Idiosyncratic Risk Premium

Junxiong Gao, Jun Liu*

December 11, 2025

Abstract

The market portfolio of U.S. stocks is highly concentrated, with market capitalization following a fat-tailed distribution. Under the Arbitrage Pricing Theory (Ross, 1976), the risk premium associated with idiosyncratic risk—the idiosyncratic risk premium (IRP)—is bounded—below and above—by the degree of market concentration, and non-zero if investors hold concentrated portfolios. We build a model to explicitly compute the IRP and show that it increases with both idiosyncratic volatility (IVol) and market capitalization. The latter dependence contrasts sharply with the negative relation between systematic risk premia and market capitalization in Fama–French factor models. For U.S. stocks, market capitalization is strongly negatively related to idiosyncratic volatility. In this case, the positive dependence of IRP on market cap implies an indirect negative relation with volatility. When this indirect effect outweighs the direct positive impact of volatility, IRP declines with IVol—resolving the IVol puzzle.

JEL-Classification: G12, G17

Keywords: Granularity, Fat Tail Distribution, Pareto Distribution, Idiosyncratic Risk Premium

*Gao: Shanghai Advanced Institute of Finance, Shanghai Jiao Tong University, jxgao@saif.sjtu.edu.cn. Liu: Rady School of Management, University of California San Diego, junliu@ucsd.edu. We thank seminar participants at UCSD for useful comments. All remaining errors are our own.

1 Introduction

In 2023, the ten largest firms in the U.S. stock market accounted for roughly one-third of total market capitalization. For comparison, in an equal-weighted market with approximately 5,000 firms, the top ten would comprise only 0.2% of total capitalization. The actual share—around one-third—is over 150 times larger, underscoring the extreme concentration of market capital. This concentration is illustrated most clearly by the recent rise of the "Magnificent Seven."[1] Despite turnover in firm identities and evolving technologies, market concentration has remained a persistent feature of the U.S. stock market.[2] Figure 1 shows that the top ten firms have consistently commanded substantial market shares since 1927, with their combined weight fluctuating between 11% and 30%. In short, the market portfolio is not well diversified.

The concentration of the market portfolio has important implications for asset pricing. In the framework of APT Ross (1976), there are two classes of risks: systematic risks, which are common to all firms and are compensated in terms of systematic risk premia; and idiosyncratic risks, which are firm-specific, with associated risk premia—known as idiosyncratic risk premia(IRP). Despite its intuitive appeal, APT only imposes a bound on total IRP. The sum of squared idiosyncratic risk premia across all stocks is bounded from both above and below by the degree of concentration in the market portfolio (Chamberlain (1983), Chamberlain and Rothschild (1983)).

When investors hold a well-diversified portfolio (Connor (1984), Ingersoll Jr (1984)), the concentration of market capital is zero. In this case, IRPs for all individual assets are zero. Empirically, most asset pricing tests adopt this benchmark and treat any deviation from zero IRP as an "anomaly." However, the observed concentration of the market portfolio—a robust and persistent empirical fact—casts doubt on the zero-IRP assumption.

The goal of this paper is to develop a model that compute IRP explicitly. We assume investors maximize expected utility over returns, following Ross (1976) and Connor (1984), which effectively assumes constant relative risk aversion. As in CAPM and APT, the investor's optimal portfolio

[1] The Magnificent Seven are Alphabet (GOOGL), Apple (AAPL), Amazon (AMZN), Meta Platforms (META), Microsoft (MSFT), Nvidia (NVDA), and Tesla (TSLA).

[2] Similar patterns—often referred to as *concentration*—have been documented in firm fundamentals such as employment, sales, and output (see, for example, Axtell (2001), Gabaix (1999), Gabaix (2011)).

is taken as exogenously given. The key assumption of our paper is that this exogenously specified portfolio is concentrated. The concentration observed in the market portfolio provides the key empirical motivation for this assumption,[3] although it need not coincide with the market index and may be either more or less concentrated.

Under this setting, the pricing kernel is determined by the return of the optimal portfolio. When the optimal portfolio is concentrated, as assumed, its return composes both systematic and idiosyncratic risk— and so does the marginal utility from stock returns. As such, the stock risk premium consists of a standard systematic component (systematic risk premium), which is proportional to factor exposure (beta), as well as an idiosyncratic component (IRP), determined by the stock's idiosyncratic variance and its weight in the optimal portfolio (again, exogenously specified). If the optimal portfolio is well diversified, each stock's portfolio weight approaches zero in a large economy, so the IRP vanishes. But when the optimal portfolio is concentrated, portfolio weights remain nonzero even as the number of assets approaches infinity, IRPs are strictly positive.

A striking feature of our model is that the idiosyncratic risk premium (IRP) increases with firm size. In contrast, the systematic size premium in the Fama–French framework is negatively related to market capitalization: smaller firms earn higher returns due to greater exposure to systematic risk (i.e., higher betas). In our setting, expected returns depends not solely by factor exposures—the idiosyncratic component emerges as an *alpha*, capturing the contribution of firm-specific shocks. This mechanism implies that larger firms command higher idiosyncratic premia, reflecting the dominant role of their idiosyncratic risks rather than systematic factor exposure.

The firm-level idiosyncratic risk premium α_n contains a leading term proportional to

$$\phi_n \sigma_n^2,$$

where ϕ_n is the optimal portfolio weight, and σ_n^2 is the idiosyncratic variance. Higher-order moments such as skewness and kurtosis also enter, but their effects decay with higher powers of ϕ_n. Therefore,

[3]If all investors share the same beliefs about return distributions, they would also choose to concentrate. In reality, heterogeneity in beliefs and preferences allows investors to arrive at different optima while still agreeing on equilibrium prices. We do not model these details explicitly and instead use the empirically observed concentration as motivation.

individual IRPs are tightly linked to portfolio weights, which in turn is detemined by the cross-sectional distribution of market capitalization distribution.

In this paper, we use a well-documnented Pareto distribution as a benchmark. We assume that portfolio weights follow a Pareto distribution with tail parameter ζ, which also serves as a Herfindahl-type measure of concentration.[4] The theoretical implications are sharp: when $\zeta > 2$, the portfolio is well diversified and all portfolio weights vanish as the number of assets grows, implying that all IRPs converge to zero. When $\zeta < 1$, a nontrivial fraction of firms retain nonzero portfolio weights even in large economies, resulting in persistent IRPs. For intermediate values $1 \leq \zeta \leq 2$, diversification improves gradually, but the convergence of IRPs remains slow. Using the market portfolio, we estimate a benchmark concentration of $\zeta \approx 1$. Investors' optimal portfolios may be either more or less concentrated, and our framework allows for this flexibility.

The key prediction of our model is that alpha is positively related to $\phi_n \sigma_n^2$, the leading term of the idiosyncratic risk premium. We estimate alphas relative to both Fama–French factor models and the Instrumented Principal Component Analysis (IPCA) factors constructed in Kelly, Pruitt, and Su (2019), and test whether $\phi_n \sigma_n^2$ explains the cross-section of alphas. Sorting firms by $\phi_n \sigma_n^2$, we find that portfolios with higher values earn significantly larger alphas. The spread between the top and bottom quantiles is both economically sizable and statistically significant. Consistent with our theoretical motivation, we find that firms with high $\phi_n \sigma_n^2$ have a lower systematic risk premium from factor models. Firm-level cross-sectional regressions confirm the same pattern: $\phi_n \sigma_n^2$ consistently and positively predicts average returns.

Our model help reconcile the idiosyncratic volatility puzzle documented by Ang et al. (2006, 2009), who find that stocks with lower idiosyncratic volatility earn higher returns not explained by factor exposures. In our framework, the IRP depends positively on both market capitalization ϕ_n and idiosyncratic volatility σ_n. This bivariate dependence can be reduced to a univariate relation if there is a stable link between ϕ_n and σ_n^2, as is the case in the data.

[4]The Pareto distribution is widely used in the literature and fits well a broad set of firm fundamentals such as employment, sales, and output.

To illustrate this, we estimate the following relationship:

$$\log \sigma_n \approx \text{constant} - \delta \log \phi_n,$$

and find $\hat{\delta} = 0.31$. This estimate confirms the intuitive pattern that firms with higher idiosyncratic volatility tend to have smaller market capitalization and hence lower ϕ_n. Moreover, it implies that the size effect dominates. Substituting this relation into the expression $\alpha_n \propto \phi_n \sigma_n^2$, yields

$$\alpha_n \propto \phi_n^{1-2\delta} = \phi_n^{0.38} = \sigma_n^{-1.23}.$$

In other words, firms with high volatility and small capitalization generate low alphas, while large firms with lower volatility command higher alphas. This mechanism reproduces the empirically observed negative relation between idiosyncratic volatility and returns. Once we adjust idiosyncratic variance for market weight, the positive relation between risk and return is restored. This highlights the importance of controlling for size in uncovering the true pricing of idiosyncratic risk. Overall, our concentration-driven IRP channel offers a resolution to the volatility puzzle and clarifies how the distribution of firm size shapes the pricing of idiosyncratic risk in the cross section.

Related Literature

Our work builds on the extensive Arbitrage Pricing Theory (APT) literature, starting with Ross (1976). Chamberlain and Rothschild (1983) and Chamberlain (1983) provide upper and lower bounds for the total IRP, both determined by the degree of market concentration. We compute the IRP for each individual asset explicitly.

The extensive literature documenting the fat-tailed nature of cross-sectional distributions of various variables provides empirical support for our assumption. The prevalence of Pareto-type tails in employment, sales, output, and other firm fundamentals is well established in Simon and Bonini (1958), Axtell (2001), Gabaix (1999), Gabaix (2011), and Zipf (2016). Motivated by the rising concentration in these distributions, a number of theoretical studies aim to explain the underlying mechanisms, including Luttmer (2011), Lucas Jr and Moll (2014), Gabaix et al. (2016), Benhabib, Perla, and Tonetti (2021), and Kwon, Ma, and Zimmermann (2024). Recent contributions such as Gabaix and Koijen (2024) explore empirical applications, proposing a "granular instrumental variable" to address endogeneity in estimating supply or demand elasticities. Our paper contributes to this literature by exploring on the asset-pricing implications of concentration in market capitalization.

We also contribute to the growing literature on the asset-pricing implications of fat-tailed market capitalization distributions. Malevergne, Santa-Clara, and Sornette (2009) study the internal inconsistency that arises when a factor model includes the market index return as a factor. They argue that this issue becomes more severe when concentration in the market portfolio is high, as residuals relative to the market index may load on an additional factor. Byun, Loudis, and Schmidt (2024) emphasize the role of concentration in shaping the time-series relation between idiosyncratic risk and returns, while we focus on the magnitude and existence of alpha in the cross-section. Kelly and Jiang (2014) fit the cross-section of realized stock returns with a Pareto distribution and use the estimated monthly Pareto exponent to measure the likelihood of extremely poorly performing stocks. They show that this measure predicts future market returns.

Our framework takes investors' concentrated portfolios as exogenous, and relates to a literature that provides mechanisms through which such portfolios arise endogenously. Merton (1987) argues that when investors are not aware of many assets, they hold under-diversified portfolios. Van Nieuwerburgh and Veldkamp (2010) show that such under-diversification can arise optimally from investors' choices about which firms to learn about. When investors are not well diversified—as is often the case in practice—they care more about the risk of the individual firms they hold and therefore place greater value on firm-specific information.

Finally, we contribute to the literature on idiosyncratic risk and its role in asset pricing. Foundational work by Campbell et al. (2001), Xu and Malkiel (2003), Goyal and Santa-Clara (2003), and Herskovic et al. (2016) examines the dynamics and pricing relevance of idiosyncratic volatility. The empirical calibration in our paper reconciles the idiosyncratic volatility puzzle documented by Ang et al. (2006) and Ang et al. (2009), who find that assets with high idiosyncratic volatility tend to earn lower returns—contradicting standard theory. According to the survey in Hou and Loh (2016), this puzzle remains unresolved.

2 Theoretical Framework

Classical Arbitrage Pricing Theory (APT) models rest on two fundamental assumptions: (i) asset returns are driven by a common factor structure, and (ii) idiosyncratic risk is diversifiable. Under these conditions, only systematic risk is priced in expected returns, as firm-specific risk is eliminated through diversification.

The notion of diversification naturally requires a large economy where the number of assets approaches infinity. We elaborate a parsimonious set of general conditions under which factor risk premiums arise in the limit of an infinite asset economy. Our assumptions are more general than those typically found in the literature, and our results nest the classical multi-factor APT model as well as the single-factor CAPM. Moreover, our framework allows for the possibility that idiosyncratic risks are not fully diversified, leading to an idiosyncratic risk premium induced by the market cap concentration observed in the data.

2.1 Settings of a Large Economy, Factor Structure, and Diversification

In the original study of APT, Ross (1976) analyzes a sequence of economies with N assets and examines the limiting case as $N \to \infty$. Our notion of a "large economy"—one with an infinite number of assets—follows the formal construction in Chamberlain and Rothschild (1983) and Chamberlain (1983), which defines asset returns within a Hilbert space. Although this framework may appear technical, its intuitive objective is to ensure that limits of portfolios are well-defined as the number of assets grows without bound.

Following this approach, we consider an economy with N assets, denoted by returns $(r_1, ..., r_N)$, and analyze its behavior as $N \to \infty$. This asymptotic perspective is essential because the foundational insight of APT—that idiosyncratic risk is diversifiable—relies on the existence of a sufficiently large cross-section of assets.

We assume the presence of a risk-free asset with constant return r_f. Consistent with the classical APT setting, each asset's return is modeled as being driven by a set of K common factors and a firm-specific idiosyncratic component. This structure is formalized below:

Assumption 1 (Factor Structure)**.** *The return of each asset $n = 1, ..., N$ satisfies:*

$$r_n - r_f = \mu_n + \beta_n' f + \epsilon_n, \tag{1}$$

where $\beta_n' f = \sum_{k=1}^{K} \beta_n(k) f(k)$, and $\{f(k)\}_{k=1}^{K}$ are common factors with zero mean, i.e., $\mathrm{E}[f(k)] = 0$*.* *The expected return is μ_n, and β_n is a K-dimensional vector of factor loadings.*
The idiosyncratic term ϵ_n has mean zero and is independent of the factors: $\mathrm{E}[\epsilon_n | f] = 0$*. We also assume that the ϵ_n are mutually independent with variance σ_n^2.*

Let $\phi = (\phi_1, ..., \phi_N)$ denote a portfolio in the N-asset economy, where ϕ_n is the weight on asset n, and the remaining weight $1 - \sum_{n=1}^{N} \phi_n$ is allocated to the risk-free asset. The portfolio return is given by:

$$r^\phi = r_f + \sum_{n=1}^{N} \phi_n (r_n - r_f) = r_f + \sum_{n=1}^{N} \phi_n (\mu_n + \beta_n' f + \epsilon_n).$$

We use the superscript ϕ to denote portfolio-level quantities. Notably, we do not require $\sum \phi_n = 1$, allowing for leveraged positions.

Under this structure, the portfolio excess return decomposes as:

$$r^\phi - r_f = \mu^\phi + (\beta^\phi)' f + \epsilon^\phi, \tag{2}$$

where

$$\mu^\phi = \sum_{n=1}^{N} \phi_n \mu_n, \quad \beta^\phi = \sum_{n=1}^{N} \phi_n \beta_n, \quad \epsilon^\phi = \sum_{n=1}^{N} \phi_n \epsilon_n.$$

By construction, common factors are not diversifiable: for most portfolios, $\beta^\phi \neq 0$.[5] In contrast, idiosyncratic risk, represented by ϵ^ϕ, does not generate systematic co-movement and is thus potentially diversifiable.

We define diversification as the condition that the variance of aggregate idiosyncratic risk vanishes in the limit:

$$\lim_{N \to \infty} \mathrm{Var}\left(\sum_{n=1}^{N} \phi_n \epsilon_n \right) = 0. \tag{3}$$

[5] Let B denote the $N \times K$ matrix of factor loadings and ϕ the $N \times 1$ vector of portfolio weights. If B has full rank K, then no nontrivial ϕ satisfies $\phi' B = 0_{1 \times K}$.

Under this condition, the idiosyncratic component of the portfolio return converges to zero in probability:

$$\lim_{N\to\infty}\sum_{n=1}^{N}\phi_n\epsilon_n = 0. \tag{4}$$

Such well-diversified portfolios only exist in the limit, as the idiosyncratic risk of any finite-N portfolio remains strictly positive. For instance, if ϵ_n are i.i.d., then a $1/N$-weighted portfolio has:

$$\mathrm{Var}\left(\sum_{n=1}^{N}\phi_n\epsilon_n\right) = \frac{1}{N}\mathrm{Var}(\epsilon_n),$$

which clearly vanishes as $N \to \infty$. This limiting result underlies the classical APT conclusion that only systematic factor exposures are priced in expected returns.

To ensure that the convergence in (4) is well-defined, we introduce the following regularity condition:

Assumption 2 (Finite Variance of Asset Returns)**.** *All individual asset returns satisfy $Var(r_n) < \infty$ for all n.*

Assumption 2 guarantees that portfolio limits are well-posed. As shown by Chamberlain and Rothschild (1983), the space of finite-variance asset returns forms a Hilbert space under the inner product operator defined by expectation and covariance. In this setting, portfolio variance uniquely determines the portfolio, and the completeness of the space ensures that a sequence of portfolios $(\phi_1, ..., \phi_N)$ has a well-defined limit as $N \to \infty$. As a result, the limiting mean and variance of portfolio returns are both meaningful and analytically tractable.

Finally, a standard condition in the literature ensures convergence of idiosyncratic risk via weak covariance of ϵ_n, as in Chamberlain and Rothschild (1983); Chamberlain (1983). The weak covariance assumption typically requires that the eigenvalues of the idiosyncratic covariance matrix remain uniformly bounded, implying that

$$\lim_{N\to\infty}\|\phi(N)\|^2 = 0 \quad \Rightarrow \quad \lim_{N\to\infty}\phi(N)'\Sigma_\epsilon^N\phi(N) = 0,$$

where $\phi(N)$ is the $N \times 1$ portfolio weight vector and Σ_ϵ^N is the $N \times N$ covariance matrix of idiosyncratic returns. The condition $\lim_{N\to\infty} \|\phi(N)\|^2 = 0$ indicates that the dispersion of portfolio weights converges to zero as the number of assets approaches infinity.

In our baseline model, we assume for tractability that ϵ_n are independent across assets, with asset-specific variances σ_n^2. Notably, this assumption does not conflict with the findings in Herskovic et al. (2016), who estimate a time-varying σ_n and document strong co-movement in the *time series* of idiosyncratic volatilities across firms. Their common factor in idiosyncratic volatility reflects a systematic component in second moments and therefore affects time-varying risk premia, but it does not generate cross-sectional dependence in the firm-specific shocks ϵ_n themselves. Our setting also allows us to retain analytical clarity while accommodating heterogeneity in idiosyncratic risk. In particular, we allow for a negative relationship between firm size and idiosyncratic variance, a feature well documented in the data.

2.2 Risk Premium and Single-Agent Optimization

Building on the framework established in the previous section, we now adopt an equilibrium approach to derive the exact form of both factor and idiosyncratic risk premiums. To facilitate tractable characterization of equilibrium outcomes, we introduce specific assumptions on investor preferences.

Although APT is classically derived from a no-arbitrage condition without relying on investor utility, we argue that utility-based restrictions are implicitly necessary for internal consistency and well-defined limiting behavior in large economies. In particular, we introduce the following assumption:

Assumption 3 (Bounded Relative Risk Aversion)**.** *All agents in the economy exhibit bounded relative risk aversion as* $N \to \infty$.

This condition addresses a key limitation in models assuming constant absolute risk aversion (CARA). When the aggregate supply of risky assets scales with the number of firms N, agents with CARA utility exhibit increasing relative risk aversion as total wealth grows. As a result, even well-diversified portfolios satisfying (4) may still assign non-zero prices of idiosyncratic risk.

To illustrate this point, consider the example in Ross (1976) with a representative investor who has CARA utility, $u(x) = -e^{-Ax}$. Suppose that asset returns are given by $r_n = \mu_n + \sigma\epsilon_n$, where $\epsilon_n \sim$ i.i.d. $\mathcal{N}(0,1)$, and hence returns are purely driven by idiosyncratic shocks. The optimal demand for asset n is then:

$$D_n = \frac{\mu_n - r_f}{A\sigma^2}.$$

Assuming that each asset has market value $M_n = M$, market clearing implies:

$$\mu_n - r_f = A\sigma^2 M,$$

so that the risk premium remains strictly positive as $N \to \infty$. In this case, the market portfolio is equally weighted and hence satisfies the diversification condition (4), yet idiosyncratic risk remains priced.

This counterexample highlights a fundamental issue: although portfolio weights ϕ_n diminish as N grows, the representative agent's total wealth increases linearly with N, which drives up the effective relative risk aversion under CARA utility. As Ross (1976) notes, "The difficulty with the constant absolute risk aversion example arises because the coefficient of relative risk aversion increases with wealth. This suggests considering risk-averse agents for whom the coefficient of relative risk aversion is uniformly bounded..."

Viewed from this perspective, the classical APT pricing result—though presented through an arbitrage-based lens—implicitly relies on a bounded relative risk aversion condition to ensure that diversification eliminates the price of idiosyncratic risk. With this in mind, we now proceed to derive equilibrium pricing formulas that make explicit the link between utility, diversification, and the emergence (or disappearance) of risk premia associated with both systematic and idiosyncratic components.

2.2.1 Single-Agent Optimization

Assumption 4 (Single Agent with Rational Expectations)**.** *There exists at least one agent with rational expectations who understands the factor structure of asset returns. This agent maximizes* *expected utility over portfolio returns:*

$$\phi^* = \arg\max \mathrm{E}[u(r^\phi)],$$

where $r^\phi = \phi' r$ *denotes the portfolio return and* ϕ^* *is the optimal portfolio weight vector.*

The APT factor pricing result is often interpreted as a consequence of diversification in the optimal portfolios of rational agents—either a representative investor or the equilibrium outcome across many agents. Assumption 4 imposes a minimal requirement: the existence of one rational investor who understands the factor structure and optimizes accordingly. This general setup nests classical models such as the CAPM, where all investors hold mean-variance efficient portfolios.[6]

Crucially, the existence of a single optimizing agent is sufficient to characterize prices, since prices must clear regardless of how many agents participate. This approach allows us to avoid the strong assumptions imposed in traditional models. For example, the CAPM assumes that all investors hold the same tangency portfolio, commonly proxied by the market index—an assumption criticized by Roll (1977), who points out that the true mean-variance frontier may be unobservable.

Although we do not observe the optimal portfolio ϕ^* directly, empirical evidence suggests that some investors do not fully diversify idiosyncratic risk. In particular, the heavy right tail in the firm size distribution implies that a small number of firms command disproportionately large market shares. This concentration, if reflected in the portfolio choices of marginal investors, means that their optimal portfolios ϕ^* are not diversified, and thus allow idiosyncratic shocks to influence pricing via the stochastic discount factor. As a result, idiosyncratic shocks can carry risk premia, and the classical APT pricing conclusion may fail.

Assumption 5 (Constant Relative Risk Aversion)**.** *The optimizing agent has constant relative risk aversion and maximizes expected utility over returns using a reduced-form utility function. We consider either the exponential utility,*

$$u(r^\phi) = -e^{-\gamma r^\phi},$$

[6]The market portfolio is efficient when two-fund separation holds—either under return normality or quadratic utility. When two-fund separation fails, optimal portfolios may differ across agents. In such cases, diversification ensures APT pricing results, but not necessarily CAPM.

where $\gamma = W_0 \cdot \Gamma$ is the relative risk aversion (the product of initial wealth and absolute risk aversion), or the power utility,

$$u(r^\phi) = \frac{(1 + r^\phi)^{1-\gamma}}{1-\gamma}.$$

As shown in the appendix, both utility specifications are derived from wealth-based preferences—exponential or power utility—and both preserve constant relative risk aversion in the limiting economy.

2.2.2 Well-Diversified Portfolios and Factor Risk Premium

Our framework yields a general expression for expected returns in equilibrium under utility maximization.

Theorem 1 (Solution of Expected Returns)**.** *In an economy with N assets, a rational investor chooses $\phi^* = (\phi_1^*, ..., \phi_N^*)$ to maximize expected utility over the portfolio return r^ϕ:*

$$\phi^* = \arg\max \mathrm{E}[u(r^\phi)].$$

Assume that asset returns follow the factor structure in Assumption 1. Then the expected excess return of any asset $n = 1, ..., N$ satisfies:

$$\mu_n = -\frac{1}{\mathrm{E}[u'(r^*)]}\mathrm{E}\left[u'(r^*)(\beta_n' f + \epsilon_n)\right], \tag{5}$$

where r^* is the return on the optimal portfolio:

$$r^* = r_f + \sum_{n=1}^{N} \phi_n^*(r_n - r_f).$$

Equation (5) expresses risk premia in terms of covariation with the marginal utility of consumption. The pricing kernel $u'(r^*)$ captures the investor's marginal valuation of returns in each state, and the risk premium compensates assets for their covariance with the systematic component $\beta_n' f$ and the idiosyncratic component ϵ_n.

For each asset n, the expression can be decomposed into a factor-driven and an idiosyncratic component:

$$\mu_n = \beta_n' \lambda + \alpha_n, \tag{6}$$

where

$$\lambda = -\frac{1}{\mathrm{E}[u'(r^*)]}\mathrm{E}[u'(r^*)f] \quad \text{and} \quad \alpha_n = -\frac{1}{\mathrm{E}[u'(r^*)]}\mathrm{E}[u'(r^*)\epsilon_n]$$

denote the factor risk premia and idiosyncratic risk premium, respectively.

As the number of assets grows, if the optimal portfolio ϕ^* is well-diversified in the sense of (4), then the portfolio-level idiosyncratic component $\epsilon^* = \sum_{n=1}^{N} \phi_n^* \epsilon_n$ vanishes in the limit. As a result, the marginal utility is no longer influenced by firm-specific shocks, and the idiosyncratic premium α_n converges to zero. This result recovers the classical APT pricing formula, in which only systematic risks are priced.

Theorem 2 (Diversification and Beta Pricing Result). *If the optimal portfolio ϕ^* satisfies the diversification condition* (4)*, i.e., idiosyncratic risk is fully diversified, then the APT factor pricing result holds. That is, there exists a vector λ such that:*

$$\lim_{N\to\infty} \mu_n = \beta_n' \lambda, \quad \forall n.$$

This conclusion highlights the critical role of diversification. When the portfolio held by the marginal investor fully diversifies idiosyncratic shocks, these shocks no longer enter the pricing kernel, and thus carry no risk premium. Systematic risks—captured by exposure to common factors—remain priced, consistent with the intuition of arbitrage-based asset pricing.

Our derivation shows that the classical APT result emerges as a special case of utility-based equilibrium pricing when the optimal portfolio satisfies strong diversification. In this sense, our framework generalizes APT by making explicit the behavioral and structural assumptions—on preferences, return dynamics, and portfolio concentration—under which factor pricing holds.

2.2.3 Failure of Diversification and Idiosyncratic Risk Premium

In contrast to the previous section, when idiosyncratic shocks are not fully diversified—i.e., when $\lim_{N\to\infty} \epsilon^*$ does not converge to zero—idiosyncratic risk may command a nontrivial risk premium, and the classical APT factor pricing result fails to hold.

We develop two key insights. First, rising capital concentration impairs diversification. By modeling market concentration using a Pareto distribution, we show that ϵ^* may decay very slowly, and under extreme concentration, may fail to converge altogether—even as the number of assets tends to infinity. Second, taking imperfect diversification as given, we derive closed-form expressions for the idiosyncratic risk premium α_n under different utility functions. Since the second perspective connects more directly to asset pricing and empirical implications, we present it first and defer the convergence analysis to a later section.

Recall from Assumption 5 that the investor has constant relative risk aversion. We begin with the case of exponential utility.

Theorem 3 (Idiosyncratic Risk Premium under Exponential Utility)**.** *Suppose the investor has exponential utility with risk aversion γ,*

$$u(r^\phi) = -e^{-\gamma r^\phi},$$

and that idiosyncratic shocks ϵ_n are independent with all moments finite. Then, the idiosyncratic risk premium of asset n is given by:

$$\alpha_n = \gamma\phi_n^*\sigma_n^2 + \sum_{i=2}^{\infty}(-1)^{i+1}\frac{\kappa_n(i+1)}{i!}(\gamma\phi_n^*)^i, \tag{7}$$

where σ_n^2 is the variance of ϵ_n, and $\kappa_n(i)$ denotes the i-th cumulant of ϵ_n.

The exponential utility permits a clean separation of each asset's return in the pricing kernel because the portfolio return appears linearly in the exponent. In the appendix, we show that:

$$\alpha_n = -\frac{1}{\mathrm{E}[u'(r^*)]}\mathrm{E}[u'(r^*)\epsilon_n] = -\frac{1}{\mathrm{E}[e^{-\gamma\phi_n^*\epsilon_n}]}\mathrm{E}[e^{-\gamma\phi_n^*\epsilon_n}\epsilon_n]. \tag{8}$$

Letting $\eta = \gamma\phi_n^*$, we observe that $\mathrm{E}[e^{-\eta\epsilon_n}]$ is the moment generating function (MGF) of $-\epsilon_n$, denoted $m(\eta)$, so that

$$\alpha_n = \frac{m'(\eta)}{m(\eta)} = \frac{d\log m(\eta)}{d\eta},$$

which equals the derivative of the cumulant-generating function of $-\epsilon_n$.

Assuming all moments of ϵ_n exist, this leads to a series expansion:

$$\alpha_n = \gamma\phi_n^*\sigma_n^2 - \frac{\kappa_n(3)}{2}(\gamma\phi_n^*)^2 + \frac{\kappa_n(4)}{6}(\gamma\phi_n^*)^3 - \cdots.$$

The first-order term scales with $\phi_n^*\sigma_n^2$, and higher-order terms enter as powers of ϕ_n^*. Since empirically $0 < \phi_n^* < 1$, the market weight amplifies the effect of idiosyncratic variance.

In the special case where ϵ_n is normally distributed, all higher-order cumulants vanish, and the idiosyncratic risk premium simplifies to:

$$\alpha_n = \gamma\phi_n^*\sigma_n^2.$$

If, in addition, the optimal portfolio ϕ^* coincides with the market portfolio, then expected returns are fully determined by exposure to the market factor, and the CAPM relation holds. Notably, CAPM does not require that the market portfolio be well-diversified; it only requires that all investors hold the market portfolio as their optimal choice.

However, in practice, two important departures arise. First, when market concentration is high, certain assets have large ϕ_n^*, leading to non-negligible α_n even under normality. Second, when idiosyncratic shocks exhibit skewness or excess kurtosis, higher-order cumulants do not vanish, and the risk premium α_n becomes increasingly sensitive to large ϕ_n^*. As a result, market concentration becomes a key determinant of the magnitude of α_n. In such environments, using the market portfolio alone—as implied by CAPM—fails to capture the full cross-section of expected returns. A proper decomposition of risk into factor and idiosyncratic components, as in our framework, becomes essential.

We now turn to power utility, which introduces additional complexity due to the nonlinear dependence on aggregate returns.

Theorem 4 (Idiosyncratic Risk Premium under Power Utility). *Suppose the investor has power utility with relative risk aversion γ,*

$$u(r^{\phi}) = \frac{(1+r^{\phi})^{1-\gamma}}{1-\gamma},$$

and define $R = 1 + r^* - \phi_n^* \epsilon_n$, which is independent of ϵ_n. Suppose that $R > 0$, and for all $k \geq 1$, the moments $\mathrm{E}[\epsilon_n^{k+1}]$ and $\mathrm{E}[R^{-\gamma-k}]$ are finite. Then the idiosyncratic risk premium associated with asset n, defined by

$$\alpha_n = -\frac{\mathrm{E}[(1+r^*)^{-\gamma} \cdot \epsilon_n]}{\mathrm{E}[(1+r^*)^{-\gamma}]}, \tag{9}$$

admits the following series expansion:

$$\alpha_n = \sum_{k=1}^{\infty} \frac{(-1)^{k+1}(\gamma)_k}{k!} \cdot (\phi_n^*)^k \cdot \mathrm{E}[\epsilon_n^{k+1}] \cdot \frac{\mathrm{E}[R^{-\gamma-k}]}{\mathrm{E}[(1+r^*)^{-\gamma}]}, \tag{10}$$

where $(\gamma)_k = \gamma(\gamma+1)\cdots(\gamma+k-1)$ is the rising factorial.

Unlike exponential utility, power utility does not allow the decomposition of the portfolio return into separable asset-level terms, because the return appears inside a power function rather than an exponent. This lack of separability implies that the pricing of ϵ_n depends not only on its marginal distribution, but also on its interaction with the aggregate return r^*. Nonetheless, the leading term remains proportional to $\phi_n^* \sigma_n^2$, and the higher-order contributions again depend on moments of ϵ_n and the distribution of R.

To illustrate the expressions in both utility cases, we consider a log-normal example that ensures tractability and guarantees the positivity of $1 + r^*$, a technical requirement for power utility.

Let

$$1 + r^* = e^{0.1+0.15Z}, \quad \epsilon_n = e^{-0.5\sigma^2+\sigma Z} - 1, \quad Z \sim \mathcal{N}(0,1),$$

so that ϵ_n has zero mean. Then, the i-th moment of ϵ_n is:

$$\mathrm{E}[\epsilon_n^i] = e^{(i^2-i)\sigma^2/2}, \quad i \geq 2.$$

Higher-order moments grow exponentially in i, and the cumulants in (7) scale with powers of $e^{\sigma^2/2} > 1$. This rapid growth may offset the decay from small ϕ_n^*, making higher-order contributions to α_n economically meaningful—even for firms with small market weight. As a result, fat tails and skewness in idiosyncratic shocks can generate persistent, non-negligible idiosyncratic risk premia, even in large asset markets.

2.2.4 Reconciling the Idiosyncratic Volatility Puzzle

The empirical literature on the idiosyncratic volatility puzzle (IVol puzzle), notably Ang et al. (2006, 2009), evaluates whether idiosyncratic volatility predicts asset returns using market capitalization-based portfolios. To maintain consistency with this empirical context—and in line with our earlier modeling assumption—we treat the optimal portfolio ϕ^* as equivalent to the market portfolio. That is, we use market weights ϕ_n as proxies for ϕ_n^*, under the interpretation that market weights reflect the allocations of marginal investors who determine prices. This assumption enables us to operationalize α_n using observable data and link it to capital concentration, which plays a central role in generating undiversified risk premia in our framework.

The closed-form expressions for α_n derived in equations (7) and (10) offer a natural resolution to the IVol puzzle, which finds that firms with higher idiosyncratic variance tend to earn lower average returns. Our resolution centers on the interaction between market concentration and the empirically observed negative relationship between firm size and idiosyncratic volatility. Specifically, firms with low IVol (i.e., small σ_n^2) tend to have large market weights ϕ_n, which in turn amplify the pricing impact of even modest variance levels through the first-order term in α_n.

We model this inverse size–volatility relationship using the following power-law specification:

$$\sigma_n = S(N)\phi_n^{-\delta}, \quad 0 \le \delta < 0.5. \tag{11}$$

Here, δ governs how quickly IVol declines with firm size. This specification is consistent with Pareto-distributed firm sizes and fits our data well, with an empirical estimate of $\delta \approx 0.3$.

To ensure that large firms do not have too small idiosyncratic variance, we impose the condition $0 \le \delta < 0.5$. In the appendix, we show that this condition maintains a finite value of σ_n, due to the positive lower bound of ϕ_n, which is naturally satisfied under the Pareto distribution. When $\delta = 0$, we have $\sigma_n = S$ for all assets. In this case, S represents the average level of idiosyncratic volatility. Since S may vary with the number of firms N, we assume that $S(N)$ is uniformly bounded above to preserve internal consistency.

We incorporate the size–IVol relation (11) into the expression for α_n and evaluate how idiosyncratic premia vary with firm size using a tractable log-normal example. Specifically, we calibrate the gross return $1 + r^*$ to match the empirical moments of the market portfolio:

$$1 + r^* = e^{0.1+0.15Z}, \quad Z \sim \mathcal{N}(0, 1).$$

For the idiosyncratic shock, we assume:

$$\epsilon_n = e^{-0.5\sigma_n^2+\sigma_n Z} - 1, \quad Z \sim \mathcal{N}(0, 1),$$

with σ_n following equation (11). We set $S = 0.2$, $\gamma = 5$, and $\delta = 0.3$, consistent with our empirical estimates. Based on these inputs, we compute the relevant moments of ϵ_n across 10,000 Monte Carlo simulations.

In Figure 2, we report how α_n varies with ϕ_n under both exponential and power utility. Panels A and B show results under each utility specification, respectively. In both cases, the idiosyncratic risk premium increases with market weight—even after adjusting for the size–IVol effect via equation (11).

Moreover, the magnitude of α_n is economically meaningful. For example, an asset with 2% market weight produces an annual idiosyncratic alpha between 3% and 4% in both utility settings. This result underscores the pricing power of large firms: despite having lower IVol, they carry large ϕ_n, which amplifies their contribution to undiversified pricing errors.

The simulation also yields an important analytical insight: adjusting for market weight is critical in explaining the observed return–volatility relationship. In particular, the first-order approximation of α_n implies a simple empirical proxy:

$$\alpha_n \approx \gamma \phi_n \sigma_n^2.$$

Substituting from equation (11), we obtain:

$$\alpha_n \approx \gamma S^2 \phi_n^{1-2\delta} = S^{1/\delta} \sigma_n^{2-1/\delta}.$$

When $\delta < 0.5$, the pricing impact of market weight dominates the decay in volatility. In this case, large firms exhibit higher idiosyncratic premia despite having lower variance—directly reconciling the IVol puzzle.

In the empirical literature, a common regression approach tests whether IVol explains returns:

$$\alpha_n = \text{constant} + \eta \sigma_n.$$

Empirically, the estimated coefficient $\hat{\eta}$ is often negative—a finding that challenges conventional theory. However, under our model, $\hat{\eta}$ reflects not the direct price of σ_n, but its correlation with the true driver $\phi_n \sigma_n^2$. In particular:

$$\hat{\eta} \propto \text{corr}(\phi_n \sigma_n^2, \sigma_n) < 0.$$

Therefore, regressions that ignore market weights ϕ_n risk structural misspecification.

Importantly, when diversification holds—i.e., $\phi_n \approx 1/N$ and $\alpha_n \approx 0$ across firms—ignoring size adjustment is less harmful. For instance, under uniform weights:

$$\hat{\eta} = \frac{1}{N} \gamma > 0,$$

which recovers the standard prediction of a positive risk–return tradeoff.

In summary, our framework reconciles the IVol puzzle by attributing the negative empirical relationship between idiosyncratic risk and expected return to market concentration and the inverse size–variance relationship. By highlighting how capital concentration amplifies undiversified risk exposure, we provide a theoretical foundation for α_n to persist even in large economies. In the empirical section, we test these predictions using firm-level return data and directly estimate the magnitude of the implied idiosyncratic risk premium.

2.3 When Does Diversification Fail? Quantification Using the Pareto Distribution

As emphasized in our treatment of the optimal portfolio, the source of undiversified idiosyncratic risk lies in capital concentration. In practice, the distribution of invested capital of the optimal portfolio may have a higher or lower degree of fat-tail than the market capital observed. Our analysis below aims to provide a general relation about the degree of fat-tail and diversification of idiosyncratic risks.

Nevertheless, since the optimal portfolio is proxied by the market portfolio, and ϕ_n reflects the relative size of market capitalization, our analysis of diversification naturally translates into an analysis of the distributional properties of market capitalization. We study how the fat-tailed nature of the market cap distribution governs the persistence of α_n via the failure of ϵ^* to vanish in large economies.

Market weight plays a central role in our asset pricing implications and naturally connects to the distribution of firm-level market capitalization. Let $X_{n=1,\ldots,N}$ denote the market capitalization of asset n. The market weight of asset n is then defined as:

$$\phi_n = \frac{X_n}{\sum_{n=1}^{N} X_n}.$$

In this section, we characterize the distribution of market capitalization and identify the conditions under which the presence of dominantly large firms breaks diversification, thereby inducing non-zero idiosyncratic risk premia.

As a benchmark, the following theorem establishes that if market capitalizations are thin-tailed—i.e., possess finite second moments—then idiosyncratic shocks are diversified in the limit, as stated in equation (4), and classical factor pricing results hold.

Theorem 5 (Thin-Tailed Size Distribution and Diversification of Idiosyncratic Risks). *Suppose firm market capitalizations $X_n, n = 1, \ldots, N$, are i.i.d. random variables drawn from a thin-tailed distribution such that* $\mathrm{E}[X_n^2] < \infty$*. Then, idiosyncratic risks are diversified away in the limit, such that*

$$\lim_{N \to \infty} \sum_{n=1}^{N} \phi_n \epsilon_n = 0,$$

i.e., the idiosyncratic shocks in the portfolio converge to zero in probability.

The key object governing diversification is the variance of total idiosyncratic shocks, $\mathrm{Var}\left(\sum_{n=1}^{N} \phi_n \epsilon_n\right)$. When market capitalization has finite variance, the market becomes sufficiently fragmented such that each asset's weight becomes negligible. In particular,

$$\lim_{N \to \infty} \sum_{n=1}^{N} \phi_n^2 \sigma_n^2 = 0, \quad \text{and} \quad \lim_{N \to \infty} \sum_{n=1}^{N} \phi_n \epsilon_n = 0.$$

In other words, the aggregate impact of idiosyncratic shocks converges to zero in probability, due to both their zero mean and vanishing aggregate variance.

We consider this result without incorporating the negative relation between ϕ_n and σ_n for simplicity. Intuitively, incorporating this relation only accelerates the convergence since it reduces the impact of large firms.

In contrast, when concentration in market capitalization is significant, diversification may fail. There may exist firms—such as the so-called "Magnificent Seven"—with sufficiently large market shares (ϕ_n) such that their idiosyncratic shocks are not diversified in the aggregate. The natural questions, then, are: how can we quantify the extent of concentration or fat-tailedness in the market cap distribution? And how does this concentration affect the convergence behavior of aggregate idiosyncratic shocks $\epsilon^* = \sum_{n=1}^{N} \phi_n \epsilon_n$ as $N \to \infty$?

To answer these questions, we adopt the Pareto distribution, widely used in the literature, to model the concentration of market cap and analyze the implications for the distribution of idiosyncratic risk premia. Specifically, we assume that the market capitalizations X_n are i.i.d. and follow a Pareto distribution with survival function:

$$P(X_n > x) = x^{-\zeta}, \quad x > 1, \quad \zeta < 2. \tag{12}$$

We defer a detailed introduction of the Pareto distribution to the appendix. The standard Pareto distribution has a lower bound $x_{\min}$ in the survival function; we normalize the lower bound $x_{\min} = 1$ for simplicity.

The advantage of this specification is that the entire degree of fat-tailedness is governed by a single parameter—the Pareto exponent ζ—which is both interpretable and empirically estimable. As we will show, this parameter provides a direct measure of concentration in the cross-section of market capitalization.

Our focus is on the case $\zeta < 2$, which implies that the size distribution has infinite variance due to extreme values in market capitalization, in contrast to the thin-tailed case discussed earlier. When $\zeta < 1$, even the mean of the distribution is infinite, and a few dominant firms account for nearly the entire market. In such cases, the presence of extreme values persists as N grows, and diversification fails.

We formalize this in the following result. Let $\epsilon^* = \sum_{n=1}^{N} \phi_n \epsilon_n$ denote the aggregate idiosyncratic shock. For simplicity, we assume that ϵ_n are i.i.d. and thin-tailed, and derive the limiting behavior of ϵ^* as a function of ζ:

Theorem 6 (Pareto Distribution and Convergence of ϵ^*)**.** *Suppose market capitalizations X_n follow a Pareto distribution with tail index $\zeta < 2$, as defined in (12), and the idiosyncratic shocks ϵ_n are i.i.d. with finite variance $\mathrm{E}[\epsilon_n^2] < \infty$. Then as $N \to \infty$, the convergence behavior of aggregate* *idiosyncratic shocks* $\epsilon^* = \sum_{n=1}^N \phi_n \epsilon_n$ *depends on* ζ*, with*

$$\lim_{N\to\infty} \epsilon^* = \begin{cases} \dfrac{Y_\zeta^{(\epsilon)}}{Y_\zeta} & \textit{if } \zeta < 1, \\ \lim\limits_{N\to\infty} \dfrac{Y_\zeta^{(\epsilon)}}{Y_\zeta + N^{1-1/\zeta}\mathbb{E}[X]} \to 0 & \textit{if } \zeta > 1, \\ \lim\limits_{N\to\infty} \dfrac{Y_\zeta^{(\epsilon)}}{Y_\zeta + \log N} \to 0 & \textit{if } \zeta = 1. \end{cases} \tag{13}$$

Theorem 6 shows that under a Pareto distribution for market cap, the aggregate idiosyncratic shock $\epsilon^* = \sum_{n=1}^N \phi_n \epsilon_n$ converges to a ratio of two random variables: $Y_\zeta^{(\epsilon)}$ and Y_ζ. Here, $Y_\zeta^{(\epsilon)}$ represents the weighted sum of idiosyncratic shocks, and Y_ζ represents the total market capitalization, with both quantities inheriting the heavy-tailed properties of the Pareto distribution through the common tail index ζ. These two terms follow stable distributions, and their ratio determines the limiting behavior of ϵ^*.

The emergence of stable distributions is a direct consequence of the Generalized Central Limit Theorem: when summing i.i.d. random variables with infinite variance—as is the case with Pareto-distributed market capitalization for $\zeta < 2$—the normalized sum converges to a stable law with the same index ζ, rather than a Gaussian. A stable distribution is defined by four parameters: the tail index $\zeta \in (0,2]$, location $\mu \in \mathbb{R}$, scale $\sigma > 0$, and skewness $\beta \in [-1,1]$. The parameter ζ governs tail thickness and the existence of moments. When $\zeta < 2$, the variance is infinite; when $\zeta < 1$, the mean also fails to exist. The characteristic function of a stable distribution is

$$\varphi_Y(t) = \exp\left\{it\mu - \sigma|t|^\zeta \left(1 + i\beta\,\mathrm{sign}(t)\cdot\omega_\zeta(t)\right)\right\},$$

where $\omega_\zeta(t) = \tan\left(\frac{\pi\zeta}{2}\right)$ for $\zeta \neq 1$, and $\omega_1(t) = -\frac{2}{\pi}\log|t|$ for $\zeta = 1$.

In our context, the tail index ζ is not merely a feature of the limiting distribution; it is inherited directly from the market cap distribution and governs the extent to which idiosyncratic shocks are diversified in the cross-section. When $\zeta > 1$, the mean of the market cap distribution exists, and although large firms still exert outsized influence, the aggregate idiosyncratic shock ϵ^* converges to zero as $N \to \infty$. However, the convergence is unusually slow—much slower than under thin-tailed assumptions—because of the persistence of granular weights. In contrast, when $\zeta < 1$, the mean of the market cap distribution is infinite, and the market is dominated by a small number of extremely large firms. In this case, the normalization fails to dilute the influence of the largest $\phi_n \epsilon_n$ terms, and ϵ^* remains non-degenerate in the limit. That is, the idiosyncratic shocks of dominant firms persist in the aggregate and prevent convergence to zero.

Quantitatively, this failure of diversification is reflected in the decay rate of ϵ^*. Under thin-tailed assumptions, idiosyncratic shocks diversify at a canonical rate of $1/N$. However, when market capitalizations follow a Pareto distribution with tail index $\zeta < 2$, the decay rate becomes $N^{1/\zeta - 1}$, which slows dramatically as ζ falls. For instance, if $\zeta = 1.05$ and $N = 5000$, roughly matching the CRSP universe, then

$$N^{1/\zeta-1} = 5000^{1/1.05-1} \approx 0.6,$$

compared to the classical rate of

$$1/N = \frac{1}{5000} = 0.0002.$$

This stark difference illustrates how idiosyncratic shocks remain quantitatively significant even in large samples when concentration is present. In such economies, diversification does not fully wash out firm-specific risks—especially for the largest firms—and idiosyncratic pricing distortions persist despite arbitrage-free conditions. We illustrate this point with the following theorem, which characterizes the portfolio weight of the largest firm when market capitalizations follow a Pareto distribution.

Theorem 7. *If firm sizes X_n are i.i.d. and follow a Pareto distribution with tail index ζ, i.e.,*

$$P(X_n > x) = x^{-\zeta}, \quad x > 1,$$

then the portfolio weight of the largest firm is given by

$$\phi_{\max} = \frac{X_{\max}}{\sum_{n=1}^{N} X_n}.$$

As $N \to \infty$, $\phi_{\max}$ converges in distribution as follows:

$$\lim_{N\to\infty} \phi_{\max} = \begin{cases} \dfrac{F_\zeta}{Y_\zeta} & \text{if } \zeta < 1, \\ \dfrac{F_\zeta}{Y_\zeta + \log N} & \text{if } \zeta = 1, \\ \dfrac{F_\zeta}{Y_\zeta + N^{1-1/\zeta}\,\mathbb{E}[X]} & \text{if } \zeta > 1, \end{cases} \tag{14}$$

where F_ζ follows a Fréchet distribution and Y_ζ is a stable random variable as defined in Theorem 6.

This result characterizes how the magnitude of $\phi_{\max}$ evolves with the number of assets N, and aligns with the convergence behavior of aggregate idiosyncratic shocks. As ζ increases, the Pareto tail becomes thinner and $\phi_{\max}$ converges to zero more quickly, reflecting more effective diversification in economies with less granular firm size distributions.

To validate the theoretical result in Theorem 7, Figure 3 plots the average maximum portfolio weight $\phi_{\max}$ across 10,000 simulated economies for each $\zeta \in \{0.5, 1.0, 1.5, 2.5\}$. In each simulation, firm sizes follow a Pareto$(\zeta, 1)$ distribution and $\phi_{\max} = \max_n X_n / \sum_j X_j$.

When $\zeta = 0.5$ (Panel A), the distribution has infinite mean, and $\phi_{\max}$ remains persistently high (~ 0.63), consistent with the non-vanishing Fréchet ratio limit. At $\zeta = 1.0$ (Panel B), $\phi_{\max}$ decays slowly, roughly at $1/\log N$, still exceeding 20% even with $N = 10{,}000$. With $\zeta = 1.5$ (Panel C), the decay accelerates, consistent with the $N^{1-1/\zeta}$ rate, and $\phi_{\max}$ drops below 5%. Finally, for thin-tailed $\zeta = 2.5$ (Panel D), $\phi_{\max}$ declines rapidly, approaching near-uniform diversification. In addition, the results for $\phi_{\max}$ hold for the few largest firms since the Pareto distribution indicates a linear relationship between the logarithm of a firm's rank k and its sorted value (market cap X_k):

$$\log k \approx -\zeta \log X_k.$$

This log-log relationship implies a rapid decay in market cap with rank. Specifically, the size of the k-th largest firm follows:

$$X_k \sim k^{-1/\zeta} X_{\max}.$$

For instance, if $\zeta = 1$, the second-largest firm is approximately half the size of the largest, the third-largest is about one-third the size of the largest, and so on. These patterns confirm that fat-tailed firm size distributions induce persistent market-concentration in capitalization, even in large economies.

In addition, one can extend the convergence result for $\epsilon^* = \sum_{n=1}^{N} \phi_n \epsilon_n$ from the case where all ϵ_n have identical variance to a more general setting that incorporates the empirically observed negative size–variance relationship. Specifically, we consider the decomposition:

$$\lim_{N\to\infty} \epsilon^* = \lim_{N\to\infty} \sum_{n=1}^{N} \phi_n \sigma_n e_n,$$

where e_n are i.i.d. idiosyncratic shocks with mean zero and unit variance, and σ_n denotes the idiosyncratic volatility of asset n. Under the size–variance relationship specified in equation (11), the limiting behavior of ϵ^* is shaped by both the tail index ζ and the decay parameter δ.

By the properties of the Pareto distribution, if $\sigma_n^{-1} \propto \phi_n^{\delta}$, then σ_n^{-1} also follows a Pareto distribution, with implied tail index ζ/δ, suggesting a much lower level of fat-tail. This tractable structure allows us to derive the convergence rate of ϵ^*. We present the asymptotic variance of ϵ^* and characterize its convergence rate explicitly.

Theorem 8 (Pareto Distribution and Convergence of $\mathrm{Var}(\epsilon^*)$)**.** *Suppose that market capitalizations X_n follow a Pareto distribution with tail index ζ, and that $\epsilon_n = \sigma_n e_n$, where e_n are i.i.d. with zero mean and unit variance. Let $\sigma_n^{-1} \propto \phi_n^{\delta}$ with $0 < \delta < 0.5$. Then the asymptotic variance of ϵ^* satisfies:*

$$\lim_{N\to\infty} \sum_{n=1}^{N} \phi_n^2 \sigma_n^2 = \frac{\lim_{N\to\infty} \sum X_n^{2-2\delta}}{\left(\lim_{N\to\infty} \sum X_n\right)^{2-2\delta}} = \begin{cases} \dfrac{Y_{\zeta/(2-2\delta)}}{(Y_\zeta)^2} & \text{if } \zeta < 1, \\ \lim\limits_{N\to\infty} \dfrac{Y_{\zeta/(2-2\delta)}}{\left(Y_\zeta + N^{1-1/\zeta}\mathbb{E}[X]\right)^{2-2\delta}} = 0 & \text{if } 1 < \zeta < 2-2\delta, \\ \lim\limits_{N\to\infty} \dfrac{\mathbb{E}[X^{2-2\delta}]}{N^{2-2\delta-1}\mathbb{E}[X]} = 0 & \text{if } 2-2\delta < \zeta < 2, \\ \lim\limits_{N\to\infty} \dfrac{Y_{\zeta/(2-2\delta)}}{(Y_\zeta + \log N)^{2-2\delta}} = 0 & \text{if } \zeta = 1. \end{cases}$$

Similar to the results in Theorem 6, when $\zeta < 1$, the aggregate idiosyncratic risk does not vanish even as $N \to \infty$, due to the dominance of a few large firms.

The expression $\sum_n \phi_n^2 \sigma_n^2$ in Theorem 8 also serves as a scaled proxy for the total idiosyncratic risk premium (IRP) in the economy. Intuitively, since exponential utility delivers a certainty-equivalent correction proportional to $\phi_n \sigma_n$, the total pricing distortion from idiosyncratic shocks scales with the quadratic sum $\sum \phi_n^2 \sigma_n^2$. This is consistent with the core implication of the APT: when diversification is effective, the aggregate contribution of idiosyncratic risk to expected returns—i.e., the summed IRP—should vanish in large economies.

To validate this implication, we simulate the total idiosyncratic pricing distortion $\sum \phi_n^2 \sigma_n^2$ under different Pareto exponents ζ and a varying number of firms N. For each $\zeta \in \{0.5, 1.0, 1.5, 2.5\}$ and each N, we generate 10,000 simulated economies with firm sizes $X_n \sim \text{Pareto}(\zeta, 1)$, compute $\phi_n = X_n / \sum_j X_j$ and $\sigma_n = S \phi_n^{-\lambda}$, and plot the average of $\sum_n \phi_n^2 \sigma_n^2$ across simulations.

As shown in Figure 4, the behavior of this sum is highly sensitive to the tail index. When $\zeta = 0.5$ (Panel A), the sum remains essentially constant around 0.027 despite growing N, consistent with persistent concentration. For $\zeta = 1.0$ (Panel B), the sum decays slowly, roughly following the rate implied by the ζ-stable central limit behavior. When $\zeta = 1.5$ and 2.5 (Panels C and D), the decay becomes much faster, confirming that thinner tails enable the idiosyncratic risk premium to wash out as the economy becomes large. These patterns support the conclusion that summed IRP only converges to zero under sufficient tail thinness.

Beyond the summed idiosyncratic risk premium, our framework also characterizes the full cross-sectional distribution of firm-level alpha across the economy. Following our IRP formulation, we take $\phi_n \sigma_n^2$ as the leading term in firm-level alpha and examine its cross-sectional distribution. This provides a direct measure of how idiosyncratic risk premia are concentrated across firms of different sizes.

In Figure 5, we simulate a single cross-section of $N = 5000$ firms—comparable to the number of traded equities in the CRSP universe—and compute $\phi_n \sigma_n^2$ under different Pareto tail exponents ζ. The plotted curves sort firms from largest to smallest risk contribution, offering a rank-size view of distortion concentration.

When $\zeta = 0.5$ (Panel A), the distribution is extremely skewed: a handful of firms dominate the aggregate distortion, with the top few exhibiting $\phi_n \sigma_n^2 > 0.03$. As ζ increases to 1.0 (Panel B), the distribution remains heavy-tailed but less concentrated. By $\zeta = 1.5$ and 2.5 (Panels C and D), the distribution becomes markedly thinner, and most firms contribute little to overall IRP. These patterns reinforce the insight that thin-tailed firm size distributions diffuse idiosyncratic shocks more evenly, while granular economies concentrate them in the largest firms.

Taken together, these results show that when concentration is sufficiently strong ($\zeta < 1$), idiosyncratic shocks are not diversified away, regardless of how large the economy becomes. This failure of diversification implies the persistence of idiosyncratic risk premia α_n, even in the limit. In practice, even when the market contains a finite number of assets, our results imply that idiosyncratic premia remain quantitatively significant whenever ζ is small and the decay parameter δ is modest. Importantly, both the Pareto tail index ζ and the size–variance decay parameter δ are readily estimable in empirical applications.

2.4 Remarks and Extensions

2.4.1 CAPM, APT Factor Pricing, and Idiosyncratic Risk Premium

As a benchmark, we derive the cross-sectional pricing results in equations (7) and (10). Notably, using exponential utility nests the CAPM model when returns are normal and some investors choose to hold the market portfolio. In this case, the expected returns have a single-beta representation, yet the decomposition of factor risk premium and idiosyncratic risk premium still holds. In other words, given CAPM holds, the APT factor pricing results might still fail when concentration is significant in the market. From this special case, one can see CAPM and APT as methods to derive a beta-presentation with different requirements: CAPM requires a normal distribution of returns (or quadratic utility), whereas APT requires diversification.

In terms of empirical implications, the APT is more robust to misspecification since one does not need to know the true factors and can apply statistical techniques to find proxies that sufficiently span all the true factors. From this perspective, our analysis based on violation of diversification complements the extensive study in estimating factor pricing models and provides a more comprehensive framework to understand asset prices. Further, different from the factor risk premium as a linear function of beta, our results in equation (7) show that the idiosyncratic risk premium highly depends on portfolio weight ϕ_n and hence links tightly to the distribution of market capitalization.

2.4.2 Time-varying Risk Premium

Without loss of generality, the concentration-driven alpha channel we test remains if we extend to a more comprehensive equilibrium setup such as Merton (1973)'s intertemporal CAPM, which nests the time-varying factor loading models, e.g., Jagannathan and Wang (1996). In a more general setting, it is usually more convenient to work on the logarithm of the pricing kernel m_{t+1} and log-normal asset returns, such that,

$$m_{t+1} - E_t[m_{t+1}] = -\gamma \left((\beta^*)' f_{t+1} + \epsilon^*_{t+1} \right) + \text{shocks in time-varying risk/returns.}$$

Here, the priced risk could also come from time-varying factor risk exposure and time-varying factor risk premia. We focus on the price of concentration-induced idiosyncratic components and take an empirical approach to control for time-varying risk components as factors. Specifically, in our empirical tests, we use the IPCA approach in Kelly, Pruitt, and Su (2019) to tease out the time-varying common factors and focus on the price of idiosyncratic risks. We argue that this approach is robust to misspecification and is well aligned with the nature of testing the factor zoo, which effectively treats factors as sources of risk that drive common movements in the asset return space.

3 Empirical Test

3.1 Data

We use monthly returns and market capitalization data from CRSP, supplemented with firm characteristics from COMPUSTAT to construct control variables. The datasets are merged by aligning monthly CRSP returns with quarterly COMPUSTAT characteristics, filling missing quarterly data with annual observations when necessary. To maintain consistent timing, we apply a standard six-month lag between the end of the quarter in which characteristics are measured and the monthly returns used for portfolio construction.

To measure idiosyncratic risk and factor exposures, we use Fama-French factors obtained from the Kenneth French data library, along with a comprehensive dataset of firm-level characteristics provided by Jensen, Kelly, and Pedersen (2023). Our empirical tests utilize all available firm fundamentals and associated factors over the sample period from 1963 to 2023.

Given our focus on large firms and their idiosyncratic risk, we exclude microcap firms—defined as the smallest 20% of firms—throughout our analysis. Following Hou, Xue, and Zhang (2015), this approach mitigates the influence of anomalies concentrated in small firms, which typically exhibit high transaction costs and low liquidity. While one common alternative in portfolio construction is to use NYSE breakpoints rather than quantiles of the entire sample, we do not adopt this method here, as our analysis centers on the market-cap distribution itself.

3.2 Evidence of Granularity in the Stock Market

As discussed in the theory section, we use the market portfolio as a proxy for the concentrated optimal portfolio. For notational simplicity, we drop the superscript $*$ and use ϕ_n to denote the market weight of each asset.

Figure 6 displays the distribution of firm weights in the stock market, highlighting the ten largest firms by ticker symbol. The distribution exhibits a pronounced fat tail: the ten largest firms account for nearly 30% of total market capitalization in the 2023 CRSP database, which contains approximately 4,000 firms.

Consistent with our theoretical framework, we find that the market-cap distribution closely follows a Pareto distribution. A key diagnostic of a Pareto distribution is a linear relationship between the logarithm of a firm's rank k and its sorted value (market cap X_k):

$$\log k \approx -\zeta \log X_k. \tag{15}$$

This log-log relationship implies a rapid decay in market cap with rank. Specifically, the size of the k-th largest firm follows

$$X_k \sim k^{-1/\zeta} X_{\max}. \tag{16}$$

For instance, if $\zeta = 1$, the second-largest firm is approximately half the size of the largest, the third-largest is about one-third the size of the largest, and so on.

Empirical studies have estimated ζ using this rank-size relationship (e.g., the Hill estimator in Hill (1975)) across various firm size metrics, such as employment and output, typically yielding values around 1. We apply the method in Gabaix and Ibragimov (2011) to correct the small-sample bias in the Hill estimator. Using this approach on market capitalization data for the top 5% of firms in 2023 (ranked by size), we estimate $\hat{\zeta} = 1.08$. Figure 7 shows the fitted log-log relationship, confirming that our estimate fits the data well. These results support the presence of "Zipf's Law" in the market-cap distribution, indicating a robust fat-tailed pattern with ζ close to 1.

Beyond 2023, we observe that this fat-tailed structure persists throughout our sample period and across multiple dimensions.

First, in a manner similar to Figure 6, we track the aggregate market share of the ten largest firms in CRSP each month from 1926 to 2023. This measure of market concentration averages around 20% over time and exhibits an upward trend in recent years.

Second, we apply the Hill estimator monthly to the top 5% of firms to construct a time series of estimated Pareto coefficients ζ. Figure 8 plots these estimates, which consistently range between 1 and 2, with values closer to 1 in recent decades.

Lastly, Table 1 reports the average market share of each firm in each decade, listing the ten largest firms, their combined market weight, and the total number of firms in that decade's sample. We find persistent concentration over time, even as the composition of leading firms evolves with technological change. This measure is also affected by the total number of firms in the market. For example, during the 1990s internet boom, the number of small tech listings surged, temporarily lowering concentration measures. Nonetheless, market concentration remained substantial, with the top ten firms still accounting for over 10% of the market, despite there being more than 10,000 firms.

3.3 Alpha and Size-Adjusted Idiosyncratic Risk

We test the theoretical predictions in (7) and (10) by approximating the idiosyncratic risk premium as the first-order effect of the market weight ϕ_n on the idiosyncratic risk premium α_n.

Specifically, the first-order approximation

$$\alpha_n \approx \gamma \phi_n \sigma_n^2$$

implies that size-adjusted idiosyncratic variance (SAIV), defined as $\phi_n \sigma_n^2$, should play a central role in explaining the idiosyncratic risk premium.

We adopt a standard methodology from the literature. Idiosyncratic variance (IV) is estimated monthly relative to the Fama-French three factors (FF3). We then compute SAIV for each firm in the sample and sort firms by their lagged SAIV into quintiles to construct portfolios. For simplicity, we use IV to represent the idiosyncratic variance σ_n^2, and refer to its square root σ_n as idiosyncratic volatility (IVol).

To isolate alpha—that is, abnormal returns unexplained by systematic factors—we apply the Instrumented Principal Component Analysis (IPCA) method developed by Kelly, Pruitt, and Su (2019). This approach also allows for time-varying factor risk premia. We define post-sample alpha as the average realized return minus the IPCA-fitted factor premium.

Panel A of Table 2 reports key statistics for the five SAIV-sorted portfolios. Firms in the top SAIV quintile account for 79% of total market capitalization, indicating that SAIV is highly concentrated in a small subset of the market. The average SAIV across quintiles displays a fat-tailed distribution. While average returns decline modestly from low to high SAIV, alpha increases with SAIV. The high-minus-low alpha spread is statistically significant, with an annualized difference of 2.73%.

To examine the cross-sectional relationship between α_n and $\phi_n \sigma_n^2$ more closely, we expand the analysis to 100 value-weighted portfolios sorted by SAIV percentiles. This finer partition provides a more accurate approximation of the sample distribution and better captures market-cap dispersion, while still leaving roughly 30–40 firms in each portfolio. Portfolio-level aggregation reduces noise in individual asset returns. In Figure 9, we plot the IPCA-fitted factor premium and corresponding alpha across these portfolios. A clear pattern emerges: as SAIV increases, alpha rises, while the factor premium declines.

Importantly, our approach to studying alpha and under-diversified idiosyncratic risk differs from traditional tests of idiosyncratic risk compensation, which typically use IV alone as the sorting variable. Prior studies, such as Ang et al. (2006, 2009) and Hou and Loh (2016), document that firms with high IV tend to earn lower expected returns—a phenomenon known as the "idiosyncratic volatility puzzle" (IVol puzzle). In contrast, our finding of a positive relationship between alpha and SAIV suggests a rational risk-return tradeoff once market weight is taken into account.

To reconcile our results with the existing literature, we perform a double sort. First, we divide firms into three IV buckets based on their previous-month IV: the bottom 30%, the middle 40%, and the top 30% of the cross-sectional IV distribution. Within each IV bucket, we then sort firms into SAIV quintiles, yielding a 3-by-5 portfolio structure. Panels B, C, and D of Table 2 present the results for the low, medium, and high IV groups, respectively.

Across these panels, we observe a negative relationship between market cap and IV. For example, within the top SAIV quintile, the shares of total market capitalization for firms with low, medium, and high IV are approximately 58%, 19%, and 6%, respectively.

Panel B (low IV group) reveals that the high-minus-low SAIV alpha spread is 4.36% and statistically significant—consistent with our theory that concentration in size contributes to higher alpha. This effect is strongest among firms with the largest size.

In contrast, Panel C (medium IV group) shows an alpha spread that is not statistically significant, indicating a weaker relationship between alpha and SAIV. In Panel D (high IV group), the alpha spread turns negative and remains statistically insignificant, suggesting that the size-risk interaction weakens among firms with very high idiosyncratic variance.

Overall, the results in Table 2 provide empirical support for the concentration-driven alpha. This effect is concentrated among large firms and does not appear in small firms with high IV. Our findings complement existing evidence that asset pricing anomalies tend to concentrate in small-cap stocks (see, for example, Hou, Xue, and Zhang (2015)). At the same time, our results highlight a novel mechanism: the interaction between market cap and risk generates a return premium even among large firms. As we will discuss in the next section, this channel also provides a resolution to the IVol puzzle by accounting for the role of market cap—an element often overlooked in prior studies.

3.4 Reconciling the Idiosyncratic Volatility Puzzle

We begin by replicating the standard portfolio-sorting results commonly cited in the idiosyncratic volatility puzzle literature to establish a baseline for comparison. Following Ang et al. (2006), we compute idiosyncratic volatility σ_n using daily returns and the Fama-French three-factor (FF3) model on a monthly basis, and define idiosyncratic variance as σ_n^2. We use FF3 rather than IPCA factors, as in our main analysis, to ensure comparability with the existing literature. Notably, using IPCA factors produces qualitatively similar results. Our cross-sectional findings are also robust to using longer estimation windows for idiosyncratic variance (available upon request). Assets are sorted into quintiles based on their lagged idiosyncratic variance $\sigma_{n,t-1}^2$, and we form five value-weighted portfolios accordingly.

The results, presented in Table 3, align with the empirical patterns documented by Ang et al. (2006). In Panel A, we report the mean and annualized volatility (in percent) of excess returns for each portfolio, as well as the market-weighted share of the assets in each group, which reflects the average market cap. We find that the portfolio with the highest idiosyncratic risk earns a significantly lower return than the lowest-risk portfolio. The annualized return spread between the lowest and highest IVol portfolios is −6.16% and statistically significant. Moreover, the lowest-IVol portfolio accounts for roughly 53% of the total market capitalization, highlighting a notable size difference—consistent with concentration. As idiosyncratic risk increases across quintiles, average market cap declines accordingly.

Consistent with prior studies, this negative return spread cannot be easily explained by traditional factor models. In Panel B, we report post-sample alpha and idiosyncratic volatility relative to the CAPM, which we use as the benchmark. The alpha spread between the lowest and highest IVol portfolios is a striking −11.14%. In Panels C and D, we repeat the analysis using the FF3 and FF5 models, respectively. Across all specifications, we consistently observe the same pattern: portfolios sorted by increasing idiosyncratic variance exhibit significantly negative alpha spreads. Additionally, higher-IVol portfolios contain smaller firms, as seen by their declining market share.

This negative relation between risk and size—derived in our theoretical framework—is central to reconciling the IVol puzzle. As we demonstrate, incorporating size-adjusted risk is critical for uncovering a positive risk-return relationship. By contrast, specifications that use IVol alone omit the negative relationship between risk and market cap, leading to a misleading negative slope between risk and return.

To further explore this argument, we extend the IRP framework using 100 value-weighted portfolios, which more precisely approximate the firm-level distribution. We treat each of these portfolios as a pseudo-asset, measuring size as the total market weight and idiosyncratic risk using portfolio-level returns. For ease of notation, we continue to refer to each portfolio as n.

We begin by testing the concentration-driven alpha using the following regression:

$$\alpha_n = \text{constant} + \hat{\gamma}\,\phi_n \sigma_n^2. \tag{17}$$

We estimate $\hat{\gamma} = 8.13$, with a t-statistic of 8.83. Although the high t-statistic partially reflects the portfolio smoothing effect, the magnitude and sign of $\hat{\gamma}$ indicate a plausible risk-aversion coefficient and a positive risk-return relation, consistent with our model.

Next, we compare this to the standard IV-only specification used in the IVol puzzle literature:

$$\alpha_n = \text{constant} + \hat{\eta}\,\sigma_n^2. \tag{18}$$

Here, we estimate $\hat{\eta} = -3.03$, with a significant t-statistic, confirming the well-known negative relationship between idiosyncratic variance and alpha in the cross-section.

In our framework, this negative estimate of $\hat{\eta}$ arises due to the negative size-risk relationship, as described in equation (11). To validate this specification, we estimate the log-linear relation:

$$\log \sigma_n \approx \text{constant} - \delta \log \phi_n. \tag{19}$$

We find $\hat{\delta} = 0.31$, confirming that larger firms tend to have lower idiosyncratic risk and echoing our theoretical assumptions.

Figure 10 plots logged size and volatility, along with the fitted linear relation. The fit is remarkably tight and highlights an avenue for future research. Substituting this size-risk relation into the alpha formula yields:

$$\alpha_n \approx \gamma \phi_n \sigma_n^2 \approx \gamma \sigma_n^{-1.23},$$

implying that alpha decreases with increasing IV when size is not explicitly accounted for. Hence, models that ignore size can yield a negative slope between alpha and IV—even when the underlying risk-return relationship is positive.

Consistent with this logic, we find that the correlation between ϕ_n and σ_n^2 is -0.66, while the correlation between $\phi_n \sigma_n^2$ and σ_n^2 is -0.78. These negative correlations explain why regressions using σ_n^2 alone yield misleading conclusions.

As discussed in Table 2, the alpha associated with SAIV is concentrated in large firms with low IV. This channel complements, rather than contradicts, the anomaly-based findings in small firms with high IV. To formally compare the explanatory power of SAIV and IV, we normalize both to have unit variance and estimate a constrained regression:

$$\alpha_n = \text{constant} + \theta\, \phi_n \sigma_n^2 + (1 - \theta)\, \sigma_n^2.$$

We estimate $\hat{\theta} = 3.03$, which is significantly greater than one, suggesting that SAIV explains a larger portion of alpha than IV alone. The implied coefficient on σ_n^2 is -2.03, consistent with an anomaly channel through IV. Intuitively, $\theta > 1$ indicates that the true priced component loads more heavily on the size-adjusted term $\phi_n \sigma_n^2$ than on raw idiosyncratic variance.

In sum, while alpha may arise through various mechanisms, the concentration in market cap and the negative size-risk relation jointly generate a negative relationship between IV and expected return—thereby helping to reconcile the IVol puzzle.

3.5 Robustness

3.5.1 Firm-level Tests

To ensure robustness, we extend our portfolio-level analysis to the firm level. While the portfolio results already highlight the importance of concentration in market cap for shaping the risk-return relationship, firm-level tests offer greater cross-sectional granularity and allow a more detailed investigation of the size distribution's role in asset pricing.

Following Ang et al. (2009), we replicate their specification for asset returns:

$$r_{n,t} = \mu_0 + \sum_{k=1}^{K} \beta_{n,k,t} \left(f_{k,t} + \mu_k \right) + \eta \sigma^2_{n,t-1} + \epsilon_{n,t}. \tag{20}$$

We then extend this to include size-adjusted idiosyncratic risk, producing the following specification:

$$r_{n,t} = \mu_0 + \sum_{k=1}^{K} \beta_{n,k,t} \left(f_{k,t} + \mu_k \right) + \gamma \phi_{n,t-1} \sigma^2_{n,t-1} + \epsilon_{n,t}. \tag{21}$$

This form emerges from extending our model's single-period competitive equilibrium to a multi-period setting, akin to Merton (1973).[7] The size-adjusted term $\phi_{n,t-1}\sigma^2_{n,t-1}$ approximates the covariance between firm-level idiosyncratic shocks $\epsilon_{n,t}$ and their market-weighted average, mirroring time-varying betas.

To compare with the IVol puzzle literature, we estimate $\hat{\eta} < 0$ and contrast it with our model's prediction of $\hat{\gamma} > 0$. We also estimate a constrained regression to compare the explanatory power of SAIV and IV:

$$r_{n,t} = \mu_0 + \sum_{k=1}^{K} \beta_{n,k,t} \left(f_{k,t} + \mu_k \right) + \theta \phi_{n,t-1} \sigma^2_{n,t-1} + (1-\theta) \sigma^2_{n,t-1} + \epsilon_{n,t}. \tag{22}$$

We estimate $\beta_{n,k,t}$ using IPCA, and idiosyncratic risk is computed from daily return volatility each month. Using the two-step Fama-MacBeth approach, as in Ang et al. (2009), we first estimate exposures and SAIV at the monthly level, then compute $\hat{\gamma}$ as the time-series average:

$$\hat{\gamma} = \frac{1}{T} \sum_{t=1}^{T} \hat{\gamma}_t.$$

We also control for lagged firm characteristics, including the book-to-market ratio and six-month momentum, as in Daniel and Titman (1997) and Jegadeesh and Titman (1993).

Panel A of Table 4 presents the results using IPCA factor exposures. Column 1 shows a

[7]We assume that $\beta_{n,k,t}, \sigma^2_{n,t}$, and other parameters evolve i.i.d. over time and are not driven by a state variable, resulting in the cross-sectional specification in (21).

significantly negative coefficient on IV ($\hat{\eta} = -1.87$, t-statistic = -1.65), which persists after controlling for market cap in Column 2. In contrast, Column 3 shows a strongly positive and significant SAIV coefficient ($\hat{\gamma} = 8.60$, t-statistic = 9.44). Column 4 includes both IV and SAIV; both remain significant ($\hat{\eta} = -2.07$, $\hat{\gamma} = 8.67$). Column 5 reports the constrained regression in (22), yielding $\hat{\theta} = 0.83$, confirming that SAIV is a significant and complementary predictor of returns alongside IV.

Panel B replicates the analysis using FF3 factors instead of IPCA. The findings are consistent. Column 1 shows a strongly negative IV effect ($\hat{\eta} = -9.69$, t-statistic = -6.76), and SAIV remains significantly positive ($\hat{\gamma} = 5.92$, t-statistic = 7.16) in Column 3. In the dual-variable specification (Column 4), both coefficients remain robust ($\hat{\eta} = -10.33$, $\hat{\gamma} = 6.15$). The constrained regression in Column 5 estimates $\hat{\theta} = 0.95$, again underscoring the dominant role of SAIV in explaining expected returns.

To complement our earlier portfolio-based double-sorting results, we divide firms into three groups based on their IV: bottom 30%, middle 40%, and top 30%. We then apply the same specifications—(20), (21), and (22)—within each IV group.

Panel A of Table 5 shows that among low-IV firms, the coefficient on IV is unexpectedly positive ($\hat{\eta} = 97.95$, t-statistic = 5.32), even after controlling for size, suggesting a distinct risk-return relation among large, low-risk firms. SAIV remains significantly positive ($\hat{\gamma} = 2.63$, t-statistic = 3.98; $\hat{\gamma} = 2.46$, t-statistic = 3.80). The constrained regression yields $\hat{\theta} = 0.62$, confirming SAIV's relevance.

For middle-IV firms (Panel B), IV loses significance, but SAIV remains strongly significant ($\hat{\gamma} = 11.55$, t-statistic = 8.10; $\hat{\gamma} = 11.40$, t-statistic = 7.90). The constrained regression yields $\hat{\theta} = 0.81$, showing that SAIV dominates in this group.

Panel C shows results for high-IV firms. IV becomes significantly negative, consistent with the IVol puzzle literature, yet SAIV continues to exert a positive and significant influence. This may reflect rare instances in which large firms temporarily exhibit high idiosyncratic volatility.

Overall, these results reaffirm the importance of size-adjusted idiosyncratic risk in explaining expected returns. While the influence of raw IV weakens or reverses in high-IV groups, SAIV remains a consistently robust predictor across firm types.

3.5.2 Subsample Evidence and Time-varying Market Concentration

We further test robustness by examining how the effects of SAIV evolve over time with changing market concentration. Using the summed monthly market weight of the ten largest firms (see Figure 1) as a proxy for concentration, we observe a declining trend from the 1970s to the early 2000s. This was driven by the expansion of listed firms during the NASDAQ boom and the proliferation of tech firms in the 1990s and 2000s.

As concentration declines, the impact of SAIV is expected to diminish. To test this, we split the sample by decade and re-estimate equations (20), (21), and (22) within each subsample. Results in Table 6 offer further insight into the temporal dynamics of risk compensation.

In Panel A (IPCA-adjusted), the SAIV coefficient (γ) remains significant across decades. However, in the 1970s—a period of rapid firm entry—the IV coefficient η becomes insignificant, consistent with our model's prediction that failing to adjust for size can obscure the true risk-return relationship. The role of SAIV weakens temporarily in the 2000s, during the tech boom, when concentration was diluted by the expansion of large firms. Panel B (FF3-adjusted) confirms these patterns.

This subsample analysis further validates our framework: as market concentration changes, so does the explanatory power of SAIV. When concentration weakens, the strength of the size-adjusted risk-return relationship fades.

In sum, our empirical evidence underscores the value of incorporating market cap when analyzing idiosyncratic risks. The documented interaction between size, idiosyncratic risk, and expected returns offers a deeper understanding of cross-sectional asset pricing in the presence of market concentration.

4 Conclusion

This paper re-examines the classical asset pricing insight that expected returns depend solely on factor exposures. We derive a closed-form expression for IRPs, assuming investors hold concentrated portfolios. The IRP increases with both idiosyncratic volatility and market capitalization. This size channel stands in sharp contrast to the size premium in Fama–French factor models, where small firms earn higher expected returns. In our setting, expected returns depend not solely on betas.

Our calibration confirms a strong negative relation between firm size and idiosyncratic volatility, consistent with well-known empirical regularities. Importantly, this size–volatility relation clarifies how idiosyncratic volatility and expected returns can be negatively correlated in the data, as documented by Ang et al. (2006, 2009).

References

Ang, Andrew, Robert J Hodrick, Yuhang Xing, and Xiaoyan Zhang, 2006, The cross-section of volatility and expected returns, *The journal of finance* 61, 259–299.

Ang, Andrew, Robert J Hodrick, Yuhang Xing, and Xiaoyan Zhang, 2009, High idiosyncratic volatility and low returns: International and further us evidence, *Journal of Financial Economics* 91, 1–23.

Axtell, Robert L, 2001, Zipf distribution of us firm sizes, *science* 293, 1818–1820.

Benhabib, Jess, Jesse Perla, and Christopher Tonetti, 2021, Reconciling models of diffusion and innovation: A theory of the productivity distribution and technology frontier, *Econometrica* 89, 2261–2301.

Breiman, Leonard, 1965, On some limit theorems similar to the arc-sin law, *Theory of Probability & Its Applications* 10, 323–331.

Byun, Sung Je, Johnathan Loudis, and Lawrence Schmidt, 2024, The idiosyncratic financial factor: An explanation for the role of size factors and the weak intertemporal risk-return relation, *Available at SSRN 3362066* .

Campbell, John Y, Martin Lettau, Burton G Malkiel, and Yexiao Xu, 2001, Have individual stocks become more volatile? an empirical exploration of idiosyncratic risk, *The journal of finance* 56, 1–43.

Chamberlain, Gary, 1983, Funds, factors, and diversification in arbitrage pricing models, *Econometrica: Journal of the Econometric Society* 1305–1323.

Chamberlain, Gary, and Michael Rothschild, 1983, Arbitrage, factor structure, and mean-variance analysis on large asset markets, *Econometrica: Journal of the Econometric Society* 1281–1304.

Connor, Gregory, 1984, A unified beta pricing theory, *Journal of Economic Theory* 34, 13–31.

Daniel, Kent, and Sheridan Titman, 1997, Evidence on the characteristics of cross sectional variation in stock returns, *the Journal of Finance* 52, 1–33.

Durrett, Rick, 2019, *Probability: theory and examples*, volume 49 (Cambridge university press).

Gabaix, Xavier, 1999, Zipf's law for cities: an explanation, *The Quarterly journal of economics* 114, 739–767.

Gabaix, Xavier, 2011, The granular origins of aggregate fluctuations, *Econometrica* 79, 733–772.

Gabaix, Xavier, and Rustam Ibragimov, 2011, Rank- 1/2: a simple way to improve the ols estimation of tail exponents, *Journal of Business & Economic Statistics* 29, 24–39.

Gabaix, Xavier, and Ralph SJ Koijen, 2024, Granular instrumental variables, *Journal of Political Economy* 132, 2274–2303.

Gabaix, Xavier, Jean-Michel Lasry, Pierre-Louis Lions, and Benjamin Moll, 2016, The dynamics of inequality, *Econometrica* 84, 2071–2111.

Goyal, Amit, and Pedro Santa-Clara, 2003, Idiosyncratic risk matters!, *The journal of finance* 58, 975–1007.

Herskovic, Bernard, Bryan Kelly, Hanno Lustig, and Stijn Van Nieuwerburgh, 2016, The common factor in idiosyncratic volatility: Quantitative asset pricing implications, *Journal of Financial Economics* 119, 249–283.

Hill, Bruce M, 1975, A simple general approach to inference about the tail of a distribution, *The annals of statistics* 1163–1174.

Hou, Kewei, and Roger K Loh, 2016, Have we solved the idiosyncratic volatility puzzle?, *Journal of Financial Economics* 121, 167–194.

Hou, Kewei, Chen Xue, and Lu Zhang, 2015, Digesting anomalies: An investment approach, *The Review of Financial Studies* 28, 650–705.

Ingersoll Jr, Jonathan E, 1984, Some results in the theory of arbitrage pricing, *The Journal of Finance* 39, 1021–1039.

Jagannathan, Ravi, and Zhenyu Wang, 1996, The conditional capm and the cross-section of expected returns, *The Journal of finance* 51, 3–53.

Jegadeesh, Narasimhan, and Sheridan Titman, 1993, Returns to buying winners and selling losers: Implications for stock market efficiency, *The Journal of finance* 48, 65–91.

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, and Hao Jiang, 2014, Tail risk and asset prices, *The Review of Financial Studies* 27, 2841–2871.

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

Kwon, Spencer Y, Yueran Ma, and Kaspar Zimmermann, 2024, 100 years of rising corporate concentration, *American Economic Review* 114, 2111–2140.

Lucas Jr, Robert E, and Benjamin Moll, 2014, Knowledge growth and the allocation of time, *Journal of Political Economy* 122, 1–51.

Luttmer, Erzo GJ, 2011, On the mechanics of firm growth, *The Review of Economic Studies* 78, 1042–1068.

Malevergne, Yannick, Pedro Santa-Clara, and Didier Sornette, 2009, Professor zipf goes to wall street, Technical report, National Bureau of Economic Research.

Merton, 1987, A simple model of capital market equilibrium with incomplete information .

Merton, Robert C, 1973, An intertemporal capital asset pricing model, *Econometrica: Journal of the Econometric Society* 867–887.

Roll, Richard, 1977, A critique of the asset pricing theory's tests part i: On past and potential testability of the theory, *Journal of financial economics* 4, 129–176.

Ross, Stephen, 1976, The arbitrage theory of capital asset pricing, *Journal of Economic Theory* 13, 341–360.

Simon, Herbert A, and Charles P Bonini, 1958, The size distribution of business firms, *The American economic review* 48, 607–617.

Van Nieuwerburgh, Stijn, and Laura Veldkamp, 2010, Information acquisition and under-diversification, *The Review of Economic Studies* 77, 779–805.

Xu, Yexiao, and Burton G Malkiel, 2003, Investigating the behavior of idiosyncratic volatility, *The Journal of Business* 76, 613–645.

Zipf, George Kingsley, 2016, *Human behavior and the principle of least effort: An introduction to human ecology* (Ravenio Books).

5 Tables and Figures

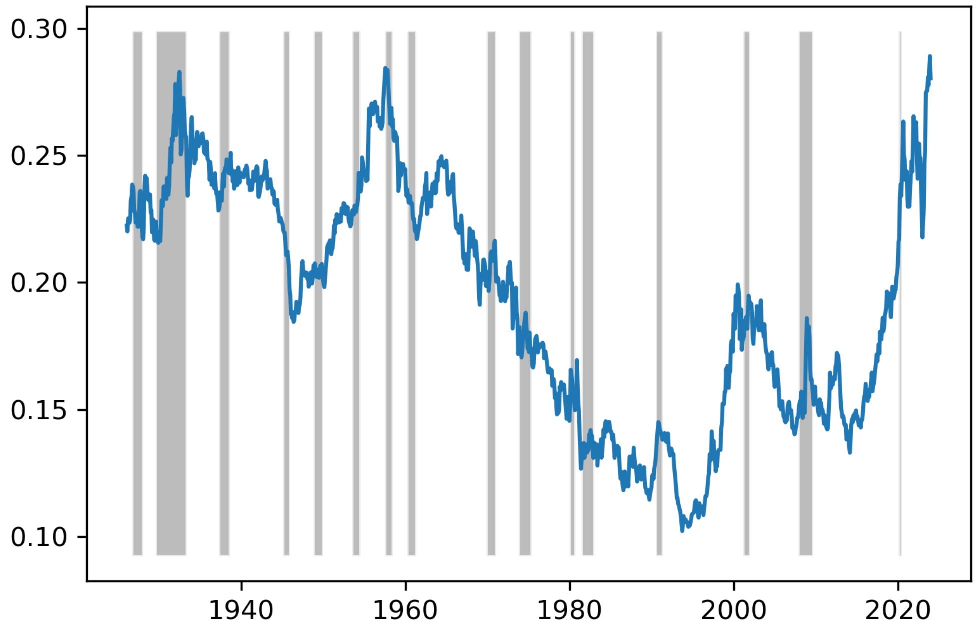
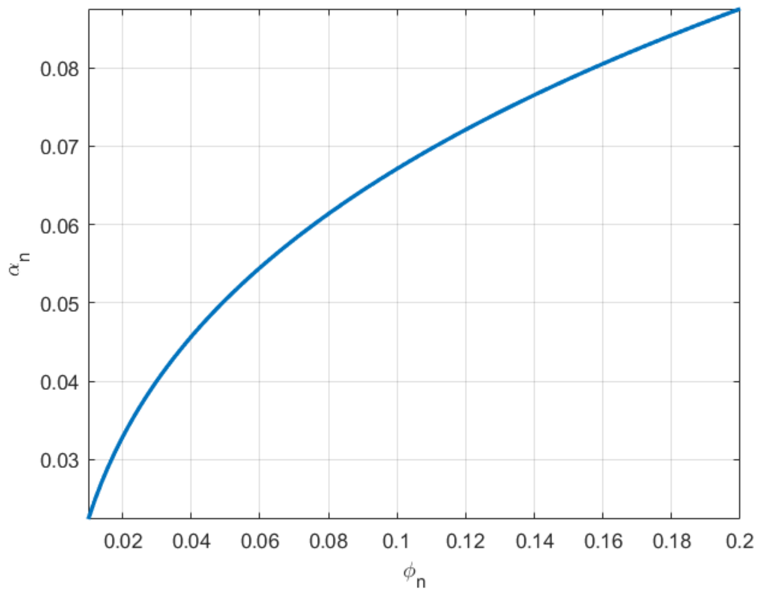
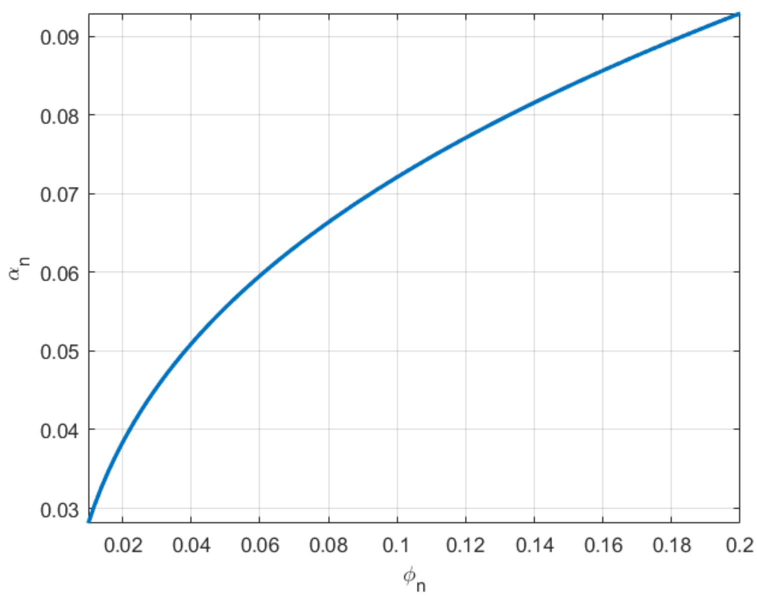


Figure 1: **Summed Market Weight of the Ten Largest Firms**
This figure plots the total market weight of the ten largest firms in the Center for Research in Security Prices (CRSP) database, measured monthly from 1927 to 2023. Shaded areas indicate NBER-defined recession periods. The magnitude of this concentration reflects a highly skewed distribution of market capitalization. For comparison, if all firms had equal market capitalization, the top ten would account for only ≈ 10/5000 = 0.2% of the total market.

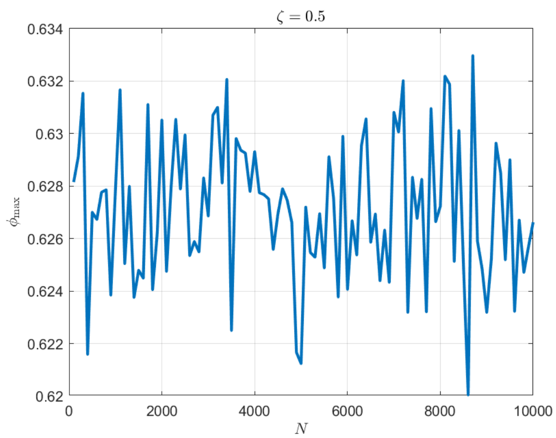


(A) Solution for $u(r^{\phi}) = -e^{-\gamma r^{\phi}}$

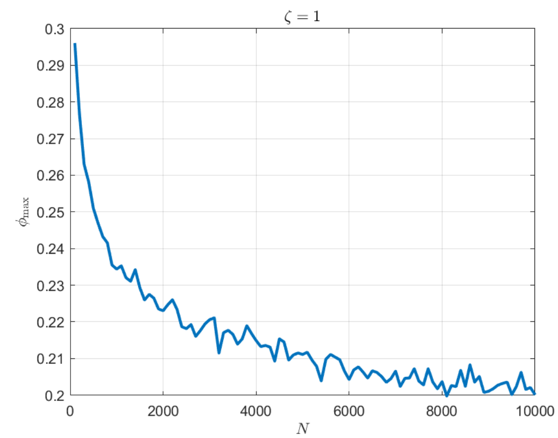


(B) Solution for $u(r^{\phi}) = \frac{(1+r^{\phi})^{1-\gamma}}{1-\gamma}$

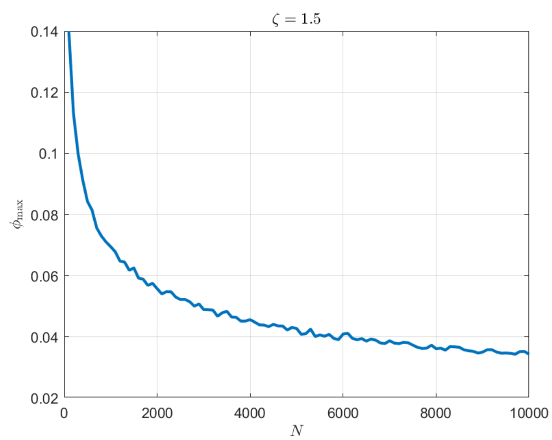
FIGURE 2: **Idiosyncratic Risk Premium and optimal portfolio Weight**
This figure illustrates the relationship between an asset's optimal portfolio weight ϕ_n and its idiosyncratic risk premium α_n, based on a tractable numerical implementation of our model. We incorporate the empirically motivated size-IV relation in equation (11) into the expression for α_n, and simulate a log-normal market environment to evaluate how idiosyncratic premia vary with firm size. Specifically, we calibrate the gross return to match empirical moments of the market portfolio via $1+r^* = e^{0.1+0.15Z}$, with $Z \sim \mathcal{N}(0,1)$. The idiosyncratic shock is modeled as $\epsilon_n = e^{-0.5\sigma_n^2+\sigma_n Z}-1$, with σ_n decaying with market cap per equation (11). Using $S = 0.2$, $\gamma = 5$, and $\delta = 0.3$, we compute α_n according to the solution in equation (8) and (9) across 10,000 Monte Carlo simulations. The resulting curve demonstrates how limited diversification in large economies leads to a rising idiosyncratic risk premium as ϕ_n increases.



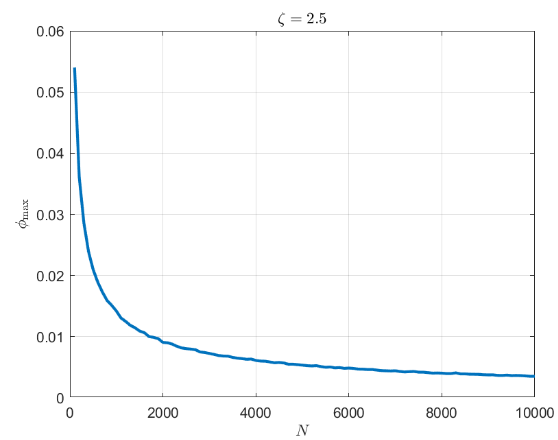
(A) $\zeta = 0.5$



(B) $\zeta = 1.0$

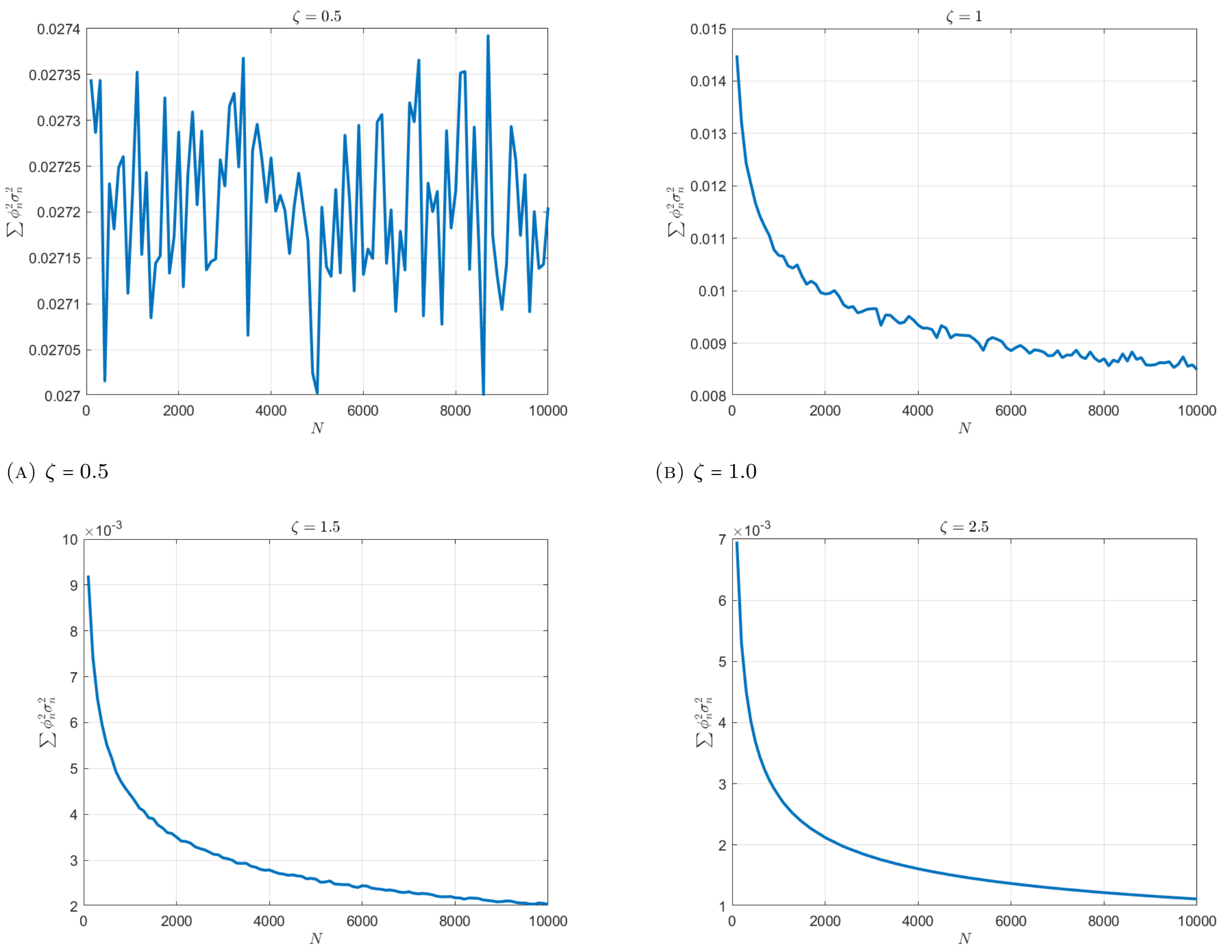


(C) $\zeta = 1.5$



(D) $\zeta = 2.5$

FIGURE 3: **Maximum portfolio weight $\boldsymbol{\phi_{\max}}$ under different Pareto exponents.** Each panel reports the average value of $\phi_{\max} = \max_n \phi_n$, where ϕ_n denotes the portfolio weight of firm n computed as the normalized market capitalization $\phi_n = X_n / \sum_j X_j$. For each value of ζ, we simulate 10,000 economies with N firms whose market capitalizations follow a Pareto$(\zeta, x_{\min} = 1)$ distribution. As ζ increases, the tail becomes thinner and $\phi_{\max}$ becomes less concentrated.



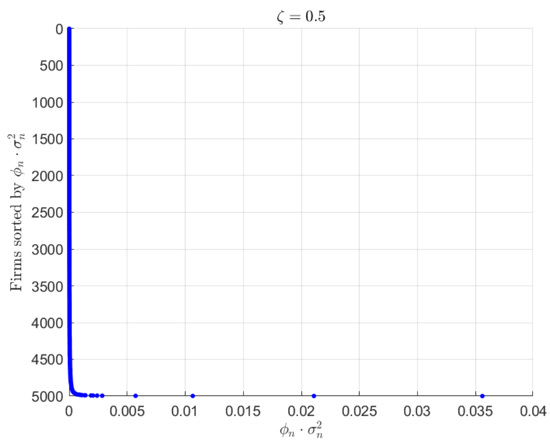
(A) $\zeta = 0.5$

(B) $\zeta = 1.0$

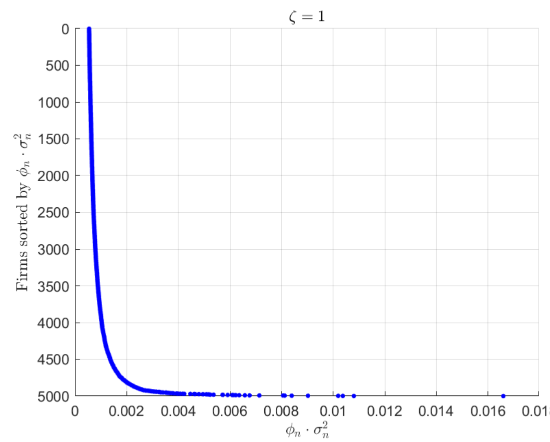
(C) $\zeta = 1.5$

(D) $\zeta = 2.5$

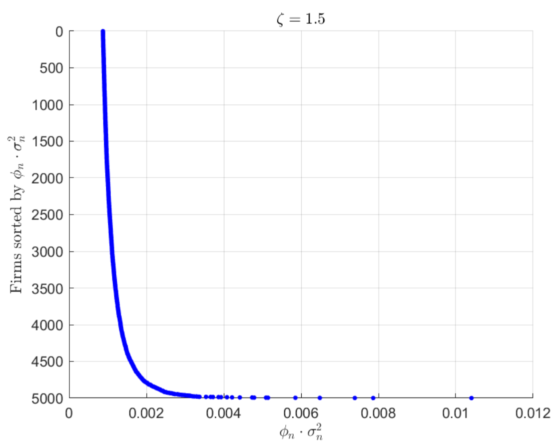
FIGURE 4: **Summed idiosyncratic risk premium $\sum \phi_n^2 \sigma_n^2$ under different Pareto tails.** Each panel reports the average value of $\sum_n \phi_n^2 \sigma_n^2$ across simulated economies with N firms, where $\phi_n = X_n / \sum_j X_j$ denotes the normalized market share and $\sigma_n = S\phi_n^{-\lambda}$ captures size-dependent risk exposure. Firm sizes X_n are drawn from a Pareto$(\zeta, 1)$ distribution. When $\zeta = 0.5$, the sum remains nearly constant as N increases, indicating persistent granularity. For higher ζ, especially $\zeta > 1$, the sum decays rapidly with N, reflecting the diminishing influence of idiosyncratic shocks and the effectiveness of diversification. These patterns validate the theoretical convergence results in Theorem 8.



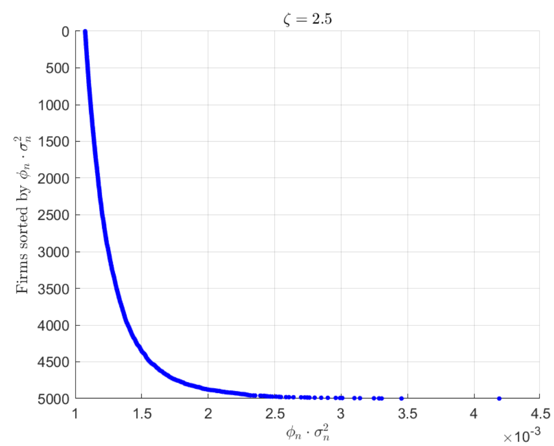
(A) $\zeta = 0.5$



(B) $\zeta = 1.0$



(C) $\zeta = 1.5$



(D) $\zeta = 2.5$

FIGURE 5: **Cross-sectional distribution of $\phi_n\sigma_n^2$ under different Pareto tails.** Each panel shows the sorted values of $\phi_n\sigma_n^2$ across $N = 5000$ firms in a single simulated economy. Based on our IRP formulation, $\phi_n\sigma_n^2$ represents the leading term in firm-level alpha, capturing the interaction between market weight and idiosyncratic volatility. Market shares ϕ_n are computed from Pareto-distributed capitalizations with tail exponent ζ and minimum value 1, and volatilities are scaled as $\sigma_n = S\phi_n^{-\lambda}$. The rank-size curves show that lower values of ζ lead to more extreme distortion concentration in the top-ranked firms, whereas thinner tails result in more diffuse pricing distortions.

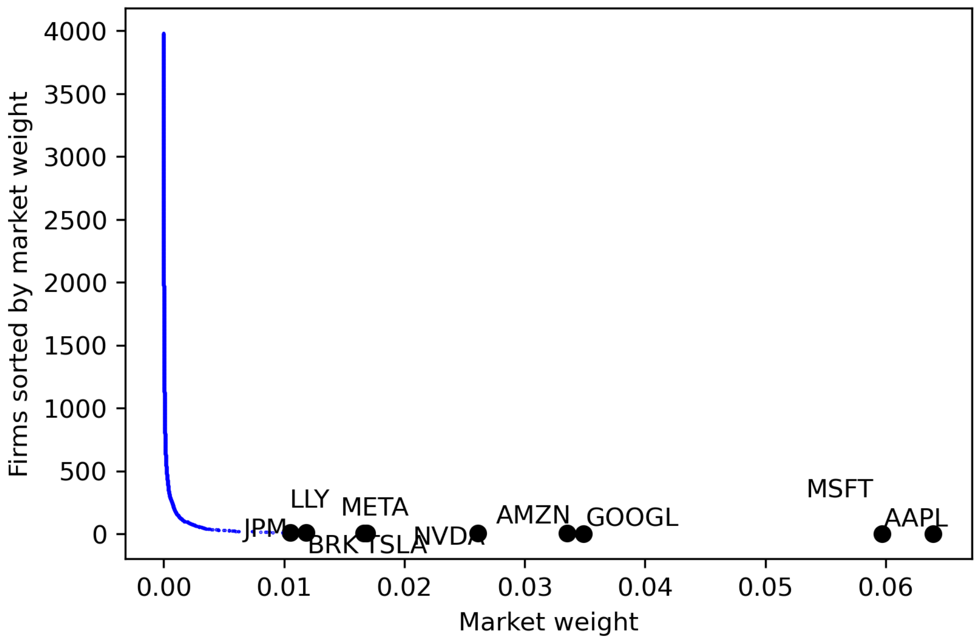


FIGURE 6: **Firm Market Weight Distribution at the End of 2023**
This figure illustrates the pronounced right-skewness in market capital, measured by each firm's relative weight in the market portfolio. The ten largest firms, highlighted in the figure, collectively account for nearly 30% of total market capitalization in the 2023 CRSP database, which includes approximately 4,000 firms.

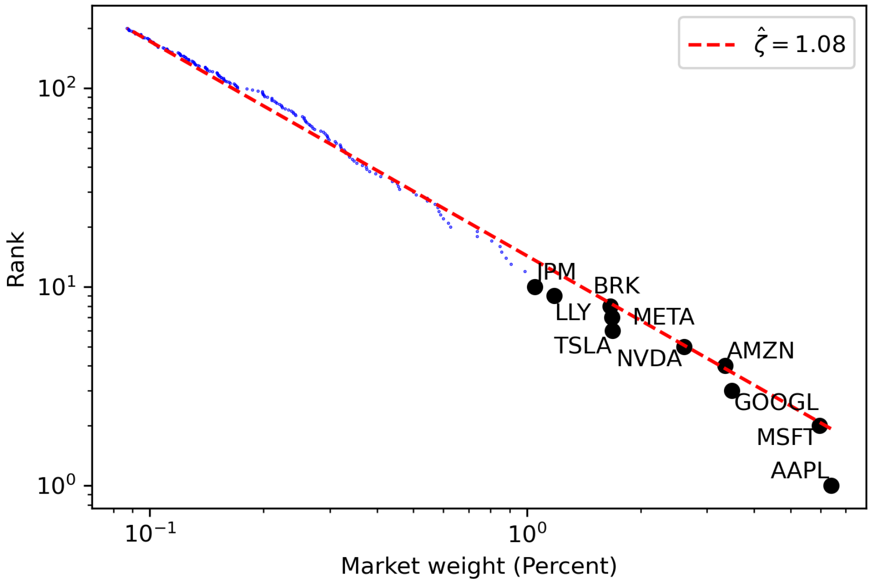


FIGURE 7: **Logged Rank-Size Plot of the Top 5% Firms in December 2023**
This figure presents the log rank-size relation for the largest 5% of firms in December 2023. The red dashed line represents the fitted relationship in equation (15) implied by a Pareto distribution with estimated tail index $\hat{\zeta} = 1.08$, i.e. the Hill estimator in Hill (1975). We apply the method in Gabaix and Ibragimov (2011) to correct the small-sample bias in Hill estimator. The ten largest firms are labeled for emphasis. Consistent with the theoretical rank-size rule, the linear relationship in log-log scale confirms a fat-tailed distribution of market capitalization—known as Zipf's Law—highlighting the disproportionate size of top firms in the cross-section.

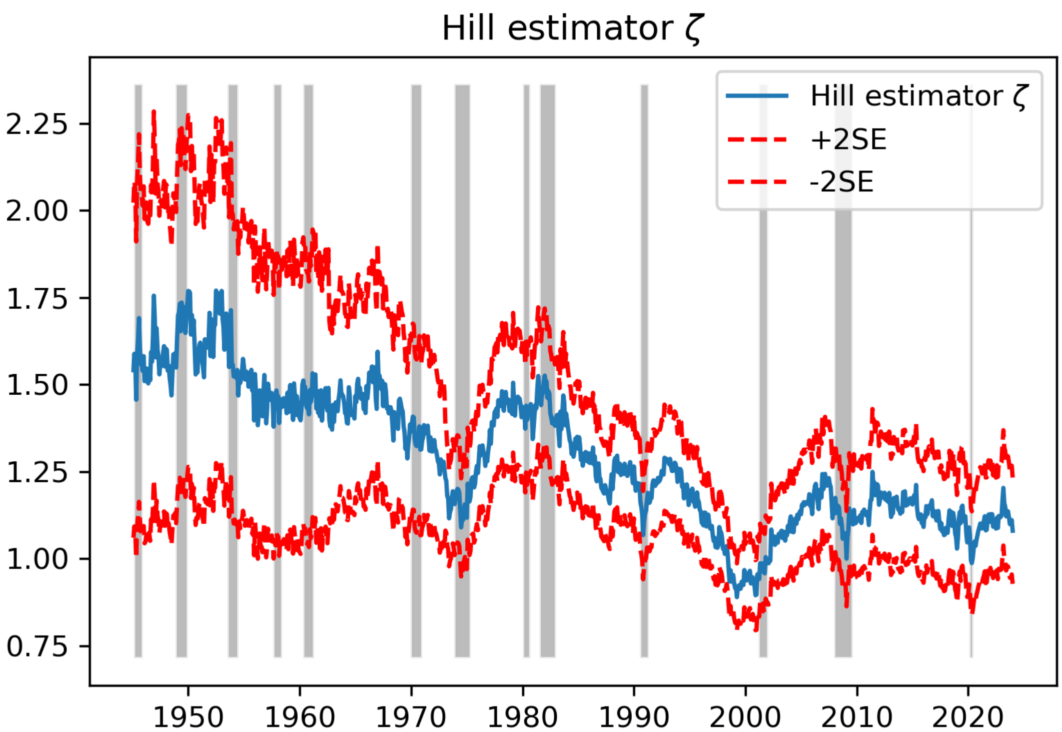


FIGURE 8: **Pareto Coefficient Estimate of Market Value per Month**
Each month, the tail parameter ζ of the Pareto distribution is estimated using the Hill estimator, as discussed in Hill (1975). We apply the method in Gabaix and Ibragimov (2011) to correct the small-sample bias in Hill estimator. This estimation focuses on the largest 5% of firms to balance the Hill estimator's bias-variance trade-off. The estimated ζ_t is depicted by a blue line, while its confidence interval (calculated as +/- two standard errors of ζ_t, assuming a maximum likelihood estimation) is represented by two red lines. NBER recession periods are indicated by shaded areas.

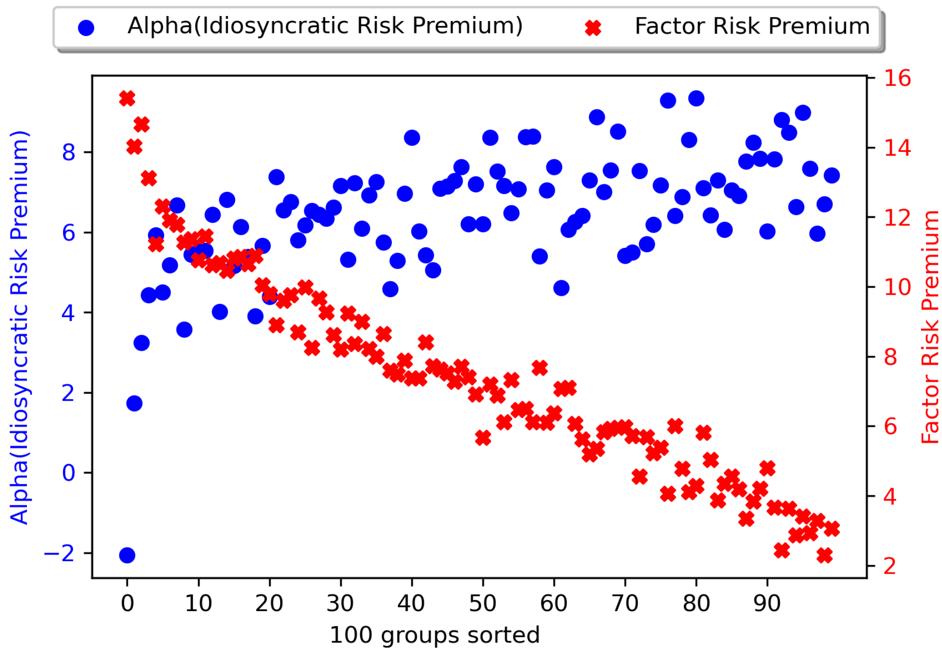


FIGURE 9: **Alpha and Factor Risk Premium of 100 Portfolios Sorted by Size-adjusted Idiosyncratic Variance.**
This figure presents the relationship between alpha and factor risk premium across 100 portfolios sorted by size-adjusted idiosyncratic variance (SAIV). The portfolios are constructed by sorting all firms into 100 percentiles based on their SAIV in the previous month, where SAIV is defined as the product of a firm's market weight and its idiosyncratic variance (IV). Portfolios are ordered from low to high SAIV. The blue dots represent the estimated alpha (idiosyncratic risk premium) for each portfolio, while the red crosses indicate the factor risk premium, measured using the Instrumented Principal Component Analysis (IPCA) approach.

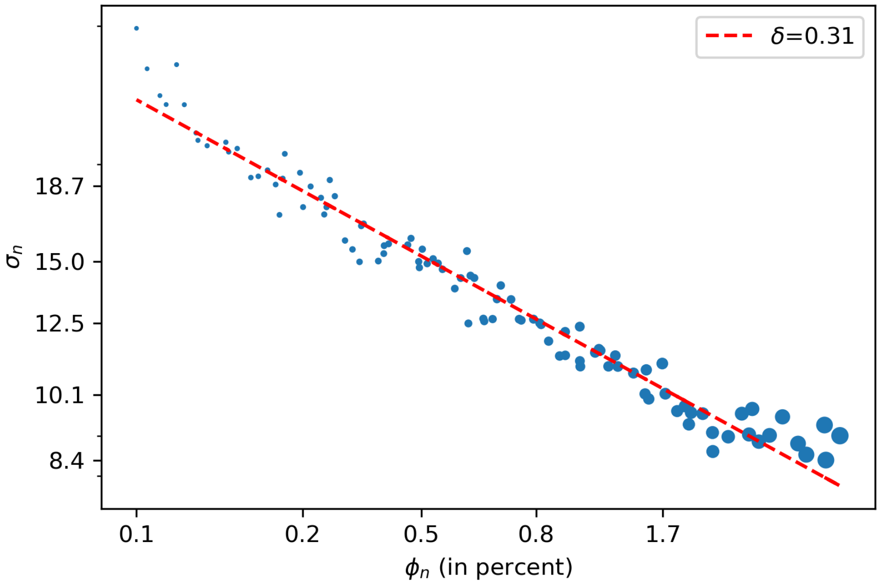


FIGURE 10: **Negative Relation between Market Weight and Idiosyncratic Variance.**
This figure illustrates the negative relationship between market weight ϕ_n and idiosyncratic volatility σ_n using 100 value-weighted portfolios sorted by idiosyncratic variance (IV). The portfolios are constructed by sorting all firms into 100 percentiles based on their IV in the previous month. Portfolio size ϕ_n is measured as the summed market weight of all firms within each portfolio, while idiosyncratic volatility σ_n is computed from the portfolio-level return residuals relative to the Fama-French three-factor model. The scatter plot shows the observed data, while the red dashed line represents the fitted regression,$\log \phi_n \approx \text{constant} - \delta \log \sigma_n$. We estimate a $\delta = 0.31$.

Table 1: **Evidence of granularity over decades.**
This table presents the names of the ten largest firms in 1920s-2020s until the end 2023. The size is measured by a company's average market weight in each period. We report the summed weight and also the total number of firms in each period.

	1920	1930	1940	1950
1	STANDARD OIL N J(0.03)	GENERAL ELECTRIC CO(0.03)	GENERAL MOTORS CORP(0.05)	GENERAL MOTORS CORP(0.05)
2	STANDARD OIL CALIFORNIA(0.02)	STANDARD OIL N J(0.04)	DU PONT & CO(0.04)	STANDARD OIL N J(0.05)
3	GENERAL ELECTRIC CO(0.03)	GENERAL MOTORS CORP(0.05)	STANDARD OIL N J(0.04)	DU PONT & CO(0.04)
4	WOOLWORTH F W CO(0.02)	CONSOLIDATED GAS CO NY(0.02)	GENERAL ELECTRIC CO(0.02)	GENERAL ELECTRIC CO(0.02)
5	GENERAL MOTORS CORP(0.05)	PHILADELPHIA CO(0.02)	COCA COLA CO(0.01)	STANDARD OIL CALIFORNIA(0.02)
6	PENNSYLVANIA RAILROAD CO(0.02)	DU PONT & CO(0.03)	TEXAS CO(0.02)	TEXAS CO(0.02)
7	NEW YORK CENT RR CO(0.02)	PENNSYLVANIA RAILROAD CO(0.01)	PROCTER & GAMBLE CO(0.01)	GULF OIL CORP(0.02)
8	SOUTHERN PACIFIC CO(0.01)	NEW YORK CENT RR CO(0.01)	STANDARD OIL IND(0.01)	COCA COLA CO(0.00)
9	UNITED STATES STEEL CORP(0.02)	UNITED STATES STEEL CORP(0.01)	WOOLWORTH F W CO(0.01)	STANDARD OIL IND(0.01)
10	UNION PACIFIC RAILROAD CO(0.01)	STANDARD OIL CALIFORNIA(0.02)	KENNECOTT COPPER CORP(0.01)	KENNECOTT COPPER CORP(0.01)
summed weight	0.23	0.24	0.22	0.24
total number	762	919	1019	1215
	1960	1970	1980	1990
1	STANDARD OIL N J(0.04)	IBM(0.05)	IBM(0.04)	EXXON CORP(0.02)
2	GENERAL MOTORS CORP(0.04)	GENERAL MOTORS CORP(0.02)	EXXON CORP(0.02)	IBM(0.01)
3	DU PONT & CO(0.02)	STANDARD OIL N J(0.03)	STANDARD OIL IND(0.01)	GENERAL ELECTRIC CO(0.02)
4	IBM(0.05)	EASTMAN KODAK CO(0.02)	MOBIL CORP(0.01)	AMOCO CORP(0.01)
5	GENERAL ELECTRIC CO(0.02)	TEXACO INC(0.01)	GENERAL MOTORS CORP(0.01)	MERCK & CO INC(0.01)
6	TEXACO INC(0.02)	XEROX CORP(0.01)	ATLANTIC RICHFIELD CO(0.01)	BRISTOL MYERS SQUIBB CO(0.01)
7	GULF OIL CORP(0.01)	MINNESOTA MINING & MFG CO(0.01)	STANDARD OIL CALIFORNIA(0.01)	DU PONT & CO(0.01)
8	EASTMAN KODAK CO(0.01)	GULF OIL CORP(0.01)	SHELL OIL CO(0.01)	BELLSOUTH CORP(0.01)
9	STANDARD OIL CALIFORNIA(0.01)	GENERAL ELECTRIC CO(0.01)	GENERAL ELECTRIC CO(0.01)	MOBIL CORP(0.01)
10	MINNESOTA MINING & MFG CO(0.01)	AVON PRODUCTS INC(0.01)	TEXACO INC(0.01)	WAL MART STORES INC(0.01)
summed weight	0.23	0.18	0.14	0.11
total number	2995	6718	10365	12376
	2000	2010	2020	2023
1	MICROSOFT CORP(0.02)	EXXON MOBIL CORP(0.02)	APPLE INC(0.06)	APPLE INC(0.06)
2	CISCO SYSTEMS INC(0.01)	MICROSOFT CORP(0.02)	MICROSOFT CORP(0.05)	MICROSOFT CORP(0.06)
3	GENERAL ELECTRIC CO(0.02)	WAL MART STORES INC(0.01)	AMAZON COM INC(0.04)	ALPHABET INC(0.03)
4	INTEL CORP(0.01)	PROCTER & GAMBLE CO(0.01)	ALPHABET INC(0.03)	AMAZON COM INC(0.03)
5	EXXON MOBIL CORP(0.03)	BERKSHIRE HATHAWAY INC DEL(0.02)	BERKSHIRE HATHAWAY INC DEL(0.02)	NVIDIA CORP(0.03)
6	WAL MART STORES INC(0.02)	APPLE INC(0.03)	FACEBOOK INC(0.01)	TESLA INC(0.02)
7	IBM(0.01)	JOHNSON & JOHNSON(0.01)	JPMORGAN CHASE & CO(0.01)	META PLATFORMS INC(0.02)
8	CITIGROUP INC(0.02)	IBM(0.01)	JOHNSON & JOHNSON(0.01)	BERKSHIRE HATHAWAY INC DEL(0.02)
9	MERCK & CO INC(0.01)	JPMORGAN CHASE & CO(0.01)	VISA INC(0.01)	LILLY ELI & CO(0.01)
10	ORACLE CORP(0.01)	WELLS FARGO & CO NEW(0.01)	WALMART INC(0.01)	JPMORGAN CHASE & CO(0.01)
summed weight	0.16	0.15	0.24	0.29
total number	10449	6892	5652	3982

Table 2: **Five Portfolios Double-sorted by Size-adjusted Idiosyncratic Variance and Idiosyncratic Variance.**
This table reports the performance of five portfolios sorted by SAIV, the size-adjusted idiosyncratic variance, which is calculated as the product of market weight and idiosyncratic variance (IV). Panel A presents the full sample results, while Panels B, C, and D show results within low, medium, and high IV deciles, respectively. For each portfolio, we report the average return, volatility, alpha estimated using Instrumented Principal Component Analysis (IPCA), t-statistics, market weight, and SAIV. The portfolios are formed based on last-month SAIV rankings, with quintile 5 (Dec5) representing the highest SAIV firms. All summary statistics are reported at an annualized frequency.

Panel A: Five Portfolios sorted by SAIV, full sample						
	Dec1	Dec2	Dec3	Dec4	Dec5	H-L
Meam	11.51	10.28	9.52	7.90	6.26	-5.25
Volatility	15.67	17.26	17.17	16.40	15.71	10.35
α_{IPCA}	0.18	1.72	2.61	2.46	2.91	2.73
T-stat	0.16	1.97	3.50	3.11	1.95	1.84
Mkt Weight	0.02	0.03	0.06	0.13	0.76	
SAIV	0.00	0.01	0.02	0.03	0.27	
Panel B: Five Portfolios sorted by SAIV, low IV decile						
	Dec1	Dec2	Dec3	Dec4	Dec5	H-L
Mean	10.36	9.85	9.26	7.95	6.81	-3.55
Volatility	13.50	15.70	15.37	14.90	14.37	10.30
α_{IPCA}	-1.70	0.58	1.75	1.71	2.66	4.36
T-stat	-1.37	0.55	1.87	1.73	1.63	2.85
Mkt Weight	0.00	0.01	0.03	0.08	0.58	
SAIV	0.00	0.00	0.01	0.03	0.17	
Panel C: Five Portfolios sorted by SAIV, medium IV decile						
	Dec1	Dec2	Dec3	Dec4	Dec5	H-L
Mean	15.51	12.67	11.36	9.87	6.78	-8.73
Volatility	20.37	21.33	20.86	20.87	19.58	13.10
α_{IPCA}	3.44	3.27	3.91	4.34	3.68	0.24
T-stat	2.88	4.00	6.37	7.09	2.72	0.13
Mkt Weight	0.00	0.00	0.01	0.02	0.19	
SAIV	0.00	0.01	0.01	0.03	0.16	
Panel D: Five Portfolios sorted by SAIV, high IV decile						
	Dec1	Dec2	Dec3	Dec4	Dec5	H-L
Mean	12.46	8.80	6.94	4.87	2.10	-10.36
Volatility	25.70	27.67	27.97	28.12	26.47	15.88
α_{IPCA}	3.72	3.10	2.71	2.93	2.47	-1.25
T-stat	2.91	2.79	3.06	3.67	1.36	-0.58
Mkt Weight	0.00	0.00	0.00	0.01	0.06	
SAIV	0.01	0.01	0.02	0.03	0.19	

TABLE 3: **Portfolios Sorted by Idiosyncratic Variance.**
This table presents the performance of five portfolios sorted by idiosyncratic variance (IV), following the methodology of Ang et al. (2006). Panel A reports summary statistics, including mean excess returns, volatility, and market weight for each portfolio. Panels B, C, and D display alphas estimated relative to the CAPM, Fama-French three-factor (FF3), and five-factor (FF5) models, respectively, along with corresponding t-statistics and the square root of IV used in each specification. Portfolios are formed based on last-month IV rankings, with quintile H representing the highest IV firms. All summary statistics are reported at an annualized frequency.

Panel A: Summary Statistics						
	L	2	3	4	H	H-L
Mean	7.09	7.40	8.14	6.45	0.93	-6.16
Volatility	13.70	16.66	20.01	24.11	29.03	21.93
Mkt Weight	0.53	0.24	0.13	0.07	0.03	0.00
Panel B: alpha relative to CAPM						
	L	2	3	4	H	H-L
α_{CAPM}	1.35	0.30	-0.24	-3.07	-9.79	-11.14
T-stat	2.23	0.62	-0.30	-1.99	-4.79	4.37
$\sqrt{IV}_{CAPM}$	3.98	3.93	5.91	10.43	15.65	18.77
Panel C: alpha relative to FF3						
	L	2	3	4	H	H-L
α_{FF3}	1.08	0.12	-0.07	-2.50	-9.26	-10.34
T-stat	2.29	0.25	-0.09	-2.14	-6.19	5.59
$\sqrt{IV}_{FF3}$	3.12	3.86	5.06	7.67	11.40	13.38
Panel D: alpha relative to FF5						
	L	2	3	4	H	H-L
α_{FF5}	-0.05	-0.37	1.19	0.09	-5.02	-4.97
T-stat	-0.11	-0.67	1.74	0.10	-4.12	3.30
$\sqrt{IV}_{FF5}$	2.74	3.79	4.84	6.76	9.83	11.23

TABLE 4: **Fama-MacBeth Results, Individual Asset Level**

This table reports the results of firm-level Fama-MacBeth regressions examining the relationship between expected returns and idiosyncratic variance (IV) and size-adjusted idiosyncratic variance (SAIV). Panel A presents results controlling for IPCA factor exposures, while Panel B controls for FF3 factor exposures. The regressions follow the specification in Equations (20) and (21), estimating the coefficients on IV ($\hat{\eta}$) and SAIV ($\hat{\gamma}$) separately and jointly. The final column reports results from the constrained-least-square (CLS) model in Equation (22), decomposing the risk-return relationship into IV and SAIV components using a constrained estimation. Control variables include factor loadings and firm characteristics including the lagged book-to-market ratio $\mathrm{b/m_{i,t-1}}$ and momentum factor $\mathrm{mom_{i,t-1}}$, calculated as the sum of returns over the past six months.

Panel A: Cross-sectional regression, control for IPCA factors					
	$r_{n,t}$	$r_{n,t}$	$r_{n,t}$	$r_{n,t}$	$r_{n,t}$ (constrained)
IV	-1.87	-1.92		-2.07	0.17
T-stat	-1.65	-1.69		-1.85	5.82
Mkt weight		-0.37	-1.94	-1.96	-3.95
T-stat		-3.16	-7.10	-7.19	-22.13
SAIV			8.60	8.67	0.83
T-stat			9.44	9.52	28.90
Panel B: Cross-sectional regression, control for FF3 factors					
	$r_{n,t}$	$r_{n,t}$	$r_{n,t}$	$r_{n,t}$	$r_{n,t}$ (constrained)
IV	-9.69	-9.87		-10.33	0.05
T-stat	-6.76	-6.98		-7.19	1.91
Mkt weight		-0.28	-1.02	-1.33	-4.16
T-stat		-1.86	-3.13	-4.27	-21.83
SAIV			5.92	6.15	0.95
T-stat			7.16	7.48	34.20

TABLE 5: **Fama-MacBeth Results, Individual Asset Level, with Three IV Groups.**
This table extends the firm-level analysis by categorizing firms into three IV-based groups: low-IV (bottom 30%), middle-IV (middle 40%), and high-IV (top 30%). Within each group, we estimate the cross-sectional regressions specified in Equations (20), (21), and (22) (the constrained-least-square (CLS)), assessing the relationship between expected returns, idiosyncratic variance (IV), and size-adjusted idiosyncratic variance (SAIV). Panel A presents results for low-IV firms, Panel B for middle-IV firms, and Panel C for high-IV firms. The final column in each panel reports the constrained regression decomposing the effects of IV and SAIV. Control variables include factor loadings and firm characteristics including the lagged book-to-market ratio $\mathrm{b/m_{i,t-1}}$ and momentum factor $\mathrm{mom_{i,t-1}}$, calculated as the sum of returns over the past six months.

Panel A: Cross-sectional regression, low-IV firms					
	$r_{n,t}$	$r_{n,t}$	$r_{n,t}$	$r_{n,t}$	$r_{n,t}$ (constrained)
IV	97.95	95.01		88.00	0.38
T-stat	5.32	5.27		4.97	19.09
Mkt weight		-0.11	-0.55	-0.44	-1.98
T-stat		-1.36	-3.16	-2.62	-19.56
SAIV			2.63	2.46	0.62
T-stat			3.98	3.80	31.26
Panel B: Cross-sectional regression, middle-IV firms					
	$r_{n,t}$	$r_{n,t}$	$r_{n,t}$	$r_{n,t}$	$r_{n,t}$ (constrained)
IV	-1.54	-4.70		-10.73	0.19
T-stat	-0.24	-0.76		-1.72	7.33
Mkt weight		-1.12	-5.66	-5.62	-10.34
T-stat		-2.65	-5.59	-5.54	-16.73
SAIV			11.55	11.40	0.81
T-stat			8.10	7.90	31.31
Panel C: Cross-sectional regression, high-IV firms					
	$r_{n,t}$	$r_{n,t}$	$r_{n,t}$	$r_{n,t}$	$r_{n,t}$ (constrained)
IV	-1.67	-1.47		-3.24	-0.21
T-stat	-1.20	-1.00		-2.20	-3.49
Mkt weight		-4.16	-40.33	-42.64	-49.09
T-stat		-1.64	-6.97	-7.41	-14.04
SAIV			34.41	35.35	1.21
T-stat			10.36	10.59	19.75

Table 6: **Fama-MacBeth Results, Individual Asset Level, Sub-sample Results Separated by Decades.** This table examines the temporal dynamics of granularity in firm size by estimating the cross-sectional regressions specified in Equations (20), (21), and (22) separately for each decade. Panel A reports results controlling for IPCA factor exposures, while Panel B controls for FF3 factor exposures. The coefficients $\hat{\eta}$ and $\hat{\gamma}$ capture the impact of idiosyncratic variance (IV) and size-adjusted idiosyncratic variance (SAIV), respectively, on expected returns across different periods. The estimated $\hat{\theta}$ in the constrained-least-square (CLS) model reflects the relative contribution of SAIV versus IV in explaining risk premiums over time.

Panel A: Fama-Macbeth regression by decades, control for IPCA factors

	1960	1970	1980	1990	2000	2010	2020
$\hat{\eta}$	4.17	-1.65	-5.52	-2.57	-2.34	-2.24	-1.07
T-stat	0.50	-0.74	-4.86	-4.05	-2.02	-1.85	-0.52
$\hat{\gamma}$	11.26	13.49	11.71	-1.03	5.64	11.94	6.23
T-stat	4.73	6.42	5.95	-0.56	2.34	4.21	2.42
$\hat{\theta}$	0.84	0.87	1.02	0.72	0.64	0.90	0.90
T-stat	10.64	14.96	17.01	10.62	8.76	9.88	7.62
total firm number	2995	6718	10365	12376	10449	6892	5652

Panel B: Fama-Macbeth regression by decades, control for FF3 factors

	1960	1970	1980	1990	2000	2010	2020
$\hat{\eta}$	-20.17	-9.32	-18.24	-6.45	-6.21	-3.30	-3.75
T-stat	-2.15	-2.74	-8.71	-4.75	-3.83	-2.00	-2.15
$\hat{\gamma}$	8.82	8.67	7.52	-2.99	4.55	9.81	5.47
T-stat	5.05	4.47	4.20	-1.56	1.93	4.08	2.27
$\hat{\theta}$	1.00	0.91	1.21	0.87	0.73	0.94	1.05
T-stat	12.54	14.38	21.06	12.16	10.11	13.07	11.47
total firm number	2995	6718	10365	12376	10449	6892	5652

Appendix A Derivations of Factor and Idiosyncratic Risk Premium

Proof of Theorem 1. The optimality condition for the optimal portfolio is

$$\mathbb{E}[u'(r^*)(r_n - r_f)] = \mathbb{E}[u'(r^*)(\mu_n + \beta_n' f + \epsilon_n)] = 0,$$

where

$$r^* - r_f = \mu^* + (\beta^*)' f + \epsilon^*,$$

and

$$\mu^* = \sum_{n=1}^{N} \phi_n^* \mu_n, \quad \beta^* = \sum_{n=1}^{N} \phi_n^* \beta_n, \quad \epsilon^* = \sum_{n=1}^{N} \phi_n^* \epsilon_n.$$

Reorganizing terms yields

$$\mu_n = \mathbb{E}[r_n - r_f] = -\frac{1}{\mathbb{E}[u'(r^*)]}\mathbb{E}[u'(r^*)(\beta_n' f + \epsilon_n)] = \beta_n' \lambda + \alpha_n,$$

where

$$\lambda = -\frac{1}{\mathbb{E}[u'(r^*)]}\mathbb{E}[u'(r^*) f] \quad \text{and} \quad \alpha_n = -\frac{1}{\mathbb{E}[u'(r^*)]}\mathbb{E}[u'(r^*)\epsilon_n].$$

Proof of Theorem 2. Our diversification definition as $N \to \infty$ is compatible with taking expectations and covariances. We therefore derive the limiting result at $N \to \infty$ by setting $\epsilon^* = 0$ and following the same steps as in the proof of Theorem 1. In this case, r^* depends only on f, so

$$\lim_{N\to\infty} \mathbb{E}[u'(r^*)\epsilon_n] = \lim_{N\to\infty} \mathbb{E}[u'(r^*)\mathbb{E}[\epsilon_n \mid r^*]] = \lim_{N\to\infty} \mathbb{E}[u'(r^*)] \cdot 0 = 0.$$

The first equality follows from the Law of Iterated Expectations, and the second equality reflects the factor structure. Thus

$$\mu_n = -\lim_{N\to\infty} \frac{1}{\mathbb{E}[u'(r^*)]}\mathbb{E}[u'(r^*)(\beta_n' f + \epsilon_n)] = \beta_n' \lambda,$$

with

$$\lambda = -\frac{1}{\mathbb{E}[u'(r^*)]}\mathbb{E}[u'(r^*) f].$$

The Constant Relative Risk Aversion Condition

In Assumption 5, we assume constant relative risk aversion in the limiting case. We consider return-based utility in either exponential or power format and show that both forms are induced by a wealth-based utility. For CARA utility,

$$-u(W(N)) = -e^{-\Gamma(N)W(N)} = -e^{-\Gamma(N)W_0(N)(1+r^\phi)} = -e^{-\Gamma(N)W_0(N)} \cdot e^{-\Gamma(N)W_0(N)r^\phi}.$$

We consider an economy with N assets, where the initial wealth $W_0(N)$ can increase with N

because the total supply may grow with the size of the economy. In the case $N \to \infty$, we require

$$\lim_{N\to\infty} \Gamma(N)W_0(N) = \gamma.$$

This condition implies that when total asset supply increases with N, absolute risk aversion absorbs this change and converges to a constant relative risk aversion γ. The convergence condition ensures consistency with Assumption 3, which requires bounded relative risk aversion.

For CRRA utility, constant relative risk aversion is imposed by construction, and the wealth-based utility is proportional to the return-based utility:

$$\frac{W^{1-\gamma}}{1-\gamma} = \frac{W_0^{1-\gamma}(1+r^{\phi})^{1-\gamma}}{1-\gamma}.$$

Proof of Theorem 3. Consider

$$\alpha_n = -\frac{1}{\mathbb{E}[e^{-\gamma r^*}]}\mathbb{E}[e^{-\gamma r^*}\epsilon_n].$$

Recall that

$$r^* - r_f = \mu^* + (\beta^*)'f + \epsilon^*.$$

By independence of idiosyncratic shocks and factors and by the Law of Iterated Expectations,

$$\mathbb{E}[e^{-\gamma r^*}] = \mathbb{E}\Big[e^{-\gamma(\mu^*+(\beta^*)'f)}\Big]\prod_{m=1}^{N}\mathbb{E}[e^{-\gamma\phi_m\epsilon_m}],$$

$$\begin{aligned}\mathbb{E}[e^{-\gamma r^*}\epsilon_n] &= \mathbb{E}\Big[e^{-\gamma(\mu^*+(\beta^*)'f+\sum_{m\neq n}\phi_m\epsilon_m)}\,\mathbb{E}[e^{-\gamma\phi_n\epsilon_n}\epsilon_n \mid \epsilon_n]\Big] \\ &= \mathbb{E}\Big[e^{-\gamma(\mu^*+(\beta^*)'f)}\Big]\prod_{m\neq n}\mathbb{E}[e^{-\gamma\phi_m\epsilon_m}]\mathbb{E}[e^{-\gamma\phi_n^*\epsilon_n}\epsilon_n].\end{aligned}$$

Therefore,

$$\alpha_n = -\frac{1}{\mathbb{E}[e^{-\gamma r^*}]}\mathbb{E}[e^{-\gamma r^*}\epsilon_n] = -\mathbb{E}[e^{-\gamma\phi_n^*\epsilon_n}\epsilon_n].$$

Let $m(\eta)$ denote the moment generating function of ϵ_n:

$$m(\eta) = \mathbb{E}[e^{\eta\epsilon_n}].$$

Then

$$m'(\eta) = \mathbb{E}[e^{\eta\epsilon_n}\epsilon_n],$$

so

$$\alpha_n = -\frac{m'(\eta_n)}{m(\eta_n)}, \quad \text{where} \quad \eta_n = -\gamma\phi_n^*,$$

which is the market price of ϵ_n risk.

Note that

$$\frac{m'(\eta)}{m(\eta)} = \frac{d}{d\eta}\log m(\eta).$$

The function $\log m(\eta)$ is the cumulant generating function:

$$K(\eta) = \log m(\eta) \quad \Rightarrow \quad K'(\eta) = \frac{m'(\eta)}{m(\eta)}.$$

The Taylor expansion of $K(\eta)$ yields

$$K(\eta) = \sum_{i=1}^{\infty} \frac{\kappa_n(i)}{i!}\eta^i \quad \Rightarrow \quad \frac{m'(\eta)}{m(\eta)} = \frac{d}{d\eta}K(\eta) = \sum_{i=1}^{\infty} \frac{\kappa_n(i)}{(i-1)!}\eta^{i-1},$$

where $\kappa_n(i)$ is the i-th cumulant of ϵ_n. Since ϵ_n has zero mean, substituting $\eta_n = \gamma\phi_n^*$ gives

$$\alpha_n = \sum_{i=1}^{\infty}(-1)^{i+1}\frac{\kappa_n(i+1)}{i!}(\gamma\phi_n^*)^i.$$

Proof of Theorem 4. Let $1 + r^* = R + \phi_n\epsilon_n$ with R independent of ϵ_n. Then

$$(1 + r^*)^{-\gamma} = R^{-\gamma}\left(1 + \frac{\phi_n\epsilon_n}{R}\right)^{-\gamma}.$$

Using the generalized binomial expansion for real exponents,

$$(1 + x)^{-\gamma} = \sum_{k=0}^{\infty} \frac{(-1)^k(\gamma)_k}{k!}x^k,$$

which is valid for $|x| < 1$. In our case, we require the portfolio gross return to be positive,

$$1 + r^* = R + \phi_n\epsilon_n > 0.$$

When $R > 0$,

$$|\phi_n\epsilon_n| < |R| \quad \Rightarrow \quad \left|\frac{\phi_n\epsilon_n}{R}\right| < 1.$$

The condition $R > 0$ indicates that the impact of $\phi_n\epsilon_n$ on the aggregate portfolio return is not large enough to determine the sign of $R + \phi_n\epsilon_n$. Under this condition,

$$(1 + r^*)^{-\gamma} = \sum_{k=0}^{\infty} \frac{(-1)^k(\gamma)_k}{k!} \cdot \phi_n^k \cdot \epsilon_n^k \cdot R^{-\gamma-k}.$$

Multiplying both sides by ϵ_n and taking expectations gives

$$\mathbb{E}[(1 + r^*)^{-\gamma}\epsilon_n] = \sum_{k=0}^{\infty} \frac{(-1)^k(\gamma)_k}{k!}\phi_n^k\mathbb{E}[\epsilon_n^{k+1}]\mathbb{E}[R^{-\gamma-k}].$$

Since $\mathbb{E}[\epsilon_n] = 0$, the $k = 0$ term vanishes, and

$$\alpha_n = -\frac{\mathbb{E}[(1 + r^*)^{-\gamma}\epsilon_n]}{\mathbb{E}[(1 + r^*)^{-\gamma}]} = \sum_{k=1}^{\infty} \frac{(-1)^{k+1}(\gamma)_k}{k!}\phi_n^k\mathbb{E}[\epsilon_n^{k+1}]\frac{\mathbb{E}[R^{-\gamma-k}]}{\mathbb{E}[(1 + r^*)^{-\gamma}]}.$$

Appendix B Derivation Using a Pareto Distribution

Building on the derivation of the idiosyncratic risk premium, we now examine the behavior of a firm's market weight under a Pareto distribution, which generates a granularity-driven idiosyncratic risk premium. We first restate the thin-tailed case in Theorem 5.

Proof of Theorem 5. Recall that

$$\phi_n = \frac{X_n}{\sum_{n=1}^{N} X_n}.$$

Hence

$$\lim_{N\to\infty} \sum_{n=1}^{N} \phi_n^2 = \lim_{N\to\infty} \sum_{n=1}^{N} \frac{X_n^2}{\left(\sum_{n=1}^{N} X_n\right)^2} = \lim_{N\to\infty} \frac{1}{N} \frac{\frac{1}{N}\sum_{n=1}^{N} X_n^2}{\left(\frac{1}{N}\sum_{n=1}^{N} X_n\right)^2}.$$

If the first and second moments of X_n are finite, then

$$\lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} X_n^2 = \mathbb{E}[X^2], \qquad \lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} X_n = \mathbb{E}[X].$$

Therefore

$$\lim_{N\to\infty} \sum_{n=1}^{N} \phi_n^2 = \lim_{N\to\infty} \frac{1}{N} \frac{\mathbb{E}[X^2]}{\mathbb{E}[X]^2} = 0. \tag{B1}$$

When all ϵ_n have the same variance,

$$\lim_{N\to\infty} \sum_{n=1}^{N} \phi_n^2 \sigma^2 = 0.$$

A generalization follows from (11), where large firms have lower idiosyncratic variance. In this case, the impact of large firms is further diluted and the limit still converges to zero.

The Pareto Distribution

We assume that market capitalizations X_n are i.i.d. and follow a Pareto distribution with survival function

$$P(X_n > x) = \left(\frac{x}{x_{\min}}\right)^{-\zeta}, \quad x > x_{\min}. \tag{B2}$$

The Pareto distribution parsimoniously describes tail thickness through a single parameter $\zeta > 0$. The Pareto coefficient ζ determines how fast the probability that firm size exceeds $x_{\min}$ decays as $x \to \infty$. A higher Pareto coefficient ζ implies lower granularity. The i-th moment of X is

$$\begin{aligned} \mathbb{E}[X^i] &= \infty, \quad \zeta \le i, \\ \mathbb{E}[X^i] &= \frac{\zeta x_{\min}^i}{\zeta - i}, \quad \zeta > i. \end{aligned} \tag{B3}$$

When $\zeta > 2$, the distribution is thin-tailed: the first and second moments of X are finite and the diversification result in (B1) applies. A small $\zeta < 2$ implies a high probability of extremely large firms and a high degree of fat-tailedness. As a result, the moments of firm size diverge and the sample averages of X_n and X_n^2 in (B1) do not converge to finite limits.

The Negative Size–Variance Relation

Based on the Pareto distribution, in (11) we assume a negative size–variance relation

$$\sigma_n = S\phi_n^{-\delta}.$$

Note that

$$\phi_n = \frac{X_n}{\sum_{n=1}^N X_n} > \frac{x_{\min}}{\sum_{n=1}^N X_n},$$

so

$$\sigma_n = S\phi_n^{-\delta} < S\left(\frac{\sum_{n=1}^N X_n}{x_{\min}}\right)^{\delta}.$$

To ensure a finite upper bound for σ_n, $x_{\min}$ must increase at the same rate as $\sum_{n=1}^N X_n$ as $N \to \infty$.

This condition delivers tractability because the portfolio weight ϕ_n is invariant to the scale of $x_{\min}$. In other words, we can scale market capitalization by $x_{\min}$ and obtain the same results. In addition, this condition is consistent with the empirical observation that emerging markets with fewer listed firms have lower listing thresholds.

Without loss of generality, we normalize the lower bound to $x_{\min} = 1$, which matches the presentation in the main text:

$$P(X_n > x) = x^{-\zeta}, \quad x > 1, \quad \zeta < 2.$$

The convergence of $\sum X_n$ when $\zeta < 2$ follows from the Stable Law, a general convergence theorem for infinite-variance random variables. We present a textbook version (Durrett (2019), Theorem 3.8.2).

Theorem (Stable Law)**.** *Suppose $X_1, X_2, \ldots$ are i.i.d. with a distribution that satisfies:*
(i) $\lim_{x\to\infty} P(X_1 > x)/P(|X_1| > x) = \theta \in [0,1]$,
(ii) $P(|X_1| > x) = x^{-\alpha}L(x)$, *where $\alpha < 2$ and $L(x)$ is slowly varying.*
Let $S_N = \sum_{n=1}^N X_n$, $a_N = \inf\{x : P(|X_1| > x) \le N^{-1}\}$ *and* $b_N = N\mathbb{E}(X_1 \mathbf{1}_{\{|X_1| \le a_N\}})$.
As $N \to \infty$, $(S_N - b_N)/a_N \Rightarrow Y$, where Y follows a stable distribution with characteristic function

$$\varphi_Y(\mu, \sigma, \beta, \zeta, t) = \exp\{t\mu i - \sigma|t|^{\zeta}\left(1 + \beta\,\mathrm{sign}(t) w_\zeta(t) i\right)\}. \tag{B4}$$

The characteristic function is parametrized by location μ and scale σ, which are determined by the underlying Pareto law; the skewness parameter $\beta = 2\theta - 1$; and shape parameter ζ, the Pareto coefficient of X. The function $\mathrm{sign}(t)$ *is the sign function and $w_\zeta(t)$ is*

$$\begin{aligned} w_\zeta(t) &= \tan(\pi\zeta/2), \quad \zeta \neq 1, \\ w_\zeta(t) &= \pi/2 \log|t|, \quad \zeta = 1. \end{aligned}$$

We now show that (B1) fails when market capitalization follows a Pareto distribution with $\zeta < 2$. Applying the Stable Law theorem to the Pareto distribution in (B2), we have $\theta = 1$, $\alpha = \zeta$, and $L(x) = 1$, so

$$a_N = N^{1/\zeta}, \tag{B5}$$

and

$$b_N = N \int_1^{N^{1/\zeta}} \zeta x^{-\zeta}\, dx.$$

The magnitude of b_N depends on the value of ζ:

$$b_N = \begin{cases} N^{1/\zeta} + N\dfrac{\zeta}{\zeta - 1} = O(a_N) & \zeta < 1, \\ N^{1/\zeta} + N\dfrac{\zeta}{\zeta - 1} = N^{1/\zeta} + N\mathbb{E}[X] = O(N) & \zeta > 1, \\ N \log N & \zeta = 1. \end{cases} \tag{B6}$$

Using these expressions, we obtain the convergence of $\sum_{n=1}^{N} X_n$:

$$\lim_{N\to\infty} \sum_{n=1}^{N} X_n = \lim_{N\to\infty} (a_N Y_\zeta + b_N),$$

so

$$\lim_{N\to\infty} \sum_{n=1}^{N} X_n = \begin{cases} \lim\limits_{N\to\infty} N^{1/\zeta}(Y_\zeta + 1) + N\dfrac{\zeta}{\zeta - 1} & \zeta < 1, \\ \lim\limits_{N\to\infty} N^{1/\zeta}(Y_\zeta + 1) + N\mathbb{E}[X] & \zeta > 1, \\ \lim\limits_{N\to\infty} NY_\zeta + N\log N & \zeta = 1. \end{cases} \tag{B7}$$

Note that X_n^2 also follows a Pareto distribution, with coefficient $\zeta/2$, so

$$\lim_{N\to\infty} \sum_{n=1}^{N} X_n^2 = \lim_{N\to\infty} N^{2/\zeta}(Y_{\zeta/2} + 1) + N\frac{\zeta/2}{\zeta/2 - 1}. \tag{B8}$$

Based on these results,

$$\lim_{N\to\infty} \sum_{n=1}^{N} \phi_n^2 = \lim_{N\to\infty} \sum_{n=1}^{N} \frac{X_n^2}{\left(\sum_{n=1}^{N} X_n\right)^2} = \begin{cases} \dfrac{Y_{\zeta/2}}{(Y_\zeta)^2} & \zeta < 1, \\ \lim\limits_{N\to\infty} \dfrac{Y_{\zeta/2}}{\left(Y_\zeta + N^{1-1/\zeta}\mathbb{E}[X]\right)^2} = 0 & \zeta > 1, \\ \lim\limits_{N\to\infty} \dfrac{Y_{\zeta/2}}{\left(Y_\zeta + \log N\right)^2} = 0 & \zeta = 1. \end{cases}$$

Here Y_ζ and $Y_{\zeta/2}$ are stable random variables with characteristic function given in (B4). We suppress constants in the numerator and denominator by shifting the means of Y_ζ and $Y_{\zeta/2}$.

We now proceed to prove Theorem 6. Note that

$$\epsilon^* = \sum_{n=1}^{N} \phi_n \epsilon_n = \frac{\sum_{n=1}^{N} X_n \epsilon_n}{\sum_{n=1}^{N} X_n}. \tag{B9}$$

The numerator $\sum_{n=1}^{N} X_n \epsilon_n$ also converges to a stable law with tail index ζ provided that ϵ_n is independent of X_n and satisfies $\mathbb{E}[|\epsilon_n|^{\zeta+\rho}] < \infty$ for some $\rho > 0$.

We first state a general result characterizing the tail behavior of the product of a heavy-tailed variable with a light-tailed one.

Theorem (Tail behavior of the product process)**.** *Let $X \geq 0$ and $\epsilon \in \mathbb{R}$ be independent random variables. Suppose X is regularly varying with tail index $\zeta > 0$, i.e.,*

$$\mathbb{P}(X > x) = x^{-\zeta} L(x), \quad x \to \infty,$$

for some slowly varying function $L(x)$. Suppose further that

$$\mathbb{E}[|\epsilon|^{\zeta+\rho}] < \infty \quad \textit{for some } \rho > 0.$$

Then, as $x \to \infty$,

$$\begin{aligned}\mathbb{P}(X\epsilon > x) &\sim \mathbb{E}[\epsilon_+^{\zeta}] \cdot \mathbb{P}(X > x),\\ \mathbb{P}(X\epsilon < -x) &\sim \mathbb{E}[\epsilon_-^{\zeta}] \cdot \mathbb{P}(X > x),\end{aligned}$$

where $\epsilon_+ := \epsilon \mathbf{1}_{\{\epsilon>0\}}$ and $\epsilon_- := -\epsilon \mathbf{1}_{\{\epsilon<0\}}$.

This result is a two-sided extension of the theorem in Breiman (1965), which focuses on the positive tail $\mathbb{P}(X\epsilon > x)$ for $x > 0$. The argument for $X\epsilon < -x$ is symmetric and examines the negative tail. A fully rigorous proof requires tools such as the uniform convergence theorem for slowly varying functions and is available upon request.

We now use this result to prove the convergence of $\epsilon^* = \sum_{n=1}^{N} \phi_n \epsilon_n$ when ϵ_n is i.i.d.

Proof of Theorem 6. By the above result, $X_n\epsilon_n$ inherits regular variation with tail index ζ. Its right and left tail weights are proportional to

$$c_+ = \mathbb{E}[\epsilon_+^{\zeta}], \qquad c_- = \mathbb{E}[\epsilon_-^{\zeta}].$$

Define normalized tail ratio parameters

$$\theta_+ := \frac{c_+}{c_+ + c_-}, \qquad \theta_- := \frac{c_-}{c_+ + c_-}. \tag{B10}$$

Applying this to $X_n\epsilon_n$ yields a two-sided extension of the Stable Law:

$$\lim_{x\to\infty} \frac{P(X_n\epsilon_n > x)}{P(|X_n\epsilon_n| > x)} = \theta_+, \tag{B11}$$

$$\lim_{x\to\infty} \frac{P(X_n\epsilon_n < -x)}{P(|X_n\epsilon_n| > x)} = \theta_-. \tag{B12}$$

The partial sums converge as

$$\sum_{n=1}^{N} X_n\epsilon_n = a_N^{(\epsilon)} Y_\zeta^{(\epsilon)} + b_N^{(\epsilon)}, \tag{B13}$$

where

$$a_N^{(\epsilon)} = \inf\{x > 0 : \mathbb{P}(|X_n\epsilon_n| > x) \leq 1/N\} = \left(\mathbb{E}[|\epsilon|^{\zeta}]\right)^{1/\zeta} a_N,$$

and

$$b_N^{(\epsilon)} \coloneqq N \cdot \mathbb{E}[X_n \epsilon_n \mathbf{1}_{\{|X_n \epsilon_n| \le a_N^{(\epsilon)}\}}],$$

while $Y_\zeta^{(\epsilon)}$ is a stable random variable with skewness parameter $\beta = 2\theta_+ - 1 = \frac{c_+ - c_-}{c_+ + c_-}$.

Note that

$$b_N^{(\epsilon)} = N \cdot \mathbb{E}\big[\epsilon_n \, \mathbb{E}[X_n \mathbf{1}_{\{|X_n \epsilon_n| \le a_N^{(\epsilon)}\}} \mid \epsilon_n]\big].$$

Define $T(\epsilon) \coloneqq \mathbb{E}[X_n \mathbf{1}_{\{X_n \le a_N^{(\epsilon)}/|\epsilon|\}}]$. Then

$$b_N^{(\epsilon)} = N \cdot \mathbb{E}[\epsilon_n \, T(\epsilon_n)].$$

Since $X_n \sim \text{Pareto}(1, \zeta)$ with density $f_X(x) = \zeta x^{-\zeta-1}$,

$$T(\epsilon) = \zeta \int_1^{a_N^{(\epsilon)}/|\epsilon|} x x^{-\zeta-1} dx = \zeta \int_1^{a_N^{(\epsilon)}/|\epsilon|} x^{-\zeta} dx.$$

Thus

$$T(\epsilon) = \begin{cases} \dfrac{\zeta}{1-\zeta} \left(\left(\dfrac{a_N^{(\epsilon)}}{|\epsilon|} \right)^{1-\zeta} - 1 \right), & \zeta \neq 1, \\ \log \left(\dfrac{a_N^{(\epsilon)}}{|\epsilon|} \right) = \log N + \log \mathbb{E}[|\epsilon|] - \log |\epsilon|, & \zeta = 1. \end{cases}$$

When $\zeta \neq 1$,

$$b_N^{(\epsilon)} = N \cdot \mathbb{E}\left[\epsilon_n \cdot \frac{\zeta}{1-\zeta} \left(\left(\frac{a_N^{(\epsilon)}}{|\epsilon_n|} \right)^{1-\zeta} - 1 \right) \right].$$

Splitting terms,

$$b_N^{(\epsilon)} = N \cdot \frac{\zeta}{1-\zeta} \big(a_N^{(\epsilon)}\big)^{1-\zeta} \mathbb{E}\big[\epsilon |\epsilon|^{\zeta-1}\big] - N \cdot \frac{\zeta}{1-\zeta} \mathbb{E}[\epsilon].$$

Since $\mathbb{E}[\epsilon_n] = 0$, we have

$$b_N^{(\epsilon)} = b_\zeta \cdot a_N, \quad b_\zeta = \frac{\zeta}{1-\zeta} \mathbb{E}\big[\epsilon|\epsilon|^{\zeta-1}\big] \big(\mathbb{E}[|\epsilon|^\zeta]\big)^{1/\zeta - 1}.$$

Similarly, when $\zeta = 1$,

$$b_N^{(\epsilon)} = N \cdot \mathbb{E}[-\epsilon \log(|\epsilon|)].$$

Therefore

$$\lim_{N \to \infty} \sum_{n=1}^N X_n \epsilon_n = \lim_{N \to \infty} a_N \big(a_\zeta Y_\zeta^{(\epsilon)} + b_\zeta \big), \tag{B14}$$

where

$$a_\zeta = \big(\mathbb{E}[|\epsilon|^\zeta]\big)^{1/\zeta}, \quad b_\zeta = \begin{cases} \dfrac{\zeta}{1-\zeta} \mathbb{E}\big[\epsilon|\epsilon|^{\zeta-1}\big] \big(\mathbb{E}[|\epsilon|^\zeta]\big)^{1/\zeta-1}, & \zeta \neq 1, \\ \mathbb{E}[-\epsilon \log(|\epsilon|)], & \zeta = 1. \end{cases}$$

Notably, when ϵ is symmetric, $b_\zeta = 0$. Since $a_\zeta Y_\zeta^{(\epsilon)} + b_\zeta$ is again stable, we rename it as $Y_\zeta^{(\epsilon)}$ for notational simplicity. Combining this with

$$\sum_{n=1}^{N} X_n \sim a_N Y_\zeta + b_N,$$

we obtain

$$\lim_{N\to\infty} \epsilon^* = \begin{cases} \dfrac{Y_\zeta^{(\epsilon)}}{Y_\zeta} & \zeta < 1, \\ \lim\limits_{N\to\infty} \dfrac{Y_\zeta^{(\epsilon)}}{Y_\zeta + N^{1-1/\zeta}\mathbb{E}[X]} = 0 & \zeta > 1, \\ \lim\limits_{N\to\infty} \dfrac{Y_\zeta^{(\epsilon)}}{Y_\zeta + \log N} = 0 & \zeta = 1. \end{cases} \tag{B15}$$

Although ϵ^* is non-degenerate and converges to a random variable when $\zeta < 1$, the limiting random fraction $\frac{Y_\zeta^{(\epsilon)}}{Y_\zeta}$ has a well-defined mean. Moreover, ϵ^* must have zero mean since it is a convex combination of ϵ_n with thin-tailed distributions:

$$|\epsilon^*| \leq \sum_{n=1}^{N} |\phi_n||\epsilon_n| \leq \max_n |\epsilon_n|.$$

Hence

$$\mathbb{E}[\epsilon^*] = \mathbb{E}\left[\sum_{n=1}^{N} \phi_n \epsilon_n\right] = \sum_{n=1}^{N} \mathbb{E}[\phi_n]\mathbb{E}[\epsilon_n] = \mathbb{E}[1] \cdot 0 = 0.$$

The non-degenerate limit when $\zeta < 1$ is driven by the fact that idiosyncratic shocks of dominant firms persist in the aggregate and prevent convergence to zero. To illustrate the impact of large firms, we prove Theorem 7.

Proof of Theorem 7. The result is a direct implication of the Fisher–Tippett–Gnedenko theorem. By definition,

$$F(a_N x) = 1 - (a_N x)^{-\zeta} = 1 - \frac{1}{N} x^{-\zeta}.$$

The limiting distribution of $X_{\max}/a_N$ is

$$\lim_{N\to\infty} P(X_{\max}/a_N \leq x) = \lim_{N\to\infty} F(a_N x)^N = \lim_{N\to\infty} \left(1 - \frac{1}{N} x^{-\zeta}\right)^N = e^{-x^{-\zeta}}.$$

Combining this with the convergence of $\sum_{n=1}^{N} X_n$ yields

$$\lim_{N\to\infty} \phi_{\max} = \lim_{N\to\infty} \frac{X_{\max}}{\sum_{n=1}^{N} X_n} = \begin{cases} \dfrac{F_\zeta}{Y_\zeta} & \zeta < 1, \\ \lim\limits_{N\to\infty} \dfrac{F_\zeta}{Y_\zeta + \log N} & \zeta = 1, \\ \lim\limits_{N\to\infty} \dfrac{F_\zeta}{Y_\zeta + N^{1-1/\zeta}\mathbb{E}[X]} & \zeta > 1, \end{cases} \tag{B16}$$

where F_ζ denotes the Fréchet limit of the maximum.

In other words, when fat tails are sufficiently strong, the portfolio weight of the largest firm does not converge to zero, and the idiosyncratic shocks of large firms do not wash out even as $N \to \infty$. We now consider the case with the negative size–variance relation (11).

Extension of Theorem 6 with a negative size–variance relation
With

$$\sigma_n = S\phi_n^{-\delta}, \quad 0 \le \delta < 0.5, \qquad \Rightarrow \qquad \phi_n \sigma_n = S\left(\frac{X_n}{\sum_{n=1}^{N} X_n}\right)^{1-\delta},$$

set $\epsilon_n = \sigma_n e_n$, so

$$\epsilon^* = S\sum_{n=1}^{N} \phi_n^{1-\delta} e_n = S\frac{\sum_{n=1}^{N} X_n^{1-\delta} e_n}{\left(\sum_{n=1}^{N} X_n\right)^{1-\delta}}.$$

Similarly,

$$\lim_{N\to\infty} \sum_{n=1}^{N} X_n^{1-\delta} e_n = \lim_{N\to\infty} N^{(1-\delta)/\zeta} Y^{(e)}_{\zeta/(1-\delta)},$$

and combining with

$$\lim_{N\to\infty} \sum_{n=1}^{N} X_n = \begin{cases} N^{1/\zeta} Y_\zeta + N\dfrac{\zeta}{\zeta - 1} & \zeta < 1, \\ N^{1/\zeta} Y_\zeta + N\mathbb{E}[X] & \zeta > 1, \\ NY_\zeta + N\log N & \zeta = 1, \end{cases}$$

and absorbing the bounded constant S into the stable random variable, we obtain

$$\lim_{N\to\infty} \epsilon^* = \begin{cases} \dfrac{Y^{(e)}_{\zeta/(1-\delta)}}{Y_\zeta} & \zeta < 1, \\ \lim\limits_{N\to\infty} \dfrac{Y^{(e)}_{\zeta/(1-\delta)}}{\left(Y_\zeta + N^{1-1/\zeta}\mathbb{E}[X]\right)^{1-\delta}} = 0 & \zeta > 1, \\ \lim\limits_{N\to\infty} \dfrac{Y^{(e)}_{\zeta/(1-\delta)}}{\left(Y_\zeta + \log N\right)^{1-\delta}} = 0 & \zeta = 1. \end{cases} \tag{B17}$$

This result is presented in the main text in terms of $\mathrm{Var}(\epsilon^*)$, as in Theorem 8.

Proof of Theorem 8. The sum converges as follows:

$$\lim_{N\to\infty}\sum_{n=1}^{N}\phi_n^2\sigma_n^2 = S^2\frac{\lim_{N\to\infty}\sum_{n=1}^{N}X_n^{2-2\delta}}{\left(\lim_{N\to\infty}\sum_{n=1}^{N}X_n\right)^{2-2\delta}}.$$

The process $X_n^{2-2\delta}$ has Pareto coefficient $\zeta/(2-2\delta)$. If $\delta < 0.5$ (the empirically relevant case), we only need $\zeta < 1$; if the variance decays faster, we require $\zeta < 2-2\delta$. The qualitative convergence result does not change much when $\delta \geq 0.5$. We have

$$\lim_{N\to\infty}\sum_{n=1}^{N}X_n^{2-2\delta} = \begin{cases} N^{(2-2\delta)/\zeta}Y_{\zeta/(2-2\delta)} + N\dfrac{\zeta/(2-2\delta)}{\zeta/(2-2\delta)-1} & \zeta < 2-2\delta, \\ N^{(2-2\delta)/\zeta}Y_{\zeta/(2-2\delta)} + N\mathbb{E}[X^{2-2\delta}] & 2-2\delta < \zeta < 2, \\ NY_{\zeta/(2-2\delta)} + N\log N & \zeta = 2-2\delta. \end{cases}$$

Combining with

$$\lim_{N\to\infty}\sum_{n=1}^{N}X_n = \begin{cases} N^{1/\zeta}Y_\zeta + N\dfrac{\zeta}{\zeta-1} & \zeta < 1, \\ N^{1/\zeta}Y_\zeta + N\mathbb{E}[X] & \zeta > 1, \\ NY_\zeta + N\log N & \zeta = 1, \end{cases}$$

and letting the stable random variable absorb the constant S^2, we obtain

$$\lim_{N\to\infty}\sum_{n=1}^{N}\phi_n^2\sigma_n^2 = S^2\frac{\lim_{N\to\infty}\sum_{n=1}^{N}X_n^{2-2\delta}}{\left(\lim_{N\to\infty}\sum_{n=1}^{N}X_n\right)^{2-2\delta}} = \begin{cases} \dfrac{Y_{\zeta/(2-2\delta)}}{(Y_\zeta)^2} & \zeta < 1, \\ \lim_{N\to\infty}\dfrac{Y_{\zeta/(2-2\delta)}}{\left(Y_\zeta + N^{1-1/\zeta}\mathbb{E}[X]\right)^{2-2\delta}} = 0 & 1 < \zeta < 2-2\delta, \\ \lim_{N\to\infty}\dfrac{\mathbb{E}[X^{2-2\delta}]}{N^{2-2\delta-1}\mathbb{E}[X]} = 0 & 2-2\delta < \zeta < 2, \\ \lim_{N\to\infty}\dfrac{Y_{\zeta/(2-2\delta)}}{\left(Y_\zeta + \log N\right)^{2-2\delta}} = 0 & \zeta = 1. \end{cases}$$