STATISTICAL ARBITRAGE IN INDIAN EQUITIES

Implementation of Avellaneda and Lee (2008) in Nifty 100 Constituents

Nitesh Rai

Introduction to Mean Reversion & Statistical Arbitrage

Mean Reversion in Financial Markets

- **Definition:** Mean reversion refers to the tendency for extreme values to be followed by observations that move closer to the average.
- In financial markets, asset prices often follow geometric random walks rather than mean-reverting behavior.
- Returns (not prices) show mean-reverting tendencies centered around zero, but trading directly on this is impractical.

Statistical Arbitrage (StatArb) Overview

- StatArb engineers mean-reverting price series by combining multiple non-mean-reverting series resulting in a portfolio whose net market value exhibits mean reversion.
- Three key characteristics:
  - **Systematic Trading** – Rule-based rather than fundamentals-driven.
  - **Market Neutrality** – Zero beta with the market.
  - **Statistical Methods** – Generating excess returns through modeling.
- StatArb approaches (Krauss, 2017): distance-based for identifying co-moving assets, cointegration-based exploiting mean-reversion in linearly combined series, stochastic control for optimal portfolio holdings, and model-driven methods.

Simplified Avellaneda & Lee (2008)

- Trading signals are generated by decomposing stock returns into systematic factor returns and idiosyncratic returns.
- Modeling the idiosyncratic returns—the component of stock returns unexplained by systematic factors—as mean-reverting processes, leading to contrarian trading signals.
- Consider a technology stock S, where its price movements can be decomposed into systematic and idiosyncratic returns:

  $$\frac{dS_t}{S_t} = \alpha\, dt + \beta_M \frac{dM_t}{M_t} + \beta_T \frac{dT_t}{T_t} + dX_t$$

  where:
  - α is the drift term.
  - β_M, β_T represent the stock's sensitivities to market and sector factors.
  - M and T are broad market and sector trends.
  - dX_t is the idiosyncratic (mean-reverting) component.
- By constructing portfolios that neutralize exposure to systematic factors, traders can isolate and profit from the mean-reversion of the idiosyncratic component dX_t.
- This forms the basis of a factor-based trading strategy: traders take long positions in undervalued stocks (relative to factor exposure) and short positions in overvalued stocks, exploiting mean reversion.

Principal Component Analysis: Mathematical Framework

Question: How do we define these risk factors ?

- PCA extracts dominant correlation patterns from high-dimensional data.
- Transforms data into new coordinate system ordered by variance explanation.
- Data centering: $\tilde{X} = X - \bar{X}$, where $X \in \mathbb{R}^{n \times p}$ is the data matrix.
- Covariance matrix: $\Sigma = \frac{1}{n-1} \tilde{X}^T \tilde{X}$.
- Eigendecomposition of the covariance matrix: $\Sigma v_i = \lambda_i v_i$, where λ_i are the eigenvalues in descending order and v_i are the corresponding orthonormal eigenvector.
- Eigenvectors represent the principal components, which describe independent sources of variation in the dataset (each represnting a distinct market pattern).
- The first principal component v_1 corresponds to the direction of maximum variance in the data (broad market component), with subsequent components $v_2, v_3, ...$ being orthogonal and capturing decreasing amounts of variance(sector specific trends).
- The importance of each principal component is quantified by its explained variance ratio $r_i = \frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j}$, which represents the proportion of total variance captured by that component.

Dataset & Return Definitions

- **NIFTY 100 Stocks**: Covers 66.4% of the free float market capitalization of stocks listed on the National Stock Exchange (Includes 17 major industry sectors).
- **Data**: Daily adjusted closing prices and daily volume from Yahoo Finance, accounting for corporate actions like stock splits and dividend distributions.
- **Study period**: 2012-2024-limited data availability for most stocks prior to 2012.
- **Final sample**: 84 stocks, filtered for consistent data during the study period.

Standard Daily Return:

$$R_{it} = \frac{S_{it} - S_{i,t-1}}{S_{i,t-1}}$$

where S_{it} is the price of stock i on day t, adjusted for dividends.

Volume-Weighted Return:

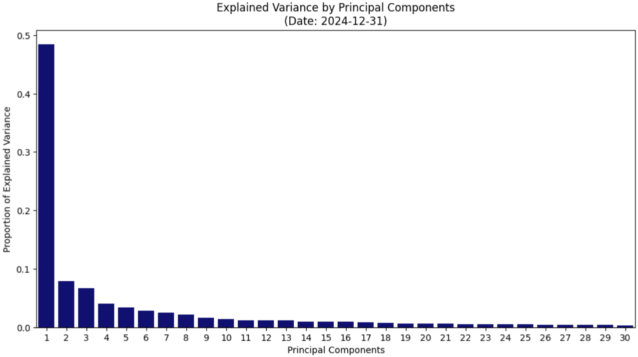
$$\bar{R}_{it} = R_{it} \cdot \frac{\langle \delta V_i \rangle}{|V_{it} - V_{i,t-1}|}$$

$\langle \delta V_i \rangle$ is typical daily trading volume
ΔV_{it} is the change in volume.

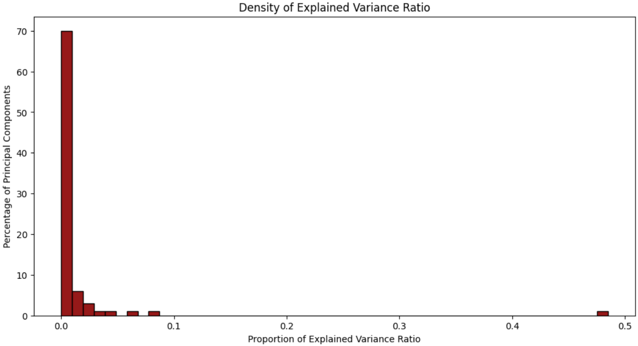
- Impact of including volume in return calculation:
  - $\bar{R}_{it} \approx R_{it}$ when trading volume is $\Delta V_{it} \approx \langle \delta V_i \rangle$.
  - $\bar{R}_{it} > R_{it}$ when trading volume is $\Delta V_{it} < \langle \delta V_i \rangle$.
  - $\bar{R}_{it} < R_{it}$ when trading volume is $\Delta V_{it} > \langle \delta V_i \rangle$.
  - Helps suppress false signals during abnormal volume spikes.

Eigen Decomposition

- The return data is standardized before eigen decompoistion to ensure results are not skewed by varying stock volatilities: $Y_{it} = \frac{R_{it} - \bar{R}_i}{\bar{\sigma}_i}$
- The empirical correlation matrix is calculated as $\rho_{ij} = \frac{1}{M-1} \sum_{t=1}^{M} Y_{it} Y_{jt}$
- We chose a one-year rolling window (252 trading days) to estimate ρ_{ij}.
- Eigenvalues represent the variance explained by each component: $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_N \geq 0$ ordered in decreasing magnitude.
- Each eigenvector at time t is composed of N elements: $v_{jt} = \left(v_{1jt}, v_{2jt}, \ldots, v_{Njt}\right)$ where v_{ijt} represents the loading of stock i in the j-th principal component at time t.



Eigen Decomposition (II)

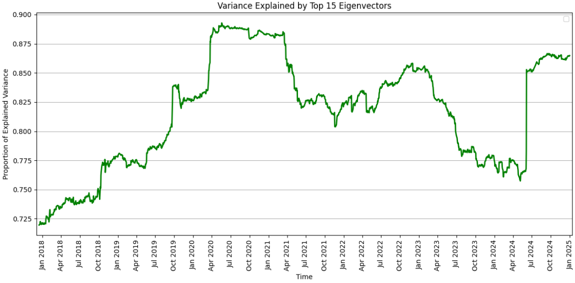


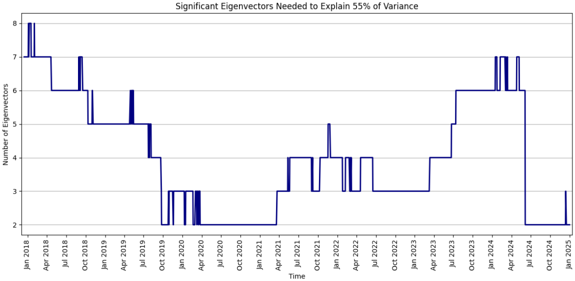
- Density distribution of the explained variance ratios highlight that the significant eigenvalues that stand distinctly separated from the near zero 'noise spectrum' represent the dominant market factors whereas near-zero eigenvalues capture less important market dynamics. The boundary between 'significant' and 'noise' eigenvalues is somewhat blurred and corresponds to the edge of the 'bulk spectrum.
- Two approaches for factor extraction: fixed (industry sector count) or variable selection based on cumulative explained variance threshold.

Interpretation of Eigenvectors & Eigenvalues

During market stress periods, fewer eigenvectors explain more total variance as assets move together, while in calmer times, assets move independently based on industry factors and individual characteristics, requiring more eigenvectors.

Market stress periods like COVID-19 crash and June 2024 Indian election results day showed higher explained variance by top 15 eigenvectors and fewer vectors needed for 55% variance coverage, indicating increased asset correlation during uncertainty.

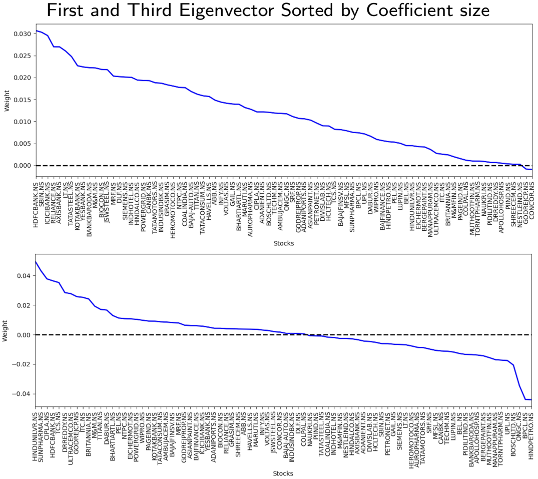

Variance Explained by Top 15 Eigenvectors


Significant Eigenvectors Needed to Explain 55% of Variance

Interpretation of Eigenvectors & Eigenvalues (II)

The first eigenvector, with uniformly positive coefficients, can be interpreted as the market portfolio with stocks within showing unified market movements. Subsequent eigenvectors have negative components confirming orthogonality.

Higher-order eigenvectors capture increasingly specific sector patterns, becoming more localized and less representative of broad market movements.



Higher-ranking eigenvectors exhibit 'coherence' where similar coefficient values are usually from same industry groups. This sector-based clustering weakens in lower-ranked eigenvectors, where patterns become more random and industry associations fade.

Eigenportfolios: Risk Factors

- Eigenportfolios are constructed for indices $j = 1, \dots, m$ by investing in stocks $i = 1, \dots, N$ with weights:

$$Q_{ijt} = \frac{v_{ijt}}{\bar{\sigma}_{it}}$$

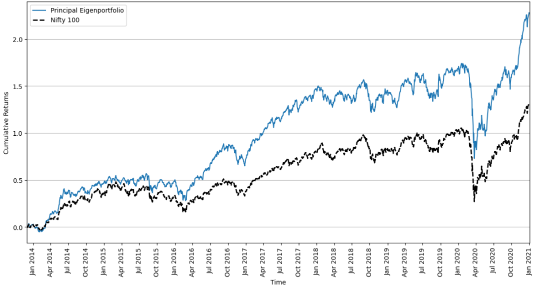
  The weights' inverse relationship with stock volatility aligns with capitalization weighting principles, as larger-cap stocks typically have lower volatilities.

- The principal eigenportfolio: proxy to capitalization weighted benchmark.

- The eigenportfolio returns are calculated as:

$$F_{jt} = \sum_{i=1}^{N} \frac{v_{ijt}}{\bar{\sigma}_{it}} \; R_{it}$$

.



The PCA approach delivers a natural set of risk-factors that can be used to decompose our stock returns into systematic and idiosyncratic components.

Mean Reversion Model: Ornstien Uhlembeck Process

- In a multi-factor model, stock returns follow the system of a SDE:

$$\frac{dS_{it}}{S_{it}} = \alpha_i \, dt + \sum_{j=1}^{m} \beta_{ijt} \frac{dI_{jt}}{I_{jt}} + dX_{it}, \quad \begin{cases} \beta_{ijt} & \text{Factor Sensitivities} \\ dI_{jt}/I_{jt} & j\text{-th Eigenportfolio Returns} \\ \alpha_i \, dt + dX_{it} & \text{Idiosyncratic Component} \end{cases}$$

- Assuming $\alpha_i = 0$, the residual dX_{it} ,is assumed to be the increment of a stationary stochastic process and is modelled using Ornstein-Uhlembeck process which is stationary, mean reverting and auto-regressive with lag 1 process:

$$dX_{it} = \kappa_i (m_i - X_{it}) \, dt + \sigma_i \, dW_{it}, \quad \begin{cases} \kappa_i > 0 & \text{Speed of mean reversion} \\ m_i & \text{Long run mean} \\ \sigma_i & \text{Diffusion coefficient} \\ dW_{it} & \text{Weinner Process} \end{cases}$$

  The parameters $(\kappa_i, m_i, \sigma_i)$are specific to each stock and are assumed to vary slowly relative to the increments of the Brownian motion dW_{it} over the chosen window.

- The increment dX_{it} has unconditional mean zero and conditional mean equal to $\kappa_i(m_i - X_{it})$. The forecast of expected daily returns is positive or negative according to the sign of $m_i - X_{it}$. It forecasts a negative return if X_{it} is high and a positive return if X_{it} is low.

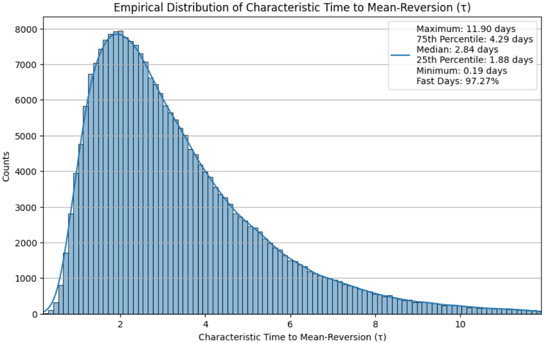
Used a 60-day trailing window to estimate both dX_{it} & parameters $(\kappa_i, m_i, \sigma_i)$.

Speed of Mean Reversion

- The speed of mean reversion κ_i or characteristic time-scale $\tau_i = 1/\kappa_i$ determine how quickly stocks return to their mean. For fast mean-reverting stocks:

$$\frac{1}{\kappa_i} \ll T_1 \quad \text{or} \quad \tau_i \ll T_1$$

- Using a 60-trading-day estimation window ($T_1 = 60/252$), we only select stocks with $\kappa > 252/30 = 8.4$, ensuring mean-reversion times below half the cycle.
- Over 95% of stocks met this criterion, with maximum mean-reversion times at 1/5 the period. When κ_i is large (i.e., $\kappa_i \gg 1$), the effect of drift α_i is negligible

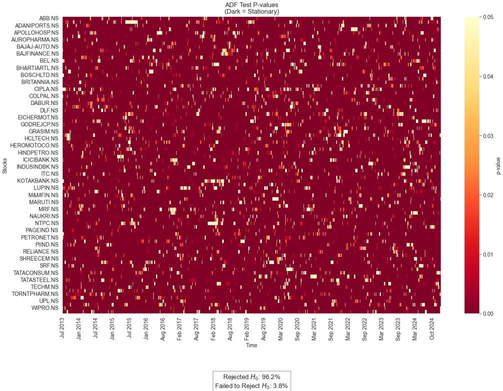


Augmented Dickey-Fuller (ADF) Test

- To validate stationarity, we conducted Augmented Dickey-Fuller (ADF) tests on residual process X_{it} for each 60-day estimation window using:

$$\Delta X_t = \alpha + \gamma X_{t-1} + \sum_{i=1}^{p-1} \delta_i \Delta X_{t-i} + \epsilon_t$$

- Testing $H_0 : \gamma = 1$ against $H_1 : |\gamma| < 1$ for the PCA strategy with 15 eigenportfolios showed 96.2% of stocks rejected the null hypothesis (p-value < 0.05), confirming residuals follow an Ornstein-Uhlembeck process.



Trading Signal

- The s-score represents the deviation of the process from its mean in units of standard deviation.
- It is theoretically defined as

$$s_i = \frac{X_{it} - m_i}{\sigma_{\text{eq},i}}$$

- Since the regression forces the residuals to have zero mean, $s_i = -\frac{m_i}{\sigma_{\text{eq,i}}}$.
- We correct for finite-sample bias in the estimated mean $\bar{m}_i = m_i - \langle m_i \rangle$.
- The final s-score, which serves as our trading signal, is then given by:

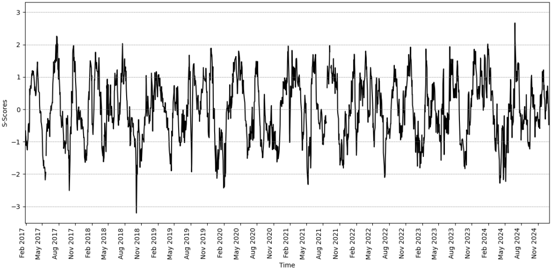
$$s_i = -\frac{\bar{m}_i}{\sigma_{\text{eq,i}}}$$

Action	**Condition**
Open a long position	$\boldsymbol{s_i} < -\bar{\boldsymbol{s}}_{bo}$
Open a short position	$\boldsymbol{s_i} > +\bar{\boldsymbol{s}}_{so}$
Close a short position	$\boldsymbol{s_i} < +\bar{\boldsymbol{s}}_{bc}$
Close a long position	$\boldsymbol{s_i} > -\bar{\boldsymbol{s}}_{sc}$

Cutoff values :

$\bar{s}_{bo} = \bar{s}_{so} = 1.25$

$\bar{s}_{bc} = 0.75$ and $\bar{s}_{sc} = 0.50$.



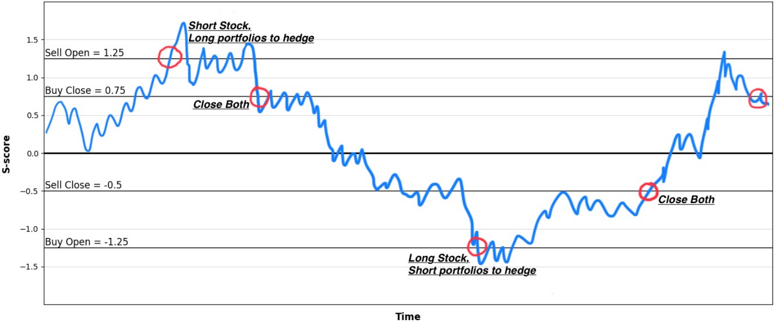
Trading Implementation

For a long position (triggered when $s_i < -\bar{s}_{bo}$):

- Long position: $1\$$ in stock i.
- Hedge position: Short β_{ijt} \$ of each portfolio j, where portfolio j consists of stocks weighted according to eigenweights v_{ijt}.

For a short position (when $s_i > \bar{s}_{so}$):

- Short position: $1\$$ in stock i.
- Hedge position: Long β_{ijt} \$ of each factor portfolio j, where portfolio j consists of stocks weighted according to eigenweights v_{ijt}.



The trading strategy that was used is 'bang-bang', there is no continuous trading.

Position Sizing & Risk Management

- The total investment Q_{it} for stock i at time t satisfies:

$$Q_{it} = \Lambda E_t \quad \begin{cases} \Lambda & \text{Fraction of portfolio value} \\ E_t & \text{Total portfolio value at time } t \end{cases}$$

- For long/short positions with along with the hedges, we can rewrite:

$$Q_{it} = k_{it} \left(1 + \sum_{j=1}^{m} \left[\frac{v_{ijt}}{\bar{\sigma}_{it}} \left(\beta_{ijt} \right) \right] \right) = \Lambda E_t$$

- Dollar position k_{it} represents the amount allocated to stock i to achieve target portfolio fraction Λ:

$$k_{it} = \frac{\Lambda E_t}{\left(1 + \sum_{j=1}^{m} \left[\frac{v_{ijt}}{\bar{\sigma}_{it}} \left(\beta_{ijt} \right) \right] \right)}$$

- This dynamic position sizing mechanism:
  - Adjusts positions based on portfolio value and factor exposures.
  - Reduces exposure to stocks with high factor sensitivities.
  - Ensures consistent risk-adjusted position sizing across the portfolio.

PnL Calculation

- The resulting profit and loss equations are:

$$PnL_{it} = k_{it} \cdot (R_{it} - \sum_{j=1}^{m} \beta_{ijt} F_{jt}), \quad \textit{when long}$$

$$PnL_{it} = k_{it} \cdot (-R_{it} + \sum_{j=1}^{m} \beta_{ijt} F_{jt}), \quad \textit{when short}$$

- The total factor exposure $\sum_{j=1}^{m} \beta_{ijt} F_{jt}$ is explicitly accounted in the PnL.
- Portfolio value updates at the beginning of each day:

$$E_{t+1} = E_t + \sum_{i=1}^{N} PnL_{it} - \sum_{i=1}^{N} |Q_{i,t+1} - Q_{it}|\delta$$

  where δ represents transaction costs accounted through a slippage factor.

- The leverage structure is maintained through $\Lambda = L/N$ where L represents our target leverage and N is the number of stocks in our universe. Increasing the Leverage increases the capital allocation.

Backtesting

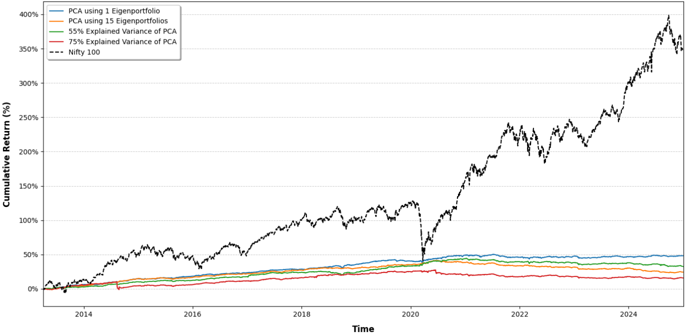
The variants tested include:

- **Using Standard Daily Returns:**
  - Using a single eigenportfolio as a risk factor
  - Using the first 15 eigenportfolios
  - Using enough eigenportfolios to explain 55% of the variance
  - Using enough eigenportfolios to explain 75% of the variance
- **Using Volume-weighted Returns:**
  - Using a single eigenportfolio as a risk factor
  - Using the first 15 eigenportfolios
  - Using enough eigenportfolios to explain 55% of the variance
- For each implementation, we evaluate the cumulative returns and the Sharpe Ratios. The Sharpe Ratio is calculated as:

$$\text{Sharpe Ratio} = \frac{\langle R_p - R_B \rangle}{\sigma(R_p - R_B)} \quad \begin{cases} R_P & \text{returns of the trading strategy} \\ R_B & \text{returns of Nifty 100 ETF (benchmark)} \end{cases}$$

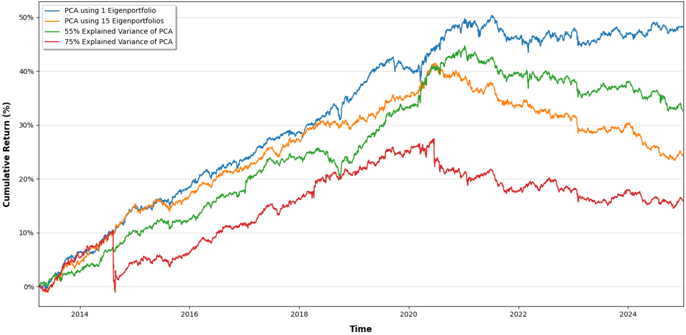
- This version of the Sharpe ratio measures the risk-adjusted performance of a strategy relative to a benchmark, with returns calculated assuming an initial capital of 1$ for both the trading strategy and investment in Nifty 100 ETF.

Backtesting Results: Using Standard Daily Returns



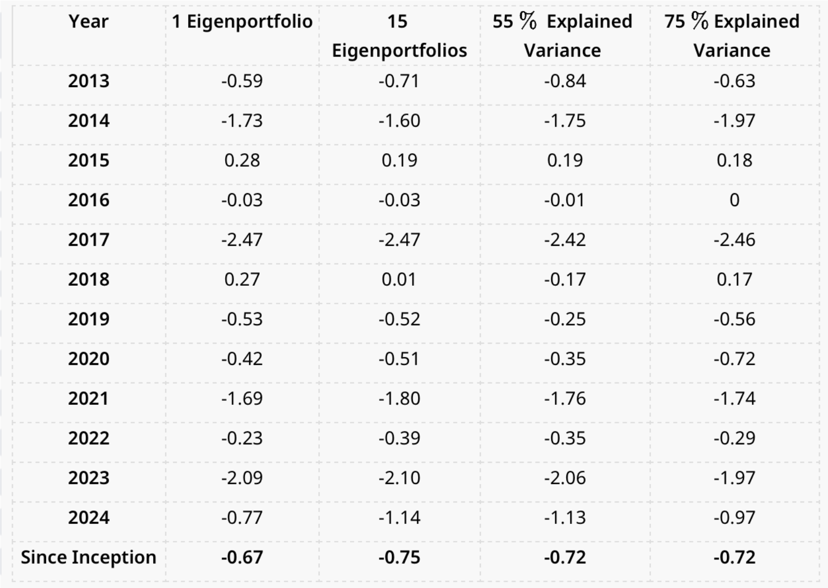
- The original Avellaneda & Lee study was conducted prior to the 2008 global financial crisis when market were more predictable, suggesting potential overfitting.
- The public dissemination of Avellaneda & Lee's methodology and optimized parameters led to widespread adoption, diminishing their effectiveness.
- The persistent bull market conditions during the period significantly limited the strategy's performance, as it diminished the shorting opportunities.
- The emergence of advanced high frequency trading reduced effectiveness of traditional StatArb approaches due to faster price discovery.

Backtesting: Using Standard Daily Returns



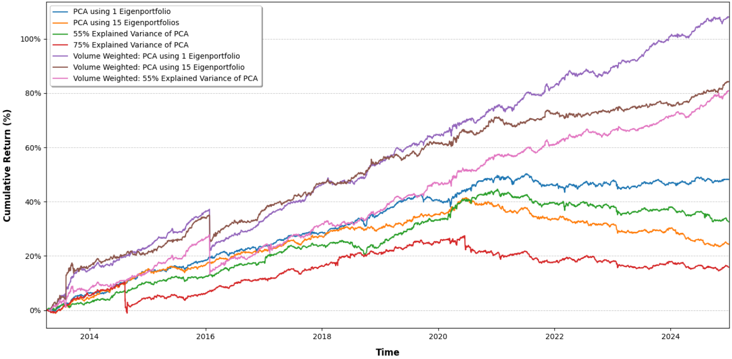
- The single eigenportfolio strategy outperforms multi-factor approaches, suggesting that in this particular market environment and timeframe, the marginal benefit of additional factors may be offset by increased complexity and trading costs.
- Additional noise trading from higher-order components lead to increased losses.
- Strategies incorporating higher percentages of explained variance, particularly the 75% threshold, demonstrate notably poor performance.
- This is primarily due to transaction costs dominating the small residual signals that remain in the system after removing the dominant factors.

Sharpe Ratio: Using Standard Daily Returns



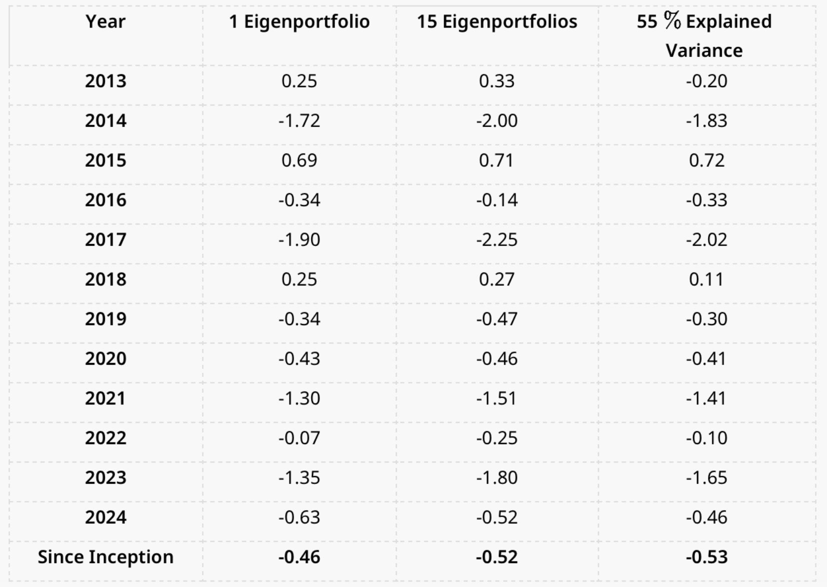
Year	1 Eigenportfolio	15 Eigenportfolios	55 % Explained Variance	75 % Explained Variance
2013	-0.59	-0.71	-0.84	-0.63
2014	-1.73	-1.60	-1.75	-1.97
2015	0.28	0.19	0.19	0.18
2016	-0.03	-0.03	-0.01	0
2017	-2.47	-2.47	-2.42	-2.46
2018	0.27	0.01	-0.17	0.17
2019	-0.53	-0.52	-0.25	-0.56
2020	-0.42	-0.51	-0.35	-0.72
2021	-1.69	-1.80	-1.76	-1.74
2022	-0.23	-0.39	-0.35	-0.29
2023	-2.09	-2.10	-2.06	-1.97
2024	-0.77	-1.14	-1.13	-0.97
Since Inception	**-0.67**	**-0.75**	**-0.72**	**-0.72**

Backtesting: Using Volume Weighted Returns



- The volume-weighted strategies demonstrated 50-60% excess returns compared to their standard daily return counterparts.
- The incorporation of trading volume information yielded significant performance improvements suggesting that volume contains important information.
- Volume-weighted single eigenportfolio strategy achieves approximately 100% cumulative returns.

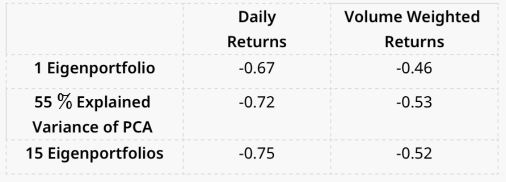
Sharpe Ratio: Using Volume Weighted Returns



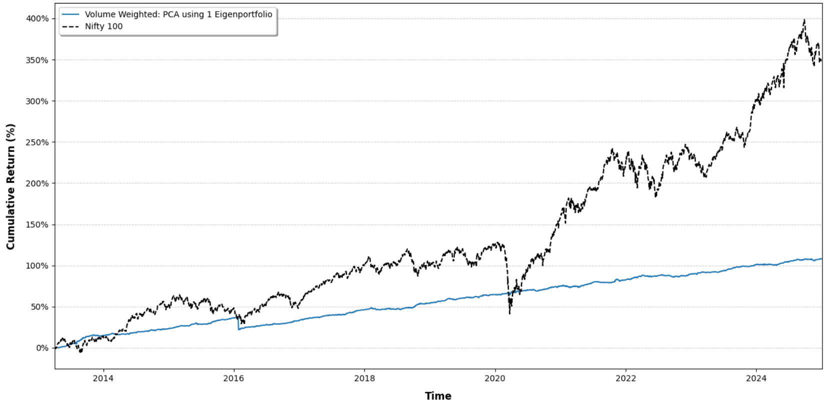
Year	1 Eigenportfolio	15 Eigenportfolios	55 % Explained Variance
2013	0.25	0.33	-0.20
2014	-1.72	-2.00	-1.83
2015	0.69	0.71	0.72
2016	-0.34	-0.14	-0.33
2017	-1.90	-2.25	-2.02
2018	0.25	0.27	0.11
2019	-0.34	-0.47	-0.30
2020	-0.43	-0.46	-0.41
2021	-1.30	-1.51	-1.41
2022	-0.07	-0.25	-0.10
2023	-1.35	-1.80	-1.65
2024	-0.63	-0.52	-0.46
Since Inception	**-0.46**	**-0.52**	**-0.53**

Backtesting: Best Performing Strategy

- The volume-weighted single eigenportfolio strategy achieved the highest Sharpe ratio, though still significantly under performing the Nifty 100 benchmark.

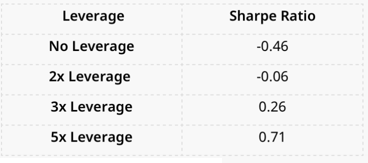


	Daily Returns	Volume Weighted Returns
1 Eigenportfolio	-0.67	-0.46
55 % Explained Variance of PCA	-0.72	-0.53
15 Eigenportfolios	-0.75	-0.52

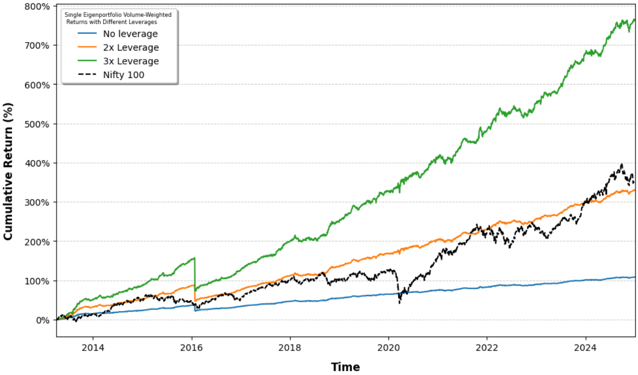


Backtesting: Impact of Leverage

- Increased leverage improves PnL and Sharpe ratios but introduces implementation challenges.
- Position sizing scales directly with leverage (2x leverage = 2x allocation per position).



Leverage	Sharpe Ratio
No Leverage	-0.46
2x Leverage	-0.06
3x Leverage	0.26
5x Leverage	0.71



- Higher leverage reduces capacity for simultaneous positions, increasing opportunity costs. Strategy requires stricter position limits to manager risk within capital constraints.

Conclusion: Strategy Performance & Insights

- The underperformance can be attributed to several factors:
  - Overfitting to pre-crisis period or shift towards more integrated markets post crisis?
  - Strategy alpha eroded by widespread adoption post Avellaneda & Lee publication.
  - Emergence of HFTs leading to faster price discovery, strategy not competitive enough?
  - Persistent bull markets during 2012-2024 which limited shorting opportunities.
- Incorporation of trading volume information yielded significant performance improvements suggesting that volume contains valuable information.
- Single eigenportfolio strategy outperformed complex approaches - optimal when few factors explain 50% variance, aligning with Avellaneda and Lee's observation.
- Using too few factors leaves market information in residuals, while too many factors reduce profits due to increased complexity and trading costs.
- Market stress increases asset correlation requiring fewer factors to explain more variance , while low volatility regimes require more factors.

Conclusion: Limitations & Future Research Directions

- PCA methodology faces dimensionality challenge, correlation matrix entries exceed data points: Asymptotic PCA (Tsay, 2010)?
- Given that our strategy fundamentally relies on identifying stationary residuals, cointegration analysis could identify all possible stationary linear combinations, Unlike PCA, which approximates the systematic risk of the entire stock universe using a limited set of eigenportfolios, residuals.
- Linear regression assumes Gaussian residuals and stationary spreads despite known stock return characteristics such as volatility clustering, heteroskedasticity & fat tails.
- Static OU process parameters for simplification may be too restrictive for financial markets where true parameter constancy is rare.
- Excluding non-stationary stocks from consideration to maintain model validity inherently restricts our investment universe, Developing modified versions of the Ornstein-Uhlembeck process that can account for non-stationary stocks could expand our trading opportunities and potentially enhance the strategy's performance by including a broader set of securities .

References

Avellaneda, M., & Lee, J. (2008). Statistical Arbitrage in the U.S. Equities Market. *Quantitative Finance*, 10(8), 761–782.

Krauss, C. (2017). Statistical Arbitrage Pairs Trading Strategies: Review and Outlook. *Journal of Economic Surveys*, 31(2), 513–545.

Di Nosse, D. M. (2022). Application of score-driven models in statistical arbitrage. (Master's thesis). Department of Physics, University of Pisa, Pisa, Italy.

Soo, C., Lian, Z., Yang, H., & Lou, J. (2017). Statistical Arbitrage. MSE448 Project, Stanford University.

Tsay, R. S. (2010). *Analysis of Financial Time Series* (3rd ed.). Wiley Series in Probability and Statistics. John Wiley & Sons, Hoboken, NJ.

Alexander, C. (2001). *Market Models: A Guide to Financial Data Analysis*. Wiley, Chichester, UK, and New York, NY.