Use PCA on Real Stock Returns

Week 1 Task 1

Author: Rohan Biswas
July 13, 2025

Contents

List of Algorithms

List of Tables

Chapter 1

Principal Component Analysis - PCA

1.1 Introduction

What exactly is PCA ? PCA is a dimensionality reduction technique, that seeks to find the directions in which data varies most.

For most returns, each stock can be considered a feature, and each day's return is a data point in high dimensional space.

1.2 Technicalties

Let,
$X \in \mathbb{R}^{T \times N}$: centered data matrix
T : Time points (days) N : Assets (stocks)

Algorithm 1 PCA

1. Center the data :

$$\text{Let, } \mu = \frac{1}{T}\sum_{t=1}^{T} X_t \in \mathbb{R}^{T\times N}$$

$$\text{Define, } X_{centered} = X - \mu \text{ [columnwise]}$$

2. Covariance Matrix :

$$\sum = \frac{1}{T-1} X_{centered}^T X_{centered} \in \mathbb{R}^{N\times N}$$

   Each element $\sum_{ij}$ is the sample covariance between stocks i and j .

3. Eigen-decomposition of covariance matrix :

$$\sum = V\Lambda V^T$$

   where,
   $V \in \mathbb{R}^{N\times N}$: orthonormal eigenvectors.
   $\Lambda = \text{diag}(\lambda_1, \lambda_2, \cdots \lambda_N)$: eigenvalues, such that : $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$

4. Principal components :
   Project data onto top-k eigenvectors:

$$Z = X_{centered} V_K$$

   where, $V_k \in \mathbb{R}^{N\times N}$, the top K eigenvectors.

Important points to note :

- Variance by i^{th} PC : $\lambda_i / \sum_j \lambda_j$.
- First PC along max variance.

Chapter 2

Stock Market Indices

2.1 Introduction

It's a statistical measure that reflects the performance of a group of stocks, serving as a benchmark for a country's economic sector or market segment. Formally, a set of stocks with prices P_i and associated weights w_i, give index value I_t at time t as :

$$I_t = \sum_{i=1}^{m} w_i \cdot P_{i,t}$$

Different indices use different weighing schemes :

1. Market Capitalization (Market-Cap) weighting.
2. Price weighted.
3. Equal weighted.
4. Fundamental weighting (Revenue weighted).
5. Volatile weighted (Min-Variance Portfolio).
6. Risk Parity (Equal Risk Contribution).
7. Factor weighted.
8. Thematic/ESG weighted.

2.2 Market-Capitalization

$$\text{Market cap} = \text{Share Price} \times \text{Total Shares Outstanding}$$

It reflects total values the market assigns to a company. It changes continuously as stock prices fluctuate.
Most indices (including S& P 500 and NIFTY 50) use free-float market cap as it includes shares that are publicly available for trading, excluding : promoter shares, govt. held shares, strategic investor holdings, ESDPs locked in.

$$\text{Free Float Market Cap}, M_i = \text{Market Cap} \times \text{Free Float Factor}$$

It's a decimal value (e.g., 0.75 if 75% shares are tradeable).
If N companies, and each company i has free-float market M_i, weight in index is :

$$w_i = \frac{M_i}{\sum_{j=1}^{N} M_j}$$

So, index at time t is :

$$\text{Index}_t = \sum_{i=1}^{N} w_i \cdot P_{i,t}$$

where $P_{i,t}$ is stock's normalized price or return.
Example :

Table 2.1: Company Market Capitalization

Company	**Price in INR**	**Shares**	**Free Float %**	**Free float Market-cap in INR**
A	100	1B	80	80B
B	200	0.5B	100	100B
C	50	2B	50	50B

Weights :

$$w_A = \frac{80}{80 + 100 + 50} = 34.78$$

$$w_B = \frac{10}{80 + 100 + 50} = 43.47$$

$$w_C = \frac{50}{80 + 100 + 50} = 21.74$$

Now, even if share count for C is high, B dominates as B's full float make its market cap highest.
Pros :

- Self rebalancing - If stock's value rises, weight naturally increases.
- Reflects actual market structure.
- Passive investing easy (e.g., tracking SPY or NIFTYBEES).

Cons:

- Biased towards large-cap companies.
- Overweights momentum trends.
- Underweights value or small-cap stocks.

Market-cap indices are replication targets for ETFs and mutual funds.

2.3 Price Weighted

It gives each stock a weight proportional to its price per share, regardless of market cap or size. Higher priced stocks have greater influence, even if the company is smaller.
Let there be N stocks with prices $P_1, P_2, \cdots, P_N$. So,

$$\text{Index} = \frac{1}{N} \sum_{i=1}^{N} w_i \cdot P_{i,t}$$

Or if a divisor D is used (like in DJIA) to adjust stock splits, mergers, etc) :

$$\text{Index} = \frac{1}{D} \sum_{i=1}^{N} w_i \cdot P_{i,t}$$

$D = N$ at start (unadjusted), but gets modified over time for corporate actions.

$$\text{weight}, w_i = \frac{P_i}{\sum_{j=1}^{N} P_j}; \text{so proportional to price.}$$

Example : Take above example .

$$\sum_i P_i = 100 + 200 + 50 = 350$$

$$w_A = \frac{100}{350} = 28.57\% \quad ; \quad w_B = \frac{200}{350} = 57.14\% \quad ; \quad w_C = \frac{50}{350} = 14.29\%$$

So, even if C has highest share count, B has higher price so it dominates.
Pros :

- Simple to compute.
- Historically used (e.g., Dow Jones).

Cons :

- Distorted by stock splits. If stock A splits 2-for-1 then price drops to INR 50 and weight halves (as in example above). If stock B splits:
  If B split: New price : [100,100,50] $\rightarrow$ sum INR 250
  New Index : 250/3 = 83.33
  New Divisor = 350 (Old sum of price)/116.67 (Old Index) $\approx$ 3
- Price $\neq$ Size.
- Manipulation share price structure by comparing.

2.4 Equal Weighted

Gives each stock on identical weight regardless of Price, Market cap, Revenue, Fundamentals, etc.
Every company has equal influence on index return.
Let there be N companies in index, which gets us

$$w_i = \frac{1}{N}$$

Index return at time t :

$$R_t = w_i \sum_{i=1}^{N} R_{i,t}$$

where $R_{i,t}$ in return of stock i at time t. Example :

Table 2.2: Company Equal Weighted

Company	**Price in INR**	**Shares**	**Free Float %**	**Market cap**
A	100	1B	80	100
B	200	0.5B	100	100
C	50	2B	50	100

Equal weighted index,

$$w_A = w_B = w_C = \frac{1}{3} = 33.33\%$$

Assume today's price,

$$A : \text{INR } 100 \to \text{INR } 110 \ (10\% \text{ return})$$

$$B : \text{INR } 200 \to \text{INR } 180 \ (-10\% \text{ return})$$

$$C : \text{INR } 50 \to \text{INR } 52 \ (4\% \text{ return})$$

$$R = \frac{1}{3}(10\% + (-10\%) + 4\%) = \frac{4\%}{3} = 1.33\%$$

So each stock contributes equally.
Pros :

- Equal exposure to all stocks.
- Easy to understand and implement.

Cons :

- Requires frequent rebalancing.
- High turnover = High Transaction costs.
- May over exposure to volatile / small-cap stocks.
- Not self rebalancing.

It often outperforms market-cap indices in bull markets, used to benchmark active fund managers, core to factor investing and smart beta ETFs.

2.5 Fundamental weighting (Revenue- Weighted)

It assigns weights to stocks based on company fundamentals like :
Revenue, Earnings, Book Value, Dividends, Cash Flows. It reflects economic importance, not market opinion (price).
Why Revenue tho ?
Universal reported, harder to manipulate, and size indicative.
But here's a thing : composite (average of 4 metrics) exists and is used by FTSE RAFI.
Let,

$$R_i : \text{Revenue of stock } i \quad , \quad R_{total} = \sum_{j=1}^{N} R_j$$

Then weight is :

$$w_i = \frac{R_i}{R_{total}}$$

Fixed until revenue updated, regardless of stock price movement. Example :

Table 2.3: Company Fundamental weighting

Company	**Revenue in INR**	**Return**
A	100	+5%
B	200	+3%
C	50	+10%

Here,

$$R_{total} = 300 \quad ; \quad w_A = \frac{100}{300} = 33.33\%$$

$$w_B = \frac{150}{300} = 50\%$$

$$w_C = \frac{50}{350} = 16.67\%$$

So, Index return :

$$R = (0.3333 \times 5\%) + (0.5 \times 3\%) + (0.1667 \times 10\%) = 4.83\%$$

Pros :

- Real business metrics, not hype.
- Avoids over-weighting overvalued (bubble) stocks.
- Value tilts : Tends to outperform growth in recovery cycles.

Cons :

- Needs regular fundamental updates.
- Ignores market sentiment and liquidity.
- May lag during momentum-driven rallies.
- No automatic rebalancing.

2.6 Volatility Weighted (Min-Variance Portfolio)

A volatility-weighted index gives more weight to less volatile stocks, under the logic that lower risk = more reliable return.
Goal is to reduce portfolio variance by down-weighting noises, risky stocks.
Let σ_i be the volatility (standard deviation) of returns for stock i. Then :

$$w_i = \frac{1/\sigma_i^2}{\sum_{j=1}^{N} 1/\sigma_j^2}$$

This is inverse-variance weighting.
Variance - it corresponds directly to risk in mean-variance optimization theory.
Example :

Table 2.4: Company Volatility weighted (Min-Variance Portfolio)

Company	**Volatility (σ)**	**$1/\sigma^2$**	**Return**
A	0.10	100	+2%
B	0.20	25	+4%
C	0.15	44.44	-1%

Total inverse variance is then :

$$\sum_i \frac{1}{\sigma_i^2} = 100 + 25 + 44.44 = 169.44$$

Compute weights :

$$w_A = \frac{1/\sigma_A^2}{\sum_{j=1}^N 1/\sigma_j^2} = \frac{100}{169.44} \approx 59.02\%$$

$$w_B = \frac{1/\sigma_B^2}{\sum_{j=1}^N 1/\sigma_j^2} = \frac{25}{169.44} \approx 14.75\%$$

$$w_C = \frac{1/\sigma_C^2}{\sum_{j=1}^N 1/\sigma_j^2} = \frac{44.44}{169.44} \approx 26.33\%$$

Index return for the day :

$$R = (0.5902 \times 2\%) + (0.1475 \times 4\%) + (0.2623 \times (-1\%)) = 1.5081\%$$

So, most volatile stock B had highest return but contributed less than the low volatile A.
Pros :

- Reduces portfolio volatility.
- Mitigates impact of risky or speculative stocks.
- Often provides superior risk-adjusted return (Sharpe ratio).
- Simpler approximation of minimum variance portfolio.

Cons :

- May over-concentrate in defensives (utilities, stables).
- Could underperform in high- momentum bull markets.
- No accounting for return expectations-only risk.
- Requires accurate volatility estimation.

It's based on Markowitz Mean-Variance Optimization. If returns are assumed equal, inverse variance gives minimum variance solution:

$$\min_w w^T \Sigma w \ , \text{ such that } \sum_i w_i = 1$$

If Σ is diagonal, inverse-variance solution is exact.

2.7 Risk Parity (Equal Risk Contribution)

Assigns weights such that each asset contributes the same amount of risk (volatility) to the overall portfolio.
Let,
w_i : weight of asset i , σ_i : volatility of asset i,
P_{ij} : correlation between assets , Σ : covariance matrix of returns.
Portfolio volatility :

$$\sigma_p = \sqrt{w^T \ \Sigma \ w}$$

Marginal risk contribution (MRC) of asset i:

$$\text{MRC}_i = \frac{\partial \sigma_p}{\partial w_i} = \frac{(\Sigma w)_i}{\sigma_p}$$

Total risk contribution (TRC) of asset i :

$$\text{TRC}_i = w_i \cdot \text{MRC}_i = \frac{w_i (\Sigma w)_i}{\sigma_p}$$

Risk parity condition :

$$\text{TRC}_i = \text{TRC}_j \ \forall i, j$$

Case : Uncorrelated assets, $P_{ij} = 0$ so Σ is diagonal, Then,

$$\sigma_P^2 = \Sigma w_i^2 \sigma_i^2$$

$$\text{and TRC}_i = \frac{w_i^2 \sigma_i^2}{\sigma_P}$$

Considering Risk parity condition :

$$w_i^2 \sigma_i^2 = w_j^2 \sigma_j^2$$

$$\frac{w_i}{w_j} = \frac{\sigma_j}{\sigma_i} => w_i \propto \frac{1}{\sigma_i}$$

Example :

Table 2.5: Company Risk Parity (Equal Risk Contribution)

Company	**Volatility (σ)**	**Return**
A	0.10	+2%
B	0.20	+4%
C	0.15	-1%

Here,

$$\frac{w_A}{w_B} = \frac{0.2}{0.1} = 2 => w_A = 2 \cdot w_B \quad , \quad \frac{w_A}{w_C} = \frac{0.15}{0.1} = 1.5 => w_A = 1.5 \cdot w_B$$

If $w_B = x$ then, $w_A = 2x$ and $w_C = \frac{2}{1.5}x = 1.333x$.
Now,

$$\sum_i w_i = 1$$

So, $x = 0.2308$.Then,

$$w_A = 46.15\% \quad , \quad w_B = 23.08\% \quad , \quad w_C = 30.77\%$$

Index return,

$$R = (0.4615 \times 2\%) + (0.2308 \times 4\%) + (0.3077 \times (-1\%)) = 1.538\%$$

Pros :

- More Balanced Risk Distrubution.
- No single asset dependency.

Cons :

- Needs accurate covariance estimation.
- Can overweight low risk assets excessively.

2.8 Factor Weighted

Involves tilting portfolio weights toward certain "factors" that explain stock returns.
Factors : Value, Momentum, Quality, Size, Low Volatility, ESG/Sentiment/Growth.
These are systematic risk premia - persistent, explainable sources of return.
Let each stock i have a factor score, like momentum or value score. We construct index as :

$$w_i = \frac{F_i}{\sum_j F_j}$$

Just like revenue weighting but F is now factor score.
Can also apply non linear transforms like :

$$w_i \propto \exp(F_i) \quad \text{or} \quad \max(F_i, 0)$$

Or even rank based weighting :

$$w_i \propto \text{rank}(F_i)$$

Let, $F = [F_1, F_2, \cdot, F_N]$: factor scores.
Returns :

$$R_t = \sum_i w_i R_{i,t}$$

Emphasizes exposure to stocks scoring high on that factor. Example :

Table 2.6: Company Factor Weight

Company	**Momentum score**	**Return**
A	0.8	+3%
B	0.5	+5%
C	0.7	+1%

Total score, $\sum_i F_i = 0.8 + 0.5 + 0.7 = 2.0$.

$$w_A = \frac{0.8}{2.0} = 0.4 = 40\% \ , \ w_B = \frac{0.5}{2.0} = 0.25 = 25\% \ , \ w_C = \frac{0.7}{2.0} = 0.35 = 35\%$$

$$R = (0.4 \times 3\%) + (0.25 \times 5\%) + (0.35 \times 1\%) = 2.8\%$$

Pros :

- Emperically proven performance drivers.
- Highly customizable.
- Used in Smart Beta ETFs.

Cons :

- Can underperform in short bursts.
- Sensitive to factor score definition.

- Requires clean, updated data.
- Needs rebalancing and signal maintenance.

Multi factor weighting :

$$F_i = \alpha \cdot \text{Momentum}_i + \beta \cdot \text{Value}_i + \gamma \cdot \text{Quality}_i$$

2.9 Thematic / ESG Weighted

Thematic : Companies assigned with themes like renewable energy, AI and robotics, EVs, Gender equality.
ESG : Dealing with Environmental, Social, and Governance stuffs like Emission, Board diversity, etc.
It is a form of multi factor weighting with Themes or ESGs as factors.
It attracts socially conscious investors, but it may lead to underexposure to profitable but dirty sectors like oil and mining.
How ESGs/ thematics are used ?

Table 2.7: Index vs Theme

Index/ETF	**Theme**
Nifty 100 ESG	ESG in India
MSCI ACWI ESG Universal	Global ESG tilt
S&F Global Clean Energy Index	Renewable Energy
iShares Global Water ETF	Water Sustainability

Chapter 3

Conclusion

Why PCA ?
It's cause each Principal component (PC) :

- Captures latent risk driver in the market.
- Is a portfolio (a vector of weights).
- The 1^{st} PC often resembles market factor.

Table 3.1: PC vs Index Comparison

Index type	1^{st} **PC or PC1**	**% Variance in PC1**	**Higher PCs meaning**
Market Cap	Large-cap market factor	High (40-60)	Sector and Idiosyncratic risks
Equal Weight	Broad market	Medium (25-40)	Size, Value, Volatility
Price Weight	Price driven pseudo factor	Skewed	Unreliable
Fundamental	Real business economy	Medium	Factor tilts, sectoral
Risk Parity	Smoothed and subtle structure	Lower (15-25)	Diversified hidden risks
Factor Weight	That Factor (e.g., Momentum)	Medium to High	Other uncorrelated factors
ESG/Thematic	ESG Sensitive market	Variable	Sectoral or Thematic shifts

Now, y'all might be questioning why I used log returns directly with values obtained from dataset (in project) that is the Price variable. Well that's cause its not the internal stuff.
If we were to create our own index then we might need raw data for weight and return calculation, but we ain't dealing with that. We have the already polished information that provider published in form of single time-series P_t that reflects the total effect of all constituents, weights, dividends, etc.

Links

To access the code repository, visit:
https://github.com/RohanBiswas67/Quant/tree/main/week1task1_pca

To share any feedback, you may reach out to me at: rohanbiswas031@gmail.com

OR

Visit portfolio at : https://rohan-biswas-portfolio.vercel.app