Portfolio Granularity and Demand Estimation

Ryan Cheng*

May 31, 2026

Click here for the latest draft.

Abstract

Previous applications of structural demand systems to asset pricing often imply puzzlingly inelastic, or even upward-sloping, demand curves. I show that these pathologies arise mechanically from the granular-portfolio condition: the softmax demand link weights each position's contribution to identification by its squared portfolio share, so the effective information is governed by portfolio concentration ($n^{\text{eff}} \approx 30$–$60$), not the raw number of holdings. Standard practice aggregating heterogeneous mandates into a single observed portfolio therefore does not increase statistical power, but instead induces convexity. Linear specifications applied to this convex object distort price coefficients and overstate latent demand. A granular estimator yields well-behaved, downward-sloping demand curves without imposing sign restrictions; its implied aggregate price elasticities are two to four times larger than previous estimates.

JEL Codes: G11, G12, G23, L13.

Keywords: Portfolio choice, cumulants, demand-elasticity, demand-system asset pricing.

*INSEAD. Email: ryan.cheng@insead.edu. I thank Frederico Belo, Manav Chaudhary, Sergei Glebkin, Peter Kondor and Motohiro Yogo for valuable feedback. I further thank participants at the INSEAD–Wharton Doctoral Consortium for helpful comments. All errors are my own.

Introduction

Structural demand estimation has deep roots in industrial organization (e.g., Berry et al., 1995; Nevo, 2001). Its recent application to asset pricing, pioneered by Koijen and Yogo (2019) and extended by Koijen et al. (2024), has produced striking empirical conclusions: aggregate demand appears extremely inelastic and, absent additional restrictions, estimated demand curves can even slope upward. A large share of return volatility is then assigned to an unobserved "latent demand" component, raising a basic concern about how much economic content remains in the decomposition.

A growing theoretical literature interprets these findings as evidence of misspecification. In this view, standard logit-style structures omit forces that are central in financial markets, including rich cross-asset substitution, spillovers, and dynamic trading considerations (Fuchs et al., 2025; He et al., 2026; van Binsbergen et al., 2025). This paper takes a different approach. Rather than asking first whether the structural framework is too simple for finance, I ask how the framework's mechanics change when it is brought from product markets into asset markets.

I argue that many of the empirical pathologies emphasized in recent work are mechanical consequences of the granular-portfolio condition (Condition G), a feature that distinguishes asset demand from standard product demand. In a logit-style demand system, each position contributes to identification in proportion to its squared portfolio share. The relevant sample size is therefore not the raw number of holdings N, but the inverse Herfindahl index,

$$n^{\text{eff}} = 1/\text{HHI},$$

which measures portfolio concentration. In institutional equity portfolios, N typically runs from hundreds to thousands, while n^{eff} is often only 30–60. Most nominal holdings therefore contribute essentially no identifying power. This wedge between portfolio breadth and effective information is a distinctly financial feature of the data, and it matters for both identification and inference.

Section 1 formalizes this idea and derives the corresponding information identity: identification is proportional to portfolio concentration, or equivalently to $1/n^{\text{eff}}$. Adding more positions does not expand the identification budget unless those positions materially change the portfolio's concentration, and empirically they rarely do. The effective amount of cross-sectional information in a portfolio is therefore governed by concentration rather than by the raw holdings count.

A natural response to limited within-portfolio identification is to enlarge the observed cross-section. This logic appears throughout the demand-system literature in several forms: using consolidated institutional portfolios that span thousands of positions rather than underlying mandates that hold hundreds, pooling investors with sparse holdings, or grouping them by type to stabilize estimation. The common premise is that a larger observed portfolio provides more identifying variation.

Proposition 1 shows that this premise is false when the added positions carry negligible squared-share mass. Expanding the list of holdings within a portfolio does not meaningfully increase information unless it changes concentration. Worse, the most common form of enlargement in practice, consolidating heterogeneous mandates into a single observed portfolio, is not merely unhelpful but distortive. This is the default in consolidated 13F data, but the issue is broader: it arises whenever the observational unit bundles distinct allocation decisions, such as growth and value sleeves, active and passive strategies, or internal and external sub-advisors. Even when each underlying mandate follows a standard logit-style, log-linear demand rule, consolidation produces an institution-level demand object that is mechanically convex in prices. The dispersion and higher moments of mandate-level price sensitivities do not wash out in aggregation; instead, they enter the consolidated demand function directly and distort the concentrated positions that drive identification. Fitting a linear specification to this convex object can attenuate the price coefficient toward zero, push it into the upward-sloping region, and mechanically inflate measured latent demand. Section 2 formalizes these results.

Estimating at the mandate level removes this aggregation curvature by construction. But the information constraint established in Section 1 continues to bind even at the correct observational unit. A single cross-section of mandate-level holdings still provides at most about $n^{\text{eff}} \approx 30$–$60$ effective observations for the price coefficient, regardless of how many nominal positions appear in the portfolio. That is the hard ceiling on what one snapshot of holdings can reveal about demand parameters.

This limitation is especially important in asset-pricing applications, where the demand system is often asked to recover time-specific coefficients quarter by quarter in order to decompose contemporaneous price movements. In product-market settings, by contrast, demand systems are typically used to estimate time-averaged parameters by pooling across many markets or periods. In finance, researchers can partially alleviate the tension by borrowing strength across time for the same mandate, exploiting the persistence of fund-level demand styles. But the basic constraint remains unchanged: the statistical content of any single portfolio snapshot is governed by n^{eff}, not by the nominal number of holdings. Section 3 develops these implications and distinguishes the numerical conditioning problem, which can be improved by preconditioning, from the statistical limitation, which cannot.

Having identified the mechanical roots of these distortions, Section 4 develops a constructive estimation strategy tailored to financial markets. The central principle is simple: estimate demand at the level where allocation decisions are actually made, not at an aggregate level that bundles heterogeneous mandates. In U.S. equities, this means individual mutual fund portfolios rather than consolidated 13F institutions.

In other settings, the relevant unit may be a mandate, sleeve, account, or desk-level portfolio. A long-tail normalization improves numerical conditioning, and pooled shrinkage stabilizes inference within the limits imposed by n^{eff}.

The resulting granular estimates are well behaved and economically interpretable without sign restrictions. Passive broad-market trackers cluster near $\alpha \approx 1$, while active mandates exhibit systematically lower price coefficients and clearer cross-sectional heterogeneity. At the aggregate level, the stock-level price elasticity recovered from the granular system is two to four times larger than in consolidated 13F implementations (Koijen and Yogo, 2019). It declines from roughly 0.4 in 2010 to 0.2 by 2024, consistent with rising passive ownership compressing the average elasticity. The early-sample magnitudes align closely with quasi-experimental benchmarks from flow-induced trading and index-inclusion studies (e.g., Wurgler and Zhuravskaya, 2002; Lou, 2012; Chang et al., 2015), and the subsequent decline is itself economically interpretable.

Section **??** discusses the relationship between partial- and general-equilibrium analysis in demand-system asset pricing. The estimates in this paper are strictly partial equilibrium, but that is not a limitation of convenience; it is the appropriate starting point. In product-market demand estimation, researchers first recover demand elasticities and only then, when the application requires it, impose a supply-side or market-clearing closure for counterfactual analysis. The same logic applies here. A well-identified partial-equilibrium demand system is already economically meaningful and informative for questions such as flow-driven reallocation and portfolio fragility, and it is the necessary benchmark against which broader general-equilibrium critiques should be evaluated.

More broadly, the contribution of the paper is constructive and econometric: many of the headline pathologies in existing applications arise not because demand systems are inherently uninformative in finance, but because the estimator is applied to the wrong observational unit and in the wrong statistical regime. Once demand is estimated at the mandate level, the framework recovers economically meaningful elasticities and becomes a credible foundation for subsequent general-equilibrium analysis.

Related literature. This paper connects to three strands of work.

Structural demand-system estimation. The structural demand-system approach to asset pricing originates with Koijen and Yogo (2019) and has been extended to multiple asset classes and institutional settings: U.S. equities (Koijen et al., 2024; Haddad et al., 2025a), corporate bonds (Bretscher et al., 2026), euro-area government bonds (Koijen et al., 2021), and international portfolios spanning equities, bonds, and currencies (Koijen and Yogo, 2026). These applications vary in their proximity to the primitive decision unit. Some work with consolidated 13F institutions (Koijen and Yogo, 2019; Koijen et al., 2024), others with fund- or account-level data (Bretscher et al., 2026; Gabaix et al., 2025), and still others pool heterogeneous investors by type. I position this paper as a direct methodological contribution to this literature. The aggregation and conditioning effects I formalize apply, in principle, whenever the observed unit bundles heterogeneous mandates—their severity governed by a measurable quantity, the within-entity dispersion of price sensitivities. Applications that already work closer to the mandate level face these effects less acutely; those that rely on consolidated institutional data face them most. By providing a diagnostic framework and an estimation procedure that explicitly respects the delegation structure of managed portfolios, I offer both a practical remedy and a portable principle for demand-system implementations across asset classes.

Theoretical foundations for asset demand-system identification. Second, this paper relates to a recent theoretical literature that clarifies the conditions under which demand-system elasticities can be given a structural interpretation in financial markets. One set of papers emphasizes cross-asset substitution, spillovers, and the limits of low-dimensional static specifications in large opportunity sets (Fuchs et al., 2025; Haddad et al., 2025b). Another shows that in dynamic settings, residual-supply shocks can shift and tilt investors' demand curves through changes in endogenous risk and investment opportunities, so that measured elasticities need not recover the conceptual slope of demand (He et al., 2026). Related work also highlights the distinction between price-level variation and expected-return variation when using supply shocks as instruments (van Binsbergen et al., 2025). I view my econometric approach as complementary to, but distinct from, these theoretical contributions. While their focus is on whether the economic assumptions of the framework capture key features of asset pricing, my focus is on a prior econometric question: whether the estimation itself is mathematically well-posed given the granular, delegated structure of institutional portfolios. Even if a researcher were to adopt the richer substitution patterns or dynamic corrections proposed by this literature, the aggregation and conditioning effects documented here would remain unless the estimator explicitly respects the mandate-level data-generating process.

Reduced-form evidence on price impact. At the same time, empirical work on asset demand has long proceeded without a fully specified structural system. A final strand this paper relates to is the reduced-form literature estimating local price effects from plausibly exogenous ownership shifts, including index reconstitutions, inclusions, and flow-induced trading (e.g., Wurgler and Zhuravskaya, 2002; Lou, 2012; Chang et al., 2015). These studies do not estimate a full cross-asset demand system, but they provide disciplined benchmark magnitudes for local price impact. Related recent work further shows that measured price impact depends on the relevant substitution margin. Chaudhary et al. (2025) demonstrate that in corporate bonds, security-level multipliers are near zero once heterogeneous substitutability across close and distant bonds is accounted for, but rise monotonically with the level of portfolio aggregation as close substitutes become scarce. More broadly, this result shows that measured price impact is not a single primitive, but depends on the relevant substitution set and the level at which assets are aggregated.

The present paper identifies an analogous sensitivity along the investor dimension. Where those papers show that aggregating assets into broader portfolios raises measured inelasticity because substitutes become scarcer, I show that aggregating mandates into broader institutions can also raise measured inelasticity because heterogeneous demand schedules are conflated within the observed unit. The two aggregation margins are logically distinct: the first reflects the market's underlying substitution structure, while the second arises from estimating demand at a level that bundles heterogeneous delegated portfolios. But both imply that demand-system estimates are sensitive to the observational unit. The structural approach contributes by going beyond a single treatment effect: it recovers substitution patterns across assets, allows counterfactuals under alternative shocks, and links equilibrium price responses back to underlying portfolio demand. My results show that once estimation is brought closer to the primitive decision unit and adjusted for aggregation and long-tail effects, the implied elasticities move materially closer to the reduced-form benchmarks established by this literature.

1 The Granular-Portfolio Condition

Figure 1 plots the cross-sectional share distribution for four representative index funds in 2024Q4. The pattern is striking and immediate. Each portfolio exhibits a steep rank-size profile: a handful of large positions carry the bulk of portfolio value, followed by a long, flat tail of hundreds to over a thousand positions clustered near the reporting floor of 10^{-4} (one basis point). The top 50 holdings account for 57–64% of total portfolio mass, even though the fund holds 449 to over 1,100 names. This is not specific to index funds. Table 2 documents the same pattern across the mutual-fund sector: by 2024, 67% of fund holdings carry weight below 1% of the portfolio, and positions below 0.1% account for roughly 7% of portfolio mass despite comprising 31% of all holdings.

This weight distribution has no analog in the product markets where structural demand estimation originated. In a cereal market with 10 brands, every brand carries meaningful share. In an institutional equity portfolio, the share distribution is orders of magnitude more skewed. The question is how to summarize this skewness in a way that is economically informative, not merely descriptive, but tied to identification.

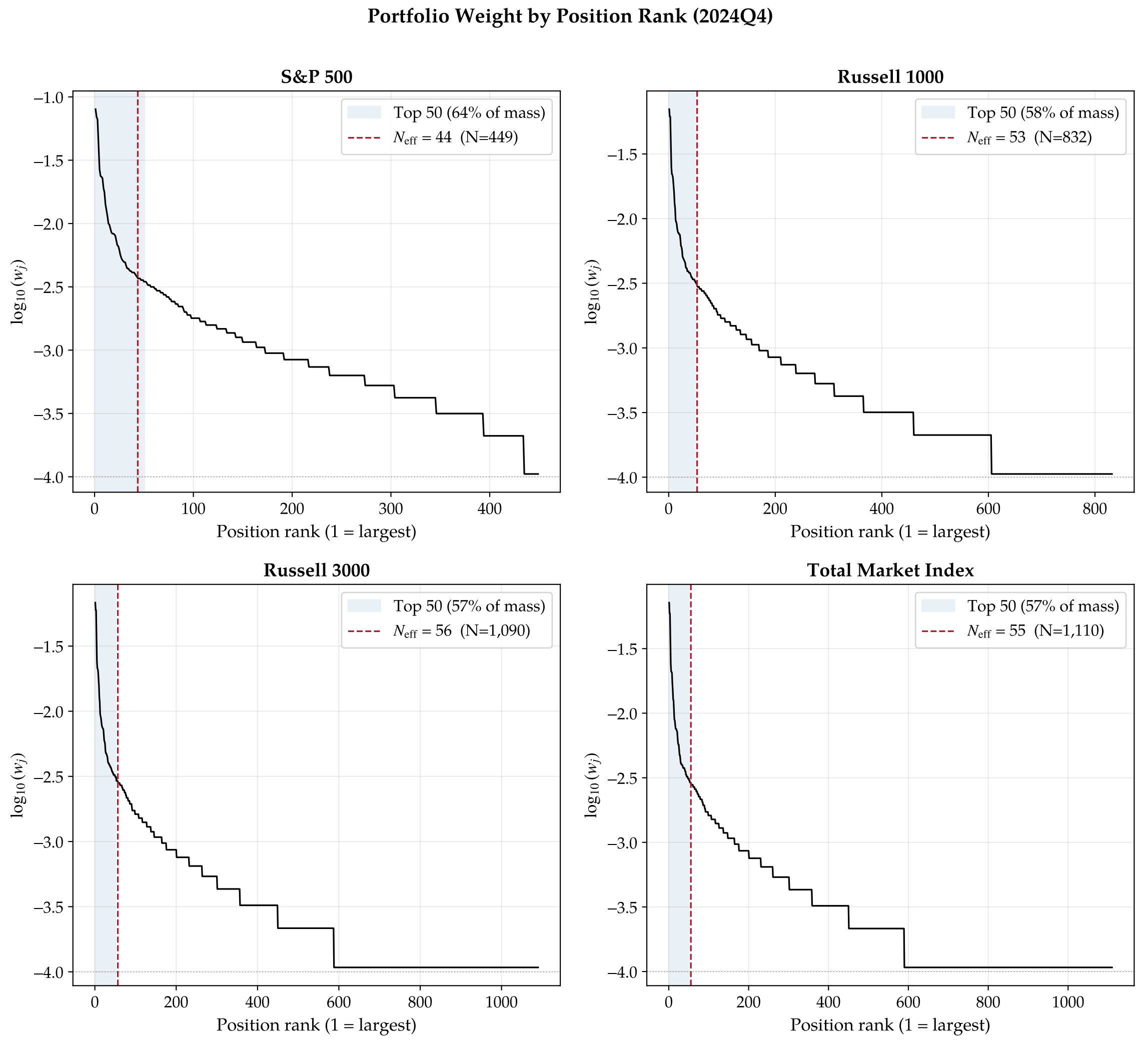

Figure 1: **Portfolio weight profiles exhibit a long tail of tiny, discretized positions.** Each panel plots cross-sectional portfolio weights for a representative index fund in 2024Q4, sorted by position size (rank 1 = largest) on a $\log_{10}$ scale. The shaded region marks the top 50 holdings, which concentrate roughly 57–64% of total portfolio mass despite the fund holding hundreds to over a thousand names. The dashed vertical line indicates the effective number of positions $n^{\text{eff}} \equiv 1/\sum_j w_j^2$, which is an order of magnitude smaller than the raw holdings count N, reflecting substantial concentration and a long tail of economically negligible weights. The horizontal reference at 10^{-4} (1 bp) highlights that many tail positions cluster near a common minimum increment, consistent with discretization or rounding in reported holdings.

The effective number of positions. Throughout this section, w_j denotes the within-universe portfolio share of stock j, normalized so that $\sum_j w_j = 1$.[1] For any normalized share vector, define

$$\text{HHI} \;\equiv\; \sum_{j=1}^{N} w_j^2, \qquad n^{\text{eff}} \;\equiv\; \frac{1}{\text{HHI}}. \tag{1}$$

At the descriptive level, n^{eff} has a clean interpretation: it is the number of equal-share positions that would generate the same concentration as the actual portfolio. If a fund held n positions each with share $1/n$, then $\text{HHI} = n \cdot (1/n)^2 = 1/n$, so $n^{\text{eff}} = n$. When the share distribution is skewed, n^{eff} falls below the raw count N, reflecting the fact that many nominal holdings contribute negligibly.

For the four index funds in Figure 1, n^{eff} ranges from 44 (S&P 500) to 56 (Russell 3000), even though the raw holdings count N ranges from 449 to over 1,100. The dashed vertical lines in the figure mark n^{eff} for each fund, and they fall roughly where the share profile transitions from its steep descent to the flat tail, consistent with the interpretation that positions beyond this point contribute little to the portfolio's statistical content.

But n^{eff} is more than a descriptive summary. The same concentration object reappears endogenously in the local curvature of the demand estimator.

Why n^{eff} governs identification. Consider a multinomial logit demand specification, the workhorse of the structural demand-system literature. I work with a single fund-quarter for clarity; Section 4.2 presents the full specification. Let $\mathcal{U}$ denote the fund's estimation universe with $|\mathcal{U}| = N$. The model assigns fitted shares

$$\mu_j(\theta) \;=\; \frac{\exp(\eta_j)}{\sum_{k\in\mathcal{U}} \exp(\eta_k)}, \qquad \eta_j = \alpha\, p_j + x_j'\beta, \tag{2}$$

so that $\sum_{j\in\mathcal{U}} \mu_j(\theta) = 1$. The NLS estimator minimizes $Q(\theta) = \sum_{j\in\mathcal{U}} (w_j - \mu_j(\theta))^2$. Its score with respect to α is

$$\frac{\partial Q}{\partial \alpha} \;=\; -2 \sum_{j\in\mathcal{U}} \left(w_j - \mu_j\right) \mu_j\, p_j^c, \tag{3}$$

where $p_j^c \equiv p_j - \bar{p}^\mu$ and $\bar{p}^\mu \equiv \sum_{k\in\mathcal{U}} \mu_k\, p_k$. The μ_j factor arises from the chain rule through the logit link: $\partial \mu_j / \partial \alpha = \mu_j\, p_j^c$.

This μ_j factor is the key object. Every stock's contribution to the gradient is weighted by its fitted share. The Gauss–Newton information measure, which governs local iden-

[1] If the original within-universe weights sum to $S \neq 1$, define $w_j \equiv \tilde{w}_j / S$ where $\tilde{w}_j$ is the raw weight. The unnormalized least-squares criterion differs from the normalized one only by the multiplicative constant S^2, so the estimator is unchanged.

tification strength, inherits this weighting *squared*:

$$\mathcal{I}(\alpha) \;\equiv\; \sum_{j\in\mathcal{U}} \mu_j^2\,(p_j^c)^2. \tag{4}$$

Information is governed by concentration, not the number of positions. Apply the same concentration measure to the fitted shares μ_j. Since $\sum_j \mu_j = 1$, the definitions (1) apply directly, giving $\text{HHI} = \sum_j \mu_j^2$ and $n^{\text{eff}} = 1/\text{HHI}$. Factoring this out of (4) gives

$$\mathcal{I}(\alpha) \;=\; \sum_{j=1}^{N} \mu_j^2\,(p_j^c)^2 \;=\; \frac{1}{n^{\text{eff}}}\sum_{j=1}^{N} \underbrace{\frac{\mu_j^2}{\text{HHI}}}_{\nu_j}(p_j^c)^2 \;=\; \frac{\bar{\sigma}_p^2}{n^{\text{eff}}}, \tag{5}$$

where ν_j are the induced *information weights* and $\bar{\sigma}_p^2 \equiv \sum_j \nu_j\,(p_j^c)^2$ is the concentration-weighted second moment of centered prices. This decomposition makes clear that raw holdings count N has no separate first-order role: identification depends on concentration in the fitted shares and on price dispersion among the positions carrying economically meaningful mass.

Two things are immediate. First, information is proportional to concentration: $\mathcal{I}(\alpha) \propto 1/n^{\text{eff}}$, provided $\bar{\sigma}_p^2$ remains bounded away from zero and infinity. Second—and this contrasts with standard demand settings where adding observations typically contributes information—adding positions to the portfolio does not increase information unless those positions materially change the concentration of the fitted portfolio.

The following proposition formalizes this.

Proposition 1 (Invariance to nominal holdings expansion)**.** *Consider a sequence of fitted portfolios indexed by N, with normalized fitted shares μ_{Nj} summing to 1 and $n_N^{\text{eff}} = \left(\sum_{j=1}^{N}\mu_{Nj}^2\right)^{-1}$. If the additional positions contribute vanishing squared-share mass, so that n_N^{eff} remains bounded, and $\bar{\sigma}_{p,N}^2 \to \bar{\sigma}_p^2 \in (0,\infty)$, then*

$$\mathcal{I}_N(\alpha) \;\longrightarrow\; \frac{\bar{\sigma}_p^2}{n^{\text{eff}}},$$

and information does not grow with N.

Proof. By (5), $\mathcal{I}_N(\alpha) = \bar{\sigma}_{p,N}^2/n_N^{\text{eff}}$. If the additional positions contribute vanishing squared-share mass, then $\sum_{j=1}^{N}\mu_{Nj}^2$ converges to a positive finite limit, so n_N^{eff} remains bounded. With $\bar{\sigma}_{p,N}^2 \to \bar{\sigma}_p^2$, the result follows immediately. □

Convergence of $\bar{\sigma}_{p,N}^2$ requires that price variation is roughly homogeneous across the cross-section, so that the concentration reweighting does not materially change the second moment. This is empirically mild: in my data, the cross-sectional variance of centered log-prices is stable across size deciles.

The mechanism is worth emphasizing. The logit link generates a μ_j^2 weighting in the information measure. For a position with fitted share μ_j, its contribution to identification is $\mu_j^2\,(p_j^c)^2$—quadratic in portfolio share. A position that carries 1% of the portfolio contributes 10,000 times more identification information than a position carrying 0.01%. This quadratic penalty is not imposed by the econometrician; it is a structural consequence of the exponential link in the logit demand system.

The proposition is visible in Figure 1. The iShares Russell 1000 ETF holds $N = 832$ names with $n^{\text{eff}} = 53$, while the iShares Russell 3000 ETF holds $N = 1{,}090$ names with $n^{\text{eff}} = 56$. Despite a 30% increase in raw holdings, effective information is nearly unchanged because the additional positions contribute negligible squared-share mass. Section 3 confirms this empirically: the two funds exhibit comparable estimation precision after preconditioning.

Formalizing the granular-portfolio condition. The empirical pattern in Figure 1 and the identification result in Proposition 1 together motivate a formal characterization of the regime that distinguishes financial portfolios from product-market settings.

Condition (Condition G — Granular portfolios). The portfolio share distribution has two features that jointly distinguish financial portfolios from product-market settings:

(a) **Diffuse core.** Portfolio mass is spread across a moderately large number of positions:

$$n^{\text{eff}} \gg 1, \qquad \max_j w_j \;=\; O\!\left(\frac{1}{n^{\text{eff}}}\right).$$

(b) **Long tail.** The number of holdings is much larger than the effective number of positions:

$$N \;\gg\; n^{\text{eff}}.$$

In product markets, both conditions are rarely satisfied simultaneously: with 10 brands, $n^{\text{eff}} \approx N$ and every product contributes materially to identification. Condition G requires both features jointly, and this conjunction is the empirical reality of institutional equity portfolios.

Part (a), the diffuse core, will play a central role in Section 2, where it validates a clean characterization of how institutional aggregation shapes demand estimates: by Condition G(a), $\max_m w_{jm} \;=\; O(1/n^{\text{eff}})$, so indirect effects and IIA corrections are $O(1/n^{\text{eff}})$, making consolidated demand tractably convex. Part (b), the long tail, drives the inference constraints formalized in Section 3: by Proposition 1, adding positions beyond n^{eff} does not expand the identification budget. The effective degrees of freedom for demand parameters are therefore tight: portfolio concentration, not raw holdings count, is the binding constraint on what the demand system can learn.

2 Granularity and Aggregation

Section 1 establishes that identification is governed by portfolio concentration (n^{eff}), not the number of holdings N, and that enlarging the cross-section within a portfolio does not expand the identification budget. This section shows that one particular form of enlargement—observing the consolidated institutional portfolio rather than its constituent mandates—does not simply fail to help; it can reshape the information that Proposition 1 identifies as the sole source of identification.

The institution observed in standard data—a 13F filer or fund family—is almost always an internal portfolio-of-portfolios: it allocates capital across heterogeneous mandates, and I observe only the sum. Each mandate maintains its own portfolio, and under Condition G each mandate's portfolio has its own diffuse core of roughly n_m^{eff} positions carrying weight $O(1/n_m^{\text{eff}})$.

For a stock j that sits in the institution's core—meaning the consolidated weight W_{ji} is large enough to contribute meaningfully to identification—the observed weight is generically a mixture of mandate-level weights that reflect *different* price sensitivities α_m. The institutional weight on Apple, say, combines a growth sleeve's large overweight (low α, negative β^{BM}) with an index sleeve's mechanical position ($\alpha \approx 1$, $\beta^{BM} \approx 0$) and perhaps a value sleeve's modest underweight. The consolidated W_{ji} is a single number that combines these heterogeneous demand primitives. Institutional consolidation does not reduce n^{eff}—the institution may still have $\sim$50 large positions—but it *mixes* heterogeneous signals into each one.

I now formalize this mixing. Section 2.1 introduces the cumulant generating function (CGF) representation that makes the aggregation structure precise. Section 2.2 derives the slope and curvature of consolidated demand under Condition G. Section 2.3 develops the consequences for standard linear estimation.

2.1 The setup and the CGF representation

Consider an institution i that aggregates mandates $m \in \mathcal{M}_i$. Each mandate allocates across a universe of assets plus an outside option (cash or the rest of the portfolio). Write the mandate-level share system in log-odds form:

$$\log \frac{w_{jm}}{w_{0m}} = \delta_{jm}, \qquad \delta_{jm} = \alpha_m p_j + x_j' \beta_m + \xi_j. \tag{6}$$

Here p_j is a price-like regressor used in empirical demand estimation (e.g., log market equity or log market-to-book), x_j are observable characteristics, ξ_j is an unobserved asset attribute, and α_m captures the mandate's effective price sensitivity (so $\alpha_m < 1$ corresponds to downward-sloping demand in the usual sense).

Let $a_{mi} \geq 0$ denote the institution's capital share allocated to mandate m, with $\sum_{m \in \mathcal{M}_i} a_{mi} = 1$. The observed institution-level weight in asset j is a mixture:

$$W_{ji} = \sum_{m \in \mathcal{M}_i} a_{mi} w_{jm}, \qquad W_{0i} = \sum_{m \in \mathcal{M}_i} a_{mi} w_{0m}. \tag{7}$$

Using $w_{jm} = w_{0m} e^{\delta_{jm}}$, the institution-level odds ratio becomes

$$\frac{W_{ji}}{W_{0i}} = \sum_{m \in \mathcal{M}_i} \pi_{mi}\, e^{\delta_{jm}}, \qquad \pi_{mi} \equiv \frac{a_{mi} w_{0m}}{\sum_{\ell \in \mathcal{M}_i} a_{\ell i} w_{0\ell}}. \tag{8}$$

Taking logs yields the central object:

$$\Lambda_{ji} \equiv \log \frac{W_{ji}}{W_{0i}} = \log \sum_{m \in \mathcal{M}_i} \pi_{mi}\, e^{\delta_{jm}}. \tag{9}$$

This is a *log-sum-exp* (LSE) function: the log of a weighted sum of exponentials of mandate-level indices. The LSE form is highly tractable because it admits a direct probabilistic representation.

Result 1 (CGF representation of institutional demand)**.** Define a discrete random variable Δ_j that takes the value δ_{jm} with probability π_{mi}, where π_{mi} are the institution's endogenous mandate weights. The consolidated institutional demand can be written exactly as:

$$\Lambda_{ji} = \log \sum_{m \in \mathcal{M}_i} \pi_{mi}\, e^{\delta_{jm}} = \log \mathbb{E}_\pi \left[e^{\Delta_j} \right]. \tag{10}$$

The institutional log-odds Λ_{ji} is the *cumulant generating function* (CGF) of the internal mandate-level index Δ_j, evaluated at 1.

This representation turns aggregation into a familiar statistical object: consolidated demand is not "a demand curve with average parameters," but a smooth transform of the entire within-institution distribution of mandate tastes.

The CGF form also provides an interpretable expansion. Let $\kappa_r(\Delta_j)$ denote the r-th cumulant of Δ_j under π. A cumulant (Taylor) expansion of the CGF gives

$$\Lambda_{ji} = \kappa_1(\Delta_j) + \frac{1}{2}\kappa_2(\Delta_j) + \frac{1}{6}\kappa_3(\Delta_j) + \frac{1}{24}\kappa_4(\Delta_j) + \cdots . \tag{11}$$

The first term is the mean mandate-level index; the second term is the variance; the third and fourth terms correspond to skewness and kurtosis. Economically, this means the consolidated institution does not behave like an "average mandate": higher-order moments of internal heterogeneity enter mechanically and alter the elasticity one would infer from aggregated holdings.

2.2 Convexity of consolidated demand

The CGF representation (10) is exact for any portfolio. To draw out its implications for estimation, I derive the slope and curvature of consolidated demand and show that Condition G makes the results particularly tractable.

The mixing weights π_{mi} defined in (8) are themselves functions of prices: each π_{mi} depends on the outside-option share w_{0m}, which shifts when any asset price moves. The total derivative of Λ_{ji} with respect to p_j therefore has both a *direct channel* (how each mandate's index responds to p_j, holding the mandate mix fixed) and an *indirect channel* (how the mandate mix shifts as p_j changes outside-option shares). Condition G is what makes the indirect channel negligible.

Direct effect. Holding the mixing weights π_{mi} fixed and differentiating the log-sum-exp in (9) yields

$$\left.\frac{\partial \Lambda_{ji}}{\partial p_j}\right|_{\pi \text{ fixed}} = \sum_{m \in \mathcal{M}_i} \tilde{\pi}_{mi}\, \alpha_m \equiv \mathbb{E}_{\tilde{\pi}}[\alpha_m], \tag{12}$$

where

$$\tilde{\pi}_{mi} \equiv \frac{\pi_{mi}\, e^{\delta_{jm}}}{\sum_{\ell \in \mathcal{M}_i} \pi_{\ell i}\, e^{\delta_{j\ell}}} \tag{13}$$

are exponentially tilted weights that up-weight mandates finding asset j relatively attractive.

Indirect effect. Because $\pi_{mi} = a_{mi} w_{0m} / \sum_\ell a_{\ell i} w_{0\ell}$ and $\partial w_{0m}/\partial p_j = -\alpha_m w_{jm} w_{0m}$, a price change in asset j shifts the mandate mix:

$$\frac{\partial \pi_{mi}}{\partial p_j} = \pi_{mi} \left(-\alpha_m w_{jm} + \sum_{\ell \in \mathcal{M}_i} \pi_{\ell i}\, \alpha_\ell\, w_{j\ell} \right). \tag{14}$$

The total derivative is

$$\frac{d\Lambda_{ji}}{dp_j} = \underbrace{\sum_{m \in \mathcal{M}_i} \tilde{\pi}_{mi}\, \alpha_m (1 - w_{jm})}_{\text{direct}} + \underbrace{\sum_{m \in \mathcal{M}_i} \pi_{mi}\, \alpha_m\, w_{jm}}_{\text{indirect}}, \tag{15}$$

where the indirect effect can be isolated as

$$\frac{d\Lambda_{ji}}{dp_j} - \left.\frac{\partial \Lambda_{ji}}{\partial p_j}\right|_{\pi \text{ fixed}} = \sum_{m \in \mathcal{M}_i} (\pi_{mi} - \tilde{\pi}_{mi})\, \alpha_m\, w_{jm}. \tag{16}$$

Economically, the indirect channel captures a *compositional shift*: when asset j's price rises, mandates with high exposure to j see their outside-option share fall relative to other mandates, tilting the institution's mandate mix. In a cereal market where a leading brand commands a 15% share, this compositional shift can be first-order. In an institutional equity portfolio, the effect is bounded by $w_{jm} = O(1/n^{\text{eff}})$.

Simplification under Condition G. Every term in the indirect effect (16) carries a factor of w_{jm}. Under Condition G, this entire channel is $O(1/n^{\text{eff}})$ and becomes negligible.[2] The factors $(1 - w_{jm})$ in the direct channel collapse to unity, and the total derivative reduces to the clean tilted average:

$$\frac{d\Lambda_{ji}}{dp_j} \approx \mathbb{E}_{\tilde{\pi}}[\alpha_m]. \tag{17}$$

The same logic applied to the second derivative yields[3]

$$\frac{d^2\Lambda_{ji}}{dp_j^2} \approx \mathrm{Var}_{\tilde{\pi}}(\alpha_m) \geq 0. \tag{18}$$

Result 2 (Consolidated demand is structurally convex)**.** Under Condition G, the consolidated institutional demand Λ_{ji} satisfies, up to $O(1/n^{\text{eff}})$ corrections:

(i) **Slope.** $d\Lambda_{ji}/dp_j = \mathbb{E}_{\tilde{\pi}}[\alpha_m]$, the exponentially tilted mean of mandate-level price sensitivities.

(ii) **Curvature.** $d^2\Lambda_{ji}/dp_j^2 = \mathrm{Var}_{\tilde{\pi}}(\alpha_m) \geq 0$. Curvature equals the tilted variance of price sensitivities across mandates. It vanishes if and only if within-institution price sensitivities are degenerate.

The moment an institution combines genuinely heterogeneous sleeves—growth and value, active and passive—dispersion appears ($\mathrm{Var} > 0$), and consolidated

[2]Formally, by Condition G(a), $\max_m w_{jm} = O(1/n^{\text{eff}})$, so $|\text{Indirect}| \leq 2 \max_m |\alpha_m| \cdot \max_m w_{jm} = O(1/n^{\text{eff}})$. The same argument extends to all higher-order indirect terms in the second derivative; see Appendix A.

[3]The exact second derivative without Condition G is $d^2\Lambda_{ji}/dp_j^2 = \mathrm{Var}_{\tilde{\pi}}(\alpha_m(1 - w_{jm})) - \mathrm{Var}_{\pi}(\alpha_m w_{jm}) + \sum_m (\pi_{mi} - \tilde{\pi}_{mi})\alpha_m^2 w_{jm}(1 - w_{jm})$. Under Condition G, the last two terms are $O(1/n^{\text{eff}})$ and the first reduces to $\mathrm{Var}_{\tilde{\pi}}(\alpha_m)$. See Appendix A.

demand mechanically curves. A representative mandate does not exist: unless all mandates share identical price sensitivity, any linear coefficient estimated on consolidated holdings is a projection of a nonlinear aggregator, not the mean of underlying mandate coefficients.

This convexity operates on the n^{eff} concentrated positions that Proposition 1 identifies as the sole source of identification—confirming that the aggregation wedge cannot be offset by adding tail stocks to the holdings.

2.3 Consequences for linear estimation

Result 2 establishes that consolidated institutional demand is generically convex. What happens when an econometrician fits a linear specification to this object? Two corollaries follow. First, the price coefficient is a projection coefficient that can be attenuated, or even sign-reversed, by within-institution heterogeneity. Second, the residual "latent demand" component absorbs the curvature that the linear model cannot represent, inflating its measured importance.

One clarification is useful before stating these results, because the convexity in Result 2 and the sign ambiguity below might appear to be in tension. They are not. Result 2 signs the curvature *in the scalar-price direction*: the second derivative of consolidated demand in p_j equals $\mathrm{Var}_{\tilde{\pi}}(\alpha_m) \geq 0$. What a linear estimator recovers, however, is the projection of this curvature onto *residualized* price—price after partialling out the characteristics x_j. That projection is what is sign-indeterminate: it depends on the cross-sectional geometry of (p_j, x_j), not on the sign of the curvature itself. Positive curvature is therefore fully consistent with a price-coefficient wedge of either sign.

Under Condition G, up to second order, the consolidated demand object can be written as a linear component plus the curvature term from Result 2. Suppressing the institution subscript i for readability:

$$\Lambda_j \;\approx\; \tilde{\xi}_j + c + \bar{\alpha}\, p_j + x_j'\bar{\beta} \;+\; \Omega_j, \tag{19}$$

where $\bar{\alpha}$ and $\bar{\beta}$ are mandate-weighted averages, and

$$\Omega_j \;\equiv\; \tfrac{1}{2}\,\mathrm{Var}(\alpha)\, p_j^2 + p_j\, x_j'\,\mathrm{Cov}(\alpha,\beta) + \tfrac{1}{2}\, x_j'\,\mathrm{Var}(\beta)\, x_j. \tag{20}$$

All variance/covariance operators are within institution i, taken over the mandate distribution. The key point is that Ω_j is not noise: it is a deterministic component implied by heterogeneity and the CGF structure.

Corollary 1 (Linear projection wedge in $\hat{\alpha}$)**.** *Suppose the econometrician estimates the canonical linear consolidated specification*

$$\Lambda_j = \hat{\alpha}\, p_j + x_j'\hat{\beta} + \text{residual}. \tag{21}$$

Let $\tilde{p}_j$ denote the residual from projecting p_j on x_j (the Frisch–Waugh–Lovell residual). Then the deviation of the estimated price coefficient from the first-order "average mandate" slope satisfies

$$\hat{\alpha} - \bar{\alpha} = \frac{\text{Cov}(\tilde{p}_j,\, \Omega_j)}{\text{Var}(\tilde{p}_j)}. \tag{22}$$

Substituting the second-order form of Ω_j yields

$$\hat{\alpha} - \bar{\alpha} \;\approx\; \tfrac{1}{2}\,\text{Var}(\alpha)\,\frac{\text{Cov}(\tilde{p}_j,\, p_j^2)}{\text{Var}(\tilde{p}_j)} \;+\; \text{Cov}(\alpha,\beta)'\,\frac{\text{Cov}(\tilde{p}_j,\, p_j x_j)}{\text{Var}(\tilde{p}_j)} \;+\; \tfrac{1}{2}\,\frac{\text{Cov}\big(\tilde{p}_j,\, x_j'\text{Var}(\beta)x_j\big)}{\text{Var}(\tilde{p}_j)}. \tag{23}$$

The projection wedge scales with within-institution heterogeneity; its sign depends on the cross-sectional geometry of (p_j, x_j). Even with well-behaved mandate primitives, a linear consolidated fit can be attenuated toward zero or pushed past zero into the upward-sloping region.

Remark 1 (When aggregation shifts the consolidated slope upward)**.** The projection wedge in (22) is not signed in general, but the scalar-price case is instructive. With common characteristic loadings the leading curvature term is $\Omega_j = \frac{1}{2}\,\text{Var}(\alpha)\, p_j^2$, so

$$\hat{\alpha} - \bar{\alpha} = \tfrac{1}{2}\,\text{Var}(\alpha)\,\frac{\text{Cov}(\tilde{p}_j,\, p_j^2)}{\text{Var}(\tilde{p}_j)}. \tag{24}$$

Aggregation thus shifts the consolidated price coefficient upward whenever residualized price loads positively on squared price; without controls and with centered p_j, this reduces to positive skewness of the price-like regressor. The sign is therefore a property of the realized stock cross-section and the characteristic controls. In the empirical section below, this projection term is predominantly positive across fund families and over the sample period, so aggregation shifts consolidated estimates toward the near-unit, mechanical-tracking benchmark.

Corollary 2 (Measured latent demand is inflated)**.** *Define the linear-model residual*

$$\xi_j^{\text{lin}} \;\equiv\; \Lambda_j - \big(\hat{c} + \hat{\alpha}\, p_j + x_j'\hat{\beta}\big). \tag{25}$$

This decomposes as

$$\xi_j^{\text{lin}} \;\approx\; \tilde{\xi}_j + g_j, \qquad g_j \equiv (I - P_{p,x})\,\Omega_j, \tag{26}$$

where $P_{p,x}$ is the linear projection operator onto $(1, p, x)$. The measured latent-demand variance satisfies

$$\text{Var}(\xi^{\text{lin}}) \;=\; \text{Var}(\tilde{\xi}) + \text{Var}(g) + 2\,\text{Cov}(\tilde{\xi}, g) \;\geq\; \text{Var}(\tilde{\xi}) \tag{27}$$

whenever $\text{Cov}(\tilde{\xi}, g) \geq 0$. A large "latent demand" component need not reflect economically large unobserved shocks; it can arise because the linear model forces a convex object onto a flat index.

Remark 2 (IIA is second-order under Condition G)**.** The logit structure underlying (6) imposes independence of irrelevant alternatives: substitution away from an asset whose price rises is distributed across all other alternatives in proportion to their current shares. In product markets with a small number of products and large interior shares, this restriction is a first-order concern.

Under Condition G, the IIA restriction is second-order for the objects of primary interest. The logit cross-price effect at the mandate level is

$$\frac{\partial w_{km}}{\partial p_j} = -\alpha_m\, w_{km}\, w_{jm}, \qquad k \neq j. \tag{28}$$

A richer substitution model would replace w_{jm} with a similarity-adjusted term, producing an IIA "error" of order $O(w_{jm}\, w_{km}\, \rho_{jk})$ per pair. Under Condition G, $w_{jm} = O(1/n^{\text{eff}})$, so pairwise errors are $O(1/(n^{\text{eff}})^2)$.

Crucially, the *total* substitution away from asset j,

$$\sum_{k \neq j} \left| \frac{\partial w_{km}}{\partial p_j} \right| = \alpha_m\, w_{jm}(1 - w_{jm}) \;\approx\; \alpha_m\, w_{jm} = O(1/n^{\text{eff}}), \tag{29}$$

is pinned by the adding-up constraint and is the same under logit and any substitution-flexible alternative. IIA governs the *distribution* of this total across alternatives, not its *level*. For own-price elasticities, aggregate demand slopes, and price-impact measures—the primary objects in demand-system asset pricing—the total is what matters. The IIA error is a reshuffling of an $O(1/n^{\text{eff}})$ quantity across $O(N)$ alternatives: economically negligible for aggregate demand measurement, even if it matters for pairwise substitution questions involving near-duplicate securities or concentrated mega-cap positions (see, e.g., Gabaix et al., 2025).

This does not dismiss substitution critiques entirely. Rather, it separates the margins: fine substitution patterns matter for specific applications (near-duplicates, concentrated exposures), while the dominant effects for aggregate demand estimation stem from the aggregation curvature formalized above. Condition G is the organizing principle for both conclusions.

3 Granularity and Inference

Estimating demand at the mandate level removes aggregation curvature by construction. Yet Condition G continues to bind. Proposition 1 shows that identification is governed by portfolio concentration ($1/n^{\text{eff}}$) together with concentration-weighted price dispersion $\bar{\sigma}_p^2$, not by the raw holdings count N. This section develops two practical consequences: estimation precision inherits this concentration structure (Section 3.1), and the contrast between the logit's curvature geometry and the portfolio's long tail creates both a numerical conditioning problem and a statistical limitation with different remedies (Section 3.2).

3.1 Standard errors and portfolio concentration

Proposition 1 shows that identification is governed by portfolio concentration, summarized by n^{eff}, rather than by the raw number of holdings N. The direct implication for inference is that estimation precision inherits the same concentration structure.

Proposition 2 (Inference reflects n^{eff}, not N)**.** *Under the conditions of Proposition 1, the variance of $\hat{\alpha}$ satisfies*

$$\mathrm{Var}(\hat{\alpha}) = \mathcal{I}(\alpha)^{-1}\,\mathcal{V}(\alpha)\,\mathcal{I}(\alpha)^{-1}, \tag{30}$$

where $\mathcal{V}(\alpha) = \sum_j \mu_j^2\,(p_j^c)^2\,\sigma_j^2$. Under conditionally homoskedastic residuals ($\sigma_j^2 = \sigma_r^2$ for all j), this simplifies to

$$\mathrm{Var}(\hat{\alpha}) = \frac{\sigma_r^2}{\mathcal{I}(\alpha)} = \frac{\sigma_r^2}{\bar{\sigma}_p^2}\,n^{\text{eff}}. \tag{31}$$

In particular, $\mathrm{Var}(\hat{\alpha})$ *inherits the same concentration structure as $\mathcal{I}(\alpha)$: expanding the portfolio by adding positions that do not materially change n^{eff} or $\bar{\sigma}_p^2$ does not reduce the variance of $\hat{\alpha}$. More fundamentally, diluting concentration—spreading portfolio weight across more names while holding $\bar{\sigma}_p^2$ fixed—*raises *n^{eff} and therefore* increases *variance, the structural opposite of the effect of adding observations in linear regression.*

Proof. The sandwich form (30) is standard for NLS (see, e.g., Newey and McFadden, 1994). Under homoskedasticity, $\mathcal{V}(\alpha) = \sigma_r^2\,\mathcal{I}(\alpha)$, so the sandwich collapses to

$\sigma_r^2 \mathcal{I}(\alpha)^{-1}$. Substituting the information identity from Proposition 1, $\mathcal{I}(\alpha) = \bar{\sigma}_p^2/n^{\text{eff}}$, gives (31). The invariance claim is immediate: if n^{eff} and $\bar{\sigma}_p^2$ are unchanged, $\text{Var}(\hat{\alpha})$ is unchanged. □

Remark 3 (Structural inversion of the OLS information logic)**.** In OLS, information takes the form $\mathcal{I}_{\text{OLS}} = N \cdot \text{Var}(x)$: it is the product of *breadth* (sample size N) and regressor dispersion, and both factors help. In the NLS logit, equation (5) gives $\mathcal{I}(\alpha) = \text{HHI} \cdot \bar{\sigma}_p^2$: information is the product of *concentration* ($\text{HHI} = 1/n^{\text{eff}}$) and concentration-weighted price dispersion. Both factors again help, but the scale parameter has inverted: where OLS rewards breadth, the logit rewards concentration.

The inversion arises from the μ_j^2 weighting in the logit curvature. Each position's contribution to identification is quadratic in its portfolio share. Spreading the same total weight across more names dilutes each share and therefore reduces each curvature contribution faster than linearly. A portfolio with $n^{\text{eff}} = 50$ is not "like having 50 observations" in the OLS sense; it is like having an information budget that *shrinks* as the portfolio becomes more diffuse.

For the iShares Russell 3000 ETF in 2024Q4, $N = 1{,}090$ and $n^{\text{eff}} \approx 56$; for the S&P 500 tracker, $N = 449$ and $n^{\text{eff}} \approx 44$. The S&P 500 fund holds fewer names *and* has the lower n^{eff}, so it is, all else equal, *more* precisely estimated—an ordering that would be impossible under OLS-style information scaling, where more holdings always means more information.

This limitation motivates the pooled estimation and shrinkage strategy in Section 4.3: since a single fund–quarter's precision is determined by its concentration profile and price dispersion, not by the number of names in the portfolio, I gain statistical power by pooling across time (the rolling four-quarter baseline) and by shrinking noisy quarter-specific estimates toward each fund's own historical demand pattern.

3.2 Empirical manifestation and limitation

Propositions 1 and 2 have two distinct consequences for estimation—one numerical and one statistical. The numerical problem is solvable; the statistical one is not.

The numerical problem: poorly conditioned optimization. The raw NLS problem is badly scaled because the local curvature terms are weighted by μ_j^2. In long-tail portfolios, typical tail shares are tiny, so most observations contribute vanishing curvature, producing a poorly conditioned objective for standard optimizers. The estimand is well-defined and the estimator is consistent, but the optimizer cannot find it.

Figure 2 illustrates the instability. Without long-tail normalization (blue crosses), the broadest-universe portfolios (Russell 3000, Total Market) produce extreme $\hat{\alpha}$ values and inflated latent demand; the poorly conditioned objective surface means that

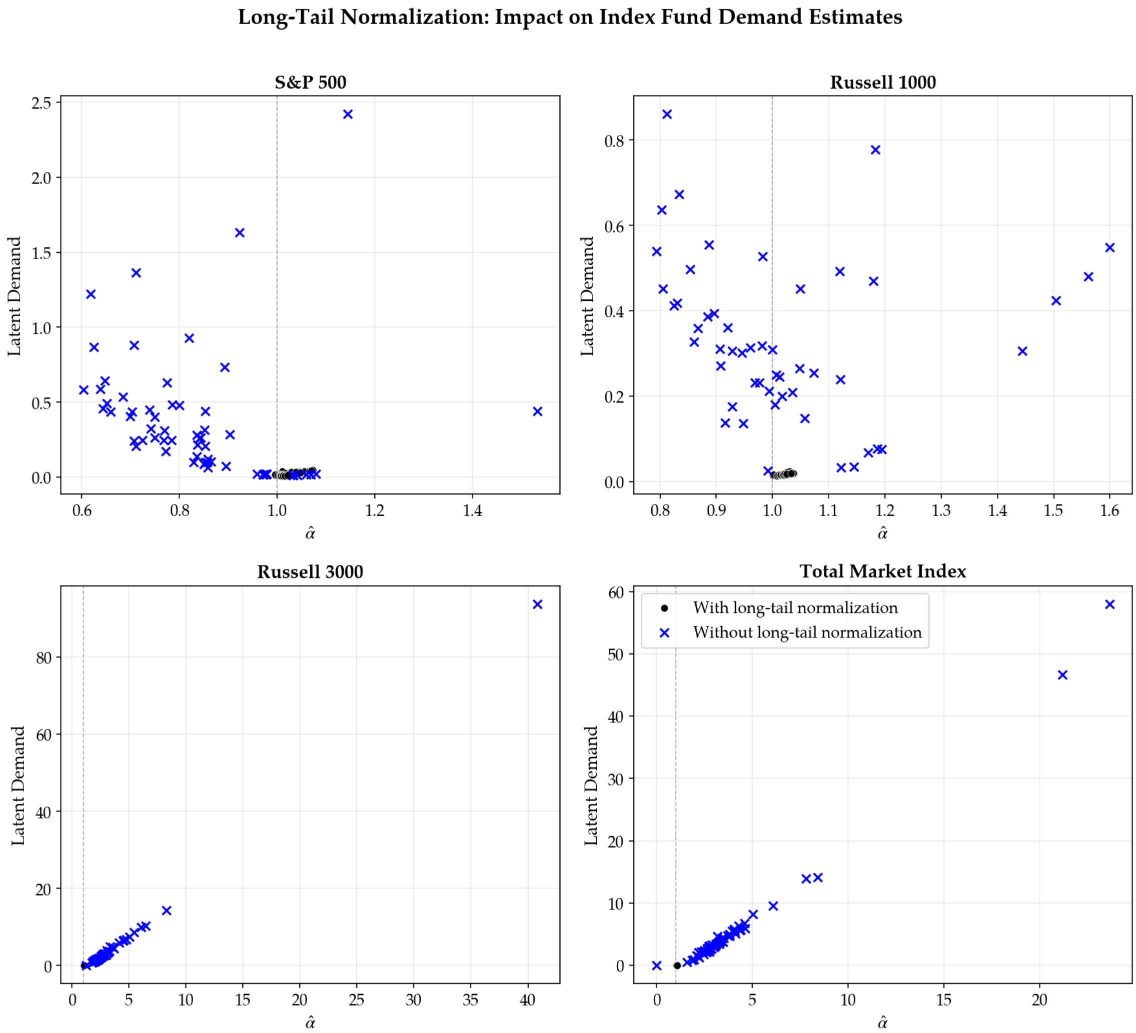

Figure 2: **The numerical problem: without preconditioning, broad-universe estimation is unstable.** Each panel plots fund-quarter estimates for index-like mutual funds matched to the indicated benchmark. Horizontal axis: estimated price coefficient $\hat{\alpha}$. Vertical axis: estimated latent-demand component. Blue crosses: estimates without long-tail normalization; black circles: with normalization. Without normalization, broad-universe portfolios (Russell 3000, Total Market) exhibit extreme $\hat{\alpha}$ values and inflated latent demand. With normalization, estimates concentrate near $\hat{\alpha} \approx 1$ and latent demand is small. The preconditioning resolves the numerical instability completely.

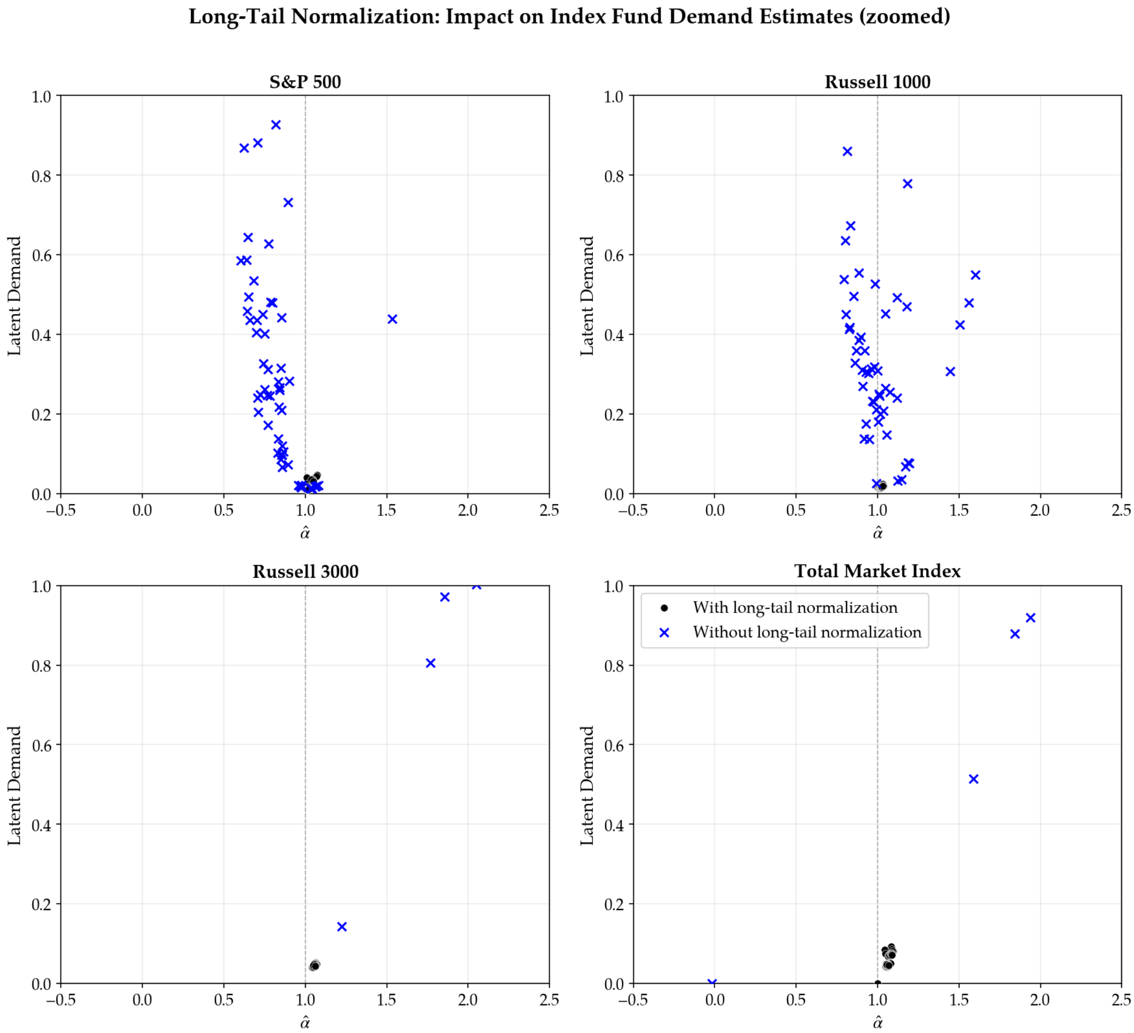

Figure 3: **The statistical limitation: residual dispersion is governed by n^{eff}, not N.** Same data as Figure 2, restricted to $\hat{\alpha} \in [-0.5, 2.5]$ and latent demand $\in [0, 1]$. After preconditioning, estimates are numerically well-behaved but exhibit irreducible dispersion. Narrow-universe funds (S&P 500, Russell 1000) cluster tightly; broad-universe funds (Russell 3000, Total Market) show wider spread. The point is not that normalization failed, but that dispersion remains even after preconditioning, consistent with the inference constraint in Proposition 2.

small perturbations in the few high-weight positions induce large movements in the estimate.

The solution is preconditioning. The long-tail normalization introduced in Section 4.3 (equation (42)) multiplies all moments for fund-quarter (f, t) by $s_{ft} = n^{\text{eff}}_{ft}$, rescaling gradients and Hessians to $O(1)$ without altering the minimizer. Because s_{ft} is constant across all $j \in \mathcal{U}_{ft}$, it preserves the relative weighting of core versus tail observations—the logit gradient's built-in μ_j downweighting is unchanged. Applying this normalization (black circles) collapses the instability: estimates concentrate near $\hat{\alpha} \approx 1$ and latent demand becomes small and stable.

The statistical limitation: concentration is the hard ceiling. Preconditioning solves the numerical problem, but it cannot relax the information bound in Proposition 1. The precision attainable from a single cross-section of holdings is determined by the portfolio's concentration profile, not by its nominal dimension, regardless of normalization.

Figure 3 makes this visible. After preconditioning, the S&P 500 and Russell 1000 panels show tight clustering near $\hat{\alpha} = 1$. The Russell 3000 and Total Market panels exhibit wider residual dispersion. The point is not that normalization failed, but that dispersion remains even after preconditioning, consistent with the inference constraint that Proposition 2 formalizes.

The distinction between the two problems determines the estimation design in the next section. The numerical problem is addressed by the n^{eff} normalization; the statistical problem—precision bounded by portfolio concentration rather than holdings count—is permanent, and must be addressed through pooling across time or shrinkage toward stable demand patterns.

4 Granular Demand Estimation Design

The preceding sections establish two consequences of Condition G for demand estimation. Section 2 shows that consolidated institutional demand is a convex aggregation of underlying mandates, making linear estimation on the aggregate structurally misspecified. Section 3 shows that even at the mandate level, the long tail of tiny portfolio weights limits the effective information content of the estimator to $n^{\text{eff}} = 1/\text{HHI}$ positions and creates a severe numerical conditioning problem.

This section provides a constructive estimation approach that addresses both issues. Section 4.1 describes the shift in observational unit from consolidated 13F institutions to individual mutual fund portfolios, which removes aggregation curvature by construction. Section 4.2 presents the demand specification and the interpretation of its key coefficients. Section 4.3 collects implementation details: the regressor basis, instrumental variables, long-tail normalization, pooled estimation and shrinkage, and identification filters.

4.1 Shift in observational unit

A note on scope. The estimates describe the U.S. mutual-fund sector—the institutional universe for which mandate-level holdings are observable in the CRSP Mutual Fund Database, and (Table 1) roughly 29% of market capitalization. This sector is heavily indexed and, when consolidated, sits among the most inelastic parts of the 13F universe; consolidating the same mandates into synthetic institutions reproduces the near-mechanical, low-elasticity estimates familiar from institution-level work (Koijen and Yogo, 2019), which validates the within-sample benchmark used throughout. The remaining 13F institutions, outside the mutual-fund panel, are known to exhibit higher elasticities, so the absolute magnitudes here should be read as the mutual-fund sector rather than the whole market. This scoping is a strength on two counts. First, the aggregation wedge is identified within the same holdings, so it is not an artifact of which institutions enter the sample. Second, because this sector is the hardest case for finding elastic demand, recovering more elasticity even here is conservative for the mechanism.

To isolate the role of financial holdings while maintaining comparability with canonical demand-system applications in asset pricing, I keep the core empirical environment as close as possible to the Koijen and Yogo (2019) setting. Stock prices, returns, dividends, shares outstanding, and identifiers are drawn from the CRSP Monthly Stock Database. I impose standard sample restrictions consistent with the prior literature (e.g., common shares and major exchanges), and I construct stock characteristics following the same templates used in recent demand-system implementations. In particular, the characteristic set is motivated by the Fama–French five-factor model and includes size-related measures, value (book-to-market or book equity), profitability, and investment, along with additional controls such as market beta and dividend measures. Throughout, the construction of characteristics and lag structure is aligned with the benchmark pipeline so that differences in results can be attributed to the observational unit and estimation design rather than to a different characteristic set.

The central change is the observational unit of holdings. Canonical applications often rely on consolidated institutional portfolios from SEC 13F filings, where the observed "institution" is typically a fund family or manager that aggregates multiple underlying mandates. In contrast, I use the CRSP Mutual Fund Database and treat each mutual fund portfolio as a mandate-like decision unit. Conceptually, this shift moves the data closer to the primitive decision maker and removes aggregation across heterogeneous sleeves by construction. Empirically, it replaces the institution-level rollup with fund-level holdings that are still large and diffuse, but substantially closer to a single investment style than a 13F filer.

I restrict attention to equity funds using Lipper classification codes, focusing on funds whose primary mandate is U.S. equity (Lipper `EQ`). This restriction serves two purposes. First, it improves comparability across portfolios by excluding multi-asset funds whose holdings and constraints differ qualitatively. Second, it sharpens the interpretation of demand coefficients as reflecting equity reallocation rather than cross-asset allocation decisions.

The CRSP Mutual Fund Database reports holdings at the share-class level, where multiple share classes (e.g., retail vs. institutional) may correspond to the same underlying portfolio but differ in fee structure, distribution channel, and investor eligibility. For the purposes of demand estimation, these share classes implement the same portfolio decision. I therefore aggregate share classes to the portfolio level (CRSP portfolio number) and compute portfolio AUM as the sum of AUM across share classes within the same portfolio. Holdings are correspondingly treated as portfolio holdings. This aggregation reduces mechanical duplication, aligns the observational unit with the portfolio choice, and avoids overweighting portfolios that simply offer many share classes.

I construct each fund's effective universe to mirror the benchmark demand-system setup. In each quarter t, the universe for a fund is defined as the union of securities observed in the fund's holdings at t and the securities that appeared in the fund's holdings over the preceding eleven quarters. This rolling-union universe matches the intuition that portfolio choice is made over a persistent opportunity set and ensures that the estimator does not mechanically treat previously held securities as unavailable options. It also maintains continuity with the large-universe logic in Koijen and Yogo (2019): funds choose from a broad set of securities, and the econometrician must account for the fact that most weights are small.

A limitation of the mutual-fund holdings data is coverage. The CRSP mutual fund holdings panel is available reliably from 2000 onward, and characteristic construction and lag requirements imply that the effective estimation sample begins in 2003Q4. While this window is shorter than what is available for 13F-based studies, it spans multiple market regimes (including the financial crisis and the post-2008 period) and is sufficient to make the paper's central point: mandate-level estimation delivers economically interpretable demand coefficients, and the remaining estimation challenges in finance stem from the granularity of portfolio weights rather than from institutional aggregation alone.

Tables 1 and 2 summarize the sample. Table 1 covers the mutual-fund sector from 2003–2024, reporting the number of funds, aggregate AUM, and the fraction of market capitalization held by mutual funds overall and by size bin. Table 2 reports portfolio granularity and tail-concentration statistics: AUM per fund, rolling-universe size and number of holdings (AUM-weighted mean and P90), and the fraction of holdings with weight below 1% and 0.1% together with the portfolio mass those positions represent. Its long-tail features motivate the remainder of Section 4: by 2024 an average fund holds over 300 names in a universe of roughly 600 securities, with 67% of positions carrying weight below 1% of the portfolio. This highly diffused portfolio geometry implies that the logit link is locally flat for much of the cross-section, weakening identification and creating conditioning challenges under exponential links—a feature I formalize in Section 4.3.

Table 1: Sample summary: coverage and mutual-fund ownership (2003–2024)

	# Funds	Mkt cap	% Market held (MF sector)					Total AUM
Year		($bn)	All	Large	Mid	Small	Micro	($bn)
2003	3,720	11,243.1	0.8	0.8	0.9	0.7	0.3	3,133.5
2004	3,936	13,487.0	0.6	0.6	0.7	0.6	0.3	4,146.2
2005	4,190	14,416.4	0.6	0.6	0.6	0.6	0.3	4,901.6
2006	4,591	15,692.9	0.6	0.6	0.6	0.6	0.4	5,937.5
2007	5,056	17,029.7	3.5	3.5	4.1	3.8	2.1	7,207.2
2008	6,053	13,100.5	14.5	14.1	18.6	17.4	8.3	6,253.4
2009	6,911	10,934.0	18.8	18.1	25.2	22.9	12.2	5,797.6
2010	7,070	13,169.3	24.5	23.7	31.1	29.8	16.5	7,217.2
2011	7,453	14,370.8	26.7	25.7	34.0	34.0	19.4	8,000.5
2012	7,752	15,639.6	26.7	25.8	33.7	33.5	18.9	8,665.0
2013	7,865	18,704.2	27.1	26.1	33.6	34.6	18.5	10,490.9
2014	8,182	21,745.3	27.6	26.6	34.2	34.9	18.1	12,377.7
2015	8,541	21,882.9	27.7	26.7	35.0	34.1	16.6	12,684.2
2016	8,782	22,223.0	28.3	27.4	36.7	34.7	17.2	13,116.0
2017	8,736	25,649.1	28.6	27.7	36.9	35.4	17.3	15,370.3
2018	8,754	27,506.4	29.1	28.2	37.4	36.8	17.2	16,192.5
2019	8,751	29,627.0	29.5	28.7	38.6	36.5	16.1	17,433.1
2020	8,584	32,191.8	29.1	28.5	37.1	34.1	13.7	17,817.3
2021	8,540	45,428.5	28.5	28.1	35.4	32.6	10.7	23,545.4
2022	8,807	39,561.2	28.5	28.0	37.7	30.4	7.9	20,665.3
2023	8,929	43,310.4	28.8	28.2	38.6	33.4	11.8	22,213.2
2024	9,007	54,291.4	29.0	28.4	38.8	34.8	14.6	26,798.0

Notes: Entries are annual averages across quarters within each year. "% Market held" is mutual-fund dollar holdings divided by total market capitalization in the corresponding stock universe. Size buckets follow the paper's size breakpoints. Initial estimations encompass all available years, limitations in data coverage restrict the analysis in Section 4.3 onward to the period beginning in the first quarter of 2010.

Table 2: Sample summary: portfolio granularity and tail concentration (2003–2024)

	AUM/fund	Universe size		# Holdings		$w < 1\%$		$w < 0.1\%$	
Year	Mean($mn)	Mean	P90	Mean	P90	% Hld	% Mass	% Hld	% Mass
2003	840.9	183	470	150	467	46.7	26.8	9.8	2.24
2004	1,052.9	236	473	157	467	49.0	29.0	11.4	2.48
2005	1,169.0	228	469	127	399	44.7	25.6	8.6	1.69
2006	1,292.8	226	474	139	461	44.2	25.2	11.1	2.30
2007	1,425.2	184	459	157	456	53.9	30.1	13.5	2.58
2008	1,056.8	268	473	184	450	55.7	31.4	15.8	3.19
2009	838.8	365	593	218	463	55.2	32.7	17.0	4.03
2010	1,020.6	358	662	216	460	54.7	33.3	16.1	4.13
2011	1,074.5	370	682	219	450	55.1	33.6	16.4	4.25
2012	1,117.8	381	700	223	450	55.0	32.8	17.3	4.42
2013	1,333.2	411	734	236	485	55.3	34.4	17.6	4.78
2014	1,512.7	432	810	244	545	55.5	34.9	18.1	4.93
2015	1,485.9	445	924	252	574	56.8	35.6	19.2	5.14
2016	1,493.6	462	1,157	263	698	57.8	36.1	20.7	5.43
2017	1,759.7	481	1,288	274	881	58.4	36.8	21.5	5.71
2018	1,849.9	499	1,584	280	1,035	58.7	36.6	22.9	6.02
2019	1,992.4	539	1,980	298	1,129	62.1	38.3	25.1	6.47
2020	2,077.4	536	1,993	291	1,083	62.7	37.3	26.5	6.21
2021	2,756.6	592	2,157	318	1,187	63.8	39.1	27.4	6.98
2022	2,347.9	618	2,351	313	1,127	64.4	39.3	27.7	6.81
2023	2,487.6	615	2,592	315	1,097	65.0	38.6	28.2	6.74
2024	2,974.9	615	2,610	317	1,034	67.3	38.1	30.8	6.99

Notes: Entries are annual averages across quarters within each year. Universe size is the rolling-union universe (current quarter plus prior 11 quarters); holdings counts are within the same universe. Mean and P90 for universe size, holdings, and concentration columns are AUM-weighted averages across funds. "% Hld" is the fraction of holdings with weight below the indicated threshold; "% Mass" is the total portfolio weight in those positions.

4.2 Specification and interpretation

Variables and characteristics. Let f index a mutual fund, j index stocks, and t index quarters. For each fund-quarter (f, t), let $\mathcal{U}_{ft}$ denote the fund's effective universe and let $w_{fjt} \geq 0$ be the observed portfolio weight for stock $j \in \mathcal{U}_{ft}$. Define the within-universe total

$$S_{ft} \equiv \sum_{j \in \mathcal{U}_{ft}} w_{fjt}. \tag{32}$$

When funds hold cash or assets outside $\mathcal{U}_{ft}$, S_{ft} need not equal one; in that case I normalize the holdings by their within-universe total so that $\sum_{j \in \mathcal{U}_{ft}} w_{fjt} = 1$. The estimator thus identifies demand parameters from relative allocations within $\mathcal{U}_{ft}$.

The characteristic vector x_{jt} collects stock attributes motivated by the Fama–French five-factor model:

- $\log(\mathrm{B/M})_{jt}$: log book-to-market ratio;
- OP_{jt}: operating profitability (revenue minus cost of goods sold, SGA, and interest, divided by lagged book equity, annualized over a trailing four-quarter window);
- Inv_{jt}: asset growth (change in total assets over four quarters, scaled by lagged total assets);
- $\beta^{mkt}_{j,t-1}$: market beta estimated from a trailing 36-month rolling regression of excess stock returns on excess market returns;
- I^{div}_{jt}: a dividend-payer indicator (equals one if the firm paid a positive dividend within the trailing twelve months); and
- $\log(1 + D/\mathrm{BE})_{jt}$: log dividend magnitude among payers (dividends to book equity, set to zero for non-payers).

All continuous characteristics are winsorized at the 2.5th and 97.5th percentiles each quarter and then standardized to zero mean and unit variance. The dividend specification decomposes the extensive margin (payer vs. non-payer) from the intensive margin (how much, conditional on paying), following the observation that a large fraction of stocks pay no dividends and a single dividend variable would be dominated by the mass point at zero.

The demand specification. My specification follows the relative-demand component of Koijen and Yogo (2019)'s equation (10). For each (f, t) and $j \in \mathcal{U}_{ft}$, define a linear index

$$\eta_{fjt}(\theta_{ft}) = \alpha_{ft} \log \mathrm{ME}_{jt} + \beta^{BM}_{ft} \log(\mathrm{B/M})_{jt} + X'_{jt}\gamma_{ft}, \tag{33}$$

where $\theta_{ft} \equiv (\alpha_{ft}, \beta_{ft}^{BM}, \gamma_{ft})$ are the parameters of interest, and the implied within-equity share

$$\mu_{fjt}(\theta_{ft}) = \frac{\exp(\eta_{fjt})}{\sum_{k \in \mathcal{U}_{ft}} \exp(\eta_{fkt})}, \tag{34}$$

so that $\sum_{j \in \mathcal{U}_{ft}} \mu_{fjt}(\theta_{ft}) = 1$ for all θ_{ft}. The residual is $u_{fjt}(\theta_{ft}) = w_{fjt} - \mu_{fjt}(\theta_{ft})$.

The object estimated here is therefore *within-equity demand*: given the fund's allocation to the equity universe, how does it allocate across stocks? The main implementation difference from the benchmark outside-option specification is that I do not estimate a free intercept. In Koijen and Yogo (2019), the intercept governs the level of inside-equity demand relative to the outside option and is identified through the residual normalization. Once holdings are normalized within the equity universe, any common level shift cancels from the logit (equation (36)) and is not identified from within-equity share moments. The next paragraph derives the exact mapping from the outside-option formulation to the within-equity specification used here.

Scope of identification: within-equity demand. It is worth making explicit the relation between this within-equity specification and the outside-option formulation in Koijen and Yogo (2019). Write Koijen and Yogo (2019)'s equation (10) compactly as

$$\frac{w_i(n)}{w_i(0)} = \exp\{c_i + r_n' \theta_i\}\, \epsilon_i(n),$$

where $c_i \equiv \beta_{K,i}$ collects the intercept and $r_n' \theta_i$ collects the nonconstant characteristics. The budget constraint over the outside option and the inside assets implies

$$w_i(n) = \frac{\exp\{c_i + r_n' \theta_i\}\, \epsilon_i(n)}{1 + \sum_{k \in \mathcal{N}_i} \exp\{c_i + r_k' \theta_i\}\, \epsilon_i(k)}, \quad w_i(0) = \frac{1}{1 + \sum_{k \in \mathcal{N}_i} \exp\{c_i + r_k' \theta_i\}\, \epsilon_i(k)}.$$

The intercept c_i shifts total inside demand relative to the outside option. It is not identified from relative inside shares; it is identified by the outside-option odds together with the residual normalization $\mathbb{E}[\epsilon_i(n) \mid Z_i, x] = 1$.

Conditioning on the inside equity universe eliminates this margin algebraically. The latent residuals enter the conditional shares unchanged, while c_i cancels exactly:

$$\frac{w_i(n)}{\sum_{\ell \in \mathcal{N}_i} w_i(\ell)} = \frac{\exp\{r_n' \theta_i\}\, \epsilon_i(n)}{\sum_{\ell \in \mathcal{N}_i} \exp\{r_\ell' \theta_i\}\, \epsilon_i(\ell)}. \tag{35}$$

The conditional equity shares therefore identify the relative slopes θ_i, but not the equity-vs-outside intercept c_i. The model-implied conditional shares $\mu_{fjt}(\theta_{ft})$ in equation (34) correspond to the same expression evaluated at $\epsilon_i \equiv 1$, so the residual $u_{fjt} = w_{fjt} - \mu_{fjt}$ inherits the latent residual variation.

The distinction is conceptual rather than merely numerical. A free intercept is meaningful in the Koijen and Yogo (2019) model because the left-hand side is $w_i(n)/w_i(0)$ and the outside option provides a level against which inside equity demand is measured. Once the data are normalized within the equity universe, the same intercept is not identified by share moments: for any scalar a,

$$\frac{\exp\{a + r'_{jt}\theta\}}{\sum_k \exp\{a + r'_{kt}\theta\}} = \frac{\exp\{r'_{jt}\theta\}}{\sum_k \exp\{r'_{kt}\theta\}}. \tag{36}$$

Because the paper's aggregation results concern within-equity slopes, I do not reintroduce a Koijen and Yogo (2019)-style intercept, which would require either an additional outside-option measurement or an auxiliary normalization on the latent residual.

Interpreting α: the price coefficient. The specification (33) separates two economically distinct dimensions: $\log \mathrm{ME}$ captures market capitalization, while $\log(\mathrm{B/M})$ captures valuation. Because $\log \mathrm{ME}_{jt} = \log P_{jt} + \log \mathrm{Shares}_{jt}$ and shares outstanding move slowly relative to prices, cross-sectional and short-horizon time-series variation in $\log \mathrm{ME}$ is dominated by price variation. Following Koijen and Yogo (2019), I therefore interpret α as a *price coefficient*: it governs how portfolio weights respond to price movements, holding fundamentals fixed (since β^{BM} absorbs valuation).

A mechanically passive value-weighted mandate—one that holds each stock in proportion to its market capitalization and does not rebalance—implies $\alpha = 1$: a one-percent increase in a stock's price raises its dollar weight by approximately one percent.[4] Values $\alpha < 1$ indicate that the fund reduces its *relative* quantity allocation as prices rise, partially offsetting the mechanical price effect—i.e., downward-sloping demand.

This object is the natural benchmark for the sanity checks in Section 5.1: I expect passive large-cap trackers to cluster near $\alpha \approx 1$, with active mandates systematically below.

Interpreting β^{BM}: the valuation tilt. Holding market capitalization fixed, variation in $\log(\mathrm{B/M})$ captures differences in valuation. The coefficient β^{BM}_{ft} governs the fund's sensitivity to this dimension:

- $\beta^{BM} > 0$: overweights high-B/M (value) stocks $\Rightarrow$ value tilt;

[4]This interpretation is exact in the frictionless mechanical benchmark where the fund does not adjust quantities when prices move and the relevant universe is fixed. For style and segmented indices (e.g., mid-cap or small-cap mandates), constituent changes and reconstitution can generate quantity adjustments when firms migrate across bins, so α need not equal one even for "passive" products. In these cases α should still be close to one if most variation is within-bin price movement rather than membership changes.

- $\beta^{BM} < 0$: overweights low-B/M (growth) stocks $\Rightarrow$ growth tilt;
- $\beta^{BM} \approx 0$: no valuation tilt (broad market trackers).

Separating α from β^{BM} avoids conflating the price response with valuation tilts: a coefficient on $\log \text{ME}$ alone would mechanically mix the fund's response to price movements with its growth-versus-value preference.

4.3 Implementation

This section collects the practical details of the estimation procedure. The baseline specification (33)–(34) defines the model; here I describe how it is taken to data.

Regressor implementation: $\log \textbf{ME} + \log \textbf{BE}$. The specification (33) uses $\log \text{ME}$ and $\log(\text{B/M})$ as the two size-related regressors. In implementation, I replace $\log(\text{B/M})$ with $\log \text{BE}$ and estimate the index as

$$\eta_{fjt} \;=\; \alpha^*_{ft} \log \text{ME}_{jt} \;+\; \beta^{BE}_{ft} \log \text{BE}_{jt} \;+\; X'_{jt}\gamma_{ft}, \tag{37}$$

where the asterisk on α^* distinguishes the coefficient in this basis from the conceptual α in (33). Since $\log(\text{B/M}) = \log \text{BE} - \log \text{ME}$, the two specifications are algebraically equivalent, with the identification

$$\alpha_{ft} = \alpha^*_{ft} + \beta^{BE}_{ft}, \qquad \beta^{BM}_{ft} = \beta^{BE}_{ft}. \tag{38}$$

The practical advantage of the $\{\log \text{ME}, \log \text{BE}\}$ basis is that $\log \text{BE}$ is predetermined and not mechanically linked to the endogenous price regressor $\log \text{ME}$. This simplifies instrumentation: $\log \text{BE}$ serves as its own instrument, whereas $\log(\text{B/M})$ would inherit the endogeneity of $\log \text{ME}$ and require separate instrumentation. I report all results in the interpretable basis (α, β^{BM}) using the conversion (38).

In implementation, all regressors are standardized to zero mean and unit variance each quarter. The price coefficient in raw (log-dollar) units is

$$\alpha_{ft} \;=\; \frac{\hat{\alpha}^{*,\text{std}}_{ft}}{\sigma_{\log \text{ME}}} \;+\; \frac{\hat{\beta}^{BE,\text{std}}_{ft}}{\sigma_{\log \text{BE}}}, \tag{39}$$

with standard error from the delta method.[5]

[5] Let V_{ft} denote the estimated variance–covariance matrix of the standardized coefficient vector and let indices 1 and k correspond to $(\alpha^{*,\text{std}}, \beta^{BE,\text{std}})$. Then $\text{se}(\alpha_{ft}) \;=\; \left(V_{11}/\sigma^2_{\log \text{ME}} + V_{kk}/\sigma^2_{\log \text{BE}} + 2V_{1k}/(\sigma_{\log \text{ME}}\sigma_{\log \text{BE}})\right)^{1/2}$. In practice, V_{1k} is often materially negative because $\log \text{ME}$ and $\log \text{BE}$ are highly correlated, so α can be more precisely estimated than α^* and β^{BE} individually.

Instrumental variables. Because portfolio weights w_{fjt} and stock prices p_{jt} are jointly determined in equilibrium, α is subject to the standard simultaneity concern. Following Koijen and Yogo (2019), I instrument for $\log \mathrm{ME}_{jt}$ using a measure of the stock's common fund flow pressure:

$$\mathrm{cf}_{jt}^{\mathrm{me}} = \sum_{f:\, j \in \mathcal{U}_{ft}} \frac{A_{ft}}{1 + |\mathcal{U}_{ft}|}, \tag{40}$$

where A_{ft} is fund f's total AUM. The instrument is $z_{jt} = \log(\mathrm{cf}_{jt}^{\mathrm{me}})$, standardized each quarter. To address the reflection problem, I use a leave-one-out construction: for each fund-stock pair (f, j), the instrument z_{fjt}^{loo} is the analogue of z_{jt} computed after excluding fund f's contribution to the common-flow sum in (40). The moment conditions are $\mathbb{E}[z_{fjt}^{\mathrm{loo}}\,(w_{fjt} - \mu_{fjt}(\theta_{ft}))] = 0$, with all exogenous characteristics serving as their own instruments.

Estimation criterion. Fund-quarter parameters are estimated by minimizing a quadratic form in the IV moment conditions:

$$\widehat{\theta}_{ft} \in \arg\min_{\theta} \left\| \sum_{j \in \mathcal{U}_{ft}} z_{fjt} \left(w_{fjt} - \mu_{fjt}(\theta) \right) \right\|^2, \tag{41}$$

where $z_{fjt} = (z_{fjt}^{\mathrm{loo}}, x_{jt}')$ stacks the leave-one-out price instrument with the exogenous characteristics. Standard errors use the heteroskedasticity-robust sandwich formula.

Long-tail normalization. As established in Proposition 1, score gradients scale with $\mu_j = O(1/N)$ for tail positions, creating a severe numerical conditioning problem even though the estimand is well-defined. To stabilize estimation, I apply a fund-specific effective-number normalization that rescales the entire moment system without changing the minimizer. Define

$$\mathrm{HHI}_{ft} \equiv \sum_{j \in \mathcal{U}_{ft}} \left(\frac{w_{fjt}}{S_{ft}} \right)^2, \qquad n_{ft}^{\mathrm{eff}} \equiv \frac{1}{\mathrm{HHI}_{ft}}, \qquad s_{ft} \equiv n_{ft}^{\mathrm{eff}}. \tag{42}$$

Multiplying all weights by s_{ft} rescales the objective without changing its minimizer under the quadratic criterion. This is a pure preconditioning step that aligns the numerical scale of the optimization with its statistical scale (see Section 3 for the empirical demonstration).

Pooled estimation and shrinkage. Following the pooled-estimation approach of Koijen et al. (2024), I stabilize coefficient panels using a two-stage empirical-Bayes procedure that exploits the time-series persistence of fund-level demand.

Stage 1: Rolling pooled baseline. For each fund-quarter (f, t), I first estimate a fund-specific baseline using the four-quarter holdings panel ending in quarter t. Specifically, I stack the fund's holdings over this rolling window and estimate a single coefficient vector $\bar{\theta}_{ft}$ by applying the logit IV-GMM criterion (41) to the stacked data, while maintaining separate logit normalizations within each quarter. This pooled estimate is used only as a fund-specific baseline; no shrinkage is applied at this stage.

Stage 2: Quarter-t estimates with shrinkage. I then estimate the quarter-t coefficient vector $\widehat{\theta}_{ft}$ using only the fund's quarter-t holdings, and shrink it toward the fund's own rolling baseline:

$$\widetilde{\alpha}_{ft} \;=\; (1-\lambda_{ft})\,\widehat{\alpha}_{ft} \;+\; \lambda_{ft}\,\bar{\alpha}_{ft}, \tag{43}$$

where $\lambda_{ft} = \lambda_\alpha/(\lambda_\alpha + N_{ft})$ and N_{ft} denotes the number of stock-level observations in the fund-quarter. Noisier estimates from smaller effective samples are therefore pulled more strongly toward the fund's own baseline, while estimates from richer cross-sections remain closer to the raw single-quarter estimate. The same empirical-Bayes shrinkage rule is applied component by component to the remaining characteristic coefficients, with coefficient-specific shrinkage parameters.

Identification filters. Not all fund-quarters are well-identified. I apply three conservative filters before aggregation. First, fund-quarters with a joint significance statistic below 10 (Wald/F-type test that all slopes are zero) are flagged as weakly identified. Second, fund-quarters with $n^{\text{eff}}_{ft} < 20$ are excluded.[6] Third, for each characteristic x^k, I compute the weight-weighted dispersion

$$\text{Disp}_{ft}(x^k) \;\equiv\; \left(\sum_{j\in\mathcal{U}_{ft}} \frac{w_{fjt}}{S_{ft}} \left(x^k_{jt} - \bar{x}^k_{ft}\right)^2\right)^{1/2} \quad ; \tag{44}$$

if this falls below 0.1 in standardized units, the corresponding $\widehat{\beta}^k_{ft}$ is set to missing. These filters exclude roughly 15–20% of fund-quarters, but are concentrated among portfolios with small AUM.

Section 5 uses the resulting panel $\{\widehat{\alpha}_{ft}, \widehat{\beta}^{BM}_{ft}, \widehat{\gamma}_{ft}\}$ (and its shrunk counterpart when stated) to benchmark estimates against well-known funds and to document the cross-sectional distribution of characteristic demand.

[6] This threshold removes sector specialists, extremely concentrated mandates, and mechanically thin universes where cross-sectional moments are unstable.

5 Estimation Results and the Aggregation Wedge

The preceding sections establish two points. First, under the granular-portfolio condition, identification and precision are governed by the concentration of economically meaningful holdings rather than by the raw number of holdings. Second, when heterogeneous mandates are aggregated into an institution-level portfolio, the resulting demand object is no longer linear in characteristics, so a standard linear demand regression estimates a projection of this curved object rather than the underlying average taste. This section shows how these theoretical predictions appear in the data.

The argument proceeds in three steps. First, I show that the aggregation wedge is systematic and quantitatively large: a same-data, different-estimator comparison reveals that consolidation shifts the price coefficient upward for 24 of the 30 largest fund families (80%), with a mean wedge of $+0.12$ (Section 5.1). Second, having established what the wedge looks like, I show what demand estimates look like once it is removed: mandate-level coefficients are well-behaved, economically interpretable, and imply stock-level elasticities two to four times larger than existing institution-level estimates (Section 5.2). Third, the granular residuals reveal economically structured cross-fund disagreement—contested versus consensus names—that a scalar institution-level residual would collapse to near zero (Section 5.3).

5.1 Within-institution heterogeneity and the aggregation wedge

The central empirical claim of the paper is that the extreme inelasticity often reported in demand-system asset pricing reflects institutional aggregation rather than a primitive feature of investor demand. The cleanest test is therefore a same-data, different-estimator comparison: start from the same underlying mandate-level holdings, estimate demand at the fund level, then aggregate those same mandates into a synthetic institution and re-estimate the linear model on the consolidated portfolio. If the theory in Section 2 is right, the consolidated estimate should be systematically shifted toward the near-unit region, even though the underlying mandate-level estimates are well behaved. Consolidating these mandates also reproduces the near-mechanical, low-elasticity coefficients reported for institution-level 13F holdings (Koijen and Yogo, 2019), which validates the synthetic benchmark and ensures the comparison isolates aggregation rather than differences in sample or method.

Case study: BlackRock. BlackRock is useful because it makes the mechanism visible. What appears in 13F data as a single institution is, in practice, a collection of heterogeneous products: broad index trackers, active mandates, style-specific portfolios, dividend strategies, and minimum-volatility funds. Figures 4 and 5 report mandate-level

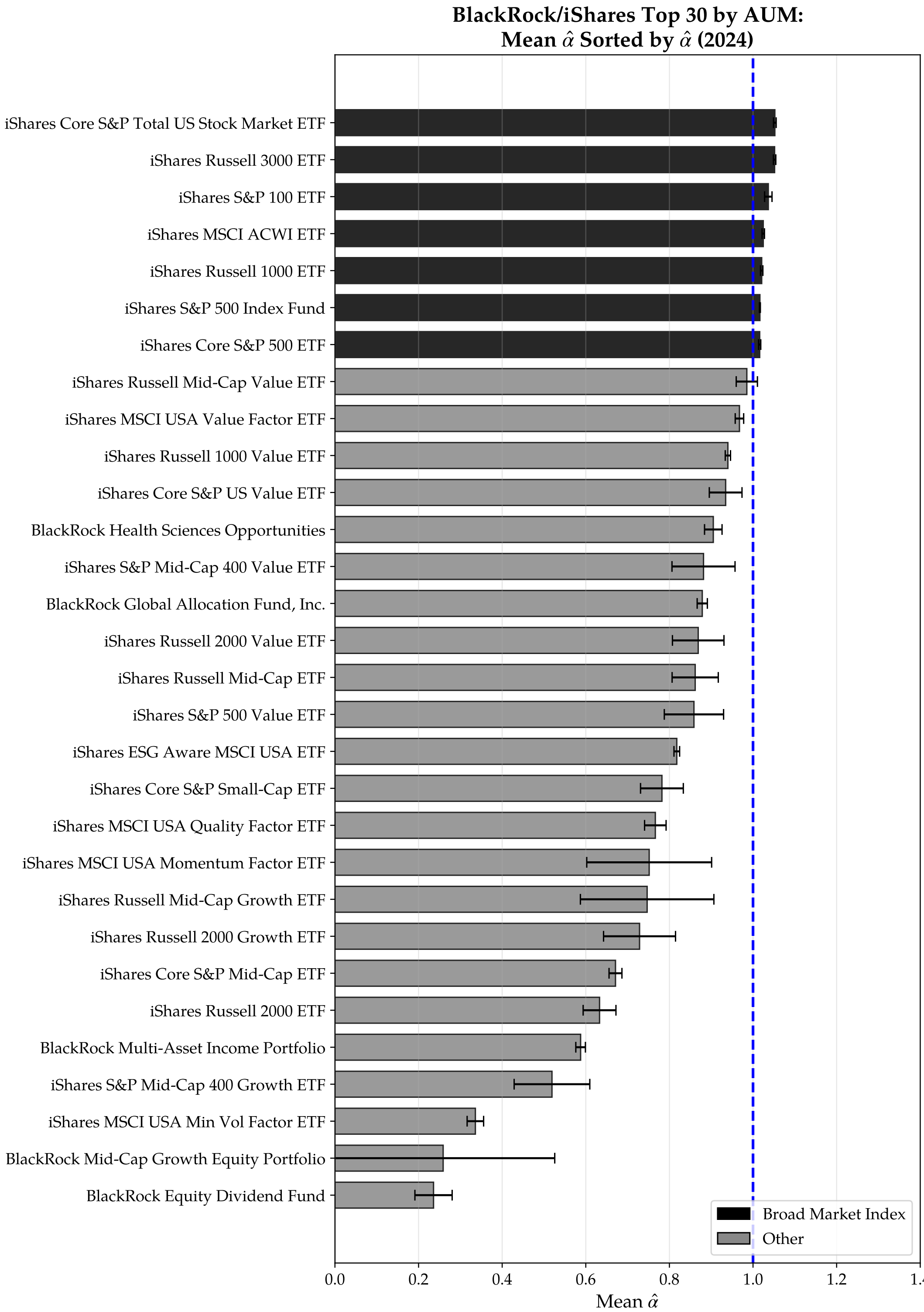

Figure 4: Mandate-level price coefficient estimates for BlackRock/iShares flagship funds (2024). Each fund's mean estimated $\hat{\alpha}$ (horizontal bars) with 95% confidence intervals (whiskers), for the top 30 products ranked by AUM. Funds are ordered by $\hat{\alpha}$; the dashed vertical line marks $\hat{\alpha} = 1$. Broad market index funds (black) are highlighted relative to other strategies (gray).

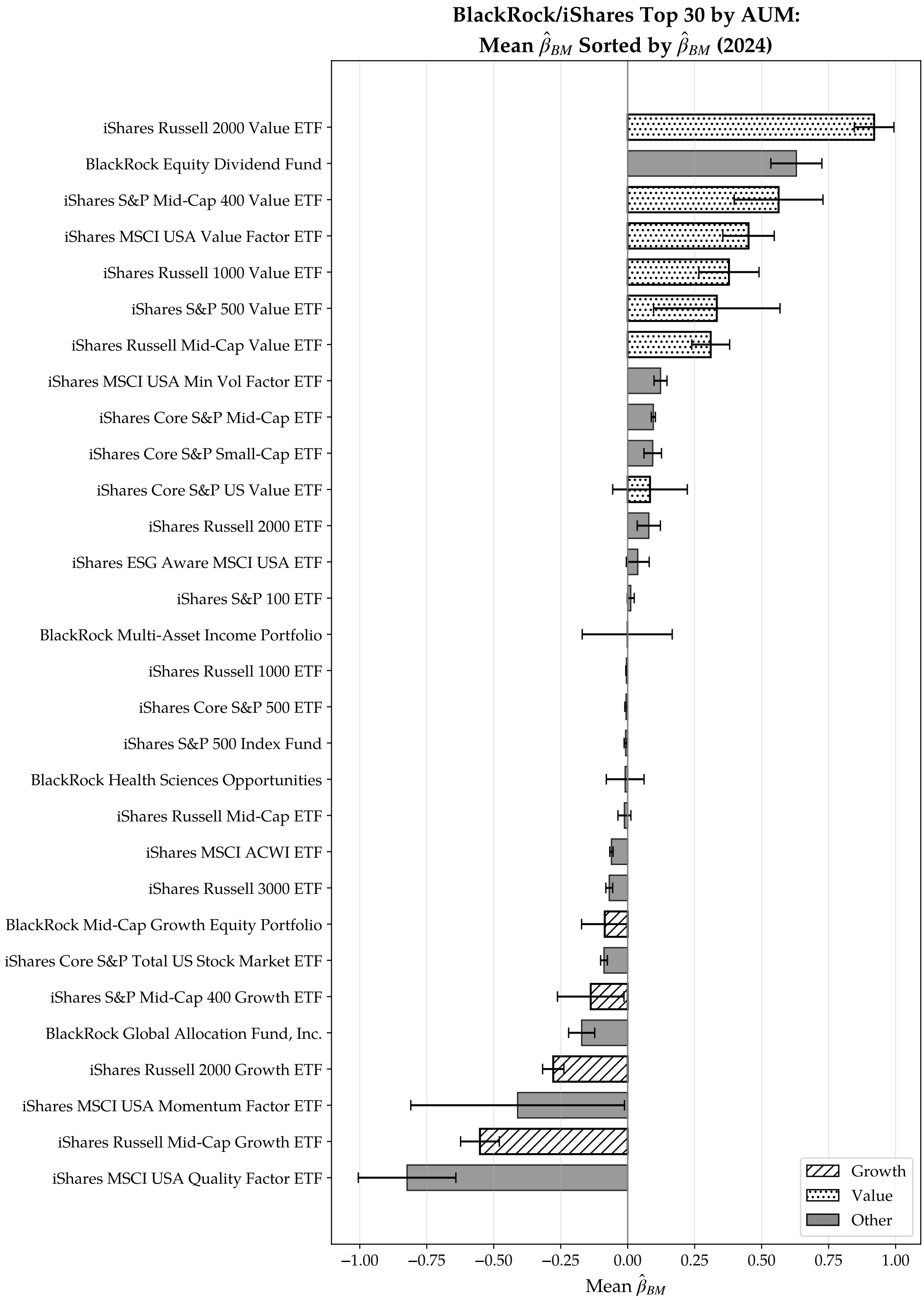

Figure 5: Mandate-level valuation-tilt coefficient estimates for BlackRock/iShares flagship funds (2024). Each fund's mean estimated $\hat{\beta}^{BM}$ with 95% confidence intervals. Value funds exhibit positive loadings, growth funds exhibit negative loadings, and broad market trackers cluster near zero.

estimates for the top 30 BlackRock/iShares products by AUM in 2024. Figure 4 shows substantial dispersion in the estimated price coefficient $\hat{\alpha}$: large-cap passive products cluster near $\hat{\alpha} \approx 1$, consistent with near-mechanical tracking, while actively tilted mandates lie systematically below one, indicating more elastic reallocation. Figure 5 shows that the valuation-tilt coefficient $\hat{\beta}^{BM}$ is also economically structured: value funds load positively, growth funds negatively, and broad trackers remain close to zero. The point of the case study is not merely that coefficients differ across funds, but that the within-institution heterogeneity is economically organized rather than idiosyncratic.

The wedge across fund families. The BlackRock example illustrates the mechanism in one institution. I now scale the comparison to the matched 2024Q4 sample. For each fund family, I compare two objects computed from the same underlying mandate-level holdings: (i) the AUM-weighted mean granular estimate, $\bar{\alpha}$, across the family's funds, and (ii) the synthetic institution estimate, $\hat{\alpha}^{13F}$, obtained by consolidating all mandates in the family into one portfolio and re-estimating the linear specification.

Table 3 reports this comparison for the 30 largest families. The AUM-weighted mean granular estimate is $\bar{\alpha} = 0.84$, which implies clearly downward-sloping demand. The corresponding synthetic 13F estimate is $\hat{\alpha}^{13F} = 0.96$, implying demand much closer to the near-mechanical benchmark. The AUM-weighted mean wedge, $\hat{\alpha}^{13F} - \bar{\alpha}$, is therefore $+0.12$, and the wedge is positive for 24 of the 30 largest families. The pattern is broad rather than driven by a few outliers. Vanguard shows a modest wedge of $+0.07$, consistent with its large passive core. Capital Group exhibits a larger shift of $+0.26$. BlackRock is slightly negative at -0.04, because its AUM is heavily concentrated in a few very large index mandates, so its AUM-weighted granular mean is already close to one. By contrast, families with substantial active AUM such as Franklin Templeton ($+0.61$), Victory ($+0.44$), and T. Rowe Price ($+0.34$) display much larger wedges.

Where the curvature appears. The family-level comparison also clarifies why the wedge need not be monotone in a simple scalar measure of within-institution heterogeneity. The curvature term generated by mandate-level dispersion can affect the consolidated estimate through two channels. The component that projects onto the price regressor shifts the estimated coefficient, while the orthogonal remainder is absorbed by the residual and appears as inflated latent demand. The split between these two channels depends on the joint distribution of prices and characteristics in the institution's stock universe. As a result, two families with similar underlying heterogeneity can show different coefficient wedges. The theory does not predict a monotone family-by-family mapping from heterogeneity to the $\hat{\alpha}$ wedge alone. It predicts instead that the total curvature effect—the coefficient wedge plus residual inflation—rises with

Table 3: Top 30 fund families by matched AUM: granular versus synthetic 13F estimates, 2024Q4

Institution	N	AUM ($bn)	Granular $\bar{\alpha}$	13F $\hat{\alpha}$	Wedge
Vanguard	72	5161.6	0.968	1.040	+0.072
Capital Group	20	1785.4	0.719	0.983	+0.264
BlackRock	114	1714.4	0.882	0.841	−0.041
Fidelity	138	1623.8	0.870	0.960	+0.089
State Street	49	887.0	0.972	1.045	+0.073
Invesco	106	646.8	0.610	1.032	+0.422
Charles Schwab	21	398.7	0.932	0.911	−0.021
JPMorgan	45	283.1	0.594	0.952	+0.358
T. Rowe Price	45	273.1	0.508	0.852	+0.343
Dimensional	27	262.8	0.781	0.667	−0.115
Franklin Templeton	44	196.4	0.414	1.022	+0.607
MFS	19	131.0	0.307	0.650	+0.343
Jackson National	37	108.5	0.596	0.869	+0.274
Advanced Series	17	95.7	0.775	0.921	+0.146
Nuveen	29	90.3	0.804	0.864	+0.060
Equitable	44	77.2	0.871	0.817	−0.053
Columbia	31	72.7	0.529	0.916	+0.387
American Century	34	71.1	0.510	0.546	+0.036
John Hancock	25	67.5	0.311	0.651	+0.340
Janus	11	67.4	0.530	0.662	+0.133
First Trust	54	64.1	0.230	0.534	+0.304
Principal	15	63.2	0.922	0.745	−0.177
Victory	41	58.5	0.172	0.614	+0.442
AllianceBernstein	23	48.3	0.536	0.721	+0.186
ProShares†	21	44.8	0.862	1.116	+0.254
Goldman Sachs	34	43.7	0.643	0.855	+0.212
Pacer Financial	15	40.3	0.587	0.402	−0.185
Allspring	25	39.9	0.313	0.521	+0.208
Hartford	22	39.0	0.419	0.838	+0.418
Lincoln	35	38.9	0.673	0.836	+0.163
AUM-weighted mean			0.837	0.959	+0.122

Notes: The table reports the 30 largest fund families in the matched 2024Q4 sample, ranked by family-level AUM. N is the number of matched granular mandates. AUM is the matched AUM: only mandates classified as equity funds that pass the identification filters (sufficient n^{eff}, F-test for coefficient significance) contribute; the same filtered mandates are then consolidated into a synthetic 13F portfolio so that the only variation between the two columns is the level of aggregation. $\bar{\alpha}$ is the AUM-weighted mean of fund-level granular estimates of α within the family. $\hat{\alpha}$ is the corresponding estimate from the synthetic 13F aggregation. *Wedge* is defined as $\hat{\alpha}^{13F} - \bar{\alpha}^{\text{granular}}$, so positive values indicate that aggregation raises the estimated coefficient relative to the matched granular benchmark. †13F estimate at the winsorization bound (1st/99th percentile).

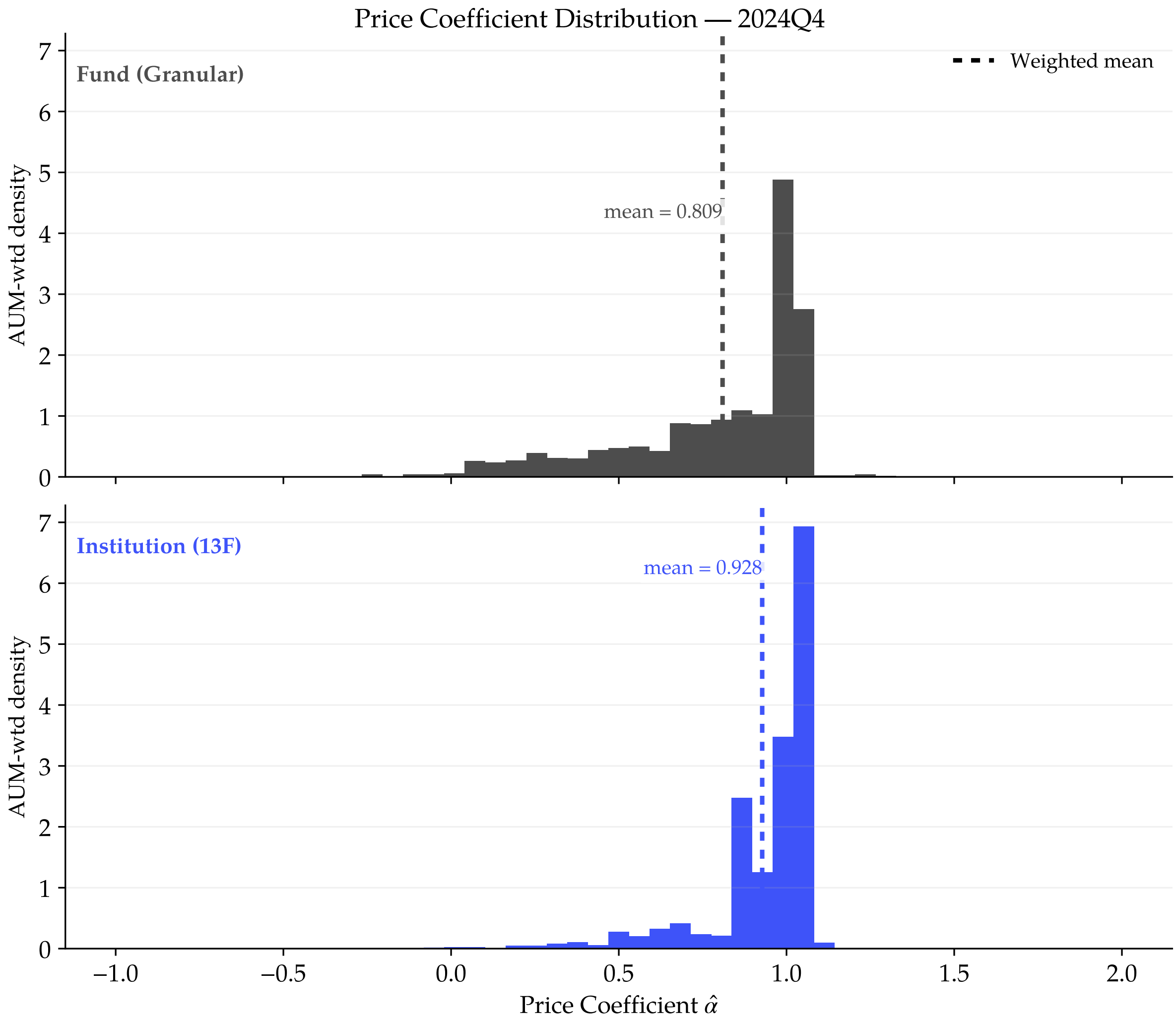

Figure 6: AUM-weighted distribution of the price coefficient $\hat{\alpha}$ at the fund level (top) versus the synthetic 13F institution level (bottom), 2024Q4. The rightward shift of the 13F distribution toward values near or above one illustrates the aggregation wedge: demand that is clearly downward-sloping at the mandate level appears near-mechanical at the institution level.

within-institution dispersion, and that the coefficient channel is directionally shifted toward attenuation.

Distributional implication. Figure 6 shows the aggregate consequence. It plots the AUM-weighted distribution of $\hat{\alpha}$ across all granular fund-level estimates and compares it with the distribution obtained from synthetic 13F aggregation in 2024Q4. The granular distribution centers around 0.81 and has a substantial left tail extending into clearly active mandates. The synthetic 13F distribution shifts rightward, with mass concentrated near 1.0 and a much thinner left tail. The AUM-weighted mean rises from 0.81 to 0.93, a roughly 15% shift toward the near-unit range that the literature has often interpreted as implausibly inelastic demand. Because both distributions are constructed from the same underlying holdings, the shift is entirely a consequence of aggregation.

Taken together, these results support the mechanism developed in Section 2. Institution-level demand estimates are not simply noisier versions of mandate-level demand. Consolidation changes the object being estimated. By mixing heterogeneous sleeves into one observed portfolio, it pushes the linear price coefficient toward the near-unit region and shifts the remainder of the curvature into measured latent demand.

5.2 Mandate-level estimates and price elasticity

Having established that the aggregation wedge is systematic and quantitatively large, I now examine what demand estimates look like once the wedge is removed.

Coefficient distributions over time. Figure 7 summarizes the distribution of characteristic-demand coefficients across funds over time. For each quarter, I estimate fund-level demand parameters using the granular (mandate-level) GMM procedure described in Section 4, and report the cross-sectional median (solid line) together with the interquartile range (shaded band, 25th–75th percentiles).

The panel for the price coefficient (log market equity) is particularly informative. Recall that α governs how a fund's dollar weight responds to price movements. Values near one correspond to nearly mechanical tracking of value weights (and therefore very low implied quantity adjustment), while lower values indicate more downward-sloping demand in the sense that quantities are reduced as prices rise. The estimates imply substantial heterogeneity: across most of the sample, the interquartile range of α lies roughly between 0.1 and 0.6, with the median in the middle of that band. These values imply economically meaningful but not extreme price sensitivity—far less inelastic than the near-unit or upward-sloping behavior obtained from institution-level estimation on consolidated holdings.

The remaining characteristic loadings exhibit similarly interpretable patterns. Market-to-book and book-equity panels show persistent nonzero medians, consistent with stable demand tilts toward (or away from) value-related attributes. Profitability and investment loadings are generally positive in the median. The market-beta loading is typically negative, consistent with mandates that penalize high-beta exposure once other characteristics are controlled for. Despite this cross-sectional heterogeneity, the median coefficients are relatively stable over time, suggesting that the granular procedure recovers persistent style-driven tilts rather than absorbing broad market dynamics into residual "latent demand."

Stock-level price elasticity. Aggregating fund-level price coefficients across holder funds with ownership weights, I construct a stock-level mutual-fund price elasticity E_{jt}^{MF} each quarter. Figure 8 plots the dollar-weighted mean across stocks over time at two levels of aggregation, both computed from the same underlying mandate-level holdings: the mandate (fund) level, and a synthetic institution built by consolidating each family's mandates and re-estimating the linear specification. Although the estimation sample begins in 2003Q4, I plot the elasticity series from 2010, when CRSP mutual-fund coverage becomes dense enough for stable dollar-weighted aggregation (see Table 1).

Two features of Figure 8 support the aggregation mechanism of Section 2. First, the synthetic 13F series sits in the 0.05–0.10 range throughout the sample. This coincides with the institution-level mutual-fund elasticities reported in Figure 3 of Koijen and Yogo (2019), which estimates the same specification on actual 13F filings rather than on a synthetic construction from mandate-level data. The close correspondence is a validation check: the synthetic aggregation reproduces the institution-level estimates obtained directly from 13F data, so the comparison with the mandate-level series isolates the role of the observational unit rather than differences in sample, specification, or estimator. Second, the mandate-level series is two to four times larger than the synthetic 13F series and trends gradually downward, from roughly 0.4 at the start to about 0.2 by 2024. The trend is consistent with the rising passive share of mutual-fund assets and with reduced-form evidence on flow-induced price pressure (Wurgler and Zhuravskaya, 2002; Lou, 2012; Chang et al., 2015).

5.3 Latent demand at the stock level

A natural diagnostic of the granular estimation is to examine the residual portfolio tilts it produces. For each fund f, stock j, and quarter t, the residual $\hat{e}_{fjt} = w_{fjt} - \hat{\mu}_{fjt}$ measures the difference between the fund's observed within-universe weight and the weight implied by its estimated demand parameters. Scaling by fund equity assets

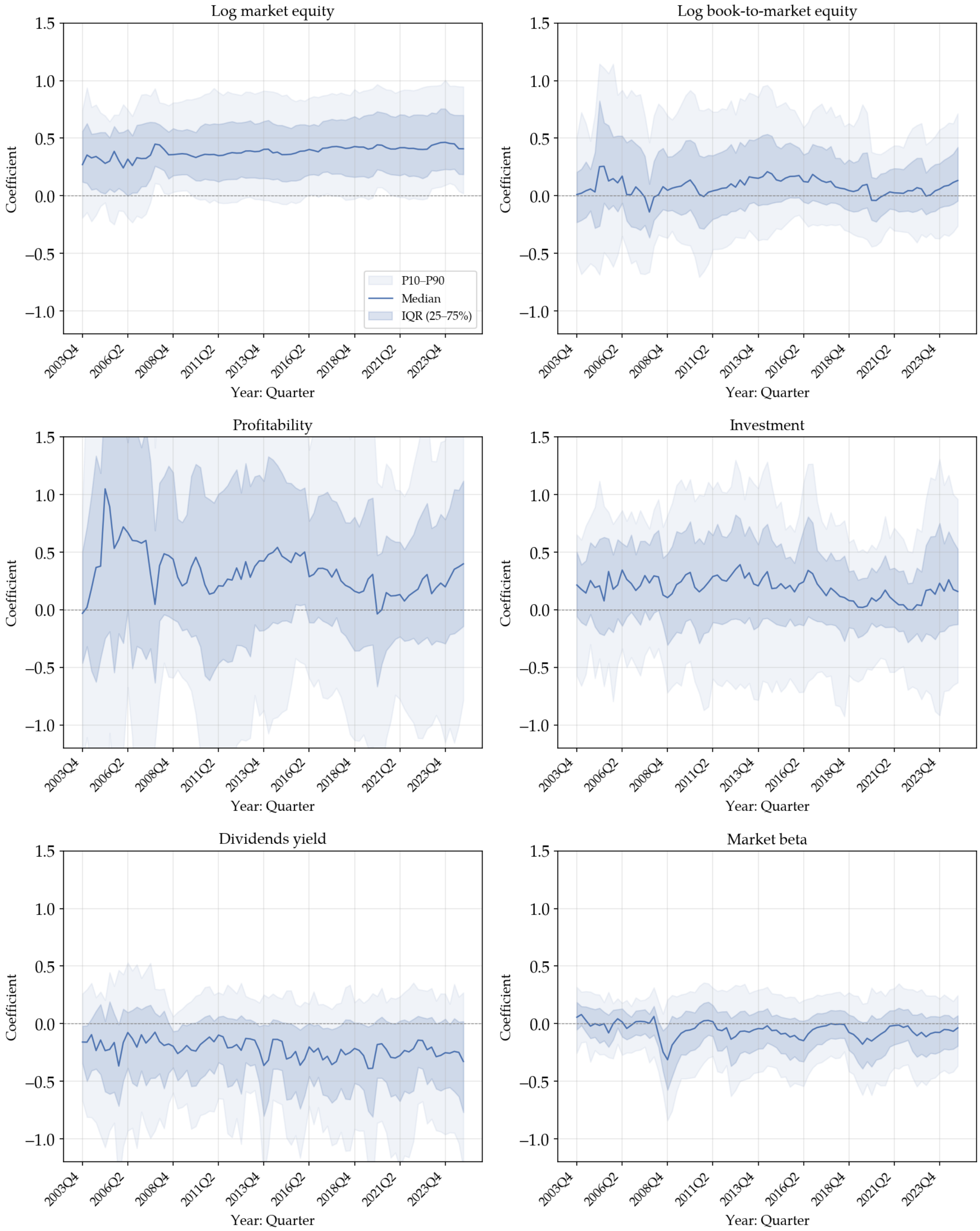

Figure 7: Coefficients on characteristics. Each panel plots the cross-sectional median (solid line), the interquartile range (dark shaded band, 25th–75th percentiles), and the 10th–90th percentile range (light shaded band) of mandate-level coefficient estimates each quarter. Coefficients are estimated quarter-by-quarter from mandate holdings using the baseline demand specification; the dashed horizontal line marks zero.

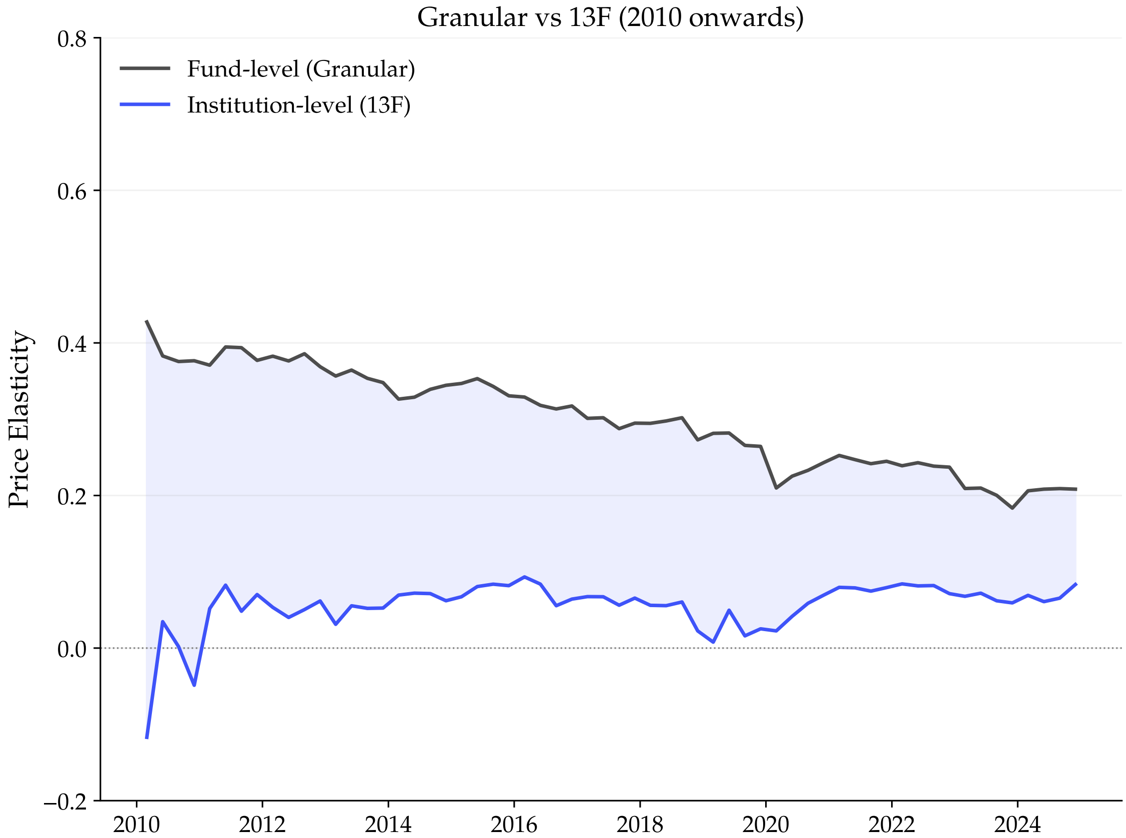

Figure 8: Dollar-weighted mean stock-level mutual-fund price elasticity E_{jt}^{MF} over time, computed from the same underlying mandate-level holdings at two levels of aggregation: the mandate (fund) level, and a synthetic institution formed by consolidating each family's mandates and re-estimating the linear specification. The mandate-level series is two to four times larger and trends downward with the rising passive share; the synthetic 13F series sits in the 0.05–0.10 range and coincides with the institution-level mutual-fund elasticities in Koijen and Yogo (2019), Figure 3, providing a validation check that the comparison isolates the role of the observational unit.

gives a dollar tilt $\hat{e}^{\$}_{fjt} = A_{ft}\phi_{ft}\,\hat{e}_{fjt}$.

A stock-level net residual can be formed by summing dollar tilts across funds, $L^{MF}_{jt} = \sum_f \hat{e}^{\$}_{fjt}$, but this aggregate is mechanically attenuated because each fund's residuals approximately sum to zero within its universe. The economically informative content therefore lies in the *shape* of the distribution: how much funds disagree about a given stock, whether that disagreement is symmetric or concentrated in one tail, and which stocks attract the most polarized active tilts.

Cross-fund disagreement and consensus. Table 4 summarizes the cross-fund residual distribution for the thirty largest stocks by market capitalization in 2024Q4. For each stock I report the number of holder funds, the AUM-weighted mean and standard deviation of fund-level residual weights $\hat{e}_{fjt}$ across holder funds (in basis points of fund equity), the ratio of net-to-predicted excess demand, and a *consensus index* defined as

$$C_{jt} = \frac{\left|\sum_f \hat{e}^{\$}_{fjt}\right|}{\sum_f \left|\hat{e}^{\$}_{fjt}\right|}, \tag{45}$$

which ranges from zero (funds' positive and negative tilts cancel perfectly—pure disagreement) to one (all funds tilt in the same direction—full consensus). This ratio is invariant to the number of holder funds and to cross-stock differences in market capitalization, making it directly comparable across the panel.

The table reveals substantial heterogeneity along two economically distinct dimensions. The first is the *direction* of net excess demand, captured by the column NetExc/Pred. Stocks such as MSFT, JPM, and NFLX attract net positive residual demand well above what firm characteristics predict, while TSLA, WMT, and ORCL are systematically underweighted. The second, and more novel, dimension is the *degree of fund agreement* behind these net positions.

Three archetypal patterns emerge:

(i) **Consensus underweights** (C_{jt} high, net demand negative). TSLA ($C = 0.83$), WMT (0.83), TMUS (0.87), and ORCL (0.77) are stocks that the broad majority of active funds tilt away from. These names share low cross-fund dispersion, indicating that the underweight is not driven by a few large funds but reflects a near-unanimous active bet.

(ii) **Contested names** (C_{jt} low, dispersion high). META ($C = 0.17$, $\sigma = 126$ bps), GOOGL (0.05, 80 bps), and AAPL (0.32, $\sigma = 164$ bps) exhibit net residual demand close to zero but large cross-fund dispersion. Funds disagree sharply: some overweight these stocks well beyond what characteristics justify, while others underweight by similar magnitudes. A consolidated institution-level estimator would report that the model fits META well—the net residual is near zero—when in fact it fits well only because large positive and negative active

Table 4: Cross-fund latent-demand summary for the thirty largest stocks (2024Q4). For each stock I report the number of holder funds, the AUM-weighted mean and standard deviation of fund-level residual weights $\hat{e}_{fjt}$ (basis points of fund equity), the ratio of aggregate net excess demand to total model-predicted demand (NetExc/Pred), and the consensus index C_{jt} from equation (45). Residuals are winsorized at the 1st and 99th percentiles within each stock before computing summary statistics.

	Ticker	ME ($B)	#Funds	Mean (bps)	σ (bps)	NetExc/Pred	C
1	AAPL	3,785	1,298	−44.1	163.6	2.0%	0.32
2	NVDA	3,289	1,255	−25.5	109.6	−3.7%	0.15
3	MSFT	3,134	1,446	28.5	119.0	17.3%	0.50
4	AMZN	2,307	1,249	−4.5	92.4	8.2%	0.12
5	TSLA	1,296	977	−53.7	74.8	−12.2%	0.83
6	META	1,276	1,392	16.9	125.8	−0.1%	0.17
7	GOOGL	1,106	1,325	−0.9	80.2	−4.4%	0.05
8	AVGO	1,086	1,275	10.8	68.3	−1.8%	0.03
9	LLY	733	1,210	−11.4	72.4	−19.0%	0.37
10	WMT	726	1,072	−63.8	52.8	−22.6%	0.83
11	JPM	675	1,195	19.6	71.4	18.7%	0.54
12	BRK	602	868	30.0	59.6	−4.4%	0.42
13	V	547	1,171	1.3	53.6	2.8%	0.28
14	MA	480	1,118	2.6	50.2	11.0%	0.12
15	XOM	473	1,222	2.2	63.2	2.3%	0.15
16	ORCL	466	1,114	−31.3	32.8	−24.0%	0.77
17	UNH	466	1,359	17.7	63.2	5.0%	0.49
18	COST	407	1,013	−2.6	44.9	4.6%	0.01
19	PG	395	1,117	−13.7	55.5	1.3%	0.12
20	HD	386	1,150	4.4	41.7	−14.7%	0.14
21	NFLX	381	1,060	17.2	52.6	17.0%	0.49
22	JNJ	348	1,170	−6.9	57.8	12.9%	0.05
23	BAC	337	1,167	−5.3	43.6	4.1%	0.20
24	CRM	320	1,217	−0.4	47.6	1.7%	0.21
25	ABBV	314	1,202	−2.9	54.3	17.4%	0.02
26	KO	268	1,049	−10.6	37.8	−14.7%	0.45
27	CVX	260	1,098	−5.8	36.1	−13.0%	0.52
28	TMUS	256	999	−28.9	35.6	−30.8%	0.87
29	MRK	252	1,299	−4.2	39.7	−11.0%	0.17
30	CSCO	236	1,188	−0.2	42.4	23.1%	0.17

bets cancel across mandates. The scalar stock-level residual obscures the economically active disagreement underneath.

(iii) **Consensus overweights** (C_{jt} high, net demand positive). MSFT ($C = 0.50$), JPM (0.54), and NFLX (0.49) attract broad positive tilts with moderate dispersion, suggesting that fund managers collectively view these stocks as underpriced relative to the characteristic-based benchmark, with limited offsetting underweights.

That these patterns align with market narratives in 2024Q4—broad skepticism toward rate-sensitive or idiosyncratic-risk names (TSLA, WMT), active disagreement over AI-exposed mega-caps (META, GOOGL, AAPL), and consensus overweighting of perceived quality (MSFT, JPM)—provides reassurance that the granular residuals carry economically interpretable information rather than estimation noise.

6 Conclusion

This paper has argued that demand estimation in asset markets is shaped by two structural features of institutional portfolios that are absent in product markets. First, institutions delegate across heterogeneous mandates, and the econometrician typically observes only the consolidated position. When each mandate follows a logit-style share system, institutional demand is a log-sum-exp aggregation of micro demands—a cumulant generating function that is intrinsically convex, with curvature equal to the cross-mandate dispersion in price sensitivities. Linear demand specifications applied to the aggregate are therefore projections of a curved object, and the resulting coefficient estimates can be extreme, attenuated, or sign-reversed. Second, even at the mandate level, portfolios span large universes with a long tail of tiny weights. The effective sample size for demand parameters is governed by portfolio concentration ($n^{\text{eff}} = 1/\text{HHI}$), not the number of assets in the choice set, meaning that conventional inference systematically overstates precision.

Empirically, estimating demand at the granular fund level—with preconditioning that respects the information structure of long-tail portfolios—delivers three results. First, fund-level price elasticities are overwhelmingly well-behaved and consistent with reduced-form benchmarks, without imposing ad hoc monotonicity constraints. Second, aggregating these micro estimates into synthetic 13F institutions reproduces the patterns observed in the literature, confirming that extreme inelasticity reflects aggregation rather than an economic primitive. Third, the granular specification substantially reduces the residual "latent demand" component, allowing the demand system to account for a larger share of observed prices. Moreover, the granular residuals reveal economically structured disagreement across funds—contested names where large positive and negative active bets cancel, versus consensus over- or underweights—that a scalar institution-level residual would collapse to near zero.

These findings suggest that the apparent tension between structural demand estimates and reduced-form evidence from quasi-experiments is not a paradox about investor behavior. It is a consequence of applying product-market econometrics to a setting whose data differ fundamentally from product-market choice sets: the universe is vast, the weight distribution is extreme, and the observed unit is an aggregation of the true decision maker. Recognizing these features—and adapting the estimator accordingly—yields demand estimates that are economically interpretable and empirically stable. The remaining open question is whether richer theoretical structures—incorporating cross-asset substitution, dynamic considerations, or equilibrium feedback—alter the picture once these aggregation effects are removed. This paper provides the structurally sound baseline against which that question can be addressed.

More broadly, the issues studied here are specific to the geometry of portfolio data. Observed shares are nonnegative and add up to one, so quantity moments do not behave like unconstrained price or return moments; their statistical content is shaped by this geometry. Whenever the observational unit bundles heterogeneous primitive decisions, the map from portfolio data to structural parameters is generically nonlinear, and the aggregation mechanism studied here is one economically important instance. The estimates here are partial-equilibrium by design and provide a baseline on which richer market-clearing analyses can build.

A Full Derivatives of Consolidated Demand

This appendix derives the exact total derivatives of the consolidated institutional demand Λ_{ji} with respect to p_j, without imposing the granular-portfolio approximation $w_{jm} \ll 1$ used in the main text.

Setup and notation

Recall that $w_{jm} = e^{\delta_{jm}} / (1 + \sum_k e^{\delta_{km}})$ denotes the mandate-level logit share and $w_{0m} = 1/(1 + \sum_k e^{\delta_{km}})$ is the outside-option share. The institutional mixing weights are $\pi_{mi} = a_{mi} w_{0m} / \sum_\ell a_{\ell i} w_{0\ell}$. Define the tilted weights $\tilde{\pi}_{mi} = \pi_{mi} e^{\delta_{jm}} / \sum_k \pi_{ki} e^{\delta_{jk}}$.

First derivative

Using the decomposition $\Lambda_{ji} = \log(\sum_m a_{mi} w_{jm}) - \log(\sum_\ell a_{\ell i} w_{0\ell})$, the total derivative is

$$\frac{d\Lambda_{ji}}{dp_j} = \sum_{m \in \mathcal{M}_i} \tilde{\pi}_{mi}\, \alpha_m (1 - w_{jm}) + \sum_{m \in \mathcal{M}_i} \pi_{mi}\, \alpha_m\, w_{jm}. \tag{46}$$

This can be rewritten as

$$\frac{d\Lambda_{ji}}{dp_j} = \mathbb{E}_{\tilde{\pi}}[\alpha_m] + \sum_{m\in\mathcal{M}_i} (\pi_{mi} - \tilde{\pi}_{mi})\, \alpha_m\, w_{jm},$$

where the second term is the indirect effect from the endogenous response of the mixing weights.

Second derivative

The exact total second derivative is

$$\frac{d^2\Lambda_{ji}}{dp_j^2} = \mathrm{Var}_{\tilde{\pi}}\big(\alpha_m(1-w_{jm})\big) - \mathrm{Var}_{\pi}\big(\alpha_m w_{jm}\big) + \sum_{m\in\mathcal{M}_i} (\pi_{mi} - \tilde{\pi}_{mi})\, \alpha_m^2\, w_{jm}(1-w_{jm}). \quad (47)$$

Granular-portfolio limit

When $w_{jm} \ll 1$ for all m:

- $(1-w_{jm}) \to 1$, so $\alpha_m(1-w_{jm}) \to \alpha_m$;
- all terms carrying w_{jm} (indirect effects, variance corrections) vanish at rate $O(\max_m w_{jm}) = O(1/n^{\text{eff}})$ by Condition G(a).

The first derivative reduces to $\mathbb{E}_{\tilde{\pi}}[\alpha_m]$ and the second to $\mathrm{Var}_{\tilde{\pi}}(\alpha_m)$, as stated in the main text [equations (17)–(18)].

B Robustness to Data Contamination in the Tail

Financial holdings data are reported at discrete precision and subject to minimum reporting thresholds. In the CRSP mutual fund holdings data, the effective weight floor is approximately $\underline{w} = 10^{-4}$ (one basis point), and weights move in discrete increments of $\Delta = 10^{-4}$. Tail positions with true weight below $\underline{w}$ are censored, and positions near the floor are coarsely quantized. This appendix shows that the logit gradient structure provides a natural robustness property: data contamination in the tail has negligible impact on the NLS estimator.

Figure 9 illustrates the long-tail structure directly. Each panel plots observed portfolio weights against market-cap weights for four representative index funds in 2022Q4. The discrete flooring is visible as horizontal bands at the bottom of each panel: positions below the reporting threshold are censored to $\underline{w} \approx 10^{-4}$, creating a mass of observations at the floor. Figure 10 shows the same data as the ratio $w_{\text{hat}}/w_{\text{mcap}}$ against market-cap rank. For the S&P 500 and Russell 1000, the ratio is close to one across most of the cross-section, consistent with near-mechanical tracking. For the Russell 3000 and Total Market, the ratio rises sharply for small-cap names (high rank), reflecting both the flooring artifact and genuine overweighting of small positions relative to market-cap weights.

Proposition 3 (Robustness of the logit NLS to tail contamination)**.** *Suppose the observed weight satisfies $w_j^{\text{obs}} = w_j^{\text{true}} + c_j$, where c_j is a contamination term (from censoring, quantization, or other measurement error). Let $\hat{\alpha}$ denote the NLS estimator using observed weights and α_0 the true parameter. Then the first-order bias satisfies*

$$\hat{\alpha} - \alpha_0 \;\approx\; \mathcal{I}(\alpha_0)^{-1} \sum_{j \in \mathcal{U}} c_j \, \mu_j(\alpha_0) \, p_j^c. \tag{48}$$

(i) ***Floor censoring.*** *If K tail positions are censored at $\underline{w}$, so that $c_j = \underline{w} - \mu_j \leq \underline{w}$ for censored stocks and $c_j = 0$ otherwise, then*

$$|\text{bias}_{\text{floor}}| \;\leq\; n^{\text{eff}} \, K \, \underline{w}^2 \, \frac{\overline{|p^c|}_{\text{cens}}}{\bar{\sigma}_p^2}. \tag{49}$$

With $n^{\text{eff}} = 50$, $K = 500$, $\underline{w} = 10^{-4}$, and $\overline{|p^c|}_{\text{cens}} / \bar{\sigma}_p^2 \sim O(1)$, this is $O(10^{-4})$ — negligible relative to typical estimates of $\alpha \sim 0.3$–1.0.

(ii) ***Quantization.*** *If c_j is approximately uniform on $[-\Delta/2, \Delta/2]$ and independent across stocks, the expected bias is zero and the variance contribution to $\hat{\alpha}$ from quantization is*

$$\text{Var}_{\text{quant}}(\hat{\alpha}) \;=\; \mathcal{I}(\alpha_0)^{-2} \, \frac{\Delta^2}{12} \sum_{j \in \mathcal{U}} \mu_j^2 \, (p_j^c)^2 \;=\; \frac{\Delta^2}{12} \, \mathcal{I}(\alpha_0)^{-1}. \tag{50}$$

This adds variance of order $\Delta^2 \cdot n^{\text{eff}}$ to the sampling variance of $\hat{\alpha}$, which is negligible for $\Delta = 10^{-4}$ and $n^{\text{eff}} \sim 50$.

Proof. Equation (48) follows from a first-order Taylor expansion of the NLS first-order condition around α_0, using $\partial \mu_j / \partial \alpha = \mu_j \, p_j^c$.

Part (i): For censored stocks, $\mu_j \leq \underline{w}$, so each term in the numerator of (48) satisfies $|c_j \, \mu_j \, p_j^c| \leq \underline{w}^2 \, |p_j^c|$. Summing over K censored stocks and dividing by $\mathcal{I}(\alpha_0) = O(1/n^{\text{eff}})$ yields the stated bound.

Part (ii): Independence of c_j across stocks and $E[c_j] = 0$ imply zero expected bias. The variance follows from $E[c_j^2] = \Delta^2/12$ and the sandwich formula. □

The key mechanism is the μ_j factor in the logit score. Each tail observation's influence on $\hat{\alpha}$ is weighted by its fitted share μ_j, which provides one full power of natural downweighting beyond what a linear model would deliver. In a linear regression of w_j

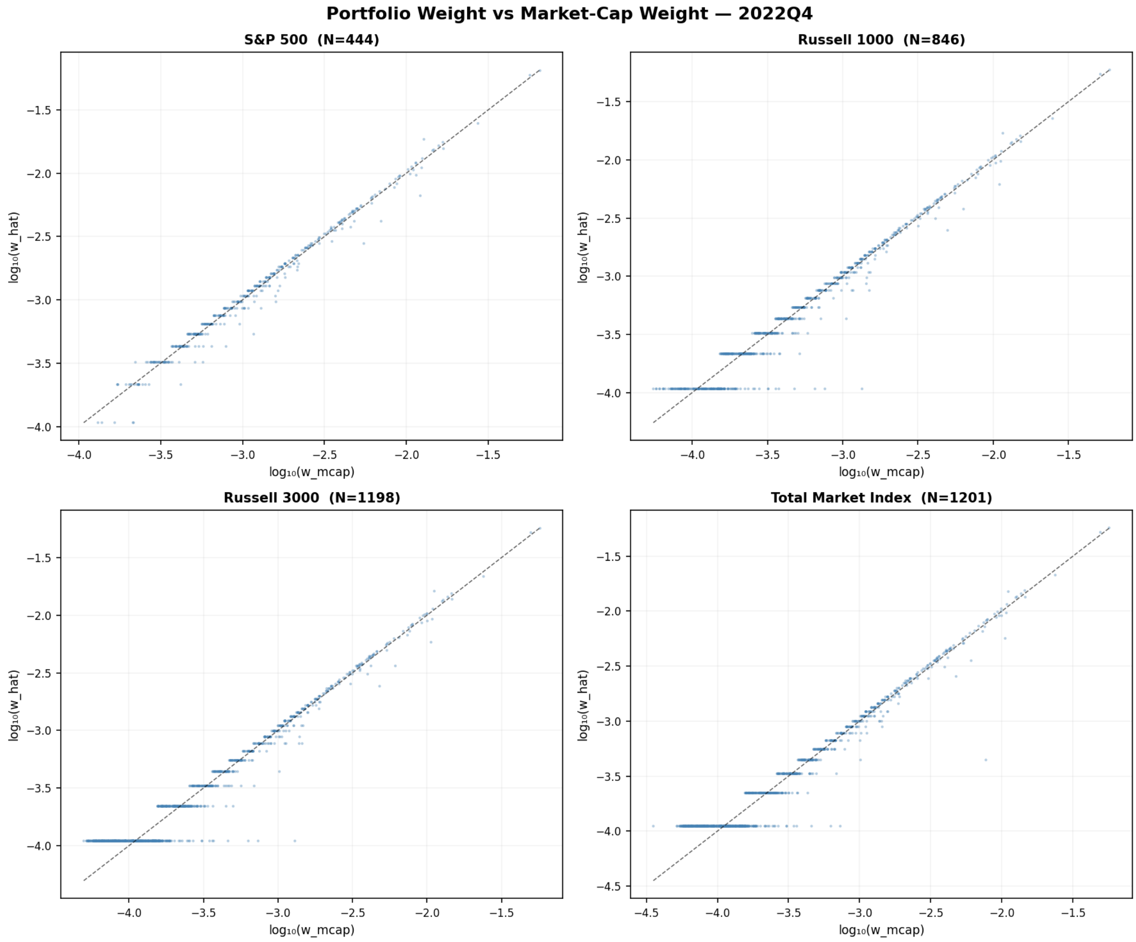

Figure 9: **Portfolio weight vs. market-cap weight for index funds (2022Q4).** Each panel plots $\log_{10}(w_{\text{hat}})$ against $\log_{10}(w_{\text{mcap}})$ for the indicated index fund. The dashed line is the 45-degree line. Horizontal banding at the bottom reflects the discrete reporting floor ($\underline{w} \approx 10^{-4}$). The number of securities in each fund's effective universe is shown in parentheses.

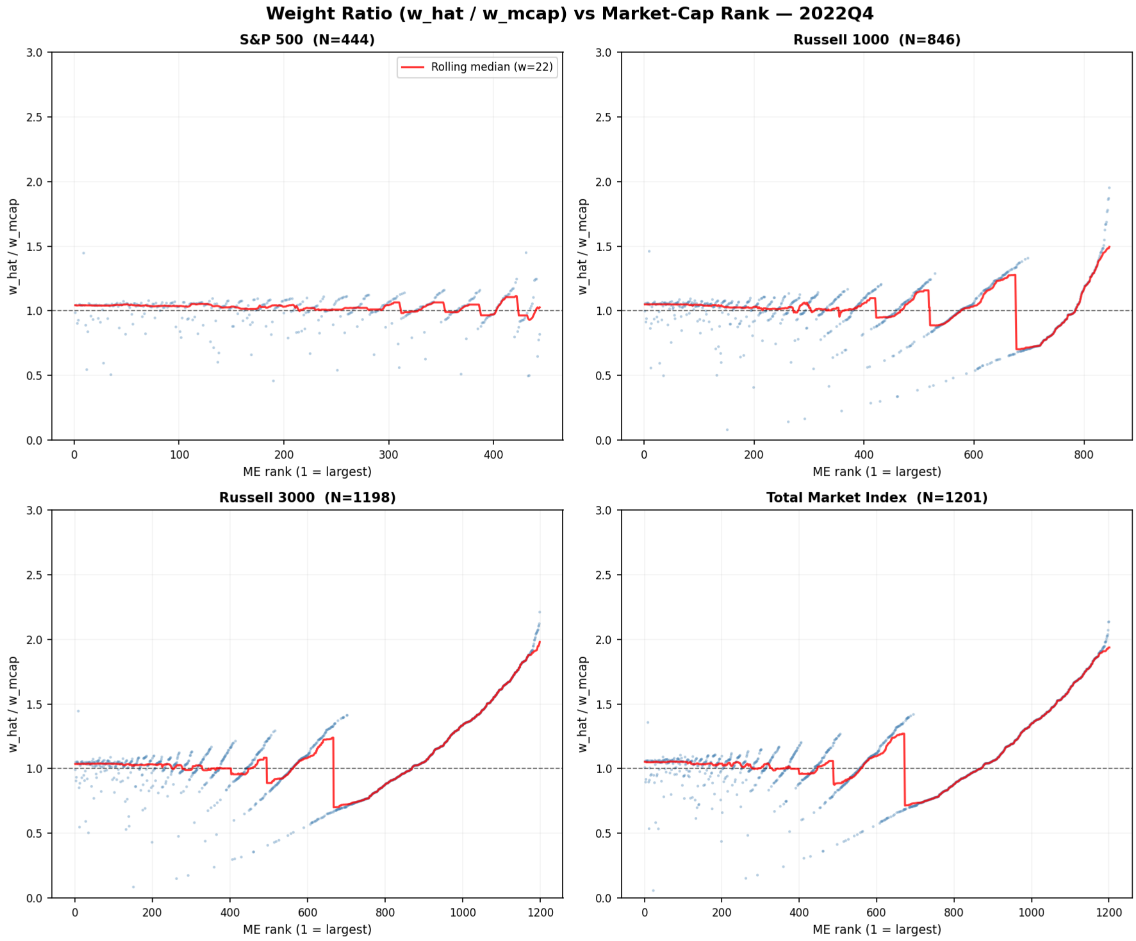

Figure 10: **Weight ratio vs. market-cap rank for index funds (2022Q4).** Each panel plots $w_{\text{hat}}/w_{\text{mcap}}$ against market-cap rank (1 = largest) for the indicated index fund. The red line is a rolling median (window = 22). For narrow-universe funds (S&P 500, Russell 1000), the ratio is close to one across most ranks. For broad-universe funds (Russell 3000, Total Market), the ratio rises sharply for small-cap names, reflecting both the reporting floor and genuine small-cap overweighting. The discrete quantization of reported weights is visible as horizontal striations.

on p_j, contamination c_j in tail observations would contribute $c_j\, p_j^c$ to the score—with no μ_j dampening—and the cumulative effect of K censored observations could be of first order. The logit link's built-in dampening is therefore not merely a convenience; it is the structural reason that logit demand estimation remains well-behaved in the presence of the data limitations that are endemic to financial holdings databases.

References

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

Bretscher, Lorenzo, Lukas Schmid, Ishita Sen, and Varun Sharma, "Institutional Corporate Bond Demand," *Review of Financial Studies*, 2026. Forthcoming.

Chang, Yen-Cheng, Harrison Hong, and Inessa Liskovich, "Regression Discontinuity and the Price Effects of Stock Market Indexing," *Review of Financial Studies*, 2015, *28* (1), 212–246.

Chaudhary, Manav, Zhiyu Fu, and Jian Li, "Demand Elasticities in Corporate Bonds: Substitution, Aggregation, and Price Impact," 2025. Working Paper.

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

Gabaix, Xavier, Ralph S. J. Koijen, Robert J. Richmond, and Motohiro Yogo, "Asset Embeddings," 2025. Working Paper.

Haddad, Valentin, 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* (1), 197–232.

_, Zhiguo He, Paul Huebner, Peter Kondor, and Erik Loualiche, "Causal Inference for Asset Pricing," 2025. Working Paper.

He, Zhiguo, Peter Kondor, and Xing Li, "Demand Elasticity in Dynamic Asset Pricing," 2026. Working Paper.

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," 2026. NBER Working Paper 27342, revised 2026.

___, **François Koulischer, Benoît Nguyen, and Motohiro Yogo**, "Inspecting the Mechanism of Quantitative Easing in the Euro Area," *Journal of Financial Economics*, 2021, *140* (1), 26–54.

___, **Robert J. Richmond, and Motohiro Yogo**, "Which Investors Matter for Equity Valuations and Expected Returns?," *Review of Economic Studies*, 2024, *91* (4), 2377–2408.

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

Nevo, Aviv, "Measuring Market Power in the Ready-to-Eat Cereal Industry," *Econometrica*, 2001, *69* (2), 307–342.

Newey, Whitney K. and Daniel McFadden, "Large Sample Estimation and Hypothesis Testing," in Robert F. Engle and Daniel McFadden, eds., *Handbook of Econometrics*, Vol. 4, Elsevier, 1994, pp. 2111–2245.

van Binsbergen, Jules H., Joel M. David, and Christian C. Opp, "How (Not) to Identify Demand Elasticities in Dynamic Asset Markets," 2025. Working Paper.

Wurgler, Jeffrey and Ekaterina Zhuravskaya, "Does Arbitrage Flatten Demand Curves for Stocks?," *Journal of Business*, 2002, *75* (4), 583–608.