A Note on Asset Demand Systems

Zhiguo He*

October 2026

Abstract

Understanding the link between prices and quantities requires treating asset demand as a system. This note works through how the characteristics-based asset demand system of Koijen and Yogo (2019) arises from a canonical mean-variance problem with a factor structure: first the endogenous scalars that serve as the coefficients on characteristics, moving with the entire price vector yet invariant across assets; then the two steps that carry the linear system into a logit form, whose coefficient restrictions turn out to impose a strong structure on the expected payoffs of the investor underneath. I then discuss four issues in demand-system estimation: substitution, learning from prices, dynamics, and whether imposing structure is a concession or a necessity.

1 Asset Demand System

Studying prices and quantities together is less a new subject than a neglected one. An equilibrium is by construction a statement about prices *and* holdings jointly, and a demand schedule is the first object one writes down when solving for the equilibrium. From the 1970s onward, though, finance set itself apart from the rest of economics by leaning on no-arbitrage — replication, the stochastic discount factor, Black–Scholes–Merton — and along that route prices are pinned down without asking who holds what. That is exact in a representative-agent endowment economy (Lucas, 1978), but not once agents are heterogeneous and markets frictional (Allen and Gale, 1994; Basak and Cuoco, 1998), as demand-based option pricing showed in the market

*Graduate School of Business, Stanford University, hezhg@stanford.edu; NBER and FTG. This note grew out of discussions with Valentin Haddad, Péter Kondor, and Marco Loseto. Thanks to Lorenzo Garlappi and Tyler Muir for their careful reading. All errors are mine.

where replication looked tightest (Gârleanu, Pedersen and Poteshman, 2009). Post-crisis macro-finance has since made the interdependence its central theme (He and Krishnamurthy, 2013; Brunnermeier and Sannikov, 2014).

The demand-system approach of Koijen and Yogo (2019, hereafter KY19) and Koijen, Richmond and Yogo (2023, hereafter KRY23) goes a step further. They do not treat quantities as one more diagnostic for a pricing model but take the demand system itself as the goal, improving on the older flow literature (Coval and Stafford, 2007; Lou, 2012), which regressed returns on flows and called the coefficient "price impact". They estimate it from the increasingly rich data on who holds what, by the method familiar from industrial organization: impose enough structure that demand can be recovered from observed characteristics, then use the estimate for counterfactuals. Concretely, their key insight for obtaining tractability in estimation is that the loadings on characteristics may depend on the entire price vector, so long as they are common across assets. That is what makes the system estimable from a single cross-section of holdings; Sections 2 and 3 work through how.

At the core of demand-system estimation is the word *system*, whether the goods are stocks, bonds, or detergents. Write aggregate demand as $\boldsymbol{D}(\boldsymbol{p})$, with elasticity matrix $\boldsymbol{\mathcal{E}} = -\partial \boldsymbol{D}/\partial \boldsymbol{p}'$.[1] Portfolio choice (Markowitz, 1952) and the basic theory of demand (Debreu, 1959) imply that demand for any one asset responds not to its own price alone, but to the price of every asset — so most questions worth asking need the whole matrix $\boldsymbol{\mathcal{E}}$, not any single own-price elasticity $\boldsymbol{\mathcal{E}}_{nn}$, as the next example shows.

Take quantitative easing. The Fed buys ten-year Treasuries (asset 1), hoping to affect the ten-year yield. Suppose the economy also has five-year bonds (asset 2) and the S&P 500 (asset 3) trading alongside. Since only asset 1 is bought, the supply shock is $\boldsymbol{\eta} = (\eta, 0, 0)'$ with $\eta < 0$. Cross-elasticities are obviously needed once the Fed asks about the S&P 500, which enters only through terms like $\boldsymbol{\mathcal{E}}_{31}$. Less obvious: the whole matrix is generically needed even when only the ten-year yield is of interest. Prices move by

$$\Delta \boldsymbol{p} = -\boldsymbol{\mathcal{E}}^{-1}\boldsymbol{\eta},$$

so the price impact on any one asset is a row of $\boldsymbol{\mathcal{E}}^{-1}$ — here $\left(\boldsymbol{\mathcal{E}}^{-1}\right)_{11}\eta$, and $\left(\boldsymbol{\mathcal{E}}^{-1}\right)_{11}$ is not $1/\boldsymbol{\mathcal{E}}_{11}$. Unless $\boldsymbol{\mathcal{E}}$ is diagonal, inverting it mixes in every cross-elasticity. One number will not do.

Estimating an $N \times N$ matrix is impractical — in principle it requires N instruments — so the literature turns to standard mean-variance portfolio theory for restrictions

[1] For clearer exposition I leave units general: the elasticity may be (log) quantity on (log) price. In fact, I would rather call $\boldsymbol{\mathcal{E}}$ a slope matrix—a slope of quantity response when price changes.

that make it tractable, as KY19 do. The goal of this note, stated once at the outset, is to show what reaching that logit form in Ky19 costs: Corollary 1 imposes a strong structure on the expected payoffs of the investor underneath, one the specification itself leaves unsaid. This note (up to Section 4) therefore follows that single thread, and traces how the literature has evolved along it; it is not an exhaustive review.

2 A Canonical Asset Demand System

Consider the classic mean-variance problem (Markowitz, 1952) underlying the asset demand system used by KY19 and KRY23. A CARA investor with risk aversion γ chooses portfolio weights $\boldsymbol{w}_{N\times 1}$ over N risky assets, with $\boldsymbol{\mu}_{N\times 1}$ the vector of expected payoffs, $\boldsymbol{p}_{N\times 1}$ the vector of prices, and $\boldsymbol{\Sigma}_{N\times N}$ the covariance matrix of time-1 payoffs. Impose the one-factor structure, $\boldsymbol{\Sigma} = \sigma_\epsilon^2 \boldsymbol{I}_{N\times N} + \sigma_f^2(\boldsymbol{\beta}\boldsymbol{\beta}')_{N\times N}$, and the Sherman–Morrison formula gives

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{\sigma_\epsilon^2}\left(\boldsymbol{I}_N - \frac{\sigma_f^2}{\sigma_\epsilon^2 + \sigma_f^2 \boldsymbol{\beta}'\boldsymbol{\beta}}\boldsymbol{\beta}\boldsymbol{\beta}'\right). \tag{1}$$

Throughout, a *primitive* is a parameter that does not move with the endogenous price vector $\boldsymbol{p}$. I take $\boldsymbol{\beta}$ as primitive for illustration, but $\boldsymbol{\mu}$ need not stay one, a point Section 3.1 turns on.

2.1 Asset Demand in Unconstrained Problem

In the unconstrained problem,

$$\boldsymbol{w} = \frac{1}{\gamma}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \boldsymbol{p}). \tag{2}$$

Using (1), and noticing $\boldsymbol{\beta}_{N\times 1}\boldsymbol{\beta}'_{1\times N}(\boldsymbol{\mu} - \boldsymbol{p})_{N\times 1} = (\boldsymbol{\beta}'(\boldsymbol{\mu} - \boldsymbol{p}))\boldsymbol{\beta}_{N\times 1}$, this gives

$$\boldsymbol{w}_{N\times 1} = \frac{1}{\gamma\sigma_\epsilon^2}(\boldsymbol{\mu} - \boldsymbol{p}) - \underbrace{\frac{1}{\gamma\sigma_\epsilon^2}\frac{\sigma_f^2 \overbrace{\boldsymbol{\beta}'(\boldsymbol{\mu} - \boldsymbol{p})}^{\text{scalar}}}{\sigma_\epsilon^2 + \sigma_f^2\boldsymbol{\beta}'\boldsymbol{\beta}}}_{\theta(\boldsymbol{p}),\ \text{scalar}}\boldsymbol{\beta} = -\frac{1}{\gamma\sigma_\epsilon^2}\boldsymbol{p} + \frac{1}{\gamma\sigma_\epsilon^2}\boldsymbol{\mu} - \theta(\boldsymbol{p})\boldsymbol{\beta}. \tag{3}$$

Finally, the weight on the risk-free (outside) asset follows from the budget constraint,

$$w_0(\boldsymbol{p}) = 1 - \sum_{n=1}^{N} w_n = 1 - \frac{(\boldsymbol{\mu} - \boldsymbol{p})'\mathbf{1}}{\gamma\sigma_\epsilon^2} + \theta(\boldsymbol{p})\boldsymbol{\beta}'\mathbf{1}. \tag{4}$$

A word on units is in order. KRY23 take CARA utility with normal payoffs (Admati, 1985), so $\boldsymbol{w}$ collects dollar positions and $\boldsymbol{p}$ is the price level; KY19 take log utility with log-normal returns (Campbell and Viceira, 2002), so $\boldsymbol{w}$ collects portfolio weights and $\boldsymbol{p}$ is log market equity. The leverage constraint of Section 3.1, $\mathbf{1}'\boldsymbol{w} \leq 1-\alpha$, reads accordingly: a ceiling on the total dollar position in the first case, on the fraction of wealth held in risky assets in the second. Nothing below turns on which reading one takes: the algebra is identical, and the two facts this note leans on — that the constraint multiplies a column of ones, so it shifts every asset alike, and that the coefficients on characteristics are scalars moving with the entire price vector — are properties of the form of demand, not of the units in which $\boldsymbol{w}$ and $\boldsymbol{p}$ are measured.

2.2 What Have We Learned So Far?

Five things follow from (3). First, demand is affine in $\boldsymbol{\mu} - \boldsymbol{p}$ plus a single loading on $\boldsymbol{\beta}$, with scalar $\theta(\boldsymbol{p})$. This scalar is endogenous—it can be read as the price of risk in this economy—because it depends on the *entire* price vector; yet it is common across assets. In its second line, with n indexing the N assets, the weight on asset n reads

$$w_n = \underbrace{-\frac{1}{\gamma\sigma_\epsilon^2}}_{\text{slope on own price}} p_n + \underbrace{\frac{1}{\gamma\sigma_\epsilon^2}}_{\text{exogenous scalar}} \mu_n - \underbrace{\theta(\boldsymbol{p})}_{\text{endogenous scalar}} \beta_n, \tag{5}$$

so a cross-sectional regression of w_n on its own price p_n, controlling for the expected payoff μ_n and the beta β_n, recovers the slope $-1/(\gamma\sigma_\epsilon^2)$.[2] The prices of other assets enter only through $\theta(\boldsymbol{p})$ — and because that scalar is the same for every n, it is absorbed into the coefficient on β_n rather than showing up as variation across the cross-section.

Second, one feature of that regression deserves emphasis: it compares assets at a given price vector. The variation is across assets, comparing w_m with w_n for two assets whose prices differ, $p_m \neq p_n$, while the entire vector $\boldsymbol{p}$ is unchanged. That is not the partial derivative with respect to p_n, holding the other assets' prices $\boldsymbol{p}_{-n}$ fixed — which does change the price vector.[3] The distinction matters for counterfactuals,

[2]Running it requires an instrument that moves prices — investment mandates in KY19. Instruments are not the center of this note, but what the exclusion restriction demands is taken up in Sections 5.1 and 5.2.

[3]van der Beck, Bretscher and Fu (2026) decompose the same system differently, writing each asset's demand against its price *relative* to a substitute portfolio — in the mean-variance case, the portfolio of the other assets that best replicates asset n. Their slope is then an own elasticity, governed by the part of asset n's payoff variance the other assets cannot hedge, $\sigma_D^2(n) = 1/\left[\boldsymbol{\Sigma}^{-1}\right]_{nn}$, rather than by σ_ϵ^2 alone. The fourth point below makes the same observation from the other side:

which are about how the entire vector $\boldsymbol{p}$ moves in response to a supply shock.

Third, although KY19 and KRY23 do not run the regression (3) suggests, their method (explained below) embodies its core idea: an endogenous scalar $\theta(\boldsymbol{p})$, common across assets yet absorbing the entire price vector. More generally, coefficients and intercepts that are constant across assets but functions of $\boldsymbol{p}$ are central to KY19, and that point carries through every step of this note.

Fourth, what is the economics of the slope on an asset's own price in KY19 and KRY23? In Haddad, He, Huebner, Kondor and Loualiche (2026, hereafter HHHKL26) we call it the relative elasticity, here $\mathcal{E}_{rel} = \frac{1}{\gamma\sigma_\epsilon^2}$: the own-price response net of the cross-price response, for two assets that differ in price but share the relevant characteristics. That it involves the idiosyncratic variance σ_ϵ^2 and not the factor variance σ_f^2 is what one should expect: two such assets differ only in idiosyncratic risk.

Fifth, $\mathcal{E}_{rel}$ is only one part of the elasticity matrix $\boldsymbol{\mathcal{E}}$; the rest of it is encoded in $\theta(\boldsymbol{p})$. Adding β_n as a control in a single cross-sectional regression like (5) absorbs $\theta(\boldsymbol{p})\beta_n$ but says nothing about how $\theta(\boldsymbol{p})$ varies with asset prices $\boldsymbol{p}$; the latter sits at the core of counterfactual analysis on the asset demand system.

3 Logit Form in KY19

This section works through how KY19 derive the logit form from the linear demand system. Unpacking the steps one at a time is part of the aim, so that readers can form their own intuitions and set them against what they already know of KY19.

3.1 Linear Demand in Characteristics

This section adds two things to the mean-variance problem: a constraint at the investor level, and a projection of expected payoffs onto observable characteristics. The projection is where $\boldsymbol{\mu}$ stops being a primitive, and that matters for what follows — once the coefficients of that projection may themselves move with prices, the linearity that mean-variance choice appears to deliver has no reason to keep holding.

Investor-level constraint KY19 introduce a leverage constraint at the investor level, $\mathbf{1}'\boldsymbol{w} \leq 1$. (Recall $\boldsymbol{w}$ holds the risky-asset weights and $w_0 = 1 - \mathbf{1}'\boldsymbol{w}$ the risk-free weight.) Short-sale constraints are also present in their paper, but are set aside here

what governs the slope is the variance the other assets cannot absorb.

for simplicity.[4] The asset demand then becomes

$$\boldsymbol{w}_{N\times 1} = -\frac{1}{\gamma\sigma_\epsilon^2}\boldsymbol{p}_{N\times 1} + \frac{1}{\gamma\sigma_\epsilon^2}\boldsymbol{\mu}_{N\times 1} - \underbrace{\theta\left(\boldsymbol{p}\right)}_{\text{scalar}}\boldsymbol{\beta}_{N\times 1} + \underbrace{\pi\left(\boldsymbol{p}\right)}_{\text{scalar}}\mathbf{1}_{N\times 1}, \tag{6}$$

where $\pi\left(\boldsymbol{p}\right)$ captures the Lagrange multiplier for the leverage constraint.[5]

Two points matter later. First, compared with (3), the new term in (6) multiplies a column of ones, so it shifts every asset by the same amount. That follows from the constraint sitting at the portfolio level;[6] a constraint on an individual asset — a short-sale constraint, say — carries one multiplier per asset and so cannot produce a constant common to all of them. Second, π is endogenous: like $\theta\left(\boldsymbol{p}\right)$ it depends on the entire price vector, which is why both are written as functions of $\boldsymbol{p}$.

One refinement will matter in Section 3.2. I work with a slightly tighter version, $\mathbf{1}'\boldsymbol{w} \leq 1-\alpha$ with $1-\alpha \in (0,1)$ a primitive, so that $w_0\left(\boldsymbol{p}\right) \geq \alpha > 0$: the investor must always keep a cash buffer of at least α. The strict bound is what lets the logit exist at all. Equation (10) divides by w_0 and then takes a logarithm, so w_0 must stay away from zero, whereas under the plain constraint $\mathbf{1}'\boldsymbol{w} \leq 1$ a binding investor holds $w_0 = 0$ and w_n/w_0 is not defined. Nothing else changes: λ is the multiplier on the tighter constraint, and binding now pins w_0 at α instead of at zero.

Characteristics-based KY19's Assumption 1 makes assets' expected payoff affine in observable characteristics, prices included. In my simplified framework the characteristics are the asset's own price, its beta, and the constant, the last following KY19's own convention that one characteristic is a column of ones.[7] Projecting the expected payoff $\boldsymbol{\mu}$ on them, and taking that projection to be exact, gives

$$\boldsymbol{\mu} = b_p\boldsymbol{p} + b_\beta\boldsymbol{\beta} + b_1\mathbf{1}, \tag{7}$$

[4]Adding short-sale constraint would change the form of the new term, because such a constraint hits an individual asset. With multipliers $\boldsymbol{\nu} \geq \mathbf{0}$ on $\boldsymbol{w} \geq \mathbf{0}$, optimal demand is $\boldsymbol{w} = \gamma^{-1}\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu} - \boldsymbol{p} - \lambda\mathbf{1} + \boldsymbol{\nu}\right)$, so what enters alongside $\pi\left(\boldsymbol{p}\right)\mathbf{1}$ is an asset-level vector $\gamma^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{\nu}$, not a scalar multiplying $\mathbf{1}$. Positions that bind at zero also sit badly with the logarithm Step 2 of Section 3.2 takes; see Kocherlakota (2025).

[5]Explicitly, if $\lambda \geq 0$ denotes the multiplier on $\mathbf{1}'\boldsymbol{w} \leq 1-\alpha$, then $\boldsymbol{w} = \gamma^{-1}\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu} - \boldsymbol{p} - \lambda\mathbf{1}\right)$, with $\pi\left(\boldsymbol{p}\right) = -\lambda\left(\boldsymbol{p}\right)/\left(\gamma\sigma_\epsilon^2\right) \leq 0$. (Opposite signs of π and λ only matter when we think about direction). Also, $\theta\left(\cdot\right)$ is evaluated at the shadow-adjusted excess payoff $\boldsymbol{\mu} - \boldsymbol{p} - \lambda\mathbf{1}$ rather than $\boldsymbol{\mu} - \boldsymbol{p}$ (see (3)) — the same structure as KY19's κ in their (A6). Either way θ remains a function of $\boldsymbol{p}$, since λ is.

[6]Not all constraints will; the key is that each position contributes equally in $\mathbf{1}'\boldsymbol{w} = \sum_{n=1}^{N} w_n \leq 1$.

[7]I simplify their Assumption 1 in two ways. First, theirs also covers the covariance matrix, which I leave out. Second, what KY19 project on is not the characteristic list itself but $\boldsymbol{y}_n$, the Mth-order polynomial expansion of that list — $\hat{\boldsymbol{x}}_n$, defined in Section 3.2, then all its products. The case $M = 1$ is harmless here, but Step 2 needs $M \to \infty$.

with b_p, b_β and b_1 three scalars; each subscript names the object its coefficient multiplies.[8] Then plugging (7) into (6) gives

$$\boldsymbol{w}_{N\times 1} = \underbrace{-\frac{1-b_p}{\gamma\sigma_\epsilon^2}}_{\mathcal{E}_{rel}}\boldsymbol{p}_{N\times 1} + \underbrace{\left[\frac{1}{\gamma\sigma_\epsilon^2}b_\beta - \theta\left(\boldsymbol{p}\right)\right]}_{\tilde{\theta}(\boldsymbol{p}),\text{ scalar}}\boldsymbol{\beta}_{N\times 1} + \underbrace{\left[\frac{1}{\gamma\sigma_\epsilon^2}b_1 + \pi\left(\boldsymbol{p}\right)\right]}_{\tilde{\pi}(\boldsymbol{p}),\text{ scalar}}\mathbf{1}_{N\times 1}. \tag{8}$$

Asset demand now has three parts: an own-price term whose coefficient is the relative elasticity $\mathcal{E}_{rel}$; a term linear in the characteristic $\boldsymbol{\beta}$, carrying the endogenous coefficient $\tilde{\theta}\left(\boldsymbol{p}\right)$; and the endogenous intercept $\tilde{\pi}\left(\boldsymbol{p}\right)$ on the constant characteristic.

Discussion on loadings in (7): endogenous coefficients or not? In (7), $b_p < 1$ lets an asset's expected payoff μ_n load on its own endogenous price p_n, which is why it enters the relative elasticity in (8): only the net price effect reaches demand. Odd as that looks, b_p reads naturally as the fraction of a price rise the market takes as news about payoffs — the learning channel of Section 5.1. The other two read the same way: b_β as compensation for factor risk, b_1 as a baseline payoff common to every asset.

The real content of KY19's projection lies elsewhere: in whether these coefficients are primitives. Nothing prevents b_p, b_β and b_1 from depending endogenously on the entire price vector $\boldsymbol{p}$ — and once they may, Section 3.3 shows that the logit needs that dependence to take one very particular form.

To anticipate the later discussion a bit, take b_1 as the leading case. Does the baseline expected payoff common to all assets move with the price vector $\boldsymbol{p}$? My own reading is that it should not: b_1 is a primitive, the intercept of expected payoffs once $\boldsymbol{p}$ and $\boldsymbol{\beta}$ are accounted for, and fixed as prices vary. But KY19 require only that b_1 be constant *across assets*, which opens the door to $b_1\left(\boldsymbol{p}\right)$ — and the logit form requires $b_1\left(\boldsymbol{p}\right)$ to take some particular form.

One could ask the same question of b_p and b_β. For that matter, the dependence of $b_p(\boldsymbol{p})$ on $\boldsymbol{p}$ and $b_\beta(\boldsymbol{p})$ on $\boldsymbol{\beta}$ are more reasonable,[9] but again as we will see shortly the logit form requires a very particular dependence of b_p and b_β on $\boldsymbol{p}$.

[8] In KY19's notation b_1 is their scalar intercept $\phi_{i,t}$, which feeds their demand intercept $\pi_{i,t}$ exactly as b_1 feeds $\tilde{\pi}\left(\boldsymbol{p}\right)$ here; b_p and b_β are entries of their slope vector $\boldsymbol{\Phi}_{i,t}$.

[9] That is, b_p can move through learning and b_β through the risk premium, both of which respond to the entire price vector as an aggregate state.

3.2 From Linear to Logit

Everything up to this point is linear, with endogenous coefficients — each a function of the entire price vector $\boldsymbol{p}$. KY19 do not stop there: the demand system they take to the data is a logit, delivered by their Corollary 1, and its coefficients are treated as primitives, invariant to $\boldsymbol{p}$. That step does not come from portfolio optimization: it adds assumptions to their Proposition 1, and the logit form is bought with them. Working from the corollary's proof in their online appendix, I isolate the two ingredients it requires.

Write (8) asset by asset. The portfolio weight on asset n is

$$w_n = -\mathcal{E}_{rel}\, p_n + \tilde{\theta}\left(\boldsymbol{p}\right)\beta_n + \tilde{\pi}\left(\boldsymbol{p}\right), \tag{9}$$

an affine function of that asset's own two characteristics, with coefficients that are common across assets even though each depends on the entire price vector. Corollary 1 of KY19 derives a logit form,

$$\frac{w_n}{w_0} = \exp\left(c_p p_n + c_\beta \beta_n\right) \Leftrightarrow \log\frac{w_n}{w_0} = c_p p_n + c_\beta \beta_n. \tag{10}$$

The coefficients c_p and c_β are taken as primitives, and (13) below collects them as $\boldsymbol{c}$.[10] Two steps lead from (9) to (10).

Step 1 The first is to assume that

$$\tilde{\pi}\left(\boldsymbol{p}\right) = w_0\left(\boldsymbol{p}\right), \tag{11}$$

which sets the intercept of the linear system equal to the portfolio weight on the outside (risk-free) asset given by (4). Note the resemblance between the two: both are common across assets, and both move with the entire price vector; but Step 1 requires exact equality. The equality could be relaxed to $\tilde{\pi}\left(\boldsymbol{p}\right) = C_0 w_0\left(\boldsymbol{p}\right)$ with C_0 a primitive,[11] but the key logic explained in Section 3.3 does not change.

Granting it, divide (9) through by w_0:

$$\frac{w_n}{w_0} = 1 - \frac{\mathcal{E}_{rel}}{w_0\left(\boldsymbol{p}\right)} p_n + \frac{\tilde{\theta}\left(\boldsymbol{p}\right)}{w_0\left(\boldsymbol{p}\right)}\beta_n, \tag{12}$$

[10] The latent demand term ϵ is suppressed, since estimation is not the focus of this note.

[11] This amounts to including a column of ones in characteristics $\boldsymbol{x}$, as KY19 do: with $C_0 = e^{c_0}$, then (15), toward the end of Step 2, becomes $w_n/w_0 = C_0 \exp\left(\hat{\boldsymbol{x}}_n'\boldsymbol{c}\right) = \exp\left(\hat{\boldsymbol{x}}_n'\boldsymbol{c} + c_0\right)$.

so that the intercept is now exactly one — as Step 2 will show, that is what the exponential series needs.

Two features of (12) are worth recording before the second ingredient arrives. First, the coefficient on an asset's own price is no longer $-\mathcal{E}_{rel}$ as in (9) but $-\mathcal{E}_{rel}/w_0$, the structural slope divided by the cash weight. Second, the coefficients in front of p_n and β_n are common across assets, so every bit of cross-sectional variation in w_n/w_0 comes from asset characteristics (prices and betas).

Step 2 The second ingredient turns the right-hand side of (12), which is of the form $1+\cdots$, into an exponential (hence logit). In the proof of their Corollary 1 (appendix, p. 4) KY19 do this exactly rather than by approximation — but the exactness is purchased with a functional-form assumption on the characteristic set and on the coefficients that multiply it.

Following KY19, let $\boldsymbol{x}_n$ hold the characteristics other than price — here the beta and the constant — and let $\hat{\boldsymbol{x}}_n = (p_n,\ \beta_n)'$ add the price.[12] In that notation (9) reads $w_n = \hat{\boldsymbol{x}}_n'\boldsymbol{\Pi}(\boldsymbol{p}) + \tilde{\pi}(\boldsymbol{p})$ with slope vector $\boldsymbol{\Pi}(\boldsymbol{p}) = \left(-\mathcal{E}_{rel},\ \tilde{\theta}(\boldsymbol{p})\right)'$, and (12) reads $w_n/w_0 = 1 + \hat{\boldsymbol{x}}_n'\boldsymbol{c}$, where

$$\boldsymbol{c} \equiv \frac{\boldsymbol{\Pi}(\boldsymbol{p})}{w_0(\boldsymbol{p})} = \left(-\frac{\mathcal{E}_{rel}}{w_0(\boldsymbol{p})},\ \frac{\tilde{\theta}(\boldsymbol{p})}{w_0(\boldsymbol{p})}\right)' = (c_p,\ c_\beta)' \tag{13}$$

is nothing new — it is the pair of logit coefficients already in (10).

The right-hand side $1+\hat{\boldsymbol{x}}_n'\boldsymbol{c}$ already looks like the opening of an exponential series; KY19 then impose further structure to make it exactly that. Corollary 1 asks for two things. First, what enters the demand system is not $\hat{\boldsymbol{x}}_n$ but its polynomial expansion up to order M: $\boldsymbol{y}_n = (\hat{\boldsymbol{x}}_n,\ \mathrm{vec}(\hat{\boldsymbol{x}}_n\hat{\boldsymbol{x}}_n'),\ \ldots)'$. Second, the slope on each block is the matching Taylor coefficient of the single index $\hat{\boldsymbol{x}}_n'\boldsymbol{c}$:

$$\frac{\boldsymbol{\Pi}}{w_0} = \left(\boldsymbol{c},\ \frac{1}{2}\mathrm{vec}(\boldsymbol{c}\boldsymbol{c}'),\ \ldots,\ \frac{1}{m!}\mathrm{vec}\left(\boldsymbol{c}^{\otimes m}\right),\ \ldots\right)'. \tag{14}$$

This is a different projection from (7), and the coefficients on the higher-order blocks are not free but pinned to products of the first-order ones. The linear system is then

[12] This is KY19's split. Their $\boldsymbol{x}_t(n)$ is the list of observed characteristics *excluding* price — log book equity, profitability, investment, market beta, and constant — while price enters only through the augmented list $\hat{\boldsymbol{x}}_{i,t}(n)$, which stacks log market equity, their $\boldsymbol{x}_t(n)$, and latent demand (unobserved by the econometrician). My $\hat{\boldsymbol{x}}_n$ plays the role of their $\hat{\boldsymbol{x}}_{i,t}(n)$ with latent demand suppressed.

Table 1: Response of the two sides of (11) to a small uniform price rise $d\boldsymbol{p} = \delta\mathbf{1}$.

	$d\tilde{\pi}/d\delta$	$dw_0/d\delta$
Leverage constraint binds	$\mathcal{E}_{rel} = \frac{1-b_p}{\gamma\sigma_\epsilon^2} > 0$	0
Leverage constraint slack	0	$(1-b_p)\,\gamma^{-1}\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1} > 0$

literally the truncated series,

$$\frac{w_n}{w_0} = 1 + \boldsymbol{y}_n' \frac{\boldsymbol{\Pi}}{w_0} = \sum_{m=0}^{M} \frac{(\hat{\boldsymbol{x}}_n'\boldsymbol{c})^m}{m!} \longrightarrow \exp(\hat{\boldsymbol{x}}_n'\boldsymbol{c}) \quad \text{as } M \to \infty, \tag{15}$$

which is their Corollary 1.[13] KY19 are explicit that it is adopted, not derived: the restriction is imposed, in their words, "as a matter of specification." In fact Davis (2025) gets cosine demand from a different restriction on the same expansion.

3.3 How Restrictive Are These Two Steps?

Step 1 This step asks two functions of the endogenous price vector to coincide. Using (8), (11) reads

$$w_0(\boldsymbol{p}) = \tilde{\pi}(\boldsymbol{p}) = \frac{b_1}{\gamma\sigma_\epsilon^2} + \pi(\boldsymbol{p}). \tag{16}$$

If b_1 were a primitive, the two sides would generically fail to agree, as I now show. Intuitively, $\pi(\boldsymbol{p})$ carries the multiplier on the leverage constraint, and that constraint binds exactly when $w_0(\boldsymbol{p}) = \alpha$. Complementary slackness then keeps the two from moving together: whichever side responds to prices, the other is pinned.

To see this formally, raise every price by the same small amount $d\boldsymbol{p} = \delta\mathbf{1}$. Exactly one side of (11) responds, and which side depends on the regime. Where the constraint is slack, $\pi = 0$, so $\tilde{\pi} = \frac{b_1}{\gamma\sigma_\epsilon^2}$ is constant while the risk-free asset absorbs the rise. Where it binds, $w_0 = \alpha$ at every price vector, while the Lagrange multiplier (opposite sign of $\pi(\boldsymbol{p})$) must fall by $(1-b_p)\,\delta$ to hold $\mathbf{1}'\boldsymbol{w}$ at $1-\alpha$: a uniform price rise and an increase in the multiplier are the same perturbation of $\boldsymbol{\mu} - \boldsymbol{p}$. So $\tilde{\pi}$ climbs at rate $\mathcal{E}_{rel}$.[14] In sum, as Table 1 records, only one side is ever active, so the two cannot agree on any interval of prices—they may cross at isolated points but (11) asks for an identity.

Now suppose b_1 — and for that matter b_p and b_β — were allowed to move with

[13] My simplified model already fails this—(9) has $M = 1$ without squared or interaction terms.

[14] With $\mathbf{1}'\boldsymbol{w} = 1-\alpha$ binding and λ its Lagrange multiplier, $d\boldsymbol{p} = \delta\mathbf{1}$ moves $\boldsymbol{\mu} - \boldsymbol{p} - \lambda\mathbf{1}$ by $[-(1-b_p)\,\delta - d\lambda]\,\mathbf{1}$ under (7), and holding $\mathbf{1}'\boldsymbol{w}$ fixed forces $d\lambda = -(1-b_p)\,\delta$. By footnote 5, π rises by $\mathcal{E}_{rel}\delta$.

prices. Nothing in KY19's Assumption 1 forbids it: the assumption disciplines the cross-section only, and *common across assets* is not *constant across different sets of price vectors*. This is how it works.

Where the constraint is slack, $\pi = 0$ and (11) is one linear equation in b_1 whose coefficient is constant; with its right-hand side being affine in $\boldsymbol{p}$, we obtain a solution $b_1(\boldsymbol{p})$ to be affine in $\boldsymbol{p}$ as well. When the constraint binds, $w_0(\boldsymbol{p}) = \alpha$ and $b_1(\boldsymbol{p})$ alone can no longer do the job; the role passes to $b_\beta(\boldsymbol{p})$, which can be chosen so that $\tilde{\pi} = \alpha = w_0$ at every price vector.[15] Either way, some coefficient of the payoff projection has to move with prices.

To be clear, the simplified model here is only meant to show how the Step 1 restriction in (16) connects to the price-dependent coefficients in (7) that KY19's Assumption 1 permits. They are what make KY19 flexible, and this should remind readers of the phenomenon met earlier with $\theta(\boldsymbol{p})$ in Section 2.

The two price dependencies are not of the same kind, though. That $\theta(\boldsymbol{p})$ moves with prices is an *outcome* of the factor model: $\boldsymbol{p}$ enters the first-order condition through $\boldsymbol{\mu} - \boldsymbol{p}$, so it sits inside θ in (3) whether one wants it there or not, and that is precisely how the factor structure delivers substitution across assets. Holding the payoff coefficients fixed, as Section 4 does in deriving (20), is then the ordinary assumption that beliefs are primitives. The logit asks for the opposite trade: hold the demand coefficients $\boldsymbol{c}$ fixed, and some payoff coefficient has to bend until (16) holds.

Where this bites depends on the variation one uses. In the cross-section the price-dependent terms are common to every asset and do no harm; the KY19 approach is powerful precisely there. What it cannot supply is the economics: $b_1(\boldsymbol{p})$ and $b_\beta(\boldsymbol{p})$ are backed out from the very equation one wants to hold. The problem surfaces in the time series, where the counterfactual questions live: one would need to know whether and why the expected payoff shifts with prices when a shock hits, and a b_1 or b_β constructed this way is silent on that.

Step 2 This step imposes a restriction of a different kind, though similar in spirit: it too requires something that moves with prices to stay fixed. By (13) the logit coefficients $\boldsymbol{c} = \boldsymbol{\Pi}(\boldsymbol{p})/w_0(\boldsymbol{p})$ are a ratio of two objects that both move with prices, and (14) then ties every higher-order coefficient to that same ratio. Reading the slope

[15] By (6), b_1 and the Lagrange multiplier λ enter demand only through $b_1 - \lambda$. When the constraint is slack, $\lambda = 0$ and solving (11) gives $b_1(\boldsymbol{p}) = \frac{1-\gamma^{-1}b_\beta \mathbf{1}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta} + \gamma^{-1}(1-b_p)\mathbf{1}'\boldsymbol{\Sigma}^{-1}\boldsymbol{p}}{(\gamma\sigma_\epsilon^2)^{-1} + \gamma^{-1}\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1}}$. When it binds, fixing everything else $\mathbf{1}'\boldsymbol{w} = 1 - \alpha$ already pins down $b_1 - \lambda$, so raising b_1 raises λ one for one and leaves $\tilde{\pi}$ untouched — the multiplier absorbs b_1 completely. But we can vary b_β; fixing b_1 and solving $\mathbf{1}'\boldsymbol{w} = 1 - \alpha$ with $b_1 - \lambda = \alpha\gamma\sigma_\epsilon^2$, we have $b_\beta(\boldsymbol{p}) = \frac{\gamma(1-\alpha) + (1-b_p)\mathbf{1}'\boldsymbol{\Sigma}^{-1}\boldsymbol{p} - \alpha\gamma\sigma_\epsilon^2\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1}}{\mathbf{1}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}}$, valid for any $b_1 \geq \alpha\gamma\sigma_\epsilon^2$ (so $\lambda \geq 0$).

vector off (8),

$$\boldsymbol{c} = \frac{\boldsymbol{\Pi}(\boldsymbol{p})}{w_0(\boldsymbol{p})} = \frac{1}{w_0(\boldsymbol{p})}\left(-\frac{1-b_p}{\gamma\sigma_\epsilon^2},\ \ \frac{b_\beta}{\gamma\sigma_\epsilon^2} - \theta(\boldsymbol{p})\right)', \tag{17}$$

in which $w_0(\boldsymbol{p})$ and $\theta(\boldsymbol{p})$ both move with the price vector. For $\boldsymbol{c}$ to be the fixed pair $(c_p,\ c_\beta)'$ of (13), those movements have to cancel — and they can be made to, by letting the payoff coefficients absorb them:

$$b_p(\boldsymbol{p}) = 1 + \gamma\sigma_\epsilon^2\, c_p\, w_0(\boldsymbol{p}), \qquad b_\beta(\boldsymbol{p}) = \gamma\sigma_\epsilon^2\left[c_\beta w_0(\boldsymbol{p}) + \theta(\boldsymbol{p})\right].$$

So a $b_p(\boldsymbol{p})$ and a $b_\beta(\boldsymbol{p})$ making (17) hold exactly always exist. Together with the slack-regime $b_1(\boldsymbol{p})$ of footnote 15, these pin down the belief function the logit implies,

$$\boldsymbol{\mu}(\boldsymbol{p}) = b_p(\boldsymbol{p})\,\boldsymbol{p} + b_\beta(\boldsymbol{p})\,\boldsymbol{\beta} + b_1(\boldsymbol{p})\,\mathbf{1}.$$

What expected payoffs that function describes is worth knowing, and one simple case solves in closed form. Switch the factor off, $\boldsymbol{\beta} = \mathbf{0}$, so that $\boldsymbol{\Sigma} = \sigma_\epsilon^2\boldsymbol{I}$ and $\theta(\boldsymbol{p}) = 0$, and keep the leverage constraint slack; Step 1 and (13) then give $b_1 = \gamma\sigma_\epsilon^2/(N + 1 + c_p\mathbf{1}'\boldsymbol{p})$ and $b_p = 1 + c_p b_1$. Prices enter only through the level $\mathbf{1}'\boldsymbol{p}$, and since $c_p = -\mathcal{E}_{rel}/w_0 < 0$, b_1 *rises* with that level: the investor must expect more from every asset exactly when the market has become more expensive.[16] And (14) asks every higher-order coefficient to line up as well.

Ultimately, any structural model, including logit here, needs some policy-invariant primitives. In KY19, $\boldsymbol{c}$ is exactly what has to be held fixed for the system to be used for policy. With $\boldsymbol{s}$ the shares outstanding, $\boldsymbol{x}$ the characteristics, $\boldsymbol{c}$ the matrix of coefficients on characteristics and $\boldsymbol{\epsilon}$ latent demand, KY19's market clearing defines price implicitly as (a simplified version of KY19's equation (22), p. 1502)

$$\boldsymbol{p}(\boldsymbol{s}, \boldsymbol{x}, \boldsymbol{c}, \boldsymbol{\epsilon}).$$

The coefficients $\boldsymbol{c}$ enter as an argument here: prices move, they do not.

Summary The net outcome of the two steps is the key logit specification (10), which imposes strong structural assumptions on investors' asset demand that are at odds with the standard way mean-variance optimization is often set up. None of this

[16] I owe this calculation to Marco Loseto.

is special to the one-factor case.[17] KY19 are aware of it; they state that their model can be used for policy experiments "only under the null that it is a structural model of asset demand that is policy invariant" (p. 1487), and they set the alternative aside as the Lucas critique: that the coefficients "ultimately capture beliefs or constraints that change with policy."

4 IO Techniques in Finance

Why, then, is the logit everywhere? Section 3 asked what reaching the logit assumes; this section asks what it buys. The answer, and the cost, are both about substitution — how a portfolio is rearranged when a single asset's price moves. Industrial organization met that question long before finance did, and with the same functional form, so what follows sets the two side by side.

4.1 Why Is Logit So Convenient?

The logit is ubiquitous in the IO literature for one reason: paired with observed portfolio shares, it reduces the entire $N \times N$ elasticity matrix to a single parameter. Given $w_n/w_0 = \exp(\hat{\boldsymbol{x}}_n'\boldsymbol{c})$ and $w_0 + \sum_n w_n = 1$,

$$w_n = \frac{\exp(\hat{\boldsymbol{x}}_n'\boldsymbol{c})}{1+\sum_{m=1}^{N}\exp(\hat{\boldsymbol{x}}_m'\boldsymbol{c})}, \qquad w_0 = \frac{1}{1+\sum_{m=1}^{N}\exp(\hat{\boldsymbol{x}}_m'\boldsymbol{c})},$$

so that, with c_p the price coefficient of (10),

$$\frac{\partial \log w_n}{\partial p_m} = c_p\left(\mathbb{1}_{n=m} - w_m\right), \qquad \text{that is} \qquad \frac{\partial \log \boldsymbol{w}}{\partial \boldsymbol{p}'} = c_p\left(\boldsymbol{I}_{N\times N} - \mathbf{1}\boldsymbol{w}'\right). \tag{18}$$

Scaling row n by w_n turns this into the elasticity matrix of Section 1,

$$\boldsymbol{\mathcal{E}} = -\frac{\partial \boldsymbol{w}}{\partial \boldsymbol{p}'} = -c_p\left[\operatorname{diag}(\boldsymbol{w}) - \boldsymbol{w}\boldsymbol{w}'\right]. \tag{19}$$

One scalar, together with holdings the researcher already observes, pins down every entry. The other $K-1$ coefficients never appear: they set the *level* of the weights, and so matter for predicting holdings under new characteristics, but given those holdings the elasticities need only c_p. One cross-sectional regression of $\log(w_n/w_0)$ on characteristics therefore delivers the whole matrix.

[17] Proposition A4 of HHHKL26: among all *stable* factor models — $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ functions of observables but not of $\boldsymbol{p}$ — logit demand is generically not a valid approximation to factor demand.

The convenience and the restriction are the same object. In levels the off-diagonal response is $\partial w_n/\partial p_m = -c_p w_n w_m$: when one asset's price moves, the investor reallocates across everything else in proportion to existing holdings, and in no other pattern. Equivalently $w_n/w_m = \exp\left((\hat{\boldsymbol{x}}_n - \hat{\boldsymbol{x}}_m)'\,\boldsymbol{c}\right)$ is untouched by any third asset's price, which is independence of irrelevant alternatives.

Why is this restrictive? Under standard mean-variance logic, investors substitute toward assets that are close in risk.[18] Among stocks, an investor holding Ford, GM and Nvidia, the last two in similar weights, sells Ford as its price rises; logit splits the proceeds evenly between GM and Nvidia, while mean-variance logic would send most of them to GM, the closer substitute in risk. In the bond market, an investor whose 30-year Treasury richens should be glad to switch into the 29-year bond, all but the same asset — yet if that position starts at zero, logit moves nothing into it, because the cross-response is proportional to w_n. And the near-zero holding is no accident: it is precisely because the two bonds are such close substitutes that there was no reason to hold both in the first place. Holdings-proportional substitution is therefore least reliable exactly where substitution is strongest; it cannot open a position that is not already open.

None of this makes the logit the wrong demand system; it makes the mean-variance factor model an uncomfortable place to derive it.

4.2 Where Does the Literature Stand?

KY19 was the pioneering step, not the last one. The limits of the logit are well understood in industrial organization, where the standard remedies accommodate much richer substitution. In finance, Fuchs, Fukuda and Neuhann (2025) have put it at the centre: because a logit ties substitution to portfolio shares alone, assets holding equal shares must substitute identically, and the complementarities of portfolio choice are absent. This bites at any degree of imperfect substitutability.

Nested logit, BLP, and others Nested logit groups the alternatives: substitution stays proportional to holdings within a nest and is damped across nests, which relaxes independence of irrelevant alternatives but leaves the pattern holdings-driven, merely block by block. Random-coefficient logit, in the manner of Berry, Levinsohn and Pakes (1995), instead lets tastes for characteristics differ across agents, so that two

[18] Whether investors in fact substitute this way is an empirical question, which I take up later in this section.

alternatives compete most closely when they are near each other in characteristic space rather than when they are large.

Finance has begun to import both. Koijen and Yogo (2024) build a global demand system in which the nesting across countries and asset classes carries much of the weight. Chaudhary, Fu and Li (2026) estimate a two-tier structure of just this kind for corporate bonds: once heterogeneous substitutability is allowed for, security-level multipliers are essentially zero, while portfolio multipliers rise monotonically with the level of aggregation. Bretscher, Schmid, Sen and Sharma (2026) likewise split corporate bonds by rating, so that substitution runs within and across the investment-grade and high-yield segments. Viewing a sector's holdings as an aggregate of heterogeneous logit agents, Fang and Xiao (2025) add random coefficients to a KY19 demand system for US bond holdings. This delivers cross-elasticity patterns that allow for transmission — substitution between Treasuries and mortgage-backed securities in particular.

HHHKL26: Factor structure HHHKL26 take a different route. In essence, by taking seriously the factor structure that is widely used in the finance literature, they keep asset demand linear as in equation (9) and estimate the matrix $\boldsymbol{\mathcal{E}}$. Substitution that lines up with risk is exactly what a factor structure delivers, since reallocation then runs through $\boldsymbol{\beta}$ rather than through the size of existing positions.

Following the standard assumption that b_p, b_β and b_1 are primitive parameters, and taking the leverage constraint to be slack so that $\tilde{\pi}$ does not move with $\boldsymbol{p}$,[19] differentiating (8) gives

$$\boldsymbol{\mathcal{E}} = -\frac{\partial \boldsymbol{w}}{\partial \boldsymbol{p}'} = \underbrace{\frac{1-b_p}{\gamma\sigma_\epsilon^2}\boldsymbol{I}}_{\text{relative elasticity}} - \underbrace{\boldsymbol{\beta}\frac{\partial \tilde{\theta}}{\partial \boldsymbol{p}'}}_{\text{substitution}} = \mathcal{E}_{rel}\boldsymbol{I} + \underbrace{\boldsymbol{\beta}}_{N\times 1}\underbrace{\mathcal{E}_X}_{\text{scalar}}\underbrace{\boldsymbol{\beta}'}_{1\times N}, \tag{20}$$

where $\mathcal{E}_X = -\frac{(1-b_p)\sigma_f^2}{\gamma\sigma_\epsilon^2\left(\sigma_\epsilon^2+\sigma_f^2\boldsymbol{\beta}'\boldsymbol{\beta}\right)} < 0$ is a primitive scalar, no longer a function of $\boldsymbol{p}$. All cross-asset substitution therefore runs through $\mathcal{E}_X$ on the outer product $\boldsymbol{\beta}\boldsymbol{\beta}'$, an $N \times N$ matrix of rank one. Off the diagonal $\mathcal{E}_{nm} = \mathcal{E}_X\beta_n\beta_m$, so the cross-effect takes the sign of $-\beta_n\beta_m$: two assets are substitutes when their betas share a sign and complements when they do not. This simple model is the case $K = 1$: one factor, hence one loading vector $\boldsymbol{\beta}$. HHHKL26 introduce K-dimensional observables to model

[19] Slackness is for convenience only. In HHHKL26 (Proposition A2) we let linear constraints bind and show that the elasticity matrix keeps the form of (20), a relative-elasticity term plus a factor term, once the characteristics entering the binding constraints are added to the observables.

substitution, raising that rank to K and reducing the problem from $N \times N$ to $K \times K$. Seen through (20), that move is structure on $\tilde{\theta}(\boldsymbol{p})$. The whole substitution block is $\boldsymbol{\beta}\, \partial\tilde{\theta}/\partial\boldsymbol{p}'$, so modelling substitution is modelling how the coefficient moves with the price vector; letting that gradient load on the K observables is a parsimonious way to model the potential price dependence.

Set beside (19), one of the key differences from (20) is a parameter count: the factor route of (20), which HHHKL26 take, carries two — one for the relative elasticity, one for substitution (because $K = 1$) — while the logit collapses both into the single coefficient c_p.

Working with (20), HHHKL26 stress that the relative elasticity $\mathcal{E}_{rel}$, which captures the substitution between assets with identical betas,[20] is then identified from the cross-section, while substitution across betas, captured by $\mathcal{E}_X$, requires time-series variation in the risk premium. HHHKL26 call this obstacle the missing-coefficient problem.[21] Because the effect of substitution is itself proportional to the observables, a cross-sectional regression like (3) absorbs it into the very slope coefficients it estimates — the price-dependent $\tilde{\theta}(\boldsymbol{p})$ among them — leaving it indistinguishable from a shift in demand. Where the only observable is a constant this is the familiar missing-intercept problem; with other observables (say β), which enter linearly, the missing-coefficient problem is the general form of it.

Cross-sectional identification of $\mathcal{E}_{rel}$ has a limit of its own. Fuchs, Fukuda and Neuhann (2025) press the basic no-arbitrage force: as two assets approach perfect substitutes their relative elasticity diverges, and the bias can diverge with it. In the notation here that limit is $\sigma_\epsilon = 0$, where no arbitrage ties the two prices together exactly and no relative price movement is left to exploit. This framework never reaches it — assets always differ a little through idiosyncratic risk, $\sigma_\epsilon > 0$ — and HHHKL26 record the case as a weak first stage: a control asset should be similar enough to satisfy the symmetry assumptions but not identical, since it is idiosyncratic risk that lets relative prices move without creating arbitrage. Their wider warning stands: when the data generating process is not exactly right and assets are close substitutes, the potential bias is large, and an estimated elasticity has to be treated with corresponding care.

[20] The economic content of $\mathcal{E}_{rel}$ is visible in (20): when $\beta_n = \beta_m$ the $\boldsymbol{\beta}\boldsymbol{\beta}'$ term contributes the same amount to $\mathcal{E}_{nn}$ and to $\mathcal{E}_{nm}$ and so cancels in the difference, leaving $\mathcal{E}_{nn} - \mathcal{E}_{nm} = \mathcal{E}_{rel}$ exactly.

[21] One could run cross-sectional regressions only, assuming the coefficients $\tilde{\theta}(\boldsymbol{p})$ are invariant to policy. Constant coefficients are the convention well beyond KY19; for instance, the parametric portfolio policies of Brandt, Santa-Clara and Valkanov (2009) fix them across assets and dates. For an asset demand system this is where the Lucas critique bites: coefficients that are not primitives move when policy moves prices, which is just what writing $\tilde{\theta}(\boldsymbol{p})$ records.

Evidence: substitution runs along risk If substitution runs through $\boldsymbol{\beta}$, as the factor term in (20) implies, then reallocation, and the price responses it produces, should line up with risk rather than with the size of existing positions. When the overall supply of Treasuries changes, the response of yields lines up with maturity even though the composition of the supply change does not (Greenwood and Vayanos, 2014). In the euro area the cross-country response of bond yields to central bank purchases tracks the riskiness of each country, even though the purchases are close to proportional to country size: in some interventions the ECB bought no Greek debt at all, yet Greek spreads fell by more than any other country's (Haddad, Moreira and Muir, 2024). Under substitution proportional to holdings, a small and untouched position should barely have moved. Substitution in the data runs along risk, which is precisely the dimension a logit elasticity matrix cannot see.

Demand systems with this factor structure have also been estimated directly: HHHKL26 apply theirs to corporate bond multipliers, and at least two other papers follow the same route. Under the framework of Vayanos and Vila (2021), Jansen, Li and Schmid (2024) let preferred-habitat investors substitute across safe assets in the Treasury market, and so estimate a linear demand system with an elasticity matrix rather than a logit. They find strong substitution across maturities, identified with separate instruments for a bucket's own yield and for the yields of the others. More recently, Davis, Kargar, Li and Silva (2026) carry HHHKL26's factor-plus-own decomposition into a dynamic setting, applying it at each demand horizon, and the two pieces come apart: stock-specific demand is nearly myopic, while substitution through portfolio exposures is forward-looking.

How large the substitution term is, relative to $\mathcal{E}_{rel}$, remains an empirical question, and the answer differs across markets. Accounting for substitution is quantitatively important in corporate bond, Treasury, and currency markets (Chaudhary, Fu and Li, 2026; Jansen, Li and Schmid, 2024; An and Huber, 2024). In the cross section of US stocks, by contrast, Chaudhry and Davis (2026) find substitution two to three orders of magnitude weaker than a frictionless mean-variance investor's: when a stock becomes cheaper, investors buy it while selling little of its close comovers, and cross-price elasticities are two to four orders of magnitude below own-price ones.[22]

What is the criterion? Which model is "more" right is not settled by fit in sample, but by out-of-sample performance, by conceptual coherence, and by whether the

[22]Small as each cross-price elasticity is, it applies to every other stock, so summed across the market the cross effects are not negligible.

model captures the institutional forces that actually shape asset demand. Ptolemy's epicycles matched the observed planetary motions well, and could always be made to match better by adding one more; Copernicus, at first, fit no better at all. A specification that can always be arranged to reproduce what is observed is not thereby vindicated, which is the force of the argument above: a demand system whose coefficients are free to absorb prices will fit any cross-section and still say nothing about the counterfactual. The question worth asking is not whether a demand system fits, but whether the coefficients it delivers still mean the same thing once prices move. That is the Lucas critique in demand-system form.

5 Beyond Substitution

Three issues come up often in discussions of this literature, and I take them up briefly here. The first is what an instrument has to be when investors learn from prices; the second is what happens when the characteristics themselves — betas, volatilities, covariances — are endogenous in a dynamic setting; the third is whether any of this can be identified without imposing structure in the first place. The following are what I learned in the last two years.

5.1 Learning: What Does an Instrument Require?

Learning is often thought to sit badly with instrumental variables: if prices convey information, whatever moves prices also moves beliefs, and nothing seems excludable from demand. The worry is real but usually misplaced, and seeing why says exactly what an instrument has to be. The argument that follows is from Appendix B.6.1 of HHHKL26, which generalizes the single-asset derivation in Haddad, Huebner and Loualiche (2025).

Following Admati (1985), consider N risky assets with payoff $\boldsymbol{m} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_m)$ and price vector $\boldsymbol{p}$. A continuum of competitive CARA traders with risk aversion γ each observe a private signal $\boldsymbol{v}_i = \boldsymbol{m} + \boldsymbol{\zeta}_i$ together with the price $\boldsymbol{p}$. Noise traders supply $\boldsymbol{u} \sim N(\boldsymbol{s}, \boldsymbol{\Sigma}_u)$, with a mean that matters below: investors know $\boldsymbol{s}$ (public news) but not the shock $\boldsymbol{u} - \boldsymbol{s}$. CARA-normal optimization and symmetry (so $\boldsymbol{\mathcal{E}}_i = \boldsymbol{\mathcal{E}}$ and $\boldsymbol{A}_i = \boldsymbol{A}$) give the individual investor's demand curve for assets

$$\boldsymbol{d}_i = \frac{\boldsymbol{\Sigma}_{m|v,p}^{-1}}{\gamma}\left(\underbrace{\mathbb{E}\left[\boldsymbol{m} \mid \boldsymbol{v}_i, \boldsymbol{p}\right]}_{\text{learning}} - \underbrace{\boldsymbol{p}}_{\text{budget}}\right) = -\boldsymbol{\mathcal{E}}\boldsymbol{p} + \underbrace{\boldsymbol{A}\boldsymbol{v}_i}_{\epsilon_i,\ \text{latent demand}}. \tag{21}$$

The elasticity matrix $\boldsymbol{\mathcal{E}}$ reflects how individual agents react to the equilibrium price $\boldsymbol{p}$: the standard budget effect and the learning effect. So learning does not destroy the linear demand system; it determines its coefficients. In Haddad, Huebner and Loualiche (2025) it does more still: because price informativeness is governed by the aggregate elasticity, information acquisition makes each investor's slope depend on everyone else's.

Now identification. Recovering $\boldsymbol{\mathcal{E}}$ needs variation in $\boldsymbol{p}$ uncorrelated with $\boldsymbol{A}\boldsymbol{v}_i$, the latent demand for investor i—that is, with what investors know. The noise-trader supply *shock* $\boldsymbol{u} - \boldsymbol{s}$—liquidity-driven sales or fund redemptions, say—is a natural candidate. And it works: $\boldsymbol{u} - \boldsymbol{s}$ moves $\boldsymbol{p}$ but is uncorrelated with $\boldsymbol{v}_i$ by assumption (Haddad, Huebner and Loualiche, 2025).[23]

But the exclusion restriction here is informational, not merely statistical: what matters is not that the shock is non-fundamental, but that investors do not know its origin. That cuts against some of the most obvious candidates. A shock investors *know* to be non-fundamental need not be valid: say the Federal Reserve announces a random purchase, or an index publishes a rebalancing. Both are better interpreted, in my model, as changes in $\boldsymbol{s}$ rather than as realizations of the noise shock $\boldsymbol{u}-\boldsymbol{s}$. The distinction is important: investors who recognize them as non-fundamental no longer read the price move as news, and so respond more elastically than (21) implies.

Letting $\boldsymbol{s}$ vary *within* the model, rather than holding it as a parameter, makes the contrast sharp. Investors know $\boldsymbol{s}$, so high supply explains away part of a low price: less of it is attributed to $\boldsymbol{m}$, and demand at any given price is higher. Writing that positive loading as $\boldsymbol{C} > 0$,

$$\boldsymbol{d}_i = -\boldsymbol{\mathcal{E}}\boldsymbol{p} + \underbrace{\boldsymbol{A}\boldsymbol{v}_i + \boldsymbol{C}\boldsymbol{s}}_{\text{intercept}}, \tag{22}$$

with (21) the special case of $\boldsymbol{s}$ being fixed at $\boldsymbol{0}$: $\boldsymbol{\mathcal{E}}$ and $\boldsymbol{A}$ are the same in both. A change in $\boldsymbol{s}$ therefore moves the intercept; $\boldsymbol{u}-\boldsymbol{s}$ does not, since investors cannot condition on what they do not observe. Hence $\boldsymbol{u}-\boldsymbol{s}$ works not by being non-fundamental but by being indistinguishable from fundamental movement, leaving investors nothing but their priors; fundamentalness is not the criterion, the information set is. That statistical orthogonality is not sufficient, and that the information set belongs in the exclusion restriction, is a point highlighted by Bastianello (2026), who explores it in a dynamic context and draws out an implication worth noting: the coarser investors'

[23] In rational expectation equilibrium $\boldsymbol{p}$ is a linear combination of $\boldsymbol{m}$ and noise supply $\boldsymbol{u}$; see HHHKL26 Appendix B.6.1.

information, the wider the set of valid instruments.

5.2 Dynamic GE with Endogenous Characteristics

Unlike the goods of the industrial-organization literature, an asset's characteristics — its beta, its volatility, its covariance with the market — are themselves endogenous. A single-asset dynamic economy is enough to see the difficulty: repeat the static model of Section 2 at date $t+1$ and the date-$t+1$ payoff is p_{t+1}, precisely the object the model is meant to determine. He, Kondor and Li (2026) work this through: investors retrade, so demand depends on the joint distribution of current and future returns, and a residual supply shock changes that distribution as investors absorb it. The shock moves the very demand curve it is meant to trace out. van Binsbergen, David and Opp (2026) press the same point on the price path itself — an instrumented shift that persists or builds up moves future prices too, and the resulting bias is roughly the fraction of the shift still unresolved at the investor's next trading date. Relatedly, Davis, Kargar and Li (2025) decompose the demand elasticity into a price pass-through term, how far a price change forecasts the next return, and a substitutability term, and empirically quantify the pass-through in US stocks.

Volatility is a time-series object, so one might hope the cross-section escapes. It does not: switching supply shocks across two otherwise identical assets switches their volatilities too, and He, Kondor and Li (2026) show that comparing them within a cross-section—the definition closest in spirit to KY19—yields much the same object as the time series, so the bias persists.

The object also moves with who is in the market. In Haddad, Huebner and Loualiche (2025) aggregate elasticity is an equilibrium outcome of investor composition: when some investors turn passive the rest trade more aggressively, but only partly, and the shift toward passive investing over the last twenty years left demand for individual stocks some eleven percent more inelastic. An elasticity estimated under one investor composition need not survive a change in it.

What to do about it is less settled. Controls are treacherous: if the instrument is a shift in residual supply, any control for return variance or for the intertemporal hedging term is itself moved by that instrument — a bad control in the sense of Angrist and Pischke (2009) rather than a remedy. Separate instruments for each endogenous argument would work but are hard to find, and the structural route trades that difficulty for a strong stance on the environment. A more useful response may be to change the target: a central bank asking what buying Treasuries does to yields

wants the aggregate multiplier, equilibrium adjustment included. The precedent is the fiscal multiplier, which is no structural parameter — its magnitude depends on the monetary stance, the financing scheme and the persistence of the shock. Macroeconomics did not conclude that such estimates are useless; it reads them as *identified moments*, causal effects that do not transport but sharply discipline the set of admissible models. The measured slope can play that role, and its value lies in disciplining a model rather than in serving as a parameter of one.

5.3 Is Structure a Concession?

Writing asset demand as a fundamental demand over state-contingent payoffs composed with a latent mapping from payoffs to holdings, Fuchs, Fukuda and Neuhann (2026) put the difficulty in its most general form. The estimand they have in view throughout is an individual asset's elasticity, and the objection is that no such estimate survives the cross-price effects the latent mapping ties to it. Their conclusion is a negative one: estimated demand elasticities are contingent, model-specific constructs that reflect one's prior assumptions about investors rather than validate them.

As a theorist I have no quarrel with that last clause — what comes out of an estimate does reflect the a priori assumptions put into it. But I read it as the premise, not the conclusion. As Section 1 set out, the object was never one asset's elasticity but the whole matrix $\boldsymbol{\mathcal{E}}$: even a question as simple as what quantitative easing does to the ten-year yield needs all of it. An $N \times N$ matrix cannot be recovered freely from the variation any dataset contains, so imposing structure — reducing $N \times N$ to $K \times K$, say — is not a concession but the only way to attack it at all.

The two issues before this one are the same argument. Learning makes the elasticity matrix a product of the information environment; endogenous characteristics make it move with the shock used to trace it. Both are filed as obstacles to identification; both are reasons to impose structure rather than to abandon it. The question is only which structure, and on what grounds.

References

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

Allen, Franklin and Douglas Gale, "Limited Market Participation and Volatility of Asset Prices," *American Economic Review*, 1994, *84* (4), 933–955.

An, Yu and Amy Huber, "Demand Propagation Through Traded Risk Factors," 2024. Working paper, The Wharton School.

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

Basak, Suleyman and Domenico Cuoco, "An Equilibrium Model with Restricted Stock Market Participation," *Review of Financial Studies*, 1998, *11* (2), 309–341.

Bastianello, Francesca, "Expectations and Inelastic Markets," 2026. Working paper, University of Chicago Booth School of Business.

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

Brandt, Michael W., Pedro Santa-Clara, and Rossen Valkanov, "Parametric Portfolio Policies: Exploiting Characteristics in the Cross-Section of Equity Returns," *Review of Financial Studies*, 2009, *22* (9), 3411–3447.

Bretscher, Lorenzo, Lukas Schmid, Ishita Sen, and Varun Sharma, "Institutional Corporate Bond Pricing," *Review of Financial Studies*, 2026, *39* (3), 605–660.

Brunnermeier, Markus K. and Yuliy Sannikov, "A Macroeconomic Model with a Financial Sector," *American Economic Review*, 2014, *104* (2), 379–421.

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

Chaudhary, Manav, Zhiyu Fu, and Jian Li, "Corporate Bond Multipliers: Substitutes Matter," *Review of Financial Studies*, 2026. Forthcoming.

Chaudhry, Aditya and Carter Davis, "The Origins of the Factor Zoo: Investors Weakly Substitute Across Stocks," 2026. Working paper, October 2026.

Coval, Joshua and Erik Stafford, "Asset Fire Sales (and Purchases) in Equity Markets," *Journal of Financial Economics*, 2007, *86* (2), 479–512.

Davis, Carter, "Classical Versus Demand System Asset Pricing," 2025. Working paper, April 2025, SSRN 5232345.

_ , Mahyar Kargar, and Jiacui Li, "Why Do Portfolio Choice Models Predict Inelastic Demand?," *Journal of Financial Economics*, 2025, *172*, 104096.

_ , _ , _ , and Dejanir Silva, "On the Recovery of Demand Elasticities in Dynamic Settings," 2026. Working paper, September 2026.

Debreu, Gerard, *Theory of Value: An Axiomatic Analysis of Economic Equilibrium*, Yale University Press, 1959.

Fang, Chuck and Kairong Xiao, "What Do $40 Trillion of Portfolio Holdings Say about Monetary Policy Transmission?," 2025. Working paper.

Fuchs, William, Satoshi Fukuda, and Daniel Neuhann, "Demand-System Asset Pricing: Theoretical Foundations," 2025. Working paper.

_ , _ , and _ , "A Trilemma for Asset Demand Estimation," 2026. Working paper.

Gârleanu, Nicolae and Lasse Heje Pedersen, "Margin-Based Asset Pricing and Deviations from the Law of One Price," *Review of Financial Studies*, 2011, *24* (6), 1980–2022.

_ , _ , and Allen M. Poteshman, "Demand-Based Option Pricing," *Review of Financial Studies*, 2009, *22* (10), 4259–4299.

Greenwood, Robin and Dimitri Vayanos, "Bond Supply and Excess Bond Returns," *Review of Financial Studies*, 2014, *27* (3), 663–713.

Haddad, Valentin, Alan Moreira, and Tyler Muir, "Asset Purchase Rules in the Euro Area and Their Effect on Bond Markets," 2024. Working paper.

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

_ , Zhiguo He, Paul Huebner, Péter Kondor, and Erik Loualiche, "Causal Inference for Asset Pricing," 2026. Working paper, June 2026.

He, Zhiguo and Arvind Krishnamurthy, "Intermediary Asset Pricing," *American Economic Review*, 2013, *103* (2), 732–770.

_ , Péter Kondor, and Jessica Shi Li, "Demand Elasticity in Dynamic Asset Pricing," 2026. Working paper, July 2026.

Jansen, Kristy A. E., Wenhao Li, and Lukas Schmid, "Granular Treasury Demand with Arbitrageurs," 2024. NBER Working Paper 33243.

Jr., Robert E. Lucas, "Asset Prices in an Exchange Economy," *Econometrica*, 1978, *46* (6), 1429–1445.

Kocherlakota, Narayana R., "Log-Linear Relative Asset Demand," 2025. NBER Working Paper 34395.

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

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

_ , Robert J. Richmond, and Motohiro Yogo, "Which Investors Matter for Equity Valuations and Expected Returns?," *Review of Economic Studies*, 2023. Advance access publication 21 August 2023, doi:10.1093/restud/rdad083.

Lou, Dong, "A Flow-Based Explanation for Return Predictability," *Review of Financial Studies*, 2012, *25* (12), 3457–3489.

Markowitz, Harry, "Portfolio Selection," *Journal of Finance*, 1952, *7* (1), 77–91.

van Binsbergen, Jules H., Benjamin David, and Christian C. Opp, "How (Not) to Identify Demand Elasticities in Dynamic Asset Markets," 2026. NBER Working Paper 34528.

van der Beck, Philippe, Lorenzo Bretscher, and Julie Zhiyu Fu, "A Bound on Price Impact and Investor Heterogeneity," 2026. Working paper, September 2026.

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