On the Recovery of Demand Elasticities in Dynamic Settings*

Carter Davis† Mahyar Kargar‡ Jiacui Li§ Dejanir Silva¶

June 12, 2026

(first draft: April 16, 2016)

Abstract

We study how to recover and use demand elasticities in dynamic asset markets. In demand-system asset pricing, a price movement is accompanied by changes in future expected prices, returns, and risk, so there is no single context-free elasticity. The instrumented elasticity, which captures the reduced-form response to a shock that moves prices together with these future objects, is shock-specific. The structural elasticity, which holds the other arguments of demand fixed, is transportable across shock environments. We show that the inverse of the instrumented elasticity is the price multiplier and develop a recovery theory that uses impulse responses of flows, returns, and risk to extract structural demand coefficients from shock-specific reduced-form estimates. We then implement the recovery in an orthonormal factor model that accommodates unbalanced stock panels. In weekly order-flow data, the naive inverse-multiplier benchmark is about 0.35, while recovered own-price structural elasticities are about 5 without risk and 6 to 7 when factor-level risk is included.

Keywords: demand elasticity, price impact, demand system asset pricing

JEL Classification: G11, G12, G14.

*We thank Ehsan Azarmsa, Aditya Chaudhry, William Fuchs, Xavier Gabaix, Andrei Gonçalves, Valentin Haddad, Péter Kondor, Tyler Muir, Daniel Neuhann, Zhaoxin Shi, Philippe van der Beck, and conference and seminar participants at NBER New Developments in Long-Term Asset Management Spring 2026 Meeting, SFS Cavalcade, Boston College, Carnegie Mellon University, and The Ohio State University, for their helpful comments and suggestions.

†Fisher College of Business, The Ohio State University; carter.davis@fisher.osu.edu

‡Gies College of Business, University of Illinois Urbana-Champaign; kargar@illinois.edu

§David Eccles School of Business, University of Utah; jiacui.li@eccles.utah.edu

¶Mitchell Daniels School of Business, Purdue University; dejanir@purdue.edu

1 Introduction

Demand system asset pricing (DSAP) has become a central framework for studying how investor flows shape asset prices and for conducting counterfactual analysis. By modeling investor demand as a function of prices, DSAP has been used to study settings such as ESG (Koijen, Richmond, and Yogo, 2024; Van der Beck, 2021), passive intermediation (Haddad, Huebner, and Loualiche, 2025b), household participation (Davis, Knüpfer, Soerlie Kvaerner, Sen Dogan, and Vokata, 2024), and anomaly returns (Tamoni, Sokolinski, and Li, 2024). The framework is attractive because, once estimated, it can map changes in investor demand into equilibrium price effects and counterfactual repricings. However, DSAP methods (see, among others, Gabaix and Koijen, 2022; Koijen et al., 2024) have focused on a particular type of elasticity: the response to a shock that moves prices together with the entire expected path of future prices and returns. In dynamic asset markets, that is a natural object to study, but it is not the only one: many distinct elasticities are consistent with the same underlying demand system.

This multiplicity is a feature of dynamic asset markets, not a measurement problem. The Campbell-Shiller identity requires that any increase in today's price be absorbed by some combination of higher expected future dividends, lower expected future returns at various horizons, or both. A price change therefore cannot be studied in isolation from the path of cash flows, discount rates, and risk that moves with it. Because there are many ways for a price to move, there is no single demand elasticity in dynamic asset markets: different price processes imply different elasticities even under the same underlying demand system. A low measured elasticity, in particular, need not mean that investors are unwilling to buy when prices fall. It may simply reflect that the identifying shock is persistent and therefore moves not just the current price but the entire expected path of future prices, returns, and risk.

The distinction matters whenever the counterfactual of interest involves a shock with different

persistence, risk profile, or systematic exposure than the shock used for identification. We therefore distinguish two elasticities. The *structural* demand elasticity is the response of holdings to current prices, holding fixed all expected future prices, dividends, and risk. The *instrumented* demand elasticity is the reduced-form response of holdings to a shock that moves prices together with the expected path of future prices and returns. The instrumented elasticity is shock-specific: it depends on the attributes of the identifying variation. The structural elasticity is not. For most DSAP applications, which study the price impact of persistent or equilibrium flows, the instrumented elasticity is the natural object. The structural elasticity is the transportable primitive that lets a researcher move from one shock environment to another.

Recent work has sharpened this distinction by emphasizing three related difficulties. First, financial prices are dynamic: moving the price today changes resale values and therefore the term structure of expected returns, not only the next-period expected return, as discussed by Davis, Kargar, and Li (2025) and Binsbergen, David, and Opp (2025). Second, expected returns and risk move together. When prices fall, expected returns tend to rise, but return volatility, covariance, and hedging risks may rise as well. He, Kondor, and Li (2025) show that these offsetting channels can make the "observed slope" of demand understate the dynamic slope from investors' first-order conditions. Third, shocks to one asset generally spill over to other assets through substitution and no-arbitrage linkages. These cross-asset effects violate the usual stable unit treatment value assumption (SUTVA) needed for identification, as discussed by Fuchs, Fukuda, and Neuhann (2023).

Taken literally, these critiques would suggest that structural demand estimation is nearly hopeless: valid identification would require instruments that move current prices without changing future prices, leave risk fixed, and generate no cross-asset spillovers. Such instruments are unlikely to exist. Our view is more constructive. That an instrument moves future prices, risk, and other

assets is not an obstacle to identification. It is what makes identification possible. A single identified shock generates a path of flow, return, and risk responses across horizons and across assets, and the linearized demand system links those responses to the structural demand coefficients through equilibrium restrictions. Persistence, risk responses, and cross-asset spillovers are therefore features of the design, not threats to it: without movement in these dimensions, dynamic preferences over expected returns, risk, and substitution cannot be separately recovered.

This recovery is not model-free. We impose a log-linear sequential demand system with finite-dimensional coefficient matrices. The specification is general enough to nest static DSAP, mean–variance demand, intertemporal portfolio choice with Epstein–Zin preferences (Merton, 1973; Campbell and Viceira, 2002), and the adjustment-cost motive of Gârleanu and Pedersen (2013). It also nests no-arbitrage demand models in which a mean–variance arbitrageur trades against habitat-style or clientele demand, as in Vayanos and Vila (2021), so no-arbitrage is not in conflict with a demand-system approach. But structure matters. Fuchs, Fukuda, and Neuhann (2025) argue that neither demand nor expected returns can be recovered in a fully model-free way in no-arbitrage environments, and we agree. Our empirical implementation reflects this: to handle unbalanced stock panels, we impose an orthonormal factor model that collapses the cross-section. This restriction is disciplined and transparent but not innocuous. Factor-model restrictions have known limits, as emphasized by Baba Yara, Boyer, and Davis (2021) and Fuchs et al. (2025).

We make three contributions. Our first contribution concerns the instrumented elasticity and its connection to price multipliers. We show that, under linearized market clearing, the inverse of the instrumented elasticity is the price multiplier for the identifying shock. We then show how the instrumented elasticity can be decomposed into structural demand coefficients and shock-specific pass-through terms. When richer impulse responses are observed, our recovery theory goes further and uncovers the underlying structural parameters.

Our second contribution is a recovery theory for the structural demand system. If the measured elasticity depends on the shock process, can one still recover the underlying structural demand? We show that the answer is yes, and that persistent shocks are in fact useful for this purpose: they generate the dynamic impulse responses needed to identify forward-looking demand coefficients. A single contemporaneous multiplier is generally insufficient, but the full path of flow, return, and risk responses identifies the structural demand matrices through a convolution equation that links them to the reduced-form impulse responses.

Our third contribution is a practical implementation of this recovery in stock-level data. The unrestricted recovery system is written for a balanced panel of N assets, but empirical panels are unbalanced and high-dimensional. We therefore work with an orthonormal factor representation. The stock-level local projections identify the own component of demand, while the factor projection captures common substitution and risk channels. This makes the recovery feasible without changing the economic object: the factor model is a low-dimensional representation of the same market-clearing equation.

The empirical recovery results show that this distinction is quantitatively large. In weekly order-flow data, the naive inverse-multiplier benchmark, which ignores all forward-looking terms and treats the contemporaneous price response as sufficient, implies an own-price elasticity of about 0.35. Once we use the dynamic path of flow and forward-return responses to recover structural demand, the implied own-price elasticity is about 5. Including factor-level risk raises the estimate to about 6 to 7, depending on the factor representation. The reduced-form elasticity is low because the identifying shock is persistent; the structural elasticity is much larger because investors are more responsive to a price decline expected to reverse than to a persistent repricing that changes the entire path of future expected returns and risk. What looks like weak demand in reduced form is therefore consistent with much stronger structural demand.

The rest of the paper is organized as follows. Section 2 sets up the model, including a general log-linearized demand system that nests canonical asset pricing models, and defines equilibrium. Section 3 distinguishes structural from instrumented elasticities and links the latter to the price multiplier. Section 4 develops the unrestricted recovery theory. Section 5 adapts the recovery to an orthonormal factor model. Section 6 presents the empirical implementation and results. Section 7 concludes.

2 Model

This section sets up the economic environment: an economy in which a representative elastic investor trades against an inelastic outside investor. Section 2.1 introduces the assets, a general dynamic log-linear demand system for the elastic investor, and the definition of equilibrium. Section 2.2 discusses the interpretation of the demand system. Finally, Section 2.3 shows that the demand system nests several canonical models in the asset pricing literature, including static demand, mean–variance demand, adjustment costs, intertemporal hedging, and models with behavioral and informational frictions.

2.1 Environment

Assets. Consider a discrete-time economy with N risky assets indexed by $j = 1, \ldots, N$ and a risk-free asset with gross return R^f. Let $P_t \in \mathbb{R}^N$ and $\mathcal{D}_t \in \mathbb{R}^N$ denote the vectors of prices and dividends at time t, and $r_{t+1} \in \mathbb{R}^N$ denote the vector of one-period log returns.

Let $\nu_t = (\nu_{t,1}, \ldots, \nu_{t,K})$ denote a K-dimensional vector of structural shocks at time t, and let $\nu^t = (\nu_0, \nu_1, \ldots, \nu_t)$ denote the history of shocks up to time t. The vector ν_t collects all primitive

sources of randomness in the economy. In particular, dividends are an exogenous function of the history of shocks: $\mathcal{D}_t = \mathcal{D}(\nu^t, t)$.

Investors and demand. Let $Q_t \in \mathbb{R}^N$ denote the portfolio holdings of a representative elastic investor. Fix a reference point $(\bar{Q},\bar{P},\bar{\mathcal{D}})$ with strictly positive components and define log deviations[1]

$$q_t \equiv \log Q_t - \log \bar{Q}, \qquad p_t \equiv \log P_t - \log \bar{P}, \qquad d_t \equiv \log \mathcal{D}_t - \log \bar{\mathcal{D}}.$$

For each horizon $s \geq 1$, define the conditional first moments

$$\mu^p_{t,t+s} \equiv \mathbb{E}_t[p_{t+s}], \qquad \mu^d_{t,t+s} \equiv \mathbb{E}_t[d_{t+s}].$$

Define the return covariance matrix $V^r_{t,t+s} \equiv \mathbb{V}\mathrm{ar}_t(r_{t+s})$, which is a symmetric $N \times N$ matrix. Let $m \equiv N(N+1)/2$ and define its half-vectorization $v^r_{t,t+s} \equiv \mathrm{vech}(V^r_{t,t+s}) \in \mathbb{R}^m$. Working with the $\mathrm{vech}(\cdot)$ operator ensures that the second-moment state variables are vectors of m unique covariance terms rather than redundant $N \times N$ matrices.

The investor's demand depends on current prices, expected future prices and dividends, and conditional second moments of returns. Throughout the paper, we focus on the log-linearized demand system:

$$q_t = -A_0\, p_t + \sum_{s=1}^{S} A_s\, \mu^p_{t,t+s} + \sum_{s=1}^{S} B_s\, \mu^d_{t,t+s} - \sum_{s=1}^{S} D_s\, v^r_{t,t+s} + \xi_t, \tag{1}$$

where $A_0,A_s \in \mathbb{R}^{N\times N}$, $B_s \in \mathbb{R}^{N\times N}$, and $D_s \in \mathbb{R}^{N\times m}$. The matrix A_0 governs the contemporaneous

[1]For simplicity, we assume that prices, dividends, and holdings are stationary, so there is a well-defined time-invariant reference point. It is straightforward to generalize the results to the case where only prices scaled by a stochastic trend, such as the price-earnings ratio, are stationary.

response of holdings to current prices. The matrices A_s and B_s capture how expectations of future prices and dividends at horizon s influence current demand. The matrices D_s capture the effect of expected future return risk; under this sign convention, larger D_s means riskier future states depress demand more strongly. The vector ξ_t is a latent demand shifter, unobserved and potentially correlated with equilibrium prices, in the sense of Koijen and Yogo (2019).[2] We focus on the case where the latent demand shifter is an exogenous function of the shocks: $\xi_t = \xi(\nu^t, t)$.

Outside investor and market clearing. Let $Z_t \in \mathbb{R}^N$ denote the (inelastic) holdings of an outside investor, with reference point $\bar{Z}$. Outside holdings are an exogenous function of the shocks: $Z_t = Z(\nu^t, t)$. Total shares outstanding are fixed and normalized so that

$$Q_t + Z_t = \mathbf{1}_N.$$

Linearizing around $(\bar{Q},\bar{Z})$ with $\bar{Q} + \bar{Z} = \mathbf{1}_N$ yields

$$\mathrm{Diag}(\bar{Q})\, q_t + z_t = 0, \qquad z_t \equiv Z_t - \bar{Z}. \tag{2}$$

Define $\Lambda_Q \equiv \mathrm{Diag}(\bar{Q})^{-1}$. Then market clearing implies

$$q_t = -\Lambda_Q\, z_t. \tag{3}$$

Substituting (1) into (3) gives the equilibrium condition

$$A_0\, p_t - \sum_{s=1}^{S} A_s\, \mu^p_{t,t+s} = \Lambda_Q\, z_t + \sum_{s=1}^{S} B_s\, \mu^d_{t,t+s} - \sum_{s=1}^{S} D_s\, v^r_{t,t+s} + \xi_t. \tag{4}$$

[2] Appendix H.3 maps several leading models of belief distortions and dispersed information into this formulation and discusses when ξ_t reflects behavioral belief shocks, omitted information, or linearization residuals.

This is a forward-looking equation linking current prices and expected future prices to the exogenous variables. Its solution, and thus the equilibrium price process, depends on the stochastic environment, in particular on the history of all shocks that drive the forcing terms on the right-hand side.

Definition 1 (Equilibrium)**.** *Given exogenous processes for outside holdings* $\{z_t\}$*, latent demand* $\{\xi_t\}$*, and dividends* $\{d_t\}$*, an equilibrium is a pair of stochastic processes* $\{p_t, q_t\}$ *adapted to the filtration generated by* v^t *such that:*

1. Demand optimality. *The investor's holdings satisfy the demand system* (1) *at every date* t*.*

2. Market clearing. *The market clearing condition* $Diag(\bar{Q})\, q_t + z_t = 0$ *holds at every date* t*.*

2.2 Interpreting the demand system

The demand system (1) is the primitive of our analysis, and this subsection discusses the interpretation that justifies this role. We make three points. First, the demand system is structural: its coefficients answer counterfactual questions, with each coefficient defined by what is held fixed when prices or expectations move. Second, the demand system is cast in sequence space, as a function of the expected paths of prices, dividends, and risk rather than of an assumed set of state variables; this choice is what makes the counterfactuals well-defined and allows us to remain agnostic about the states investors actually track. Third, the demand coefficients exhaust the preference information that is relevant for counterfactual analysis, so taking demand, rather than utility, as the primitive entails no loss of generality.

A structural demand system. The demand system (1) is *structural* in the sense of Haavelmo (1943): it specifies behavior in structural form and is assumed to be a policy-invariant (autonomous)

function of its arguments.[3] Hence, the demand coefficients $\{A_0, A_s, B_s, D_s\}$ answer counterfactual questions about how demand would respond to changes in the environment. For instance, A_0 provides the elasticity of demand with respect to current prices, holding all future expected prices, dividends, and risk fixed. In practice, price movements are never observed in isolation, which makes recovering the structural demand coefficients challenging.

There are alternative formulations of the demand system, typically obtained by restricting the coefficients of (1), that effectively answer different counterfactual questions. For instance, instead of conditioning on the sequence of expected prices, we could express the demand in terms of the path of expected returns:

$$q_t = \sum_{s=1}^{S} C_s \, \mathbb{E}_t[r_{t+s}] - \sum_{s=1}^{S} D_s \, v^r_{t,t+s} + \xi_t. \tag{5}$$

In this formulation, the coefficients C_s capture how current holdings respond to a change in expected returns at horizon s, holding all other expected returns and risk fixed. The formulation above is a special case of the demand system (1). To see this, consider the Campbell–Shiller approximation relating log returns to log prices and log dividends:

$$r_{t+1} = \kappa + \rho \, p_{t+1} + (1 - \rho) \, d_{t+1} - p_t,$$

where $\rho \in (0,1)$ and κ are log-linearization constants.

Taking conditional expectations of the expression above at each horizon and substituting into (5), we can recover the coefficients (A_0, A_s, B_s) from the return-based coefficients (C_s):

$$A_0 = C_1, \qquad A_s = \rho \, C_s - C_{s+1} \;\text{ for } s = 1, \ldots, S, \qquad B_s = (1 - \rho) \, C_s \;\text{ for } s = 1, \ldots, S, \tag{6}$$

[3] See, e.g., Heckman and Pinto (2024) for a similar definition of a structural equation.

with the convention $C_{S+1} \equiv 0$, and where the constant terms involving κ are absorbed into ξ_t.

An implication of the fact that the structural demand admits different representations is that we cannot talk about price elasticities without specifying all the other arguments of the demand system, that is, precisely what is kept fixed in the counterfactual. Moreover, knowing the demand coefficients (A_0, A_s, B_s, D_s), one can compute the demand response to *any* counterfactual path of its arguments.

Sequence-space vs. state-space representations. An important feature of equation (1) is that it represents demand as a function of the entire sequence of expected prices, dividends, and risk. This specification corresponds to the *sequence-space representation* of the demand system.[4] This is in contrast to virtually all of the literature on dynamic portfolio choice, which represents demand as a function of a state vector that summarizes the model dynamics, corresponding to the *state-space representation* of the demand system.[5]

There are several advantages to the sequence-space representation. First, it enables us to cleanly define the counterfactual response of demand to a change in current prices, holding future expected prices fixed. Movements in state variables typically affect the entire path of current and future expected prices. Hence, there is no direct way to isolate the effect of a change in current prices on demand from movements in future expected prices. Second, we do not need to assume that the state variables are observable or identifiable by the econometrician. Given the substantial debate about the state variables that are relevant for portfolio choice, the sequence-space representation allows

[4]See Ljungqvist and Sargent (2018) for a discussion of the sequence-space vs. state-space representation of dynamic models. Appendix D provides a detailed discussion in the context of our setting.

[5]The dynamic programming approach delivers portfolio rules as functions of the state variables; see, e.g., Merton (1969, 1973) and Campbell and Viceira (2002). The martingale approach of Karatzas, Lehoczky, and Shreve (1987) and Cox and Huang (1989) is closer in spirit to the sequence-space representation, but recovering the portfolio policy that implements the optimal wealth profile still requires the dynamics of the state variables.

us to remain agnostic about the choice of state variables.[6] Third, the sequence-space representation will be instrumental in devising identification strategies for the demand coefficients in a dynamic setting: the equilibrium condition (4) links the demand coefficients to the impulse responses of prices and flows, which are estimable from the data, as we show in Section 4.

Demand vs. utility. For convenience, we take the demand system as our primitive. This allows us to specify a flexible demand system that is consistent with a rich set of possible microfoundations. From the perspective of consumer theory, one can either take the utility function as the primitive and demands as the derived object, or take the demand system as the primitive and the utility function as the derived representation (see, e.g., Berry and Haile, 2021, for a discussion). Given the coefficient matrices (A_0, A_s, B_s, D_s), one can recover the preference parameters, up to monotone transformations of the utility function, from the integrability theorem in classical demand theory.[7] This provides a formal justification for treating the structural demand coefficients as the primitive objects of interest: they encode all the preference information that is relevant for counterfactual analysis.

2.3 Nesting canonical models

The demand system (1) is deliberately general. By restricting the coefficient matrices (A_0, A_s, B_s, D_s), it nests the main demand specifications used in the asset pricing and DSAP literatures.

[6] See, for instance, Welch and Goyal (2008) for a discussion of the limited agreement on the set of relevant state variables that drive expected returns.

[7] See Mas-Colell, Whinston, and Green (1995, Chapter 3). Appendix C states the theorem and discusses its implications for the demand coefficients. Since demand is linearized, the recovered utility corresponds to the quadratic expansion of preferences, as in linear-quadratic economies (Hansen and Sargent, 2013).

Static demand. The simplest case sets $S = 0$, so that no future expectations enter demand:

$$q_t = -A_0\, p_t + \xi_t, \tag{7}$$

abstracting from movements in expected dividends and volatility, for simplicity. This case nests the specification that underlies much of the early DSAP literature, including Koijen and Yogo (2019) and Koijen et al. (2024), as well as the richer specification in Haddad, He, Huebner, Kondor, and Loualiche (2025a). In this model, investors respond only to current prices; A_0 corresponds to the matrix of own- and cross-price elasticities, and there are no intertemporal links.[8]

Mean–variance (myopic) demand. Setting $S = 1$ restricts the investor to care only about one-period-ahead prices and dividends:

$$q_t = -A_0\, p_t + A_1\, \mu^p_{t,t+1} + B_1\, \mu^d_{t,t+1} + \xi_t, \tag{8}$$

assuming that second moments are constant. The investor is myopic: only next-period expectations influence current holdings.

A concrete microfoundation corresponds to the mean–variance investor with constant risk aversion $\gamma > 0$ and perceived (constant) dollar covariance matrix $\Sigma \succ 0$. Exact demand in levels is

$$Q_t = \frac{1}{\gamma}\, \Sigma^{-1} \big(\mathbb{E}_t[P_{t+1} + \mathcal{D}_{t+1}] - R^f\, P_t \big).$$

[8] Relative to Koijen and Yogo (2019), the specification in (7) allows for a more flexible pattern of substitution: as discussed in Appendix H, the substitution pattern in their demand system is pinned down by the own-price elasticity and the vector of portfolio shares.

Log-linearizing around $(\bar{Q},\bar{P},\bar{\mathcal{D}})$ delivers (8) with

$$A_0 = \mathrm{Diag}(\bar{Q})^{-1}\frac{R^f}{\gamma}\,\Sigma^{-1}\mathrm{Diag}(\bar{P}), \qquad A_1 = \beta\, A_0, \qquad B_1 = \mathrm{Diag}(\bar{Q})^{-1}\frac{1}{\gamma}\,\Sigma^{-1}\mathrm{Diag}(\bar{\mathcal{D}}), \tag{9}$$

where $\beta = 1/R^f$. The relation $A_1 = \beta\, A_0$ reflects the fact that, in this model, the investor cares about prices and dividends only through their effect on next-period expected returns.

Adjustment costs. Adjustment costs provide a motivation for demand to depend on the entire path of future expected returns. Gârleanu and Pedersen (2013) study the optimal trading strategy of a mean–variance investor who faces quadratic transaction costs. The key insight is that the optimal portfolio depends on the "aim" portfolio, a forward-looking weighted average of future mean-variance portfolios, toward which the investor trades partially each period.

In the Gârleanu and Pedersen (2013) model, the investor's position Q_t evolves according to

$$Q_t - Q_{t-1} = \lambda_{tc}\big(\mathrm{Aim}_t - Q_{t-1}\big), \tag{10}$$

where $\lambda_{tc} \in (0,1)$ governs the speed of adjustment to the aim portfolio.

The aim portfolio corresponds to a weighted average of future myopic demands:

$$\mathrm{Aim}_t \propto \sum_{s=0}^{\infty} \psi_{tc}^{s}\,\Sigma^{-1}\,\mathbb{E}_t\big[P_{t+s+1} + \mathcal{D}_{t+s+1} - R^f P_{t+s}\big],$$

where the decay rate $\psi_{tc} \in (0,1)$ is a function of the transaction cost parameter. In the log-linear demand notation, the geometric weighting of future frictionless optima generates price–dividend coefficients that decay geometrically:

$$A_s \propto \psi_{tc}^{s-1}\,A_1, \qquad B_s \propto \psi_{tc}^{s-1}\,B_1, \qquad s = 1,2,\ldots,$$

where A_1 and B_1 are the myopic (frictionless) coefficients, and the latent demand ξ_t captures the dependence on the previous portfolio choice q_{t-1}.

Although the underlying preference is myopic, transaction costs make demand dynamic: $A_s \neq 0$ and $B_s \neq 0$ for all $s > 1$, because adjusting today is costly and the investor optimally anticipates future desired positions. Higher transaction costs imply slower adjustment (lower λ_{tc}) and a longer effective horizon (higher ψ_{tc}). The partial-adjustment structure in (10) is what generates the dependence on the lagged portfolio Q_{t-1} captured by ξ_t.

Intertemporal hedging demand. Merton (1973) shows that the portfolio share of a long-horizon investor combines the myopic mean–variance portfolio with a *hedging* portfolio that insures against adverse shifts in future investment opportunities. Hedging demands are typically written as functions of the state variables driving those opportunities. We show next that they can be equivalently expressed in terms of the path of future expected returns, as in the demand system (1).

To see this, consider the case of a single risky asset with constant return variance σ_r^2, where expected excess returns are affine in an S-dimensional vector of state variables X_t:

$$\mathbb{E}_t[r^e_{t+1}] = \mu_r + \psi_r^\top X_t,$$

where $\mu_r \in \mathbb{R}$ is the average excess return and $\psi_r \in \mathbb{R}^S$ is a vector of coefficients. The state vector follows a VAR(1):

$$X_t = \Phi X_{t-1} + \epsilon_t, \qquad \epsilon_t \sim \mathcal{N}(0, \Sigma_\epsilon).$$

For this environment, Campbell and Viceira (2002) derive a log-linear approximation of the

portfolio share α_t of a long-horizon investor with risk aversion γ:

$$\alpha_t = \underbrace{\frac{\mathbb{E}_t[r^e_{t+1}] + \sigma_r^2/2}{\gamma\sigma_r^2}}_{\text{myopic component}} + \underbrace{A_{\alpha,0} + A_{\alpha,1}^\top X_t}_{\text{hedging component}},$$

where the coefficients $A_{\alpha,0} \in \mathbb{R}$ and $A_{\alpha,1} \in \mathbb{R}^S$ characterize the hedging demand. This is the standard representation of the portfolio share as an affine function of the state variables.

To express the portfolio share in terms of the path of expected returns instead, note that the VAR structure implies

$$\mathbb{E}_t[r^e_{t+s}] = \mu_r + \psi_r^\top \Phi^{s-1} X_t, \qquad s \geq 1.$$

Assuming that the vector $[\mathbb{E}_t[r^e_{t+1}], \ldots, \mathbb{E}_t[r^e_{t+S}]]^\top$ spans the state vector X_t, there exist coefficients $\tilde{C}_s$, $s = 1, \ldots, S$, such that $A_{\alpha,1}^\top X_t = \sum_{s=1}^S \tilde{C}_s(\mathbb{E}_t[r^e_{t+s}] - \mu_r)$, and hence[9]

$$\alpha_t = \frac{\mathbb{E}_t[r^e_{t+1}] + \sigma_r^2/2}{\gamma\sigma_r^2} + \tilde{A}_{\alpha,0} + \sum_{s=1}^S \tilde{C}_s \mathbb{E}_t[r^e_{t+s}],$$

where $\tilde{A}_{\alpha,0} = A_{\alpha,0} - \sum_{s=1}^S \tilde{C}_s \mu_r$ absorbs the constant terms. Hence, hedging demand implies that the investor cares not only about one-period-ahead returns but also about expected returns further into the future.

It remains to map the portfolio share into quantities. Demand satisfies $Q_t = \alpha_t W_t(1 - C_t)/P_t$, where W_t is wealth and C_t is the consumption-wealth ratio, so linearizing q_t requires expanding both W_t and C_t. An argument analogous to the one above shows that the consumption-wealth ratio

[9] In more general settings, the state variables may drive both expected excess returns and volatility, in which case the joint path of first and second moments of returns would be required to span the state vector.

is also a function of the path of expected returns. Appendix H shows that the linearized demand is

$$q_t = -A_0^w p_t + \sum_{s=1}^{S} C_s \mathbb{E}_t[r_{t+s}] + \xi_t,$$

where the coefficients C_s combine the contributions of the hedging demand and of the consumption-wealth ratio, and the latent demand ξ_t captures the dependence on the previous portfolio choice q_{t-1} implicit in wealth. Since the risk-free rate is constant, expected returns and expected excess returns differ by a constant, which is absorbed into ξ_t. This generalizes the return-based representation (5): the coefficient A_0^w captures the direct effect of current prices on holdings, through the conversion of portfolio shares into quantities and through the wealth channel.

An important advantage of this sequence-space representation of hedging demands is that it does not require the econometrician to directly observe the state variables considered by the investor. Instead, the demand system can remain agnostic about the underlying state variables and model explicitly the dependence of demand on the entire path of future expected returns.

Hedging demand is economically important for the elasticity analysis. Because the hedging component responds to longer-horizon expectations, a persistent shock that shifts the entire expected path of future prices elicits a different demand response than a transitory shock that moves only next-period expectations.

Behavioral and informational frictions. We have considered so far full-information rational-expectations (FIRE) models. We show next that our demand system also nests models with behavioral frictions and informational frictions.

To accommodate these models, suppose investors choose their portfolios using subjective beliefs

about future prices, dividends, and risk:

$$q_t = -A_0\, p_t + \sum_{s=1}^{S} \hat{A}_s\, \hat{\mu}^p_{t,t+s} + \sum_{s=1}^{S} \hat{B}_s\, \hat{\mu}^d_{t,t+s} - \sum_{s=1}^{S} \hat{D}_s\, \hat{v}^r_{t,t+s}, \tag{14}$$

where $\hat{\mu}^p_{t,t+s}$ and $\hat{\mu}^d_{t,t+s}$ denote the investor's perceived paths of prices and dividends at horizon s, $\hat{v}^r_{t,t+s}$ denotes perceived conditional return risk, and $\hat{A}_s$,$\hat{B}_s \in \mathbb{R}^{N\times N}$ and $\hat{D}_s \in \mathbb{R}^{N\times m}$ are subjective demand coefficients.

We assume that subjective beliefs are related to the objective expectations through

$$\hat{\mu}^p_{t,t+s} = M^p_s \mu^p_{t,t+s} + \xi^p_{t,t+s}, \qquad \hat{\mu}^d_{t,t+s} = M^d_s \mu^d_{t,t+s} + \xi^d_{t,t+s}, \qquad \hat{v}^r_{t,t+s} = M^r_s v^r_{t,t+s} + \xi^r_{t,t+s}, \tag{15}$$

where M^p_s,$M^d_s \in \mathbb{R}^{N\times N}$ and $M^r_s \in \mathbb{R}^{m\times m}$ summarize systematic distortions of beliefs, while $\xi^p_{t,t+s}$, $\xi^d_{t,t+s}$, and $\xi^r_{t,t+s}$ capture residual belief or information wedges.

Appendix H.3 shows how to map a range of models to the formulation above. The matrices M_s and wedges $\xi_{t,t+s}$ admit different interpretations depending on the model. In behavioral-inattention models (Gabaix, 2019, 2023), the matrices M_s capture the degree of attention paid to different variables or horizons. In diagnostic-expectations models (Bordalo, Gennaioli, and Shleifer, 2018), they capture extrapolation from recent news. In noisy rational-expectations models, the wedges $\xi_{t,t+s}$ arise because investors condition on information that is not observed by the econometrician.

Substituting (15) into (14) yields

$$A_s = \hat{A}_s M^p_s, \qquad B_s = \hat{B}_s M^d_s, \qquad D_s = \hat{D}_s M^r_s,$$

plus a composite latent-demand term that absorbs the wedges $\xi^p_{t,t+s}$, $\xi^d_{t,t+s}$, and $\xi^r_{t,t+s}$.

Two caveats are worth emphasizing. First, without direct information on subjective beliefs, one

Model	Horizon S	Coefficient restrictions	References
Static	0	$A_s = B_s = 0$ for all $s \geq 1$	Koijen and Yogo (2019), Gabaix and Koijen (2022)
Mean–variance	1	$A_1 = \beta A_0$; $A_s = B_s = 0$ for $s \geq 2$	Mean–variance CAPM
Behavioral and info. frictions	S≥ 1	$A_s = \hat{A}_s M_s^p$; $B_s = \hat{B}_s M_s^d$; $D_s = \hat{D}_s M_s^r$	Gabaix (2019), Bordalo et al. (2018)
Hedging demand	$S > 1$	A_s, B_s unrestricted	Merton (1973), Campbell and Viceira (2002)
Adjustment costs	∞	$A_s \propto \psi^{s-1} A_1$; $B_s \propto \psi^{s-1} B_1$	Gârleanu and Pedersen (2013)

Table 1. Nesting canonical demand models.
Each row shows the restrictions on the demand coefficients $\{A_s, B_s\}_{s=1}^{S}$ that reduce the general demand system (1) to a canonical special case. The static model ignores all future expectations. The mean–variance model responds only to one-period-ahead expected prices and dividends. The behavioral row covers models in which investors act on subjective expected paths: belief distortions are captured through matrices (M_s^p,M_s^d,M_s^r), and belief shocks absorbed in ξ_t. Hedging demand in Merton (1973); Campbell and Viceira (2002) adds unrestricted sensitivity to longer-horizon expectations. Gârleanu and Pedersen (2013) generate geometrically decaying coefficients through the aim-portfolio structure induced by quadratic adjustment costs.

can only identify the composite coefficients such as $A_s = \hat{A}_s M_s^p$, which combine the investor's response to subjective expectations with the mechanism generating those expectations. Second, the structural interpretation of the recovered coefficients is relative to the class of interventions that leave the matrices M_s unchanged. Whether this restriction is plausible depends on the economic environment and the counterfactual under consideration.[10]

Taking stock. The examples above illustrate that the demand system (1) provides a common language for a wide range of models used in asset pricing and the demand-system asset-pricing literature. Table 1 summarizes the coefficient restrictions that map the general demand system into several canonical specifications. Appendix H shows that a number of additional models, including the preferred-habit model of Vayanos and Vila (2021), can also be expressed in this form. We next show how these coefficients can be recovered from the data.

[10]Hence, the demand system (1) is structural relative to a set of possible interventions. This view is consistent, e.g., with Hurwicz (1962) who claims that *"the concept of structure is relative to the domain of modifications anticipated."*

3 Structural vs. Instrumented Elasticities

In dynamic asset markets, any price movement is accompanied by changes in future expected prices, returns, and risk, so there is no single context-free demand elasticity. The Campbell–Shiller identity requires that a price change today be absorbed by some combination of higher expected future dividends, higher expected returns at various horizons, or both. Because there are many possible configurations of how expectations of future variables accompany the current price movement, the measured elasticity depends on which configuration the identifying variation generates, even under the same underlying demand system. A low measured elasticity need not mean that investors are intrinsically unwilling to buy when prices fall; it may instead reflect that the identifying shock is persistent and therefore changes not just the current price, but the entire expected path of future returns and risk.

This section draws a sharp line between two objects. The *structural* demand elasticity is the response of holdings to current prices, holding fixed all future expected prices, dividends, and risk. The *instrumented* demand elasticity is the reduced-form response to a shock that moves prices together with the expected path of future prices, returns, and risk. The instrumented elasticity is shock-specific: it depends on the persistence, risk profile, and systematic exposure of the identifying variation. The structural elasticity is not; it is the transportable primitive that allows a researcher to move from one counterfactual scenario to another. We show that the instrumented elasticity equals the inverse of the price multiplier. We then decompose it into the structural demand coefficients and characterize the special cases that arise when the identifying shock operates through different channels.

3.1 Structural demand elasticities

Contemporaneous structural elasticity. Because the analysis in this section concerns the price channel of demand, it is convenient to absorb expected dividends, risk, and latent demand into a composite shifter and write the demand system (1) as

$$q_t = -A_0\, p_t + \sum_{s=1}^{S} A_s\, \mu^p_{t,t+s} + f_t, \qquad f_t \equiv \sum_{s=1}^{S} B_s\, \mu^d_{t,t+s} - \sum_{s=1}^{S} D_s\, v^r_{t,t+s} + \xi_t.$$

The shifter f_t is a composite equilibrium object rather than an exogenous forcing process: while latent demand and dividends are exogenous, the conditional return variances are determined in equilibrium. Holding fixed all future expected prices $\mu^p_{t,t+s}$ and the composite shifter f_t, the local elasticity of log holdings with respect to current log prices is

$$-\frac{\partial q_t}{\partial p_t^\top} = A_0.$$

This is a partial-equilibrium object: the slope of demand with respect to current prices when every other argument of the demand system is held constant.

At first glance, A_0 might seem like the only "true" demand elasticity, since it is the derivative of current holdings with respect to current prices holding all future objects fixed. But that interpretation is too narrow for dynamic asset markets. The Campbell–Shiller identity (introduced in detail in Section 3.2 below) imposes an accounting constraint: if current prices move, then some combination of expected future returns and expected future dividends must move as well for beliefs to remain internally consistent. Once that consistency requirement is imposed, many consistent structural elasticities become relevant, each corresponding to a different assumption about how a price movement is absorbed.

3.2 Instrumented elasticity and the price multiplier

Instrumented elasticity. The instrumented elasticity (IE) matrix is the ratio of the reduced-form demand response to the first-stage price response:

$$\mathcal{E}_t(z_t) \equiv \underbrace{-\frac{dq_t}{dz_t^\top}}_{\text{reduced form}} \overbrace{\left(\frac{dp_t}{dz_t^\top}\right)}^{\text{first stage}}{}^{-1}. \tag{16}$$

This is the population analog of an instrumental-variables estimator: demand regressed on prices, instrumented by the demand shock z_t.

Price multiplier. The first-stage matrix in (16) defines the price multiplier:

$$\mathcal{M}_t(z_t) \equiv \frac{dp_t}{dz_t^\top}.$$

3.3 Demand shock exogeneity and the main equivalence

Assumption 1 (Latent demand exogeneity)**.** *Let the identifying outside-demand process admit the moving-average representation*

$$z_t = \bar{z} + \sum_{h=0}^{\infty} \Psi_h^z \varepsilon_{t-h},$$

where ε_t is the identified exogenous innovation in outside demand at date t, and Ψ_h^z is the horizon-h impulse response of z_{t+h} to ε_t. The latent demand shifter is exogenous to both the current outside-demand state z_t and the identified innovations $\{\varepsilon_t\}$:

$$\frac{d\xi_t}{dz_t^\top} = 0.$$

Moreover, in the linear projection of ξ_t on the shock history $\{\varepsilon_{t-h}\}_{h\geq 0}$,

$$\xi_t = \bar{\xi} + \sum_{h=0}^{\infty} \Psi_h^{\xi} \varepsilon_{t-h}, \qquad \Psi_h^{\xi} = 0 \quad \textit{for all } h \geq 0.$$

This assumption requires that the identifying shock does not directly shift the unobserved component of demand. In the language of Koijen and Yogo (2019), it rules out a correlation between the instrument and the latent demand for unobserved characteristics. Assumption 1 ensures that the instrument affects demand only through the equilibrium price path. Appendix H.3 discusses micro-founded sources of ξ_t—behavioral belief shocks and omitted-information wedges—and when this exclusion restriction applies.

Equivalence result. Proposition 1 links the instrumented elasticity to the price multiplier.

Proposition 1 (Instrumented elasticity and the multiplier)**.** *Suppose the price multiplier $\mathcal{M}_t(z_t)$ is invertible, linearized market clearing* (3) *holds, and Assumption 1 holds. Then*

$$\mathcal{E}_t(z_t) = \Lambda_Q \, \mathcal{M}_t(z_t)^{-1}. \tag{17}$$

Define the aggregate elasticity $\tilde{\mathcal{E}}_t(z_t) \equiv \Lambda_Q^{-1} \mathcal{E}_t(z_t) = Diag(\bar{Q}) \, \mathcal{E}_t(z_t)$. Then

$$\tilde{\mathcal{E}}_t(z_t) = \mathcal{M}_t(z_t)^{-1}. \tag{18}$$

See Appendix B for the proof. The aggregate elasticity $\tilde{\mathcal{E}}_t(z_t)$ rescales the investor-level IE into

shares-outstanding units, matching the units of z_t. In those units, the aggregate elasticity is the inverse of the price multiplier.[11]

3.4 What the instrumented elasticity loads on

The instrumented demand elasticity is a reduced form quantity. It is a weighted combination of the structural demand coefficients, with weights determined by how the identifying shock passes through to future expected prices, dividends, and risk. For each horizon s, define the shock-specific pass-through matrices

$$\Pi^{p}_{t,s}(z_t) \equiv \frac{d\mu^{p}_{t,t+s}}{dz_t^\top}\left(\frac{dp_t}{dz_t^\top}\right)^{-1}, \qquad \Pi^{d}_{t,s}(z_t) \equiv \frac{d\mu^{d}_{t,t+s}}{dz_t^\top}\left(\frac{dp_t}{dz_t^\top}\right)^{-1}, \qquad \Pi^{v}_{t,s}(z_t) \equiv \frac{dv^{r}_{t,t+s}}{dz_t^\top}\left(\frac{dp_t}{dz_t^\top}\right)^{-1}.$$

These objects are the characteristic analogs of the instrumented elasticity itself: each is the reduced-form response of the corresponding characteristic to the shock, divided by the first-stage price response.

Proposition 2 (Instrumented elasticity and shock-specific pass-through)**.** *Suppose the price multiplier $\mathcal{M}_t(z_t) = dp_t/dz_t^\top$ is invertible and Assumption 1 holds. Then the instrumented elasticity satisfies*

$$\mathcal{E}_t(z_t) = A_0 - \sum_{s=1}^{S} A_s\,\Pi^{p}_{t,s}(z_t) - \sum_{s=1}^{S} B_s\,\Pi^{d}_{t,s}(z_t) + \sum_{s=1}^{S} D_s\,\Pi^{v}_{t,s}(z_t), \tag{19}$$

See Appendix B for the proof.

The IE therefore loads on three channels: pass-through of the price movement to future expected prices, to future expected dividends, and to future return risk. The structural elasticity A_0 is the

[11] With multiple elastic investors, the aggregate IE is the holdings-weighted sum of their instrumented elasticities in shares-outstanding units. The outside investor contributes zero because its demand is inelastic. Section 4.3 gives the extension with heterogeneous investors.

baseline; the remaining terms capture how the instrument propagates through each characteristic at each horizon.

3.5 Special cases

Equation (19) shows that the IE generically depends on all the structural demand coefficients and all the pass-through matrices induced by the instrument. Only under stronger restrictions on the shock does the IE isolate narrower structural objects.

Static IE. If the instrument induces a purely transitory price change with no pass-through to future expected prices, dividends, or risk ($\Pi^p_{t,s}(z_t) = \Pi^d_{t,s}(z_t) = \Pi^v_{t,s}(z_t) = 0$ for all s), then

$$\mathcal{E}^{\text{stat}}_t(z_t) = A_0.$$

The IE coincides with the structural elasticity. This is the only case in which the instrument identifies A_0 directly.

Variance IE. If the instrument changes conditional second moments while leaving expected future prices and dividends unchanged ($\Pi^p_{t,s}(z_t) = \Pi^d_{t,s}(z_t) = 0$ for all s), then

$$\mathcal{E}^{\text{var}}_t(z_t) = A_0 + \sum_{s=1}^{S} D_s \, \Pi^v_{t,s}(z_t).$$

This covers shocks that operate through future return variance, dividend variance, or price-dividend covariance.

Immediate IE. If the instrument generates a pure one-period price movement that passes through entirely to expected prices one period ahead and nothing else ($\Pi^p_{t,1}(z_t) = I$, $\Pi^p_{t,s}(z_t) = 0$ for $s \geq 2$, $\Pi^d_{t,s}(z_t) = \Pi^v_{t,s}(z_t) = 0$), then

$$\mathcal{E}^{\text{imm}}_t(z_t) = A_0 - A_1.$$

The IE picks up the contemporaneous price sensitivity net of the one-period-ahead expected price sensitivity. Under myopic mean-variance demand with $A_1 = \beta A_0$, this reduces to $(1 - \beta) A_0$.

Dynamic IE. If the instrument shifts the entire expected path of future prices and possibly future risk, then

$$\mathcal{E}^{\text{dyn}}_t(z_t) = A_0 - \sum_{s=1}^{S} A_s \Pi^p_{t,s}(z_t) - \sum_{s=1}^{S} B_s \Pi^d_{t,s}(z_t) + \sum_{s=1}^{S} D_s \Pi^v_{t,s}(z_t).$$

This is the empirically relevant case. Persistent flow instruments generally identify neither A_0 nor any single coefficient matrix; they identify a weighted combination of all the structural demand coefficients, with weights determined by how the shock propagates through expected prices, dividends, and risk at each horizon.

3.6 There is no single price elasticity

These special cases make clear why searching for "the" demand elasticity is misguided in dynamic asset markets. Different instruments generate different pass-through patterns across horizons and across the price, dividend, and risk channels, and therefore recover distinct objects even when all are consistent with the same underlying structural demand system. The multiplicity of IEs is not a defect of any particular instrument; it is a consequence of the fact that each IE bundles the structural demand slopes with the propagation of that specific shock through the full set of demand-relevant state variables.

This is why counterfactual analysis requires the structural coefficients, not an IE. Much of the DSAP literature estimates price multipliers using persistent instruments that shift not just current prices but the entire expected path of future prices and returns. Such designs do not recover a ceteris-paribus elasticity that holds expectations fixed. They recover IEs that blend structural demand slopes with the propagation of the instrument through prices, dividends, and risk at each horizon. The practical problem is immediate: when the counterfactual of interest involves a shock with different persistence, risk profile, or systematic exposure than the shock used for identification, the IE estimated from the data cannot be transported to the new scenario. The structural demand coefficients $\{A_s, B_s, D_s\}$ can, because they are properties of the investor's demand function rather than of any particular shock process.

Asset demand is indirect: investors do not value prices per se, but the return and cash-flow characteristics that prices encode. Once that pass-through is made explicit, the multiplicity of elasticities is no longer puzzling. Each elasticity corresponds to a distinct ceteris-paribus clause about how a price movement reallocates those characteristics across horizons. The challenge, taken up in Section 4, is to recover the transportable structural coefficients from the shock-specific reduced-form evidence.

4 Recovering Demand in a Dynamic Model

Given that the inverse of an equilibrium multiplier is generally not the contemporaneous structural elasticity, a separate recovery step is needed. In this section, we show that reduced-form price responses do contain enough information to recover the underlying intertemporal demand system. The key observation is that, in the unified log-linear notation, the equilibrium restriction links the structural demand matrices to the reduced-form impulse responses through a

linear convolution equation. We state a general recovery result and then develop low-dimensional restrictions that can be taken to the data.

4.1 Impulse responses

We begin from the unrestricted equilibrium condition implied by (1) and linearized market clearing:

$$A_0 p_t - \sum_{s=1}^{S} A_s \mu^p_{t,t+s} = \Lambda_Q z_t + \sum_{s=1}^{S} B_s \mu^d_{t,t+s} - \sum_{s=1}^{S} D_s v^r_{t,t+s} + \xi_t. \tag{20}$$

Suppose an identified demand innovation ε_t induces the impulse-response representations

$$z_t = \bar{z} + \sum_{h=0}^{\infty} \Psi^z_h \varepsilon_{t-h}, \tag{21}$$

$$p_t = \bar{p} + \sum_{h=0}^{\infty} \Upsilon_h \varepsilon_{t-h}, \tag{22}$$

$$\mu^d_{t,t+s} = \bar{\mu}^d_s + \sum_{h=0}^{\infty} \Phi^d_{s,h} \varepsilon_{t-h}, \qquad s = 1, \ldots, S,$$

$$v^r_{t,t+s} = \bar{v}_s + \sum_{h=0}^{\infty} \Omega_{s,h} \varepsilon_{t-h}, \qquad s = 1, \ldots, S. \tag{23}$$

Here Ψ^z_h is the impulse response of outside holdings, Υ_h is the impulse response of prices, $\Phi^d_{s,h}$ is the impulse response of expected dividends at horizon s, and $\Omega_{s,h}$ is the impulse response of the risk state at horizon s. These are population infinite-horizon impulse responses; the recovery results below use only the finite subset of horizons that is observed or estimated in practice. Taking expectations in (22) gives

$$\mu^p_{t,t+s} = \mathbb{E}_t[p_{t+s}] = \bar{p} + \sum_{h=0}^{\infty} \Upsilon_{h+s} \varepsilon_{t-h}.$$

Under Assumption 1, ξ_t contributes only its constant component $\bar{\xi}$ to these projections, so it drops out of the coefficient restrictions. Substituting these impulse responses into (20) and matching coefficients on ε_{t-h} yields

$$A_0 \Upsilon_h - \sum_{s=1}^{S} A_s \Upsilon_{h+s} = \Lambda_Q \Psi_h^z + \sum_{s=1}^{S} B_s \Phi_{s,h}^d - \sum_{s=1}^{S} D_s \Omega_{s,h}, \qquad h \geq 0. \tag{24}$$

Equation (24) is the unrestricted recovery equation. It says that the structural demand parameters must jointly reconcile the observed impulse responses of prices and outside holdings: for each horizon h, the price response Υ_h and the outside-holdings response Ψ_h^z are treated as knowns, and the structural coefficients—current-price elasticity, expected-price elasticity, expected-dividend elasticity, and risk elasticity—are the unknowns. Each horizon contributes an $N \times N$ matrix restriction. Thus horizons $h = 0, \ldots, H$ give $(H+1)N^2$ scalar restrictions. The recovery problem is to choose the structural coefficients so that these dynamic restrictions hold.

4.2 Recovery from unrestricted impulse responses

For a given horizon h, equation (24) is linear in the unknown coefficient matrices. Stacking the vectorized horizon-by-horizon restrictions yields a system of the form

$$m_H = M_H \, g(\theta), \tag{25}$$

where m_H stacks the observed reduced-form impulse responses through horizon H, $g(\theta)$ stacks the structural coefficient matrices implied by the primitive parameter vector $\theta \in \Theta \subset \mathbb{R}^{K_\theta}$, and M_H collects the known coefficient blocks from the vectorized horizon-by-horizon equations.

Appendix G gives one explicit construction of this stacked system and the associated counting formulas.

Equation (25) is best viewed as a bookkeeping device. The unrestricted benchmark takes θ to collect every free element of $(A_0, \ldots, A_S, B_1, \ldots, B_S, D_1, \ldots, D_S)$, so that g is the identity stacking map. Empirical work may instead use a lower-dimensional parameterization: θ contains the economically meaningful primitive parameters, while $g(\theta)$ maps those primitives into the full coefficient system.

To write the system in standard GMM notation, define the stacked moment condition

$$\varphi_H(\theta) \equiv m_H - M_H g(\theta).$$

The true parameter θ_0 satisfies $\varphi_H(\theta_0) = 0$.

Equation (24) contributes one $N \times N$ matrix restriction at each horizon, so stacking horizons $h = 0, \ldots, H$ yields $(H+1)N^2$ scalar equations. In the fully unrestricted benchmark, the parameter vector is very large, especially because each risk block D_s is $N \times m$ with $m = N(N+1)/2$. Unrestricted $(A_0, \ldots, A_S)$ contribute $(S+1)N^2$ parameters, unrestricted $(B_1, \ldots, B_S)$ contribute SN^2, and unrestricted $(D_1, \ldots, D_S)$ contribute $SN^2(N+1)/2$. Thus the unrestricted system has $N^2\{1 + S(N+5)/2\}$ unknowns, so a necessary condition for identification is $H \geq S(N+5)/2$. Additional structure lowers this burden. For instance, if the risk component of demand depends only on each asset's own conditional variance and F factor variances, then each D_s has only $N(1+F)$ free parameters; with unrestricted A_s and B_s blocks, the necessary condition becomes $H \geq 2S + S(1+F)/N$. Appendix G reports the unrestricted count and two simple special cases.

One may also aggregate the cross-sectional restrictions, producing fewer than $N^2(H+1)$ moments. The fully disaggregated system preserves all cross-sectional restrictions, but it is high-

dimensional and potentially noisy in finite samples. Aggregating moments across assets reduces dimensionality, improves numerical stability, and naturally accommodates unbalanced panels by forming moments from all available observations at each date. The cost is that some cross-sectional variation is discarded. The next proposition states the resulting identification condition.

Proposition 3 (Recovery from parameterized impulse responses)**.** *Suppose Assumption 1 holds, the impulse responses* $\{\Psi_h^z, \Upsilon_h\}_{h=0}^{H+S}$, $\{\Phi_{s,h}^d\}_{s=1,\dots,S;\, h=0,\dots,H}$, *and* $\{\Omega_{s,h}\}_{s=1,\dots,S;\, h=0,\dots,H}$ *are observed, and let* θ_0 *denote the true parameter vector. If the moment condition* $\varphi_H(\theta) = 0$ *implies* $\theta = \theta_0$*, then* θ_0 *is identified. If* g *is continuously differentiable, a sufficient condition for local identification at* θ_0 *is*

$$\operatorname{rank}\left(\frac{\partial \varphi_H(\theta_0)}{\partial \theta^\top}\right) = K_\theta.$$

This is the standard identification condition for nonlinear GMM problems, so no proof is given. Proposition 3 is deliberately general. The unrestricted system is the special case in which θ collects every free element of the coefficient matrices. The empirical implementation below does not attempt fully unrestricted recovery; it imposes the return-based restrictions in (6), so that current prices affect demand only through expected returns, and the low-dimensional factor structure introduced in Section 5.

It is important to note that recovery requires a rich time-series structure of shocks and responses. Mathematically, identification becomes easier as the available horizon length H increases, because each additional horizon contributes another block of dynamic restrictions.

This differs from the identification problem emphasized by Binsbergen et al. (2025). Their point is that persistent or predictable instruments generally do *not* identify a ceteris-paribus price elasticity (i.e. A_0 here), because they move the entire future path of resale prices. Our point here is different: once the object of interest is the structural *dynamic* demand system, those same persistent

responses are useful because they generate the impulse-response variation needed to recover the forward-looking coefficients. For recovery, persistent shocks are therefore a feature, not a bug. And even recovering the contemporaneous ceteris-paribus elasticity A_0 would not by itself recover the full dynamic demand system, because counterfactuals in dynamic settings also depend on the continuation coefficients governing responses to future expected returns, dividends, and risk.

The rank condition has the usual GMM interpretation: the impulse responses must generate enough independent variation across horizons and cross-sectional directions to distinguish the primitive coefficient vector from nearby alternatives.

4.3 Recovery with institutional holdings data

Demand system asset pricing is often focused on using institutional-level holdings data, like 13F filings, to recover the demand system at the investor level. We present a model with a representative investor and an empirical implementation that does not use holdings data beyond the construction of the outside-flow measure z_t. However, we briefly trace out how the same recovery method can be used to identify investor-specific structural demand coefficients with investor-level holdings data.

Consider investor i with demand

$$q_{i,t} = -A_{0,i}\, p_t + \sum_{s=1}^{S} A_{s,i}\, \mu^{p}_{t,t+s} + \sum_{s=1}^{S} B_{s,i}\, \mu^{d}_{t,t+s} - \sum_{s=1}^{S} D_{s,i}\, v^{r}_{t,t+s} + \xi_{i,t}. \tag{26}$$

Let the same identified demand shock ε_t induce the additional impulse-response representation

$$q_{i,t} = \bar{q}_i + \sum_{h=0}^{\infty} \Psi^{q_i}_{h}\, \varepsilon_{t-h}.$$

Assume likewise that the investor-specific latent demand shifter $\xi_{i,t}$ is exogenous to ε_t, so its impulse response is zero. Taking impulse responses of (26) with respect to ε_t and rearranging yields the

investor-level recovery equation:

$$A_{0,i}\,\Upsilon_h - \sum_{s=1}^{S} A_{s,i}\,\Upsilon_{h+s} = \sum_{s=1}^{S} B_{s,i}\,\Phi^d_{s,h} - \sum_{s=1}^{S} D_{s,i}\,\Omega_{s,h} - \Psi^{q_i}_h, \qquad h \geq 0. \tag{27}$$

This is structurally identical to (24), with $-\Psi^{q_i}_h$ in place of $\Lambda_Q \Psi^z_h$ on the right-hand side. The price impulse responses Υ_h, $\Phi^d_{s,h}$, and $\Omega_{s,h}$ are aggregate objects estimated from price and dividend data exactly as before; the only additional input is $\Psi^{q_i}_h$, the impulse response of investor i's log holdings to the demand shock. The same GMM system applies, and the same rank condition recovers the investor-specific coefficients $(A_{0,i},\ldots,A_{S,i},B_{1,i},\ldots,B_{S,i},D_{1,i},\ldots,D_{S,i})$. This derivation imposes homogeneous beliefs: the aggregate objects $\mu^d_{t,t+s}$ and $v^r_{t,t+s}$ are assumed common across investors, so the aggregate IRFs $\Phi^d_{s,h}$ and $\Omega_{s,h}$ apply to each investor's demand equation.

Consistency with the aggregate equation follows from market clearing. With I investors, total shares satisfy $\sum_i Q_{i,t} + Z_t = \mathbf{1}_N$. Log-linearizing around $(\bar{Q}_i, \bar{Z})$ gives the multi-investor clearing condition:

$$\sum_i \mathrm{Diag}(\bar{Q}_i)\, q_{i,t} + z_t = 0,$$

which in impulse-response form is $\sum_i \mathrm{Diag}(\bar{Q}_i)\,\Psi^{q_i}_h = -\Psi^z_h$. Premultiplying (27) by $\mathrm{Diag}(\bar{Q}_i)$ and summing over i yields

$$\begin{aligned}
&\Big(\sum_i \mathrm{Diag}(\bar{Q}_i) A_{0,i}\Big)\Upsilon_h - \sum_{s=1}^{S}\Big(\sum_i \mathrm{Diag}(\bar{Q}_i) A_{s,i}\Big)\Upsilon_{h+s} \\
&= \sum_{s=1}^{S}\Big(\sum_i \mathrm{Diag}(\bar{Q}_i) B_{s,i}\Big)\Phi^d_{s,h} - \sum_{s=1}^{S}\Big(\sum_i \mathrm{Diag}(\bar{Q}_i) D_{s,i}\Big)\Omega_{s,h} + \Psi^z_h.
\end{aligned}$$

Premultiplying by $\Lambda_Q = \mathrm{Diag}(\bar{Q})^{-1}$, where $\bar{Q} = \sum_i \bar{Q}_i$, this is exactly (24) with aggregate structural parameters defined as share-weighted averages, e.g. $A_0 \equiv \Lambda_Q \sum_i \mathrm{Diag}(\bar{Q}_i) A_{0,i}$.

4.4 Estimating dynamic causal effects of demand shocks

The recovery theory takes as inputs the impulse responses of outside holdings, prices or returns, expected dividends, and risk to an identified demand innovation. The objects Ψ_h^z, Υ_h, $\Phi_{s,h}^d$, and $\Omega_{s,h}$ in (21)–(23) are therefore dynamic causal effects of the demand shock on the economic quantities that enter the recovery equation. We refer to estimating these impulse responses as the first stage. We refer to mapping them into structural demand coefficients through (24) as the second-stage recovery. This terminology is only meant to separate the estimation of impulse response inputs from the structural recovery step.

Following the external instrument literature on dynamic causal effects, these impulse responses can be estimated either with proxy structural VARs or with IV local projections (LPs) (Jordà, 2005; Mertens and Ravn, 2013; Stock and Watson, 2018). In this terminology, the proxy is an observed variable that is correlated with the unobserved structural demand shock and orthogonal to the other structural shocks. A proxy VAR imposes a finite-dimensional law of motion for the outcome vector and obtains impulse responses by iterating the estimated transition equation, using this proxy to identify the column of the impact matrix associated with the demand shock. IV LPs instead estimate each horizon directly, using the same proxy as an instrument in a horizon-by-horizon projection. The choice between these estimators is conceptually separate from the recovery step. Both target the same structural impulse responses under valid external instrument assumptions. Plagborg-Møller and Wolf (2021) show that, with unrestricted lag structures, LPs and VARs estimate the same population impulse responses. In finite samples, the choice matters because the two estimators impose different restrictions. A VAR can be more efficient when its law of motion is correctly specified, since all horizons are tied together by the same estimated transition equation. LPs impose less cross-horizon structure and are therefore typically more robust to dynamic misspecification, especially at longer horizons.

The identifying restrictions are the standard external instrument restrictions used by proxy VARs and LP-IV, written in the notation of the recovery system. Let ε_t denote the normalized identified demand innovation and let χ_t collect all non-demand structural shocks. The impact response of outside holdings to this innovation is

$$\frac{\partial z_t}{\partial \varepsilon_t} = \Psi_0^z.$$

Relevance requires that this response be nonzero:

$$\Psi_0^z \neq 0,$$

contemporaneous exogeneity with respect to non-demand shocks,

$$\mathbb{E}[\varepsilon_t \chi_t^\top] = 0,$$

and lead-lag exogeneity,

$$\mathbb{E}[\varepsilon_t \chi_{t+j}^\top] = 0 \qquad \text{for all } j \neq 0,$$

after conditioning on the predetermined controls included in the LP or VAR. In addition, ε_t must affect the recovery outcomes only through the identified demand innovation, up to the normalization of the shock. Assumption 1 is the corresponding demand-system exclusion restriction: it rules out a direct projection of the latent demand shifter ξ_t on the identified shock history. Under these restrictions, the projection coefficient of any recovery outcome on the normalized demand innovation is the dynamic causal effect that enters (24).

The factor-model implementation below uses the LP route in the first stage. The stock-level innovation $\varepsilon_{n,t}$ is the external instrument for a stock-level outside-demand shock; week fixed effects absorb common shocks; and the interactions between stock loadings and factor shocks recover the

factor-mediated spillovers. Thus the LPs in (29), (31), and (33) estimate the dynamic causal-effect inputs to the recovery system. The second-stage GMM step is not an alternative identification of the shock; it is the structural inversion that maps those estimated dynamic causal effects into the demand primitives (C_s, D_s). Appendix F gives the formal population conditions for this identification argument.

5 Factor Model Recovery

The recovery result in Section 4 is written for a balanced panel of N assets. That is the right population object, but it is not the right object to take literally to stock-level data. The set of stocks changes over time, some stocks have missing returns or missing flow observations, and a fully unrestricted $N_t \times N_t$ recovery system is too large to estimate with any precision. The usual solution in empirical asset pricing is to work with a low-dimensional factor representation. Thus, in this section, we parametrize our stock-level model using a factor model structure to make subsequent estimation feasible.

At each date t, let N_t be the number of stocks in the estimation sample and let

$$X_t \in \mathbb{R}^{N_t \times M},$$

be an orthonormal loading matrix, so $X_t' X_t = I_M$. We write $x_{n,t}'$ for row n of X_t, so $x_{n,t} \in \mathbb{R}^M$ is stock n's vector of factor loadings at date t. The columns of X_t span the factor space used to summarize the cross-section at date t. The construction can come from characteristics, from principal components, or from an instrumented principal components design in the spirit of Kelly, Pruitt, and Su (2019). The same orthonormal matrix can be interpreted as characteristics, factor

loadings, and factor-portfolio weights, as discussed in Chaudhry and Davis (2026) and Baba Yara et al. (2021). Importantly, M is small relative to N_t.

We use two linear maps throughout. First, for any stock-level vector $y_t \in \mathbb{R}^{N_t}$, its factor projection is $y_t^F = X_t' y_t$. In particular,

$$f_{t+h} = X_t' r_{t+h}, \qquad z_{t+h}^F = X_t' z_{t+h}, \qquad s_t = X_t' \varepsilon_t,$$

where r_{t+h} is the vector of excess returns, z_{t+h} is outside demand, ε_t is the identified stock-level demand shock, and s_t is the corresponding factor shock. Second, for any stock-level response matrix $R_t \in \mathbb{R}^{N_t \times N_t}$, the factor-level compression is $R_t^F = X_t' R_t X_t$. Conversely, for any factor-level matrix $K \in \mathbb{R}^{M \times M}$, the stock-level lift is $X_t K X_t'$. These are only changes of representation: the compression records how a stock-level response acts on factor portfolios, while the lift maps a factor response back to the cross-section of stocks.

The structural return-sensitivity matrix is parameterized as a stock-level lift plus an own-stock component:

$$C_s(X_t) = a_{C,s} I_{N_t} + X_t \Gamma_s^C X_t', \qquad \Gamma_s^C \in \mathbb{R}^{M \times M}. \tag{28}$$

The scalar $a_{C,s}$ is the component that acts on each stock's own return. The low-rank term $X_t \Gamma_s^C X_t'$ captures common substitution through the factor space. Compressing (28) gives the corresponding factor-level matrix

$$C_s^F \equiv a_{C,s} I_M + \Gamma_s^C,$$

since $X_t' C_s(X_t) X_t = a_{C,s} I_M + \Gamma_s^C$. The distinction matters for estimation. A factor-level local projection alone would identify only the composite object $a_{C,s} I_M + \Gamma_s^C$. Estimating the flow and return responses in stock space separately identifies the own component $a_{C,s} I_{N_t}$ because the regression includes the stock's own shock $\varepsilon_{n,t}$ separately from the shock's factor projection s_t.

The outside-demand and return local projections are estimated as stock-week panel regressions. For each horizon h, we pool all available stock-week observations (n,t) in the unbalanced panel. The date-specific intercepts below absorb week fixed effects; equivalently, the outcomes and regressors can be demeaned by week before estimating the common slope coefficients. This panel structure is essential. The own-shock coefficient and the factor-spillover coefficients are identified from cross-sectional variation in which stocks are shocked, which stocks load on the shocked factor directions, and which stocks receive the spillovers. An aggregate time series would collapse these objects into a single response. This is the sense in which the design is related to Haddad et al. (2025a): estimating causal spillovers through a factor structure requires panel data that observe both the shock exposure and the cross-sectional response. We retain only weeks with at least 25 usable stock observations for the relevant horizon.

For outside demand, the horizon-h stock-level local projection is

$$z_{n,t+h} = \alpha^z_{h,t} + \phi^{z\prime}_h x_{n,t} + a^{(h)}_z \varepsilon_{n,t} + \sum_{a=1}^{M}\sum_{b=1}^{M} g^{z,(h)}_{ab} x_{n,a,t} s_{b,t} + u^z_{n,t+h}, \tag{29}$$

where $\alpha^z_{h,t}$ is a date-specific intercept, $\phi^z_h \in \mathbb{R}^M$ controls for the factor loadings of stock n, $a^{(h)}_z$ is the direct own-stock effect of the shock $\varepsilon_{n,t}$, and $u^z_{n,t+h}$ is the projection residual. The interaction term is equivalently $x'_{n,t} G^{(h)}_z s_t$, where

$$G^{(h)}_z = \left(g^{z,(h)}_{ab}\right)^M_{a,b=1}.$$

It allows the shock to spill over across assets through the factor space: the index b records the factor direction of the shock, while the index a records the factor loading of the receiving stock. This is a flexible factor-spillover specification, similar in spirit to the factor-based spillover structure in Haddad et al. (2025a).

Because $s_t = X_t'\varepsilon_t$, the fitted response of the full vector z_{t+h} to a stock-level shock vector ε_t is

$$\widehat{\Psi}_h^z(X_t) = \hat{a}_z^{(h)} I_{N_t} + X_t \widehat{G}_z^{(h)} X_t', \qquad \widehat{\Psi}_h^{z,F} = \hat{a}_z^{(h)} I_M + \widehat{G}_z^{(h)}. \tag{30}$$

The first expression is the stock-level flow-response matrix. The second is its factor-level compression, $X_t'\widehat{\Psi}_h^z(X_t)X_t$.

The return local projection uses the same structure:

$$r_{n,t+h} = \alpha_{h,t}^r + \phi_h^{r\prime} x_{n,t} + a_r^{(h)} \varepsilon_{n,t} + \sum_{a=1}^{M}\sum_{b=1}^{M} g_{ab}^{r,(h)} x_{n,a,t} s_{b,t} + u_{n,t+h}^r, \tag{31}$$

where $r_{n,t+h}$ is the forward excess return of stock n at horizon h, $\alpha_{h,t}^r$ is a date-specific intercept, ϕ_h^r controls for factor loadings, $a_r^{(h)}$ is the direct own-stock return response, $G_r^{(h)} = (g_{ab}^{r,(h)})_{a,b=1}^M$ captures factor-mediated return spillovers, and $u_{n,t+h}^r$ is the residual. The implied stock-level return-response matrix and factor-level compression are

$$\widehat{\Upsilon}_h(X_t) = \hat{a}_r^{(h)} I_{N_t} + X_t \widehat{G}_r^{(h)} X_t', \qquad \widehat{\Upsilon}_h^F = \hat{a}_r^{(h)} I_M + \widehat{G}_r^{(h)}. \tag{32}$$

These are future returns, not cumulative returns. Thus $\widehat{\Upsilon}_1$ is the return from week $t+1$ associated with a shock in week t.

Risk is estimated directly in factor space because a stock-level covariance matrix has $N_t(N_t+1)/2$ entries. Let $f_d \in \mathbb{R}^M$ be the daily factor excess-return vector on trading day d, constructed from the same orthonormal basis used in the weekly LPs. If $\mathcal{W}_{t+h}$ denotes the set of trading days in week $t+h$, the model risk object is the expected realized factor second moment over that week:

$$\Sigma_{t,t+h}^F \equiv \mathbb{E}_t\left[\sum_{d\in\mathcal{W}_{t+h}} f_d f_d'\right] \in \mathbb{S}^M, \qquad \sigma_{t,t+h}^F \equiv \operatorname{vech}(\Sigma_{t,t+h}^F) \in \mathbb{R}^J, \qquad J = \frac{M(M+1)}{2}.$$

The empirical outcome in the risk LP is the realized analogue

$$\widetilde{\Sigma}^F_{t,t+h} \equiv \sum_{d \in \mathcal{W}_{t+h}} f_d f_d', \qquad \widetilde{\sigma}^F_{t,t+h} \equiv \mathrm{vech}(\widetilde{\Sigma}^F_{t,t+h}).$$

This is a weekly realized second moment, not a cumulative return and not a demeaned within-week sample covariance. For $M = 1$ it is the sum of squared daily factor excess returns in the target week; for $M > 1$ it is the full matrix of summed daily factor-return cross-products. For each horizon h, the factor-risk local projection is

$$\widetilde{\sigma}^F_{t,t+h} = \alpha^\Omega_h + \Omega_h s_t + u^\Omega_{t+h}, \qquad \Omega_h \in \mathbb{R}^{J \times M}. \tag{33}$$

Here $\alpha^\Omega_h \in \mathbb{R}^J$ is an intercept, Ω_h maps the M factor shocks into the J factor-covariance states, and u^Ω_{t+h} is the risk-projection residual. The risk LP is estimated over the weekly factor series after the stock-level shocks have been aggregated to $s_t = X_t' \varepsilon_t$. It is a constant-coefficient factor-level LP: the coefficient matrix Ω_h does not vary with X_t. The role of X_t is to construct factor returns and factor shocks from the stock panel, and later to lift the recovered factor object back to stock space.

Using the return-based demand system and market clearing $q_t = -z_t$, the factor-space recovery equation is

$$-\widehat{\Psi}^{z,F}_h = \sum_{s=1}^{S} C^F_s \widehat{\Upsilon}^F_{h+s} - \sum_{s=1}^{S} D^F_s \widehat{\Omega}_{h+s}, \qquad h = 0, \ldots, H,$$

where $D^F_s \in \mathbb{R}^{M \times J}$ maps factor covariance states into factor demand. The corresponding stock-space residual is

$$\mathcal{E}_h(\theta; X_t) \equiv \widehat{\Psi}^z_h(X_t) + \sum_{s=1}^{S} C_s(X_t; \theta) \widehat{\Upsilon}_{h+s}(X_t) - \sum_{s=1}^{S} X_t D^F_s(\theta) \widehat{\Omega}_{h+s} X_t'.$$

The population restriction is $\mathcal{E}_h(\theta_0; X_t) = 0$. In estimation, we use two low-dimensional summaries of this residual:

$$\frac{1}{N_t}\operatorname{tr}\big(\mathcal{E}_h(\theta; X_t)\big), \qquad \frac{M}{N_t}X_t'\mathcal{E}_h(\theta; X_t)X_t.$$

The first moment is one scalar equation: the average stock-level own residual. This scalar moment is important because it is what separately identifies the stock-level identity component $a_{C,s}$. The second object is an $M \times M$ matrix, giving M^2 equations for the full factor-block residual. Thus each horizon contributes $1 + M^2$ moments. We scale the factor block by M/N_t so that the factor equations are expressed in the same stock-level units as the trace moment.[12]

Writing the GMM vector explicitly, define

$$g_h(\theta; X_t) \equiv \begin{bmatrix} N_t^{-1}\operatorname{tr}\big(\mathcal{E}_h(\theta; X_t)\big) \\ \operatorname{vec}\left(\dfrac{M}{N_t}X_t'\mathcal{E}_h(\theta; X_t)X_t\right) \end{bmatrix} \in \mathbb{R}^{1+M^2}, \qquad g(\theta) \equiv \begin{bmatrix} g_0(\theta; X_t) \\ \vdots \\ g_H(\theta; X_t) \end{bmatrix}.$$

Thus, $g(\theta)$ has $(H + 1)(1 + M^2)$ entries, which are the GMM equations used to identify the parameters.

Using (30)–(32), the factor-block moment can be written as

$$\frac{M}{N_t}X_t'\mathcal{E}_h(\theta; X_t)X_t = \frac{M}{N_t}\left[\widehat{\Psi}_h^{z,F} + \sum_{s=1}^{S} C_s^F(\theta)\widehat{\Upsilon}_{h+s}^F - \sum_{s=1}^{S} D_s^F(\theta)\widehat{\Omega}_{h+s}\right].$$

This is the factor-model version of the recovery equation. It preserves the economic content of

[12] Because $X_t'X_t = I_M$, an unscaled factor compression $X_t'\mathcal{E}_h X_t$ is naturally in factor-portfolio units. The trace of a lifted factor matrix satisfies $\operatorname{tr}(X_t K X_t') = \operatorname{tr}(K)$, while the average stock-level diagonal contribution is $\operatorname{tr}(K)/N_t$. Multiplying the full factor block by M/N_t puts a representative factor entry on the same order as an average stock-level response. This normalization affects the relative weighting of moments, not the population zero of the restrictions.

the unrestricted N-asset system but makes the empirical problem feasible in an unbalanced stock panel.

It is useful to be explicit about the number of unknowns. For each horizon s, the return-sensitivity block has one scalar $a_{C,s}$ and an unrestricted $M \times M$ matrix Γ_s^C, for $1 + M^2$ parameters. The risk-loading block has

$$D_s^F \in \mathbb{R}^{M \times J}, \qquad J = \frac{M(M+1)}{2},$$

so it contains $MJ = M^2(M+1)/2$ unrestricted parameters. With S unrestricted structural horizons, the no-risk factor recovery has

$$K_C = S(1 + M^2)$$

unknowns, while the risk-augmented recovery has

$$K_{C,D} = S\left(1 + M^2 + M\frac{M(M+1)}{2}\right)$$

unknowns before any sign or symmetry restrictions are imposed. A necessary order condition is therefore $(H + 1)(1 + M^2) \geq K$, where K is the number of free parameters in the chosen specification. The corresponding local identification condition is the usual GMM rank condition:

$$\operatorname{rank}\left(\frac{\partial g(\theta_0)}{\partial \theta'}\right) = K.$$

Economically, this condition requires the horizon profile of flow, forward-return, and risk responses to move enough independent directions in factor space to distinguish the own-stock identity component, the factor substitution block, and the risk-loading block.

6 Empirical Results

6.1 Data

We implement the recovery using weekly U.S. stock data from February 5, 1993, through December 30, 2022. The outside-demand measure is order-flow imbalance (OFI), obtained from WRDS intraday indicators. OFI classifies trades as buyer- or seller-initiated using the Lee and Ready (1991) algorithm, aggregates signed trading to the stock-week level, and scales by lagged shares outstanding. This normalization puts OFI in the same units as the outside-demand shock in the model: a unit is a fraction of shares outstanding. Li and Lin (2023) provide evidence that OFI is useful for studying price impact at asset-pricing frequencies. We use weekly excess log returns, with the risk-free rate from the Ken French daily factors compounded to the week.

For the five-factor implementation, we also use monthly stock characteristics from the Jensen-Kelly-Pedersen global factor data (Jensen, Kelly, and Pedersen, 2023). Appendix A and Appendix Table A.1 list the 120 characteristics, all of which are not a function of prices or market equity. We use only characteristics that are not a function of prices or market equity. The characteristic timing is ex ante: characteristics are lagged by one month when matched to returns and OFI. Following Kelly et al. (2019) and Kozak, Nagel, and Santosh (2020), we rank each characteristic within month, map the ranks into percentiles in $[0,1]$, and subtract 0.5. Remaining missing characteristic values are assigned the cross-sectional median rank before recentering, so they enter the centered characteristic matrix as zeros. This transformation puts all characteristics on a common bounded scale in $[-0.5,0.5]$.

The identified demand shock is the innovation in weekly OFI. For each stock, we estimate the forecast of OFI using four weekly lags of OFI and four weekly lags of excess returns, with stock fixed effects removed before estimation. The residual is the stock-level shock $\varepsilon_{n,t}$. We winsorize the

shock residuals and use them in the local projections described in Section 5. All return and OFI local projections are pooled stock-week panel regressions with week fixed effects, estimated separately by horizon. The final sample contains 7,402,271 stock-week observations, 17,779 distinct stocks, and 1,561 weeks.

Variable	Mean	SD	P25	P50	P75
Weekly OFI	-0.053	0.696	-0.207	-0.019	0.148
OFI shock	0.001	0.619	-0.167	0.012	0.196
Weekly excess return	0.189	8.666	-3.425	-0.100	3.227

Table 2. Summary statistics for the recovery sample
This table reports stock-week summary statistics in percent. OFI is signed order-flow imbalance scaled by lagged shares outstanding. The OFI shock is the residual from the within-stock forecasting regression used to identify demand innovations.

We use a block bootstrap over weeks to compute standard errors. Each bootstrap draw resamples 12-week blocks, re-estimates the first-stage local projections, and then re-estimates the second-stage recovery system. Reported standard errors for the structural summaries are delta-method standard errors computed from the bootstrap covariance matrix of the primitive second-stage coefficient vector. Unless stated otherwise, tables report point estimates with bootstrap standard errors in parentheses on the row immediately below. Very small entries are reported as one-sided bounds at the displayed precision.

The recovery tables report two second-stage objects. The first is a naive IV benchmark. This benchmark treats the contemporaneous price response as sufficient and ignores all continuation terms. Formally, if demand is static,

$$q_t = \bar{q} - A_0 p_t + \xi_t, \qquad A_s = 0 \text{ for } s \geq 1, \qquad D_s = 0 \text{ for all } s, \tag{34}$$

with no dividend pass-through from the identified shock, then market clearing implies

$$A_0\Upsilon_0 = \Lambda_Q\Psi_0^z, \qquad A_0 = \Lambda_Q\Psi_0^z\Upsilon_0^{-1}.$$

In this ultra-myopic special case, the inverse contemporaneous multiplier is the structural elasticity. Outside this special case, it is only a reduced-form benchmark.

Our main second-stage estimates are one-step estimates. In the factor recovery system, one-step means that current demand loads on next-period expected returns and, when included, next-period risk:

$$C_s^F = 0, \quad D_s^F = 0 \qquad \text{for all } s \geq 2.$$

Thus the estimated objects in the main tables are C_1^F and, in the risk specifications, D_1^F.

6.2 A one-factor own/other basis

We first use the simplest possible factor basis:

$$X_t = x_t = \frac{1}{\sqrt{N_t}}\iota_t, \qquad x_t'x_t = 1.$$

With one factor, the stock-level return-sensitivity matrix is

$$C_s(X_t) = a_{C,s}I_{N_t} + \frac{\Gamma_s^C}{N_t}\iota_t\iota_t'.$$

This is an own/other basis in a particularly transparent form. The diagonal entry is $a_{C,s} + \Gamma_s^C/N_t$, while every off-diagonal entry is Γ_s^C/N_t. Thus $a_{C,s}$ is the own-minus-other component, and Γ_s^C/N_t is the common cross-stock component. The one-factor risk state is the expected realized second

LP object	Intercept	Avg. diag. response	Avg. off-diag. response
Returns LP, $h = 0$	3.089 (0.063)	-0.001 (0.003)	>-0.0001 (<0.0001)
Returns LP, $h = 1$	-0.169 (0.010)	>-0.0001 (0.001)	>-0.0001 (<0.0001)
Flows LP, $h = 0$	1.053 (0.001)	-0.001 (0.001)	>-0.0001 (<0.0001)
Risk LP, $h = 1$	3.060 (0.182)	-26.161 (4.204)	

Table 3. First-stage impulse responses in the one-factor own/other basis
This table reports summaries of the stock-level local projections in the equal-weight one-factor basis. "Intercept" is the stock-level own-shock coefficient $a^{(h)}$ for returns and flows, and the baseline factor-variance intercept for the risk row. Average diagonal and off-diagonal responses summarize the factor interaction contribution. Blank cells are not applicable.

moment of the equal-weight factor return, measured empirically as the sum of squared daily equal-weight factor excess returns within the target week. The risk-loading coefficient is therefore a scalar.

Table 3 reports the first-stage objects. The impact return response is positive, 3.089, while the one-week-ahead forward-return response is negative, -0.169. The contemporaneous OFI response is positive, 1.053. The one-factor risk projection has baseline variance intercept 3.060 and horizon-one variance response -26.161. These signs are the basic ingredients for positive recovery: market clearing requires investors to absorb the outside demand shock, and the next-week return response gives the expected-return incentive that induces them to do so.

Table 4 reports the one-factor recovery. The naive inverse-multiplier benchmark is 0.341. The one-step no-risk specification gives an average diagonal C_1 of 5.520 and an average off-diagonal of -0.0076. Adding the scalar risk channel raises the diagonal to 6.741. In this one-factor basis, the same risk channel also raises the common off-diagonal component to 1.1544 because the single factor loads equally on all stocks.

Table 5 shows that the own-minus-other component $a_{C,1}$ is stable across the no-risk and risk

Specification	Avg. diag. C_1	Avg. off-diag. C_1
Naive IV benchmark	0.341 (0.007)	
One-step, no risk	5.520 (0.397)	-0.00762 (0.00663)
One-step, with risk	6.741 (0.460)	1.15437 (0.39042)

Table 4. Structural recovery in the one-factor own/other basis
This table reports the stock-level average diagonal and average off-diagonal entries of $C_1(X_t)$ for the naive benchmark and the one-step recovery specifications. The naive IV benchmark is the inverse contemporaneous multiplier and is structural only under the ultra-myopic restriction in (34).

Parameter	No risk	With risk	Parameter	No risk	With risk
$a_{C,1}$	5.528 (0.399)	5.515 (0.398)	Γ_1^C	-39.731 (34.564)	6020.001 (2036.031)

Table 5. Primitive C_1 parameters in the one-factor own/other basis
The table reports the primitive coefficients in $C_1(X_t) = a_{C,1} I_{N_t} + \Gamma_1^C \iota_t \iota_t' / N_t$. Bootstrap standard errors are in parentheses. The common-factor coefficient Γ_1^C is scaled by $1/N_t$ in stock space, so its magnitude should be interpreted through the stock-level averages in Table 4.

Specification	Factor-variance D_1^F
One-step, with risk	30.218 (11.199)

Table 6. Risk loading in the one-factor own/other basis
This table reports the one-step risk-loading coefficient. With one factor, the only risk state is the variance of the equal-weight factor return, so the reported coefficient is the scalar D_1^F.

specifications, at about 5.5. The risk channel mainly changes the common factor component Γ_1^C, which is why the one-factor average off-diagonal entry rises when risk is included.

The recovered risk loading in Table 6 is large and positive, 30.218. This is why the risk-augmented one-factor estimate is higher: part of the observed flow response is attributed to changes in future risk rather than expected returns alone.

6.3 Estimating Dynamic Preferences

The one-step specification is myopic in the return-risk representation: demand responds to next-period expected returns and next-period risk, but not directly to longer-horizon objects. We also estimate two simple dynamic restrictions to ask whether the data favor a more persistent structural coefficient sequence. The two-step restriction is

$$C_1^F = C^F, \quad D_1^F = D^F, \qquad C_2^F = \delta C^F, \quad D_2^F = \delta D^F, \qquad C_s^F = D_s^F = 0 \text{ for } s \geq 3, \tag{35}$$

where δ is estimated by minimizing the second-stage GMM criterion. The geometric restriction is

$$C_s^F = \delta^{s-1} C^F, \qquad D_s^F = \delta^{s-1} D^F, \qquad s = 1, \ldots, S. \tag{36}$$

Both restrictions nest the one-step specification at $\delta = 0$.

The data select this boundary. Table 7 reports the estimated decay parameters for the two-step and geometric restrictions. In both the one-factor and five-factor bases, the estimates are $\hat{\delta} = 0$, with and without risk. Since these specifications reproduce the one-step elasticities, we do not report separate elasticity tables for them. Empirically, the best-fitting investor in this class is myopic: demand responds to next-period expected returns and risk, with no additional persistence in the structural coefficient sequence. The last two estimates show this holds for the richer five-factor basis as well, which is introduced below.

6.4 A five-factor PCA basis

The one-factor basis is useful because it is transparent. We next use a richer five-factor basis constructed from the characteristic panel described in Section 6.1. Let Z_t be the $N_t \times 120$ matrix

Basis	Dynamic restriction	No risk	With risk
One-factor	Two-step	0.000 (0.010)	0.000 (0.052)
One-factor	Geometric	0.000 (0.005)	0.000 (0.027)
Five-factor	Two-step	0.000 (< 0.001)	0.000 (0.128)
Five-factor	Geometric	0.000 (< 0.001)	0.000 (0.021)

Table 7. Estimated dynamic-preference decay
This table reports the estimated decay parameter δ for the dynamic restrictions in (35) and (36). Bootstrap standard errors are in parentheses. The one-step specification is nested at $\delta = 0$.

of centered monthly characteristics and let

$$W_t = \begin{bmatrix} \iota_t & Z_t \end{bmatrix}$$

collect the intercept and the 120 characteristics. We right-whiten this matrix each month,

$$U_t = W_t (W_t' W_t)^{-1/2}, \qquad U_t' U_t = I.$$

For each daily return vector r_τ^d whose signal month is t, we form characteristic-managed returns

$$m_\tau = U_t' r_\tau^d.$$

We then run PCA on the pooled time series $\{m_\tau\}$ and keep the first five principal components. If $\widehat{V}_5$ denotes the associated 121×5 loading matrix, the stock-level factor basis is

$$X_t = U_t \widehat{V}_5, \qquad X_t' X_t = I_5.$$

This is the orthonormal version of the Kelly et al. (2019) IPCA construction. With orthonormal instruments, IPCA collapses to a one-step PCA estimator on characteristic-managed portfolios: in the model $r_\tau^d = U_t C f_\tau + e_\tau$ with $C'C = I$, the least-squares factor estimate is $f_\tau = C'U_t' r_\tau^d = C' m_\tau$, so choosing C is equivalent to choosing the leading principal components of $\sum_\tau m_\tau m_\tau'$. Thus the characteristics discipline the cross-sectional loading space, while the final five factors are selected from the covariance structure of the managed returns.

The return and flow local projections are the same stock-level panel equations as before, now with five factor loadings and five factor shocks. The risk state is the full factor realized second-moment matrix. At horizon h, we measure this object by summing daily factor-return outer products within week $t + h$ and storing the result in vech form:

$$\widetilde{\Sigma}^F_{t,t+h} = \sum_{d \in \mathcal{W}_{t+h}} f_d f_d', \qquad \widetilde{\sigma}^F_{t,t+h} = \mathrm{vech}(\widetilde{\Sigma}^F_{t,t+h}).$$

The daily factor returns f_d are excess returns on the five orthonormal factor portfolios. We do not subtract a within-week mean, so this is a realized second moment rather than a sample covariance. Since $M = 5$, $\widetilde{\sigma}^F_{t,t+h}$ has 15 entries, and the first-stage risk response Ω_h is a 15×5 matrix estimated from the weekly factor-level LP. We project each shock-specific factor-covariance response onto the negative semidefinite cone before the second stage. This restriction imposes that the demand shock lowers future factor risk in the estimated direction, consistent with the sign of the reduced-form risk response.

For the second-stage risk loading, we do not impose elementwise nonnegativity on all entries of D_1^F. Such a restriction is not invariant to sign changes in the factor basis: covariance entries change sign when a factor is multiplied by -1. Instead, for each factor-demand row i, we interpret

LP object	Intercept	Avg. diag. response	Avg. off-diag. response
Returns LP, $h = 0$	2.937 (0.055)	13.422 (0.354)	1.501 (0.354)
Returns LP, $h = 1$	-0.184 (0.011)	0.677 (0.270)	0.421 (0.251)
Flows LP, $h = 0$	1.051 (0.001)	0.116 (0.010)	-0.015 (0.006)
Risk LP, $h = 1$	0.880 (0.050)	-7.449 (1.404)	-0.627 (0.207)

Table 8. First-stage impulse responses in the five-factor basis
This table reports summaries of the stock-level return and OFI local projections and the factor-level covariance-risk projection in the five-factor basis. "Intercept" is the stock-level own-shock coefficient $a^{(h)}$ for returns and flows, and the average diagonal baseline covariance intercept for the risk row. The risk response columns summarize the horizon-one projected factor-covariance response, averaging diagonal and off-diagonal entries across shock slices.

the row of D_1^F as a symmetric kernel $\mathcal{B}_i$ satisfying

$$d_i' \operatorname{vech}(\Sigma) = \operatorname{tr}(\mathcal{B}_i \Sigma) \qquad \text{for every } \Sigma \in \mathbb{S}^5, \tag{37}$$

and impose $\mathcal{B}_i \succeq 0$. This row-positive-semidefinite (row-PSD) restriction says that higher covariance risk in any factor direction weakly lowers demand, while allowing individual covariance-coordinate coefficients to be positive or negative depending on the orientation of the factor basis.

The five-factor first-stage responses are similar in the scalar direction and richer in the factor block. The impact return response is 2.937, the one-week-ahead return response is -0.184, and the contemporaneous OFI response is 1.051. The baseline covariance intercept averages 0.880 on the diagonal. The horizon-one factor-risk response is negative on average, with average diagonal -7.449 and average off-diagonal -0.627.

Table 9 reports the five-factor recovery. The naive inverse-multiplier benchmark is 0.358. The one-step no-risk specification gives an average diagonal C_1 of 5.167 and an average off-diagonal of -0.00096. Adding the factor-covariance risk block with the row-PSD restriction raises the diagonal

Specification	Avg. diag. C_1	Avg. off-diag. C_1
Naive IV benchmark	0.358 (0.007)	
One-step, no risk	5.167 (0.308)	-0.00096 (0.00006)
One-step, with risk	6.421 (0.484)	-0.00093 (0.00038)

Table 9. Structural recovery in the five-factor basis
This table reports the stock-level average diagonal and average off-diagonal entries of $C_1(X_t)$ for the naive benchmark and the one-step recovery specifications. The five-factor risk specification uses the row-PSD restriction in (37).

to 6.421, while the average off-diagonal remains essentially zero at −0.00093. The small stock-level cross-price elasticities are consistent with Chaudhry and Davis (2026), who measure substitution across individual stocks using several complementary approaches and find that cross-price effects are tiny relative to own-price effects.

Table 10 reports the full primitive C_1 block. As in the one-factor case, the identity component $a_{C,1}$ accounts for most of the stock-level own effect. The factor block Γ_1^C is economically important for fitting the projected factor moments, but the stock-level averages in Table 9 show that these factor interactions nearly cancel on average off the diagonal.

Table 11 summarizes the risk-loading operator. The average own-variance loading is 3.830, while the average loading on the remaining covariance entries is 2.538. The individual covariance-coordinate signs are not themselves structural because they depend on factor orientation. The row-PSD kernel is the economically meaningful restriction.

Table 12 shows why we use the row-PSD restriction. Constraining only variance columns to be nonnegative raises the with-risk diagonal from 5.167 to 5.458. The row-PSD restriction raises it to 6.421. The difference is economically natural: risk aversion is a statement about variance in any factor direction, not about the sign of a particular covariance coordinate.

Taken together, the two implementations deliver the same message. The naive inverse-multiplier

Parameter	No risk	With risk	Parameter	No risk	With risk
$a_{C,1}$	5.172 (0.308)	6.426 (0.484)	$(\Gamma_1^C)_{3,3}$	-4.241 (0.806)	-3.345 (3.214)
$(\Gamma_1^C)_{1,1}$	-5.379 (0.305)	-7.561 (1.874)	$(\Gamma_1^C)_{3,4}$	4.654 (1.207)	-0.639 (5.481)
$(\Gamma_1^C)_{1,2}$	0.487 (0.291)	2.581 (3.407)	$(\Gamma_1^C)_{3,5}$	-1.050 (2.432)	-3.570 (6.763)
$(\Gamma_1^C)_{1,3}$	-0.153 (0.265)	-0.752 (5.949)	$(\Gamma_1^C)_{4,1}$	-0.358 (0.247)	2.172 (1.074)
$(\Gamma_1^C)_{1,4}$	1.688 (0.453)	9.152 (9.420)	$(\Gamma_1^C)_{4,2}$	2.057 (0.932)	0.972 (1.833)
$(\Gamma_1^C)_{1,5}$	-0.923 (0.954)	-6.992 (9.465)	$(\Gamma_1^C)_{4,3}$	-0.656 (0.971)	-3.587 (2.880)
$(\Gamma_1^C)_{2,1}$	0.770 (0.229)	4.078 (1.224)	$(\Gamma_1^C)_{4,4}$	-5.351 (1.522)	-8.899 (5.115)
$(\Gamma_1^C)_{2,2}$	-6.736 (0.711)	-5.394 (2.699)	$(\Gamma_1^C)_{4,5}$	1.448 (2.721)	2.876 (5.632)
$(\Gamma_1^C)_{2,3}$	-1.927 (0.604)	0.649 (3.226)	$(\Gamma_1^C)_{5,1}$	-0.641 (0.320)	1.665 (1.131)
$(\Gamma_1^C)_{2,4}$	-0.638 (0.985)	-7.959 (5.858)	$(\Gamma_1^C)_{5,2}$	-0.019 (0.973)	-3.874 (1.989)
$(\Gamma_1^C)_{2,5}$	2.816 (2.046)	-6.778 (7.495)	$(\Gamma_1^C)_{5,3}$	-0.574 (0.940)	-2.811 (2.488)
$(\Gamma_1^C)_{3,1}$	-1.271 (0.263)	0.634 (1.267)	$(\Gamma_1^C)_{5,4}$	4.349 (1.448)	-0.370 (4.474)
$(\Gamma_1^C)_{3,2}$	0.837 (0.754)	3.012 (2.407)	$(\Gamma_1^C)_{5,5}$	-3.672 (3.309)	-1.523 (5.451)

Table 10. Primitive C_1 parameters in the five-factor basis
This table reports the primitive coefficients in $C_1(X_t) = a_{C,1} I_{N_t} + X_t \Gamma_1^C X_t'$. Bootstrap standard errors are in parentheses. Parameters are shown for the one-step no-risk specification and the one-step row-PSD risk specification.

Specification	Avg. own-variance D_1	Avg. other-covariance D_1
One-step, with risk	3.830 (3.443)	2.538 (1.152)

Table 11. Risk loading in the five-factor basis
The own-variance column averages entries of D_1^F that load factor i demand on factor i variance. The other-covariance column averages the remaining entries of the 5×15 risk-loading operator.

benchmark is around 0.35. Once we use forward-return and flow impulse responses to recover structural demand, the average own elasticity is around five without risk and between six and seven with the risk channel included. Average off-diagonal elasticities are close to zero in the five-

D restriction	No-risk avg. diag. C_1	With-risk avg. diag. C_1
Variance nonnegative only	5.167 (0.308)	5.458 (0.319)
Row-PSD risk kernel	5.167 (0.308)	6.421 (0.484)

Table 12. Five-factor recovery under alternative risk-loading restrictions
This table compares one-step estimates under two restrictions on D_1^F, holding the first-stage LPs fixed. Bootstrap standard errors are in parentheses. The variance-nonnegative restriction constrains only variance columns of D_1^F to be nonnegative and leaves covariance columns unrestricted. The row-PSD restriction is the baseline five-factor specification.

factor basis, in line with the weak stock-level substitution evidence in Chaudhry and Davis (2026), while the one-factor basis mechanically assigns part of the risk channel to a common cross-stock component. The economically stable conclusion is that the structural own-price elasticity is far larger than the naive inverse multiplier, and that accounting for factor-level risk raises it further. Appendix E gives a scalar numerical illustration of the algebra behind these estimates.

7 Conclusion

This paper studies how to recover and use demand elasticities in dynamic asset markets. Because a price movement today must be accompanied by some path of future expected dividends, returns, or risk, there is no single context-free elasticity. The object recovered by an empirical design depends on the process generating the price movement. This is not a defect of demand-system asset pricing. It is the accounting structure of dynamic asset markets.

We formalize this point by distinguishing structural elasticities from instrumented elasticities. The structural elasticity is the demand response holding fixed the other arguments of the demand system. The instrumented elasticity is the response induced by a particular shock, including the way that shock moves future expected returns and risk. Under linearized market clearing, the inverse

of the instrumented elasticity is the price multiplier for the identifying shock. The multiplier is therefore a valid reduced-form object, but it is not generally a transportable structural parameter.

The recovery result shows how to move from the reduced-form object back to structural demand. The path of flow, return, and risk impulse responses imposes a system of restrictions on the structural coefficients. For stock data, we implement this idea with an orthonormal factor representation that accommodates unbalanced panels and keeps the recovery problem low-dimensional. The stock-level local projections identify the own identity component of demand, while the factor projection captures common substitution and risk channels.

In our data, the quantitative gap is large. The naive inverse-multiplier benchmark is about 0.35. The recovered average own elasticity is about 5.2 without risk. When we include factor-level covariance risk, the recovered own elasticity rises to 6.7 in the one-factor own/other basis and 6.4 in the five-factor basis. The five-factor implementation also finds average off-diagonal elasticities close to zero. Thus the data point to a strong own-price structural elasticity, limited average cross-stock substitution, and an economically meaningful risk channel.

References

Baba Yara, Fahiz, Brian H. Boyer, and Carter Davis, 2021, The limits of factor model spanning, SSRN working paper, Available at SSRN.

Berry, Steven T, and Philip A Haile, 2021, Foundations of demand estimation, in *Handbook of Industrial Organization*, volume 4, 1–62 (Elsevier).

Bianchi, Francesco, Cosmin Ilut, and Hikaru Saijo, 2024, Diagnostic business cycles, *The Review of Economic Studies* 91, 129–162.

Binsbergen, Jules H van, Benjamin David, and Christian C Opp, 2025, How (not) to identify demand elasticities in dynamic asset markets, Working paper.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer, 2018, Diagnostic expectations and credit cycles, *The Journal of Finance* 73, 199–227.

Campbell, John Y, and Luis M Viceira, 2002, *Strategic Asset Allocation: Portfolio Choice for Long-Term Investors* (Oxford University Press).

Chaudhry, Aditya, and Carter Davis, 2026, The origins of the factor zoo: Investors weakly substitute across stocks, Working paper.

Cox, John C, and Chi-fu Huang, 1989, Optimal consumption and portfolio policies when asset prices follow a diffusion process, *Journal of Economic Theory* 49, 33–83.

Davis, Carter, Mahyar Kargar, and Jiacui Li, 2025, Why do portfolio choice models predict inelastic demand?, *Journal of Financial Economics* 172, 104096.

Davis, Carter, Samuli Knüpfer, Jens Soerlie Kvaerner, Bahar Sen Dogan, and Petra Vokata, 2024, Do households matter for asset prices?, Working paper.

Fuchs, William, Satoshi Fukuda, and Daniel Neuhann, 2023, Demand-system asset pricing: Theoretical foundations, Working paper.

Fuchs, William, Satoshi Fukuda, and Daniel Neuhann, 2025, A trilemma for asset demand estimation, Working paper.

Gabaix, Xavier, 2019, Behavioral inattention, in *Handbook of Behavioral Economics: Applications and Foundations 2*, volume 2, 261–343 (Elsevier).

Gabaix, Xavier, 2023, Behavioral macroeconomics via sparse dynamic programming, *Journal of the European Economic Association* 21, 2327–2376.

Gabaix, Xavier, and Ralph SJ Koijen, 2022, In search of the origins of financial fluctuations: The inelastic markets hypothesis, Working paper.

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

Grossman, Sanford J, and Joseph E Stiglitz, 1980, On the impossibility of informationally efficient markets, *The American Economic Review* 70, 393–408.

Haavelmo, Trygve, 1943, The statistical implications of a system of simultaneous equations, *Econometrica* 11, 1–12.

Haddad, Valentin, Zhiguo He, Paul Huebner, Péter Kondor, and Erik Loualiche, 2025a, Causal inference for asset pricing, Working paper.

Haddad, Valentin, Paul Huebner, and Erik Loualiche, 2025b, How competitive is the stock market? theory, evidence from portfolios, and implications for the rise of passive investing, *American Economic Review* 115, 975–1018.

Hansen, Lars Peter, and Thomas J Sargent, 2013, *Recursive Models of Dynamic Linear Economies* (Princeton University Press).

He, Zhiguo, Péter Kondor, and Jessica S Li, 2025, Demand elasticity in dynamic asset pricing, Working paper.

Heckman, James J, and Rodrigo Pinto, 2024, Econometric causality: The central role of thought experiments, *Journal of Econometrics* 243, 105719.

Hurwicz, Leonid, 1962, On the structural form of interdependent systems, in Ernest Nagel, Patrick Suppes, and Alfred Tarski, eds., *Logic, Methodology and Philosophy of Science: Proceedings of the 1960 International Congress*, 232–239 (Stanford University Press, Stanford).

Jensen, Theis Ingerslev, Bryan Kelly, and Lasse Heje Pedersen, 2023, Is there a replication crisis in finance?, *Journal of Finance* 78, 2465–2518.

Jordà, Òscar, 2005, Estimation and inference of impulse responses by local projections, *American Economic Review* 95, 161–182.

Karatzas, Ioannis, John P Lehoczky, and Steven E Shreve, 1987, Optimal portfolio and consumption decisions for a "small investor" on a finite horizon, *SIAM Journal on Control and Optimization* 25, 1557–1586.

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

Koijen, Ralph SJ, Robert J Richmond, and Motohiro Yogo, 2024, Which investors matter for equity valuations and expected returns?, *Review of Economic Studies* 91, 2387–2424.

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

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

Lee, Charles MC, and Mark J Ready, 1991, Inferring trade direction from intraday data, *The Journal of Finance* 46, 733–746.

Li, Jiacui, and Zihan Lin, 2023, Prices are less elastic at more aggregate levels, Working paper.

Ljungqvist, Lars, and Thomas J Sargent, 2018, *Recursive Macroeconomic Theory*, fourth edition (MIT Press).

Mas-Colell, Andreu, Michael D Whinston, and Jerry R Green, 1995, *Microeconomic Theory* (Oxford University Press).

Mertens, Karel, and Morten O Ravn, 2013, The dynamic effects of personal and corporate income tax changes in the United States, *American Economic Review* 103, 1212–1247.

Merton, Robert C, 1969, Lifetime portfolio selection under uncertainty: The continuous-time case, *The Review of Economics and Statistics* 51, 247–257.

Merton, Robert C, 1973, An intertemporal capital asset pricing model, *Econometrica* 41, 867–887.

Plagborg-Møller, Mikkel, and Christian K Wolf, 2021, Local projections and VARs estimate the same impulse responses, *Econometrica* 89, 955–980.

Stock, James H, and Mark W Watson, 2018, Identification and estimation of dynamic causal effects in macroeconomics using external instruments, *The Economic Journal* 128, 917–948.

Tamoni, Andrea, Stanislav Sokolinski, and Yizhang Li, 2024, Which investors drive anomaly returns and how?, Working paper.

Van der Beck, Philippe, 2021, Flow-driven ESG returns, Technical report, Swiss Finance Institute.

Vayanos, Dimitri, and Jean-Luc Vila, 2021, A preferred-habitat model of the term structure of interest rates, *Econometrica* 89, 77–112.

Welch, Ivo, and Amit Goyal, 2008, A comprehensive look at the empirical performance of equity premium prediction, *The Review of Financial Studies* 21, 1455–1508.

Appendix

A Characteristic Definitions

This appendix lists the 120 monthly stock characteristics used to construct the five-factor basis. We use 106 characteristics directly from the Jensen-Kelly-Pedersen global factor data and 14 additional variables constructed from Jensen-Kelly-Pedersen inputs so that these transformed variables are not themselves price-scaled characteristics. All underlying Jensen-Kelly-Pedersen variable names and definitions follow Jensen et al. (2023) and the accompanying Global Factor Data Documentation.

Variable	Description
Jensen-Kelly-Pedersen characteristics used directly	
cowc_gr1a	Change in current operating working capital
oaccruals_at	Operating accruals scaled by assets
oaccruals_ni	Operating accruals scaled by net income
taccruals_at	Total accruals scaled by assets
taccruals_ni	Total accruals scaled by net income
capex_abn	Abnormal corporate investment
debt_gr3	Total debt growth over three years
fnl_gr1a	Change in financial liabilities over one year
ncol_gr1a	Change in non-current operating liabilities over one year
nfna_gr1a	Change in net financial assets over one year
ni_ar1	Lagged net income to assets
noa_at	Net operating assets scaled by total assets
aliq_at	Liquidity scaled by lagged assets
at_gr1	Asset growth over one year
be_gr1a	Change in book equity scaled by assets
capx_gr1	Capital expenditure growth over one year
capx_gr2	Capital expenditure growth over two years
capx_gr3	Capital expenditure growth over three years
coa_gr1a	Change in current operating assets over one year
col_gr1a	Change in current operating liabilities over one year
emp_gr1	Employee growth over one year
inv_gr1	Inventory growth over one year
inv_gr1a	Change in inventory over one year
lnoa_gr1a	Change in long-term net operating assets
mispricing_mgmt	Management-based mispricing
ncoa_gr1a	Change in non-current operating assets over one year
nncoa_gr1a	Change in net non-current operating assets over one year
noa_gr1a	Change in net operating assets over one year
ppeinv_gr1a	Change in property, plant, equipment, and inventories
sale_gr1	Sales growth over one year
sale_gr3	Sales growth over three years
saleq_gr1	Quarterly sales growth

Table A.1. Stock characteristics used in the five-factor recovery basis

Variable	Description
Jensen-Kelly-Pedersen characteristics used directly, continued	
age	Firm age
at_be	Book leverage
bidaskhl_21d	21-day bid-ask high-low spread
cash_at	Cash scaled by assets
ni_ivol	Earnings volatility
rd_sale	Research and development scaled by sales
rd5_at	Research and development capital scaled by assets
tangibility	Asset tangibility
beta_60m	60-month CAPM beta
beta_dimson_21d	21-day Dimson beta
betabab_1260d	Betting-against-beta
betadown_252d	252-day downside beta
earnings_variability	Earnings variability
ivol_capm_21d	21-day CAPM idiosyncratic volatility
ivol_capm_252d	252-day CAPM idiosyncratic volatility
ivol_ff3_21d	21-day Fama-French three-factor idiosyncratic volatility
ivol_hxz4_21d	21-day HXZ four-factor idiosyncratic volatility
ocfq_saleq_std	Volatility of quarterly operating cash flow to sales
rvol_21d	21-day return volatility
turnover_126d	126-day share turnover
zero_trades_21d	21-day zero-trade measure
zero_trades_126d	126-day zero-trade measure
zero_trades_252d	252-day zero-trade measure
dsale_dinv	Change in sales minus change in inventory
dsale_drec	Change in sales minus change in receivables
dsale_dsga	Change in sales minus change in SG&A
niq_at_chg1	Change in quarterly net income scaled by assets
niq_be_chg1	Change in quarterly net income scaled by book equity
niq_su	Quarterly earnings surprise
ocf_at_chg1	Change in operating cash flow scaled by assets
sale_emp_gr1	Sales growth relative to employee growth
saleq_su	Quarterly revenue surprise
tax_gr1a	Change in effective tax rate over one year

Table A.2. Stock characteristics used in the five-factor recovery basis, continued

Variable	Description
Jensen-Kelly-Pedersen characteristics used directly, continued	
dolvol_var_126d	126-day dollar-volume volatility
ebit_bev	EBIT scaled by business enterprise value
ebit_sale	EBIT scaled by sales
f_score	Piotroski F-score
ni_be	Net income scaled by book equity
niq_be	Quarterly net income scaled by book equity
o_score	O-score
ocf_at	Operating cash flow scaled by assets
ope_be	Operating profitability scaled by book equity
ope_bel1	Operating profitability scaled by lagged book equity
turnover_var_126d	126-day turnover volatility
at_turnover	Asset turnover
cop_at	Cash-based operating profitability scaled by assets
cop_atl1	Cash-based operating profitability scaled by lagged assets
dgp_dsale	Change in gross profit minus change in sales
gp_at	Gross profits scaled by assets
gp_atl1	Gross profits scaled by lagged assets
ni_inc8q	Number of consecutive earnings increases
niq_at	Quarterly net income scaled by assets
op_at	Operating profits scaled by assets
op_atl1	Operating profits scaled by lagged assets
opex_at	Operating leverage

Table A.3. Stock characteristics used in the five-factor recovery basis, continued

Variable	Description
Jensen-Kelly-Pedersen characteristics used directly, continued	
qmj	Quality minus junk
qmj_growth	Quality minus junk: growth
qmj_prof	Quality minus junk: profitability
qmj_safety	Quality minus junk: safety
sale_bev	Sales scaled by business enterprise value
corr_1260d	Market correlation
coskew_21d	Coskewness
dbnetis_at	Net debt issuance scaled by assets
lti_gr1a	Change in long-term investments over one year
pi_nix	Taxable income scaled by net income
sti_gr1a	Change in short-term investments over one year
ami_126d	Amihud illiquidity measure
iskew_capm_21d	21-day CAPM idiosyncratic skewness
iskew_ff3_21d	21-day Fama-French three-factor idiosyncratic skewness
iskew_hxz4_21d	21-day HXZ four-factor idiosyncratic skewness
rskew_21d	21-day return skewness
chcsho_12m	12-month change in shares outstanding
eqnetis_at	Equity net issuance scaled by assets
netis_at	Net issuance scaled by assets

Table A.4. Stock characteristics used in the five-factor recovery basis, continued

Variable	Description
Other variables (not as market-equity-scaled ratios as in Jensen et al., 2023)	
`netdebt`	Net debt
`rd`	Research and development expense
`at`	Total assets
`be`	Book equity
`debt`	Total debt
`div12m`	Trailing 12-month dividends
`ebitda`	EBITDA
`eqnpo`	Equity net payout
`eqpo`	Net equity payout
`fcf`	Free cash flow
`ival`	Intrinsic value
`ni`	Net income
`ocf`	Operating cash flow
`sale`	Sales

Table A.5. Additional accounting characteristics used in the five-factor recovery basis

Notes: The first three tables report the 106 characteristics taken directly from the Jensen et al. (2023) global factor data. The final table reports the 14 additional characteristics we construct by recovering underlying accounting quantities from Jensen-Kelly-Pedersen variables that are distributed in market-equity-scaled form. These transformations ensure that the variables in the final table enter the factor-model instrumentation as underlying accounting quantities rather than price-scaled ratios. Characteristics are lagged by one month when matched to returns, and each month we apply the rank-to-percentile transform and recenter at zero.

B Proofs

Proof of Proposition 1. By linearized market clearing (3),

$$q_t = -\Lambda_Q z_t.$$

Differentiating with respect to z_t gives

$$\frac{dq_t}{dz_t^\top} = -\Lambda_Q.$$

By definition, the price multiplier matrix is

$$\mathcal{M}_t(z_t) = \frac{dp_t}{dz_t^\top}.$$

Invertibility of $\mathcal{M}_t(z_t)$ then allows us to substitute into (16):

$$\mathcal{E}_t(z_t) = -\frac{dq_t}{dz_t^\top}\mathcal{M}_t(z_t)^{-1} = \Lambda_Q \mathcal{M}_t(z_t)^{-1}.$$

This proves (17).

Now define the aggregate demand elasticity matrix

$$\tilde{\mathcal{E}}_t(z_t) \equiv \Lambda_Q^{-1}\mathcal{E}_t(z_t).$$

Using (17),

$$\tilde{\mathcal{E}}_t(z_t) = \Lambda_Q^{-1}\Lambda_Q \mathcal{M}_t(z_t)^{-1} = \mathcal{M}_t(z_t)^{-1},$$

which proves (18). □

Proof of Proposition 2. Total differentiation of (1) with respect to z_t gives

$$\frac{dq_t}{dz_t^\top} = A_0\left(-\frac{dp_t}{dz_t^\top}\right) + \sum_{s=1}^{S} A_s \frac{d\mu^p_{t,t+s}}{dz_t^\top} + \sum_{s=1}^{S} B_s \frac{d\mu^d_{t,t+s}}{dz_t^\top} - \sum_{s=1}^{S} D_s \frac{dv^r_{t,t+s}}{dz_t^\top} + \frac{d\xi_t}{dz_t^\top}.$$

Assumption 1 sets the last term to zero. Therefore,

$$\frac{dq_t}{dz_t^\top} = A_0\left(-\frac{dp_t}{dz_t^\top}\right) + \sum_{s=1}^{S} A_s \frac{d\mu^p_{t,t+s}}{dz_t^\top} + \sum_{s=1}^{S} B_s \frac{d\mu^d_{t,t+s}}{dz_t^\top} - \sum_{s=1}^{S} D_s \frac{dv^r_{t,t+s}}{dz_t^\top}.$$

Premultiplying by a minus sign and postmultiplying by $\mathcal{M}_t(z_t)^{-1} = (dp_t/dz_t^\top)^{-1}$ as in (16) gives

$$\mathcal{E}_t(z_t) = A_0 - \sum_{s=1}^{S} A_s \frac{d\mu^p_{t,t+s}}{dz_t^\top}\left(\frac{dp_t}{dz_t^\top}\right)^{-1} - \sum_{s=1}^{S} B_s \frac{d\mu^d_{t,t+s}}{dz_t^\top}\left(\frac{dp_t}{dz_t^\top}\right)^{-1} + \sum_{s=1}^{S} D_s \frac{dv^r_{t,t+s}}{dz_t^\top}\left(\frac{dp_t}{dz_t^\top}\right)^{-1},$$

which is (19) by the definition of the shock-specific pass-through matrices. □

C Integrability and Demand Coefficients

The integrability theorem (Proposition 3.H.1 in Mas-Colell et al., 1995) states that a continuously differentiable demand function can be rationalized by a continuous, locally nonsatiated, and strictly quasiconcave utility function if and only if: (i) the budget constraint holds with equality (Walras' law), (ii) demand is homogeneous of degree zero in prices and wealth, and (iii) the Slutsky substitution matrix is symmetric and negative semi-definite. The symmetry of the Slutsky matrix is the economically substantive condition. It requires that the compensated cross-price effects be equal: the compensated effect on the demand for asset i of a change in the price of asset j must equal the compensated effect on asset j of a change in the price of asset i.

In our setting, the demand arguments are current prices p_t and the sequences of conditional expectations $(\mu^p_{t,t+s}, \mu^d_{t,t+s}, v^r_{t,t+s})$. The Jacobian of q_t with respect to these arguments is composed of the coefficient matrices: $-A_0$ is the own-price derivative, A_s is the derivative with respect to $\mu^p_{t,t+s}$, B_s is the derivative with respect to $\mu^d_{t,t+s}$, and $-D_s$ is the derivative with respect to $v^r_{t,t+s}$. The integrability conditions therefore impose cross-equation restrictions across these matrices. For instance, if demand is generated by myopic mean-variance preferences with a symmetric covariance matrix Σ, then A_0 inherits the symmetry and positive definiteness of Σ^{-1}. More generally, symmetry of the compensated substitution effects constrains the relationship between the own-price matrix A_0 and the intertemporal price coefficients A_s.

Example: CARA/mean–variance demand. To make the integrability conditions concrete, consider the myopic CARA investor from Section 2.3. The investor maximizes $\mathbb{E}_t[-\exp(-\gamma W_{t+1})]$ subject to $W_{t+1} = R^f(W_t - P_t^\top Q_t) + (P_{t+1} + \mathcal{D}_{t+1})^\top Q_t$, where $\gamma > 0$ is the coefficient of absolute risk aversion. Under conditional normality of returns with covariance $\Sigma \succ 0$, the first-order condition

yields

$$Q_t = \frac{1}{\gamma}\,\Sigma^{-1}\big(\mathbb{E}_t[P_{t+1} + \mathcal{D}_{t+1}] - R^f\,P_t\big).$$

We verify the integrability conditions for this demand and trace the restrictions they impose on the coefficient matrices.

Slutsky symmetry and negative semi-definiteness. Under CARA preferences demand does not depend on current wealth W_t, so the income effect vanishes: $\partial Q_t/\partial W_t = 0$. The Slutsky substitution matrix therefore coincides with the Marshallian (uncompensated) price derivative:

$$S \;\equiv\; \frac{\partial Q_t}{\partial P_t^\top} + \frac{\partial Q_t}{\partial W_t}\,Q_t^\top \;=\; \frac{\partial Q_t}{\partial P_t^\top} \;=\; -\frac{R^f}{\gamma}\,\Sigma^{-1}.$$

Because Σ is symmetric and positive definite, so is Σ^{-1}. Therefore $S = -(R^f/\gamma)\,\Sigma^{-1}$ is (i) symmetric and (ii) negative definite, verifying the two economically substantive integrability conditions. Symmetry says that the compensated effect on demand for asset i of a change in the price of asset j equals the compensated effect on asset j of a change in the price of asset i. This is a testable restriction whenever $N \geq 2$.

Cross-equation restrictions on the log-linear coefficients. Log-linearizing the demand around steady state $(\bar{Q}, \bar{P}, \bar{\mathcal{D}})$ yields the coefficient matrices:

$$A_0 = \Lambda_Q\,\frac{R^f}{\gamma}\,\Sigma^{-1}\,\mathrm{Diag}(\bar{P}), \qquad A_1 = \beta\,A_0, \qquad B_1 = \Lambda_Q\,\frac{1}{\gamma}\,\Sigma^{-1}\,\mathrm{Diag}(\bar{\mathcal{D}}),$$

where $\Lambda_Q \equiv \mathrm{Diag}(\bar{Q})^{-1}$ and $\beta = 1/R^f$. All higher-order coefficients vanish: $A_s = B_s = D_s = 0$ for $s \geq 2$. These coefficients are not free: the CARA utility imposes several cross-equation restrictions.

First, the relation $A_1 = \beta\,A_0$ ties the current-price and expected-future-price coefficients together. Because the investor cares about prices only through one-period expected excess payoffs $\mathbb{E}_t[P_{t+1} + \mathcal{D}_{t+1}] - R^f P_t$, a unit increase in p_t with an offsetting R^f-unit increase in $\mathbb{E}_t[p_{t+1}]$ has no effect on demand.

Second, the rescaled Jacobian $\tilde{A}_0 \equiv \Lambda_Q^{-1}\,A_0\,\mathrm{Diag}(\bar{P})^{-1} = (R^f/\gamma)\,\Sigma^{-1}$ must be symmetric. This is the Slutsky symmetry condition expressed in the log-linear coefficients, and it holds because Σ^{-1} is symmetric. When A_0 is estimated freely with N^2 parameters, the $N(N-1)/2$ symmetry conditions $(\tilde{A}_0)_{ij} = (\tilde{A}_0)_{ji}$ for $i \neq j$ are overidentifying restrictions that can be tested empirically.

Third, the vanishing of all coefficients beyond $s = 1$ is itself a restriction. A myopic investor

has no hedging motive and faces no adjustment costs, so expectations beyond the next period do not enter demand.

Recovering preferences from demand coefficients. The integrability theorem says not only that the conditions above are necessary, but that they are sufficient to recover the generating preferences. For the CARA case the recovery is explicit. Given A_0 and the observables $(\bar{Q},\bar{P}, R^f)$, one obtains

$$\frac{1}{\gamma}\,\Sigma^{-1} \;=\; \frac{1}{R^f}\,\Lambda_Q^{-1}\,A_0\,\mathrm{Diag}(\bar{P})^{-1},$$

which identifies the composite object $\gamma^{-1}\Sigma^{-1}$. If the return covariance Σ is separately observable from return data, inverting gives the risk-aversion parameter γ directly. One may also use B_1 to form a second expression, $\gamma^{-1}\Sigma^{-1} = \Lambda_Q^{-1}\,B_1\,\mathrm{Diag}(\bar{\mathcal{D}})^{-1}$, and test whether the two routes to $\gamma^{-1}\Sigma^{-1}$ agree. Rejection would indicate a departure from the CARA specification.

The example illustrates how integrability turns observed demand sensitivities into statements about preferences. Symmetry of the Slutsky matrix disciplines the shape of the investor's risk/return trade-off. The restriction $A_1 = \beta\,A_0$ pins down the discount rate. And the fact that all higher-order coefficients vanish tells us the investor's planning horizon is one period.

D Sequential and Recursive Equilibrium Representations

This appendix expands on the discussion in Section 2.2 regarding the sequential and recursive representations of equilibria and the sense in which the sequential formulation is more general.

Two representations of complete-market equilibria. A classical result in dynamic economic theory is that, under complete markets, competitive equilibria can be characterized in two ways (Ljungqvist and Sargent, 2018, Chapter 8). In the Arrow–Debreu formulation, all agents meet at date 0 and trade a complete set of state-contingent claims, one for every date-event pair (t,s^t). Each agent solves a single lifetime problem subject to a single budget constraint. In the *sequential* formulation, agents trade period by period in spot markets, choosing portfolios of one-period Arrow securities at each date t and history s^t. Each agent solves a dynamic programming problem, and equilibrium requires market clearing at every date-event pair.

Under complete markets the two formulations deliver identical allocations and prices. The Arrow–Debreu formulation is conceptually clean but requires specifying the full probability tree

up front; the sequential formulation is more natural for dynamic models and connects directly to recursive methods.

The sequential representation in detail. The demand system (1) is written in sequential form. At each date t, the investor conditions on the filtration generated by ν^t and forms expectations about the entire future path of prices $\{\mu^p_{t,t+s}\}_{s=1}^S$, dividends $\{\mu^d_{t,t+s}\}_{s=1}^S$, and return risk $\{v^r_{t,t+s}\}_{s=1}^S$. The demand system is therefore a map from the full sequence of conditional expectations to current holdings.

This representation has two important features. First, the entire sequence of expected prices, dividends, and volatilities shows up explicitly as arguments of the demand function. There is no assumption that these sequences are generated by a low-dimensional state vector; they are treated as free inputs. Second, the equilibrium condition (4) directly links the structural demand coefficients $\{A_s\}$ to the impulse responses of equilibrium prices through the convolution equation (24). This equation must hold at every horizon h and provides the basis for recovering the structural demand matrices from observed impulse responses.

The recursive competitive equilibrium alternative. An alternative approach, when applicable, is to cast the equilibrium as a *recursive competitive equilibrium* (RCE). In the RCE formulation, one assumes that the relevant state of the economy at date t can be summarized by a finite-dimensional Markov state vector s_t. Value functions and policy functions are then expressed as functions of s_t alone, and the equilibrium is characterized as a fixed point in the space of policy and pricing functions.

As an example of the recursive approach applied to the study of demand elasticities in dynamic settings, Binsbergen et al. (2025) study a continuous-time economy with CRRA investors and a risk-free asset plus J risky assets. Absent exogenous price-shifter shocks, asset prices follow diffusions with aggregate and idiosyncratic Brownian components. A Markov "shifter state" $s_t \in \Omega$ drives exogenous log price shifts $\beta_j s_t$ (interpreted as arising from noise traders, central banks, or share issuances), so the actual trading price of asset j is $P_{j,t} = e^{\beta_j s_t} \tilde{P}_{j,t}$. Events arrive at Poisson rate λ; at each event date the shifter state transitions according to $q(s'|s)$, agents may consume and be replaced, and portfolios are rebalanced. The value function takes the CRRA form $V(W,s) = u(W)\, A(s)$, and optimal portfolio shares $\theta^*(s)$ solve the HJB equation jointly with the function $A(s)$.

The key insight of Binsbergen et al. (2025) is that the shifter state s_t alters the entire future path of resale prices, so a shock that shifts the contemporaneous price while also moving future prices need not identify a structural price elasticity. They show that, under continuous trading, true own-price elasticities are infinite (Proposition 1 of their paper), while discrete trading frictions produce finite elasticities that are governed by the trading frequency and the resolution probability of the price shift. The estimated "shifter-process elasticity" is specific to the dynamics of the particular shock process and therefore does not represent a structural object.

The RCE approach has clear advantages: once the state vector and functional forms are specified, the model can be solved numerically and used for counterfactual analysis. However, it requires the researcher to commit to a particular set of state variables and to specify how they enter preferences and technology. The choice of state vector effectively determines the model.

Why the sequential representation. The sequential representation is strictly more general than the recursive alternative, and this generality is central to our approach.

First, the sequential demand system (1) does not require specifying or identifying the state variables that drive expectations. The structural coefficients $\{A_s, B_s, D_s\}$ characterize how demand responds to expectations at each horizon, regardless of which underlying states generate those expectations. Any RCE with a well-defined demand function is a special case of the sequential formulation, but the converse is not true: the sequential form accommodates settings where no finite Markov state suffices. This is what makes the structural elasticities transportable across shock environments. They are properties of the investor's demand function, not of a particular equilibrium model.

Second, the sequential representation connects directly to the impulse-response objects that are estimable from data. The convolution equation (24) links the structural demand matrices to the reduced-form impulse responses of prices and flows. Because the sequential form keeps the entire expected path as arguments, the dynamic structure of the impulse responses can be exploited for identification without assuming a particular law of motion for the state.

Third, nothing is lost by working at this level of generality. Any recursive model that satisfies the conditions for the Ljungqvist–Sargent equivalence admits an equivalent sequential representation in which the demand coefficients $\{A_s, B_s, D_s\}$ are well-defined. The structural elasticities recovered from the sequential representation are therefore the same objects that would emerge from a correctly specified recursive model.

E Numerical Illustration of Return-Based Recovery

This appendix gives a scalar version of the one-step recovery calculation. The purpose is only to make the sign convention and the role of the return impulse response transparent. The empirical recovery in Section 6 uses the full factor GMM system, all included horizons, and both the trace and factor-block moments.

Suppress factor notation and average the stock-level diagonal restrictions. In the one-step specification without risk, the horizon-zero recovery equation is

$$0 = \bar{\Psi}_0^z + C_1 \bar{\Upsilon}_1, \qquad C_1 = -\frac{\bar{\Psi}_0^z}{\bar{\Upsilon}_1}.$$

Here $\bar{\Psi}_0^z$ is the average contemporaneous outside-demand response, and $\bar{\Upsilon}_1$ is the forward return from week $t+1$ associated with a shock in week t. It is not the cumulative return through week $t+1$. This distinction is central. The impact return is the contemporaneous price response to the shock; the one-step structural coefficient is identified from the next-period expected return that investors anticipate when choosing demand at date t.

The five-factor first-stage estimates in Table 8 give

$$\bar{\Psi}_0^z = 1.051, \qquad \bar{\Upsilon}_1 = -0.184.$$

The scalar calculation is therefore

$$C_1 \approx -\frac{1.051}{-0.184} = 5.71.$$

This is close to the full five-factor no-risk estimate of 5.167 in Table 9. The two numbers need not be identical because the full estimate uses all moments in the factor GMM system rather than only the single averaged diagonal equation. The same calculation in the one-factor basis uses $\bar{\Psi}_0^z = 1.053$ and $\bar{\Upsilon}_1 = -0.169$, giving

$$-\frac{1.053}{-0.169} = 6.23,$$

again matching the scale of the full one-factor no-risk estimate of 5.520 in Table 4.

With risk, the scalar one-step equation becomes

$$0 = \bar{\Psi}_0^z + C_1 \bar{\Upsilon}_1 - \overline{D_1^F \Omega_1}, \qquad C_1 = \frac{\overline{D_1^F \Omega_1} - \bar{\Psi}_0^z}{\bar{\Upsilon}_1}. \tag{A.1}$$

In the five-factor estimates, the projected factor-risk response is negative on average and the row-PSD restriction makes the economically relevant risk loading positive. Thus $\overline{D_1^F \Omega_1}$ is negative, so the numerator in (A.1) is more negative than in the no-risk calculation. Since $\bar{\Upsilon}_1 < 0$, including risk raises the recovered value of C_1. This is the basic algebra behind the increase from 5.167 without risk to 6.421 with risk in Table 9, and from 5.520 to 6.741 in Table 4.

F Identification of Dynamic Causal Effects with External Instruments

This appendix states the population conditions under which the impulse-response inputs used in the recovery system are identified as dynamic causal effects of a demand shock. The notation follows Stock and Watson (2018) but is specialized to the demand-system setting of the paper.

F.1 Population objects

Let y_t collect the recovery outcomes of interest. In the unrestricted notation,

$$y_t \equiv \left(z_t^\top, p_t^\top, \{\mu_{t,t+s}^{d\,\top}\}_{s=1}^S, \{v_{t,t+s}^{r\,\top}\}_{s=1}^S \right)^\top.$$

In the return-based empirical implementation, p_t is replaced by forward excess returns and the risk state is the factor realized second moment. Suppose the recovery outcomes admit the structural moving-average representation

$$y_t = \bar{y} + \sum_{\ell=0}^{\infty} \Theta_\ell^{\varepsilon} \varepsilon_{t-\ell} + \sum_{\ell=0}^{\infty} \Theta_\ell^{\chi} \chi_{t-\ell}. \tag{A.2}$$

The scalar ε_t is the identified outside-demand innovation used throughout the recovery section, and χ_t collects all other structural shocks. The dynamic causal effect of a one-unit demand innovation

on the recovery outcomes at horizon h is

$$\mathrm{IRF}_h^d \equiv \frac{\partial y_{t+h}}{\partial \varepsilon_t} = \Theta_h^{\varepsilon}. \tag{A.3}$$

The components of (A.3) are the population objects denoted by Ψ_h^z, Υ_h, $\Phi_{s,h}^d$, and $\Omega_{s,h}$ in Section 4.

F.2 External-instrument conditions

Identification requires the following population restrictions on the identified innovation ε_t. In the panel implementation, the corresponding object is the stock-level innovation $\varepsilon_{n,t}$ after residualizing with respect to the controls in the LP.

Assumption 2 (External-instrument validity)**.** *The identified demand innovation ε_t satisfies:*

(i) Relevance: $\Psi_0^z \neq 0$*, so the innovation moves outside demand on impact.*

(ii) Innovation normalization: $\mathbb{E}[\varepsilon_t] = 0$*,* $\mathbb{E}[\varepsilon_t^2] > 0$*, and* $\mathbb{E}[\varepsilon_t \varepsilon_{t+j}] = 0$ *for every* $j \neq 0$*.*

(iii) Contemporaneous exclusion: $\mathbb{E}[\varepsilon_t \chi_t^\top] = 0$*.*

(iv) Lead-lag exclusion: $\mathbb{E}[\varepsilon_t \chi_{t+j}^\top] = 0$ *for every* $j \neq 0$*.*

(v) Predetermined controls: *when the projection includes controls w_t, the preceding orthogonality conditions hold for the residualized innovation $\tilde{\varepsilon}_t = \varepsilon_t - \mathbb{E}[\varepsilon_t \mid w_t]$ and the residualized outcome.*

(vi) Demand-system exclusion: *the latent demand shifter has no projection on the instrumented demand-shock history, so $\Psi_h^{\xi} = 0$ for all $h \geq 0$ in Assumption 1.*

The first four restrictions are the standard external-instrument assumptions, written after normalizing the observed proxy into the identified innovation ε_t. Relevance says the innovation moves outside demand. Contemporaneous exclusion rules out correlation with non-demand structural shocks at date t. Lead-lag exclusion rules out anticipation and delayed correlation with non-demand shocks at other dates. The fifth condition is the controlled version used in empirical LPs and VARs. The final condition is specific to the demand-system recovery: ε_t may move equilibrium prices, returns, and risk through the identified demand innovation, but it cannot directly move the unobserved elastic-investor demand shifter ξ_t.

F.3 IV LPs

For a scalar outcome component $y_{i,t+h}$, define the population LP coefficient

$$\beta_{i,h}^{LP} \equiv \frac{\mathbb{E}[y_{i,t+h}\tilde{\varepsilon}_t]}{\mathbb{E}[\tilde{\varepsilon}_t^2]},$$

where $\tilde{\varepsilon}_t = \varepsilon_t$ if no predetermined controls are included. Substituting the moving-average representation (A.2) gives

$$\mathbb{E}[y_{i,t+h}\tilde{\varepsilon}_t] = \sum_{\ell=0}^{\infty} e_i^\top \Theta_\ell^\varepsilon \mathbb{E}[\varepsilon_{t+h-\ell}\tilde{\varepsilon}_t] + \sum_{\ell=0}^{\infty} e_i^\top \Theta_\ell^\chi \mathbb{E}[\chi_{t+h-\ell}\tilde{\varepsilon}_t].$$

By Assumption 2, every term is zero except the demand-innovation term with $\ell = h$. Therefore

$$\beta_{i,h}^{LP} = e_i^\top \Theta_h^\varepsilon.$$

Thus LP-IV identifies the structural impulse response at each horizon in the units of the identified innovation ε_t. If a raw external proxy is used instead, its normalization only rescales all impulse responses by a common constant. The recovery equation is invariant to this common shock scaling because the flow, return, dividend, and risk responses all refer to the same normalized innovation.

In practice, the LP also includes lagged outcomes or other predetermined variables. These controls are not part of the causal object; they residualize the innovation and outcome so that Assumption 2 holds conditionally. Under the same population restrictions, adding sufficient lag controls leaves the target impulse response unchanged, as in Stock and Watson (2018) and Plagborg-Møller and Wolf (2021).

F.4 Proxy structural VAR

Alternatively, suppose the recovery outcomes are represented by a finite-order reduced-form VAR,

$$y_t = \sum_{p=1}^{P} A_p y_{t-p} + u_t, \qquad u_t = b_\varepsilon \varepsilon_t + B_\chi \chi_t,$$

where b_ε is the impact effect of the identified demand innovation on the VAR innovations. Under Assumption 2,

$$\mathbb{E}[\varepsilon_t u_t] = \mathbb{E}[\varepsilon_t^2]\, b_\varepsilon.$$

Hence the covariance between the identified innovation and the reduced-form VAR residual identifies the demand-shock impact vector after dividing by $\mathbb{E}[\varepsilon_t^2]$. Iterating the VAR then gives

$$\mathrm{IRF}_h^{VAR} = J\mathcal{A}^h J^\top b_\varepsilon,$$

where $\mathcal{A}$ is the companion matrix and J selects the contemporaneous block. This is the proxy-SVAR estimator used in the external-instrument literature.

Plagborg-Møller and Wolf (2021) show that, with unrestricted lag structures, VARs and LPs estimate the same population impulse responses for any identification scheme, including external instruments. With a finite lag order P, a correctly specified VAR can be more efficient because it imposes a law of motion, while LPs estimate each horizon directly and are less sensitive to dynamic misspecification. This tradeoff affects the precision and robustness of the impulse-response inputs, not the definition of the recovery equation.

F.5 Panel implementation

The empirical design applies the same external-instrument logic in an unbalanced stock panel. For an outcome $y_{n,t+h}$ equal to outside demand or returns, the population version of the stock-level LP is

$$y_{n,t+h} = \alpha_{h,t}^y + \phi_h^{y\top} x_{n,t} + a_y^{(h)} \varepsilon_{n,t} + x_{n,t}^\top G_y^{(h)} s_t + u_{n,t+h}^y, \qquad s_t = X_t^\top \varepsilon_t.$$

This specification imposes that the stock-level response matrix lies in the factor-plus-own class

$$\mathrm{IRF}_h^y(X_t) = a_y^{(h)} I_{N_t} + X_t G_y^{(h)} X_t^\top.$$

The date fixed effect $\alpha_{h,t}^y$ absorbs aggregate shocks common to all stocks in week t, the loading control removes systematic differences associated with predetermined factor exposures, the own coefficient identifies the direct stock-level demand-shock effect, and $G_y^{(h)}$ identifies factor-mediated spillovers. The factor-risk projection (33) is the corresponding factor-level LP applied to the realized second-moment outcome after stock-level demand innovations have been aggregated to s_t.

The panel analogue of Assumption 2 requires the residualized stock-level innovation $\varepsilon_{n,t}$ to be relevant for stock n's outside-demand shock and orthogonal, after week fixed effects and predetermined loading controls, to non-demand shocks affecting all recovery outcomes at all leads and lags. It also requires that the factor-plus-own response class is correctly specified for the dynamic causal effects used in recovery. If the true response matrix has components outside this class, the LPs identify the projection of the dynamic causal effects onto the factor-plus-own space, and the second-stage GMM recovers the demand coefficients for that projected system.

Finally, identification of the demand coefficients requires more than identification of the dynamic causal effects. The external-instrument assumptions identify the first-stage impulse-response inputs. The second-stage recovery additionally requires the rank condition in Proposition 3, or its factor-model analogue in Section 5: the estimated paths of flow, return, dividend, and risk responses must contain enough independent horizon and cross-sectional variation to distinguish the primitive demand coefficients.

G Stacked Recovery System

For a given horizon h, equation (24) is linear in the unknown coefficient matrices. Applying $\operatorname{vec}(AB) = (B^\top \otimes I)\operatorname{vec}(A)$ gives

$$\begin{aligned}\operatorname{vec}(\Lambda_Q \Psi^z_h) = (\Upsilon_h^\top \otimes I)\operatorname{vec}(A_0) - \sum_{s=1}^{S}(\Upsilon_{h+s}^\top \otimes I)\operatorname{vec}(A_s) \\ - \sum_{s=1}^{S}\left((\Phi^d_{s,h})^\top \otimes I\right)\operatorname{vec}(B_s) - \sum_{s=1}^{S}\left((\Omega_{s,h})^\top \otimes I\right)\operatorname{vec}(D_s).\end{aligned} \tag{A.4}$$

Let $\theta \in \Theta \subset \mathbb{R}^{K_\theta}$ denote the primitive parameter vector. We write $A_s(\theta)$ for example to denote A_s as a function of the parameters. The vector $g(\theta)$ maps the parameters to the vectorized coefficient matrices, but if all matrices are unrestricted, then $\theta = g(\theta)$, and thus the dimensionality of θ is simply the sum of the number of elements in all A_s, B_s, and D_s matrices. Define

$$g_A(\theta) \equiv \left(\operatorname{vec}(A_0(\theta))^\top, \ldots, \operatorname{vec}(A_S(\theta))^\top\right)^\top, \qquad g_B(\theta) \equiv \left(\operatorname{vec}(B_1(\theta))^\top, \ldots, \operatorname{vec}(B_S(\theta))^\top\right)^\top,$$

$$g_D(\theta) \equiv \left(\operatorname{vec}(D_1(\theta))^\top, \ldots, \operatorname{vec}(D_S(\theta))^\top\right)^\top, \qquad g(\theta) \equiv \left(g_A(\theta)^\top, g_B(\theta)^\top, g_D(\theta)^\top\right)^\top.$$

Also define

$$m_H \equiv \left(\text{vec}(\Lambda_Q \Psi_0^z)^\top, \ldots, \text{vec}(\Lambda_Q \Psi_H^z)^\top\right)^\top, \qquad M_H \equiv \begin{bmatrix} M_0 \\ M_1 \\ \vdots \\ M_H \end{bmatrix},$$

where the horizon-h block is

$$M_h \equiv \begin{bmatrix} M_h^A & M_h^B & M_h^D \end{bmatrix},$$

with

$$M_h^A \equiv \left[\Upsilon_h^\top \otimes I, \; -\Upsilon_{h+1}^\top \otimes I, \; \ldots, \; -\Upsilon_{h+S}^\top \otimes I\right],$$

$$M_h^B \equiv \left[-(\Phi_{1,h}^d)^\top \otimes I, \; \ldots, \; -(\Phi_{S,h}^d)^\top \otimes I\right], \qquad M_h^D \equiv \left[-(\Omega_{1,h})^\top \otimes I, \; \ldots, \; -(\Omega_{S,h})^\top \otimes I\right].$$

Stacking (A.4) across horizons $h = 0, \ldots, H$ yields equation (25) in the main text.

The unrestricted benchmark is useful for counting. Each horizon contributes one $N \times N$ restriction, so stacking $h = 0, \ldots, H$ gives $(H+1)N^2$ scalar equations. If every coefficient entry is free, the unknown blocks contribute

$$A_0 : \; N^2, \qquad \{A_s\}_{s=1}^S : \; SN^2, \qquad \{B_s\}_{s=1}^S : \; SN^2, \qquad \{D_s\}_{s=1}^S : \; SNm,$$

where $D_s \in \mathbb{R}^{N \times m}$ and $m = N(N+1)/2$. The unrestricted system therefore contains

$$N^2(1 + 2S) + SNm$$

scalar unknowns, so a necessary condition for identification is

$$(H+1)N^2 \geq N^2(1 + 2S) + SNm,$$

or equivalently

$$H + 1 \geq 1 + 2S + \frac{Sm}{N} = 1 + 2S + \frac{S(N+1)}{2}.$$

This grows quickly with both horizon length and cross-sectional dimension, which is why the unrestricted system is mainly a benchmark.

Two special cases help calibrate the counting. If $B_s = D_s = 0$ and $A_s = 0$ for all $s \geq 1$, then only A_0 is unknown, giving N^2 parameters and requiring only $H + 1 \geq 1$. If $B_s = D_s = 0$ but $(A_1, \ldots, A_S)$ are retained, then the unknowns are $(A_0, \ldots, A_S)$ only, for a total of $(S + 1)N^2$ parameters, and the necessary condition becomes $H + 1 \geq S + 1$.

Restrictions on the risk block can also lower the dimension substantially. As discussed in the text, if investors care only about each asset's own conditional variance and its conditional covariance with F common factors, then each risk matrix is $N \times (1 + F)$ rather than $N \times m$. The unrestricted count then becomes

$$N^2(1 + 2S) + SN(1 + F),$$

which is much smaller when $F \ll N$.

H Nesting Different Demand Specifications

This appendix derives how our demand system (1) nests a wide range of models considered in the literature. We consider the static Koijen and Yogo (2019) system, intertemporal hedging demand (Merton, 1973; Campbell and Viceira, 2002), models with belief distortions (Gabaix, 2019; Bordalo et al., 2018) and dispersed information (Grossman and Stiglitz, 1980), and preferred-habitat no-arbitrage specifications (Vayanos and Vila, 2021). Each subsection linearizes a canonical model and maps it into our notation, making explicit the implied restrictions on (A_0, A_s, B_s, D_s) and on the latent demand shifter ξ_t.

H.1 Static models

This is the $S = 0$ case of (1): demand depends only on current prices, so $A_s = B_s = D_s = 0$ for all $s \geq 1$, as in (7).

Koijen and Yogo (2019). We consider a simplified version of their demand system that abstracts from observed characteristics other than the latent shifter $\Lambda_{j,t}$:

$$\frac{\omega_{j,t}}{\omega_{0,t}} = \exp\left(\beta_0 + \beta_1 p_{j,t}\right) \Lambda_{j,t} \equiv \delta_{j,t},$$

for assets $j = 1, \ldots, N$, where $\omega_{j,t}$ denotes the portfolio weight of asset j at time t, and $\Lambda_{j,t}$ is the latent demand shifter. Asset $j = 0$ is the outside asset, whose price we normalize to one, so $p_{0,t} = 0$; in our setting, we take it to be the risk-free asset with constant price.

The portfolio weight of asset j is given by the multinomial logit formula:

$$\omega_{j,t} = \frac{\delta_{j,t}}{1 + \sum_{i=1}^{N} \delta_{i,t}}.$$

Given the portfolio weight, we can solve for the quantity of asset j held $Q_{j,t}$ using the expression: $\omega_{j,t} \equiv \frac{P_{j,t} Q_{j,t}}{W_t (1 - C_t)}$, where $P_{j,t}$ denotes the price of asset j in levels, W_t denotes the wealth of the investor, and C_t denotes the consumption-wealth ratio. For simplicity, we assume the consumption-wealth ratio is constant in this static model.

We consider a linearization of the demand system around the approximation point $(\overline{P}, \overline{Q}, \overline{W}, \overline{C})$. We assume that the price and wealth are stationary, so that the approximation point is time-invariant. It is straightforward to generalize the results to the case where prices and wealth are scaled by a stochastic trend, so normalized prices and wealth are stationary.

Linearizing the demand for asset $j = 0$ gives:

$$q_{0,t} = \overline{q}_0 - \sum_{j=1}^{N} \frac{\overline{\delta}_j}{1 + \sum_{i=1}^{N} \overline{\delta}_i} (\beta_1 p_{j,t} + \log \Lambda_{j,t}) + w_t - \overline{w},$$

where $\overline{q}_0$ collects the constant terms.

Abstracting from dividends, the wealth is given by $W_t = \sum_{j=0}^{N} P_{j,t} Q_{j,t-1}$. Linearizing this expression, we obtain:

$$w_t = \overline{w} + \sum_{j=0}^{N} \frac{\overline{P}_j \overline{Q}_j}{\overline{W}} \left[(p_{j,t} - \overline{p}_j) + (q_{j,t-1} - \overline{q}_j) \right]$$

where $\overline{W}$ is the approximation point for wealth. Note that the $j = 0$ term contributes only through $q_{0,t-1}$, since $p_{0,t} = 0$.

Combining the expressions above, we obtain:

$$q_{0,t} = \overline{q}_0 + \chi_0^\top p_t + \xi_{0,t},$$

where $p_t = (p_{1,t}, \ldots, p_{N,t})^\top$, and the vector χ_0 and shifter $\xi_{0,t}$ are given by:

$$\chi_0 = \left(\frac{\overline{P}_1 \overline{Q}_1}{\overline{W}} - \frac{\overline{\delta}_1 \beta_1}{1 + \sum_{i=1}^N \overline{\delta}_i}, \ldots, \frac{\overline{P}_N \overline{Q}_N}{\overline{W}} - \frac{\overline{\delta}_N \beta_1}{1 + \sum_{i=1}^N \overline{\delta}_i} \right)^\top, \qquad \xi_{0,t} = \sum_{j=0}^N \frac{\overline{P}_j \overline{Q}_j}{\overline{W}} q_{j,t-1} - \sum_{j=1}^N \frac{\overline{\delta}_j}{1 + \sum_{i=1}^N \overline{\delta}_i} \log \Lambda_{j,t},$$

where the remaining constant terms are absorbed into $\overline{q}_0$. Each element of χ_0 combines a wealth effect, as a higher price $p_{j,t}$ raises wealth and hence the demand for the outside asset, and a substitution effect, as a higher price changes the relative attractiveness of asset j.

Taking logs of the demand for asset $j \in \{1, \ldots, N\}$, and using $\omega_{j,t}/\omega_{0,t} = P_{j,t} Q_{j,t} / (P_{0,t} Q_{0,t})$ with $p_{0,t} = 0$, gives the exact relation:

$$q_{j,t} = q_{0,t} + \beta_0 + (\beta_1 - 1) p_{j,t} + \log \Lambda_{j,t}.$$

Combining the expressions above, we obtain:

$$q_t = -A_0 p_t + \xi_t, \tag{A.5}$$

where

$$A_0 = (1 - \beta_1) I_N - \mathbf{1}_N \times \chi_0^\top, \qquad \xi_{j,t} = \overline{q}_0 + \beta_0 + \xi_{0,t} + \log \Lambda_{j,t}.$$

Differentiating (A.5) gives the own- and cross-price elasticities

$$\frac{\partial q_{i,t}}{\partial p_{j,t}} = -(1 - \beta_1) \mathbf{1}_{\{i=j\}} - \chi_{0,j}.$$

Thus the cross-price elasticity of asset i with respect to $p_{j,t}$ is $-\chi_{0,j}$ for $i \neq j$, while the own-price elasticity is $-(1 - \beta_1) - \chi_{0,i}$. The matrix A_0 therefore encodes the contemporaneous own- and cross-price responses, while ξ_t collects the latent demand shifters $\log \Lambda_{j,t}$, lagged holdings, and constants.

H.2 Intertemporal hedging demands

We next nest intertemporal hedging demand from the Campbell and Viceira (2002) log-linearization of Merton (1973). In the homoskedastic case, the conditional covariance matrix Σ is constant, so demand responds to the path of expected returns while higher moments are fixed; in (1), this corresponds to $D_s = 0$ for all s (or, equivalently, to risk terms that enter only through

constants absorbed into ξ_t). The coefficients $(A_s, B_s)_{s=1}^{S}$ are unrestricted beyond what is implied by the return-based representation below.

We focus on the case in which expected returns are affine in a K-dimensional state X_t,

$$\mathbb{E}_t[r_{t+1}] = \mu_r + \Psi X_t,$$

where μ_r is the unconditional mean of returns, Ψ is an $N \times K$ matrix, and the state follows the VAR(1) process

$$X_t = \Phi_1 X_{t-1} + \epsilon_t, \tag{A.6}$$

where ϵ_t is a K-dimensional vector of white-noise shocks.

Campbell and Viceira (2002) linearize the portfolio share equation as

$$\alpha_t = \underbrace{\frac{1}{\gamma}\Sigma^{-1}\left(\mathbb{E}_t[r_{t+1}] + \frac{\sigma^2}{2}\right)}_{\text{myopic demand}} + \underbrace{A_{\alpha,0} + A_{\alpha,X} X_t}_{\text{hedging demand}},$$

where σ^2 denotes the vector of diagonal elements of Σ, so the Jensen correction is part of the myopic demand and hedging demand vanishes in the log-utility case.

Portfolio shares are therefore affine in X_t. Equivalently, they can be written as a function of the sequence of expected returns. Since

$$\mathbb{E}_t[r_{t+s}] = \mu_r + \Psi \Phi_1^{s-1} X_t,$$

we stack expected returns over horizons $1, \ldots, S$:

$$\mathcal{R}_t \equiv \begin{bmatrix} \mathbb{E}_t[r_{t+1}] \\ \vdots \\ \mathbb{E}_t[r_{t+S}] \end{bmatrix} = \mu_{\mathcal{R}} + \widetilde{\Psi} X_t,$$

where $\mu_{\mathcal{R}} = \mathbf{1}_S \otimes \mu_r$ and $\widetilde{\Psi} = [\Psi^\top, (\Psi\Phi_1)^\top, \ldots, (\Psi\Phi_1^{S-1})^\top]^\top$ is an $NS \times K$ matrix. If S is large enough that $\operatorname{rank}(\widetilde{\Psi}) = K$, the state can be recovered from the expected-return sequence,

$$X_t = (\widetilde{\Psi}^\top \widetilde{\Psi})^{-1} \widetilde{\Psi}^\top (\mathcal{R}_t - \mu_{\mathcal{R}}).$$

If expected returns alone do not span the state—for example, because volatility also moves with X_t—the joint path of expected returns and risk would be required instead.

Substituting into the portfolio-share equation and defining

$$\tilde{A}_{\alpha,0} \equiv A_{\alpha,0} - A_{\alpha,X}(\widetilde{\Psi}^\top \widetilde{\Psi})^{-1}\widetilde{\Psi}^\top \mu_{\mathcal{R}},$$

we obtain

$$\alpha_t = \tilde{A}_{\alpha,0} + \frac{1}{\gamma}\Sigma^{-1}\left(\mathbb{E}_t[r_{t+1}] + \frac{\sigma^2}{2}\right) + A_{\alpha,X}(\widetilde{\Psi}^\top \widetilde{\Psi})^{-1}\widetilde{\Psi}^\top \mathcal{R}_t.$$

Defining the consecutive $N \times N$ blocks of $A_{\alpha,X}(\widetilde{\Psi}^\top \widetilde{\Psi})^{-1}\widetilde{\Psi}^\top$ as $\{\tilde{C}_s\}_{s=1}^S$, we obtain

$$\alpha_t = \tilde{A}_{\alpha,0} + \frac{1}{\gamma}\Sigma^{-1}\left(\mathbb{E}_t[r_{t+1}] + \frac{\sigma^2}{2}\right) + \sum_{s=1}^{S} \tilde{C}_s \mathbb{E}_t[r_{t+s}].$$

The same linearization also implies $C_t = A_{C,0} + A_{C,X}X_t$, and hence that C_t can be expressed as a function of $\mathcal{R}_t$. Using $Q_{j,t} = \alpha_{j,t}(1 - C_t)W_t/P_{j,t}$ and linearizing around the steady state as in the static subsection above gives

$$q_t = -A_0^w p_t + \sum_{s=1}^{S} C_s \mathbb{E}_t[r_{t+s}] + \xi_t,$$

where $A_0^w = I_N - \mathbf{1}_N \overline{s}^\top$, with $\overline{s} = (\overline{P}_1\overline{Q}_1/\overline{W}, \ldots, \overline{P}_N\overline{Q}_N/\overline{W})^\top$ the vector of steady-state wealth shares, and where C_s combine the contributions of α_t and C_t. The shifter ξ_t collects lagged holdings and constants from the wealth expansion; since the risk-free rate is constant, expected returns and expected excess returns differ only by constants absorbed into ξ_t. This is the multivariate generalization of the return-based demand system (5), augmented by the contemporaneous price term $-A_0^w p_t$ through the wealth channel.

Under log utility, hedging demand vanishes and the consumption-wealth ratio is constant, so $C_s = 0$ for all $s > 1$ and demand depends only on the current expected return. With Epstein–Zin preferences, intertemporal links can remain even when the EIS is one (through hedging) or when risk aversion is one (through the consumption-wealth ratio). The sequence-space representation therefore nests Merton hedging while allowing the econometrician to work directly with the path $\{\mathbb{E}_t[r_{t+s}]\}_{s=1}^S$ without observing X_t.

H.3 Behavioral models and informational frictions

The preceding subsections nest canonical demand systems under full-information rational expectations. However, a large literature studies models with belief distortions, where expectations deviate from the rational benchmark. Our demand system can accommodate a range of behavioral models, as well as rational models in which investors face informational frictions.

Starting from (1), suppose the investor's demand depends on subjective rather than objective expected paths. To capture belief distortions and informational frictions jointly, we write demand in terms of the *subjective* or perceived paths of prices, dividends, and risk:

$$q_t = -A_0\, p_t + \sum_{s=1}^{S} \hat{A}_s\, \hat{\mu}^p_{t,t+s} + \sum_{s=1}^{S} \hat{B}_s\, \hat{\mu}^d_{t,t+s} - \sum_{s=1}^{S} \hat{D}_s\, \hat{v}^r_{t,t+s}, \tag{A.7}$$

where $\hat{\mu}^p_{t,t+s}$ and $\hat{\mu}^d_{t,t+s}$ are the investor's perceived paths of prices and dividends at horizon s, $\hat{v}^r_{t,t+s}$ is the perceived conditional return risk at horizon s, and $\hat{A}_s$,$\hat{B}_s \in \mathbb{R}^{N\times N}$ and $\hat{D}_s \in \mathbb{R}^{N\times m}$ are the subjective demand coefficients.

We assume that the subjective values are related to the objective ones through the following relationship:

$$\hat{\mu}^p_{t,t+s} = M^p_s \mu^p_{t,t+s} + \xi^p_{t,t+s}, \qquad \hat{\mu}^d_{t,t+s} = M^d_s \mu^d_{t,t+s} + \xi^d_{t,t+s}, \qquad \hat{v}^r_{t,t+s} = M^r_s v^r_{t,t+s} + \xi^r_{t,t+s}, \tag{A.8}$$

where M^p_s,$M^d_s \in \mathbb{R}^{N\times N}$, $M^r_s \in \mathbb{R}^{m\times m}$, and $\xi^p_{t,t+s}$,$\xi^d_{t,t+s} \in \mathbb{R}^N$ and $\xi^r_{t,t+s} \in \mathbb{R}^m$ are belief shocks, dated t and indexed by the horizon $t+s$ they refer to. Other latent demand shifters can be subsumed in these shocks or appear after substituting (A.8) into (A.7).

We show next that a range of belief distortions and informational frictions proposed in the literature fit this general framework.

Behavioral inattention. In the behavioral inattention model of Gabaix (2019) and Gabaix (2023), investors solve a dynamic problem with a perceived law of motion for the state variables that differs from the true one. Gabaix (2023) represents this distortion in state space: if the true law of motion is (A.6), the investor perceives

$$X_t = M_X \Phi_1 X_{t-1} + M_\epsilon \epsilon_t,$$

where M_X is a $K \times K$ matrix and M_ϵ is a $K \times K$ diagonal matrix of attention coefficients. In the baseline Gabaix representation, inattention is deterministic and operates through (M_X, M_ϵ); we allow additional idiosyncratic belief shocks $\xi^{p,d,r}_{t,t+s}$ in (A.8) on top of this attenuation.

Under the hedging setup above, $\mu^p_{t,t+s}$, $\mu^d_{t,t+s}$, and $v^r_{t,t+s}$ are affine in $\Phi_1^{s-1} X_t$, so attenuating the state transition and shock loading maps into horizon-dependent matrices M^p_s, M^d_s, M^r_s in (A.8). Alternatively, one can view (A.8) as the direct belief distortions for an investor who solves a sequential problem. Current prices are observed, so A_0 is unchanged.

Substituting (A.8) into (A.7) yields the objective demand system

$$q_t = -A_0\, p_t + \sum_{s=1}^{S} A_s\, \mu^p_{t,t+s} + \sum_{s=1}^{S} B_s\, \mu^d_{t,t+s} - \sum_{s=1}^{S} D_s\, v^r_{t,t+s} + \xi_t,$$

with composite coefficients

$$A_s = \hat{A}_s M^p_s, \qquad B_s = \hat{B}_s M^d_s, \qquad D_s = \hat{D}_s M^r_s,$$

and shifter

$$\xi_t = \sum_{s=1}^{S} \left(\hat{A}_s \xi^p_{t,t+s} + \hat{B}_s \xi^d_{t,t+s} - \hat{D}_s \xi^r_{t,t+s} \right). \tag{A.9}$$

Two implications follow. First, only the composite coefficients (A_s, B_s, D_s) are identified from data on objective expectations: a behavioral investor who misperceives future paths is observationally equivalent to a rational investor with different intertemporal elasticities to prices, dividends, and risk. For counterfactuals that move entire expected paths of prices, dividends, and risk, these composites are the relevant objects, so the recovery analysis applies unchanged. Second, when belief shocks $\xi^{p,d,r}_{t,t+s}$ are present, (A.9) provides a micro-foundation for the latent demand shifter ξ_t.

Diagnostic expectations. Diagnostic expectations (Bordalo et al., 2018) provide a second example in which behavioral beliefs can be mapped into our sequential demand representation. The key difference relative to behavioral inattention is that diagnostic beliefs depend not only on the current forecast, but also on a reference forecast formed in the past. To keep the equilibrium finite dimensional, we follow the "shadow rational-expectations" implementation in Bianchi, Ilut, and Saijo (2024): investors form diagnostic forecasts by distorting forecasts from a shadow rational-

expectations economy rather than recursively distorting the forecasts generated by the diagnostic economy itself.

In the shadow economy, the primitive state follows (A.6). Under shadow rational expectations, any demand-relevant object is affine in the state. For concreteness, consider expected returns,

$$\mathbb{E}_t^{RE}[r_{t+s}] = \mu_r + \Psi\Phi_1^{s-1}X_t.$$

To keep the notation transparent, consider the case in which the reference forecast is formed one period ago. Diagnostic expectations distort the shadow forecast relative to this recent reference forecast:

$$\widehat{\mathbb{E}}_t[r_{t+s}] = \mathbb{E}_t^{RE}[r_{t+s}] + \theta\left(\mathbb{E}_t^{RE}[r_{t+s}] - \mathbb{E}_{t-1}^{RE}[r_{t+s}]\right), \tag{A.10}$$

where $\theta \geq 0$ measures the strength of diagnostic overreaction; when $\theta = 0$, beliefs coincide with rational expectations. Since the anchor is the shadow rational-expectations law of motion, the lagged reference forecast is

$$\mathbb{E}_{t-1}^{RE}[r_{t+s}] = \mu_r + \Psi\Phi_1^{s}X_{t-1}.$$

Substituting into (A.10) gives

$$\widehat{\mathbb{E}}_t[r_{t+s}] = \mu_r + (1+\theta)\Psi\Phi_1^{s-1}X_t - \theta\Psi\Phi_1^{s}X_{t-1}. \tag{A.11}$$

Thus diagnostic beliefs expand the investor's relevant state from X_t to

$$\mathcal{S}_t^D \equiv \begin{bmatrix} X_t \\ X_{t-1} \end{bmatrix}.$$

The same logic applies to prices, dividends, and risk. If the shadow rational-expectations forecast of any demand-relevant object y_{t+s} is

$$\mathbb{E}_t^{RE}[y_{t+s}] = \mu_y + \Psi_y\Phi_1^{s-1}X_t,$$

then the diagnostic forecast is

$$\widehat{\mathbb{E}}_t[y_{t+s}] = \mu_y + (1+\theta)\Psi_y\Phi_1^{s-1}X_t - \theta\Psi_y\Phi_1^{s}X_{t-1}, \tag{A.12}$$

which is affine in $\mathcal{S}_t^D = (X_t^\top, X_{t-1}^\top)^\top$. The objective equilibrium expectations derived below are affine in the same expanded state, with potentially different coefficients. Once the diagnostic equilibrium is solved and $\mathcal{S}_t^D$ is spanned by the objective expected-return sequence, subjective and objective paths can therefore be mapped into each other as in (A.8).

Consider now the portfolio-share representation. After solving the diagnostic-expectations equilibrium, let $\mathbb{E}_t[r_{t+s}]$ denote the *objective* expected return in that equilibrium—the quantity that appears in market clearing and recovery, not the shadow-RE benchmark $\mathbb{E}_t^{RE}[r_{t+s}]$ nor the diagnostic belief $\widehat{\mathbb{E}}_t[r_{t+s}]$. Prices and these objective expected returns are affine in the expanded state:

$$\mathbb{E}_t[r_{t+s}] = \mu_s^D + \Psi_s^D \mathcal{S}_t^D, \qquad s = 1, \ldots, S.$$

The investor's portfolio share is also affine in the same state,

$$\alpha_t = A_{\alpha,0}^D + A_{\alpha,S}^D \mathcal{S}_t^D,$$

with the myopic mean–variance component absorbed into the intercept $A_{\alpha,0}^D$. Stack the objective expected returns in the diagnostic equilibrium,

$$\mathcal{R}_t^D \equiv \begin{bmatrix} \mathbb{E}_t[r_{t+1}] \\ \vdots \\ \mathbb{E}_t[r_{t+S}] \end{bmatrix} = \mu_{\mathcal{R}}^D + \widetilde{\Psi}_D \mathcal{S}_t^D, \qquad \widetilde{\Psi}_D \equiv \begin{bmatrix} \Psi_1^D \\ \vdots \\ \Psi_S^D \end{bmatrix}.$$

Stack diagnostic return beliefs the same way,

$$\widehat{\mathcal{R}}_t^D \equiv \begin{bmatrix} \widehat{\mathbb{E}}_t[r_{t+1}] \\ \vdots \\ \widehat{\mathbb{E}}_t[r_{t+S}] \end{bmatrix} = \widehat{\mu}_{\mathcal{R}}^D + \widehat{\Psi}_D \mathcal{S}_t^D,$$

where $\widehat{\Psi}_D$ collects the diagnostic coefficients from (A.11). If S is large enough that $\operatorname{rank}(\widetilde{\Psi}_D) = \dim(\mathcal{S}_t^D)$, then the expanded state can be recovered from the objective expected-return sequence:

$$\mathcal{S}_t^D = (\widetilde{\Psi}_D^\top \widetilde{\Psi}_D)^{-1} \widetilde{\Psi}_D^\top (\mathcal{R}_t^D - \mu_{\mathcal{R}}^D),$$

and diagnostic beliefs can be written as linear functions of objective equilibrium expectations:

$$\widehat{\mathcal{R}}_t^D = \widehat{\mu}_{\mathcal{R}}^D + \widehat{\Psi}_D(\widetilde{\Psi}_D^\top \widetilde{\Psi}_D)^{-1}\widetilde{\Psi}_D^\top(\mathcal{R}_t^D - \mu_{\mathcal{R}}^D). \tag{A.13}$$

The same logic applies to prices, dividends, and risk through (A.12). Horizon-by-horizon, (A.13) delivers the matrices (M_s^p, M_s^d, M_s^r) in (A.8); unlike behavioral inattention, they are equilibrium objects rather than direct attention parameters on the state law of motion, and they generally couple horizons. Substituting the recovered state into the portfolio-share equation and defining

$$\widetilde{A}_{\alpha,0}^D \equiv A_{\alpha,0}^D - A_{\alpha,S}^D(\widetilde{\Psi}_D^\top \widetilde{\Psi}_D)^{-1}\widetilde{\Psi}_D^\top \mu_{\mathcal{R}}^D,$$

we obtain

$$\alpha_t = \widetilde{A}_{\alpha,0}^D + A_{\alpha,S}^D(\widetilde{\Psi}_D^\top \widetilde{\Psi}_D)^{-1}\widetilde{\Psi}_D^\top \mathcal{R}_t^D.$$

Defining the consecutive $N \times N$ blocks of

$$A_{\alpha,S}^D(\widetilde{\Psi}_D^\top \widetilde{\Psi}_D)^{-1}\widetilde{\Psi}_D^\top$$

as $\{\widetilde{C}_s^D\}_{s=1}^S$, we obtain

$$\alpha_t = \widetilde{A}_{\alpha,0}^D + \sum_{s=1}^{S} \widetilde{C}_s^D \mathbb{E}_t[r_{t+s}].$$

Therefore, diagnostic expectations nest our sequential demand system along the same return-spanning route as intertemporal hedging. The behavioral distortion changes the equilibrium mapping from states to prices and returns and expands the relevant state by introducing one lagged memory state, but once the diagnostic equilibrium is solved, portfolio demand can be written as a linear function of the sequence of objective expected future returns. The same spanning argument applies to prices, dividends, and risk through (A.12); here we display the return-based representation for brevity. Applying the same log-linearization from portfolio shares to quantities as in the hedging subsection gives

$$q_t = -A_0^w p_t + \sum_{s=1}^{S} C_s^D \mathbb{E}_t[r_{t+s}] + \xi_t^D,$$

where A_0^w is the wealth-channel price coefficient defined above, C_s^D are the log-linearized

counterparts of $\widetilde{C}_s^D$ and incorporate both intertemporal hedging motives and the diagnostic belief distortion, and ξ_t^D collects the residual terms from the log-linearization of portfolio shares into quantities—including dependence on lagged holdings—as well as any forecast wedge if the econometrician's information set is coarser than $\mathcal{S}_t^D$.

Noisy rational expectations. Noisy-information models provide another route through which our sequential demand system arises. Unlike the behavioral inattention and diagnostic-expectations examples above, investors remain fully rational conditional on their information set. The departure from full-information rational expectations comes from the fact that investors do not directly observe the state vector driving expected returns.

The true state follows (A.6), and investors observe the noisy signal

$$Y_t = HX_t + \eta_t,$$

where H maps the state into signals and η_t is a white-noise information shock, independent of ϵ_t.

In a general noisy rational-expectations equilibrium, prices themselves convey information (Grossman and Stiglitz, 1980). In a linear equilibrium, observing the price system is informationally equivalent to observing the endogenous signal

$$\tilde{X}_t = GX_t + u_t,$$

where u_t is the noise that prevents prices from being fully revealing—outside supply, for instance—and the matrix G is determined in equilibrium.

Under linear-Gaussian dynamics, investors form the filtered estimate

$$\hat{X}_t \equiv \mathbb{E}[X_t \mid Y^t, \tilde{X}^t],$$

which evolves recursively according to the Kalman filter,

$$\hat{X}_t = K_Y Y_t + K_{\tilde{X}} \tilde{X}_t + K_X \hat{X}_{t-1},$$

for matrices $(K_Y, K_{\tilde{X}}, K_X)$ given by the steady-state Kalman gain. Because the signal $\tilde{X}_t$ is

endogenous, the gains are determined as a fixed point: beliefs shape demand, demand shapes the pricing rule, and the pricing rule shapes the informativeness of prices.

Rather than describing the equilibrium from the perspective of an outside observer who sees the true state, we work directly with the investor-observable state

$$\mathcal{S}_t^{NR} \equiv \begin{bmatrix} \hat{X}_t \\ \tilde{X}_t \end{bmatrix},$$

which is a sufficient statistic for the investor's forecasts: $\hat{X}_t$ summarizes beliefs about fundamentals, while $\tilde{X}_t$ pins down the current price and the investor's estimate of the current noise.

The investor's conditional expectations of future returns are then affine in this state:

$$\widehat{\mathbb{E}}_t[r_{t+s}] \equiv \mathbb{E}[r_{t+s} \mid Y^t, \tilde{X}^t] = \mu_s^{NR} + \Psi_s^{NR}\mathcal{S}_t^{NR}, \qquad s = 1, \ldots, S.$$

Likewise, portfolio shares are affine in the same state,

$$\alpha_t = A_{\alpha,0}^{NR} + A_{\alpha,S}^{NR}\mathcal{S}_t^{NR},$$

with the myopic mean–variance component absorbed into $A_{\alpha,0}^{NR}$. Stacking the investor's expected returns gives

$$\mathcal{R}_t^{NR} = \mu_{\mathcal{R}}^{NR} + \widetilde{\Psi}_{NR}\mathcal{S}_t^{NR},$$

where

$$\widetilde{\Psi}_{NR} \equiv \begin{bmatrix} \Psi_1^{NR} \\ \vdots \\ \Psi_S^{NR} \end{bmatrix}.$$

If $\widetilde{\Psi}_{NR}$ has full column rank, the investor-observable state can be recovered from the investor's expected-return sequence,

$$\mathcal{S}_t^{NR} = (\widetilde{\Psi}_{NR}^{\top}\widetilde{\Psi}_{NR})^{-1}\widetilde{\Psi}_{NR}^{\top}(\mathcal{R}_t^{NR} - \mu_{\mathcal{R}}^{NR}),$$

and portfolio demand can be written as a linear function of $\widehat{\mathbb{E}}_t[r_{t+s}]$ using the same spanning argument as above.

It is natural to assume that the econometrician does not observe the full information set $\{Y^t, \tilde{X}^t\}$

used by investors. For instance, the econometrician may not observe all private signals, filtered beliefs, or informational variables that enter investors' forecasts. Let $\mathcal{I}_t^E \subset \{Y^t, \tilde{X}^t\}$ denote the information set available to the econometrician and define

$$\mathbb{E}_t[r_{t+s}] \equiv \mathbb{E}[r_{t+s} \mid \mathcal{I}_t^E]$$

for the econometrician's forecast of future returns. Define the forecast wedge

$$\xi_{t,t+s}^{NR} \equiv \widehat{\mathbb{E}}_t[r_{t+s}] - \mathbb{E}[r_{t+s} \mid \mathcal{I}_t^E], \tag{A.14}$$

so that $\widehat{\mathbb{E}}_t[r_{t+s}] = \mathbb{E}_t[r_{t+s}] + \xi_{t,t+s}^{NR}$. By the law of iterated expectations—using that $\mathcal{I}_t^E$ is nested in the investor's information set—$\mathbb{E}[\xi_{t,t+s}^{NR} \mid \mathcal{I}_t^E] = 0$: the wedge is orthogonal to the econometrician's information set. This orthogonality delivers the exclusion restriction in Assumption 1, provided the identified shocks are measurable with respect to $\mathcal{I}_t^E$. When $\mathcal{I}_t^E = \{Y^t, \tilde{X}^t\}$, the wedge vanishes and the representation collapses to standard sequential demand with no latent shifter.

Relative to the econometrician's measured expectations, (A.8) therefore takes the special case $M_s^p = M_s^d = M_s^r = I$; the entire departure from measured expectations operates through the forecast wedges $\xi_{t,t+s}^{NR}$, not through attenuation of objective paths. Substituting (A.14) into the portfolio-share equation and defining

$$\widetilde{A}_{\alpha,0}^{NR} \equiv A_{\alpha,0}^{NR} - A_{\alpha,S}^{NR}(\widetilde{\Psi}_{NR}^\top \widetilde{\Psi}_{NR})^{-1} \widetilde{\Psi}_{NR}^\top \mu_{\mathcal{R}}^{NR},$$

we obtain

$$\alpha_t = \widetilde{A}_{\alpha,0}^{NR} + \sum_{s=1}^{S} \widetilde{C}_s^{NR}\, \mathbb{E}_t[r_{t+s}] + \xi_t^{NR}, \qquad \xi_t^{NR} \equiv \sum_{s=1}^{S} \widetilde{C}_s^{NR}\, \xi_{t,t+s}^{NR},$$

where the consecutive $N \times N$ blocks of $A_{\alpha,S}^{NR}(\widetilde{\Psi}_{NR}^\top \widetilde{\Psi}_{NR})^{-1} \widetilde{\Psi}_{NR}^\top$ are $\{\widetilde{C}_s^{NR}\}_{s=1}^S$. Thus ξ_t^{NR} is not a behavioral belief distortion: it is an omitted-information term arising because investors condition on a richer information set than the researcher. Applying the same log-linearization from portfolio shares to quantities as in the hedging and diagnostic subsections gives

$$q_t = -A_0^w p_t + \sum_{s=1}^{S} C_s^{NR} \mathbb{E}_t[r_{t+s}] + \xi_t^{NR},$$

where C_s^{NR} are the log-linearized counterparts of $\widetilde{C}_s^{NR}$.

This example illustrates an important source of latent demand. If the econometrician observed the same information set as investors, the latent demand component would disappear, because all relevant information would already be incorporated into measured expected returns; it arises only because the econometrician observes a coarser information set than investors. From the investors' perspective, $\widehat{\mathbb{E}}_t[r_{t+s}]$ is the relevant expectation; from the econometrician's perspective, the component not spanned by $\mathbb{E}_t[r_{t+s}]$ appears as an unobserved demand shifter.

Together, these examples show that the sequential demand system (1) and the recovery exercise in Section 4 apply once expectations enter through their measured, objective counterparts. Behavioral inattention reshapes the identified intertemporal elasticities through the matrices M_s^p, M_s^d, M_s^r but leaves the recovery machinery unchanged. Diagnostic expectations change the equilibrium mapping from states to prices and returns while preserving the return-spanning representation. Noisy information leaves the structural coefficients on measured expectations unchanged but introduces ξ_t^{NR} whenever the econometrician conditions on a coarser information set than investors. In all three cases, Assumption 1 remains the relevant exclusion restriction whenever ξ_t is not directly shifted by the identifying outside-demand shock.

H.4 Preferred-habitat models

Finally, we nest preferred-habitat no-arbitrage models (Vayanos and Vila, 2021). In (1), this is the $S = 1$ mean–variance case with $D_s = 0$, $A_1 = \beta A_0$, and $A_s = B_s = 0$ for $s \geq 2$: a myopic arbitrageur responds only to next-period expected prices and dividends while second moments are constant. Habitat investors enter either through outside holdings or through an additive contribution to A_0; the latent shifter ξ_t collects linearization constants and does not represent habitat demand, which is absorbed into z_t or A_0^{hab}.

Vayanos and Vila (2021) develop the leading example for the term structure, where habitat investors prefer specific bond maturities and competitive arbitrageurs absorb residual demand; the same logic applies whenever a price-elastic mean–variance arbitrageur clears against an inelastic or asset-specific clientele. There are N assets, habitat demand that may be price-inelastic or price-elastic, and a representative arbitrageur who maximizes $\mathbb{E}_t[W_{t+1}] - \frac{\gamma}{2}\mathbb{V}\text{ar}_t(W_{t+1})$ with risk aversion $\gamma > 0$ and conditional covariance $\Sigma \succ 0$. The arbitrageur's first-order condition delivers

(8)–(9), so A_0^{arb} is proportional to Σ^{-1} up to the diagonal scaling in (9).[13] When habitat demand is price-inelastic, it is subsumed into the outside-holdings vector Z_t defined in Section 2.1; when it is price-elastic, the aggregate own-price coefficient decomposes as $A_0 = A_0^{\text{arb}} + A_0^{\text{hab}}$. Market clearing then reduces to (2).

Combined with the arbitrageur's pricing condition—expected returns move with $\gamma\Sigma$ times the residual portfolio the arbitrageur must hold—this delivers the familiar mapping from supply shocks in z_t to risk premia in the affine equilibrium of Vayanos and Vila (2021). No-arbitrage is therefore not in tension with a demand-system representation: the structural coefficients (A_0, A_1) encode the arbitrageur's effective price of risk, with γ and Σ determining the cross-asset spillovers that accompany any habitat shock.

[13] Vayanos and Vila (2021) work in continuous time with the analogous instantaneous mean–variance objective $\mathbb{E}_t[dW_t] - \frac{a}{2}\mathbb{V}\text{ar}_t(dW_t)$ in place of the one-period objective used here. Both formulations deliver the same mean–variance demand: arbitrageur holdings proportional to Σ^{-1} times the conditional risk premium.