Principal Component Analysis of the S&P 500:

Eigenportfolio Construction and Performance vs. Passive Investing

March 2021 to March 2026

Sonly Nguyen

Marshall School of Business

University of Southern California

Los Angeles, CA 90037

sonlyngu@usc.edu

March 2026

Executive Summary

This paper applies Principal Component Analysis (PCA) across five years of daily adjusted closing prices for S&P 500 constituents, from March 2021 through March 2026, to find hidden factors behind equity returns. After analyzing principal components and backtesting several strategies against the S&P 500 as a passive benchmark, three of four strategies delivered superior risk-adjusted returns. The highest result was PC4, which captured AI infrastructure versus the housing market and achieved a Sharpe ratio of 0.76 compared to 0.38 for SPY. A combined four-factor eigenportfolio produced the lowest drawdown of any strategy tested, at -14.7%.

1. Introduction

Passive investing has become the main approach to equity investors as it's low-cost, diversified, and difficult to consistently beat. The S&P 500 weights its constituents by market capitalization, meaning the largest companies contain the highest portfolio weights, regardless of their expected return or risk profile.

This strategy takes a different approach. It asks whether a data-driven model can identify specific exposures within the S&P 500 that generate better risk-adjusted returns than simply buying the index. It utilizes Principal Component Analysis, a statistical technique that breaks down a large return matrix into orthogonal factors, each representing an independent source of movement across stocks.

A market-cap weighted index gives you exposure to all factor risks simultaneously, with no ability to bias toward attractive ones or away from unattractive ones. PCA isolates independent sources of variation, manually interpret them, and build portfolios that concentrate in specific factors. These are called eigenportfolios.

2. Data and Methodology

2.1 Data

Daily adjusted closing prices were pulled from Yahoo Finance using the yfinance Python library and verified through FactSet's Universal Screener. The data covered March 9, 2021 through March 9, 2026, which resulted in a total of 1,259 trading days. Any tickers missing >10% of observations were dropped. After cleaning, the dataset covered 495 stocks. All prices were converted to daily log returns using:

$$r = \ln\left(\frac{P_t}{P_t + 1}\right)$$

Log returns are symmetric, additive over time, and more normally distributed than simple percentage returns.

2.2 Standardization

Before running PCA, the return matrix was standardized so every stock had a mean of 0 and a standard deviation of 1. Without standardization, high-volatility stocks like semiconductors would dominate the components simply because their numbers are larger, not because they share more meaningful movement with the rest of the market.

2.3 Principal Component Analysis

PCA was applied to the standardized 1,259 by 495 return matrix. The top 50 principal components explained about 67% of total variance. Each component comes with a loading vector indicating how much each stock contributes to that factor and in which direction. To build an eigenportfolio, I take those loadings as raw weights, zero out the negatives for the long-only portfolio constraint, and normalize the rest to sum to one.

3. PCA Results

3.1 Variance Decomposition

The first principal component alone explains 30.2% of total return variance, confirming a dominant market-wide factor. From there the drop is steep: PC2 explains 6.6%, PC3 explains 3.5%, and each subsequent component contributes less. Even with 50 components, cumulative variance only reaches 67%. The remaining 33% is mostly stock-specific noise that a diversified portfolio handles naturally.

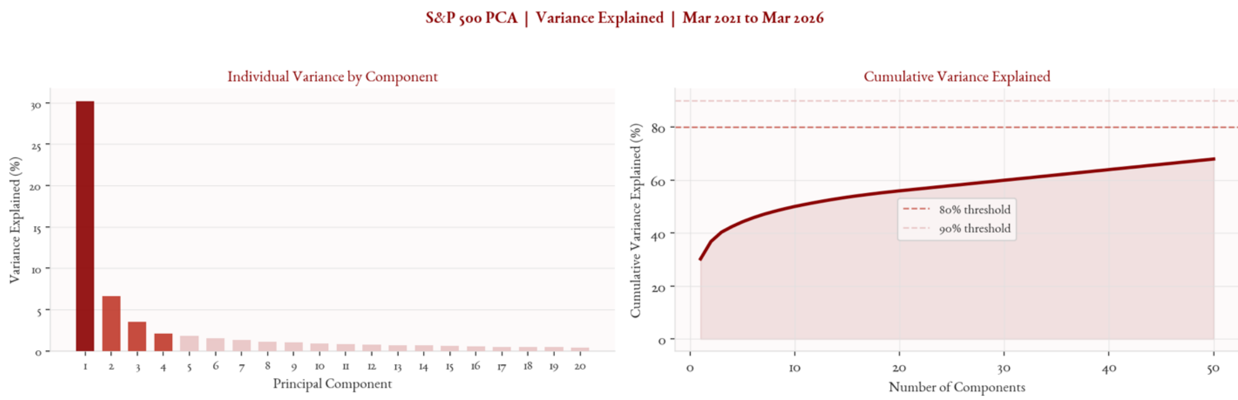

Figure 1. Variance explained by the top 50 principal components extracted from S&P 500 daily log returns, March 2021 to March 2026. The left panel shows individual variance per component; the right panel shows cumulative variance. PC1 alone accounts for 30.2% of total variance, reflecting the dominant market factor. The curve flattens after approximately 10 components, with 50 components collectively explaining 67% of total variance.

3.2 Economic Interpretation of Components

PC1 (30.2%) is the market factor. It loads positively on almost every stock, with high weights on asset managers and industrials like BLK, AMP, and ITW. The lowest-loading names are defensive staples like KR, DG, and MCK. PC1 loadings are essentially broad market beta.

PC2 (6.6%) separates utilities from semiconductors. Consolidated Edison, Duke Energy, and other regulated utilities load positively, while NVDA, LRCX, and AMAT load negatively. The split could reflect interest rate sensitivities.

PC3 (3.5%) captures energy versus software. The top positive names are oil and gas producers (COP, XOM, HAL, SLB), while the negatives are high-multiple software platforms (INTU, NOW, ADBE). Rising inflation boosted energy cash flows while reducing software multiples.

PC4 (2.1%) is the most interesting finding given the sample period. Positive loadings cluster around AI infrastructure and power grid companies (VST, WMB, NVDA, CRWD, ANET). Negative loadings are concentrated in homebuilders and materials (LEN, DHI, SWK, MAS). This component captures the defining macroeconomic trends from 2022 through 2025.

PC5 (1.8%) separates property and casualty insurers (ACGL, AJG, CB, PGR) from rate-sensitive utilities and REITs (DLR, NEE, CMS).

4. Eigenportfolio Backtest Results

4.1 Portfolio Construction

For each component, I built a long-only eigenportfolio by treating the PC loadings as raw weights, setting negatives to zero, and normalizing to sum to one. I tested four single-component portfolios and one combined portfolio against SPY. The risk-free rate is set at 4% per. Weights are fixed at the start of the period with no rebalancing.

4.2 Performance Results

Table 1 summarizes the full performance statistics. Three of four eigenportfolios beat SPY on a Sharpe ratio basis.

Portfolio	Ann. Return	Volatility	Sharpe Ratio	Max Drawdown	Total Return	vs SPY Sharpe
SPY (Passive)	11.2%	17.3%	0.38	-24.5%	62.4%	Benchmark
PC2: Defensives vs Semis	6.6%	13.2%	0.18	**-17.4%**	33.9%	<SPY
PC3: Energy vs Software	**15.9%**	19.8%	**0.54**	**-21.3%**	**96.5%**	**>SPY ***
PC4: AI Infra vs Housing	**18.4%**	17.0%	**0.76**	**-18.9%**	**116.9%**	**>SPY ***
PC2+3+4+5	**14.4%**	14.1%	**0.67**	**-14.7%**	**85.3%**	**>SPY ***

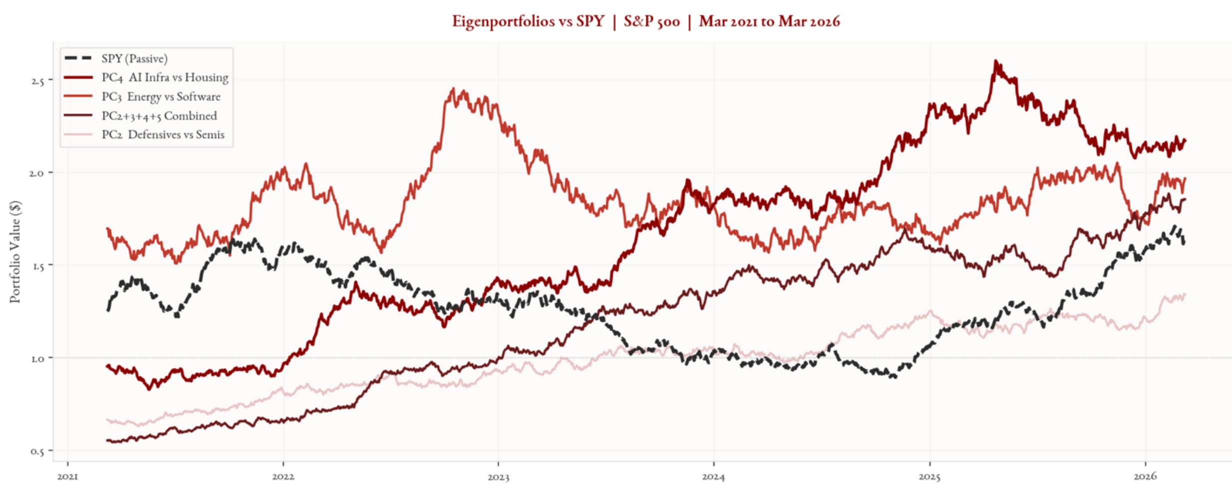

Figure 2. Cumulative return of four eigenportfolios versus SPY as the passive benchmark, showing the growth of $1 invested on March 9, 2021 through March 9, 2026. Eigenportfolio weights are derived from PCA loadings computed on S&P 500 daily adjusted closing prices, with a long-only constraint and no rebalancing. Three of four strategies outperform SPY over the full period.

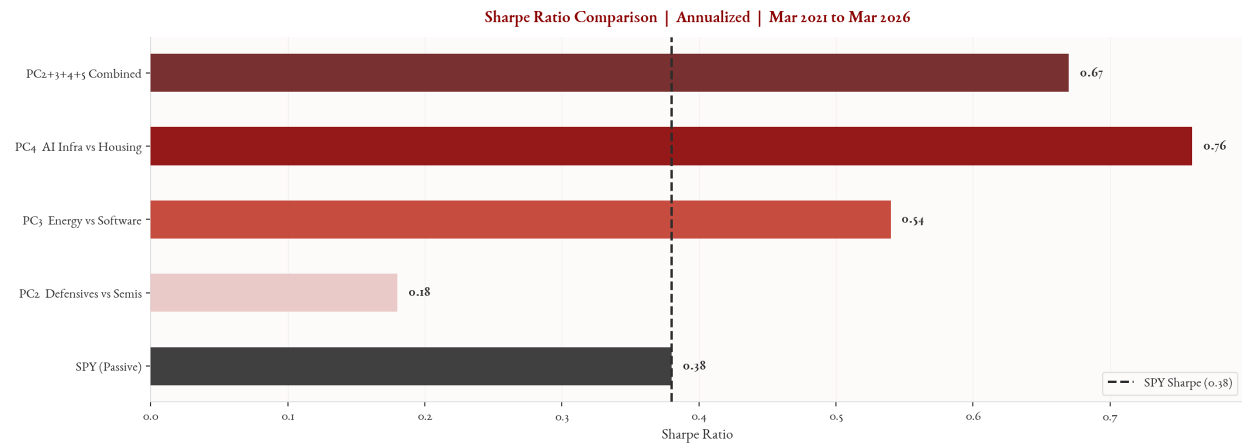

Figure 3. Annualized Sharpe ratios for each eigenportfolio strategy versus SPY, computed using a risk-free rate of 4% per year over the full backtest period. The dashed vertical line marks the SPY benchmark Sharpe of 0.38. PC4 (AI Infrastructure vs. Housing) achieves the highest single-factor Sharpe at 0.76, while the combined four-factor portfolio reaches 0.67 with the lowest maximum drawdown of any strategy.

4.3 Discussion

PC4 stands out as the strongest single-factor result. It generated 18.4% annualized return with a Sharpe ratio of 0.76 and a maximum drawdown of -18.9%, better than SPY's -24.5%. It delivered more return with less downside risk. Overweighting AI infrastructure and power grid names while avoiding homebuilders resulted in the highest total return throughout the backtest period.

PC3 also outperformed on a risk-adjusted basis with a Sharpe of 0.54, benefiting from the commodity cycle in 2021 to 2022 and the rerating of energy cash flows as inflation rose.

PC2 did not perform so well. A Sharpe of 0.18 and total return of 33.9% versus SPY's 62.4% is meaningful underperformance. Utilities got hit hard by rising rates throughout 2022 and 2023, and this illustrates the central risk: choosing the wrong factor at the wrong point in the cycle can produce worse outcomes than passive investing.

The combined PC2+3+4+5 portfolio is arguably the most practically useful result. Blending four orthogonal components brought the Sharpe to 0.67 and the maximum drawdown to -14.7%, the lowest of any strategy including SPY. That is what genuine factor diversification looks like in practice.

5. Limitations

The backtest is in-sample. PCA loadings were computed on the full 2021 to 2026 dataset and the same period was used to evaluate performance. A rigorous test would use a rolling walk-forward approach: compute PCA on a 36-month training window, test on the next 12 months, then roll forward.

Transaction costs are ignored. Eigenportfolios need periodic rebalancing as factor loadings shift. At typical institutional trading costs and 30 to 50% annual turnover, net returns would be meaningfully lower.

Survivorship bias is present. The constituent list is based on the current S&P 500, so companies that were delisted or dropped during the period are absent from the data. This biases every strategy upward.

6. Conclusion

Running PCA on five years of S&P 500 daily returns surfaces factor structures that map onto the macro themes that defined this period: the AI infrastructure buildout, the rate hike cycle's effects on housing and utilities, the energy commodity supercycle, and the insurance pricing environment. These are not arbitrary statistical constructs.

Three of four eigenportfolios beat SPY on a Sharpe ratio basis. PC4 produced the best single-factor result with a Sharpe of 0.76 and lower maximum drawdown than the index. The combined four-factor portfolio achieved the lowest drawdown of any strategy while still outperforming SPY on a risk-adjusted basis.

The takeaway is not that eigenportfolios automatically beat the market. PC2's underperformance is a real reminder that bad factor timing can produce worse outcomes than just buying the index. The more defensible point is that PCA gives you a rigorous, data-grounded way to translate a macro thesis into a portfolio. If your view on the dominant economic regime turns out to be right, the eigenportfolio framework is an efficient way to act on it.