# Statistical Arbitrage in Indian Equities

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

*Nitesh Rai*

## Abstract

This paper implements and critically examines Avellaneda and Lee's (2008) PCA-based statistical arbitrage strategy in the Indian equity market, using constituents of the NIFTY 100 index from 2012 to 2024. Given the limited availability of data for Indian stocks prior to 2012, and considering the strategy's documented decline in effectiveness post-2008, we opted for full sample backtesting rather than parameter optimization on historical data, as utilizing pre-2008 data (when the strategy was most effective) was not possible in the Indian market context. The strategy constructs market-neutral portfolios by decomposing stock returns into systematic (identified through PCA) and idiosyncratic returns. The key insight is modeling these idiosyncratic returns as mean-reverting Ornstein-Uhlembeck processes generating contrarian trading signals. We examine multiple variants of the strategy: implementing both fixed and adaptive methods of selecting risk factors—the latter based on cumulative explained variance thresholds—while also exploring the impact of volume-weighted returns. Our analysis reveals that while the basic strategy generates positive returns ranging from 20% to 50%, it substantially underperforms the benchmark's almost 350% return. The volume-weighted modifications yield superior results, with returns increased by 50-60% compared to their basic counterparts and the single eigenportfolio approach achieving approximately 100% cumulative returns. Notably, the single eigenportfolio strategy consistently outperforms multi-factor implementations, while strategies incorporating higher explained variance thresholds demonstrate poor performance due to transaction costs overwhelming smaller residual signals and increased noise from higher-order components. The strategy's overall underperformance can be attributed to several factors: potential overfitting to the pre-2008 period when markets were more predictable and less globally integrated, rapid erosion of arbitrage opportunities following the strategy's widespread adoption and public dissemination, limited shorting opportunities during our sample period's persistent bull market (with Nifty 100 gaining almost 350%), and the emergence of sophisticated algorithmic trading techniques that have reduced the effectiveness of traditional statistical arbitrage approaches.

## 1 Introduction

Mean reversion in financial markets refers to the tendency for extreme values to be followed by observations that move closer to the average. While this concept is well-documented for random variables, its application in financial markets requires nuanced analysis. Most asset price series exhibit characteristics of geometric random walks (integrated processes) rather than mean-reverting behavior. Instead, it is typically the returns that demonstrate a distribution centered around a zero mean and exhibit mean-reverting tendencies. However, directly trading on return's mean reversion is impractical.

The strength of Statistical Arbitrage (StatArb) lies in its ability to engineer mean-reverting price series by combining multiple non-mean-reverting series into a portfolio with a net market value that is mean-reverting. StatArb strategies share three key characteristics: (i) systematic trading signals based on rules rather than fundamentals, (ii) market-neutral portfolios with zero beta with market and (iii) statistical methods for generating excess returns. These strategies aim to achieve consistent returns with low volatility through diversified positions across multiple stocks, with holding periods varying from seconds to weeks.

StatArb strategies in equity markets can be broadly classified into following categories, as noted by Krauss (2017) [2]: distance-based approaches which use metrics to identify co-moving assets, cointegration-based strategies that exploit mean-reversion in linearly combined time series, time series approaches that model spreads as mean-reverting processes, stochastic control approaches focusing on optimal portfolio holdings, or model-driven approaches including Principal Component Analysis (PCA) for factor decomposition.

We focus on implementing the model-driven approach from Avellaneda and Lee (2008) [1], where trading signals are generated by decomposing stock returns into systematic factor returns identified through Principal Component Analysis (PCA) and idiosyncratic returns. The key insight of their approach is modeling the idiosyncratic returns—the component of stock returns unexplained by systematic component—as mean-reverting processes, which naturally leads to contrarian trading signals.

For example, consider a technology stock $S$ whose price movements can be decomposed into systematic and idiosyncratic returns. The stock's returns might be driven by broad market movements $M$, technology sector trends $T$, and company-specific factors. This relationship can be modeled as:

$$\frac{dS_t}{S_t} = \alpha\, dt + \beta_M \frac{dM_t}{M_t} + \beta_T \frac{dT_t}{T_t} + dX_t, \tag{1}$$

where $\alpha$ is the drift term, $\beta_M$ and $\beta_T$ represent the stock's sensitivities to market, while $M$ and $T$ represent the proxies for price idices for the broad market and technology sector at time t respectively, and $dX_t$ is the idiosyncratic component. However, actual price movements may deviate from this theoretical relationship due to imperfect correlations or company-specific factors. By constructing portfolios that neutralize exposure to these common factors, traders can isolate and profit from the mean-reversion of the idiosyncratic component $dX_t$.

This forms the foundation of a factor-based trading strategy: it assumes that stock prices will eventually align with their factor-implied levels. Traders can exploit this convergence by taking long positions in stocks that are undervalued relative to their factor exposures and short positions in those that are overvalued. As prices revert to their expected values, the resulting spread offers a profit opportunity.

For $N$ stocks in a portfolio, where stocks returns depend on returns of multiple factors $F_j$, this decomposition generalizes to:

$$R_i = \sum_{j=1}^{m} \beta_{ij} F_j + \tilde{R}_i \tag{2}$$

where $\beta_{ij}$ is the sensitivity of stock $i$ to the $j$-th factor $F_j$, and $\tilde{R}_i$ is the idiosyncratic return.

To construct a market-neutral portfolio, which is exclusively affected by idiosyncratic returns, the dollar amounts $\{Q_i\}_{i=1}^{N}$ invested in each of the N stocks must satisfy:

$$\bar{\beta}_j = \sum_{i=1}^{N} \beta_{ij} Q_i = 0 \quad \text{for all} \quad j = 1, 2, \ldots, m \tag{3}$$

where the coefficients $\bar{\beta}_j$ correspond to the portfolio betas, representing the projections of the portfolio returns on the different factors.

A market-neutral portfolio exhibits vanishing portfolio betas, making it uncorrelated with both the market portfolio and the factors driving market returns. Due to the market-neutral condition, the portfolio returns reduce to:

$$\sum_{i=1}^{N} Q_i R_i = \sum_{i=1}^{N} Q_i \tilde{R}_i \tag{4}$$

This demonstrates that a market-neutral portfolio's returns are determined solely by the idiosyncratic component $\tilde{R}_i$ of each stock's return.

## 2 Principal Component Analysis

Principal component analysis (PCA) is a fundamental technique in statistics, where dominant correlation patterns are extracted from high-dimensional data. At its core, PCA transforms high-dimensional data into a new coordinate system where the axes (principal components) are ordered by the amount of variance they explain in the data. The process begins with data centering by subtracting the mean: $\tilde{X} = X - \bar{X}$, where $X \in \mathbb{R}^{n \times p}$ is the original data matrix with n observations and p features. The covariance matrix is then computed as $\Sigma = \frac{1}{n-1} \tilde{X}^T \tilde{X}$, which captures the relationships between features. The key step in PCA involves finding the eigenvalues and eigenvectors of this covariance matrix through the equation $\Sigma v_i = \lambda_i v_i$, where $\lambda_i$ are the eigenvalues in descending order and $v_i$ are the corresponding orthonormal eigenvectors.

These eigenvectors represent the principal components, which describe independent sources of variation in the dataset. The first principal component $v_1$ corresponds to the direction of maximum variance in the data, with subsequent components $v_2, v_3, ...$ being orthogonal and capturing decreasing amounts of variance. The transformed data points are obtained through projection: $Z = \tilde{X}V$, where $V = [v_1, v_2, ..., v_p]$ is the orthogonal matrix of eigenvectors and $Z$ contains the projected data points. 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 and sums to unity $\sum_{i=1}^{p} r_i = 1$. The original data can be approximately reconstructed using k principal components through $X \approx \bar{X} + ZV_k^T$, where $V_k$ contains the first $k$ eigenvectors, allowing for effective dimensionality reduction while preserving the most important features of the data.

In their paper [1], Avellaneda and Lee applied PCA to decompose the correlation matrix of standardized daily stock returns into ranked components, each representing a distinct market pattern. The first principal component, accounting for the largest variance, typically captures broad market movements affecting all stocks. Subsequent components reveal more nuanced patterns like sector-specific trends or other systematic factors, enabling identification of underlying market dynamics not visible in raw price data.

The PCA approach involves using historical daily price data for a cross-section of $N$ stocks over an $M$-day period. To convert a non-stationary price series to a stationary one, we compute returns. The return for stock $i$ on day $t$, denoted as $R_{it}$, is calculated as:

$$R_{it} = \frac{S_{it} - S_{i,t-1}}{S_{i,t-1}} \quad \text{for} \quad t = 1, \dots, M \quad \text{and} \quad i = 1, \dots, N$$

where $S_{it}$ is the price of stock $i$ on day $t$, adjusted for dividends. This represents the standard daily return or percentage change in price.

We also calculated volume-weighted returns incorporating trading volume into the strategy. The volume-weighted return for stock $i$ on day $t$, denoted as $\bar{R}_{it}$, adjusts the classical return $R_{it}$ by accounting for the relationship between typical daily trading volume, $\langle \delta V_i \rangle$, and recent volume changes, $\Delta V_{it} = V_{it} - V_{i,t-1}$. Specifically, the volume-weighted returns are defined as:

$$\bar{R}_{it} = R_{it} \cdot \frac{\langle \delta V_i \rangle}{|V_{it} - V_{i,t-1}|} \quad \text{for} \quad t = 1, \dots, M \quad \text{and} \quad i = 1, \dots, N$$

We considered a 10-day lookback period for calculating the typical daily volume, $\langle \delta V_i \rangle$, using a rolling mean window. The modified returns are equivalent to the classical returns when the daily trading volume is typical, i.e., when $\Delta V_{it} \approx \langle \delta V_i \rangle$. If the trading volume is low, the adjustment factor $\frac{\langle \delta V_i \rangle}{\Delta V_{it}}$ becomes greater than unity, resulting in $\bar{R}_{it} > R_{it}$. Conversely, if the trading volume is high, the adjustment factor is less than unity, leading to $\bar{R}_{it} < R_{it}$.

This volume-weighted adjustment takes into account mean-reversion strategy's sensitivity to trading activity preceding signal triggers. For example, when a stock rallies on high volume, while a classical return calculation might trigger an open-to-short signal, the modified return would be significantly reduced under these conditions, potentially suppressing the shorting signal. Similarly, the approach discourages buying stocks that experience sharp drops on high volume, as the modified returns account for these volume spikes.

For this analysis, we selected the constituents of NIFTY 100 index of Indian Market that covers 66.4% of the free float market capitalization of the stocks listed in National Stock Exchange (NSE). It encompasses 17 major industry sectors, providing comprehensive coverage of India's economic landscape. The data used is daily adjusted closing prices and daily volume from Yahoo Finance, which accounts for corporate actions such as stock splits and dividend distributions. After filtering for stocks with consistent data availability during the study period (2012-2024), the final sample consists of 84 stocks out of 100 stocks. The reason for choosing this period was that data prior to 2012 was available for very few stocks, limiting the scope for reliable analysis.

The return data is standardized to ensure PCA results are not skewed by varying stock volatilities. Since PCA maximizes variance, standardization removes scale dependencies by equalizing the impact of each stock's volatility on the analysis. The standardized return $Y_{it}$ for each stock $i = 1, \ldots, N$ is computed as:

$$Y_{it} = \frac{R_{it} - \bar{R}_i}{\bar{\sigma}_i}$$

where $\bar{R}_i$ is the mean return and $\bar{\sigma}_i$ is the standard deviation of returns for stock $i$:

$$\bar{R}_i = \frac{1}{M} \sum_{t=1}^{M} R_{it}$$

$$\bar{\sigma}_i^2 = \frac{1}{M-1} \sum_{t=1}^{M} (R_{it} - \bar{R}_i)^2$$

The empirical correlation matrix $\rho_{ij}$ is calculated as:

$$\rho_{ij} = \frac{1}{M-1} \sum_{t=1}^{M} Y_{it} Y_{jt} \tag{5}$$

We have a 84 by 84 correlation matrix, which corresponds to the 84 stocks included in the final sample. A key challenge is selecting the optimal estimation window. The correlation matrix estimated is relatively small compared to typical dimensions of 500 by 500 or 1000 by 1000 of a correlation matrix, we still face the fundamental trade-off in estimation windows. A long window risks including economically irrelevant market relationships from the distant past, while a short window must contend with having more correlation matrix entries than data points. Following the literature, we chose a one-year rolling window (252 trading days) to estimate the correlation matrix. We employed PCA to decompose the correlation matrix into its constituent eigenvectors and eigenvalues. In this eigendecomposition, the eigenvalues are ordered in decreasing magnitude, representing the variance explained by each principal component:

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_N \geq 0$$

A larger eigenvalue indicates that its corresponding eigenvector (principal component) captures more of the total variance in the return series. Each eigenvector at time $t$ is composed of $N$ elements:

$$v_{jt} = \big(v_{1jt}, v_{2jt}, \ldots, v_{Njt}\big),$$

where $v_{ijt}$ represents the loading of stock $i$ in the $j$-th principal component at time $t$. The time dependence of the eigenvectors, $v_{jt}$, arises from the rolling window approach used in the correlation matrix estimation.

We can see that the first component captures the largest proportion of total variance (Figure 1). Each subsequent component explains progressively less variance and are orthogonal to each other, as per the construction of PCA, ensuring that they represent independent sources of variance in the data. The density distribution of the explained variance ratios highlights a characteristic 'bulk spectrum' centered near zero, with a few significant eigenvalues clearly separated from this bulk (Figure 2). These significant eigenvalues correspond to dominant market factors, while the eigenvalues centered near zero represent the 'noise spectrum' that captures less important dynamics.

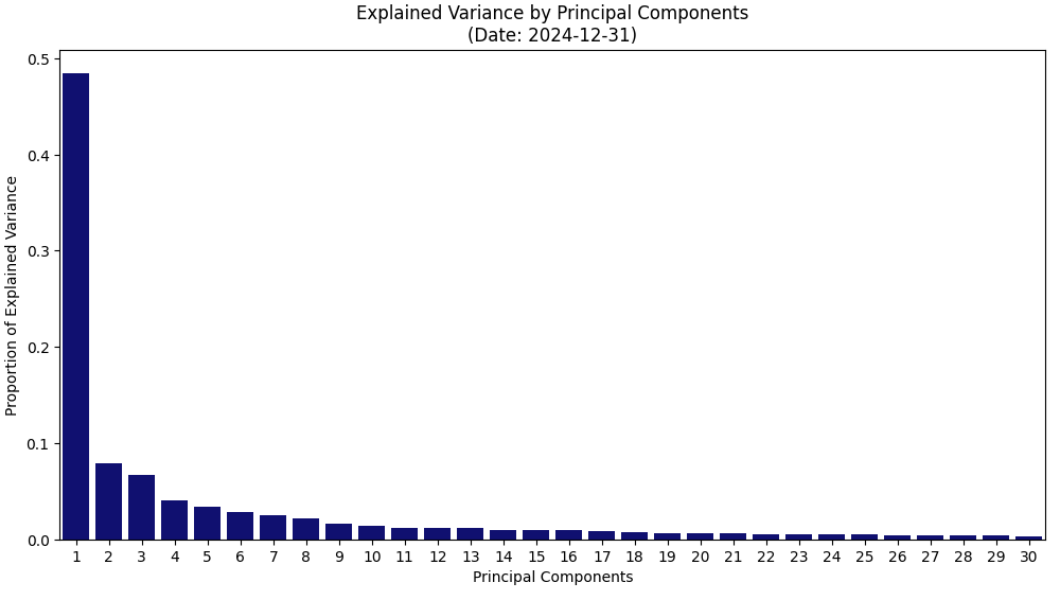


Fig. 1: Top 30 Eigenvalues of the correlation matrix of standard daily returns computed at the close of December 31 2024 estimated using a 1-year window and a universe of 84 stocks.(Eigenvalues measured as proportion of explained variance).

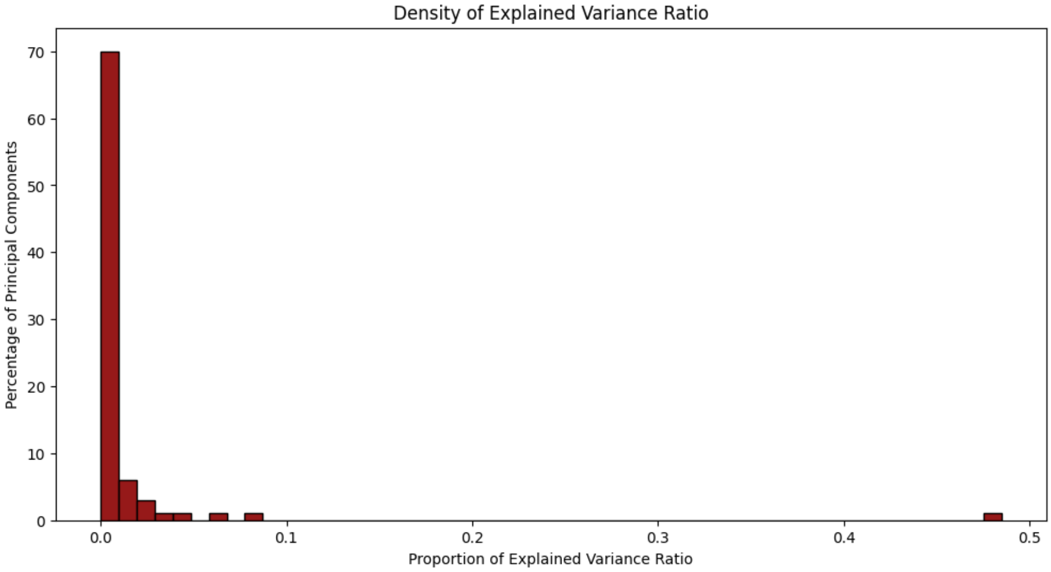


Fig. 2: Density of Explained Variance Ratio estimated using a 1-year window corresponding to the same data used to generate Figure 1.

As we can see that there are fewer 'detached' eigenvalues than industry sectors in our universe, suggesting that the boundary between 'significant' and 'noise' eigenvalues is somewhat blurred and corresponds to the edge of the 'bulk spectrum'. This presents two potential approaches for factor extraction: selecting a fixed number of eigenvalues approximately matching the number of industry sectors, or adopting a variable approach where eigenvalues are retained until their cumulative sum exceeds a specified proportion of explained variance. The methodological choice between fixed and variable eigenvalue selection becomes crucial when analyzing market dynamics.

During periods of market stress, we observe that fewer eigenvectors explain a larger proportion of total variance, indicating a market environment where assets move more uniformly. In contrast, during calmer periods, assets tend to move independently, driven by industry-specific factors, company size, and individual characteristics, requiring more eigenvectors to capture market movements. This pattern emerges consistently across major market events, as shown in Figure 3 and 4.

The variance explained by the top 15 eigenvectors exhibits notable increases during the COVID-19 market crash and the June 2024 Indian election's result day crash, periods and events marked by heightened market uncertainty. Correspondingly, these same periods show a significant reduction in the number of eigenvectors needed to explain 55 % of total variance, demonstrating how market stress influences the collective behavior of assets.

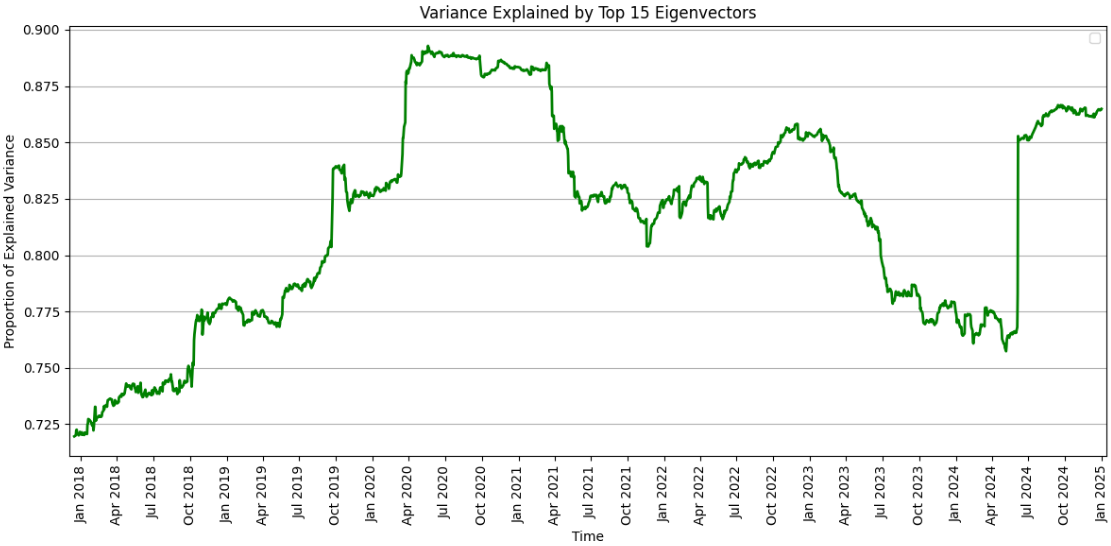


Fig. 3: Proportion of Variance Explained by Top 15 Eigenvectors: 2018-2024 (Notice the increase during COVID-19 crash and the spike in June 2024)

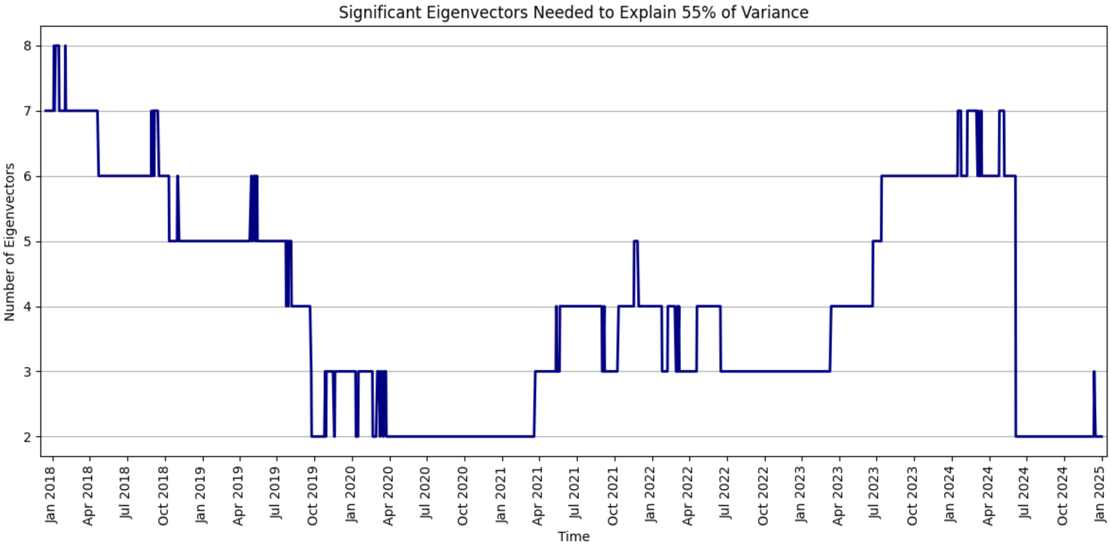


Fig. 4: Number of significant eigen vectors needed to explain the variance of correlation matrix estimated using a 252 days window at 55 % level : 2018-2024

Coming back to the first eigenvector, it is typically characterized by uniformly positive coefficients (Figure 5), is interpreted as the market's systematic component—essentially the 'market portfolio'. The stocks in this eigenvector move in the same direction, suggesting a unified market trend. All subsequent eigenvectors must contain negative components (Figure 6 and 7) to maintain orthogonality with the first eigenvector. These components may reflect sectoral trends or other unique patterns within the market, becoming progressively narrower in scope and less representative of overall market dynamics.

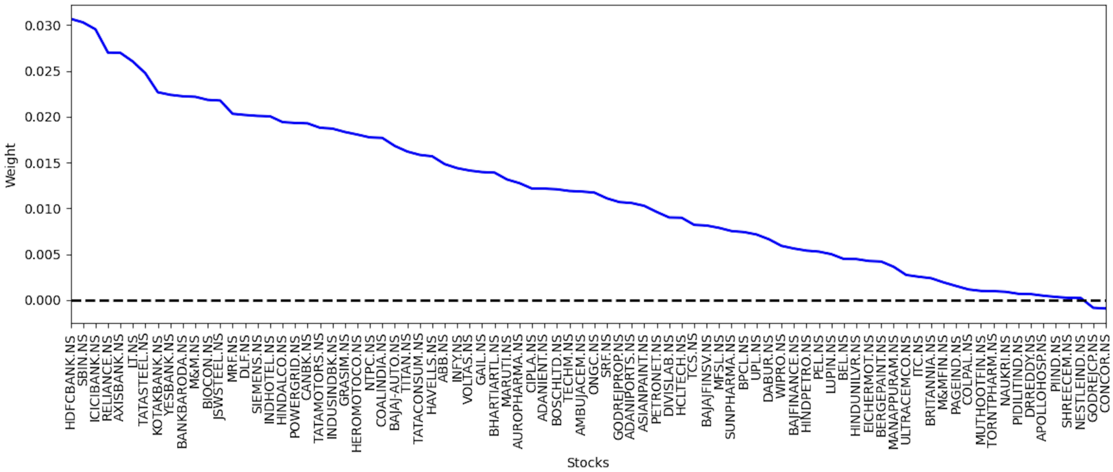


Fig. 5: First eigenvector sorted by coefficient size.

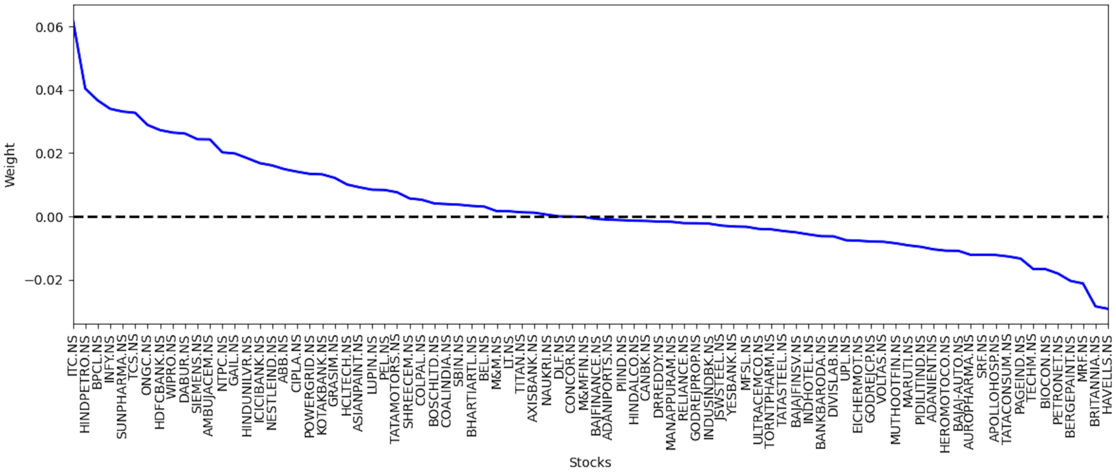


Fig. 6: Second eigenvector sorted by coefficient size.

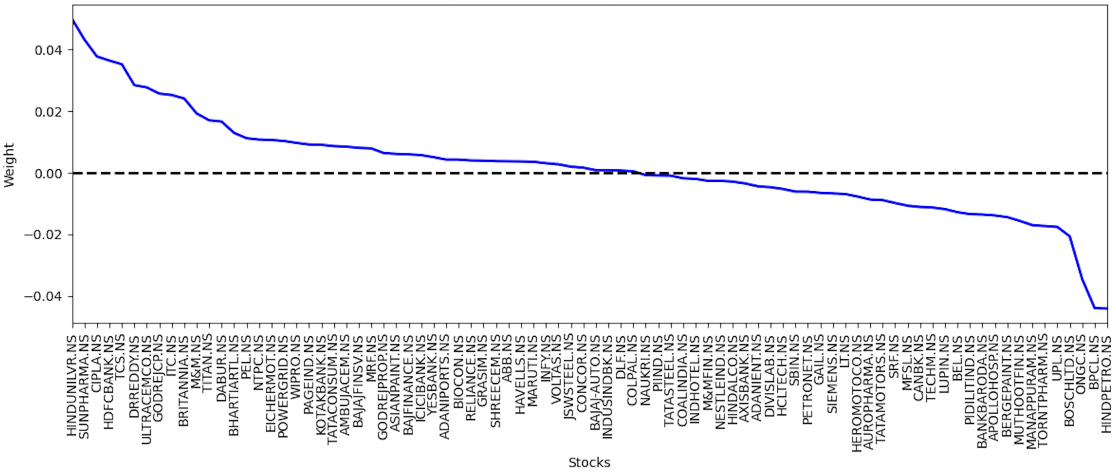


Fig. 7: Third eigenvector sorted by coefficient size.

A key characteristic of these higher-ranking eigenvectors is their 'coherence' property - stocks with similar coefficient values tend to be from the same industry groups. This coherence is particularly strong in the high-ranking components, but gradually weakens as we move towards eigenvectors associated with smaller eigenvalues in the noise spectrum, where coefficient patterns become more scattered and industry relationships less distinct. As we can observe, in the first eigenvector, the top 10 stocks predominantly belong to the Banking sector, whereas the bottom 10 are primarily from the Fast-moving consumer goods (FMCG) sector. Similarly, in the third eigenvector, the top 10 stocks are dominated by the Pharmaceutical sector, while the bottom 10 are from Oil and Gas.

These eigenvectors are used to construct eigenportfolios, for each index $j = 1, \ldots, m$, the corresponding eigenportfolio is constructed by the investing in the respective stocks $i = 1, \ldots, N$ with weights proportional to the elements of the eigenvector:

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

We can see the weights are inversely proportional to the volatility of the stocks. This approach aligns with the principle of capitalization-weighting, as companies with larger capitalizations generally exhibit lower volatilities. As discussed that the first eigenvector is interpreted as the 'market portfolio', the principal (first) eigenportfolio and capitalization-weighted portfolios serve as effective proxies for one another (Figure 8). With the weights obtained, we can now calculate the returns for each eigenportfolio using the following equation:

$$F_{jt} = \sum_{i=1}^{N} \frac{v_{ijt}}{\bar{\sigma}_{it}} R_{it} \tag{6}$$

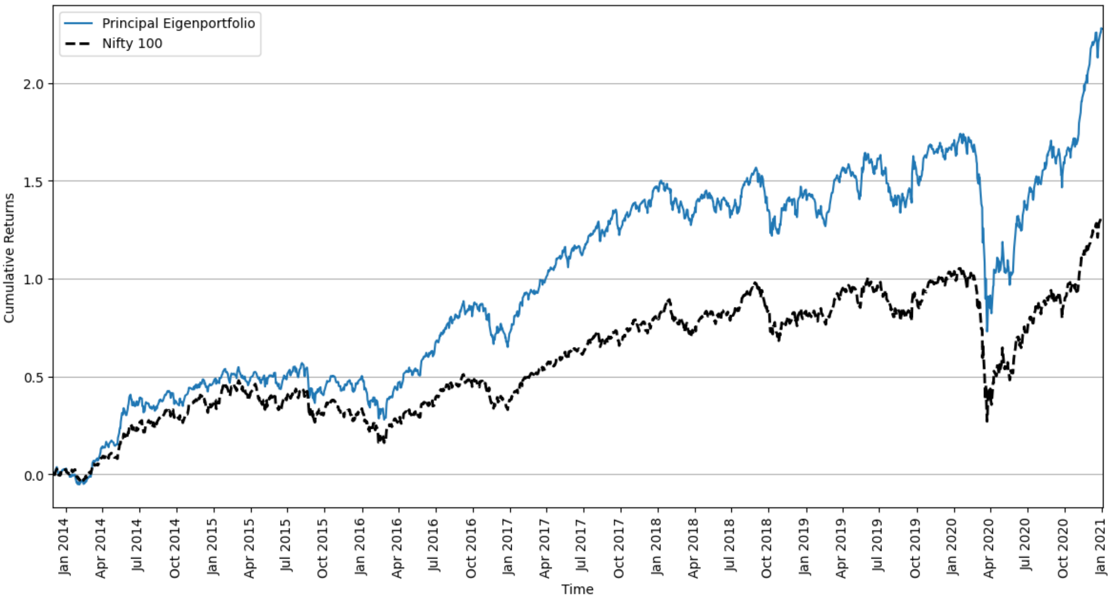


Fig. 8: Cumulative returns of the principal eigenportfolio and capitalization-weighted benchmark Nifty 100

Each stock return in the investment universe can be decomposed into its projection on the $m$ factors and a residual, as in Equation [2]. Thus, the PCA approach delivers a natural set of risk-factors that can be used to decompose our stock returns into systematic and idiosyncratic components.

## 3 Mean Reversion Modelling

As shown in Equation [2], in a multi-factor model, stock returns follow the system of stochastic differential equations:

$$\frac{dS_{it}}{S_{it}} = \alpha_i \, dt + \sum_{j=1}^{m} \beta_{ijt} F_{jt} + dX_{it}, \tag{7}$$

where the term

$$\sum_{j=1}^{m} \beta_{ijt} F_{jt}$$

captures the systematic component of returns and $\beta_{ijt}$ denotes the factor loadings. Here, $F_{jt}$ denotes the $j$-th eigenportfolio returns.

The idiosyncratic component of the return is given by:

$$\alpha_i \, dt + dX_{it},$$

where $\alpha_i$ represents the drift of the idiosyncratic component i.e. term $\alpha_i \, dt$ reflects the excess rate of return of stock $i$ relative to the systematic components over the relevant period, for simplicity we will only discuss the case with $\alpha_i = 0$, as its usually small in practice. The residual $dX_{it}$ is assumed to be the increment of a stationary stochastic process that models fluctuations in the stock price not explained by the systematic components.

What interests us is this residual, as it represents the stock's unique, idiosyncratic fluctuations. By analyzing the behavior of $dX_{it}$ , we can gain a deeper understanding of how $X_{it}$ , the stock's deviation from its equilibrium value, evolves over time.

Avellaneda and Lee assumed that the $dX_{it}$ is a stationary process and thus modelled it using Ornstein-Uhlembeck process which is stationary, mean reverting and auto-regressive process with lag 1:

$$dX_{it} = \kappa_i (m_i - X_{it}) \, dt + \sigma_i \, dW_{it}, \qquad \kappa_i \ > 0 \tag{8}$$

In particular the increment $dX_i(t)$ has unconditional mean zero and conditional mean equal to $\kappa_i(m_i - X_{it})$ and forecast of expected daily returns is positive or negative according to the sign of $m_i - X_{it}$.

The parameters drift $\alpha_i$, speed of mean reversion $\kappa_i$, long run mean $m_i$ and diffusion coefficient $\sigma_i$ are specific to each stock and are assumed to vary slowly relative to the increments of the Brownian motion $dW_{it}$ over the chosen time window. The first term in Equation [8] $\kappa_i(m_i - X_{it})$ reflects the model's prediction: it forecastes a negative return if $X_{it}$ is high and a positive return if $X_{it}$ is low.

The estimation of OU parameters (A detailed derivation and discussion of the calculations are provided in **Appendix of** [1]) is done over a window assuming that parameter are constant over that window. This is done by fitting a one-lag regression model allowing us to calculate the parameters :

$$\kappa_i = -\log(b_i) \times 252, \quad m_i = \frac{a_i}{1 - b_i}, \quad \tau_i = \frac{1}{\kappa_i}$$

$$\sigma_i = \sqrt{\frac{\text{Variance}(\zeta_i) \times 2\kappa_i}{1 - b_i^2}}, \quad \sigma_{\text{eq},i} = \frac{\sigma_i}{\sqrt{1 - b_i^2}}$$

Here, $\kappa_i$ represents the speed of mean reversion and $\tau_i$ is characteristic time-scale for mean reversion. If $\kappa_i$ is large (i.e., $\kappa_i \gg 1$), the stock reverts quickly to its mean, making the effect of drift $\alpha_i$ negligible. Given that the parameters are assumed constant, we are particularly interested in stocks where the mean-reversion speed is fast, such that:

$$\frac{1}{\kappa_i} \ll T_1,$$

To align with the estimation procedure that assumes constant parameters, we focus on stocks that exhibit fast mean reversion, meaning those where

$$\tau_i \ll T_1,$$

Here, $T_1$ represents the estimation window for residuals, measured in years.

Similar to Avellaneda and Lee, we chose a 60-business-day estimation window ($T_1 = 60/252$), which spans at least one earnings cycle for the company, capturing price fluctuations within that period. We focused on stocks with mean-reversion times less than half the cycle i.e.

$$\kappa > 252/30 = 8.4$$

The mean reversion time filter helps minimize the duration of holding positions, reducing the risk of our OU model assumptions and parameter estimation process. Across all backtesting experiments, the mean-reversion criterion was consistently met by over 95% of stocks in the sample as also shown in Figure 9. In the paper, they select stocks with mean-reversion times less than 1/2 the period. For us, the longest mean-reversion time is 1/5 the period. Thus, we are clearly in the safe zone.

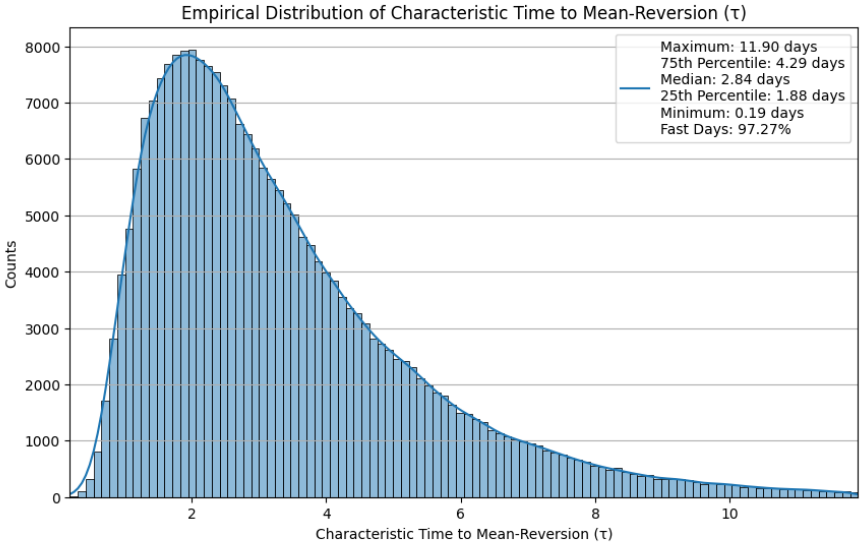


Fig. 9: Empirical Distribution of characteristic time to mean reversion for the variant: PCA with 15 eigenportfolios using standard daily returns along with its Descriptive Statistics

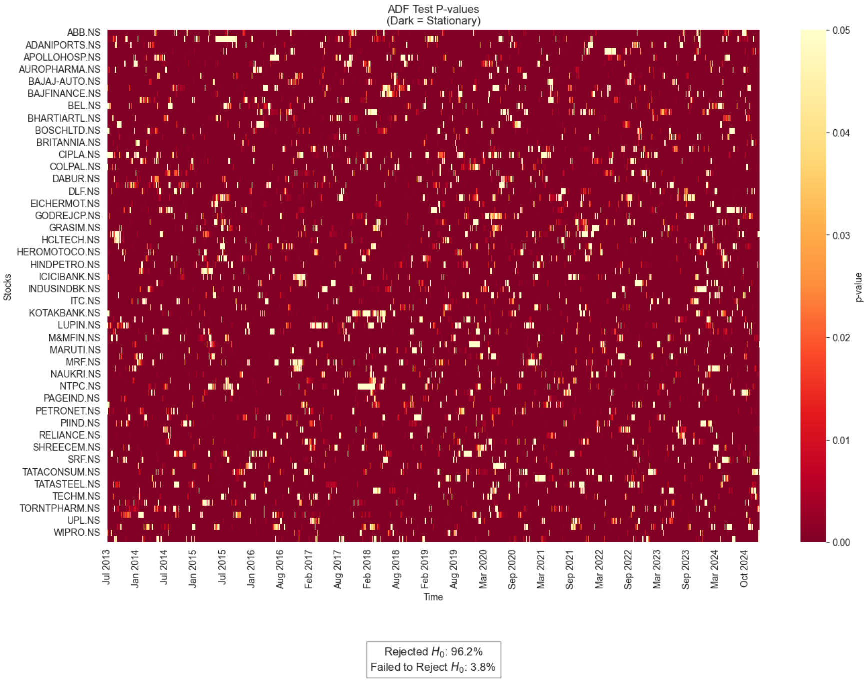


Fig. 10: Heat Map of p-values obtained from ADF Test computed on residuals using a estimation window length of 60 days for the variant: PCA with 15 eigenportfolios using standard daily returns.

Additionaly, to validate our stationarity assumption, we performed the Augmented Dickey-Fuller (ADF) test on the residual process $X_{it}$ for each 60-day estimation window. The test examines whether the time series contains a unit root, based on the regression equation:

$$\Delta X_t = \alpha + \gamma X_{t-1} + \sum_{i=1}^{p} \delta_i \Delta X_{t-i} + \epsilon_t \tag{9}$$

The null hypothesis of a unit root ($H_0 : \gamma = 1$) was tested against the alternative hypothesis of stationarity ($H_1 : |\gamma| < 1$). While testing on residuals for strategy with PCA using 15 eigenportfolios , the analysis showed that 96.2% of the stocks (Figure 10) in the sample consistently rejected the null hypothesis (p-value $< 0.05$) across their estimation windows, confirming the suitability of modeling these residuals as an Ornstein-Uhlembeck process.

## 4 Trading Implementation

The s-score is theoretically defined as:

$$s_{it} = \frac{X_{it} - m_i}{\sigma_{\text{eq},i}} \tag{10}$$

neglecting the drift $\alpha_i$ . Since $X_{it} = X_{60} = 0$ as the regression forces the residuals to have zero mean, although it wont be the same in case if we use different estimation windows for regression and residual processes, the s-score simplifies to:

$$s_{it} = -\frac{m_i}{\sigma_{\text{eq,i}}} = \frac{-a_i\sqrt{1-b_i^2}}{(1-b_i)\sqrt{\text{Variance}(\zeta_i)}}.$$

To correct for finite-sample bias in the estimated mean $m_i$, we adjust it by subtracting the cross-sectional average across all stocks:

$$\bar{m}_i = \frac{a_i}{1-b_i} - \langle \frac{a_i}{1-b_i} \rangle,$$

where $\bar{m}_i$ is the bias-adjusted mean for stock $i$ and $\langle \cdot \rangle$ denotes the cross-sectional average operator across all stocks. The final s-score, which serves as our trading signal, is then given by:

$$s_{it} = -\frac{\bar{m}_i}{\sigma_{\text{eq,i}}} \tag{11}$$

The s-score quantifies how far a residual deviates from its equilibrium value in terms of standard deviations. In other words, it indicates the distance of a stock from its expected equilibrium value according to our model. A positive s-score indicates potential negative mean reversion, while a negative s-score suggests potential positive mean reversion, creating distinct trading opportunities in both directions. In Figure 11 we can see the evolution of the s-score for Reliance Industries Ltd.

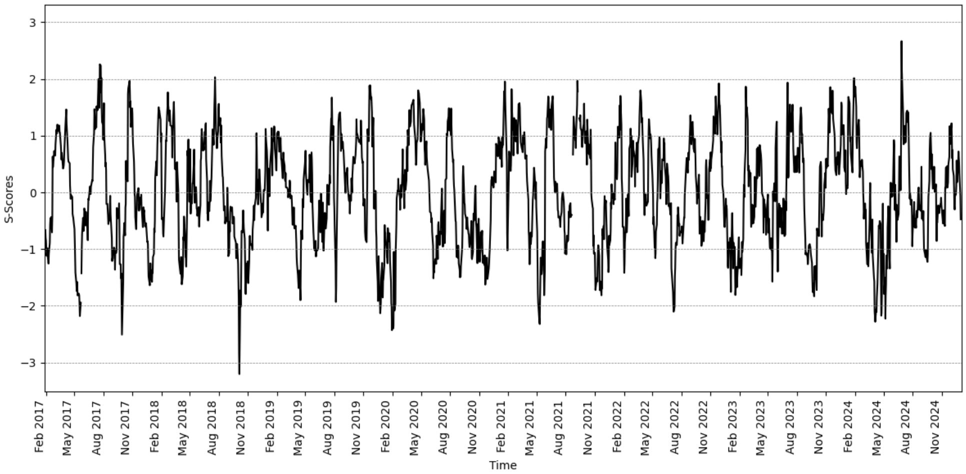
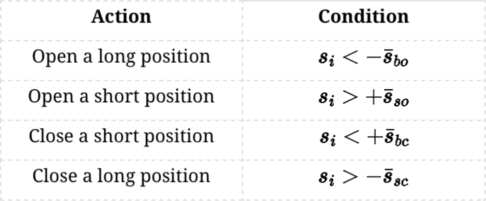


Fig. 11: Evolution of s-score of Reliance Industries Limited from January 2017 to December 2024

This s-score serves as the basis for generating trading signals. The trading rules are summarized in the following table:



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

Tab. 1: Trade Signal Conditions

The cutoff values we used are : $\bar{s}_{bo} = \bar{s}_{so} = 1.25$, $\bar{s}_{bc} = 0.75$, and $\bar{s}_{sc} = 0.50$, similar to those that are used in the paper. They optimized these values on the same data set on which they tested the strategy, this is not the best procedure to implement in order to find optimal hyper parameters as it leads to overfitting. The rationale behind this strategy is to enter trades only when the s-score reflects a substantial deviation from the equilibrium, as detected by a value exceeding 1.25 in absolute terms. Closing trades at $\bar{s}_{bc}$ and $\bar{s}_{sc}$ before reaching zero ensures better performance by acknowledging that stocks typically revert to equilibrium. By focusing on significant deviations and assuming these deviations will revert to the mean within a time frame roughly equal to the mean-reversion time $\tau_i$, we aim to capture the profitable portion of the mean-reversion process while minimizing the risk of prolonged exposure to market.

The implementation of trading positions and their corresponding hedges follows this framework:

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

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

When closing positions (triggered by $s_i > -\bar{s}_{sc}$ for long positions or $s_i < \bar{s}_{bc}$ for short positions), we simultaneously unwind both the stock position and its corresponding hedge components. This involves reversing the initial trades: selling (buying) the long (short) stock position and buying (selling) back the hedge eigenportfolio positions (Figure 12). This implementation achieves multi-factor neutrality by simultaneously hedging against all significant eigenportfolios.

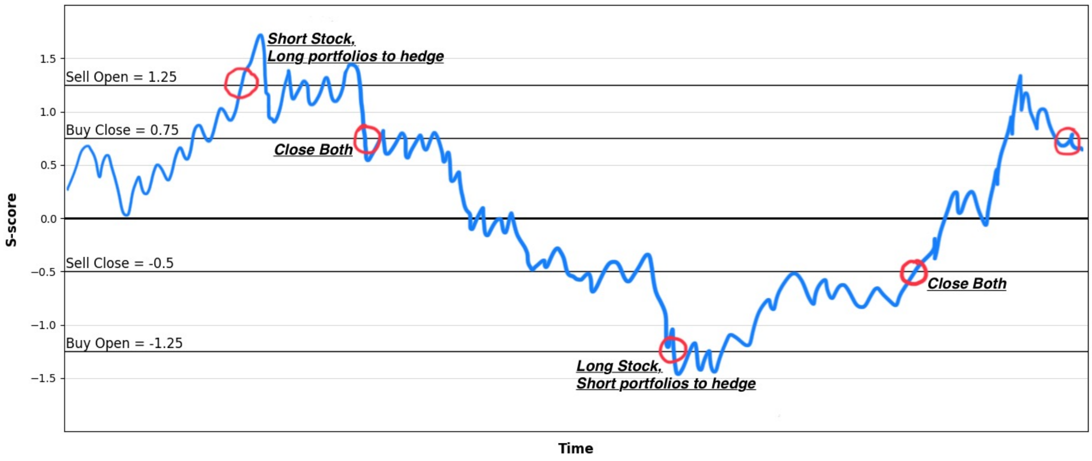


Fig. 12: Schematic evolution of the s-score and the associated signal, or trading rule

The trading strategy that was used is 'bang-bang', there is no continuous trading. Instead, the full amount is invested on the stock once the signal is active (buy-to-open, short-to-open) and the position is unwound when the s-score indicates a closing signal and all the positions are entered and exited at daily closing prices.

After determining when to enter and exit trades based on the s-score thresholds, the next crucial step is determining the appropriate position size for each trade. We account for both the total portfolio value and the risk exposure from factor sensitivities to determine the position size.

The total investment $Q_{it}$ for stock $i$ at time $t$ must satisfy:

$$Q_{it} = \Lambda E_t \tag{12}$$

where $\Lambda$ is the fraction of the portfolio value invested in each trade and $E_t$ represents the total portfolio value at time $t$. Since the algorithm goes long (short) one unit on the stock and short (long) the portfolios proportionally to $\beta_{ijt}$, we can rewrite the equation:

$$Q_{it} = k_{it} \left( 1 + \sum_{j=1}^{m} \left[ \frac{v_{ijt}}{\bar{\sigma}_{it}} (\beta_{ijt}) \right] \right) = \Lambda E_t \tag{13}$$

The dollar position $k_{it}$ represents the amount allocated to stock $i$ to achieve the target portfolio fraction $\Lambda$ while accounting for factor risk exposures [3].

$$k_{it} = \frac{\Lambda E_t}{\left( 1 + \sum_{j=1}^{m} \left[ \frac{v_{ijt}}{\bar{\sigma}_{it}} (\beta_{ijt}) \right] \right)} \tag{14}$$

This dynamic position sizing mechanism automatically reduces exposure to stocks with high factor sensitivities, maintains consistent risk-adjusted position sizes across the portfolio and scales positions based on current portfolio value and factor exposures.

The profit and loss calculation for each stock position, incorporating the position sizing, is:

$$PnL_{it} = k_{it} \cdot ((-1)^{1-p_i} R_{it} + \sum_{j=1}^{m} (-1)^{p_i} \beta_{ijt} F_{jt}) \tag{15}$$

The resulting profit and loss for a long position when $p_i = 1$ (Equation 16) and short position when $p_i = 0$ (Equation 17) are:

$$PnL_{it} = k_{it} \cdot (R_{it} - \sum_{j=1}^{m} \beta_{ijt} F_{jt}) \tag{16}$$

$$PnL_{it} = k_{it} \cdot (-R_{it} + \sum_{j=1}^{m} \beta_{ijt} F_{jt}) \tag{17}$$

The risk management framework incorporates two key hedging mechanisms:

- The total factor exposure $\sum_{j=1}^{m} \beta_{ijt} F_{jt}$ is explicitly accounted in the PnL calculation
- Dynamic risk adjustment via position sizing $k_{it}$ that scales inversely with factor exposures.

The portfolio value is updated at the beginning of each day based on:

$$E_{t+1} = E_t + \sum_{i=1}^{N} PnL_{it} - \sum_{i=1}^{N} |Q_{i,t+1} - Q_{it}| \delta \tag{18}$$

where we incorporate transaction costs through a slippage factor $\delta$=0.0005 similar to the paper. 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. For instance, in a portfolio targeting 5x leverage across 100 stocks, $\Lambda$ would be set to 5/100, increasing allocation by 5x while maintaining the desired risk exposure.

# 5 Backtesting Results

We implement several variants of the strategy that differ in how the risk factors are determined and weighted. Beyond using a fixed number of eigenportfolios, we also investigate adaptive factor selection approaches where the number of eigenportfolios is determined by the cumulative explained variance. This allows the model to dynamically adjust the number of factors based on the underlying market structure. Additionally, we examine whether incorporating trading volume information into the return calculations improves the strategy's performance.

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 Ratio as a measure of risk-adjusted performance. The Sharpe Ratio is calculated using the following formula:

$$\text{Sharpe Ratio} = \frac{\langle R_p - R_B \rangle}{\sigma(R_p - R_B)} \tag{19}$$

where $R_p$ represents the returns of the trading strategy, and $R_B$ represents the returns of Nifty 100 ETF (benchmark). The term $\langle R_p - R_B \rangle$ denotes the mean of the excess returns $(R_p - R_B)$, while $\sigma(R_p - R_B)$ represents the standard deviation of the excess returns $(R_p - R_B)$. 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.

## Using Standard Daily Returns

Figure 13 compares the cumulative returns of all standard daily return variants against the Nifty 100 benchmark. While the strategy variants generated profits ranging from 20% to 50%, they substantially underperformed the benchmark's 350% return.

This underperformance can be attributed to several factors:

- The original Avellaneda & Lee study was conducted prior to the 2008 global financial crisis, suggesting potential overfitting to a period when markets were more predictable and less exposed to global risks than in recent years.
- The public dissemination of Avellaneda & Lee's methodology and optimized parameters led to widespread adoption. Consequently, these arbitrage opportunities were quickly identified and exploited by market participants, diminishing their effectiveness.
- The persistent bull market conditions during the backtesting period (with Nifty 100 gaining 350%) significantly limited the strategy's performance, as it diminished the shorting opportunities.
- The emergence of advanced algorithmic trading techniques and more sophisticated implementations of statistical arbitrage strategies has significantly reduced the effectiveness of traditional approaches.

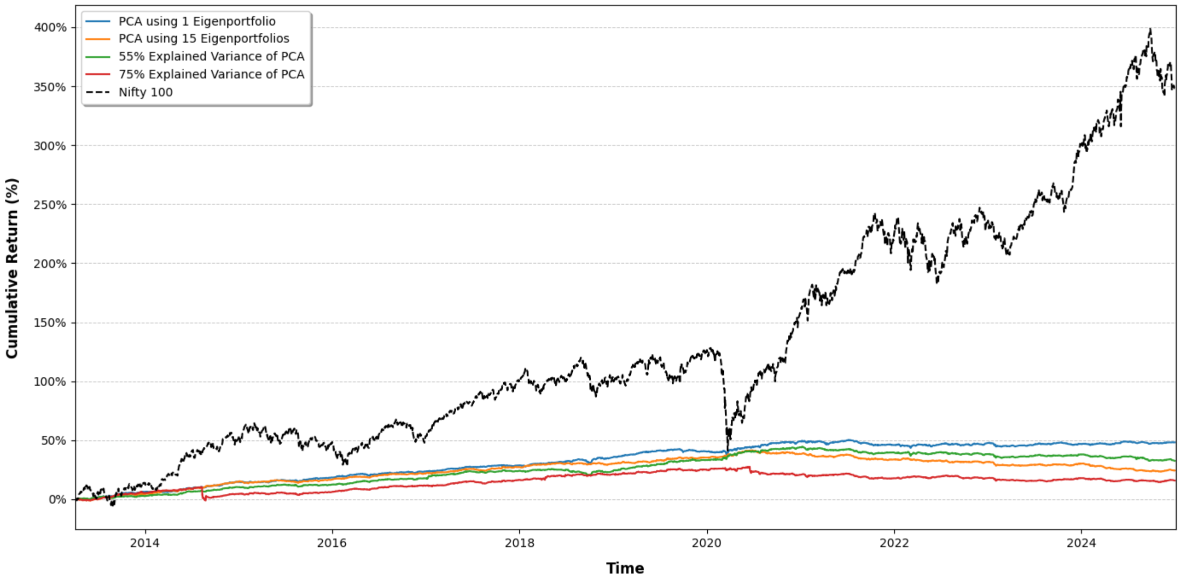


Fig. 13: Historical Cumulative Returns of strategies for all the different variants of risk factor determination using standard daily returns, compared with benchmark Nifty 100: 2012-2024

To better examine the relative performance of different strategy variants, Figure 14 presents the cumulative returns excluding the benchmark. Table 2 presents the Sharpe ratios for all the backtest variants using standard daily return.

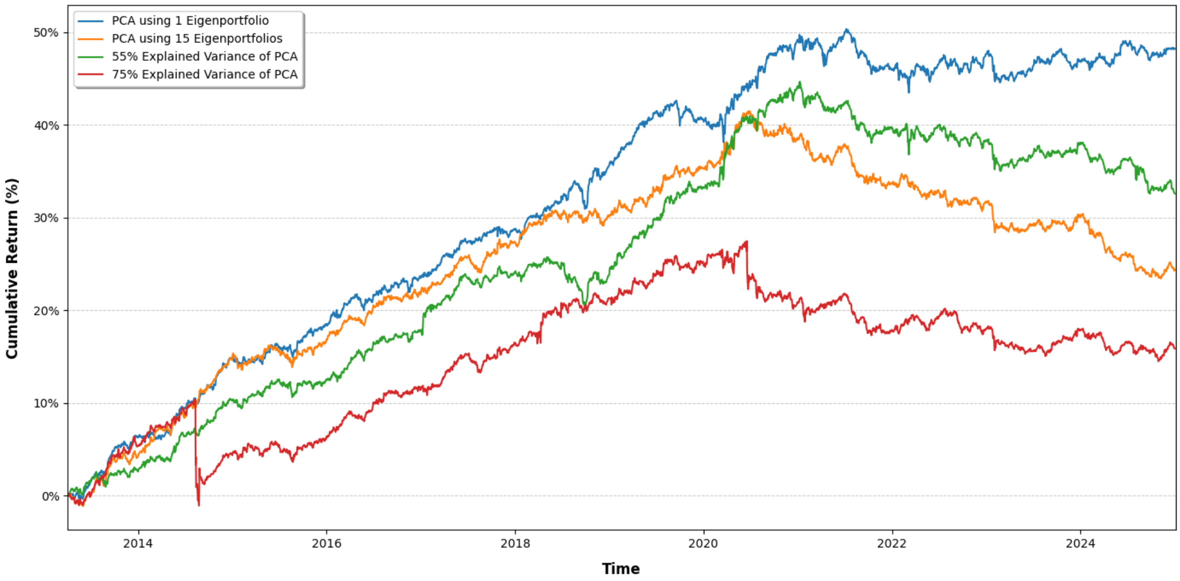
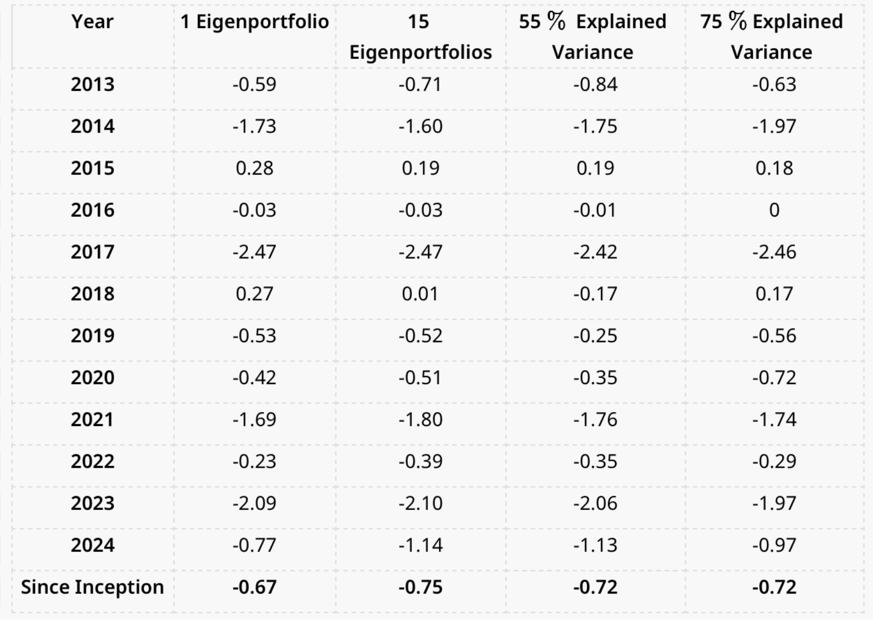


Fig. 14: Historical Cumulative Returns of strategies for all the different variants of risk factor determination using standard daily returns: 2012-2024

The comparative analysis reveals distinct patterns in strategy performance:

- 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.
- 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, combined with additional noise trading from these higher-order components leading to increased losses.



| 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** |

Tab. 2: Sharpe ratios of strategies for all the different variants of risk factor determination using standard daily returns

## Using Volume-Weighted Returns

To assess whether incorporating trading volume information improves strategy performance, Figure 15 compares the cumulative returns of volume-weighted implementations directly against their standard daily return counterparts.

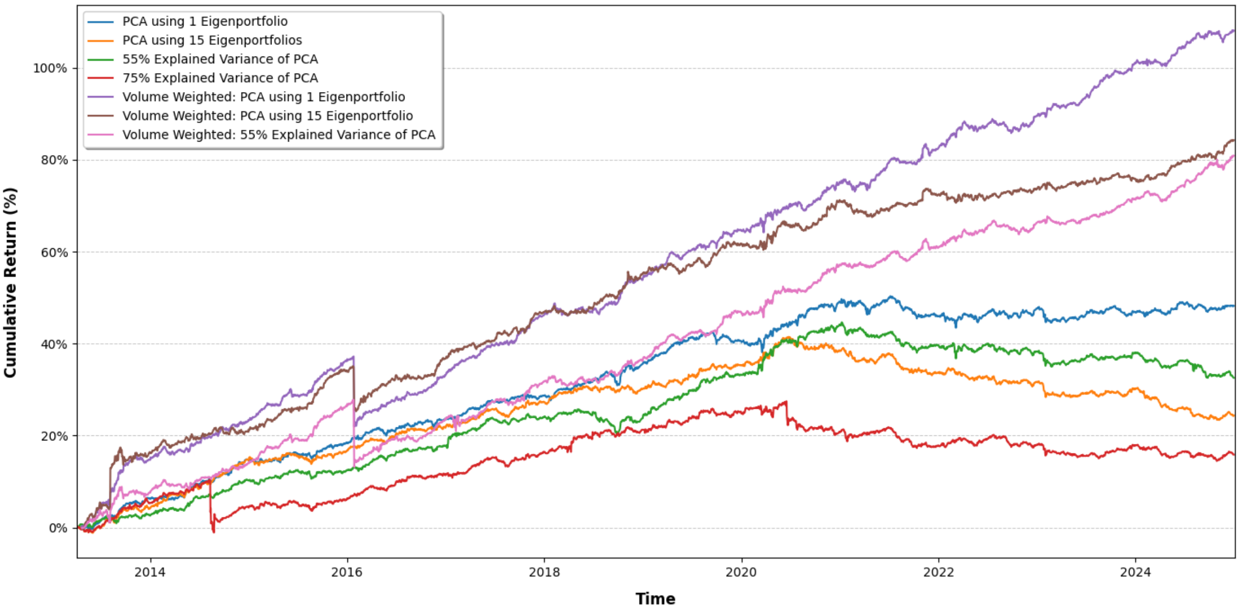
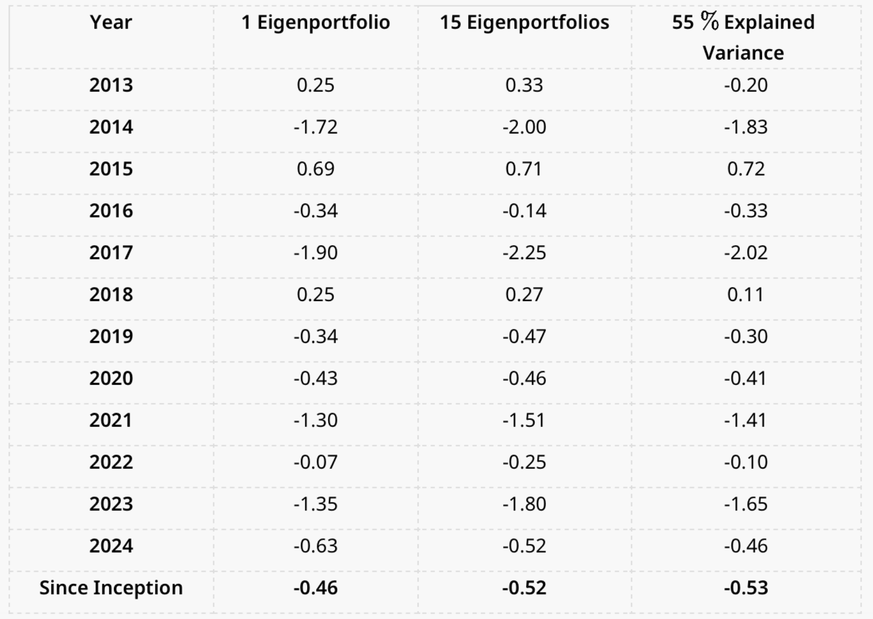


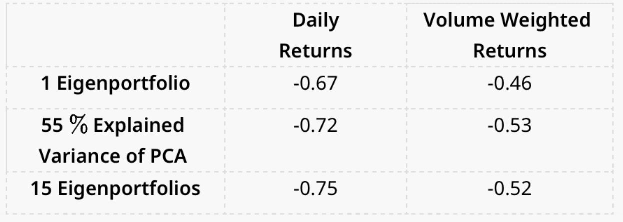
Fig. 15: Comparison of Historical Cumulative Returns of strategies for all the different variants of risk factor determination using both volume weighted returns returns and standard daily returns: 2012-2024

The volume-weighted strategies demonstrated 50-60% excess returns compared to their standard daily return counterparts, with the volume-weighted single eigenportfolio strategy achieving approximately 100% cumulative returns. The incorporation of trading volume information yielded significant performance improvements suggesting that volume contains important information. Table 3 presents the Sharpe ratios for the backtest variants 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** |

Tab. 3: Sharpe ratios of strategies for all the different the variants of risk factor determination using volume weighted returns



| | 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 |

Tab. 4: Comparison of Since Inception Sharpe ratios of all the strategies

Among all variants tested, the volume-weighted single eigenportfolio strategy achieved the highest Sharpe ratio (Table 4), though still significantly underperforming the Nifty 100 benchmark (Figure 15).

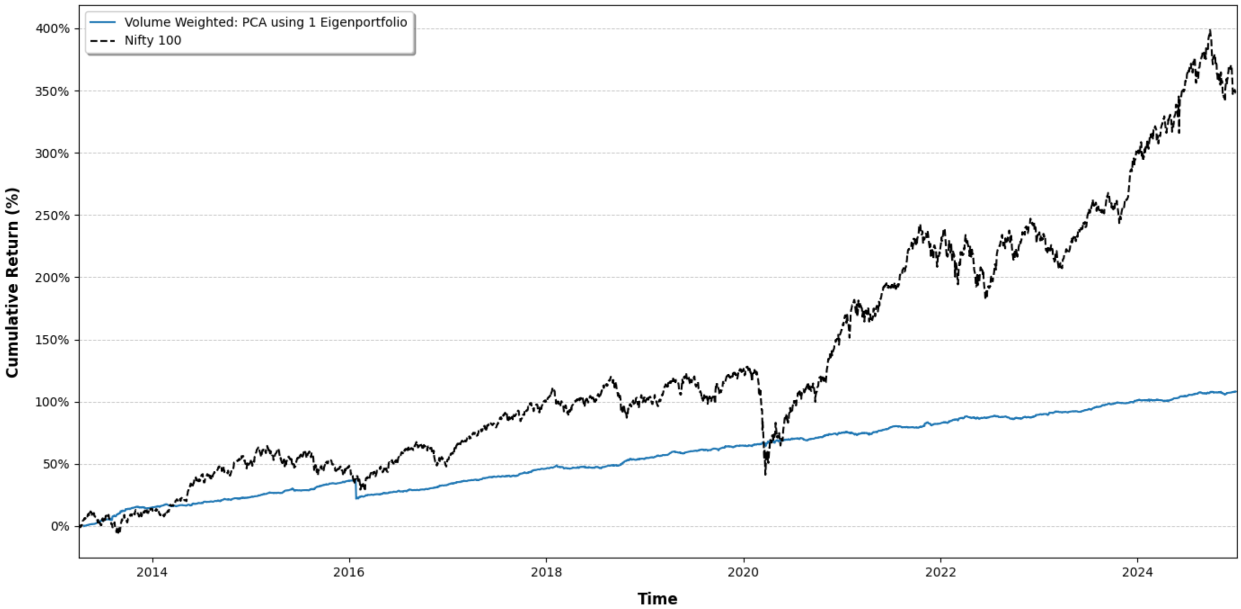


Fig. 16: Comparison of Historical Cumulative Returns of Volume Weighted Single PCA and Benchmark Nifty 100: 2012-2024

## Impact of Leverage

Given the strategy's underperformance relative to the benchmark, we explored whether increasing portfolio leverage could potentially match market performance. While $\Lambda$ is not a hyperparameter and can be adjusted, this modification introduces important practical constraints. As discussed before $\Lambda$ directly impacts position sizing - for example, with 2x leverage, each position would receive twice the capital allocation compared to the base strategy, while 5x leverage would quintuple the allocation per position.

The relationship between $\Lambda$ and position capacity follows an inverse pattern: while increasing $\Lambda$ scales both PnL and Sharpe ratios proportionally, it limits the number of positions that can be opened and hence also increasing opportunity costs as fewer simultaneous positions can be maintained. With fixed capital constraints, the maximum number of open positions must be carefully managed to avoid capital shortfalls.

I tested the volume-weighted single eigenportfolio strategy (our best-performing variant) with different leverage levels (2x and 3x) to evaluate the impact on performance. Figure 17 presents the cumulative returns under different leverage scenarios. Table 5 shows the corresponding Sharpe ratios for different leverage levels.

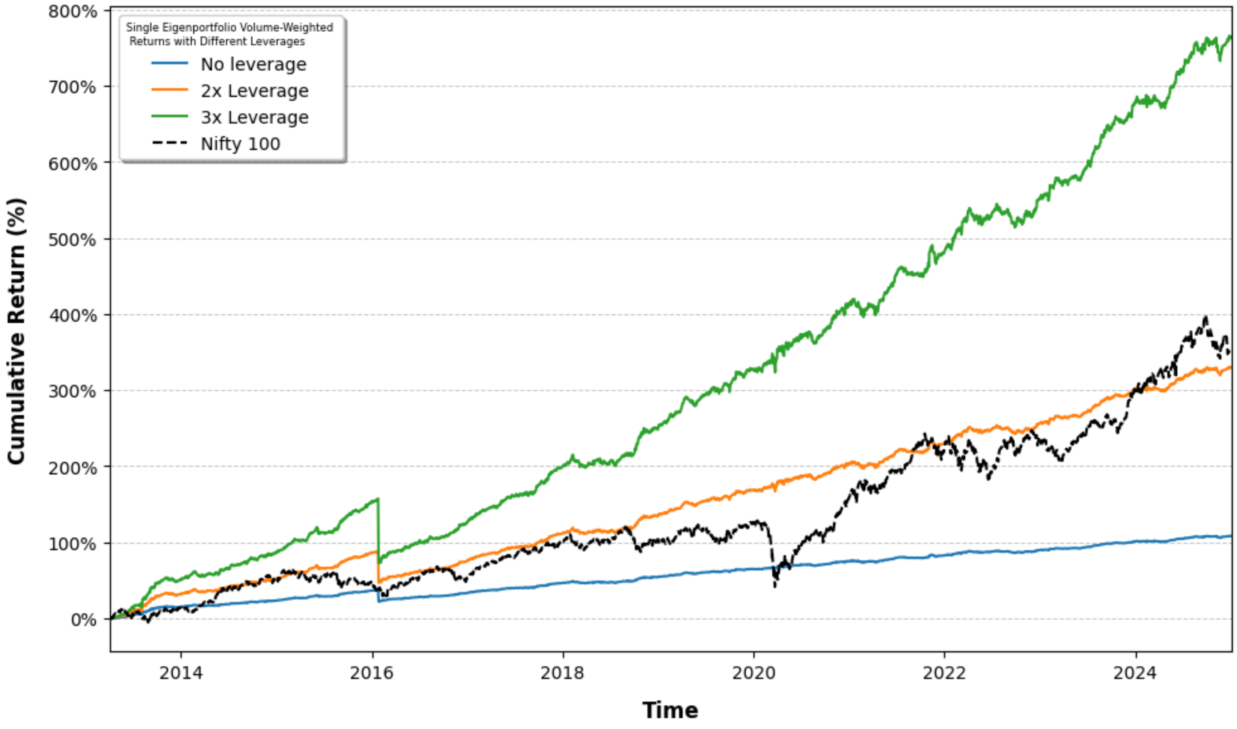
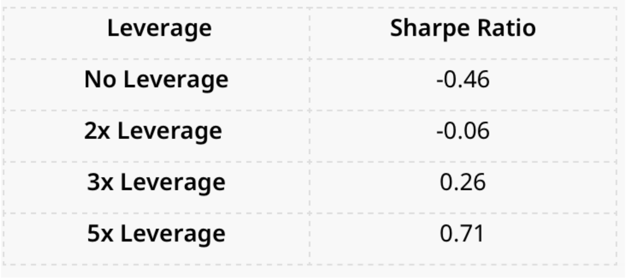


Fig. 17: Comparison of Cumulative returns of Volume Weighted Single PCA strategy with Different Leverages



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

Tab. 5: Sharpe ratios of Volume Weighted Single PCA strategy with Different Leverages

Although increasing leverage improved returns, it necessitated stricter position sizing limits to maintain risk management within capital availability.

## 6 Conclusion

This study provides empirical evidence on the evolution and current state of statistical arbitrage opportunities in the Indian equity market, while highlighting several methodological considerations and potential areas for improvement. Our implementation of the PCA-based strategy reveals several important findings.

First, while the strategy generated profits ranging from 20-50% for standard implementations, it significantly underperformed the benchmark's 350% return. This underperformance can be attributed to several structural changes in market dynamics: the post-2008 shift toward more globally integrated markets, the rapid erosion of arbitrage opportunities following public dissemination of Avellaneda & Lee's methodology, the emergence of sophisticated algorithmic trading techniques leading to faster price discovery, and the persistent bull market conditions during 2012-2024 which limited shorting opportunities. Notably, the incorporation of trading volume information yielded significant performance improvements, with volume weighted strategies demonstrating enhanced returns, suggesting that volume contains valuable information.

The analysis of eigenportfolio decomposition highlights a fundamental limitation of PCA methodology: there are considerably more entries in the correlation matrix than data points, potentially affecting the statistical reliability of our results. While implementing asymptotic PCA as suggested by Tsay (2010) could address this dimensionality challenge, a more fundamental methodological shift towards cointegration framework might be beneficial. Unlike PCA, which approximates the systematic risk of the entire stock universe using a limited set of eigenportfolios, cointegration analysis could identify all possible stationary linear combinations of the price series (Alexander, 2001). Given that our strategy fundamentally relies on identifying stationary residuals, such a cointegration approach might provide a more theoretically appropriate framework for capturing mean-reversion opportunities [2].

The framework's application of linear regression models at crucial junctures assumes Gaussian residuals and stationary spread processes. However, stock returns exhibit well-documented characteristics such as volatility clustering, heteroskedasticity, and fat tails, suggesting potential benefits from more sophisticated modeling approaches. Yet, implementing such sophisticated models present fundamental challenges, as they would fundamentally change how the spread process should be modeled, conflicting with our current mean-reversion framework. The assumption that OU process parameters vary slowly relative to Brownian motion increments, for simplifying estimation procedures, may be overly restrictive for financial markets where true parameter constancy is rare. Although, while we verify stationarity using the Augmented Dickey-Fuller test and exclude non-stationary stocks from consideration to maintain model validity, this approach 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 [4].

Our analysis also revealed that simpler implementations, particularly the single eigenportfolio strategy, outperformed multi-factor approaches. This aligns with Avellaneda & Lee's observation that mean-reversion strategies perform better when a small number of factors can explain a significant portion (approximately 50%) of return variance. When the true number of explanatory factors is large, using fewer factors leaves market information in the residuals, while using too many factors reduces profit opportunities as the marginal benefits of additional factors may be offset by increased complexity and trading costs, we observed that the optimal number of factors varies inversely with market volatility, requiring fewer factors during crisis periods and more during low-volatility regimes.

While increasing leverage could theoretically improve returns, it introduces significant practical challenges. Position sizing must be constrained by available capital to avoid external capital dependencies, which would introduce unwanted market exposure despite the strategy's statistical arbitrage focus. Although limiting positions according to capital availability reduces potential profits, it provides better risk management by avoiding losses from market movements.

## References

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

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

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

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

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

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