

Portfolio Granularity and Demand Estimation

Ryan Cheng*

July 30, 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 logit demand link weights each position's contribution to identification by its squared portfolio share, so the effective information is governed by portfolio concentration ($n^{\text{eff}} \approx 30\text{--}60$), 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, Péter 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 [Kojen and Yogo \(2019\)](#) and extended by [Kojen 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.

These findings have prompted a reassessment of what holdings-based estimation recovers. One line of work asks whether the underlying economic model is too simple for financial markets, and looks for the answer in richer substitution, dynamics, or equilibrium feedback ([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 *portfolio granularity* (Section 1), a feature that distinguishes asset demand from standard product demand. In a logit-style demand system, each position’s contribution to the noise in the estimating moment scales with 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: precision is governed by effective breadth $n^{\text{eff}} = 1/\text{HHI}$, not by the holdings count N . 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 shift the price coefficient toward the near-unit, mechanical-tracking benchmark—or past 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 $n^{\text{eff}} \approx 30\text{--}60$ 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. 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.

Having identified the mechanical roots of these distortions, Section 3 makes two changes to the empirical design. First, it shifts the observation unit from consolidated 13F institutions to individual mutual-fund portfolios, bringing the data closer to the mandate at which allocation decisions are made. Second, it replaces the inverse-share ratio moment with a level-share moment. The ratio moment mechanically restores equal weight to the full holdings count by amplifying tiny positions; the level moment instead accepts the genuine n^{eff} information ceiling and is robust to rounding and reporting noise in the long tail. The remaining specification and implementation details are reported in the appendix.

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 (Kojen 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.

A final remark concerns 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.

Demand-system asset pricing. This paper contributes directly to the demand-system approach to asset pricing introduced by [Kojien and Yogo \(2019\)](#) and developed by [Kojien et al. \(2024\)](#). The approach uses investor-level holdings to estimate demand curves and quantify how changes in investor demand affect asset prices, expected returns, and the allocation of wealth. I study two choices that are logically prior to these economic applications: the observational unit at which portfolio demand is measured and the weighting of positions in the estimating moments. The first determines whether the observed portfolio represents a primitive allocation decision or a nonlinear aggregation of heterogeneous mandates. The second determines whether identification is drawn from the economically meaningful core of the portfolio or from its long tail of small positions. These choices become first-order when demand systems are applied to granular, delegated portfolios.¹

Investor heterogeneity and the composition of demand. A broader literature studies how heterogeneous investors generate trading and shape asset prices. Differences in investors' interpretations of common information can produce disagreement, heterogeneous positions, and trading volume ([Kandel and Pearson, 1995](#); [Hong and Stein, 2007](#)). Direct evidence documents large and persistent heterogeneity in expected returns and shows that these beliefs are reflected in the portfolios of both households and professional asset managers ([Giglio et al., 2021](#); [Dahlquist and Ibert, 2024](#)). The composition of the investor base consequently matters for nonfundamental volatility, price impact, and aggregate demand elasticity ([Greenwood and Thesmar, 2011](#); [Gabaix and Kojien, 2021](#); [Kojien et al., 2024](#); [Haddad et al., 2025](#)). [van der Beck et al. \(2025\)](#) further show that investor disagreement, portfolio turnover, and price impact are jointly restricted by observable quantity and price moments.

¹Applications of the demand-system approach include U.S. equities ([Kojien et al., 2024](#)), euro-area government bonds ([Kojien et al., 2021](#)), corporate bonds ([Bretscher et al., 2026](#)), and international portfolios spanning equities, bonds, and currencies ([Kojien and Yogo, 2026](#)).

My contribution concerns how such heterogeneous demands are represented in holdings data. Whereas this literature asks how heterogeneous investors jointly determine trading and prices, I ask what an econometrician recovers when heterogeneous allocation decisions are consolidated before estimation. In a logit share system, heterogeneity does not generally average out: the variance, covariances, and higher cumulants of mandate-level tastes enter the consolidated demand function. The composition of demand therefore affects not only market outcomes, but also the structural object recovered from observed holdings.

Reduced-form evidence on price impact. Finally, this paper relates to the reduced-form literature documenting downward-sloping demand for stocks using plausibly exogenous ownership shifts. This work studies price responses to index inclusions and reconstitutions, flow-induced trading, and limits to arbitrage (e.g., Shleifer, 1986; Wurgler and Zhuravskaya, 2002; Lou, 2012; Chang et al., 2015; Greenwood and Sammon, 2025). These studies do not estimate a full cross-asset demand system, but they provide disciplined benchmarks for the magnitude of local price impact. The granular estimates in this paper move toward these reduced-form magnitudes, while consolidating the same underlying holdings into institution-level portfolios reproduces the much smaller elasticities obtained from conventional structural implementations. The comparison therefore helps reconcile the reduced-form and structural evidence without changing either the underlying holdings or the economic shocks being studied.

1 Granular Portfolios and the Limits of Count-Based Identification

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). A dozen or so positions above 1% of the portfolio, together with roughly 150–160 positions between 0.1% and 1%, account for 76–86% of total portfolio mass, even though the fund holds 449 to over 1,100 names. This is not specific to index funds. Table 4 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 not merely descriptive, but tied to identification.

The effective number of positions. Throughout this section, w_j denotes the within-universe portfolio share of stock j , normalized so that $\sum_j w_j = 1$.² For any normalized share vector,

²If 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. Multiplying all within-universe weights by the same constant rescales the level moment but does not create cross-sectional information.

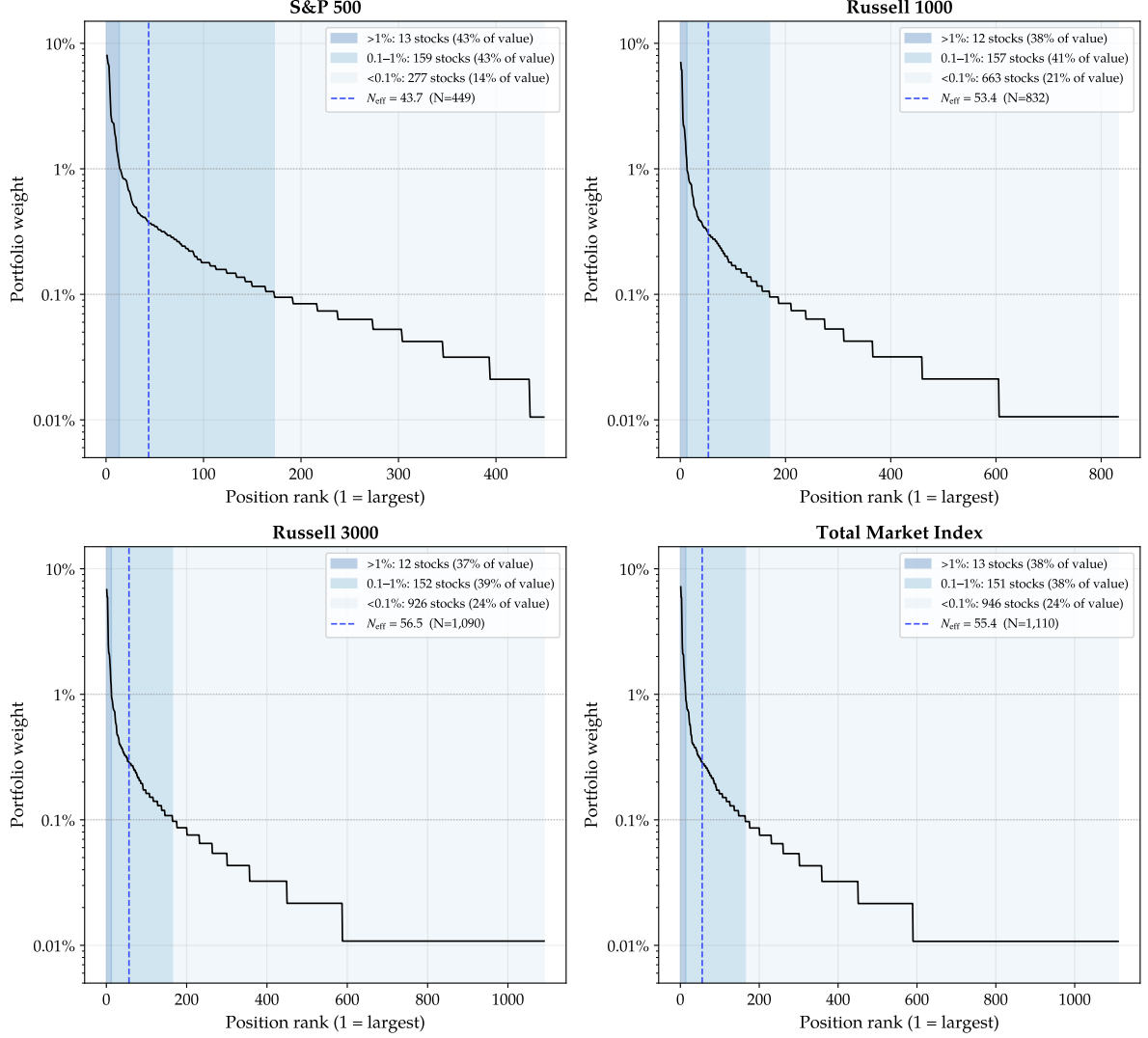


Figure 1: Portfolio weight profiles exhibit a long tail of tiny, discretized positions. Each panel plots cross-sectional portfolio weights for a representative index fund in 2024Q4, sorted by position size (rank 1 = largest) on a $\log_{10}$ scale. Shaded bands partition holdings into three weight tiers—above 1% of the portfolio, between 0.1% and 1%, and below 0.1%—with the legend reporting the number of stocks and the share of portfolio value in each tier. A dozen or so positions above 1%, together with roughly 150–160 positions in the middle tier, account for most of portfolio value, while hundreds to nearly a thousand sub-0.1% positions make up the bulk of the raw holdings count but little of its value. The dashed vertical line marks the effective number of positions, $n^{\text{eff}} \equiv 1/\sum_j w_j^2$, reported alongside the raw count N ; it falls well inside the middle tier, an order of magnitude below N , reflecting the negligible contribution of the long tail.

define

$$\text{HHI} \equiv \sum_{j=1}^N w_j^2, \quad n^{\text{eff}} \equiv \frac{1}{\text{HHI}}. \quad (1)$$

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 $\text{HHI} = 1/n$ and $n^{\text{eff}} = n$. 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 43.7 (S&P 500) to 56.5 (Russell 3000), even though the raw holdings count N ranges from 449 to over 1,100. The dashed vertical lines mark this effective count directly in each panel; it falls well inside the tier of moderate-weight positions (0.1–1% of the portfolio) and well short of the long tail of sub-0.1% holdings that make up most of the raw count but contribute little to the portfolio’s statistical content.

The rest of the paper uses n^{eff} in this disciplined sense: not merely as a metaphor for concentration, but as the benchmark size of the fund-quarter’s identifying cross-section. A portfolio with one thousand reported names but fifty economically meaningful positions is statistically much closer to a fifty-observation cross-section than to a thousand-observation one. The next two paragraphs make that claim precise with one elementary derivation; Appendix A supplies the general version.

The information contributed by one position. A portfolio is one fixed amount of capital divided across assets. Adding another reported holding does not append an independent observation in the usual regression sense; it reallocates the same unit mass across another share.

To see what this means for estimation, fix one fund-quarter and write demand in the form used throughout the paper:

$$\log q_j = \alpha x_j + \varepsilon_j, \quad (2)$$

where x_j is the characteristic whose coefficient is of interest and ε_j is latent demand, independent across positions with variance σ^2 . Let w_j denote the observed portfolio share and let $\mu_j(\alpha)$ denote the share implied by the model at a trial value of α . The paper estimates α by matching the portfolio’s observed and model-implied exposure to x_j :

$$\sum_j x_j \{w_j - \mu_j(\alpha)\} = 0. \quad (3)$$

Let $\mu_{0j} \equiv \mu_j(\alpha_0)$ denote the model-implied share at the true parameter. The characteristic

is weighted-centered and weighted-standardized with respect to this local share profile,³

$$\sum_j \mu_{0j} x_j = 0, \quad \sum_j \mu_{0j} x_j^2 = 1. \quad (4)$$

Because portfolio shares add to one, a position moves only relative to the rest of the portfolio. Differentiating the model-implied share at the true parameter gives

$$\left. \frac{\partial \mu_j(\alpha)}{\partial \alpha} \right|_{\alpha=\alpha_0} = \mu_{0j} \left(x_j - \sum_k \mu_{0k} x_k \right) = \mu_{0j} x_j, \quad (5)$$

where the final equality uses weighted centering. A latent-demand shock similarly changes the observed share, to first order, by

$$w_j - \mu_{0j} \approx \mu_{0j} \left(\varepsilon_j - \sum_k \mu_{0k} \varepsilon_k \right). \quad (6)$$

Thus both the response to α and the response to latent-demand noise are scaled by the position's own portfolio share. A one-basis-point holding carries very little signal, but it also contributes very little structural noise. Counting it as one additional observation therefore exaggerates the identifying content of the portfolio.

The effective information of the whole portfolio. Holding the local share profile and characteristics fixed, linearizing the level moment (3) around α_0 gives

$$\hat{\alpha} - \alpha_0 \approx \frac{\sum_j \mu_{0j} x_j \varepsilon_j}{\sum_j \mu_{0j} x_j^2}, \quad (7)$$

where weighted centering removes the portfolio-average shock term $\sum_k \mu_{0k} \varepsilon_k$ from the numerator exactly, not approximately (Appendix A). Independence of the position-level shocks then implies the first-order conditional variance⁴ is

$$\text{Var}_{\text{lin}}(\hat{\alpha}) = \sigma^2 \frac{\sum_j \mu_{0j}^2 x_j^2}{\left(\sum_j \mu_{0j} x_j^2 \right)^2}. \quad (8)$$

The two sums respond differently to the share distribution. The denominator carries each share once: it is the weighted mean of x_j^2 , equal to one under the standardization. Adding names does not mechanically enlarge it. The random contribution of position j to the estimating moment is $\mu_{0j} x_j \varepsilon_j$. Its contribution to the variance therefore contains the squared share, μ_{0j}^2 . A position that is ten times smaller contributes one hundred times less to the variance of the es-

³In the application, x_j is the price characteristic after partialling out the remaining characteristics; Appendix A gives the multi-characteristic and instrumented version. The formal local calculation below is evaluated at the model shares μ_0 ; empirically, observed shares provide the natural plug-in for this share profile. Because the standardization rescales the characteristic, coefficients are expressed per weighted standard deviation of x_j ; the variance of the coefficient in original units rescales accordingly.

⁴Because the estimator is nonlinear, this is a first-order approximation to the finite-sample variance of $\hat{\alpha}$, not an exact expression; Appendix A gives the exact linearization and its regularity conditions.

timating moment. Tiny positions consequently add little to the portfolio-level signal-to-noise calculation.

For a generic characteristic direction whose squared values are not systematically related to position size, $\sum_j \mu_{0j}^2 x_j^2 \approx \sum_j \mu_{0j}^2$. Using observed shares as the empirical plug-in for the local share profile, equation (8) then reduces to

$$\text{Var}_{\text{lin}}(\hat{\alpha}) \approx \sigma^2 \sum_j w_j^2 = \sigma^2 \text{HHI} = \frac{\sigma^2}{n^{\text{eff}}}. \quad (9)$$

The textbook regression is the equal-weight special case. If $\mu_{0j} = 1/N$ for every position, weighted standardization implies

$$\sum_j \mu_{0j}^2 x_j^2 = \frac{1}{N},$$

so the first-order variance is σ^2/N exactly. Equal weights are the case in which the holdings count and the number of economically meaningful positions coincide. Away from that case the two diverge, and that divergence is the whole content of the result: because the standardization holds the denominator of (8) at one, the first-order variance is exactly the squared-share-weighted dispersion in its numerator, $\sum_j \mu_{0j}^2 x_j^2$. The following proposition records this as the section's identification benchmark.

Proposition 1 (Identification is governed by concentration, not holdings count). *Under the conditions above, the first-order conditional variance of the level-moment estimator is*

$$\text{Var}_{\text{lin}}(\hat{\alpha}) = \sigma^2 \sum_j \mu_{0j}^2 x_j^2. \quad (10)$$

If characteristic extremeness is not systematically related to position size, then

$$\text{Var}_{\text{lin}}(\hat{\alpha}) \approx \frac{\sigma^2}{n^{\text{eff}}}. \quad (11)$$

Thus the generic identifying content of a fund-quarter is governed by its concentration structure. The nominal holdings count N has no independent role: it matters only through the concentration structure.

Remark 1 (Realized precision is direction-specific). Proposition 1 is a generic benchmark, not a claim that every characteristic is equally well identified. Concentration in characteristic-extreme positions raises the squared-share variance factor and lowers effective sample size; concentration in characteristic-typical positions does the reverse. Appendix A isolates this alignment factor, and the empirical sections report realized identification diagnostics for the residualized price direction.

The mechanism, against the OLS benchmark. It is worth setting this against the ordinary regression the reader has in mind, because the source of precision is reversed. In OLS with a standardized regressor, precision comes from the denominator: identifying information is the *unweighted* sum $\sum_j (x_j - \bar{x})^2$, which grows with the number of observations, so the sam-

pling variance $\sigma^2 / \sum_j (x_j - \bar{x})^2$ falls to zero as the count rises. The analogous object here is *share-weighted*, $\sum_j \mu_{0j} x_j^2$, and the standardization pins it at one; a weighted average is unmoved by appending negligible-weight terms, so lengthening the holdings list cannot inflate it. The count channel that drives OLS precision is therefore closed. What remains is the numerator: rather than a growing denominator beaten down toward zero, the variance is a fixed unit denominator divided into the squared-share mass $\sum_j \mu_{0j}^2 x_j^2 \approx 1/n^{\text{eff}}$. Precision is bought by de-concentration, not by counting, and a portfolio does not de-concentrate by adding names to its tail.

The distinction is immediate in the data. Moving from the iShares Russell 1000 to the iShares Russell 3000 ETF raises the reported holdings count from 832 to 1,090, about 31%, while effective breadth rises only from 53.4 to 56.5, under 6% (Figure 1). The count moves substantially; the information barely does. A small effective core inside a large nominal support is the structural feature that separates institutional equity portfolios from the product-market settings where structural demand estimation originated, and it is worth naming.

Definition 1 (Granular portfolio). I call a portfolio *granular* when it lies in the empirical regime

$$1 \ll n^{\text{eff}} \ll N.$$

Portfolio mass is spread across a nontrivial effective core, while the reported holdings count is much larger than the number of positions carrying economically meaningful weight.

Nothing in the definition is specific to equities: the same regime arises in any share system with large nominal support and a concentrated core, and canonical product-market applications simply sit outside it, with n^{eff} close to N because most products carry meaningful share.

This reframes where identifying information can come from. Lengthening a single portfolio's holdings list changes N but leaves its effective core, and hence its generic identifying content, essentially unchanged; one snapshot therefore delivers a fixed and finite n^{eff} however many names it reports.

From information to estimator design. The result of this section is an information constraint, not yet a choice of estimator. One can mechanically undo the share damping by dividing the share residual by its fitted weight and thereby treat the N listed holdings as equally weighted observations. Whether that is desirable depends on the quality of the long tail and is therefore an estimation-design question rather than part of the concentration result itself. I defer the comparison between ratio and level moments to Section 3.2, where it forms the second of the paper's two empirical design changes.

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 ex-

pand 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 each mandate’s portfolio is granular (Section 1), with 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 ~ 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. 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:

$$\log \frac{w_{jm}}{w_{0m}} = \delta_{jm}, \quad \delta_{jm} = \alpha_m p_j + x_j' \beta_m + \xi_j. \quad (12)$$

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:

$$W_{ji} = \sum_{m \in \mathcal{M}_i} a_{mi} w_{jm}, \quad W_{0i} = \sum_{m \in \mathcal{M}_i} a_{mi} w_{0m}. \quad (13)$$

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

$$\frac{W_{ji}}{W_{0i}} = \sum_{m \in \mathcal{M}_i} \pi_{mi} e^{\delta_{jm}}, \quad \pi_{mi} \equiv \frac{a_{mi} w_{0m}}{\sum_{\ell \in \mathcal{M}_i} a_{\ell i} w_{0\ell}}. \quad (14)$$

Taking logs yields the central object:

$$\Lambda_{ji} \equiv \log \frac{W_{ji}}{W_{0i}} = \log \sum_{m \in \mathcal{M}_i} \pi_{mi} e^{\delta_{jm}}. \quad (15)$$

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:

$$\Lambda_{ji} = \log \sum_{m \in \mathcal{M}_i} \pi_{mi} e^{\delta_{jm}} = \log \mathbb{E}_\pi [e^{\Delta_j}]. \quad (16)$$

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

$$\Lambda_{ji} = \kappa_1(\Delta_j) + \frac{1}{2}\kappa_2(\Delta_j) + \frac{1}{6}\kappa_3(\Delta_j) + \frac{1}{24}\kappa_4(\Delta_j) + \dots \quad (17)$$

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 (16) is exact for any portfolio. To draw out its implications for estimation, I derive the slope and curvature of consolidated demand and show that, in the granular regime, the leading behavior takes a particularly tractable form.

The mixing weights π_{mi} defined in (14) 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).

Direct effect. Holding the mixing weights π_{mi} fixed and differentiating the log-sum-exp in (15) yields

$$\left. \frac{\partial \Lambda_{ji}}{\partial p_j} \right|_{\pi \text{ fixed}} = \sum_{m \in \mathcal{M}_i} \tilde{\pi}_{mi} \alpha_m \equiv \mathbb{E}_{\tilde{\pi}} [\alpha_m], \quad (18)$$

where

$$\tilde{\pi}_{mi} \equiv \frac{\pi_{mi} e^{\delta_{jm}}}{\sum_{\ell \in \mathcal{M}_i} \pi_{\ell i} e^{\delta_{j\ell}}} \quad (19)$$

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:

$$\frac{\partial \pi_{mi}}{\partial p_j} = \pi_{mi} \left(-\alpha_m w_{jm} + \sum_{\ell \in \mathcal{M}_i} \pi_{\ell i} \alpha_{\ell} w_{j\ell} \right). \quad (20)$$

The total derivative is

$$\frac{d\Lambda_{ji}}{dp_j} = \underbrace{\sum_{m \in \mathcal{M}_i} \tilde{\pi}_{mi} \alpha_m (1 - w_{jm})}_{\text{direct}} + \underbrace{\sum_{m \in \mathcal{M}_i} \pi_{mi} \alpha_m w_{jm}}_{\text{indirect}}, \quad (21)$$

where the indirect effect can be isolated as

$$\left. \frac{d\Lambda_{ji}}{dp_j} - \frac{\partial \Lambda_{ji}}{\partial p_j} \right|_{\pi \text{ fixed}} = \sum_{m \in \mathcal{M}_i} (\pi_{mi} - \tilde{\pi}_{mi}) \alpha_m w_{jm}. \quad (22)$$

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 $w_{jm} = O(1/n^{\text{eff}})$.

The direct effect dominates. Every term in the indirect effect (22) carries a factor of w_{jm} . In the granular regime, this entire channel is of relative order $1/n^{\text{eff}}$.⁵ Conversely, the factors $(1 - w_{jm})$ in the direct channel differ from unity only at relative order $1/n^{\text{eff}}$, and the total derivative reduces to the clean tilted average:

$$\frac{d\Lambda_{ji}}{dp_j} \approx \mathbb{E}_{\tilde{\pi}}[\alpha_m]. \quad (23)$$

The same logic applied to the second derivative yields⁶

$$\frac{d^2 \Lambda_{ji}}{dp_j^2} \approx \text{Var}_{\tilde{\pi}}(\alpha_m) \geq 0. \quad (24)$$

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

⁵Formally, in the granular regime, $\max_m w_{jm} = O(1/n^{\text{eff}})$, 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 B.

⁶The exact second derivative is $d^2 \Lambda_{ji} / dp_j^2 = \text{Var}_{\tilde{\pi}}(\alpha_m (1 - w_{jm})) - \text{Var}_{\tilde{\pi}}(\alpha_m w_{jm}) + \sum_m (\pi_{mi} - \tilde{\pi}_{mi}) \alpha_m^2 w_{jm} (1 - w_{jm})$. In the granular regime, the last two terms are $O(1/n^{\text{eff}})$ and the first reduces to $\text{Var}_{\tilde{\pi}}(\alpha_m)$. See Appendix B.

- (i) **Slope.** $d\Lambda_{ji}/dp_j = \mathbb{E}_{\tilde{\pi}}[\alpha_m]$, the exponentially tilted mean of mandate-level price sensitivities.
- (ii) **Curvature.** $d^2\Lambda_{ji}/dp_j^2 = \text{Var}_{\tilde{\pi}}(\alpha_m) \geq 0$. 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 ($\text{Var} > 0$), and consolidated 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 $\text{Var}_{\tilde{\pi}}(\alpha_m) \geq 0$. 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.

In the granular regime, 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:

$$\Lambda_j \approx \tilde{\xi}_j + c + \bar{\alpha} p_j + x_j' \bar{\beta} + \Gamma_j, \quad (25)$$

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

$$\Gamma_j \equiv \frac{1}{2} \text{Var}(\alpha) p_j^2 + p_j x_j' \text{Cov}(\alpha, \beta) + \frac{1}{2} x_j' \text{Var}(\beta) x_j. \quad (26)$$

All variance/covariance operators are within institution i , taken over the mandate distribution. The key point is that Γ_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*

$$\Lambda_j = \hat{\alpha} p_j + x_j' \hat{\beta} + \text{residual}. \quad (27)$$

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

$$\hat{\alpha} - \bar{\alpha} = \frac{\text{Cov}(\tilde{p}_j, \Gamma_j)}{\text{Var}(\tilde{p}_j)}. \quad (28)$$

Substituting the second-order form of Γ_j yields

$$\hat{\alpha} - \bar{\alpha} \approx \frac{1}{2} \text{Var}(\alpha) \frac{\text{Cov}(\tilde{p}_j, p_j^2)}{\text{Var}(\tilde{p}_j)} + \text{Cov}(\alpha, \beta)' \frac{\text{Cov}(\tilde{p}_j, p_j x_j)}{\text{Var}(\tilde{p}_j)} + \frac{1}{2} \frac{\text{Cov}(\tilde{p}_j, x_j' \text{Var}(\beta) x_j)}{\text{Var}(\tilde{p}_j)}. \quad (29)$$

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 pulled away from the average mandate slope in either direction—including toward, or past, the near-unit mechanical-tracking benchmark that the literature reads as inelastic or upward-sloping demand.

Remark 2 (When aggregation shifts the consolidated slope upward). The projection wedge in (28) is not signed in general, but the scalar-price case is instructive. With common characteristic loadings the leading curvature term is $\Gamma_j = \frac{1}{2} \text{Var}(\alpha) p_j^2$, so

$$\hat{\alpha} - \bar{\alpha} = \frac{1}{2} \text{Var}(\alpha) \frac{\text{Cov}(\tilde{p}_j, p_j^2)}{\text{Var}(\tilde{p}_j)}. \quad (30)$$

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*

$$\zeta_j^{\text{lin}} \equiv \Lambda_j - (\hat{c} + \hat{\alpha} p_j + x_j' \hat{\beta}). \quad (31)$$

This decomposes as

$$\zeta_j^{\text{lin}} \approx \tilde{\zeta}_j + g_j, \quad g_j \equiv (I - P_{p,x}) \Gamma_j, \quad (32)$$

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

$$\text{Var}(\xi^{\text{lin}}) = \text{Var}(\tilde{\xi}) + \text{Var}(g) + 2\text{Cov}(\tilde{\xi}, g) \geq \text{Var}(\tilde{\xi}) \quad (33)$$

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.

3 Estimation Design: Observation Unit and Moment Condition

The theory points to two changes in how the demand system should be taken to holdings data. The first is to estimate at the level at which the portfolio decision is made, rather than on a consolidated institution that bundles heterogeneous mandates. The second is to estimate with a moment condition that respects the information content of portfolio shares, rather than mechanically treating every reported holding as an equally precise observation. This section develops those two changes. Data construction, characteristic definitions, basis conversions, numerical preconditioning, and screening rules are collected in Appendix D.

3.1 Change in observation unit: from institutions to mandates

The first change is the observational unit. A consolidated 13F filer is typically a portfolio of portfolios: broad index products, active growth and value funds, sector mandates, and externally managed sleeves can all appear as one institutional position. Section 2 shows that estimating a linear demand curve on that rollup does not recover an average mandate. It recovers a projection of the nonlinear mixture.

I therefore use individual mutual-fund portfolios in the CRSP Mutual Fund Database as mandate-like decision units. Multiple share classes that implement the same portfolio are combined at the CRSP portfolio-number level, but economically distinct funds within the same family remain separate. The sample is restricted to U.S. equity mandates, so the estimated slopes describe relative allocation within the equity universe. The mutual-fund sector covers roughly 29% of U.S. market capitalization in 2024 and is sufficiently passive that it provides a conservative setting for recovering downward-sloping demand.

This design also permits a same-holdings validation exercise. After estimating the constituent mandates, I reconsolidate the same funds into synthetic families and re-estimate the institution-level model. Any wedge between the two estimates is then generated by the observational unit, not by a different stock sample, characteristic set, or investor population. Appendix D.1 describes the sample, rolling choice universe, share-class aggregation, and coverage in detail.

3.2 Change in moment condition: from ratios to levels

The second change concerns how each position enters the estimating moment. Let w_{fjt} be the observed within-universe share, $\mu_{fjt}(\theta)$ the fitted logit share, and z_{fjt} the instrument vector.

The level moment used in this paper is

$$g_{ft}^{\text{lev}}(\theta) \equiv \sum_{j \in \mathcal{U}_{ft}} z_{fjt} \{w_{fjt} - \mu_{fjt}(\theta)\} = 0. \quad (34)$$

The ratio moment used in the canonical implementation can be written as

$$g_{ft}^{\text{rat}}(\theta) \equiv \sum_{j \in \mathcal{U}_{ft}} z_{fjt} \left\{ \frac{w_{fjt}}{\mu_{fjt}(\theta)} - 1 \right\} = \sum_{j \in \mathcal{U}_{ft}} \frac{z_{fjt}}{\mu_{fjt}(\theta)} \{w_{fjt} - \mu_{fjt}(\theta)\} = 0. \quad (35)$$

Thus the difference is a position weighting. The level moment matches observed and fitted portfolio exposures and lets the simplex naturally dampen tiny positions. The ratio moment divides out that damping, so a one-basis-point holding can enter with the same first-order weight as a one-percent holding.

Under perfectly measured shares, inverse-share weighting can make the nominal holdings count N appear as the effective cross-section. Section 1, however, shows what this gain requires: the extra precision is purchased almost entirely from the long tail that carries little portfolio mass. In holdings data that tail is also where rounding, reporting floors, stale positions, and residual allocations are most important. A fixed absolute reporting error is multiplied by $1/\mu_{fjt}$ in the ratio moment but enters the level moment without that amplification. The level design therefore accepts the genuine information ceiling n^{eff} in exchange for robustness to the part of the data least capable of supporting equal weighting. Appendix A gives the fixed-design variance calculation, and Appendix C quantifies the effect of floor censoring and grid quantization.

Figure 2 shows the distinction in a setting where the economic benchmark is transparent. For S&P 500 and Russell 1000 trackers, whose relevant positions are comparatively well measured, the two moments yield similar estimates near the passive benchmark $\hat{\alpha} \approx 1$. For Russell 3000 and total-market trackers, the ratio estimates disperse sharply and assign more variation to latent demand, while the level estimates remain tightly concentrated near the mechanical benchmark. The empirical difference appears exactly where the $1/\mu_j$ weighting is most aggressive: broad portfolios with long, discretized tails.

The moment change is therefore not a cosmetic efficiency adjustment. It chooses the statistical regime appropriate for portfolio data. The ratio moment treats the support of the portfolio as N equally reliable demand observations; the level moment treats the portfolio as one unit of capital whose informative cross-section is its effective core.

3.3 Baseline specification and implementation summary

For each fund-quarter, the systematic demand index is

$$\eta_{fjt}(\theta_{ft}) = \alpha_{ft} \log \text{ME}_{jt} + \beta_{ft}^{BM} \log(\text{B/M})_{jt} + X'_{jt} \gamma_{ft}, \quad (36)$$

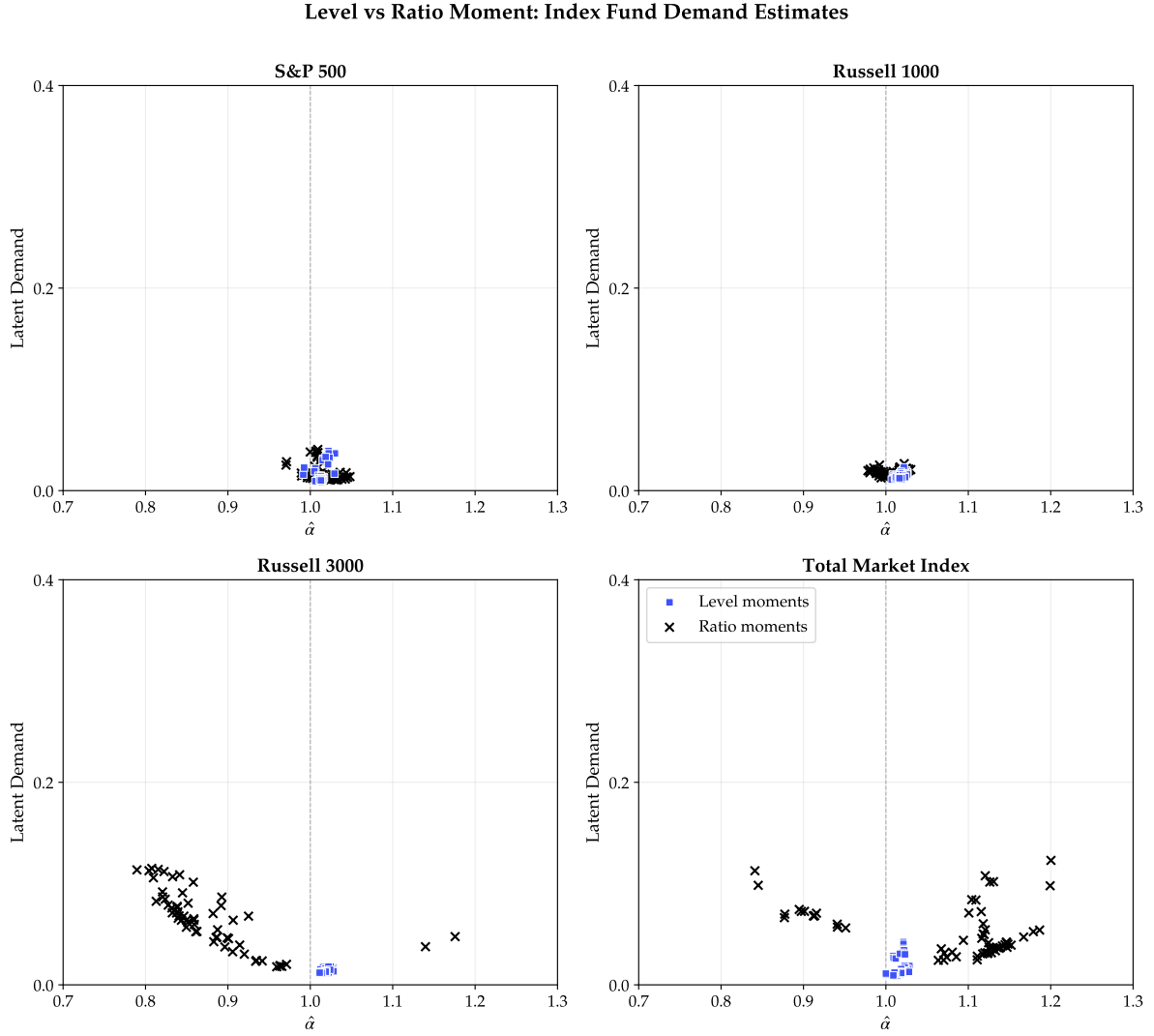


Figure 2: Equal weighting reads the contaminated tail; level weighting does not. Each panel plots fund-quarter estimates for index-like mutual funds matched to the indicated benchmark. Horizontal axis: the price coefficient $\hat{\alpha}$ recovered by each weighting. Vertical axis: the corresponding latent-demand component. Blue squares: level-weighted estimates; black crosses: equally weighted estimates. On narrow universes (S&P 500, Russell 1000) the two weightings agree closely, clustering near $\hat{\alpha} \approx 1$ with small latent demand. On broad, heavy-tailed universes (Russell 3000, Total Market Index) the equally weighted estimates scatter across a wide range of $\hat{\alpha}$ and imply substantially larger latent demand, while the level-weighted estimates remain tightly clustered near $\hat{\alpha} \approx 1$: the $1/\mu_j$ weighting amplifies tail-share measurement error that the level weighting discounts.

with fitted within-equity shares

$$\mu_{fjt}(\theta_{ft}) = \frac{\exp\{\eta_{fjt}(\theta_{ft})\}}{\sum_{k \in \mathcal{U}_{ft}} \exp\{\eta_{fkt}(\theta_{ft})\}}. \quad (37)$$

I instrument log market equity with the leave-one-out common-fund demand shifter of [Kojen and Yogo \(2019\)](#); exogenous characteristics serve as their own instruments. Parameters are estimated separately by fund-quarter using the level moments in equation (34), without sign restrictions. The implementation uses an algebraically equivalent $\{\log \text{ME}, \log \text{BE}\}$ basis, an effective-number preconditioner that does not change the minimizer, and conservative identification screens. Appendix D provides the full specification and construction.

4 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 4.1). 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 4.2). 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 4.3).

4.1 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 ([Kojen and Yogo, 2019](#)), which validates the synthetic benchmark and ensures the comparison isolates aggregation rather than differences in sample or method.

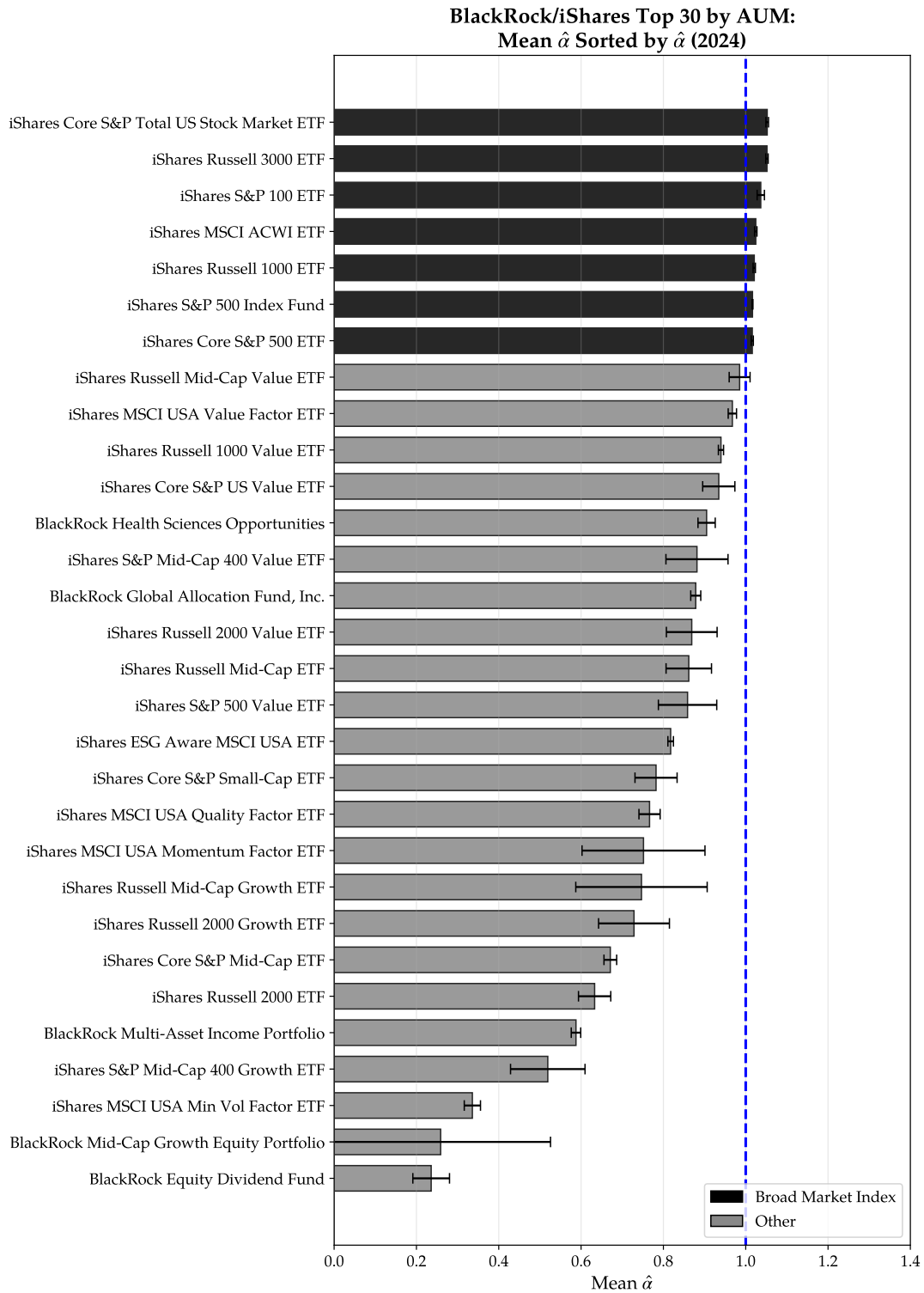


Figure 3: Mandate-level price coefficient estimates for BlackRock/iShares flagship funds (2024). Each fund's mean estimated $\hat{\alpha}$ (horizontal bars) with 95% confidence intervals (whiskers), for the top 30 products ranked by AUM. Funds are ordered by $\hat{\alpha}$; the dashed vertical line marks $\hat{\alpha} = 1$. Broad market index funds (black) are highlighted relative to other strategies (gray).

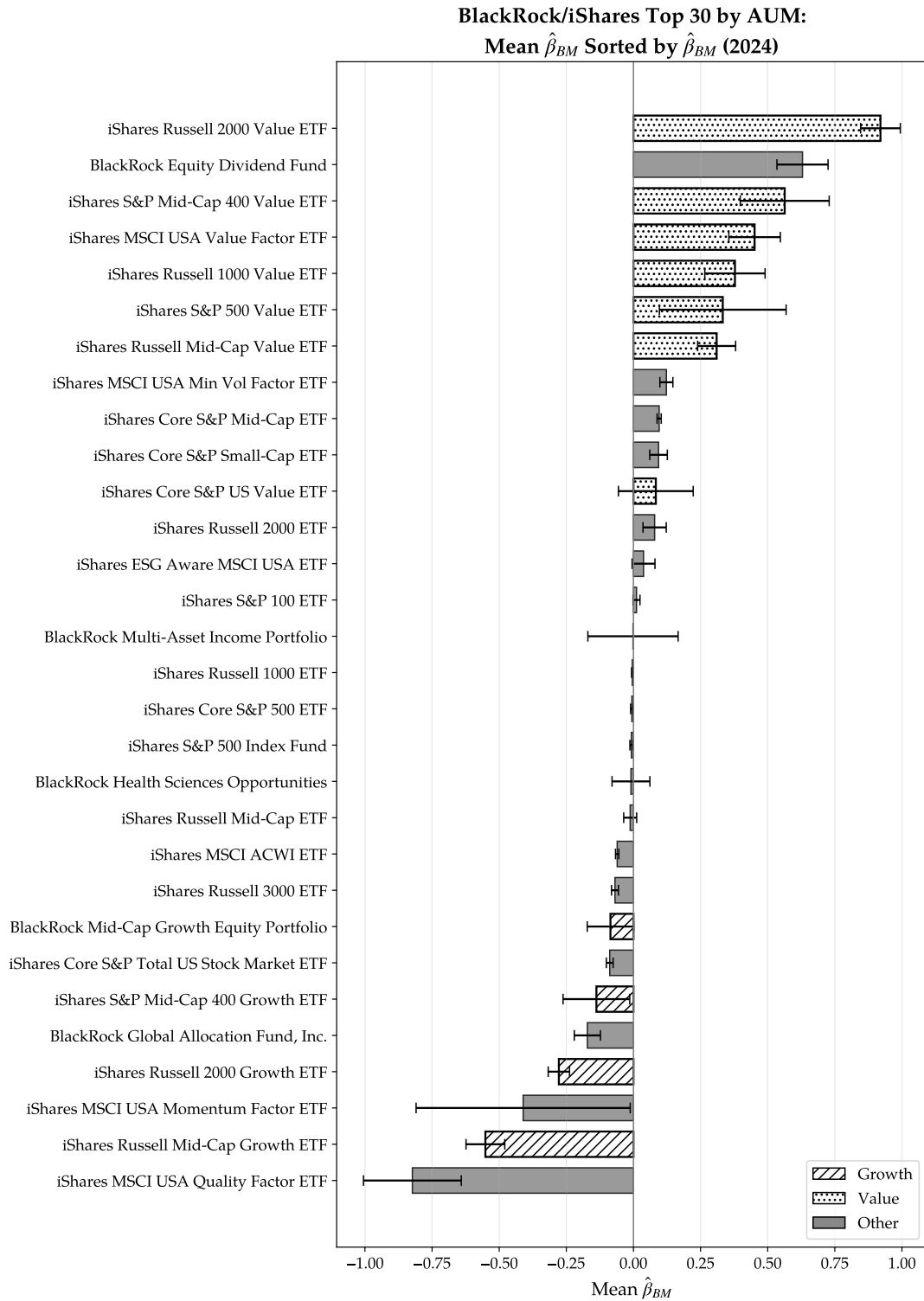


Figure 4: Mandate-level valuation-tilt coefficient estimates for BlackRock/iShares flagship funds (2024). Each fund's mean estimated $\hat{\beta}^{BM}$ with 95% confidence intervals. Value funds exhibit positive loadings, growth funds exhibit negative loadings, and broad market trackers cluster near zero.

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 3 and 4 report mandate-level estimates for the top 30 BlackRock/iShares products by AUM in 2024. Figure 3 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 4 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 1 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 within-institution dispersion; in this sample the coefficient channel is directionally

Table 1: 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 win-sorization bound (1st/99th percentile).

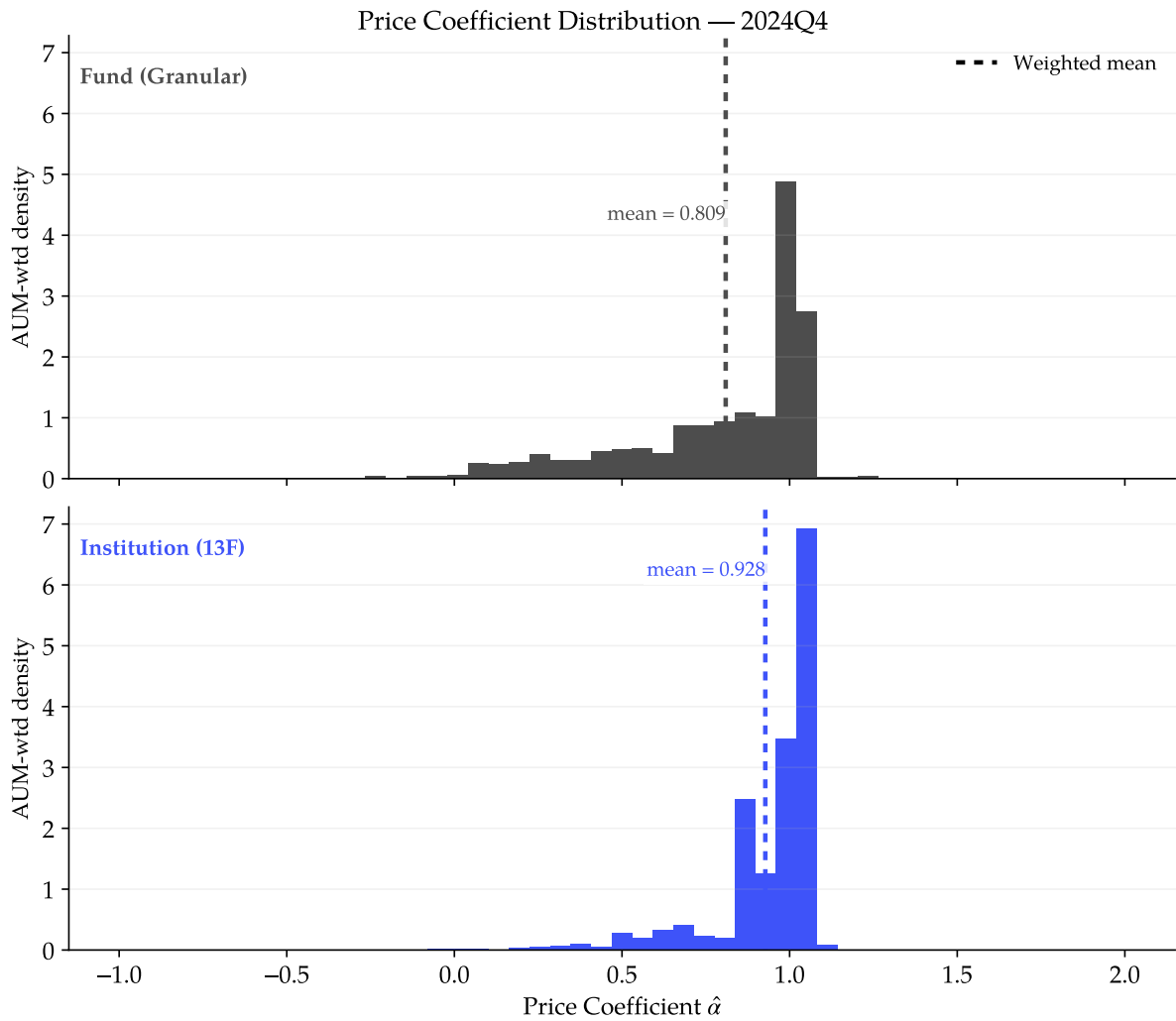


Figure 5: AUM-weighted distribution of the price coefficient $\hat{\alpha}$ at the fund level (top) versus the synthetic 13F institution level (bottom), 2024Q4. The rightward shift of the 13F distribution toward values near or above one illustrates the aggregation wedge: demand that is clearly downward-sloping at the mandate level appears near-mechanical at the institution level.

shifted toward the near-unit benchmark (Remark 2).

Distributional implication. Figure 5 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. (These means sit slightly below the top-30-family figures of Table 1, 0.84 and 0.96, because the distribution here covers all matched funds rather than only the largest families.) 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.

4.2 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 6 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 3, 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 7 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 3).

Two features of Figure 7 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](#)).

4.3 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 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, $L_{jt}^{MF} = \sum_f \hat{e}_{fjt}^{\$}$, 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 2 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-

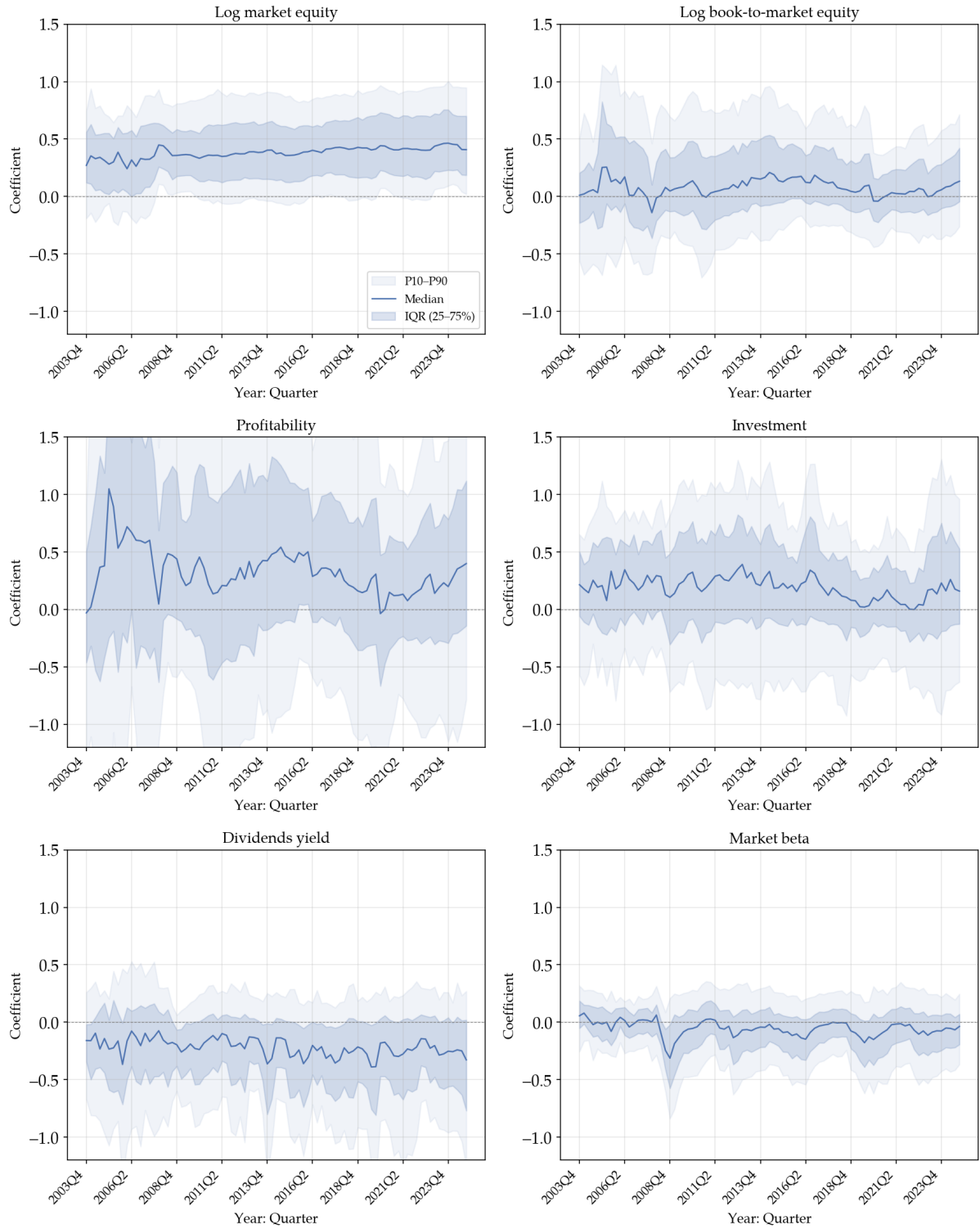


Figure 6: Coefficients on characteristics. Each panel plots the cross-sectional median (solid line), the interquartile range (dark shaded band, 25th–75th percentiles), and the 10th–90th percentile range (light shaded band) of mandate-level coefficient estimates each quarter. Coefficients are estimated quarter-by-quarter from mandate holdings using the baseline demand specification; the dashed horizontal line marks zero.

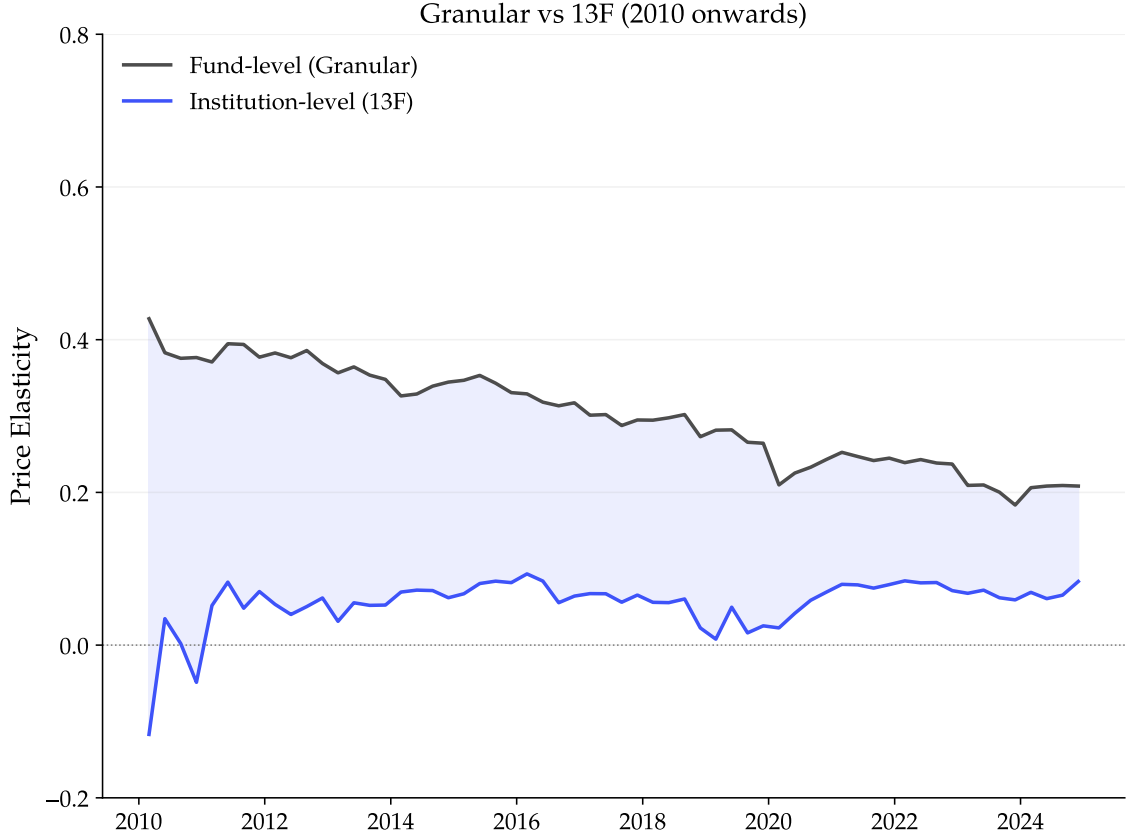


Figure 7: Dollar-weighted mean stock-level mutual-fund price elasticity E_{jt}^{MF} over time, computed from the same underlying mandate-level holdings at two levels of aggregation: the mandate (fund) level, and a synthetic institution formed by consolidating each family's mandates and re-estimating 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.

predicted excess demand, and a *consensus index* defined as

$$C_{jt} = \frac{|\sum_f \hat{e}_{fjt}^s|}{\sum_f |\hat{e}_{fjt}^s|}, \quad (38)$$

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 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.

5 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 de-

Table 2: 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 (38). 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

mand 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 are pulled away from the average mandate slope—in this sample, toward the near-unit region the literature reads as extreme inelasticity. 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 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.

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.
- Dahlquist, Magnus and Markus Ibert**, “Equity Return Expectations and Portfolios: Evidence from Large Asset Managers,” *Review of Financial Studies*, 2024, 37 (6), 1887–1928.
- Fuchs, William, Satoshi Fukuda, and Daniel Neuhann**, “Demand-System Asset Pricing: Theoretical Foundations,” 2025. Working Paper.
- Gabaix, Xavier and Ralph S. J. Koijen**, “In Search of the Origins of Financial Fluctuations: The Inelastic Markets Hypothesis,” NBER Working Paper 28967, National Bureau of Economic Research 2021.
- Giglio, Stefano, Matteo Maggiori, Johannes Stroebe, and Stephen Utkus**, “Five Facts about Beliefs and Portfolios,” *American Economic Review*, 2021, 111 (5), 1481–1522.
- Greenwood, Robin and David Thesmar**, “Stock Price Fragility,” *Journal of Financial Economics*, 2011, 102 (3), 471–490.
- and **Marco Sammon**, “The Disappearing Index Effect,” *Journal of Finance*, 2025, 80 (2), 657–698.
- 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 (3), 975–1018.
- He, Zhiguo, Peter Kondor, and Xing Li**, “Demand Elasticity in Dynamic Asset Pricing,” 2026. Working Paper.
- Hong, Harrison and Jeremy C. Stein**, “Disagreement and the Stock Market,” *Journal of Economic Perspectives*, 2007, 21 (2), 109–128.
- Kandel, Eugene and Neil D. Pearson**, “Differential Interpretation of Public Signals and Trade in Speculative Markets,” *Journal of Political Economy*, 1995, 103 (4), 831–872.
- 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.
- Shleifer, Andrei**, “Do Demand Curves for Stocks Slope Down?,” *Journal of Finance*, 1986, 41 (3), 579–590.
- van Binsbergen, Jules H., Joel M. David, and Christian C. Opp**, “How (Not) to Identify Demand Elasticities in Dynamic Asset Markets,” 2025. Working Paper.
- van der Beck, Philippe, Lorenzo Bretscher, and Julie Zhiyu Fu**, “A Bound on Price Impact and Disagreement,” Working Paper 26-053, Harvard Business School 2025. Also circulated as Swiss Finance Institute Research Paper 25-77.
- Wurgler, Jeffrey and Ekaterina Zhuravskaya**, “Does Arbitrage Flatten Demand Curves for Stocks?,” *Journal of Business*, 2002, 75 (4), 583–608.

A The Level Moment and the Effective Sample Size

This appendix supplies the estimator-level calculation behind Proposition 1. It keeps separate the observed share vector, the fitted share vector, the residualized price characteristic, and the instrument that the main text suppresses for readability.

A.1 Model, shares, and the local evaluation point

Fix a fund-quarter. Latent demand in log quantities is

$$\log q_j = \alpha p_j + x_j' \beta + \varepsilon_j, \quad \mathbb{E}[\varepsilon_j] = 0, \quad \text{Var}(\varepsilon_j) = \sigma^2, \quad (39)$$

with ε_j independent across positions. Observed within-universe shares are $w_j = q_j / \sum_k q_k$. Let $\mu(\alpha, \beta)$ denote the multinomial-logit shares implied by the systematic index, and write $\mu_0 \equiv \mu(\alpha_0, \beta_0)$ at the local evaluation point.

The share map differentiates to

$$d\mu = \Omega d \log q, \quad \Omega \equiv \text{diag}(\mu_0) - \mu_0 \mu_0'. \quad (40)$$

For any vector v ,

$$(\Omega v)_j = \mu_{0j}(v_j - \bar{v}_\mu), \quad \bar{v}_\mu \equiv \sum_k \mu_{0k} v_k.$$

The local share model is therefore

$$w - \mu_0 = \Omega \varepsilon + O(\|\varepsilon\|^2), \quad \left. \frac{\partial \mu}{\partial \alpha} \right|_{\theta_0} = \Omega p. \quad (41)$$

All derivatives and conditional variance calculations below are evaluated at μ_0 . The main text writes μ_{0j} for this local share vector and uses observed shares w_j as its empirical plug-in.

A.2 Partialling, centering, and the exact linearization

Let X_- collect the non-price characteristics. Define the residualized price characteristic in the share metric by

$$\tilde{p} \equiv p - X_- (X_-' \Omega X_-)^{-1} X_-' \Omega p. \quad (42)$$

Center and rescale $\tilde{p}$ so that

$$\sum_j \mu_{0j} \tilde{p}_j = 0, \quad \sum_j \mu_{0j} \tilde{p}_j^2 = 1. \quad (43)$$

The main text writes $x \equiv \tilde{p}$ after this normalization.

The level moment for the price coefficient is

$$z' \{w - \mu(\alpha)\} = 0, \quad (44)$$

where z is the instrument for $\tilde{p}$. Its first-order score is

$$z'(w - \mu_0) = z'\Omega\varepsilon + O(\|\varepsilon\|^2) \quad (45)$$

$$= \sum_j \mu_{0j}(z_j - \bar{z}_\mu)\varepsilon_j + O(\|\varepsilon\|^2). \quad (46)$$

If z is weighted-centered, $\bar{z}_\mu = 0$, so the portfolio-average shock cancels exactly and the first-order score becomes $\sum_j \mu_{0j}z_j\varepsilon_j$.

Linearizing (44) around the truth gives

$$\hat{\alpha} - \alpha_0 \approx \frac{z'\Omega\varepsilon}{z'\Omega\tilde{p}}, \quad \text{Var}(\hat{\alpha} \mid \mu_0, \tilde{p}, z) = \sigma^2 \frac{z'\Omega^2 z}{(z'\Omega\tilde{p})^2}. \quad (47)$$

The corresponding realized effective sample size is

$$N_{\text{eff}}(z, \tilde{p}; \mu_0) \equiv \frac{\sigma^2}{\text{Var}(\hat{\alpha} \mid \mu_0, \tilde{p}, z)} = \frac{(z'\Omega\tilde{p})^2}{z'\Omega^2 z}. \quad (48)$$

This is a fixed-design, finite-sample calculation. It does not invoke an asymptotic argument in the number of listed holdings.

A.3 The exogenous case and the alignment factor

Set $z = \tilde{p} \equiv x$ and use the weighted centering and standardization in (43). Then

$$x'\Omega x = \sum_j \mu_{0j}x_j^2 = 1,$$

and

$$x'\Omega^2 x = \sum_j \mu_{0j}^2 x_j^2.$$

Hence

$$\text{Var}(\hat{\alpha} \mid \mu_0, x) = \sigma^2 \sum_j \mu_{0j}^2 x_j^2. \quad (49)$$

Let

$$S_2 \equiv \sum_j \mu_{0j}^2, \quad \kappa \equiv \sum_j \frac{\mu_{0j}^2}{S_2} x_j^2. \quad (50)$$

Then

$$\text{Var}(\hat{\alpha}) = \sigma^2 S_2 \kappa, \quad N_{\text{eff}} = \frac{1}{S_2 \kappa} = \frac{n^{\text{eff}}}{\kappa}. \quad (51)$$

The factor κ contains the direction-specific part of the problem. It exceeds one when the squared characteristic is concentrated in the large positions, which raises variance and lowers effective sample size. It falls below one when characteristic extremeness is concentrated in smaller positions. The approximation in the main text is $\kappa \approx 1$.

A.4 The generic-direction benchmark

Let x have independent entries with mean zero and unit variance, independent of the portfolio weights. For any fixed matrix A , $\mathbb{E}[x'Ax] = \text{tr}(A)$. Writing $S_k \equiv \sum_j \mu_{0j}^k$, the trace identities are

$$\text{tr}(\Omega) = 1 - S_2, \quad \text{tr}(\Omega^2) = S_2 - 2S_3 + S_2^2. \quad (52)$$

Replacing the realized quadratic forms in (48) by these design expectations gives

$$\bar{N}_{\text{eff}} = \frac{\{\text{tr}(\Omega)\}^2}{\text{tr}(\Omega^2)} = \frac{(1 - S_2)^2}{S_2 - 2S_3 + S_2^2} \approx \frac{1}{S_2} = n^{\text{eff}}, \quad (53)$$

where the approximation holds when no single position dominates the portfolio, i.e., in the granular regime of Definition 1.

A.5 Equal weights

Let $\mu_{0j} = 1/N$, $z = x$, $\mathbf{1}'x = 0$, and $N^{-1}x'x = 1$. With $P \equiv I - \mathbf{1}\mathbf{1}'/N$, equal weights imply $\Omega = P/N$. Therefore

$$x'\Omega x = \frac{x'Px}{N} = 1, \quad x'\Omega^2 x = \frac{x'Px}{N^2} = \frac{1}{N},$$

so

$$N_{\text{eff}} = N, \quad \text{Var}(\hat{\alpha}) = \frac{\sigma^2}{N}. \quad (54)$$

The textbook regression variance is therefore recovered exactly when all portfolio slices are equal.

A.6 Remarks

First, the quadratic form $x'\Omega^2 x$ is the score-noise term of the level moment, not its information denominator. The information object is the signal-to-noise ratio in (48). Treating $x'\Omega^2 x$ as information reverses the concentration comparative static.

Second, zero positions enter the level moment without modification through $w_j = 0$. Moments that divide by small fitted shares place fundamentally different weight on the tail; this weighting choice is taken up in Section 3.2.

Third, when $z \neq x$, instrument alignment enters alongside the characteristic-alignment factor in (51). The empirical implementation therefore reports realized, partialled identification diagnostics rather than relying only on $1/\text{HHI}$.

Finally, all statements are per fund-quarter. Information accumulates across fund-quarters, not through nominal expansion of a single portfolio's long tail.

B 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

$$\frac{d\Lambda_{ji}}{dp_j} = \sum_{m \in \mathcal{M}_i} \tilde{\pi}_{mi} \alpha_m (1 - w_{jm}) + \sum_{m \in \mathcal{M}_i} \pi_{mi} \alpha_m w_{jm}. \quad (55)$$

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

$$\frac{d^2 \Lambda_{ji}}{dp_j^2} = \text{Var}_{\tilde{\pi}}(\alpha_m (1 - w_{jm})) - \text{Var}_{\pi}(\alpha_m w_{jm}) + \sum_{m \in \mathcal{M}_i} (\pi_{mi} - \tilde{\pi}_{mi}) \alpha_m^2 w_{jm} (1 - w_{jm}). \quad (56)$$

Granular-portfolio limit

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

- $(1 - w_{jm}) \rightarrow 1$, so $\alpha_m (1 - w_{jm}) \rightarrow \alpha_m$;
- all terms carrying w_{jm} (indirect effects, variance corrections) vanish at rate $O(\max_m w_{jm}) = O(1/n^{\text{eff}})$ in the granular regime.

The first derivative reduces to $\mathbb{E}_{\tilde{\pi}}[\alpha_m]$ and the second to $\text{Var}_{\tilde{\pi}}(\alpha_m)$, as stated in the main text [equations (23)–(24)].

C 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 quantifies the claim made in Section 3.2: contamination on the reporting grid perturbs the level-moment estimator by an amount an order of magnitude below its sampling error, while the same contamination enters the ratio moment at first order.

Throughout, notation follows Appendix A: x_j is the residualized price characteristic, weighted-centered and weighted-standardized with respect to the local share profile μ_0 as in (43), and the level moment is $\sum_j x_j \{w_j - \mu_j(\alpha)\} = 0$.

Figure 8 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 9 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 2 (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*

$$\hat{\alpha} - \alpha_0 \approx \mathcal{I}(\alpha_0)^{-1} \sum_{j \in \mathcal{U}} c_j \mu_j(\alpha_0) p_j^c. \quad (57)$$

- (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*

$$|\text{bias}_{\text{floor}}| \leq n^{\text{eff}} K \underline{w}^2 \frac{\overline{|p^c|}_{\text{cens}}}{\bar{\sigma}_p^2}. \quad (58)$$

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*

$$\text{Var}_{\text{quant}}(\hat{\alpha}) = \mathcal{I}(\alpha_0)^{-2} \frac{\Delta^2}{12} \sum_{j \in \mathcal{U}} \mu_j^2 (p_j^c)^2 = \frac{\Delta^2}{12} \mathcal{I}(\alpha_0)^{-1}. \quad (59)$$

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 (57) 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 (57) satisfies $|c_j \mu_j p_j^c| \leq \underline{w}^2 |p_j^c|$. Summing over K censored stocks and dividing by $\mathcal{I}(\alpha_0) = O(1/n^{\text{eff}})$

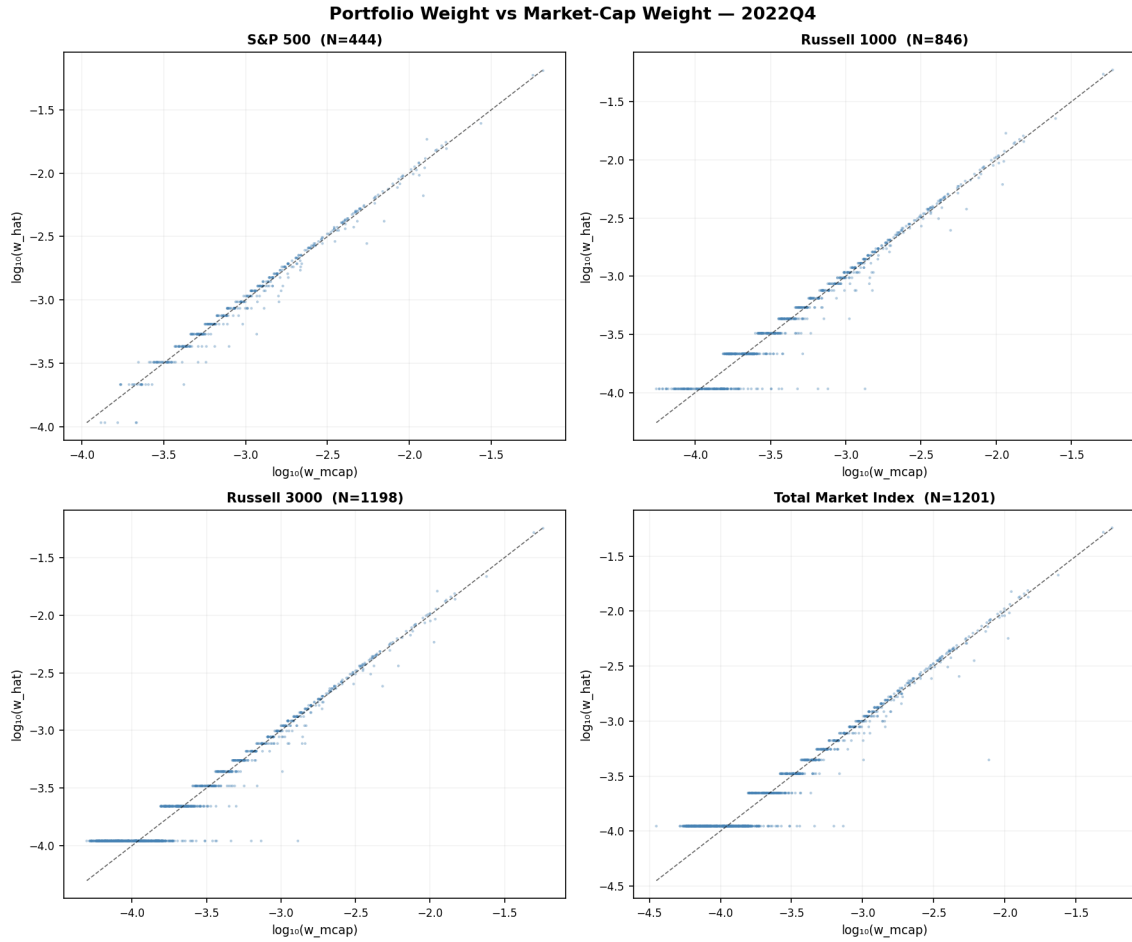


Figure 8: **Portfolio weight vs. market-cap weight for index funds (2022Q4).** Each panel plots $\log_{10}(w_{\text{hat}})$ against $\log_{10}(w_{\text{mcap}})$ for the indicated index fund. The dashed line is the 45-degree line. Horizontal banding at the bottom reflects the discrete reporting floor ($w \approx 10^{-4}$). The number of securities in each fund's effective universe is shown in parentheses.

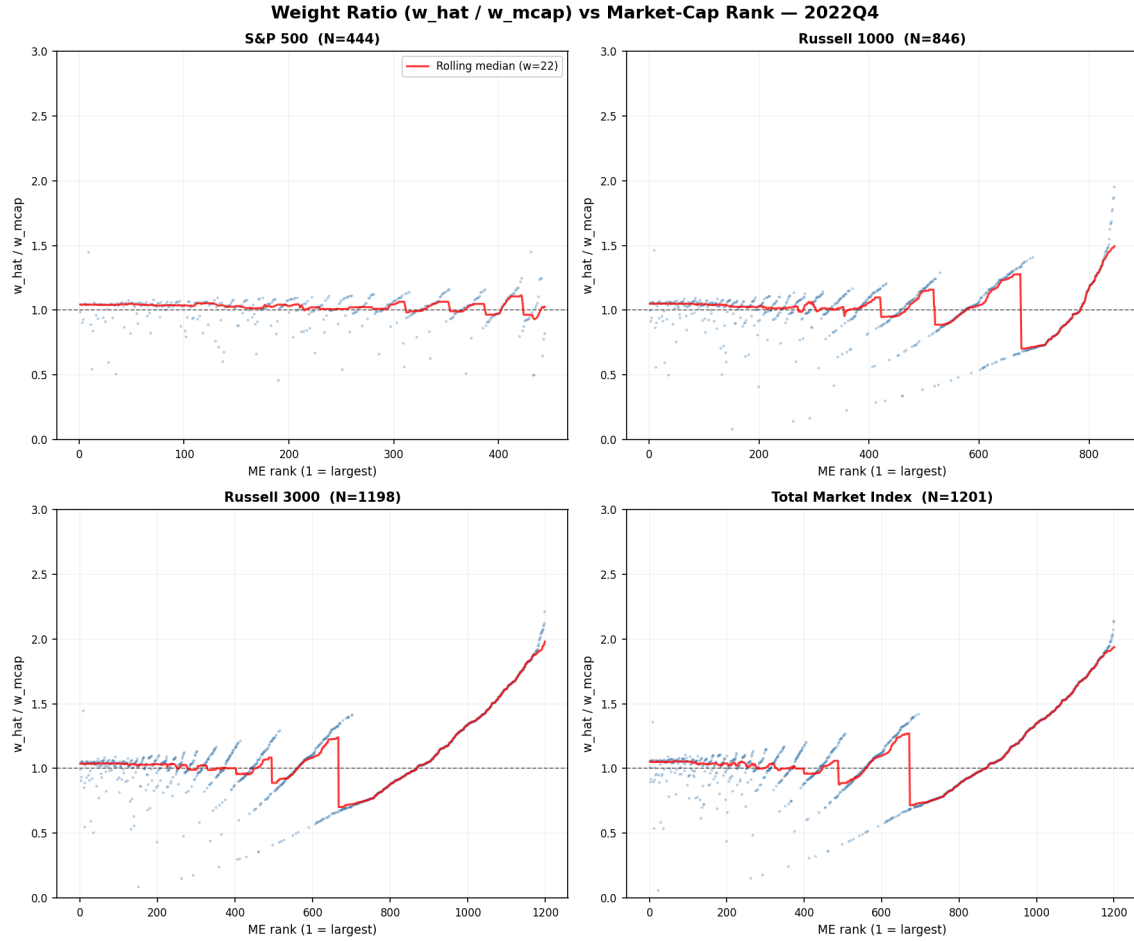


Figure 9: **Weight ratio vs. market-cap rank for index funds (2022Q4).** Each panel plots $w_{\text{hat}}/w_{\text{mcap}}$ against market-cap rank (1 = largest) for the indicated index fund. The red line is a rolling median (window = 22). For narrow-universe funds (S&P 500, Russell 1000), the ratio is 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.

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. $\square$

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 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.

D Data, Specification, and Implementation Details

This appendix collects the data construction and implementation details that are deliberately suppressed in Section 3, where the main text focuses on the two substantive design changes: the observation unit and the moment condition.

D.1 Data and observation-unit construction

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 3) 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 [Kojen 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 3 and 4 summarize the sample. Table 3 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 4 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 3: 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 addressed by the level-moment design in Section 3.2.

Table 3: Sample summary: coverage and mutual-fund ownership (2003–2024)

Year	# Funds	Mkt cap	% Market held (MF sector)					Total AUM
		(\$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 4: Sample summary: portfolio granularity and tail concentration (2003–2024)

Year	AUM/fund	Universe size		# Holdings		$w < 1\%$		$w < 0.1\%$	
	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.

D.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

$$S_{ft} \equiv \sum_{j \in \mathcal{U}_{ft}} w_{fjt}. \quad (60)$$

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:

- $\log(\text{B/M})_{jt}$: 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);
- $\beta_{j,t-1}^{\text{mkt}}$: market beta estimated from a trailing 36-month rolling regression of excess stock returns on excess market returns;
- I_{jt}^{div} : a dividend-payer indicator (equals one if the firm paid a positive dividend within the trailing twelve months); and
- $\log(1 + D/\text{BE})_{jt}$: 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 [Kojen and Yogo \(2019\)](#)'s equation (10). For each (f, t) and $j \in \mathcal{U}_{ft}$, define a linear index

$$\eta_{fjt}(\theta_{ft}) = \alpha_{ft} \log \text{ME}_{jt} + \beta_{ft}^{\text{BM}} \log(\text{B/M})_{jt} + X'_{jt} \gamma_{ft}, \quad (61)$$

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

$$\mu_{fjt}(\theta_{ft}) = \frac{\exp(\eta_{fjt})}{\sum_{k \in \mathcal{U}_{ft}} \exp(\eta_{fkt})}, \quad (62)$$

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 (64)) 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:

$$\frac{w_i(n)}{\sum_{\ell \in \mathcal{N}_i} w_i(\ell)} = \frac{\exp\{r'_n \theta_i\} \epsilon_i(n)}{\sum_{\ell \in \mathcal{N}_i} \exp\{r'_\ell \theta_i\} \epsilon_i(\ell)}. \quad (63)$$

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 (62) 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 ,

$$\frac{\exp\{a + r'_{jt} \theta\}}{\sum_k \exp\{a + r'_{kt} \theta\}} = \frac{\exp\{r'_{jt} \theta\}}{\sum_k \exp\{r'_{kt} \theta\}}. \quad (64)$$

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 (61) separates two economically distinct dimensions: $\log \text{ME}$ captures market capitalization, while $\log(\text{B}/\text{M})$ captures valuation. Because $\log \text{ME}_{jt} = \log P_{jt} + \log \text{Shares}_{jt}$ and shares outstanding move slowly relative to prices, cross-sectional and short-horizon time-series variation in $\log \text{ME}$ 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.⁷ 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 4.1: 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 $\log(\text{B}/\text{M})$ captures differences in valuation. The coefficient β_{ft}^{BM} governs the fund’s sensitivity to this dimension:

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

D.3 Implementation

This section collects the practical details of the estimation procedure. The baseline specification (61)–(62) defines the model; here I describe how it is taken to data.

Regressor implementation: $\log \text{ME} + \log \text{BE}$. The specification (61) uses $\log \text{ME}$ and $\log(\text{B}/\text{M})$ as the two size-related regressors. In implementation, I replace $\log(\text{B}/\text{M})$ with $\log \text{BE}$ and estimate the index as

$$\eta_{fjt} = \alpha_{ft}^* \log \text{ME}_{jt} + \beta_{ft}^{BE} \log \text{BE}_{jt} + X'_{jt} \gamma_{ft}, \quad (65)$$

where the asterisk on α^* distinguishes the coefficient in this basis from the conceptual α in (61). Since $\log(\text{B}/\text{M}) = \log \text{BE} - \log \text{ME}$, the two specifications are algebraically equivalent, with the identification

$$\alpha_{ft} = \alpha_{ft}^* + \beta_{ft}^{BE}, \quad \beta_{ft}^{BM} = \beta_{ft}^{BE}. \quad (66)$$

⁷This 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.

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 $\log(\text{B}/\text{M})$ 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 (66).

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

$$\alpha_{ft} = \frac{\hat{\alpha}_{ft}^{*,\text{std}}}{\sigma_{\log \text{ME}}} + \frac{\hat{\beta}_{ft}^{BE,\text{std}}}{\sigma_{\log \text{BE}}}, \quad (67)$$

with standard error from the delta method.⁸

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 [Kojien and Yogo \(2019\)](#), I instrument for $\log \text{ME}_{jt}$ using a measure of the stock's common fund flow pressure:

$$\text{cf}_{jt}^{\text{me}} = \sum_{f: j \in \mathcal{U}_{ft}} \frac{A_{ft}}{1 + |\mathcal{U}_{ft}|}, \quad (68)$$

where A_{ft} is fund f 's total AUM. The instrument is $z_{jt} = \log(\text{cf}_{jt}^{\text{me}})$, 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 (68). The moment conditions are $\mathbb{E}[z_{fjt}^{\text{loo}}(w_{fjt} - \mu_{fjt}(\theta_{ft}))] = 0$, 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:

$$\hat{\theta}_{ft} \in \arg \min_{\theta} \left\| \sum_{j \in \mathcal{U}_{ft}} z_{fjt} (w_{fjt} - \mu_{fjt}(\theta)) \right\|^2, \quad (69)$$

where $z_{fjt} = (z_{fjt}^{\text{loo}}, x'_{jt})$ 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 nor-

⁸Let 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 $\text{se}(\alpha_{ft}) = (V_{11}/\sigma_{\log \text{ME}}^2 + V_{kk}/\sigma_{\log \text{BE}}^2 + 2V_{1k}/(\sigma_{\log \text{ME}}\sigma_{\log \text{BE}}))^{1/2}$. 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.

malization that rescales the entire moment system without changing the minimizer. Define

$$\text{HHI}_{ft} \equiv \sum_{j \in \mathcal{U}_{ft}} \left(\frac{w_{fjt}}{S_{ft}} \right)^2, \quad n_{ft}^{\text{eff}} \equiv \frac{1}{\text{HHI}_{ft}}, \quad s_{ft} \equiv n_{ft}^{\text{eff}}. \quad (70)$$

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.

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 $n_{ft}^{\text{eff}} < 20$ are excluded.⁹ Third, for each characteristic x^k , I compute the weight-weighted dispersion

$$\text{Disp}_{ft}(x^k) \equiv \left(\sum_{j \in \mathcal{U}_{ft}} \frac{w_{fjt}}{S_{ft}} \left(x_{jt}^k - \bar{x}_{ft}^k \right)^2 \right)^{1/2}; \quad (71)$$

if this falls below 0.1 in standardized units, the corresponding $\hat{\beta}_{ft}^k$ is set to missing. These filters exclude less than 5.6% of AUM each quarter.

Section 4 uses the resulting panel $\{\hat{\alpha}_{ft}, \hat{\beta}_{ft}^{BM}, \hat{\gamma}_{ft}\}$ to benchmark estimates against well-known funds and to document the cross-sectional distribution of characteristic demand.

⁹This threshold removes sector specialists, extremely concentrated mandates, and mechanically thin universes where cross-sectional moments are unstable.