The Origins of the Factor Zoo: Investors Weakly Substitute Across Stocks*

Aditya Chaudhry and Carter Davis†

August 27, 2026

ABSTRACT

We propose a simple explanation for the "factor zoo" in expected stock returns: investors weakly substitute across individual stocks. In classical asset-pricing models, substitution is strong: investors readily reallocate across assets that comove, so differences in expected returns are explained by the few factors that create the most comovement across stocks. In contrast, we show that substitution is empirically weak using evidence from cross-sectional return predictability, cross-price multipliers, and cross-price elasticities. This weak substitution gives rise to the factor zoo: expected returns depend on many factors, including weak factors that create little comovement. Moreover, weak substitution rationalizes why equilibrium Sharpe ratios remain moderate even though weak factors contribute to expected returns. Additionally, weak substitution implies that misspecified substitution patterns create only small biases in price elasticity and multiplier estimates.

JEL classification: G11, G12

Keywords: Demand-System Asset Pricing, Cross-Price Elasticities, Factor Zoo, Factor Models

*We thank Ehsan Azarmsa, Darrell Duffie, Joey Engelberg, Julie Fu, Mahyar Kargar, Ralph Koijen, Jiacui Li, Stefan Nagel, Sangmin Oh, and Moto Yogo for helpful comments.

†The Ohio State University Fisher College of Business. Chaudhry email: chaudhry.127@osu.edu, Davis email: carter.davis@fisher.osu.edu.

1 Introduction

One of the central questions in asset pricing is: why do some assets have higher expected returns than others? Classical models offer a simple answer: comovement. Since idiosyncratic risks cancel out in a diversified portfolio, investors should only demand compensation for common risks that create comovement across assets. Thus, two assets that comove strongly with each other should earn similar expected returns. Otherwise an investor could earn extreme Sharpe ratios by buying the cheap asset and shorting the expensive one, pocketing the expected return spread while bearing little risk since the common movements cancel. Since a few common risk factors dominate comovement across stocks, in classical models those few factors should explain most of the cross-sectional variation in expected returns (Ross, 1976; Kozak et al., 2018a).

However, empirically a "zoo" of hundreds of factors is required to explain cross-sectional differences in expected stock returns (Cochrane, 2011; Lettau and Pelger, 2020; Lopez-Lira and Roussanov, 2023; Kelly et al., 2024). Moreover, even though many of these factors are "weak" — they create little comovement across assets — empirical Sharpe ratios remain moderate.

We propose an explanation for the factor zoo: investors weakly substitute across stocks. Substitution refers to how readily investors reallocate capital across stocks that comove when their expected returns diverge. In classical models where investors have mean-variance preferences, substitution is strong: if expected returns diverged between two stocks that strongly comove, investors would aggressively buy the cheap one and short the expensive one, since doing so barely raises portfolio volatility. In contrast, we show empirically that substitution is two to three orders of magnitude weaker than implied by unconstrained mean-variance optimization. Weak substitution arises because, in reality, investors face many considerations beyond portfolio expected return and volatility (e.g., uncertainty about expected returns or frictions such as benchmarking). These other considerations reduce the emphasis investors place on minimizing portfolio volatility, thereby reducing the incentive to reallocate between stocks that strongly comove.

Without aggressive reallocation across stocks that strongly comove, such stocks can sustain different expected returns in equilibrium. As a result, many factors, including weak factors that create little comovement, can contribute to expected return variation.

Crucially, this situation arises not from investors leaving profitable trades on the table, but because investors demand less compensation for comovement. Under weak substitution, investors place less emphasis on minimizing portfolio volatility, and so the equilibrium premium on non-diversifiable risk is small. As a result, expected returns are low and dispersed across many factors rather than high and concentrated in a few. Thus, unlike other potential explanations for the factor zoo (e.g., limits to arbitrage), weak substitution allows a factor zoo to coexist with moderate Sharpe ratios.

We establish that weak substitution gives rise to the factor zoo in two steps. First, we show theoretically that a factor zoo with moderate Sharpe ratios arises if and only if investors treat stocks as weak substitutes. Second, we provide evidence from three sources that substitution is empirically weak enough to generate the observed factor zoo: the cross-section of expected returns, cross-stock spillovers of demand shocks ("cross-price multipliers"), and cross-stock reallocations in holdings data ("cross-price elasticities").

Overall, our results suggest that comovement-based explanations for the cross-section of expected stock returns are fundamentally incomplete. Investors' willingness to substitute across stocks shapes which assets earn higher expected returns.

We begin by presenting a parsimonious model to formalize the link between substitution and equilibrium expected returns (in Section 2). In particular, we extend the model of Kozak et al. (2018a). A representative investor has mean-variance preferences but, crucially, is also subject to an ℓ_2 penalty that dissuades her from taking extreme positions in any single stock. This penalty arises in many different microfoundations, including uncertainty about expected returns or covariances, frictions such as benchmarking or adjustment costs, or behavioral biases such as correlation neglect or narrow framing. We label all of these deviations from mean-variance preferences and full information rational expectations as "portfolio-choice frictions".

These frictions introduce a reason — beyond risk aversion — for not fully exploiting

high expected return opportunities. Suppose the investor seeks to take a larger position in one stock because its expected return has idiosyncratically risen. The per unit cost to do so is comprised of a risk cost — due to raising portfolio volatility by taking on more idiosyncratic risk — and a friction cost — due to taking on larger positions.

The composition of this idiosyncratic position cost determines how strongly the investor substitutes across stocks. The overall degree of substitution is controlled by a single parameter, which we denote by ψ. Formally, ψ is the proportion of the total idiosyncratic position cost that comes from the risk cost:

$$\psi = \frac{\text{Risk Cost}}{\text{Risk Cost} + \text{Friction Cost}}.$$

Intuitively, in mean-variance preferences, substitution arises from the investor's desire to minimize portfolio volatility. Two assets are substitutable if the investor can swap one for the other without significantly altering portfolio volatility, which occurs if two assets strongly comove. Thus, the greater the investor's desire to minimize volatility — that is, the greater the risk cost — the greater is the degree of substitution. Note that while ψ captures the *overall* degree of substitution, the investor's willingness to substitute across two particular stocks also depends on the covariance between those stocks: substitution is asset-specific and our model allows for rich heterogeneity in substitution across stocks.[1]

The core comparative static we evaluate throughout this paper is: holding all else fixed (i.e., holding the total idiosyncratic position cost fixed), what happens as substitution weakens (i.e., as the friction cost replaces the risk cost so ψ decreases)? We show that as substitution weakens, the number of factors required to explain cross-sectional variation in expected stock returns grows and the maximum attainable Sharpe ratio shrinks.

When substitution is strong, investors aggressively trade away differences in expected returns between stocks that strongly comove, so such stocks must earn similar equilibrium expected returns. As a result, the few factors that contribute most to comovement

[1]For example, Apple and Google, which strongly comove, are always better substitutes than Apple and Caterpillar. But holding the covariance between any two stocks fixed, lower ψ means the investor is less willing to substitute between them.

between stocks also explain most differences in expected returns. However, as substitution weakens, stocks that strongly comove can earn different expected returns. As a result, expected-return variation becomes dispersed across many factors, including those that contribute little to return comovement. Moreover, the relationship is if and only if: weak substitution implies many factors are required to explain expected returns, and a high-dimensional factor structure in expected returns can only be sustained in equilibrium if investors weakly substitute across stocks.

Moreover, equilibrium Sharpe ratios remain moderate despite this dispersion in expected returns across many factors. As substitution weakens, the risk cost is replaced with the friction cost and so the investor cares less about minimizing volatility. As a result, the investor demands less compensation for bearing non-diversifiable risk, which lowers expected returns for all factors. Thus, the maximum attainable Sharpe ratio decreases, preventing near arbitrage opportunities.

We next provide empirical evidence that weak substitution gives rise to the factor zoo. As a benchmark for the degree of substitution, we use the ψ value that arises in the classical mean-variance benchmark with no ℓ_2 penalty: $\psi = 1$.

First, we empirically show that the factor structure in expected returns is consistent with weak substitution (in Section 4). We use Kelly et al. (2019) instrumented principal components analysis (IPCA) to estimate the factor structure in realized returns. In particular, we use as stock characteristics nonlinear transformations of 120 monthly stock characteristics built using Random Fourier Features (RFF). This procedure delivers 200 latent factors and associated factor variances that explain realized return comovement. Our results prove robust to using standard PCA (using the deflated heteroskedastic PCA of Zhou and Chen, 2025) on daily returns for individual stocks. In theory, under strong substitution, the few factors that contribute most to comovement (i.e., those with largest factor variances) should explain most variation in expected returns.

We test this theoretical prediction by running cross-sectional predictive regressions of future returns on stock-level factor loadings. We find that the explanatory power (R^2) of these regressions empirically rises much more slowly than it would under strong

substitution. While under the classical strong substitution benchmark the first 9 factors would explain 75% of cross-sectional expected return variation, empirically 139 factors are required. Theoretically, our model implies this situation can only arise under weak substitution. Based on the empirical dispersion in explanatory power across factors, we estimate the degree of substitution to be $\psi = 0.011$ — two orders of magnitude smaller than the benchmark $\psi = 1$.

Is such a weak degree of substitution realistic? Yes. Next, we directly measure the degree of substitution in four different ways, and find ψ estimates that are quantitatively consistent with the $\psi = 0.011$ our model requires to rationalize the empirical factor zoo.

First, we estimate the degree of substitution from the empirical distribution of factor-level Sharpe ratios (in Section 5). As explained above, our model implies that, ceteris paribus, weaker substitution implies lower equilibrium Sharpe ratios. This creates a natural "pick your poison" dilemma: either Sharpe ratios are counterfactually large or substitution is weak. Using a factor model estimated through 1999 and returns from 2000 through 2023, the cross-factor pattern of Sharpe ratios implies $\psi = 0.025$, far below the classical benchmark $\psi = 1$ and similar to the $\psi = 0.011$ our model requires to rationalize the empirical factor zoo. Given our estimated factor model, the maximum attainable out-of-sample annualized Sharpe ratio is 1.8. In contrast, if, ceteris paribus, substitution were as strong as in classical models ($\psi = 1$), then the maximum attainable Sharpe ratio would be 26.6, which is unrealistically high.

Second, we estimate the degree of substitution from the relative magnitudes of price multipliers at the stock ("micro") and aggregate ("macro") levels (in Section 6). At the micro level, a multiplier of 2 means that buying 1% of shares outstanding raises a stock's price by 2%. At the macro level, a multiplier of 2 means that a 1% purchase of the aggregate stock market raises the overall market by 2%. Strong substitution implies large cross-price multipliers, which capture how much a stock's price rises when 1% of shares outstanding of *another* stock are purchased. If micro multipliers are 1 and cross-price multipliers are 0.1, then buying 1% of the aggregate stock market raises its price by $1\% + 0.1\% \times (N - 1)$ where N is the number of stocks. For example, with 2,000

stocks this would yield a macro multiplier of about 200, far larger than previous empirical estimates (Gabaix and Koijen, 2021). We generalize this argument to a setting with heterogeneous cross-price multipliers; the intuition remains the same. This presents a "pick your poison" dilemma: either macro multipliers are unrealistically large or substitution is weak. The relative magnitudes of micro- and macro-multipliers from previous work imply $\psi = 0.0014$, three orders of magnitude smaller than the benchmark $\psi = 1$ and similar to the $\psi = 0.011$ our model requires to rationalize the empirical factor zoo.

Third, we estimate the degree of substitution from cross-price multipliers (in Section 7). We use three plausibly uninformed demand shocks from previous work to measure cross-price multipliers: changes in benchmarking intensity around Russell index reconstitutions (Pavlova and Sikorskaya, 2023), mutual fund flow induced trading (Lou, 2012), and order flow imbalance (Li and Lin, 2025). Consistent with previous work, we find that own-price multipliers are large: buying 1% of shares outstanding for a given stock raises its price by 1% to 3% across the three shocks. However, we find that cross-price multipliers are very small: two to four orders of magnitude smaller than own-price multipliers. Through the lens of our model, these small cross-price multipliers map to ψ values from 0.001 to 0.007, two to three orders of magnitude smaller than the benchmark $\psi = 1$ and similar to the $\psi = 0.011$ our model requires to rationalize the empirical factor zoo.

Fourth, we estimate the degree of substitution from cross-price elasticities in institutional investor holdings data (in Section 8). We estimate cross-price elasticities of demand for institutional investors in the U.S. stock market using holdings data from the SEC Form 13F. We examine how investors adjust their holdings for a particular stock in response to exogenous price variation in other stocks. We find that cross-price elasticities are small — two to four orders of magnitude smaller than own-price micro elasticities. Through the lens of our model, these small cross-price elasticities map to $\psi = 0.001$, three orders of magnitude smaller than the benchmark $\psi = 1$ and similar to the $\psi = 0.011$ our model requires to rationalize the empirical factor zoo.

While we do not take a strong stance on the primitive source of weak substitution, we provide one potential mechanism that can quantitatively rationalize our results: optimal

shrinkage of portfolio weights due to estimation error. An investor who seeks to maximize her out-of-sample Sharpe ratio but who understands that she faces estimation error in expected returns and covariances will optimally choose to shrink extreme position sizes (Jorion, 1986; DeMiguel et al., 2009a,b). We find that the optimal level of shrinkage is quite high and implies a weak degree of substitution around $\psi = 0.019$, similar to the $\psi = 0.011$ our model requires to rationalize the empirical factor zoo.

Thus, we provide a parsimonious explanation for the factor zoo that is quantitatively consistent with the empirical behavior of prices and quantities. We do not necessarily rule out all other potential explanations for the factor zoo. Nevertheless, we make two observations. First, our explanation does not require unobservable "dark matter" like hedging demand for non-tradable risks (Merton, 1973), and can be rationalized with Sharpe-ratio maximizing behavior (Chen et al., 2019). Second, while other explanations for the factor zoo — like hedging demand, limits to arbitrage (Novy-Marx and Velikov, 2016; Chu et al., 2020; Gonçalves et al., 2025), or priced idiosyncratic risk (Merton et al., 1987; Dello Preite et al., 2024) — also raise the maximum attainable Sharpe ratio, weaker substitution *lowers* Sharpe ratios (Kozak et al., 2018a). Thus, weak substitution explains how a factor zoo can arise without creating near arbitrage opportunities.

More generally, our results provide guidance on how to interpret observed portfolios and evaluate positive portfolio choice models. Recent work in demand-system asset pricing argues that models that impose limited or homogeneous substitution across assets can bias own-price elasticities downward, thereby overstating the role of flows in moving prices (Chaudhary et al., 2023; Fuchs et al., 2023, 2025; Haddad et al., 2025). In contrast, our results lend support to such models: they may over-restrict substitution, but they do so in a way that is broadly consistent with the data, and the resulting own-price elasticity estimates are likely credible. In Sections 7.3.3 and 8.3.3 we show that, given our small estimates of ψ, failing to control for substitution creates only small biases in own-price elasticity and multiplier estimates.

The rest of the paper proceeds as follows. Section 2 develops our theoretical framework. Section 3 describes the data used in the paper. Section 4 empirically demonstrates

that far more factors are required to explain cross-sectional variation in expected returns than would theoretically be required under strong substitution. Section 5 measures the degree of substitution using the distribution of factor-level Sharpe ratios. Section 6 measures the degree of substitution from the relative magnitudes of market-level macro and stock-level micro multipliers. Sections 7 and 8 estimate the degree of substitution from cross-price multipliers and cross-price elasticities, respectively. Section 9 demonstrates that Sharpe-maximizing behavior in the presence of estimation error can quantitatively rationalize the weak degree of substitution we find. Section 10 concludes.

1.1 Related Literature

This paper relates to three strands of literature: work demonstrating that many factors are required to explain cross-sectional variation in expected stock returns, work studying why investors do not trade away these expected return differences, and work on demand-system asset pricing.

First, a large literature studies cross-sectional variation in expected stock returns (Fama and French, 1993, 2015; Hou et al., 2014; Barillas and Shanken, 2018). Recent work argues that high-dimensional factor models explain much more cross-sectional expected return variation than more parsimonious models (Lettau and Pelger, 2020; Gu et al., 2021; Lopez-Lira and Roussanov, 2023; Kelly et al., 2024; Dello Preite et al., 2024; Baba Yara et al., 2021; Bessembinder et al., 2025). Our contribution to this literature is to provide an economic rationale for how such a "factor zoo" can emerge in equilibrium: we show theoretically that a factor zoo arises if and only if substitution is weak.

Second, previous work studies how portfolio-choice frictions can give rise to a factor zoo by preventing investors from trading away differences in expected returns between stocks. One strand of this literature argues that uncertainty about expected returns can prevent investors from trading away the expected returns earned by anomaly portfolios (Martin and Nagel, 2019; Da et al., 2024; Gonçalves et al., 2025). A separate strand emphasizes trading frictions like short sales constraints (Jones and Lamont, 2002; Nagel,

2005; Drechsler and Drechsler, 2016; Chu et al., 2020) or trading costs (Novy-Marx and Velikov, 2016; Patton and Weller, 2020; Frazzini et al., 2018). These are the types of portfolio-choice frictions that the ℓ_2 penalty in our model captures in reduced form.[2] Our explanation for the factor zoo differs from these existing explanations because weak substitution is fundamentally about risk aversion being small, not about these portfolio-choice frictions being large. More severe frictions lower own-price elasticities and so raise Sharpe ratios. In contrast, lower risk aversion can weaken substitution while holding fixed own-price elasticities, thereby giving rise to a factor zoo while lowering Sharpe ratios.

Third, we contribute to the literature on demand system asset pricing. A large literature finds that own-price elasticities are small (Koijen and Yogo, 2019; Gabaix and Koijen, 2021; Davis et al., 2025; Haddad et al., 2022; van der Beck, 2022), or equivalently that own-price multipliers are large (Shleifer, 1986; Harris and Gurel, 1986; Chang et al., 2014; Hartzmark and Solomon, 2022; Li, 2022; Huebner, 2023; Pavlova and Sikorskaya, 2023; Chaudhry and Li, 2025). However, in theory, failing to properly account for substitution patterns across assets can bias elasticity and multiplier estimates (Haddad et al., 2025; Chaudhary et al., 2023; Fuchs et al., 2023, 2025). Indeed, previous work shows that properly accounting for substitution is quantitatively important in corporate bond, Treasury, and currency markets (Chaudhary et al., 2023; Jansen et al., 2024; An and Huber, 2024; Chaudhary et al., 2026). Our contribution to this literature is to measure the degree of substitution in the cross section of stocks. While previous work suggests that substitution may be far from perfect in the cross section of stocks (Wurgler and Zhuravskaya, 2002), this paper is the first to directly measure cross-price elasticities and cross-price multipliers for individual stocks. We find substitution across stocks is weak, and, accordingly, demonstrate that biases in elasticity and multiplier estimates are small. We also establish a link between the degree of substitution across assets and the number of factors required to explain cross-sectional variation in expected returns.

[2]In contemporaneous work Chaudhary et al. (2026) study the implications of the tradeoff between portfolio volatility versus portfolio-choice frictions for comovement between yields and arbitrage spreads in fixed income markets.

2 Theoretical Framework

In this section, we present a simple model to characterize the relationship between stock-level substitution and equilibrium expected returns. At a high level, the model features a representative investor who has two motives for not taking large positions in high expected return assets: she dislikes portfolio volatility due to risk-aversion, and she also faces a direct ℓ_2 penalty on large position sizes. This penalty captures in reduced-form many different portfolio-choice frictions. Substitution across stocks is governed by how strongly the investor weighs portfolio volatility relative to the portfolio-choice frictions: lowering the weight on portfolio volatility weakens substitution. We show that weaker substitution increases the number of return factors required to explain cross-sectional expected return variation and simultaneously lowers the maximum attainable Sharpe ratio.

2.1 Model Setup

We extend the setup from Kozak et al. (2018a).

Assets There are N stocks and two periods. Each stock n pays a risky dividend in the second period:

$$D_n = g_n + \boldsymbol{\theta}_n' \boldsymbol{F} + \epsilon_n.$$

$\boldsymbol{F} \in \mathbb{R}^{K+1}$ is a vector of common, orthogonal factors with $\mathbb{V}[\boldsymbol{F}] = \mathbf{E} \equiv \text{diag}(e_k)_{k=0}^K$. The factor variances are in decreasing order: $e_0 \geq \cdots \geq e_K$. ϵ_n is the idiosyncratic return with $\mathbb{V}[\epsilon_n] = \sigma^2$.

The factor loadings $\boldsymbol{\theta}_n$ are orthonormal: $\boldsymbol{\Theta} = [\boldsymbol{\theta}_n']_{n=1}^N$ satisfies $\boldsymbol{\Theta}'\boldsymbol{\Theta} = \boldsymbol{I} \in \mathbb{R}^{K+1 \times K+1}$. The first factor $k = 0$ is a "market" factor, so the first column of $\boldsymbol{\Theta}$ is $[1/\sqrt{N}]_{n=1}^N$. This market factor does not contribute to cross-sectional variation. Since the first factor is a market factor, orthonormality implies that all factors $k > 0$ are zero-exposure long-short portfolios: all columns $k > 0$ of $\boldsymbol{\Theta}$ satisfy $\boldsymbol{\iota}'\boldsymbol{\Theta}_{k,\cdot} = 0$, where $\boldsymbol{\iota}$ is a vector of ones (as in

Kozak et al., 2018a).

Let P_n be the price of asset n. Define $\boldsymbol{D} = [D_n]_{n=1}^N$ and $\boldsymbol{P} = [P_n]_{n=1}^N$ to be the vectors of dividends and prices across all stocks.

Each risky asset has supply of $1/N + \boldsymbol{\theta}_n'\boldsymbol{\pi}$ shares, where $\boldsymbol{\pi} \in \mathbb{R}^{K+1}$ controls the supply of each factor.[3] This supply should be interpreted as the residual supply absorbed by the representative investor: the total fixed supply less the demand from noise traders outside the model. Higher π_k means a larger supply of that factor must be absorbed by the representative investor.

The exogenous risk-free rate is normalized to zero.

Agents and Preferences There is a representative investor who has mean-variance preferences over terminal wealth, but faces an ℓ_2 penalty on the size of her positions:

$$\max_{\boldsymbol{Q}} \boldsymbol{Q}'\mathbb{E}\left[\boldsymbol{D} - \boldsymbol{P}\right] - \frac{\gamma}{2}\boldsymbol{Q}'\mathbb{V}\left[\boldsymbol{D}\right]\boldsymbol{Q} - \frac{\lambda}{2}\left|\left|\boldsymbol{Q}\right|\right|_2^2. \tag{1}$$

For analytical simplicity, we follow Kozak et al. (2018a) in using the mean-variance formulation associated with CARA preferences over dollar wealth. The main difference between this model and Kozak et al. (2018a) is that the latter has $\lambda = 0$. Appendix A derives the entire model starting with CRRA preferences over wealth and shows that all of our theoretical results continue to hold.

Moreover, our theoretical results do not depend on this specific functional form for preferences. We find similar results in a generalized model in Appendix B.4 that features non-pecuniary preferences (representing hedging demand, sentiment, or tastes for particular assets) and shrinkage toward an arbitrary target portfolio (e.g., a benchmark).

An ℓ_2 penalty arises in many realistic models of portfolio choice. There are at least three broad classes of microfoundations that give rise to this behavior.

First, the classic motivation for penalizing large weights is uncertainty about expected returns or covariances. A rational Bayesian agent will optimally shrink portfolio weights

[3] The supply of each factor $k > 0$ is $\boldsymbol{\Theta}_k\left(\frac{1}{N}\boldsymbol{\iota} + \boldsymbol{\Theta}\boldsymbol{\pi}\right) = \pi_k$ by the orthonormality of $\boldsymbol{\Theta}$ and the fact that $\boldsymbol{\iota}'\boldsymbol{\Theta}_{k,\cdot} = 0$.

toward a conservative benchmark, with the penalty strength governed by the degree of posterior variance (Barry, 1974; Klein and Bawa, 1976; Frost and Savarino, 1986; Black and Litterman, 1990; Pástor, 2000; Ledoit and Wolf, 2004; Kozak et al., 2018b).[4] Moreover, since empirical estimates of expected returns and covariances involve significant estimation error, previous work shows that this shrinkage improves out-of-sample portfolio performance (Jorion, 1986; DeMiguel et al., 2009a,b).

Second, a penalty for large weights arises from many frictions. Inventory costs directly penalize large positions (Amihud and Mendelson, 1980; Ho and Stoll, 1981). Benchmarking (Roll, 1992; Petajisto, 2009; Pavlova and Sikorskaya, 2023) and adjustment costs (Gârleanu and Pedersen, 2013, 2016) penalize deviations from a target portfolio (either the benchmark index or the lagged portfolio). Even short-sales constraints can be interpreted through shrinkage (Jagannathan and Ma, 2003).

Third, behavioral biases can also give rise to this penalty. Correlation neglect leads investors to act as if the covariance matrix is more diagonal than in reality (Ellis and Piccione, 2017; Enke and Zimmermann, 2019; Guo and Wachter, 2025). Narrow framing induces investors to consider each asset as a separate mental account, leading them to dislike large positions in any asset regardless of its covariance properties (Barberis et al., 2006; Barberis and Huang, 2009). Lastly, naive diversification pushes investors to hold equally sized positions across all (risky) assets (Benartzi and Thaler, 2001).

Though these mechanisms appear diverse, they share a common implication: they all introduce "portfolio-choice frictions" that give investors an additional reason — beyond risk aversion — to not exploit high expected return opportunities. Crucially, these portfolio-choice frictions are unrelated to covariances.

Moreover, our goal is not to isolate a particular microfoundation for the ℓ_2 penalty. Indeed, many of these mechanisms likely operate simultaneously. Rather, our goal is to characterize how the relative strength of volatility concerns versus frictions weakens investors' willingness to substitute across stocks and shapes equilibrium expected returns.

[4]This penalty can also arise from non-Bayesian uncertainty, such as from ambiguity aversion (Maenhout, 2004; Garlappi et al., 2007).

2.2 Optimal Portfolio Choice & Elasticities

We now define the own- and cross-price elasticities of demand and develop the intuition for how they shape the cross section of expected returns, which the next section formalizes.

The investor's optimal portfolio choice is given by:

$$\boldsymbol{Q} = \left[\gamma \boldsymbol{\Theta} \mathbf{E} \boldsymbol{\Theta}' + \left(\lambda + \gamma \sigma^2\right) \mathbf{I}\right]^{-1} \mathbb{E}\left[\boldsymbol{D} - \boldsymbol{P}\right]. \tag{2}$$

By the Woodbury matrix identity, the investor's own-price and cross-price elasticities for all stocks are given by the following elasticity matrix:

$$\begin{aligned} \boldsymbol{\zeta} &\equiv -\left[\gamma \boldsymbol{\Theta} \mathbf{E} \boldsymbol{\Theta}' + \left(\lambda + \gamma \sigma^2\right) \mathbf{I}\right]^{-1} \\ &= -\frac{1}{\lambda + \gamma \sigma^2} \mathbf{I} + \frac{\gamma}{\left(\lambda + \gamma \sigma^2\right)^2} \boldsymbol{\Theta} \left[\mathbf{E}^{-1} + \frac{\gamma}{\lambda + \gamma \sigma^2} \mathbf{I}\right]^{-1} \boldsymbol{\Theta}'. \end{aligned} \tag{3}$$

We follow Haddad et al. (2025) and label the derivative of the level of holdings with respect to the level of prices as the elasticity matrix, since these are the natural units to measure price sensitivity in our setting. In Appendix A we derive the full model starting from CRRA preferences and define elasticities with log-quantity and log-price units (the natural units in that setting): the key expressions that we map to the data are unchanged.

The "micro" (Gabaix and Koijen, 2021) or "relative" (Haddad et al., 2025) own-price elasticity is[5]

$$\zeta^{Own} \equiv \frac{1}{\lambda + \gamma \sigma^2}. \tag{4}$$

The micro elasticity captures how aggressively the investor exploits a purely idiosyncratic expected return deviation. Consider two stocks n and m that have the same factor

[5]For analytical simplicity, the main model uses share quantities and dollar expected returns, while the empirics use log shares and log prices. Appendix A derives the log-share elasticities from a standard CRRA problem. Their cross-to-micro ratio converges to (6) as N becomes large. The appendix also allows for a common dynamic price pass-through (i.e., a 1% change in price does not necessarily create a 1% change in the one-period expected return): it changes the levels of the elasticities but not their ratio.

loadings but different expected returns.[6] Intuitively, the investor would want to long the high expected return stock and short the low expected return stock. The cost (i.e., the per-unit compensation the investor requires) to do this idiosyncratic trade is

$$\mathcal{C}^{idio} \equiv \underbrace{\gamma\sigma^2}_{\text{Risk Cost}} + \underbrace{\lambda}_{\text{Friction Cost}}, \tag{5}$$

which reflects two components.[7] First, the risk cost $\gamma\sigma^2$ compensates the investor for the additional portfolio volatility this trade creates, which arises purely from idiosyncratic volatility since the stocks' common movements cancel out because they have the same factor loadings. Second, the friction cost λ compensates the investor for the larger positions she will hold as a result of this trade. The micro elasticity is the inverse of this cost: higher $\mathcal{C}^{idio}$ makes the investor less willing to take larger idiosyncratic positions and so renders demand more inelastic.

The cross-price elasticities $\zeta^{Cross}_{n,m} \equiv \frac{\partial Q_n}{\partial P_m}$ are the off-diagonal elements in (3):

$$\frac{\zeta^{Cross}_{n,m}}{\zeta^{Own}} = \sum_{k=0}^{K} \theta_{n,k}\theta_{m,k}\frac{\psi\tilde{e}_k}{1+\psi\tilde{e}_k} \tag{6}$$

$$\psi \equiv \frac{\gamma\sigma^2}{\lambda+\gamma\sigma^2}. \tag{7}$$

$\tilde{e}_k = e_k/\sigma^2$ are the factor variances scaled by idiosyncratic variance. Cross price elasticities reflect how aggressively the investor reallocates from stock n to stock m when the expected return on stock m changes.

Which stocks the investor substitutes between is determined by the stocks' factor loadings. Two stocks have a larger cross-price elasticity (i.e., are more substitutable) if they have similar factor loadings (i.e., $\theta_{n,k}$ and $\theta_{m,k}$ have the same sign), especially for

[6] Formally, the micro elasticity measures the change in quantity demanded for stock n relative to another stock with the same factor loadings m due to a relative price change: $\zeta^{Own} \equiv -\frac{\partial(Q_n - Q_m)}{\partial P_n}$.

[7] Consider stocks n and m with identical factor loadings and let the investor trade q shares in each leg: long q shares of stock n, short q shares of stock m. From (1), this position costs her $C(q) = \left(\gamma\sigma^2 + \lambda\right) q^2 = cq^2$, since the common movements cancel and only the idiosyncratic risk and the size of the two positions are penalized. The first-order condition $\Delta\mu = C'(q) = 2 \cdot c \cdot q$ states that the investor requires compensation of c per gross unit of the trade (i.e., trading q shares in each leg involves gross exposure of $2q$ shares).

the high-variance factors. Intuitively, if the price of stock m goes up, the investor will want to substitute to "similar" stocks — those that comove strongly with stock m.

The overall degree of substitution is controlled by the parameter ψ in (7). For any pair of stocks, smaller ψ indicates smaller cross-price (relative to micro) elasticities and weaker substitution. ψ captures how strongly the investor weighs portfolio volatility relative to portfolio-choice frictions. It is a unitless number between 0 and 1 that reflects the proportion of the idiosyncratic position cost $\mathcal{C}^{idio}$ that comes from the risk cost $\gamma\sigma^2$.

Intuitively, large ψ indicates that the investor places significant weight on minimizing portfolio volatility. In this case, the investor will fund purchases of stock n using funds from stocks that comove strongly with n, since doing so minimizes volatility. This is the strong substitution case. We define the classical benchmark as $\psi = 1$, which arises when $\gamma\sigma^2/\lambda \to \infty$ (e.g., as in standard mean-variance preferences where $\lambda = 0$).

In contrast, small ψ indicates that the investor places more weight on frictions. In this case, the investor will fund purchases of stock n using funds from many other stocks (and the risk-free asset) in order to avoid taking extreme (short) positions in any particular stock. This is the weak substitution case.

Thus, the main comparative static we study is: what happens if, ceteris paribus, substitution is weaker? That is, what happens if we hold fixed the total idiosyncratic position cost ($\mathcal{C}^{idio} = \gamma\sigma^2 + \lambda$), reallocate its composition from risk ($\gamma\sigma^2$) to frictions (λ), and so move from the classical benchmark ($\psi = 1$) toward the weak substitution regime ($\psi \to 0$)? This comparative static isolates the impact of substitution: it changes how the investor funds her trades — from close comovers versus from many stocks — without changing how aggressively she exploits expected return deviations overall. In other words, this comparative static holds the micro elasticity fixed and shrinks the cross-price elasticities.[8]

[8] In terms of the model's primitives, this comparative static lowers γ while adjusting λ to maintain the fixed total idiosyncratic position cost: $\lambda = \mathcal{C}^{idio} - \gamma\sigma^2$.

2.3 Equilibrium Expected Returns

As shown in Appendix B.1.1, market clearing

$$\boldsymbol{Q} = \frac{1}{N}\boldsymbol{\iota} + \boldsymbol{\Theta}\boldsymbol{\pi}$$

yields the following equilibrium expected returns as the number of assets N grows large:

$$\mathbb{E}[\boldsymbol{D}] - \boldsymbol{P} = \frac{1}{\zeta^{Own}}\boldsymbol{\Theta}\left[\pi_k\left(\psi\tilde{e}_k + 1\right)\right]_{k=0}^{K}. \tag{8}$$

Consider simulating this economy many times and drawing a new factor supply vector $\boldsymbol{\pi}$ each time, where each element π_k is i.i.d. both within and across draws. Appendix B.1.2 proves that, across many realizations, the proportion of cross-sectional expected return variance explained by (exposures to) the k-th factor (for $k > 0$) is on average

$$f_k \equiv \frac{\left(\psi\tilde{e}_k + 1\right)^2}{\sum_{j=1}^{K}\left(\psi\tilde{e}_j + 1\right)^2}. \tag{9}$$

Note that f_k increases with the scaled factor variance $\tilde{e}_k$. Thus, the factors with higher variances — those that explain more comovement across stocks — explain more of the cross-sectional variation in expected returns, as in Kozak et al. (2018a).

However, the relative importance of the high-variance factors decreases as substitution weakens. If substitution is strong — so ψ approaches 1 — we have:

$$f_k \xrightarrow{\psi\to 1} \frac{\left(1 + \tilde{e}_k\right)^2}{\sum_{j=1}^{K}\left(1 + \tilde{e}_j\right)^2}. \tag{10}$$

In this case, the higher variance factors explain a greater proportion of expected return variation. Intuitively, since the higher variance factors create more comovement across stocks, they contribute more to portfolio volatility, and so the investor demands more compensation for exposure to these factors. This situation arises when the risk cost is large relative to the friction cost and so the investor cares a lot about portfolio volatility

(and so about covariances) .

By contrast, if substitution is weak ($\psi \to 0$), we have:

$$f_k \xrightarrow{\psi \to 0} \frac{1}{K}.$$

In this case, all factors explain the same proportion of expected return variation, regardless of their variances. Intuitively, shutting down substitution leads the investor to not demand compensation for comovement across stocks. Instead, the investor just demands compensation for holding large quantities of a given factor (since $\lambda > 0$). This situation arises when the risk cost is small relative to the friction cost and so the investor places little emphasis on portfolio volatility (and so on covariances).

Thus, as substitution weakens, so does the link between equilibrium expected returns and covariances. Weak substitution spreads out expected return variation across many factors, including those that explain little comovement across realized returns. Therefore, a factor zoo in expected stock returns arises if and only if substitution is weak, as formalized in the following proposition (proven in Appendix B.1.3).

Proposition 1 (A Factor Zoo If and Only If Substitution is Weak)**.** *Let the factor variances be in decreasing order ($e_0 \geq \cdots \geq e_K$). The average proportion of cross-sectional expected return variation explained by the first J factors (after the market factor) $\sum_{k=1}^{J} f_k$ is increasing in ψ. Equivalently, the number of factors required to explain a given proportion of expected return variation increases as ψ decreases.*

Crucially, weak substitution does not create near arbitrage opportunities (i.e., very high Sharpe ratios), as established in the next proposition (proven in Appendix B.2).

Proposition 2 (Weaker Substitution Lowers Sharpe Ratios)**.** *The maximum attainable squared Sharpe ratio in this economy is*

$$\textit{Max Sharpe Ratio}^2 = \left(\mathcal{C}^{idio}\right)^2 \sum_{k=0}^{K} \pi_k^2 \frac{(1+\psi\tilde{e}_k)^2}{(\tilde{e}_k+1)\,\sigma^2}, \tag{11}$$

which decreases as ψ decreases and substitution weakens (as ψ decreases while holding the idiosyncratic position cost $\mathcal{C}^{idio}$ fixed).

Intuitively, lowering ψ while holding the idiosyncratic position cost $\mathcal{C}^{idio}$ fixed reallocates the cost from risk to frictions. As a result, the investor demands less compensation for bearing non-diversifiable risk, which lowers expected returns for all factors (as can be seen in (8)). Thus, the maximum attainable Sharpe ratio decreases.

Figure 1 graphically illustrates both propositions. The vertical axis measures the risk cost $\gamma\sigma^2$ while the horizontal axis measures the friction cost λ. The dashed blue "iso-cost" curve displays all combinations of risk and friction costs that yield the same total idiosyncratic position cost $\mathcal{C}^{idio}$, and so the same micro own-price elasticity. The classical benchmark is the point where the blue line intersects the vertical axis (i.e., where $\lambda = 0$).

The solid red line is an "iso-substitution" curve: it displays all risk and friction cost combinations that yield the same degree of substitution ψ. Moving down along the iso-cost curve replaces the risk cost with the friction cost, and so lowers ψ and weakens substitution. Thus, per Proposition 1, moving down along the iso-cost curve raises the number of factors required to explain cross-sectional expected return variation.

The solid green line is an "iso-max Sharpe" curve: it displays all risk and friction cost combinations that yield the same maximum attainable Sharpe ratio. Moving down along the iso-cost curve replaces the risk cost with the friction cost, which reduces how much the investor cares about minimizing volatility, and so lowers compensation required to bear non-diversifiable risk. Thus, per Proposition 2, moving down along the iso-cost curve lowers the maximum attainable Sharpe ratio.

These two solid lines create four quadrants that represent whether or not a factor zoo exists and whether or not the maximum attainable Sharpe ratio is unrealistically high. The green shaded area indicates the region where a factor zoo with moderate Sharpe ratios arises. Crucially, these two phenomena coexist if and only if substitution is weak. Classical models without frictions in which comovement alone shapes the cross section of expected returns cannot rationalize a factor zoo with moderate Sharpe ratios.

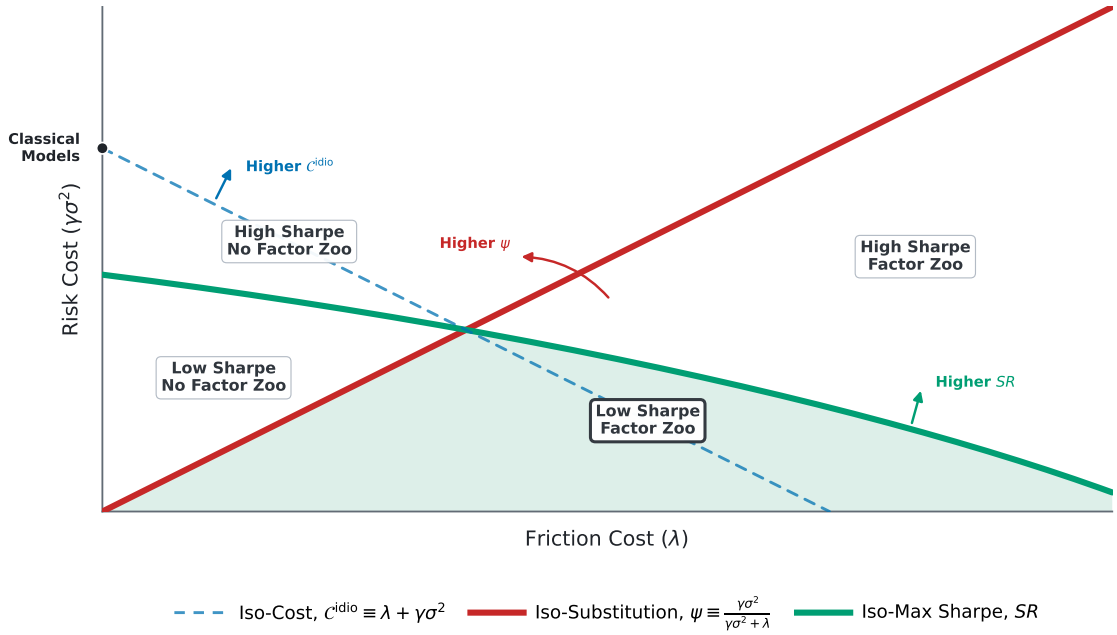


Figure 1: The blue iso-cost curve holds fixed total idiosyncratic position cost $\mathcal{C}^{idio} \equiv \lambda + \gamma\sigma^2$, and thus the micro own-price elasticity. The red iso-substitution curve holds fixed the degree of substitution ψ. The green iso-max Sharpe curve holds fixed the maximum attainable Sharpe ratio. Moving down and to the right along a blue curve replaces the risk cost with the friction cost. As a result, substitution weakens, expected returns spread across more factors, and the maximum attainable Sharpe ratio falls. The green shaded area indicates the region where a factor zoo with moderate Sharpe ratios arises.

2.4 Comparison to Other Explanations for the Factor Zoo

We do not necessarily rule out all other potential explanations for the factor zoo. Nevertheless, we make two observations. First, our explanation does not require unobservable "dark matter" like hedging demand for non-tradable risks (Merton, 1973; Chen et al., 2019). Second, whereas many other channels that give rise to a factor zoo — like hedging demand, sentiment (Kozak et al., 2018a), or priced idiosyncratic risk (Merton et al., 1987; Dello Preite et al., 2024) — also raise the maximum attainable Sharpe ratio, weaker substitution *lowers* the maximum attainable Sharpe ratio (Kozak et al., 2018a). Intuitively, these channels create stock-specific demands unrelated to covariances. These are isomorphic to large stock-specific supply shocks that create large expected returns unrelated to exposures to tradable risk factors. As a result, investors can earn high Sharpe ratios in the space of tradable assets. The generalized model featuring non-pecuniary preferences in Appendix B.4 formalizes this argument in Proposition 3.

Similarly, other explanations generate a factor zoo by adding limits to arbitrage that prevent investors from exploiting high Sharpe ratio opportunities (Novy-Marx and Velikov, 2016; Chu et al., 2020; Gonçalves et al., 2025). As these limits grow stronger, the expected returns on weak factors grow larger, thereby allowing weak factors to contribute more to cross-sectional expected return variation. However, in doing so, these limits to arbitrage also raise the maximum attainable Sharpe ratio and can create near arbitrage opportunities. This situation corresponds to starting at the "Classical Models" point in Figure 1 and moving directly right by increasing λ.

In contrast to all these other explanations, weak substitution *lowers* expected returns on all factors, especially those that contribute most to comovement. In doing so, weak substitution simultaneously generates a factor zoo while keeping Sharpe ratios moderate. This situation corresponds to moving down along the iso-cost curve in Figure 1.

3 Data and Factor Structure in Returns

We use three sets of data in this paper: stock characteristics and returns, plausibly exogenous demand shocks, and institutional holdings data. Table 1 displays the summary statistics for the key variables.

				Percentiles						
	Obs	Mean	StDev	1%	5%	25%	50%	75%	95%	99%
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Quarterly Return (%)	479,317	4.63	25.21	-51.54	-31.33	-8.59	2.98	15.20	44.19	87.31
OFI (%)	330,138	0.25	3.87	-10.71	-4.86	-1.10	0.03	1.38	5.97	13.50
FIT (%)	470,178	0.06	0.56	-1.33	-0.69	-0.14	0.00	0.19	0.97	2.01
ΔBMI (%)	25,461	0.08	0.66	-1.84	-0.79	-0.15	0.01	0.24	1.22	2.35
Shares Held (Rescaled, Basis Points)	59,174,552	65.70	459	0.00	0.03	0.49	3.30	19.02	190	848
Log Market Equity	59,174,552	8.50	1.93	4.47	5.52	7.08	8.39	9.87	11.88	12.73
Koijen-Yogo (2019) Instrument	59,174,552	8.01	1.26	4.28	5.65	7.33	8.19	8.94	9.69	10.08

Table 1: **Summary statistics**

This table reports summary statistics. Column (1) reports the total number of stock-quarter observations for the quarterly return and the OFI, FIT, and ΔBMI demand shocks, and the total number of institution-stock-quarter observations for the Shares Held (which has been rescaled such that each stock has one share outstanding), log market equity, and the Koijen and Yogo (2019) market equity instrument. As described in the paper, we have rescaled ΔBMI so it is expressed as a fraction of total shares outstanding. The sample for the three demand shocks spans 1998 to 2018 for ΔBMI, and 1993 to 2022 for FIT and OFI. The BMI sample is limited by the period for which we observe Russell index constituents. The OFI sample is limited by the availability of the signed OFI measure from WRDS. We constrain the FIT sample to match that of OFI. The sample for the holdings data spans 1980 to 2021, with the start date limited by the availability of the 13F data from Refinitiv.

3.1 Stock Characteristics and Returns

We use the CRSP daily returns data from the Center for Research in Security Prices (CRSP) from 1980 onward. We retain firms with at least one year of data and drop the bottom decile of market capitalization each month.

To estimate the factor model in returns, we use 120 monthly stock characteristics built from the Jensen-Kelly-Pedersen global factor data (Jensen et al., 2023). Appendix C lists all 120 variables. We only use characteristics that are not functions of stock prices.

All characteristics are lagged as in Kelly et al. (2019). We follow Kelly et al. (2019) and Kozak et al. (2020) by normalizing characteristics each month into percentiles in $[0, 1]$ and subtracting 0.5. This transformation ensures characteristics are on a comparable, centered scale with bounded influence.

3.2 Estimating Factor Structure in Realized Returns

We estimate a nonlinear factor model in daily returns using Kelly et al. (2019) instrumented principal components analysis (IPCA) on nonlinear transformations of the 120 monthly stock characteristics described in Section 3.1.

Following Kelly et al. (2024), we map characteristics into a high-dimensional Random Fourier Features basis. This kind of high-dimensional nonlinear specification is natural here because, when expected returns are dispersed across many weak directions, simple low-dimensional predictor sets can miss economically important return-predictive structure that richer machine-learning specifications can recover (Kelly et al., 2024). In particular, we use $M = 3{,}000$ features, which is slightly less than the number of stocks in the month with the fewest stocks. We orthonormalize the features using a standard right-whitening to form our final set of stock characteristics.

We then do IPCA with these characteristics, which has a convenient closed-form solution due to the orthonormalization. The first factor is the market factor that we constrain all stocks in each cross-section to have the same loading on. We use $K = 200$ cross-sectional factors after this, giving us 201 factors in total. We use a large number of factors K to allow for the possibility that weak factors explain much expected return variation. Using a five-factor model, for example, with just a few characteristics would force a strong factor model to hold. The factor betas, written as an $M \times (K+1)$ matrix B_t, are naturally orthonormal. Appendix D gives the details of the feature map, whitening step, and the construction of factor returns and variances.

Since the estimated factors are not necessarily orthogonal, to map more closely to the model in Section 2, we eigendecompose the systematic covariance matrix $B_t \Sigma_F B_t'$, where Σ_F is the covariance matrix of estimated factor returns and B_t is the matrix of factor loadings. In particular, we can write $B_t \Sigma_F B_t' = \Theta_t E_t \Theta_t'$, where Θ_t represents the month-specific orthonormal loadings and E_t is a diagonal matrix with factor variances $e_{k,t}$ in descending order.

Given this estimated factor model, we find the average monthly per-stock idiosyncratic

variance is about 0.0078, which we round to 0.01 for simplicity.

3.3 Plausibly Exogenous Demand Shocks

We use three plausibly uninformed demand shocks from previous work to measure own- and cross-price multipliers in Section 7. Here, we briefly explain the construction and logic for these shocks; we refer the reader to the original papers for further details.

Benchmarking Intensity (BMI). Our first demand shock is exogenous changes in benchmarking intensity around Russell index reconstitutions from Pavlova and Sikorskaya (2023). Russell index reconstitution occurs each June and is a mechanical function of market capitalization. Each May, Russell ranks eligible stocks by market capitalization. Stocks above a specified rank cutoff join the Russell 1000, while those below join the Russell 2000.[9] The Russell 2000 has historically attracted more benchmarked institutional capital than the Russell 1000. Hence, a stock that starts in the Russell 1000 but falls below the cutoff in May moves to Russell 2000 and experiences inflows of benchmarked capital in June. Conversely, stocks that move from the Russell 2000 to 1000 experience outflows. Conditional on the market cap as of the May ranking date, Russell index membership in June is exogenous to June cash-flow news, and so these reconstitution-driven flows are uninformed demand shocks (Chang et al. (2014); Crane et al. (2016); Glossner (2019)).

Moreover, stocks can move across multiple Russell indices simultaneously. Every stock in the Russell 2000 Blend index is also in the Russell 2000 Value or Growth indices, which have different levels of benchmarked capital. Every stock in the Russell 1000 Blend index is also in the Russell 1000 Value or Growth indices, and some — those with market cap rank below 200 — are in the Russell Midcap Blend, Value, and Growth indices. Thus, a stock moving from the Russell 1000 Value to the 2000 Value experiences different inflows

[9]More specifically, prior to 2007 the rank cutoff was the 1,000th stock. To reduce turnover, since 2007 Russell has used a "banding policy" under which there are two separate cutoffs for stocks starting in the Russell 1000 and 2000 pre-reconstitution, both of which are still mechanical functions of the firm size distribution. Thus, there is a "band" of market caps including stocks from the Russell 1000 and 2000. Appendix E.1 explains the Russell methodology we use to calculate these cutoffs.

of benchmarked capital than a stock moving from the 1000 Growth to the 2000 Growth.

The Pavlova and Sikorskaya (2023) benchmarking intensity (BMI) measure captures this heterogeneity in Russell index reconstitution-induced demand shocks:

$$BMI_{n,t} = \sum_{\text{Index } j} \frac{\begin{array}{c}\text{Institutional AUM Benchmarked to Index } j \text{ in Month } t \\ \times \text{ Weight of Stock } n \text{ in Index } j \text{ in Month } t\end{array}}{\text{Stock } n \text{ Market Value in Month } t}.$$

BMI depends on which indices stock n is part of and on the proportion of n's market value that is held by benchmarked investors. BMI is calculated from thirty-four indices, including the nine Russell indices, that cover 90% of mutual fund and ETF assets under management.[10]

We focus on exogenous changes in BMI due to the Russell index reconstitution each June. Specifically, we consider June BMI changes, denoted ΔBMI, for stocks in a narrow window around Russell 1000/2000 reconstitution thresholds (500 stocks on both sides of the threshold). The *change* ΔBMI for stocks in this window is driven by Russell index membership changes, which are exogenous to June cash flow news conditional on the May rank-date market cap.

We use the BMI and Russell index constituent data provided by Pavlova and Sikorskaya (2023). All of our specifications include the stock-level controls used by Pavlova and Sikorskaya (2023): May rank-date log market cap, one-year monthly average bid-ask percentage spread, and the banding controls from Appel et al. (2019).[11] Whereas Pavlova and Sikorskaya (2023) use the proprietary Russell market cap, we calculate market cap from standard databases using the method of Ben-David et al. (2019).[12]

[10]Since ΔBMI measures the Russell index reconstitution-induced change in demand for a stock by only benchmarked mutual funds and ETFs — not the total change in demand by all institutional investors — we scale it to have the same units as OFI and FIT. Pavlova and Sikorskaya (2023) find that a 1% reconstitution-induced BMI increase raises total institutional ownership by 0.2%, so we multiply the raw ΔBMI by 0.2. See Column 3 of Table 3 in Pavlova and Sikorskaya (2023) for justification for this scaling.

[11]The banding controls include an indicator for rank-date market cap in the "band" with stocks from the Russell 1000 and 2000, an indicator for being in the 2000 in May, and the interaction of these indicators.

[12]Since Russell ranks stocks using a proprietary market cap that we lack access to, we use the method of Ben-David et al. (2019) to approximate this market cap using standard databases. Doing so accurately predicts assignment to the Russell indices. We use May, not June, market caps to calculate the reconstitution thresholds to avoid selection bias (e.g., Chang et al., 2014; Appel et al., 2021; Wei and Young, 2021).

Flow-Induced Trading (FIT). Our second demand shock is flow-induced trading (FIT) from Lou (2012), which captures the predicted, mechanical stock-level trading of mutual funds and ETFs in response to flows.

Funds tend to scale their pre-existing holdings in a non-discretionary fashion in response to flows (Coval and Stafford, 2007; Lou, 2012). For example, if Apple's weight is 5% in a fund's portfolio, a $1 inflow (outflow) induces the fund to allocate about five more (fewer) cents to Apple. This predicted mechanical component of the cross-sectional trading induced by flows is uninformed.

To construct FIT, we first calculate the quarterly percentage flow to fund i as

$$f_{i,t} = \frac{TNA_{i,t} - TNA_{i,t-1} \cdot (1 + \text{Ret}_{i,t})}{\text{TNA}_{i,t-1}}.$$

where $TNA_{i,t}$ and $\text{Ret}_{i,t}$ are fund i's end-of-quarter-t total net assets and return from end of quarter $t-1$ to t, respectively.

The predicted non-discretionary trading by fund i in stock n due to this flow is proportional to $(\text{Shares Held}_{i,n,t-1}) \cdot f_{i,t}$. Following Chaudhry and Li (2025), we allow for heterogeneous flow-to-trade relationships, by estimating a scaling factor $b(w_{i,n,t-1}, f_{i,t})$ that is allowed to flexibly vary with realized fund flow $f_{i,t}$ and pre-existing portfolio weight $w_{i,n,t-1}$.[13]

Aggregating across all funds and dividing by shares outstanding yields:

$$\text{FIT}_{n,t} = \sum_{\text{fund } i} b(w_{i,n,t-1}, f_{i,t}) \cdot \underbrace{\frac{\text{Shares Held}_{i,n,t-1}}{\text{Shares Outstanding}_{n,t-1}}}_{\equiv S_{i,n,t-1}} f_{i,t}, \tag{12}$$

where $S_{i,n,t-1}$ is the proportion of stock n's shares owned by mutual fund i in quarter $t-1$.

Previous work argues that FIT is unrelated to cash flow news because a large fraction of flows are driven by unsophisticated retail investors who do not respond to cash flow

[13] Empirical estimates of this scaling factor are approximately one and vary little across funds. We refer the reader to Appendix A.2. in Chaudhry and Li (2025) for further details.

news (Lou, 2012; Li and Lin, 2022; Huang et al., 2024). Moreover, previous work also shows that the price impact of FIT reverts over time, which suggests these shocks are unrelated to cash flow news.

Order flow imbalance (OFI). Our third demand shock is signed order flow imbalance (OFI) from Li and Lin (2025). We classify all stock-level trades as buyer- or seller-initiated using the Lee and Ready (1991) algorithm. Specifically, we download daily Lee-Ready signed OFI data from WRDS intraday indicators, aggregate to quarterly frequencies, and then normalize by lagged shares outstanding.

OFI only captures all aggressively executed trades; slowly and passively executed trades by sophisticated institutional investors are not captured. While OFI has been used extensively in the microstructure literature to study price impact at daily or higher frequencies, recent work also uses OFI to measure price multipliers at monthly or lower frequencies (Li and Lin, 2025; Chaudhry and Li, 2025). Li and Lin (2025) conduct extensive tests and do not find evidence that OFI is related to various measures of cash-flow news. However, they acknowledge that one cannot be fully certain that OFI is unrelated to cash flow news.

3.4 Institutional Holdings Data

The own- and cross-price elasticity estimation in Section 8 uses institutional holdings data from the SEC Form 13F provided by Refinitiv through WRDS. The SEC requires all institutional investors with at least $100 million in assets under management (AUM) to report stock-level long positions each quarter. Following Koijen and Yogo (2019), we construct a residual "household" sector that comprises direct holdings by households and those by non-13F institutions (i.e., institutions with less than $100 million AUM).

4 Evidence from Returns: Many Factors are Required to Explain Expected Returns

In this section, we measure the degree of substitution ψ by measuring the proportion of cross-sectional expected return variation explained by different sets of return factors.

From (9) in Section 2, the proportion of expected return variation explained by the first J factors (after the market factor) is:

$$R^2_\psi(J) \equiv \sum_{k=1}^{J} f_k = \frac{\sum_{k=1}^{J} (\psi \tilde{e}_k + 1)^2}{\sum_{k=1}^{K} (\psi \tilde{e}_k + 1)^2}. \tag{13}$$

We can measure this proportion using cross-sectional predictive regressions of returns on factor loadings. Realized returns are expected returns plus unforecastable noise: $D_n - P_n = \mathbb{E}[D_n] - P_n + v_n$, where $v_n \sim N(0, \sigma_v^2)$ is i.i.d. Thus, the R^2 from a cross-sectional regression of returns on factor loadings for the first J factors

$$D_n - P_n = \sum_{k=1}^{J} \beta_k \theta_{n,k} + v_n \tag{14}$$

is

$$R^2_{reg}(J) = \frac{\sum_{k=1}^{J} (\psi \tilde{e}_k + 1)^2}{\sum_{k=1}^{K} (\psi \tilde{e}_k + 1)^2 + \sigma_v^2} \tag{15}$$

on average across many realizations of factor supply $\boldsymbol{\pi}$ (as proven in Appendix B.3).

Note that $R^2_{reg}(J)$ is biased downward from $R^2_\psi(J)$ due to the variance of unexpected returns σ_v^2. However, since this bias is also present in the R^2 from the predictive regression with all K factors, we can cancel it out and recover $R^2_\psi(J)$ using the ratio

$$R^2_{data}(J) \equiv \frac{R^2_{reg}(J)}{R^2_{reg}(K)} = R^2_\psi(J). \tag{16}$$

This normalization in the denominator of (16) provides a simple and powerful way to

measure explained variation in expected returns using realized returns, despite the presence of unforecastable noise. Comparing $R^2_{\psi}(J)$ with a high ψ value to $R^2_{data}(J)$ thus provides a clean way to assess the degree to which weak factors govern expected returns (i.e., the extent to which there is a factor zoo).

Empirically, we estimate the empirical $R^2_{reg}(J)$ as the average R^2 from monthly cross-sectional predictive regressions of returns on the first J factor loadings $\boldsymbol{\theta}^{(J)}_{n,t}$:

$$r_{n,t+1} = \boldsymbol{\beta}_t \boldsymbol{\theta}^{(J)}_{n,t} + e_{n,t}.$$

We show that $\forall J < K$, $R^2_{data}(J) \ll R^2_{100}(J)$. Thus, far more factors are required to explain cross-sectional expected return variation than would be required under the classical benchmark of $\psi = 1$. We find that $R^2_{data}(J) = R^2_{\psi}(J)$ for only small values of ψ. To address the concern that $R^2_{data}(J)$ is inflated by overfitting, we re-estimate the factor model out-of-sample in Section 4.2 and show that the results are similar.

4.1 Empirical Results

Figure 2 displays the results. The data-based regression curve $R^2_{data}(J)$ lies far below the benchmark $R^2_{100}(J)$ curve. Under the classical benchmark of $\psi = 1$, the first factor would explain about 54% of cross-sectional expected-return variation, the first 9 factors about 75%, the first 50 factors about 87%, and the first 100 factors about 94%. This concentration arises because the first few factors explain far more comovement (i.e., have far larger factor variances) than the remaining factors. In contrast, empirically the first factor accounts for only 6.7%, the first 10 factors for 16.1%, the first 50 factors for 35.2%, and even the first 100 factors for only 58.1% of expected return variation. Thus, many factors are needed to explain the cross section of expected returns.

The figure also compares the $R^2_{data}(J)$ curve to the model-implied curve obtained by choosing ψ to best match the shape of the empirical R^2 path. The best-fitting value is approximately $\psi = 0.011$, with a quarter-block bootstrap standard error of 0.0025. This estimate is two orders of magnitude smaller than the classical benchmark $\psi = 1$, and

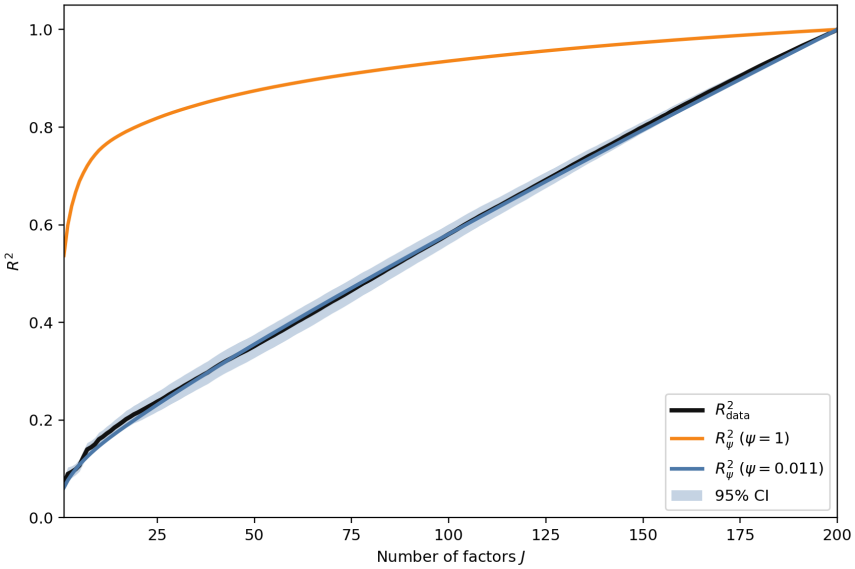


Figure 2: The black data curve is the normalized cross-sectional R^2 from regressing next month's average excess stock returns on the first J orthonormal loading directions, averaged across months. The shaded region is a 95% confidence band based on 100 quarter-block bootstrap replications. The orange benchmark curve is R^2_ψ evaluated at the classical benchmark of $\psi = 1$. The blue curve is the R^2_ψ curve for the best-fitting value of $\psi = 0.011$.

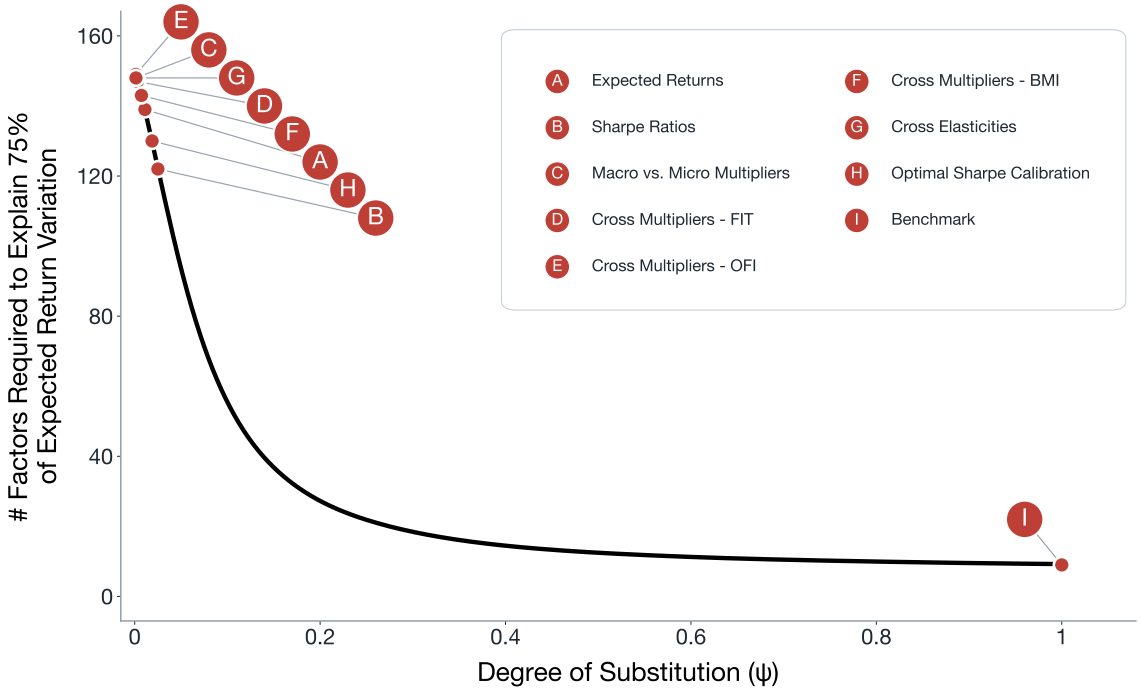
the implied curve closely tracks the $R^2_{data}(J)$ curve. The shaded region reports a 95% confidence band based on the same 100 bootstrap replications.

Figure 3 displays the relationship implied by the model in Section 2 between the number of factors required to explain cross-sectional expected return variation and the degree of substitution ψ. Under the classical benchmark of $\psi = 1$, only 9 factors would be required to explain 75% of expected return variation. In contrast, empirically we find that 139 factors are required to explain 75% of expected return variation, as displayed in Figure 2.

Thus, empirically, far more factors are required to explain cross-sectional expected return variation than suggested by classical models with strong substitution.

4.2 Robustness

To demonstrate the robustness of the results in Figure 2, we estimate the factor structure in realized returns in several alternate ways. Across these alternate specifications, we continue to find that far more factors are required to explain cross-sectional expected



	Empirical Strategy	ψ	Number of Factors
A	Expected Returns	0.011	139
B	Sharpe Ratios	0.025	122
C	Macro vs. Micro Multipliers	0.001	148
D	Cross Multipliers - FIT	0.003	147
E	Cross Multipliers - OFI	0.001	149
F	Cross Multipliers - BMI	0.007	143
G	Cross Elasticities	0.001	148
H	Optimal Sharpe Calibration	0.019	130
I	Benchmark	1.000	9

Figure 3: This figure in Panel A plots the model-implied number of factors required to explain 75% of expected return variation as a function of the degree of substitution ψ. The overlaid points correspond to ψ values and required numbers of factors estimated across the different empirical strategies, as detailed in the table in Panel B.

return variation than in classical models with strong substitution.

Out-of-Sample Estimation Because the factor model is estimated from returns, an in-sample regression may overfit. To address this concern, we also estimate the factor model on the pre-2000 sample and then run the return-prediction regressions out-of-sample from 2000–2023. Appendix F.1 Figure F.1 shows that the normalized out-of-sample data curve remains far below the strong-substitution benchmark at $\psi = 1$, and the best-fitting ψ remains small. In summary, the result that the factor structure of equilibrium expected returns is consistent with weak substitution is not an artifact of in-sample overfitting.

Daily RFF Appendix F.2 Figure F.2 repeats the expected-return spanning exercise using the daily returns. The same qualitative pattern appears: the daily data curve lies far below the $\psi = 1$ benchmark, and the best-fitting value of ψ remains small.

Standard PCA To address the concern that the IPCA construction with RFF characteristics imposes too much structure on the factor loadings, Appendix F Figure F.3 repeats the analysis using standard PCA (specifically, we use the deflated heteroskedastic PCA method of Zhou and Chen (2025)). The same qualitative pattern appears: the daily data curve lies far below the $\psi = 1$ benchmark, and the best-fitting value of ψ remains small.

5 Measuring Substitution Using Sharpe Ratios

In this section we measure the degree of substitution ψ using the distribution of factor-level Sharpe ratios. We then argue that strong substitution would generate unrealistically high Sharpe ratios.

As proven in Appendix G, the annualized squared Sharpe ratio of factor k is:

$$SR_k^2 = C g_k(\psi), \qquad g_k(\psi) = \frac{(1+\psi\tilde{e}_k)^2}{(1+\tilde{e}_k)\,\sigma^2}, \qquad C = 12\,\mathbb{E}[\pi_k^2]\left(\frac{1}{\zeta^{Own}}\right)^2. \tag{17}$$

The common scale C moves all Sharpe ratios together. ψ determines how quickly Sharpe ratios rise with factor variance.[14] The relative magnitudes of Sharpe ratios across factors identify ψ. Intuitively, stronger substitution means investors care more about minimizing portfolio volatility, and so require greater compensation for bearing non-diversifiable risk. As a result, higher ψ implies that factor Sharpe ratios rise faster as factor variances rise.

Thus, we estimate ψ using a nonlinear regression of relative squared factor Sharpe ratios on factor variances:

$$\widehat{\psi} = \underset{\psi}{\operatorname{argmin}} \sum_{k=0}^{K} \left[\frac{SR_k^2}{\sum_j^K SR_j^2} - \frac{g_k(\psi)}{\sum_j^K g_j(\psi)} \right]^2$$

Given ψ, we can then identify C using the maximum-attainable Sharpe ratio. From (11) in Proposition 2,

$$\widehat{C} = \frac{\text{Max Annualized Sharpe Ratio}^2}{\sum_{k=0}^{K} g_k(\widehat{\psi})} \tag{18}$$

Appendix G provides estimation details.

To calculate Sharpe ratios empirically, we use the same calendar split as in the out-of-sample expected-return exercise in Section 4.2: we estimate the factor model through 1999 and evaluate factor-level moments out-of-sample from 2000–2023. We also estimate the max-Sharpe (i.e., mean-variance-optimal) portfolio across the 201 factors using the pre-2000 sample and hold its weights fixed thereafter. This portfolio's out-of-sample annualized Sharpe ratio is 1.812. The standard error from 1,000 calendar-quarter block bootstrap samples of the post-2000 period is 0.197.

Through this estimation procedure, we find that substitution is weak. We find $\widehat{\psi} = 0.02484$ (standard error 0.00702) and $\widehat{C} = 0.001361$ (standard error 0.000429), as displayed in Appendix Table G.1. This ψ estimate is two orders of magnitude smaller than the classical benchmark $\psi = 1$. Figure 3 shows that this ψ estimate implies that 122 factors are required to explain 75% of expected-return variation. This is close to the

[14] The factor of 12 annualizes the squared Sharpe ratio; the factor (e_k) and idiosyncratic (σ^2) variances remain in monthly units.

139 factors we find are empirically required to explain 75% of expected return variation, and far more than the 9 factors required under the classical benchmark of $\psi = 1$.

We emphasize that this procedure provides an independent estimate of ψ separate from that in Section 4.1, since each exercise uses different empirical moments. Section 4 uses cumulative R^2s from cross-sectional return prediction regressions. In contrast, here we use factor-level Sharpe ratios, and so incorporate data on factor variances. This variance adjustment is the critical difference.

Moreover, the $\widehat{C}$ estimate implies that strong substitution would generate unrealistically high Sharpe ratios. We hold $\widehat{C}$ fixed and calculate the counterfactual model-implied maximum Sharpe ratio that would arise under different ψ values. This exercise corresponds to moving up the blue iso-cost curve in Figure 1. Following (18), the fixed-$\widehat{C}$ counterfactual maximum Sharpe ratio is

$$SR(\psi) = \sqrt{\widehat{C} \sum_{k=0}^{K} g_k(\psi)} \tag{19}$$

Moving toward higher ψ along an iso-cost curve shifts a fixed total idiosyncratic position cost from frictions to risk. As a result, investors care more about minimizing portfolio volatility, and so demand more compensation for bearing non-diversifiable risk. Thus, the maximum annualized Sharpe ratio increases, reaching an unrealistically high 26.63 (standard error 3.92) at the classical benchmark of $\psi = 1$, as shown in Figure 4.

This result gives quantitative content to Figure 1. Weak substitution creates a factor zoo not by raising expected returns on weak factors, but by reducing the equilibrium compensation for non-diversifiable risk and compressing the spread in expected returns across factors. Stock-level expected returns, therefore, spread across many factors while attainable Sharpe ratios remain moderate.

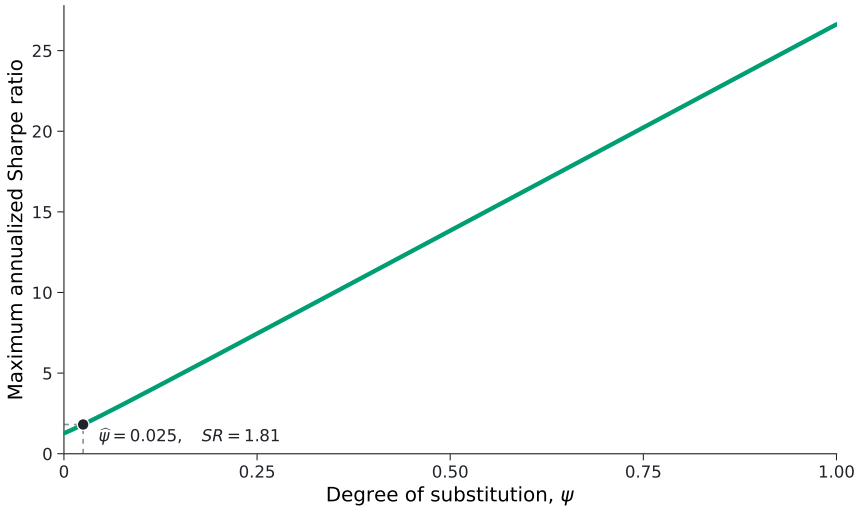


Figure 4: Maximum Sharpe ratios along the fixed-cost counterfactual. The point combines $\widehat{\psi} = 0.02484$, estimated from the relative pattern of factor-level Sharpe ratios, with the 1.812 annualized out-of-sample Sharpe ratio of the mean-variance optimal portfolio whose weights are estimated over 1980–1999 and held fixed over 2000–2023. The curve holds $\widehat{C}$ and the post-2000 factor eigenvalues fixed as ψ varies.

6 Measuring Substitution Using Micro- and Macro-Multipliers

In this section we measure the degree of substitution ψ using the relative magnitudes of stock-level and market-level price multipliers.

The economics are familiar from the distinction between demand for a brand and demand for a product category. Suppose the quantity of Cheerios available to consumers falls by one percent. Its price rises, but consumers substitute toward other cereals, limiting the price response. This substitution also creates cross-price effects: a shortage of Corn Flakes redirects some consumers toward Cheerios and therefore raises the price of Cheerios. If the quantity of every cereal falls simultaneously, the price of Cheerios reflects both its own shortage and the cross-price effects created by shortages of every other cereal. Thus, saying that consumers cannot absorb a category-wide shortage by substituting across brands is equivalent to saying that the cross-price effects accumulate under a category-wide shock. This is why demand for an individual brand is more elastic than demand for the category as a whole.

The same logic applies to stocks. A *micro multiplier* measures the price response to buying one percent of a single stock. A *macro multiplier* measures the market response to buying one percent of every stock. The macro response combines each stock's own-price response with the cross-price effects generated by purchases of all other stocks. The difference between the micro and macro multipliers therefore measures the strength of substitution across stocks.

A simple example shows why these spillovers matter. Suppose the own-price multiplier is m_{own} and, for illustration, every pair of distinct stocks has the same cross-price multiplier m_{cross}. With N stocks, an equal purchase of every stock produces the macro multiplier

$$M^{\text{macro}} = m_{\text{own}} + (N-1)m_{\text{cross}}. \tag{20}$$

Thus, even a small pairwise spillover can become large after it is summed across thousands of stocks. If $N = 2{,}000$, $m_{\text{own}} = 1$, and $m_{\text{cross}} = 0.1$, then the macro multiplier is about 201, meaning a purchase of one percent of the market portfolio generates a price response of about 201 percent. Conversely, a macro multiplier of 5 and a micro multiplier of 2 imply an average cross-price multiplier of only $(5-2)/(2{,}000-1) \approx 0.0015$. Moderate macro multipliers are therefore compatible with observed micro multipliers only when pairwise cross-price multipliers are very small.

The model in Section 2 allows these effects to vary across pairs of stocks according to their factor exposures. Nevertheless, the general mapping is nearly as simple as (20). Let $M^{\text{micro}} = \lambda + \gamma\sigma^2$ denote the micro multiplier. The model implies

$$\frac{M^{\text{macro}}}{M^{\text{micro}}} = 1 + \psi\kappa_M, \tag{21}$$

where κ_M measures the systematic-risk exposure of a marketwide demand shock, normalized by idiosyncratic variance. We estimate $\kappa_M = 1044.45$ from the factor loadings and normalized factor variances in Section 3.2; Appendix H.1 gives the exact construction

and shows how (21) follows from the model.

Equation (21) allows multiplier evidence to discipline the mechanism behind our explanation of the factor zoo. A larger ψ means stronger substitution and therefore larger cross-price multipliers. Because a marketwide purchase activates the cross-price effects associated with purchases of all stocks, these effects accumulate into a large macro multiplier. Conversely, the relatively modest macro multipliers observed in the data can be reconciled with sizable stock-level micro multipliers only if cross-price effects, and hence ψ, are small.

We discipline ψ using existing multiplier estimates. Stock-level studies generally find micro multipliers in the neighborhood of 1 to 3.[15] Appendix Table H.1 reports these estimates study by study. We use 2 as our representative micro multiplier. Gabaix and Koijen (2021, Table 1) estimate a macro multiplier between 2.2 and 6.5, with a point estimate of 5. Substituting these two point estimates and $\kappa_M = 1044.45$ into (21) gives

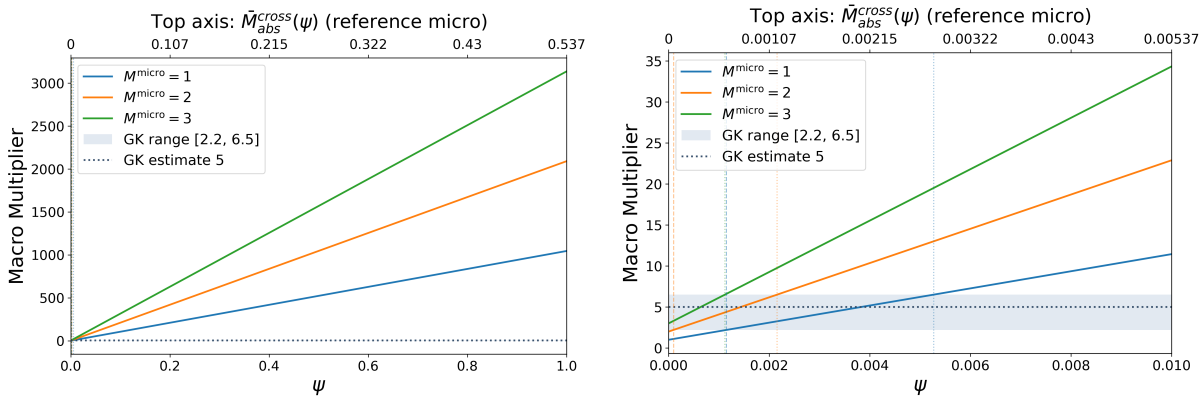
$$\widehat{\psi} = \frac{5/2 - 1}{1044.45} = 0.0014. \tag{22}$$

Figure 5 shows how the implied macro multiplier varies with ψ for micro multipliers of 1, 2, and 3.

The sensitivity analysis reinforces the point-estimate result. At the classical strong-substitution benchmark $\psi = 1$, the model implies macro multipliers above 1,000 for each micro multiplier in Figure 5. Across the full set of micro estimates and the empirical macro range, Appendix Table H.1 implies small values of ψ and average absolute cross-price multipliers of only 0.0001 to 0.0016.

Figure 3 displays that under the $\psi = 0.0014$ estimate in Figure 5, the model in Section 2 implies that 148 factors would be required to explain 75% of expected return variation. This is consistent with the 139 factors we find are empirically required to explain 75% of expected return variation, and far more than the 9 factors required under the classical

[15] We use "micro" rather than "own-price" deliberately. Under the identifying assumptions in Haddad et al. (2025), which are satisfied by our general factor structure for returns, standard stock-level regressions identify the micro, or relative, multiplier $M^{\text{micro}} = \lambda + \gamma\sigma^2$, rather than the full own-price multiplier $M^{\text{own},n} = M^{\text{micro}} + \gamma\sigma^2\widetilde{\tau}_{n,n}$. See also Fuchs et al. (2023).



(a) Full range: $\psi \in [0, 1]$

(b) Empirically relevant range: $\psi \in [0, 0.01]$

Figure 5: Model-implied macro multipliers. Each curve starts from a stock-level micro multiplier and uses the estimated factor structure to calculate the macro multiplier at each value of ψ. The shaded band is the macro-multiplier range of 2.2 to 6.5 estimated by Gabaix and Koijen (2021, Table 1). Panel (a) shows that the strong-substitution benchmark $\psi = 1$ implies extremely large macro multipliers. Panel (b) shows that the empirical macro range can be matched only when substitution is weak. The top axis reports the corresponding average absolute cross-price multiplier for the $M^{\text{micro}} = 2$ curve.

benchmark of $\psi = 1$. Thus, the empirical weak degree of substitution implies that far more factors are required to explain cross-sectional expected return variation than suggested by classical models.

7 Measuring Substitution Using Cross-Price Multipliers

In this section, we estimate the degree of substitution ψ from own- and cross-price multipliers. We find ψ is empirically two to four orders of magnitude smaller than the benchmark $\psi = 1$, which indicates investors weakly substitute across stocks.

7.1 Empirical Approach

As proven in Appendix B.1.1, we can write the equilibrium price for stock n as

$$P_n = \mathbb{E}[D_n] + \underbrace{\left(\lambda + \gamma\sigma^2\right)}_{\equiv M^{Own}} Q_n^{noise} + \sum_m \underbrace{\gamma\tau_{n,m}}_{\equiv M^{cross}_{n,m}} Q_m^{noise}, \tag{23}$$

where $\tau_{n,m} = \sum_{k=0}^{K} e_k \theta_{n,k} \theta_{m,k}$ is the systematic covariance between stocks n and m, and $Q_n^{noise} = -\boldsymbol{\theta}_n' \boldsymbol{\pi}$ is the total (negative) supply shock to stock n. Note that, under this sign convention, larger Q_n^{noise} reduces the number of shares held by the representative investor and so raises the price.

The "micro" (Gabaix and Koijen, 2021) or "relative" (Haddad et al., 2025) own-price multiplier is $M^{Own} = \lambda + \gamma\sigma^2$. It captures the direct price impact on stock n of a supply shock to stock n. For $m \neq n$, the cross-price multiplier $M_{n,m}^{Cross} = \gamma\tau_{n,m}$ captures the price impact on n of a noise shock to stock m. Intuitively, cross-price multipliers are smaller for less "similar" stocks — those with lower covariances. Thus, given residual supply shocks that are unrelated to expected cash flows, we can estimate the own- and cross-price multipliers by regressing the price of stock n on shocks to all stocks.

Since these multipliers are functions of λ and γ, they jointly identify the degree of substitution ψ:

$$\frac{M_{n,m}^{Cross}}{M^{Own}} = \psi \frac{\tau_{n,m}}{\sigma^2} \tag{24}$$

If cross-price multipliers are small relative to own-price multipliers, then γ is small relative to λ, ψ is small, and substitution is weak. Intuitively, if the investor weakly substitutes across stocks, then she does not reallocate from stock m to stock n when the price of stock m increases due to noise buying. As a result, the price impact of a shock to stock m stays localized and does not spill over to stock n, resulting in small cross-price multipliers.

Thus, we run the following regression motivated by (23):

$$\Delta p_{n,t} = \beta^{own} d_{n,t} + \beta^{Cross} \underbrace{d_{n,t}^{cross}}_{\equiv \sum_m \frac{\tau_{n,m,t}}{\sigma^2} d_{m,t}} + \boldsymbol{c}_b' \boldsymbol{x}_{n,t} + e_{n,t}. \tag{25}$$

$\Delta p_{n,t}$ is the percentage change in the price of stock n in quarter t. $d_{n,t}$ is one of the three plausibly exogenous demand shocks discussed in Section 3.3 (FIT, OFI, or BMI), expressed as a percentage of total shares outstanding. Mapping to the model in Section 2, these demand shocks are the *residual* supply shocks Q_n^{noise} that are uncorrelated with

expected cash flows: they reduce the number of shares held by the representative investor. $d_{n,t}^{cross} = \sum_m (\tau_{n,m,t}/\sigma^2) d_{m,t}$ is the covariance-weighted sum of demand shocks to all stocks. We measure the covariance $\tau_{n,m,t}$ between all pairs of stocks n and m using the factor model from Section 3.2.[16]

Per (24), we identify ψ as

$$\frac{\beta^{Cross}}{\beta^{Own}} = \frac{\gamma\sigma^2}{\lambda + \gamma\sigma^2} \equiv \psi.$$

7.2 Empirical Results

We find that cross-price multipliers are empirically small, which indicates that substitution is weak and ψ is small. Table 2 displays the results of regression (25).[17]

The own-price coefficients β^{Own} range from about 1 to 4 across the three shocks. These estimates imply that buying 1% of shares outstanding for a given stock raises its price by 1% to 4%, consistent with estimates from previous work.

The cross-price coefficients β^{Cross} are smaller and imply small ψ values. We find ψ values of 0.003 for FIT, 0.001 for OFI, and 0.007 for BMI. Thus, across all three shocks ψ is two to three orders of magnitude smaller than the benchmark $\psi = 1$. Therefore, substitution is empirically far weaker than implied by the classical benchmark.

Figure 3 displays that under the $\psi = 0.001$ to $\psi = 0.007$ estimates in Table 2, the model in Section 2 implies that 143 to 149 factors would be required to explain 75% of expected return variation. This is consistent with the 139 factors we find are empirically required to explain 75% of expected return variation, and far more than the 9 factors required under the classical benchmark of $\psi = 1$. Thus, the empirical weak degree of substitution implies that far more factors are required to explain cross-sectional expected

[16] Since FIT and OFI are quarterly shocks, we use the quarterly covariance matrix and idiosyncratic variance in those regressions. Since BMI is a monthly shock, we use the monthly counterparts.

[17] We account for estimation uncertainty in the covariances $\tau_{n,m,t}$ by drawing bootstrapped samples of quarters, estimating the factor model from Section 3.2 in each sample of quarters, computing the implied covariances $\tau_{n,m,t}$ in each sample, and running regression (25) in each sample. Thus, we obtain bootstrapped sampling distributions of β^{Own} and β^{Cross}. We report the standard deviations of those bootstrapped sampling distributions as the standard errors for these coefficients.

	FIT	OFI	BMI
β^{Own}	4.030***	0.896***	1.024***
	(0.382)	(0.069)	(0.139)
β^{Cross}	0.012**	0.001	0.008
	(0.006)	(0.001)	(0.011)
$\psi = \frac{\beta^{Cross}}{\beta^{Own}}$	0.003*	0.001	0.007
	(0.002)	(0.001)	(0.013)
Observations	470,178	330,138	17,091
R^2	0.027	0.022	0.008

Table 2: This table displays the results of regression (25) for the three demand shocks: the own-price coefficient β^{own}, the cross-price coefficient β^{cross}, and their ratio ψ. We report quarterly bootstrapped standard errors over 100 simulations that account for uncertainty in the estimated systematic covariances $\tau_{n,m,t}$. Levels of significance are presented as follows: *p<0.1; **p<0.05; ***p<0.01.

return variation than suggested by classical models.

7.3 Robustness and Associated Results

7.3.1 Addressing Measurement Error in Estimated Covariances

One may be concerned that classical measurement error in the estimated covariances $\tau_{n,m,t}$ could lead to attenuation bias in β^{Cross} that gives the misleading impression that ψ is small. We address this concern in two ways and find that the amount of measurement error is small and does not account for the small ψ estimates in Table 2.

Quantifying measurement error in estimated covariances In Appendix I.1, we conduct bootstrap simulations to quantify the amount of measurement error in the cross regressor $d_{n,t}^{cross}$ and find that it is small. We construct $d_{n,t}^{cross,(s)} = \sum_m (\tau_{n,m,t}^{(s)}/\sigma^2) d_{m,t}$ using $\tau_{n,m,t}^{(s)}$ estimated from different bootstrapped samples of quarters indexed by s. Since this bootstrapping procedure injects estimation error into $d_{n,t,(s)}^{cross}$ compared to the full-sample cross term $d_{n,t}^{cross}$, it allows us to quantify how much measurement error this estimation error can create. In particular, we calculate the classical measurement error attenuation factor $\text{Var}[d_{n,t}^{Cross,(s)}]/\text{Var}[d_{n,t}^{Cross}]$, which represents how much measurement

error this estimation error adds on top of the baseline variation in $d_{n,t}^{Cross}$.[18] A higher attenuation factor ratio indicates more measurement error. A ratio of one indicates no measurement error.

Appendix I.1 Table I.1 displays the results of this simulation exercise. The attenuation factors are small for all three demand shocks. For the FIT and OFI shocks, the attenuation factors are 1.01, implying that β^{Cross} (and so also ψ) is underestimated by one percent. Due to the much smaller sample size of the BMI shock (only stocks in a narrow window around Russell reconstitution thresholds in June of each year), the attenuation factor is slightly larger at 1.39, implying that β^{Cross} (and so also ψ) is underestimated by about 40%. But even accounting for this attenuation would yield a $\psi = 0.007 \times 1.39 = 0.01$, still far below the benchmark of $\psi = 1$.

Alternate estimation approach based on nonparametric sorts In Appendix I.2, we estimate ψ from own- and cross-price multipliers using an alternate approach based on nonparametric sorts. For each stock n, we sort all other stocks m into 100 bins based on their pairwise covariances, and estimate bin-specific cross-price multipliers. We then find the ψ value that best aligns the model-implied cross-price multipliers from (23) with the empirical cross-price multipliers across all bins. As we explain in Appendix I.2, because this approach does not use the estimated covariances $\tau_{n,m,t}$ directly in any regression, the associated ψ estimates cannot suffer from attenuation bias due to classical measurement error in the covariances $\tau_{n,m,t}$.

Appendix I.2 Table I.2 shows that this approach yields similar ψ estimates to those in Table 2: $\psi = 0.005, 0.002$, and 0.046 (with large standard errors on 0.046) for FIT, OFI, and BMI versus $\psi = 0.003, 0.001$, and 0.007, respectively, in Table 2.

7.3.2 Cross-Price Multiplier Estimates

Given the estimates of ψ from Table 2, we can compute the model-implied structural cross-price multipliers. From (24), we see that the ratio of the cross- to micro-price

[18] In particular, we calculate $\text{Var}[\tilde{d}_{n,t}^{Cross,(s)}]/\text{Var}[\tilde{d}_{n,t}^{Cross}]$, where the tilde variables have been residualized with respect to the own-stock demand shock $d_{n,t}$.

multipliers is given by $\psi\frac{\tau_{n,m}}{\sigma^2}$. Thus, for each stock n, we sort all other stocks m into $B = 100$ bins based on their absolute covariance $|\tau_{n,m}|$. Then we compute the average absolute scaled cross-price multiplier for all stocks in each bin. Appendix I.4 Figure I.3 displays the results. For the $\psi = 0.007$ estimate from the BMI demand shock, the scaled absolute cross-price multipliers range from 0.001 to 0.01 across bins, one to two orders of magnitude smaller than the 0.04 to 0.4 range for $\psi = 1$. For the $\psi = 0.003$ and $\psi = 0.001$ estimates from the FIT and OFI demand shocks, the scaled absolute cross-price multipliers range from 0.00005 to 0.002 across bins, two to four orders of magnitude smaller than the range implied by $\psi = 1$.

7.3.3 Implications for Biases in Estimated Multipliers

Previous work argues that failing to appropriately control for substitution can bias own-price multiplier estimates (Fuchs et al., 2023, 2025; Haddad et al., 2025). In Appendix I.3 we quantify these biases. We find that, given the weak degree of substitution indicated by the small ψ estimates in Table 2, these biases are quantitatively small. Even after removing the bias from failing to control for substitution, own-price multipliers remain large. Moreover, since these biases are small, full own-price multipliers and micro multipliers are nearly identical, and methods from previous work to control for substitution yield nearly unbiased estimates of these parameters.[19]

8 Measuring Substitution Using Cross-Price Elasticities

In this section we estimate the degree of substitution ψ from own- and cross-price elasticities in institutional investor holdings data from the SEC Form 13F. We find that ψ is three orders of magnitude smaller than the benchmark $\psi = 1$, which indicates investors weakly substitute across stocks.

[19] Our results are consistent with those in Chaudhary et al. (2023) Appendix J, which shows that grouping stocks on Fama-French three-factor characteristics or on 27 GICS industry and controlling for group-quarter fixed effects has little impact on the estimated own-price multipliers.

8.1 Empirical Approach

From (2) in Section 2, we can write the quantity demanded for stock n as

$$Q_n = \underbrace{\frac{1}{\lambda + \gamma\sigma^2}}_{\equiv \zeta^{Own}} \left(\mathbb{E}[D_n] - P_n\right) - \sum_{m=1}^{N} \underbrace{\frac{1}{\lambda + \gamma\sigma^2} \left(\sum_{k=0}^{K} \theta_{n,k}\theta_{m,k} \frac{\psi\tilde{e}_k}{1 + \psi\tilde{e}_k} \right)}_{\equiv \zeta_{n,m}^{Cross}} \left(\mathbb{E}[D_m] - P_m\right). \quad (26)$$

Given exogenous price variation that is unrelated to the investor's expected cash flows (i.e., price variation driven by the exogenous supply shocks), we can estimate the own- and cross-price elasticities via a regression of the quantity demanded for stock n on the instrumented prices for all stocks. Then from (6), we can infer the degree of substitution ψ from the relative magnitudes of own- and cross-price elasticities:

$$\frac{\zeta_{n,m}^{Cross}}{\zeta_{n}^{Own}} = \sum_{k=0}^{K} \theta_{n,k}\theta_{m,k} \frac{\psi\tilde{e}_k}{1 + \psi\tilde{e}_k}$$

Thus, we estimate the following demand curve

$$\log Q_{i,n,t} = \zeta^{own} \left(-\log P_{n,t} + W_{i,n,t}\left(\psi\right)\right) + \sum_{k=1}^{K} \beta_{k,t}\theta_{n,k,t} + \epsilon_{i,n,t} \quad (27)$$

$$W_{i,n,t}\left(\psi\right) = \sum_{m \in \mathcal{I}_{i,t}} \sum_{k=0}^{K} \underbrace{b_k\left(\psi\right)}_{\equiv \frac{\psi\tilde{e}_k}{1+\psi\tilde{e}_k}} \theta_{n,k,t}\theta_{m,k,t} \log P_{m,t}.$$

$\log Q_{i,n,t}$ is the log quantity of shares held by investor i in stock n in quarter t. $\log P_{n,t}$ is the log market equity. $\mathcal{I}_{n,t}$ is investor i's investment universe in quarter t, which we define as all stocks held in this quarter or in the past 11 quarters following Koijen and Yogo (2019).

Importantly, the log prices in (27) are not exogenous to the residual demand shocks $e_{i,n,t}$. Through market clearing, increases in this investor's demand for stock n raise the price of stock n. Thus, we instrument for the log prices using the investment mandate

instrument from Koijen and Yogo (2019). Specifically, the instrument is

$$Z_{i,n,t} = \log\left(\sum_{j\neq i} A_{j,t-1}\frac{1_{j,t}(n)}{1+\sum_{m=1}^{N} 1_{j,t}(m)}\right),$$

where $1_{j,t}(n)$ is an indicator for if stock n falls into the universe of investor j and $A_{j,t-1}$ is the lagged assets under management of investor j. This instrument represents the counterfactual market equity of stock n if all investors held an equal-weighted portfolio of the stocks in their investment universe. This instrument exploits cross-sectional variation in which investors' investment universes stock n falls into. Investment universe membership is plausibly exogenous to the residual demand shocks in this quarter because universes are defined by investment mandates: predetermined rules that do not respond to current demand shocks. Thus, if stock n exogenously falls into the investment universe of more or larger investors, it will face greater demand and will have greater market equity.

Using this instrument, we estimate ζ^{Own} and ψ using generalized method of moments (GMM) and the following two moment conditions:

$$\mathbb{E}_t\left[\left(\log Q_{i,n,t} - \zeta^{own}\left(-\log P_{n,t} + W_{i,n,t}(\psi)\right)\right) Z_{i,n,t} \mid \boldsymbol{\theta}_{n,t}\right] = 0 \tag{28}$$

$$\mathbb{E}_t\left[\left(\log Q_{i,n,t} - \zeta^{own}\left(-\log P_{n,t} + W_{i,n,t}(\psi)\right)\right) Z^{cross}_{i,n,t} \mid \boldsymbol{\theta}_{n,t}\right] = 0 \tag{29}$$

where $Z^{cross}_{i,n,t} \equiv \sum_{m\in\mathcal{I}_{i,t}}\sum_{k=0}^{K}\theta_{n,k,t}\theta_{m,k,t}Z_{i,m,t}$. That is, we use two instruments ($Z_{i,n,t}$ and $Z^{cross}_{i,n,t}$) and two moment conditions ((28) and (29)) to identify two parameters (ζ^{Own} and ψ). Appendix J.1 provides additional estimation details.

8.2 Empirical Results

We find that ψ is empirically small, which means cross-price elasticities are small and substitution is weak.

Table 3 displays the GMM estimation results.[20] We find $\zeta^{Own} = 0.54$, meaning an

[20] We account for estimation uncertainty in the covariances $\tau_{n,m,t}$ by drawing bootstrapped samples of quarters, estimating the factor model from Section 3.2 in each sample of quarters, and conducting the GMM estimation of moment conditions (28) and (29) in each sample. Thus, we obtain bootstrapped

exogenous 1% increase in price for stock n lowers the quantity demanded by 0.54%. More importantly, we find $\psi = 0.001$, three orders of magnitude smaller than the benchmark $\psi = 1$. Therefore, substitution is empirically far weaker than implied by the classical benchmark.

	ζ^{own}	ψ
Average Point Estimate	0.540^{***}	0.001350^{***}
	(0.004)	(0.000143)

Table 3: This table displays the results of estimating ζ^{Own} and ψ in demand curve (27) with GMM and moment conditions (28) and (29). We estimate these parameters in each quarter and report the time series average. We report quarterly bootstrapped standard errors over 100 simulations that account for uncertainty in the estimated factor loadings $\boldsymbol{\theta}_{n,t}$ and scaled factor variances $\tilde{e}_k$. Levels of significance are presented as follows: *p<0.1; **p<0.05; ***p<0.01.

Figure 3 displays that under the $\psi = 0.001$ estimate in Table 3, the model in Section 2 implies that 148 factors would be required to explain 75% of expected return variation. This is consistent with the 139 factors we find are empirically required to explain 75% of expected return variation, and far more than the 9 factors required under the classical benchmark of $\psi = 1$. Thus, the empirical weak degree of substitution implies that far more factors are required to explain cross-sectional expected return variation than suggested by classical models.

8.3 Robustness and Associated Results

8.3.1 Addressing Measurement Error in Estimated Covariances

One may be concerned that classical measurement error in the estimated factor loadings $\boldsymbol{\theta}_{n,t}$ and factor variances e_k could lead to attenuation bias in ψ that gives the misleading impression that substitution is weak. We address this concern in two ways and find that the amount of measurement error is small and does not account for the small ψ estimates in Table 3.

sampling distributions of ζ^{Own} and ψ. We report the standard deviations of those bootstrapped sampling distributions as the standard errors for these coefficients.

Quantifying attenuation bias due to measurement error in estimated factor loadings In Appendix J.2, we conduct bootstrap simulations to quantify how much attenuation bias measurement error in $\boldsymbol{\theta}_{n,t}$ and factor variances e_k can create in ψ. We find this attenuation is small. We conduct the GMM estimation of ψ using factor loadings and variances estimated from different bootstrapped samples of quarters indexed by s. Since this bootstrapping procedure injects estimation error in the factor loadings and variances compared to their full-sample counterparts, it allows us to quantify how much attenuation bias this estimation error can create. In particular, we measure $\mathbb{E}[\hat{\psi}_s] - \hat{\psi}$, where $\hat{\psi}$ is the full-sample estimate of ψ and $\mathbb{E}[\hat{\psi}_s]$ is the average estimate of ψ across all bootstrapped samples. A more negative difference indicates greater attenuation. As displayed in Appendix J.2 Table J.1, estimation error into the factor loadings and variances creates almost no attenuation.

Alternate estimation approach based on nonparametric sorts In Appendix J.3, we estimate ψ from own- and cross-price elasticities using an alternate approach based on nonparametric sorts. For each stock n, we sort all other stocks m into 100 bins based on their pairwise factor loading similarity, and estimate bin-specific cross-price elasticities. We then find the ψ value that best aligns the model-implied cross-price elasticities from (26) with the empirical cross-price elasticities across all bins. As we explain in Appendix J.3, because this approach does not use the estimated factor loadings directly in any regression, the associated ψ estimates cannot suffer from attenuation bias due to measurement error in the loadings.

Appendix J.3 Table J.2 shows that this approach yields similar ψ estimates to those in Table 3: $\psi = 0.0017$ versus $\psi = 0.0014$ in Table 3.

8.3.2 Cross-Price Elasticity Estimates

Given the estimates of ψ from Table 3, we can compute the model-implied structural cross-price elasticities. From (6), we see that the ratio of the cross- to own-price elasticities is given by $\sum_{k=0}^{K} \theta_{n,k}\theta_{m,k}\frac{\psi\tilde{e}_k}{1+\psi\tilde{e}_k}$. Thus, for each stock n, we sort all other stocks m into

$B = 100$ bins based on the absolute value of this summation. Then we compute the average absolute scaled cross-price elasticity for all stocks in each bin. Appendix J.5 Figure J.3 displays the results. For the $\psi = 0.0014$ estimate, the scaled absolute cross-price elasticities are two orders of magnitude smaller and much less dispersed (ranging from 0.0002 to 0.0003 across bins) than under the benchmark $\psi = 1$ (where they range from 0.00008 to 0.02).

8.3.3 Implications for Biases in Estimated Elasticities

Previous work argues that failing to appropriately control for substitution can bias own-price elasticity estimates (Fuchs et al., 2023, 2025; Haddad et al., 2025). In Appendix J.4 we quantify these biases. We find that, given the weak degree of substitution indicated by the small ψ estimates in Table 3, these biases are quantitatively small. Even after removing the bias from failing to control for substitution, own-price elasticities remain small. Moreover, since these biases are small, full own-price elasticities and micro elasticities are virtually identical, and methods from previous work to control for substitution yield nearly unbiased estimates of these parameters.[21]

9 A Potential Mechanism: Optimal Shrinkage Due to Estimation Error

We do not seek to distinguish among the potential sources of weak substitution. We nevertheless briefly examine one force—estimation error—to ask whether it alone can rationalize the small ψ values we estimate. An investor who seeks to maximize her out-of-sample Sharpe ratio but who understands that she faces estimation error in expected returns and covariances will optimally choose to shrink extreme position sizes (Jorion, 1986; DeMiguel et al., 2009a,b). If we express the model in Section 2 in terms of portfolio

[21] Our results are consistent with those in Tamoni et al. (2024) Appendix B.2.2, which shows that controlling for up to 49 Fama-French industries interacted with time fixed effects has little impact on the estimated elasticities.

weights, we can write the portfolio choice expression (2) as

$$\widetilde{\boldsymbol{w}}_t(\psi) \propto \left(\mathbf{I} + \psi \frac{\widehat{\boldsymbol{\Sigma}}_t^{com}}{\sigma^2} \right)^{-1} \widehat{\boldsymbol{\mu}}_t. \tag{30}$$

This is the expression for the vector of portfolio weights demanded, up to an overall scale factor that does not affect Sharpe ratios.

Here $\widehat{\boldsymbol{\mu}}_t$ is the expected-return forecast, $\widehat{\boldsymbol{\Sigma}}_t^{com}$ is the estimated common, or systematic, component of the covariance matrix excluding idiosyncratic variance, and σ^2 is idiosyncratic variance. In our baseline analysis, we construct $\widehat{\Sigma}_t^{com}$ from the estimated factor loadings and variances from the factor model in Section 4. Appendix K details the exact construction and provides an alternate covariance estimate.

Lower ψ places less weight on this common covariance structure and shrinks the portfolio toward a rule that treats stocks independently. This raises a simple question: how much shrinkage would an investor who seeks to maximize out-of-sample Sharpe ratio choose, given the empirical difficulty of estimating expected returns and covariances?

We estimate expected returns and the factor-model covariance structure before 2000, form portfolios strictly ex ante from 2000 through 2023, and choose ψ to maximize the realized hold-out Sharpe ratio. Figure 6 reports the benchmark exercise using the pre-2000 Fama–MacBeth expected-return forecast.

The Sharpe-maximizing value is $\psi = 0.0188$, compared with the classical benchmark $\psi = 1$. Estimation error therefore gives investors a conventional reason to choose portfolio behavior associated with weak substitution. Appendix K provides the full construction and considers nine alternative combinations of mean and covariance estimates. Eight select a ψ below 0.0188, and several select $\psi = 0$, making this baseline $\psi = 0.0188$ estimate conservative about the amount of shrinkage favored by the data.

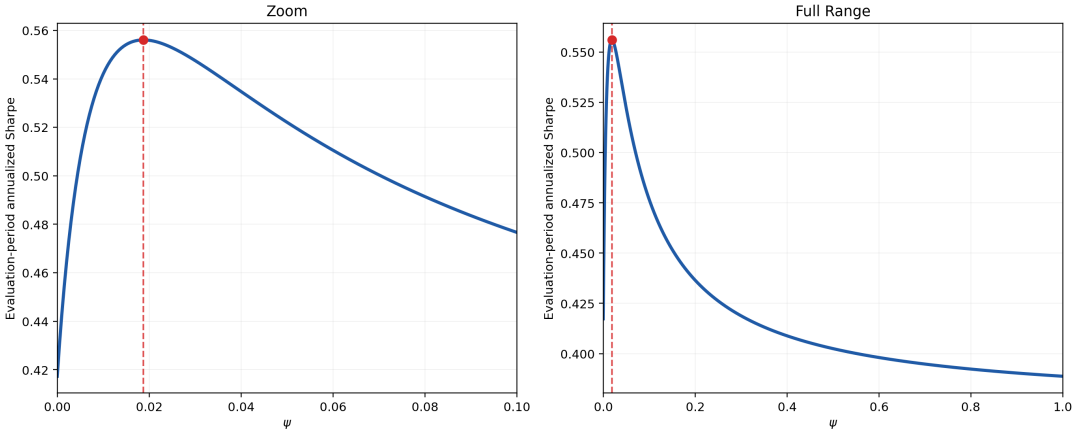


Figure 6: Out-of-sample Sharpe ratio by ψ in the baseline specification. The Sharpe-maximizing value is $\widehat{\psi}^{SR} = 0.0188$, well below the classical benchmark $\psi = 1$.

10 Conclusion

We argue that weak substitution across individual stocks provides a simple, unified explanation for the existence of the factor zoo in expected stock returns. In our theoretical framework, an otherwise standard mean–variance investor faces an ℓ_2 penalty on portfolio positions. Lower risk aversion relative to this ℓ_2 penalty weakens substitution without generating near-arbitrage opportunities. When substitution is strong, the factors that contribute most to return comovement explain most of the cross-sectional variation in expected returns. As substitution weakens, expected returns disperse across many factors, including many weak factors that contribute little to comovement. The model thus maps the degree of substitution to the strength of the factor structure in expected returns and shows how a genuine factor zoo arises in equilibrium.

Our empirical work shows the U.S. equity market lies firmly in the weak-substitution region described by the model. First, predictive regressions using factor loadings estimated with both IPCA and standard PCA on individual stock returns confirm that many factors are needed to explain the cross section of expected returns, far more than theory would predict under strong substitution. Second, the distribution of factor Sharpe ratios implies that substitution must be weak, since stronger substitution would generate implausibly large maximum attainable Sharpe ratios. Third, a calibration linking mi-

cro and macro price multipliers shows that cross-price spillovers must be tiny — and so substitution must be weak — in order for aggregate market multipliers to have reasonable magnitudes. Fourth, direct estimates of cross-price multipliers using three plausibly exogenous demand shocks are far smaller than own-price multipliers, implying ψ values far below the classical benchmark. Finally, cross-price elasticities in institutional holdings data are far smaller than own-price elasticities, which reinforces the conclusion that investors only weakly substitute across stocks.

While we do not take a strong stance on the mechanism behind the weak substitution we document, we provide one candidate. An investor who seeks to maximize her out-of-sample Sharpe ratio but who understands that she faces estimation error in expected returns and covariances will optimally choose to shrink extreme position sizes. We find the optimal level of shrinkage leads to a weak degree of substitution that is quantitatively consistent with our empirical results.

These findings have implications for how asset-pricing models are specified and how empirical evidence on anomalies is interpreted. Portfolio choice models that impose limited substitution or coarse substitution patterns, which might seem overly restrictive relative to the classical mean-variance benchmark, appear well aligned with the data once weak substitution is taken seriously. In a market where investors only partially substitute when prices change, it is natural that expected returns do not collapse onto a small set of dominant risk factors and that a large number of characteristics and portfolio strategies appear to earn distinct premia.

References

Amihud, Yakov, and Haim Mendelson, 1980, Dealership market: Market-making with inventory, *Journal of Financial Economics* 8, 31–53.

An, Yu, and Amy Huber, 2024, Demand propagation through traded risk factors, *Jacobs Levy Equity Management Center for Quantitative Financial Research Paper, The Wharton School Research Paper* .

Appel, Ian, Todd A Gormley, and Donald B Keim, 2021, Identification using russell 1000/2000 index assignments: A discussion of methodologies, *Critical Finance Review, Forthcoming* .

Appel, Ian R, Todd A Gormley, and Donald B Keim, 2019, Standing on the shoulders of giants: The effect of passive investors on activism, *The Review of Financial Studies* 32, 2720–2774.

Baba Yara, Fahiz, Brian H. Boyer, and Carter Davis, 2021, Evaluating the anomaly zookeepers, *working paper* .

Barberis, Nicholas, and Ming Huang, 2009, Preferences with frames: A new utility specification that allows for the framing of risks, *Journal of Economic Dynamics and Control* 33, 1555–1576.

Barberis, Nicholas, Ming Huang, and Richard Thaler, 2006, Individual preferences, monetary gambles, and stock market participation: A case for narrow framing, *American Economic Review* 96, 1069–1090.

Barillas, Francisco, and Jay Shanken, 2018, Comparing asset pricing models, *Journal of Finance* 73, 715–754.

Barry, Christopher B., 1974, Portfolio analysis under uncertain means, variances, and covariances, *The Journal of Finance* 29, 515–522.

Ben-David, Itzhak, Francesco Franzoni, and Rabih Moussawi, 2019, An improved method to predict assignment of stocks into russell indexes, Technical report, National Bureau of Economic Research.

Benartzi, Shlomo, and Richard H. Thaler, 2001, Naive diversification strategies in defined contribution saving plans, *American Economic Review* 91, 79–98.

Bessembinder, Hendrik, Aaron Burt, and Christopher M Hrdlicka, 2025, More factors matter and factors matter more than you might think: The role of time variation in factor premia, *Available at SSRN 3992041* .

Black, Fischer, and Robert Litterman, 1990, Asset allocation: combining investor views with market equilibrium, *Goldman Sachs Fixed Income Research* 115, 7–18.

Campbell, John Y., Yeung Lewis Chan, and Luis M. Viceira, 2003, A multivariate model of strategic asset allocation, *Journal of Financial Economics* 67, 41–80.

Campbell, John Y., and Robert J. Shiller, 1988, The dividend-price ratio and expectations of future dividends and discount factors, *The Review of Financial Studies* 1, 195–228.

Chang, Yen-Cheng, Harrison Hong, and Inessa Liskovich, 2014, Regression discontinuity and the price effects of stock market indexing, *The Review of Financial Studies* 28, 212–246.

Chaudhary, Manav, Julie Zhiyu Fu, and Jian Li, 2026, Which intermediary costs matter for asset prices?, Working Paper 6517218, Social Science Research Network, SSRN Working Paper No. 6517218.

Chaudhary, Manav, Zhiyu Fu, and Jian Li, 2023, Corporate bond multipliers: Substitutes matter, *Available at SSRN* .

Chaudhry, Aditya, and Jiacui Li, 2025, Endogenous elasticities: Price multipliers are smaller for larger demand shocks .

Chen, Hui, Winston Wei Dou, and Leonid Kogan, 2019, Measuring "dark matter" in asset pricing models, Technical report, National Bureau of Economic Research.

Chu, Yongqiang, David Hirshleifer, and Liang Ma, 2020, The causal effect of limits to arbitrage on asset pricing anomalies, *Journal of Finance* 75, 2631–2672.

Coval, Joshua, and Erik Stafford, 2007, Asset fire sales (and purchases) in equity markets, *Journal of Financial Economics* 86, 479–512.

Crane, Alan D, Sébastien Michenaud, and James P Weston, 2016, The effect of institutional ownership on payout policy: Evidence from index thresholds, *The Review of Financial Studies* 29, 1377–1408.

Da, Rui, Stefan Nagel, and Dacheng Xiu, 2024, The statistical limit of arbitrage, Technical report, National Bureau of Economic Research.

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

Dello Preite, Massimo, Raman Uppal, Paolo Zaffaroni, and Irina Zviadadze, 2024, Cross-sectional asset pricing with unsystematic risk, *Available at SSRN 4135146* .

DeMiguel, Victor, Lorenzo Garlappi, Francisco J. Nogales, and Raman Uppal, 2009a, A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms, *Management Science* 55, 798–812.

DeMiguel, Victor, Lorenzo Garlappi, and Raman Uppal, 2009b, Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy?, *Review of Financial Studies* 22, 1915 – 1953.

Drechsler, Itamar, and Qingyi Drechsler, 2016, Shorting, arbitrage, and return predictability, *Review of Financial Studies* 29, 637–672.

Ellis, Andrew, and Michele Piccione, 2017, Correlation misperception in choice, *American Economic Review* 107, 1048–1076.

Enke, Benjamin, and Florian Zimmermann, 2019, Correlation neglect in belief formation, *The Review of Economic Studies* 86, 313–332.

Fama, Eugene F., and Kenneth R. French, 1993, Common risk factors in the returns on stocks and bonds, *Journal of Financial Economics* 33, 3–56.

Fama, Eugene F., and Kenneth R. French, 2015, A five-factor asset pricing model, *Journal of Financial Economics* 116, 1 – 22.

Frazzini, Andrea, Ronen Israel, and Tobias J. Moskowitz, 2018, Trading costs of asset pricing anomalies, *Journal of Financial Economics* 127, 349–373.

Frost, Peter A., and James E. Savarino, 1986, An empirical bayes approach to efficient portfolio selection, *Journal of Financial and Quantitative Analysis* 21, 293–305.

Fuchs, William, Satoshi Fukuda, and Daniel Neuhann, 2023, Demand-system asset pricing: Theoretical foundations, *Available at SSRN 4672473* .

Fuchs, William, Satoshi Fukuda, and Daniel Neuhann, 2025, A trilemma for asset demand estimation, *Available at SSRN* .

Gabaix, Xavier, and Ralph S. J. Koijen, 2021, In search of the origins of financial fluctuations: The inelastic markets hypothesis, Working Paper 28967, National Bureau of Economic Research.

Garlappi, Lorenzo, Raman Uppal, and Tan Wang, 2007, Portfolio selection with parameter and model uncertainty: A multi-prior approach, *The Review of Financial Studies* 20, 41–81.

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

Gârleanu, Nicolae, and Lasse Heje Pedersen, 2016, Dynamic portfolio choice with frictions, *Journal of Economic Theory* 165, 487–516.

Glossner, Simon, 2019, Russell index reconstitutions, institutional investors, and corporate social responsibility, *Available at SSRN 3180776* .

Gonçalves, Andrei, Johnathan Loudis, and Richard Ogden, 2025, Out-of-sample alphas post-publication, *Fisher College of Business Working Paper* 002.

Greenwood, Robin, and Marco C. Sammon, 2023, The disappearing index effect, *Working paper* .

Gu, Shihao, Bryan Kelly, and Dacheng Xiu, 2021, Autoencoder asset pricing models, *Journal of Econometrics* 222, 429–450, Annals Issue:Financial Econometrics in the Age of the Digital Economy.

Guo, Hongye, and Jessica A Wachter, 2025, Correlation neglect in asset prices, *Available at SSRN* .

Haddad, Valentin, Zhiguo He, Paul Huebner, Péter Kondor, and Erik Loualiche, 2025, Causal inference for asset pricing, *Available at SSRN 5187305* .

Haddad, Valentin, Paul Huebner, and Erik Loualiche, 2022, How competitive is the stock market? theory, evidence from portfolios, and implications for the rise of passive investing.

Harris, Lawrence, and Eitan Gurel, 1986, Price and volume effects associated with changes in the s&p 500 list: New evidence for the existence of price pressures, *the Journal of Finance* 41, 815–829.

Hartzmark, Samuel M, and David H Solomon, 2022, Predictable price pressure, Technical report, National Bureau of Economic Research.

Ho, Thomas, and Hans R. Stoll, 1981, Optimal dealer pricing under transactions and return uncertainty, *Journal of Financial Economics* 9, 47–73.

Hou, Kewei, Chen Xue, and Lu Zhang, 2014, Digesting Anomalies: An Investment Approach, *The Review of Financial Studies* 28, 650–705.

Huang, Shiyang, Yang Song, and Hong Xiang, 2024, Noise trading and asset pricing factors, *Management Science* .

Huebner, Paul, 2023, The making of momentum: A demand-system perspective, in *Proceedings of the EUROFIDAI-ESSEC Paris December Finance Meeting*.

Jagannathan, Ravi, and Tongshu Ma, 2003, Risk reduction in large portfolios: Why imposing the wrong constraints helps, *The Journal of Finance* 58, 1651–1683.

Jansen, Kristy AE, Wenhao Li, and Lukas Schmid, 2024, Granular treasury demand with arbitrageurs, Technical report, National Bureau of Economic Research.

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

Jones, Charles M., and Owen A. Lamont, 2002, Short-sale constraints and stock returns, *Journal of Financial Economics* 66, 207–239.

Jorion, Philippe, 1986, Bayes-stein estimation for portfolio analysis, *Journal of Financial and Quantitative Analysis* 21, 279–292.

Kelly, Bryan, Semyon Malamud, and Kangying Zhou, 2024, The virtue of complexity in return prediction, *The Journal of Finance* 79, 459–503.

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

Klein, Roger W., and Vijay S. Bawa, 1976, The effect of estimation risk on optimal portfolio choice, *Journal of Financial Economics* 3, 215–231.

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

Kozak, Serhiy, Stefan Nagel, and Shrihari Santosh, 2018a, Interpreting factor models, *The Journal of Finance* 73, 1183–1223.

Kozak, Serhiy, Stefan Nagel, and Shrihari Santosh, 2018b, Interpreting factor models, *The Journal of Finance* 73, 1183–1223.

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

Ledoit, Olivier, and Michael Wolf, 2004, A well-conditioned estimator for large-dimensional covariance matrices, *Journal of Multivariate Analysis* 88, 365–411.

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

Lettau, Martin, and Markus Pelger, 2020, Factors that fit the time series and cross-section of stock returns, *The Review of Financial Studies* 33, 2274–2325.

Li, Jiacui, 2022, What Drives the Size and Value Factors?, *The Review of Asset Pricing Studies* 12, 845–885.

Li, Jiacui, and Zihan Lin, 2022, Price multipliers are larger at more aggregate levels, *Available at SSRN 4038664* .

Li, Jiacui, and Zihan Lin, 2025, Price multipliers are larger for less diversifiable order flows, *Working paper* .

Lopez-Lira, Alejandro, and Nikolai L Roussanov, 2023, Do common factors really explain the cross-section of stock returns, *Available at SSRN* .

Lou, Dong, 2012, A flow-based explanation for return predictability, *The Review of Financial Studies* 25, 3457–3489.

Maenhout, Pascal J., 2004, Robust portfolio rules and asset pricing, *The Review of Financial Studies* 17, 951–983.

Martin, Ian, and Stefan Nagel, 2019, Market efficiency in the age of big data, *NBER Working Paper* No. w26586.

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

Merton, Robert C, et al., 1987, A simple model of capital market equilibrium with incomplete information .

Nagel, Stefan, 2005, Short sales, institutional investors and the cross-section of stock returns, *Journal of Financial Economics* 78, 277–309.

Novy-Marx, Robert, and Mihail Velikov, 2016, A taxonomy of anomalies and their trading costs, *Review of Financial Studies* 29, 104–147.

Patton, Andrew J., and Brian M. Weller, 2020, What you see is not what you get: The costs of trading market anomalies, *Journal of Financial Economics* 134, 416–442.

Pavlova, Anna, and Taisiya Sikorskaya, 2023, Benchmarking intensity, *The Review of Financial Studies* 36, 859–903.

Petajisto, Antti, 2009, Why do demand curves for stocks slope down?, *Journal of Financial and Quantitative Analysis* 44, 1013–1044.

Pástor, Ľuboš, 2000, Portfolio selection and asset pricing models, *The Journal of Finance* 55, 179.

Roll, Richard, 1992, A mean/variance analysis of tracking error, *The Journal of Portfolio Management* 18, 13–22.

Schmickler, Simon, 2020, Identifying the price impact of fire sales using high-frequency surprise mutual fund flows, Technical report, Working Paper.

Shao, Shidong, Chen Yao, Mao Ye, and Jason J. Zou, 2026, Asset pricing with supply shocks, Technical report, Working Paper.

Shleifer, Andrei, 1986, Do demand curves for stocks slope down?, *The Journal of Finance* 41, 579–590.

Tamoni, Andrea, Stanislav Sokolinski, and Yizhang Li, 2024, Which investors drive anomaly returns and how?, *Available at SSRN 4242745* .

van der Beck, Philippe, 2022, Short-versus long-run demand elasticities in asset pricing, *Asset Pricing (May 16, 2022). Swiss Finance Institute Research Paper* .

Wei, Wei, and Alex Young, 2021, Selection bias or treatment effect? a re-examination of russell 1000/2000 index reconstitution, *Critical Finance Review* 1000.

Wurgler, Jeffrey, and Ekaterina Zhuravskaya, 2002, Does arbitrage flatten demand curves for stocks?, *The Journal of Business* 75, 583–608.

Zhou, Yuchen, and Yuxin Chen, 2025, Deflated heteropca: Overcoming the curse of ill-conditioning in heteroskedastic pca, *The Annals of Statistics* 53, 91–116.

Appendix

A CRRA Equilibrium in Log-Share Units

The main text uses a static mean-variance problem in share quantities because it makes the economic mechanism transparent. This appendix embeds that problem in a standard CRRA portfolio-choice model and carries the model through market clearing. We write demand in log shares and prices in logs, matching the units used in the empirical analysis. The CRRA economy delivers the same substitution parameter, elasticity and multiplier restrictions, equilibrium expected-return equation, expected-return spanning result, and Sharpe-ratio result as the main model. We first derive the investor's objective, then map portfolio weights into log shares, and finally solve for equilibrium.

A.1 From CRRA Utility to Portfolio Demand

Returns. For stocks $n = 1, \ldots, N$, let x_n be stock n's next-period log payoff and p_n its current log price. Its log excess return and simple excess return are

$$\ell_n^e = x_n - p_n, \qquad r_n^e = \exp(\ell_n^e) - 1. \tag{31}$$

Let $\boldsymbol{\mu} \equiv \mathbb{E}[\boldsymbol{r}^e]$ and

$$\boldsymbol{\Sigma} = \mathbb{V}[\boldsymbol{\ell}^e] = \boldsymbol{\Theta}\boldsymbol{E}\boldsymbol{\Theta}' + \sigma^2\boldsymbol{I}. \tag{32}$$

Here the columns $\boldsymbol{\Theta}_{\cdot,k}$, for $k = 0, \ldots, K$, are orthonormal factor-loading vectors. The diagonal matrix $\boldsymbol{E}$ has element e_k for factor k, and σ^2 is idiosyncratic variance. As in the main text, $\boldsymbol{\Theta}_{\cdot,0} = \boldsymbol{\iota}/\sqrt{N}$ is the market direction, where $\boldsymbol{\iota}$ is the N-vector of ones; every column $k > 0$ is orthogonal to $\boldsymbol{\iota}$.

A second-order expansion of the exponential in (31) gives

$$\boldsymbol{\mu} \approx \mathbb{E}[\boldsymbol{\ell}^e] + \frac{1}{2}\text{diag}(\boldsymbol{\Sigma}) = \boldsymbol{\nu} - \boldsymbol{p}, \tag{33}$$

where $\boldsymbol{\nu}$ collects expected log payoffs and the Jensen adjustment from log to simple returns.[22] Equation (33) implies $-\partial\boldsymbol{\mu}/\partial\boldsymbol{p}' = \boldsymbol{I}$ in the static model.

CRRA objective. Normalize the risk-free gross return to one. Let $\boldsymbol{W}$ be the vector of risky-asset portfolio weights; the risk-free asset finances the remaining position. The portfolio's simple excess return is $r_p^e = \boldsymbol{W}'\boldsymbol{r}^e$. The log-share approximation below is taken around positive reference weights, as in standard log-share demand systems. A second-order expansion of $\ell_p = \log(1 + r_p^e)$ gives

$$\mathbb{E}[\ell_p] \approx \boldsymbol{W}'\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{W}'\boldsymbol{\Sigma}\boldsymbol{W}, \tag{34}$$

$$\mathbb{V}[\ell_p] \approx \boldsymbol{W}'\boldsymbol{\Sigma}\boldsymbol{W}. \tag{35}$$

Let terminal wealth be $\mathcal{W}_1 = \exp(\ell_p)$ and utility be $u(\mathcal{W}_1) = \mathcal{W}_1^{1-\gamma}/(1-\gamma)$, with relative risk aversion $\gamma > 0$ and the log-utility limit at $\gamma = 1$. Under conditional lognormality,

$$\mathbb{E}[\mathcal{W}_1^{1-\gamma}] = \exp\left\{(1-\gamma)\mathbb{E}[\ell_p] + \frac{(1-\gamma)^2}{2}\mathbb{V}[\ell_p]\right\}.$$

It follows that expected utility is increasing in $\mathbb{E}[\ell_p] + (1-\gamma)\mathbb{V}[\ell_p]/2$. Combining (34) and (35) therefore gives the standard Campbell–Viceira objective (Campbell et al., 2003):

$$\boldsymbol{W}'\boldsymbol{\mu} - \frac{\gamma}{2}\boldsymbol{W}'\boldsymbol{\Sigma}\boldsymbol{W}. \tag{36}$$

The $-1/2$ variance term in expected log returns and the $(1-\gamma)/2$ variance term from CRRA utility combine to produce the coefficient $-\gamma/2$.

[22] This is the standard Campbell–Viceira approximation (Campbell et al., 2003). Under short-horizon scaling, expected log returns and return variances are first order, whereas squared expected log returns are second order; the approximation therefore retains the Jensen variance adjustment and omits squared expected returns.

As in the main text, the investor also faces an ℓ_2 penalty on portfolio positions, with penalty strength $\lambda \geq 0$. Her portfolio problem is therefore

$$\max_{\boldsymbol{W}} \quad \boldsymbol{W}'\boldsymbol{\mu} - \frac{\gamma}{2}\boldsymbol{W}'\boldsymbol{\Sigma}\boldsymbol{W} - \frac{\lambda}{2}\|\boldsymbol{W}\|_2^2. \tag{37}$$

The first-order condition and portfolio-weight demand are

$$\boldsymbol{\mu} = (\gamma\boldsymbol{\Sigma} + \lambda\boldsymbol{I})\boldsymbol{W}, \tag{38}$$

$$\boldsymbol{W} = \left[\gamma\boldsymbol{\Theta}\boldsymbol{E}\boldsymbol{\Theta}' + (\lambda + \gamma\sigma^2)\boldsymbol{I}\right]^{-1}\boldsymbol{\mu}. \tag{39}$$

Thus, the mean-variance problem in the main text is the standard quadratic representation of CRRA portfolio choice.

A.2 From Portfolio Weights to Log Shares

Let Q_n be the number of shares held, P_n the price per share, and A assets under management. Portfolio weights satisfy the accounting identity

$$W_n = \frac{P_n Q_n}{A}. \tag{40}$$

Writing $q_n \equiv \log Q_n$, $p_n \equiv \log P_n$, $a \equiv \log A$, and $\omega_n \equiv \log W_n$, the same identity becomes

$$q_n = a + \omega_n - p_n. \tag{41}$$

This price term matters. Even if an investor holds a portfolio weight fixed, a higher price requires fewer shares.

Fix reference AUM $\bar{a}$, reference log prices $\bar{\boldsymbol{p}}$, and reference weights $\boldsymbol{W}^{ref}$. The corresponding reference log shares are

$$\bar{q}_n = \bar{a} + \log W_n^{ref} - \bar{p}_n. \tag{42}$$

Define deviations from this reference by

$$\Delta^{ref} q_n \equiv q_n - \bar{q}_n, \qquad \Delta^{ref} p_n \equiv p_n - \bar{p}_n, \qquad \Delta^{ref} a \equiv a - \bar{a},$$
$$\Delta^{ref} \omega_n \equiv \omega_n - \log W_n^{ref}.$$

Subtracting the reference version of (41) from its current version gives the exact relationship

$$\Delta^{ref} \boldsymbol{\omega} = \Delta^{ref} \boldsymbol{q} + \Delta^{ref} \boldsymbol{p} - \Delta^{ref} a \, \boldsymbol{\iota}. \tag{43}$$

Portfolio weights satisfy

$$\boldsymbol{W} = \boldsymbol{W}^{ref} \odot \exp(\Delta^{ref} \boldsymbol{\omega}) \approx \boldsymbol{W}^{ref} + \mathrm{diag}(\boldsymbol{W}^{ref}) \Delta^{ref} \boldsymbol{\omega}, \tag{44}$$

where the exponential and $\odot$ operate element by element.

We use the equal-weight reference, $W_n^{ref} = 1/N$. This choice gives every stock the same conversion between an active portfolio weight and an active log-share position. It does not require the investor's equilibrium portfolio to remain equal weighted. Under this reference,

$$\boldsymbol{W} - \boldsymbol{W}^{ref} \approx \frac{1}{N} \Delta^{ref} \boldsymbol{\omega}. \tag{45}$$

Subtract the first-order condition at the reference portfolio from (38). Terms involving the reference portfolio lie in the market direction because $\boldsymbol{W}^{ref} = \boldsymbol{\iota}/N$. We absorb those terms into the expected return on the market and use $\boldsymbol{\mu}$ below for expected-return deviations from the reference. The results for the long-short factors $k > 0$ are then immediate. We include $k = 0$ in the formulas below by absorbing the reference-market term into the definition of the market-factor supply. Equations (38) and (45) then imply

$$\boldsymbol{\mu} = \frac{1}{N} \left[\gamma \boldsymbol{\Theta} \boldsymbol{E} \boldsymbol{\Theta}' + (\lambda + \gamma \sigma^2) \boldsymbol{I} \right] \Delta^{ref} \boldsymbol{\omega}. \tag{46}$$

Solving this equation for active log weights and using the accounting identity (43) gives raw log-share demand:

$$\Delta^{ref}\boldsymbol{q} = N\left[\gamma\boldsymbol{\Theta}\boldsymbol{E}\boldsymbol{\Theta}' + (\lambda+\gamma\sigma^2)\boldsymbol{I}\right]^{-1}\boldsymbol{\mu} - \Delta^{ref}\boldsymbol{p} + \Delta^{ref}a\,\boldsymbol{\iota}. \tag{47}$$

The three terms have distinct economic meanings. The first is the investor's portfolio response to expected returns. The second converts a desired portfolio weight into shares as prices change. The third is the common change in shares induced by AUM. Under the equal-weight reference, the AUM term lies entirely in the market direction and drops out of all long-short comparisons below.

A.3 Elasticities and the Degree of Substitution

Define, as in the main text,

$$c \equiv \lambda + \gamma\sigma^2, \qquad \psi \equiv \frac{\gamma\sigma^2}{\lambda+\gamma\sigma^2}, \qquad \widetilde{e}_k \equiv \frac{e_k}{\sigma^2}. \tag{48}$$

Thus $\psi \in [0,1]$. The classical CRRA benchmark has no portfolio-choice penalty, $\lambda = 0$, and hence $\psi = 1$; lowering ψ weakens substitution while holding total idiosyncratic position cost c fixed. The matrix that maps expected returns into portfolio weights has the spectral representation

$$\begin{aligned}&[\gamma\boldsymbol{\Theta}\boldsymbol{E}\boldsymbol{\Theta}' + c\boldsymbol{I}]^{-1} \\ &\quad = \frac{1}{c}\left[\boldsymbol{I} - \boldsymbol{\Theta}\mathrm{diag}\left(\frac{\psi\widetilde{e}_k}{1+\psi\widetilde{e}_k}\right)_{k=0}^{K}\boldsymbol{\Theta}'\right].\end{aligned} \tag{49}$$

To see why, note that the curvature in factor-k portfolios is $c(1+\psi\widetilde{e}_k)$, while curvature in a portfolio orthogonal to every factor is simply c. Equation (49) takes the reciprocal of each of these eigenvalues.

All large-N limits below hold K and the primitives c, ψ, σ^2, and $\{e_k\}_{k=0}^K$ fixed. More generally, the finite-N corrections vanish along any sequence satisfying

$$\max_{0\le k\le K} \frac{c(1+\psi\widetilde{e}_k)}{N} \longrightarrow 0. \tag{50}$$

When the primitives vary with N, the limiting formulas below are understood as asymptotic equivalents under (50). The exact finite-N formulas do not require this condition. In the dynamic extension, we additionally hold the common pass-through $\phi > 0$ fixed.

Hold AUM and expected payoffs fixed. Since $d\boldsymbol{\mu} = -d\boldsymbol{p}$ in the static model, differentiating (47) gives the finite-N log-share elasticity matrix

$$\frac{\partial \boldsymbol{q}}{\partial \boldsymbol{p}'} = -N\left[\gamma \boldsymbol{\Theta} \boldsymbol{E} \boldsymbol{\Theta}' + c\boldsymbol{I}\right]^{-1} - \boldsymbol{I}. \tag{51}$$

The $-\boldsymbol{I}$ term is the mechanical share adjustment from (41); it affects the same-stock response but not cross-stock responses.

The literal own-price elasticity is the magnitude of the diagonal derivative:

$$\begin{aligned}\zeta_{N,n}^{Own} &\equiv -\frac{\partial q_n}{\partial p_n} \\ &= 1 + \frac{N}{c} - \frac{N}{c}\sum_{k=0}^{K} \theta_{n,k}^2 \frac{\psi\widetilde{e}_k}{1+\psi\widetilde{e}_k}.\end{aligned} \tag{52}$$

The first two terms are the covariance-independent response. The final term removes the part of the response that would change the portfolio's systematic risk.

The micro, or relative, elasticity isolates the covariance-independent response. Consider stock n and a hypothetical stock $\widetilde{n}$ with the same factor loadings. Raise p_n and lower $p_{\widetilde{n}}$ by equal amounts, so the experiment changes only their relative price. Their systematic responses are identical and cancel. Therefore

$$\zeta_N^{Micro} \equiv -\left.\frac{\partial(q_n - q_{\widetilde{n}})}{\partial(p_n - p_{\widetilde{n}})}\right|_{\boldsymbol{\theta}_n = \boldsymbol{\theta}_{\widetilde{n}}} = 1 + \frac{N}{c}. \tag{53}$$

Because an active log-share position is N times its active portfolio-weight position under the equal-weight reference,

$$\lim_{N\to\infty} \frac{\zeta_N^{Micro}}{N} = \frac{1}{c} = \zeta^{Own}, \tag{54}$$

which is the covariance-independent portfolio response in (4).

For $m \neq n$, the cross-price elasticity is the off-diagonal derivative

$$\begin{aligned} \zeta_{N,nm}^{Cross} &\equiv \frac{\partial q_n}{\partial p_m} \\ &= \frac{N}{c} \sum_{k=0}^{K} \theta_{n,k}\theta_{m,k} \frac{\psi\widetilde{e}_k}{1+\psi\widetilde{e}_k}. \end{aligned} \tag{55}$$

Dividing (55) by (53) yields

$$\begin{aligned} \frac{\zeta_{N,nm}^{Cross}}{\zeta_N^{Micro}} &= \frac{N}{N+c} \sum_{k=0}^{K} \theta_{n,k}\theta_{m,k} \frac{\psi\widetilde{e}_k}{1+\psi\widetilde{e}_k} \\ &\xrightarrow{N\to\infty} \sum_{k=0}^{K} \theta_{n,k}\theta_{m,k} \frac{\psi\widetilde{e}_k}{1+\psi\widetilde{e}_k}. \end{aligned} \tag{56}$$

The limiting expression is exactly (6). Thus, the ratio of cross-price to micro elasticities identifies the same ψ in the CRRA log-share model as in the main model.

A.4 Price Multipliers

The multiplier matrix maps a reduction in the shares the representative investor must absorb into the resulting price increase. It is

$$\boldsymbol{M}_N \equiv -\frac{\partial \boldsymbol{p}}{\partial \boldsymbol{q}'}. \tag{57}$$

Inverting the negative of the elasticity matrix in (51) gives

$$\boldsymbol{M}_N = \left\{ N\left[\gamma \boldsymbol{\Theta} \boldsymbol{E} \boldsymbol{\Theta}' + c\boldsymbol{I}\right]^{-1} + \boldsymbol{I} \right\}^{-1}. \tag{58}$$

This is a full matrix inverse, not the element-by-element reciprocal of the elasticity matrix.

For a relative supply change between two stocks with identical factor loadings, the systematic part again cancels. The micro multiplier is

$$M_N^{Micro} = \frac{c}{N+c}. \tag{59}$$

This multiplier is expressed per unit of raw log-share supply. Since an active log-share position is N times its active portfolio-weight position,

$$\lim_{N\to\infty} N M_N^{Micro} = c = M^{\text{micro}}, \tag{60}$$

which is the micro multiplier used in the main model. Thus both the elasticity and multiplier levels differ across the two representations only by their quantity normalization. Using the eigenvalues described after (49), the full multiplier matrix can be written as

$$\boldsymbol{M}_N = M_N^{Micro}\left[\boldsymbol{I} + \boldsymbol{\Theta}\text{diag}\left(\frac{N\psi\widetilde{e}_k}{N + c(1+\psi\widetilde{e}_k)}\right)_{k=0}^{K}\boldsymbol{\Theta}'\right]. \tag{61}$$

The literal own-price multiplier is the diagonal element of this matrix,

$$M_{N,n}^{Own} = M_N^{Micro}\left[1 + \sum_{k=0}^{K}\theta_{n,k}^2\frac{N\psi\widetilde{e}_k}{N + c(1+\psi\widetilde{e}_k)}\right]. \tag{62}$$

For $m \neq n$, the off-diagonal element is the cross-price multiplier. The common-factor covariance between stocks n and m is

$$\tau_{n,m} \equiv \sum_{k=0}^{K}\theta_{n,k}\theta_{m,k}e_k.$$

Dividing the cross-price multiplier by the micro multiplier gives

$$\begin{aligned}\frac{M_{N,nm}^{Cross}}{M_N^{Micro}} &= \sum_{k=0}^{K} \theta_{n,k}\theta_{m,k} \frac{N\psi\widetilde{e}_k}{N + c(1+\psi\widetilde{e}_k)} \\ &\xrightarrow{N\to\infty} \psi \sum_{k=0}^{K} \theta_{n,k}\theta_{m,k}\widetilde{e}_k = \psi\frac{\tau_{n,m}}{\sigma^2}. \end{aligned} \tag{63}$$

This is the multiplier restriction in (24). Hence the elasticity and multiplier evidence identify the same ψ, even though elasticities and multipliers involve different covariance kernels.

A.5 Market Clearing and Equilibrium Expected Returns

Residual supply is most naturally expressed in the same log-share units as demand. Let

$$\boldsymbol{q} = \bar{\boldsymbol{q}} + N\boldsymbol{\Theta}\boldsymbol{\pi}, \tag{64}$$

where $\bar{\boldsymbol{q}}$ is the reference log-share portfolio and $\boldsymbol{\pi} \in \mathbb{R}^{K+1}$ collects the factor-supply tilts the representative investor must absorb. The factor N puts these tilts on the portfolio-weight scale at the equal-weight reference. Equivalently, market clearing requires

$$\Delta^{ref}\boldsymbol{q} = N\boldsymbol{\Theta}\boldsymbol{\pi}. \tag{65}$$

Set the reference expected-return intercept to zero and suppress the pure-market AUM term. With expected payoffs held fixed, (33) gives $\boldsymbol{\mu} = -\Delta^{ref}\boldsymbol{p}$. Substituting this relationship into log-share demand (47) gives

$$\Delta^{ref}\boldsymbol{q} = N\left[\gamma\boldsymbol{\Theta}\boldsymbol{E}\boldsymbol{\Theta}' + c\boldsymbol{I}\right]^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}. \tag{66}$$

The first term is the desired portfolio adjustment. The second appears because the price decline that raises expected returns also increases the number of shares associated with a given portfolio weight.

Equating demand (66) with supply (65) gives

$$N\boldsymbol{\Theta}\boldsymbol{\pi} = \left\{ N\left[\gamma\boldsymbol{\Theta}\boldsymbol{E}\boldsymbol{\Theta}' + c\boldsymbol{I}\right]^{-1} + \boldsymbol{I} \right\} \boldsymbol{\mu}. \tag{67}$$

This equation can be solved factor by factor. The curvature matrix has eigenvalue $c(1+\psi\widetilde{e}_k)$ for factor k. Therefore factor k's equilibrium expected return is

$$\mu_{N,k} \equiv \boldsymbol{\Theta}'_{\cdot,k}\boldsymbol{\mu} = b_{N,k}\pi_k, \tag{68}$$

$$b_{N,k} \equiv \frac{Nc(1+\psi\widetilde{e}_k)}{N + c(1+\psi\widetilde{e}_k)}. \tag{69}$$

The numerator is the compensation implied by portfolio curvature and factor supply. The denominator accounts for the fact that a price change alters both expected returns and the number of shares associated with a given portfolio weight.

Under (50), the finite-N correction vanishes factor by factor:

$$\frac{b_{N,k}}{c(1+\psi\widetilde{e}_k)} = \frac{N}{N + c(1+\psi\widetilde{e}_k)} \xrightarrow{N\to\infty} 1. \tag{70}$$

Consequently,

$$\boldsymbol{\mu} \xrightarrow{N\to\infty} \frac{1}{\zeta^{Own}} \boldsymbol{\Theta} \left[\pi_k(1+\psi\widetilde{e}_k)\right]_{k=0}^{K}, \tag{71}$$

because $1/\zeta^{Own} = c$. This is exactly the equilibrium expected-return equation (8) in the main text.

Expected-return variance shares. Orthonormality of $\boldsymbol{\Theta}$ implies that factor k contributes $b_{N,k}^2\pi_k^2$ to the squared norm of equilibrium expected returns. Suppose factor supplies are independent and have a common second moment. Averaging each factor's

contribution over supply realizations before normalizing gives the finite-N share

$$\begin{aligned} f_{N,k} &\equiv \frac{\mathbb{E}[b_{N,k}^2 \pi_k^2]}{\sum_{j=1}^K \mathbb{E}[b_{N,j}^2 \pi_j^2]} \\ &= \frac{b_{N,k}^2}{\sum_{j=1}^K b_{N,j}^2}, \qquad k > 0. \end{aligned} \tag{72}$$

Using (70),

$$f_{N,k} \xrightarrow{N\to\infty} \frac{(1+\psi\widetilde{e}_k)^2}{\sum_{j=1}^K (1+\psi\widetilde{e}_j)^2} = f_k. \tag{73}$$

Thus Proposition 1 applies to the CRRA equilibrium: weaker substitution spreads equilibrium expected-return variation across more factors.

Expected-return spanning. The same calculation connects the CRRA equilibrium to the return-spanning regressions. Suppose realized returns equal equilibrium expected returns plus a mean-zero residual $\boldsymbol{v}$ that is orthogonal to the factor-loading space in population. Under the same equal-second-moment assumption used above, averaging over factor-supply realizations implies that a cross-sectional projection on factors $1, \ldots, J$ captures expected-return variation proportional to $\sum_{k=1}^J b_{N,k}^2$. The total cross-sectional variance of realized returns is the same in the regressions with J and K factors. It therefore cancels when their R^2 values are divided, leaving

$$\frac{R^2_{realized,N}(J)}{R^2_{realized,N}(K)} = \frac{\sum_{k=1}^J b_{N,k}^2}{\sum_{k=1}^K b_{N,k}^2} \xrightarrow{N\to\infty} \sum_{k=1}^J f_k. \tag{74}$$

This is the normalized spanning restriction used in the empirical analysis.

Maximum attainable Sharpe ratio. At finite N, the expected-return vector has factor coefficient $b_{N,k}\pi_k$. Because

$$\boldsymbol{\Sigma}^{-1}\boldsymbol{\Theta}_{\cdot,k} = \frac{1}{e_k + \sigma^2}\boldsymbol{\Theta}_{\cdot,k}, \tag{75}$$

the maximum squared Sharpe ratio is

$$\mathrm{SR}^2_{N,\max} = \boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} = \sum_{k=0}^{K} \pi_k^2 \frac{b_{N,k}^2}{e_k + \sigma^2}. \tag{76}$$

Taking the factor-by-factor limit and using $e_k + \sigma^2 = (1 + \widetilde{e}_k)\sigma^2$ yields the main-text expression:

$$\mathrm{SR}^2_{N,\max} \xrightarrow{N\to\infty} c^2 \sum_{k=0}^{K} \pi_k^2 \frac{(1 + \psi\widetilde{e}_k)^2}{(1 + \widetilde{e}_k)\sigma^2}. \tag{77}$$

Holding c and factor supplies fixed, each $b_{N,k}$ is increasing in ψ. Therefore weaker substitution lowers the maximum attainable Sharpe ratio in both the finite-N CRRA economy and its limiting representation.

A.6 Common Dynamic Price Pass-Through

The baseline model is static: holding the next-period payoff fixed, a one-percent increase in the current price lowers the expected return one for one. With a long-lived asset, however, the current price may also change investors' beliefs about the future price and dividend. Campbell–Shiller log-linearization makes the mechanical and belief effects explicit.

Let $P_{n,t}$ denote stock n's ex-dividend price, let $D_{n,t+1}$ be its next dividend, and write $p_{n,t} \equiv \log P_{n,t}$ and $d_{n,t} \equiv \log D_{n,t}$. The stock's realized gross log return is

$$\ell_{n,t+1} \equiv \log\left(\frac{P_{n,t+1} + D_{n,t+1}}{P_{n,t}}\right). \tag{78}$$

Linearizing this identity around a constant reference log dividend-price ratio gives (Campbell and Shiller, 1988)

$$\ell_{n,t+1} \approx \kappa^{CS} + \rho^{CS} p_{n,t+1} + (1 - \rho^{CS}) d_{n,t+1} - p_{n,t}, \tag{79}$$

where

$$\rho^{CS} \equiv \frac{\bar{P}}{\bar{P}+\bar{D}} = \left[1+\exp(\bar{d}-\bar{p})\right]^{-1}, \quad \kappa^{CS} \equiv -\rho^{CS}\log\rho^{CS} - (1-\rho^{CS})\log(1-\rho^{CS}). \tag{80}$$

Both ρ^{CS} and κ^{CS} are constants of the log-linearization, not time-varying state variables. The coefficient $\rho^{CS} \in (0,1)$ is the reference continuation-price share of next period's cum-dividend payoff. Equivalently, the reference gross payoff-price ratio is $(\bar{P}+\bar{D})/\bar{P} = 1/\rho^{CS}$; it is the reciprocal of the coefficient in (79).

Let $\ell^e_{n,t+1} \equiv \ell_{n,t+1} - \ell_{f,t}$ and $\mu^\ell_{n,t} \equiv \mathbb{E}_t[\ell^e_{n,t+1}]$, where $\ell_{f,t}$ is the log risk-free return. Under the normalization above, $\ell_{f,t} = 0$; retaining it here makes the distinction between log and simple excess returns explicit. The risk-free return is known at time t and is locally unaffected by stock n's own price. Holding the Jensen adjustment locally fixed, the price derivative of $\mu^\ell_{n,t}$ is also the first-order derivative of the expected simple excess return $\mu_{n,t}$ used in the portfolio problem. Taking conditional expectations in (79) therefore gives the local price pass-through of prices into expected returns (see Davis et al., 2025, for more discussion on this pass-through term):

$$\phi_n \equiv -\frac{\partial \mu_{n,t}}{\partial p_{n,t}} \approx -\frac{\partial \mu^\ell_{n,t}}{\partial p_{n,t}} = 1 - \rho^{CS}\frac{\partial \mathbb{E}_t[p_{n,t+1}]}{\partial p_{n,t}} - (1-\rho^{CS})\frac{\partial \mathbb{E}_t[d_{n,t+1}]}{\partial p_{n,t}}. \tag{81}$$

If expected future prices and dividends are held fixed, both belief-response terms are zero. The current price then enters (79) with coefficient minus one, so

$$\frac{\partial \mu^\ell_{n,t}}{\partial p_{n,t}} = -1, \qquad \phi_n = 1. \tag{82}$$

This is the mechanical, static benchmark. It is not generally the investor's full response: an observed price movement is informative, and a rational investor may revise beliefs about the continuation price and dividends.

To isolate the continuation-price channel, consider a price movement unrelated to cash-flow news, so that expected future dividends remain fixed, and define its perceived

persistence by

$$\chi_n \equiv \frac{\partial \mathbb{E}_t[p_{n,t+1}]}{\partial p_{n,t}}, \qquad 0 \leq \chi_n \leq 1. \tag{83}$$

Equation (81) reduces to

$$\phi_n = 1 - \rho^{CS}\chi_n. \tag{84}$$

If the price movement is expected to reverse fully by $t+1$, then $\chi_n = 0$ and $\phi_n = 1$. If it is expected to persist partly, then $0 < \chi_n < 1$ and pass-through lies strictly between the two bounds. If the future price is expected to move one for one with the current price, then $\chi_n = 1$ and $\phi_n = 1 - \rho^{CS}$. Hence

$$1 - \rho^{CS} \leq \phi_n \leq 1. \tag{85}$$

The lower endpoint is the dividend share of next period's payoff, $1 - \rho^{CS} = \bar{D}/(\bar{P} + \bar{D})$, which is approximately the dividend-price ratio when that ratio is small. Even with one-for-one continuation-price persistence, this dividend component prevents the expected-return response from disappearing. Davis et al. (2025) develop and estimate this price pass-through object in depth and emphasize the same interpretation: lower pass-through means that price movements associated with expected-return variation are more persistent.

For the comparative statics below, assume a common local own-price pass-through $\phi > 0$ and no cross-stock belief response. In matrix form,

$$-\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{p}'} = \phi \boldsymbol{I}, \qquad \phi \in [1 - \rho^{CS}, 1]. \tag{86}$$

This is a local restriction on expected-return responses; it does not require expected returns to be globally linear in prices. A smaller ϕ means that a given price change

produces a smaller change in expected returns. Differentiating (47) now gives

$$\frac{\partial \boldsymbol{q}}{\partial \boldsymbol{p}'} = -\phi N \left[\gamma \boldsymbol{\Theta} \boldsymbol{E} \boldsymbol{\Theta}' + c\boldsymbol{I}\right]^{-1} - \boldsymbol{I}. \tag{87}$$

Accordingly,

$$\zeta_N^{Micro}(\phi) = 1 + \frac{\phi N}{c}, \tag{88}$$

$$\zeta_{N,nm}^{Cross}(\phi) = \frac{\phi N}{c} \sum_{k=0}^{K} \theta_{n,k}\theta_{m,k} \frac{\psi \widetilde{e}_k}{1 + \psi \widetilde{e}_k}. \tag{89}$$

Their ratio is

$$\begin{aligned} \frac{\zeta_{N,nm}^{Cross}(\phi)}{\zeta_N^{Micro}(\phi)} &= \frac{\phi N}{\phi N + c} \sum_{k=0}^{K} \theta_{n,k}\theta_{m,k} \frac{\psi \widetilde{e}_k}{1 + \psi \widetilde{e}_k} \\ &\xrightarrow{N\to\infty} \sum_{k=0}^{K} \theta_{n,k}\theta_{m,k} \frac{\psi \widetilde{e}_k}{1 + \psi \widetilde{e}_k}. \end{aligned} \tag{90}$$

The dynamic multiplier matrix is the inverse of the negative derivative in (87). Its micro multiplier and finite-N cross-to-micro ratio are

$$M_N^{Micro}(\phi) = \frac{c}{\phi N + c}, \tag{91}$$

$$\begin{aligned} \frac{M_{N,nm}^{Cross}(\phi)}{M_N^{Micro}(\phi)} &= \sum_{k=0}^{K} \theta_{n,k}\theta_{m,k} \frac{\phi N \psi \widetilde{e}_k}{\phi N + c(1 + \psi \widetilde{e}_k)} \\ &\xrightarrow{N\to\infty} \psi \frac{\tau_{n,m}}{\sigma^2}. \end{aligned} \tag{92}$$

Thus a common dynamic pass-through changes the levels of elasticities and multipliers, but not the limiting relative-response restrictions that identify ψ. Heterogeneous or cross-asset pass-through need not cancel in this way.

The CRRA model therefore reproduces the paper's complete economic mechanism in the units used by the empirical analysis. The parameter ψ controls substitution, market clearing maps substitution into the factor structure of expected returns, and weaker substitution simultaneously produces a factor zoo and lowers the maximum attainable

Sharpe ratio.

B Theoretical Proofs and Extensions

B.1 Equilibrium Expected Returns and the Factor Zoo

B.1.1 Proof of (8)

$$\begin{aligned}
\mathbb{E}[\boldsymbol{D}] - \boldsymbol{P} &= \left[\gamma \boldsymbol{\Theta}\mathbf{E}\boldsymbol{\Theta}' + \left(\lambda + \gamma\sigma^2\right)\mathbf{I}\right]\left(\frac{1}{N}\boldsymbol{\iota} + \boldsymbol{\Theta}\boldsymbol{\pi}\right) \\
&\xrightarrow{N\to\infty} \left[\gamma \boldsymbol{\Theta}\mathbf{E}\boldsymbol{\Theta}' + \left(\lambda + \gamma\sigma^2\right)\mathbf{I}\right]\boldsymbol{\Theta}\boldsymbol{\pi} \\
&= \left(\lambda + \gamma\sigma^2\right)\boldsymbol{\Theta}\left[\psi\frac{1}{\sigma^2}\mathbf{E} + \mathbf{I}\right]\boldsymbol{\pi} \\
&= \boldsymbol{\Theta}\underbrace{\left[\left(\lambda + \gamma\sigma^2\right)\pi_k\left(\psi\tilde{e}_k + 1\right)\right]_{k=0}^{K}}_{\equiv\boldsymbol{\nu}},
\end{aligned} \tag{93}$$

where the third line follows from the orthonormality of $\boldsymbol{\Theta}$.

B.1.2 Proof of (9)

Since $\boldsymbol{\Theta}$ is orthonormal, the cross-sectional variance of expected returns is:

$$\begin{aligned}
\mathbb{V}[\mathbb{E}[D_n] - P_n] &= \mathbb{E}[(\mathbb{E}[D_n] - P_n)^2] - (\mathbb{E}[\mathbb{E}[D_n] - P_n])^2 \\
&= \frac{1}{N}\boldsymbol{\nu}'\boldsymbol{\Theta}'\boldsymbol{\Theta}\boldsymbol{\nu} - \left(\frac{1}{N}\boldsymbol{\iota}'\boldsymbol{\Theta}\boldsymbol{\nu}\right)^2 \\
&= \frac{1}{N}\boldsymbol{\nu}'\boldsymbol{\nu} - \frac{1}{N}\left(\left(\lambda + \gamma\sigma^2\right)\pi_0\left(\psi\tilde{e}_0 + 1\right)\right)^2 \\
&= \frac{1}{N}\left(\lambda + \gamma\sigma^2\right)^2\sum_{k=1}^{K}\pi_k^2\left(\psi\tilde{e}_k + 1\right)^2
\end{aligned} \tag{94}$$

The third line follows from the orthonormality of $\boldsymbol{\Theta}$ and because the first column of $\boldsymbol{\Theta}$ is $[1/\sqrt{N}]_{n=1}^{N}$ while all columns $k > 0$ of $\boldsymbol{\Theta}$ satisfy $\boldsymbol{\iota}'\boldsymbol{\Theta}_{k,\cdot} = 0$.

The proportion of cross-sectional expected return variance explained by (exposures

to) the k-th factor (for $k > 0$) is:

$$\tilde{f}_k \equiv \frac{\pi_k^2 \left(\psi \tilde{e}_k + 1\right)^2}{\sum_{j=1}^K \pi_j^2 \left(\psi \tilde{e}_j + 1\right)^2}. \tag{95}$$

Consider simulating this economy many times and drawing a new $\boldsymbol{\pi}$ each time, where each element π_k is i.i.d. both within and across draws. Then we can linearize (95) around $\pi_k^2 = \mathbb{E}\left[\pi_k^2\right], \forall k$ to obtain:

$$\begin{aligned}\tilde{f}_k \approx & \frac{\left(\psi \tilde{e}_k + 1\right)^2}{\sum_{j=1}^K \left(\psi \tilde{e}_j + 1\right)^2} + \frac{\left(\psi \tilde{e}_k + 1\right)^2}{\sum_{j=1}^K \mathbb{E}\left[\pi_j^2\right] \left(\psi \tilde{e}_j + 1\right)^2} \left(\pi_k^2 - \mathbb{E}\left[\pi_k^2\right]\right) \\ & - \sum_{j=1}^K \frac{\mathbb{E}\left[\pi_k^2\right] \left(\psi \tilde{e}_k + 1\right)^2 \left(\psi \tilde{e}_j + 1\right)^2}{\left[\sum_{l=1}^K \mathbb{E}\left[\pi_l^2\right] \left(\psi \tilde{e}_l + 1\right)^2\right]^2} \left(\pi_j^2 - \mathbb{E}\left[\pi_j^2\right]\right).\end{aligned}$$

Taking the expectation $\mathbb{E}[\tilde{f}_k]$ yields (9):

$$\begin{aligned} f_k \equiv & \mathbb{E}[\tilde{f}_k] \\ = & \frac{\mathbb{E}[\pi_k^2] \left(\psi \tilde{e}_k + 1\right)^2}{\sum_{j=1}^K \mathbb{E}[\pi_k^2] \left(\psi \tilde{e}_j + 1\right)^2} \\ = & \frac{\left(\psi \tilde{e}_k + 1\right)^2}{\sum_{j=1}^K \left(\psi \tilde{e}_j + 1\right)^2} \end{aligned} \tag{96}$$

B.1.3 Proof of Proposition 1

We use the following lemma, which is proven in Appendix B.1.4.

Lemma 1. *As substitution increases (i.e., as ψ increases), the importance of stronger factors relative to weaker ones in explaining cross-sectional variation in expected returns increases. Specifically, for any $e_k > e_j$, the ratio f_k / f_j is increasing in ψ.*

Fix $J < K$ and let $\psi_1 < \psi_2$. For any $k \leq J < m$, we have $e_k \geq e_m$ by ordering, so Lemma 1 implies that

$$\frac{f_k(\psi_2)}{f_m(\psi_2)} > \frac{f_k(\psi_1)}{f_m(\psi_1)}.$$

Summing this inequality over $k = 1, \ldots, J$ gives

$$\begin{aligned}
\frac{\sum_{k=1}^{J} f_k(\psi_2)}{f_m(\psi_2)} &> \frac{\sum_{k=1}^{J} f_k(\psi_1)}{f_m(\psi_1)} \\
\leftrightarrow \sum_{k=1}^{J} f_k(\psi_2) \cdot f_m(\psi_1) &> \sum_{k=1}^{J} f_k(\psi_1) \cdot f_m(\psi_2) \\
\rightarrow \sum_{k=1}^{J} f_k(\psi_2) \cdot \sum_{m=J+1}^{K} f_m(\psi_1) &> \sum_{k=1}^{J} f_k(\psi_1) \cdot \sum_{m=J+1}^{K} f_m(\psi_2) \\
\leftrightarrow \sum_{k=1}^{J} f_k(\psi_2) \cdot \left(1 - \sum_{k=1}^{J} f_k(\psi_1)\right) &> \sum_{k=1}^{J} f_k(\psi_1) \cdot \left(1 - \sum_{k=1}^{J} f_k(\psi_2)\right) \\
\leftrightarrow \sum_{k=1}^{J} f_k(\psi_2) &> \sum_{k=1}^{J} f_k(\psi_1),
\end{aligned}$$

as desired.

B.1.4 Proof of Lemma 1

Let $e_k > e_j$. Then

$$\frac{f_k}{f_j} = \left(\frac{\psi\tilde{e}_k + 1}{\psi\tilde{e}_j + 1}\right)^2.$$

Since $(\psi\tilde{e}_k + 1)/(\psi\tilde{e}_j + 1) > 0$, it suffices to show that

$$\frac{\partial}{\partial\psi}\frac{\psi\tilde{e}_k + 1}{\psi\tilde{e}_j + 1} > 0.$$

Note that

$$\begin{aligned}
\frac{\partial}{\partial\psi}\frac{\psi\tilde{e}_k + 1}{\psi\tilde{e}_j + 1} &= \frac{\tilde{e}_k - \tilde{e}_j}{(\psi\tilde{e}_j + 1)^2} \\
&> 0
\end{aligned}$$

since $\tilde{e}_k > \tilde{e}_j$ because $e_k > e_j$.

B.2 Absence of Near Arbitrage Opportunities

The maximum Sharpe ratio portfolio is

$$\boldsymbol{Q}_* = \Sigma^{-1}\mathbb{E}[\mathbf{D}-\mathbf{P}].$$

So the maximum attainable (squared) Sharpe ratio is

$$\begin{aligned}
\text{Max Sharpe Ratio}^2 &= \frac{\left(\boldsymbol{Q}_*'\mathbb{E}[\mathbf{D}-\mathbf{P}]\right)^2}{\boldsymbol{Q}_*'\Sigma\boldsymbol{Q}_*} \\
&= \frac{\left(\mathbb{E}[\mathbf{D}-\mathbf{P}]'\Sigma^{-1}\mathbb{E}[\mathbf{D}-\mathbf{P}]\right)^2}{\mathbb{E}[\mathbf{D}-\mathbf{P}]'\Sigma^{-1}\Sigma\Sigma^{-1}\mathbb{E}[\mathbf{D}-\mathbf{P}]} \\
&= \mathbb{E}[\mathbf{D}-\mathbf{P}]'\Sigma^{-1}\mathbb{E}[\mathbf{D}-\mathbf{P}] \\
&= \left(\lambda+\gamma\sigma^2\right)^2\boldsymbol{\pi}'\boldsymbol{\Theta}'\left[\mathbf{I}+\psi\frac{1}{\sigma^2}\boldsymbol{\Theta}\mathbf{E}\boldsymbol{\Theta}'\right]\Sigma^{-1}\left[\mathbf{I}+\psi\frac{1}{\sigma^2}\boldsymbol{\Theta}\mathbf{E}\boldsymbol{\Theta}'\right]\boldsymbol{\Theta}\boldsymbol{\pi}
\end{aligned}$$

where the fourth line follows from the matrix expression for equilibrium expected returns (93) from Appendix B.1.1. Since $\Sigma = \boldsymbol{\Theta}\mathbf{E}\boldsymbol{\Theta}' + \sigma^2\mathbf{I}$, we have

$$\begin{aligned}
\Sigma\boldsymbol{\Theta} &= \left(\boldsymbol{\Theta}\mathbf{E}\boldsymbol{\Theta}' + \sigma^2\mathbf{I}\right)\boldsymbol{\Theta} \\
&= \boldsymbol{\Theta}\left(\mathbf{E}+\sigma^2\mathbf{I}\right),
\end{aligned}$$

using the orthonormality of $\boldsymbol{\Theta}$. Therefore,

$$\Sigma^{-1}\boldsymbol{\Theta} = \boldsymbol{\Theta}\left(\mathbf{E}+\sigma^2\mathbf{I}\right)^{-1},$$

and so

$$\begin{aligned}
\text{Max Sharpe Ratio}^2 &= \left(\lambda+\gamma\sigma^2\right)^2\boldsymbol{\pi}'\left(\mathbf{I}+\psi\frac{1}{\sigma^2}\mathbf{E}\right)\boldsymbol{\Theta}'\Sigma^{-1}\boldsymbol{\Theta}\left(\mathbf{I}+\psi\frac{1}{\sigma^2}\mathbf{E}\right)\boldsymbol{\pi} \\
&= \left(\lambda+\gamma\sigma^2\right)^2\boldsymbol{\pi}'\left(\mathbf{I}+\psi\frac{1}{\sigma^2}\mathbf{E}\right)\left(\mathbf{E}+\sigma^2\mathbf{I}\right)^{-1}\left(\mathbf{I}+\psi\frac{1}{\sigma^2}\mathbf{E}\right)\boldsymbol{\pi} \\
&= \left(\frac{1}{\zeta^{Own}}\right)^2\sum_{k=0}^{K}\pi_k^2\frac{\left(1+\psi\tilde{e}_k\right)^2}{\left(\tilde{e}_k+1\right)\sigma^2}
\end{aligned}$$

Note that we have reparameterized the model in terms of the micro elasticity ζ^{Own} and degree of substitution ψ instead of in terms of risk aversion γ and the ℓ_2 penalty λ. Since ζ^{Own} and ψ are parameters of both γ and λ, any comparative static that involves changing ψ while holding ζ^{Own} fixed can be expressed as a comparative static that involves changing γ (to change ψ) while adjusting λ appropriately (to maintain fixed ζ^{Own}).

Thus, if we hold the micro elasticity ζ^{Own} fixed (i.e. hold $\lambda + \gamma\sigma^2$ fixed) and reduce substitution ψ by lowering γ (and increasing λ appropriately), the maximum Sharpe *decreases*.

B.3 Measuring Proportion of Cross Sectional Expected Return Variation Explained

B.3.1 Proof of (15)

We can measure the proportion of cross-sectional expected return variation explained by the first J factors ($\sum_{k=1}^{J} f_k$) using cross-sectional predictive regressions of returns on factor loadings.

Realized returns are expected returns plus unforecastable noise: $D_n - P_n = \mathbb{E}[D_n] - P_n + v_n$, where $v_n \sim N(0, \sigma_v^2)$ is i.i.d. Then the variation in returns (following (94)) is:

$$\begin{aligned}\mathbb{V}[D_n - P_n] &= \mathbb{V}[\mathbb{E}[D_n] - P_n] + \mathbb{V}[v_n] \\ &= \frac{1}{N}\left(\lambda + \gamma\sigma^2\right)^2 \sum_{k=1}^{K} \pi_k^2 \left(\psi\tilde{e}_k + 1\right)^2 + \sigma_\nu^2\end{aligned}$$

In this case, the R^2 from a cross-sectional regression of returns on factor loadings for the first J factors

$$D_n - P_n = \sum_{k=1}^{J} \beta_k \theta_{n,k} + v_n$$

is for $A = \frac{1}{N}\left(\lambda + \gamma\sigma^2\right)^2$

$$
\begin{aligned}
R^2_{realized}(J) &= \frac{A\sum_{k=1}^{J}\pi_k^2\left(\psi\tilde{e}_k+1\right)^2}{A\sum_{k=1}^{K}\pi_k^2\left(\psi\tilde{e}_k+1\right)^2+\sigma_v^2} \\
&\approx \frac{A\mathbb{E}\left[\pi_k^2\right]\sum_{k=1}^{J}\left(\psi\tilde{e}_k+1\right)^2}{A\mathbb{E}\left[\pi_k^2\right]\sum_{k=1}^{K}\pi_k^2\left(\psi\tilde{e}_k+1\right)^2+\sigma_v^2} + \frac{A\mathbb{E}\left[\pi_k^2\right]\sum_{k=1}^{J}\left(\psi\tilde{e}_k+1\right)^2}{A\mathbb{E}\left[\pi_k^2\right]\sum_{k=1}^{K}\pi_k^2\left(\psi\tilde{e}_k+1\right)^2+\sigma_v^2}\left(\pi_k^2-\mathbb{E}\left[\pi_k^2\right]\right) \\
&\quad -\sum_{j=1}^{K}\frac{A\mathbb{E}\left[\pi_k^2\right]\sum_{k=1}^{J}\left(\psi\tilde{e}_k+1\right)^2\left(\psi\tilde{e}_j+1\right)^2}{\left[A\mathbb{E}\left[\pi_k^2\right]\sum_{k=1}^{K}\left(\psi\tilde{e}_k+1\right)^2+\sigma_v^2\right]^2}\left(\pi_j^2-\mathbb{E}\left[\pi_k^2\right]\right) \\
&= \frac{\sum_{k=1}^{J}\left(\psi\tilde{e}_k+1\right)^2}{\sum_{k=1}^{K}\left(\psi\tilde{e}_k+1\right)^2+\frac{\sigma_v^2}{A\mathbb{E}\left[\pi_k^2\right]}} + \frac{\sum_{k=1}^{J}\left(\psi\tilde{e}_k+1\right)^2}{\sum_{k=1}^{K}\left(\psi\tilde{e}_k+1\right)^2+\frac{\sigma_v^2}{A\mathbb{E}\left[\pi_k^2\right]}}\left(\pi_k^2-\mathbb{E}\left[\pi_k^2\right]\right) \\
&\quad -\sum_{j=1}^{K}\frac{\sum_{k=1}^{J}\left(\psi\tilde{e}_k+1\right)^2\left(\psi e_j+1\right)^2}{\left[\sum_{k=1}^{K}\left(\psi\tilde{e}_k+1\right)^2+\frac{\sigma_v^2}{A\mathbb{E}\left[\pi_k^2\right]}\right]^2}\left(\pi_j^2-\mathbb{E}\left[\pi_k^2\right]\right),
\end{aligned}
$$

where the second line follows from a linearization around $\pi_k^2 = \mathbb{E}\left[\pi_k^2\right]$. Taking the expectation over many draws of $\boldsymbol{\pi}$ we have on average

$$
R^2_{realized}(J) = \frac{\sum_{k=1}^{J}\left(\psi\tilde{e}_k+1\right)^2}{\sum_{k=1}^{K}\left(\psi\tilde{e}_k+1\right)^2+\frac{\sigma_v^2}{A\mathbb{E}\left[\pi_k^2\right]}}.
$$

The following ratio of $R^2_{realized}(J)/R^2_{realized}(K)$ recovers the desired sum:

$$
\begin{aligned}
\frac{R^2_{realized}(J)}{R^2_{realized}(K)} &= \frac{\sum_{k=1}^{J}\left(\psi\tilde{e}_k+1\right)^2}{\sum_{k=1}^{K}\left(\psi\tilde{e}_k+1\right)^2} \\
&= \sum_{k=1}^{J} f_k.
\end{aligned}
$$

B.4 Generalized Model

In this appendix we present a generalized version of the model from Section 2 that allows for a more flexible specification of preferences. We show that our main theoretical results continue to hold in this generalized model.

Assets are as specified in Section 2.1.

The representative investor has a more general utility function:

$$\max_{\boldsymbol{Q}} \ \boldsymbol{Q}'\big(\mathbb{E}[\boldsymbol{D}] - \boldsymbol{P}\big) - \frac{\gamma}{2}\boldsymbol{Q}'\boldsymbol{\Sigma}\boldsymbol{Q} - \frac{\lambda}{2}\left\|\boldsymbol{Q} - \bar{\boldsymbol{Q}}\right\|_2^2 + \boldsymbol{\delta}'\boldsymbol{Q},$$

This setup generalizes the preferences in Section 2.1 in two ways. First, we now allow for an arbitrary shrinkage target $\bar{\boldsymbol{Q}}$, whereas our baseline model shrinks toward zero. The target $\bar{\boldsymbol{Q}}$ can naturally be interpreted as a benchmark portfolio or the investor's lagged portfolio in a dynamic model with quadratic adjustment costs. Second, we now allow for non-pecuniary preferences, as captured by the vector $\boldsymbol{\delta}$. For example, each δ_n can represent the investor's preference for green stocks times stock n's greenness characteristic. $\boldsymbol{\delta}$ can also represent hedging demand for particular assets or sentiment. If parameterized as $\delta_n = k\sigma_n^2$, where σ_n^2 is stock n's idiosyncratic variance, then idiosyncratic risk will be priced and this model can nest the channels of Merton et al. (1987); Dello Preite et al. (2024).

Given these preferences, the optimal quantity demanded of each stock is given by:

$$\boldsymbol{Q} = \big[\gamma\boldsymbol{\Theta}\mathbf{E}\boldsymbol{\Theta}' + (\lambda + \gamma\sigma^2)\mathbf{I}\big]^{-1}\big(\mathbb{E}[\boldsymbol{D}] - \boldsymbol{P} + \lambda\bar{\boldsymbol{Q}} + \boldsymbol{\delta}\big).$$

Elasticity matrix is unchanged The derivative of quantity demanded with respect to price is

$$\frac{\partial \boldsymbol{Q}}{\partial \boldsymbol{P}'} = -\big[\gamma\boldsymbol{\Theta}\mathbf{E}\boldsymbol{\Theta}' + (\lambda + \gamma\sigma^2)\mathbf{I}\big]^{-1}, \tag{97}$$

which matches the elasticity matrix in the baseline model (3).

Equilibrium expected returns Imposing market clearing $\boldsymbol{Q} = \boldsymbol{\iota}/N + \boldsymbol{\Theta}\boldsymbol{\pi}$ and letting $N \to \infty$ yields

$$\mathbb{E}[\boldsymbol{D}] - \boldsymbol{P} = (\lambda + \gamma\sigma^2)\,\boldsymbol{\Theta}(\mathbf{I} + \psi\frac{1}{\sigma^2}\mathbf{E})\boldsymbol{\pi} \ - \ \lambda\bar{\boldsymbol{Q}} \ - \ \boldsymbol{\delta}. \tag{98}$$

We can decompose the target portfolio $\bar{\boldsymbol{Q}}$ and taste vector $\boldsymbol{\delta}$ into factor-spanned and orthogonal components:

$$\bar{\boldsymbol{Q}} = \boldsymbol{\Theta}\bar{\boldsymbol{q}} + \bar{\boldsymbol{Q}}^{\perp}, \qquad \boldsymbol{\delta} = \boldsymbol{\Theta}\boldsymbol{d} + \boldsymbol{\delta}^{\perp},$$

where $\bar{\boldsymbol{q}} = \boldsymbol{\Theta}'\bar{\boldsymbol{Q}} \in \mathbb{R}^{K+1}$ and $\boldsymbol{d} = \boldsymbol{\Theta}'\boldsymbol{\delta} \in \mathbb{R}^{K+1}$ are the projections onto the factor space, and $\bar{\boldsymbol{Q}}^{\perp} = \boldsymbol{M}_{\Theta}\bar{\boldsymbol{Q}}$, $\boldsymbol{\delta}^{\perp} = \boldsymbol{M}_{\Theta}\boldsymbol{\delta}$ with $\boldsymbol{M}_{\Theta} \equiv \mathbf{I} - \boldsymbol{\Theta}\boldsymbol{\Theta}'$ are the orthogonal residuals. Substituting into (98),

$$\mathbb{E}[\boldsymbol{D}] - \boldsymbol{P} = \boldsymbol{\Theta}\left[\nu_k\right]_{k=0}^{K} \; - \; \lambda\bar{\boldsymbol{Q}}^{\perp} \; - \; \boldsymbol{\delta}^{\perp},$$

where ν_k is the expected return on factor-k:

$$\nu_k \;\equiv\; (\lambda + \gamma\sigma^2)(1 + \psi\tilde{e}_k)\pi_k \;-\; \lambda\bar{q}_k \;-\; d_k. \tag{99}$$

Thus, the target portfolio and non-pecuniary preference generate two wedges in factor-k's expected return beyond the supply-driven term from (8) in the baseline model.

To simplify notation, redefine factor supply as

$$\tilde{\pi}_k \;\equiv\; \pi_k \;-\; \frac{\lambda\bar{q}_k + d_k}{(\lambda + \gamma\sigma^2)(1 + \psi\tilde{e}_k)}. \tag{100}$$

Then (99) simplifies to

$$\nu_k = (\lambda + \gamma\sigma^2)(1 + \psi\tilde{e}_k)\,\tilde{\pi}_k, \tag{101}$$

which is analogous to (8) in the baseline model. Thus, we can express stock-level expected returns as:

$$\mathbb{E}[\boldsymbol{D}] - \boldsymbol{P} = \boldsymbol{\Theta}\big[(\lambda + \gamma\sigma^2)(1 + \psi\tilde{e}_k)\tilde{\pi}_k\big]_{k=0}^{K} - \lambda\bar{\boldsymbol{Q}}^{\perp} - \boldsymbol{\delta}^{\perp}. \tag{102}$$

Proportion of Cross-Sectional Expected Return Variation Explained Writing factor k's second moment as

$$m_{\tilde{\pi},k} \equiv \mathbb{E}[\tilde{\pi}_k^2], \tag{103}$$

equation (101) immediately yields $\mathbb{E}[\nu_k^2] = (\lambda + \gamma\sigma^2)^2(1 + \psi\tilde{e}_k)^2 m_{\tilde{\pi},k}$. The total cross-sectional expected-return variance following (102) and the orthonormality of $\boldsymbol{\Theta}$ is then:

$$\mathbb{V}[\mathbb{E}[D_n] - P_n] = \sum_{k=1}^{K} \nu_k^2 + \|\lambda\bar{\boldsymbol{Q}}^{\perp} + \boldsymbol{\delta}^{\perp}\|^2. \tag{104}$$

The orthogonal residuals $\bar{\boldsymbol{Q}}^{\perp}$ and $\boldsymbol{\delta}^{\perp}$ generate expected-return variance that is purely idiosyncratic — it does not load on any factor — but contributes to total cross-sectional variance.

The proportion of total cross-sectional expected-return variance explained by factor k is therefore

$$f_k = \frac{(1 + \psi\tilde{e}_k)^2 \, m_{\tilde{\pi},k}}{\sum_{j=1}^{K}(1 + \psi\tilde{e}_j)^2 \, m_{\tilde{\pi},j} + \left\|(1 - \psi)\bar{\boldsymbol{Q}}^{\perp} + \zeta^{\text{Own}}\boldsymbol{\delta}^{\perp}\right\|^2}. \tag{105}$$

When $\bar{\boldsymbol{q}} = \boldsymbol{d} = \boldsymbol{0}$ and $\bar{\boldsymbol{Q}}^{\perp} = \boldsymbol{\delta}^{\perp} = \boldsymbol{0}$, $\tilde{\pi}_k = \pi_k$ and $\rho_{\perp} = 0$, so (105) reduces to (9) of the paper exactly. The factor shares f_k in general sum to strictly less than one because the target portfolio and non-pecuniary preference generate a non-zero idiosyncratic component of expected-return variance.

The economic content of target portfolio and non-pecuniary preferences is captured by how they make $m_{\tilde{\pi},k}$ vary across factors. Expanding this second moment gives

$$m_{\tilde{\pi},k} = \mathbb{E}[\pi_k^2] \; - \; \frac{2\,\mu_\pi(\lambda\bar{q}_k + d_k)}{(\lambda + \gamma\sigma^2)(1 + \psi\tilde{e}_k)} \; + \; \frac{(\lambda\bar{q}_k + d_k)^2}{(\lambda + \gamma\sigma^2)^2(1 + \psi\tilde{e}_k)^2}. \tag{106}$$

We now reparameterize in terms of ζ^{Own} and ψ. Substituting $\lambda + \gamma\sigma^2 = 1/\zeta^{\text{Own}}$ and

$\lambda = (1-\psi)/\zeta^{\text{Own}}$ into the intercept term $\lambda\bar{q}_k + d_k$ yields

$$\frac{\lambda\bar{q}_k + d_k}{\lambda + \gamma\sigma^2} = (1-\psi)\,\bar{q}_k + \zeta^{\text{Own}}\,d_k \;\equiv\; h_k(\psi),$$

where the function $h_k(\psi)$ collects the target portfolio and non-pecuniary preference pieces. Equation (106) then simplifies to

$$m_{\tilde{\pi},k}(\psi) = \mathbb{E}[\pi_k^2] \;-\; \frac{2\,\mu_\pi\,h_k(\psi)}{1+\psi\tilde{e}_k} \;+\; \frac{h_k(\psi)^2}{(1+\psi\tilde{e}_k)^2}, \tag{107}$$

which depends on ψ only through $h_k(\psi)$ and the $(1+\psi\tilde{e}_k)$ denominators, and on ζ^{Own} only through $h_k(\psi)$.

Heterogeneity in $\bar{q}_k$ or d_k across factors mechanically generates heterogeneity in $m_{\tilde{\pi},k}$, which generates heterogeneity in the explained variance share f_k beyond what the factor variances alone would predict.

A factor zoo arises in equation (105) when $m_{\tilde{\pi},k}$ is large for weak-variance factors and small for strong-variance factors.

Note that in this general case, we empirically identify in (16):

$$\frac{\sum_{k=1}^{J} f_k}{\sum_{k=1}^{K} f_k} = \frac{\sum_{k=1}^{J}(1+\psi\tilde{e}_k)^2\,m_{\tilde{\pi},k}}{\sum_{k=1}^{K}(1+\psi\tilde{e}_k)^2\,m_{\tilde{\pi},k}}.$$

This ratio is the proportion of total *factor-spanned* expected return variation explained by the first J factors. This ratio is *greater than* the proportion of total expected return variation explained $\sum_{k=1}^{J} f_k$

$$\sum_{k=1}^{J} f_k = \frac{\sum_{k=1}^{J}(1+\psi\tilde{e}_k)^2\,m_{\tilde{\pi},k}}{\sum_{j=1}^{K}(1+\psi\tilde{e}_j)^2\,m_{\tilde{\pi},j} + \left\|(1-\psi)\bar{\boldsymbol{Q}}^{\perp} + \zeta^{\text{Own}}\boldsymbol{\delta}^{\perp}\right\|^2}.$$

$\sum_{k=1}^{J} f_k$ is smaller because the orthogonal components of the target portfolio and non-pecuniary preferences add expected-return variance that is not spanned by the factors.

Maximum attainable Sharpe Ratio The maximum attainable squared Sharpe ratio is that attained by the mean-variance efficient portfolio $\boldsymbol{Q}_* = \boldsymbol{\Sigma}^{-1}(\mathbb{E}[\boldsymbol{D}] - \boldsymbol{P})$. Using the orthogonality $\boldsymbol{\Sigma}\bar{\boldsymbol{Q}}^{\perp} = \sigma^2\bar{\boldsymbol{Q}}^{\perp}$ (since $\boldsymbol{\Theta}\mathbf{E}\boldsymbol{\Theta}'\bar{\boldsymbol{Q}}^{\perp} = \mathbf{0}$), and likewise for $\boldsymbol{\delta}^{\perp}$, the factor and orthogonal pieces decouple. Substituting (101), the factor-space term inherits the same structure as (11):

$$\text{Max SR}^2 = \left(\zeta^{\text{Own}}\right)^{-2} \sum_{k=0}^{K} \frac{(1+\psi\tilde{e}_k)^2\,\tilde{\pi}_k^2}{(\tilde{e}_k+1)\,\sigma^2} \;+\; \frac{\left\|\lambda\bar{\boldsymbol{Q}}^{\perp} + \boldsymbol{\delta}^{\perp}\right\|^2}{\sigma^2}.$$

Setting $\bar{\boldsymbol{Q}} = \boldsymbol{\delta} = \mathbf{0}$ recovers (11) of the paper. The first term collects factor-level contributions written in terms of redefined supply $\tilde{\boldsymbol{\pi}}$; the second term collects idiosyncratic alpha created by orthogonal components of the target portfolio and non-pecuniary preferences.

In expectation over $\boldsymbol{\pi}$,

$$\mathbb{E}[\text{Max SR}^2] = \left(\zeta^{\text{Own}}\right)^{-2} \sum_{k=0}^{K} \frac{(1+\psi\tilde{e}_k)^2\,m_{\tilde{\pi},k}}{(\tilde{e}_k+1)\,\sigma^2} \;+\; \frac{\left\|\lambda\bar{\boldsymbol{Q}}^{\perp} + \boldsymbol{\delta}^{\perp}\right\|^2}{\sigma^2}. \tag{108}$$

Can target portfolios generate a factor zoo without weak substitution? No.

Set $\boldsymbol{d} = \mathbf{0}$ and consider only the target portfolio channel. If substitution is strong, then from (107) we have

$$m_{\tilde{\pi},k} \xrightarrow{\psi\to 1} \mathbb{E}[\pi_k^2].$$

Thus, from (105) we have (using that π_k is i.i.d.)

$$f_k \xrightarrow{\psi\to 1} \frac{(\tilde{e}_k+1)^2}{\sum_j(\tilde{e}_j+1)^2}$$

This expression is the same as the strong substitution limit in the baseline model (10).

Thus, if substitution is strong, then factor-level explained variance shares are tightly linked to factor variances, which precludes a factor zoo: the few factors that explain the most comovement will also explain most expected-return variation. Therefore, if substitution is strong, then a non-zero target portfolio cannot generate a factor zoo.

Intuitively, the strong substitution limit emerges when λ is small (relative to γ); that is, when the entire ℓ_2 penalty is small. If the ℓ_2 penalty is small, then it doesn't matter what the shrinkage target is.

However, conditional on substitution being weak, then a non-zero target portfolio can *exacerbate* the factor zoo. As can be seen from (105), a non-zero target portfolio can lower all f_k (e.g., $\bar{\boldsymbol{Q}}^{\perp} \neq \mathbf{0}$ adds expected-return variation that cannot be spanned by the factors).

Can non-pecuniary preferences generate a factor zoo without weak substitution? Yes, but in doing so they inflate the maximum attainable Sharpe ratio.

Set $\bar{\boldsymbol{q}} = \mathbf{0}$ and consider only the non-pecuniary preference channel. The maximum attainable Sharpe ratio is given by (108) with $\bar{\boldsymbol{q}} = \mathbf{0}$ and substituting (107):

$$\mathbb{E}[\text{Max SR}^2] = \underbrace{\left(\zeta^{\text{Own}}\right)^{-2}\mathbb{E}[\pi_k^2]\sum_{k=0}^{K}\frac{(1+\psi\tilde{e}_k)^2}{(\tilde{e}_k+1)\,\sigma^2}}_{\text{From Baseline Model}} \underbrace{-\,2\mathbb{E}[\pi_k]\left(\zeta^{\text{Own}}\right)^{-1}\sum_{k=0}^{K}\frac{(1+\psi\tilde{e}_k)\,d_k}{(\tilde{e}_k+1)\,\sigma^2}}_{\text{Cross-Term}}$$
$$+ \underbrace{\sum_{k=0}^{K}\frac{d_k^2}{(\tilde{e}_k+1)\,\sigma^2}}_{\text{Non-pecuniary Alpha, Factor Part}} + \underbrace{\frac{\|\boldsymbol{\delta}^{\perp}\|^2}{\sigma^2}}_{\text{Non-pecuniary Alpha, Idiosyncratic Part}}. \tag{109}$$

The first term is the maximum squared Sharpe ratio in the baseline model (11). The cross-term can be either sign depending on $\text{sgn}(\mathbb{E}[\pi_k]d_k)$; the third and fourth terms are always positive.

Before stating the main proposition, it is useful to identify the configuration of $\boldsymbol{d}$ that minimizes $\mathbb{E}[\text{Max SR}^2]$. The first-order condition $\partial_{d_k}\mathbb{E}[\text{Max SR}^2] = 0$ in (109) yields

$$d_k = \alpha_k \equiv \frac{\mathbb{E}[\pi_k](1+\psi\tilde{e}_k)}{\zeta^{\text{Own}}}.$$
$$\delta_n^{\perp} = 0.$$

At this profile, substituting into (107) with $\bar{q}_k = 0$ gives

$$m_{\tilde{\pi},k}\big|_{d_k=\alpha_k} = \mathbb{E}[\pi_k^2] - (\mathbb{E}[\pi_k])^2 = \mathbb{V}[\pi_k], \tag{110}$$

which is constant across k. The associated minimum possible maximum squared Sharpe ratio is

$$S_{\min} \equiv \mathbb{E}[\text{Max SR}^2]\Big|_{\boldsymbol{d}=\boldsymbol{\alpha},\boldsymbol{\delta}^\perp=\mathbf{0}} = \left(\zeta^{\text{Own}}\right)^{-2}\mathbb{V}[\pi_k]\sum_{k=0}^{K}\frac{(1+\psi\tilde{e}_k)^2}{(\tilde{e}_k+1)\,\sigma^2}. \tag{111}$$

Let $F_J(\boldsymbol{\delta}, \boldsymbol{\delta}^\perp) = \sum_{k=1}^{J} f_k$ be the cumulative expected return variance share of the first J factors given the preference profile $(\boldsymbol{\delta}, \boldsymbol{\delta}^\perp)$. Let F_J^0 be the baseline F_J when $\boldsymbol{d} = \mathbf{0}$ and $\boldsymbol{\delta}^\perp = \mathbf{0}$:

$$F_J^0 \equiv F_J(\mathbf{0}, \mathbf{0}) = \frac{\sum_{k=1}^{J}(1+\psi\tilde{e}_k)^2}{\sum_{j=1}^{K}(1+\psi\tilde{e}_j)^2}. \tag{112}$$

It turns out that F_J^0 is the same as $F_J(\boldsymbol{\alpha}, \mathbf{0})$, the cumulative variance share under the Sharpe-minimizing preference profile $\boldsymbol{d} = \boldsymbol{\alpha} \equiv [\alpha_k]_{k=1}^K$:

$$F_J(\boldsymbol{\alpha}, \mathbf{0}) = F_J^0. \tag{113}$$

This is because at $\boldsymbol{d} = \boldsymbol{\alpha}$, the numerator of f_k in (105) is $(1+\psi\tilde{e}_k)^2(\mathbb{E}[\pi_k^2] - (\mathbb{E}[\pi_k])^2)$, which is proportional to the numerator that prevails when $\boldsymbol{d} = \mathbf{0}$: $(1+\psi\tilde{e}_k)^2\mathbb{E}[\pi_k^2]$. Since $\boldsymbol{\delta}^\perp = \mathbf{0}$ in both the baseline and in the Sharpe-minimizing preference profile, the idiosyncratic-alpha term in the denominator of f_k vanishes in both cases. As a result,

the ratio defining F_J is unchanged:

$$\begin{aligned} F_J(\boldsymbol{\alpha}, \mathbf{0}) &= \frac{\sum_{k=1}^{J}(1+\psi\tilde{e}_k)^2(\mathbb{E}[\pi_k^2] - (\mathbb{E}[\pi_k])^2)}{\sum_{j=1}^{K}(1+\psi\tilde{e}_j)^2(\mathbb{E}[\pi_j^2] - (\mathbb{E}[\pi_j])^2)} \\ &= \frac{\sum_{k=1}^{J}(1+\psi\tilde{e}_k)^2}{\sum_{j=1}^{K}(1+\psi\tilde{e}_j)^2} \\ &= \frac{\sum_{k=1}^{J}(1+\psi\tilde{e}_k)^2\mathbb{E}[\pi_k^2]}{\sum_{j=1}^{K}(1+\psi\tilde{e}_j)^2\mathbb{E}[\pi_j^2]} \\ &= F_J^0. \end{aligned}$$

Now let $\tilde{d}_k \equiv d_k - \alpha_k$ denote the deviation from the Sharpe-minimizing preference profile. We have the identity

$$S(\boldsymbol{d}, \boldsymbol{\delta}^{\perp}) - S_{\min} = \sum_{k=0}^{K} \frac{\tilde{d}_k^2}{(\tilde{e}_k + 1)\,\sigma^2} + \frac{\|\boldsymbol{\delta}^{\perp}\|^2}{\sigma^2} \;\geq\; 0, \tag{114}$$

with equality iff $\tilde{\boldsymbol{d}} = \mathbf{0}$ and $\boldsymbol{\delta}^{\perp} = \mathbf{0}$.

The following proposition demonstrates that any non-pecuniary preference profile that strictly flattens f_k (i.e., that spreads out the explanatory power over expected return variance across more factors) raises Sharpe. That is, larger non-pecuniary preferences (i.e., larger $|\boldsymbol{d}|$) raise the number of factors required to explain expected return variation, but in doing so *raise* the maximum attainable Sharpe ratio. Note that this is the opposite of what weak substitution accomplishes: ceteris paribus, weaker substitution raises the number of factors required to explain expected return variation, but in doing so *lowers* the maximum attainable Sharpe ratio.

Proposition 3 (Flattening forces Sharpe inflation)**.** *Take* $\bar{\boldsymbol{q}} = \mathbf{0}$ *and* $\bar{\boldsymbol{Q}}^{\perp} = \mathbf{0}$*, and let* J *satisfy* $e_J > e_{J+1}$*. With* $S_{\min}$ *and* F_J^0 *defined in* (111) *and* (112)*, if a configuration* $(\boldsymbol{d}, \boldsymbol{\delta}^{\perp})$ *strictly flattens* f_k *in the sense*

$$F_J(\boldsymbol{d}, \boldsymbol{\delta}^{\perp}) < F_J^0, \tag{115}$$

then

$$S(\boldsymbol{d}, \boldsymbol{\delta}^{\perp}) - S_{\min} \;>\; 0. \tag{116}$$

Therefore, under strong substitution, non-pecuniary preferences can create a factor zoo, but only if the maximum attainable Sharpe ratio is high. This is exactly the point made in Kozak et al. (2018a) (where they label $\boldsymbol{\delta}$ as sentiment shocks instead of non-pecuniary preferences). In the model in that paper, large $|\boldsymbol{\delta}|$ creates a factor zoo, but also gives rise to near arbitrage opportunities.

Proof of Proposition 3. We work in the variables $\tilde{d}_k = d_k - \alpha_k$. Let V_k be the numerator of f_k in (105):

$$V_k = (1 + \psi\tilde{e}_k)^2 m_{\tilde{\pi},k}.$$

Let $V_k^0 = (1+\psi\tilde{e}_k)^2\mathbb{E}[\pi_k^2]$ be the baseline numerator when $\boldsymbol{d} = \mathbf{0}$ and $\boldsymbol{\delta}^{\perp} = \mathbf{0}$. Substituting $d_k = \alpha_k + \tilde{d}_k$ into the expression for V_k,

$$V_k = V_k^0 - 2\mathbb{E}[\pi_k]\zeta^{\text{Own}}(1 + \psi\tilde{e}_k)d_k + (\zeta^{\text{Own}})^2 d_k^2 = V_k^0 + (\zeta^{\text{Own}})^2\big[\tilde{d}_k^2 - \alpha_k^2\big],$$

where we used $2\mathbb{E}[\pi_k]\zeta^{\text{Own}}(1 + \psi\tilde{e}_k)\alpha_k = 2\mu_\pi^2(1 + \psi\tilde{e}_k)^2 = 2(\zeta^{\text{Own}})^2\alpha_k^2$.

Note that $\alpha_k^2 = (\mathbb{E}[\pi_k])^2(1 + \psi\tilde{e}_k)^2/(\zeta^{\text{Own}})^2$, which is proportional to V_k^0. Hence $\sum_{k=1}^J \alpha_k^2 / \sum_{k=1}^K \alpha_k^2 = \sum_{k=1}^J V_k^0 / \sum_{k=1}^K V_k^0 = w$.

The flattening condition (115) expands to

$$\frac{\sum_{k=1}^J V_k}{\sum_{k=1}^K V_k + \rho_\perp} < w, \tag{117}$$

where $\rho_\perp = (\zeta^{\text{Own}})^2\|\boldsymbol{\delta}^{\perp}\|^2$. Substituting $V_k - V_k^0 = (\zeta^{\text{Own}})^2[\tilde{d}_k^2 - \alpha_k^2]$ and using $\sum_{k=1}^J \alpha_k^2 =$

$w \sum_{k=1}^{K} \alpha_k^2$, (117) simplifies (after cross-multiplication and division by $(\zeta^{\text{Own}})^2$) to

$$w \sum_{k=J+1}^{K} \tilde{d}_k^2 + w\|\boldsymbol{\delta}^{\perp}\|^2 \;>\; (1-w) \sum_{k=1}^{J} \tilde{d}_k^2. \tag{118}$$

For (116), we bound $S - S_{\min}$ from below using (114):

$$\begin{aligned} S - S_{\min} &= \sum_{k=0}^{J} \frac{\tilde{d}_k^2}{(\tilde{e}_k + 1)\,\sigma^2} + \sum_{k=J+1}^{K} \frac{\tilde{d}_k^2}{(\tilde{e}_k + 1)\,\sigma^2} + \frac{\|\boldsymbol{\delta}^{\perp}\|^2}{\sigma^2} \\ &\geq 0 + \frac{1}{(\tilde{e}_{J+1} + 1)\,\sigma^2} \sum_{k=J+1}^{K} \tilde{d}_k^2 + \frac{\|\boldsymbol{\delta}^{\perp}\|^2}{\sigma^2} \\ &> \frac{1}{(\tilde{e}_J + 1)\,\sigma^2} \left[\sum_{k=J+1}^{K} \tilde{d}_k^2 + \|\boldsymbol{\delta}^{\perp}\|^2 \right] \\ &> 0. \end{aligned}$$

The second line follows from dropping the non-negative $\sum_{k=0}^{J} \frac{\tilde{d}_k^2}{(\tilde{e}_k+1)\sigma^2}$ term and using $\tilde{e}_k \leq \tilde{e}_{J+1}$ for $k > J$. The third line follows from $1/((\tilde{e}_{J+1}+1)\,\sigma^2) > 1/((\tilde{e}_J+1)\,\sigma^2)$ (strict since $\tilde{e}_J > \tilde{e}_{J+1}$) and $1/\sigma^2 > 1/((\tilde{e}_J+1)\sigma^2)$. The final strict inequality follows from (118), which gives $\sum_{k=J+1}^{K} \tilde{d}_k^2 + \|\boldsymbol{\delta}^{\perp}\|^2 > \frac{1-w}{w} \sum_{k=1}^{J} \tilde{d}_k^2 \geq 0$.

□

C Characteristic Definitions

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

Table C.1: Stock Characteristics Used in the Returns Analysis

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

Continued on next page

Variable	Description
ncoa_gr1a	Change in non-current operating assets over one year
nncoa_gr1a	Change in net non-current operating assets over one year
noa_gr1a	Change in net operating assets over one year
ppeinv_gr1a	Change in property, plant, equipment, and inventories
sale_gr1	Sales growth over one year
sale_gr3	Sales growth over three years
saleq_gr1	Quarterly sales growth
age	Firm age
at_be	Book leverage
bidaskhl_21d	21-day bid-ask high-low spread
cash_at	Cash scaled by assets
ni_ivol	Earnings volatility
rd_sale	Research and development scaled by sales
rd5_at	Research and development capital scaled by assets
tangibility	Asset tangibility
beta_60m	60-month CAPM beta
beta_dimson_21d	21-day Dimson beta
betabab_1260d	Betting-against-beta
betadown_252d	252-day downside beta
earnings_variability	Earnings variability
ivol_capm_21d	21-day CAPM idiosyncratic volatility
ivol_capm_252d	252-day CAPM idiosyncratic volatility
ivol_ff3_21d	21-day Fama-French three-factor idiosyncratic volatility
ivol_hxz4_21d	21-day HXZ four-factor idiosyncratic volatility
ocfq_saleq_std	Volatility of quarterly operating cash flow to sales
rvol_21d	21-day return volatility
turnover_126d	126-day share turnover

Continued on next page

Variable	Description
zero_trades_21d	21-day zero-trade measure
zero_trades_126d	126-day zero-trade measure
zero_trades_252d	252-day zero-trade measure
dsale_dinv	Change in sales minus change in inventory
dsale_drec	Change in sales minus change in receivables
dsale_dsga	Change in sales minus change in SG&A
niq_at_chg1	Change in quarterly net income scaled by assets
niq_be_chg1	Change in quarterly net income scaled by book equity
niq_su	Quarterly earnings surprise
ocf_at_chg1	Change in operating cash flow scaled by assets
sale_emp_gr1	Sales growth relative to employee growth
saleq_su	Quarterly revenue surprise
tax_gr1a	Change in effective tax rate over one year
dolvol_var_126d	126-day dollar-volume volatility
ebit_bev	EBIT scaled by business enterprise value
ebit_sale	EBIT scaled by sales
f_score	Piotroski F-score
ni_be	Net income scaled by book equity
niq_be	Quarterly net income scaled by book equity
o_score	O-score
ocf_at	Operating cash flow scaled by assets
ope_be	Operating profitability scaled by book equity
ope_bel1	Operating profitability scaled by lagged book equity
turnover_var_126d	126-day turnover volatility
at_turnover	Asset turnover
cop_at	Cash-based operating profitability scaled by assets
cop_atl1	Cash-based operating profitability scaled by lagged assets
dgp_dsale	Change in gross profit minus change in sales

Continued on next page

Variable	Description
gp_at	Gross profits scaled by assets
gp_atl1	Gross profits scaled by lagged assets
ni_inc8q	Number of consecutive earnings increases
niq_at	Quarterly net income scaled by assets
op_at	Operating profits scaled by assets
op_atl1	Operating profits scaled by lagged assets
opex_at	Operating leverage
qmj	Quality minus junk
qmj_growth	Quality minus junk: growth
qmj_prof	Quality minus junk: profitability
qmj_safety	Quality minus junk: safety
sale_bev	Sales scaled by business enterprise value
corr_1260d	Market correlation
coskew_21d	Coskewness
dbnetis_at	Net debt issuance scaled by assets
lti_gr1a	Change in long-term investments over one year
pi_nix	Taxable income scaled by net income
sti_gr1a	Change in short-term investments over one year
ami_126d	Amihud illiquidity measure
iskew_capm_21d	21-day CAPM idiosyncratic skewness
iskew_ff3_21d	21-day Fama-French three-factor idiosyncratic skewness
iskew_hxz4_21d	21-day HXZ four-factor idiosyncratic skewness
rskew_21d	21-day return skewness
chcsho_12m	12-month change in shares outstanding
eqnetis_at	Equity net issuance scaled by assets
netis_at	Net issuance scaled by assets
Other variables (not as a ratio of market equity as in Jensen et al., 2023)	
netdebt	Net debt

Continued on next page

Variable	Description
`rd`	Research and development expense
`at`	Total assets
`be`	Book equity
`debt`	Total debt
`div12m`	Trailing 12-month dividends
`ebitda`	EBITDA
`eqnpo`	Equity net payout
`eqpo`	Net equity payout
`fcf`	Free cash flow
`ival`	Intrinsic value
`ni`	Net income
`ocf`	Operating cash flow
`sale`	Sales

Notes: The first block reports the 106 characteristics taken directly from the Jensen et al. (2023) global factor data. The second block reports the 14 additional characteristics we construct by recovering underlying accounting quantities from Jensen-Kelly-Pedersen variables that are distributed in market-equity-scaled form. These transformations ensure that the variables in the second block enter the factor-model instrumentation as underlying accounting quantities rather than price-scaled ratios. Characteristics are lagged by one month when matched to returns, and each month we apply the rank-to-percentile transform and recenter at zero, following Kelly et al. (2019) and Kozak et al. (2020).

D RFF Factor-Model Construction

This appendix provides additional details for the RFF orthonormal factor model introduced in Section 3.2. Let $z_{n,t} \in \mathbb{R}^1 20$ denote the vector of rank-normalized characteristics for stock n in month t. We first standardize characteristics using moments from the train-

ing sample and then apply a Random Fourier Features map that approximates an RBF kernel:

$$\phi(z) = \sqrt{\frac{2}{M}} \left[\cos(w_1' z + b_1), \sin(w_1' z + b_1), \ldots, \cos(w_{M/2}' z + b_{M/2}), \sin(w_{M/2}' z + b_{M/2})\right],$$

where the draws (w_j, b_j) are fixed across all months. Stacking these rows gives the month-t feature matrix Φ_t.

Because the raw RFF columns are not orthonormal, we right-whiten within each month,

$$U_t = \Phi_t (\Phi_t' \Phi_t)^{-1/2},$$

so $U_t' U_t = I$ exactly. We then estimate a standard IPCA model with U_t. The proposition below shows that with orthonormal characteristics, IPCA just collapses to PCA on characteristic-managed portfolios $x_{t+1} = U_t' r_{t+1}$.

Proposition 4 (Orthonormal IPCA Collapses to PCA)**.** *Suppose returns satisfy*

$$r_{t+1} = U_t C f_{t+1} + \varepsilon_{t+1},$$

where $U_t' U_t = I$ for every t and $C'C = I$. Then the IPCA objective

$$\min_{C, \{f_{t+1}\}} \sum_t \|r_{t+1} - U_t C f_{t+1}\|^2$$

is equivalent to

$$\max_{C: C'C = I} \sum_t \|C' x_{t+1}\|^2 \qquad \text{where } x_{t+1} = U_t' r_{t+1}.$$

Hence C is given by the top eigenvectors of $\sum_t x_{t+1} x_{t+1}'$, so estimation is just PCA on the managed-portfolio time series x_{t+1}. The resulting loadings satisfy $B_t = U_t C$ and are automatically orthonormal because

$$B_t' B_t = C' U_t' U_t C = C'C = I.$$

Proof. For fixed C, the first-order condition for f_{t+1} is

$$f_{t+1} = (C'U_t'U_tC)^{-1}C'U_t'r_{t+1} = C'U_t'r_{t+1} = C'x_{t+1}.$$

Substituting back into the objective gives

$$\sum_t \|r_{t+1}\|^2 - \sum_t \|C'x_{t+1}\|^2,$$

so minimizing squared residuals is equivalent to maximizing $\sum_t \|C'x_{t+1}\|^2$ subject to $C'C = I$. This is the standard PCA variational problem, solved by the leading eigenvectors of $\sum_t x_{t+1}x_{t+1}'$. □

As usual with IPCA, the resulting stock-level loading matrix is $B_t = U_tC$. These loadings remain orthonormal because $C'C = I$. Factor realizations are therefore

$$f_{t+1} = B_t'r_{t+1} = C'U_t'r_{t+1} = C'x_{t+1}.$$

Their covariance matrix Σ_F determines the systematic covariance matrix $B_t\Sigma_F B_t'$. We use the eigendecomposition of this matrix to recover the orthonormal loading directions Θ_t and factor variances $e_{k,t}$ that enter the theoretical and empirical R^2 curves in the main text.

E Supplements for Section 3

E.1 Description of Russell Banding Methodology Starting in 2007

This appendix describes the methodology used to calculate the market cap rank cutoffs for the annual Russell index reconstitutions.

Prior to 2007, Russell reconstitution rules were simple. Firms with market capitalizations on the May rank date that fell between ranks 1 and 1000 were assigned to the

Russell 1000, and those with market caps ranked between 1001 and 3000 were assigned to the Russell 2000.

To reduce turnover, since 2007 FTSE Russell has used a "banding policy" under which there are two separate cutoffs for stocks in the Russell 1000 and 2000 in the previous year, both of which are based on a mechanical function of the firm size distribution in the year. Under this policy:

- Stocks in the Russell 2000 in the previous year are assigned to the Russell 1000 if their rank date market cap ranks fall between 1 and $1000 - c_1$.
- Stocks in the Russell 1000 in the previous year are assigned to the Russell 2000 if their rank date market cap ranks fall between $1000 + c_2$ and 3000.

To calculate c_1 and c_2 Russell first computes the cumulative market cap of the largest 1000 stocks (i.e., those with ranks 1 through 1000). Let $C(N)$ represent the cumulative market cap of the largest N stocks. c_1 is calculated such that $C(1000-c_1) = 0.95 \cdot C(1000)$. c_2 is calculated such that $C(1000+c_2) = 1.05 \cdot C(1000)$. That is, the band of stocks between ranks $1000 - c_1$ and $1000 + c_2$ constitutes a 5% band around the cumulative market cap of the largest 1000 stocks.

Thus, even after the introduction of the banding policy, assignment to the Russell 1000 or 2000 is still based on a mechanical rule. After the introduction of the banding policy, this mechanical rule changes each year with the distribution of firm sizes.

F Supplements for Section 4

F.1 Out-of-Sample Return-Spanning Exercise

This subsection reports the out-of-sample analogue of Figure 2. We estimate the RFF orthonormal factor model on the training sample from January 1980 through December 1999, and then construct both the return-spanning and model-implied curves using only the out-of-sample period from January 2000 through December 2023. As in the main

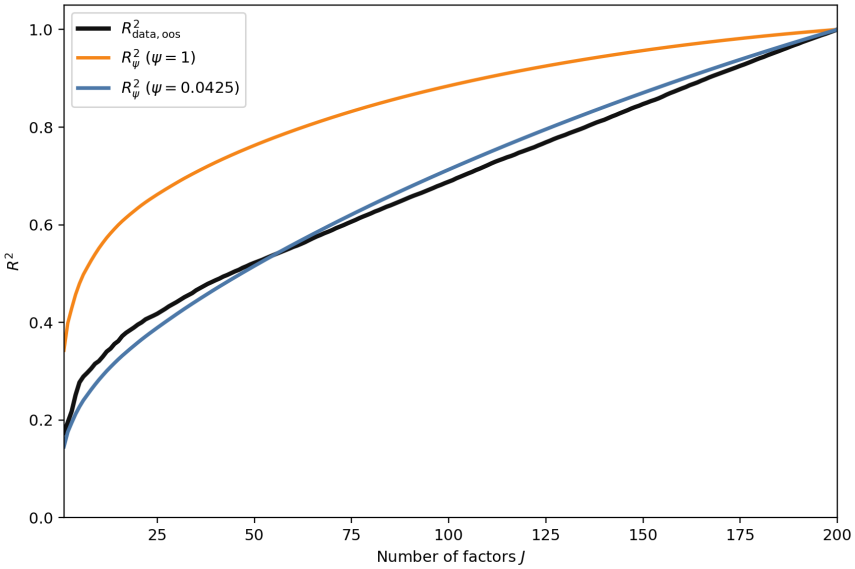


Figure F.1: Out-of-sample RFF orthonormal factor-model evidence on the factor structure of expected returns. The factor model is estimated on January 1980–December 1999, and both the data and model-implied curves use January 2000–December 2023. The black curve is the out-of-sample R^2_{data} path, the orange curve is the strong-substitution benchmark R^2_ψ evaluated at $\psi = 1$, and the blue curve is the out-of-sample best-fitting R^2_ψ path, which corresponds to $\psi \approx 0.0425$.

text, we normalize the resulting R^2 curve at $J = 200$ factors. Figure F.1 shows that the out-of-sample data curve still lies far below the strong-substitution benchmark at $\psi = 1$, while the best-fitting ψ curve remains much flatter than the classical benchmark, with a best-fitting value of approximately $\psi = 0.0425$.

One subtlety is that the normalized out-of-sample curve is somewhat higher than the in-sample one. This does not mean the full return-spanning fit improves out of sample. In raw terms, the regression R^2 is slightly higher out of sample at $J = 1$ (20.2% versus 18.7%), but lower thereafter: for example, it falls from 26.8% to 21.9% at $J = 10$, from 44.3% to 23.7% at $J = 50$, and from 79.5% to 27.6% at $J = 200$. So the out-of-sample exercise should be read as a robustness check on the shape restriction, not as evidence that the absolute level of return-spanning fit is stronger after 2000.

F.2 Daily-Return RFF Exercise

This subsection repeats the expected-return spanning exercise using the daily RFF factor model. We estimate the factor space from daily excess returns, force the first loading to

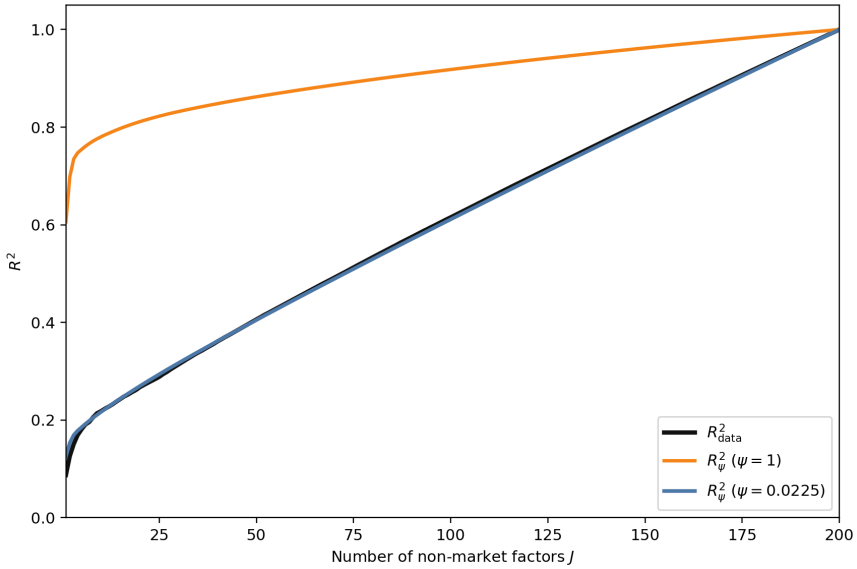


Figure F.2: Full-sample daily RFF factor-model evidence on the factor structure of expected returns. The black curve is the normalized empirical R^2_{data} path, the orange curve is R^2_ψ evaluated at $\psi = 1$, and the blue curve is the best-fitting R^2_ψ path, which corresponds to $\psi \approx 0.0225$.

be equal across assets in each cross section, and then compare the normalized empirical R^2 curve to the model-implied R^2_ψ curves exactly as in the monthly analysis. Figure F.2 shows that the daily data curve also lies far below the strong-substitution benchmark at $\psi = 1$, while the best-fitting value of ψ remains small at approximately 0.0225.

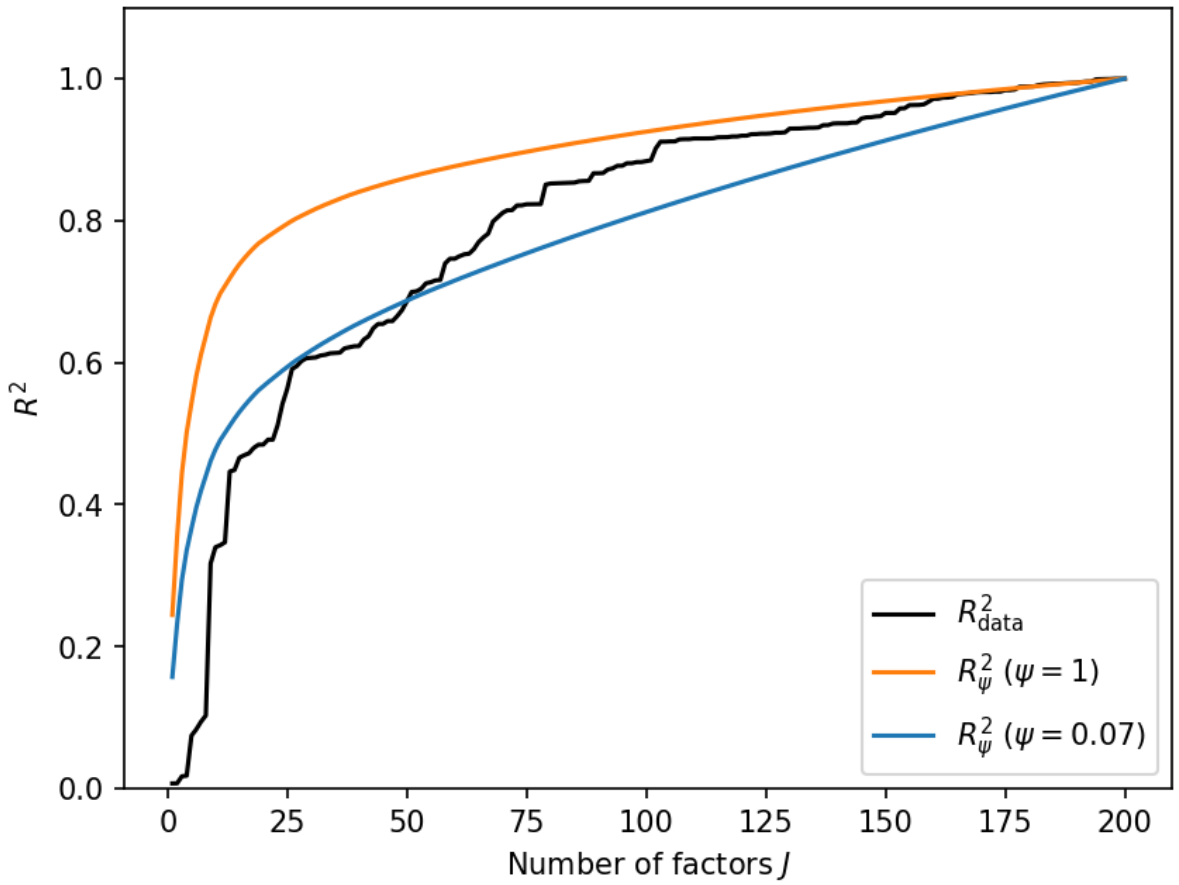


Figure F.3: Standard PCA on daily CRSP stock returns.

G Supplements for Section 5

G.1 Factor-level Sharpe ratios and the fixed-micro counterfactual

Consider the factor-mimicking portfolio associated with factor k. From (8), its expected return is

$$\mu_k \equiv \theta_k'(\mathbb{E}[\mathbf{D}] - \mathbf{P}) = \frac{1}{\zeta^{Own}}\pi_k(1 + \psi\tilde{e}_k), \tag{119}$$

while its variance is

$$\text{Var}(\theta_k'(\mathbf{D} - \mathbf{P})) = (1 + \tilde{e}_k)\sigma^2. \tag{120}$$

Consequently, conditional on the factor-supply shock π_k, the monthly squared Sharpe ratio is

$$\text{Sharpe}^2_{k,\text{monthly}} = \pi_k^2 \left(\frac{1}{\zeta^{Own}}\right)^2 \frac{(1+\psi\tilde{e}_k)^2}{(1+\tilde{e}_k)\sigma^2}. \tag{121}$$

Taking expectations over i.i.d. factor-supply shocks and annualizing squared Sharpe gives (17), with

$$C = 12\,\mathbb{E}[\pi_k^2]\left(\frac{1}{\zeta^{Own}}\right)^2 = 12\,\mathbb{E}[\pi_k^2]\left(M^{\text{micro}}\right)^2. \tag{122}$$

Our counterfactual holds fixed both the second moment of primitive factor-supply shocks and the micro multiplier, and therefore holds C fixed. Because the micro elasticity and micro multiplier are inverses, this is equivalently a fixed-micro-elasticity counterfactual. Under the maintained fixed supply-shock distribution, fixing C in the empirical exercise implements this restriction, so varying ψ changes only the covariance-based amplification of a systematic shock.

In the model, summing the factor contributions gives the annualized maximum attainable squared Sharpe ratio:

$$\text{MaxSharpe}^2(\psi) = C\sum_{k=0}^{K} \frac{(1+\psi\tilde{e}_k)^2}{(1+\tilde{e}_k)\sigma^2}. \tag{123}$$

The exact time-specific expression replaces the common scale C with $12\alpha_{k,t}^2$, where $\alpha_{k,t} \equiv (\lambda+\gamma\sigma^2)\pi_{k,t}$, inside the sum. The counterfactual holds these primitive coefficients fixed rather than holding fixed reduced-form factor premia, which themselves change with ψ.

The summation in (123) is a model result. At a given date, the model's covariance eigenportfolios satisfy

$$\text{Var}(f \mid \Theta, E) = \Theta'\Sigma\Theta = E + \sigma^2 I, \tag{124}$$

so their conditional covariance is diagonal. Across time, however,

$$\mathrm{Cov}(f_t) = \mathbb{E}[\mathrm{Var}(f_t \mid \mathcal{F}_t)] + \mathrm{Var}(\mathbb{E}[f_t \mid \mathcal{F}_t]). \tag{125}$$

The first term is diagonal, but the second need not be. The empirical maximum squared Sharpe ratio is therefore generally $\mu_F' \Sigma_F^{-1} \mu_F$, not the sum of the individual squared Sharpe ratios. We use the factor-level Sharpe ratios only to identify ψ (which is pinned down by their relative magnitudes) and obtain the aggregate Sharpe anchor from actual mean-variance-optimal portfolio returns.

G.2 Estimation

We estimate the RFF factor model through December 1999 and reserve January 2000 through December 2023 for evaluation. For each evaluation month, let $f_{k,t} = \Theta_{t,k}' r_t$ denote the return on the month-specific covariance eigenportfolio and let $e_{k,t}$ denote its eigenvalue. We calculate

$$\widehat{SR}_k^{\mathrm{raw}} = \sqrt{12}\,\frac{\overline{f}_k}{\mathrm{sd}(f_k)}, \qquad e_k = \frac{1}{T}\sum_{t=1}^{T} e_{k,t}, \qquad g_k(\psi) = \frac{(1 + \psi e_k/\sigma^2)^2}{e_k + \sigma^2}, \tag{126}$$

where $\sigma^2 = 0.01$ and both variance objects remain monthly. To remove the unidentified common scale, define the empirical and model-implied shares

$$s_k = \frac{(\widehat{SR}_k^{\mathrm{raw}})^2}{\sum_j (\widehat{SR}_j^{\mathrm{raw}})^2}, \qquad m_k(\psi) = \frac{g_k(\psi)}{\sum_j g_j(\psi)}. \tag{127}$$

The shape estimate is

$$\widehat{\psi} = \operatorname*{argmin}_{\psi \in [0,1]} \sum_{k=1}^{201} [s_k - m_k(\psi)]^2 = 0.02484. \tag{128}$$

We identify the level separately. Let F_t be the returns on the 201 fixed factors from the pre-2000 estimation. Using 1980–1999, we calculate $w = \widehat{\Sigma}_F^{-1} \widehat{\mu}_F$. We then hold w

fixed and evaluate $w'F_t$ over 2000–2023. Its annualized Sharpe ratio is

$$\widehat{SR}^{\text{OOS}} = \sqrt{12}\,\frac{\overline{w'F_t}}{\text{sd}(w'F_t)} = 1.812. \tag{129}$$

Writing $G(\psi) = \sum_k g_k(\psi)$, the implied annualized scale is

$$\widehat{C} = \frac{(\widehat{SR}^{\text{OOS}})^2}{G(\widehat{\psi})} = 0.001361. \tag{130}$$

Finally, we hold $\widehat{C}$ and the post-2000 mean eigenvalues fixed and evaluate

$$\text{MaxSharpe}(\psi; \widehat{C}) = \sqrt{\widehat{C}G(\psi)} = \widehat{SR}^{\text{OOS}}\sqrt{\frac{G(\psi)}{G(\widehat{\psi})}}. \tag{131}$$

Re-estimating C as ψ changes would hold Sharpe fixed by construction. Freezing $\widehat{C}$ instead holds fixed the micro multiplier and primitive factor-supply-shock scale, giving the iso-cost counterfactual in the main text.

Table G.1: Sharpe-ratio estimation results

Calibration	$\widehat{\psi}$	$\widehat{C}$	$SR(\widehat{\psi})$	$SR(0.1)$	$SR(1)$
Fixed-split OOS	0.02484	0.001361	1.812	3.648	26.629
	(0.00702)	(0.000429)	(0.197)	(0.536)	(3.92)

Note: The factor model and fixed factor-portfolio weights are estimated over 1980–1999; factor-level Sharpe ratios and the fixed portfolio are evaluated over 2000–2023. Parentheses report calendar-quarter block bootstrap standard errors. The realized portfolio Sharpe $SR(\widehat{\psi}) = 1.812$ anchors $\widehat{C}$. The last two columns hold $\widehat{C}$ and the post-2000 mean eigenvalues fixed while changing ψ.

G.3 Concentration of Sharpe-ratio contribution

To make the concentration logic explicit, define the model-implied share of total Sharpe ratio squared accruing to factor k at a candidate ψ by

$$q_k(\psi) = \frac{\frac{(1+\psi\tilde{e}_k)^2}{(\tilde{e}_k+1)\sigma^2}}{\sum_{j=0}^{K}\frac{(1+\psi\tilde{e}_j)^2}{(\tilde{e}_j+1)\sigma^2}}. \tag{132}$$

This object strips out the common scale C and isolates how substitution reallocates Sharpe-ratio mass across factors. Small ψ makes the distribution of $q_k(\psi)$ diffuse across many factors; large ψ shifts that mass toward a small number of high-variance factors.

Table G.2: Concentration of model-implied Sharpe-ratio contribution across factors

ψ	Top 10 share	Top 50 share	Factors for 50%	Factors for 90%
0.02484	6.9%	23.5%	109	185
0.01	3.2%	17.5%	121	187
0.1	19.2%	39.7%	74	175
1.0	32.4%	54.8%	40	156

Note: This table uses the post-2000 mean RFF factor eigenvalues and reports concentration statistics for the normalized factor-share object $q_k(\psi)$. "Top 10 share" and "Top 50 share" are the fractions of total model-implied Sharpe ratio squared contributed by the 10 and 50 largest-variance factors. "Factors for 50%" and "Factors for 90%" are the numbers of factors needed to reach 50% and 90% of total Sharpe ratio squared.

H Supplements for Section 6

H.1 Derivation and construction of the macro–micro mapping

From optimal demand (2) and the market-clearing condition in Section 2.3, equilibrium prices satisfy

$$\boldsymbol{P} = \mathbb{E}[\boldsymbol{D}] - \left[\left(\lambda + \gamma\sigma^2\right)\mathbf{I} + \gamma\boldsymbol{\Theta}\mathbf{E}\boldsymbol{\Theta}'\right]\left(\frac{1}{N}\boldsymbol{\iota} + \boldsymbol{\Theta}\boldsymbol{\pi}\right). \tag{133}$$

Define

$$\boldsymbol{Q}^{\text{noise}} \equiv -\boldsymbol{\Theta}\boldsymbol{\pi}.$$

The vector $\boldsymbol{\Theta}\boldsymbol{\pi}$ is the factor-driven residual supply that the representative investor must hold. The minus sign converts this supply shock into an outside-purchase shock: a positive element of $\boldsymbol{Q}^{\text{noise}}$ reduces the shares available to the representative investor and therefore raises prices. Under the same large-N approximation used to derive (8), we

drop the equal-weight component $\boldsymbol{\iota}/N$, so (133) becomes

$$\boldsymbol{P} = \mathbb{E}[\boldsymbol{D}] + \left[\left(\lambda + \gamma\sigma^2\right)\mathbf{I} + \gamma\boldsymbol{\Theta}\mathbf{E}\boldsymbol{\Theta}'\right]\boldsymbol{Q}^{\text{noise}}. \tag{134}$$

Demand shocks are expressed as fractions of shares outstanding. Differentiating (134) with respect to $\boldsymbol{Q}^{\text{noise}}$ gives the multiplier matrix

$$\mathbf{M} \equiv \frac{\partial \boldsymbol{P}}{\partial \boldsymbol{Q}^{\text{noise}'}} = \left(\lambda + \gamma\sigma^2\right)\mathbf{I} + \gamma\boldsymbol{\Theta}\mathbf{E}\boldsymbol{\Theta}' \tag{135}$$

$$= M^{\text{micro}}\left(\mathbf{I} + \psi\widetilde{\boldsymbol{\tau}}\right), \tag{136}$$

where

$$M^{\text{micro}} \equiv \lambda + \gamma\sigma^2, \qquad \psi \equiv \frac{\gamma\sigma^2}{\lambda + \gamma\sigma^2}, \qquad \widetilde{\boldsymbol{\tau}} \equiv \frac{1}{\sigma^2}\boldsymbol{\Theta}\mathbf{E}\boldsymbol{\Theta}' = \boldsymbol{\Theta}\widetilde{\mathbf{E}}\boldsymbol{\Theta}', \tag{137}$$

where $\widetilde{\mathbf{E}} = \mathbf{E}/\sigma^2$ is the diagonal matrix of factor variances normalized by idiosyncratic variance. We use $\sigma^2 = 0.01$, the same rounded idiosyncratic-variance calibration used elsewhere in the paper.

The entries of $\mathbf{M}$ make the distinction between micro, own-price, and cross-price multipliers explicit. For stock n, the full same-stock response is

$$M^{\text{own},n} \equiv M_{n,n} = M^{\text{micro}} + \gamma\sigma^2\widetilde{\tau}_{n,n} = M^{\text{micro}}\left(1 + \psi\widetilde{\tau}_{n,n}\right), \tag{138}$$

whereas, for $m \neq n$, the cross-price multiplier is

$$M^{\text{cross}}_{n,m} \equiv M_{n,m} = \gamma\sigma^2\widetilde{\tau}_{n,m} = M^{\text{micro}}\psi\widetilde{\tau}_{n,m}. \tag{139}$$

The relative-price regressions described in Haddad et al. (2025) hold fixed the systematic component of the price response and therefore identify M^{micro}, not the full diagonal element $M^{\text{own},n}$, under their identifying assumptions. This is why the stock-level estimates used in the calibration map directly to M^{micro}. Related distinctions appear in Fuchs et al.

(2023, 2025). Empirically, the average same-stock systematic moment is 1.0831, so the full own-price and micro multipliers are nevertheless nearly identical in the low-ψ region.

We next derive the multiplier for an arbitrary portfolio and demand shock. Let $\boldsymbol{\nu}$ be a vector of purchases and let $\boldsymbol{w}$ be the weights of the portfolio whose price response is measured. Normalize the response by the portfolio-weighted size of the purchase, $\boldsymbol{w}'\boldsymbol{\nu}$. Equations (136) and (137) imply

$$M(\boldsymbol{w}, \boldsymbol{\nu}) \equiv \frac{\boldsymbol{w}'\mathbf{M}\boldsymbol{\nu}}{\boldsymbol{w}'\boldsymbol{\nu}} \tag{140}$$

$$= M^{\text{micro}} \left[1 + \psi\kappa(\boldsymbol{w}, \boldsymbol{\nu})\right], \tag{141}$$

$$\kappa(\boldsymbol{w}, \boldsymbol{\nu}) \equiv \frac{\boldsymbol{w}'\widetilde{\boldsymbol{\tau}}\boldsymbol{\nu}}{\boldsymbol{w}'\boldsymbol{\nu}}. \tag{142}$$

Thus, $\kappa(\boldsymbol{w}, \boldsymbol{\nu})$ is the systematic covariance exposure of the demand shock per unit of portfolio-weighted shock size, normalized by idiosyncratic variance.

For the macro experiment in month t, the shock purchases the same fraction x of every stock, $\boldsymbol{\nu}_t^{\text{macro}} = x\mathbf{1}$, and the response is measured using market weights $\boldsymbol{w}_t^{\text{macro}}$. Since the market weights sum to one, $\boldsymbol{w}_t^{\text{macro}\prime}\boldsymbol{\nu}_t^{\text{macro}} = x$. Therefore,

$$M_t^{\text{macro}}(\psi) = M^{\text{micro}} \left(1 + \psi\kappa_{M,t}\right), \tag{143}$$

$$\kappa_{M,t} \equiv \boldsymbol{w}_t^{\text{macro}\prime}\boldsymbol{\Theta}_t\widetilde{\mathbf{E}}_t\boldsymbol{\Theta}_t'\mathbf{1}. \tag{144}$$

The shock size x cancels, so a one-percent shock and a unit shock imply the same multiplier. We compute $\kappa_{M,t}$ from the estimated factor loadings, normalized factor variances, and market weights in each of the $T = 528$ sample months. Averaging across months gives

$$\overline{\kappa}_M \equiv \frac{1}{T}\sum_{t=1}^{T}\kappa_{M,t} = 1044.45. \tag{145}$$

Because the micro anchor is held fixed across months, the average model-implied macro

multiplier is consequently

$$\overline{M}^{\text{macro}}(\psi) = M^{\text{micro}}\left(1 + \psi\overline{\kappa}_M\right). \tag{146}$$

Solving this equation for ψ, for any macro target m, yields

$$\psi(m) = \frac{m/M^{\text{micro}} - 1}{\overline{\kappa}_M}. \tag{147}$$

In particular, $M^{\text{micro}} = 2$, $m = 5$, and $\overline{\kappa}_M = 1044.45$ imply $\widehat{\psi} = 0.00144$, which we report as 0.0014 in the text.

The homogeneous calculation in the text is a special case. If every diagonal element of $\mathbf{M}$ equals m_{own} and every off-diagonal element equals m_{cross}, then an equal-fraction purchase of all N stocks raises every stock price by $m_{\text{own}} + (N-1)m_{\text{cross}}$. The estimated factor structure relaxes this symmetry and allows each pairwise spillover to depend on the stocks' factor exposures through (139).

Finally, the average absolute cross-price multiplier reported in Table H.1 and on the top axis of Figure 5 is

$$\bar{M}^{\text{cross}}_{\text{abs}}(\psi; M^{\text{micro}}) \equiv \frac{1}{T}\sum_{t=1}^{T}\frac{1}{N_t(N_t-1)}\sum_{n\neq m}\left|M^{\text{micro}}\psi\widetilde{\tau}_{n,m,t}\right|. \tag{148}$$

It summarizes the magnitude of a typical pairwise price spillover while allowing the spillovers to vary across pairs through their factor exposures. To avoid constructing every element of the large $N_t \times N_t$ matrix, we estimate this average using a fixed random sample of 1,000 off-diagonal pairs in every twelfth month. The same fixed-seed approximation is used for the top axis of Figure 5 and the last two columns of Table H.1.

Table H.1: Implied degree of substitution and average absolute cross-price multipliers under Gabaix and Koijen (2021) macro multiplier estimates

Source	Method	M^{micro}	$\psi(2.2)$	$\psi(6.5)$	$\bar{M}^{\text{cross}}_{\text{abs}}(2.2)$	$\bar{M}^{\text{cross}}_{\text{abs}}(6.5)$
Greenwood and Sammon (2023)	Additions (2010–2020)	0.4	0.00431	0.01460	0.000463	0.001568
Chang et al. (2014)	Russell RD	0.7	0.00205	0.00793	0.000386	0.001491
Greenwood and Sammon (2023)	Deletions (2010–2020)	0.7	0.00205	0.00793	0.000386	0.001491
Schmickler (2020)	Mutual-fund flow surprises	0.8	0.00168	0.00682	0.000360	0.001465
Harris and Gurel (1986)	S&P 500 additions BOE	1.1	0.000957	0.00470	0.000283	0.001388
Pavlova and Sikorskaya (2023)	Benchmarking-intensity shocks	1.5	0.000447	0.00319	0.000180	0.001285
Li and Lin (2025)	OFI idiosyncratic component	1.6	0.000359	0.00293	0.000154	0.001260
Chaudhry and Li (2025)	BMI shock	1.8	0.000213	0.00250	0.000103	0.001208
Chang et al. (2014)	Passive-asset estimate	2.5	-0.000115	0.00153	–	0.001028
Shao et al. (2026)	Tick Size Pilot	2.6	-0.000147	0.00144	–	0.001003
Chaudhry and Li (2025)	OFI shock	3.1	-0.000278	0.00105	–	0.000874
Chaudhry and Li (2025)	FIT shock	3.9	-0.000417	0.000638	–	0.000668

Note: For each row, M^{micro} is the micro, or relative, multiplier from the cited study. Under the mapping in (146), the implied $\psi(m)$ solves $m = M^{\text{micro}}(1 + \psi\overline{\kappa}_M)$, where $m \in \{2.2, 6.5\}$ is the macro target and $\overline{\kappa}_M = 1044.45$. We compute $\overline{\kappa}_M$ from monthly RFF factor loadings and normalized factor variances using a market-cap-weighted portfolio and an equal-fraction marketwide shock, as described in Appendix H.1. The reported $\bar{M}^{\text{cross}}_{\text{abs}}(m)$ is the corresponding average absolute off-diagonal multiplier defined in (148). The Harris and Gurel (1986) value is $3.13\%/2.96\% \approx 1.1$, using their announcement-day price increase and approximate index-fund share. "–" indicates that no nonnegative ψ matches the 2.2 target for that micro anchor.

I Supplements for Section 7

I.1 Measurement Error Simulation Results

	Reliability Ratio	Attenuation Correction Factor
FIT	0.990	1.010
OFI	0.988	1.012
BMI	0.721	1.386

Table I.1: This table displays the reliability ratio $\text{Var}[d^{Cross}_{n,t}/\text{Var}[d^{Cross,(s)}_{n,t}$ and attenuation factor $\text{Var}[d^{Cross}_{n,t}/\text{Var}[d^{Cross,(s)}_{n,t}$ for each of the three demand shocks, calculated as described in Section 7.3.

I.2 Alternate Estimation Approach Based on Nonparametric Sorts

In this section, we estimate the degree of substitution ψ using an alternate estimation approach based on non-parametric sorts.

I.2.1 Empirical Approach

We empirically sort stocks into bins based on covariances, and infer the degree of substitution from the magnitudes of own- and cross-price multipliers in each bin. Specifically, we measure the covariance $\tau_{n,m,t}$ between all pairs of stocks n and m using the factor model from Section 3.2. We then sort all other stocks m into $B = 100$ bins based on $\tau_{n,m,t}$. Thus, the $b = 1$st bin contains the 1% most "similar" stocks to stock n (i.e., the 1% of stocks with the largest covariances with n), while the $b = 100$th bin contains the 1% most "dissimilar" stocks.

We then run the following regression for each bin b:

$$\Delta p_{n,t} = \beta_b^{own} d_{n,t} + \beta_b^{cross} \underbrace{D_{n,t}(b)}_{\equiv \sum_{m \in S_{n,t}(b)} d_{m,t}} + \boldsymbol{c}_b' \boldsymbol{x}_{n,t} + e_{b,n,t}, \tag{149}$$

$\Delta p_{n,t}$ is the percentage change in the price of stock n in quarter t. $d_{n,t}$ is one of the three plausibly exogenous demand shocks discussed in Section 3.3, expressed as a percentage of total shares outstanding. Mapping to the model in Section 2, these demand shocks are the *residual* supply shocks Q_n that are uncorrelated with expected cash flows: they reduce the number of shares held by the representative investor. $S_{n,t}(b)$ is the set of stocks $m \neq n$ such that $\tau_{n,m,t}$ is in one of the top b bins (e.g., $S_{n,t}(10)$ contains the 10% most "similar" stocks to stock n). $D_{n,t}(b)$ is the sum of all shocks $d_{m,t}$ for $m \in S_{n,t}(b)$. $\boldsymbol{x}_{n,t}$ is a vector of stock-level controls (none for FIT and OFI, and the banding controls discussed in Section 3.3 for BMI).

In reduced form, β_b^{own} reflects the own-price multiplier: the impact of demand shocks to stock n on the price of stock n. β_b^{cross} reflects the cross-price multiplier: the impact of demand shocks to other stocks m in the top b bins on the price of stock n. Weaker substitution implies smaller β_b^{cross}.

Yet these reduced-form coefficients are *not* the structural own- and cross-price param-

eters. (23) implies the following structural forms for own- and cross-price multipliers[23]

$$\Delta p_{n,t} = \left(\lambda + \gamma\left(\sigma^2 + \mathbb{E}[\tau_{n,n,t}]\right)\right) d_{n,t} + \gamma W_{n,t} + \epsilon_{n,t} \tag{150}$$
$$W_{n,t} = \sum_{m \neq n} \tau_{n,m,t} d_{m,t}.$$

Thus, we can express β_b^{own} and β_b^{cross} in terms of the underlying structural parameters:

$$\beta_b^{own} = \lambda + \gamma\left(\sigma^2 + \mathbb{E}[\tau_{n,n,t}]\right) + \gamma \underbrace{\kappa_b^{own}}_{\equiv \frac{Cov\left(W_{n,t}, \tilde{d}_{n,t}(b)\right)}{\mathbb{V}\left[\tilde{d}_{b,n,t}(b)\right]}} \tag{151}$$

$$\beta_b^{cross} = \gamma \underbrace{\kappa_b^{cross}}_{\equiv \frac{Cov\left(W_{n,t}, \tilde{D}_{n,t}(b)\right)}{\mathbb{V}\left[\tilde{D}_{n,t}(b)\right]}}, \tag{152}$$

where $\tilde{d}_{n,t}(b)$ and $\tilde{D}_{n,t}(b)$ are the residuals from regressing each of $d_{n,t}$ and $D_{n,t}(b)$ on the other variable and all controls. See Appendix I.2.3 for the derivation.

Intuitively, β_b^{own} reflects λ ($\mathbb{E}[\tau_{n,n,t}]$ is quantitatively small) and β_b^{cross} reflects γ. Yet both reduced-form coefficients also reflect the correlations between the residualized demand shocks and the true substitute term $W_{n,t}$.

The term κ_b^{own} captures the correlation between the own-stock shock $d_{n,t}$ and the shocks to other stocks in $W_{n,t}$. So $\kappa_b^{own} = 0$ if $d_{n,t}$ is uncorrelated across stocks.[24]

The term κ_b^{cross} captures the correlation between the cross-stock shocks $D_{n,t}(b)$ and $W_{n,t}$, which is essentially the average covariance for stocks in the top b bins.[25] Thus, intuitively, β_b^{cross} should decline as b increases and more dissimilar stocks are included in the regression.

Crucially, since we can construct $W_{n,t}$ from all $d_{n,t}$ and $\tau_{n,m,t}$, we can measure the

[23] In particular, this form arises if $Cov\left(\tau_{n,n,t}, \tilde{d}_{n,t}\right) = 0$. A sufficient condition for this covariance to be zero is that the residualized demand shocks are independent of the covariances: $\tilde{d}_{n,t} \perp \tau_{n,n,t}$.

[24] Hence, in this uncorrelated case, $\kappa_b^{own} = \frac{Cov\left(W_{n,t}, \tilde{d}_{n,t}(b)\right)}{\mathbb{V}\left[\tilde{d}_{b,n,t}(b)\right]} = 0 \rightarrow \beta_b^{own} = \lambda + \gamma(\sigma^2 + \mathbb{E}[\tau_{n,n,t}])$

[25] If $\tilde{d}_{n,t}$ is uncorrelated and has the same variance across stocks, then κ_b^{cross} is exactly the average $\tau_{n,m,t}$ for all stocks in $S_{n,t}(b)$: $\kappa_b^{cross} = \frac{Cov\left(W_{n,t}, \tilde{D}_{n,t}(b)\right)}{\mathbb{V}\left[\tilde{D}_{n,t}(b)\right]} = \mathbb{E}\left[\frac{1}{|S_{n,t}(b)|}\sum_{m \in S_{n,t}(b)} \tau_{n,m,t}\right]$. In general, κ_b^{cross} is a linear combination of the $\tau_{n,m,t}$ for all stocks in $S_{n,t}(b)$, with higher weights on stocks m that have larger covariances of $\tilde{d}_{m,t}$ with $\tilde{D}_{n,t}(b)$.

correlation terms $\kappa_{own}(b)$ and $\kappa_{cross}(b)$ empirically. Moreover, classical measurement error in the covariances $\tau_{n,m,t}$, and so in $W_{n,t}$, does not bias the estimates of these terms since it is by definition uncorrelated with the demand shocks $d_{n,t}$.

Given this structure, we can compare our empirical estimates of β_b^{cross} to the theoretical values $\beta_b^{cross,theoretical}(\psi)$ that would prevail under different values of ψ:

$$\begin{aligned}\beta_b^{cross,theoretical}(\psi) &= \gamma(\psi)\kappa_b^{cross} \\ \gamma(\psi) &= \mathbb{E}_b\left[\frac{\beta_b^{own}\psi}{\sigma^2 + \psi\left(\mathbb{E}\left[\tau_{n,n,t}\right] + \kappa_b^{own}\right)}\right].\end{aligned} \tag{153}$$

The second equation follows from (151) and $\mathbb{E}_b$ is the average across all bins b. Intuitively, larger own-price coefficients β_b^{own} imply larger λ values, and hence large γ values for any given ψ. These higher γ values imply higher cross-price coefficients β_b^{cross}.

Hence, we can estimate ψ with generalized method of moments (GMM) and this moment condition[26]:

$$\mathbb{E}_b\left[\beta_b^{cross} - \gamma(\psi)\kappa_b^{cross}\right] = 0.$$

I.2.2 Empirical Results

We find that cross-price multipliers are empirically small, which indicates that substitution is weak and ψ is small.

Figure I.1 displays empirical estimates of the reduced-form cross-price coefficients β_b^{cross} for the three demand shocks. We compare the empirical estimates of β_b^{cross} to the theoretical values that arise under smaller values of ψ (holding the estimated β_b^{own} fixed). For FIT and OFI, the empirical estimates are precise enough to statistically distinguish various small degrees of ψ. For BMI, although the β_b^{cross} estimates are precise enough to

[26] In particular, we solve

$$\min_{\psi} \boldsymbol{g}(\psi)' \mathbf{W} \boldsymbol{g}(\psi),$$

where $\boldsymbol{g}(\psi) = \left[\beta_b^{cross} - \beta_b^{cross,theoretical}(\psi)\right]_{b=1}^{B}$ and $\mathbf{W}$ is the optimal two-step GMM weighting matrix.

reject $\psi = 1$, since BMI has far fewer observations than the other shocks, these estimates are not precise enough to distinguish various small degrees of ψ. The GMM estimation discussed in Section I.2.1 simply finds the value of ψ that minimizes the distance between the empirical β_b^{cross} coefficients (displayed in the solid red line) and their theoretical counterparts (displayed in the dashed lines).

Table I.2 displays the GMM estimates of ψ for the three demand shocks. We find ψ values of 0.005 for FIT, 0.002 for OFI, and 0.046 for BMI (with large standard errors). Thus, across all three shocks ψ is two to three orders of magnitude smaller than the benchmark $\psi = 1$, which indicates that investors weakly substitute across stocks.

	FIT	OFI	BMI
ψ	0.00542***	0.00209*	0.04631
	(0.00166)	(0.00117)	(0.18876)

Table I.2: This table displays the GMM estimates of ψ implied by the reduced-form own- and cross-price coefficient estimates β_b^{own} and β_b^{cross} for the three demand shocks. The second row displays quarterly block-bootstrapped standard errors over 500 simulations. Levels of significance are presented as follows: *p<0.1; **p<0.05; ***p<0.01.

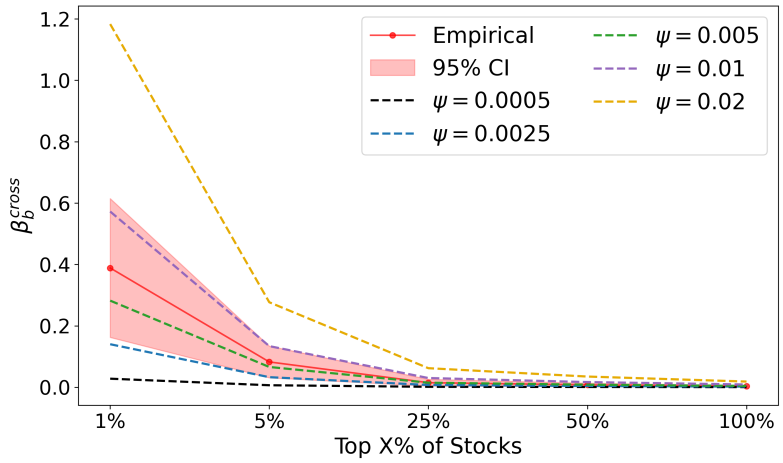
Therefore, substitution is empirically far weaker than implied by the classical benchmark.

I.2.3 Proofs

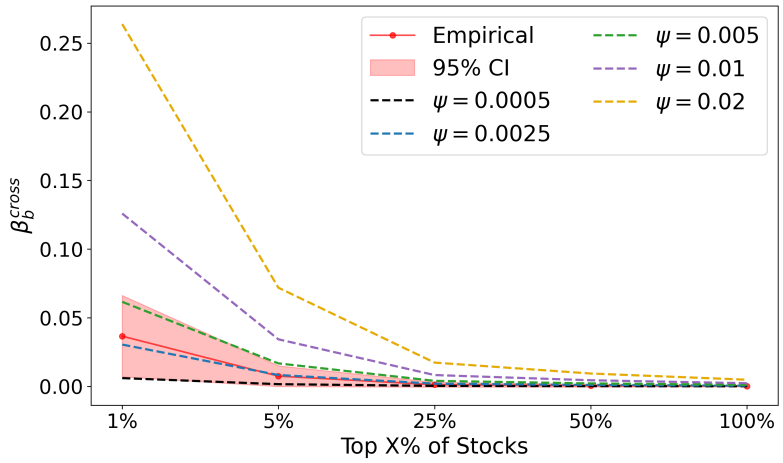
Let $\tilde{d}_{n,t}(b)$ and $\tilde{D}_{n,t}(b)$ be the residuals from the regressions of each of $d_{n,t}$ and $D_{n,t}(b)$ on the other variable and all controls.

Given the reduced-form equation (149),

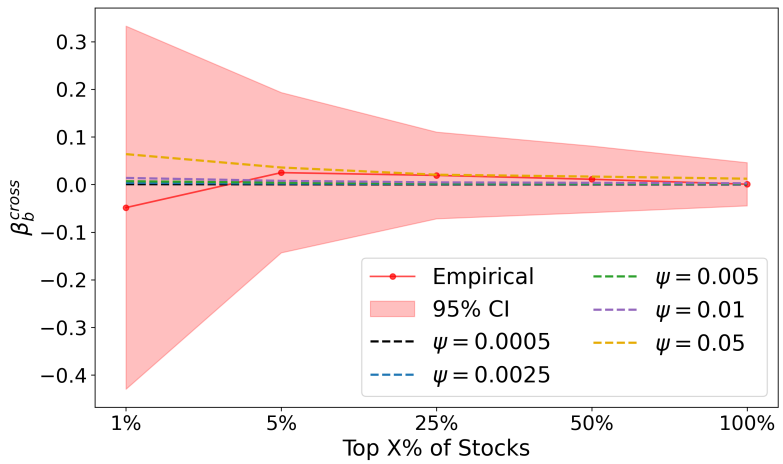
$$\Delta p_{n,t} = \beta_b^{own} d_{n,t} + \beta_b^{cross} \underbrace{D_{n,t}(b)}_{\equiv \sum_{m \in S_{n,t}(b)} d_{m,t}} + \boldsymbol{c}_b' \boldsymbol{x}_{n,t} + e_{b,n,t},$$



(a) FIT

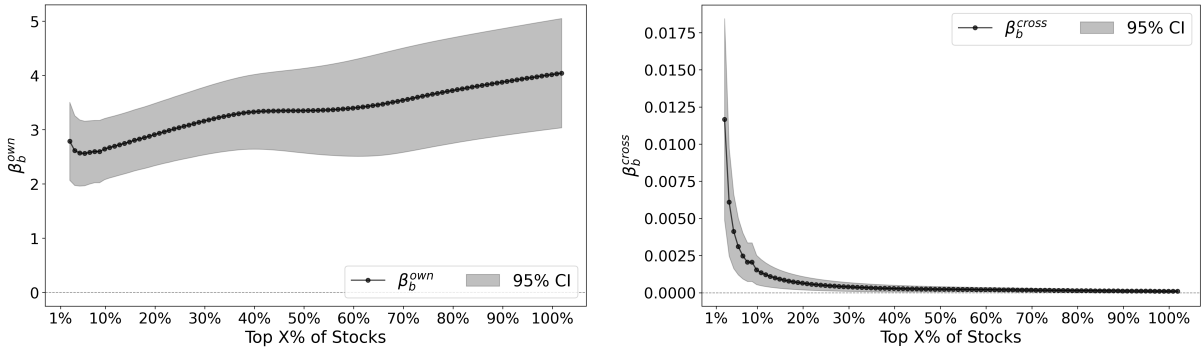


(b) OFI

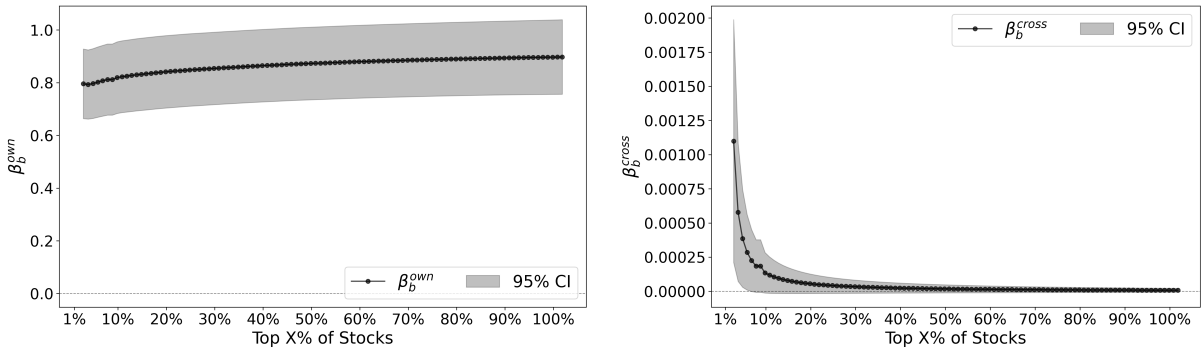


(c) BMI

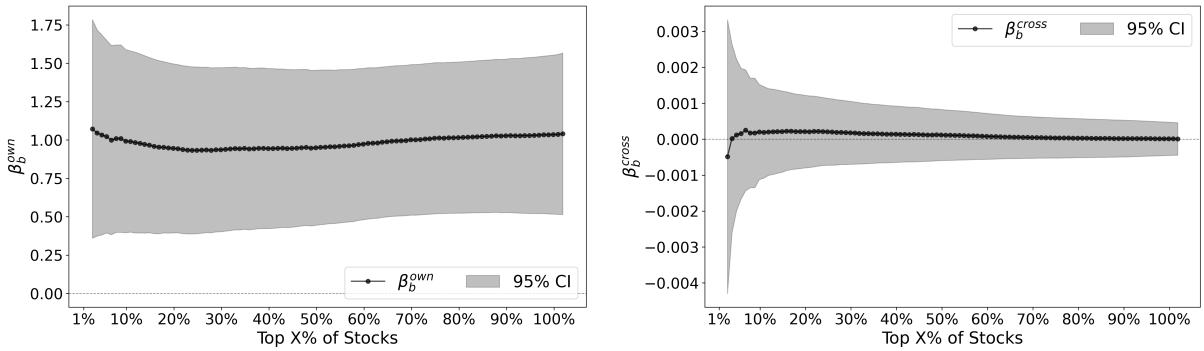
Figure I.1: This figure displays the reduced-form cross-price coefficients $\beta_{cross,b}$ from regression (149) for selected bins b. The three rows display the results for the FIT, OFI, and BMI demand shocks, in order. The left column compares the empirical $\beta_{cross,b}$ estimates to their theoretical counterparts under the benchmark $\psi = 1$. The right column compares the empirical $\beta_{cross,b}$ estimates to the theoretical counterparts under smaller values of ψ. Shaded regions indicate 95% confidence intervals.



(a) FIT



(b) OFI



(c) BMI

Figure I.2: This figure displays the reduced-form own- (left) and cross-price (right) coefficients β_b^{own} and β_b^{cross} from regression (149) for all bins b. The three rows display the results for the FIT, OFI, and BMI demand shocks, in order. Shaded regions indicate 95% confidence intervals.

we have by the Frisch-Waugh-Lovell theorem that

$$\begin{aligned}
\beta_{own,b} &= \frac{Cov\left(\Delta p_{n,t}, \tilde{d}_{n,t}(b)\right)}{\mathbb{V}\left[d^{\perp}_{n,t}(b)\right]} \\
&= \left(\lambda + \gamma\mathbb{E}\left[\tau_{n,n,t}\right] + \gamma\sigma^2\right)\frac{Cov\left(d_{n,t}, \tilde{d}_{n,t}(b)\right)}{\mathbb{V}\left[\tilde{d}_{n,t}(b)\right]} + \gamma\frac{Cov\left(W_{n,t}, \tilde{d}_{n,t}(b)\right)}{\mathbb{V}\left[\tilde{d}_{n,t}(b)\right]} \qquad (154)\\
&= \lambda + \gamma\mathbb{E}\left[\tau_{n,n,t}\right] + \gamma\sigma^2 + \underbrace{\gamma\frac{Cov\left(W_{n,t}, \tilde{d}_{n,t}(b)\right)}{\mathbb{V}\left[\tilde{d}_{n,t}(b)\right]}}_{\equiv\kappa_b^{own}} \qquad (155)
\end{aligned}$$

$$\begin{aligned}
\beta_{cross,b} &= \frac{Cov\left(\Delta p_{n,t}, \tilde{D}_{n,t}(b)\right)}{\mathbb{V}\left[\tilde{D}_{n,t}(b)\right]} \\
&= \lambda\frac{Cov\left(d_{n,t}, \tilde{D}_{n,t}(b)\right)}{\mathbb{V}\left[\tilde{D}_{n,t}(b)\right]} + \gamma\frac{Cov\left(W_{n,t}, \tilde{D}_{n,t}(b)\right)}{\mathbb{V}\left[\tilde{D}_{n,t}(b)\right]} \qquad (156)\\
&= \underbrace{\gamma\frac{Cov\left(W_{n,t}, \tilde{D}_{n,t}(b)\right)}{\mathbb{V}\left[\tilde{D}_{n,t}(b)\right]}}_{\equiv\kappa_b^{cross}} \qquad (157)
\end{aligned}$$

(154) and (156) follow from substituting in the structural equation (150)

$$\begin{aligned}
\Delta p_{n,t} &= \left(\lambda + \gamma\mathbb{E}[\tau_{n,n,t}] + \gamma\sigma^2\right)d_{n,t} + \gamma W_{n,t} + \epsilon_{n,t} \\
W_{n,t} &= \sum_{m\neq n}\tau_{n,m,t}d_{m,t}.
\end{aligned}$$

(155) follows since $Cov\left(d_{n,t}, \tilde{d}_{n,t}(b)\right) = \mathbb{V}\left[\tilde{d}_{n,t}(b)\right]$. (157) follows since $Cov\left(d_{n,t}, \tilde{D}_{n,t}(b)\right) = 0$.

Note that if $Cov\left(d_{n,t}, d_{m,t}\right) = 0$ for all $m \neq n$, then

$$Cov\left(W_{n,t}, \tilde{d}_{n,t}(b)\right) = \sum_{m\neq n}\tau_{n,m,t}Cov\left(d_{m,t}, \tilde{d}_{n,t}(b)\right) = 0,$$

and so $\kappa_b^{own} = 0$. In this case, we also have:

$$Cov\left(W_{n,t}, \tilde{D}_{n,t}(b)\right) = \sum_{m \in S_{n,t}(b)} \tau_{n,m,t} \mathbb{V}[d_{m,t}]$$
$$\mathbb{V}\left[\tilde{D}_{n,t}(b)\right] = \sum_{m \in S_{n,t}(b)} \mathbb{V}[d_{m,t}].$$

If all stocks have the same demand shock variance

$$\mathbb{V}[d_{m,t}] = \sigma_d^2,$$

then we have:

$$\kappa_b^{cross} = \mathbb{E}\left[\frac{\sum_{m \neq n} \tau_{n,m,t}}{|S_{n,t}(b)|}\right].$$

I.3 Implications for Biases in Estimated Multipliers

Previous work argues that failing to appropriately control for substitution can bias own-price multiplier estimates (Fuchs et al., 2023, 2025; Haddad et al., 2025). However, the weak degree of substitution indicated by the small ψ estimates in Table 2 suggests that this bias is likely small. The bias in own-price multiplier estimates is controlled by the $\gamma\kappa_b^{own}$ term in (151), which shrinks as ψ decreases since smaller ψ implies smaller γ relative to λ.

We quantify the magnitude of this bias given the ψ estimates in Table I.2. We regress price changes $\Delta p_{n,t}$ on the three demand shocks $d_{n,t}$ without controlling for any substitute terms

$$\Delta p_{n,t} = \beta^{own} d_{n,t} + \boldsymbol{c}' \boldsymbol{x}_{n,t} + e_{n,t}. \tag{158}$$

Following (151) from Appendix I.2, we have

$$\beta^{own} = \lambda + \gamma\left(\sigma^2 + \mathbb{E}[\tau_{n,n,t}]\right) + \gamma \underbrace{\kappa^{own}}_{\equiv \frac{Cov(W_{n,t}, \tilde{d}_{n,t}(b))}{\mathbb{V}\left[\tilde{d}_{b,n,t}(b)\right]}}$$

$$W_{n,t} = \sum_{m \neq n} \tau_{n,m,t} d_{m,t},$$

where $\tilde{d}_{n,t}$ is the demand shock residualized with respect to all characteristic controls $\boldsymbol{x}_{n,t}$ discussed in Section 3. Given the ψ estimates, we then compute the implied $\gamma(\psi)$ as in (153). We can then compare the estimated β^{own} to the de-biased own-price and micro multipliers

$$\text{Own-Price Multiplier} = \lambda + \gamma(\psi)\left(\sigma^2 + \mathbb{E}\left[\tau_{n,n,t}\right]\right) = \beta^{own} - \gamma(\psi)\kappa_b^{own} \tag{159}$$

$$\text{Micro Multiplier} = \lambda + \gamma\sigma^2 = \beta^{own} - \gamma(\psi)\left(\mathbb{E}\left[\tau_{n,n,t}\right] + \kappa_b^{own}\right). \tag{160}$$

Appendix Table I.3 demonstrates that even after removing the bias from failing to control for substitution, own-price multipliers remain large, ranging from 0.9 to 4.1. Since the $\gamma(\psi)\mathbb{E}\left[\tau_{n,n,t}\right]$ term is small given the small ψ estimates, full own-price multipliers and micro multipliers are virtually identical. The bias in β^{own} varies across demand shocks based on how strongly correlated $d_{n,t}$ is across stocks, which controls κ^{own}.

We also run the regression suggested by Haddad et al. (2025) and often used in previous work (e.g., Chaudhry and Li, 2025) that uses time fixed effects and characteristics interacted with time fixed effects to control for substitution:

$$\Delta p_{n,t} = \beta_{HHHKL}^{own} d_{n,t} + \boldsymbol{c}_l^{'} \boldsymbol{x}_{n,t} + FE_t + e_{n,t}, \tag{161}$$

using the following characteristics: accruals, asset growth, beta, gross profitability, 1 year issuance, 5 year issuance, and net operational assets. Table I.3 shows that the β_{HHHKL}^{own} estimates are close to the de-biased own-price and micro multipliers.

Thus, own-price multipliers are virtually identical to the micro multipliers, both pa-

	OFI	FIT	BMI
β^{own}	1.002***	6.139***	1.064***
	(0.109)	(0.993)	(0.394)
β^{own}_{HHHKL}	0.871***	3.563***	0.857***
	(0.073)	(0.343)	(0.243)
Debiased Full Own-Price Multiplier	0.915***	4.091***	1.046***
	(0.100)	(0.745)	(0.387)
Debiased Micro Multiplier	0.912***	4.070***	1.033***
	(0.100)	(0.742)	(0.382)

Table I.3: This table displays in the first two rows the estimated β^{own} and β^{own}_{HHHKL} from (158) and (161) for the three demand shocks. The last two rows display the de-biased full own-price and micro multipliers using the $\gamma(\psi)$ estimates from (159) and (160) given the ψ estimates in Table I.2. Quarterly block-bootstrapped standard errors over 500 simulations are displayed in parentheses. Levels of significance are presented as follows: *p<0.1; **p<0.05; ***p<0.01.

rameters remain large even after controlling for substitution, and methods from previous work to control for substitution yield nearly unbiased estimates of these parameters.[27]

I.4 Additional Empirical Results

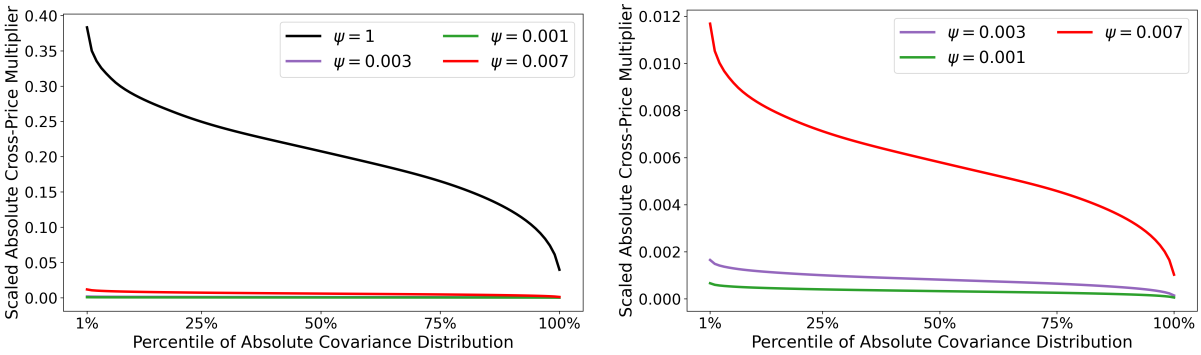


Figure I.3: This figure displays the model-implied ratios of absolute structural cross-price multipliers to micro own-price multipliers under different values of ψ. The right panel compares these ratios under the empirical estimates of ψ for the FIT, OFI, and BMI demand shocks from Table 2 while the left panel overlays the ratios that would emerge under the benchmark $\psi = 1$.

[27] Our results are consistent with those in Chaudhary et al. (2023) Appendix J, which shows that grouping stocks on Fama-French three-factor characteristics or on 27 GICS industries and controlling for group-quarter fixed effects has little impact on the estimated own-price multipliers.

J Supplements for Section 8

J.1 GMM Estimation Details

We start with demand curve (27)

$$\log Q_{i,n,t} = \zeta^{own}\left(-\log P_{n,t} + W_{i,n,t}\left(\psi\right)\right) + \sum_{k=1}^{K} \beta_{k,t}\theta_{n,k,t} + \epsilon_{i,n,t}$$

$$W_{i,n,t}\left(\psi\right) = \sum_{m\in\mathcal{I}_{i,t}} \sum_{k=0}^{K} \underbrace{b_k\left(\psi\right)}_{\equiv\frac{\psi\tilde{e}_k}{1+\psi\tilde{e}_k}} \theta_{n,k,t}\theta_{m,k,t} \log P_{m,t}.$$

and the instruments

$$Z_{i,n,t} = \log\left(\sum_{j\neq i} A_{j,t-1}\frac{1_{j,t}(n)}{1+\sum_{m=1}^{N} 1_{j,t}(m)}\right)$$

$$Z^{cross}_{i,n,t} = \sum_{m\in\mathcal{I}_{i,t}} \sum_{k=0}^{K} \theta_{n,k,t}\theta_{m,k,t} Z_{i,m,t}.$$

For any candidate ψ, we first residualize $\log Q_{i,n,t}$, $Z_{i,n,t}$, $X_{i,n,t}(\psi) = -\log P_{n,t} + W_{i,n,t}\left(\psi\right)$ and $Z^{cross}_{i,n,t}$ with respect to the factor loadings $\boldsymbol{\theta}_{n,t}$ within each quarter to obtain $\tilde{y}_{i,n,t}$, $\tilde{Z}_{i,n,t}$, $\tilde{X}_{i,n,t}(\psi)$ and $\tilde{Z}^{cross}_{i,n,t}$, respectively.

We then compute the following two sample moment conditions in each quarter t:

$$g_{1,t}\left(\zeta^{own},\psi\right) = \frac{1}{\sum_i |\mathcal{I}_{i,t}|}\sum_i \sum_{n\in\mathcal{I}_{i,t}} \left(\tilde{y}_{i,n,t} - \zeta^{own}\tilde{X}_{i,n,t}\left(\psi\right)\right)\tilde{Z}_{i,n,t}$$

$$g_{2,t}\left(\zeta^{own},\psi\right) = \frac{1}{\sum_i |\mathcal{I}_{i,t}|}\sum_i \sum_{n\in\mathcal{I}_{i,t}} \left(\tilde{y}_{i,n,t} - \zeta^{own}\tilde{X}_{i,n,t}\left(\psi\right)\right)\tilde{Z}^{cross}_{i,n,t}\left(\psi\right).$$

Stacking these moment conditions in quarter t

$$\mathbf{g}\left(\zeta^{own},\psi\right) = \left[g_1\left(\zeta^{own},\psi\right), g_2\left(\zeta^{own},\psi\right)\right]$$

we find (ζ_t^{Own}, ψ_t) in quarter t that minimize the following GMM objective function:

$$\min_{\zeta^{own},\psi} \mathbf{g}\left(\zeta^{own}, \psi\right)' \mathbf{g}\left(\zeta^{own}, \psi\right).$$

We then report the average (ζ_t^{Own}, ψ_t) across all quarters t as the (ζ^{Own}, ψ) point estimate.

J.2 Measurement Error Simulation Results

	ψ
Full-sample mean	0.001344
Bootstrap mean	0.001350
Measurement-error bias	$6.0e-6$

Table J.1: This table displays the estimated attenuation bias due to measurement error arising from estimation noise in the estimated factor loadings $\boldsymbol{\theta}_{n,t}$ and variances e_k, calculated as described in Section 8.3.

J.3 Alternate Estimation Approach Based on Nonparametric Sorts

In this section, we estimate the degree of substitution ψ using an alternate estimation approach based on non-parametric sorts.

J.3.1 Empirical Approach

From (2) in Section 2, we can write the quantity demanded for stock n as (26):

$$Q_n = \underbrace{\frac{1}{\lambda+\gamma\sigma^2}}_{\equiv\zeta^{Own}}\left(\mathbb{E}[D_n]-P_n\right) - \sum_{m=1}^{N} \underbrace{\frac{1}{\lambda+\gamma\sigma^2}\left(\sum_{k=0}^{K}\theta_{n,k}\theta_{m,k}\frac{\psi\tilde{e}_k}{1+\psi\tilde{e}_k}\right)}_{\equiv\zeta_{n,m}^{Cross}}\left(\mathbb{E}[D_m]-P_m\right).$$

Given exogenous price variation that is unrelated to the investor's expected cash flows (i.e., price variation driven by the exogenous supply shocks), we can estimate the own- and cross-price elasticities via a regression of the quantity demanded for stock n on the

instrumented prices for all stocks. Then from (6), we can infer the degree of substitution ψ from the relative magnitudes of own- and cross-price elasticities:

$$\frac{\zeta_{n,m}^{Cross}}{\zeta_{n}^{Own}} = \sum_{k=0}^{K} \theta_{n,k}\theta_{m,k}\frac{\psi\tilde{e}_k}{1+\psi\tilde{e}_k}$$

Thus, we empirically sort stocks into bins based on similarity of factor loadings, and infer the degree of substitution from the magnitudes of own- and cross-price elasticities in each bin. Specifically, we measure each stock n's factor loadings using the factor model in Section 3.2. For each investor i in each quarter t we then sort all other stocks *in her investment universe* into $B = 100$ bins based on factor loading similarity: $\delta_{n,m,t} \equiv \boldsymbol{\theta}_{n,t}^{'}\boldsymbol{\theta}_{m,t}$.[28]

Thus, the $b = 1$st bin contains the 1% most similar stocks to stock n (i.e., the 1% of stocks with the largest similarities with n), while the $b = 100$th bin contains the 1% most dissimilar stocks.[29] Following Koijen and Yogo (2019), an investor's universe is all stocks held in this quarter or in the past 11 quarters.

In each quarter t, we then run the following two-stage least squares regression for each bin b pooled across all investors and stocks:

$$\log Q_{i,n,t} = \alpha_{b,t}^{own}\log P_{n,t} + \alpha_{b,t}^{cross}\underbrace{\sum_{m\in S_{i,n,t}(b)}\log P_{m,t}}_{\equiv P_{i,n,t}^{cross}(b)} + \boldsymbol{c}_{b,t}^{'}\boldsymbol{\theta}_{n,t} + e_{b,i,n,t} \tag{162}$$

$$\begin{bmatrix}\log P_{n,t}\\ \log P_{i,n,t}^{cross}\end{bmatrix} = \begin{bmatrix}\pi_{1,b,t}^{own} & \pi_{1,b,t}^{cross}\\ \pi_{2,b,t}^{own} & \pi_{2,b,t}^{cross}\end{bmatrix}\begin{bmatrix}Z_{i,n,t}\\ \sum_{m\in S_{i,n,t}(b)} Z_{i,m,t}\end{bmatrix} + \begin{bmatrix}\boldsymbol{b}_{1,b,t}^{'}\\ \boldsymbol{b}_{2,b,t}^{'}\end{bmatrix}\boldsymbol{\theta}_{n,t} + \begin{bmatrix}u_{b,i,n,t}\\ v_{b,i,n,t}\end{bmatrix} \tag{163}$$

In (162), $\log Q_{i,n,t}$ is the log quantity of shares held by investor i in stock n in quarter

[28] In particular, we re-orthonormalize the loading directions within each quarter after restricting to the stocks that appear in the holdings sample. We compute the covariance matrix of the latent factors across all time periods Σ_F, and then in each quarter take the eigendecomposition of the systematic covariance matrix $\boldsymbol{\Theta}_t\mathbf{E}\boldsymbol{\Theta}_t^{'} = \check{\boldsymbol{\Theta}}_t\check{\mathbf{E}}_t\check{\boldsymbol{\Theta}}_t^{'}$. We use $\check{\boldsymbol{\theta}}_{n,t}$ as the factor loadings for stock n in quarter t and the diagonal elements $\check{e}_{k,t}$ from $\check{\mathbf{E}}_t$ as the factor variances. This step is needed because orthonormality is preserved in the full return panel, but not automatically after dropping stocks that are absent from holdings data.

[29] If an investor has $N < B$ stocks in her investment universe in this quarter, we assign one stock to bins $b = 1, 2, \ldots, N$, and no stocks to the remaining bins.

t. $\log P_{n,t}$ is the log market equity. We normalize all stocks to have one share outstanding, so price equals market equity. $S_{i,n,t}(b)$ is the set of stocks $m \neq n$ in investor i's universe such that $\delta_{n,m,t}$ is in one of the top b bins (e.g., $S_{i,n,t}(10)$ contains the 10% most similar stocks to stock n).

Importantly, the log prices in (162) are not exogenous to the residual demand shocks $e_{i,n,t}$. Through market clearing, increases in this investor's demand for stock n raise the price of stock n. Thus, we instrument for the log prices in the first stage regression (163) using the investment mandate instrument from Koijen and Yogo (2019). Specifically, the instrument is

$$Z_{i,n,t} = \log\left(\sum_{j \neq i} A_{j,t-1} \frac{1_{j,t}(n)}{1 + \sum_{m=1}^{N} 1_{j,t}(m)}\right),$$

where $1_{j,t}(n)$ is an indicator for if stock n falls into the universe of investor j and $A_{j,t-1}$ is the lagged assets under management of investor j. This instrument represents the counterfactual market equity of stock n if all investors held an equal-weighted portfolio of the stocks in their investment universe. This instrument exploits cross-sectional variation in which investors' investment universes stock n falls into. Investment universe membership is plausibly exogenous to the residual demand shocks in this quarter because universes are defined by investment mandates: predetermined rules that do not respond to current demand shocks. Thus, if stock n exogenously falls into the investment universe of more or larger investors, it will face greater demand and will have greater market equity.

Note that although we use the instrument from Koijen and Yogo (2019), we do *not* estimate a logit demand system. In contrast, our empirical approach allows for potentially rich substitution patterns, leaving the actual degree of substitution to be determined by the data.

In reduced form, $\alpha_{b,t}^{own}$ reflects the own-price elasticity: the impact of exogenous price changes to stock n on this investor's holdings of stock n. $\alpha_{b,t}^{cross}$ reflects the cross-price elasticity: the impact of exogenous price changes to other stocks m in the top b bins on this investor's holdings of stock n. Weaker substitution implies smaller $\alpha_{b,t}^{cross}$.

However, these reduced-form coefficients are *not* the structural own- and cross-price elasticities. (26) implies the following structural forms for own- and cross-price elasticities:

$$\log Q_{i,n,t} = -\frac{1}{\lambda+\gamma\sigma^2}\left(1-\mathbb{E}_t\left[\sum_{k=0}^{K}\underbrace{b_k}_{\equiv\frac{\psi\tilde{e}_k}{1+\psi\tilde{e}_k}}\theta_{n,k,t}^2\right]\right)\log P_{n,t}+\frac{1}{\lambda+\gamma\sigma^2}W_{i,n,t}(\psi)+\epsilon_{i,n,t}, \tag{164}$$

$$W_{i,n,t}(\psi) = \sum_{m\neq n\in\mathcal{I}_{i,t}}\sum_{k=0}^{K} b_k\theta_{n,k,t}\theta_{m,k,t}\log P_{m,t}.$$

where $\mathcal{I}_{i,t}$ is investor i's investment universe in quarter t. Thus, we can express $\alpha_{b,t}^{own}$ and $\alpha_{b,t}^{cross}$ in terms of the underlying structural parameters[30]:

$$\alpha_{b,t}^{own} = -\frac{1}{\lambda+\gamma\sigma^2}\left(1-\mathbb{E}_t\left[\sum_{k=0}^{K} b_k\theta_{n,k,t}^2\right]\right)+\frac{1}{\lambda+\gamma\sigma^2}\underbrace{\omega_{b,t}^{own}}_{\equiv\frac{Cov_t\left(W_{i,n,t}(\psi),\log\tilde{P}_{i,n,t}(b)\right)}{\mathbb{V}\left[\log\tilde{P}_{i,n,t}(b)\right]}} \tag{165}$$

$$\alpha_{b,t}^{cross} = \frac{1}{\lambda+\gamma\sigma^2}\underbrace{\omega_{b,t}^{cross}}_{\equiv\frac{Cov_t\left(W_{i,n,t}(\psi),\log\tilde{P}_{i,n,t}^{cross}(b)\right)}{\mathbb{V}\left[\log\tilde{P}_{i,n,t}^{cross}(b)\right]}}, \tag{166}$$

where $\log\tilde{P}_{i,n,t}(b)$ and $\log\tilde{P}_{i,n,t}^{cross}(b)$ are the residuals from regressing each of the *instrumented* $\log P_{n,t}$ and $\log P_{i,n,t}^{cross}$ on the other variable and all controls. See Appendix J.3.3 for the derivation.

Intuitively, $\alpha_{b,t}^{own}$ reflects the micro elasticity $1/(\lambda+\gamma\sigma^2)$ and $\omega_{b,t}^{cross}$ reflects the cross-price elasticities $1/(\lambda+\gamma\sigma^2)\sum_{k=0}^{K} b_k\theta_{n,k,t}\theta_{m,k,t}$. Yet both reduced-form coefficients also reflect the correlations between the residualized instrumented prices and the true substitute term $W_{i,n,t}(\psi)$.

The term $\omega_{b,t}^{own}$ captures the correlation between the own-stock instrumented price $\log\tilde{P}_{i,n,t}(b)$ and the prices of other stocks in $W_{i,n,t}(\psi)$. So $\omega_{b,t}^{own}=0$ if $Z_{i,n,t}$ only impacts the price of stock n.[31].

[30] In particular, this form arises if $Cov_t\left(\sum_{k=1}^{K} b_{k,t}\theta_{n,k,t}^2, \log\tilde{P}_{i,n,t}(b)\right)=0$.

[31] Hence, in this no spillover case, $\omega_{b,t}^{own} = \frac{Cov_t\left(W_{i,n,t}(\psi),\log\tilde{P}_{i,n,t}(b)\right)}{\mathbb{V}\left[\log\tilde{P}_{i,n,t}(b)\right]} = 0 \rightarrow \alpha_{b,t}^{own} =$

The term $\omega_{b,t}^{cross}$ captures the correlation between the cross-stock price term $\log \tilde{P}_{i,n,t}^{cross}(b)$ and $W_{i,n,t}(\psi)$, which is essentially the average weighted similarity $\sum_{k=0}^{K} b_k \theta_{n,k,t} \theta_{m,k,t}$ for stocks in the top b bins.[32] Thus, intuitively, $\alpha_{b,t}^{cross}$ should decline as b increases and more dissimilar stocks are included in the regression.

Crucially, since given a ψ we can construct $W_{i,n,t}(\psi)$ from all $\log P_{n,t}, \theta_{n,k,t}$, and $\theta_{m,k,t}$, we can measure these correlation terms $\omega_{own}(b)$ and $\omega_{cross}(b)$ empirically. Moreover, classical measurement error in the factor loadings $\theta_{n,k,t}$, and so in $W_{i,n,t}(\psi)$, does not bias the estimates of these correlation terms: $W_{i,n,t}(\psi)$ only appears in the numerators and classical measurement error is by definition uncorrelated with the residualized instrumented prices $\log \tilde{P}_{i,n,t}(b)$ and $\log \tilde{P}_{i,n,t}^{cross}(b)$.

Given this structure, we can compare our empirical estimates of $\alpha_{b,t}^{cross}$ to the theoretical values $\alpha_{b,t}^{cross,theoretical}(\psi)$ that would prevail under different values of ψ:

$$\alpha_{b,t}^{cross,theoretical}(\psi) = \frac{1}{L_t(\psi)} \omega_{b,t}^{cross} \tag{167}$$

$$L_t(\psi) = \mathbb{E}_b \left[\frac{-1 + \mathbb{E}_t \left[\sum_{k=0}^{K} b_k \theta_{n,k,t}^2 \right] + \omega_{b,t}^{own}}{\alpha_{b,t}^{own}} \right]. \tag{168}$$

Given ψ, $1/L_t(\psi)$ is micro elasticity $1/(\lambda + \gamma\sigma^2)$ implied by $\alpha_{b,t}^{own}$, following (165). $\mathbb{E}_b$ is the average across all bins b. Intuitively, larger magnitude own-price coefficients $\alpha_{b,t}^{own}$ imply larger micro elasticities $1/L_t(\psi)$, which in turn imply higher cross-price coefficients $\alpha_{b,t}^{cross}$.

Hence, we can estimate ψ with generalized method of moments (GMM) and this

$-\frac{1}{\lambda+\gamma\sigma^2}\left(1 - \mathbb{E}_t\left[\sum_{k=0}^{K} b_k \theta_{n,k,t}^2\right]\right)$

[32] If the instrumented prices $\log \tilde{P}_{i,n,t}(b)$ are uncorrelated and have the same variance across stocks, then $\omega_{b,t}^{cross}$ is exactly the average weighted similarity $\sum_{k=0}^{K} b_k \theta_{n,k,t} \theta_{m,k,t}$ for all stocks in $S_{i,n,t}(b)$: $\omega_{b,t}^{cross} = \frac{Cov_t\left(W_{i,n,t}(\psi), \log \tilde{P}_{i,n,t}^{cross}(b)\right)}{\mathbb{V}\left[\log \tilde{P}_{i,n,t}^{cross}(b)\right]} = \mathbb{E}\left[\frac{1}{|S_{i,n,t}(b)|} \sum_{m \in S_{i,n,t}(b)} \sum_{k=0}^{K} b_k \theta_{n,k,t} \theta_{m,k,t}\right]$. A sufficient condition for this uncorrelated case is that each instrument $Z_{i,n,t}$ only impacts the price of stock n and the instruments are uncorrelated across stocks. In general, $\omega_{b,t}^{cross}$ is a linear combination of the weighted similarities for all stocks in $S_{i,n,t}(b)$, with higher weights on stocks m that have larger covariances of $\log \tilde{P}_{i,n,t}(b)$ with $\log \tilde{P}_{i,n,t}^{cross}(b)$.

moment condition[33]:

$$\mathbb{E}\left[\alpha_{b,t}^{cross} - \alpha_{b,t}^{cross,theoretical}\left(\psi_t\right)\right] = 0, \tag{169}$$

where the expectation is taken across all bins b within each quarter t. We then report the time series average of ψ_t.

J.3.2 Empirical Results

We find that cross-price elasticities are empirically small, which indicates that substitution is weak and ψ is small.

Figure J.1 displays the time-series average empirical estimates of the reduced-form cross-price coefficients $\alpha_{b,t}^{cross}$. These reduced-form coefficients are statistically and economically smaller than the time-series average theoretical values $\alpha_{b,t}^{cross,theoretical}\left(\psi\right)$ that would prevail under the benchmark $\psi = 1$ (holding the estimated average $\alpha_{b,t}^{own}$ fixed), even among the most similar stocks. Empirically, the average $\alpha_{b,t}^{cross}$ ranges from 0.00012 to 0.012, statistically significantly smaller than the 0.00023 to 0.023 range that prevails under $\psi = 1$. Appendix Figure J.2 displays the $\alpha_{b,t}^{cross}$ and $\alpha_{b,t}^{own}$ results for all bins. Thus, we can reject the null hypothesis of maximum substitution: that γ is far larger than λ and $\psi = 1$.

Next, we quantify ψ. Figure J.1 also compares the empirical estimates of α_{b}^{cross} to the theoretical values that arise under smaller values of ψ (holding the estimated average $\alpha_{b,t}^{own}$ fixed). The GMM estimation discussed in Section J.3.1 simply finds the value of ψ that minimizes the distance between the empirical α_{b}^{cross} coefficients (displayed in the solid red line) and their theoretical counterparts (displayed in the dashed lines). The theoretical curves are close to one another because the instrumented prices $\log \tilde{P}_{i,n,t}(b)$ are strongly correlated across stocks, which leads to large $\omega_{b,t}^{own}$ values in (167). Nevertheless, since

[33] In particular, we solve

$$\min_{\psi} \boldsymbol{g}(\psi)' \boldsymbol{g}(\psi),$$

where $\boldsymbol{g}(\psi) = \left[\alpha_{b,t}^{cross} - \alpha_{b,t}^{cross,theoretical}\left(\psi\right)\right]_{b=1}^{B}$.

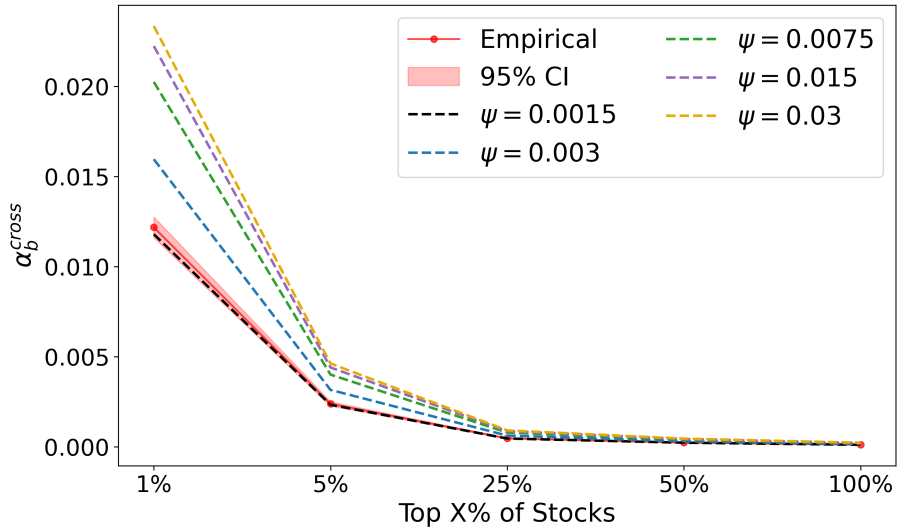


Figure J.1: This figure compares the time-series averages of the reduced-form cross-price coefficients $\alpha_{b,t}^{cross}$ from regression (162) for selected bins b to their theoretical counterparts under different ψ values. The shaded regions indicate the 95% confidence interval using Fama-MacBeth standard errors.

α_b^{cross} are precisely estimated, we can statistically distinguish between the different values of ψ.

Table J.2 displays the GMM estimate of $\psi = 0.0017$. This estimate is three orders of magnitude smaller than the benchmark $\psi = 1$, which indicates that investors weakly substitute across stocks.

	ψ
Average Point Estimate	0.0017***
	(6.2e-05)

Table J.2: This table displays the GMM estimates of ψ implied by the reduced-form own- and cross-price coefficient estimates $\alpha_{b,t}^{own}$ and $\alpha_{b,t}^{cross}$. We estimate ψ_t in each quarter t following (169) and report the time series average. The second row displays the Fama-MacBeth standard error. Levels of significance are presented as follows: *p<0.1; **p<0.05; ***p<0.01.

J.3.3 Proofs

Let $\log \tilde{P}_{i,n,t}(b)$ and $\log \tilde{P}_{i,n,t}^{cross}(b)$ be the residuals from regressing each of the *instrumented* $\log P_{n,t}$ and $\log P_{i,n,t}^{cross}$ on the other variable and all controls. Let $b_k = \frac{\psi \tilde{e}_k}{1+\psi \tilde{e}_k}$.

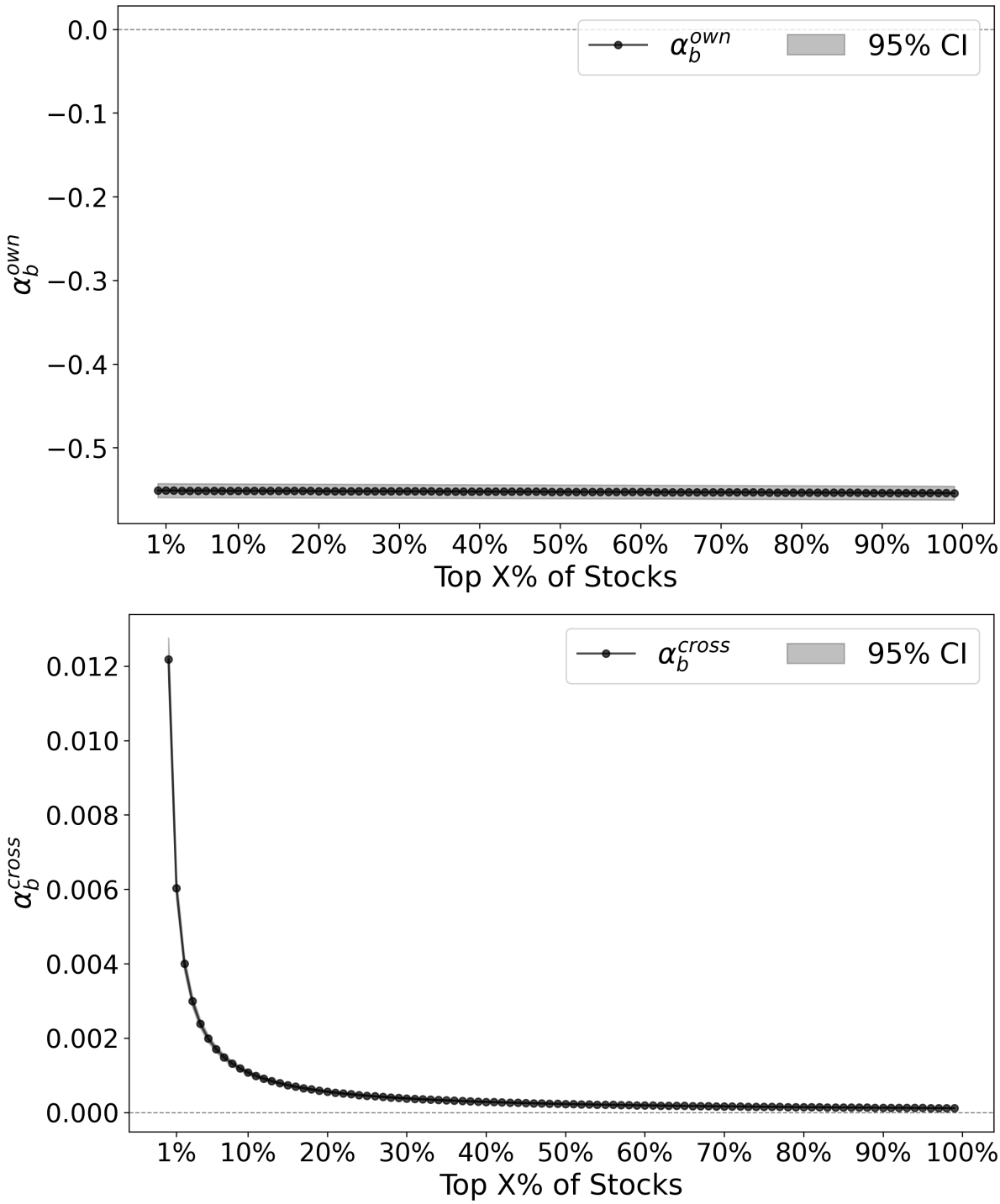


Figure J.2: This figure displays the time-series averages of the reduced-form own- (left) and cross-price (right) coefficients $\alpha_{b,t}^{own}$ and $\alpha_{b,t}^{cross}$ from regression (170) for all bins b. Shaded regions indicate 95% confidence intervals calculated using Fama-MacBeth standard errors (which are very small for $\alpha_{b,t}^{own}$).

Given the reduced-form two-stage least squares equation (162),

$$\log Q_{i,n,t} = \alpha_{b,t}^{own} \log P_{n,t} + \alpha_{b,t}^{cross} \underbrace{\sum_{m \in S_{i,n,t}(b)} \log P_{m,t}}_{\equiv P_{i,n,t}^{cross}(b)} + \boldsymbol{c}_{b,t}^{'} \boldsymbol{\theta}_{n,t} + e_{i,n,t} \tag{170}$$

$$\begin{bmatrix} \log P_{n,t} \\ \log P_{i,n,t}^{cross} \end{bmatrix} = \begin{bmatrix} \pi_{1,b,t}^{own} & \pi_{1,b,t}^{cross} \\ \pi_{2,b,t}^{own} & \pi_{2,b,t}^{cross} \end{bmatrix} \begin{bmatrix} Z_{i,n,t} \\ \sum_{m \in S_{i,n,t}(b)} \log Z_{i,m,t} \end{bmatrix} + \begin{bmatrix} \boldsymbol{b}_{1,b,t}^{'} \\ \boldsymbol{b}_{2,b,t}^{'} \end{bmatrix} \boldsymbol{\theta}_{n,t} + \begin{bmatrix} u_{b,i,n,t} \\ v_{b,i,n,t} \end{bmatrix} \tag{171}$$

we have by the Frisch-Waugh-Lovell theorem that

$$\begin{aligned} \alpha_{b,t}^{own} &= \frac{Cov_t\left(\log Q_{i,n,t}, \log \tilde{P}_{i,n,t}(b)\right)}{\mathbb{V}\left[\log \tilde{P}_{i,n,t}(b)\right]} \\ &= \frac{1}{\lambda + \gamma\sigma^2}\left(-1 + \mathbb{E}_t\left[\sum_{k=0}^{K} b_k \theta_{n,k,t}^2\right]\right) \frac{Cov_t\left(\log P_{n,t}, \log \tilde{P}_{i,n,t}(b)\right)}{\mathbb{V}\left[\log \tilde{P}_{i,n,t}(b)\right]} + \frac{1}{\lambda + \gamma\sigma^2} \frac{Cov_t\left(W_{i,n,t}(\psi), \log \tilde{P}_{i,n,t}(b)\right)}{\mathbb{V}\left[\log \tilde{P}_{i,n,t}(b)\right]} \end{aligned} \tag{172}$$

$$= \frac{1}{\lambda + \gamma\sigma^2}\left(-1 + \mathbb{E}_t\left[\sum_{k=0}^{K} b_k \theta_{n,k,t}^2\right]\right) + \frac{1}{\lambda + \gamma\sigma^2} \underbrace{\frac{Cov_t\left(\hat{W}_{i,n,t}(\psi), \log \tilde{P}_{i,n,t}(b)\right)}{\mathbb{V}\left[\log \tilde{P}_{i,n,t}(b)\right]}}_{\equiv \omega_{b,t}^{own}(\psi)} \tag{173}$$

$$\begin{aligned} \alpha_{b,t}^{cross} &= \frac{Cov_t\left(\log Q_{i,n,t}, \log \tilde{P}_{i,n,t}^{cross}(b)\right)}{\mathbb{V}\left[\log \tilde{P}_{i,n,t}^{cross}(b)\right]} \\ &= -\frac{1}{\lambda + \gamma\sigma^2} \frac{Cov_t\left(\log P_{n,t}, \log \tilde{P}_{i,n,t}^{cross}(b)\right)}{\mathbb{V}\left[\log \tilde{P}_{i,n,t}^{cross}(b)\right]} + \frac{1}{\lambda + \gamma\sigma^2} \frac{Cov_t\left(\hat{W}_{i,n,t}(\psi), \log \tilde{P}_{i,n,t}^{cross}(b)\right)}{\mathbb{V}\left[\log \tilde{P}_{i,n,t}^{cross}(b)\right]} \end{aligned} \tag{174}$$

$$= \frac{1}{\lambda + \gamma\sigma^2} \underbrace{\frac{Cov_t\left(\hat{W}_{i,n,t}(\psi), \log \tilde{P}_{i,n,t}^{cross}(b)\right)}{\mathbb{V}\left[\log \tilde{P}_{i,n,t}^{cross}(b)\right]}}_{\equiv \omega_{b,t}^{cross}(\psi)}. \tag{175}$$

(172) and (174) follow from substituting in the structural equation below:

$$\log Q_{i,n,t} = -\frac{1}{\lambda + \gamma\sigma^2}\left(1 - \mathbb{E}_t\left[\sum_{k=0}^{K} \underbrace{b_k}_{\equiv \frac{\psi \tilde{e}_k}{1+\psi \tilde{e}_k}} \theta_{n,k,t}^2\right]\right) \log P_{n,t} + \frac{1}{\lambda + \gamma\sigma^2} W_{i,n,t}(\psi) + \epsilon_{i,n,t},$$

$$W_{i,n,t}(\psi) = \sum_{m \neq n \in \mathcal{I}_{i,t}} \sum_{k=0}^{K} b_k \theta_{n,k,t} \theta_{m,k,t} \log P_{m,t}.$$

(173) follows since $Cov_t\left(\log P_{n,t}, \log \tilde{P}_{i,n,t}(b)\right) = \mathbb{V}_t\left[\log \tilde{P}_{i,n,t}(b)\right]$. (175) follows since $Cov_t\left(\log P_{n,t}, \log \tilde{P}^{cross}_{i,n,t}(b)\right) = 0$.

Note that if each instrument $Z_{i,n,t}$ only impacts the price of stock n, then the (residualized) instrumented price $\log \tilde{P}_{i,n,t}(b)$ is uncorrelated with all of the other prices $\log P_{m,t}$ for $m \neq n$ in $W_{i,n,t}(\psi)$. Therefore,

$$Cov_t\left(W_{i,n,t}(\psi), \log \tilde{P}_{i,n,t}(b)\right) = \sum_{m \neq n \in \mathcal{I}_{i,t}} \sum_{k=0}^{K} b_k \theta_{n,k,t} \theta_{m,k,t} Cov_t\left(\log P_{m,t}, \log \tilde{P}_{i,n,t}(b)\right) = 0,$$

and so $\omega^{own}_{b,t} = 0$.

If the (residualized) instrumented prices $\log \tilde{P}_{i,n,t}(b)$ are uncorrelated across stocks, then:

$$Cov_t\left(W_{i,n,t}(\psi), \log \tilde{P}^{cross}_{i,n,t}(b)\right) = \sum_{m \in S_{i,n,t}(b)} \sum_{k=0}^{K} b_k \theta_{n,k,t} \theta_{m,k,t} \mathbb{V}_t\left[\log \tilde{P}_{i,n,t}(b)\right]$$

$$\mathbb{V}_t\left[\log \tilde{P}^{cross}_{i,n,t}(b)\right] = \sum_{m \in S_{i,n,t}(b)} \mathbb{V}_t\left[\log \tilde{P}_{i,n,t}(b)\right].$$

If all stocks have the same instrumented price variance

$$\mathbb{V}_t\left[\log \tilde{P}_{i,n,t}(b)\right] = \sigma^2_P,$$

then we have:

$$\omega^{cross}_{b,t} = \mathbb{E}_t\left[\frac{1}{|S_{i,n,t}(b)|} \sum_{m \in S_{i,n,t}(b)} \sum_{k=0}^{K} b_k \theta_{n,k,t} \theta_{m,k,t}\right].$$

J.4 Implications for Biases in Estimated Elasticities

Previous work argues that failing to appropriately control for substitution can bias own-price elasticity estimates (Fuchs et al., 2023, 2025; Haddad et al., 2025). However, the weak degree of substitution indicated by the small ψ estimates in Table 3 suggests that this bias is likely small. The bias in own-price elasticity estimates is controlled by the

$\frac{1}{\lambda+\gamma\sigma^2}\omega_{b,t}^{own}$ term in (165), which shrinks as ψ decreases. This is a similar analysis to that in Appendix I.3, except for cross-price elasticities instead of multipliers.

We quantify the magnitude of this bias given the ψ estimate in Table 3. We run in each quarter, pooled across all investors and stocks, the following two-stage least squares regression of log quantities held on instrumented prices without controlling for any substitute terms:

$$\log Q_{i,n,t} = \alpha_t^{own} \log P_{n,t} + \boldsymbol{c}_t^{'}\boldsymbol{\theta}_{n,t} + e_{b,i,n,t} \tag{176}$$

$$\log P_{n,t} = \pi_t^{own} Z_{i,n,t} + \boldsymbol{b}_t^{'}\boldsymbol{\theta}_{n,t} + u_{b,i,n,t}. \tag{177}$$

Given the ψ estimate, we compute $\omega_t^{own} = \frac{Cov_t(W_{i,n,t}(\psi), \log \tilde{P}_{i,n,t})}{\mathbb{V}_t[\log \tilde{P}_{i,n,t}]}$, where $\tilde{P}_{i,n,t}$ is the instrumented log price residualized with respect to all controls. We compute the implied $L_t(\psi)$ as in (168). We can then compare the time-series averages for the estimated α_t^{own} and the de-biased own-price and micro elasticities

$$\text{Own-Price Elasticity}_t = -\frac{1}{L_t(\psi)}\left(1 - \mathbb{E}_t\left[\sum_{k=0}^{K} b_k \theta_{n,k,t}^2\right]\right) = \alpha_t^{own} - \frac{1}{L_t(\psi)}\omega_t^{own} \tag{178}$$

$$\text{Micro Elasticity}_t = -\frac{1}{L_t(\psi)} = \alpha_t^{own} - \frac{1}{L_t(\psi)}\left(\omega_t^{own} + \mathbb{E}_t\left[\sum_{k=0}^{K} b_k \theta_{n,k,t}^2\right]\right). \tag{179}$$

Appendix Table J.3 demonstrates that even after removing the bias from failing to control for substitution, own-price elasticities remain small, around -0.53. In fact, the de-biased own-price elasticity is even more inelastic (i.e., smaller in magnitude) than the average α_t^{own} because ω_t^{own} is empirically negative in our setting. Since the $\mathbb{E}_t\left[\sum_{k=0}^{K} b_k \theta_{n,k,t}^2\right]$ term is small given the small ψ estimates, full own-price elasticities and micro elasticities are virtually identical.[34]

Thus, own-price elasticities are virtually identical to the micro elasticities and both parameters remain small in magnitude even after controlling for substitution.[35] Thus,

[34] Moreover, note that (176) and (177) correspond to the regression suggested by Haddad et al. (2025) since we control for the factor loadings interacted with time fixed effects.

[35] Our results are consistent with those in Tamoni et al. (2024) Appendix B.2.2, which shows that controlling for up to 49 Fama-French industries interacted with time fixed effects has little impact on the estimated elasticities.

asset demand systems in stocks that place strong assumptions on substitution do not lead to large biases in estimated own-price elasticities.

	Estimate
α^{own}	-0.700^{***}
	(0.040)
Own-Price Elasticity	-0.528^{***}
	(0.058)
Micro Elasticity	-0.529^{***}
	(0.058)

Table J.3: This table displays in the first row the estimated α^{own} from (176). The last two rows display the de-biased full own-price and micro elasticities using the ψ estimate in Table J.2. Quarterly block-bootstrapped standard errors over 500 simulations are displayed in parentheses. Levels of significance are presented as follows: *p<0.1; **p<0.05; ***p<0.01.

J.5 Additional Empirical Results

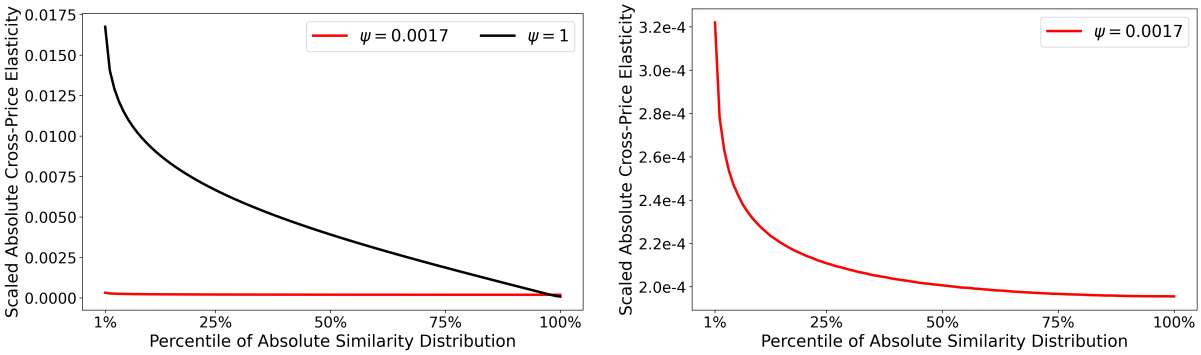


Figure J.3: This figure displays the model-implied ratios of absolute structural cross-price elasticities to own-price elasticities under different values of ψ. The right panel displays these ratios under the empirical estimate of ψ from Table J.2 while the left panel overlays the ratios that would emerge under the benchmark $\psi = 1$.

K Technical Details for the Estimation Error and Shrinkage Tests

This appendix records the mechanics of the estimation error and shrinkage tests summarized in Section 9.

Why the exercise can be written in terms of ψ alone. Demand is given by equation (39) in Appendix A:

$$\boldsymbol{w}_t = \left[\gamma \boldsymbol{\Theta}_t \mathbf{E}_t \boldsymbol{\Theta}_t' + \left(\lambda + \gamma\sigma^2\right)\mathbf{I}\right]^{-1} \boldsymbol{\mu}_t. \tag{180}$$

Now use the paper's notation $\zeta^{Own} = 1/(\lambda + \gamma\sigma^2)$ from (4) and $\psi = \gamma\sigma^2/(\lambda + \gamma\sigma^2)$ from (7). Then

$$\boldsymbol{w}_t = \zeta^{Own}\left(\mathbf{I} + \psi\frac{\boldsymbol{\Theta}_t \mathbf{E}_t \boldsymbol{\Theta}_t'}{\sigma^2}\right)^{-1} \boldsymbol{\mu}_t. \tag{181}$$

Thus, ζ^{Own} only affects overall leverage. Since multiplying all portfolio weights by a positive scalar does not change the Sharpe ratio, Sharpe comparisons across candidate ψ's depend only on the portfolio direction

$$\widetilde{w}_t(\psi) \propto \left(\mathbf{I} + \psi\frac{\boldsymbol{\Theta}_t \mathbf{E}_t \boldsymbol{\Theta}_t'}{\sigma^2}\right)^{-1} \widehat{\mu}_t. \tag{182}$$

This is the benchmark factor-model version of the rule. More generally, in the empirical implementations below we replace the factor-model common covariance object $\boldsymbol{\Theta}_t \mathbf{E}_t \boldsymbol{\Theta}_t'$ by an estimated common covariance component $\widehat{\Sigma}_t^{com}$. This yields the generic empirical rule

$$\widetilde{w}_t(\psi) \propto \left(\mathbf{I} + \psi\frac{\widehat{\Sigma}_t^{com}}{\sigma^2}\right)^{-1} \widehat{\mu}_t, \tag{183}$$

where $\widehat{\mu}_t$ is the expected-return forecast, $\widehat{\Sigma}_t^{com}$ is the common covariance component, and $\sigma^2 = 0.01$ is the paper's monthly idiosyncratic-variance calibration. Small ψ means heavy shrinkage, while $\psi = 1$ gives the classical benchmark under this calibration.

Sample split. The hold-in sample is 1980 through the end of 1999. The hold-out sample is 2000 through 2023. For each specification, the portfolio weights used in month t are formed strictly ex ante from information available at t, and realized returns are then

evaluated in the hold-out sample.

Mean and covariance specifications. We consider the same ten specifications summarized in Table K.1. On the covariance side, we use two alternatives:

$$\widehat{\Sigma}_t^{com,\,factor} = \Theta_t \widehat{E}_t \Theta_t',$$
$$\widehat{\Sigma}_t^{com,\,daily} = \left(\widehat{\Sigma}_t - \widehat{\sigma}_t^2 I\right)_+,$$

where $(A)_+$ denotes the projection of a symmetric matrix onto the positive-semidefinite cone, obtained by setting negative factor variances to zero.[36] Both covariance specifications are in monthly units, and we normalize both by the paper's calibrated monthly idiosyncratic variance, $\sigma^2 = 0.01$. Thus the new parameter is exactly 0.01 times the parameter in the original unnormalized exercise, leaving every portfolio and Sharpe ratio unchanged. Here $\widehat{\Sigma}_t$ is the rolling 60-month daily covariance estimate, and

$$\widehat{\sigma}_t^2 = \frac{1}{N_t - K} \operatorname{tr}\Big((I - \widehat{P}_t)\widehat{\Sigma}_t(I - \widehat{P}_t)\Big)$$

is the average covariance in the residual space orthogonal to Θ_t, with $\widehat{P}_t = \Theta_t \Theta_t'$. Note that N_t is the number of stocks and K is the number of factors. This is a natural estimate of the idiosyncratic variance component: under the model covariance decomposition $\Sigma_t = \Theta_t E_t \Theta_t' + \sigma_t^2 I$, the residual-space covariance is exactly $\sigma_t^2(I - \widehat{P}_t)$, so the trace formula recovers σ_t^2. Since finite-sample noise can make $\widehat{\Sigma}_t - \widehat{\sigma}_t^2 I$ slightly indefinite, we clip negative estimated common-covariance factor variances to zero before applying the portfolio rule.

[36] Equivalently, if $A = U \operatorname{diag}(d_j) U'$, then $A_+ = U \operatorname{diag}(\max\{d_j, 0\}) U'$. This is the Frobenius-norm projection of A onto the positive-semidefinite cone.

On the expected-return side, we consider five forecasts:

$$\begin{aligned}
\widehat{\mu}_t^{FM} &= \Theta_t \widehat{\lambda}_{FM}^{\Theta}, \\
\widehat{\mu}_t^{NN8} &= f_8(\Theta_t), \\
\widehat{\mu}_t^{NN32} &= f_{32}(\Theta_t), \\
\widehat{\mu}_t^{NN64\times 64} &= f_{64\times 64}(\Theta_t), \\
\widehat{\mu}_t^{Hist60} &= \frac{1}{60}\sum_{j=1}^{60} r_{t+1-j}.
\end{aligned}$$

These correspond to a fixed pre-2000 Fama-MacBeth forecast, three unregularized neural-network forecasts fit on the pre-2000 stock-month sample with architectures (8), (32), and $(64, 64)$, and a rolling 60-month stock-level average return. Aside from the rolling 60-month daily covariance estimate and the rolling historical-average forecast, the inputs are estimated once in the hold-in sample and then carried forward unchanged.

Evaluation. For each specification and each candidate ψ, we use (183) to generate a time series of ex ante portfolio weights and corresponding realized hold-out returns. This produces a hold-out Sharpe ratio for each ψ. We then choose

$$\widehat{\psi}^{SR} = \arg\max_{0\le\psi\le 1} \widehat{SR}^{eval}(\psi). \tag{184}$$

Table K.1 reports the results for all ten combinations of mean and covariance estimates. The first two columns list the mean forecast and covariance-side specification; the remaining columns report the Sharpe-maximizing ψ, the Sharpe ratio at that value, the Sharpe ratio at $\psi = 0$, and the Sharpe ratio at the classical benchmark $\psi = 1$. Within each covariance-side block, moving to more flexible expected-return forecasts generally lowers the Sharpe-maximizing ψ. Intuitively, a more flexible mean forecast creates more scope for estimation error, making greater shrinkage valuable.

Because the same hold-out sample is used both to choose $\widehat{\psi}^{SR}$ and to report performance, the exercise should be interpreted as a fixed-split portfolio diagnostic, not as a

Mean forecast	Covariance side	$\widehat{\psi}^{SR}$	Sharpe ($\widehat{\psi}^{SR}$)	Sharpe ($\psi = 0$)	Sharpe ($\psi = 1$)
Fama-MacBeth	Factor model	0.0188	0.5561	0.4173	0.3887
NN (8)	Factor model	0.0027	0.5329	0.4642	0.1218
NN (32)	Factor model	0.0004	0.3482	0.3468	-0.1073
NN (64,64)	Factor model	0.0007	0.2120	0.2050	-0.0415
Historical Avg (60m)	Factor model	0.00	0.3744	0.3744	0.1766
Fama-MacBeth	Rolling 5-year daily	0.0375	0.7931	0.5705	0.4840
NN (8)	Rolling 5-year daily	0.0065	0.6389	0.6041	0.2943
NN (32)	Rolling 5-year daily	0.0000	0.5225	0.5225	0.0069
NN (64,64)	Rolling 5-year daily	0.0000	0.3666	0.3666	-0.1008
Historical Avg (60m)	Rolling 5-year daily	0.0000	0.3649	0.3649	0.0524

Table K.1: Sharpe-based portfolio test across mean-forecast and covariance-side specifications. Within each covariance-side block, more flexible expected-return forecasts generally require more shrinkage to maximize hold-out Sharpe, i.e. a lower Sharpe-maximizing ψ. The last column reports Sharpe at the classical benchmark $\psi = 1$. For the rolling daily covariance rows, we project the estimated common covariance component onto the positive-semidefinite cone by clipping negative estimated factor variances to zero.

fully out-of-sample structural estimate of ψ.