# Risk and Return in Asset Demand Systems *

**Ozan E. Akbas** †  **Ao Wang** ‡

First Version: 15 December 2024
This Version: 18 July 2025

**Abstract**

We develop a characteristic-based asset demand model in which cross-asset risk-return trade-offs vary with asset characteristics. The model relaxes the uniform substitution structure of the multinomial logit (MNL), accommodates large price elasticities, and enables recovery of investor-specific primitives, including alphas and factor loadings, from structural demand estimates. Applied to U.S. institutional equity holdings from 2000 to 2022, the model reveals meaningful deviations from MNL substitution patterns, particularly along the market equity dimension. The estimated average own-price elasticity is 77 percent higher than under the MNL, driven largely by investors whose portfolios imply cross-asset complementarity. Nonetheless, both elasticity estimates are substantially lower than those implied by CAPM calibrations. The model also uncovers heterogeneity in investor alphas: hedge funds earn near-zero alphas, while brokers earn up to five basis points annually.

**JEL Codes:** C51, G11, G23
**Keywords:** asset demand systems, flexible substitution, cross-asset complementarity

*We are thankful for useful comments by Andrea Gamba, Ruggero Jappelli, Noah Lyman, Lars Nesheim, Yushi Peng, Giulio Rossetti and seminar participants at University of Warwick, Tilburg University, Shanghai University of Finance and Economics, Finance Theory Group, and Bristol Econometrics Study Group.
†Warwick Business School, University of Warwick. E-mail: ozan.akbas@warwick.ac.uk
‡Department of Economics, University of Warwick / CAGE Research Centre. E-mail: ao.wang@warwick.ac.uk

## 1. Introduction

The demand-based approach pioneered by Koijen and Yogo (2019) (KY henceforth) provides a new way to understand investor heterogeneity and asset pricing using portfolio holdings data. Assuming characteristics-based factor structures in asset returns and factor loadings, they derive a multinomial logit asset demand model (MNL) from the mean-variance portfolio choice (Markowitz, 1952). Augmenting the MNL with instruments for potentially endogenous asset prices, one can back out investors' tastes and beliefs, and study the equilibrium consequences of counterfactual shocks in a bottom-up fashion.

This methodology has been gaining popularity in recent empirical asset pricing research, as applied various setups such as in corporate bond markets (e.g., Bretscher et al., 2025), in sovereign debt markets (e.g., Fang et al., 2025) or in quantitative easing (e.g., Koijen et al., 2021). However, the implications of the MNL on investor's asset demand are underdiscussed. In particular, the independence of irrelevant alternatives (IIA), an intrinsic property of the MNL (McFadden, 1974), implies that the investor perceives all her assets to be either uniformly substitutes, uniformly complements, or independent to each other in demand. Such uniform substitution pattern rules out, for instance, the possibility that the investor sees Tesla and General Motors stocks as substitutes, while Tesla and Panasonic as complements.[1] As a consequence, the MNL may fail to capture the stock holding adjustments following a negative Tesla shock where the investor increases her holding of General Motors but decreases the holding of Panasonic.

In this paper, we explore the economic origin of the MNL in asset pricing and develop a new characteristic-based asset demand model that relaxes its inherent uniform substitution pattern. First, we show that the new model shrinks to the MNL if and only if investor's perceived asset risk-return trade-off, i.e., the ratio between asset factor loadings and return, is homogeneous across assets. When this condition holds, the common risk-return trade-off is canceled out from the relative asset weights. Asset returns solely determine asset weights, leading to the IIA property and the uniform asset substitution dictated by the price coefficient in returns. Second, when the risk-return trade-off is heterogeneous across assets, it enters the relative asset weights. We decompose the resulting cross-asset demand

[1] Panasonic is among Tesla's major battery suppliers.

dependence into a diversion effect (as in the MNL), an intra-portfolio effect, and a portfolio-level effect. The latter two are linked to asset characteristics through the new parameters in the risk-return trade-off. As a result, asset substitution is jointly determined by the price coefficient in returns and that in the risk-return trade-off, breaking the uniform substitution pattern in the MNL. Third, we show that both the MNL and our model have the capacity to generate large price elasticities and small price effects. The MNL achieves this via highly negative price coefficients, while our model can attain it with more moderate coefficients, thanks to the additional flexibility introduced by the risk-return trade-off parameters. Crucially, this shared ability suggests that the empirical findings of low price elasticities in demand-based approaches (much lower than those implied by calibrated CAPMs) are not driven by model constraints, but reflect empirical realities. More in general, our model bridges the parameters in asset demand and the primitives in the underlying factor structure, such as investor's alpha. It allows us to back out these primitives from the demand parameters and shed light on new insights into the heterogeneity in investors' tastes and beliefs.

The identification and estimation of asset demand now involve both the parameters in asset returns and in risk-return trade-offs. The latter ones enter nonlinearly into the log relative asset weights. Besides, similarly to KY, the endogeneity concern still exists. We propose a two-step inverse-style identification strategy to deal with the non-linearity and price endogeneity. Relying on the market clearing condition, we show that the mandate-based instrument in KY can be seen as an asset popularity measure and a *cost-shifter* type argument: the more popular an asset is among investors, the higher one needs to pay to buy it. Moreover, we illustrate that the cross-sectional variation in the interactions of asset characteristics (e.g., quadratic terms) provides identification power for the risk-return trade-off parameters. This intuition via higher-order terms joins the arguments for identifying the *non-linear* parameters, i.e., the distribution of random coefficients, in BLP-type demand models (e.g., Gandhi and Houde (2019); Wang (2023, 2024)). Finally, our identification strategy leads to a practical GMM estimation procedure. In particular, we use the model's financial micro-foundation to reduce the demand inverse to an one-dimensional problem, solving the typical dimensionality challenge in the number of assets.

We apply our method to investigate institutional investors' equity demand using the U.S. stock market data from 2000 Q1 to 2022 Q4. We first derive from our model a reduced-form test of the if-and-only-if condition for the MNL asset demand. Its implementation relies on the estimates of classic CAPM variables—market-level asset returns and factor loadings—and does not require estimating the model of asset demand. We find significant deviations from the MNL and heterogeneous risk-return trade-off across assets via characteristics such as market equity. Our demand estimates are coherent with this suggestive evidence: more than half of the investor-quarter-level risk-return parameter estimates for market equity are significantly different from zero. Besides, our model outperforms the MNL in terms of out-of-sample prediction for the masked second-largest holding à la Gabaix et al. (2025a,b): with only one additional risk-return trade-off parameter for market equity, it reduces the root-mean-square error by around $30\%$.

Second, we find that the MNL underestimates the asset-level own-price elasticity ($0.45$) relative to our model ($0.79$). The difference is mainly driven by the investors who perceive their assets as complements under the MNL. For these investors, the MNL substantially underestimates their overall own-price elasticity ($0.20$) relative to ours ($0.50$), and also significantly overestimates their overall cross-asset complementarity ($0.80$ in the MNL versus $0.50$ in our model). In contrast, both models' predictions of the overall own- and cross-asset elasticities are similar for investors who perceive their assets as substitutes. Despite these differences, both models predict much smaller asset-level price elasticities than theoretically calibrated ones in the literature (e.g., Petajisto (2009)), supporting the inelastic market hypothesis in finance (Gabaix and Koijen, 2024).

Finally, we back out investor's alpha (relative to her outside asset) from the demand estimates. These alphas vary across investor types and stock groups, with hedge funds' averaging near-zero alphas, while brokers earning annualized alphas up to 5 basis points, which for example translates to approximately $\$15$ million on a $\$32$ billion portfolio for Morgan Stanley across 2000.Q1 and 2022.Q4. We find that the dispersion in the residual performance is mild across stock styles. For instance, brokers earn a steady 1 bp per quarter regardless of stock style, likely pointing to an intermediation edge rather than stock-selection skill.

**Related Literature.** This paper relates to two strands of research in finance and economics: (i) the literature on demand systems in asset pricing and (ii) the empirical industrial organization (IO) literature on demand. Koijen and Yogo (2019) propose the MNL framework for demand-based asset pricing, which follow-up work predominantly relies on. Applications across different asset classes include equities (Koijen et al., 2024; Haddad et al., 2025), corporate bonds (Koijen and Yogo, 2023; Chaudhary et al., 2023; Siani, 2025; Bretscher et al., 2025), government bonds (Koijen et al., 2021; Jansen et al., 2024; Eren et al., 2024), and currencies (Koijen and Yogo, 2024; Jiang et al., 2024, 2025). Our work contributes to this rising literature by proposing an asset demand model that builds on the same mean-variance microfoundation as the MNL but relaxes its inherent uniform asset substitution restriction. Our model is suitable for settings in which investor demand exhibits complex interdependence across assets—for example, where some assets are substitutes and others are complements. This flexibility is relevant in empirical settings where investors reallocate capital asymmetrically across assets—for instance, during market stress, sectoral rotations, or shifts in risk sentiment—leading to substitution patterns that the MNL may not capture.

Our paper joins recent work that seeks to improve the methodological foundations of demand-based asset pricing and to better understand the gap between the MNL and conventional asset pricing approaches. Advances include the nested logit model in corporate bond markets (Chaudhary et al., 2023) and the Almost Ideal Demand System applied to Canadian Treasury bonds (Allen et al., 2024). Like these approaches, our model allows for more flexible substitution patterns including both substitutability and complementarity. Differently, our model introduces richer substitution structures while remaining anchored in mean-variance optimization. As a result, it enables us to link demand parameters to investor-level financial primitives, such as alpha, offering new insights into investor behavior. Fuchs et al. (2024) and Davis et al. (2025) also point out the MNL's substitution limitations and discuss implications for asset complementarity and elasticity. Our model contributes to these discussions from within the mean-variance framework. In particular, because it nests the MNL as a special case, it can be used to test whether—and how—observed investor holdings deviate from MNL predictions, a feature not shared by the aforementioned alternatives.

Our paper also contributes to the ongoing discussion about the elasticity gap in asset pricing—the discrepancy between the low price elasticities estimated in demand-based models and the much higher ones implied by calibrated structural models such as the CAPM (Petajisto, 2009). We show that both the MNL and our model are theoretically capable of generating large elasticities, suggesting that the gap is not driven by model constraints. Our empirical estimates, based on a more flexible substitution structure, are larger than but of similar magnitude to those obtained under the MNL. This finding provides further support for the inelastic market hypothesis in the U.S. equity market (Koijen and Yogo, 2019; Gabaix and Koijen, 2024; Davis et al., 2025; Haddad et al., 2025).

Finally, our work bridges the asset pricing and empirical IO literatures by linking identification and estimation strategies across both fields. Our demand-inverse approach resembles the method pioneered by Berry (1994); Berry et al. (1995) and extended by Wang (2024), who incorporate product complementarity into the BLP framework. We show that the investment mandate instruments proposed by Koijen and Yogo (2019) function similarly to cost shifters in BLP-type settings, while the identification of risk-return trade-off parameters parallels the identification of random coefficient distributions. However, unlike BLP, our model is grounded in mean-variance portfolio choice rather than utility maximization. This distinction results in different economic origins for substitution and yields a demand inverse that avoids the curse of dimensionality typical of IO models with a large number of products. We demonstrate that our inverse mapping reduces to a univariate root-finding problem that remains tractable regardless of the number of assets.

**Outline.** The remainder of the paper is organized as follows. Section 2 presents the theoretical framework and derives the structural asset demand model. Section 3 discusses identification and estimation. Section 4 assesses our model relative to the MNL in capturing institutional investors' equity demand. Section 5 concludes.

## 2. Model

Throughout the paper, let the indices $i = 1, \ldots, I$, $n = 1, \ldots, N$, and $t$ represent investors, assets, and time (e.g., quarter), respectively. Denote by $\mathcal{N}_{i,t} \subseteq \{1, \ldots, N\}$ investor $i$'s investment universe at time $t$ determined by her investment mandate. We refer to the $N_{i,t}$ assets in investor $i$'s portfolio at time $t$ as *inside assets*; all wealth outside of the inside assets are referred to as *outside asset*. We use lowercase letters to denote the natural logarithms of corresponding uppercase variables. For example, let $q_{i,t}(n) = \ln(Q_{i,t}(n))$ represent the natural logarithm of investor $i$'s holdings of asset $n$ at time $t$. The corresponding vector forms are denoted in bold, i.e., $\mathbf{q}_{i,t} = \ln(\mathbf{Q}_{i,t})$.

Investor $i$ at time $t$ allocates her wealth $A_{i,t}$ across inside assets and the outside asset. Each inside asset $n$ has a gross return $R_{t+1}(n)$ from time $t$ to time $t+1$. The outside asset's gross return from time $t$ to $t+1$ is $R_{t+1}(0)$. The investor puts $\mathbf{w}_{i,t} = (w_{i,t}(n))_{n \in \mathcal{N}_{i,t}}$ of her wealth into the inside assets. This leads to portfolio returns $\mathbf{w}'_{i,t}\mathbf{R}_{t+1} + (1 - \mathbf{w}'_{i,t}\mathbf{1})R_{t+1}(0) = R_{t+1}(0) + \mathbf{w}'_{i,t}(\mathbf{R}_{t+1} - R_{t+1}(0)\mathbf{1})$. The wealth is then governed by the following intertemporal budget constraint:

$$A_{i,t+1} = A_{i,t}\left(R_{t+1}(0) + \mathbf{w}'_{i,t}(\mathbf{R}_{t+1} - R_{t+1}(0)\mathbf{1})\right). \tag{1}$$

Furthermore, the investor faces short-sale constraints:

$$\mathbf{w}_{i,t} \geq \mathbf{0}, \quad \mathbf{1}'\mathbf{w}_{i,t} \leq 1. \tag{2}$$

The investor chooses asset weights $\mathbf{w}_{i,t}$ at each time to maximize her expected log utility over her terminal wealth at time $T$ subject to constraints (2):[2]

$$\max_{\mathbf{w}_{i,t}} \mathbb{E}_{i,t}(\ln(A_{i,T})) \quad \text{s.t.} \quad \begin{cases} A_{i,t+1} = A_{i,t}\left(R_{t+1}(0) + \mathbf{w}'_{i,t}(\mathbf{R}_{t+1} - R_{t+1}(0)\mathbf{1})\right), \\ \mathbf{w}_{i,t} \geq \mathbf{0}, \quad \mathbf{1}'\mathbf{w}_{i,t} \leq 1. \end{cases} \tag{3}$$

As shown in Lemma 1 and Appendix A in KY, the mean-variance approximate solution to (3) (for long positions) can be written as

$$\mathbf{w}_{i,t} \approx \Sigma_{i,t}^{-1}(\mu_{i,t} - \lambda_{i,t}\mathbf{1}), \tag{4}$$

where $\lambda_{i,t}$ is the Lagrange multiplier for the short-sale constraint $\mathbf{1}'\mathbf{w}_{i,t} \leq 1$ in (2) (see (A5) in Appendix A of KY for its formula).

Investor $i$ at time $t$ forms expectations about an asset $n$'s returns and risks based on its characteristics. Some characteristics are observed to the researcher; these

[2] As pointed out in footnote 3 of Koijen and Yogo (2019), the log utility form (or more generally the isoelastic utility shown by Samuelson (1969)) reduces the multi-period portfolio optimization to a one-period problem in which the optimal portfolio decision is independent of wealth at each $t$ and hedging demand is absent. Relaxing the log utility specification may introduce dynamics in asset demand: a demand shock of today affects current asset prices and consequently next period's asset weights via tomorrow's wealth.

characteristics include log market equity, $\mathbf{me}_t(n)$, as well as $K-1$ other ones, denoted by $x_{k,t}(n)$, $k = 1, ..., K-1$. Examples include log book equity, dividend to book equity, operating profits to book equity, log growth of assets, and market beta. Besides, some asset characteristics are unobserved to the researcher. We summarize the $i, t$-specific effect of these characteristics of asset $n$ into latent demand term $\varepsilon_{i,t}(n)$ for investor $i$ at time $t$. Let $\mathbf{x}_{i,t}(n)$ denote the stacked vector of non-constant characteristics and $\mathbf{y}_{i,t}(n)$ their polynomials in $i$'s portfolio at time $t$:

$$\mathbf{x}_{i,t}(n) = \begin{bmatrix} \mathbf{me}_t(n) \\ x_{1,t}(n) \\ \vdots \\ x_{K-1,t}(n) \\ \ln\left(\varepsilon_{i,t}(n)\right) \end{bmatrix} \quad \text{and} \quad \mathbf{y}_{i,t}(n) = \begin{bmatrix} \mathbf{x}_{i,t}(n) \\ \operatorname{vec}\left(\mathbf{x}_{i,t}(n)\mathbf{x}_{i,t}(n)'\right) \\ \vdots \end{bmatrix}. \tag{5}$$

Following KY, we adopt a one-factor structure in investor's log excess returns, and the expected excess returns and factor loadings are affine in asset characteristics.

**Assumption 1 (Factor Structure of Returns and Characteristic Dependence).** *Investor $i$'s log excess asset returns exhibit a one-factor structure:*

$$\mathbf{r}_{i,t+1} - r_{i,t+1}(0)\mathbf{1} = \left(\alpha_i(n)\right)_{i=1}^{N_{i,t}} + \Gamma_{i,t} f_{t+1} + \epsilon_{i,t}, \tag{6}$$

*where $\mathbb{V}ar_{i,t}\left(f_{t+1}\right) = 1$, $\mathbb{E}_{i,t}\left(\epsilon_{i,t}\right) = 0$, $\mathbb{V}ar_{i,t}\left(\epsilon_{i,t}\right) = \gamma_{i,t}\mathbf{I}$, and $\mathbb{C}ov_{i,t}\left(f_{t+1}, \epsilon_{i,t}\right) = 0$. Moreover, the conditional expected log excess returns and factor loadings are:*

$$\mu_{i,t}(n) = \phi_{i,t} + \mathbf{y}_{i,t}(n)'\mathbf{\Phi}_{i,t}, \tag{7}$$

$$\Gamma_{i,t}(n) = \psi_{i,t} + \mathbf{y}_{i,t}(n)'\mathbf{\Psi}_{i,t}, \tag{8}$$

*where $\mathbf{\Phi}_{i,t}$ and $\mathbf{\Psi}_{i,t}$ are vectors and $\phi_{i,t}$ and $\psi_{i,t}$ are constant scalars across assets.*

KY show that the mean-variance solution (4) and Assumption 1 imply the following characteristics-based asset demand (see their Proposition 1):

$$w_{i,t}(n) = \pi_{i,t} + \mathbf{y}'_{i,t}(n)\mathbf{\Pi}_{i,t}, \tag{9}$$

where

$$\mathbf{\Pi}_{i,t} = (\mathbf{\Phi}_{i,t} - \mathbf{\Psi}_{i,t}\kappa_{i,t})/\gamma_{i,t}, \tag{10}$$

$$\pi_{i,t} = (\phi_{i,t} - \lambda_{i,t} - \psi_{i,t}\kappa_{i,t})/\gamma_{i,t}, \tag{11}$$

and $\kappa_{i,t} = \frac{\Gamma'_{i,t}(\mu_{i,t} - \lambda_{i,t}\mathbf{1})}{\Gamma'_{i,t}\Gamma_{i,t} + \gamma_{i,t}}$ reflects the portfolio exposure to the systematic risk.

In the next result, we derive the characteristics-based asset demand from (9) and characterize the if-and-only-if condition for the MNL asset demand in KY. The proof is in Supplemental Appendix A.1.

**Proposition 1 (If-and-Only-If Condition for MNL Asset Demand).**

- *Suppose that Assumption 1 holds.*
- *Suppose that the Lagrange multiplier* $\lambda_{i,t} = 0$ *(or equivalently* $w_{i,t}(0) > 0$*). Then,*

$$\begin{aligned}\frac{w_{i,t}(n)}{w_{i,t}(0)} &= \exp\left(\beta_{K,i,t}\right)\left(1 + \mathbf{y}'_{i,t}\frac{\mathbf{\Phi}_{i,t}}{\phi_{i,t}} + \mathbf{y}'_{i,t}\left(\frac{\mathbf{\Phi}_{i,t}}{\phi_{i,t}} - \frac{\mathbf{\Psi}_{i,t}}{\psi_{i,t}}\right)\frac{\tilde{\kappa}_{i,t}}{1-\tilde{\kappa}_{i,t}}\right) \\ &= \exp\left(\beta_{K,i,t}\right)\left(\frac{\mu_{i,t}(n)}{\phi_{i,t}} + \left(\frac{\Gamma_{i,t}(n)}{\psi_{i,t}} - \frac{\mu_{i,t}(n)}{\phi_{i,t}}\right)\frac{\tilde{\kappa}_{i,t}}{1-\tilde{\kappa}_{i,t}}\right),\end{aligned} \tag{12}$$

  *where* $\beta_{K,i,t} = -\ln\left(\frac{w_{i,t}(0)}{\pi_{i,t}}\right)$ *and* $\tilde{\kappa}_{i,t} = \left(\sum_{n\in\mathcal{N}_{i,t}}\frac{\mu_{i,t}}{\phi_{i,t}}\frac{\Gamma_{i,t}}{\psi_{i,t}}\right)\Big/\left(\frac{\gamma_{i,t}}{\psi^2_{i,t}} + \sum_{n\in\mathcal{N}_{i,t}}\left(\frac{\Gamma_{i,t}}{\psi_{i,t}}\right)^2\right)$.

- *For any* $n \in \mathcal{N}_{i,t}$*:* $\mu_{i,t}(n) \approx \left(\alpha_i(n) + \frac{\gamma_{i,t}}{2}\right) + \Gamma_{i,t}(n)\,\mathbb{E}_{i,t}\left(f_{t+1}\right) + \frac{1}{2}\Gamma^2_{i,t}(n)$.
- *Suppose that* $\lambda_{i,t} = 0$ *and* $\left[\mathbf{y}'_{i,t}(1);\ldots;\mathbf{y}'_{i,t}(N_{i,t})\right]$ *is of full column rank. Then,* (12) *has the following MNL form*

$$w_{i,t}(n) = \frac{\exp\left(\ln\left(\frac{\mu_{i,t}(n)}{\phi_{i,t}}\right)\right)}{\exp\left(-\beta_{K,i,t}\right) + \sum_{m\in\mathcal{N}_{i,t}}\exp\left(\ln\left(\frac{\mu_{i,t}(m)}{\phi_{i,t}}\right)\right)} \tag{13}$$

  *if and only if* $\frac{\mathbf{\Phi}_{i,t}}{\phi_{i,t}} = \frac{\mathbf{\Psi}_{i,t}}{\psi_{i,t}}$.

The first result of Proposition 1 characterizes the demand system derived from Assumption 1 and the condition $\lambda_{i,t} = 0$, or equivalently $w_{i,t}(0) > 0$. The latter is motivated by the fact that realistic portfolios rarely comprise only the modeled subset of assets, e.g., a large institutional investor allocating her entire wealth to inside assets. Institutional investors often maintain positive outside wealth. For example, in the FactSet Ownership dataset used by Koijen et al. (2024), all investors allocate some weight to the outside asset - averaging $6\%$ - and approximately $85\%$ hold at least $1\%$. These outside positions may reflect liquidity needs, regulatory requirements, or unobserved investment opportunities.

The second result of Proposition 1 bridges the primitives of the factor structure (6) and the objects of interest in the estimation of (12), $\mu_{i,t}(n)$ and $\Gamma_{i,t}(n)$. The approximation originates from the fact that (12) is derived from the mean-variance approximation in (4). Built on this relationship, for each $(i,t)$, one can run a cross-sectional regression of estimated $\mu_{i,t}(n)$ on $\Gamma_{i,t}(n)$, $\Gamma^2_{i,t}(n)$, and an intercept, to back out investor's alpha for each asset (relative to the outside option) and its distribution across investors. This property of (12) helps us gain insights on the drivers of differentiated portfolios across investors. In Section 4.5, we implement this post-estimation regression and illustrate these insights on institutional investors.

The third results gives the if-and-only-if condition for MNL asset demand. Intuitively, the relationship $\mathbf{\Phi}_{i,t}/\phi_{i,t} = \mathbf{\Psi}_{i,t}/\psi_{i,t}$ implies that the risk-return trade-off, $\Gamma_{i,t}(n)/\mu_{i,t}(n) = \psi_{i,t}/\phi_{i,t}$, does not depend on asset characteristics and is an investor-time specific constant.[3] This homogeneity in risk-return trade-off means that only expected excess returns affect relative demand in (12), leading to the $\frac{w_{i,t}(n)}{w_{i,t}(m)} = \frac{\mu_{i,t}(n)}{\mu_{i,t}(m)}$ for any $n \neq m$, i.e., the independence of irrelevant alternatives (IIA). The full column rank is a regularity condition. It requires sufficient cross-asset variation in the polynomial vector $\mathbf{y}_{i,t}(n)$.

Building on (12), we derive an asset demand model with a log-linear specification on $\mu_{i,t}(n)$ and $\Gamma_{i,t}(n)$. Without loss of generality, we normalize $\phi_{i,t} = \psi_{i,t} = 1$.[4]

**Proposition 2 (Portfolio Weights).** *Suppose that both $\mu_{i,t}$ and $\Gamma_{i,t}$ have a log-linear form. For $n \in \mathcal{N}_{i,t}$ and $\mathbf{x}_t(n) = (\mathbf{me}_t(n), x_{1,t}(n), ..., x_{K-1,t}(n))'$,*

$$\mu_{i,t}(n) = \exp\left(\, \mathbf{x}_t'(n)\beta_{i,t} \,\right)\, \varepsilon_{i,t}(n), \tag{14}$$

$$\Gamma_{i,t}(n) = \exp\left(\, \mathbf{x}_t'(n)\left(\beta_{i,t} + \Delta_{i,t}\right) \right)\, \varepsilon_{i,t}(n). \tag{15}$$

[3] To see this, note that $\frac{\Gamma_{i,t}(n)}{\mu_{i,t}(n)} = \frac{\psi_{i,t}+\mathbf{y}_{i,t}(n)'\mathbf{\Psi}_{i,t}}{\phi_{i,t}+\mathbf{y}_{i,t}(n)'\mathbf{\Phi}_{i,t}} = \frac{\psi_{i,t}(1+\mathbf{y}_{i,t}(n)'\mathbf{\Psi}_{i,t}/\psi_{i,t})}{\phi_{i,t}(1+\mathbf{y}_{i,t}(n)'\mathbf{\Phi}_{i,t}/\phi_{i,t})} = \frac{\psi_{i,t}}{\phi_{i,t}}$.

[4] This is because $(\mu_{i,t}, \Gamma_{i,t}, \phi_{i,t}, \psi_{i,t}, \gamma_{i,t})$ and $(\mu_{i,t}/\phi_{i,t}, \Gamma_{i,t}/\psi_{i,t}, 1, 1, \gamma_{i,t}/\psi^2_{i,t})$ deliver the same asset demand in (12).

*Then, portfolio weights* (12) *become:*

$$\frac{w_{i,t}(n)}{w_{i,t}(0)} = \exp\left(\mathbf{x}_t'(n)\beta_{i,t} + \beta_{K,i,t} + \ln\left(1 + (1 - \exp(\mathbf{x}_t'(n)\Delta_{i,t}))\frac{\kappa_{i,t}}{1-\kappa_{i,t}}\right)\right)\varepsilon_{i,t}(n), \quad (16)$$

*where*

$$\kappa_{i,t} = \left(\sum_{n=1}^{N}\mu_{i,t}(n)\Gamma_{i,t}(n)\right) \Big/ \left(\gamma_{i,t} + \sum_{n=1}^{N}\Gamma_{i,t}^2(n)\right). \quad (17)$$

**Interpretation of $\Delta_{i,t}$.** The log-linear specifications (14) and (15) lead to empirical tractability without losing the economic content when we extend asset demand beyond the MNL asset demand (13). Note that

$$\ln\left(\frac{\Gamma_{i,t}(n)}{\mu_{i,t}(n)}\right)_{n=1}^{N_{i,t}} = (\mathbf{x}_t'(1); \ldots; \mathbf{x}_t'(N_{i,t}))\Delta_{i,t}, \quad (18)$$

The new parameters $\Delta_{i,t}$ in (16) determine the dependence of investor's perceived risk-return trade-off on asset characteristics and its heterogeneity across assets. Along the lines of Proposition 1, our model (16) reduces to the MNL in KY when $\Delta_{i,t} = \mathbf{0}$, i.e., homogeneous risk-return trade-off across assets. When one of the components in $\Delta_{i,t}$ is non-zero, say $\Delta_{\mathbf{me},i,t} \neq 0$, increasing the market equity of asset $n$ increases by $1\%$, will lead to $\Delta_{\mathbf{me},i,t}\%$ increase in its risk-return trade-off.

This interpretation of $\Delta_{i,t}$ is useful for assessing (in a first-order fashion) MNL's adequacy for investors' asset demand. To see this point, we develop the first-order Taylor approximation of $\ln\Gamma_{i,t}(n)$ and $\ln\mu_{i,t}(n)$ around their market means:

$$\ln(\Gamma_{i,t}(n)) \approx \ln(\mathbb{E}_t(\Gamma_{i,t}(n))) + \frac{\Gamma_{i,t}(n) - \mathbb{E}_t(\Gamma_{i,t}(n))}{\mathbb{E}_t(\Gamma_{i,t}(n))},$$
$$\ln(\mu_{i,t}(n)) \approx \ln(\mathbb{E}_t(\mu_{i,t}(n))) + \frac{\mu_{i,t}(n) - \mathbb{E}_t(\mu_{i,t}(n))}{\mathbb{E}_t(\mu_{i,t}(n))},$$

where $\mathbb{E}_t$ refers to the expectation over investors at time $t$. Plugging the Taylor expansions of $\ln(\Gamma_{i,t}(n))$ and $\ln(\mu_{i,t}(n))$ into (18) and applying $\mathbb{E}_t$ on both sides:

$$\begin{aligned} &\ln\left(\frac{\mathbb{E}_t(\Gamma_{i,t}(n))}{\mathbb{E}_t(\mu_{i,t}(n))}\right) \approx x_t'(n)\,\mathbb{E}_t(\Delta_{i,t}), \quad \forall n \in \mathcal{N}_t = \cup_i \mathcal{N}_{i,t} \\ &\implies \mathbb{E}_t(\Delta_{i,t}) \approx (\mathbf{X}_t'\mathbf{X}_t)^{-1}\mathbf{X}_t' \ln\left(\frac{\mathbb{E}_t(\Gamma_{i,t}(n))}{\mathbb{E}_t(\mu_{i,t}(n))}\right)_{n\in\mathcal{N}_t}, \end{aligned} \quad (19)$$

where $\mathbf{X}_t = (\mathbf{x}_t(1), ..., \mathbf{x}_t(N_t))$. This relationship relates the market average of $\Delta_{i,t}$ to the classic CAPM variables. In fact, by assuming that investors' perceptions of asset factor loadings and expected log excess return are correct on average, $\mathbb{E}_t\left(\Gamma_{i,t}(n)\right)$ and $\mathbb{E}_t\left(\mu_{i,t}(n)\right)$ are equal to the asset-time specific factor loadings and expected log excess return in the classic CAPM version of (6) and can be readily estimated without estimating asset demand (16). If the estimates of $\mathbb{E}_t\left(\Delta_{i,t}\right)$ from (19) are significantly different from zero, then at least a non-negligible portion of investors have non-zero $\Delta_{i,t}$, suggesting potential deviation from the MNL. In Section 4.2, we will use (19) to assess such deviations and their directions (i.e., along which asset characteristics such deviations occur) in the real data.

**Price Elasticities and $\Delta_{i,t}$.** When $\Delta_{i,t} = \mathbf{0}$, only the parameters in expected excess returns are relevant in the resulting price elasticities. When $\Delta_{i,t} \neq \mathbf{0}$, asset $n$'s relative demand depends additionally and nonlinearly on its asset characteristics $\mathbf{x}_t(n)$ via $\Delta_{i,t}$ as well as other assets' characteristics via $\kappa_{i,t}$. We discuss two implications of this new feature for the price elasticities in (16): (1). how it relaxes the restricted uniform asset substitution pattern in the MNL and (2). how it enables to generate large price elasticities and small price effects. We first give the price elasticities formula in the next result. The proof is in Supplemental Appendix A.3.

**Corollary 1 (Price Elasticity).** *Let $\mathbf{q}_{i,t} = \ln\left(A_{i,t}\mathbf{w}_{i,t}\right) - \mathbf{p}_t$ be the vector of log shares held by investor $i$. Then, from (16), we derive the price elasticity of demand for an individual investor $i$ as*

$$-\frac{\partial \mathbf{q}_{i,t}(n)}{\partial \mathbf{p}_t(j)} = \underbrace{\mathbf{1}\{n=j\} - \beta_{\mathbf{me},i,t}\left[\mathbf{1}\{n=j\} - w_{i,t}(j)\right]}_{\text{Diversion Effect: MNL}} \tag{20}$$

$$+ \underbrace{\left[\mathbf{1}\{n=j\} - w_{i,t}(j)\right] \frac{\exp\left(\mathbf{x}_t'(j)\Delta_{i,t}\right)\kappa_{i,t}}{1-\exp\left(\mathbf{x}_t'(j)\Delta_{i,t}\right)\kappa_{i,t}} \Delta_{\mathbf{me},i,t}}_{\text{Intra-Portfolio Effect: via asset } j\text{'s risk-return trade-off}} \tag{21}$$

$$\underbrace{-\frac{\partial \kappa_{i,t}}{\partial \mathbf{p}_t(j)} \sum_{r=0}^{N_{i,t}} w_{i,t}(r) \frac{\exp\left(\mathbf{x}_t'(r)\Delta_{i,t}\right) - \exp\left(\mathbf{x}_t'(n)\Delta_{i,t}\right)}{\left(1-\kappa_{i,t}\exp\left(\mathbf{x}_t'(r)\Delta_{i,t}\right)\right)\left(1-\kappa_{i,t}\exp\left(\mathbf{x}_t'(n)\Delta_{i,t}\right)\right)}}_{\text{Portfolio-Level Effect: via changes in the portfolio exposure to systematic risk}} \tag{22}$$

The diversion effect in (20) originates from the expected excess return, a channel already captured by the MNL demand. In this component, the own-price part of asset $n$ is equal to $1 - \beta_{\mathbf{me},i,t}(1 - w_{i,t}(n))$. Given the observed asset weights, the magnitudes of own-price part is determined by $\beta_{\mathbf{me},i,t}$. Besides, the cross-price part in (20) is equal to $\beta_{\mathbf{me},i,t}w_{i,t}(j)$ for $j \neq n$. Its sign (i.e., asset complementarity, substitutability, or independence) and magnitude are determined by $\beta_{\mathbf{me},i,t}$: $\beta_{\mathbf{me},i,t}w_{i,t}(j) < 0$ ($n$ and $j$ are substitutes) if and only if $\beta_{\mathbf{me},i,t} < 0$. We summarize the substitution restrictions in the MNL demand in the following corollary.

**Corollary 2 (Substitution pattern in the MNL).** *Suppose that $\Delta_{i,t} = \mathbf{0}$. Then, all assets are uniformly substitutes, complements, or independent if and only if $\beta_{\mathbf{me},i,t} < 0$, $\beta_{\mathbf{me},i,t} > 0$, or $\beta_{\mathbf{me},i,t} = 0$.*

Corollary 2 implies that an MNL investor perceives all her assets as substitutes, complements, or independent, depending on her $\beta_{\mathbf{me},i,t}$. This uniform substitution rules out the possibility that the investor sees some assets as substitutes (e.g., two AI stocks) and some others as complements (blue-chip and growth stocks).

In contrast, when $\Delta_{i,t} \neq \mathbf{0}$, we have two new channels in (16) that determine price elasticities: *intra-portfolio effects* (21) and *portfolio-level effects* (22). The intra-portfolio component (21) is the effect of market equity via the risk-return trade-off $\exp(\mathbf{x}_t'(j)\Delta_{i,t})$. This adjustment captures how asset-specific characteristics impact the investor's portfolio rebalancing. For instance, if $\Delta_{\mathbf{me},i,t} > 0$ and $\exp(\mathbf{x}_t'(j)\Delta_{i,t}) < \kappa_{i,t}^{-1}$, the intra-portfolio effect (21) is then positive, increasing the own-price elasticity relative to the MNL (20). Intuitively, this corresponds to a scenario where an asset's risk rises faster than its return when its price increases, prompting the investor to cut back more aggressively. Hence, this component ensures that elasticities accurately reflect the nuanced contributions of asset characteristics to both returns and risk-return trade-offs. The portfolio-level component (22) captures the effect due to the change in portfolio-level risk exposure, $\frac{\partial \kappa_{i,t}}{\partial \mathbf{p}_t(j)}$. It highlights the interconnected nature of the optimal portfolio choice that adjustments to one asset ripple through the entire portfolio and could influence the demand for seemingly unrelated assets. When $n \neq j$, different from (20) and (21), (22) depends on $n$'s as well as other assets' characteristics.

With these mechanics in place, we next examine how our model (16) generates large own-price elasticities and small price effects. First, it is worth noting that the MNL ($\Delta_{i,t} = \mathbf{0}$) can already generate arbitrarily large own-price elasticities and small price effects by having sufficiently negative $\beta_{\mathbf{me},i,t}$. To shed light on this point, suppose that all price coefficients $\beta_{\mathbf{me},i,t}$ are of order $\beta$ and that there is a shock $\delta$ that shifts asset prices according to the market clearing conditions:

$$\mathbf{me}_t(n) = \ln\left(\sum_{i=1}^{I} A_{i,t} w_{i,t}(n)\right) \qquad \forall n = 1, ..., N_{i,t}. \tag{23}$$

Then, the MNL demand system implies that the price effects are approximately bounded by the order $\frac{1}{1-\beta}$ when $\beta \ll 0$ (see Supplemental Appendix A.4):

$$\left|\frac{\partial \mathbf{p}_t(n)}{\partial \delta}\right| = O\left(\frac{1}{1-\beta}\right), \tag{24}$$

i.e., when $\beta_{\mathbf{me},i,t} \ll 0$, the MNL can deliver large price elasticities (see (20)) and consequently small price effects. As a comparison, a back-of-envelop calculation indicates that $\beta \approx -6,172$ could generate the magnitude of the price change in the calibrated CAPM of Petajisto (2009) ($\approx 0.16$bp) due to a $-10\%$ supply shock.[5]

More in general, whether or not $\Delta_{i,t} = \mathbf{0}$, asset demand system (16) has the ability of generating arbitrarily large price elasticities and small price effects. Differently from the MNL, a model (16) with $\Delta_{i,t} \neq \mathbf{0}$ does not necessarily require $\beta_{\mathbf{me},i,t} \ll 0$ for this purpose. Figure 1 illustrates this point via a simple example. We consider (16) with identical investors (the same demand parameters) and symmetric assets (the same asset characteristics), boiling the relationship between $(\beta_{\mathbf{me}}, \Delta_{\mathbf{me}})$ and the own-price elasticity $\eta^{\text{own}}$ down to a 2-D surface. The red one represents the relationship in the MNL (corresponding to $\Delta_{\mathbf{me}} = 0$) and the blue one the relationship without restricting $\Delta_{\mathbf{me}} = 0$. For instance, at $\beta_{\mathbf{me}} = -30$, the MNL predicts $\eta^{\text{own}} \approx 30$. In contrast, when $\Delta_{\mathbf{me}} \approx 0.4$, $\beta_{\mathbf{me}} \approx -5$ can already deliver an elasticity of similar magnitude.

To summarize, the new parameters $\Delta_{i,t}$ characterize the dependence of investor's perceived risk-return trade-offs on asset characteristics. In the MNL asset demand, $\Delta_{i,t}$ are zero and the risk-return trade-off is uniform across assets. This feature leads to the IIA substitution pattern. Importantly, investor perceives all

[5]Concretely, in Section II.A of Petajisto (2009)'s calibrated CAPM, a $-10\%$ supply shock generates $0.00162\%$ increase in asset prices. (24) implies an approximate magnitude of $\frac{1}{1-\beta}\%$ price increase due to $1\%$ change in the shock ($\delta$ is the log of the shock). Then, $\beta \approx 1 - 10/0.00162 = -6,172$.

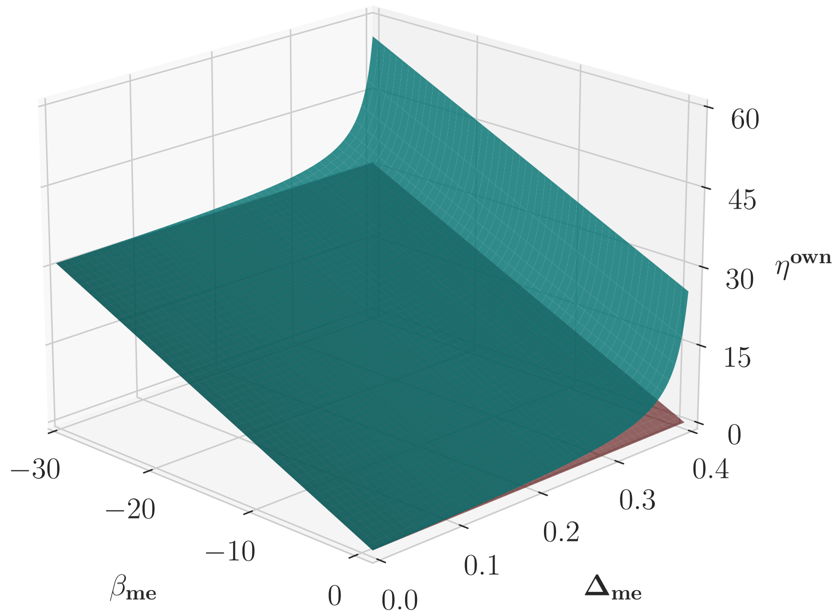


**Figure 1: Asset Demand Systems' Ability to Generate Large Elasticities**

her assets to be either substitutes, complements, or independent (Corollary 2). Unlike the MNL in which $\beta_{\mathbf{me},i,t}$ governs both own- and cross-price elasticities, our model (16) provides additional flexibility: $\Delta_{i,t}$ determine the diagonals and the off-diagonals of $\frac{\partial \mathbf{q}_{i,t}}{\partial \mathbf{p}_t}$ jointly with $\beta_{\mathbf{me},i,t}$, thereby relaxing the uniform substitution in the MNL and capturing investors who may see some assets as substitutes and others as complements. Furthermore, the MNL has the ability of generating large price elasticities and small price effects by having sufficiently negative $\beta_{\mathbf{me},i,t}$. Our model (16) with $\Delta_{i,t} \neq \mathbf{0}$ has the same property, but can also deliver large price elasticities via a much less negative $\beta_{\mathbf{me},i,t}$ and non-zero $\Delta_{\mathbf{me},i,t}$.

## 3. Identification and Estimation

In this section, we discuss our two-step inverse-style strategy for identifying and estimating the structural parameters $(\beta_{i,t}, \Delta_{i,t}, \gamma_{i,t}, \beta_{K,i,t})$ in our model (16). We assume that the researcher observes the portfolio weights $(w_{i,t}(n))_{i=1}^{N_{i,t}}$ and characteristics $(\mathbf{x}_t(n))_{n=1}^{N_{i,t}}$. While the discussion focuses on investor-time specific parameters, the arguments naturally extend to the case of pooled parameters across investors or time.

To give some intuition, we start with formulating this strategy in the MNL setting. In this case, the relative weights are simplified to:

$$\frac{w_{i,t}(n)}{w_{i,t}(0)} = \exp\left( \mathbf{x}_t'(n)\beta_{i,t} + \beta_{K,i,t} \right) \varepsilon_{i,t}(n).$$

Given $(\beta_{i,t}, \beta_{K,i,t})$, one can invert $\frac{w_{i,t}(n)}{w_{i,t}(0)}$ to $\varepsilon_{i,t}(n)$ by $\varepsilon_{i,t}(n) = \frac{w_{i,t}(n)}{w_{i,t}(0)} \exp\{-\mathbf{x}_t'(n)\beta_{i,t} - \beta_{K,i,t}\}$. Suppose that asset characteristics $(x_{1,t}(n), ..., x_{K-1,t}(n))$ are orthogonal to $\varepsilon_{i,t}(n)$, whereas, $\mathbf{me}_t(n)$ is potentially correlated to $\varepsilon_{i,t}(n)$ but there is a valid instrument $z_{i,t}(n)$. We can then construct moment conditions

$$\mathbb{E}_t \left( \frac{w_{i,t}(n)}{w_{i,t}(0)} \exp\left( -\mathbf{x}_t'(n)\beta_{i,t} - \beta_{K,i,t} \right) \middle| z_{i,t}, (x_{k,t}(n))_{k\geq 1} \right) = 1, \tag{25}$$

that identify $(\beta_{i,t}, \beta_{K,i,t})$ and estimate them via a GMM procedure.

When $\Delta_{i,t} \neq \mathbf{0}$, the inverse from $\left(\frac{w_{i,t}(n)}{w_{i,t}(0)}\right)_{n=1}^{N_{i,t}}$ to the vector of latent demand $(\varepsilon_{i,t}(n))_{n=1}^{N_{i,t}}$ is no longer straightforward, because the right-hand side of (16) involves also $\varepsilon_{i,t}(m)$ with $m \neq n$ via $\kappa_{i,t}$, leading to a simultaneous system in $(\varepsilon_{i,t}(n))_{n=1}^{N_{i,t}}$. Moreover, it becomes less clear if and how the moment conditions along the lines of (25) separably identify $\beta_{i,t}$ in returns and $\Delta_{i,t}$ in risk-return trade-offs, both of which govern the elasticities.

Thus, using an approximate form of our model (16), we first shed light on the source for this separable identification. Taking $\kappa_{i,t}$ as fixed and developing the Taylor expansion of $\ln\left(\frac{w_{i,t}(n)}{w_{i,t}(0)}\right)$ around $\Delta_{i,t} = \mathbf{0}$, we obtain:

$$\ln\left(\frac{w_{i,t}(n)}{w_{i,t}(0)}\right) \approx \beta_{K,i,t} + \mathbf{x}_t'(n)\left(\beta_{i,t} - \frac{\kappa_{i,t}}{1-\kappa_{i,t}}\Delta_{i,t}\right) - \frac{\kappa_{i,t}(1-2\kappa_{i,t})}{2(1-\kappa_{i,t}^2)}\Delta_{i,t}'(\mathbf{x}_t(n)\mathbf{x}_t'(n))\Delta_{i,t} + \ln\left(\varepsilon_{i,t}(n)\right).$$

First, $\beta_{i,t} - \frac{\kappa_{i,t}}{1-\kappa_{i,t}}\Delta_{i,t}$ is identified via exogenous variation in $\mathbf{x}_t(n)$. Such variation in $(x_{1,t}(n), ..., x_{K-1,t}(n))$ is guaranteed to be exogenous because of the orthogonality of these asset characteristics with respect to $\varepsilon_{i,t}(n)$. To have exogenous variation in

$\mathbf{me}_t(n)$, one can rely on the investment mandate based instrument in KY:

$$z_{i,t}^{\mathbf{me}}(n) = \ln\left(\sum_{j\neq i} A_{j,t}\frac{\mathbf{1}\{n\in\mathcal{N}_{j,t}\}}{1+N_{j,t}}\right), \quad \forall n = 1,\ldots,N.$$

These instruments are exogenous given predetermined, persistent constraints on investors' investment universes, such as sector-specific funds or index-tracking mandates, and the exogeneity of other investors' wealth.[6] Intuitively, $z_{i,t}^{\mathbf{me}}(n)$ can be seen as a measure of asset popularity. In fact, given $(i,t)$, the variation in $z_{i,t}^{\mathbf{me}}(n)$ across assets is mainly driven by the log frequency of asset $n$ in investor's wealth weighted investment universe.[7] Suppose that two assets $m$ and $n$ have similar observed non-price characteristics. If the frequency for asset $m$ is higher than asset $n$, then $m$ appears more often in investor's investment universe than $n$, leading to greater aggregate investment in $m$ than $n$, i.e., $\sum_{i=1}^{I} A_{i,t}w_{i,t}(m) > \sum_{i=1}^{I} A_{i,t}w_{i,t}(n)$. Thus, according to the market clearing condition (23), the price of $m$ must rise relative to $n$, i.e., a greater left-hand side of (23) for asset $m$, to clear the market for asset $m$. The more popular $n$, the more an investor has to pay per share, validating the first-stage predictive power of $z_{i,t}^{\mathbf{me}}(n)$ for $\mathbf{me}_t(n)$. From this perspective, the mandate-based instruments mimic the role of cost shifters in the identification of demand models in IO: greater asset popularity among investors mean higher cost of investing in this asset, and hence higher price per share.

Next, we turn to the identification of $\Delta_{i,t}$, which govern how risk-return trade-offs vary with asset characteristics. Since $\Delta_{i,t}$ enters the model through interactions like $\mathbf{x}_t(n)\mathbf{x}_t'(n)$, its identification can be achieved via exogenous variation in

[6] KY, Koijen et al. (2024) and Bretscher et al. (2025) show investors' holdings exhibit high persistence over time, with $95\%$ their investment universes remaining unchanged over 11-quarters. When the inclusion in investor's mandate, $\mathbf{1}\{n\in\mathcal{N}_{j,t}\}$, is positively correlated to latent demand $\varepsilon_{j,t}(n)$ (i.e., $n$ is more likely to appear in $j$'s mandate if the latent demand is higher), the orthogonality of $z_{i,t}^{\mathbf{me}}(n)$ with respect to $\varepsilon_{i,t}(n)$ would require the independence of $\varepsilon_{i,t}(n)$ among all $i$ at time $t$. Besides, Fuchs et al. (2024) highlight that the exclusion restriction for such instruments might be violated under price spillovers.

[7] Note that $z_{i,t}^{\mathbf{me}}(n) = \ln\left(1 - \frac{\frac{A_{i,t}\mathbf{1}\{n\in\mathcal{N}_{j,t}\}}{1+N_{i,t}}}{\sum_{j=1}^{N} A_{j,t}\frac{\mathbf{1}\{n\in\mathcal{N}_{j,t}\}}{1+N_{i,t}}}\right) + \ln\left(\sum_{j=1}^{N} A_{j,t}\frac{\mathbf{1}\{n\in\mathcal{N}_{j,t}\}}{1+N_{i,t}}\right)$. When $\frac{A_{i,t}\mathbf{1}\{n\in\mathcal{N}_{j,t}\}}{1+N_{i,t}}$ is small relative to $\sum_{j=1}^{N} A_{j,t}\frac{\mathbf{1}\{n\in\mathcal{N}_{j,t}\}}{1+N_{i,t}}$, the first term will be approximately $\frac{\frac{A_{i,t}\mathbf{1}\{n\in\mathcal{N}_{j,t}\}}{1+N_{i,t}}}{\sum_{j=1}^{N} A_{j,t}\frac{\mathbf{1}\{n\in\mathcal{N}_{j,t}\}}{1+N_{i,t}}}$ that is dominated by $\ln\left(\sum_{j=1}^{N} A_{j,t}\frac{\mathbf{1}\{n\in\mathcal{N}_{j,t}\}}{1+N_{i,t}}\right)$.

these higher-order terms of $n$'s characteristics.[8] One can use quadratic-form instruments, $x_{k,t}(n)x_{l,t}(n)$ with $k, l \geq 1$, for $\Delta_{i,t}$. Their validity is ensured by the orthogonality of $x_{k,t}(n)$ for $k \geq 1$. Along the lines of this reasoning, $z^{\mathbf{me}}_{i,t}(n)x_{k,t}(n)$ with $k \geq 1$ are also valid instruments for $\Delta_{i,t}$ because of the validity of $z^{\mathbf{me}}_{i,t}(n)$. The intuition of identification via higher-order terms of exogenous characteristics joins the arguments for identifying the distribution of random coefficients in BLP-type demand models (e.g. Gandhi and Houde, 2019; Wang, 2023, 2024); risk-return trade-off parameters $\Delta_{i,t}$ in (16) mimic the *nonlinear* parameters in BLP.

We formalize our two-step strategy that relies on a key reformulation of our demand model (16). The proof is in Appendix A.5.

**Proposition 3 (Demand Inverse).** *The asset demand system* (16) *can be rewritten for all* $n = 1, \ldots, N_{i,t}$*:*

$$\frac{w_{i,t}(n)}{w_{i,t}(0)} = \exp\left(\, \mathbf{x}'_t(n)\beta_{i,t} + \beta_{K,i,t}\,\right)\left(1 + (1 - \exp\left(\, \mathbf{x}'_t(n)\Delta_{i,t}\,\right))G_{i,t}\right)\varepsilon_{i,t}, \tag{26}$$

*where* $G = G_{i,t}$ *satisfies the following equation:*

$$\tilde{\gamma}_{i,t}G = \sum_{m=1}^{N_{i,t}} \frac{\left(\dfrac{w_{i,t}(m)}{w_{i,t}(0)}\right)^2 \exp\left(\, \mathbf{x}'_{i,t}(m)\Delta_{i,t}\,\right)}{1 + (1 - \exp\left(\, \mathbf{x}'_t(m)\Delta_{i,t}\,\right))G} \tag{27}$$

*with* $\tilde{\gamma}_{i,t} = \gamma_{i,t}\exp\left(\, \beta_{K,i,t}\,\right)$*.*

Proposition 3 implies an important property of the inverse of the asset demand (16): as long as $G_{i,t}$ is solved from (27), one can implement the inverse by

$$\varepsilon_{i,t}(n) = \frac{\dfrac{w_{i,t}(n)}{w_{i,t}(0)} \exp\left(\, -\mathbf{x}'_t(n)\beta_{i,t} - \beta_{K,i,t}\,\right)}{1 + \left(1 - \exp\left(\, \mathbf{x}'_t(n)\Delta_{i,t}\,\right)\right) G_{i,t}}. \tag{28}$$

Consequently, the invertibility of asset demand (16) is equivalent to the uniqueness of the solution of (27). Because (27) is uni-dimensional, one can establish its invertibility by single-cross conditions which further guarantee the invertibility of asset demand (16). One set of such conditions is $\mathbf{x}'_t(n)\Delta_{i,t} \leq 0$ for $1 \leq n \leq N_{i,t}$. Under this condition, the right-hand side of (27) is non-increasing in $G$, leading to

[8] Once $\Delta_{i,t}$ is identified, $\beta_{i,t}$ is identified from previously recovered $\beta_{i,t} - \frac{\kappa_{i,t}}{1-\kappa_{i,t}}\Delta_{i,t}$ and $\Delta_{i,t}$.

a single crossing with the left-hand side when $G$ increases from $0$ to $+\infty$.

Since asset demand (16) is non-separable in $(\varepsilon_{i,t}(n))_{n=1}^{N_{i,t}}$, directly solving its inverse could be computationally intense especially when the number of assets is large.[9] This numerical hurdle is more pronounced if one estimates the parameters of interest for each investor-time combination. Proposition 3 shrinks a $N-$dimensional inverse in (16) to a uni-dimensional one in (27), an almost closed-form formula, largely alleviating the numerical challenge of demand estimation with many assets. We will discuss the implementation details in the next section.

We now turn to the second step and use moment conditions to identify the parameters of interest and construct the GMM estimator. The following assumption summarizes the required orthogonality conditions.

**Assumption 2 (Exogeneity of Asset Characteristics).** *There exists random variables $z_{i,t}(n)$ such that*

$$\mathbb{E}\left(\, \varepsilon_{i,t}(n) \mid z_{i,t}(n),\, x_{1,t}(n), \ldots, x_{K-1,t}(n)\,\right) = 1.$$

That non-price characteristics are exogenous is common in the demand system asset pricing literature that builds on KY. This assumption treats investors as atomistic, implying that their individual demand shocks do not shape aggregate conditions. However, as Kim (2025) highlights, correlated demand shocks driven by institutional rebalancing or procyclical risk-taking can introduce factor structures into latent demand. Such correlations may violate exogeneity.

Using the inverse established in (28) and Assumption 2, we now formulate the moment conditions of the true parameters $(\beta_{i,t}^0, \Delta_{i,t}^0, \tilde{\gamma}_{i,t}^0, \beta_{K,i,t}^0)$. From (27) we write $G_{i,t}$ as a function $G\left(\left(w_{i,t}^2(n)/w_{i,t}^2(0), \mathbf{x}_t(n)\right)_{n=1}^{N_{i,t}}; \Delta_{i,t}, \tilde{\gamma}_{i,t}\right)$. Then, for $n = 1, \ldots, N_{i,t}$,

$$\mathbb{E}_{i,t}\left( \frac{\frac{w_{i,t}(n)}{w_{i,t}(0)} \exp\left(-\mathbf{x}_t'(n)\beta_{i,t}^0 - \beta_{K,i,t}^0\right)}{1 + \left(1 - \exp\left(\mathbf{x}_t'(n)\Delta_{i,t}^0\right)\right) G\left(\left(\frac{w_{i,t}^2(n)}{w_{i,t}^2(0)}, \mathbf{x}_t(n)\right)_{n=1}^{N_{i,t}}; \Delta_{i,t}^0, \tilde{\gamma}_{i,t}^0\right)} \,\middle|\, z_{i,t}(n), (x_{k,t}(n))_{k=1}^{K-1} \right) = 1. \quad (29)$$

Due to its parametric nature, one can combine (29) and the classical arguments such as the rank conditions (e.g. Rothenberg, 1971) to identify $(\beta_{i,t}^0, \Delta_{i,t}^0, \tilde{\gamma}_{i,t}^0, \beta_{K,i,t}^0)$.

[9] For example, in the FactSet data, the Vanguard Group holds $\sim 3000$ assets between 2000-2022 per quarter.

These conditions could be achieved by having sufficient cross-asset variation in $z_{i,t}(n)$ and $(x_{k,t}(n))_{k=1}^{K-1}$. For the variation in linear terms of $z_{i,t}(n)$ and $(x_{k,t}(n))_{k=1}^{K-1}$, we have $K$ moments (one per asset characteristics) for the identification of K-dimensional $\beta_{i,t}^0$. As pointed out in our identification analysis, the variation in the quadratic terms $x_{k,t}(n)x_{r,t}(n)$ and $z_{i,t}(n)x_{k,t}$ with $k, r \geq 1$ is useful for the identification of (up to) K-dimensional $\Delta_{i,t}^0$ and $\tilde{\gamma}_{i,t}^0$, adding at least $K(K+1)/2+1$ moments and at most $K+1$ parameters. $\beta_{K,i,t}^0$ is identified from the residual demand (the total weight of inside assets relative to the outside one), adding one more moment. We have $K(K+1)/2+K+2$ moments versus at most $2K+2$ parameters: the number of moments exceeds that of parameters whenever there is at least one asset characteristic.

To simplify the exposition, we denote by $\mathbf{z}_{i,t}(n)$ the vector of instruments and exogenous variables. Given $\hat{\theta}_{i,t} = (\hat{\beta}_{i,t}, \hat{\Delta}_{i,t}, \hat{\tilde{\gamma}}_{i,t}, \hat{\beta}_{K,i,t}, \hat{G}_{i,t})$, the GMM estimator can be formulated as

$$\hat{\theta}_{i,t} = \arg\min_{\theta_{i,t}\in\Theta} \left\| \frac{1}{N_{i,t}} \sum_{n=1}^{N_{i,t}} \frac{\frac{w_{i,t}(n)}{w_{i,t}(0)} \exp\left(-\mathbf{x}_t'(n)\beta_{i,t} - \beta_{K,i,t}\right)}{1 + (1 - \exp\left(\mathbf{x}_t'(n)\Delta_{i,t}\right)) G_{i,t}} \mathbf{z}_{i,t}(n) \right\|_{W_{N_{i,t}}} \tag{30}$$

$$\text{(inverse constraint) subject to } \tilde{\gamma}_{i,t} G_{i,t} = \sum_{m=1}^{N_{i,t}} \frac{\left(\frac{w_{i,t}(m)}{w_{i,t}(0)}\right)^2 \exp\left(\mathbf{x}_{i,t}'(m)\Delta_{i,t}\right)}{1 + (1 - \exp\left(\mathbf{x}_t'(m)\Delta_{i,t}\right)) G_{i,t}},$$

where $\Theta$ is the domain of $\theta_{i,t}$, $W_{N_{i,t}}$ is a symmetric weighting matrix that converges to some $W > 0$, and $\|\cdot\|_{W_{N_{i,t}}}$ is defined as $\|v\|_{W_{N_{i,t}}} = v' W_{N_{i,t}} v$. The inverse constraint on the structural parameters and $G_{i,t}$ is uni-dimensional, reducing the dimensionality of the inverse of asset demand system from $N_{i,t}$ to 1. We can further simplify the constrained GMM (30) by a reparameterization. Note that the objective function (30) only involves $(\beta_{i,t}, \Delta_{i,t}, \beta_{K,i,t}, G_{i,t})$ but not $\tilde{\gamma}_{i,t}$. Moreover, one can straightforwardly back out $\tilde{\gamma}_{i,t}$ from the inverse constraint and $(\beta_{i,t}, \Delta_{i,t}, \beta_{K,i,t}, G_{i,t})$:

$$\tilde{\gamma}_{i,t} = \frac{1}{G_{i,t}} \sum_{m=1}^{N_{i,t}} \frac{\left(\frac{w_{i,t}(m)}{w_{i,t}(0)}\right)^2 \exp\left(\mathbf{x}_{i,t}'(m)\Delta_{i,t}\right)}{1 + (1 - \exp\left(\mathbf{x}_t'(m)\Delta_{i,t}\right)) G_{i,t}}.$$

Thus, given $\hat{\theta}_{-\tilde{\gamma},i,t} = (\hat{\beta}_{i,t}, \hat{\Delta}_{i,t}, \hat{\beta}_{K,i,t}, \hat{G}_{i,t})$, one can reparametrize (30):

$$\widehat{\theta}_{-\tilde{\gamma},i,t} = \arg \min_{\theta_{-\tilde{\gamma},i,t} \in \Theta_{-\tilde{\gamma}}} \left\| \frac{1}{N_{i,t}} \sum_{n=1}^{N_{i,t}} \frac{\frac{w_{i,t}(n)}{w_{i,t}(0)} \exp\left( -\mathbf{x}_t'(n)\beta_{i,t} - \beta_{K,i,t} \right)}{1 + \left(1 - \exp\left( \mathbf{x}_t'(n)\Delta_{i,t} \right)\right) G_{i,t}} \mathbf{z}_{i,t}(n) \right\|_{W_{N_{i,t}}}$$

$$\widehat{\tilde{\gamma}}_{i,t} = \frac{1}{\widehat{G}_{i,t}} \sum_{m=1}^{N_{i,t}} \frac{\left(\frac{w_{i,t}(m)}{w_{i,t}(0)}\right)^2 \exp\left( \mathbf{x}_{i,t}'(m)\widehat{\Delta}_{i,t} \right)}{1 + \left(1 - \exp\left( \mathbf{x}_t'(m)\widehat{\Delta}_{i,t} \right)\right) \widehat{G}_{i,t}},$$

where $\Theta_{-\tilde{\gamma}}$ is the domain of $(\beta_{i,t}, \Delta_{i,t}, \beta_{K,i,t}, G_{i,t})$. This reparametrized GMM estimation has the practical advantage of avoiding any demand inverse implementation in the course of numerical optimization. These inverse operations are instead replaced by a simple plug-in after the numerical optimization to back out the structural estimate $\widehat{\tilde{\gamma}}_{i,t}$.

# 4. Empirical Application: US Equity Market

In this section, we asses our model relative to the MNL in describing institutional investors' equity demand in the U.S from 2000.Q1 to 2022.Q4.

## 4.1. Data

We source institutional holdings from FactSet Ownership, which compiles quarterly 13F filings on long positions. Prices and shares outstanding are from Center for Research in Security Prices (CRSP). Firm balance sheet data are from Compustat. We follow Koijen et al. (2024) in constructing the final panel with asset holdings and asset characteristics; we apply the same data filters, investor classifications, and the investment universe definition. Similarly, we aggregate residual demand into a household sector to ensure market clearing in each quarter. Supplemental Appendix B.1 provides further details on data construction.

### 4.2. Deviation From the MNL: Reduced-Form Evidence

We use the reduced form approximation (19) to assess potential deviation from the MNL. Specifically, for each quarter $t$, we estimate the cross-sectional regression,

$$\ln\left(\frac{\mathbb{E}_t\left(\Gamma_{i,t}(n)\right)}{\mathbb{E}_t\left(\mu_{i,t}(n)\right)}\right) \sim \mathbf{x}_t(n), \quad \forall n = 1, \ldots, N_t, \tag{31}$$

where the corresponding coefficients approximate the market-average $\mathbb{E}_t\left(\Delta_{i,t}\right)$, summarizing the average deviation from the MNL risk–return trade-off. The regressor $\mathbf{x}_t(n)$ contains the six asset characteristics as in KY: log market equity, log book equity, profitability, investment, dividend-to-book, and the market-beta.

We implement the CAPM version of (6) to obtain the asset-time specific factor loadings $\mathbb{E}_t\left(\Gamma_{i,t}(n)\right)$ and expected log excess return $\mathbb{E}_t\left(\mu_{i,t}(n)\right)$. To that end, for each stock, we regress daily excess returns over the risk-free rate onto daily excess market returns and extract both the market factor loadings $\Gamma_t(n)$ and the average excess returns $\mu_t(n)$ within each quarter ($\sim 60$ observations per quarter).[10]

Figure 2 presents the estimates of $\mathbb{E}_t\left(\Delta_{i,t}\right)$ for each quarter. The top panel illustrates the corresponding adjusted $R^2$. The other panels display the estimates of each component in $\mathbb{E}_t\left(\Delta_{i,t}\right)$, one corresponding to an asset characteristic. We find that the coefficients for log market equity, log book equity, and market beta are overall the most statistically significantly different from zero (red plots). We see the dispersion of $\mathbb{E}_t\left(\Delta_{i,t}\right)$ for these three dominant characteristics roughly doubling in 2008–09 and again during the COVID-19 pandemic, consistent with investors demanding heterogeneous compensation when market frictions bite hardest. Interestingly, the adjusted $R^2$ in the top panel frequently exceeds $40\%$, peaking above $60\%$ during episodes of market stress. This suggests that the asset characteristics explain a substantial share of the cross-sectional variation in $\ln\left(\frac{\mathbb{E}_t\left(\Gamma_{i,t}(n)\right)}{\mathbb{E}_t\left(\mu_{i,t}(n)\right)}\right)$ which is otherwise constant in the MNL asset demand. These pieces of preliminary evidence hint on potential violation of the uniform risk-return trade-off implied by the MNL and motivate asset demand model (16) that allows for non-zero $\Delta_{i,t}$.

[10] While market beta enters on the right-hand side of (31) as a characteristic, it plays a conceptually different role from the estimated CAPM factor loading on the left-hand side. Following KY, the former is a long-run characteristic estimated from a 60-month rolling regression reflecting an asset's risk style, while the latter captures realized pricing behavior within the quarter.

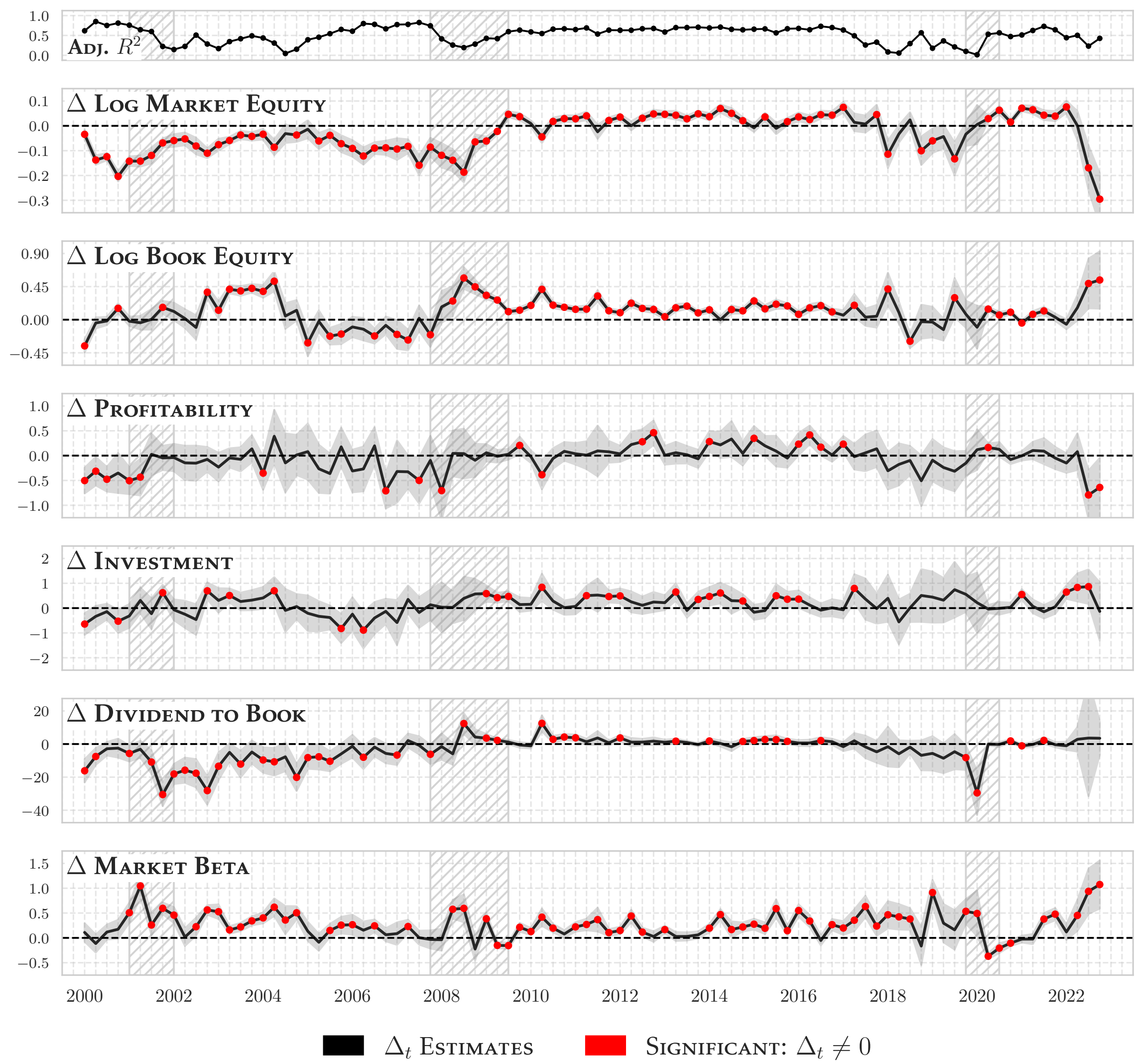


**Figure 2: Deviations From the MNL: Market-Level $\mathbb{E}_t\left(\Delta_{i,t}\right)$ Estimates**

This figure reports the market-level reduced-form evidence for deviations from the multinomial logit (MNL). From 2000.Q1 to 2022.Q4, we run a cross-sectional regression of log ratios of the market factor loadings $\Gamma_t(n)$ and the average excess returns $\mu_t(n)$ on a vector of asset characteristics including log market equity, log book equity, profitability, investment, dividend to book equity, market beta. To obtain $\mu_t(n)$ and $\Gamma_t(n)$, we regress daily excess returns over the risk-free rate onto daily excess market returns within each quarter ($\sim$ 60 observations per quarter). The top panel is the adjusted $R^2$ of each cross-sectional regression; the rest of the panels show the estimated $\Delta_t$ coefficients of each asset characteristic. The gray areas around the point estimates represent the 95% confidence intervals. The red dots highlight the point estimates that are statistically different from zero at 5% significance. The hatched areas indicate NBER recession periods.

## 4.3. Structural Asset Demand Estimation

We estimate our structural asset demand (16) using the GMM procedure (30). Following our identification analysis, we use the mandate-based instrument from KY for $\beta_{\mathbf{me},i,t}$ and the quadratic form of the exogenous asset characteristics for $\Delta_{i,t}$. As the baseline, we specify $\Delta_{i,t} = \Delta_{\mathbf{me},i,t}$, as market equity appears to be among the most significant asset characteristics that explains the cross-sectional variation in risk-return trade-off (see Figure 2). We also use the specification $\Delta_{i,t} = (\Delta_{\mathbf{me},i,t}, \Delta_{\mathbf{be},i,t}, \Delta_{\text{marketbeta},i,t})$ as a robustness check. As a comparison, we estimate the MNL asset demand that imposes $\Delta_{i,t} = \mathbf{0}$.

### Table 1: Structural Demand Estimates

This table reports the estimates of $(\beta_{i,t}, \Delta_{i,t})$ across 2000.Q1 - 2022.Q4 from the model:

$$\frac{w_{i,t}(n)}{w_{i,t}(0)} = \exp\left(\mathbf{x}_t'(n)\beta_{i,t} + \beta_{K,i,t} + \ln\left(1 + (1 - \exp(\mathbf{x}_t'(n)\Delta_{i,t}))\frac{\kappa_{i,t}}{1-\kappa_{i,t}}\right)\right)\varepsilon_{i,t}(n).$$

We specify $\mathbf{x}_t(n)$ as a vector of log market equity, log book equity, profitability, investment, dividend to book equity, and market beta. Both the estimates and their standard errors are computed using a GMM procedure as detailed in Supplemental Appendix B.2. Reported estimates are weighted equally across quarters and by asset under management across investors. Column (1) reports the multinomial logit specification. Columns (2) and (3) report the model specifications with $\Delta_{i,t}$ on log market equity only and on log market equity, log book equity, and market beta, respectively. To benchmark out-of-sample predictive accuracy, we report root mean square errors (RMSE) from a masked asset prediction, detailed in Supplemental Appendix B.3.

| | | $\Delta_{i,t} = 0$ | $\Delta_{i,t} \neq 0$ | |
|---|---|---|---|---|
| | | (1) | (2) | (3) |
| $\beta_{i,t}$ | Log Market Equity | 0.72 (0.10) | 0.71 (0.04) | 0.71 (0.06) |
| | Log Book Equity | 0.40 (0.09) | 0.39 (0.06) | 0.48 (0.08) |
| | Profitability | 0.08 (0.29) | 0.38 (0.19) | 0.26 (0.21) |
| | Investment | 0.10 (0.32) | 0.31 (0.23) | 0.23 (0.25) |
| | Dividend to Book | -2.17 (1.90) | -1.57 (0.40) | -1.53 (0.49) |
| | Market Beta | 0.17 (0.11) | 0.26 (0.09) | 0.26 (0.14) |
| $\Delta_{i,t}$ | Log Market Equity | | -0.42 (0.02) | -0.43 (0.05) |
| | Log Book Equity | | | -0.08 (0.11) |
| | Market Beta | | | 0.02 (0.18) |
| Number of $(i,t)$ Observations | | 207,634 | 207,634 | 207,634 |
| RMSE ($\times 10^{-2}$) | | 9.50 | 6.69 | 7.59 |

Table 1 summarizes the estimates of the MNL (Column 1) and our model (16) with $\Delta_{i,t} = \Delta_{\mathbf{me},i,t}$ (Column 2) and $\Delta_{i,t} = (\Delta_{\mathbf{me},i,t}, \Delta_{\mathbf{be},i,t}, \Delta_{\mathbf{marketbeta},i,t})$ (Column 3). Each cell reports the average parameter estimates and their standard errors. We take an AUM weighted average across investors and an equally weighted average across quarters. Switching on a single deviation parameter $\Delta_{\mathbf{me},i,t} \neq 0$ has significant effects; a large, negative, and tightly estimated deviation from the implied risk-return trade-off in MNL. In Column 3, we see that further degrees of freedom change very little. In the bottom panel of Table 1, we adopt the out-of-sample benchmarks proposed by Gabaix et al. (2025a,b) and report the root-mean-square error (RMSE) for each model when used to predict the masked second-largest asset holding in each investor-quarter combination. The model in Column 2—with only one risk-return trade-off parameter for market equity—already reduces the RMSE by around $30\%$ relative to the MNL.

While the average estimates of $\beta_{i,t}$ look quantitatively similar across the models, there is a statistically significant difference for a large portion of investor-quarter observations. In the top panels of Figure 3, we plot the distribution of each component of $\beta_{i,t}$ estimated by our model with $\Delta_{\mathbf{me},i,t} \neq 0$ of Table 1. The red segments represent the proportion of investor-quarter combinations for which our estimates statistically differ from those of the MNL. For $\beta_{me,i,t}$ (first panel), such difference seems to occur across the entire range of the estimates. For other asset characteristics such as market beta (last panel), the difference happens more frequently at the right tail. The bottom panel of Figure 3 summarizes our estimates of $\Delta_{\mathbf{me},i,t}$. Despite a mass around zero, most estimated $\Delta_{\mathbf{me},i,t}$ are negative and about $55\%$ of them are significantly different from zero (red segment).

Next, we study the heterogeneity in estimates across investor types in Figure 4. Blue dots represent our estimates for each quarter and red ones represent the MNL estimates (the corresponding $\bar{\Delta}_{\mathbf{me},t}$ are zero by construction). While the average estimates of $\beta_{\mathbf{me},i,t}$ seem to be close across the two models (the projections of blue dots on $x$ axis overlap with the red ones), the average $\bar{\Delta}_{\mathbf{me},t}$ of some investor types are non-trivially distant from zero, e.g., private banking.

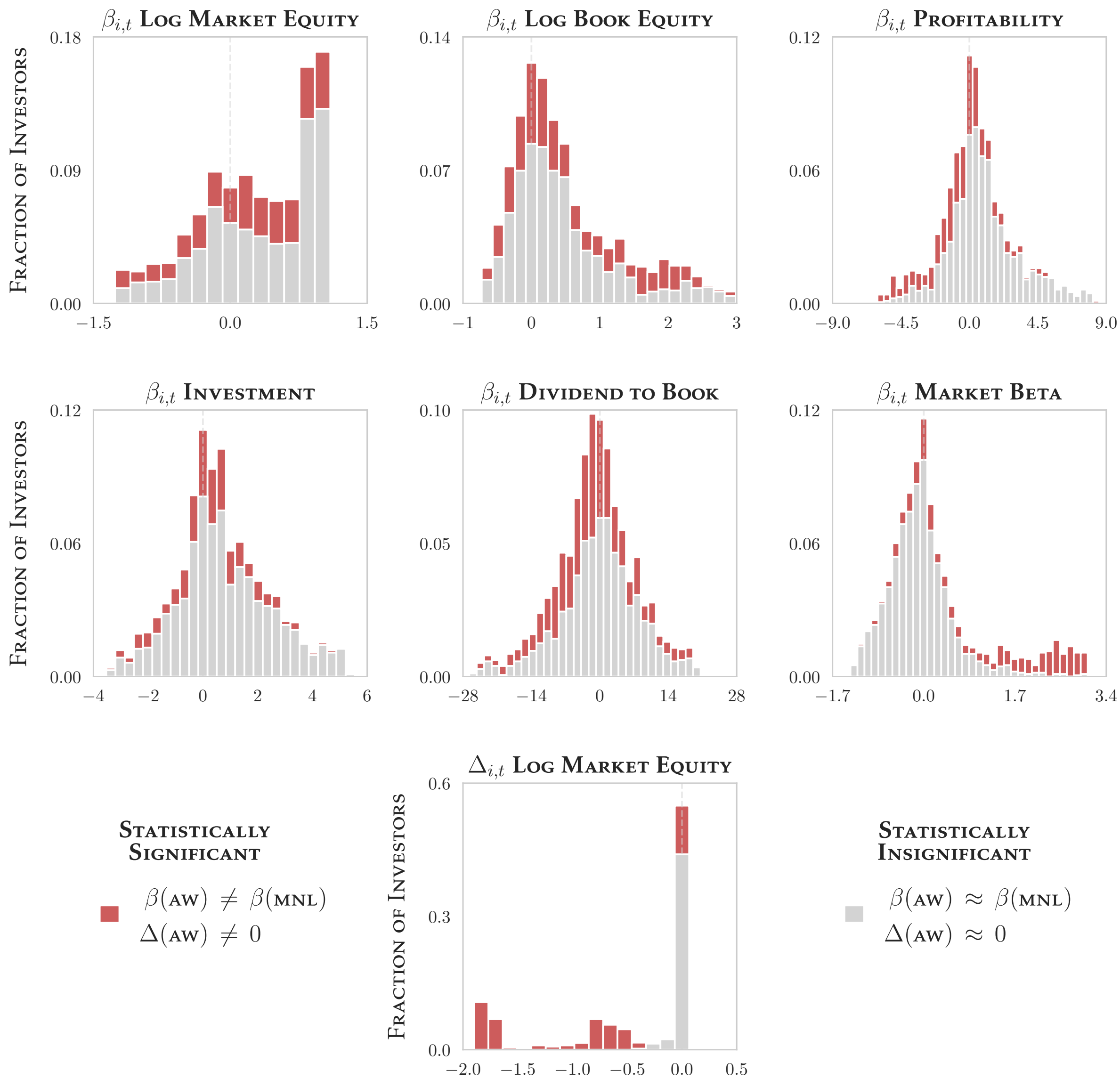


**Figure 3: Univariate Distribution of $\beta_{i,t}$ and $\Delta_{i,t}$ Estimates**

This figure reports the univariate distributions of investor-quarter specific $\beta$ and $\Delta$ coefficient estimates for 2011.Q3. The model specification allows $\Delta_{\mathbf{me},i,t} \neq 0$ relative to the multinomial logit. The top two rows show the fraction of investors binned by their $\beta_{i,t}$ estimates for log market equity, log book equity, profitability, investment, dividend to book equity and market beta. The third row show the fraction of investors binned by their $\Delta_{i,t}$ estimates for log market equity. In all panels, we trim the lowest and highest 5% of observations, and choose bin widths in proportion to each sample's spread and size so as to balance detail and smoothness. Bars shaded in red highlight the coefficient estimates that are statistically significant at the 5% level: $\beta(\text{AW}) \neq \beta(\text{MNL})$ in the top two rows and $\Delta(\text{AW}) \neq 0$ in the third row); gray bars denote non-significant estimates.


**Figure 4: Heterogeneity in Estimated $\beta_{i,t}$ and $\Delta_{i,t}$ Across Investor Types**

This figure highlights the investor heterogeneity in US equity demand. Each panel plots the joint distribution of average $\beta_{\mathbf{me},i,t}$ and $\Delta_{\mathbf{me},i,t}$ for the quarters 2000 Q1–2022 Q4. We take an AUM weighted average across each investor within types. Each dot reflects a quarter; red dots are estimates from the multinomial logit demand model, and blue dots are from our demand model with $\Delta_{\mathbf{me},i,t} \neq 0$.

## 4.4. Price Elasticities and Substitution Patterns

As in Corollary 2, an MNL investor perceives her assets to be either substitutes, complements, or independent, depending on her $\beta_{\mathbf{me},i,t}$, while our model (16) can relax this uniform substitution pattern with non-zero $\Delta_{i,t}$. We assess the differences in predicted price elasticities and the implications on asset substitution.

**A Measure of Asset Substitution.** To measure overall substitution in investor's asset demand, we construct a portfolio-level price elasticity that accounts for asset substitution when aggregating asset-level demand. This investor-quarter-specific elasticity, denoted by $\bar{\eta}_{i,t}$ describes the overall change in her asset holding due to a uniform $1\%$ increase in her assets' prices:

$$\bar{\eta}_{i,t} = \sum_{n \in \mathcal{N}_{i,t}} \frac{Q_{i,t}(n)}{\sum_{m \in \mathcal{N}_{i,t}} Q_{i,t}(m)} \sum_{j \in \mathcal{N}_{i,t}} \left[ -\frac{\partial \mathbf{q}_{i,t}(n)}{\partial \mathbf{p}_t(j)} \right] \tag{32}$$

In Supplemental Appendix B.4, we provide a proof for (32) and generalize it to assessing overall substitution among a group of assets (e.g., AI stocks). To see how $\bar{\eta}_{i,t}$ measures the extent of overall asset substitution, we further decompose it into:

$$\bar{\eta}_{i,t} = \underbrace{\sum_{n \in \mathcal{N}_{i,t}} \frac{Q_{i,t}(n)}{\sum\limits_{m \in \mathcal{N}_{i,t}} Q_{i,t}(m)} \left[ -\frac{\partial \mathbf{q}_{i,t}(n)}{\partial \mathbf{p}_t(n)} \right]}_{\bar{\eta}_{i,t}^{\text{own}}} + \underbrace{\sum_{n \in \mathcal{N}_{i,t}} \frac{Q_{i,t}(n)}{\sum\limits_{m \in \mathcal{N}_{i,t}} Q_{i,t}(m)} \sum_{j \in \mathcal{N}_{i,t}: j \neq n} \left[ -\frac{\partial \mathbf{q}_{i,t}(n)}{\partial \mathbf{p}_t(j)} \right]}_{\bar{\eta}_{i,t}^{\text{cross}}}. \tag{33}$$

The first term $\bar{\eta}_{i,t}^{\text{own}}$ in (33) is a weighted average of asset-level own-price elasticities. The second term $\bar{\eta}_{i,t}^{\text{cross}}$ captures the overall asset-level demand diversion to other assets following the price increase. In the MNL, $-\frac{\partial \mathbf{q}_{i,t}(n)}{\partial \mathbf{p}_t(j)} < 0$ (substitutes) if and only if $\beta_{\mathbf{me},i,t} < 0$ (see Corollary 2). Consequently, $\bar{\eta}_{i,t} < \bar{\eta}_{i,t}^{\text{own}}$ when assets are overall substitutes, and the opposite holds when assets are overall complements. Following this intuition, we use $\bar{\eta}_{i,t}$ and its magnitude relative to $\bar{\eta}_{i,t}^{\text{own}}$ as a proxy for investor's overall perceived substitution in general settings when $\Delta_{i,t} \neq \mathbf{0}$. In what follows, we call an investor substitutes (complements) investor (as per the MNL) if her estimate $\hat{\beta}_{\mathbf{me},i,t}^{\text{MNL}} < 0$ ($\hat{\beta}_{\mathbf{me},i,t}^{\text{MNL}} > 0$).

Figure 5 summarizes the averages of the own-price elasticities estimated by the

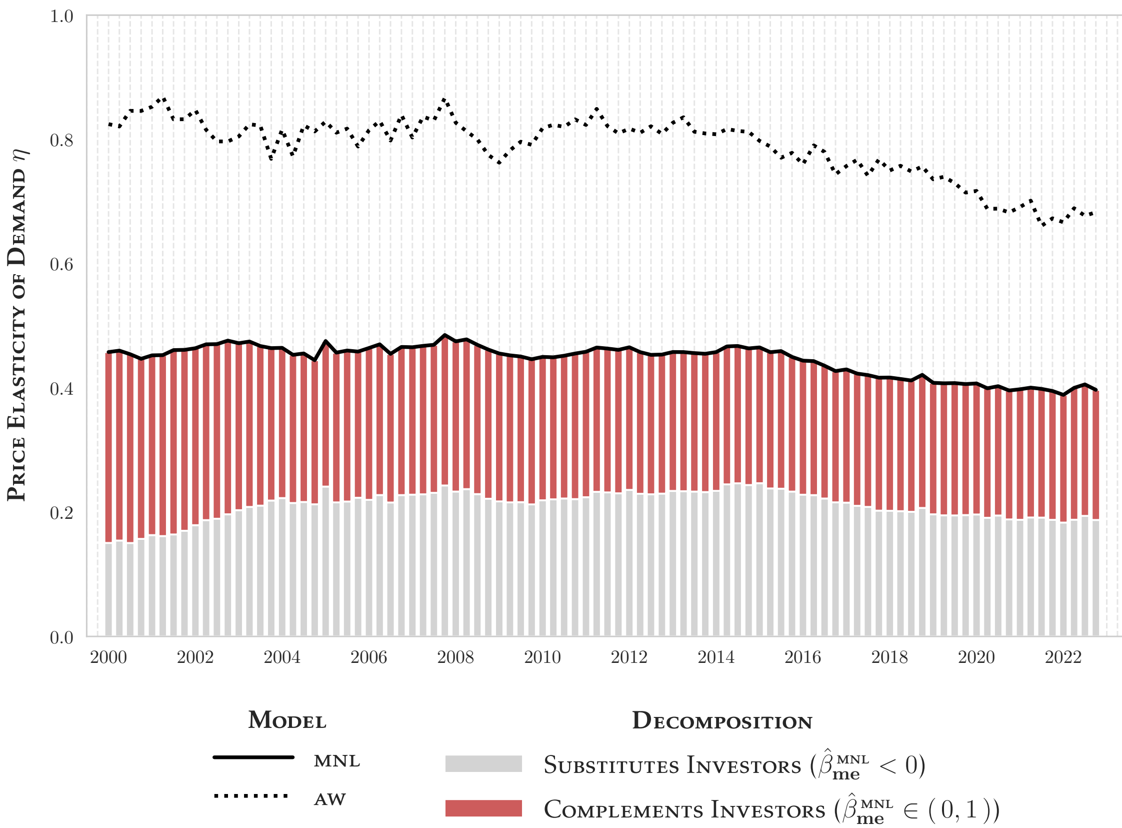


**Figure 5: Own-Price Elasticities of Demand: Investor Aggregation**

This figure tracks the quarterly averages of the own-price elasticities of demand from 2000.Q1 to 2022.Q4. We aggregate across investors taking into account their assets under management and then report the mean across own-price elasticities in the aggregate investor matrix. The solid and dotted lines plot the aggregate elasticity implied by the multinomial logit (AW) and by our model (AW), respectively. In each quarter, we decompose the aggregate elasticities into contributions by the substitutes investors (with $\hat{\beta}^{\text{MNL}}_{\mathbf{me}} < 0$) in gray and by the complements investors (with $\hat{\beta}^{\text{MNL}}_{\mathbf{me}} \in (0, 1)$) in red. The hatched areas indicate NBER recession periods.

MNL and our model (16). Similar to the MNL estimates, our model predicts a downward trend in the own-price elasticities over time. We predict a greater overall own-price elasticity ($0.79$) than the MNL ($0.45$), but both are of the same magnitude and far smaller than the one in the calibrated CAPM in Petajisto (2009) (by an order of three magnitudes). Given the MNL and our model's ability of generating large price elasticities (see Section 2) and their established econometric properties (Section 3), our elasticity estimates seem to support "the inelastic markets hypothesis" in the equity market, joining the recent literature that investigates small price elasticity in finance (e.g., Gabaix and Koijen (2024)).

To understand the origin of the difference between our estimated elasticities and those by the MNL, we decompose investors into substitutes and complements ones (as per the MNL). We find that both groups contribute around half to the overall own-price elasticities ($47\%$ vs $53\%$) in the MNL. In contrast, the same group of complements investors contribute significantly more to the own-price elasticity in our model (16): their average own-price elasticities make up around $73\%$ of the global average ($0.57$ out of $0.79$), while substitutes investors account for the remaining $27\%$ ($0.21$ out of $0.79$). Interestingly, the difference in the global averages is almost entirely driven by complements investors: the increase in their contribution to the own-price elasticities ($0.34 = 0.57 - 0.23$) accounts for virtually all of the difference between our model (16) and the MNL ($0.34 = 0.79 - 0.45$).

Next, in Figure 6, we assess the institutional investors' overall asset substitution using $\bar{\eta}_{i,t}$ (32) and its magnitude relative to $\bar{\eta}_{i,t}^{\text{own}}$ in (33). In the upper panel, we plot the quarter-average $\bar{\eta}_{i,t}$ (solid) and $\bar{\eta}_{i,t}^{\text{own}}$ (dotted) estimated by our model (blue) and the MNL (red) for substitutes investors. In the lower panel, we plot these estimates for complements investors. Both models predict similar quarter-average $\bar{\eta}_{i,t}$ ($\approx 1$) for both types of investors. In other words, an average substitutes/complements investor reduces around $1\%$ of her asset holding if the overall price of her assets increases by $1\%$. Moreover, both predict similar $\bar{\eta}_{i,t}^{\text{own}}$ for substitutes investors ($1.44$ vs $1.45$), i.e., both models predicts similar overall asset substitutability for substitutes investors. In contrast, relative to our model, the MNL substantially underestimates $\bar{\eta}_{i,t}^{\text{own}}$ for complements investors ($0.20$ vs $0.50$), and overestimates $\bar{\eta}_{i,t}^{\text{cross}} = \bar{\eta}_{i,t} - \bar{\eta}_{i,t}^{\text{own}}$, i.e., the overall complementarity among assets.

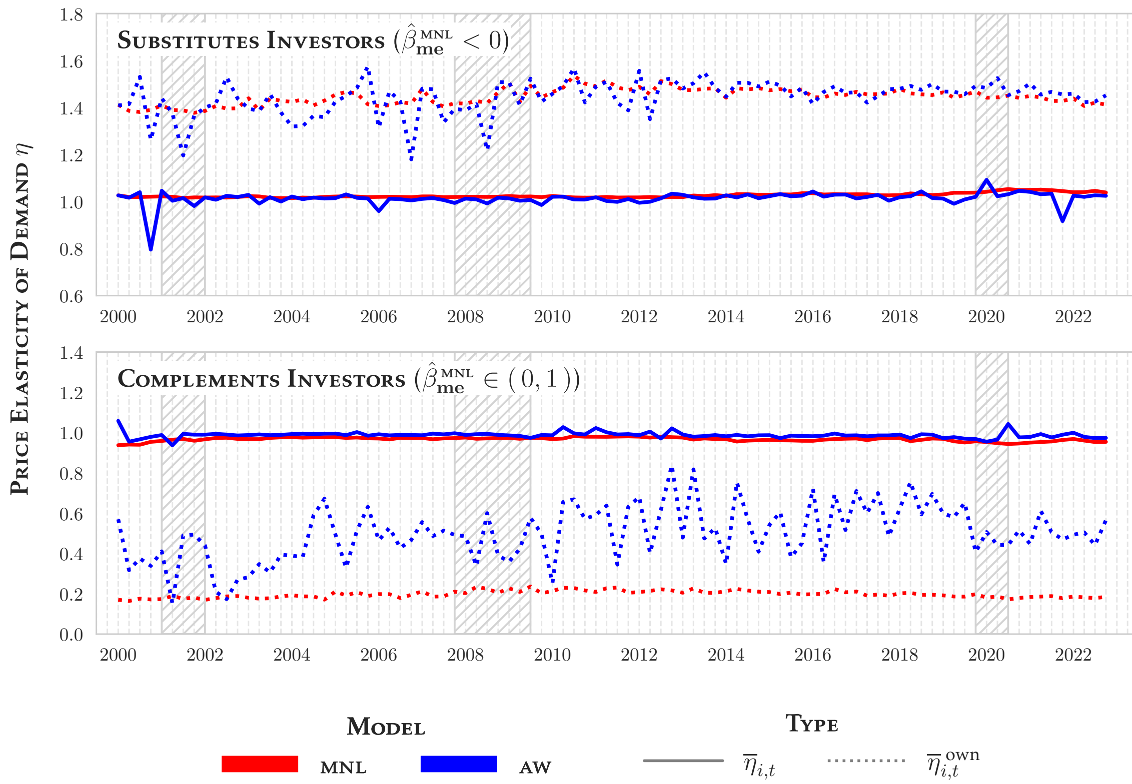


**Figure 6: Own-Price Elasticities of Demand: Stock Aggregation**

This figure reports the quarterly average price elasticities of demand for two investor subgroups over 2000.Q1–2022.Q4. In each quarter, we look at each investor's full elasticity matrix and its diagonals (own-price). Next, we aggregate the stocks for each case: for each investor, we end up with a single elasticity number for total price elasticity, and a single elasticity number for own-price elasticity. Finally, we take a weighted average across investors taking their asset under management into account. The top panel covers substitutes investors (those with negative MNL estimated $\beta$ on log market equity), and the bottom panel covers complements investors (those with positive MNL estimated $\beta$ on log market equity). Within each panel, solid lines plot the overall elasticities $\overline{\eta}_{i,t}$ and dashed lines plot the own-price elasticities $\overline{\eta}_{i,t}^{\text{own}}$. Blue highlights results from our structural model (aw); red highlights the multinomial logit (mnl) results. The hatched areas indicate NBER recession periods.

## 4.5. Investor Level Relative Alpha Estimation

In this section, we now use Proposition 1 to estimate investors' alpha from the demand estimates in column (2) of Table 1. First, we write a more primitive factor model that describes both the inside and the outside assets' return (relative to the risk-free asset) and implies (6):

$$r_{i,t+1}(n) - rf_{i,t+1} = \tilde{\alpha}_i(n) + \tilde{\Gamma}_{i,t}(n) f_{t+1} + \epsilon_{i,t}(n), \qquad \forall n = 0, 1, ..., N_{i,t}.$$

Then, we obtain:

$$r_{i,t+1}(n) - r_{i,t+1}(0) = \underbrace{\tilde{\alpha}_i(n) - \tilde{\alpha}_i(0) - \epsilon_{i,t}(0)}_{=:\alpha_{i,t}(n)} + [\underbrace{\tilde{\Gamma}_{i,t}(n) - \tilde{\Gamma}_{i,t}(0)}_{\Gamma_{i,t}(n)}] f_{t+1} + \epsilon_{i,t}(n).$$

Suppose that $\epsilon_{i,t}(0)$ (with a zero mean over time) is in investor $i$'s information set at time $t$ and $\mathbb{E}_{i,t}[f_{t+1}] = f_t$. The latter means that the investor's perception of the next quarter's market factor is equal to the one of the current quarter. Technically, this amounts to absorbing the $it$-specific $\mathbb{E}_{i,t}[f_{t+1}]/f_t$ in the definition of $\Gamma_{i,t}(n)$. Moreover, recall that we normalize $\phi_{i,t} = \psi_{i,t} = 1$ in (12) (see Footnote 4). Consequently, the formula in the second result of Proposition 1 becomes: for each $(i,t)$

$$\hat{\mu}_{i,t}(n) \approx \underbrace{\frac{\alpha_{i,t}(n)}{\phi_{i,t}} + \hat{\gamma}_{i,t}\frac{\psi_{i,t}^2}{2\phi_{i,t}}}_{=:\, c_{i,t,0}} + \hat{\Gamma}_{i,t}(n) \underbrace{\frac{\psi_{i,t}}{\phi_{i,t}} f_t}_{=:\, c_{i,t,1}} + \hat{\Gamma}_{i,t}^2(n) \underbrace{\frac{\psi_{i,t}^2}{2\phi_{i,t}}}_{=:\, c_{i,t,2}},$$

where $\hat{\mu}_{i,t}(n)$ and $\hat{\Gamma}_{i,t}(n)$ are the plug-in estimates of $\mu_{i,t}(n)$ and $\Gamma_{i,t}(n)$ in (14) and (15). For each $(i,t)$, we run a cross-asset regression of $\hat{\mu}_{i,t}(n)$ over a constant term, $\hat{\Gamma}_{i,t}(n)$, and $\hat{\Gamma}_{i,t}^2(n)$, and obtain the estimates $(\hat{c}_{i,t,0}, \hat{c}_{i,t,1}, \hat{c}_{i,t,2})$. Then, we back out

$$\frac{\alpha_{i,t}(n)}{\phi_{i,t}} = \hat{c}_{i,t,0} - \hat{\gamma}_{i,t}\hat{c}_{i,t,2}, \quad \frac{f_t}{\psi_{i,t}} = \frac{\hat{c}_{i,t,1}}{2\hat{c}_{i,t,2}}, \quad \frac{\psi_{i,t}^2}{\phi_{i,t}} = 2\hat{c}_{i,t,2},$$

and

$$\alpha_{i,t}(n) = \left(\frac{2 f_t \hat{c}_{i,t,2}}{\hat{c}_{i,t,1}}\right)^2 \frac{\hat{c}_{i,t,0} - \hat{\gamma}_{i,t}\hat{c}_{i,t,2}}{2\hat{c}_{i,t,2}} = \frac{2 f_t^2 \hat{c}_{i,t,2}(\hat{c}_{i,t,0} - \hat{\gamma}_{i,t}\hat{c}_{i,t,2})}{\hat{c}_{i,t,1}^2} = \tilde{\alpha}_i(n) - \tilde{\alpha}_i(0) - \epsilon_{i,t}(0).$$

We estimate investor's alpha of asset $n$ relative to the outside asset $\tilde{\alpha}_i(n) - \tilde{\alpha}_i(0)$ as the average of estimated $\alpha_{i,t}(n)$ over time. The results are reported in Figure 7.

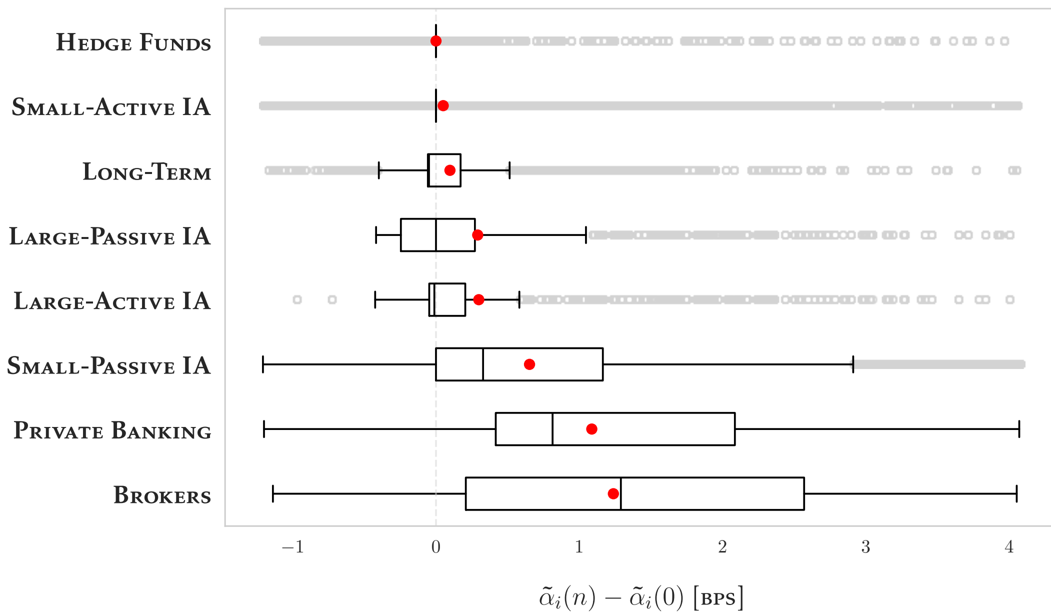


**Figure 7: Distribution of Investors' Relative Alphas by Institutional Type**

This figure reports, for each institutional type in the FactSet Ownership dataset, the weighted distribution of each investor's inside asset alphas relative to each investor's outside assets (cash, bonds, etc.). We take a weighted average across investors taking their asset under management into account. The alphas are expressed in basis points. In each panel, the horizontal box spans the 25th–75th percentiles of that category's alpha distribution, the central line marks the median, the diamond denotes the AUM-weighted mean, whiskers extend to 1.5 times the interquartile range, and light-gray circles indicate outliers. Panels are arranged from lowest to highest mean alpha.

Our estimates indicate that hedge funds exhibit near-zero alphas, i.e., there is no persistent wedge that remains relative to what they could have earned by parking the same amount of capital in bonds, cash, FX or any other outside asset. Consistent with the existing studies (e.g. Agarwal and Naik, 2004; Patton and Ramadorai, 2013; Ardia et al., 2024), we show that enriching the factor structure to capture time-varying and nonlinear exposures absorbs what a naïve CAPM would misattribute to persistent alpha. Our structural factor loadings, $\Gamma_{i,t}(n)$, capture hedge funds' sophisticated, nonlinear, and time-varying risk exposures, hence effectively reclassifying apparent excess returns as factor-driven compensation.

In contrast, investor types such as brokers display non-zero alphas relative to the outside assets. These investors either persistently earn premiums in the equity market relative to the outside options, or employ strategies that remain orthogonal to our modeled market factor. Such strategies might include profits earned from bid–ask spreads, payment-for-order-flow and other liquidity sources, which is an unmodeled risk for our factor loadings and thus would highlight meaningful variation in alpha attribution.

Overall, the estimated alphas may appear small in magnitude, but even modest excess returns translate into economically significant gains when viewed in monetary terms. For instance, among brokers, Morgan Stanley achieves approximately $15 million annually from an estimated quarterly alpha of 1.33 bp.

| Investor Type | Value | Neutral | Growth |
|---|---|---|---|
| Hedge Funds | -0.0016 bp | -0.0009 bp | 0.0037 bp |
| Small-Active IA | 0.0612 bp | 0.0504 bp | 0.0390 bp |
| Long-Term | 0.0910 bp | 0.0968 bp | 0.1072 bp |
| Large-Active IA | 0.2924 bp | 0.3042 bp | 0.2984 bp |
| Large-Passive IA | 0.2632 bp | 0.2853 bp | 0.3473 bp |
| Small-Passive IA | 0.6834 bp | 0.6557 bp | 0.6078 bp |
| Private Banking | 1.1051 bp | 1.0951 bp | 1.0492 bp |
| Brokers | 1.2135 bp | 1.2510 bp | 1.2514 bp |

Asset Type

**Figure 8: Heatmap of Investor Alphas: Investor Types vs Asset Types**

This figure reports the weighted average of investor alphas by different asset style buckets. We take a weighted average across investors taking their asset under management into account. The asset buckets are formed each quarter by computing each stock's book-to-market, ranking all stocks into percentiles, and then assigning the top 30% to value, the middle 40% to neutral and the bottom 30% to growth. Each cell shows the mean $\alpha_i(n$ (relative to the outside asset) for that investor-type/style combination, with investor types ordered by their overall mean alpha. Cell colors highlight within-column rank from lowest (blue) to highest (red), and the annotated numbers report the exact alpha values in basis points.

We further study how residual performance varies not just by investor but by the style of assets they hold. Grouping stocks into Value, Neutral, and Growth buckets in Figure 8 shows that residual performance is driven far more by the investor than by the style of stock held. For hedge funds, we still see statistically indistinguishable alphas from zero in every bucket, confirming once again that their returns are fully captured by our structural model of systematic factor exposures. Brokers, by contrast, earn a steady 1 bp per quarter regardless of style; likely pointing to an intermediation edge rather than stock-selection skill. Among advisers, the pattern splits by scale: small active and small passive managers pick up most of their alpha in value names, whereas large passive funds and long-term institutions (e.g., pension funds) realize what little alpha they have in growth stocks, which is plausibly a natural outcome of cap-weighted index and benchmark mechanics. In short, investor identity and not value-versus-growth positioning explains the bulk of residual performance.

Our framework produces investor-level alphas that compare each investor's inside-asset positions with the rest of the opportunity set, moving beyond the usual single-portfolio or asset-level measures implied by representative-agent models. The same factor-model regressions also deliver the structural demand parameters that underlie these alphas.

## 5. Conclusion

In this paper, we revisit the demand-based asset-pricing framework of KY in light of its assumptions on asset substitution, which require investors to view all assets uniformly as either substitutes, complements, or independent. We ask whether a more flexible asset demand model can capture the cross-asset risk-return trade-offs observed in real-world portfolios, and what such flexibility implies for price elasticities, substitution patterns, and alphas across investor types.

Our answer is a flexible characteristic-based asset demand system that nests the MNL model while allowing the risk-return trade-off to vary with observable asset characteristics. Methodologically, we (i) derive closed-form portfolio weights that decompose demand into diversion, intra-portfolio, and portfolio-level channels; (ii) construct a one-dimensional inverse mapping that sidesteps the usual *curse of dimensionality*; and (iii) identify the model using mandate-based and higher-order characteristic instruments, estimated via a GMM procedure. Applied to U.S. institutional equity holdings from 2000 to 2022, the model reveals meaningful deviations from MNL predictions. It yields an average own-price elasticity of 0.79, higher than the MNL's 0.45, driven almost entirely by investors who, under MNL, perceive their assets as complements. The model also uncovers non-trivial heterogeneity in investor-specific alphas.

Moreover, we demonstrate that both the MNL and our model are capable of generating large price elasticities and small price effects. This property suggests that a flexible demand-based approach should be able to recover high (or low) elasticities in finite samples (subject to statistical errors), provided the true elasticities are indeed high (or low). In our empirical analysis of the U.S. equity market, both the MNL and our model produce elasticity estimates of similar magnitude—substantially lower than those implied by calibrated CAPMs. Given the validity of our identification strategy and estimation framework, this finding provides empirical support for the inelastic market hypothesis, at least within the equity market we examine.

Built on our framework, two extensions could enhance the research on demand system asset pricing. First, embedding a multi-factor return structure would allow different characteristic mappings for different factor loadings, enabling tests of whether substitution heterogeneity persists once multiple sources of priced risk are incorporated—for instance, separating factor exposures for green premia and liquidity premia. Second, it is beneficial to develop suitable asymptotic settings and formally discuss the inference of demand parameters estimated by our two-step strategy. One candidate is many-asset asymptotics: the number of assets increases to infinity, while the number of investors is given. It is conceptually analogue to the large-market setting in IO (e.g., Armstrong (2016)) and fits the empirical applications with investor-time specific demand parameters. Another one is many-investor (or time periods) setting where the number of investors increases to infinity while the number of assets is bounded. Intuitively, it is suitable when one pools demand parameters across (some) investors and/or time periods.

- References -


Agarwal, V. and N. Y. Naik (2004): "Risks and Portfolio Decisions Involving Hedge Funds," *Review of Financial Studies*, 17, 63–98.

Allen, J., J. Kastl, and M. Wittwer (2024): "Estimating Demand Systems with Bidding Data," *Working Paper*.

Ardia, D., L. Barras, P. Gagliardini, and O. Scaillet (2024): "Is It Alpha or Beta? Decomposing Hedge Fund Returns When Models Are Misspecified," *Journal of Financial Economics*, 154, 103805.

Armstrong, T. B. (2016): "Large Market Asymptotics for Differentiated Product Demand Estimators With Economic Models of Supply," *Econometrica*, 84, 1961–1980.

Berry, S., J. Levinsohn, and A. Pakes (1995): "Automobile Prices in Market Equilibrium," *Econometrica*, 63, 841–890.

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

Bretscher, L., L. Schmid, I. Sen, and V. Sharma (2025): "Institutional Corporate Bond Pricing," *Forthcoming in Review of Financial Studies*.

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

Chaudhary, M., Z. Fu, and J. Li (2023): "Corporate Bond Multipliers: Substitutes Matter," *Working Paper*.

Cremers, K. J. M. and A. Petajisto (2009): "How Active Is Your Fund Manager? A New Measure That Predicts Performance," *Review of Financial Studies*, 22, 3329–3365.

Davis, C., M. Kargar, and J. Li (2025): "Why Do Portfolio Choice Models Predict Inelastic Demand?" *Journal of Financial Economics*, 172, 104096.

Eren, E., A. Schrimpf, and D. Xia (2024): "The Demand for Government Debt," *Working Paper*.

Fang, X., B. Hardy, and K. K. Lewis (2025): "Who Holds Sovereign Debt and Why It Matters," *Fortcoming in Review of Financial Studies*.

Fuchs, W. M., S. Fukuda, and D. Neuhann (2024): "Demand-System Asset Pricing: Theoretical Foundations," *Working Paper*.

Gabaix, X., R. Koijen, R. J. Richmond, and M. Yogo (2025a): "Asset Embeddings," *Working Paper*.

——— (2025b): "Upgrading Credit Pricing and Risk Assessment through Embeddings," *Working Paper*.

Gabaix, X. and R. S. J. Koijen (2024): "In Search of the Origins of Financial Fluctuations: The Inelastic Markets Hypothesis," *Working Paper*.

Gandhi, A. and J.-F. Houde (2019): "Measuring Substitution Patterns in Differentiated-Products Industries," *Working Paper*.

Haddad, V., P. Huebner, and E. Loualiche (2025): "How Competitive Is the Stock Market? Theory, Evidence from Portfolios, and Implications for the Rise of Passive Investing," *American Economic Review*, 115, 975–1018.

Jansen, K. A., W. Li, and L. Schmid (2024): "Granular Treasury Demand with Arbitrageurs," *Working Paper*.

Jiang, Z., R. J. Richmond, and T. Zhang (2024): "A Portfolio Approach to Global Imbalances," *Journal of Finance*, 79, 2025–2076.

——— (2025): "Understanding the Strength of the Dollar," *Forthcoming in Journal of Financial Economics*.

Kim, B. (2025): "Correlated Demand Shocks and Asset Pricing," *Working Paper*.

Koijen, R. S., F. Koulischer, B. Nguyen, and M. Yogo (2021): "Inspecting the Mechanism of Quantitative Easing in the Euro Area," *Journal of Financial Economics*, 140, 1–20.

Koijen, R. S. J., R. J. Richmond, and M. Yogo (2024): "Which Investors Matter for Equity Valuations and Expected Returns?" *Review of Economic Studies*, 91, 2387–2424.

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

——— (2023): "Understanding the Ownership Structure of Corporate Bonds," *American Economic Review: Insights*, 5, 73–91.

——— (2024): "Exchange Rates and Asset Prices in a Global Demand System," *Working Paper*.

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

McFadden, D. (1974): "Conditional Logit Analysis of Qualitative Choice Behavior," *Frontiers in Econometrics*.

Patton, A. J. and T. Ramadorai (2013): "On the High-Frequency Dynamics of Hedge Fund Risk Exposures," *Journal of Finance*, 68, 597–635.

Petajisto, A. (2009): "Why Do Demand Curves for Stocks Slope Down?" *Journal of Financial and Quantitative Analysis*, 44.

Rothenberg, T. J. (1971): "Identification in Parametric Models," *Econometrica*, 577–591.

Samuelson, P. A. (1969): "Lifetime Portfolio Selection By Dynamic Stochastic Programming," *Review of Economics and Statistics*, 51, 239–246.

Siani, K. (2025): "Raising Bond Capital in Segmented Markets," *Forthcoming in Review of Financial Studies*.

Wang, A. (2023): "Sieve BLP: A Semi-Nonparametric Model of Demand for Differentiated Products," *Journal of Econometrics*, 235, 325–351.

——— (2024): "Demand with Complementarity: A Market-Share Approach," *Working Paper*.

# Supplemental Appendix

to

# "Risk and Return in Asset Demand Systems"

# A. Proofs

## A.1. Proof of Proposition 1

**For the first statement.** From (9), we have

$$\begin{aligned}\frac{w_{i,t}(n)}{w_{i,t}(0)} &= \frac{\pi_{i,t}}{w_{i,t}(0)}\left(1+\mathbf{y}_{i,t}'(n)\frac{\mathbf{\Phi}_{i,t}-\mathbf{\Psi}_{i,t}\kappa_{i,t}}{\phi_{i,t}-\psi_{i,t}\kappa_{i,t}}\right)\\ &= \exp\left(\beta_{K,i,t}\right)\left(1+\mathbf{y}_{i,t}'\frac{\mathbf{\Phi}_{i,t}}{\phi_{i,t}}+\mathbf{y}_{i,t}'\left(\frac{\mathbf{\Phi}_{i,t}}{\phi_{i,t}}-\frac{\mathbf{\Psi}_{i,t}}{\psi_{i,t}}\right)\frac{\kappa_{i,t}}{\phi_{i,t}/\psi_{i,t}-\kappa_{i,t}}\right)\\ &= \exp\left(\beta_{K,i,t}\right)\left(1+\mathbf{y}_{i,t}'\frac{\mathbf{\Phi}_{i,t}}{\phi_{i,t}}+\mathbf{y}_{i,t}'\left(\frac{\mathbf{\Phi}_{i,t}}{\phi_{i,t}}-\frac{\mathbf{\Psi}_{i,t}}{\psi_{i,t}}\right)\frac{\tilde{\kappa}_{i,t}}{1-\tilde{\kappa}_{i,t}}\right),\end{aligned}$$

where

$$\tilde{\kappa}_{i,t}=\frac{\sum_{n\in\mathcal{N}_{i,t}}\frac{\mu_{i,t}}{\phi_{i,t}}\frac{\Gamma_{i,t}}{\psi_{i,t}}}{\frac{\gamma_{i,t}}{\psi_{i,t}^2}+\sum_{n\in\mathcal{N}_{i,t}}\left(\frac{\Gamma_{i,t}}{\psi_{i,t}}\right)^2},\qquad \kappa_{i,t}=\tilde{\kappa}_{i,t}\frac{\phi_{i,t}}{\psi_{i,t}}.$$

Note that $\frac{\mu_{i,t}(n)}{\phi_{i,t}}=1+\mathbf{y}_{i,t}'\frac{\mathbf{\Phi}_{i,t}}{\phi_{i,t}}$ and $\frac{\Gamma_{i,t}(n)}{\psi_{i,t}}=1+\mathbf{y}_{i,t}'\frac{\mathbf{\Psi}_{i,t}}{\psi_{i,t}}$. □

**For the second statement.** Note that $\mu_{i,t}(n)=\mathbb{E}_{i,t}\left[r_{t+1}(n)-r_{t+1}(0)\right]+\frac{\sigma_{i,t}^2(n)}{2}$ derived from the mean-variance approximation (see p.1481 of Koijen and Yogo (2019) and equation 2.23 of Campbell and Viceira (2002)), where $\sigma_{i,t}^2(n)=\Gamma_{i,t}^2(n)+\gamma_{i,t}$. We then obtain the desired result by combining the factor structure (6) and the expression of $\mu_{i,t}(n)$. □

**For the third statement.**

- **The if part.**

  Using (12), when $\frac{\mathbf{\Phi}_{i,t}}{\phi_{i,t}} = \frac{\mathbf{\Psi}_{i,t}}{\psi_{i,t}}$:

$$\frac{w_{i,t}(n)}{w_{i,t}(0)} = \exp\left(\,\beta_{K,i,t}\,\right)\left(1 + \mathbf{y}'_{i,t}\frac{\mathbf{\Phi}_{i,t}}{\phi_{i,t}}\right) = \exp\left(\,\beta_{K,i,t}\,\right)\mu_{i,t}(n)/\phi_{i,t}.$$

  Then,

$$w_{i,t}(n) = \frac{\mu_{i,t}(n)/\phi_{i,t}}{\exp\left(\,-\beta_{K,i,t}\,\right) + \sum_{m\in\mathcal{N}_{i,t}}\mu_{i,t}(m)/\phi_{i,t}} \qquad \text{(A.1.1)}$$

  □

- **The only if part.**

  Suppose that the asset weight in (9) has the form (A.1.1). Then for any $n$:

$$\frac{w_{i,t}(n)}{w_{i,t}(0)} = \exp\left(\,\beta_{K,i,t}\,\right)\mu_{i,t}(n)/\phi_{i,t} \implies w_{i,t}(n) = \pi_{i,t}\left(1 + \mathbf{y}'_{i,t}(n)\frac{\mathbf{\Phi}_{i,t}}{\phi_{i,t}}\right).$$

  From (9), we have $w_{i,t}(n) = \pi_{i,t} + \mathbf{y}'_{i,t}(n)\mathbf{\Pi}_{i,t}$. Consequently, we obtain $\mathbf{y}'_{i,t}(n)\mathbf{\Pi}_{i,t} = \mathbf{y}'_{i,t}(n)\pi_{i,t}\mathbf{\Phi}_{i,t}/\phi_{i,t}$ for any $n$. Moreover, since $(\mathbf{y}'_{i,t}(1); ...; \mathbf{y}'_{i,t}(N_{i,t}))$ is of full column rank, we have: $\mathbf{\Pi}_{i,t} = \mathbf{\Phi}_{i,t}/\gamma_{i,t} - \mathbf{\Psi}_{i,t}\kappa_{i,t}/\gamma_{i,t} = \pi_{i,t}\mathbf{\Phi}_{i,t}/\phi_{i,t} = \phi_{i,t}\mathbf{\Phi}_{i,t}/(\phi_{i,t}\gamma_{i,t}) - \psi_{i,t}\mathbf{\Phi}_{i,t}\kappa_{i,t}/(\phi_{i,t}\gamma_{i,t}) = \mathbf{\Phi}_{i,t}/\gamma_{i,t} - \psi_{i,t}\mathbf{\Phi}_{i,t}\kappa_{i,t}/(\phi_{i,t}\gamma_{i,t})$. Thus, $\mathbf{\Phi}_{i,t}/\phi_{i,t} = \mathbf{\Psi}_{i,t}/\psi_{i,t}$.

  □

## A.2. Proof of Proposition 2

We obtain the desired expression of asset weight by plugging $\mu_{i,t}(n)$ and $\Gamma_{i,t}(n)$ into (12).

## A.3. Proof of Corollary 1

Note that $-\frac{\partial \mathbf{q}_{i,t}(n)}{\partial \mathbf{p}_t(j)} = -\frac{\partial \ln w_{i,t}(n)}{\partial \text{me}_t(j)} + \mathbf{1}\{n=j\}$, where

$$\ln w_{i,t}(n) = \mathbf{x}_t'(n)\beta_{i,t} + \ln\left(1 + \left(1 - \exp\left(\mathbf{x}_t'(n)\Delta_{i,t}\right)\right)\frac{\kappa_{i,t}}{1-\kappa_{i,t}}\right) + \ln\varepsilon_{i,t}(n)$$
$$-\ln\left(\exp\left(-\beta_{K,i,t}\right) + \sum_{m=1}^{N_{i,t}} \exp\left(\mathbf{x}_t'(m)\beta_{i,t}\right)\left(1 + \left(1 - \exp\left(\mathbf{x}_t'(m)\Delta_{i,t}\right)\right)\frac{\kappa_{i,t}}{1-\kappa_{i,t}}\right)\varepsilon_{i,t}(m)\right).$$

The diversion effect comes from the derivative via $\mathbf{x}_t'(n)\beta_{i,t}$:

$$\mathbf{1}\{n=j\} - \beta_{\text{me},i,t}\mathbf{1}\{n=j\} + w_{i,t}(j)\beta_{\text{me},i,t} = \mathbf{1}\{n=j\} - \beta_{\text{me},i,t}\left[\mathbf{1}\{n=j\} - w_{i,t}(j)\right].$$

The intra-portfolio part originates from the derivative via $\mathbf{x}_t'(n)\Delta_{i,t}$:

$$-\left[\begin{array}{c}\mathbf{1}\{n=j\}\dfrac{-\exp\left(\mathbf{x}_t'(n)\Delta_{i,t}\right)\frac{\kappa_{i,t}}{1-\kappa_{i,t}}}{1+\left(1-\exp\left(\mathbf{x}_t'(n)\Delta_{i,t}\right)\right)\frac{\kappa_{i,t}}{1-\kappa_{i,t}}} \\ -\dfrac{\exp\left(\mathbf{x}_t'(j)\beta_{i,t}\right)\left(-\exp\left(\mathbf{x}_t'(j)\Delta_{i,t}\right)\right)\frac{\kappa_{i,t}}{1-\kappa_{i,t}}\varepsilon_{i,t}(j)}{\exp\left(-\beta_{K,i,t}\right)+\sum_{m=1}^{N_{i,t}}\exp\left(\mathbf{x}_t'(m)\beta_{i,t}\right)\left(1+\left(1-\exp\left(\mathbf{x}_t'(m)\Delta_{i,t}\right)\right)\frac{\kappa_{i,t}}{1-\kappa_{i,t}}\right)\varepsilon_{i,t}(m)}\end{array}\right]\Delta_{\text{me},i,t}$$
$$=$$
$$-\left[\mathbf{1}\{n=j\}\frac{-\exp\left(\mathbf{x}_t'(n)\Delta_{i,t}\right)\frac{\kappa_{i,t}}{1-\kappa_{i,t}}}{1+\left(1-\exp\left(\mathbf{x}_t'(n)\Delta_{i,t}\right)\right)\frac{\kappa_{i,t}}{1-\kappa_{i,t}}} + \frac{w_{i,t}(j)}{1+\left(1-\exp\left(\mathbf{x}_t'(j)\Delta_{i,t}\right)\right)\frac{\kappa_{i,t}}{1-\kappa_{i,t}}}\right]\Delta_{\text{me},i,t}$$
$$=$$
$$\left[\mathbf{1}\{n=j\} - w_{i,t}(j)\right]\frac{\exp\left(\mathbf{x}_t'(j)\Delta_{i,t}\right)\kappa_{i,t}}{1-\exp\left(\mathbf{x}_t'(j)\Delta_{i,t}\right)\kappa_{i,t}}\Delta_{\text{me},i,t}.$$

The portfolio-level effect is due to the derivative via $\kappa_{i,t}$.

## A.4. Proof of the MNL Price Effect

We ignore notation $t$ for simplicity. First, we derive the generic formula of price effect due to the change in $\delta$. Using (23), we have: for $n = 1, ..., N_i$

$$
\begin{aligned}
\frac{\partial \mathbf{me}(n)}{\partial \delta} &= \frac{\partial \mathbf{me}(n)}{\partial \delta} \frac{\sum_i A_i \frac{\partial w_i(n)}{\partial \mathbf{me}(n)}}{\sum_i A_i w_i(n)} + \sum_{m:m\neq n} \frac{\partial \mathbf{me}(m)}{\partial \delta} \frac{\sum_i A_i \frac{\partial w_i(n)}{\partial \mathbf{me}(m)}}{\sum_i A_i w_i(n)} + \frac{\sum_i A_i \frac{\partial w_i(n)}{\partial \delta}}{\sum_i A_i w_i(n)} \\
&= \frac{\partial \mathbf{me}(n)}{\partial \delta} \frac{\sum_i A_i w_i(n) \frac{\partial \ln w_i(n)}{\partial \mathbf{me}(n)}}{\sum_i A_i w_i(n)} + \sum_{m:m\neq n} \frac{\partial \mathbf{me}(m)}{\partial \delta} \frac{\sum_i A_i w_i(n) \frac{\partial \ln w_i(n)}{\partial \mathbf{me}(m)}}{\sum_i A_i w_i(n)} + \frac{\sum_i A_i w_i(n) \frac{\partial \ln w_i(n)}{\partial \delta}}{\sum_i A_i w_i(n)}.
\end{aligned}
$$

Note that $\mathbf{me}(n) = \ln S(n) + \mathbf{p}(n)$ and $\ln w_i(n) = \mathbf{q}_i(n) + \mathbf{p}(n) - \ln A_i$. Then,

$$
\frac{\partial \mathbf{p}(n)}{\partial \delta} \underbrace{\left( \frac{\sum_i A_i w_i(n) \left[ -\frac{\partial \mathbf{q}_i(n)}{\partial \mathbf{p}(n)} \right]}{\sum_i A_i w_i(n)} \right)}_{\bar{\eta}(n,n)} + \sum_{m:m\neq n} \frac{\partial \mathbf{p}(m)}{\partial \delta} \underbrace{\frac{\sum_i A_i w_i(n) \left[ -\frac{\partial \mathbf{q}_i(n)}{\partial \mathbf{p}(m)} \right]}{\sum_i A_i w_i(n)}}_{\bar{\eta}(n,m)} = \frac{\sum_i A_i w_i(n) \frac{\partial \ln w_i(n)}{\partial \delta}}{\sum_i A_i w_i(n)},
$$

or in the matrix form:

$$
\frac{\partial \mathbf{p}}{\partial \delta} = \left[ (\bar{\eta}(n,m))_{n,m} \right]^{-1} \left( \frac{\sum_i A_i w_i(1) \frac{\partial \ln w_i(1)}{\partial \delta}}{\sum_i A_i w_i(1)}; ...; \frac{\sum_i A_i w_i(N) \frac{\partial \ln w_i(N)}{\partial \delta}}{\sum_i A_i w_i(N)} \right).
$$

In the case of MNL asset demand, i.e., $\Delta_i = \mathbf{0}$ for all $i$ in (16), we have:

$$
\bar{\eta}(n,n) = \sum_i \frac{A_i w_i(n)}{\sum_i A_i w_i(n)} \left(1 - \beta_{\text{me},i}(1 - w_i(n))\right), \quad \bar{\eta}(n,m) = \sum_i \frac{A_i w_i(n)}{\sum_i A_i w_i(n)} \beta_{\text{me},i} w_i(m).
$$

Suppose that $\beta_{\text{me},i}$ is of the order of $\beta$. Then,

$$
\bar{\eta}(n,n) \approx 1 - \beta + \beta \frac{\sum_i A_i w_i(n) w_i(n)}{\sum_i A_i w_i(n)}, \quad \bar{\eta}(n,m) \approx \beta \frac{\sum_i A_i w_i(n) w_i(m)}{\sum_i A_i w_i(n)}.
$$

Consequently,

$$
\left[ (\bar{\eta}(n,m))_{n,m} \right]^{-1} \approx \frac{1}{1-\beta} \left[ \mathbf{1}_{N\times N} + \frac{\beta}{1-\beta} \left( \frac{\sum_i A_i w_i(n) w_i(m)}{\sum_i A_i w_i(n)} \right)_{n,m} \right]^{-1}.
$$

When $\beta \ll 0$, we can further approximate $\left[(\bar{\eta}(n,m))_{n,m}\right]^{-1}$ by:

$$\left[(\bar{\eta}(n,m))_{n,m}\right]^{-1} \approx \frac{1}{1-\beta}\left[\mathbf{1}_{N\times N} - \frac{\beta}{1-\beta}\left(\frac{\sum_i A_i w_i(n) w_i(m)}{\sum_i A_i w_i(n)}\right)_{n,m}\right].$$

Now suppose that $\left|\frac{\partial w_i(n)}{\partial \delta}\right|$ is uniformly bounded by some constant $d > 0$ for $n = 1, ..., N$. Then, the price effect $\frac{\partial \mathbf{p}}{\partial \delta}$ is of the order $\frac{1}{1-\beta}$: for any $n$,

$$\left|\frac{\partial \mathbf{p}(n)}{\partial \delta}\right| \leq \frac{1}{1-\beta} d - \frac{\beta}{(1-\beta)^2} \sum_{m=1}^{N} \frac{\sum_i A_i w_i(n) w_i(m)}{\sum_i A_i w_i(n)} d \leq \frac{D}{1-\beta}$$

where $D = \frac{d(1-2\beta)}{1-\beta}$.

## A.5. Proof of Proposition 3

Using the expression of $\kappa_{i,t}$ in (17), we obtain:

$$\begin{aligned}
\sum_{n=1}^{N_{i,t}} \frac{w_{i,t}(n)}{w_{i,t}(0)} \Gamma_{i,t}(n) &= \sum_{n=1}^{N_{i,t}} \mu_{i,t}(n)\Gamma_{i,t}(n) + \sum_{n=1}^{N_{i,t}} \Gamma_{i,t}(n)\left(\mu_{i,t}(n) - \Gamma_{i,t}(n)\right) \frac{\kappa_{i,t}}{1-\kappa_{i,t}}, \\
&= \sum_{n=1}^{N_{i,t}} \mu_{i,t}(n)\Gamma_{i,t}(n) + \left(\sum_{n=1}^{N_{i,t}} \mu_{i,t}(n)\Gamma_{i,t}(n) - \sum_{n=1}^{N_{i,t}} \Gamma_{i,t}^2(n)\right) \frac{\sum_{n=1}^{N_{i,t}} \mu_{i,t}(n)\Gamma_{i,t}(n)}{\gamma_{i,t} + \sum_{n=1}^{N_{i,t}} \Gamma_{i,t}^2(n) - \sum_{n=1}^{N_{i,t}} \mu_{i,t}(n)\Gamma_{i,t}(n)}, \\
&= \gamma_{i,t} \frac{\kappa_{i,t}}{1-\kappa_{i,t}}.
\end{aligned}$$

We can then rewrite (12) as:

$$\frac{w_{i,t}(n)}{w_{i,t}(0)} = \exp\left(\beta_{K,i,t}\right) \mu_{i,t}(n) \left[1 + \frac{1}{\gamma_{i,t}}\left(1 - \frac{\Gamma_{i,t}(n)}{\mu_{i,t}(n)}\right) \sum_{m=1}^{N_{i,t}} \frac{w_{i,t}(m)}{w_{i,t}(0)} \Gamma_{i,t}(r)\right], \quad \text{(A.5.2)}$$

or equivalently under the log-linear specification,

$$\left(\frac{w_{i,t}(n)}{w_{i,t}(0)}\right)^2 = \exp\left(\beta_{K,i,t}\right) \underbrace{\left(\frac{w_{i,t}(n)}{w_{i,t}(0)} \mu_{i,t}(n)\right)}_{\tilde{\mu}_{i,t}(n)} \left[1 + \left(1 - \exp\left(\mathbf{x}_t'(n)\Delta_{i,t}\right)\right) \underbrace{\frac{\sum_{m=1}^{N_{i,t}} \tilde{\mu}_{i,t}(m) \exp\left(\mathbf{x}_t'(m)\Delta_{i,t}\right)}{\gamma_{i,t}}}_{=: G_{i,t}}\right].$$

We then express

$$\tilde{\mu}_{i,t}(n) = \frac{\left(\frac{w_{i,t}(n)}{w_{i,t}(0)}\right)^2}{\exp\left(\beta_{K,i,t}\right)\left(1 + \left(1 - \exp\left(\mathbf{x}_t'(n)\Delta_{i,t}\right)\right) G_{i,t}\right)}.$$

Plugging this expression into the definition of $G_{i,t}$, we obtain:

$$\underbrace{\gamma_{i,t}\exp\left(\beta_{K,i,t}\right)}_{\tilde{\gamma}_{i,t}} G_{i,t} = \sum_{n=1}^{N_{i,t}} \frac{\left(\frac{w_{i,t}(n)}{w_{i,t}(0)}\right)^2 \exp\left(\mathbf{x}_t'(n)\Delta_{i,t}\right)}{1 + \left(1 - \exp\left(\mathbf{x}_t'(n)\Delta_{i,t}\right)\right) G_{i,t}}.$$

Plugging $G_{i,t}$ back into (A.5.2), we obtain the desired result.

# B. Empirics

## B.1. Data Construction

**Institutional Stock Holdings Data.** Following Koijen et al. (2024), we construct quarterly equity positions for U.S. institutional investors from FactSet Ownership, which standardizes SEC 13F filings for all managers that control at least \$100 million in assets under their discretionary management. The sample spans from 2000 Q1 through 2022 Q4. For each filing, we multiply reported shares by the corresponding share price to obtain dollar holdings, and we drop the two FactSet entity identifiers 0FSVG4-E and 000V4B-E, whose filings contain persistent data errors. FactSet's *investor subtype* codes initially sort managers into six broad categories: investment advisors/mutual funds, hedge funds, long-term investors (e.g., pensions, insurance firms), private banks, brokers, and a residual *other*. Since investment advisors comprise a substantial proportion of institutional investors, we further classify them into four subgroups: large-passive, large-active, small-passive, and small-active. Specifically, each quarter we first split investment advisors into two groups based on their total equity holdings. Within each group, we further split advisors at the median active share, which is half the sum of absolute deviations between an advisor's portfolio weights and the market weights across all stocks within their investment universe (Cremers and Petajisto, 2009).

Finally, we treat any stock in the bottom decile of market capitalization, or with missing fundamentals (book equity, sales, dividends, or beta), as an outside asset, and merge any institution with fewer than ten holdings or less than $10 million in total equity holdings into the household sector to ensure that the number of shares held adds up to the number of shares outstanding.

**Stock Characteristics.** We closely follow Koijen and Yogo (2019) and Koijen et al. (2024) to construct stock characteristics. Market equities are from FactSet Ownership. Dividends and stock returns come from the Center for Research in Security Prices (CRSP), while accounting fundamentals come from the Compustat. We limit our sample to U.S.-listed ordinary common equities (CRSP share codes 10 and 11) traded on NYSE, AMEX, or NASDAQ (exchange codes 1–3) with non-missing prices and shares outstanding. Each stock is characterized by log market equity, log book equity, operating profitability (operating income divided by book equity), investment (annual log asset growth), dividend to book equity (annual split-adjusted dividends divided by book equity), and market beta (estimated from a rolling 60-month regression of monthly excess returns on the market excess return). Profitability, investment, and market beta are winsorized at the 2.5th and 97.5th percentiles. Dividend to book equity is winsorized at the 97.5th percentile.

## B.2. Estimation

**Constructing IVs.** Depending on the underlying model (MNL or our model), the identification strategy relies on two sets of a vector of instruments and exogenous variables.

- **MNL - Investment Mandates**: We use the investment mandate IV (constructed following KY) and its square, alongside the exogenous asset characteristics.

- **Our Model - Investment Mandates + Higher Order Interactions**: We form all pairwise products of the investment mandate IV with each exogenous asset characteristic, as well as all second-order interactions among these characteristics. This generates a vector of 28 instruments that accommodate both nonlinear supply effects and cross-characteristic spillovers.

**Numerical Procedure.** Due to the nonlinearity of our asset demand model (16), convergence to a global optimum crucially depends on good initial values. To address this, we first estimate the log-linear MNL specification and progressively introduce complexity using the previous estimates as initial values in numerical computations. Specifically, for each $(i, t)$, we minimize the GMM objective using MATLAB's `fmincon` solver (interior-point algorithm with analytical gradients) and enforce KY's constraint $\beta_{\mathbf{me},i,t} < 1$. For efficiency, we employ a two-step GMM procedure: initially using an identity weighting matrix, then updating it to the optimal weighting matrix (the inverse of the asymptotic variance-covariance matrix of sample moments).

**(Step 1) Linear GMM for Log-linear MNL ($\Delta_{i,t} = 0$)**

**(1.1)** We retain only weights satisfying $w_{i,t}(n) \neq 0$ and $w_{i,t}(n) \geq \text{med}(w_{i,t}(n))$.

**(1.2)** Given $\mathbf{z}_{i,t}(n)$ with only the investment mandate IVs, we compute:

$$\widehat{\beta}_{i,t}^{\text{MNL, GMM0}} = \left( \sum_{n=1}^{N_{i,t}} \mathbf{x}_t'(n)\mathbf{z}_{i,t}(n)\mathbf{z}_{i,t}'(n)\mathbf{x}_t(n) \right)^{-1} \sum_{n=1}^{N_{i,t}} \mathbf{x}_t'(n)\mathbf{z}_{i,t}(n)\mathbf{z}_{i,t}'(n) \ln\left( \frac{w_{i,t}(n)}{w_{i,t}(0)} \right).$$

**(Step 2) Nonlinear GMM for Nonlinear MNL ($\Delta_{i,t} = 0$)**

**(2.1)** We set $\widehat{\beta}_{i,t}^{\text{MNL, GMM0}}$ as starting values and minimize the GMM objective with an identity weighting matrix, obtaining the first-step GMM estimates $\widehat{\beta}_{i,t}^{\text{MNL, GMM1}}$. We then update the weighting matrix.

**(2.2)** We re-run the optimization using $\widehat{\beta}_{i,t}^{\text{MNL, GMM1}}$ as starting values and the optimal weighting matrix, yielding the efficient second-step GMM estimates, denoted by $\widehat{\theta}_{i,t}^{\text{MNL}}$.

**(Step 3) Nonlinear GMM for Our Model ($\Delta_{i,t} \neq 0$)**

**(3.1)** We set initial values as

$$\beta_{i,t}^{\text{init}} = \widehat{\beta}_{i,t}^{\text{MNL}}, \quad G_{i,t}^{\text{init}} = \sum_{n=1}^{N_{i,t}} \left( \frac{w_{i,t}(n)}{w_{i,t}(0)} \right)^2, \quad \Delta_{i,t}^{\text{init}} = \mathbf{0}_{6\times 1},$$

and minimize the GMM objective with an identity weighting matrix to obtain first-step estimates $\widehat{\beta}_{i,t}^{\text{GMM1}}$. We then update the weighting matrix.

**(3.2)** We use the first-step estimates $\widehat{\beta}_{i,t}^{\text{GMM1}}$ as new initial values, re-estimate with the optimal weighting matrix, and obtain efficient second-step estimates, denoted by $\widehat{\theta}_{i,t}$.

**Computing Standard Errors.** We compute the standard errors for both $\widehat{\theta}_{i,t}^{\text{MNL}}$ and $\widehat{\theta}_{i,t}$ from the asymptotic GMM covariance matrix (we drop the $(i,t)$ subscripts):

$$\frac{1}{N}\left(G'WG\right)^{-1}G'W\Omega WG\left(G'WG\right)^{-1}$$

where the Jacobian of the moment conditions is $G = \frac{1}{N}\sum_{n=1}^{N}\mathbf{z}(n)\frac{\partial\varepsilon(n)}{\partial\theta'}$, the sample moment covariance matrix is $\Omega = \frac{1}{N}\sum_{n=1}^{N}\left(\varepsilon(n)-1\right)^2\mathbf{z}(n)\mathbf{z}(n)'$, and the optimal weighting matrix is $W = \Omega^{-1}$. The standard errors are then the square roots of diagonal elements.

In some cases, the matrix $G'WG$ is theoretically invertible but becomes nearly singular (i.e. its condition number is very high yet finite, e.g., $\approx 10^{10}$), making the direct inverse unstable and inflating variances. Against this numerical instability, we apply a small Tikhonov (ridge) regularization $\lambda \approx 10^{-4}$ as $G'WG = G'WG + \lambda\mathbf{I}$ before inverting $G'WG$. This adjustment is performed only in the covariance stage and it effectively curbs spurious large standard errors without affecting the parameter estimates.

## B.3. Masked-Asset Prediction

Gabaix et al. (2025a) highlight the importance of economically relevant benchmarks in defining success of one asset demand system over another. They propose a *masked asset* prediction test as a simple yet powerful benchmark that measures a model's out-of-sample predictive accuracy.

Inspired by their approach, initially applied across time and investors, we employ a similar masked-asset prediction exercise for our investor–quarter specific asset demand systems. This allows an objective comparison between the MNL and our richer substitution model (16), explicitly quantifying whether additional complexity translates into meaningful gains in out-of-sample predictive performance.

Specifically, for each $(i,t)$ portfolio, our procedure is as follows:

(**Step 1**) Mask the second largest asset weight $w_{i,t}^{(2)}(n)$ from the estimation sample.

(**Step 2**) Estimate:

(**i**) MNL ($\Delta_{i,t} = 0$).

(**ii**) Our model with $\Delta_{\mathbf{me},i,t}$ free.

(**iii**) Our model with $(\Delta_{\mathbf{me},i,t}, \Delta_{\mathbf{be},i,t}, \Delta_{\mathbf{marketbeta},i,t})$ free.

(**Step 3**) Predict $\ln\left(\frac{\widehat{w_{i,t}^{(2)}}(n)}{w_{i,t}(0)}\right)$ for the masked asset weight out of sample.

For each $(i,t)$, we compute RMSE and relative RMSE (rRMSE) measures:

$$\text{RMSE} = \sqrt{\sum_{i,t}\left(\frac{A_{i,t}}{\sum_i A_{i,t}}\left(w_{i,t}^{(2)}(n) - \widehat{w_{i,t}^{(2)}}(n)\right)\right)^2}$$

$$\text{rRMSE} = \frac{\text{RMSE}}{\sum_{i,t}\frac{A_{i,t}}{\sum_i A_{i,t}} w_{i,t}^{(2)}(n)}$$

where $A_{i,t}$ is the assets under investor $i$'s management in quarter $t$.

## B.4. Substitution Measure $\bar{\eta}_{i,t}$

We generalize the substitution measure (32) to one between two groups of assets.

**Corollary 3.** *Denote by* $\mathbf{A}$ *and* $\mathbf{B}$ *two groups of assets held by investor* $i$ *at time* $t$ *and* $\bar{\eta}_{i,t}(\mathbf{A},\mathbf{B})$ *the minus percentage change in* $Q_{i,t}(\mathbf{A}) = \sum_{n\in\mathbf{A}} Q_{i,t}(n)$ *due to* $1\%$ *overall increase in the prices of assets in* $\mathbf{B}$*. Then,*

$$\bar{\eta}_{i,t}(\mathbf{A},\mathbf{B}) = \sum_{n\in\mathbf{A}} \frac{Q_{i,t}(n)}{Q_{i,t}(\mathbf{A})} \sum_{j\in\mathbf{B}} \left[-\frac{\partial \mathbf{q}_{i,t}(n)}{\partial \mathbf{p}_t(j)}\right].$$

**Proof of Corollary 3.** *Suppose that the prices of assets in group* $\mathbf{B}$ *increase relatively by* $1\%$*. The absolute change in the demand for assets in group* $\mathbf{A}$ *due to the change in asset* $j$*'s price is* $\sum_{n\in\mathbf{A}} \frac{\partial \mathbf{q}_{i,t}(n)}{\partial \mathbf{p}_t(j)} \mathbf{Q}_{i,t}(n)$*. Consequently, the total change due to the overall price increase is* $\sum_{j\in\mathbf{B}} \sum_{n\in\mathbf{A}} \frac{\partial \mathbf{q}_{i,t}(n)}{\partial \mathbf{p}_t(j)} \mathbf{Q}_{i,t}(n) = \sum_{n\in\mathbf{A}} \mathbf{Q}_{i,t}(n) \sum_{j\in\mathbf{B}} \frac{\partial \mathbf{q}_{i,t}(n)}{\partial \mathbf{p}_t(j)}$ *(the chain rule), and the minus relative change is* $-\sum_{n\in\mathbf{A}} \frac{\mathbf{Q}_{i,t}(n)}{\mathbf{Q}_{i,t}(\mathbf{A})} \sum_{j\in\mathbf{B}} \frac{\partial \mathbf{q}_{i,t}(n)}{\partial \mathbf{p}_t(j)}$*. To obtain (32), we replace* $\mathbf{A}$ *and* $\mathbf{B}$ *by* $\mathcal{N}_{i,t}$*.* □

When $\mathbf{A} = \mathbf{B}$, $\bar{\eta}_{i,t}(\mathbf{A}, \mathbf{A})$ is interpreted as the price elasticity of asset group $\mathbf{A}$. Similar to asset-level own-price elasticity, $\bar{\eta}_{i,t}(\mathbf{A}, \mathbf{A})$ describes the extent to which investor adjusts her asset holdings in $\mathbf{A}$ when the overall asset price in $\mathbf{A}$ increases by $1\%$. Note that:

$$\bar{\eta}_{i,t}(\mathbf{A}, \mathbf{A}) = -\sum_{n\in\mathbf{A}} \frac{Q_{i,t}(n)}{Q_{i,t}(\mathbf{A})} \frac{\partial \mathbf{q}_{i,t}(n)}{\partial \mathbf{p}_t(n)} + \sum_{n\in\mathbf{A}} \frac{Q_{i,t}(n)}{Q_{i,t}(\mathbf{A})} \sum_{j\in\mathbf{A}: j\neq n} \left( -\frac{\partial \mathbf{q}_{i,t}(n)}{\partial \mathbf{p}_t(j)} \right). \quad \text{(B.4.3)}$$

When assets in $\mathbf{A}$ are substitutes, i.e., $\frac{\partial \mathbf{q}_{i,t}(n)}{\partial \mathbf{p}_t(j)} > 0$ for $n \neq j$, the second term on the RHS of (B.4.3) is negative, attenuating the weighted own-price elasticities in the first term on the RHS of (B.4.3). Consequently, group-level own-price elasticity $\bar{\eta}_{i,t}(\mathbf{A}, \mathbf{A})$ is smaller than the average of asset-level own-price elasticities. In contrast, complementarities among assets in $\mathbf{A}$ $\left( \frac{\partial \mathbf{q}_{i,t}(n)}{\partial \mathbf{p}_t(j)} < 0,\ n \neq j \right)$ amplify own-price elasticity when aggregating asset demand to group level. The gap between group-level and asset-level price elasticities and its direction are determined by the overall extent of substitutability and complementarity among assets in the group.