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

Aditya Chaudhry and Carter Davis

March 16, 2026

ABSTRACT

We show that investors treat individual stocks as weak substitutes, and that this weak substitutability explains the existence of the "factor zoo" in expected stock returns. In classical asset-pricing models, substitutability is strong: when relative expected returns change, investors readily reallocate across assets that covary. In this case, there would be no factor zoo: differences in expected returns would be explained by the few factors that drive most covariance across stocks. In contrast, we show that substitutability is empirically weak in investor holdings data. We demonstrate that this weak substitutability gives rise to the factor zoo: expected returns depend on many factors, including those that drive little covariance across stocks.

JEL classification: G11, G12

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

1 Introduction

Do investors treat individual stocks as close substitutes or weak substitutes? If investors view stocks as close substitutes, then they will aggressively reallocate when one stock's expected return rises relative to another's. Yet if investors view stocks as weak substitutes, then they will only weakly adjust their portfolios as expected returns diverge. Thus, in equilibrium, the degree of substitutability shapes how expected returns vary across stocks. Asset pricing has traditionally assumed strong substitutability since stocks, unlike art, are claims to streams of ultimately fungible cash flows (Scholes (1972)). Yet, despite its importance in shaping the cross-section of expected returns, this degree of substitutability remains empirically unclear.

In this paper, we show empirically that investors treat individual stocks as weak substitutes and that this weak substitutability explains the existence of the factor zoo in expected stock returns. Intuitively, if stocks are close substitutes, then investors aggressively reallocate across assets that strongly covary, forcing them to earn similar expected returns. In this case, the few factors that explain most return comovement also explain most cross-sectional differences in expected returns. However, if stocks are weak substitutes — that is, if *cross-price elasticities are small* — this link breaks down, and expected returns disperse across many factors, including those that contribute little to comovement.

In particular, we make three contributions. First, we show theoretically that as substitutability weakens — as cross-price elasticities fall — the number of return factors required to explain cross-sectional differences in expected returns rises beyond the number of factors required to explain covariances. Second, we develop a formal test to show that, indeed, far more factors are required to explain cross-sectional differences in expected returns than are required to explain covariances. Third, we directly measure cross-price elasticities in institutional investor holdings data and find they are quantitatively small. This weak substitutability rationalizes the existence of the factor zoo.

We begin by presenting a stylized model to formalize the link between substitutabil-

ity and equilibrium expected returns (in Section 2). In particular, we extend the model of Kozak et al. (2018). Noise traders demand stocks with certain characteristics, and these demands are satisfied by a representative arbitrageur. This arbitrageur has mean-variance preferences but, crucially, is also subject to an ℓ_2 penalty that dissuades him from taking large positions in any single stock. This penalty can arise from many different microfoundations, including benchmarking, Bayesian uncertainty, or inventory or adjustment costs. When the penalty is zero, we recover classical mean-variance demand in which stocks are close substitutes and cross-price elasticities are large. Yet as this penalty increases, substitutability weakens.

We show that, the arbitrageur's willingness to substitute across stocks determines the structure of equilibrium expected returns. When the ℓ_2 penalty is small relative to risk aversion, cross-price elasticities are large, and the few factors that explain most covariances between stocks also explain most differences in expected returns. However, as the ℓ_2 penalty rises and substitutability weakens, 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 substitutability implies many factors are required to explain expected returns, and a high-dimensional factor structure in expected returns can only be sustained if stocks are weak substitutes. Importantly, we show that this dispersion in expected returns across many factors does not give rise to near arbitrage opportunities.

We next connect this theory to the data by calibrating the magnitude of cross-price elasticities implied by existing estimates of price multipliers (in Section 3). We start from a simple two-parameter structure in which the price multiplier matrix is summarized by an own-price term and a common cross-price term, as in Haddad et al. (2025). Previous work shows that a 1% idiosyncratic flow shock raises the price of the affected stock by about 2%, whereas a 1% marketwide flow shock moves the entire market by about 5% (Gabaix and Koijen, 2021; Davis, 2025; Li and Lin, 2022). Translating these multiplier estimates into elasticities using our model, we find that own-price elasticities are economically meaningful (around 0.5) while cross-price elasticities are extremely small (around 0.0001).

Thus, this calibration indicates weak substitutability.

Given these small calibrated cross-price elasticities, our model predicts that explaining the cross section of expected returns should require many factors. We test this prediction using reduced-form evidence on the factor structure of returns. First, we estimate a rich factor model using Instrumented Principal Components Analysis (IPCA) on daily stock returns with a large set of firm characteristics as instruments, following Kelly et al. (2019). This delivers a high-dimensional set of factors and associated eigenvalues that explain realized return comovement. In theory, under strong substitutability, the few factors with large eigenvalues that explain most comovement 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. That is, while only a handful of factors explain most realized return covariances, many factors are required to explain the variation in expected returns. Theoretically, this situation can only arise under weak substitutability.

We find similar patterns when we repeat the exercise with standard PCA (using the deflated heteroskedastic PCA of Zhou and Chen, 2025) on daily stock returns and on daily returns of the 56 characteristic-sorted portfolios of Kozak et al. (2018). In these settings, we continue to find that a few factors explain most covariances across assets, but many factors are needed to explain cross-sectional differences in expected returns, consistent with weak substitutability.

We then directly estimate the degree of substitutability in institutional investor holdings data, and show that it is small (in Section 5). We estimate cross-price elasticities of demand for institutional investors in the U.S. stock market using holdings data from the SEC Form 13F. Using the optimal granular instrumental variable (GIV) approach of Chaudhary et al. (2024), which extends the approach of Gabaix and Koijen (2021), we find that cross-price elasticities are small — two to four orders of magnitude smaller than own-price micro elasticities, consistent with our calibration. The weak degree of substitutability implied by these small cross-price elasticities is quantitatively consistent

with the dispersed nature of expected stock returns we document.

Why do cross-elasticities and their link to substitution patterns and weak factors matter? First, they speak directly to how investors make portfolio decisions. The insight of Markowitz (1952) was to move from choosing stocks in isolation to choosing portfolios based on covariances, a major step forward in the theory of portfolio choice. Our results suggest that, in practice, investors behave somewhere in between the pure Markowitz world and the pure isolationist world: they recognize comovement, but frictions and additional motives make large substitutions along any single covariance direction unattractive. Understanding where the data lie between these two extremes is crucial for interpreting observed portfolios and for evaluating positive portfolio models.

Second, the magnitude of cross-price elasticities lies at the heart of ongoing debates in demand-system asset pricing. Many recent models impose limited or homogeneous substitutability across asset, such as by grouping assets coarsely or restricting the dimension of substitution patterns. If true cross-price elasticities were large and heterogeneous, such models would misrepresent substitution and bias own-price elasticities downward, thereby overstating the role of flows in moving prices (Chaudhary et al., 2023; Fuchs et al., 2023; Haddad et al., 2025). Our results instead indicate that cross-price elasticities in the stock market are small. This finding lends support to models with limited substitutability: 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.

Third, the stock market's weak factor structure has been documented but not fully explained. A large empirical literature finds that parsimonious factor models explain only a modest share of the cross section of expected returns, while more complex, high-dimensional models perform much better (Lettau and Pelger, 2020; Kelly et al., 2024). Our results provide an economic rationale: when cross-price elasticities are small, expected returns simply do not concentrate on a handful of dominant risk factors, even if realized return comovement does. A genuine anomaly zoo — a situation in which many different characteristics and portfolio strategies appear to predict returns — is synonymous with a weak factor structure. Our theory shows that limited substitutability

is a natural mechanism that generates this outcome.

The rest of the paper proceeds as follows. Section 2 develops our theoretical framework, introducing the mean–variance investor with an ℓ_2 position penalty, deriving the implied elasticity matrix, and characterizing how cross-price elasticities shape the factor structure of expected returns. Section 3 presents a simple calibration of cross-price elasticities based on micro and macro estimates of price multipliers. 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 substitutability. Finally, Section 5 estimates a structural demand system that nests our theoretical framework and directly recovers cross-price elasticities from the joint behavior of prices, flows, and characteristics.

2 Theoretical Framework

In this section, we present a simple model to characterize the relationship between stock-level substitutability and equilibrium expected returns.

2.1 Model Setup

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

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

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

$\boldsymbol{F} \in \mathbb{R}^K$ is a vector of common, orthogonal factors with $\mathbb{V}[\boldsymbol{F}] = \mathbf{E} \equiv \text{diag}(e_k)_{k=1}^K$ and $K = \phi N$. The eigenvalues are in decreasing order: $e_1 \geq \cdots \geq e_k$.

The factor loadings $\boldsymbol{\theta}_n$ are orthonormal: $\boldsymbol{\Theta} = [\boldsymbol{\theta}_n']_{n=1}^N$ satisfies $\boldsymbol{\Theta}'\boldsymbol{\Theta} = \boldsymbol{I} \in \mathbb{R}^{K \times K}$. The first factor $k = 1$ is a "market" factor, so the first column of $\boldsymbol{\Theta}$ is $[1/\sqrt{N}]_{n=1}^N$.

Furthermore, all factors $k > 1$ are zero-exposure long-short portfolios, so all columns $k > 1$ of $\boldsymbol{\Theta}$ satisfy $\boldsymbol{\iota}'\boldsymbol{\Theta}_{k,\cdot} = 0$, where $\boldsymbol{\iota}$ is a vector of ones (as in Kozak et al. (2018)).

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 $1/N$ shares.

The exogenous risk-free rate is normalized to zero.

Agents and Preferences There are noise traders with random, exogenous demand for exposure to certain factors (as in Kozak et al. (2018)):

$$\boldsymbol{Q}^N = \boldsymbol{\Theta}\boldsymbol{\pi}$$

where $\boldsymbol{\pi} \in \mathbb{R}^K$ controls the level of noise demand for each factor.

A representative arbitrageur absorbs this demand. The arbitrageur has mean-variance preferences over terminal wealth but faces an ℓ_2 penalty on the size of his positions:

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

Such an ℓ_2 penalty arises from many frictions (e.g., benchmarking, adjustment costs, inventory costs) or deviations from full information rational expectations (e.g., uncertainty about expected returns or covariances). The main difference between this model and that in Kozak et al. (2018) is that the latter has $\lambda = 0$.

2.2 Optimal Portfolio Choice & Elasticities

The arbitrageur's optimal portfolio choice is given by:

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

By the Woodbury matrix identity, the arbitrageur's own-price and cross-price elastic-

ities for all stocks are given by the following elasticity matrix:

$$-\left[\gamma\mathbf{\Theta}\mathbf{E}\mathbf{\Theta}'+\lambda\mathbf{I}\right]^{-1}=-\frac{1}{\lambda}\mathbf{I}+\frac{\gamma}{\lambda^2}\mathbf{\Theta}\left[\mathbf{E}^{-1}+\frac{\gamma}{\lambda}\mathbf{I}\right]^{-1}\mathbf{\Theta}'. \tag{2}$$

The own price elasticities are the diagonal elements

$$\zeta_n\equiv\frac{\partial Q_n}{\partial P_n}=-\frac{1}{\lambda}\left(1-\sum_{k=1}^{K}\theta_{n,k}^2\frac{\gamma e_k}{\lambda+\gamma e_k}\right) \tag{3}$$

Stocks with more exposure $\theta_{n,k}$ to large eigenvalue factors (those with high e_k) have lower own-price elasticities (in magnitude). Intuitively, if n has large exposures to high eigenvalue factors, then the arbitrageur will be unwilling to raise his position in n when price falls because doing so adds a lot of non-diversifiable risk to the portfolio.

The cross-price elasticities are the off-diagonal elements in (2)

$$\zeta_{n,m}\equiv\frac{\partial Q_n}{\partial P_m}=\frac{1}{\lambda}\sum_{k=1}^{K}\theta_{n,k}\theta_{m,k}\frac{\gamma e_k}{\lambda+\gamma e_k}. \tag{4}$$

Two stocks n and m 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 the large eigenvalue factors. Intuitively, if the price of stock m goes up, the investor will want to substitute to a "similar" stock — one whose covariances with all other assets are similar to those of stock m. Stocks with similar factor loadings, especially for the high eigenvalue factors, have similar covariance properties.

Cross-price elasticities approach zero as $\gamma\to 0$ in (4). Intuitively, in mean-variance preferences substitutability 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 have similar covariances with all other assets. The less the investor cares about portfolio volatility (i.e., the lower γ is), the less substitutable she will consider individual stocks.

Since both own- and cross-price elasticities depend on both λ and γ, we let γ vary

the degree of substitutability (i.e., cross-price elasticities) while adjusting λ to maintain a constant average own-price elasticity. Appendix A.1 shows the average own-price elasticity is

$$\frac{1}{N}\sum_{n=1}^{N}\zeta_n = -\frac{1}{\lambda}\left(1-\phi\mathbb{E}^k\left[\frac{\psi e_k}{1+\psi e_k}\right]\right), \tag{5}$$

where $\psi = \gamma/\lambda$ controls the degree of substitutability: lower ψ corresponds to lower γ and weaker substitutability. The expectation $\mathbb{E}^k$ is taken across $k = 1, .., K$. Thus,

$$\lambda(\psi) \equiv \frac{1-\phi\mathbb{E}^k\left[\frac{\psi e_k}{1+\psi e_k}\right]}{-\bar{\zeta}} \tag{6}$$

fixes the own-price elasticity at $\bar{\zeta}$ as ψ varies. Note that $\lambda(\psi)$ is decreasing in ψ: as ψ rises with risk aversion γ, this investor's demand becomes more inelastic, so λ must fall to maintain an average own-price elasticity of $\bar{\zeta}$.

2.3 Equilibrium Expected Returns

All proofs for this section are in Appendix A.2. Market clearing

$$\boldsymbol{Q} + \boldsymbol{Q}^N = \frac{1}{N}\boldsymbol{\iota}, \forall n$$

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

$$\mathbb{E}[\boldsymbol{D}] - \boldsymbol{P} = -\boldsymbol{\Theta}\left[\lambda(\psi)\pi_k\left(\psi e_k + 1\right)\right]_{k=1}^{K}. \tag{7}$$

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

$$\mathbb{V}[\mathbb{E}[D_n] - P_n] = \lambda(\psi)^2\sum_{k=2}^{K}\pi_k^2\left(\psi e_k + 1\right)^2, \tag{8}$$

Across many realizations of noise demand $\boldsymbol{\pi}$, the proportion of cross-sectional expected

return variance coming from (exposures to) the k-th factor is on average

$$f_k \equiv \frac{(\psi e_k + 1)^2}{\sum_{j=2}^{K} (\psi e_j + 1)^2}. \tag{9}$$

Note that f_k increases with the eigenvalue e_k. Thus, the factors with higher eigenvalues — those that explain more co-movement across stocks — explain more of the cross-sectional variation in expected returns, as in Kozak et al. (2018).

However, the relative importance of the high-eigenvalue factors decreases as substitutability weakens.

Proposition 1. *For $e_k > e_j$, the ratio f_k/f_j is increasing in ψ.*

If substitutability is strong ($\psi \to \infty$, as in Kozak et al. (2018)), we have:

$$f_k \xrightarrow{\psi \to \infty} \frac{e_k^2}{\sum_{j=2}^{K} e_j^2}. \tag{10}$$

The larger eigenvalue factors explain a greater proportion of expected return variation (as in Kozak et al. (2018)). Intuitively, since the larger eigenvalue factors create more comovement across stocks, they contribute more to portfolio volatility, and so the arbitrageur demands more compensation for exposure to these factors.

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

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

All factor explain the same proportion of expected return variation, regardless of their eigenvalues. Intuitively, shutting down substitutability leads the investor to not demand compensation for co-movement across stocks. Instead, the investor just demands compensation for holding large quantities of a given factor (since $\lambda > 0$).

Thus, weak substitutability spreads out expected return variation across many factors, including those that explain little comovement across realized returns. That is, weak substitutability gives rise to a factor zoo: many factors are required to fully explain

cross-sectional differences in expected stock returns.

Importantly, weak substitutability does not create near arbitrage opportunities, as proven in Appendix A.3. Indeed, for large enough ψ, the maximum attainable Sharpe ratio *decreases* when ψ decreases. Intuitively, weakening substitutability by lowering γ reduces expected returns for all factors, which lowers the maximum Sharpe ratio.

2.4 Measuring The Proportion of Cross Sectional Expected Return Variation Explained

We can measure the proportion of cross-sectional expected return variation explained by the frist J factors ($\sum_{k=2}^{J} f_k$) 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. Thus, the R^2 from a cross-sectional regression of returns on factor loadings for the first J factors

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

is

$$R^2_{realized}(J) = \frac{\sum_{k=2}^{J} (\psi e_k + 1)^2}{\sum_{k=2}^{K} (\psi e_k e_k + 1)^2 + \sigma_v^2} \tag{11}$$

on average across many realizations of noise demand $\boldsymbol{\pi}$ (as proven in Appendix A.4).

Hence, the ratio of the R^2 from the cross-sectional regression with only the first J factors to the R^2 from the corresponding regression with all K factors identifies $\sum_{k=2}^{J} f_k$:

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

The proportion of cross-sectional expected return variation explained by the first J factors

increases with the degree of substitutability ψ, as formalized in the following proposition (proven in A.4.2).

Proposition 2. *With the eigenvalues in decreasing order* ($e_1 \geq \cdots \geq e_K$), $\frac{R^2_{\text{empirical}}(J)}{R^2_{\text{empirical}}(K)}$ *is increasing in* ψ.

Therefore, as ψ falls and substitutability weakens, more factors are required to explain cross-sectional variation in expected returns.

3 A Simple Calibration of Substitutability

A simple calibration based on micro and macro multiplier estimates from previous work suggests that investors view stocks as weak substitutes.

Consider a simple setting like Proposition 5 of Haddad et al. (2025) where the price multiplier matrix is parameterized simply:[1]

$$\boldsymbol{M} = M_{own}\boldsymbol{I} + M_{cross}(\mathbf{1}\mathbf{1}' - \boldsymbol{I}).$$

where M_{own} is the scalar of own-price multipliers along the diagonal and M_{cross} is the scalar of the off-diagonal cross price terms. We consider this simple calibration to give a rough sense for how large the own (i.e., on-diagonal) and cross (i.e., off-diagonal) elasticity terms would be given literature estimates of price impacts.

Define the micro or relative multiplier (Gabaix and Koijen, 2021; Haddad et al., 2025) as:

$$\widehat{M} = M_{own} - M_{cross}.$$

Haddad et al. (2025) show that in this setting, $\widehat{M}$ is how much prices go up due to a demand shock of 1% of shares outstanding.

Haddad et al. (2025) also show that the aggregate price impact due to a 1% systematic

[1] This setup follows from (2) with one factor ans homogeneous loadings across stocks.

marketwide shock is given by:

$$\overline{M} = M_{own} + (N-1)M_{cross}, \tag{12}$$

where N is the number of stocks. Solving these two equations yields

$$M_{\text{cross}} = \frac{\overline{M} - \widehat{M}}{N}, \tag{13}$$

$$M_{\text{own}} = \frac{(N-1)\,\widehat{M} + \overline{M}}{N}. \tag{14}$$

For the U.S. market, taking $N \approx 2{,}000$, $\widehat{M} \approx 2$, and $\overline{M} \approx 5$ as in Gabaix and Koijen (2021), Davis (2025), and Li and Lin (2022) gives

$$M_{\text{cross}} = \frac{5-2}{2{,}000} \approx 0.0015, \quad M_{\text{own}} = 2 + 0.0015 \approx 2.0015.$$

The implied elasticity matrix is

$$\boldsymbol{M}^{-1} = \frac{1}{\widehat{M}}\,\boldsymbol{I} - \frac{M_{\text{cross}}}{\widehat{M}\cdot\overline{M}}\,\mathbf{1}\mathbf{1}',$$

which yields own-price elasticities around 0.50 and cross-price elasticities around -0.00015.

The large gap between own and cross-price elasticities suggests that the degree of substitutability ψ is small.

Let $\lambda_j^{\text{eig}}(\boldsymbol{M})$ denote the jth eigenvalue of $\boldsymbol{M}$ (ordered largest to smallest). Since $\mathbf{1}\mathbf{1}'$ has eigenvalue N with eigenvector $\mathbf{1}$ and eigenvalue 0 on the $(N-1)$-dimensional orthogonal complement, the eigenvalues of $\boldsymbol{M}$ are

$$\lambda_1^{\text{eig}}(\boldsymbol{M}) = (M_{\text{own}} - M_{\text{cross}}) + NM_{\text{cross}} = \widehat{M} + (\overline{M} - \widehat{M}) = \overline{M}, \qquad \text{(multiplicity 1)},$$

and for $j = 2, \ldots, N$,

$$\lambda_j^{\text{eig}}(\boldsymbol{M}) = (M_{\text{own}} - M_{\text{cross}}) = \widehat{M}, \qquad \text{(multiplicity } N-1\text{)}.$$

Hence,

$$\frac{\lambda_1^{\text{eig}}(\boldsymbol{M})}{\lambda_N^{\text{eig}}(\boldsymbol{M})} = \frac{\overline{M}}{\widehat{M}}.$$

In the model (Section 2), the analogous eigenvalue ratio of the relevant operator is[2]

$$\frac{\lambda_1^{\text{eig}}}{\lambda_N^{\text{eig}}} = \frac{1+\psi e_{\max}}{1+\psi e_{\min}},$$

where $e_{\max}$ and $e_{\min}$ are the largest and smallest eigenvalues used in the R^2 decomposition (under the same normalization). Equating these ratios and solving yields

$$\frac{\overline{M}}{\widehat{M}} = \frac{1+\psi e_{\max}}{1+\psi e_{\min}} \quad \Longrightarrow \quad \psi = \frac{\frac{\overline{M}}{\widehat{M}} - 1}{e_{\max} - \frac{\overline{M}}{\widehat{M}}\, e_{\min}},$$

where $e_{\max}$ and $e_{\min}$ are the largest and smallest eigenvalues respectively. We can simplify this further by using the fact that the smallest eigenvalue accounts for very little variation, i.e. $e_{\min} \approx 0$. Plugging this in yields:

$$\psi = \frac{\frac{\overline{M}}{\widehat{M}} - 1}{e_{\max}}.$$

With $\overline{M} \approx 5$, $\widehat{M} \approx 2$ (so $\overline{M}/\widehat{M} \approx 2.5$) and $e_{\max} = 13.7$, which is the estimate of the largest eigenvalue in the empirical section below, we obtain:

$$\psi \approx \frac{2.5-1}{13.7} = \frac{1.5}{13.7} \approx 0.11,$$

which is similar in magnitude to the holdings-based structural estimate of ψ in Section 5 and the value of ψ estimated from returns.

[2]In the model, the linear map from quantities to expected-return wedges is $\boldsymbol{M}_{\text{model}} \equiv \gamma \Theta E \Theta' + \lambda I$. Since $\Theta_{\cdot,k}$ is an eigenvector with $\boldsymbol{M}_{\text{model}}\Theta_{\cdot,k} = (\lambda + \gamma e_k)\Theta_{\cdot,k}$, the factor-space eigenvalues are $\lambda + \gamma e_k = \lambda(1+\psi e_k)$ with $\psi \equiv \gamma/\lambda$. Hence $\lambda_{\max}^{\text{eig}}(\boldsymbol{M}_{\text{model}})/\lambda_{\min}^{\text{eig}}(\boldsymbol{M}_{\text{model}}) = (1+\psi e_{\max})/(1+\psi e_{\min})$.

4 Reduced-Form Evidence that Spanning the Cross Section of Expected Returns Many Factors

Given the small degree of substitutability (i.e., the small cross-price elasticities) implied by the calibration in Section 3, the model in Section 2 suggests that explaining the cross section of expected returns should require many factors. In this section, we show that this is indeed the case.

At a high level, we fit a factor model to the panel of daily stock returns and recover the covariance matrix eigenvectors (i.e., factor loadings) $\mathbf{\Theta}_t$ as well as the eigenvalues e_k for $k = 1, ..., K$.

Following (10), we calculate the theoretical proportion of cross-sectional expected return variation explained by the first J factors under the assumption of perfect substitutability (as in Kozak et al. (2018)):

$$R^2_{theory}(J) \equiv \sum_{k=1}^{J} \frac{e_k^2}{\sum_{j=1}^{K} e_j^2}.$$

Then, as described in Section 2.4, we measure the empirical proportion of cross-sectional expected return variation explained by the first J factors

$$R^2_{empirical}(J) \equiv \frac{R^2_{realized}(J)}{R^2_{realized}(J)} = \sum_{k=1}^{J} \frac{(\psi e_k + 1)^2}{\sum_{j=1}^{K} (\psi e_j + 1)^2} \tag{15}$$

using predictive cross-sectional regressions of future returns on this period's stock-level loadings $\mathbf{\Theta}_t$.

We show that $\forall J < K$, $R^2_{empirical}(J) \ll R^2_{theory}(J)$. Thus, far more factors are required to span the cross-section of expected returns than would be required if stocks were perfect substitutes.

4.1 Empirical Results

We estimate the factor models in returns using the Instrumented Principal Components Analysis (IPCA) method from Kelly et al. (2019) on daily CRSP stock returns from 1980 onward. We retain firms with at least one year of data and drop the bottom decile of market capitalization each month. The firm-level instrument set consists of an intercept and 120 monthly characteristics taken from Jensen et al. (2023); Appendix A lists the variables. We purposely exclude any characteristic that are a direct function of the stock price to avoid any issues where idiosyncratic returns are may not be orthogonal to prices. 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 using the rank-to-percentile transform—convert to cross-sectional percentiles in $[0, 1]$ and subtract 0.5—so instruments are on a comparable, centered scale with bounded influence. The final instrument vector each month t for each stock i is $Z_{it} = \left[1,\ X_{it}^{(120)}\right]$, where $X_{it}^{(120)}$ is a row vector of the 120 characteristics and the 1 is an intercept term.

We then fit a standard daily IPCA with a fully saturated set of latent factors. Since the goal of the IPCA estimation is to capture the covariance structure of realized returns, we demean returns within each stock. The vector of parameters in the IPCA estimation, denoted as Γ, has 121 rows and 120 columns. Monthly betas are calculated as $\beta_t = Z_t\Gamma$ where Z_t is the matrix of characteristics in a given period.

To map to the model in Section 2, we orthogonalize the estimated betas using the factor covariance:

$$\widetilde{\beta}_t \equiv \beta_t \, \Sigma_F^{1/2},$$

and take the thin SVD of $\widetilde{B}_t$:

$$\widetilde{\beta}_t \;=\; V_t D_t U_t', \qquad V_t'V_t = I, \quad U_t'U_t = I, \quad D_t = \mathrm{diag}(d_{1,t}, \ldots, d_{L,t}).$$

The systematic return covariance matrix implied by IPCA in month t is then

$$\Sigma_t \;\equiv\; B_t \Sigma_F B_t' \;=\; \widetilde{B}_t \widetilde{B}_t', \;=\; V_t D_t^2 V_t'.$$

which identifies the PCA decomposition $\Sigma_t = \Theta_t E_t \Theta_t'$ in the usual way,

$$\Theta_t \equiv V_t, \qquad E_t \equiv D_t^2 = \text{diag}\big(e_{1,t}, \ldots, e_{L,t}\big),$$

Let $V_t^{(J)}$ collect the first J columns of V_t and $e_{k,t}$ be the k-th eigenvalue. Then the model-implied proportion of expected return variation explained by the first J factors under the assumption of perfect substitutability is:

$$R^2_{\text{theory},t}(J) \;\equiv\; \frac{\sum_{k=1}^{J} e_{k,t}^2}{\sum_{j=1}^{L} e_{j,t}^2}$$

For the empirical measure, define the $N \times 1$ vector of average daily returns in month t, $\bar{r}_t$ (computed from *non-demeaned* daily returns). Then $R^2_{\text{empirical},t}(J)$ is the cross-sectional OLS R^2 from regressing $\bar{r}_{t+1}$ on the first J columns of V_t, scaled by the R^2 for the analogous regression using all K columns. We average both the theoretical and empirical R^2 curves across months.

Figure 1 plots the resulting curves. The orange curve ("Theory R^2") rises extremely quickly: since the first few principal components of $\widetilde{B}_m$ explain the vast majority of the cross-sectional variance, if stocks were close substitutes then these few principal components would explain the vast majority of expected returns. By contrast, the blue curve ("Empirical R^2") increases only gradually. With one component, the average empirical R^2 is only around 10%. Dozens of components are required to reach 50%. Only when we approach the full K=120 factors (i.e., allowing an extremely rich set of factors) does the empirical curve near one by construction. In short, expected returns display a *weak factor structure*: many factors are required to explain cross-sectional differences in expected returns even though only a few factors explain most of the comovement in realized returns.

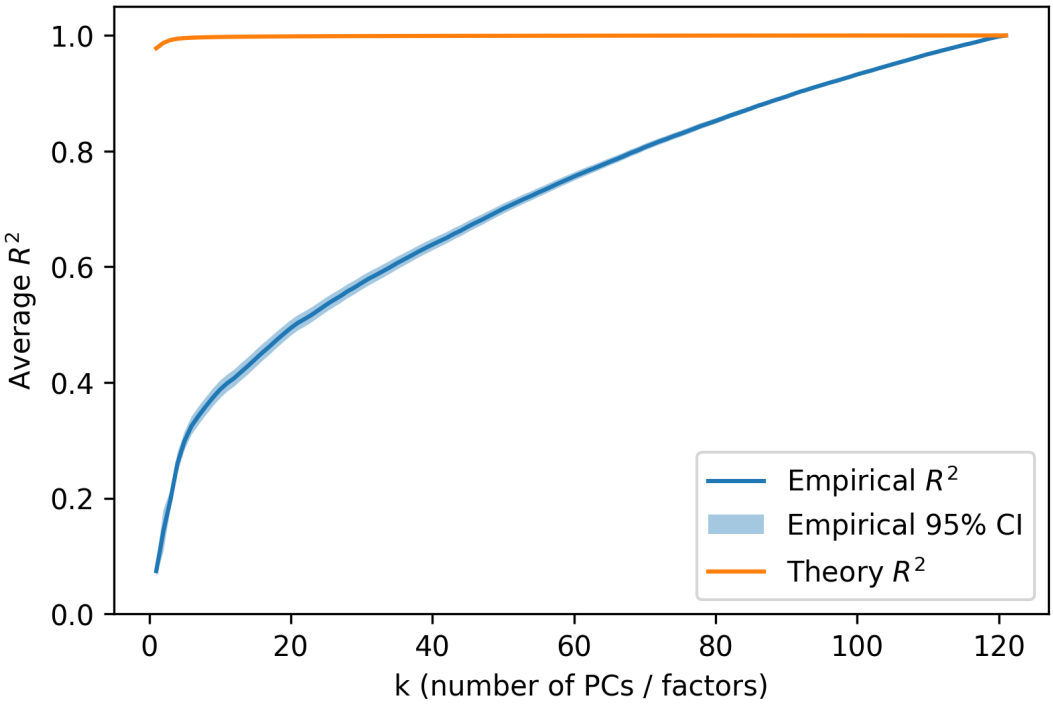


Figure 1: Empirical vs. theory R^2 from IPCA betas. The orange line displays the theoretical R^2 that would prevail under perfect substitutability; the blue line displays the average empirical R^2 across months from the cross-sectional regressions of monthly average returns on the first k principal components.

This gap between the orange and blue curves is precisely what the model in Section 2 implies under limited substitutability: when substitutability is weak (ψ is small), the mapping from comovement to compensation is "flattened," so explaining the cross section of expected returns requires many factors beyond the handful of high-eigenvalue factors.

4.2 Robustness

Out of Sample Estimation We show the results are similar if the IPCA Γ is estimated out-of-sample. Starting from January 1990, we re-estimate Γ every month based using an expanding window from the sample start to the given date. We then form $\beta_t = Z_t\Gamma_t$, where Γ_t is based on data from the beginning of the sample to t. Then β_t is used to form V_t as discussed above, and then the R^2 of regressing $\bar{r}_{t+1}$ on V_t is calculated as described above.

Figure 2 shows the results, which are similar to those from the in-sample analysis and

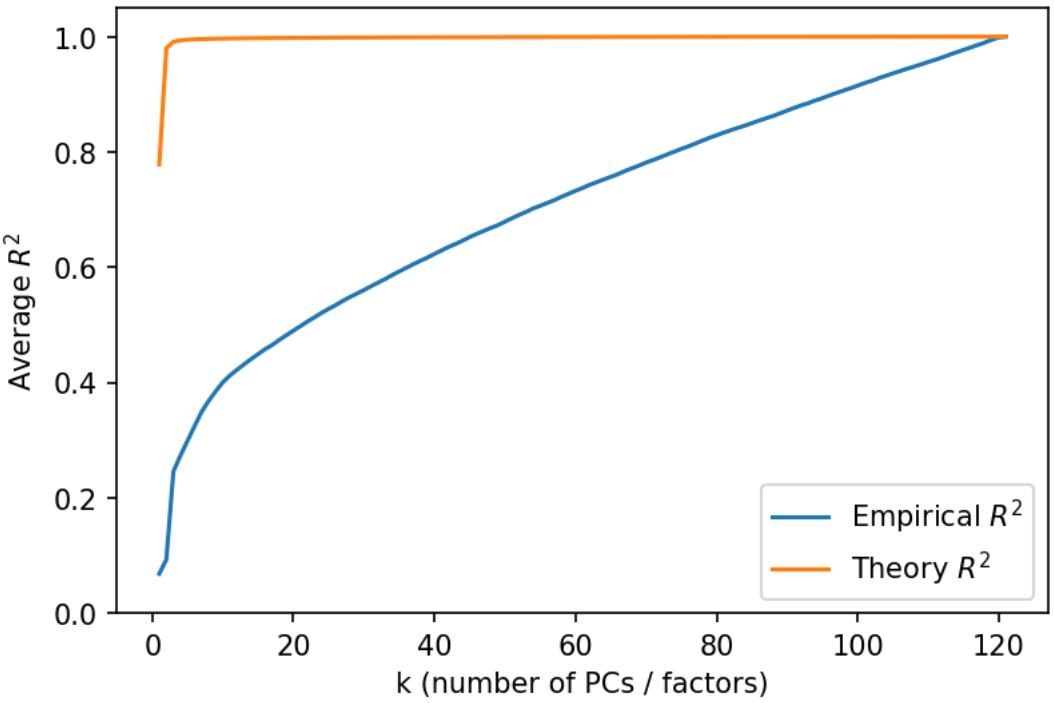


Figure 2: Out of sample empirical vs. theory R^2 from IPCA betas. The orange line displays the theoretical R^2 that would prevail under perfect substitutability; the blue line displays the average empirical R^2 across months from the cross-sectional regressions of monthly average returns on the first k principal components, which are estimated out-of-sample.

are consistent with weak substitutability.

Standard PCA To address the concern that IPCA puts too much structure on the factor loadings, which prevents them from spanning the cross section of expected returns, we repeat the analysis using standard PCA (specifically, we use the deflated heteroskedastic PCA method of Zhou and Chen (2025)).

Figure 3 demonstrates that the difference $R^2_{theory}(J) - R^2_{empirical}(J;t)$ remains large.

Factor Portfolios To address the concern that the unbalanced nature of the panel of daily stock returns leads to measurement error in the estimated factor loadings that prevents them from spanning the cross section of expected returns, we repeat the analysis using daily returns for the 56 characteristics-sorted portfolios from Kozak et al. (2018) (market portfolio and 55 anomaly portfolios). In particular, we apply the deflated heteroskedastic PCA method of Zhou and Chen (2025) instead of IPCA.

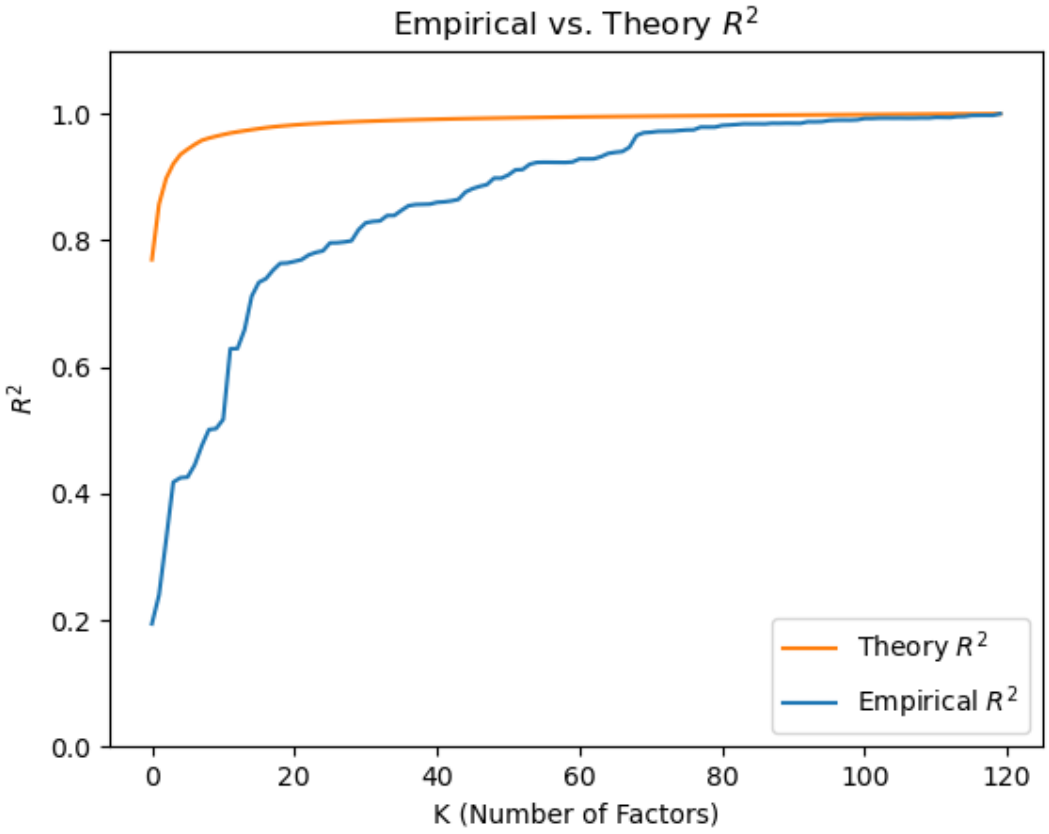


Figure 3: Standard PCA on daily CRSP stock returns.

Figure 4 Panel A displays that, until J becomes pretty large, $R^2_{empirical}(J;t) \ll R^2_{theory}(J)$ for the Kozak et al. (2018) portfolios. These portfolios have a strong factor structure in realized returns. Thus, in theory if there was perfect substitutability, it would take only 3 factors to explain 95% of the cross sectional variance in expected returns. In contrast, in reality it takes 19 factors to explain 95% of the variance.

5 Structural Estimation of Cross-Price Elasticities

In this section we estimate cross-price elasticities in institutional investor holdings data from the SEC form 13F. We find that cross-price elasticities are small, which means investors only weakly substitute across stocks. As established by the model in Section 2, this weak substitutability rationalizes why many factors are required to explain the cross section of expected returns even though only a few factors suffice to explain covariances.

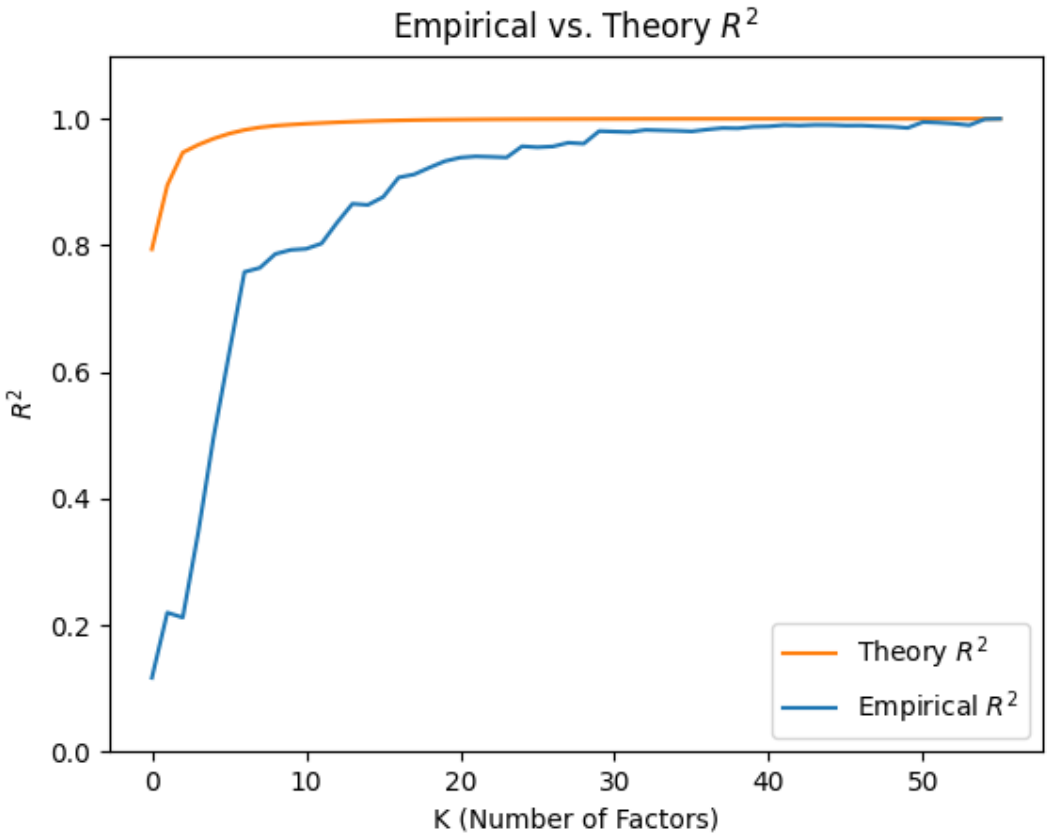


Figure 4: Standard PCA on daily returns on 56 Kozak et al. (2018) characteristic-sorted portfolios

5.1 Empirical Specification

We start with the portfolio choice expression (1) from the model in Section 2. To better match the data, we now express the model in terms of portfolio weights instead of dollar demand. The portfolio weight demanded by investor i in stock n in quarter t is

$$w_{i,n,t} = \frac{1}{\lambda_{i,t}}\mu_{i,n,t} - \frac{1}{\lambda_i}\sum_{m=1}^{N}\left(\sum_{k=1}^{K}\theta_{n,k}\theta_{m,k}\frac{\psi_i e_k}{1+\psi_i e_k}\right)\mu_{i,m,t}$$

where $\mu_{i,n,t}$ is the investor's subjective expected return.

Empirically, we work with percentage changes in the quantity of shares held instead

of with the level of portfolio weights:

$$\Delta q_{i,n,t} = \frac{1}{\lambda_i}\Delta\mu_{i,n,t} - \frac{1}{\lambda_i}\sum_{m=1}^{N}\left(\sum_{k=1}^{K}\theta_{n,k,t}\theta_{m,k,t}\frac{\psi_i e_k}{1+\psi_i e_k}\right)\Delta\mu_{i,m,t} \tag{16}$$

Linearizing the expected returns around small percentage changes in prices yields the following demand curve

$$\Delta q_{i,n,t} = -\alpha_i \Delta p_{n,t} + \alpha_i \sum_{m=1}^{N}\sum_{k=1}^{K}\left(\frac{\psi_i e_k}{1+\psi_i e_k}\theta_{n,k,t}\theta_{m,k,t}\right)\Delta p_{m,t} + \sum_{k=1}^{K} c_{i,k}\theta_{n,k,t} + \sum_{l=1}^{L} d_{i,l}X_{n,l,t} + \text{`}u_{i,n,t} \tag{17}$$

where $\Delta p_{n,t}$ is the percentage price change.

The first term in (17) captures the reduction in quantity demanded for stock n when price ceteris paribus rises, and so expected return falls. α is the micro or relative elasticity (Gabaix and Koijen (2021); Haddad et al. (2025)) and depends on how sensitive the investor is to changes in expected returns (governed by λ_i in Section 2) and how persistent she views price changes to be (i.e., how much price changes alter the one-period expected return).

The second term captures substitution. When the price of different stock m falls and its expected return rises, the investor will shift demand away from stock n and toward stock m. This substitution is stronger when stocks n and m share similar factor loadings ($\theta_{n,k,t}\theta_{m,k,t}$ is big), especially for the factors that explain more comovement (the high eigenvalue factors, which have large $\frac{\psi_i e_k}{1+\psi_i e_k}$). Moreover, the overall strength of this substitution is controlled by ψ.

The next three terms capture shifts in the demand curve. In our microfoundation from Section 2, these shifts arise from either changes in this investors' (subjective) cash flow expectations or from the noise-trader demand shocks. We model these demand shifts as depending factor loadings $\theta_{n,k,t}$ as the noise-trader demand shocks do in Section 2. Moreover, we allow the demand shifts to depend on L latent stock characteristics $X_{n,l,t}$ to

capture the factor structure in expected cash flows.[3]. We use $L = 10$ latent characteristics in the baseline specification, but the results are robust to alternative choices of L. We also allow for an uncorrelated residual shift $u_{i,n,t}$.

For estimation, it is convenient to group the cross-price terms at the factor level:

$$\Delta q_{n,t} = -\alpha_i \Delta p_{n,t} + \sum_{k=1}^{K} \underbrace{b_{i,k}}_{\equiv \alpha_i \frac{\psi_i e_k}{1+\psi_i e_k}} \theta_{n,k,t} \underbrace{\Delta p^F_{k,t}}_{\equiv \sum_{m=1}^{N} \theta_{m,k,t} \Delta p_{m,t}} + \sum_{k=1}^{K} c_{i,k} \theta_{n,k,t} + \sum_{l=1}^{L} d_{i,l} X_{n,l,t} + u_{i,n,t} \tag{18}$$

where $\Delta p^F_{k,t}$ is the percentage change in the price of factor k. Since the eigenvalues e_k are known, identifying both α_i and $b_{i,k}$ suffices to identify the degree of substitutability ψ_i. Intuitively, if the cross-price elasticities are small relative to the micro elasticity (i.e., $b_{i,k}$ is small relative to α_i), then substitutability is weak (i.e., ψ_i is small).

5.2 Identification

A regression of quantity changes on price changes does not identify $\alpha_{i,k}$ or $b_{i,k}$ because the investor's residual demand shock $u_{i,n,t}$ in (18) correlates with the price changes she faces through market clearing. Thus, to obtain exogenous variation in prices, we use the optimal granular instrumental variables methodology (GIV) of Chaudhary et al. (2024), which extends the granular instrumental variables methodology of Gabaix and Koijen (2021). At a high level, we use the uncorrelated residual demand shifts of other investors $j \neq i$ to instrument for the price changes investor i faces.

Identifying α_i and $b_{i,k}$ We assume the residual demand shocks are uncorrelated across investors conditional on the vectors of factor loadings $\boldsymbol{\theta}_{n,t}$ and latent characteristics $\boldsymbol{X}_{n,t}$

$$\forall i \neq j, \forall t, : \mathbb{E}^{CX}[u_{i,n,t} u_{j,n,t} \mid \boldsymbol{\theta}_{n,t}, \boldsymbol{X}_{n,t}] = \mathbf{0}, \tag{19}$$

[3]We estimate these latent characteristics using PCA on the pooled panel $\Delta q_{i,(n,t)}$

where $\mathbb{E}^{CX}$ indicates the cross-sectional expectation taken across stocks n. In this case, we can construct the GIV for the price change investor i faces for stock n:

$$z_{i,n,t} = \sum_{j \neq i} S_{j,n,t-1} u_{j,n,t}. \tag{20}$$

$S_{i,n,t-1}$ are lagged ownership share weights — the proportion of outstanding shares held by investor i. From these stock-level GIVs, we can construct factor-level GIVs interacted with the stock-level factor loadings:

$$Z^{(k)}_{i,n,t} = \theta_{n,k,t} \sum_{m} \theta_{m,k,t} z_{i,m,t}.$$

With these instruments, we can identify the $K+1$ parameters of interest α_i and all $b_{i,k}$ from the following $K+1$ moment conditions

$$\mathbb{E}\left[z_{i,n,t} u_{i,n,t} \mid \boldsymbol{\theta}_{n,t}, \boldsymbol{X}_{n,t}\right] = 0 \tag{21}$$

$$\mathbb{E}\left[Z^{(k)}_{i,n,t} u_{i,n,t} \mid \boldsymbol{\theta}_{n,t}, \boldsymbol{X}_{n,t}\right] = 0, \forall k = 1, \ldots, K \tag{22}$$

These moment conditions pool variation across stocks n and time periods t.

Identifying ψ_i Having identified α_i and $b_{i,k}$, ψ_i can be identified using the functional form of $b_{i,k}$ from (18): $b_{i,k} = \alpha_i \psi_i e_k / (1 + \psi e_k)]$. Given estimates $\hat{\alpha}_i$ and $\hat{b}_{i,k}$, and eigenvalues e_k from the IPCA estimation in Section 4, this moment condition identifies ψ_i:

$$\mathbb{E}\left[\frac{\hat{b}_{i,k}}{\hat{\alpha}_i} - \frac{\psi_i e_k}{1 + \psi e_k}\right] = 0 \tag{23}$$

The expectation is taken across k.

5.3 Empirical Implementation

We estimate moment conditions (21), (22), and (23) on stock-level U.S. institutional investor holdings data from the SEC form 13F (from 1980 to 2021), which are provided

by Thomson Reuters 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 allocate all stock holdings not covered by 13F institutions to a residual "household" sector, which includes both direct stock holdings by households and those by non-13F institutions (i.e., institutions with less than $100 million AUM).

For the factor loadings $\theta_{n,k,t}$, we use IPCA loadings on 120 latent factors, estimated as in Section 2.[4]

To preserve estimation efficiency, we group factors with similar eigenvalues together. That is, we do not estimate separate $b_{i,k}$ coefficients for the 120 different factors, but rather group the K factors to H groups that minimize the average within-group standard deviation of the log eigenvalues. We thus estimate:

$$\Delta q_{i,n,t} = -\alpha_i \Delta p_{n,t} + \sum_{h=1}^{H} b_{i,h} \sum_{k \in \text{Group}_h} \theta_{n,k} \Delta p^F_{k,t} + u_{i,n,t} \tag{24}$$

The baseline estimation uses $H = 2$, but our results are robust to alternative choices of H as well.

Moreover, to preserve estimation efficiency, we group investors together. We consider two specifications: a baseline specification in which α_i and $b_{i,h}$ are homogeneous across investors, and a specification in these parameters are affine functions of investor i's active share (demeaned within each quarter). For example, in the second specification we model

$$\alpha_{i,t} = \alpha_0 + \alpha_{AS} \cdot \text{Active Share}_{i,t-1}$$

[4]Since the IPCA factor loadings are not orthonormal as assumed in the model in Section 2, moment condition (23) requires a slight adjustment. Specifically, we take the K non-zero eigenvalues of the cross-price elasticity matrix $\boldsymbol{\zeta}_i = \boldsymbol{\Theta}' \text{diag}[b_{i,k}]_{k=1}^K \boldsymbol{\Theta}$, denoted $\rho_{i,k}$. Per the model Section 2, the eigendecomposition of this matrix is $\boldsymbol{\zeta}_i = \tilde{\boldsymbol{\Theta}}' \mathbf{P}_i \tilde{\boldsymbol{\Theta}}$, where $\tilde{\boldsymbol{\Theta}}$ are orthonormal and $\mathbf{P}_i$ is a diagonal matrix with non-zero eigenvalues $\rho_{i,k} = \alpha_i \frac{\psi_i e_k}{1+\psi_i e_k}$. We then use the following moment condition in place of (23) to identify ψ_i:

$$\mathbb{E}\left[\frac{\rho_{i,k}}{\alpha_i} - \frac{\psi_i e_k}{1+\psi_i e_k}\right] = 0.$$

The active share for institution i is defined as $\frac{\sum_n |w_{i,n,t} - w_{i,n,t}^{\text{benchmark}}|}{2}$, where $w_{i,n,t}$ and $w_{i,n,t}^{\text{benchmark}}$ are the actual and benchmark portfolio weights (Cremers and Petajisto (2009)). We proxy for the latter using the market-cap weights for each institution's investment universe.[5]

5.4 Empirical Results

5.4.1 Baseline Results

Table 1 displays the results of our baseline estimation of demand curve (24).

α	b_1	b_2	ψ
0.086^{***}	$1.399 \times 10^{-4**}$	$5.565 \times 10^{-5***}$	0.226^{***}
(0.012)	(6.047×10^{-5})	(1.080×10^{-5})	(0.021)

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 1: This table displays the results of the own- and cross-price elasticity estimation in institutional holdings data when imposing investor homogeneity.

The micro elasticity α is small, as documented in previous work: $\alpha = 0.086$, which means a 1% ceteris paribus rise in the price of stock n leads investors to lower their quantity of shares demanded by only 0.086%.

The cross-price coefficients b_h are positive, which implies that investors do substitute toward stock n when the prices of other stocks rise. However, the degree of this substitution is weak. Figure 5 displays the distribution of cross-price elasticities across all stocks and quarters. The mean cross-price elasticity is 0.000038, which means a 1% rise in the price of stock m leads investors to raise their demand for stock n by only 0.000038% — four orders of magnitude less than the own price reaction. Even the most extreme cross-price elasticities are two orders of magnitude smaller than the own price micro elasticity. Note that these estimated heterogeneous cross-price elasticities are of the same order of magnitude as the calibrated homogeneous cross-price elasticity in Section 3.

We map these small cross-price elasticities back to the model in Section 2 by estimating

[5] In particular, we the four-quarter rolling average of active share up to time $t-1$. Following Koijen and Yogo (2019), an institution's universe is all stocks held in the past 11 quarters.

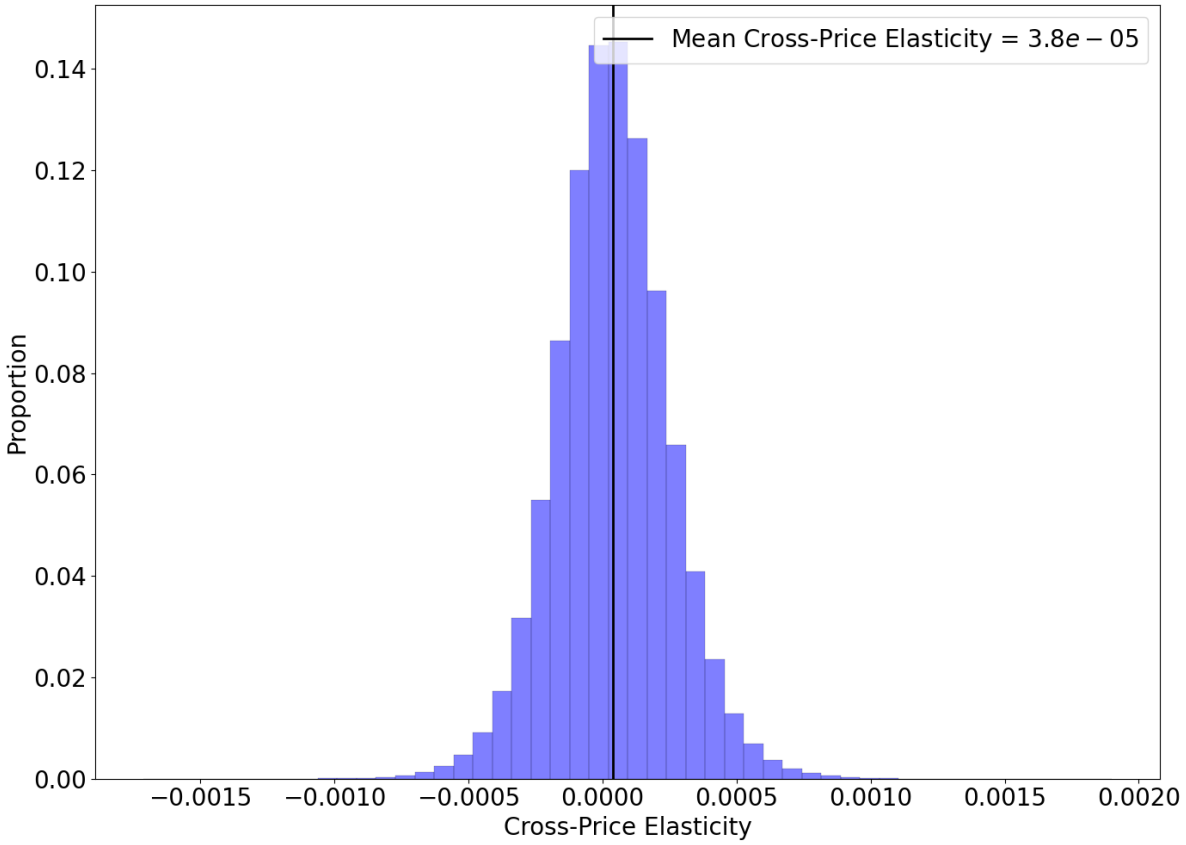


Figure 5: Distribution of cross-price elasticities across all stocks and quarters.

ψ. We find ψ is small: $\psi = 0.226$, which means empirically investors only weakly substitute across stocks.

5.4.2 Investor Heterogeneity

The baseline results in Table 1 imply that, overall, investors weakly substitute across stocks. We now study how the degree of substitutability differs across investors. To parsimoniously capture heterogeneity across investors, we parameterize α_i and each $b_{i,h}$ as an affine function of investor i's active share.

Table 2 displays the results of estimating demand curve (24) when accounting for investor heterogeneity.

Panel A: Reduced-Form Coefficients

α_0	α_{AS}	$b_{1,0}$	$b_{1,AS}$	$b_{2,0}$	$b_{2,AS}$
0.146^{***}	0.385^{*}	-0.002	-0.004	$4.535 \times 10^{-4**}$	0.001^{*}
(0.050)	(0.228)	(0.001)	(0.004)	(2.121×10^{-4})	(0.001)

Panel B: Micro Elasticity by Active-Share Percentile

10th	25th	50th	75th	90th
0.045^{***}	0.081^{***}	0.124^{***}	0.172^{***}	0.216^{**}
(0.015)	(0.014)	(0.037)	(0.065)	(0.091)

Panel C: ψ by Active-Share Percentile

10th	25th	50th	75th	90th
0.223	0.361	0.528**	0.717***	0.892***
(0.226)	(0.236)	(0.246)	(0.248)	(0.329)

Table 2: This table displays the results of the own- and cross-price elasticity estimation in institutional holdings data when accounting for heterogeneity across investors by active share.

We find that micro elasticities increase with active share but remain small across the active-share distribution. The estimated α_1 and α_{AS} parameters yield a range of micro elasticities from 0.045 to 0.216 from the 10^{th} to the 90^{th} percentile of active share.

Moreover, we find that the degree of substitutability ψ_i increases with active share, but also remains small across the active-share distribution. The estimated $b_{h,0}$ and $b_{h,AS}$

parameters yield a range of ψ_i from 0.22 to 0.89 from the 10^{th} to the 90^{th} percentile of active share. All of these ψ_i are small and are consistent with investors only weakly substitute across stocks.

We redo the R^2 plot from Figure 1 and overlay the proportion of expected return variation explained by the first J factors implied by the model in Section 2 and our estimated ψ:

$$R^2_\psi(J) = \sum_{k=1}^{J} \frac{(\psi e_k + 1)^2}{\sum_{j=1}^{K} (\psi e_j + 1)^2}.$$

In particular, we use the median ψ of 0.528. We form a 95% confidence interval based on the standard errors from Table 2. Figure 6 shows the results. The R^2_ψ aligns with the empirical R^2 estimated from returns, and is statistically significantly below the theoretical R^2 that would prevail under perfect substitutability.

In summary, the calibrated ψ from Section 3, the ψ implied by the return regressions from Section 4, and the ψ estimated from holdings data are all quantitatively similar and indicate that investors treat stocks as weak substitutes. This weak substitutability gives rise the factor zoo in the cross section of expected stock returns.

6 Conclusion

We argue that weak substitutability 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 arbitrageur faces an ℓ_2 penalty on portfolio positions, which weakens cross-price elasticities without generating near-arbitrage opportunities. When substitutability is strong, the factors that explain most return comovement also explain most of the cross-sectional variation in expected returns. As substitutability weakens, expected returns disperse across many directions in factor space, including those associated with small eigenvalues of the covariance matrix. The model thus maps the strength of cross-price elasticities into the strength of the factor structure in expected returns and shows how a genuine factor zoo arises in equilibrium.

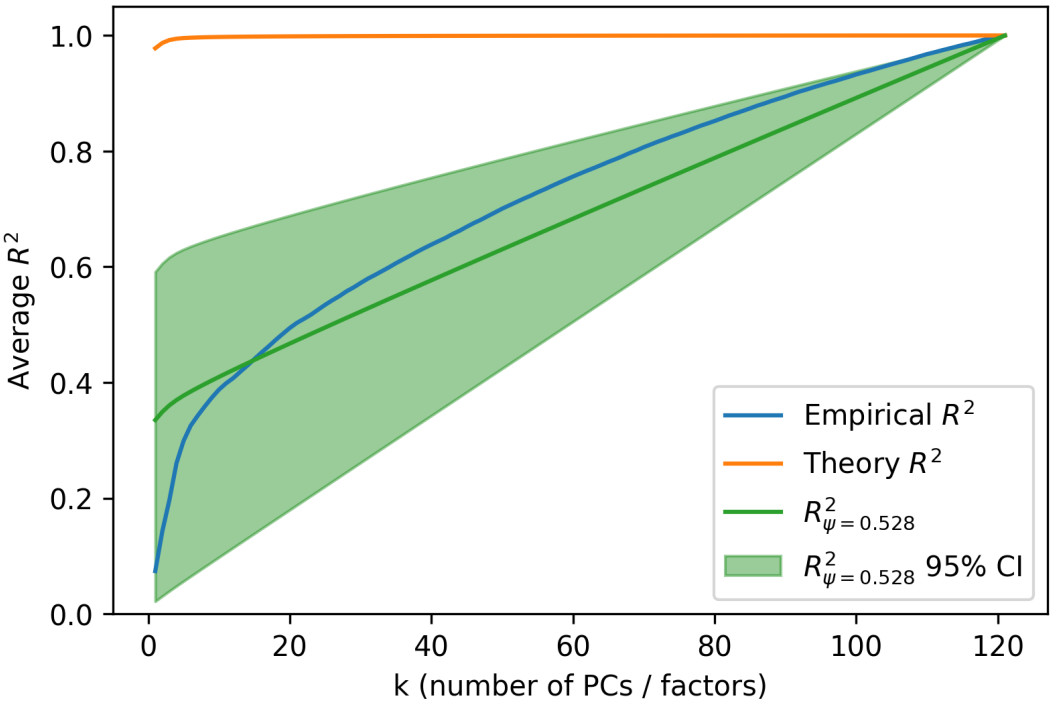


Figure 6: Empirical, theory, and ψ-specific R^2 from IPCA betas. The orange line cumulates s_k^4 ("theory" under perfect substitutability); the blue line cumulates fitted-value variance from projecting monthly average returns onto the first k orthonormal beta directions.

Our empirical work shows that the U.S. equity market lies firmly in the weak-substitutability region described by the model. A simple calibration based on micro and macro price multiplier estimates implies very small cross-price elasticities, indicating that investors view stocks as only weak substitutes. Reduced-form evidence from IPCA and standard PCA on both individual stock returns and characteristic-sorted portfolios confirms that realized returns display a strong factor structure while expected returns do not. While only a few factors are required to explain comovement across stocks, many factors are needed to explain the cross section of expected returns, far more than theory would predict under strong substitutability. Finally, our structural estimation using institutional holdings data directly recovers small cross-price elasticities, reinforcing the insight from the calibration and reduced-form exercises.

These findings have implications for how asset-pricing models are specified and how empirical evidence on anomalies is interpreted. Models that impose limited substitutabil-

ity or coarse substitution patterns, which might seem overly restrictive under the classical Markowitz benchmark, appear well aligned with the data once weak cross-price elasticities are 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

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

Chaudhary, Manav, Zhiyu Fu, and Haonan Zhou, 2024, Anatomy of the treasury market: Who moves yields?, *Available at SSRN* .

Cremers, KJ Martijn, and Antti Petajisto, 2009, How active is your fund manager? a new measure that predicts performance, *The review of financial studies* 22, 3329–3365.

Davis, Carter, 2025, Classical versus demand system asset pricing, *Available at SSRN 5232345* .

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

Gabaix, Xavier, and Ralph S.J. Koijen, 2021, In search of the origins of financial fluctuations: The inelastic markets hypothesis.

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

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

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.

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, 2018, 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.

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

Markowitz, Harry M., 1952, Portfolio selection, *The Journal of Finance* 7, 77–91.

Scholes, Myron S, 1972, The market for securities: Substitution versus price pressure and the effects of information on share prices, *The Journal of Business* 45, 179–211.

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

A Proofs for Section 2

A.1 Proof of (5)

The average own-price elasticity is

$$\begin{aligned}
\frac{1}{N}\sum_{n=1}^{N}\zeta_n &= -\frac{1}{\lambda}\frac{1}{N}\left(1-\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\theta_{n,k}^2\frac{\gamma e_k}{\lambda+\gamma e_k}\right)\\
&= -\frac{1}{\lambda}\left(1-\frac{1}{N}\sum_{k=1}^{K}\frac{\gamma e_k}{\lambda+\gamma e_k}\sum_{n=1}^{N}\theta_{n,k}^2\right)\\
&= -\frac{1}{\lambda}\left(1-\frac{K}{N}\frac{1}{K}\sum_{k=1}^{K}\frac{\gamma e_k}{\lambda+\gamma e_k}\right)\\
&\equiv -\frac{1}{\lambda}\left(1-\phi\mathbb{E}^k\left[\frac{te_k}{1+te_k}\right]\right)
\end{aligned}$$

where the third line follows from $\sum_{n=1}^{N}\theta_{n,k}^2 = 1$ by the orthonormality of $\boldsymbol{\Theta}$. The expectation $\mathbb{E}^k\left[\frac{1}{\frac{\lambda}{\gamma\cdot e_k}+1}\right]$ in the last line is taken across $k = 1, .., K$.

A.2 Proofs for Section 2.3

A.2.1 Proof of (7)

$$\begin{aligned}
\mathbb{E}[\boldsymbol{D}] - \boldsymbol{P} &= \left[\gamma\boldsymbol{\Theta}\mathbf{E}\boldsymbol{\Theta}' + \lambda\mathbf{I}\right]\left(\frac{1}{N}\boldsymbol{\iota} - \boldsymbol{\Theta}\boldsymbol{\pi}\right) \\
&\xrightarrow{N\to\infty} -\left[\gamma\boldsymbol{\Theta}\mathbf{E}\boldsymbol{\Theta}' + \lambda\mathbf{I}\right]\boldsymbol{\Theta}\boldsymbol{\pi} \\
&= -\lambda(t)\boldsymbol{\Theta}\left[t\mathbf{E} + \mathbf{I}\right]\boldsymbol{\pi} \\
&= -\boldsymbol{\Theta}\underbrace{\left[\lambda(t)\pi_k\left(te_k+1\right)\right]_{k=1}^{K}}_{\equiv\boldsymbol{\nu}},
\end{aligned} \tag{25}$$

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

A.2.2 Proof of (8)

Total cross-sectional variation in expected returns is given by:

$$\begin{aligned}
\mathbb{V}[\mathbb{E}[D_n] - P_n] &= \mathbb{E}[(\mathbb{E}[D_n] - P_n)^2] - (\mathbb{E}[\mathbb{E}[D_n] - P_n])^2 \\
&= \boldsymbol{\nu}'\boldsymbol{\Theta}'\boldsymbol{\Theta}\boldsymbol{\nu} - \left(-\frac{1}{N}\boldsymbol{\iota}'\boldsymbol{\Theta}\boldsymbol{\nu}\right)^2 \\
&= \boldsymbol{\nu}'\boldsymbol{\nu} - \left(-\lambda(t)\pi_1\left(te_1+1\right)\right)^2 \\
&= \lambda(t)^2\sum_{k=2}^{K}\pi_k^2\left(te_k+1\right)^2
\end{aligned}$$

where 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 $\boldsymbol{\iota}'\boldsymbol{\Theta}_{k,\cdot} = 0$.

A.2.3 Proof of (9)

The proportion of cross-sectional expected return variance explained by (exposures to) the k-th factor is:

$$\tilde{f}_k \equiv \frac{\pi_k^2 (\psi e_k + 1)^2}{\sum_{j=2}^{K} \pi_j^2 (\psi e_j + 1)^2}. \tag{26}$$

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 (26) around $\pi_k = \mathbb{E}[\pi_k^2], \forall k$ to obtain:

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

Taking the expectation yields (9).

A.2.4 Proof of Proposition 1

Let $e_k > e_j$. Then

$$\frac{f_k}{f_J} = \left(\frac{\psi e_k + 1}{\psi e_j + 1}\right)^2.$$

Since $(\psi e_k + 1)/(\psi e_j + 1) > 0$, it suffices to show that

$$\frac{\partial}{\partial \psi} \frac{\psi e_k + 1}{\psi e_j + 1} > 0.$$

Note that

$$\begin{aligned}\frac{\partial}{\partial \psi} \frac{\psi e_k + 1}{\psi e_j + 1} &= \frac{e_k - e_j}{(\psi e_j + 1)^2} \\ &> 0\end{aligned}$$

since $e_k > e_j$.

A.3 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

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

where the fourth line follows from the matrix expression for equilibrium expected returns (25) from Appendix A.2.1. The fifth line uses the orthonormality of $\boldsymbol{\Theta}$ and that fact that

$$\begin{aligned}
\boldsymbol{\pi}'\boldsymbol{\Theta}'\Sigma^{-1}\boldsymbol{\pi}\boldsymbol{\Theta} &= \boldsymbol{\pi}'\boldsymbol{\Theta}'\left[\boldsymbol{\Theta}\mathbf{E}\boldsymbol{\Theta}'\right]^{-1}\boldsymbol{\Theta}\boldsymbol{\pi} \\
&= \boldsymbol{\pi}'\boldsymbol{\Theta}'\boldsymbol{\Theta}\mathbf{E}^{-1}\boldsymbol{\Theta}'\boldsymbol{\Theta}\boldsymbol{\pi} \\
&= \boldsymbol{\pi}'\mathbf{E}^{-1}\boldsymbol{\pi}
\end{aligned}$$

where the second line follows by the orthonormality of $\boldsymbol{\Theta}$ and noting that the inverse and pseudo-inverse deliver identical values in the span of $\boldsymbol{\Theta}$ ($\left[\boldsymbol{\Theta}\mathbf{E}\boldsymbol{\Theta}'\right]$ is singular for $K < N$).

Thus, we have

$$\begin{aligned} SR^2 &= \sum_k \pi_k^2 \left(\gamma^2 e_k + 2\gamma\lambda + \lambda^2 e_k^{-1}\right) \\ &= \lambda(\psi)^2 \sum_k \pi_k^2 \left(\psi^2 e_k + 2\psi + e_k^{-1}\right) \\ &= \left[\frac{1 - \phi\mathbb{E}^k\left[\frac{\psi e_k}{1+\psi e_k}\right]}{-\bar{\zeta}}\right]^2 \sum_k \pi_k^2 \left(\psi^2 e_k + 2\psi + e_k^{-1}\right) \end{aligned}$$

where the last line follows from (6).

To evaluate how the maximum Sharpe ratio varies with the degree of substitutability, we differentiate with respect to ψ:

$$\frac{\partial}{\partial\psi} SR^2 = \underbrace{\frac{1 - \phi\mathbb{E}^k\left[\frac{\psi e_k}{1+\psi e_k}\right]}{-\bar{\zeta}} \sum_k \pi_k^2 \left(2\psi e_k + 2\right)}_{>0 \text{ and } \xrightarrow{\psi\to\infty} \infty} - \underbrace{\frac{\phi}{\bar{\zeta}} \mathbb{E}^k\left[\frac{e_k}{\left(1+\psi e_k\right)^2}\right] \sum_k \pi_k^2 \left(\psi^2 e_k + 2\psi + e_k^{-1}\right)}_{>0}$$

Asymptotically, the first term dominates. So for large enough ψ, $\frac{\partial}{\partial\psi} SR^2 > 0$, and the maximum Sharpe ratio decreases if substitutability weakens (i.e., if ψ decreases).

A.4 Measuring Proportion of Cross Sectional Expected Return Variation Explained

A.4.1 Proof of (11)

We can measure the proportion of cross-sectional expected return variation is 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 observe 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

(8)) is:

$$\begin{aligned}\mathbb{V}[D_n - P_n] &= \mathbb{V}[\mathbb{E}[D_n] - P_n] + \mathbb{V}[v_n] \\ &= \lambda(t)^2 \sum_{k=2}^{K} \pi_k^2 \left(te_k + 1\right)^2 + \sigma_v^2.\end{aligned}$$

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

$$\begin{aligned}R^2_{realized}(J) &= \frac{\lambda(t)^2 \sum_{k=2}^{J} \pi_k^2 \left(te_k + 1\right)^2}{\lambda(t)^2 \sum_{k=2}^{K} \pi_k^2 \left(te_k + 1\right)^2 + \sigma_v^2} \\ &\approx \frac{\lambda(t)^2 \mathbb{E}\left[\pi_k^2\right] \sum_{k=2}^{J} \left(te_k + 1\right)^2}{\lambda(t)^2 \mathbb{E}\left[\pi_k^2\right] \sum_{k=2}^{K} \pi_k^2 \left(te_k + 1\right)^2 + \sigma_v^2} + \frac{\lambda(t)^2 \mathbb{E}\left[\pi_k^2\right] \sum_{k=2}^{J} \left(te_k + 1\right)^2}{\lambda(t)^2 \mathbb{E}\left[\pi_k^2\right] \sum_{k=2}^{K} \pi_k^2 \left(te_k + 1\right)^2 + \sigma_v^2} \left(\pi_k^2 - \mathbb{E}\left[\pi_k^2\right]\right) \\ &\quad - \sum_{j=2}^{K} \frac{\lambda(t)^2 \mathbb{E}\left[\pi_k^2\right] \sum_{k=2}^{J} \left(te_k + 1\right)^2 \left(te_j + 1\right)^2}{\left[\lambda(t)^2 \mathbb{E}\left[\pi_k^2\right] \sum_{k=2}^{K} \left(te_k + 1\right)^2 + \sigma_v^2\right]^2} \left(\pi_j^2 - \mathbb{E}\left[\pi_k^2\right]\right) \\ &= \frac{\sum_{k=2}^{J} \left(te_k + 1\right)^2}{\sum_{k=2}^{K} \left(te_k + 1\right)^2 + \frac{\sigma_v^2}{\lambda(t)^2 \mathbb{E}\left[\pi_k^2\right]}} + \frac{\sum_{k=2}^{J} \left(te_k + 1\right)^2}{\sum_{k=2}^{K} \left(te_k + 1\right)^2 + \frac{\sigma_v^2}{\lambda(t)^2 \mathbb{E}\left[\pi_k^2\right]}} \left(\pi_k^2 - \mathbb{E}\left[\pi_k^2\right]\right) \\ &\quad - \sum_{j=2}^{K} \frac{\sum_{k=2}^{J} \left(te_k + 1\right)^2 \left(te_j + 1\right)^2}{\left[\sum_{k=2}^{K} \left(te_k + 1\right)^2 + \frac{\sigma_v^2}{\lambda(t)^2 \mathbb{E}\left[\pi_k^2\right]}\right]^2} \left(\pi_j^2 - \mathbb{E}\left[\pi_k^2\right]\right),\end{aligned}$$

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

$$R^2_{realized}(J) = \frac{\sum_{k=2}^{J} \left(te_k + 1\right)^2}{\sum_{k=2}^{K} \left(te_k + 1\right)^2 + \frac{\sigma_v^2}{\lambda(t)^2 \mathbb{E}\left[\pi_k^2\right]}}.$$

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

$$\begin{aligned}\frac{R^2_{realized}(J;t)}{R^2_{realized}(K;t)} &= \frac{\sum_{k=2}^{J}(te_k+1)^2}{\sum_{k=2}^{K}(te_k+1)^2} \\ &= \sum_{k=2}^{J} f_k.\end{aligned}$$

A.4.2 Proof of Proposition 2

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 Proposition 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 $k = 2, \ldots, J$ gives

$$\begin{aligned}&\frac{\sum_{k=2}^{J} f_k(\psi_2)}{f_m(\psi_2)} > \frac{\sum_{k=2}^{J} f_k(\psi_1)}{f_m(\psi_1)} \\ \leftrightarrow\ &\frac{\sum_{k=2}^{J} f_k(\psi_2)}{f_m(\psi_2)} \cdot \frac{f_m(\psi_2)}{\sum_{k=2}^{K} f_k(\psi_2)} > \frac{\sum_{k=2}^{J} f_k(\psi_1)}{f_m(\psi_1)} \cdot \frac{f_m(\psi_2)}{\sum_{k=2}^{K} f_k(\psi_2)} \\ \leftrightarrow\ &\frac{\sum_{k=2}^{J} f_k(\psi_2)}{\sum_{k=2}^{K} f_k(\psi_2)} > \frac{\sum_{k=2}^{J} f_k(\psi_1)}{\sum_{k=2}^{K} f_k(\psi_2)} \\ \leftrightarrow\ &\left.\frac{R^2(J)}{R^2(K)}\right|_{\psi_1} > \left.\frac{R^2(J)}{R^2(K)}\right|_{\psi_1},\end{aligned}$$

as desired.