PCA on different asset classes

Are there specific topics driving asset prices?

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

Sophie Grill, Sebastian Herzog, Alexei Volodin,
Arina Suhodolova, Dinara Zainullina

19.12.2023

High Level Roadmap

The research problem can be divided into the following **three steps:**
(1) Data Preparation & Brainstorming
(2) Statistical analysis & macro-economic interpretation
(3) Visualization & Next Steps

High Level Roadmap (Part 1)

I. **Data Preparation & Brainstorming**

1. **Data Preparation**

a. Define theoretically suitable Data

Economic and Capital Markets related time series across all asset classes, especially Factor-based Time-Series

b. Identify Data Sources

Bloomberg, Yahoo Finance, Kenneth R. French Library on Stock Return Factors, etc.

c. Data Preparation & Quality Assurance

2. **Statistical Methods Brainstorming**

Pro & Contra of e.g. Hidden Markov-Chains, Principal Component Analysis, Bayesian Nets, Neural Nets, etc.

Devise Long-List and most promising short list of suitable methods

High Level Roadmap (Part 2)

II. **Statistical Analysis & Macro-economic Interpretation**

a. Application of short-listed models to data, identify issues & solutions and come up with macro-economic interpretation of results

b. Time-Series Regression of Principal Components onto macro-economic/ Factor-Portfolios

III. **Visualization of Theme Evolution through time**

In progress

Outcome

- Fully integrated R Code (Data Input, Data Quality Checks, Statistical Analysis, Output)
- Sensitivity Assessment: Which asset classes are more heavily influenced by the identified topics, which are defensive safe havens?
- Interactive visualization dashboard / web application (e.g. R Shiny, Power BI) of "Driving Topics" through time (incl. conditional correlations)

1. Data Preparation

a) Define Theoretically Suitable Data

Chosen time series across multiple asset classes: commodity prices, bond indices, spread indices, equity indices, FX rates, as well as macro data such as CPI rates, unemployment rates, real GDP (%)

Time horizon: last 20 years

Frequency: daily

b) Identify Data Sources

Bloomberg and Kenneth R. French Library

c) Data Preparation and Quality Assurance

2. Statistical Methods Brainstorming

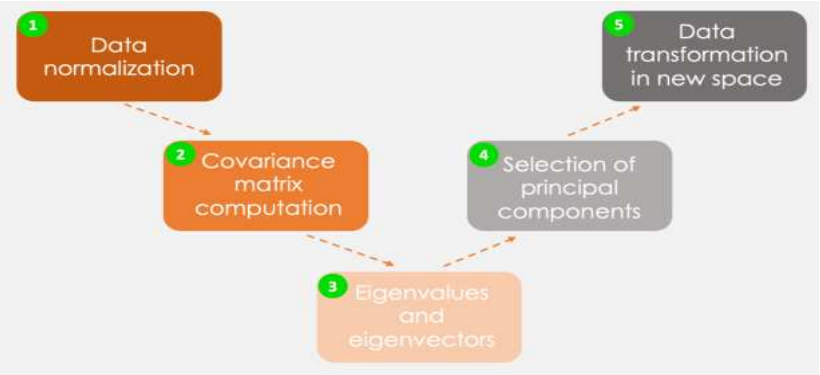
- Principal Component Analysis as a chosen statistical method for analysing high-dimensional data and capturing the most important information from it (principal components/ potential ”drivers” of asset prices)

- This is done by transforming the original data into a lower-dimensional space while collating highly correlated variables together

- Main advantages: Dimensionality Reduction, Multicollinearity Mitigation, Pattern Recognition

- Possible obstacles to be addressed: Interpretability, Sensitivity to Outliers

PCA in 5 Steps – we use a package – delete the slide?



- **Step 1 - Data normalization**
  - Created log returns and normalized them
  - Attributes them on same level, no bias
- **Step 2 - Covariance matrix**
  - symmetric matrix, each element (i, j)
  - corr. to the covariance between variables i/j.
- **Step 3 - Eigenvectors and eigenvalues**
  - **Eigenvector** represents direction. An **eigenvalue** is a number representing the amount of variance present in the data for a given direction. Each eigenvector has its corresponding eigenvalue.
- **Step 4 - Selection of principal components**
  - Data variables determine the pairs of eigenvectors and eigenvalues. In our data are 76 columns (excluding macro data), hence 76*5721 pairs. Not all the pairs are relevant. So, the eigenvector with the highest eigenvalue corresponds to the first principal component.
- **Step 5 - Data transformation in new dimensional space**
  - re-orienting the original data onto a new subspace defined by the principal components This reorientation is done by multiplying the original data by the previously computed eigenvectors.

Descriptive Data – Correlation Matrix?

- If we manage to make it look readable
- Or not
- Then delete the slide

Applying PCA – update with the new result

```
> summary(data.pca)
Importance of components:
                          Comp.1     Comp.2      Comp.3      Comp.4      Comp.5
Standard deviation     1.3718675 0.6067002 0.084690125 0.074143005 0.054761363
Proportion of Variance 0.8304466 0.1624186 0.003164849 0.002425648 0.001323232
Cumulative Proportion  0.8304466 0.9928651 0.996029989 0.998455637 0.999778868
                           Comp.6       Comp.7       Comp.8       Comp.9
Standard deviation     0.0179960562 1.310913e-02 2.331889e-03 1.220501e-08
Proportion of Variance 0.0001429032 7.582895e-05 2.399403e-06 6.573004e-17
Cumulative Proportion  0.9999217716 9.999976e-01 1.000000e+00 1.000000e+00
```

- **Nine** principal components have been generated (Comp.1 to Comp.9)
- In the **Cumulative Proportion** section, the first principal component explains almost 83% of the total variance. This implies that almost two-thirds of the data in the set of 9 variables can be represented by just the first principal component. The second one explains 16%

Loading Matrix for PC 1 & 2 – update or delete

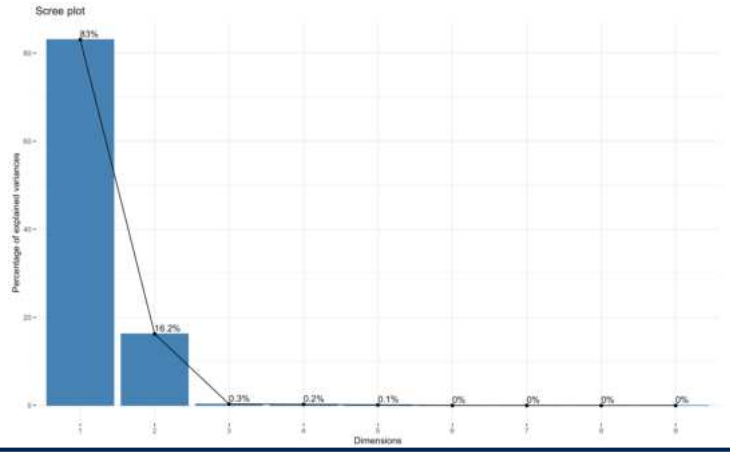
It's great to have the first two components, but what do they really mean?

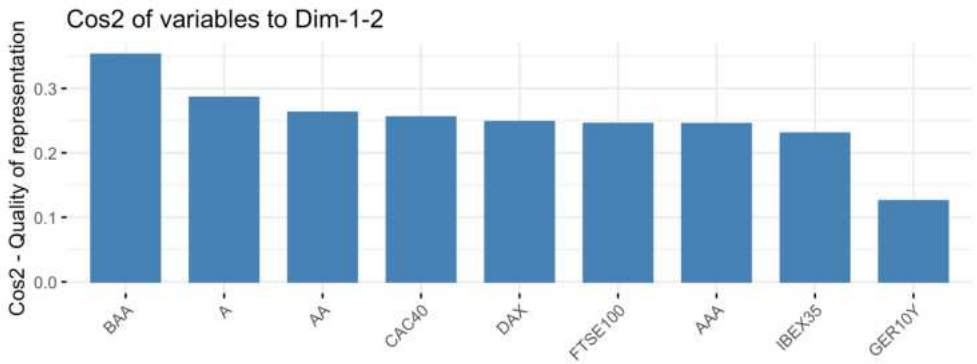
	Comp.1	Comp.2
A	0.36979515	0.2779464
AA	0.34977309	0.2964470
AAA	0.33501760	0.3027773
BAA	0.42526358	0.1820070
CAC40	-0.34155025	0.3115903
DAX	-0.34296810	0.2692544
FTSE100	-0.32849570	0.3383866
GER10Y	-0.04060333	-0.5763094
IBEX35	-0.32246346	0.3063331

- The loading matrix shows that the first principal component has high positive values for Fixed Income prices.

Visualization of the principal components

- **Scree Plot**
  - This plot shows the eigenvalues in a downward curve, from highest to lowest. The first two components can be considered to be the most significant since they contain almost 99% of the total information of the data.





- **Contribution of each variable**
  - Determines how much each variable is represented in a given component. Such a quality of representation is called the Cos2 and corresponds to the square cosine
  - A high value, on the other hand, means a good representation of the variable on that component.

Further Steps…

a) Macro-economic interpretation of the results
b) Time-series regression of PCs onto macroeconomic data/factor portfolios