2025, *forthcoming*.

Coval, Joshua and Erik Stafford, "Asset fire sales (and purchases) in equity markets," *Journal of Financial Economics*, 2007, *86* (2), 479–512.

Daniel, Kent and Sheridan Titman, "Evidence on the characteristics of cross sectional variation in stock returns," *the Journal of Finance*, 1997, *52* (1), 1–33.

Darmouni, Olivier, Kerry Siani, and Kairong Xiao, "Nonbank Fragility in Credit Markets: Evidence from a Two-Layer Asset Demand System," Technical Report 2023.

Davis, Carter, "The Elasticity of Quantitative Investment," *The Review of Financial Studies*, 2024, *forthcoming*.

___ **, Mahyar Kargar, and Jiacui Li**, "Why Do Portfolio Choice Models Predict Inelastic Demand?," *Journal of Financial Economics*, 2025, *forthcoming*.

Deaton, Angus and John Muellbauer, "An almost ideal demand system," *The American economic review*, 1980, *70* (3), 312–326.

Debreu, Gerard, *Theory of value: An axiomatic analysis of economic equilibrium*, Vol. 17, Yale University Press, 1959.

Drechsler, Itamar, Alexi Savov, Philipp Schnabl, and Dominik Supera, "Monetary Policy and the Mortgage Market," Technical Report, Working paper 2024.

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

Duffie, Darrell, *Dynamic asset pricing theory*, Princeton University Press, 2010.

Evans, Martin D. D. and Richard K. Lyons, "Order Flow and Exchange Rate Dynamics," *Journal of Political Economy*, 2002, *110* (1), 170–180.

Fama, Eugene F and James D MacBeth, "Risk, return, and equilibrium: Empirical tests," *Journal of political economy*, 1973, *81* (3), 607–636.

___ **and Kenneth R French**, "Common risk factors in the returns on stocks and bonds," *Journal of financial economics*, 1993, *33* (1), 3–56.

Fang, Chuck, *Monetary policy amplification through bond fund flows*, University of Pennsylvania, 2023.

___ **and Kairong Xiao**, "Dissecting Bond Market Transmission of Monetary Policy," *Available at SSRN 5025417*, 2024.

Froot, Kenneth A. and Tarun Ramadorai, "Currency Returns, Intrinsic Value, and Institutional-Investor Flows," *The Journal of Finance*, 2005, *60* (3), 1535–1566.

Fuchs, William, Satoshi Fukuda, and Daniel Neuhann, "Demand-System Asset Pricing: Theoretical Foundations," *Available at SSRN 4672473*, 2025.

___ **,** ___ **, and** ___ , "A Trilemma for Asset Demand Estimation," *Available at SSRN 5355589*, 2025.

Gabaix, Xavier and Ralph SJ Koijen, "In search of the origins of financial fluctuations: The inelastic markets hypothesis," Technical Report, National Bureau of Economic Research 2021.

___ **and** ___ , "Granular instrumental variables," *Journal of Political Economy*, 2024, *132* (7), 2274–2303.

Gandhi, Amit and Aviv Nevo, "Empirical models of demand and supply in differentiated products industries," in "Handbook of industrial organization," Vol. 4:1, Elsevier, 2021, pp. 63–139.

___ **and Jean-François Houde**, "Measuring substitution patterns in differentiated-products industries," *NBER Working paper*, 2019.

Goldsmith-Pinkham, Paul, Isaac Sorkin, and Henry Swift, "Bartik Instruments: What, When, Why, and How," *American Economic Review*, August 2020, *110* (8), 2586–2624.

___ **, Peter Hull, and Michal Kolesár**, "Contamination bias in linear regressions," *American Economic Review*, 2024, *114* (12), 4015–4051.

Gompers, Paul A. and Andrew Metrick, "Institutional Investors and Equity Prices*," *The Quarterly Journal of Economics*, 02 2001, *116* (1), 229–259.

Gormsen, Niels Joachim and Eben Lazarus, "Interest Rates and Equity Valuations," Technical Report, Working paper 2025.

Graves, Daniel, "What Lies Beneath Zero: Censoring, Demand Estimation, and Hidden Beliefs," Technical Report, Working paper 2025.

Greenwood, Robin and Dimitri Vayanos, "Bond supply and excess bond returns," *The Review of Financial Studies*, 2014, *27* (3), 663–713.

___ **and Marco Sammon**, "The Disappearing Index Effect," *Working paper*, 2024.

Greenwood, Robin Marc and Annette Vissing-Jorgensen, "The impact of pensions and insurance on global yield curves," Technical Report, Harvard Business School 2018.

Greenwood, Robin, Samuel G Hanson, and Gordon Y Liao, "Asset Price Dynamics in Partially Segmented Markets," *The Review of Financial Studies*, 04 2018, *31* (9), 3307–3343.

___ **, Samuel Hanson, and Dimitri Vayanos**, "Supply and demand and the term structure of interest rates," *Annual Review of Financial Economics*, 2024, *16*, 115–151.

Guren, Adam, Alisdair McKay, Emi Nakamura, and Jón Steinsson, "What Do We Learn from Cross-Regional Empirical Estimates in Macroeconomics?," *NBER Macroeconomics Annual*, 2021, *35*, 175–223.

Gürkaynak, Refet S., Brian Sack, and Jonathan H. Wright, "The U.S. Treasury yield curve: 1961 to the present," *Journal of Monetary Economics*, 2007, *54* (8), 2291–2304.

Haddad, Valentin, Alan Moreira, and Tyler Muir, "When selling becomes viral: Disruptions in debt markets in the COVID-19 crisis and the Fed's response," *The Review of Financial Studies*, 2021, *34* (11), 5309–5351.

___ **,** ___ **, and** ___ , "Asset purchase rules: How QE transformed the bond market," Technical Report, Working paper 2024.

___ **,** ___ **, and** ___ , "Asset Purchase Rules in the Euro Area and Their Effect on Bond Markets," Technical Report, Working paper 2024.

___ **,** ___ **, and** ___ , "Whatever it takes? The impact of conditional policy promises," *American Economic Review*, 2025, *115* (1), 295–329.

___ **and Tyler Muir**, "Do intermediaries matter for aggregate asset prices?," *The Journal of Finance*, 2021, *76* (6), 2719–2761.

___ **and** ___ , "Market macrostructure: Institutions and asset prices," *Annual Review of Financial Economics*, 2025, *17*.

___ **, Paul Huebner, and Erik Loualiche**, "How Competitive Is the Stock Market? Theory, Evidence from Portfolios, and Implications for the Rise of Passive Investing," *American Economic Review*, 2025, *115* (3), 975–1018.

Harris, Lawrence and Eitan Gurel, "Price and Volume Effects Associated with Changes in the S&P 500 List: New Evidence for the Existence of Price Pressures," *The Journal of Finance*, 1986, *41* (4), 815–829.

Hartzmark, Samuel M and David H Solomon, "Predictable price pressure," Technical Report, National Bureau of Economic Research 2022.

Hausman, J. A., "Specification Tests in Econometrics," *Econometrica*, 1978, *46* (6), 1251–1271.

Hausman, Jerry and Daniel McFadden, "Specification Tests for the Multinomial Logit Model," *Econometrica*, 1984, *52* (5), 1219–1240.

He, Zhiguo, Paymon Khorrami, and Zhaogang Song, "Commonality in credit spread changes: Dealer inventory and intermediary distress," *The Review of Financial Studies*, 2022, *35* (10), 4630–4673.

___ **, Peter Kondor, and Jessica Shi Li**, "Demand Elasticity in Dynamic Asset Pricing," Technical Report, Working paper 2025.

Houweling, Patrick, Albert Mentink, and Ton Vorst, "Comparing possible proxies of corporate bond liquidity," *Journal of Banking & Finance*, 2005, *29* (6), 1331–1358.

Huber, Kilian, "Estimating general equilibrium spillovers of large-scale shocks," *The Review of Financial Studies*, 2023, *36* (4), 1548–1584.

Hubert de Fraisse, Antoine, "Crowding Out Long-Term Corporate Investment: The Role of Long-Term Government Debt Supply," *Working paper*, 2022.

Huebner, Paul, "The Making of Momentum: A Demand-System Perspective," *Working paper*, 2024.

Hurwicz, Leonid, "On the structural form of interdependent systems," in "Studies in Logic and the Foundations of Mathematics," Vol. 44, Elsevier, 1966, pp. 232–239.

Imbens, Guido W and Joshua D Angrist, "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 1994, *62* (2), 467–475.

Jansen, Kristy A.E., Wenhao Li, and Lukas Schmid, "Granular Treasury Demand with Arbitrageurs," *Available at SSRN 4940397*, 2024.

Jiang, Zhengyang, Robert J Richmond, and Tony Zhang, "A portfolio approach to global imbalances," *The Journal of Finance*, 2024, *79* (3), 2025–2076.

Kelly, Bryan T, Seth Pruitt, and Yinan Su, "Characteristics are covariances: A unified model of risk and return," *Journal of Financial Economics*, 2019, *134* (3), 501–524.

—— , —— , and —— , "Instrumented principal component analysis," *Available at SSRN 2983919*, 2020.

Koijen, Ralph S. J. and Motohiro Yogo, "A Demand System Approach to Asset Pricing," *Journal of Political Economy*, 2019, *127* (4), 1475–1515.

—— and —— , "Exchange Rates and Asset Prices in a Global Demand System," *Working paper*, 2024.

Koijen, Ralph S J, Robert J Richmond, and Motohiro Yogo, "Which Investors Matter for Equity Valuations and Expected Returns?," *The Review of Economic Studies*, 08 2023, *91* (4), 2387–2424.

Kozak, Serhiy, Stefan Nagel, and Shrihari Santosh, "Shrinking the cross-section," *Journal of Financial Economics*, 2020, *135* (2), 271–292.

Krishnamurthy, Arvind and Annette Vissing-Jorgensen, "The effects of quantitative easing on interest rates: channels and implications for policy," Technical Report, National Bureau of Economic Research 2011.

Kyle, Albert S., "Informed Speculation with Imperfect Competition," *Review of Economic Studies*, 1989, *56* (3), 317–355.

Li, Jiacui and Zihan Lin, "Price Multipliers are Larger at More Aggregate Levels," *Available at SSRN 4038664*, 2022.

Li, Ziang, "Long Rates, Life Insurers, and Credit Spreads," Technical Report, Working paper 2025.

Litterman, Robert B and José Scheinkman, "Common Factors Affecting Bond Returns," *The Journal of Fixed Income*, 1991, *1* (1), 54–61.

Lopez-Lira, Alejandro and Nikolai L Roussanov, "Do common factors really explain the cross-section of stock returns?," *Jacobs Levy Equity Management Center for Quantitative Financial Research Paper*, 2020.

Lou, Dong, "A flow-based explanation for return predictability," *The Review of Financial Studies*, 2012, *25* (12), 3457–3489.

Lu, Xu and Lingxuan Wu, "Monetary Transmission and Portfolio Rebalancing: A Cross-Sectional Approach," Technical Report 2023.

Markowitz, Harry M, "Portfolio selection," *The journal of finance*, 1952, *7*, 77–91.

Merton, Robert C., "An Intertemporal Capital Asset Pricing Model," *Econometrica*, 1973, *41* (5), 867–887.

Mundlak, Yair, "On the Pooling of Time Series and Cross Section Data," *Econometrica*, 1978, *46* (1), 69–85.

Olea, José Luis Montiel and Carolin Pflueger, "A robust test for weak instruments," *Journal of Business & Economic Statistics*, 2013, *31* (3), 358–369.

Pavlova, Anna and Taisiya Sikorskaya, "Benchmarking Intensity," *The Review of Financial Studies*, 08 2022, *36* (3), 859–903.

Peng, Cameron and Chen Wang, "Factor Demand and Factor Returns," *Working paper*, 2023.

Petajisto, Antti, "Why do demand curves for stocks slope down?," *Journal of Financial and Quantitative Analysis*, 2009, *44* (5), 1013–1044.

Reynaert, Mathias and Frank Verboven, "Improving the performance of random coefficients demand models: The role of optimal instruments," *Journal of econometrics*, 2014, *179* (1), 83–98.

Ross, Stephen A, "The arbitrage theory of capital asset pricing," *Journal of Economic Theory*, 1976, *13* (3), 341–360.

Rostek, Marzena and Ji Hee Yoon, "Imperfect competition in financial markets: Recent developments," *Journal of Economic Literature*, 2025, *63* (4), 1191–1243.

Selgrad, Julia, "Testing the Portfolio Rebalancing Channel of Quantitative Easing," *Working paper*, 2023.

Shleifer, Andrei, "Do Demand Curves for Stocks Slope Down?," *The Journal of Finance*, 1986, *41* (3), 579–590.

Stock, James H and Motohiro Yogo, "Testing for weak instruments in Linear IV regression," in "Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg," Cambridge University Press, 2005, pp. 80–108.

Van der Beck, Philippe, "Flow-driven ESG returns," *Swiss Finance Institute Research Paper*, 2021, *21-71*.

Van Kippersluis, Hans and Cornelius A Rietveld, "Beyond plausibly exogenous," *The Econometrics Journal*, 2018, *21* (3), 316–331.

Vayanos, Dimitri and Jean-Luc Vila, "A Preferred-Habitat Model of the Term Structure of Interest Rates," *Econometrica*, 2021, *89* (1), 77–112.

Vazza, Diane and Nick W Kraemer, "Default, Transition, and Recovery: 2012 Annual Global Corporate Default Study and Rating Transitions," RatingsDirect, Standard & Poor's Ratings Services 3 2013. Accessed: 2026-02-05.

Welch, Ivo, "Simply Better Market Betas," *Critical Finance Review*, 2022, *11* (1), 37–64.

Wolf, Christian K, "The missing intercept: A demand equivalence approach," *American Economic Review*, 2023, *113* (8), 2232–2269.

Supplemental Appendix for "Causal Inference for Asset Pricing"

A Proofs and Derivations

A.1 Identifying the relative elasticity — Proposition 1

Start from the general demand equation with demand shocks:

$$\Delta D_i = \mathcal{E}_{ii}\Delta P_i + \sum_{j\neq i}\mathcal{E}_{ij}\Delta P_j + \epsilon_i. \tag{A1}$$

We recall the two assumptions necessary for identification:

- **Assumption A1.** For all $i, j \in \mathcal{S}, l \neq i, j$, we have $X_i = X_j \Rightarrow \mathcal{E}_{il} = \mathcal{E}_{jl} = \mathcal{E}_{\text{cross}}(X_i, X_l) = X_i'\boldsymbol{\mathcal{E}}_X X_l$, where X_i is a $K \times 1$ vector of observables, and $\boldsymbol{\mathcal{E}}_X$ is a $K \times K$ matrix.
- **Assumption A2.** For all $i \in \mathcal{S}$, $\mathcal{E}_{ii} - \mathcal{E}_{\text{cross}}(X_i, X_i) = \mathcal{E}_{relative}$.

Proposition 1 shows that under Assumptions A1 and A2 and the exogeneity condition, the IV estimator, conditioning on X_i, identifies the coefficient $\mathcal{E}_{relative}$.

Proof. Starting from equation (A1), we can rewrite the demand equation as a cross-sectional regression:

$$\begin{aligned}
\Delta D_i &= \mathcal{E}_{ii}\Delta P_i + \sum_{j\neq i}\mathcal{E}_{ij}\Delta P_j + \epsilon_i && \text{(A2)}\\
&= \mathcal{E}_{ii}\Delta P_i + \sum_{j\neq i}\mathcal{E}_{cross}(X_i, X_j)\Delta P_j + \epsilon_i && \text{(A3)}\\
&= \left(\mathcal{E}_{ii} - \mathcal{E}_{cross}(X_i, X_i)\right)\Delta P_i + \sum_{j}\mathcal{E}_{cross}(X_i, X_j)\Delta P_j + \epsilon_i && \text{(A4)}\\
&= \mathcal{E}_{relative}\Delta P_i + \sum_{j}\mathcal{E}_{cross}(X_i, X_j)\Delta P_j + \epsilon_i && \text{(A5)}\\
&= \mathcal{E}_{relative}\Delta P_i + \sum_{j} X_i'\boldsymbol{\mathcal{E}}_X X_j \Delta P_j + \epsilon_i && \text{(A6)}\\
&= \mathcal{E}_{relative}\Delta P_i + X_i'\underbrace{\left(\sum_{j}\boldsymbol{\mathcal{E}}_X X_j \Delta P_j\right)}_{\theta} + \epsilon_i && \text{(A7)}\\
&= \mathcal{E}_{relative}\Delta P_i + \theta' X_i + \epsilon_i && \text{(A8)}
\end{aligned}$$

Equation (A4) adds and subtracts $\mathcal{E}_{cross}(X_i, X_i)\Delta P_i$. Equations (A5) and (A6) use assumptions 2 and 1, respectively. Equation (A7) pulls out X_i' from the sum. The remaining

part of the sum gets absorbed into θ, a $K \times 1$ vector of cross-sectional constants. These θ are K regression coefficients on the K observables, X_{ik}.

Given the exclusion restriction that $Z_i \perp \epsilon_i | X_i$ and the relevance condition that $\text{cov}(\Delta P_i, Z_i | X_i) \neq 0$, this is the standard IV setting, and the regression coefficient $\widehat{\mathcal{E}}$ consistently (as $N \to \infty$) estimates $\mathcal{E}_{relative}$. ■

A.2 Properties of elasticity under assumptions A1 and A2

A.2.1 A matrix representation.

We derive the simple matrix representations of Section 2.1 for an elasticity matrix under assumptions A1 and A2.

Lemma A1 *Let $\boldsymbol{\mathcal{E}}$ be an elasticity matrix that satisfies assumption A1. Then it can be written as:*

$$\boldsymbol{\mathcal{E}} = \text{diag}(\boldsymbol{\mathcal{E}}_{relative}) + X \boldsymbol{\mathcal{E}}_X X', \tag{A9}$$

where $\boldsymbol{\mathcal{E}}_{relative}$ is an $N \times 1$ vector and $\boldsymbol{\mathcal{E}}_X$ is a $K \times K$ matrix.

Proof. Write out the elasticity matrix $\boldsymbol{\mathcal{E}}$:

$$\boldsymbol{\mathcal{E}} = \begin{pmatrix} \mathcal{E}_{11} & \mathcal{E}_{12} & \dots & \mathcal{E}_{1N} \\ \mathcal{E}_{21} & \mathcal{E}_{22} & \dots & \mathcal{E}_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{E}_{N1} & \mathcal{E}_{N2} & \dots & \mathcal{E}_{NN} \end{pmatrix} = \begin{pmatrix} \mathcal{E}_{11} & X_1' \boldsymbol{\mathcal{E}}_X X_2 & \dots & X_1' \boldsymbol{\mathcal{E}}_X X_N \\ X_2' \boldsymbol{\mathcal{E}}_X X_1 & \mathcal{E}_{22} & \dots & X_2' \boldsymbol{\mathcal{E}}_X X_N \\ \vdots & \vdots & \ddots & \vdots \\ X_N' \boldsymbol{\mathcal{E}}_X X_1 & X_N' \boldsymbol{\mathcal{E}}_X X_2 & \dots & \mathcal{E}_{NN} \end{pmatrix}. \tag{A10}$$

The (i,j) element of matrix $\boldsymbol{\mathcal{E}}$ is $[\boldsymbol{\mathcal{E}}]_{ij} = X_i' \boldsymbol{\mathcal{E}}_X X_j = \mathcal{E}_{cross}(X_i, X_j)$, as defined by Assumption A1, for $i \neq j$. The representation with $\boldsymbol{\mathcal{E}}_{relative,ii} = \mathcal{E}_{ii} - X_i' \boldsymbol{\mathcal{E}}_X X_i$ follows. ■

Lemma A2 *Let $\boldsymbol{\mathcal{E}}$ be an elasticity matrix that satisfies assumptions A1 and A2. Then it can be written as:*

$$\boldsymbol{\mathcal{E}} = \mathcal{E}_{relative} \mathbf{I} + X \boldsymbol{\mathcal{E}}_X X', \tag{A11}$$

where $\mathcal{E}_{relative}$ is a scalar and $\boldsymbol{\mathcal{E}}_X$ is a $K \times K$ matrix. Conversely, any matrix of the form (A11) *satisfies assumptions A1 and A2.*

Proof. Start from the representation of the previous lemma, and notice that Assumption A2 implies that $\boldsymbol{\mathcal{E}}_{relative,ii} = \mathcal{E}_{relative}$ is constant. The converse holds by direct verification. ■

A.2.2 Transforming the observables.

The following lemma shows that observables can be recombined in a linear way. In particular they could be demeaned, standardized, or orthogonalized.

Lemma A3 *Let $\boldsymbol{\mathcal{E}}$ be an elasticity matrix that satisfies assumptions A1 and A2 with respect to a set of observables X. If A is a $K \times K$ invertible matrix, $\boldsymbol{\mathcal{E}}$ satisfies assumptions A1 and A2 with respect to the recombined observables $\tilde{X} = XA$.*

Proof. Insert AA^{-1} judiciously into the decomposition of Lemma A2.

$$\boldsymbol{\mathcal{E}} = \mathcal{E}_{relative}\mathbf{I} + XAA^{-1}\boldsymbol{\mathcal{E}}_X(A')^{-1}A'X' = \mathcal{E}_{relative}\mathbf{I} + \tilde{X}\boldsymbol{\mathcal{E}}_{\tilde{X}}\tilde{X}', \tag{A12}$$

$$\text{with } \boldsymbol{\mathcal{E}}_{\tilde{X}} = A^{-1}\boldsymbol{\mathcal{E}}_X(A')^{-1}. \tag{A13}$$

■

For example, if the first observable is a constant and the other ones have mean $\bar{X}_1, \ldots, \bar{X}_{K-1}$, the following matrix demeans them:

$$A_{\text{demean}} = \mathbf{I}_K - \begin{pmatrix} 0 & \bar{X}_1 & \cdots & \bar{X}_{K-1} \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}. \tag{A14}$$

Importantly, note that there is no reason that orthogonalizing the characteristics makes the substitution matrix $\boldsymbol{\mathcal{E}}_{\tilde{X}}$ diagonal.

A.2.3 Stability by inversion — Lemma 1

Lemma 1 states that under Assumptions A1 and A2, the multiplier matrix $\boldsymbol{\mathcal{M}} = -\boldsymbol{\mathcal{E}}^{-1}$ also satisfies Assumptions A1 and A2 (and the converse as well), with $\mathcal{M}_{relative} = -1/\mathcal{E}_{relative}$.
Proof. Start from equation (A11), and apply the Woodbury matrix identity:[56]

$$-\boldsymbol{\mathcal{E}}^{-1} = -\left(\mathcal{E}_{relative}\mathbf{I}_N + X\boldsymbol{\mathcal{E}}_X X'\right)^{-1} \tag{A15}$$

$$= -\mathcal{E}_{relative}^{-1}\mathbf{I}_N + X\mathcal{E}_{relative}^{-1}\left(\mathcal{E}_{relative}\mathbf{I}_K + \boldsymbol{\mathcal{E}}_X X'X\right)^{-1}\boldsymbol{\mathcal{E}}_X X' \tag{A16}$$

$$= \mathcal{M}_{relative}\mathbf{I}_N + X\boldsymbol{\mathcal{M}}_X X'. \tag{A17}$$

By Lemma A2, this representation implies that assumptions A1 and A2 apply to $\boldsymbol{\mathcal{M}}$ with $\mathcal{M}_{relative} = -1/\mathcal{E}_{relative}$. The reverse direction is symmetric: applying the same Woodbury identity to the decomposition of $\boldsymbol{\mathcal{M}}$ recovers the decomposition of $\boldsymbol{\mathcal{E}}$, so by Lemma A2 $\boldsymbol{\mathcal{E}}$ satisfies assumptions A1 and A2. ■

A.2.4 Stability by aggregation

We show that assumptions A1 and A2 are stable by aggregation across investors.

Lemma A4 *Let $\boldsymbol{\mathcal{E}}_1$ and $\boldsymbol{\mathcal{E}}_2$ be two elasticity matrices that satisfy assumptions A1 and A2, and (λ_1, λ_2) two scalars. Then the matrix $\lambda_1\boldsymbol{\mathcal{E}}_1 + \lambda_2\boldsymbol{\mathcal{E}}_2$ satisfies assumptions A1 and A2.*

[56] We apply $(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}$, with $A = \mathcal{E}_{relative}\mathbf{I}$, $U = X$, $C = \mathbf{I}_K$, and $V = \boldsymbol{\mathcal{E}}_X X'$. Hence, we do not rely on invertibility of $\boldsymbol{\mathcal{E}}_X$.

Proof. From lemma A2 we decompose both elasticities which leads to:

$$\lambda_1 \boldsymbol{\mathcal{E}}_1 + \lambda_2 \boldsymbol{\mathcal{E}}_2 = (\lambda_1 \mathcal{E}_{relative,1} + \lambda_2 \mathcal{E}_{relative,2})\, \mathbf{I} + X\, (\lambda_1 \boldsymbol{\mathcal{E}}_{X,1} + \lambda_2 \boldsymbol{\mathcal{E}}_{X,2})\, X' \tag{A18}$$

The decomposition and the equivalence from Lemma A2 conclude the proof. ■

A.3 The missing coefficient problem

We derive the missing coefficient problem highlighted in Section 3.2. We show that without additional restrictions, substitution cannot be identified from a single cross section. Start with the structural relation: $\Delta D = \boldsymbol{\mathcal{E}} \Delta P + \epsilon$, and impose our assumptions: $\boldsymbol{\mathcal{E}} = \mathcal{E}_{relative} \mathbf{I} + X \boldsymbol{\mathcal{E}}_X X'$.

Recall that the demand shift ϵ measures all demand changes that are not caused by a price change. In particular, it can have a relation with the observables X (e.g., if beliefs about relative payoffs of assets with different values of X change) and a non-zero mean (e.g., if the investor demands more assets overall). We can separate ϵ across those components:

$$\epsilon_X = (X'X)^{-1} X' \epsilon \tag{A19}$$

$$\epsilon_{idio,i} = \epsilon_i - X_i' \epsilon_X, \tag{A20}$$

with all components of ϵ_X potentially different from 0 even in the limit of many assets (large N) and imposing no constraints on all other model quantities.

Plugging into demand, we obtain:

$$\Delta D_i = \mathcal{E}_{relative} \Delta P_i + X_i' \boldsymbol{\mathcal{E}}_X X' \Delta P + X_i' \epsilon_X + \epsilon_{idio,i} \tag{A21}$$

$$= \mathcal{E}_{relative} \Delta P_i + \epsilon_{idio,i} + X_i' \underbrace{(\boldsymbol{\mathcal{E}}_X X' \Delta P + \epsilon_X)}_{K \times 1} \tag{A22}$$

We prove Proposition 2.

Proof. Denote $\boldsymbol{\mathcal{E}}_X^{true}$ and ϵ_X^{true} the true values of $\boldsymbol{\mathcal{E}}_X$ and ϵ_X. For any other guess $\boldsymbol{\mathcal{E}}_X^{false}$, the model with $\boldsymbol{\mathcal{E}}_X = \boldsymbol{\mathcal{E}}_X^{false}$ and $\epsilon_X = \epsilon_X^{true} + (\boldsymbol{\mathcal{E}}_X^{true} - \boldsymbol{\mathcal{E}}_X^{false}) X' \Delta P$ is observationally equivalent to the true model. As long as such guesses exist, that is, as long as $\boldsymbol{\mathcal{E}}_X$ has at least one free parameter, this concludes the proof of no identification. ■

Simplest case: only a constant. Consider the simplest possible case, where X is a constant. Writing cross-sectional means at date t with a bar, and noting that $\boldsymbol{\mathcal{E}}_X$ is a scalar in this case, we have:

$$\Delta D_{i,t} = \mathcal{E}_{relative} \Delta P_{i,t} + \boldsymbol{\mathcal{E}}_X N \overline{\Delta P}_t + N \bar{\epsilon}_t + \epsilon_{idio,i,t} \tag{A23}$$

$$= (\boldsymbol{\mathcal{E}}_X N \overline{\Delta P}_t + N \bar{\epsilon}_t) + \mathcal{E}_{relative} \Delta P_{i,t} + \epsilon_{idio,i,t} \tag{A24}$$

Both substitution $\boldsymbol{\mathcal{E}}_X \overline{\Delta P}_t$ and the aggregate demand shift $\bar{\epsilon}_t$ contribute to the constant of a cross-sectional regression, and there is no way to separate them. This is a version of the missing intercept problem.

Obtaining partial identification with symmetry across observables. One way to obtain some partial identification of substitution is to impose that the same parameter drives substitution across many observables. Then, if one also assumes that the number of observables grows as the number of assets increases in the cross section, one can identify part of the substitution matrix.

We illustrate this approach when the observables are dummy variables belonging to a given group. A priori, substitution across groups could follow any matrix $\boldsymbol{\mathcal{E}}_X$. But one might want to assume that all groups substitute symmetrically, that is, $\boldsymbol{\mathcal{E}}_X = \mathcal{E}_{own-g}\mathbf{I} + \mathcal{E}_{cross-g}\mathbf{1}\mathbf{1}'$, with $\mathcal{E}_{own-g}$ an own-group elasticity and $\mathcal{E}_{cross-g}$ a cross-group elasticity. In this case, we have:

$$\Delta D_i = \mathcal{E}_{relative}\Delta P_i + \mathcal{E}_{own-g}N_g\Delta P_g + \mathcal{E}_{cross-g}N\overline{\Delta P} \text{ if } i \in g, \tag{A25}$$

where ΔP_g is the average price change for group g. A cross-sectional regression with an instrument for ΔP_g across groups allows us to recover $\mathcal{E}_{own-g}$. Notice that $\mathcal{E}_{cross-g}$ remains unidentified. This approach corresponds to repeating the relative elasticity estimation at a higher level of aggregation: the matrix $\boldsymbol{\mathcal{E}}_X$ (in contrast to $\boldsymbol{\mathcal{E}}$) satisfies assumptions A1 and A2 with only a constant as observable.

The key assumption here is not that the observables correspond to disjoint groups, but instead that there is a common substitution parameter that affects each of the observables separately. Its plausibility depends on context: for example in a simple mean variance setting, it does not apply if the groups are based on levels of factor loadings.

The nested logit model assumes such a symmetry across groups (the "nests") and additionally imposes that the missing intercept $\mathcal{E}_{cross-g}$ is pinned down by the other parameters.

A.4 Identifying substitution — Propositions 3 and 4

We consider aggregation for the generic case of an elasticity matrix $\boldsymbol{\mathcal{E}}$ that satisfies assumptions A1 and A2 for an arbitrary set of observables X. Recall that the first observable is the constant in most cases, and that we can represent $\boldsymbol{\mathcal{E}}$ as

$$\boldsymbol{\mathcal{E}} = \mathcal{E}_{relative}\mathbf{I} + X\boldsymbol{\mathcal{E}}_X X'. \tag{A26}$$

To define price and quantity aggregates along the various dimensions of the observables, we regress these vectors on X. We prove Proposition 3.

Proof. Using the relation $\Delta D = \boldsymbol{\mathcal{E}}\Delta P$ and the decomposition of $\boldsymbol{\mathcal{E}}$ under the two assumptions, we obtain

$$\begin{aligned}
(X'X)^{-1}X'\Delta D &= \left(\mathcal{E}_{relative}(X'X)^{-1}X'\mathbf{I}_N + (X'X)^{-1}X'X\boldsymbol{\mathcal{E}}_X X'\right)\Delta P && \text{(A27)}\\
&= \left(\mathcal{E}_{relative}(X'X)^{-1} + \boldsymbol{\mathcal{E}}_X\right)X'\Delta P && \text{(A28)}\\
&= \left(\mathcal{E}_{relative}\mathbf{I}_K + \boldsymbol{\mathcal{E}}_X X'X\right)(X'X)^{-1}X'\Delta P && \text{(A29)}\\
\Delta D_X &= \underbrace{\left(\mathcal{E}_{relative}\mathbf{I}_K + \boldsymbol{\mathcal{E}}_X X'X\right)}_{\breve{\boldsymbol{\mathcal{E}}}}\Delta P_X && \text{(A30)}
\end{aligned}$$

This implies that ΔD_X can be expressed as a linear combination of the K elements of ΔP_X,

as opposed to the whole N components of the changes in price ΔP.

From the definition of the idiosyncratic change in demand:

$$\begin{aligned}
\Delta D_{idio} &= \Delta D - X\Delta D_X && \text{(A31)}\\
&= \mathcal{E}_{relative}\Delta P + X\boldsymbol{\mathcal{E}}_X X'\Delta P - \left(X\mathcal{E}_{relative} + X\boldsymbol{\mathcal{E}}_X X'X\right)(X'X)^{-1}X'\Delta P && \text{(A32)}\\
&= \mathcal{E}_{relative}\left(\Delta P - X\Delta P_X\right) + X\boldsymbol{\mathcal{E}}_X X'\Delta P - X\boldsymbol{\mathcal{E}}_X X'\Delta P && \text{(A33)}\\
&= \mathcal{E}_{relative}\Delta P_{idio}. && \text{(A34)}
\end{aligned}$$

Because $\mathcal{E}_{relative}$ is scalar, the idiosyncratic component is determined asset by asset, which concludes the proof. ■

Simple case with no characteristic. We can recover the simpler cases studied in the paper. First, consider a setting with a single variable X which is constant equal to one. In this case ΔD_X has only one component equal to $\Delta D_{agg} = N^{-1}\sum_i \Delta D_i$, and $\breve{\mathcal{E}} = \mathcal{E}_{relative} + N\mathcal{E}_X$ is a scalar equal to the macro elasticity.

With one normalized characteristic. Now turn to a single standardized characteristic that we call X in a slight abuse of notation. There is still a constant. There, the regression gives two aggregate prices and quantities: the aggregate component ΔD_{agg} defined as before (the constant of the regression); the meso component $\Delta D_X = N^{-1}\sum_i X_i\Delta D_i$. Then the matrix $\breve{\boldsymbol{\mathcal{E}}}$ is 2×2 and equal to

$$\breve{\boldsymbol{\mathcal{E}}} = \begin{pmatrix} \mathcal{E}_{relative} + N[\boldsymbol{\mathcal{E}}_X]_{11} & N[\boldsymbol{\mathcal{E}}_X]_{12} \\ N[\boldsymbol{\mathcal{E}}_X]_{21} & \mathcal{E}_{relative} + N[\boldsymbol{\mathcal{E}}_X]_{22} \end{pmatrix} = \begin{pmatrix} \bar{\mathcal{E}}_{agg} & \bar{\mathcal{E}}_X \\ \tilde{\mathcal{E}}_{agg} & \tilde{\mathcal{E}}_X \end{pmatrix}. \tag{A35}$$

When observables are group dummies. Consider the case when the observables are dummy variables expressing the belonging to disjoint groups. In this situation, there is no need for a constant. The aggregate demand and price indices have a simple interpretation: they measure the average change in demand and price for each group k:

$$\Delta P_{X,k} = \frac{1}{N_k}\sum_{i\in k}\Delta P_i \tag{A36}$$

This implies that demand elasticities have a nested structure. First, there is a relative elasticity within each group $\widehat{\mathcal{E}}$ capturing the impact of changes in relative prices within a group. Then individual assets can be replaced by the aggregate portfolio of each group, and there is an elasticity matrix across these aggregate portfolios, $\breve{\boldsymbol{\mathcal{E}}}$.

Estimation. Applying the decomposition to the structural demand equation yields equation (37). Applying the standard instrumental variable conditions in the time series yields Proposition 4.

A.5 Heterogeneous relative elasticities based on observables

We maintain assumption A1. We relax assumption A2 by allowing the relative elasticity to depend linearly on observables:

$$\mathcal{E}_{ii} - \mathcal{E}_{cross}(X_i, X_i) = \mathcal{E}_{relative}(X_i) = \mathcal{E}_r' X_i. \tag{A37}$$

From section A.1 and the proof of Proposition 1 we obtain:

$$\begin{aligned}
\Delta D_i &= \mathcal{E}_{ii}\Delta P_i + \sum_{j\neq i} \mathcal{E}_{ij}\Delta P_j + \epsilon_i && \text{(A38)}\\
&= \mathcal{E}_{ii}\Delta P_i + \sum_{j\neq i} \mathcal{E}_{cross}(X_i, X_j)\Delta P_j + \epsilon_i && \text{(A39)}\\
&= \left(\mathcal{E}_{ii} - \mathcal{E}_{cross}(X_i, X_i)\right)\Delta P_i + \sum_{j} \mathcal{E}_{cross}(X_i, X_j)\Delta P_j + \epsilon_i && \text{(A40)}\\
&= \mathcal{E}_{relative}(X_i)\Delta P_i + \sum_{j} X_i' \boldsymbol{\mathcal{E}}_X X_j \Delta P_j + \epsilon_i && \text{(A41)}\\
&= \underbrace{\mathcal{E}_{relative}(X_i)}_{\mathcal{E}_r' X_i}\Delta P_i + X_i' \underbrace{\left(\sum_j \boldsymbol{\mathcal{E}}_X X_j \Delta P_j\right)}_{\theta} + \epsilon_i && \text{(A42)}\\
&= \mathcal{E}_r' X_i \Delta P_i + \theta' X_i + \epsilon_i && \text{(A43)}
\end{aligned}$$

In this case we want to identify the vector $\mathcal{E}_r$ which characterizes the relative elasticity with respect to observables. Identification must rely on a vector of instruments. It is natural to construct those instruments from a single instrument Z_i for the price interacted with the observables. The set of identification conditions is:

$$Z_i X_i \perp \epsilon_i | X_i \tag{A44}$$

Under these conditions the two-stage least squares regression proceeds as follows. First, regress each component of $X_i \Delta P_i$ on the vector of instruments $X_i Z_i$ and the observables X_i. The relevance condition is that the matrix of coefficients on the instruments is full-rank. This leads to predicted values of the change in price interacted with observables $\widehat{X_i \Delta P_i}$. Second, regress the change in demand on these predicted values and the observables. The coefficients on the predicted values recover $\mathcal{E}_r$. Finally, the relative elasticity for each asset is simply $\mathcal{E}_{relative}(X_i) = \mathcal{E}_r' X_i$.

A.6 Estimating a Local Average Treatment Effect

This section removes Assumption A2, deriving a Local Average Treatment Effect result for relative elasticities—a counterpart to Proposition 1—under stronger assumptions on instrument exogeneity.

The structural equations (data-generating process) are

$$\Delta D_i = \sum_{j=1}^{N} \mathcal{E}_{ij} \Delta P_j + \epsilon_i \tag{A45}$$

$$\Delta P_i = \sum_{j=1}^{N} \Lambda_{ij} Z_j + \mu_i, \tag{A46}$$

where equations (A45) and (A46) represent the structural demand equation and the structural impact of the instrument under an unrestricted spillover matrix $\mathbf{\Lambda}$, respectively. The elasticity matrix $\boldsymbol{\mathcal{E}}$ satisfies Assumption A1 but not A2, taking the form

$$\boldsymbol{\mathcal{E}} = \operatorname{diag}(\boldsymbol{\mathcal{E}}_{relative}) + X \boldsymbol{\mathcal{E}}_X X', \tag{A47}$$

such that relative elasticities $\mathcal{E}_{relative,i} = \mathcal{E}_{ii} - X_i' \boldsymbol{\mathcal{E}}_X X_i$ are heterogeneous across assets. Without loss of generality, the $N \times K$ matrix of observables X is of rank K, and we assume that K remains fixed as we take limits. This ensures in particular that the operator $\mathbf{Q}_X = \mathbf{I} - X(X'X)^{-1}X'$, which constructs residuals after controlling for X, is well-defined.

The econometrician estimates the same two-stage least squares (2SLS) regression as in Proposition 1:

$$\Delta D_i = \widehat{\mathcal{E}}\, \Delta P_i + X_i' \theta + e_i, \tag{A48}$$

instrumenting ΔP_i with Z_i while controlling for X_i.

We assume strict instrument exogeneity and independence, meaning that Z_i is randomly assigned and independent of all other components:[57]

1. $Z_i \perp\!\!\!\perp Z_j$ for all $i \neq j$, with $\mathbf{E}[Z_i] = 0$, $\operatorname{var}(Z_i) = \sigma_Z^2$ and $\mathbf{E}[Z_i^4] < \infty$ for all i.[58]
2. $Z_i \perp\!\!\!\perp (\mu_j, \epsilon_j)$ for all i, j.
3. $Z_i \perp\!\!\!\perp (\Lambda_{kl}, X_j, \mathcal{E}_{relative,j})$ for all i, j, k, l.

The standard exclusion restriction follows directly from these conditions. We also assume relevance, meaning that the first stage retains power after partialling out observables:

$$\operatorname{cov}(\Delta P_i, Z_i \mid X) \neq 0. \tag{A49}$$

[57] Full statistical independence is sufficient but stronger than necessary: the expectation calculations in the proof use only first- and second-moment conditions. Independence across instruments is used in the concentration argument underlying the probability limit.

[58] Zero mean is without loss of generality: because the first element of X_i is 1, an intercept is included in the controls, and $\mathbf{Q}_X$ annihilates any common instrument mean. Identical variance is assumed for notational convenience. With heterogeneous variances $\operatorname{var}(Z_i) = \sigma_{Z,i}^2$, the weights in Proposition A1 become $\lambda_i = [\mathbf{Q}_X \operatorname{diag}(\sigma_{Z,j}^2) \mathbf{\Lambda}']_{ii}$, and all results carry through, including non-negative weights under bilinear $\mathbf{\Lambda}$ with monotonicity. Intuitively, instruments with larger variance generate more first-stage price variation and receive proportionally greater weight in the estimand. The finite fourth moment ensures concentration of the quadratic forms in the 2SLS estimand (Step 4 of the proof).

Equivalently, this requires $\mathrm{tr}(\mathbf{Q}_X\mathbf{\Lambda}) \neq 0$, ensuring that the instruments generate price variation that is not fully absorbed by observables.[59]

The following proposition gives an interpretation of the 2SLS regression coefficient as a local average treatment effect (LATE).

Proposition A1 (LATE of relative elasticities) *Under instrument exogeneity, independence, relevance, and Assumption A1, the two-stage least squares estimator from regression* (A48) *identifies:*

$$\widehat{\mathcal{E}} \xrightarrow{p} \frac{\sum_{i=1}^{N} \lambda_i \, \mathcal{E}_{relative,i}}{\sum_{i=1}^{N} \lambda_i}, \quad \textit{with} \quad \lambda_i = [\mathbf{Q}_X \mathbf{\Lambda}']_{ii}, \tag{A50}$$

where $[\mathbf{Q}_X\mathbf{\Lambda}']_{ii} = \sum_j [\mathbf{Q}_X]_{ij} \Lambda_{ij}$ *and* $\mathbf{Q}_X = \mathbf{I} - X(X'X)^{-1}X'$ *is the residualization operator. The weights* λ_i *are determined by the first-stage spillover matrix* $\mathbf{\Lambda}$ *after projecting out observables.*

Moreover, if $\mathbf{\Lambda}$ *has the same bilinear structure as* $\boldsymbol{\mathcal{E}}$*—that is,* $\Lambda_{ij} = X_i'\mathbf{\Lambda}_X X_j$ *for all* $i \neq j$*, for some* $K \times K$ *matrix* $\mathbf{\Lambda}_X$*—and the monotonicity condition* $\Lambda_{relative,i} := \Lambda_{ii} - X_i'\mathbf{\Lambda}_X X_i \geqslant 0$ *holds for all* i*, then the weights are non-negative and the estimand is a convex combination of relative elasticities.*

We discuss the proposition before turning to the proof.

Role of controls. The controls X_i serve two purposes. First, they absorb the cross-elasticity spillovers $X_i'\boldsymbol{\mathcal{E}}_X X'\Delta P$, which are common across assets sharing the same characteristics. Second, the residualization $\mathbf{Q}_X$ removes the component of first-stage variation that is explained by observables, ensuring that the IV variation is "idiosyncratic" relative to X. Without controls, the IV estimand would include an additional bias term from the uncontrolled cross-elasticity channel.

No contamination. With heterogeneous treatment effects and controls, 2SLS estimands do not generally equal a convex weighted average of treatment effects: heterogeneity in how units respond to both the treatment and the controls can distort the Frisch–Waugh–Lovell residualization (Goldsmith-Pinkham et al., 2024). In our setting, absent additional structure, demand for asset i depends on the entire vector of prices through heterogeneous cross-elasticities $\mathcal{E}_{ij}$, creating an N-treatment problem where contamination is generic.

Assumption A1 resolves this by imposing homogeneous substitution conditional on observables. Because all cross-price effects take the form $X_i'\boldsymbol{\mathcal{E}}_X X_j$, the entire cross-price channel collapses to the common term $X_i'\theta$ in equation (A54). Since θ is constant across assets in the cross-section, it is fully absorbed by the controls and annihilated by $\mathbf{Q}_X$, regardless of how the own-price elasticities $\mathcal{E}_{relative,i}$ vary.

[59] This condition permits $\mathbf{\Lambda}$ (and $\mathbf{Q}_X\mathbf{\Lambda}$) to be rank-deficient: some assets may be redundant, in the sense that their price variation is fully explained by other assets and observables. Such assets have $[\mathbf{Q}_X\mathbf{\Lambda}']_{ii} = 0$ and receive zero weight $\lambda_i = 0$ in the estimand (A50). Identification requires only that *some* assets have nonzero first-stage variation after partialling out X.

As a result, once we condition on X, the model reduces to a single-treatment IV problem in the sense of Angrist (1998). Heterogeneity remains only in the relative elasticities $\mathcal{E}_{relative,i}$, allowing the 2SLS estimand to be a first-stage-weighted average free of contamination.

Classic LATE and interpretation of weights. In the standard LATE framework (Imbens and Angrist, 1994; Angrist et al., 1996), IV with heterogeneous treatment effects identifies a first-stage-weighted average. When the first stage is diagonal ($\Lambda_{ij} = 0$ for $i \neq j$), the weights reduce to $\lambda_i \propto \Lambda_{ii}$: assets whose prices are most responsive to their own instrument receive greater weight.

However, with a full spillover matrix $\mathbf{\Lambda}$, the weights become $[\mathbf{Q}_X \mathbf{\Lambda}']_{ii}$—the total instrument-induced price response of asset i, combining direct and indirect channels, after projecting out the component explained by observables.

This distinction matters when spillovers are heterogeneous. An asset with a modest direct effect Λ_{ii} can receive large weight if it is strongly exposed to other assets' instruments through off-diagonal elements of $\mathbf{\Lambda}$. Conversely, an asset with a large direct effect may receive little weight if spillovers from correlated assets are absorbed by the controls X.

Economically, the source of heterogeneous first-stage exposure here is much richer than in the diagonal case: it reflects equilibrium price transmission across assets, rather than simple, direct sensitivity to an instrument. For example, if more illiquid assets both experience larger residualized price responses $[\mathbf{Q}_X \mathbf{\Lambda}']_{ii}$ and larger relative elasticities $\mathcal{E}_{relative,i}$, the IV estimand will overstate how inelastic the typical asset is.

Sign of weights. In general, some of the weights λ_i can be negative, meaning that the estimand is not guaranteed to be a convex combination of relative elasticities. A negative weight $\lambda_i < 0$ arises when $[\mathbf{Q}_X \mathbf{\Lambda}']_{ii} < 0$: the net effect of instruments on asset i's price is opposite in sign to the average effect across assets.

The second part of Proposition A1 gives sufficient conditions for non-negative weights. When $\mathbf{\Lambda}$ has the same bilinear structure as $\boldsymbol{\mathcal{E}}$, the annihilator $\mathbf{Q}_X$ reduces the weights to $[\mathbf{Q}_X]_{ii}\, \Lambda_{relative,i}$ (Step 5 of the proof), where $\Lambda_{relative,i} = \Lambda_{ii} - X_i' \mathbf{\Lambda}_X X_i$ is the "relative" first-stage coefficient—the analog of $\mathcal{E}_{relative,i}$ for the first stage. The monotonicity condition $\Lambda_{relative,i} \geqslant 0$ then requires that each asset's own first-stage response weakly exceeds the cross-asset spillover between two assets sharing its characteristics.

Proof. *Step 1: Rewriting the demand equation.* Starting from the structural demand equation (A45) and the elasticity decomposition (A47):

$$\Delta D_i = \mathcal{E}_{ii} \Delta P_i + \sum_{j \neq i} \mathcal{E}_{ij} \Delta P_j + \epsilon_i \tag{A51}$$

$$= \left(\mathcal{E}_{relative,i} + X_i' \boldsymbol{\mathcal{E}}_X X_i\right) \Delta P_i + \sum_{j \neq i} X_i' \boldsymbol{\mathcal{E}}_X X_j \Delta P_j + \epsilon_i \tag{A52}$$

$$= \mathcal{E}_{relative,i}\, \Delta P_i + X_i' \boldsymbol{\mathcal{E}}_X \sum_{j=1}^{N} X_j \Delta P_j + \epsilon_i \tag{A53}$$

$$= \mathcal{E}_{relative,i}\, \Delta P_i + X_i' \theta + \epsilon_i, \tag{A54}$$

where $\theta = \boldsymbol{\mathcal{E}}_X \sum_j X_j \Delta P_j$ is a $K \times 1$ vector that is constant across assets in the cross-section, but varies across realizations. The third equality adds and subtracts $X_i' \boldsymbol{\mathcal{E}}_X X_i \Delta P_i$ and uses Assumption A1 to write all cross-elasticity terms as $X_i' \boldsymbol{\mathcal{E}}_X X_j$, then factors out $X_i' \boldsymbol{\mathcal{E}}_X$ from the full sum over j.

Step 2: The 2SLS estimand via Frisch–Waugh–Lovell. By the FWL theorem, the 2SLS coefficient on ΔP_i in regression (A48) equals the coefficient from the residualized regression:

$$\widehat{\mathcal{E}} = \frac{\sum_i Z_i^{\perp} \cdot \Delta D_i}{\sum_i Z_i^{\perp} \cdot \Delta P_i}, \tag{A55}$$

where $Z^{\perp} = \mathbf{Q}_X Z$ denotes the residual of Z after projecting out X.

Substituting (A54) into the numerator:

$$\begin{aligned}
\text{Numerator} &= \sum_i Z_i^{\perp} \left(\mathcal{E}_{relative,i} \, \Delta P_i + X_i' \theta + \epsilon_i \right) \\
&= \sum_i \mathcal{E}_{relative,i} \, Z_i^{\perp} \Delta P_i + \theta' \underbrace{\left(\sum_i X_i Z_i^{\perp} \right)}_{= X' \mathbf{Q}_X Z = 0} + \sum_i Z_i^{\perp} \epsilon_i \\
&= \sum_i \mathcal{E}_{relative,i} \, Z_i^{\perp} \Delta P_i + \sum_i Z_i^{\perp} \epsilon_i .
\end{aligned} \tag{A56}$$

The $X'\theta$ terms vanish because $\mathbf{Q}_X$ is the projector orthogonal to the column space of X.

Step 3: Expectations over instruments. Conditional on all structural parameters $\Theta = (\mathcal{E}_{relative,i}, \Lambda_{kl}, X_i, \mu_i, \epsilon_i)_{i,k,l}$, only Z is random. We compute the expected numerator and denominator.

Since $Z^{\perp} = \mathbf{Q}_X Z$, the i-th element is $Z_i^{\perp} = \sum_k [\mathbf{Q}_X]_{ik} Z_k$. Using the structural price change $\Delta P_i = \sum_j \Lambda_{ij} Z_j + \mu_i$, independence of Z from μ, and $\mathbf{E}[Z_k Z_j] = \sigma_Z^2 \mathbf{1}_{k=j}$ (since instruments are zero-mean and independent):

$$\begin{aligned}
\mathbf{E}\left[Z_i^{\perp} \Delta P_i \mid \Theta \right] &= \sum_k \sum_j [\mathbf{Q}_X]_{ik} \Lambda_{ij} \, \mathbf{E}[Z_k Z_j] \\
&= \sigma_Z^2 \sum_j [\mathbf{Q}_X]_{ij} \Lambda_{ij} = \sigma_Z^2 \, [\mathbf{Q}_X \boldsymbol{\Lambda}']_{ii},
\end{aligned} \tag{A57}$$

where the last equality uses the definition $[\mathbf{Q}_X \boldsymbol{\Lambda}']_{ii} = \sum_j [\mathbf{Q}_X]_{ij} [\boldsymbol{\Lambda}']_{ji} = \sum_j [\mathbf{Q}_X]_{ij} \Lambda_{ij}$.

For the denominator, summing (A57) over i:

$$\mathbf{E}\left[\text{Denominator} \mid \Theta \right] = \sigma_Z^2 \sum_i [\mathbf{Q}_X \boldsymbol{\Lambda}']_{ii} = \sigma_Z^2 \, \operatorname{tr}(\mathbf{Q}_X \boldsymbol{\Lambda}'). \tag{A58}$$

For the numerator, the ϵ term vanishes since $\epsilon_i \in \Theta$ and the instruments are zero-mean: $\mathbf{E}[Z_i^{\perp} \epsilon_i | \Theta] = \epsilon_i \sum_k [\mathbf{Q}_X]_{ik} \mathbf{E}[Z_k] = 0$. From (A56) and (A57):

$$\mathbf{E}\left[\text{Numerator} \mid \Theta \right] = \sigma_Z^2 \sum_i \mathcal{E}_{relative,i} \, [\mathbf{Q}_X \boldsymbol{\Lambda}']_{ii}. \tag{A59}$$

Step 4: The conditional estimand. The numerator and denominator are quadratic forms in Z. As $N \to \infty$, since instruments are independently assigned across assets and first-stage exposure is dispersed across many assets, cross-sectional averages satisfy a law of large numbers. Both the numerator and denominator concentrate around their conditional expectations.[60][61]

Since $\mathbf{E}[\text{Denom} \mid \Theta] = \sigma_Z^2 \operatorname{tr}(\mathbf{Q}_X \mathbf{\Lambda}') \neq 0$ by relevance, the continuous mapping theorem implies

$$\widehat{\mathcal{E}} \xrightarrow{p} \frac{\sum_i \mathcal{E}_{relative,i} \left[\mathbf{Q}_X \mathbf{\Lambda}'\right]_{ii}}{\operatorname{tr}(\mathbf{Q}_X \mathbf{\Lambda}')} = \frac{\sum_i \lambda_i \mathcal{E}_{relative,i}}{\sum_i \lambda_i}, \tag{A60}$$

with weights $\lambda_i = [\mathbf{Q}_X \mathbf{\Lambda}']_{ii}$ that depend on $(\mathbf{\Lambda}, X)$ but not on $\mathcal{E}_{relative,i}$.

Step 5: Non-negative weights under bilinear $\mathbf{\Lambda}$ with monotonicity. Suppose $\mathbf{\Lambda}_{ij} = X_i' \mathbf{\Lambda}_X X_j$ for all $i \neq j$, so that $\mathbf{\Lambda} = \operatorname{diag}(\mathit{\Lambda}_{relative,i}) + X \mathbf{\Lambda}_X X'$ with $\mathit{\Lambda}_{relative,i} = \mathbf{\Lambda}_{ii} - X_i' \mathbf{\Lambda}_X X_i$. Then $\mathbf{\Lambda}' = \operatorname{diag}(\mathit{\Lambda}_{relative,i}) + X \mathbf{\Lambda}_X' X'$, and since $\mathbf{Q}_X X = 0$:

$$\mathbf{Q}_X \mathbf{\Lambda}' = \mathbf{Q}_X \operatorname{diag}(\mathit{\Lambda}_{relative,i}). \tag{A61}$$

Thus $[\mathbf{Q}_X \mathbf{\Lambda}']_{ii} = [\mathbf{Q}_X]_{ii} \mathit{\Lambda}_{relative,i}$. Since $\mathbf{Q}_X$ is a symmetric idempotent matrix, its diagonal elements satisfy $[\mathbf{Q}_X]_{ii} \in [0, 1]$. Under the monotonicity condition $\mathit{\Lambda}_{relative,i} \geqslant 0$ for all i, we have $[\mathbf{Q}_X \mathbf{\Lambda}']_{ii} \geqslant 0$ for all i, so $\lambda_i \geqslant 0$. ■

B Elasticity inside models

B.1 When does demand elasticity measure a structural relationship?

The investor problem. Let us consider the textbook static microfounded model of portfolio decisions. An agent chooses between $N + 1$ assets, where asset 0 is the numeraire, meaning its price is always normalized to 1. The agent believes (rationally or not) that the payoffs $(\mathit{\Pi}_0, \ldots, \mathit{\Pi}_N)$ of the various assets have joint distribution F. The agent is endowed with quantities $(\bar{Q}_0, \ldots, \bar{Q}_N)$ of the various assets. They maximize their expected utility, with utility function u, taking prices $P = (P_1, \ldots, P_N)$ as given. This corresponds to:

$$\max_{Q_0, Q_1, \ldots, Q_N} \mathbb{E}\left[u\left(\sum_{i=0}^{N} Q_i \mathit{\Pi}_i \right) \right] \tag{A62}$$

$$\text{s.t. } Q_0 + \sum_{i=1}^{N} Q_i P_i \leqslant \bar{Q}_0 + \sum_{i=1}^{N} \bar{Q}_i P_i \tag{A63}$$

[60] This requires that first-stage exposure is not concentrated in a vanishing fraction of assets as N grows. Intuitively, as additional assets are added to the cross-section, sufficiently many must contribute independent residualized first-stage variation after partialling out X; no single asset may account for a non-negligible share of total variation.

[61] Independence of instruments across assets ensures that cross-terms vanish when taking expectations of quadratic forms such as $\sum_i Z_i^{\perp} \mathit{\Delta} P_i$, so that averaging across assets yields concentration.

The optimal positions define a demand function. Focusing only on positions in assets 1 to N, we define this function $D(P) = (Q_1^*(P), \ldots, Q_N^*(P))$. Demand is a function from $\mathbb{R}^N$ to $\mathbb{R}^N$.

In this microeconomic problem, the demand curve entirely characterizes how decisions change when prices change. If one focuses on small changes in prices, changes in positions are characterized by the slope of this demand curve, the Jacobian $\boldsymbol{\mathcal{E}} = \partial D/\partial P = [\partial D_i/\partial P_j]_{i,j=1,\ldots,N}$. In a slight abuse of language, we refer to $\boldsymbol{\mathcal{E}}$ as the elasticity of demand. We have immediately $\Delta Q \approx \boldsymbol{\mathcal{E}} \Delta P$ for a small change in the price vector.

One might also be interested in what happens if the distribution of payoffs changes. For example, the distribution F might be parametrized by a vector Y, and one can ask how optimal positions change when Y moves. Similarly, preferences might change too: for example the risk aversion of the investor could depend on their mood. We can also encode this in the vector Y. This extends the definition of the demand function to $D(P, Y)$. Note that while in equilibrium, the distribution of returns depends on market clearing, the demand function $D(P, Y)$ describes the agent's optimal portfolio for any arbitrary price vector P and belief system Y, effectively isolating the agent's preferences from the equilibrium constraints. We still refer to the price elasticity of demand as $\boldsymbol{\mathcal{E}} = \partial D/\partial P$.

This elasticity captures a *structural relationship* in the sense of Hurwicz (1966) for any counterfactual about this investor that involves changes in prices but not in the vector Y. The elasticity characterizes how the investor's portfolio changes when prices change locally but the vector Y does not, irrespective of any other details about what determines equilibrium prices.

Inside of an equilibrium. To make things concrete, we flesh out an equilibrium with many such standard agents. Specifically, consider an equilibrium populated by J agents indexed by $j = 1, \ldots, J$. Each agent has a demand function $D_j(P, Y_j)$ and an initial endowment $\bar{Q}_j$. Market clearing requires that the sum of demands equals the aggregate endowment. We define the aggregate demand for risky assets as $\mathcal{D}(P, Y) = \sum_{j=1}^J D_j(P, Y_j)$ and the aggregate endowment as $\bar{Q}_{\text{mkt}} = \sum_{j=1}^J \bar{Q}_j$. The equilibrium price vector P^* is the solution to:

$$\sum_{j=1}^{J} D_j(P^*, Y_j) = \bar{Q}_{\text{mkt}} \tag{A64}$$

We define the aggregate elasticity of demand as the sum of individual elasticities (Jacobians): $\boldsymbol{\mathcal{E}}_{agg} = \sum_{j=1}^J \boldsymbol{\mathcal{E}}_j$. Assuming standard conditions for uniqueness and stability, $\boldsymbol{\mathcal{E}}_{agg}$ is negative definite (invertible). We can now compute two structural counterfactuals using the implicit function theorem.

Counterfactual 1: Absorption of a large trade. Consider the arrival of a new trader who wishes to purchase a fixed portfolio Q_{new} of risky assets. This trader is endowed only with the numeraire (asset 0), so they add no new supply of risky assets to the market. Their demand represents a pure removal of shares from the existing pool. Market clearing now requires $\sum D_j(P') + Q_{\text{new}} = \bar{Q}_{\text{mkt}}$. This is equivalent to a negative supply shock of magnitude $\Delta S = -Q_{\text{new}}$ to the risky assets available to the original J agents. Linearizing around the initial equilibrium, the change in prices ΔP required to accommodate this trade satisfies

$\boldsymbol{\mathcal{E}}_{agg}\Delta P \approx -Q_{\text{new}}$. Thus, the price impact is:

$$\Delta P \approx -\boldsymbol{\mathcal{E}}_{agg}^{-1} Q_{\text{new}} \tag{A65}$$

The price response is determined entirely by the aggregate "risk-bearing capacity" of the market ($\boldsymbol{\mathcal{E}}_{\text{agg}}$) and the size of the trade. The existing agents must be induced to reduce their holdings by exactly Q_{new}, and $\boldsymbol{\mathcal{E}}_{agg}^{-1}$ describes the price concession required to achieve this reduction.

Counterfactual 2: A shock to beliefs, preferences or macrostructure. Consider a change in the characteristics Y_k of a single agent k (e.g., a change in sentiment or risk aversion), while other agents and the total supply remain fixed. The market clearing condition requires the price to adjust to offset agent k's shift in demand. Differentiating the clearing condition with respect to Y_k:

$$\boldsymbol{\mathcal{E}}_{agg}\Delta P + \frac{\partial D_k}{\partial Y_k}\Delta Y_k = 0 \implies \Delta P \approx -\boldsymbol{\mathcal{E}}_{agg}^{-1}\frac{\partial D_k}{\partial Y_k}\Delta Y_k \tag{A66}$$

This price change forces all other agents $i \neq k$ to rebalance their portfolios. Since their parameters Y_i are unchanged, their trading is purely a response to the price movement:

$$\Delta Q_i \approx \boldsymbol{\mathcal{E}}_i \Delta P \tag{A67}$$

This illustrates the mechanism of how the equilibrium changes: a shock to agent k's beliefs transmits to agent i's portfolio strictly through the price channel, governed by agent i's structural elasticity $\boldsymbol{\mathcal{E}}_i$.

B.2 From elasticity to multiplier

As we describe in Section B.3, it is sometimes more suitable to estimate demand elasticities in different units (logarithms, portfolio shares instead of quantities, ...). The inversion result of Section 4 applies to these different cases but with slightly adjusted formulas:

$$\boldsymbol{\mathcal{M}}_{\{\log P,\log Q\}} = -\boldsymbol{\mathcal{E}}_{\{\log Q,\log P\}}^{-1}, \tag{A68}$$

$$\boldsymbol{\mathcal{M}}_{\{\log P,\log Q\}} = -\left[\boldsymbol{\mathcal{E}}_{\{\log \omega,\log P\}} - (\mathbf{I} - \mathbf{1}\omega')\right]^{-1}, \tag{A69}$$

$$\boldsymbol{\mathcal{M}}_{\{\log P,\log Q\}} = -\left[\text{diag}(\omega)^{-1}\boldsymbol{\mathcal{E}}_{\{\omega,\log P\}} - (\mathbf{I} - \mathbf{1}\omega')\right]^{-1}. \tag{A70}$$

For example in the case of logit where demand elasticity is measured by regressing the log portfolio share on log price, equation (A69) gives us the multiplier in log units: by how many percent do prices move in response to a one percent change in aggregate demand. Similarly, equation (A70) is useful for the case of CRRA.

B.3 Standard models of asset demand

We discuss how standard models of asset demand relate to the identification assumptions. For each model, we derive the appropriate units and parameter restrictions under which the

	CARA	CRRA	Logit
Regression units "demand" LHS	demand D_i	portfolio shares $P_i D_i / W$	log portfolio shares $\log(P_i D_i / W)$
Regression units "price" RHS	price P_i	log price $\log P_i$	log price $\log P_i$
Own Price Elasticity $\mathcal{E}_{ii}$	$-\frac{1}{\gamma}[\Sigma^{-1}]_{ii}$	$-\frac{1}{\gamma}[\Sigma^{-1}]_{ii}$	$-\alpha(1-\omega_i)$
Cross Price Elasticity $\mathcal{E}_{ij}$	$-\frac{1}{\gamma}[\Sigma^{-1}]_{ij}$	$-\frac{1}{\gamma}[\Sigma^{-1}]_{ij}$	$\alpha\omega_j$
Relative Elasticity $\widehat{\mathcal{E}} = \mathcal{E}_{ii} - \mathcal{E}_{ji}$	$-\frac{1}{\gamma}\left([\Sigma^{-1}]_{ii} - [\Sigma^{-1}]_{ji}\right)$	$-\frac{1}{\gamma}\left([\Sigma^{-1}]_{ii} - [\Sigma^{-1}]_{ji}\right)$	$-\alpha$

Table A1: Three standard models of asset demand.

demand regression is well specified.[62] Table A1 summarizes the results.

Constant absolute risk aversion. In the mean-variance model (CARA) described in Section 1, we have seen the direct mapping between covariance matrix and the elasticity matrix when considering a relation between the level of demand and the level of prices: $\boldsymbol{\mathcal{E}} = \partial D / \partial P = -\gamma^{-1}\Sigma^{-1}$. Section 2.2.1 shows that Assumptions A1 and A2 are satisfied if the covariance matrix has a factor structure with factor loadings which depend on the observables.

Constant relative risk aversion. Preferences with constant relative risk aversion (CRRA) are the workhorse model of macro-finance. Utility in this case is given by $u(C) = C^{1-\gamma}/(1-\gamma)$, with now γ being the constant relative risk-aversion. Assume that the risk-free rate is r_f and that there are N assets with payoffs $X = \{X_i\}_i$ at time 1, with prices $\{P_i\}$. Hence, log asset returns are $R_i = \log(X_i/P_i)$.

To solve for the optimal demands, we assume that the payoffs follow a lognormal distribution: $\log X \sim \mathcal{N}(\bar{\Pi}, \Sigma)$ and log-linearize portfolio returns following Campbell and Viceira (2002).[63] For an investor with wealth W, the optimal demand is:

$$D_i = \frac{1}{\gamma}\frac{W}{P_i}\left[\Sigma^{-1}\left(\bar{\Pi} - \log P - r_f + \frac{1}{2}\operatorname{diag}(\Sigma)\right)\right]_i \tag{A72}$$

This implies that when considering the relation between portfolio weights, $\omega_i = P_i D_i / W$, and log prices, the elasticity matrix is (this relation is exact in continuous time; see Duffie,

[62] Petajisto (2009), Davis et al. (2025), and Davis (2024) quantify elasticities in these models.

[63] We log-linearize the return of portfolio ω as R_p:

$$R_p - r_f = \log\left(\omega' \exp\left(\mathbf{R} - r_f\right)\right) \simeq \omega'(\mathbf{R} - r_f) + \frac{1}{2}\omega' \operatorname{diag}(\Sigma) - \frac{1}{2}\omega'\Sigma\omega. \tag{A71}$$

2010; He et al., 2025):

$$\boldsymbol{\mathcal{E}} = \frac{\partial \omega}{\partial \log P} = -\frac{1}{\gamma}\Sigma^{-1}. \tag{A73}$$

This is the same elasticity as the CARA case, albeit with different units: portfolio weights on log prices. Therefore, our earlier discussion relating properties of the covariance matrix (here of log returns) and the identification assumptions apply to this case as well.

Constant relative risk aversion with a factor structure. Consider the factor structure of Section 1.3. We add a factor structure on log payoffs to the previous setting. Log returns inherit this structure because they have the same covariance matrix as payoffs. This corresponds to

$$\Sigma = \beta \Sigma_F \beta' + \sigma_v^2 \mathbf{I}_N, \tag{A74}$$

with Σ_F a $K \times K$ factor risk matrix, and β a $N \times K$ set of loadings. Equivalently we can represent log excess returns as:

$$R^e_{it} = R_{it} - r_f + \frac{1}{2}\Sigma_{ii} = \alpha_i + \beta_i F_t + v_{it}, \tag{A75}$$

where $v_{it} \sim \mathcal{N}(0, \sigma_v^2)$ is i.i.d. across assets and orthogonal to the factor $F_t \sim \mathcal{N}(E(F), \Sigma_F)$. The means are given by $E(F) = (\beta'\beta)^{-1}\beta' E(R^e)$ and $\alpha_i = E(R^e_i) - \beta_i E(F)$. Note that $\beta'\alpha = 0$. First the inverse of the covariance matrix multiplied by risk aversion is obtained from the Sherman-Morrison formula:

$$\gamma^{-1}\Sigma^{-1} = \frac{1}{\gamma\sigma_v^2}\mathbf{I}_N - \gamma^{-1}\sigma_v^{-2}\beta(\Sigma_F^{-1} + \beta'\sigma_v^{-2} I \beta)^{-1}\beta'\sigma_v^{-2}\mathbf{I}_N \tag{A76}$$

$$= \frac{1}{\gamma\sigma_v^2}\mathbf{I}_N - \frac{1}{\gamma\sigma_v^2}\sigma_v^{-2}\beta(\Sigma_F^{-1} + \beta'\sigma_v^{-2}\beta)^{-1}\beta' \tag{A77}$$

$$= \frac{1}{\gamma\sigma_v^2}\mathbf{I} - \beta\Psi\beta', \tag{A78}$$

where we define

$$\Psi = \frac{1}{\gamma\sigma_v^2}(\sigma_v^2\Sigma_F^{-1} + \beta'\beta)^{-1} \tag{A79}$$

Plugging in the demand equation (A72) above we obtain equation (5) in the paper:

$$\omega = \frac{1}{\gamma}\Sigma^{-1}E(R^e) \tag{A80}$$

$$= \frac{1}{\gamma\sigma_v^2}(\alpha + \beta E(F)) - \beta\Psi\beta'(\alpha + \beta E(F)) = \frac{1}{\gamma\sigma_v^2}\alpha + \beta\left(\frac{1}{\gamma\sigma_v^2}\mathbf{I}_K - \Psi\beta'\beta\right)E(F) \tag{A81}$$

$$= \frac{1}{\gamma\sigma_v^2}\alpha + \beta\Upsilon E(F), \tag{A82}$$

where we define

$$\Upsilon = \frac{1}{\gamma\sigma_v^2}\mathbf{I}_K - \Psi\beta'\beta. \tag{A83}$$

We derive the elasticity from the previous expression of demand shares:

$$\frac{d\omega}{d\log P} = \frac{1}{\gamma\sigma_v^2}\frac{d\alpha}{d\log P} + \beta\Upsilon\frac{dE(F)}{d\log P} \tag{A84}$$

$$= \frac{1}{\gamma\sigma_v^2}\frac{dE(R^e)}{d\log P} - \frac{1}{\gamma\sigma_v^2}\frac{\beta dE(F)}{d\log P} + \beta\Upsilon\frac{dE(F)}{d\log P} \tag{A85}$$

The elasticity combines two groups of effects. First changing prices of each asset directly translates into changing the alpha of each asset. Second, changing prices affects factor premia which shifts both the direct demand for the factor and indirectly the alphas (because the factors is the benchmark used to compute the alphas). Plugging in $dE(F)/d\log P = (\beta'\beta)^{-1}\beta'$, we obtain the elasticity matrix, equation (6) in the paper:

$$\boldsymbol{\mathcal{E}} = -\frac{d\omega}{d\log P} = -\frac{1}{\gamma\sigma^2}\mathbf{I}_N + \beta(\frac{\mathbf{I}_K}{\gamma\sigma^2} - \Upsilon)(\beta'\beta)^{-1}\beta' \tag{A86}$$

$$= -\frac{1}{\gamma\sigma^2}\mathbf{I}_N + \beta(\frac{\mathbf{I}_K}{\gamma\sigma^2} - \Upsilon)(\beta'\beta)^{-1}\beta' \tag{A87}$$

$$= -\frac{1}{\gamma\sigma_v^2}\mathbf{I}_N + \beta\Psi\beta'. \tag{A88}$$

Of course, we could have found this immediately from $\boldsymbol{\mathcal{E}} = -(\gamma\Sigma)^{-1}$.

Logit. Koijen and Yogo (2019) introduce a model of portfolio demand of the logit form. They show the existence of factor models giving rise to this demand for an investor with log utility. Unlike the model we just studied, these factor models feature a covariance matrix and expected returns that depend nonlinearly on prices, and hence have a different elasticity matrix; Appendix D details this distinction. The logit model is also commonly used in industrial organization, as well as in many fields in economics including trade and spatial equilibrium models. There, it is most often motivated by aggregation of a consumer discrete choice model, but can also apply to an individual's choice of consumption basket.[64]

For an investor with initial wealth W, the expenditure shares or portfolio weights are:[65]

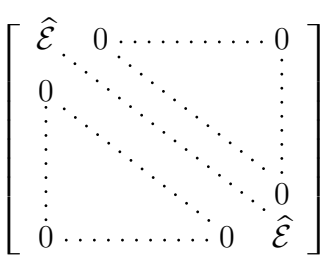
$$\omega_i = \frac{P_i D_i}{W} = \frac{\exp\left(-\alpha p_i + \theta' X_i + \epsilon_i\right)}{1 + \sum_l \exp\left(-\alpha p_l + \theta' X_l + \epsilon_l\right)}, \tag{A90}$$

where p_i is the log of the price of asset i, X_i observable demand shifters, and ϵ_i the unobserved

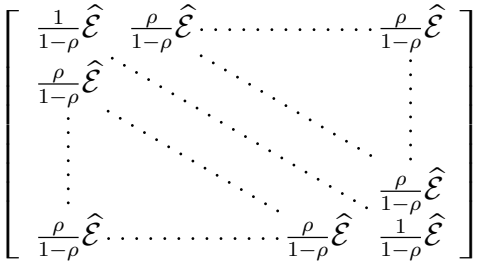
[64] Anderson et al. (1988) derive the utility that leads to logit shares as optimal demand.

[65] If there is no outside asset, the model of expenditure shares becomes:

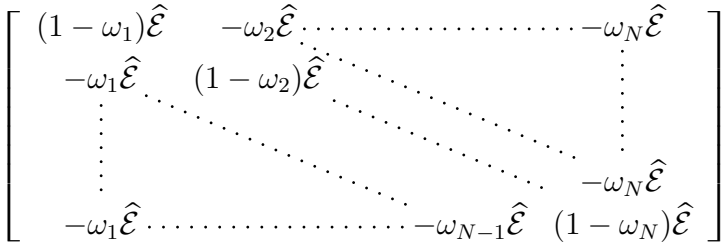
$$\omega_i = \frac{P_i D_i}{W} = \frac{\exp\left(-\alpha p_i + \theta' X_i + \epsilon_i\right)}{\sum_l \exp\left(-\alpha p_l + \theta' X_l + \epsilon_l\right)}. \tag{A89}$$



(a) Diagonal matrix.



(b) Symmetric matrix.



(c) Logit-style matrix.

Figure A1: Different elasticity matrices $\boldsymbol{\mathcal{E}}$ with the same relative elasticity $\widehat{\mathcal{E}}$

component of demand.

When considering the relation between log portfolio weights and log prices, the elasticity matrix is:

$$\boldsymbol{\mathcal{E}} = \frac{\partial \log \omega}{\partial \log P} = -\alpha \left(\mathbf{I} - \mathbf{1}\omega' \right), \tag{A91}$$

where ω is the vector of portfolio weights given in (A90). Note that $\boldsymbol{\mathcal{E}}$ in general is not symmetric in this case. The coefficient α is the only demand parameter that determines the matrix of demand elasticity, as opposed to the whole covariance matrix in the CARA and CRRA cases. Further, this matrix always satisfies assumptions A1 ($\mathcal{E}_{jk} = \mathcal{E}_{ik} = \alpha\omega_k$) and A2 ($\mathcal{E}_{ii} - \mathcal{E}_{ji} = -\alpha$), with $-\alpha$ being the relative elasticity of demand.

Same relative elasticity vs. same elasticity matrix. Note that the fact that the simple risk-based models and the logit model can both satisfy the two assumptions only implies that they lead to the same estimation of the relative elasticity. Even when these models have the same value of relative elasticity, they exhibit different elasticity matrices. Figure A1 illustrates this nuance, with three distinct elasticity matrices that share the same relative elasticity.

B.4 Demand beyond risk-based motives for substitution

Consider the problem of investors combining risk-based mean-variance demand where the covariance matrix Σ is characterized by a set of characteristics $X^{(3)}$ with a cost of holding assets that is quadratic in another set of characteristics $X^{(1)}$ and a portfolio constraint linear in yet another set of characteristics $X^{(2)}$. Section 2.2.2 is a special case of this that assumes that $X^{(1)}$ and $X^{(2)}$ each only contain one observable: carbon intensity and a bank's liquidity ratio. The proposition below generalizes this.

Proposition A2 (Mean-variance demand with quadratic cost and linear constraint) *Assume that investors choose their demand according to the problem*

$$\max_{D} \quad D'(M-P) - \frac{\gamma}{2} D'\Sigma D - \frac{\kappa}{2} D' X^{(1)} X^{(1)\prime} D \tag{A92}$$

$$\textit{such that} \quad D' X^{(2)} \leqslant \Theta, \tag{A93}$$

where D, M, and P are the $N \times 1$ vectors of investor demand, expected payoffs, and prices, γ is risk aversion, Σ the $N \times N$ covariance matrix, κ controls the quadratic cost function, $X^{(1)}$ and $X^{(2)}$ are the $N \times K_1$ and $N \times K_2$ matrices of stock characteristics, and Θ is a $1 \times K_2$ vector that controls the linear constraint.

Further assume that the risk-based component of investor demand satisfies assumptions A1 and A2, i.e.,

$$-\frac{1}{\gamma}\Sigma^{-1} = \mathcal{E}^{(3)}_{relative} I + X^{(3)} \boldsymbol{\mathcal{E}}_{X^{(3)}} X^{(3)\prime}, \tag{A94}$$

where $X^{(3)}$ is another $N \times K_3$ matrix of stock characteristics, and $\boldsymbol{\mathcal{E}}_{X^{(3)}}$ is the $K_3 \times K_3$ matrix of substitution between observables.

Then the resulting demand curve satisfies assumptions A1 and A2 conditional on the stacked observables $\mathbb{X} = [X^{(1)}, X^{(2)}, X^{(3)}]$.

Proof. By Lemma A2, to prove the proposition, we need to show that the elasticity matrix $\mathcal{E}$ can be expressed as

$$\boldsymbol{\mathcal{E}} = \mathcal{E}_{relative} I + \mathbb{X} \boldsymbol{\mathcal{E}}_{\mathbb{X}} \mathbb{X}'. \tag{A95}$$

Start by putting together equations (A92) and (A93) in the Lagrangian

$$\mathcal{L}(D, \lambda) = D'(M-P) - \frac{\gamma}{2} D'\Sigma D - \frac{\kappa}{2} D' X^{(1)} X^{(1)\prime} D - (D' X^{(2)} - \Theta)\lambda, \tag{A96}$$

where λ is the $K_2 \times 1$ Lagrange multiplier on the linear constraint.

Setting the first-order condition with respect to D to zero, and solving for D, yields

$$D = \left(\underbrace{\gamma\Sigma + \kappa X^{(1)} X^{(1)\prime}}_{\equiv \Omega} \right)^{-1} \left(M - P - X^{(2)}\lambda \right) \tag{A97}$$

$$= \Omega^{-1} \left(M - P - X^{(2)}\lambda \right), \tag{A98}$$

where

$$\Omega = \gamma\Sigma + \kappa X^{(1)} X^{(1)\prime}. \tag{A99}$$

Plugging into the linear constraint to solve for λ when constraints are binding (otherwise the corresponding components of λ are equal to 0):

$$\left(M - P - X^{(2)}\lambda \right)' \Omega^{-1} X^{(2)} = \Theta \tag{A100}$$

$$\implies (M - P)' \Omega^{-1} X^{(2)} - \lambda' X^{(2)\prime} \Omega^{-1} X^{(2)} = \Theta \tag{A101}$$

$$\implies \lambda = \left[X^{(2)\prime} \Omega^{-1} X^{(2)} \right]^{-1} \left[X^{(2)\prime} \Omega^{-1} (M - P) - \Theta' \right] \tag{A102}$$

Plugging the Lagrange multipliers back into optimal investor demand gives:

$$D = \Omega^{-1} (M - P) - \Omega^{-1} X^{(2)} \left[X^{(2)\prime} \Omega^{-1} X^{(2)} \right]^{-1} \left[X^{(2)\prime} \Omega^{-1} (M - P) - \Theta' \right] \tag{A103}$$

The elasticity matrix therefore is:

$$\frac{dD}{dP} = -\Omega^{-1} + \Omega^{-1} X^{(2)} S_b \left[S_b' X^{(2)\prime} \Omega^{-1} X^{(2)} S_b \right]^{-1} S_b' X^{(2)\prime} \Omega^{-1} \tag{A104}$$

Here, S_b is the binding constraint selection matrix, which for the first-order condition selects the columns of $X^{(2)}$ for which constraints are binding.

Start now with the part for when the inequality constraints are all non-binding:

$$-\Omega^{-1} = \mathcal{E}^{(3)}_{relative} I + X^{(3)} \boldsymbol{\mathcal{E}}_{X^{(3)}} X^{(3)\prime} \tag{A105}$$

$$+ \left(\mathcal{E}^{(3)}_{relative} I + X^{(3)} \boldsymbol{\mathcal{E}}_{X^{(3)}} X^{(3)\prime}\right) X^{(1)} \underbrace{\left[\frac{1}{\kappa} I + X^{(1)\prime} (\gamma\Sigma)^{-1} X^{(1)}\right]^{-1}}_{\equiv \mathcal{H}} X^{(1)\prime} \left(\mathcal{E}^{(3)}_{relative} I + X^{(3)} \boldsymbol{\mathcal{E}}_{X^{(3)}} X^{(3)\prime}\right) \tag{A106}$$

$$= \mathcal{E}^{(3)}_{relative} I + X^{(3)} \boldsymbol{\mathcal{E}}_{X^{(3)}} X^{(3)\prime} + \left(\mathcal{E}^{(3)}_{relative}\right)^2 X^{(1)} \mathcal{H} X^{(1)\prime} + \mathcal{E}^{(3)}_{relative} X^{(1)} \mathcal{H} X^{(1)\prime} X^{(3)} \boldsymbol{\mathcal{E}}_{X^{(3)}} X^{(3)\prime} \tag{A107}$$

$$+ \mathcal{E}^{(3)}_{relative} X^{(3)} \boldsymbol{\mathcal{E}}_{X^{(3)}} X^{(3)\prime} X^{(1)} \mathcal{H} X^{(1)\prime} + X^{(3)} \boldsymbol{\mathcal{E}}_{X^{(3)}} X^{(3)\prime} X^{(1)} \mathcal{H} X^{(1)\prime} X^{(3)} \boldsymbol{\mathcal{E}}_{X^{(3)}} X^{(3)\prime} \tag{A108}$$

$$= \mathcal{E}^{(3)}_{relative} I + \underbrace{[X^{(1)}, X^{(3)}]}_{\equiv X^{(1,3)}} \underbrace{\begin{bmatrix} \left(\mathcal{E}^{(3)}_{relative}\right)^2 \mathcal{H} & \mathcal{E}^{(3)}_{relative} \mathcal{H} X^{(1)\prime} X^{(3)} \boldsymbol{\mathcal{E}}_{X^{(3)}} \\ \mathcal{E}^{(3)}_{relative} \boldsymbol{\mathcal{E}}_{X^{(3)}} X^{(3)\prime} X^{(1)} \mathcal{H} & \boldsymbol{\mathcal{E}}_{X^{(3)}} + \boldsymbol{\mathcal{E}}_{X^{(3)}} X^{(3)\prime} X^{(1)} \mathcal{H} X^{(1)\prime} X^{(3)} \boldsymbol{\mathcal{E}}_{X^{(3)}} \end{bmatrix}}_{\equiv \mathcal{F}} [X^{(1)}, X^{(3)}]' \tag{A109}$$

$$= \mathcal{E}^{(3)}_{relative} I + X^{(1,3)} \mathcal{F} X^{(1,3)\prime} \tag{A110}$$

When the linear constraints are not binding, the elasticity matrix satisfies assumptions A1 and A2 conditional on the stacked observables $[X^{(1)}, X^{(3)}]$.

For the case where some constraints are binding, define:

$$\mathcal{G} \equiv S_b \left[S_b' X^{(2)\prime} \Omega^{-1} X^{(2)} S_b\right]^{-1} S_b' \tag{A111}$$

The elasticity matrix is

$$\frac{dD}{dP} = \mathcal{E}^{(3)}_{relative} I + X^{(1,3)} \mathcal{F} X^{(1,3)\prime} + \left(\mathcal{E}^{(3)}_{relative} I + X^{(1,3)} \mathcal{F} X^{(1,3)\prime}\right) X^{(2)} \mathcal{G} X^{(2)\prime} \left(\mathcal{E}^{(3)}_{relative} I + X^{(1,3)} \mathcal{F} X^{(1,3)\prime}\right) \tag{A112}$$

$$= \mathcal{E}_{relative} I + \mathbb{X} \boldsymbol{\mathcal{E}}_{\mathbb{X}} \mathbb{X}', \tag{A113}$$

where

$$\mathcal{E}_{relative} = \mathcal{E}^{(3)}_{relative} \tag{A114}$$

$$\mathbb{X} = \left[X^{(1)}, X^{(3)}, X^{(2)}\right] \tag{A115}$$

$$\boldsymbol{\mathcal{E}}_{\mathbb{X}} = \begin{bmatrix} \mathcal{F} + \mathcal{F} X^{(1,3)\prime} X^{(2)} \mathcal{G} X^{(2)\prime} X^{(1,3)} \mathcal{F} & \mathcal{E}^{(3)}_{relative} \mathcal{F} X^{(1,3)\prime} X^{(2)} \mathcal{G} \\ \mathcal{E}^{(3)}_{relative} \mathcal{G} X^{(2)\prime} X^{(1,3)} \mathcal{F} & \left(\mathcal{E}^{(3)}_{relative}\right)^2 \mathcal{G} \end{bmatrix}. \tag{A116}$$

The elasticity matrix satisfies assumptions A1 and A2 conditional on the stacked observ-

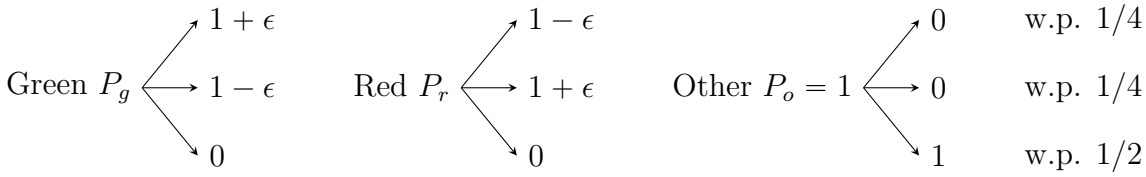
ables $\mathbb{X}$. ■

B.5 What about equilibrium spillovers?

The reader might be surprised that, so far, we have not discussed the concept of equilibrium, which is usually central in asset pricing. This is not because we assume that the world is not in equilibrium: equation (3) also holds for a change in demand in equilibrium. Instead, we can do so because identifying specific sources of variation in prices — the instrument Z_i — and assuming that an investor's demand satisfies Assumptions A1 and A2 is enough to estimate this investor's demand elasticity without understanding the entire structure of the equilibrium.

In this section, we work out a simple equilibrium model to illustrate this insight. The setting is inspired by Fuchs et al. (2025a), who point out that endogenous cross-asset spillovers can imply a low measured own-price elasticity even if the true own price elasticity is near infinite. This result considers a different regression from the causal inference framework of this paper. We show that the issue arises because the omitted variable bias that we have pointed out in Section 1.2 is present. We also explain that, because the example satisfies Assumptions A1 and A2, a standard difference-in-difference regression is unbiased, once we recognize that it recovers the relative elasticity.

Setting. The economy is populated by a representative agent with log utility. There are three assets with different payoffs in three possible states of the world, with payoffs as follows:



The "other" asset acts as a numeraire, whose price is normalized to 1. Denote the prices of the green and red assets P_g and P_r. These two assets become closer substitutes as ϵ goes towards 0. Indeed, in the limit, any price difference between P_g and P_r represents an arbitrage opportunity. The representative agent has endowments E_g, E_r, and E_o, which implies that their wealth is $W = P_g E_g + P_r E_r + E_o$.

Demand and equilibrium. We can first derive the demand function, that is, the optimal portfolio share as a function of prices:

$$\omega_g(P_g, P_r) = \frac{P_g}{2}\frac{\epsilon^2(P_r+P_g)+(P_r-P_g)}{\epsilon^2(P_r+P_g)^2-(P_r-P_g)^2}, \quad \omega_r(P_g, P_r) = \frac{P_r}{2}\frac{\epsilon^2(P_r+P_g)+(P_g-P_r)}{\epsilon^2(P_r+P_g)^2-(P_r-P_g)^2}. \tag{A117}$$

Market-clearing for the two assets, $\omega_g W = P_g E_g$ and $\omega_r W = P_r E_r$ lead to equilibrium prices as functions of the endowments:

$$P_g\left(E_o, E_g, E_r\right) = E_o \frac{\epsilon^2\left(E_g - E_r\right) - \left(E_g + E_r\right)}{\epsilon^2\left(E_g - E_r\right)^2 - \left(E_g + E_r\right)^2}, \quad P_r\left(E_o, E_g, E_r\right) = E_o \frac{\epsilon^2\left(E_r - E_g\right) - \left(E_g + E_r\right)}{\epsilon^2\left(E_g - E_r\right)^2 - \left(E_g + E_r\right)^2}. \tag{A118}$$

As an initial equilibrium, we assume that the endowments are $E_g = E_r = 1/2$ and $E_o = 1$. It is then immediate that $P_r = P_g = 1$.

Demand elasticities. We can compute the demand elasticities: how individual demand would respond to a change in prices. Because utility is CRRA, we measure the sensitivity of portfolio shares to log prices (in line with Section B.3) around the initial equilibrium values of prices:

$$\mathcal{E}_{own} = \frac{\partial \omega_g}{\partial \log P_g} = \frac{1}{8} - \frac{1}{8\epsilon^2}; \tag{A119}$$

$$\mathcal{E}_{cross} = \frac{\partial \omega_g}{\partial \log P_r} = -\frac{1}{8} + \frac{1}{8\epsilon^2}. \tag{A120}$$

The expressions for ω_r are identical. The relative elasticity is $\mathcal{E}_{own} - \mathcal{E}_{cross} = 1/4 - 1/(4\epsilon^2)$.

These measures show that when the two assets are near-identical, $\epsilon \to 0$, any deviation from parity would lead to a near-infinite increase in demand for the cheaper asset, and near-infinite decrease in demand for the more expensive one. This is the standard arbitrage argument.

Running regressions. We are interested in whether various regressions around a supply shock for one of the assets can identify these elasticities. A shift in supply of the green asset leads to the price changes

$$\frac{\partial \log P_g}{\partial E_g} = -\left(1 + \varepsilon^2\right), \qquad \frac{\partial \log P_r}{\partial E_g} = -\left(1 - \varepsilon^2\right), \tag{A121}$$

around the equilibrium.

Fuchs et al. (2025a) correctly point out that regressing the demand for the green asset on the change in its price using such a supply shock does not recover the own price elasticity. This regression corresponds to taking the ratio of the change in portfolio to the change in price across equilibria:

$$\frac{d\omega_g / dE_g}{d \log P_g / dE_g} = -\frac{1}{4}\frac{1 - \varepsilon^2}{1 + \varepsilon^2} \neq \mathcal{E}_{own}. \tag{A122}$$

In particular, when $\varepsilon \to 0$, the regression coefficient on the left-hand-side converges to $-1/4$, while the own-price elasticity goes to infinity. Unpacking the total derivative explains the

source of the bias:

$$\frac{d\omega_g/dE_g}{d\log P_g/dE_g} = \frac{\frac{\partial\omega_g}{\partial\log P_g}\frac{\partial\log P_g}{\partial E_g} + \frac{\partial\omega_g}{\partial\log P_r}\frac{\partial\log P_r}{\partial E_g}}{\frac{d\log P_g}{dE_g}} = \mathcal{E}_{own} + \mathcal{E}_{cross}\frac{\partial\log P_r/\partial E_g}{\partial\log P_g/\partial E_g} \tag{A123}$$

The change in demand for the green asset is not driven only by the change in its own price but also by the change in price of its substitute, the red asset, because $\mathcal{E}_{cross} \neq 0$. In the language of regressions, the price of the red asset is acting as a correlated omitted variable. Intuitively, the induced price drop of the green asset would lead to a large increase of its demand if the price of the red asset remained high. However, in equilibrium the price of the red asset drops too, resulting in only a moderate change in the demand for the green asset.

However, the canonical causal inference framework corresponds to a standard difference-in-difference regression in this setting with two assets. The regression coefficient is the ratio of the difference of change in portfolio weight to the difference of change in price, as in equation (25):

$$\frac{d\omega_g/dE_g - d\omega_r/dE_g}{d\log P_g/dE_g - d\log P_r/dE_g} = \frac{1}{4} - \frac{1}{4\varepsilon^2} = \mathcal{E}_{own} - \mathcal{E}_{cross} = \mathcal{E}_{relative}. \tag{A124}$$

The difference-in-difference coefficient correctly identifies the relative elasticity. Indeed, this setting satisfies Assumptions A1 and A2: because the elasticity matrix is symmetric, substitution is homogeneous and the relative elasticity is constant. Also, because the endowment shock does not have a direct effect on log investors' choice of portfolio shares, the standard exogeneity condition is satisfied. Note that the identified relative elasticity is unbounded when ϵ converges to 0, in line with the economic intuition that the relative demand for near arbitrage assets should react strongly to a change in their relative price. The estimator leads to this limit because the relative change in portfolio remains finite, while the relative change in price goes to zero in this limit due to arbitrage.

This example highlights an important conceptual point: Demand elasticities are well defined regardless of the source of the change in prices. Precisely, the demand curve maps the price vector to the quantity vector through any source of change in price that is not accompanied by changes in other drivers of demand. As a result, it is not surprising that equilibrium spillovers are not a problem for the identification of demand elasticities per se. Instead, the econometrician has to be careful that prices of other assets might introduce omitted variable bias. Assumptions A1 and A2 ensure that this is not the case for a standard difference-in-difference estimator, which is then an unbiased estimator of relative elasticity.

B.6 Learning from prices, strategic trading, dynamic trading

In the main text, we showed that the linear approximation of the structural demand function

$$\mathbf{D}_t = \boldsymbol{\mathcal{E}}\,\mathbf{P}_t + \boldsymbol{\epsilon}_t \tag{A125}$$

where $\mathbf{D}_t \in \mathbb{R}^n$ is the vector of (net) demands for n assets, $\mathbf{P}_t \in \mathbb{R}^n$ is the price vector, $\boldsymbol{\mathcal{E}} \in \mathbb{R}^{n\times n}$ is a constant matrix, and $\boldsymbol{\epsilon} \in \mathbb{R}^n$ is a price-independent intercept, is consistent with

investors trading for risk-sharing purposes, potentially in the presence of portfolio constraints and a factor structure of returns. Here we present two examples building on standard models to demonstrate that this form is also consistent with: (i) asymmetric information models where investors learn from prices, and (ii) models based on strategic trading with imperfect competition. However, we also emphasize some caveats. We explain that we implicitly require a sufficiently stable economic environment that (A125) remains meaningful to estimate. Also, in the last part of this appendix we discuss that in dynamic environments demand functions might be considerably more complex to be represented by a mapping like (A125).

B.6.1 Admati (1985): competitive traders learning from signals and prices with multiple assets

This section presents a multi-asset noisy rational expectations model following Admati (1985).

Environment. There are n risky assets with payoff vector $\mathbf{m} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_m)$ and price vector $\mathbf{p} \in \mathbb{R}^n$. A continuum of competitive, risk-averse traders indexed by $i \in [0,1]$ have CARA utility with absolute risk aversion $\gamma > 0$.

Each trader observes a private signal

$$\mathbf{s}_i = \mathbf{m} + \boldsymbol{\varepsilon}_i, \qquad \boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\varepsilon),$$

i.i.d. across i and independent of $\mathbf{m}$. Traders also observe the equilibrium price $\mathbf{p}$ and use it to update beliefs. Noise traders supply an exogenous quantity $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_u)$.

Individual demand. Trader i is a price-taker and chooses position $\mathbf{d}_i \in \mathbb{R}^n$ to maximize CARA expected utility. Standard CARA-normal optimization yields:

$$\mathbf{d}_i = \frac{1}{\gamma}\, \boldsymbol{\Sigma}_{m|s,p}^{-1}\big(\mathbb{E}[\mathbf{m} \mid \mathbf{s}_i, \mathbf{p}] - \mathbf{p}\big),$$

where $\boldsymbol{\Sigma}_{m|s,p} = \mathrm{Var}(\mathbf{m} \mid \mathbf{s}_i, \mathbf{p})$ is the posterior covariance.

Bayesian updating. In equilibrium, the price is informationally equivalent to a signal $\boldsymbol{\eta} = \mathbf{m} + \boldsymbol{\xi}$ for some noise $\boldsymbol{\xi}$ with covariance $\boldsymbol{\Sigma}_\eta$. The posterior precision combines prior and signal precisions:

$$\boldsymbol{\Sigma}_{m|s,p}^{-1} = \boldsymbol{\Sigma}_m^{-1} + \boldsymbol{\Sigma}_\varepsilon^{-1} + \boldsymbol{\Sigma}_\eta^{-1}.$$

The posterior mean is a precision-weighted average:

$$\mathbb{E}[\mathbf{m} \mid \mathbf{s}_i, \mathbf{p}] = \boldsymbol{\Sigma}_{m|s,p}\big(\boldsymbol{\Sigma}_\varepsilon^{-1}\mathbf{s}_i + \boldsymbol{\Sigma}_\eta^{-1}\boldsymbol{\eta}\big).$$

Aggregate demand. Integrating over traders and using $\int_0^1 \mathbf{s}_i\, di = \mathbf{m}$ by the law of large numbers:

$$\mathbf{D} = \int_0^1 \mathbf{d}_i\, di = \frac{1}{\gamma}\boldsymbol{\Sigma}_\varepsilon^{-1}(\mathbf{m} - \mathbf{p}) + \frac{1}{\gamma}\boldsymbol{\Sigma}_\eta^{-1}(\boldsymbol{\eta} - \mathbf{p}) + \frac{1}{\gamma}\boldsymbol{\Sigma}_m^{-1}(-\mathbf{p}).$$

Equilibrium. Market clearing $\mathbf{D} = \mathbf{u}$ determines the equilibrium. In the unique linear equilibrium, the price signal precision satisfies:

$$\boldsymbol{\Sigma}_\eta^{-1} = \frac{1}{\gamma^2}\boldsymbol{\Sigma}_\varepsilon^{-1}\boldsymbol{\Sigma}_u^{-1}\boldsymbol{\Sigma}_\varepsilon^{-1}.$$

The precision of $\mathbf{m}$ given prices alone is:

$$\boldsymbol{\Sigma}_{m|p}^{-1} = \boldsymbol{\Sigma}_m^{-1} + \boldsymbol{\Sigma}_\eta^{-1} = \boldsymbol{\Sigma}_m^{-1} + \frac{1}{\gamma^2}\boldsymbol{\Sigma}_\varepsilon^{-1}\boldsymbol{\Sigma}_u^{-1}\boldsymbol{\Sigma}_\varepsilon^{-1}.$$

Prices become fully revealing as $\gamma \to 0$ or $\boldsymbol{\Sigma}_u \to \mathbf{0}$.

Mapping to $\boldsymbol{\mathcal{E}}$ and $\boldsymbol{\epsilon}$. Writing aggregate demand as $\mathbf{D} = \boldsymbol{\mathcal{E}}\mathbf{p} + \boldsymbol{\epsilon}$:

$$\boldsymbol{\mathcal{E}} = -\frac{1}{\gamma}\big(\boldsymbol{\Sigma}_m^{-1} + \boldsymbol{\Sigma}_\varepsilon^{-1} + \boldsymbol{\Sigma}_\eta^{-1}\big) = -\frac{1}{\gamma}\boldsymbol{\Sigma}_{m|s,p}^{-1},$$

$$\boldsymbol{\epsilon} = \frac{1}{\gamma}\big(\boldsymbol{\Sigma}_\varepsilon^{-1}\mathbf{m} + \boldsymbol{\Sigma}_\eta^{-1}\boldsymbol{\eta}\big).$$

Cross-asset demand elasticities arise from learning: even if private signals are independent across assets ($\boldsymbol{\Sigma}_\varepsilon$ diagonal), correlated noise trading ($\boldsymbol{\Sigma}_u$ non-diagonal) creates cross-asset information linkages through $\boldsymbol{\Sigma}_\eta^{-1}$.

Identification of $\boldsymbol{\mathcal{E}}$. To identify the demand elasticity matrix $\boldsymbol{\mathcal{E}}$ empirically, one needs instruments that shift prices but are uncorrelated with the intercept $\boldsymbol{\epsilon}$. In this model, $\boldsymbol{\epsilon}$ depends on the fundamental $\mathbf{m}$ and the price signal $\boldsymbol{\eta}$—both of which carry information about payoffs. A valid instrument must therefore be a *non-informational* shock to prices. The natural candidate is noise trading $\mathbf{u}$: shocks to the supply from noise traders move equilibrium prices but are, by assumption, orthogonal to fundamentals $\mathbf{m}$ and hence to $\boldsymbol{\epsilon}$. In practice, this could correspond to index rebalancing flows, liquidity-driven sales (e.g., mutual fund redemptions), or other demand shocks uncorrelated with asset payoffs.

Importantly, while a shock to noise trading is a valid instrument, a shock that investors *learn* to originate from non-fundamental sources might *not* be valid in this model. For example, if the Federal Reserve unexpectedly announces a random purchase of certain assets, investors would know that the resulting price movement is uninformative about $\mathbf{m}$. However, such knowledge constitutes additional information not present in the model setup: here, investors assign probabilities to the sources of price changes based on their priors ($\boldsymbol{\Sigma}_m$, $\boldsymbol{\Sigma}_\varepsilon$, $\boldsymbol{\Sigma}_u$) and cannot perfectly disentangle fundamental from non-fundamental shocks. If they could identify the shock's source with certainty, their inference problem—and hence the equilibrium demand intercept $\boldsymbol{\epsilon}$—would change.

This points to a crucial, even if implicit, assumption behind our analysis. We assume that *the environment is sufficiently stable* that the $\mathbf{D}_t = \boldsymbol{\mathcal{E}}\,\mathbf{P}_t + \boldsymbol{\epsilon}_t$ relationship continues to hold at least approximately, at least for the relevant period. In this particular framework, this implies that the parameters of the model are stable.

B.6.2 Multi-asset Kyle (1989): imperfect competition with common signal

This section presents a multi-asset version of Kyle (1989)'s model of imperfect competition with demand schedules. The setup and solution method follow the unified framework developed by Rostek and Yoon (2025), who provide a comprehensive treatment of imperfectly competitive financial markets encompassing static, dynamic, centralized, and decentralized settings. Our common-signal specification is a special case of their general framework with symmetric traders and no inference from prices.

Environment. There are n risky assets with payoff vector $\mathbf{m} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_m)$ and prices $\mathbf{p}$. There are $N \geqslant 3$ risk-averse speculators with CARA utility (risk aversion $\gamma > 0$), all observing the *same* signal:

$$\mathbf{s} = \mathbf{m} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\varepsilon).$$

Noise traders supply $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_u)$. Traders submit demand schedules (not quantities), and the market clears without a market maker. Because signals are common, prices reveal no information beyond $\mathbf{s}$.

Posterior beliefs. All traders share the same posterior:

$$\mathbb{E}[\mathbf{m} \mid \mathbf{s}] = \boldsymbol{\Sigma}_m(\boldsymbol{\Sigma}_m + \boldsymbol{\Sigma}_\varepsilon)^{-1}\mathbf{s}, \qquad \boldsymbol{\Sigma}_{m|s} = \left(\boldsymbol{\Sigma}_m^{-1} + \boldsymbol{\Sigma}_\varepsilon^{-1}\right)^{-1}.$$

Equilibrium characterization. Following Rostek and Yoon (2025), equilibrium is characterized by two conditions: (i) each trader optimizes given their price impact, and (ii) price impacts are mutually consistent (a fixed point condition). In the symmetric equilibrium, each speculator k submits a demand schedule:

$$\mathbf{d}_k(\mathbf{p}, \mathbf{s}) = \mathbf{B}\mathbf{s} - \boldsymbol{\Gamma}\mathbf{p},$$

where $\boldsymbol{\Gamma}$ is the price sensitivity matrix. The individual price impact $\boldsymbol{\Lambda}_I$ (the slope of each trader's inverse residual supply) satisfies the fixed point condition:

$$\boldsymbol{\Lambda}_I = \left[(N-1)\boldsymbol{\Gamma}\right]^{-1}.$$

Optimization. Each speculator k maximizes expected utility taking price impact into account. The first-order condition equates marginal utility to marginal payment:

$$\mathbb{E}[\mathbf{m} \mid \mathbf{s}] - \gamma\boldsymbol{\Sigma}_{m|s}\mathbf{d}_k = \mathbf{p} + \boldsymbol{\Lambda}_I\mathbf{d}_k,$$

which gives the best response:

$$\mathbf{d}_k = (\gamma\boldsymbol{\Sigma}_{m|s} + \boldsymbol{\Lambda}_I)^{-1}(\mathbb{E}[\mathbf{m} \mid \mathbf{s}] - \mathbf{p}).$$

Hence $\boldsymbol{\Gamma} = (\gamma\boldsymbol{\Sigma}_{m|s} + \boldsymbol{\Lambda}_I)^{-1}$.

Solving the fixed point. Substituting $\boldsymbol{\Gamma}$ into the fixed point condition:

$$\boldsymbol{\Lambda}_I = \frac{\gamma \boldsymbol{\Sigma}_{m|s} + \boldsymbol{\Lambda}_I}{N-1} \quad \Rightarrow \quad (N-2)\boldsymbol{\Lambda}_I = \gamma \boldsymbol{\Sigma}_{m|s} \quad \Rightarrow \quad \boxed{\boldsymbol{\Lambda}_I = \frac{\gamma}{N-2}\,\boldsymbol{\Sigma}_{m|s}}$$

Equilibrium coefficients. Substituting back:

$$\boxed{\boldsymbol{\Gamma} = \frac{N-2}{\gamma(N-1)}\,\boldsymbol{\Sigma}_{m|s}^{-1} = \frac{N-2}{\gamma(N-1)}\big(\boldsymbol{\Sigma}_m^{-1} + \boldsymbol{\Sigma}_\varepsilon^{-1}\big)}$$

For the signal sensitivity $\mathbf{B}$, matching coefficients on $\mathbf{s}$ in the best response (using $\mathbb{E}[\mathbf{m} \mid \mathbf{s}] = \boldsymbol{\Sigma}_m(\boldsymbol{\Sigma}_m + \boldsymbol{\Sigma}_\varepsilon)^{-1}\mathbf{s}$):

$$\boxed{\mathbf{B} = \boldsymbol{\Gamma}\,\boldsymbol{\Sigma}_m(\boldsymbol{\Sigma}_m + \boldsymbol{\Sigma}_\varepsilon)^{-1} = \frac{N-2}{\gamma(N-1)}\,\boldsymbol{\Sigma}_\varepsilon^{-1}}$$

where the last equality uses the matrix identity $\boldsymbol{\Sigma}_{m|s}^{-1}\boldsymbol{\Sigma}_m(\boldsymbol{\Sigma}_m + \boldsymbol{\Sigma}_\varepsilon)^{-1} = \boldsymbol{\Sigma}_\varepsilon^{-1}$.

Market clearing and aggregate price impact. From market clearing $\sum_{k=1}^{N} \mathbf{d}_k = \mathbf{u}$, the equilibrium price satisfies:

$$\mathbf{p} = \boldsymbol{\Lambda}\big[N\mathbf{B}\mathbf{s} - \mathbf{u}\big], \qquad \boldsymbol{\Lambda} = (N\boldsymbol{\Gamma})^{-1} = \frac{\gamma(N-1)}{N(N-2)}\,\boldsymbol{\Sigma}_{m|s}.$$

The relationship between individual and aggregate price impact is $\boldsymbol{\Lambda}_I = \frac{N}{N-1}\boldsymbol{\Lambda}$.

Mapping to $\boldsymbol{\mathcal{E}}$ and $\boldsymbol{\epsilon}$. Aggregate demand is $\mathbf{D} = N\mathbf{d}_k = N\mathbf{B}\mathbf{s} - N\boldsymbol{\Gamma}\mathbf{p}$. Thus:

$$\boldsymbol{\mathcal{E}} = -N\boldsymbol{\Gamma} = -\frac{N(N-2)}{\gamma(N-1)}\big(\boldsymbol{\Sigma}_m^{-1} + \boldsymbol{\Sigma}_\varepsilon^{-1}\big),$$

$$\boldsymbol{\epsilon} = N\mathbf{B}\mathbf{s} = \frac{N(N-2)}{\gamma(N-1)}\boldsymbol{\Sigma}_\varepsilon^{-1}\mathbf{s}.$$

Unlike the competitive Hellwig/Admati model, cross-asset elasticities here arise purely from the covariance structure of payoffs and signals—not from learning, since all traders share the same information.

Identification of $\boldsymbol{\mathcal{E}}$. To identify $\boldsymbol{\mathcal{E}}$ in this setting, one again needs instruments that shift prices but are uncorrelated with $\boldsymbol{\epsilon}$. Here, $\boldsymbol{\epsilon}$ depends on the common signal $\mathbf{s} = \mathbf{m} + \boldsymbol{\varepsilon}$, which contains information about fundamentals. As in the Hellwig/Admati model, noise trading $\mathbf{u}$ is a valid instrument: it shifts equilibrium prices (through market clearing) but is orthogonal to $\mathbf{s}$ and hence to $\boldsymbol{\epsilon}$. An interesting feature of this model is that $\boldsymbol{\Sigma}_u$ does not appear in the equilibrium demand coefficients $(\mathbf{B}, \boldsymbol{\Gamma})$—noise trading affects price *levels* but not how traders respond to signals or prices. Nevertheless, cross-sectional or time-series variation in $\mathbf{u}$ (e.g., ETF creation/redemption activity, futures rolls, or settlement-driven flows) provides

the exogenous price variation needed to trace out demand curves. Shocks correlated with the signal—such as public news announcements that also enter $\mathbf{s}$—would be invalid instruments.

B.6.3 Dynamic environments

In all our motivating models, here and in the main text, the distribution of investors' relevant pay-offs is exogenously given. In standard dynamic settings as in Merton (1973), this is not the case. The key issue is that in dynamic models, investors' demand depends not only on current prices but on the entire distribution of future returns—including volatilities, covariances, and correlations with future investment opportunities. Simply, demand curves have higher dimensions in a dynamic environment. There is still a notion of demand curve, but demand shifts ϵ depend on other equilibrium outcomes. As He et al. (2025) argues, in these environments an instrument based on a supply shock is less trivially sufficient for identification than in static models.

C Robustness to deviations from the assumptions

We assess the robustness of the identification result of Proposition 1 with respect to deviations from the assumptions. For simplicity, we focus on a case with two assets and deviations from Assumption A2. The argument generalizes to other types of deviations.

We first show that the conclusions are robust when the first stage is economically large. Then we illustrate potential issues in the presence of weak instruments in the context of a model.

C.1 Robustness to deviations from Assumption A2

Consider a setting with two assets and an arbitrary elasticity matrix. In response to an exogenous shock, the relative change in demand is:

$$\begin{aligned} \Delta D_1 - \Delta D_2 &= \mathcal{E}_{11}\Delta P_1 + \mathcal{E}_{12}\Delta P_2 - \mathcal{E}_{22}\Delta P_2 - \mathcal{E}_{21}\Delta P_1 && \text{(A126)} \\ &= (\mathcal{E}_{11} - \mathcal{E}_{21})\Delta P_1 + (\mathcal{E}_{22} - \mathcal{E}_{12})\Delta P_2 && \text{(A127)} \end{aligned}$$

We denote the relative elasticities by $\mathcal{E}_{rel,1} = \mathcal{E}_{11} - \mathcal{E}_{21}$ and $\mathcal{E}_{rel,2} = \mathcal{E}_{22} - \mathcal{E}_{12}$. Rearranging the terms leads to

$$\Delta D_1 - \Delta D_2 = \frac{\mathcal{E}_{rel,1} + \mathcal{E}_{rel,2}}{2}(\Delta P_1 - \Delta P_2) + \frac{\mathcal{E}_{rel,1} - \mathcal{E}_{rel,2}}{2}(\Delta P_1 + \Delta P_2) \tag{A128}$$

Dividing by the relative change in price in response to the shock, we obtain the estimator:

$$\frac{\Delta D_1 - \Delta D_2}{\Delta P_1 - \Delta P_2} = \frac{\mathcal{E}_{rel,1} + \mathcal{E}_{rel,2}}{2} + \frac{\mathcal{E}_{rel,1} - \mathcal{E}_{rel,2}}{2}\frac{\Delta P_1 + \Delta P_2}{\Delta P_1 - \Delta P_2} \tag{A129}$$

The first term is the average relative elasticity. The second term is the potential bias: the heterogeneity of relative elasticities times the ratio of sum of changes in prices to their

difference. The denominator $\Delta P_1 - \Delta P_2$ is the first stage of the regression, and the relevance condition is $\Delta P_1 - \Delta P_2 \neq 0$.

It is straightforward to see that if the relevance condition is satisfied, the identification result of relative elasticity is robust to small deviations from assumption A2. Formally, consider a family of experiments (or models) indexed by a variable x, such that the experiment for $x = 0$ satisfies assumption A2 but it does not otherwise. If elasticities and changes in prices are continuous in x, then the IV estimator is also continuous in x as long as $\Delta P_1 - \Delta P_2 \neq 0$.

When the relevance condition is not satisfied, the last term of equation (A129) becomes infinite, and the bias is large relative to the actual coefficient. The same issue arises with weak instruments in the neighborhood of this limit, when $\Delta P_1 - \Delta P_2$ is small. We illustrate this in an example next. In practice, one should assess the magnitude of the price spread induced by the instrument and confirm it is economically large, not merely non-zero.

C.2 An example with a weak first stage

Setting. We consider a variation from the model of Section B.5, where the probability of the different states is asymmetric. Furthermore, we measure elasticities in the "wrong" units, portfolio share on price rather than portfolio share on log price. These two features create deviations from Assumption A2. The experiment is a shock to the endowment of the green asset E_g.

The setting is the same as before except for a change in the probability of the states:

$$\text{Green } P_g \begin{cases} \nearrow 1+\epsilon \\ \rightarrow 1-\epsilon \\ \searrow 0 \end{cases} \qquad \text{Red } P_r \begin{cases} \nearrow 1-\epsilon \\ \rightarrow 1+\epsilon \\ \searrow 0 \end{cases} \qquad \text{Other } P_o = 1 \begin{cases} \nearrow 0 & \text{w.p. } \rho/2 \\ \rightarrow 0 & \text{w.p. } (1-\rho)/2 \\ \searrow 1 & \text{w.p. } 1/2 \end{cases}$$

The optimal portfolio shares are:

$$\omega_g\left(P_g, P_r\right) = \frac{P_g\left(\left(\epsilon^2-1\right)P_g + P_r\left(4\rho\epsilon + (\epsilon-1)^2\right)\right)}{4\left(\epsilon^2+1\right)P_gP_r + 2\left(\epsilon^2-1\right)P_g^2 + 2\left(\epsilon^2-1\right)P_r^2} \tag{A130}$$

$$\omega_r\left(P_g, P_r\right) = \frac{P_r\left(P_g\left((\epsilon+1)^2 - 4\rho\epsilon\right) + \left(\epsilon^2-1\right)P_r\right)}{4\left(\epsilon^2+1\right)P_gP_r + 2\left(\epsilon^2-1\right)P_g^2 + 2\left(\epsilon^2-1\right)P_r^2} \tag{A131}$$

Equilibrium and elasticities. Assume that the endowments are $E_g = E_r = 1/2$ and $E_o = 1$. Then equilibrium prices are

$$P_g = 1 - \epsilon(1-2\rho), \qquad P_r = 1 + \epsilon(1-2\rho). \tag{A132}$$

At this equilibrium, the elasticity matrix of the portfolio shares with respect to the level

of prices for the green and red asset is:

$$\mathcal{E}_{gg} = \frac{\partial \omega_g}{\partial P_g} = \frac{\left(\epsilon^2 - 1\right)\left((2\rho - 1)\epsilon - 1\right)}{32(\rho - 1)\rho\epsilon^2}, \tag{A133}$$

$$\mathcal{E}_{rr} = \frac{\partial \omega_r}{\partial P_r} = -\frac{\left(\epsilon^2 - 1\right)\left((2\rho - 1)\epsilon + 1\right)}{32(\rho - 1)\rho\epsilon^2}, \tag{A134}$$

$$\mathcal{E}_{gr} = \frac{\partial \omega_g}{\partial P_r} = \frac{\left(\epsilon^2 - 1\right)\left((2\rho - 1)\epsilon + 1\right)}{32(\rho - 1)\rho\epsilon^2}, \tag{A135}$$

$$\mathcal{E}_{rg} = \frac{\partial \omega_r}{\partial P_g} = -\frac{\left(\epsilon^2 - 1\right)\left((2\rho - 1)\epsilon - 1\right)}{32(\rho - 1)\rho\epsilon^2}. \tag{A136}$$

This leads to the two relative elasticities:

$$\mathcal{E}_{rel,g} = \mathcal{E}_{gg} - \mathcal{E}_{rg} = \frac{\left(\epsilon^2 - 1\right)\left((2\rho - 1)\epsilon - 1\right)}{16(\rho - 1)\rho\epsilon^2}, \tag{A137}$$

$$\mathcal{E}_{rel,r} = \mathcal{E}_{rr} - \mathcal{E}_{gr} = \frac{\left(\epsilon^2 - 1\right)\left(-(2\rho - 1)\epsilon - 1\right)}{16(\rho - 1)\rho\epsilon^2} \tag{A138}$$

Then the terms from the difference-in-difference estimator are:

$$\frac{\mathcal{E}_{rel,g} + \mathcal{E}_{rel,r}}{2} = \frac{1 - \epsilon^2}{16(\rho - 1)\rho\epsilon^2} \tag{A139}$$

$$\frac{\mathcal{E}_{rel,g} - \mathcal{E}_{rel,r}}{2} = \frac{(\epsilon^2 - 1)(2\rho - 1)}{16(\rho - 1)\rho\epsilon} \tag{A140}$$

Note that constant relative elasticity, assumption A2, holds only if $\rho = \frac{1}{2}$. We work in a neighborhood of assumption A2, where $\rho \sim \frac{1}{2}$. To weaken the first stage and make the assets perfect substitutes, we take ϵ to zero. These expressions become approximately

$$\frac{\mathcal{E}_{rel,g} + \mathcal{E}_{rel,r}}{2} \approx \frac{-1}{4\epsilon^2} \tag{A141}$$

$$\frac{\mathcal{E}_{rel,g} - \mathcal{E}_{rel,r}}{2} \approx \frac{2\rho - 1}{4\epsilon} \tag{A142}$$

Equilibrium prices as a function of the endowments are

$$P_g\left(E_o, E_g, E_r\right) = \frac{E_o\left(\left(\epsilon^2 - 1\right)E_g - E_r\left(4\rho\epsilon + (\epsilon - 1)^2\right)\right)}{-2\left(\epsilon^2 + 1\right)E_g E_r + \left(\epsilon^2 - 1\right)E_g^2 + \left(\epsilon^2 - 1\right)E_r^2} \tag{A143}$$

$$P_r\left(E_o, E_g, E_r\right) = \frac{E_o\left(\left(\epsilon^2 - 1\right)E_r - E_g\left((\epsilon + 1)^2 - 4\rho\epsilon\right)\right)}{-2\left(\epsilon^2 + 1\right)E_g E_r + \left(\epsilon^2 - 1\right)E_g^2 + \left(\epsilon^2 - 1\right)E_r^2}. \tag{A144}$$

Around the initial equilibrium, the changes in prices are:

$$\Delta P_g = \frac{\partial P_g}{\partial E_g} = -4\epsilon\rho - (\epsilon - 1)^2 \tag{A145}$$

$$\Delta P_r = \frac{\partial P_r}{\partial E_g} = \epsilon^2 - 1 \tag{A146}$$

The term controlling the bias is:

$$\frac{\Delta P_g + \Delta P_r}{\Delta P_g - \Delta P_r} = \frac{2\epsilon\rho - \epsilon + 1}{\epsilon(2\rho + \epsilon - 1)} \tag{A147}$$

Putting it all together. We will study what happens around $\rho = 1/2$, so, echoing our general setup, we call $x = 2\rho - 1$. When $x = 0$, assumption A2 is satisfied, and the estimator is unbiased.

We plug all the expressions above in equation (A129):

$$\frac{\Delta D_1 - \Delta D_2}{\Delta P_1 - \Delta P_2} = \frac{\mathcal{E}_{rel,1} + \mathcal{E}_{rel,2}}{2} + \frac{\mathcal{E}_{rel,1} - \mathcal{E}_{rel,2}}{2} \frac{\Delta P_1 + \Delta P_2}{\Delta P_1 - \Delta P_2} \tag{A148}$$

$$\approx \frac{-1}{4\epsilon^2} + \frac{x}{4\epsilon} \frac{1}{\epsilon(x + \epsilon)} \tag{A149}$$

The average relative elasticity goes to infinity faster than the difference in relative elasticity (at $1/\epsilon^2$ vs. $1/\epsilon$). However, the weak first stage amplifies the bias by another order of magnitude. To visualize this issue, it is more natural to compute the relative bias of the estimator:

$$\frac{\frac{\Delta D_1 - \Delta D_2}{\Delta P_1 - \Delta P_2} - \frac{\mathcal{E}_{rel,1} + \mathcal{E}_{rel,2}}{2}}{\frac{\mathcal{E}_{rel,1} + \mathcal{E}_{rel,2}}{2}} \approx -\frac{x}{x + \epsilon} \tag{A150}$$

The bias term present when $x \neq 0$ is of equal magnitude as the correct estimate in the limit of a weak instrument.

D A non-linear framework

We derive properties for a family of non-linear demand functions which satisfy locally our assumption of homogeneous substitution conditional on observables. Doing so provides more general intuition behind our results in linear structures.

Because the non-linear structural models considered in Koijen and Yogo (2019) also belong to this family of demand functions, this framework also allows us to better understand the connection of our results with properties of those models. In particular, we explain the restrictions imposed by the logit form relative to arbitrary factor models, simple factor models (with constant variance and expected payoffs), and more general demand functions.

D.1 Basic concepts

We consider a setting with an investor, N assets indexed by i, and K observables for each asset. We start with a general demand function defined as a mapping from the vector of (log) prices $\boldsymbol{p}$ and the $N \times K$ matrix of observables $\boldsymbol{x}$ to a vector of positions D (portfolio shares in our applications):

$$D(\boldsymbol{p}, \boldsymbol{x}) : \mathbb{R}^N \times \mathbb{R}^{N \times K} \to \mathbb{R}^N$$

It will be helpful to define the following property.

Definition A1 (HCO functions) *A function $F : \mathbb{R}^N \times \mathbb{R}^{N \times K} \to \mathbb{R}^N$ is homogeneous-conditional-on-observables (HCO) if $\forall i, [F(\boldsymbol{p}, \boldsymbol{x})]_i = f(p_i, x_i; \boldsymbol{p}, \boldsymbol{x})$ for a function $f : (\mathbb{R} \times \mathbb{R}^K) \times (\mathbb{R}^N \times \mathbb{R}^{N \times K}) \to \mathbb{R}$ for each i.*

That is, for a fixed overall price and observables vector, the value for each element is given by the same (scalar-valued) function of its own price and observables.

Then, in the spirit of the discussion of properties of factor models of Koijen and Yogo (2019), we can restrict attention to a subset of general demand functions as follows.

Definition A2 (HCO demand) *A demand function is a homogeneous-conditional-on-observables demand if it is an HCO function.*

With HCO demand functions, individual positions can be written as:

$$[D(\boldsymbol{p}, \boldsymbol{x})]_i = d(p_i, x_i; \boldsymbol{p}, \boldsymbol{x}).$$

This notation emphasizes the dual role of prices and observables. On the one hand, the same function $d(\cdot, \cdot; \boldsymbol{p}, \boldsymbol{x})$ describes how the demand of each asset depends on its own price and own observables only. On the other hand, this mapping varies with the vector $(\boldsymbol{p}, \boldsymbol{x})$. Thinking of this vector as the state of the economy, an HCO demand describes a mapping which is possibly state-dependent, but identical across assets.

Naturally, the choice of observables is what gives meaningful restrictions to this definition. For example, if the observables $\boldsymbol{x}$ includes each asset's "name", i, then all demand functions are also HCO demand.

An example of HCO demand is logit:

$$[D^{logit}(\boldsymbol{p}, \boldsymbol{x})]_i = \frac{\exp(-\alpha p_i + \beta' x_i)}{1 + \sum_{j=1}^{N} \exp(-\alpha p_j + \beta' x_j)},$$

because the numerator is a function of p_i and x_i only, while the denominator is a fixed function (i.e. that does not depend on i) of $\boldsymbol{p}$ and $\boldsymbol{x}$.

D.2 Relative elasticity vs. substitution and identification

HCO demand leads to a natural decomposition of the elasticity matrix between a relative elasticity and a substitution matrix. HCO demand implies an elasticity matrix:

$$\boldsymbol{\mathcal{E}} = \frac{\partial D}{\partial p} = \underbrace{diag\left(\frac{\partial d}{\partial p_i}\left(p_i, x_i; \boldsymbol{p}, \boldsymbol{x}\right)\right)}_{\text{relative elasticity, } N\times N} + \underbrace{\begin{bmatrix} \frac{\partial d}{\partial p}\left(p_1, x_1; \boldsymbol{p}, \boldsymbol{x}\right)' \\ \vdots \\ \underbrace{\frac{\partial d}{\partial p}\left(p_N, x_N; \boldsymbol{p}, \boldsymbol{x}\right)'}_{1\times N} \end{bmatrix}}_{\text{substitution, } N\times N}$$

where the derivative in the second term is with respect to the third argument of d, not a total derivative. We call the first term relative elasticity and the second one substitution. To understand why the first one is a relative elasticity, notice that if two assets have the same price and observables, this term is equal to the difference between their own-price and cross-price elasticity:

$$\begin{aligned} \mathcal{E}_{ii} - \mathcal{E}_{ji} &= \underbrace{\left(\frac{\partial d}{\partial p_i}\left(p_i, x_i; \boldsymbol{p}, \boldsymbol{x}\right) + \left[\frac{\partial d}{\partial p}\left(p_i, x_i; \boldsymbol{p}, \boldsymbol{x}\right)\right]_i\right)}_{\mathcal{E}_{ii}} - \underbrace{\left[\frac{\partial d}{\partial p}\left(p_j, x_j; \boldsymbol{p}, \boldsymbol{x}\right)\right]_i}_{\mathcal{E}_{ji}} \\ &= \frac{\partial d}{\partial p_i}\left(p_i, x_i; \boldsymbol{p}, \boldsymbol{x}\right) \text{ if } (p_i, x_i) = (p_j, x_j) \end{aligned}$$

The substitution matrix captures how the investor reallocates between assets when their prices change. Any HCO demand satisfies homogeneous substitution conditional on all observables and the price:

$$\mathcal{E}_{il} = \mathcal{E}_{jl} = \left[\frac{\partial d}{\partial p}(p_i, x_i, \boldsymbol{p}, \boldsymbol{x})\right]_l \text{ if } (p_i, x_i) = (p_j, x_j)$$

Indeed, this corresponds to assumption A1 in the text when the observables $\boldsymbol{x}$ are variables that the econometrician can measure.

Identification. The cross-section permits identification of relative elasticity by comparing demand for two assets with the same observables but nearby prices. However, because $(\boldsymbol{p}, \boldsymbol{x})$ are fixed in a given cross-section, identification of substitution is generally impossible with the cross-section.

This limitation of the cross-section can be overcome by imposing additional restrictions. For example, when there is no substitution, that is the demand function does not depend on the price vector per se (its third argument), we have: $d\left(p_i, x_i, \boldsymbol{p}, \boldsymbol{x}\right) = d(p_i, x_i; \boldsymbol{x})$. Elasticity is relative elasticity, and hence can be estimated from the cross-section alone.

Another case is logit. Even though this model has non-zero substitution, it can be estimated from the cross-section because parameters determining substitution can be identified by measuring relative elasticity. Specifically the relative elasticity vector is $-\alpha\omega$ and the

substitution matrix is $\alpha\omega\omega'$ where $\omega = D(\boldsymbol{p}, \boldsymbol{x})$ is the realized vector of portfolio weights. This calculation also highlights that the structure of substitution is very restricted in logit: the substitution matrix is of rank 1, and the effects must be proportional to portfolio weights. To better understand how limiting these restrictions are, we compare logit to other demand models in the following section.

D.3 Logit, log utility and factor models

Our main interest in this section is to what extent the demand of an investor with a standard utility function and particular views on the dynamics of expected returns might be represented with logit demand.

For this purpose, we consider the demand of a log investor with log-normal returns (like in Section B.3) as the simplest example of a standard utility.

$$D^{log}(\boldsymbol{p}, \boldsymbol{x}) = \Sigma(\boldsymbol{p}, \boldsymbol{x})^{-1}\mu(\boldsymbol{p}, \boldsymbol{x})$$

where $\mu(\boldsymbol{p}, \boldsymbol{x})$ is the expected return vector and $\Sigma(\boldsymbol{p}, \boldsymbol{x})$ is the return covariance matrix.

For the representation of views of the investor on the structure of asset returns, we define first a class of factor models where factor loadings might be state-dependent.

Definition A3 (General factor models) *A general M-factor model is defined by functions $\mu(\boldsymbol{p}, \boldsymbol{x}) : \mathbb{R}^N \times \mathbb{R}^{N\times K} \to \mathbb{R}^N$ and $\Sigma(\boldsymbol{p}, \boldsymbol{x}) : \mathbb{R}^N \times \mathbb{R}^{N\times K} \to \mathbb{R}^{N\times N}$ describing the expected return vector and covariance matrix respectively such that*

$$\Sigma(\boldsymbol{p}, \boldsymbol{x}) = \operatorname{diag}(\underbrace{\sigma_\epsilon^2(\boldsymbol{p}, \boldsymbol{x})}_{N\times 1}) + \underbrace{\beta(\boldsymbol{p}, \boldsymbol{x})}_{N\times M}\underbrace{\Sigma_F}_{M\times M}\beta(\boldsymbol{p}, \boldsymbol{x})'$$

, where $\mu(\boldsymbol{p}, \boldsymbol{x})$, $\sigma_\epsilon^2(\boldsymbol{p}, \boldsymbol{x})$ and each column of $\beta(\boldsymbol{p}, \boldsymbol{x})$ are HCO functions and Σ_F is a covariance matrix.

A general M-factor model lets the functions mapping prices and observables to expected payoffs, to factor loadings and to idiosyncratic risk for each asset to freely vary with the state of the economy $(\boldsymbol{p}, \boldsymbol{x})$. As we show below, this is a rich enough set that for any HCO demand $D^{HCO}(\boldsymbol{p}, \boldsymbol{x})$ one can always find a particular factor model that the log investor's demand exactly corresponds to the chosen HCO demand: $D^{HCO}(\boldsymbol{p}, \boldsymbol{x}) = D^{log}(\boldsymbol{p}, \boldsymbol{x})$. This is a generalized version of Corollary 1 of Koijen and Yogo (2019), who specialize the function $D^{HCO}(\boldsymbol{p}, \boldsymbol{x})$ to be the logit demand.

Proposition A3 *Fix two functions $\sigma_\epsilon^2(\boldsymbol{p}, \boldsymbol{x})$ and an $N \times 1$ $\beta(\boldsymbol{p}, \boldsymbol{x})$ which are HCO. For any HCO demand $D^{HCO}(\boldsymbol{p}, \boldsymbol{x})$, there exists a general 1-factor model with the corresponding covariance matrix $\Sigma(\boldsymbol{p}, \boldsymbol{x}) = \operatorname{diag}(\sigma_\epsilon^2(\boldsymbol{p}, \boldsymbol{x})) + \sigma_F^2\underbrace{\beta(\boldsymbol{p}, \boldsymbol{x})}_{N\times 1}\beta(\boldsymbol{p}, \boldsymbol{x})'$, such that log utility demand with this factor model yields the same demand function.*

Proof. Choose $\mu(\boldsymbol{p},\boldsymbol{x}) = \Sigma(\boldsymbol{p},\boldsymbol{x})D^{HCO}(\boldsymbol{p},\boldsymbol{x})$ then clearly $D^{log}(\boldsymbol{p},\boldsymbol{x}) = D^{HCO}(\boldsymbol{p},\boldsymbol{x})$. Therefore, we have to show only that $\mu(\boldsymbol{p},\boldsymbol{x})$ is HCO. For this, note that

$$\left[\Sigma(\boldsymbol{p},\boldsymbol{x})D^{HCO}(\boldsymbol{p},\boldsymbol{x})\right]_i = \left[\sigma^2_\epsilon(\boldsymbol{p},\boldsymbol{x})\right]_i d^{HCO}(p_i,\boldsymbol{x}_i;\boldsymbol{p},\boldsymbol{x}) + \left[\beta(\boldsymbol{p},\boldsymbol{x})\underbrace{\beta'(\boldsymbol{p},\boldsymbol{x})D^{HCO}(\boldsymbol{p},\boldsymbol{x})}_{scalar}\right]_i$$

which, given that $\beta(\boldsymbol{p},\boldsymbol{x})$ is HCO gives the proof. ■

This result shows that for any HCO demand, and a fortiori for logit demand, there exist factor models that microfound it. However, the reverse is clearly not true: an arbitrary factor model does not give rise to logit demand. To move closer to common finance intuition, we consider a restricted class of models often used to think about portfolio choice with stable variance and expected payoffs.

Definition A4 (Stable factor model) *A stable M-factor model (based on observables) is a factor model where expected payoff, idiosyncratic risk, and factor loadings depend on observables only:* $\mu(\boldsymbol{p},\boldsymbol{x}) = M(\boldsymbol{x}) - \boldsymbol{p}$, $\sigma^2_\epsilon(\boldsymbol{p},\boldsymbol{x}) = \sigma^2_\epsilon(\boldsymbol{x})$, *and* $\beta(\boldsymbol{p},\boldsymbol{x}) = \beta(\boldsymbol{x})$ *where* $M(\boldsymbol{x})$ *and* $\sigma^2_\epsilon(\boldsymbol{x})$ *and the columns of* $\beta(\boldsymbol{x})$ *are all HCO.*

One might wonder if there are stable factor models that the demand of a log investor can be represented with the logit form. This cannot hold overall due to different functional forms: the stable factor model is linear in log prices. To make the comparison more meaningful we ask if such an equivalence holds locally. To do so, we define first-order equivalence.

Definition A5 *Two demand functions* $D^1(\boldsymbol{p},\boldsymbol{x})$ *and* $D^2(\boldsymbol{p},\boldsymbol{x})$ *are first-order equivalent around a point* $(\boldsymbol{p}_0,\boldsymbol{x}_0)$ *if they have the same value and Jacobian with respect to the price matrix at that point:*

$$\begin{aligned} D^1(\boldsymbol{p}_0,\boldsymbol{x}_0) &= D^2(\boldsymbol{p}_0,\boldsymbol{x}_0) \\ \underbrace{\frac{\partial D^1}{\partial p}(\boldsymbol{p}_0,\boldsymbol{x}_0)}_{N\times N} &= \frac{\partial D^2}{\partial p}(\boldsymbol{p}_0,\boldsymbol{x}_0) \end{aligned}$$

We obtain that in general the answer is negative.

Proposition A4 *Out of the set of all stable M-factor models, there is only one under which a logit demand model with* $\alpha > 0$ *can be first-order equivalent to the demand of the log investor. This specific model has 1 factor with identical factor loadings and idiosyncratic variance inversely proportional to demand. Under any other stable factor model, logit demand is not a valid approximation of the demand of a log investor.*

Proof. For a stable factor model, we have $D^{log}(\boldsymbol{p},\boldsymbol{x}) = \Sigma(x)^{-1}(M(x) - p)$, so $\partial D^{log}/\partial p = -\Sigma(x)^{-1}$. For logit we have: $\partial D^{logit}/\partial p = -\alpha\,\mathrm{diag}(\omega)(I - \mathbf{1}\omega') = -\alpha\,\mathrm{diag}(\omega) + \alpha\omega\omega'$, where

ω is the investor's portfolio share vector. We can invert it with the Sherman-Morrison formula and identify with Σ:

$$\begin{aligned}
-\left(\partial D^{logit}/\partial p\right)^{-1} &= \alpha^{-1}\left(I - \mathbf{1}\omega'\right)^{-1}\operatorname{diag}(\omega)^{-1} \\
&= \alpha^{-1}\left(I + \frac{1}{1-\omega'\mathbf{1}}\mathbf{1}\omega'\right)\operatorname{diag}(\omega)^{-1} \\
\Sigma &= \underbrace{\alpha^{-1}\operatorname{diag}(\omega)^{-1}}_{\text{idiosyncratic risk}} + \underbrace{\alpha^{-1}\frac{1}{1-\omega'\mathbf{1}}\mathbf{1}\mathbf{1}'}_{\text{single factor}}
\end{aligned}$$

Comparing this to the covariance matrix under a generic stable factor model,

$$\Sigma(\boldsymbol{p},\boldsymbol{x}) = \operatorname{diag}(\underbrace{\sigma_\epsilon^2(\boldsymbol{x})}_{N\times 1}) + \underbrace{\beta(\boldsymbol{x})}_{N\times M}\underbrace{\Sigma_F}_{M\times M}\beta(\boldsymbol{x})',$$

concludes the statement. ■

The proof also illustrates immediately that logit can never be the approximation of a stable multi-factor model that cannot be reduced to a single factor. In such a model, the substitution matrix is of rank equal to the number of factors. Intuitively investors substitute along portfolios corresponding to the various risk factors, differently for assets with different loading on those factors. More broadly, HCO demands include models where, following a price increase for a given position, the investor would substitute disproportionately with assets with similar observables.

A two-asset example. We illustrate that this limitation arises even in the simplest possible 2×2 example with the same variance. Fix the vectors $\boldsymbol{p}$ and $d(p_i;\boldsymbol{p}) = \omega_i$ that we are looking for first-order equivalence around. Such a position could come from a factor model with covariance matrix for any correlation ρ and variance σ^2:

$$\Sigma = \sigma^2\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

If we have a valid approximating logit model, it would feature:

$$-\left(\partial D^{logit}/\partial p\right)^{-1} = \alpha^{-1}\begin{pmatrix} \omega_1^{-1} + \frac{1}{1-\omega_1-\omega_2} & \frac{1}{1-\omega_1-\omega_2} \\ \frac{1}{1-\omega_1-\omega_2} & \omega_2^{-1} + \frac{1}{1-\omega_1-\omega_2} \end{pmatrix}$$

Clearly the two matrices can never be identical if $\omega_1 \neq \omega_2$ because the diagonal terms must be equal. Even if we assume that our point of approximation has a given value $\omega_1 = \omega_2 = \bar{\omega}$, the models are identical only if both on-diagonal and off-diagonal elements match:

$$\begin{aligned}
\sigma^2 &= \alpha^{-1}\left(\bar{\omega}^{-1} + \frac{1}{1-2\bar{\omega}}\right) \\
\sigma^2\rho &= \alpha^{-1}\frac{1}{1-2\bar{\omega}}
\end{aligned}$$

We can already see the issue: there is only one degree of freedom in logit (α) but two degrees of freedom for the covariance matrix (σ^2 and ρ). Let us construct the corresponding contradiction. Subtracting the second equation from the first one gives:

$$\sigma^2 (1-\rho) \bar{\omega} = \alpha^{-1}$$

Plugging this expression for α^{-1} in the second equation leads to:

$$\sigma^2 \rho = \sigma^2 (1-\rho) \frac{\bar{\omega}}{1-2\bar{\omega}}$$
$$\frac{\rho}{1-\rho} = \frac{\bar{\omega}}{1-2\bar{\omega}}$$

The right-hand-side is fixed, this is our point of approximation. The left-hand-side could take any value as ρ is a free parameter.

E Simulation Exercise

This appendix describes implementation details and parameter choices of the static CRRA demand model under a factor structure introduced in Section 1.3. We conduct three counterfactual experiments varying design and parameter choices. In all those cases, existing demand models do not lead to accurate predictions; in all cases, our framework does.

E.1 Setting

Consider N risky assets indexed by iLog asset payoffs are normally distributed with mean vector $\bar{\Pi}$ and covariance matrix Σ, and follow a one-factor structure:

$$\log \Pi_i = \bar{\Pi}_i + \beta_i z + v_i \tag{A151}$$
$$\Sigma = \sigma_v^2 I + \sigma_F^2 \beta\beta', \tag{A152}$$

where $z \sim \mathcal{N}(0, \sigma_F^2)$ is a common payoff shock, $v_i \sim \mathcal{N}(0, \sigma_v^2)$ is idiosyncratic, and all terms are independent. Let P denote the equilibrium price vector. A risk-free asset is in perfectly elastic supply at log rate $r_f = 0$.

There are two investors. A CRRA investor with risk aversion γ is endowed with E units of the risky assets and E_{rf} units of the risk-free asset. An investor with deep pockets (in risk-free assets) purchases A units of risky assets, parameterizing the supply variation in the experiment of Section 1.3. The residual supply for the CRRA investor is $S = E - A$. Both investors are price takers, and equilibrium clears the market for risky assets.

CRRA demand. Let W denote the wealth of the CRRA investor, D their vector of risky asset positions, and $\omega_i = D_i P_i / W$ the portfolio weight in asset i. The investor maximizes expected CRRA utility over terminal wealth with log-normally distributed payoffs. Following the log-linearization of Campbell and Viceira (2002), we use the resulting optimal static

demand as the structural demand equation in the data-generating process:

$$\boldsymbol{\omega}^{\text{CRRA}} = \frac{1}{\gamma}\,\Sigma^{-1}\big(\bar{\Pi} - \log P - r_f + \tfrac{1}{2}\,\text{diag}(\Sigma)\big), \tag{A153}$$

Initial wealth is $W = E_{rf} + E'P$.

Equilibrium. Equilibrium requires optimal demand of the CRRA investor to equal residual supply: $D = S$, i.e. $\omega_i = S_i P_i / W$. Substituting into (A153) and rearranging:

$$\log P = \bar{\Pi} - r_f + \tfrac{1}{2}\,\text{diag}(\Sigma) - \gamma\,\Sigma\,\omega, \qquad \omega = \frac{\text{diag}(S)P}{W}, \qquad W = E_{rf} + E'P. \tag{A154}$$

As an aside, at any equilibrium price vector P, the expected-return term in equation (A153) can be written in the factor-premium representation used in equation (5).[66]

The equilibrium condition defines a fixed-point map $g(\boldsymbol{\omega}) = \text{diag}(S)P(\boldsymbol{\omega})/W(\boldsymbol{\omega})$, where $P(\boldsymbol{\omega}) = \exp\big(\bar{\Pi} - r_f + \frac{1}{2}\,\text{diag}(\Sigma) - \gamma\,\Sigma\,\boldsymbol{\omega}\big)$ and $W(\boldsymbol{\omega}) = E_{rf} + E'P(\boldsymbol{\omega})$. We numerically solve for fixed points $\boldsymbol{\omega} = g(\boldsymbol{\omega})$.

Parameter choices. All parameter choices are summarized in Table A2. We choose $N = 200$ to create a rich cross-section while maintaining parsimony. This number corresponds closely to the average effective number of stocks in the U.S. market: computing the inverse Herfindahl–Hirschman index (HHI) of market-capitalization weights from CRSP each month from 1980 to 2021 and averaging over time gives an effective number of 233. We explore robustness versions at $N = 100, 300$. $E_i = S_{0,i} = 1$ and $E_{rf} = N$ are normalizations for the number of shares in each of the N risky assets and one risk-free asset. The risk-free rate is normalized to 0.

The model is static, so the single period requires an economic interpretation as a horizon. We interpret the risky assets as long-duration equity claims and set the model horizon to $d = 20$ years.[67] This value is in line with Gormsen and Lazarus (2025), who estimate a U.S. market equity duration of about 20 years from the loading of stock returns on discount-rate shocks. We explore robustness to this choice at $d = 10, 30$.

To ensure that baseline log prices are flat across β, we assume that expected log payoffs are related to β. Specifically, we assume that expected log payoffs in annualized inputs are of the form

$$\bar{\Pi}_i = \bar{\Pi}_{ref} + \bar{\Pi}_\beta \beta_i + \bar{\Pi}_{\beta\beta}\beta_i^2, \tag{A155}$$

[66] Let

$$\mu_i \equiv \overline{\Pi}_i - \log P_i - r_f + \frac{1}{2}(\sigma_v^2 + \sigma_F^2 \beta_i^2)$$

denote the Jensen-adjusted expected excess return entering demand. In the notation of equation (5):

$$\mathbb{E}[F] = \frac{\sum_j \beta_j \mu_j}{\sum_j \beta_j^2}, \qquad \alpha_i = \mu_i - \beta_i \mathbb{E}[F],$$

so that $\beta'\alpha = 0$.

[67] If interpreted as a one-year model and calibrated directly to annual returns, a frictionless CRRA economy generates unrealistically small price impacts (e.g., Petajisto, 2009; Davis et al., 2025).

and the simulation scales this vector to the d-year horizon by using $d\bar{\Pi}_i$. A common shift in payoffs mostly moves equilibrium log prices one-for-one, leaving expected returns largely unchanged, and therefore controls the baseline allocation between risky assets and the risk-free asset. By contrast, the β and β^2 terms determine how the cross-section of expected returns is split between expected payoffs and baseline prices. We choose $\bar{\Pi}_\beta = 0.080$ and $\bar{\Pi}_{\beta\beta} = -0.011$ so that baseline log prices are flat across β.[68] This makes the baseline calibration equal-weighted in stocks: variation in expected returns across β goes along with differences in expected payoffs rather than lower prices for high-β assets. In other words, this calibration ensures that risky portfolio weights are equal. We choose the payoff level $\bar{\Pi}_{ref} = -0.032$ to calibrate the risk-free portfolio share. The target is $\omega_{rf} = 0.29$, corresponding to the ratio of U.S. Treasuries (consolidated with Fed holdings) to U.S. Treasuries plus total U.S. stock market capitalization as of early 2026.

We choose β_i to be equally spaced on $[0.2, 1.8]$. Equal spacing is for parsimony, and to show most clearly how results for experiments differ across β_i. The distribution of betas is centered around 1 in line with the factor corresponding to a market factor, such that the average asset naturally has a beta of 1. The endpoints of the uniform spacing are chosen such that the cross-sectional standard deviation across betas matches the time-series average of the cross-sectional standard deviation of betas, based on the beta estimates from Welch (2022) over 1980–2021. The annual volatility of the market factor, $\sigma_F = 0.15$, is chosen to match the time-series volatility of the market excess return between 1980 and 2021 (≈ 0.1551), with data taken from the Kenneth French data library.[69] We choose $\sigma_v = 0.4$ for annual idiosyncratic volatility to match the volatility of CRSP excess returns in 1980–2021 (excluding bottom quintile micro cap stocks).[70] Specifically, average equal-weighted total volatility is 0.44, which translates to an idiosyncratic volatility of 0.41 for the average stock at $\beta_i = 1$.

In our baseline parametrization, we choose a constant relative risk aversion of $\gamma = 5$, toward the middle of commonly accepted parameter choices in asset pricing; together with the calibrated payoff process, this implies an annualized expected return of 8.27% for the $\beta = 1$ stock.[71] We show robustness versions for $\gamma = 2$ and $\gamma = 10$, toward the lower and upper end of common choices.

[68] Using annualized units, the equilibrium condition implies

$$\frac{1}{d}\log P_i = \overline{\Pi}_{ref} + \overline{\Pi}_\beta \beta_i + \overline{\Pi}_{\beta\beta}\beta_i^2 - r_f + \frac{1}{2}(\sigma_v^2 + \sigma_F^2\beta_i^2) - \gamma\sigma_v^2\omega_i - \gamma\sigma_F^2\beta_i B, \tag{A156}$$

where $B = \sum_j \beta_j \omega_j$. With flat baseline prices and $S_{0,i} = 1$, risky weights are equal, $\omega_i = (1 - \omega_{rf})/N$, and $B = (1 - \omega_{rf})\bar{\beta}$. Setting the coefficients on β_i and β_i^2 to zero gives

$$\overline{\Pi}_\beta = \gamma\sigma_F^2 B, \tag{A157}$$

$$\overline{\Pi}_{\beta\beta} = -\frac{1}{2}\sigma_F^2. \tag{A158}$$

[69] Since $r_f = 0$ in the model, the market volatility corresponds directly to the volatility of excess market returns. For comparison, the volatility of total market returns over the same period is 0.1546.

[70] Volatility parameters are annualized inputs; the simulation scales them by $\sqrt{d}$ for the d-year horizon.

[71] We report annualized $\log \mathbb{E}[R_i]$, which equals $\overline{\Pi}_i - \log P_i/d + \frac{1}{2}\sigma_{ii}^2$ under log-normal payoffs.

Table A2: **Parameter choices for the CRRA demand simulation**

Symbol	Name	Baseline	High idiosyncratic risk	Duration risk (Treasuries)
N	Number of assets	200 (100, 300)	200	29
d	Holding horizon	20 (10, 30)	20	20
E_i	Risky endowment/supply	1	1	1
E_{rf}	Risk-free endowment	N	N	1
r_f	Risk-free rate	0	0	0
$\bar{\Pi}_{ref}$	Payoff intercept	-0.032	-0.487	-0.0003
$\bar{\Pi}_{\beta}$	Linear payoff slope	0.080	0.080	0.084
$\bar{\Pi}_{\beta\beta}$	Quadratic payoff slope	-0.011	-0.011	-0.009
β_i	Factor loading	$[0.2, 1.8]$, equally spaced	$[0.2, 1.8]$	$[0.067, 1.933]$
σ_F	Factor volatility	0.15	0.15	0.132
σ_v	Idiosyncratic volatility	0.4	1.05	0.028
γ	Risk aversion	5 (2, 10)	5	5

Table A2 summarizes the parameter choices for the CRRA demand simulation described in Appendix E. Values in parentheses in the Baseline column are robustness alternatives reported alongside the headline calibration; "High idiosyncratic risk" and "Duration risk" are stand-alone calibrations with no parameter perturbations. The equity return process (Baseline and High idiosyncratic risk) is calibrated to U.S. stock market data between 1980 and 2021; the Treasury return process (Duration risk) is calibrated to the Gürkaynak–Sack–Wright nominal zero-coupon yield curve between 1986 and 2019. Further details are provided in the text.

The column "Baseline" of Table A2 summarizes these parameter choices. We consider two alternative calibrations: the first entertains unrealistic levels of idiosyncratic volatility to better match some of the evidence on relative price impact; the second matches properties of Treasury returns instead of stocks.

Specifically, the Baseline calibration leads to a relative multiplier of $\mathcal{M}_{relative} = 0.0568$ and an aggregate multiplier of $\mathcal{M}_{agg} = 1.6543$, a ratio of approximately 29.[72] The aggregate multiplier is of the same order of magnitude as estimates from the literature, but the relative multiplier is far below existing estimates, which Gabaix and Koijen (2021) summarize to be around 1. We therefore entertain a second calibration, "High idiosyncratic risk." This calibration raises annual idiosyncratic volatility to a deliberately high level of $\sigma_v = 1.05$ while keeping N, d, γ and σ_F unchanged. Recalibrating $\bar{\Pi}_{ref}$ preserves the same flat-price baseline and risk-free portfolio share and gives $\mathcal{M}_{relative} = 0.3914$, $\mathcal{M}_{agg} = 1.9889$, a ratio of about 5, closer to the estimates from Gabaix and Koijen (2021).

We additionally calibrate the model to the cross-section of U.S. Treasury zero-coupon bonds, in which β_i has a direct interpretation as duration. This corresponds to the column "Duration risk" in Table A2. Using the Gürkaynak–Sack–Wright nominal yield curve (Gürkaynak et al., 2007) for maturities $n_i \in \{2, \ldots, 30\}$ years, we construct annual log excess

[72] We have $\mathcal{M}_{relative} = \gamma d \sigma_v^2 \frac{1-\omega_{rf}}{N} = 5 \times 20 \times 0.4^2 \times \frac{1-0.29}{200} = 0.0568$ and $\mathcal{M}_{agg} = \mathcal{M}_{relative} + \gamma d \sigma_F^2 \bar{\beta}^2 (1-\omega_{rf}) = 0.0568 + 5 \times 20 \times 0.15^2 \times (1-0.29) = 1.6543$.

returns $rx_{t+1}^{(n)} = ny_t(n) - (n-1)y_{t+1}(n-1) - y_t(1)$ from the first available trading day of each calendar year between 1986 and 2019, giving $T = 33$ observations on $N = 29$ assets. We impose a duration-based loading $\beta_i = (n_i - 1)/15$, equally spaced on $[0.067, 1.933]$ with mean one.[73] A one-factor fit of the covariance matrix gives $\sigma_F = 0.132$ and $\sigma_v = 0.028$; unlike in the equity calibration, factor risk dominates idiosyncratic risk. We work in excess-return units and set $E_{rf} = 1$ with target $\omega_{rf} = 1/30$, so that all 30 portfolio weights (one risk-free plus 29 risky) coincide at the equal-weights equilibrium. The payoff coefficients are $(\bar{\Pi}_{ref}, \bar{\Pi}_{\beta}, \bar{\Pi}_{\beta\beta}) = (-0.0003,\ 0.084,\ -0.009)$.

E.2 Simulation

We next describe how we generate data that we use to fit various demand models, precisely what models we estimate, and how we conduct the counterfactuals.

Step 1: Simulate data from the CRRA model. Fix baseline supply $S_0 = E$ and solve the CRRA equilibrium (A154) to obtain baseline P_0, ω_0, W_0. Then simulate $T = 50$ periods of supply perturbations under $\sigma_S = 0.05$:[74]

$$S_t = \mathrm{diag}(S_0)\exp(\sigma_S \cdot \eta_t - \tfrac{1}{2}\sigma_S^2), \qquad \eta_t \sim \mathcal{N}(0, \mathbf{I}_N), \tag{A159}$$

and solve the equilibrium at each S_t to obtain repeated cross-sections $\{\omega_t, \log P_t\}_{t=1}^{T}$.

For all experiments, we also include a robustness version that treats wealth as fixed (e.g., Koijen and Yogo, 2019) at the initial level of W_0. Specifically, the market-clearing condition becomes $\omega_i = S_i P_i / W_0$ with W_0 held constant across all supply perturbations—both during simulation and when computing counterfactual prices.

Step 2a: Estimate a logit model. The logit model defines portfolio weights as:

$$\omega_i^{\text{Logit}} = \frac{\exp(\alpha \log P_i + \varepsilon_i)}{1 + \sum_j \exp(\alpha \log P_j + \varepsilon_j)}, \tag{A160}$$

with the risk-free share $\omega_{rf} = 1/(1 + \sum_j \exp(\alpha \log P_j + \varepsilon_j))$. The parameters of the model are α (elasticity) and ε_i (asset-specific tastes). The latter is potentially related to β_i.

We estimate the logit model by taking logs on both sides and estimating in levels (Koijen and Yogo, 2019):

$$\log \omega_{i,t} = \alpha \log P_{i,t} + c_i + a_t + b_t\, \beta_i + u_{i,t}, \tag{A161}$$

where α is the elasticity parameter estimated, c_i are asset fixed effects, a_t time fixed effects (absorbing the log outside share), and b_t time-varying slopes on β_i.

[73] This approximates the leading eigenvector of the return covariance to $R^2 = 0.998$, and the first principal component explains 96.8% of return variation.

[74] These parameters are not included in Table A2 as they are not primitives of the model but rather control the precision of estimation.

One may argue that any failure of logit here may just be the result of misspecification in functional form, especially from forcing a log on portfolio shares. We therefore estimate an alternative log-linear specification in differences,[75] which is identical to the relative elasticity estimation in Step 2c, with $\alpha_{lin} = \widehat{\mathcal{E}}$, but without the substitution component of our model. Accordingly, the counterfactual under this specification uses the same linear form as our model, $\omega_0 + \boldsymbol{\mathcal{E}} \Delta \log P$, but with the elasticity matrix set to $\widehat{\mathcal{E}} I_N$.

Similarly, one may argue that logit fails only because it imposes a common relative elasticity across assets, and that allowing the elasticity to vary with factor exposure would suffice to capture the factor structure in demand. We therefore consider a robustness specification that adds an interaction term $\alpha_\beta \log P_{i,t} (\beta_i - \bar{\beta})$ to (A161), on top of the common elasticity parameter α, so that the implied elasticity parameter $\alpha(\beta_i) = \alpha + \alpha_\beta (\beta_i - \bar{\beta})$ used in the counterfactual is free to vary linearly with β_i.

Step 2b: Estimate logit + macro model. As suggested in Gabaix and Koijen (2024), we enrich logit with a macro elasticity. Define the total risky portfolio share $\omega_{\text{TOT},t} \equiv \sum_i \omega_{i,t} = 1 - \omega_{rf,t}$ and the aggregate risky-asset price $\bar{P}_{\text{agg},t} = \frac{1}{N} \sum_i P_{i,t}$. The model, which adds the macro elasticity Θ to logit, becomes:

$$\omega_{i,t}^{\text{Logit+Macro}} = \underbrace{\omega_{\text{TOT},0} \, \exp\big(\Theta \, \Delta \log \bar{P}_{\text{agg},t}\big)}_{\omega_{\text{TOT},t}(\bar{P}_{\text{agg},t})} \times \frac{\exp\big(\alpha \log P_{i,t} + \varepsilon_i\big)}{\sum_j \exp\big(\alpha \log P_{j,t} + \varepsilon_j\big)}. \tag{A162}$$

We estimate

$$\Delta \log \omega_{\text{TOT},t} = \Theta \, \Delta \log \bar{P}_{\text{agg},t} + u_t. \tag{A163}$$

Further, Gabaix and Koijen (2021) emphasize value-weighting. Under our baseline calibration, market capitalization is equal in the cross-section, so that equal and value-weighting are identical.

Step 2c: Estimate our model. We estimate the relative elasticity $\widehat{\mathcal{E}}$ from a cross-sectional regression, controlling for β_i as the observable, interacted with time-fixed effects, as this is a repeated cross-section:

$$\Delta \omega_{i,t} = \widehat{\mathcal{E}} \Delta \log P_{i,t} + \theta_{1,t} \, \beta_i + \theta_{0,t} + e_{i,t}. \tag{A164}$$

Denote by X_i a cross-sectionally demeaned and normalized (to unit norm) β_i. We estimate substitution from the time-series by running $K = 2$ bivariate regressions, estimating four substitution parameters:

$$\begin{pmatrix} \frac{1}{\sqrt{N}} \mathbf{1}' \Delta \omega_t \\ X' \Delta \omega_t \end{pmatrix} = \breve{\boldsymbol{\mathcal{E}}} \begin{pmatrix} \frac{1}{\sqrt{N}} \mathbf{1}' \Delta \log P_t \\ X' \Delta \log P_t \end{pmatrix} + \begin{pmatrix} e_t^{(1)} \\ e_t^{(2)} \end{pmatrix}, \tag{A165}$$

[75] Differences here are defined as perturbations from the baseline, i.e., $\Delta Y_{i,t} \equiv Y_{i,t} - Y_{i,0}$ for any variable Y between simulation t and the baseline 0.

where $\check{\boldsymbol{\mathcal{E}}}$ is 2×2, combining substitution and the relative elasticity as in Proposition 3.

Step 3: Run experiments I to III. Finally, we conduct three experiments based on the estimated models. Experiment I computes counterfactual prices from the CRRA model and inverts each estimated model for the implied supply shocks. Experiments II and III fix the supply shocks and compare the predicted price changes across models.

Experiment I evaluates the effects of a uniform supply shock of 10% to all risky assets, i.e., setting $S_1 = 0.9S_0$. We solve for the CRRA model-implied counterfactual prices P_1 and wealth W_1 based on equilibrium conditions in (A154). The experiment then asks what size of supply shocks are required for each asset under different models—logit, logit + macro, and our model—to justify the counterfactual price P_1 from the CRRA model. Panel A of Figure 1 visualizes the log price change $\Delta \log P$ we require each demand model to match. For logit, we apply counterfactual prices P_1, the estimates $\hat{\alpha}$ and $\hat{\varepsilon}_i$ to Equation (A160), which defines counterfactual portfolio shares. In our model, the counterfactual portfolio share is simply the initial portfolio share plus $\boldsymbol{\mathcal{E}} \Delta \log P$, where $\boldsymbol{\mathcal{E}}$ is the estimated elasticity matrix combining the relative elasticity with substitution, i.e.,

$$\boldsymbol{\mathcal{E}} = \widehat{\mathcal{E}}\mathbf{I} + \left[\frac{1}{\sqrt{N}}\mathbf{1}, X\right](\check{\boldsymbol{\mathcal{E}}} - \widehat{\mathcal{E}}\mathbf{I}_2)\left[\frac{1}{\sqrt{N}}\mathbf{1}, X\right]'. \tag{A166}$$

The model combining logit and a macro elasticity computes counterfactual shares as the product of the predicted total risky asset portfolio share and within-risky asset portfolio shares, $\omega_i^{\text{Logit+Macro}}(P_1) = \omega_{\text{TOT}}^{\text{macro}}(P_1) \cdot \tilde{\omega}_i(P_1)$, where

$$\omega_{\text{TOT}}^{\text{macro}}(P_1) = \omega_{\text{TOT},0} \cdot \exp\big(\hat{\Theta}\,(\log \bar{P}_{\text{agg},1} - \log \bar{P}_{\text{agg},0})\big), \tag{A167}$$

$$\tilde{\omega}_i(P_1) = \frac{\tilde{\omega}_{0,i} \cdot \exp\big(\hat{\alpha}\,(\log P_{1,i} - \log P_{0,i})\big)}{\sum_j \tilde{\omega}_{0,j} \cdot \exp\big(\hat{\alpha}\,(\log P_{1,j} - \log P_{0,j})\big)}, \tag{A168}$$

with $\tilde{\omega}_{0,i}$ the within-risky assets portfolio share of asset i in the baseline equilibrium.

Experiment II evaluates the impact of an asymmetric supply shock with opposing signs for high-versus-low β assets. Specifically, it reduces supply by 20% for assets in the sixth and seventh octiles of β, and increases supply by 20% for assets in the second and third octiles of β, analogous to an "Operation Twist" type of supply shock. Panel A of Figure A2 visualizes the experiment. We then invert all estimated demand functions to obtain the predicted price change of each estimated demand model given the structure of the supply shock, and compare it to the one generated under the CRRA data-generating process. This is the inverse procedure to how Experiment I is calculated, and generally requires numerical inversion. Experiment III is similar, but reduces supply by 20% for assets with below-average β, leaving the remaining assets unchanged, as shown in Panel B of Figure A2.

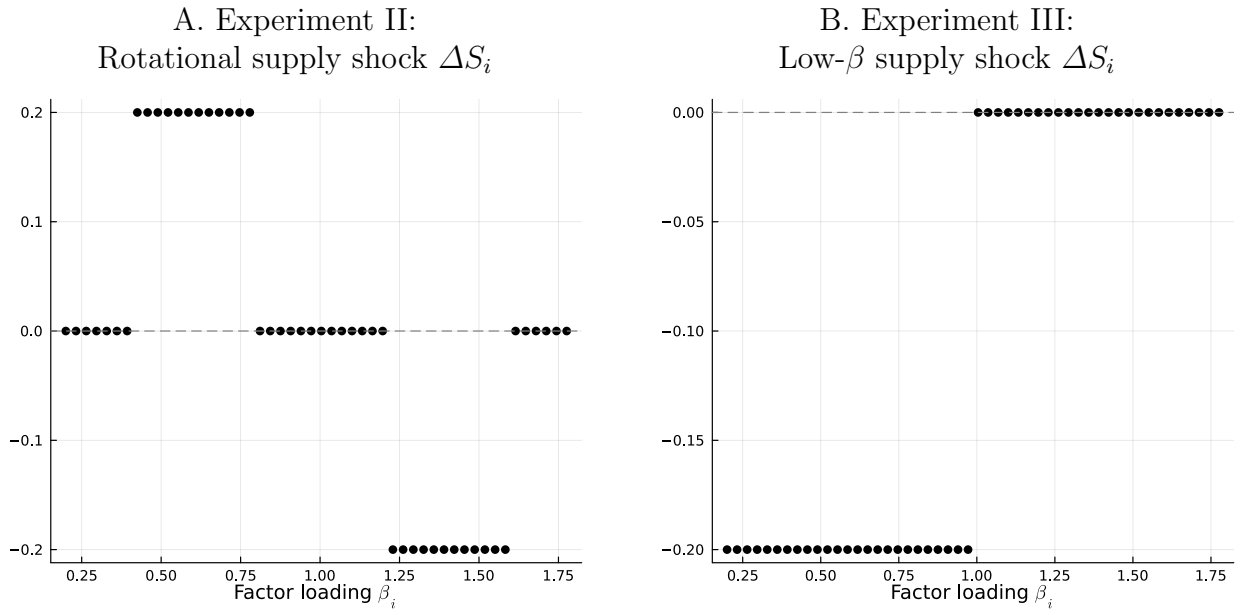


Figure A2: **Supply shocks in Experiments II and III.** Figure A2 shows the supply perturbations ΔS_i as a function of asset beta β_i for experiments II and III. Panel A displays the rotational shock, which reduces supply by 20% for sixth- and seventh-octile β assets and increases supply by 20% for second- and third-octile β assets. Panel B displays the low-β shock, which reduces supply by 20% for assets with below-average β and leaves remaining assets unchanged.

E.3 Experiments

We now discuss the results from the three experiments. For the baseline calibration, Figure 1 in the main text reports the results of experiment I, while Figure A3 reports the results of experiments II and III.

Experiment I: uniform supply shock A uniform supply shock of 10% of supply moves the price of low-β assets less than that of high-β assets, as shown in Panel A of Figure 1. How large a demand shock is needed to create that same movement in prices according to the different estimated demand models? Panel B provides the answer for each model. Our model recovers the 10% shock that is uniform across the β distribution. In contrast, logit and logit + macro attribute the stronger price response of high-β assets to a demand shock that is tilted toward high-β assets.

Panels A and B of Figure A4 report the implied supply changes under the High idiosyncratic risk calibration (which yields a higher relative multiplier) and the Duration risk calibration, respectively. The pattern is similar to our baseline.

Experiment II: rotational supply shock Experiment II is an "Operation Twist" type of supply shock, a rotation that involves reducing supply by 20% for assets in the sixth and seventh octiles of β, and increasing supply by 20% for assets in the second and third octiles of β, as seen in Panel A of Figure A2. Greenwood et al. (2024) discuss this type of experiment in the context of Treasuries and models such as Vayanos and Vila (2021).

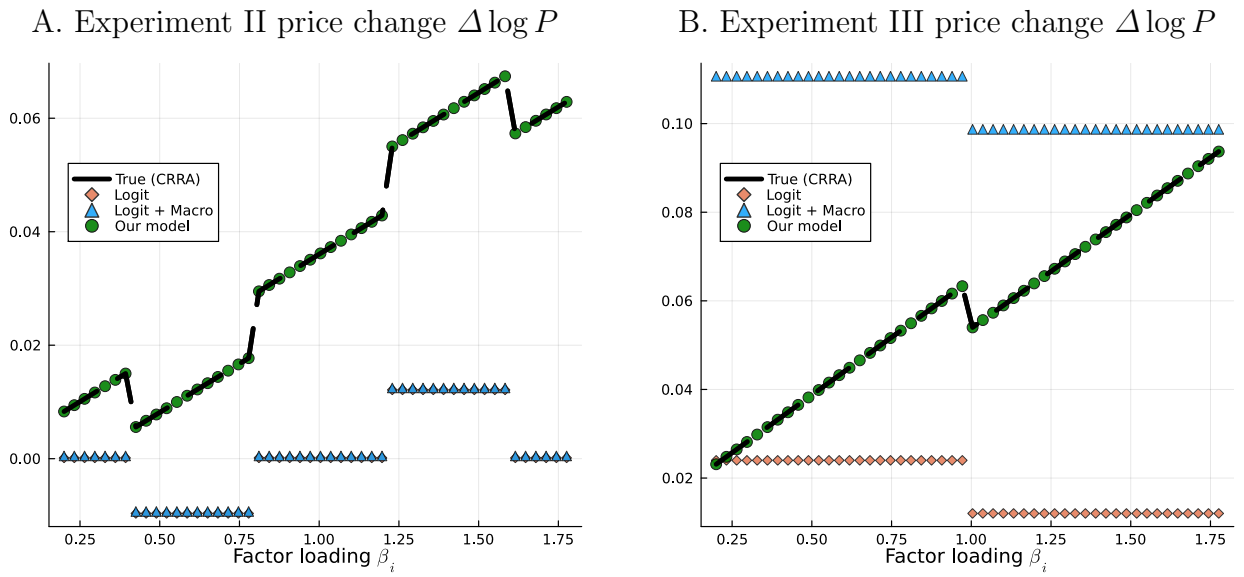


Figure A3: **Price responses in Experiments II and III.** Figure A3 shows the equilibrium log price change $\Delta \log P_i$ relative to baseline prices P_0 as a function of asset beta β_i for experiments II and III. The dashed black line is the true CRRA response; colored markers show predictions from the logit (orange), logit + macro (blue), and our model (green). Panel A displays the response to the rotational shock of experiment II. Panel B displays the response to the low-β shock of experiment III. The supply shocks are detailed in Figure A2.

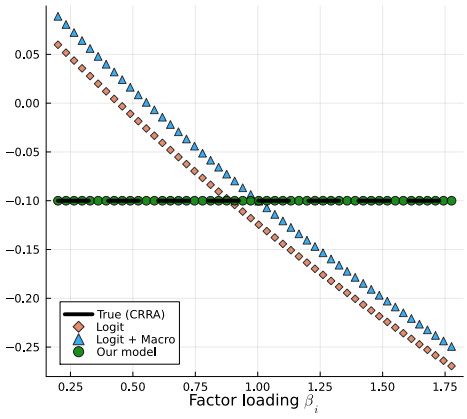
Panel A of Figure A3 shows that, under the data-generating process, the price impact of the rotational supply shock is close to linearly increasing in β, as in Experiment I, but with discrete jumps at the cutoffs induced by the supply rotation. First, the linear component reflects the same mechanism as before: reducing supply of high-β assets while increasing supply of low-β assets lowers the risk premium associated with market risk, generating a price response proportional to β_i. It implies that the assets not directly purchased or sold in the experiment experience price changes, sometimes larger than those of the assets traded in the experiment. Second, the discrete jumps are best understood by comparing assets just above and just below a cutoff. For example, the supply of assets just above the cutoff for the 6th octile is reduced by the shock, while the supply of assets just below the cutoff is not.[76] Their relative price therefore moves according to the inverse relative elasticity, generating the discrete jump. Our model reproduces both components of the response.

In contrast, logit captures the discontinuities at the relevant cutoffs but does not reproduce the linear component associated with the change in the factor risk premium. The same is true for logit + macro. Because the supply shock is approximately long-short, it generates little aggregate supply variation for the macro multiplier to absorb.

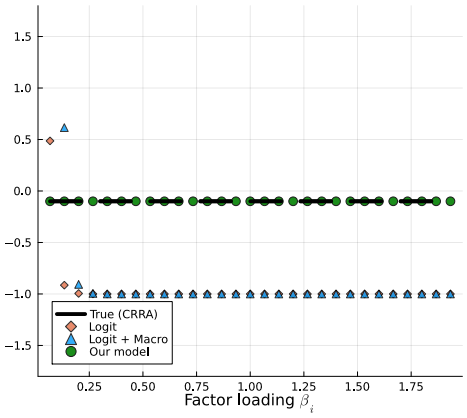
The same pattern also holds, with different quantitative magnitudes, in the High idiosyncratic risk and Duration risk calibrations in Panels C and D of Figure A4. The relative importance of the two components differs across these calibrations, with the former putting more weight on the discontinuities at the treated cutoffs and the latter putting more weight

[76] Such discrete cutoffs are similar in spirit to regression discontinuity designs.

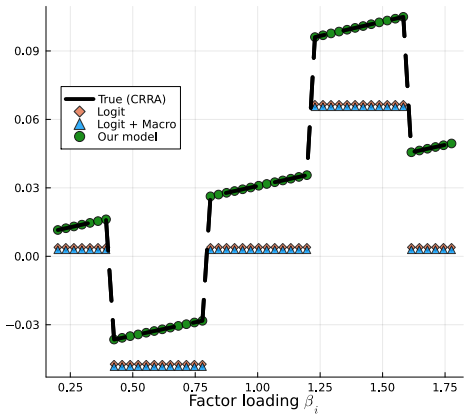
A. Experiment I supply change, High idio. risk



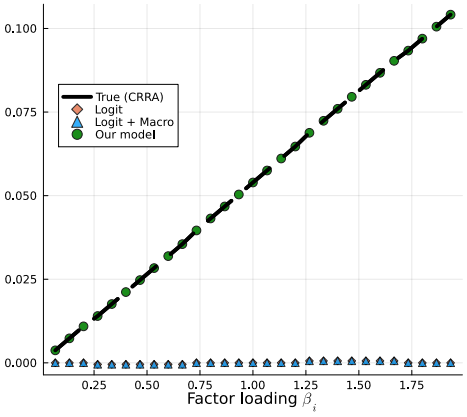
B. Experiment I supply change, Duration risk



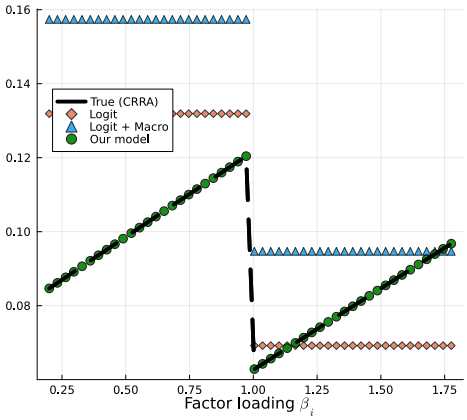
C. Experiment II price change, High idio. risk



D. Experiment II price change, Duration risk



E. Experiment III price change, High idio. risk



F. Experiment III price change, Duration risk

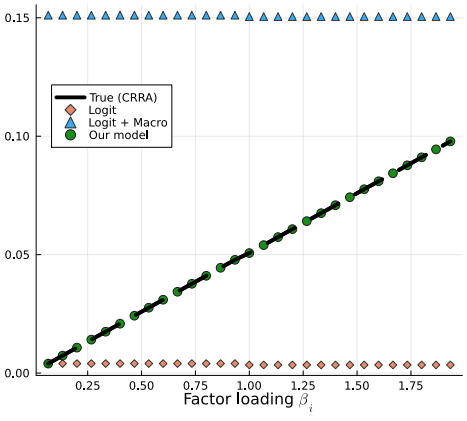


Figure A4: **Alternative calibrations for Experiments I–III.** Panels A–B repeat the implied-supply-change exercise from Experiment I in Figure 1 under two alternative calibrations. Panels C–D repeat the rotational supply shock from Experiment II, and Panels E–F repeat the low-β supply shock from Experiment III, corresponding to Figure A3. Panels A, C, E use the High idiosyncratic risk calibration, chosen to match price-multiplier magnitudes from the literature. Panels B, D, and F use the Duration risk calibration.

on the factor-premium component. This difference reflects the ratio of idiosyncratic to factor risk in the two calibrations.

Experiment III: low-β supply shock The final experiment corresponds to a supply reduction of 20% in assets with a β below 1, as seen in Panel B of Figure A2.

Panel B of Figure A3 shows that the price change under the CRRA data-generating process takes a similar form as in the previous experiment, combining a linear component in β_i with a discontinuity at $\beta_i = 1$. Assets just below the cutoff move more because their supply is reduced, whereas the supply of assets just above the cutoff is unchanged. Our model again reproduces both components of the counterfactual response.

As in the previous experiment, logit captures the discontinuity at $\beta_i = 1$ but does not reproduce the component associated with the change in factor risk premia. Logit + macro also captures the discontinuity, but its aggregate component shifts all prices by more than the largest true price change across assets. The reason is that the macro component treats the supply shock as a large aggregate shock: reducing supply by 20% for half of the assets is close, in aggregate quantity, to a 10% uniform supply shock. This aggregation abstracts from the fact that the supply reduction is concentrated in low-β assets. As a result, logit + macro attributes too much aggregate price impact to the shock.

This qualitative pattern is robust across the two alternative calibrations, as can be seen in Panels E (High idiosyncratic risk) and F (Duration risk) of Figure A4.

Robustness These conclusions are not specific to the chosen parameters or other features of the experiment. Figure A5 reproduces the results of the three experiments for the logit and logit + macro models across multiple variations in parameter choices and specification. Panels A and B report the implied supply changes in Experiment I for these alternatives, for logit and logit + macro, respectively. Panels C–F report deviations from the CRRA-implied price changes for Experiments II and III. We focus on these errors as opposed to the raw price changes, because the latter changes across specifications even for the true model.

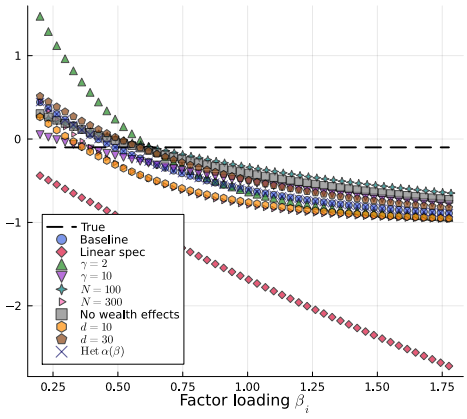
Relative to our baseline, we vary the risk aversion parameter ($\gamma = 2, 10$), the trading horizon ($d = 10, 30$), and the number of assets ($N = 100, 300$). We consider a log-linear specification instead of logit (line "Linear spec"), a specification without wealth effects (line "No wealth effects"), and a specification allowing the logit price coefficient α to depend linearly on β_i (line "Het $\alpha(\beta)$"). The pattern of deviation from correct predictions remains consistent across all these parametrizations.

F Empirical Results

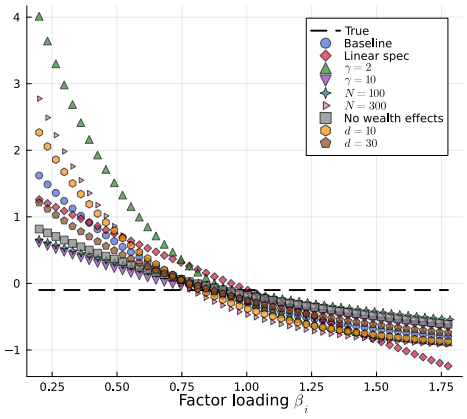
F.1 Falsifying the assumptions

Complementary to the balance-on-covariance diagnostics in Figure 4, we assess whether the estimated relative multiplier $\widehat{\mathcal{M}}$ is stable when the comparison is restricted to bonds that are more similar. Since Assumption A1 requires homogeneity in substitution after conditioning on observables, instability across narrower comparison sets would indicate that the baseline observables may omit relevant dimensions of substitution.

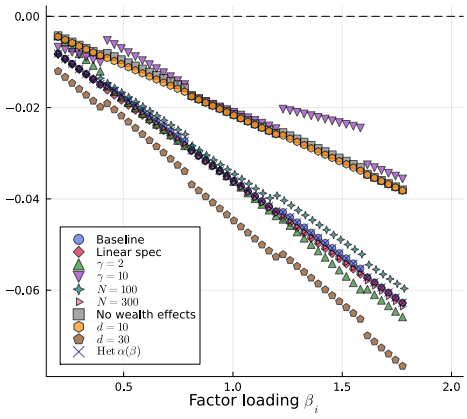
A. Experiment I: implied supply change ΔS_i for logit



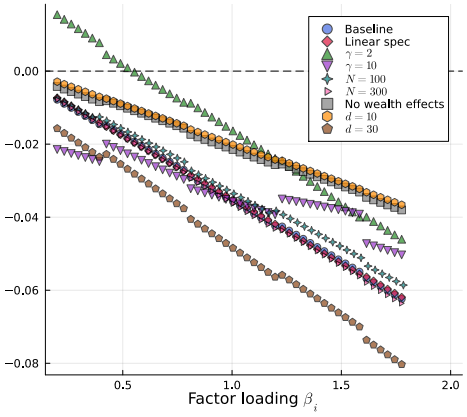
B. Experiment I: implied supply change ΔS_i for logit + macro



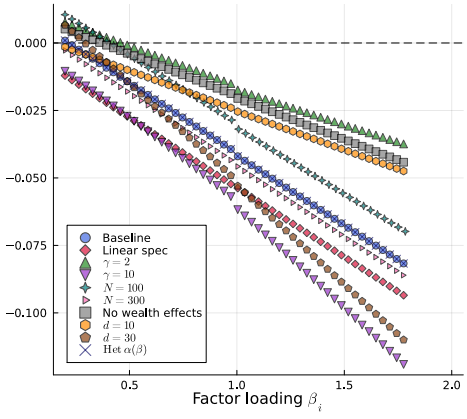
C. Experiment II: price error for logit



D. Experiment II: price error for logit + macro



E. Experiment III: price error for logit



F. Experiment III: price error for logit + macro

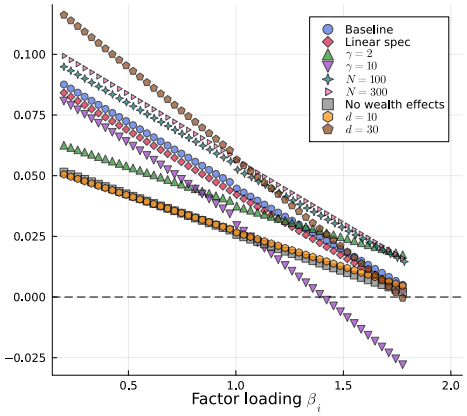


Figure A5: **Robustness of logit and logit + macro demand models across experiments.** Panels A–B plot the implied supply change, as in Figure 1, against asset beta β_i for the uniform 10% supply reduction of experiment I, under different robustness specifications for logit and logit + macro; the dashed black line shows the true supply change. Panels C–F plot the price prediction error $\Delta \log P_i^{\text{model}} - \Delta \log P_i^{\text{true}}$ relative to baseline prices P_0 for the rotation shock of experiment II (C–D) and the low-β shock of experiment III (E–F) under logit and logit + macro; the dashed line marks zero error. Each series corresponds to a different parametrization (γ, d, N; see Table A2 for details), or specification variant: using log-linear specifications, shutting down wealth effects, or letting the relative elasticity in logit depend on β.

The first diagnostic restricts identification to within-issuer variation. This comparison follows the logic of Hausman and McFadden (1984), who compare parameter estimates from a full choice set to estimates from a restricted subset. In corporate bonds, issuer is a natural restriction because bonds from the same issuer share many characteristics. We implement the comparison using the Mundlak (1978) regression approach.

Table A3: **Hausman test for the relative multiplier $\widehat{\mathcal{M}}$**

	Return R_{it}		
	(1)	(2)	(3)
Z_{it}	0.055	0.067	0.067
	(0.084)	(0.056)	(0.056)
$\bar{Z}_{issuer,it}$			-0.048
			(0.122)
Date Fixed Effects	Yes		Yes
X^1_{it} × Date Fixed Effects	Yes	Yes	Yes
X^2_{it} × Date Fixed Effects	Yes	Yes	Yes
Issuer × Date Fixed Effects		Yes	
$\bar{X}^1_{issuer,it}$ × Date Fixed Effects			Yes
$\bar{X}^2_{issuer,it}$ × Date Fixed Effects			Yes
N	1,041,985	1,041,985	1,041,985
R^2	0.464	0.783	0.466

Table A3 reports the results of a Hausman (1978) test for the relative multiplier of bond returns R_{it} on flow-induced trading demand shock Z_{it} defined in (46) for U.S. non-defaulted corporate bonds. Specification (1) is the same as specification (3) in Table 1, controlling for date fixed effects and a continuous duration variable X^1_{it} and credit risk variable based on average historical default probabilities for each S&P credit rating category X^2_{it} for each date. Specification (2) replaces date fixed effects with bond issuer × date fixed effects. Specification (3) adds the Hausman (1978) test via a Mundlak (1978) regression, testing whether the coefficients in (1) and (2) are different through the coefficient on the issuer-average of the instrument, $\bar{Z}_{issuer,it}$. The regressions weigh each date equally. The sample period is 2010:04 to 2024:03. Standard errors are clustered by date and bond issuer.

Table A3 reports the results. Specification (1) is the baseline specification, while specification (2) adds issuer-by-date fixed effects. Specification (3) implements the Mundlak regression by adding the issuer-date average of the instrument, $\bar{Z}_{issuer,it}$. The coefficient on $\bar{Z}_{issuer,it}$ tests whether the pooled and within-issuer estimates differ. The coefficient is small and statistically insignificant, suggesting that restricting identification to within-issuer variation does not materially change the estimated relative multiplier.

We next test coefficient stability with respect to an additional observable. After duration and credit risk, a natural candidate characteristic is bond liquidity. Table A4 shows that adding a proxy for bond liquidity (log outstanding amount) does not significantly change the relative multiplier.

These diagnostics are informative specification checks, but they are neither necessary nor

Table A4: **Coefficient stability test for the relative multiplier $\widehat{\mathcal{M}}$: bond liquidity**

	Return R_{it}	
	(1)	(2)
Z_{it}	0.055	0.055
	(0.084)	(0.084)
$D \times Z_{it}$		-0.022
		(0.017)
Date Fixed Effects	Yes	
X_{it}^1 × Date Fixed Effects	Yes	
X_{it}^2 × Date Fixed Effects	Yes	
$D\times$ Date Fixed Effects		Yes
X_{it}^1 × $D\times$ Date Fixed Effects		Yes
X_{it}^2 × $D\times$ Date Fixed Effects		Yes
X_{it}^3 × $D\times$ Date Fixed Effects		Yes
N	1,041,985	2,083,970
R^2	0.464	0.467

Table A4 reports the results of a coefficient stability test for the relative multiplier of bond returns R_{it} on flow-induced trading demand shock Z_{it} defined in (46) for U.S. non-defaulted corporate bonds. Specification (1) is the same as specification (3) in Table 1, controlling for date fixed effects and a continuous duration variable X_{it}^1 and credit risk variable X_{it}^2 for each date. To test whether adding a liquidity control changes the coefficient on Z_{it}, the dataset is stacked: two copies of the data are created, indexed by $D \in \{0, 1\}$, with all fixed effects and controls interacted with D so that each copy estimates its own specification. The liquidity control, standardized log bond amount outstanding (e.g., Houweling et al., 2005), enters only the $D = 1$ copy. The coefficient on the interaction $D \times Z_{it}$ directly tests whether the relative multiplier differs across the two specifications. The regressions weigh each date equally. The sample period is 2010:04 to 2024:03. Standard errors are clustered by date and bond issuer.

sufficient for Assumption A1. A rejection may reflect omitted dimensions of substitution, but it may also reflect heterogeneous relative multipliers, as discussed in Section 3.1.2.[77] For instance, if the relative multiplier differs across bond issuers, the within-issuer estimator in specification (2) may capture a different weighted average of issuer-level multipliers than the pooled estimator in specification (1), causing the Mundlak coefficient to reject even absent a violation of Assumption A1. Conversely, non-rejection cannot rule out omitted observables or unobserved dimensions of substitution.

F.2 Additional results

[77] Goldsmith-Pinkham et al. (2020) emphasize a related issue in the context of heterogeneity with Bartik instruments.

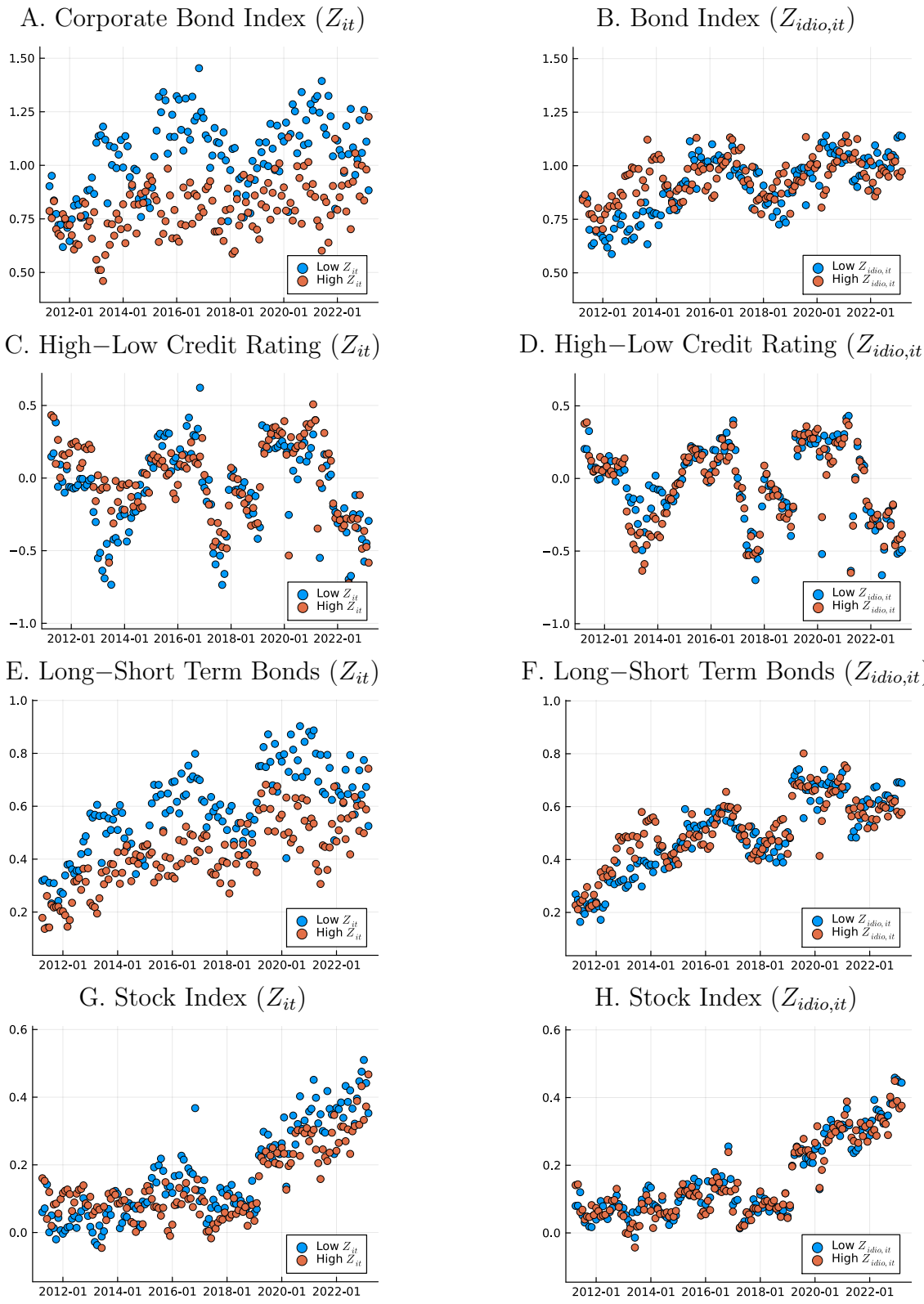


Figure A6: **Balance on covariances: exposure of portfolios sorted on demand shocks to various factors.** Figure A6 follows the exact definitions from Figure 4, but instead of showing the exposure of long-short portfolios to various factors, it shows the exposure for the long (orange) and short (blue) legs separately, sorted based on Z_{it} in the left panels and $Z_{idio,it}$ in the right panels.

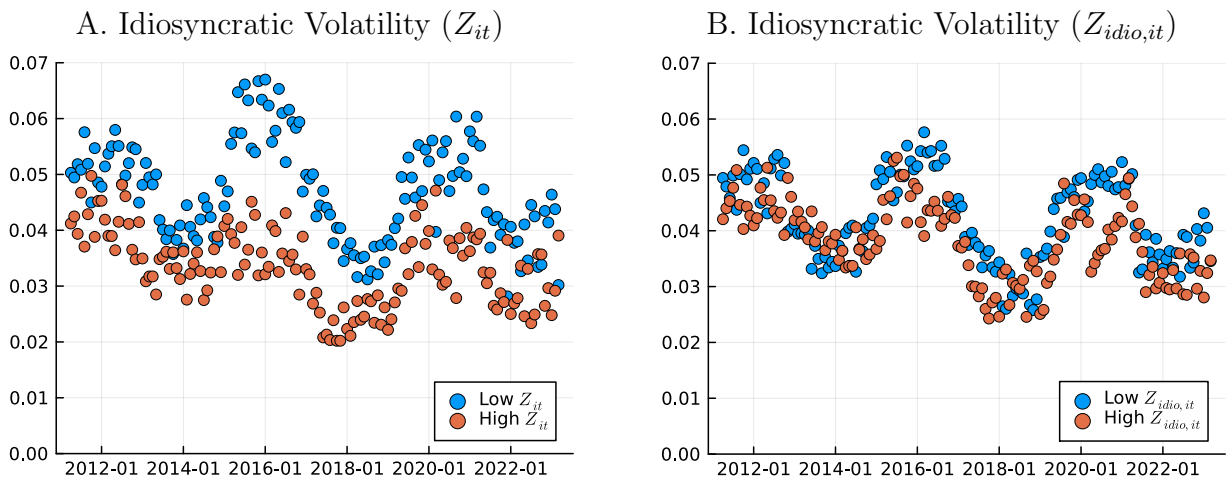


Figure A7: **Balance on variances: average idiosyncratic volatility sorted on demand shocks.** Figure A7 reports average idiosyncratic volatilities per group sorted on the values of the instrument Z_{it} (blue) or the residualized instrument $Z_{idio,it}$ (orange). Equation (46) defines Z_{it}, while $Z_{idio,it}$ is obtained by cross-sectionally orthogonalizing the instrument at each date to duration and credit risk (long-run average default probabilities based on S&P ratings). At each date, we compute idiosyncratic volatilities for each corporate bond over a two-year window centered around t, excluding t, with respect to four factors: the total return of the ICE BofA US Corporate Index, the difference in total return between the ICE BofA US High Yield Index and the ICE BofA US Corporate Index, the difference in total return between the ICE BofA 15+ Year US Corporate Index and the ICE BofA 1-3 Year US Corporate Index, and the Fama and French (1993) excess stock market return. We present the equal-weighted average idiosyncratic volatility among bonds whose instrument value Z_{it} (or $Z_{idio,it}$) is above or below the median. The data for factors is from FRED and the Kenneth French data library. The time series is from 2011:04 to 2023:03.

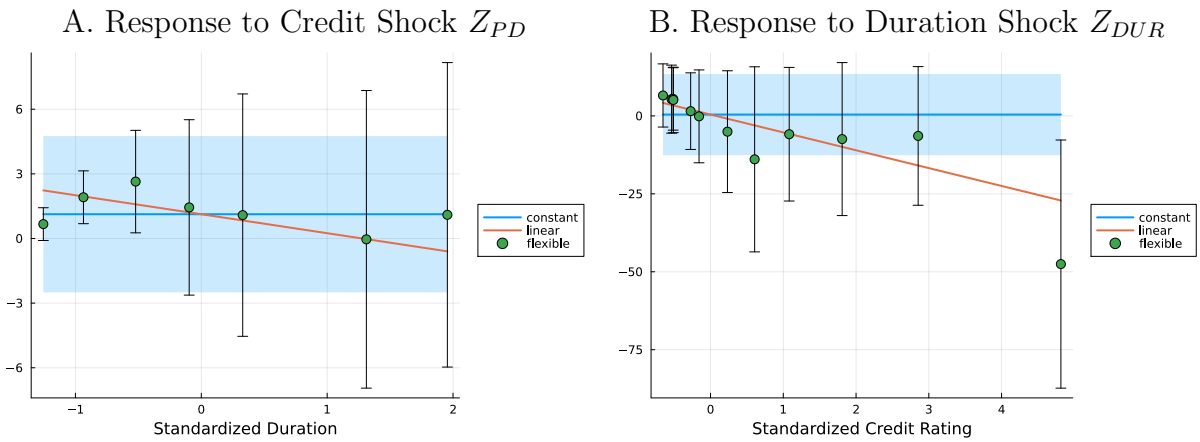


Figure A8: **Off-diagonal meso and macro multipliers in the cross-section.** Figure A8 reports the off-diagonal elements of $\widetilde{\mathcal{M}}$ omitted from Figure 5. That is, the response of bonds sorted on duration to a credit shock Z_{PD} (Panel A), and the response of bonds sorted on credit rating to a duration shock Z_{DUR} (Panel B). Bonds are grouped in seven buckets based on duration: <1 year, 1–3 years, 3–5 years, 5–7 years, 7–10 years, 10–15 years, and 15+ years. Bonds are also grouped by S&P credit rating, with individual notches from A+ through B-, and with AAA/AA and CCC/C ratings pooled at the extremes. The blue lines correspond to the estimates from column 1 of Table 2, which assume identical responses. The orange lines incorporate all results from columns 1 to 3, which include linear interaction terms with either duration or credit risk, neutralizing the other. The green dots estimate multipliers separately by duration or credit rating bucket in a pooled panel regression. The sample period is 2010:04 to 2024:03.

Table A5: **Relative multiplier $\widehat{\mathcal{M}}$ in corporate bonds in yields**

	Yield change ΔY_{it}				
	(1)	(2)	(3)	(4)	(5)
Demand shock:					
Z_{it}	-0.058			-0.281***	-0.584***
	(0.030)			(0.060)	(0.134)
$Z_{idio,it}$		-0.058	-0.058		
		(0.030)	(0.030)		
Date Fixed Effects	Yes	Yes	Yes	Yes	
Duration × Date Fixed Effects	Yes	Yes			
Credit Risk × Date Fixed Effects	Yes	Yes			
N	1,016,830	1,016,830	1,016,830	1,016,830	1,016,830
R^2	0.099	0.099	0.040	0.042	0.006

Table A5 reports the results of relative multiplier regressions of yield changes ΔY_{it} on demand shocks Z_{it} and $Z_{idio,it}$ for U.S. non-defaulted corporate bonds. Specifications (1) and (4)–(5) use the flow-induced trading demand shock Z_{it} defined in Equation (46). Specification (1) includes a time fixed effect and controls for a continuous duration variable and a continuous credit risk variable based on average historical default probabilities for each S&P credit rating category, while specification (4) uses only date fixed effects and specification (5) only a common intercept. Specifications (2)–(3) use the demand shock $Z_{idio,it}$ orthogonalized to duration and credit risk each period, with and without controlling for duration and credit risk in the regression. The regressions weigh each date equally. The sample period is 2010:05 to 2024:03. Standard errors are clustered by date and bond issuer.

Table A6: **Spillover matrix $\widetilde{\mathcal{M}}$ in corporate bonds in yields**

	Yield Change				
	$\Delta Y_{agg,t}$	$\Delta Y_{DUR,t}$	$\Delta Y_{PD,t}$	ΔY_{it}	$\Delta Y_{agg,t}$
	(1)	(2)	(3)	(4)	(5)
$Z_{agg,t}$	-2.506***	0.146	-1.475	-2.506***	-2.961***
	(0.526)	(0.150)	(0.839)	(0.523)	(0.565)
$Z_{DUR,t}$	1.142	-0.943*	3.040	1.142	
	(1.485)	(0.391)	(2.507)	(1.471)	
$Z_{PD,t}$	-0.498	0.227*	-0.939	-0.498	
	(0.330)	(0.109)	(0.491)	(0.329)	
$Z_{idio,it}$				-0.058	
				(0.030)	
N	167	167	167	1,016,830	167
R^2	0.452	0.205	0.277	0.018	0.443

Table A6 reports the results of macro- and meso multiplier regressions of yield changes on demand shocks for non-defaulted U.S. corporate bonds. Specifications (1)–(3) jointly estimate the matrix $\widetilde{\mathcal{M}}$ from Proposition 5, which together with the relative multiplier $\widehat{\mathcal{M}}$ determines the spillover matrix between observables, $\boldsymbol{\mathcal{M}}_X$. Specification (4) estimates multipliers mechanically identical to specification (1) using disaggregated, repeated cross-sectional regressions, while adding the relative multiplier $\widehat{\mathcal{M}}$. Specification (5) estimates the macro multiplier in isolation by regressing aggregate bond yield changes $\Delta Y_{agg,t}$ on the aggregated instrument $Z_{agg,t}$ in the time series. The $K = 3$ observables are a vector of ones, standardized duration, and standardized credit risk. The sample period is 2010:05 to 2024:03. Robust standard errors are used for specifications (1) to (3) and (5). For specification (4), standard errors are clustered by date and issuer, and regressions are weighted such that each date receives equal weight.

Table A7: **Spillover matrix $\widetilde{\mathcal{M}}$ in corporate bonds with ZCA whitening**

	Return					
	$R_{agg,t}$	$R_{DUR,t}$	$R_{PD,t}$	R_{it}		$R_{agg,t}$
	(1)	(2)	(3)	(4)	(5)	(6)
$Z_{agg,t}$	14.315***	7.018***	3.548**	14.315***	14.315***	15.032***
	(2.669)	(1.815)	(1.324)	(2.651)	(2.645)	(2.393)
$Z_{DUR,t}$	-0.187	5.541	-6.100	-0.187	-0.187	
	(6.729)	(4.234)	(3.901)	(6.663)	(6.658)	
$Z_{PD,t}$	1.162	-0.489	2.466	1.162	1.162	
	(2.156)	(1.443)	(1.305)	(2.138)	(2.137)	
$Z_{idio,it}$				0.055	0.055	
				(0.087)	(0.084)	
Duration DUR_{it}					-0.002	
					(0.001)	
Credit Risk PD_{it}					-0.002	
					(0.001)	
$Z_{agg,t} \times DUR_{it}$					7.018***	
					(1.793)	
$Z_{agg,t} \times PD_{it}$					3.548**	
					(1.322)	
$Z_{DUR,t} \times DUR_{it}$					5.541	
					(4.180)	
$Z_{DUR,t} \times PD_{it}$					-6.100	
					(3.885)	
$Z_{PD,t} \times DUR_{it}$					-0.489	
					(1.429)	
$Z_{PD,t} \times PD_{it}$					2.466	
					(1.305)	
N	168	168	168	1,041,985	1,041,985	168
R^2	0.343	0.125	0.386	0.100	0.138	0.342

Table A7 reports the results of spillover matrix $\widetilde{\mathcal{M}}$ regressions of bond returns on demand shocks for non-defaulted U.S. corporate bonds. It is identical to Table 2, other than that it uses ZCA whitening of standardized duration and credit risk at each date to create orthonormal observables X. It also adds another specification (5), showing that all estimates using disaggregated data are mechanically identical to those obtained from time-series regressions in specifications (1)–(3). The sample period is 2010:04 to 2024:03. Robust standard errors are used for specifications (1) to (3) and (6). For specifications (4) and (5), standard errors are clustered by date and issuer, and regressions are weighted such that each date receives equal weight.