PCA-Based Risk Models for U.S. Equity Portfolio Construction

A study of PCA and clustering on S&P 500 returns, 2000 to 2026

Rose Rahimi

Abstract

This project applies Principal Component Analysis (PCA) and clustering to daily returns of S&P 500 constituents from 2000 to 2026 to answer two questions: whether unsupervised clustering on residualized returns recovers the GICS sector taxonomy, and whether the resulting clusters yield portfolios that beat equal weight and Markowitz baselines on a risk adjusted basis. The analysis uses a survivorship bias-free, point in time universe of roughly 1,099 historical S&P 500 members and a walk forward backtest across 290 monthly rebalances.

Two findings stand out. First, clustering on market residualized returns produces partial but real agreement with GICS (mean Adjusted Rand Index (ARI) = 0.244, Normalized Mutual Information = 0.477 for k-means with K = 11); additionally stripping the sector signal (the project's Q2 panel) drops ARI by a factor of five, confirming that the residualization works as designed. Second, the naive one stock per cluster equal weight portfolio underperforms every baseline (best cluster Sharpe = 0.41, versus equal weight 0.51), while minimum variance with a Marchenko Pasteur filtered covariance matrix decisively wins (Sharpe = 0.94, with the lowest drawdown of the set). The PCA model beats the baselines, but as a covariance filter rather than as a cluster labeler. The headline reframing, from PCA as clustering to PCA as filter, is the main contribution.

1. Introduction

Modern portfolio theory begins with the observation that the dominant determinant of portfolio risk is not which individual assets one holds but how their returns covary (Markowitz, 1952). For a universe of N assets, the covariance matrix Σ is the central object of the optimization, yet on financial data Σ is notoriously hard to estimate well: a typical investible universe has roughly 500 assets while only 252 trading days of recent history are reliably representative of the current regime. The sample covariance is therefore noise dominated, and the resulting minimum variance portfolios produce extreme long short leverage with out-of-sample performance far worse than the in sample estimate would suggest. This is the curse of dimensionality for portfolio construction (Plerou et al., 1999; Qian, Hua & Sorensen, 2007, p. 64).

Industry practice solves the problem in two ways. The first is structural: classify each company by its sector under the Global Industry Classification Standard (GICS, 11 top-level sectors) and force the portfolio to hold an equal weight across sectors. The second is statistical: shrink or filter Σ to suppress noise eigenvalues before optimizing. This project asks whether the unsupervised version of the structural approach, clustering stocks based on their return histories rather than human curated labels, produces results comparable to GICS, and whether the statistical approach using random matrix theory (RMT) filtering produces portfolios that beat both GICS stratified and equal weight baselines.

Two specific research questions, originally posed in the project prospectus, guide the work:

1. Q1, Structural: Does an unsupervised PCA + clustering pipeline on residualized daily returns produce clusters that resemble the GICS sector labels, or does it identify different and potentially more informative groupings?
2. Q2, Practical: Do portfolios constructed using the data driven risk model deliver higher risk adjusted return than equal weight, GICS stratified, and Markowitz baselines?

The two questions sit on opposite sides of the same concept. If the answer to Q1 is yes (clusters track GICS), then the Q2 portfolios should approximately reproduce GICS-stratified performance. If no, then the data driven portfolios take different risk exposure, and Q2 measures whether those exposures pay off. The project answers Q1 with a qualified yes: clusters on market residualized returns show partial agreement with GICS (mean ARI = 0.244, NMI = 0.477). It answers Q2 in a manner the original framing did not predict: the data driven portfolio that wins is the Markowitz minimum variance portfolio applied to a covariance matrix denoised using random matrix theory, not any cluster based portfolio. The PCA model does outperform the baselines, but as a covariance filter rather than as a cluster labeler. This reframing, from PCA as clustering to PCA as filter, is the main empirical contribution of the project.

The remainder of this report is organized as follows. Section 2 surveys the relevant background, including the curse of dimensionality problem in covariance estimation, the Marchenko Pasteur eigenvalue distribution, and the three primary literature threads (Mantegna style hierarchical clustering, Tola et al. ultrametric filtering, and López de Prado's Hierarchical Risk Parity), plus the Avellaneda & Lee eigenportfolio perspective that shaped the methodological choices. Section 3 describes the data, the construction of three residual return panels (raw, market residual, sector residual), the walk forward PCA embedding pipeline, the clustering methods, and the portfolio construction and evaluation framework. Section 4 presents the structural results (ARI/NMI versus GICS) and the practical results (portfolio risk adjusted return and a QHS (Quin et al., 2007) style risk decomposition, including the crisis period PCR versus PCL test). Section 5 concludes, including a frank discussion of where the original framing missed and how the work should be extended.

2. Background

2.1 The covariance estimation problem and the PCA factor model duality

Markowitz (1952) formalized portfolio construction as a quadratic optimization over the covariance matrix Σ of asset returns. For N assets and T observations, the sample covariance has N(N+1)/2 distinct entries estimated from NT data points; when N is large relative to T, most of those entries are dominated by sampling noise. The resulting min-variance optimum overfits to in sample fluctuations and performs poorly out of sample, often producing extreme long/short positions that net to near zero variance in sample but blow up out of sample. This phenomenon is the curse of dimensionality in portfolio construction (Qian, Hua & Sorensen, 2007, p. 64). It motivates the entire literature on covariance matrix filtering.

PCA provides a natural angle on the problem through the spectral decomposition Σ = LPL', where L is the matrix of eigenvectors and P the diagonal matrix of eigenvalues. The decomposition has a useful duality: the principal components are simultaneously risk factors (the orthogonal directions in which return variance is maximized) and tradeable portfolios (each eigenvector L_k is a weighting that produces a portfolio whose variance is the corresponding eigenvalue). This is the textbook treatment in Qian, Hua & Sorensen (2007, Ch. 3-4) and the conceptual foundation for both interpretations of PCA explored in this project: as a feature extraction step that produces a low dimensional embedding for clustering, and as a covariance filtering step that suppresses noise eigenvalues before reassembling Σ for the min-variance optimizer.

2.2 Random matrix theory and the Marchenko Pasteur edge

Marchenko & Pastur (1967) derived the limiting eigenvalue distribution of an N × N sample correlation matrix computed from NT i.i.d. observations. As N, T → ∞ with N/T → q ∈ (0, 1] fixed, the empirical eigenvalues concentrate on the interval $[\lambda_-, \lambda_+]$ where $\lambda_\pm = (1 \pm \sqrt{q})^2$. Any eigenvalue above the upper edge λ_+ cannot be explained by pure noise and corresponds to a statistically real factor. Plerou et al. (1999) applied this to U.S. equity returns and found roughly 5-10 eigenvalues above the noise edge, with the rest indistinguishable from random correlations. The procedure gives a principled cutoff for selecting the number of factors in PCA, replacing the more arbitrary scree-plot (pca-variance plot) inspection.

2.3 Hierarchical clustering on correlation matrices

Mantegna (1999) introduced the use of hierarchical clustering for stock correlation matrices, converting each pairwise correlation ρ_{ij} to a distance via $d_{ij} = \sqrt{2(1 - \rho_{ij})}$ and building either a Minimum Spanning Tree or a single linkage dendrogram on the resulting distance matrix. Single linkage is the dominant baseline in the subsequent literature, but Marti et al. (2021) document eight known critiques, most importantly the chaining pathology, where the algorithm produces one giant cluster plus several singletons. Average and Ward linkage are more compact alternatives and are tested here.

2.4 Cluster based covariance filtering: Tola et al. (2008)

Tola, Lillo, Gallegati & Mantegna (2008), the primary source cited in the project prospectus, used hierarchical clustering as a covariance filtering technique. After building a dendrogram with single or average linkage on the Mantegna distance, they reconstructed an ultrametric correlation matrix in which each pair of stocks inherits the correlation associated with their lowest common ancestor in the tree. This matrix has only $N - 1$ distinct entries and is dramatically less noisy than the raw sample correlation. Their headline result is a reduction in the gap between predicted and realized portfolio risk: plugging the filtered correlation into the closed form Markowitz weights produces a portfolio whose realized out-of-sample volatility is closer to the optimizer's prediction than under the raw sample covariance, with average linkage outperforming single linkage. This is the cluster-as-filter baseline (MV-Tola in the results below) and the methodological anchor for the project's portfolio evaluation framework: the QHS-style risk decomposition in Section 4.3 generalizes this same predicted versus realized comparison to a per stock attribution.

2.5 Hierarchical Risk Parity: López de Prado (2016)

López de Prado (2016), the project's second primary source, argued that the min-variance optimization itself is the source of instability, regardless of how Σ is estimated, because it requires inverting Σ. His Hierarchical Risk Parity (HRP) algorithm avoids the inversion entirely. It proceeds in three steps: (1) build a single-linkage tree on the correlation distance; (2) quasi-diagonalize Σ by reordering rows and columns according to the tree's leaf order, so that similar assets sit adjacent; and (3) allocate capital top down via recursive bisection, splitting each cluster's allocation between two sub clusters in inverse proportion to their cluster variances. The original Monte Carlo experiments showed HRP delivering lower out-of-sample variance than the inverse min-variance and naive min-variance allocators. HRP is included here as a strong cluster based baseline (HRP row in the results).

2.6 The PCA-as-eigenportfolios perspective: Avellaneda & Lee (2010)

Avellaneda & Lee (2010), the outside reference required by the rubric and the paper that most directly shaped the methodological choices in this project, observed that running PCA on raw daily returns produces a first eigenvector that almost exactly matches the equal weighted S&P 500 (the market portfolio) and subsequent eigenvectors that align with sector ETFs (the so-called eigenportfolios). They proposed a statistical arbitrage strategy on the residuals after projecting out the top eigenportfolios. Two operational choices in their paper carry into this work: the 252-day rolling estimation window for the covariance matrix, and the use of sector ETFs (XLE, XLK, XLF, and so on) as the natural benchmarks for stripping the sector factor. The conceptual move (that the structurally interesting content of stock returns is what remains after removing market and sector) is the entire motivation for the Q1 and Q2 residual panels below.

2.7 The QHS textbook framework

Qian, Hua & Sorensen (2007, hereafter QHS) is the textbook used throughout this project as the canonical reference for both portfolio mathematics (Ch. 2 on minimum variance, Ch. 3 on risk decomposition) and PCA on returns (Ch. 4). QHS Eq. 2.13 gives the closed-form minimum-variance weights $w^* = \Sigma^{-1}\cdot\iota \ / \ (\iota'\cdot\Sigma^{-1}\cdot\iota)$; QHS Ch. 3 §3.2-3.3 defines Marginal Contribution to Risk (MCR), Risk Contribution (CR), and Percentage Contribution to Risk (PCR), and identifies the Herfindahl of PCRs as an effective positions count N_eff. QHS p. 71 makes the testable claim that during crises ex-post Percentage Contribution to Loss (PCL) should track ex-ante PCR; in quiet periods the two diverge. The crisis-period validation in Section 4.3 is a direct test of that claim.

3. Methods

3.1 Data

Daily adjusted close prices for S&P 500 constituents and sector ETFs were obtained from Sharadar via the Nasdaq Data Link API, covering January 1, 2000 through February 15, 2026, approximately 6,500 trading days. The universe was constructed from Sharadar's SP500 membership events table, which records adds, removes, current memberships, and historical memberships per ticker. From these events a (ticker, start, end) membership interval table was constructed: each row identifies a contiguous period during which the ticker was an active S&P 500 member, with end = NaT for currently active members. The wide price panel was then masked by these intervals so that any observation outside its ticker's membership window becomes NaN. This is the standard remedy for survivorship bias.

The resulting universe contains roughly 1,099 distinct historical and current S&P 500 members. Point-in-time correctness was verified by spot checking three known events: Lehman Brothers (Sharadar ticker LEHMQ) is present in the membership set on September 1, 2008 and absent on October 1, 2008; the price panel shows 14 non-NaN observations between September 1 and September 19, 2008 and zero non-NaN observations from October 1 onward, confirming the mask correctly drops post-bankruptcy prices. Enron (ENRNQ, end ≈ November 2001) and Silicon Valley Bank (SIVBQ, end ≈ March 2023) were verified analogously. The eleven GICS sector ETFs (XLK, XLF, XLV, XLE, XLI, XLY, XLP, XLB, XLU, XLRE, XLC) and SPY were loaded similarly. All data were persisted in ArcticDB (LMDB backend) for fast slicing across the 290 monthly rebalance dates.

3.2 Three residual return panels

Three input panels were constructed to test how much of the structural content of returns is attributable to market exposure, sector exposure, and idiosyncratic behavior:

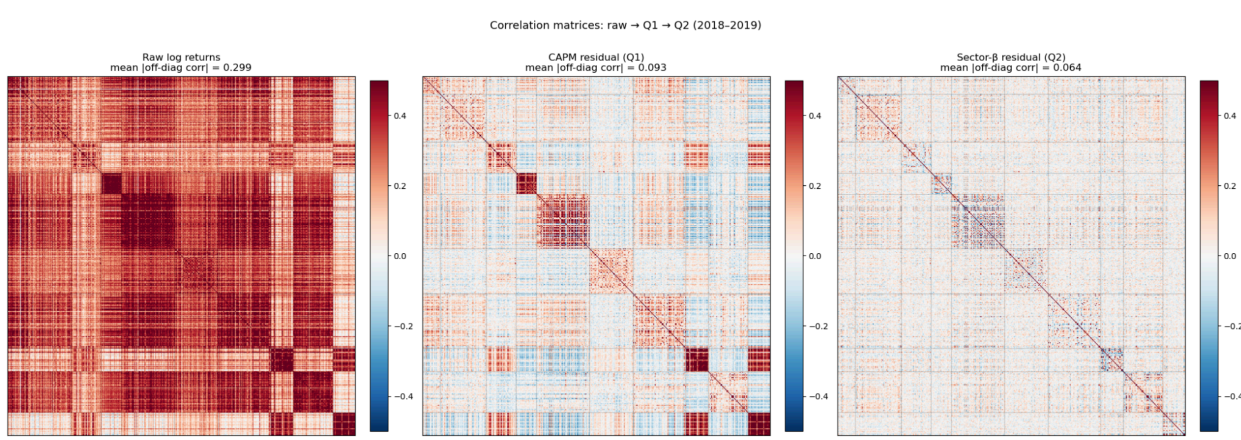
- **Q0, raw log returns:** $r_{i,t} = \log(P_{i,t} / P_{i,t-1})$. No residualization. Used as a baseline to demonstrate that clustering on raw returns is dominated by the market factor.
- **Q1, CAPM (market) residual:** for each stock i, a 252-day rolling regression against SPY produces a time varying beta $\beta_{i,t}$. The residual is $r_{i,t} - \beta_{i,t} \cdot r_{SPY,t}$. The beta is lagged by one day (.shift(1)) to eliminate look ahead bias.
- **Q2, sector-beta residual:** the same rolling regression construction, but each stock is regressed against its GICS-sector ETF (e.g., Apple against XLK, JPMorgan against XLF) rather than SPY. Strips both market and sector exposure, leaving the idiosyncratic and sub-sector component.

All three panels were cross sectionally z-scored daily (for each trading day, the row mean is subtracted and the row standard deviation divides) before any downstream use. This Avellaneda-style standardization removes day-to-day fluctuations in cross-sectional dispersion that would otherwise inflate the influence of high volatility days (such as March 2020) on the covariance estimates.

The Q2 panel has a known coverage hole. The Real Estate ETF XLRE launched on October 8, 2015 and the Communication Services ETF XLC launched on June 19, 2018, so Q2 residuals do not exist for stocks in those sectors before the corresponding ETF inception + 252 days. Affected tickers are dropped from the early rebalance windows; this incurs a small universe size shrinkage in the early years but preserves methodological purity.

Validation of residualization

Static window diagnostic: on the 2018-2019 slice, corr(PC1, SPY) = +0.058 for Q1, confirming that market exposure has been largely stripped. The per-sector analog for Q2, corr(PC1_sector, sector-ETF), is below 0.10 in absolute value for all eleven sectors. However, rolling diagnostics across regimes reveal that the corr(PC1, SPY) of Q1 swings between -0.40

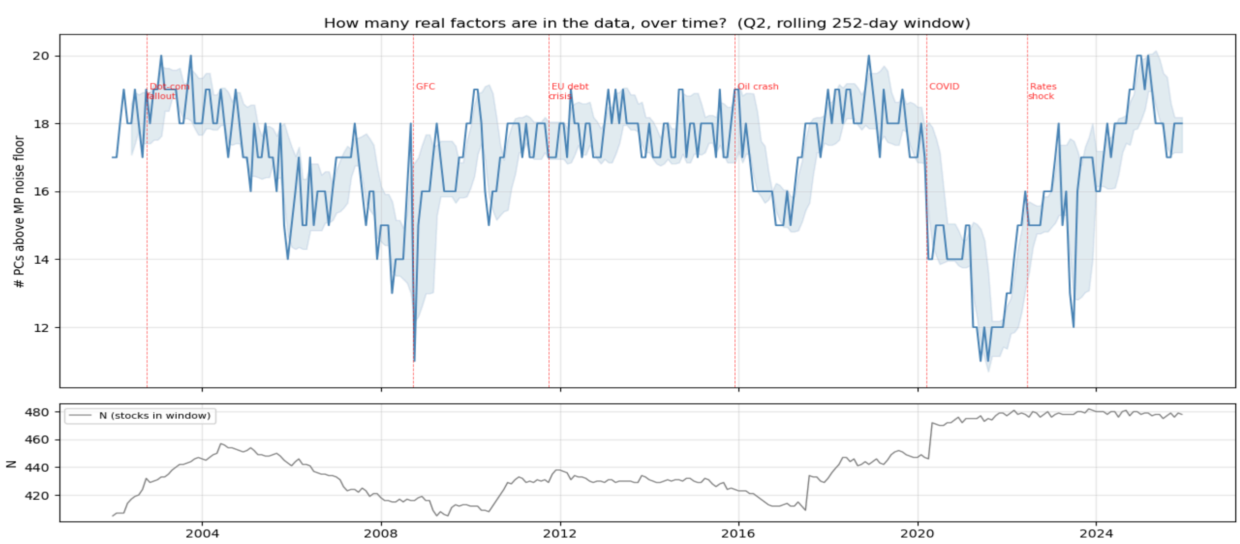


and +0.56 during the 2008-2009 Global Financial Crisis (GFC), exposing the fragility of the 252-day rolling beta when the underlying regime changes faster than the estimation window. The Q2 per sector rolling diagnostic shows analogous spikes for Financial Services and Energy during the GFC. This is a documented limitation rather than a fatal problem, but it means the clustering downstream inherits some regime dependent residual market or sector exposure during stress periods.

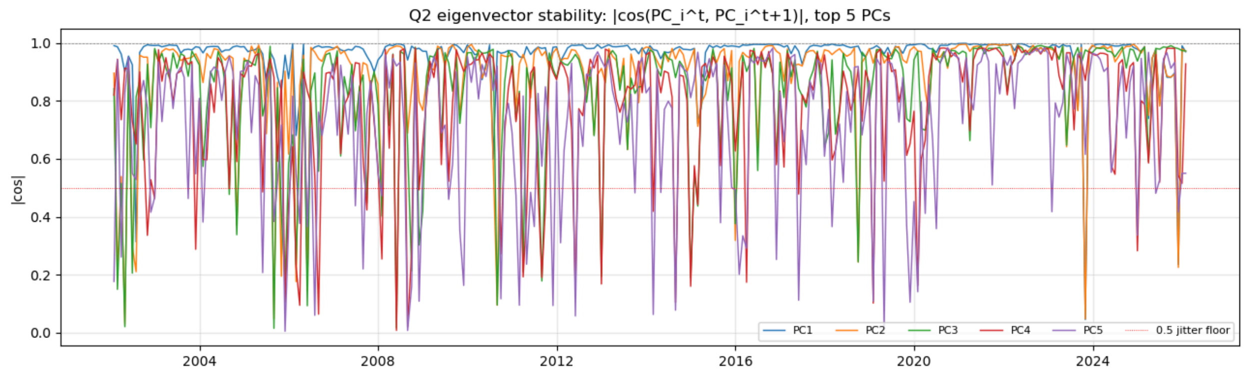
Independent recompute on the Q2 panel confirmed the persisted artifact matched a from scratch rebuild on the XLE (Energy) group to numerical precision, maximum absolute difference of 0.0 across 183,604 cells, and verified the no look-ahead property by checking that the persisted beta on day t equals the no-shift beta computed through day t-1. The validation script accompanies the notebooks.

3.3 Walk-forward PCA and Marchenko Pasteur factor count

At each of 290 monthly rebalance dates (month-start, snapped to the nearest trading day, 2002-01-01 through 2026-01-01), a PCA was fit on the prior 252-day window of the cross-sectionally z-scored panel. For each window, the Marchenko Pasteur upper edge, expressed as an explained variance ratio, $\lambda_+/N = (1 + \sqrt{(N/T)})^2 / N$, was used to count the number of statistically significant principal components, with a floor of 2. The PC dimensionality k per window ranged from 12 to 20 (mean 15) and dropped to the 11-14 range during the GFC and the 2020-2022 stress period, replicating the regime-compression effect first documented by Avellaneda & Lee (2010, Fig. 4) and an Avellaneda-style stylized fact: factor structure concentrates and effective dimensionality falls during crises.



Per-window stability of the top eigenvectors was measured as $|\cos(PC_i^t, PC_i^{t+1})|$ on the common ticker set between consecutive months. Median values: PC1 = 0.99 (very stable), PC2 = 0.96, PC3 = 0.92, PC4 = 0.89, PC5 = 0.83. Top-3 PCs are essentially time-invariant; PC5 and beyond carry meaningful month-to-month rotation that will translate into cluster churn downstream.



Embeddings are stored in ArcticDB as (rebal_date, ticker) → PC_1, PC_2, …, PC_max long-form panels, one symbol per residual panel.

3.4 Clustering

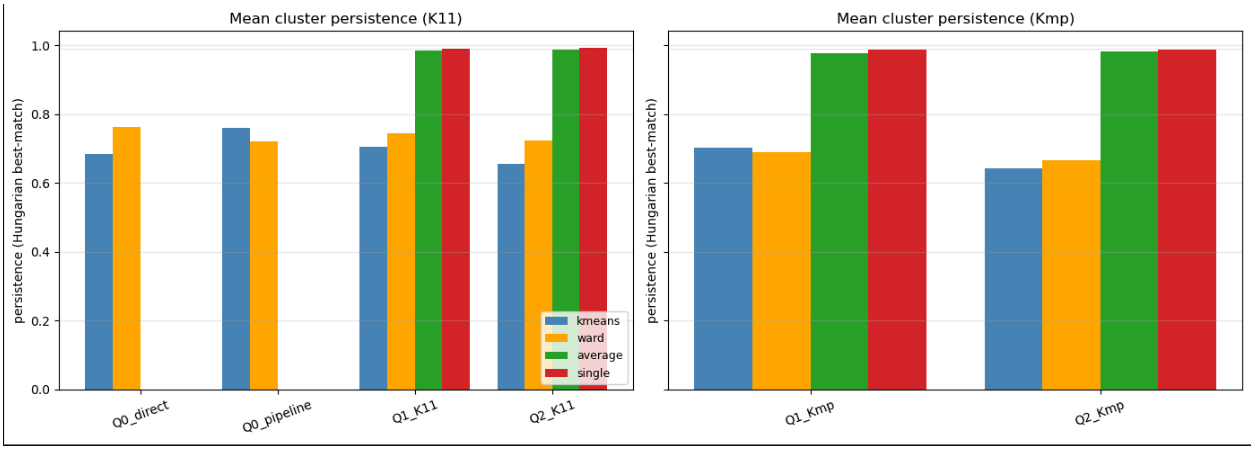
For each rebalance window, the per-PC scale was equalized, each PC column divided by its in-sample standard deviation, so that Euclidean distance reflects the angular structure of the embedding rather than being dominated by PC1's variance. Four clustering algorithms were applied:

- k-means with n_init = 20 (Sklearn KMeans)
- Ward agglomerative clustering (Euclidean linkage)
- Average-linkage agglomerative clustering (Euclidean)
- Single-linkage agglomerative clustering (Euclidean)

Each method was run twice per window: once with K = 11 (matching the GICS sector count, making ARI directly interpretable) and once with K = k_MP (the per-window MP-determined PC count). Single and average linkage exhibit the chaining pathology (one giant cluster plus K − 1 singletons) at this scale, as predicted by the Marti et al. (2021) critique; their ARI numbers are reported for completeness in Section 4.1 but they are excluded from the portfolio construction in Section 4.2.

Cluster persistence diagnostic

Per QHS Ch. 8 Eq. 8.20, the lag-1 autocorrelation of risk-adjusted forecasts governs portfolio turnover. For the cluster portfolios, the analog is cluster persistence: between consecutive rebalance dates, what fraction of tickers retain the same cluster assignment after optimal label permutation (computed via the Hungarian algorithm)? Mean persistence: k-means 65-77%, Ward similar; single and average linkage 99% (misleadingly so, because the giant cluster does not move). The persistence number for k-means/Ward is the genuine reorganization rate and connects directly to the 60% turnover observed in the cluster portfolios.



3.5 Portfolios

All portfolios share the same monthly rebalance schedule and a 252-day trailing window for any required covariance estimate. Nine portfolios were constructed:

Cluster portfolios (one per panel)

For each cluster artifact (Q0_direct, Q1 K=11, Q2 K=11) the centroid of each cluster in the equal-weight-scaled PC space was computed, and the stock whose feature vector is Euclidean closest to the centroid was selected. The 11 representatives are then equal-weighted at 1/11. This is the simplest possible cluster-based selection rule.

Six non-cluster baselines

- **EW S&P 500:** 1/N weights across all current members. Null hypothesis that structure does not matter.
- **GICS-stratified:** 1/11 across the eleven GICS sectors, equal weight within each. QHS Ch. 11 null: human-curated taxonomy is sufficient.
- **MV-raw:** closed-form minimum variance $w^* = \Sigma^{-1}\cdot\iota / (\iota'\cdot\Sigma^{-1}\cdot\iota)$ on the raw sample Σ, with a 1e-6·trace(Σ)/N ridge for numerical stability only. Curse-of-dimensionality baseline; demonstrates the failure mode that motivates all subsequent filtering.
- **MV-RMT:** same closed-form on an RMT-filtered Σ. Eigenvalues of the correlation matrix below the MP edge are replaced with their mean; the matrix is reconstructed; the diagonal is restored to 1; the covariance is then C_filt · σσ'. PCA-as-filter (Plerou et al., 1999; Avellaneda & Lee, 2010).
- **MV-Tola:** same closed-form on a Tola-style ultrametric-filtered Σ using average-linkage on the Mantegna distance. Cluster-as-filter.

- **HRP:** López de Prado (2016) recursive bisection with single-linkage quasi-diagonalization. State-of-the-art cluster allocator.

A cap weighted S&P 500 reference (SPY ETF returns) is also reported.

Return aggregation convention

Portfolio daily returns were aggregated in simple return space: weighted sum of $(\exp(r_{i,\square}) - 1)$ across positions, then log(1 + r_p,t) to convert back. The naive alternative (weighted sum of log returns) is wrong for large constituent moves. A single position dropping to zero implies a log return of approximately −5; under a 1/11 weight the naive aggregation would imply a 45% portfolio loss when the mathematical maximum from a single 9.1% position going to zero is exactly 9.1%. This bookkeeping error surfaced in an initial pass and corrupted every metric until corrected.

3.6 Evaluation framework

Structural evaluation (Q1)

At each rebalance date, cluster labels for the active universe were compared against GICS sector labels (from Sharadar metadata) via the Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI). Both metrics range from 0 (random) to 1 (perfect agreement up to label permutation); ARI is the chance corrected version of the Rand Index. Reported numbers are the means across all 290 rebalance windows.

Practical evaluation (Q2)

For each portfolio, the following metrics are reported:

- Annualized log return (CAGR), annualized volatility, and Sharpe ratio (zero risk-free rate).
- Maximum log-drawdown and Calmar ratio (CAGR / |max DD|).
- Average per-rebalance turnover (fraction of holdings rotated).
- Effective-positions count N_eff = 1 / PCR-HHI, where PCR-HHI is the Herfindahl of percentage risk contributions.
- Sector-PCR: percentage risk contributions aggregated by GICS sector.

Percentage Contribution to Risk (PCR) and Percentage Contribution to Loss (PCL).

PCR is an ex-ante quantity: given a portfolio's weights and the estimated covariance matrix at a rebalance date, PCR tells you what fraction of the portfolio's total predicted volatility is attributable to each position. The PCR values across positions sum to one and constitute the risk model's claim about where future risk lives in the portfolio. A position can carry a high PCR either because it has a large weight, because it has high standalone volatility, because it covaries strongly with other positions, or some combination of the three. PCL is the ex-post analog: after a realized return window, PCL tells you what fraction of the portfolio's actual realized loss came from each position. The PCL values also sum to one, but they are an accounting decomposition of what already happened rather than a forecast of what might happen. The reason these two quantities are useful together is that a well-calibrated risk model should predict where losses will come from, so during a real drawdown PCR and PCL should agree across positions. If they agree, the optimizer correctly identifies the risk before it materializes; if they diverge, the model was monitoring the wrong exposures. QHS p. 71 makes this comparison meaningful only during genuine crisis windows because in quiet periods total realized loss is small, individual PCL values become numerically unstable, and the correlation between the two becomes uninformative. This is the structural reason §4.3 reports the PCR-versus-PCL correlation only for the COVID-2020 and Rates-2022 windows rather than over the full backtest, and it is the reason portfolios that largely sidestepped the 2022 drawdown show near-zero correlations in Table 3.

Two crisis subperiods were defined for the ex-post Percentage Contribution to Loss (PCL) analysis: COVID-2020 (2020-02-19 to 2020-03-23, SPY peak to trough) and Rates-2022 (2022-01-03 to 2022-10-12, SPY peak to trough). For each crisis, PCL_i is the normalized realized contribution of position i to portfolio loss. The correlation between PCR (averaged across crisis-period rebalances) and PCL is reported as a direct test of the QHS p. 71 claim. Crisis windows were defined ex-post by SPY peak to trough; this is a known limitation discussed in Section 5.

4. Results

4.1 Structural results: cluster agreement with GICS

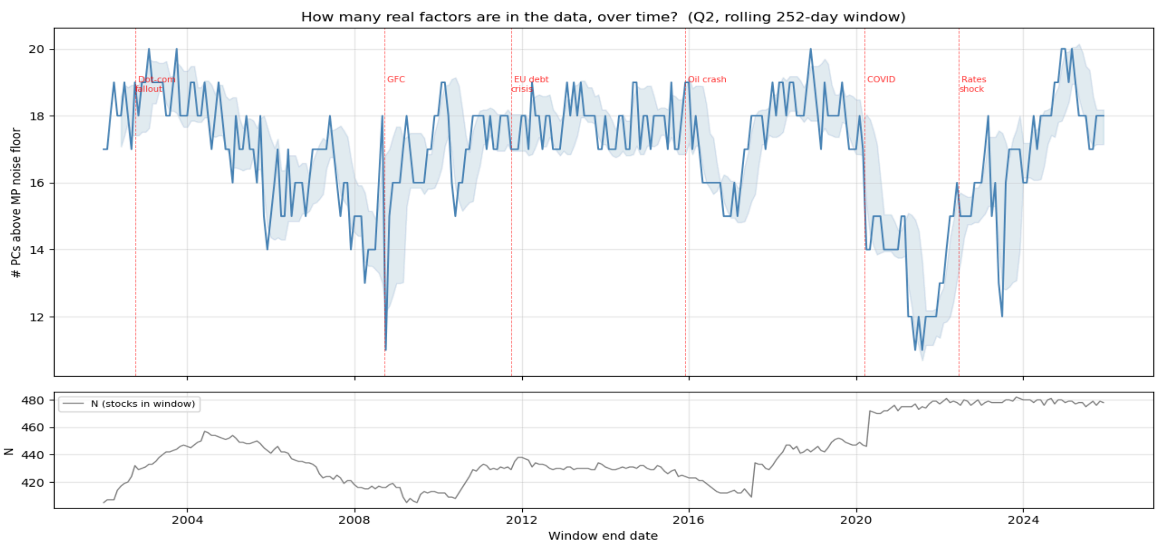
Table 1 reports mean ARI and NMI versus GICS sectors across all 290 walk-forward windows for k-means with K = 11. Single and average linkage are omitted because their ARI values are not interpretable in the presence of chaining.

Table 1. Mean ARI and NMI versus GICS sectors, k-means with K = 11, 290 windows.

Input panel	ARI	NMI	Interpretation
Q0 direct (raw 252-day vectors)	0.195	0.422	Lowest; market factor dominates
Q0 pipeline (PCA on raw)	0.234	0.468	PCA helps modestly
Q1 (CAPM residual)	**0.244**	**0.477**	**Highest GICS agreement**
Q2 (sector-β residual)	0.052	0.199	Moves away from GICS by design

The pattern is consistent across both metrics. Clustering on raw 252-day return vectors gives ARI = 0.195, well above the chance baseline of 0 because correlation with the market factor is itself partially sector-correlated. PCA + MP cutoff on raw returns improves this slightly to 0.234; PC2 and below pick up sector content that the dominant market factor was hiding. Stripping the market via CAPM residualization raises ARI to its peak of 0.244, which is the project's strongest answer to Q1: unsupervised clustering on market residualized returns produces partial but real agreement with the human-curated GICS taxonomy, achieved with no information about company business lines.

Q2's 5x drop in ARI versus Q1 is not a failure. There is direct evidence that the sector-beta residualization successfully removed the sector signal; clusters now operate on what remains, namely sub-sector and idiosyncratic structure. The correlation-matrix triptych in the data-preprocessing diagnostics confirms this: mean off-diagonal correlation falls from 0.30 (raw, market dominated) to 0.09 (Q1, sector blocks remain) to 0.06 (Q2, mostly noise plus a faint Energy block). The top-10 PC1 loaders on Q1 are 10/10 Energy stocks (APA, DVN, EOG, etc.), reproducing the Avellaneda-Lee eigenportfolio result on this universe; on Q2, that mega-factor is absorbed by XLE and PC1 instead loads on Consumer Cyclical.



ARI dips during the 2008-2010 GFC and the 2020-2022 stress period for all panels, mirroring the MP survivor count compression in Section 3.3, both metrics are facets of the same regime effect.

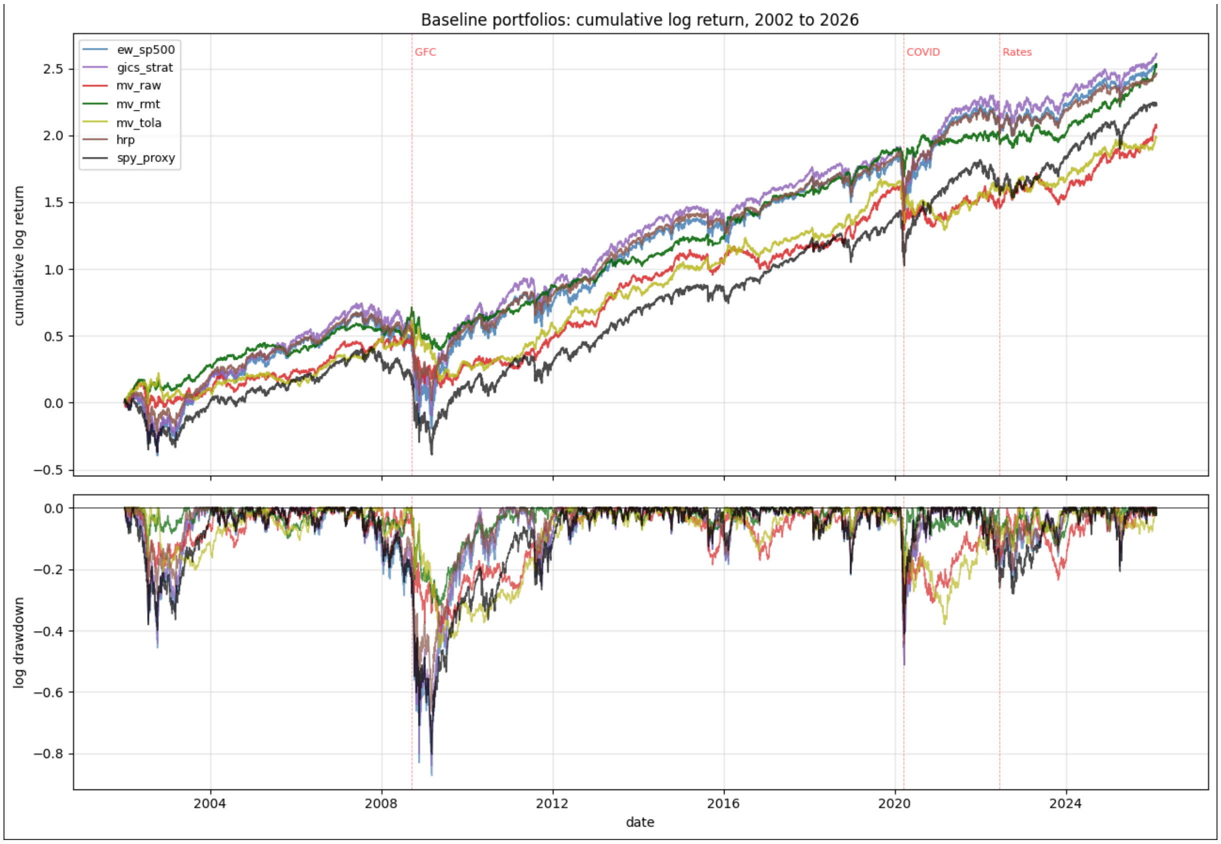
4.2 Practical results: portfolio performance

Table 2 reports the full 2002 to 2026 backtest statistics for all nine portfolios plus the SPY reference.

Table 2. Backtest statistics, monthly rebalance, 2002-01 to 2026-01.

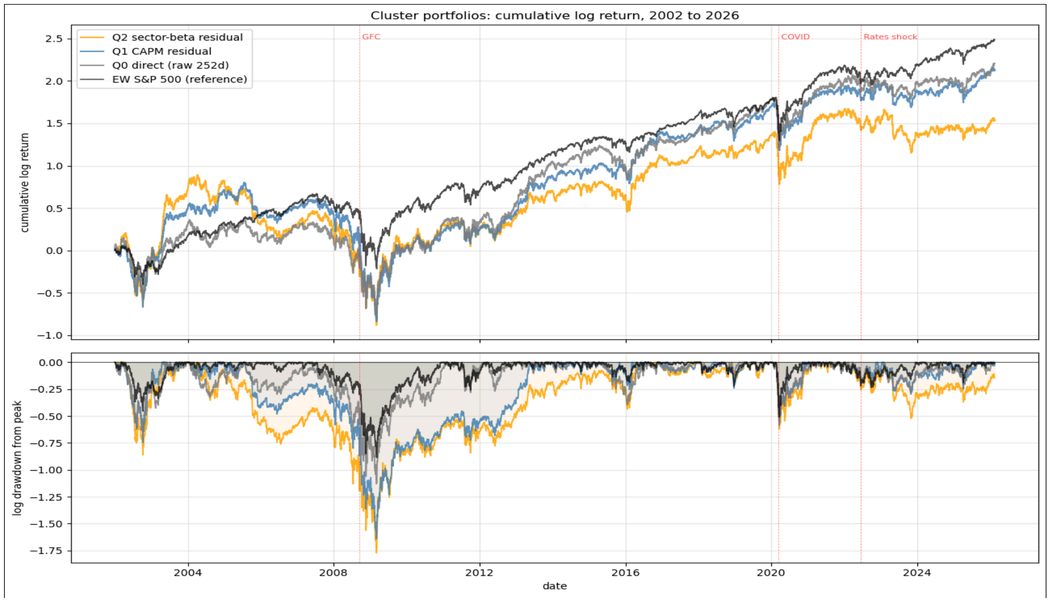
Portfolio	CAGR	Vol	Sharpe	Max DD	N_eff
Q0 cluster (raw)	+11.0%	27.1%	0.41	−121%	8.3
Q1 cluster (CAPM)	+8.5%	27.0%	0.32	−177%	8.2
Q2 cluster (sector-β)	+5.4%	29.7%	0.18	−173%	8.4
EW S&P 500	+10.5%	20.8%	0.51	−87%	414
GICS-stratified	+10.8%	20.6%	0.53	−84%	321

Portfolio	CAGR	Vol	Sharpe	Max DD	N_eff
SPY proxy (cap-wt)	+9.2%	19.0%	0.49	−80%	n/a
MV-raw	+8.5%	14.0%	0.61	−41%	≈ 0
MV-Tola	+8.2%	12.8%	0.64	−46%	43
HRP	+10.3%	16.7%	0.61	−66%	387
MV-RMT	**+10.4%**	**11.1%**	**0.94**	**−32%**	**59**



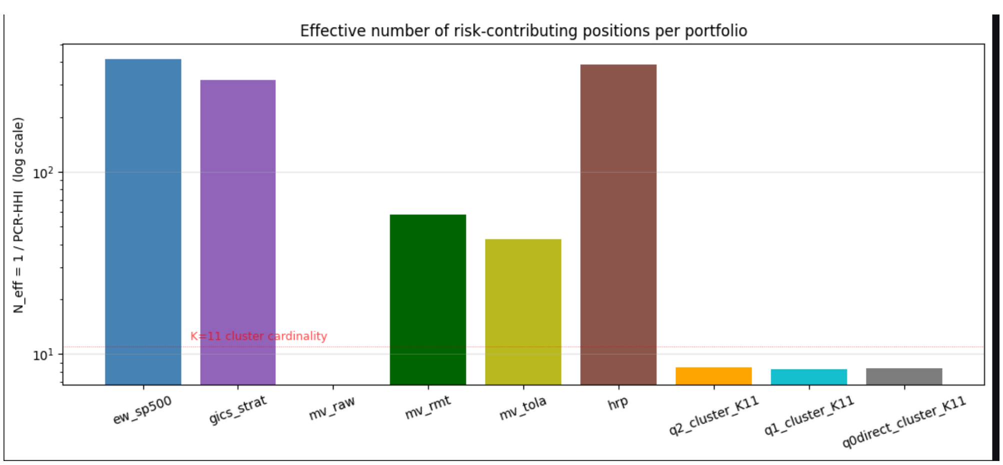
Three observations dominate this table.

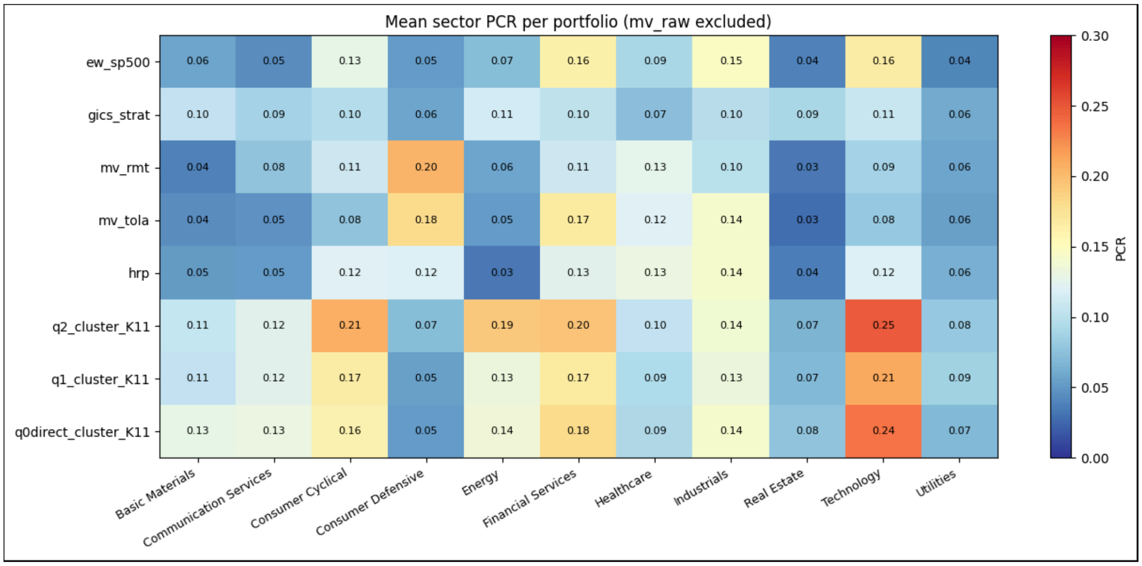
(i) The naive cluster portfolios fail. None of Q0, Q1, or Q2 cluster portfolios beats the equal-weighted S&P 500 on Sharpe. Q2, the panel originally hypothesized to be the most useful, scores worst at Sharpe 0.18, less than half of EW and less than a quarter of MV-RMT. The risk decomposition explains the mechanism: cluster portfolios concentrate to N_eff ≈ 8 (roughly eight of the 11 positions effectively drive risk), and the centroid pick within each cluster gravitates to high-volatility names. Sector-PCR shows cluster portfolios with 21-26% of risk in Technology versus 16% for the EW reference. The Q0 > Q1 > Q2 ranking on Sharpe is, in retrospect, internally consistent: Q0 implicitly groups by volatility tier (since correlation with the market is the dominant axis on raw returns) and the centroid pick of each tier behaves like a defensive diversifier; Q1's GICS-like clusters produce GICS-like portfolios; Q2's idiosyncratic clusters produce 11 specific-risk bets.



(ii) MV-raw is the curse-of-dimensionality failure mode in action. Despite a nominal Sharpe of 0.61, MV-raw's effective-positions count is approximately zero, and individual position weights routinely exceed ±100. The unconstrained closed-form min-variance solution on N ≈ 500 and T = 252 produces wild long/short leverage that nets to small in-sample variance and is not investable in any practical sense. This is the QHS p. 64 / Plerou et al. (1999) result rendered concretely on this universe, and the reason every subsequent MV variant first filters the covariance matrix.

(iii) MV-RMT wins decisively across every metric. Same CAGR as EW (10.4% vs 10.5%), roughly half the volatility (11.1% vs 20.8%), about one-third the drawdown (−32% vs −87%). Sharpe 0.94, which is 0.30 above the next best baseline (MV-Tola at 0.64) and approximately 0.76 above the best cluster portfolio. Sector PCR reveals where the risk is taken: Consumer Defensive 20% and Healthcare 13% (the sectors with the lowest in-window volatility), with a complementary underweight in Technology (8.7% PCR versus 16.5% in EW). This is what a well calibrated minimum-variance optimizer is supposed to do; the RMT filter is what makes the calibration robust enough to deliver.





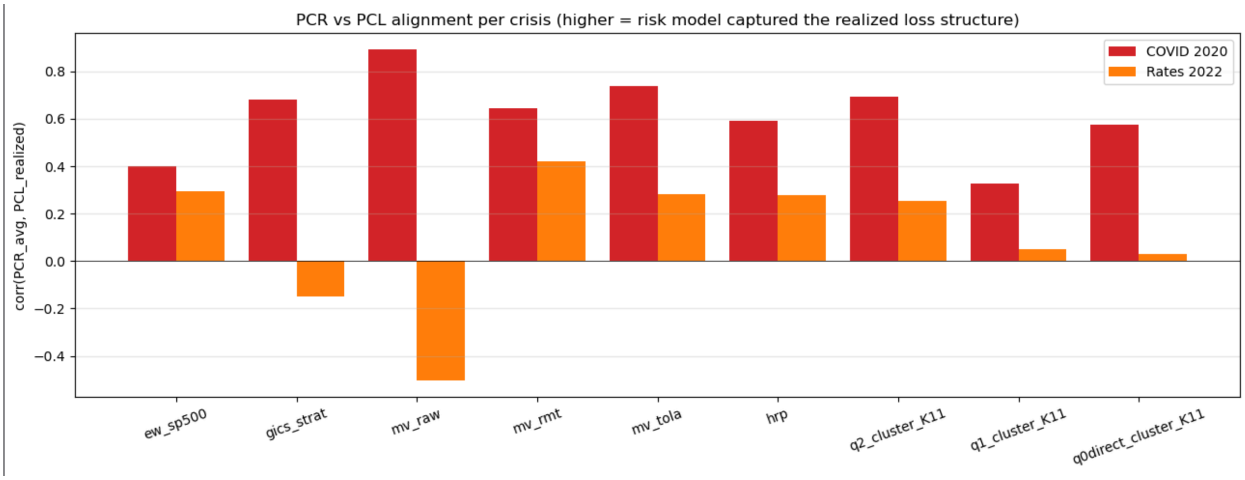
4.3 Crisis-period validation: PCR versus PCL

Section 2.4 introduced Tola et al.'s (2008) framing of portfolio quality as the gap between predicted and realized risk; the QHS risk decomposition extends this to a per stock attribution. Specifically, QHS p. 71 makes the testable claim that during crises, ex-post Percentage Contribution to Loss (PCL) should approximate ex-ante Percentage Contribution to Risk (PCR), and in quiet periods the two should diverge because the realized losses are not large enough to be informative. Table 3 reports the per portfolio correlation between average PCR over the crisis window and realized PCL, for the two crises defined in Section 3.6.

Table 3. PCR/PCL correlation per crisis window.

Portfolio	COVID-2020	Rates-2022
MV-RMT	**+0.65**	**+0.42**
MV-Tola	+0.74	+0.28
EW S&P 500	+0.40	+0.29
GICS-stratified	+0.68	−0.15
Q1 cluster	(varies)	+0.05

During COVID-2020 every portfolio shows positive correlation (range 0.32 to 0.89): the drawdown was broad enough that ex-ante risk shares correctly identified where losses would land. The Rates-2022 picture is more diagnostic. MV-RMT (+0.42) is the only portfolio with a robustly positive correlation; GICS-stratified is actually slightly negative (−0.15), meaning its risk model predicted one set of losses and realized losses landed elsewhere. Portfolios that largely avoided the 2022 drawdown (MV-Tola, Q1 cluster) show near zero correlations because their total realized loss is small and PCL is poorly defined, this is precisely the QHS quiet period caveat. The takeaway is that MV-RMT is not just statistically tighter; it is the most behaviorally informative across regimes.



5. Conclusion

5.1 Summary of findings

On Q1 (structural): unsupervised PCA + clustering on market-residualized returns recovers GICS sectors at ARI ≈ 0.24 (k-means, K = 11), which is well above the chance baseline of 0 and constitutes meaningful but partial agreement. Stripping the sector signal additionally (Q2) drops ARI by a factor of about five, confirming that the sector-beta residualization does what it is designed to do: clusters then operate on sub-sector and idiosyncratic structure rather than on GICS-aligned signal. The top-loaders analysis additionally reproduces the Avellaneda-Lee eigenportfolio finding (PC1 on Q1 loads 10/10 on Energy).

On Q2 (practical): the answer is yes. PCA-based pipeline produce a portfolio that beats every baseline. The winning portfolio, though, is not a cluster portfolio. The naive one stock per cluster equal-weight construction underperforms all baselines including equal-weight, with Q2 the worst. The winning portfolio is minimum-variance applied to a covariance matrix denoised by replacing sub-MP eigenvalues with their mean (MV-RMT), delivering Sharpe 0.94 versus 0.51 for equal-weight, half the volatility, one-third the drawdown, and the most informative ex-ante / ex-post risk decomposition across both crisis windows. The PCA pipeline is what wins; it just wins as a covariance filter, not as a cluster labeler.

5.2 Comparison with the literature

The cluster as filter result of Tola et al. (2008) is reproduced qualitatively: MV-Tola (Sharpe 0.64) clearly beats MV-raw (effectively broken). However, the RMT filter outperforms the ultrametric filter on this data and over this period, suggesting that the eigenvalue replacement approach is more aggressive and effective on a full S&P 500 universe than the cluster-based approach the original Tola paper compared against.

HRP (López de Prado, 2016) delivers Sharpe 0.61, well above the cluster portfolios but well below MV-RMT. The LdP paper's criticism that minimum-variance is unstable applies here specifically to the unfiltered version (MV-raw); once filtered with the Plerou/Avellaneda approach, MV is the strongest performer. The two ideas that worked individually, clustering for structure and RMT for filtering, were never combined in the baseline set; doing so is the most promising single follow up (see 5.4).

Avellaneda & Lee (2010) did not construct long-only portfolios from PCA; they used the eigenportfolios for statistical arbitrage on residuals. Even so, two of their core diagnostic findings are reproduced cleanly here: the temporal stability of the top eigenvector (|cos| ≈ 0.99 month-to-month) and the regime compression effect (MP survivor count dropping from ~17 to ~11 during crises).

5.3 Limitations

- **Transaction costs are not modeled.** All backtests are frictionless. The cluster portfolios show average turnover of 60% per rebalance, which would substantially erode realized Sharpe under any reasonable assumption about bid-ask spreads and market impact.
- **Crisis windows were defined ex post.** The COVID-2020 and Rates-2022 windows were chosen by inspection of SPY drawdown peaks, not pre-registered.
- **Single-market regime.** Universe is U.S. large-cap equities only. Whether the findings generalize to international markets, small caps, or fixed income is untested.
- **Beta-estimation fragility under stress.** The 252-day rolling β used for Q1 and Q2 residualization is robust on average but breaks down during regime transitions (corr(PC1, SPY) swings to ±0.5 during the GFC). EWMA β with a shorter half-life is the standard remedy and was not tested.
- **Sharpe-difference significance was not formally tested.** The differences between MV-RMT and the next-best baselines are large enough that significance is not in doubt, but the formal bootstrap is deferred.

5.4 Future work

Three extensions appear most worth pursuing, in priority order:

3. Combine the two ideas that worked. The natural follow-up is HRP applied to an RMT-filtered correlation matrix, or equivalently, MV-RMT applied to a cluster-restricted block diagonal covariance. Both versions would test whether the clustering and the filtering contribute orthogonal information or merely overlap.
4. Alternative within cluster selection rules for the cluster portfolios. The centroid + 1/K rule was the naive choice; lowest-volatility within each cluster (a clean inverse-vol within-cluster rule), or skipping clusters smaller than five members, should mute the high-vol bias documented in 4.2. The point would be to clarify whether the cluster-portfolio failure is structural or simply a consequence of an unreasonable selection rule.
5. Universe expansion and transaction-cost modeling. Extending to Russell 3000 would give PCA more dimensions to work with and lower the Marchenko Pastur N/T noise floor. Adding a basic transaction cost model (10 bps per round-trip on the cluster portfolios; 5 bps on the baselines) would test the practical viability of the higher turnover strategies and is necessary before any of these results would be taken seriously by a practitioner.

5.5 Closing

The headline finding of this project is one the prospectus did not anticipate: among the data driven methods tested, the one that beats human curated GICS stratified portfolios is not the cluster based selection rule but the eigenvalue filter step itself. The PCA + Marchenko Pasteur pipeline that the prospectus framed as a clustering preprocessing technique is more useful as a standalone covariance denoising technique, plugged into a closed form Markowitz optimizer. The cluster portfolio idea, as originally formulated with the one stock per cluster, equal-weight rule, does not survive contact with baseline comparison; the QHS risk decomposition reveals that the resulting 11 stock portfolios concentrate to $N_{eff} \approx 8$ in high-vol-sector exposures. This is a real and informative negative result, and one that points naturally toward the two-step combinations identified in 5.4.

Bibliography

Avellaneda, M., & Lee, J. H. (2010). Statistical arbitrage in the U.S. equities market. Quantitative Finance, 10(7), 761-782.

López de Prado, M. (2016). Building diversified portfolios that outperform out of sample. Journal of Portfolio Management, 42(4), 59-69.

Mantegna, R. N. (1999). Hierarchical structure in financial markets. European Physical Journal B, 11(1), 193-197.

Marchenko, V. A., & Pastur, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. Matematicheskii Sbornik, 72(4), 507-536.

Markowitz, H. (1952). Portfolio selection. Journal of Finance, 7(1), 77-91.

Marti, G., Nielsen, F., Bińkowski, M., & Donnat, P. (2021). A review of two decades of correlations, hierarchies, networks and clustering in financial markets. Signals and Communication Technology, 245-274.

Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L. A. N., & Stanley, H. E. (1999). Universal and nonuniversal properties of cross correlations in financial time series. Physical Review Letters, 83(7), 1471-1474.

Qian, E. E., Hua, R. H., & Sorensen, E. H. (2007). Quantitative Equity Portfolio Management: Modern Techniques and Applications. Chapman & Hall / CRC.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. Journal of Finance, 19(3), 425-442.

Tola, V., Lillo, F., Gallegati, M., & Mantegna, R. N. (2008). Cluster analysis for portfolio optimization. Journal of Economic Dynamics and Control, 32(1), 235-258.