Choose Your Own Project: Analysis of Banking Indicators

Towards Detecting Warning Signs in the Wider Banking System

Panagiotis Chatzakis

21 November, 2020

Introduction

This is the second project of the HarvardX Data Science Certificate. In this project, the Principal Component Analysis (PCA) method is applied on economy-wide macrofinancial indicators to identify potential sector-wide linkages between these variables.

The purpose of this project is to analyze data from a banking system's portfolio or costs in order to identify latent factors that will be used to forecast deterioration, using measurements such as: nonperforming loans (NPL) ratio, cost of funds (relative to the interbank offered rate or the overnight policy rate), and the credit default swap (CDS) spreads. The goal is to identify a cluster of indicators with potential predictive power to forecast deterioration in a banking system, developed to detect potential warning signs in the wider banking system.

Methodology

The main method used here is the Principal Component Analysis (PCA). PCA is an unsupervised, non-parametric statistical technique primarily used for dimensionality reduction in machine learning, by transforming a large set of variables into a smaller one that still contains most of the information in the large set. Reducing the number of components or features costs some accuracy and on the other hand, it makes the large data set simpler, easy to explore and visualize. Also, it reduces the computational complexity of the model which makes machine learning algorithms run faster.

Data Exploration

Data Transformation

The dataset used in this project was sourced from http://www.bnm.gov.my/index.php?ch=statistic and was transformed to produce the following variables:

- Ratios to total deposits:
  - Demand deposits ratio (`dd.deposits.r`)
  - Foreign currency deposits ratio (`fx.deposits.r`)
  - Repurchase agreements ratio (`repo.deposits.r`)
- Ratios to total loan applied:
  - Passenger car loan application ratio (`loan.app.cars.r`)

  - Construction loan application ratio (`loan.app.construction.r`)
  - Non-residential property loan application ratio (`loan.app.nonresprop.r`)
  - Residential property loan application ratio (`loan.app.resprop.r`)
  - Working capital loan application ratio (`loan.app.workingcapital.r`)
- Total loans applied growth rate (`loan.yy`)
- Liquidity capital ratio (`lcr`)

Data Balancing

The dataset is "unbalanced" due to the different starting points to which the data was first reported. For example, there are clusters of variables with different number of observations; `npl.r`, which has data as far back as January 1997, has 270 observations, whereas `lcr`, which BNM only began reporting in June 2015, only has 40 recorded observations. For this reason, the dataset had to be balanced. That means to equalize the amount of entries for each variable, by imputing missing values using regularized iterative PCA algorithm.

If PCA was done on the raw dataset, the size of dataset included in the analysis will be constrained to the size of the variable with the least amount of entries and the entries which all variables concurrently recorded an observation. To illustrate this, the following is a graph of the available nonperforming loans (NPL) ratio data (which the bank only began reporting in December 2008), which has recorded observations less than half of the variable with the longest reporting time frame in the dataset:

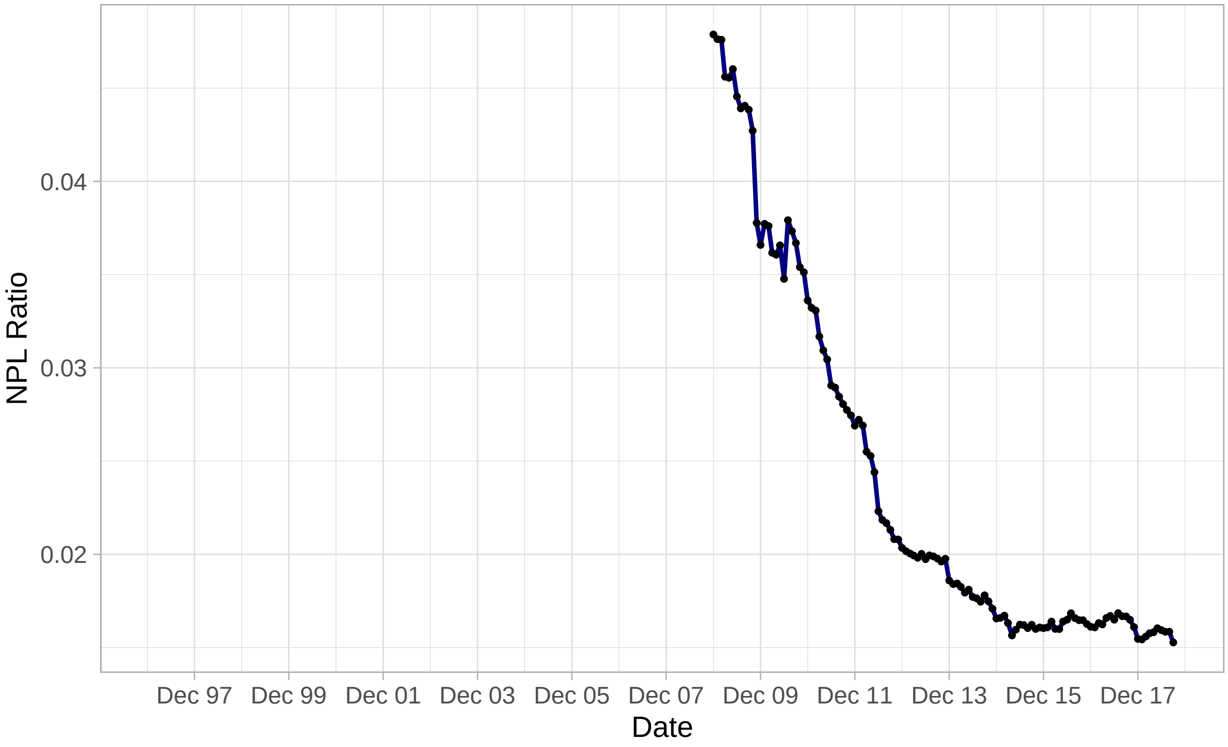

Figure 1: NPL ratio, as sourced

For this reason, the regularized iterative PCA algorithm was used for this project, in order to impute the missing values using existing principal axes and components in the dataset whilst simultaneously overcoming the issue of overfitting. The R Package `missMDA` is used to perform principal component methods on incomplete data, aiming at estimating parameters and obtaining graphical representations despite mising values. Using these algorithms yields the following graph of imputed nonperforming loans (NPL) ratio for entries before December 2008 and maintaining the original values December 2008 onwards:

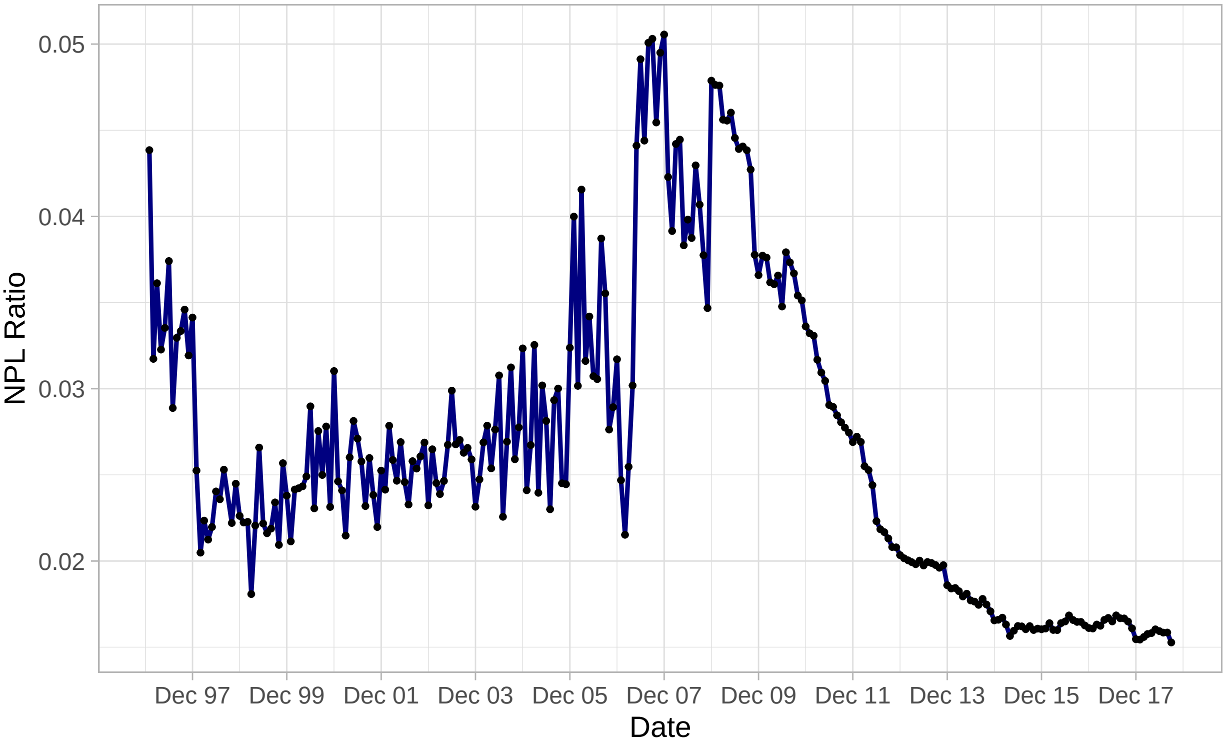

Figure 2: NPL ratio with imputed values for observations before December 2008

However, the imputed values affect the inter-variable relationships, by either strengthening existing correlations in the original dataset. The following figures illustrate that, whilst the full summary statistics of the datasets before and after the transformation are presented just before the results.

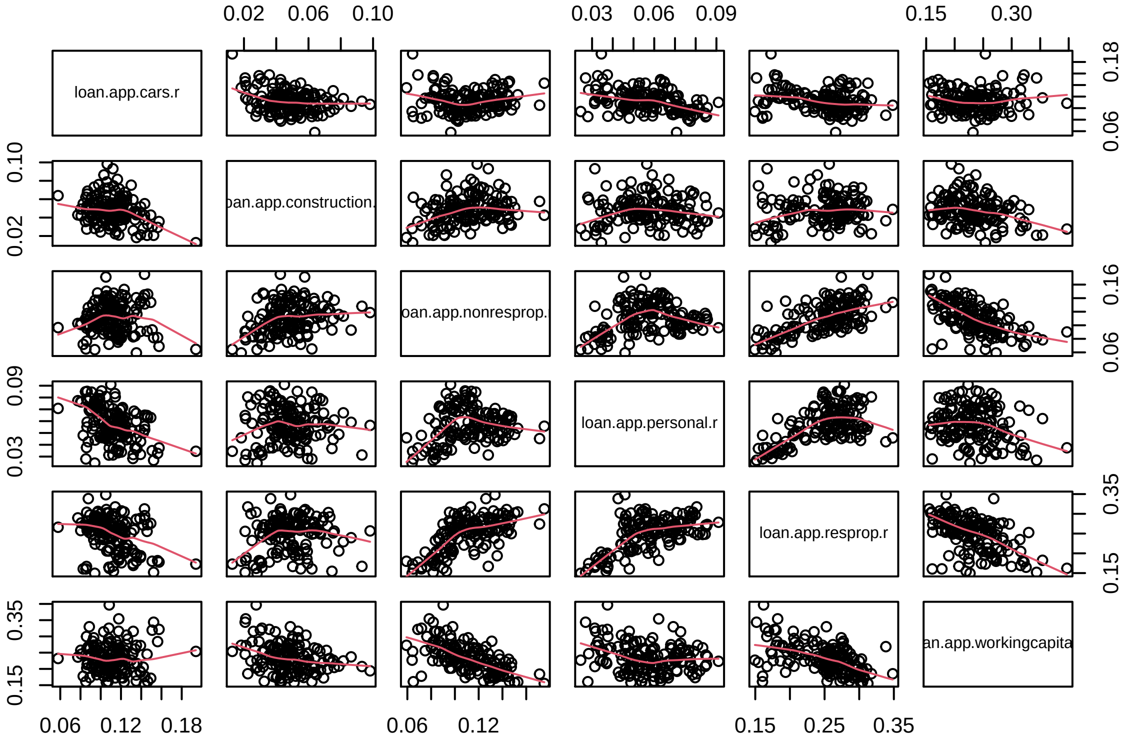

Figure 3: Relationship between loan application ratios in the raw dataset

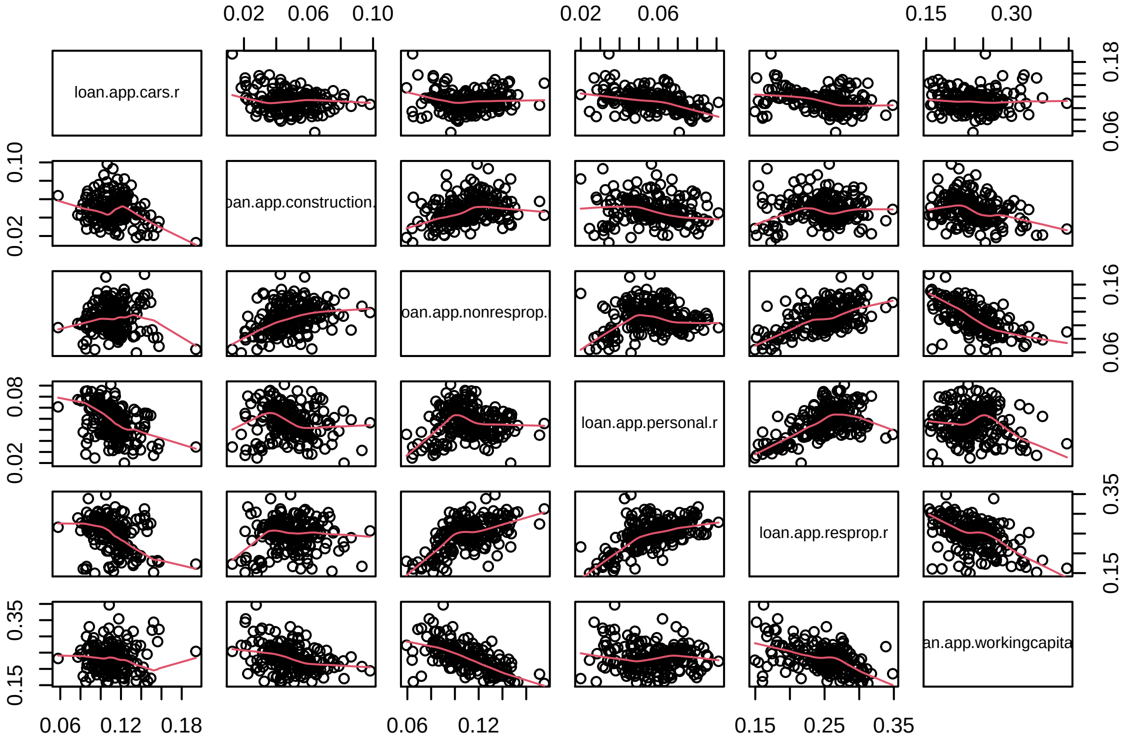

Figure 4: Relationship between loan application ratios in the imputed dataset

Analysis

Principal Component Analysis (PCA)

Eigendecomposition – eigenvalues

The goal of PCA is to reduce the dimensionality, meaning to reduce the number of variables, while retaining as much as possible of the variation present in the data.

When running the PCA on the new transformed dataset with imputed and original values, the analysis yields eigenvalues, i.e. a vector of values that provide information about the amount of variability captured by each principal component (PC). Principal components are new variables that are constructed as linear combinations or mixtures of the initial variables. Each eigenvalue covers a proportion of variation that exists in the dataset.

We explore through the following table and histogram the eigenvalues of each PC (listed as “comp”) from the PCA.

Table 1: Table of eigenvalues

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	4.0343	31.0333	31.0333
comp 2	3.4497	26.5362	57.5695
comp 3	1.2558	9.6602	67.2297
comp 4	1.1043	8.4944	75.7241
comp 5	0.9179	7.0610	82.7851
comp 6	0.5860	4.5073	87.2924
comp 7	0.4563	3.5101	90.8025
comp 8	0.3799	2.9221	93.7246
comp 9	0.3059	2.3534	96.0780
comp 10	0.1727	1.3281	97.4061
comp 11	0.1589	1.2222	98.6283
comp 12	0.1358	1.0444	99.6727
comp 13	0.0426	0.3273	100.0000

The main challenge is to determine the number of dimensions to reduce, since there is not a general rule to determine that. A useful metric to decide which PCs to retain is the amount of variance each PC covers.From the above table, the percentage of variance captured by the first two dimensions covers over half (57.6%) of the entire variation in the dataset.

The percentage variance captured by the first three dimensions covers about two-thirds of the entire variation in the dataset (67.2%). Hence, the first two PCs will be given the most focus, while the third PC is included intermittently for comparative and illustrative purposes.

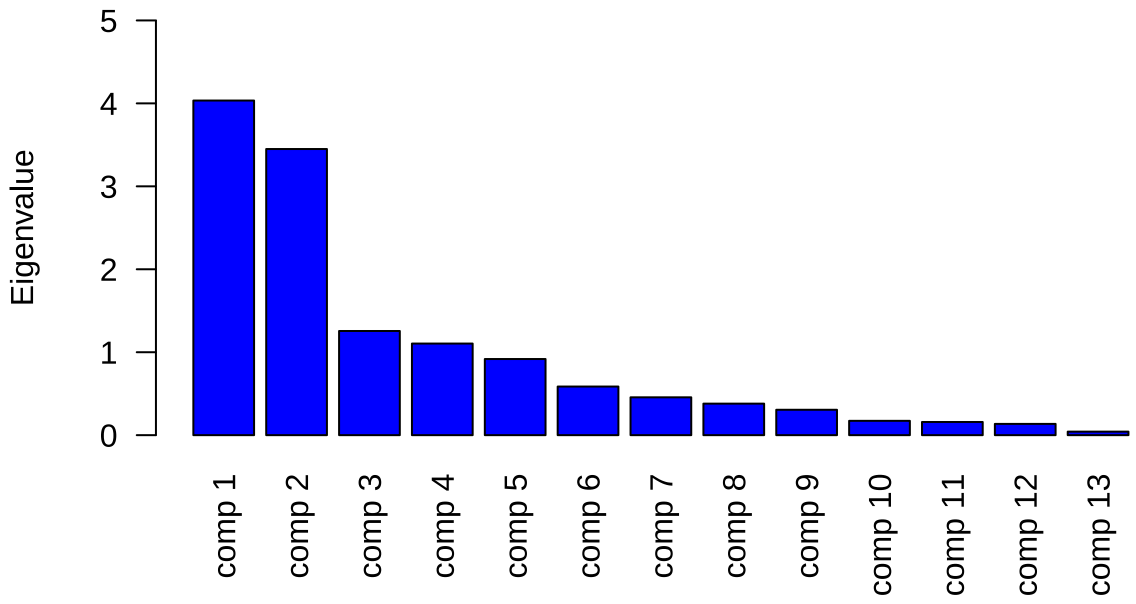

Figure 5: Screenplot of eigenvalues

Evaluation of the correlations between variables and principal components

To find out how each PC is characterized, we evaluate the correlations between the variables and the PCs as below. PCs are listed in the following table as “Dim”)

Table 2: Correlation between variables and PCs

	Dim.1	Dim.2
dd.deposits.r	0.2122	0.7983
fx.deposits.r	0.6564	0.6740
repo.deposits.r	-0.6852	0.3819
loan.app.cars.r	-0.4925	-0.0739
loan.app.construction.r	0.1110	-0.6163
loan.app.nonresprop.r	0.5065	-0.6601
loan.app.personal.r	0.7821	0.4720
loan.app.resprop.r	0.8853	-0.0696
loan.app.workingcapital.r	-0.4652	0.5913
loan.yy	-0.1376	-0.8114
nfa.yy	-0.1236	0.0423
depo.yy	-0.3876	-0.2886
lcr	0.8838	-0.2655

To further understand the correlation between them we plot the circle of correlations.

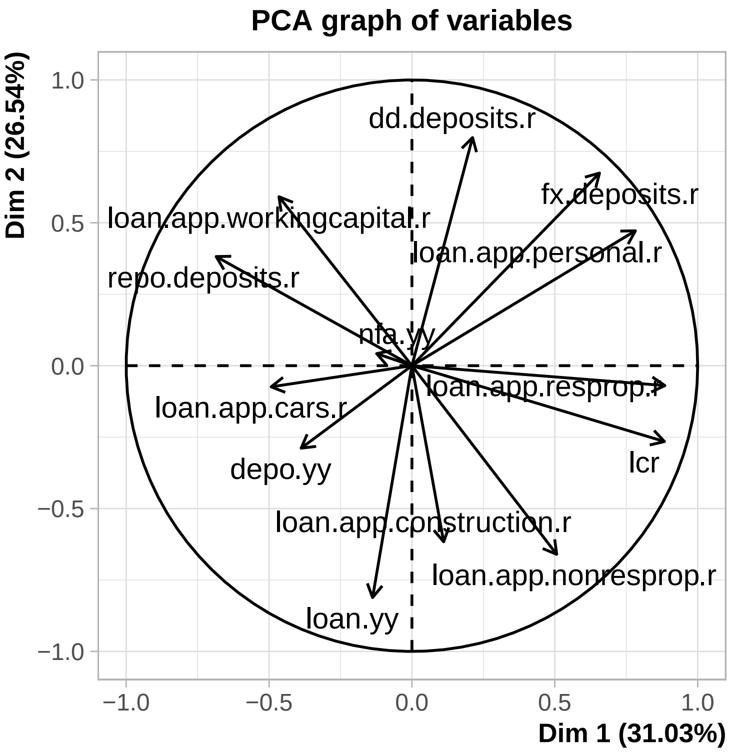

Figure 6: Variables factor map

The closer an arrow is to the circumference of the circle, the better its representation on the given axes. For example, if we take the residential property loan applications ratio (`loan.app.resprop.r`), we can notice that PC1 has a strong positive correlation with `loan.app.resprop.r` whilst registering an insignificant, negative correlation with PC2.

Contribution of each variable to each principal component

Furthermore, we evaluate how variables characterize each PCs by examining the contributions of each variable to each principal component. The following table illustrates the proportion of each variable that make up a single principal component:

Table 3: Contributions of variables on each principal component

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
dd.deposits.r	1.1161	18.4734	12.1968	0.0086	5.3174
fx.deposits.r	10.6802	13.1692	0.6559	0.7047	1.8705
repo.deposits.r	11.6391	4.2273	11.2983	0.1107	0.6166
loan.app.cars.r	6.0120	0.1585	31.2103	16.5201	3.0715
loan.app.construction.r	0.3057	11.0114	0.4777	10.3968	27.2086
loan.app.nonresprop.r	6.3592	12.6315	8.6681	0.8344	1.0770
loan.app.personal.r	15.1635	6.4581	0.1343	1.0963	0.0411
loan.app.resprop.r	19.4251	0.1403	1.4825	0.5996	1.4969
loan.app.workingcapital.r	5.3653	10.1341	13.3699	0.2713	1.6626
loan.yy	0.4695	19.0865	1.9737	0.5968	0.1237

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
nfa.yy	0.3786	0.0518	2.3971	56.5477	33.4362
depo.yy	3.7243	2.4150	13.9091	12.0576	23.1312
lcr	19.3612	2.0428	2.2263	0.2554	0.9467
TOTAL	100.0000	100.0000	100.0000	100.0000	100.0000

The regularised iterative PCA algorithm transforms the original workable dataset from the Table 4 to Table 5 (note the number of observations for each variable under the column “Count”):

Table 4: Before transformation

	Minimum	Maximum	Mean	Median	StDev	Count
dd.deposits.r	3.5349	5.9648	4.4144	4.2233	0.5808	40
fx.deposits.r	1.3918	2.6076	1.7868	1.7776	0.2625	40
repo.deposits.r	0.1380	0.3716	0.2444	0.2307	0.0559	40
loan.app.cars.r	0.0578	0.1216	0.0959	0.0962	0.0121	40
loan.app.construction.r	0.0224	0.0864	0.0462	0.0450	0.0133	40
loan.app.nonresprop.r	0.0870	0.1175	0.1026	0.1019	0.0083	40
loan.app.personal.r	0.0551	0.0910	0.0750	0.0745	0.0071	40
loan.app.resprop.r	0.2348	0.3176	0.2704	0.2674	0.0207	40
loan.app.workingcapital.r	0.1732	0.3298	0.2382	0.2374	0.0341	40
loan.yy	-1.3457	15.7542	5.7492	5.3173	4.3140	40
nfa.yy	-65.8031	495.5951	13.9990	2.8135	82.4924	40
depo.yy	-27.8895	26.5950	3.5531	3.8723	9.1832	40
lcr	116.0000	145.0000	130.3250	128.5000	8.3800	40

Table 5: After transformation

	Minimum	Maximum	Mean	Median	StDev	Count
dd.deposits.r	1.5450	5.9800	3.9257	3.9724	0.6813	260
fx.deposits.r	-0.2552	2.6076	1.2347	1.2503	0.4146	260
repo.deposits.r	-0.2619	3.3420	0.4491	0.2214	0.6687	260
loan.app.cars.r	0.0578	0.1940	0.1117	0.1109	0.0148	260
loan.app.construction.r	0.0128	0.0982	0.0474	0.0457	0.0126	260
loan.app.nonresprop.r	0.0593	0.1752	0.1099	0.1080	0.0183	260
loan.app.personal.r	0.0199	0.0910	0.0578	0.0592	0.0128	260
loan.app.resprop.r	0.1495	0.3484	0.2484	0.2522	0.0339	260
loan.app.workingcapital.r	0.1551	0.3969	0.2343	0.2345	0.0353	260
loan.yy	-20.8429	50.4676	7.9953	7.1200	8.5919	260
nfa.yy	-67.4449	495.5951	16.0721	4.9507	54.4199	260
depo.yy	-27.8895	72.0200	8.6885	7.3000	13.2417	260
lcr	108.6945	145.0000	127.1583	127.3462	5.8685	260

Focusing on the first three principal components, the contribution for each variable is plotted in the following figures:

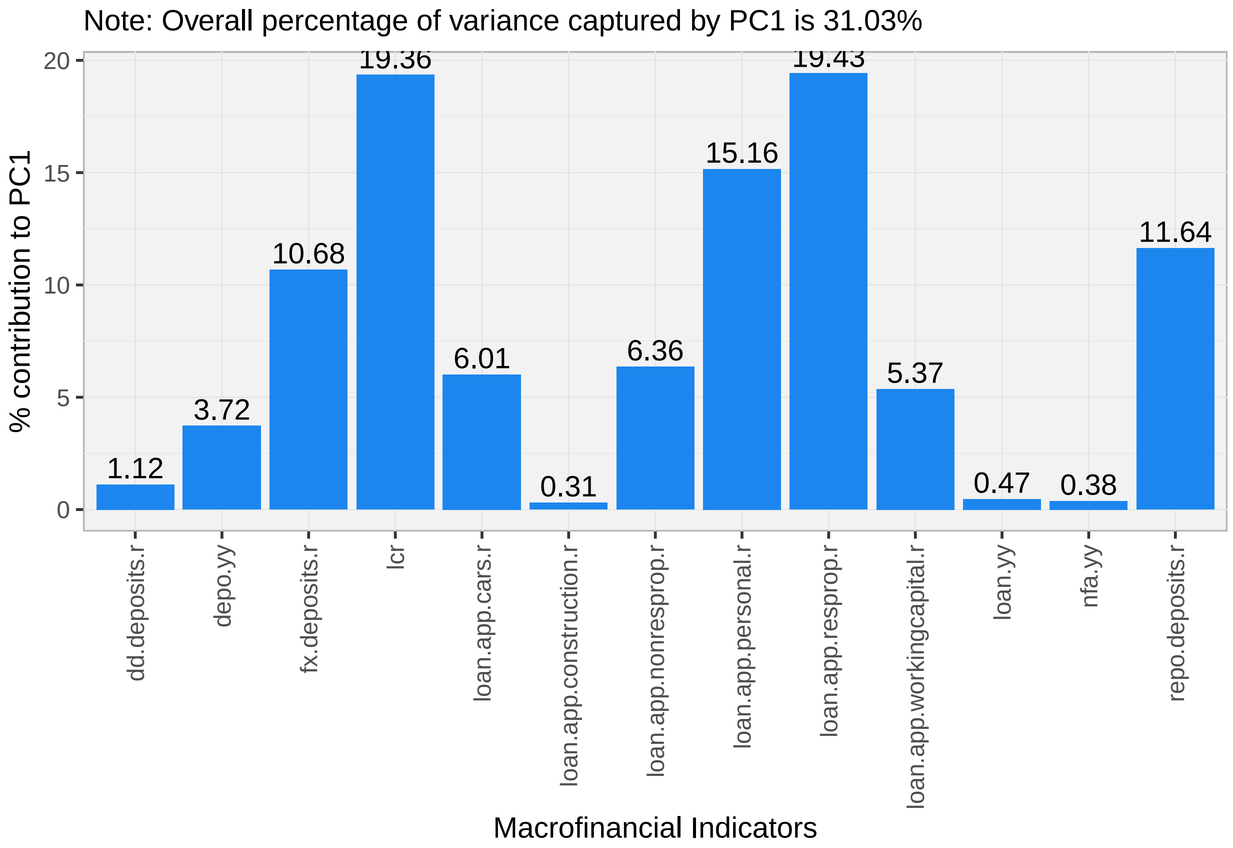
Note: Overall percentage of variance captured by PC1 is 31.03%

Figure 7: Influence of each macrofinancial indicator on PC1

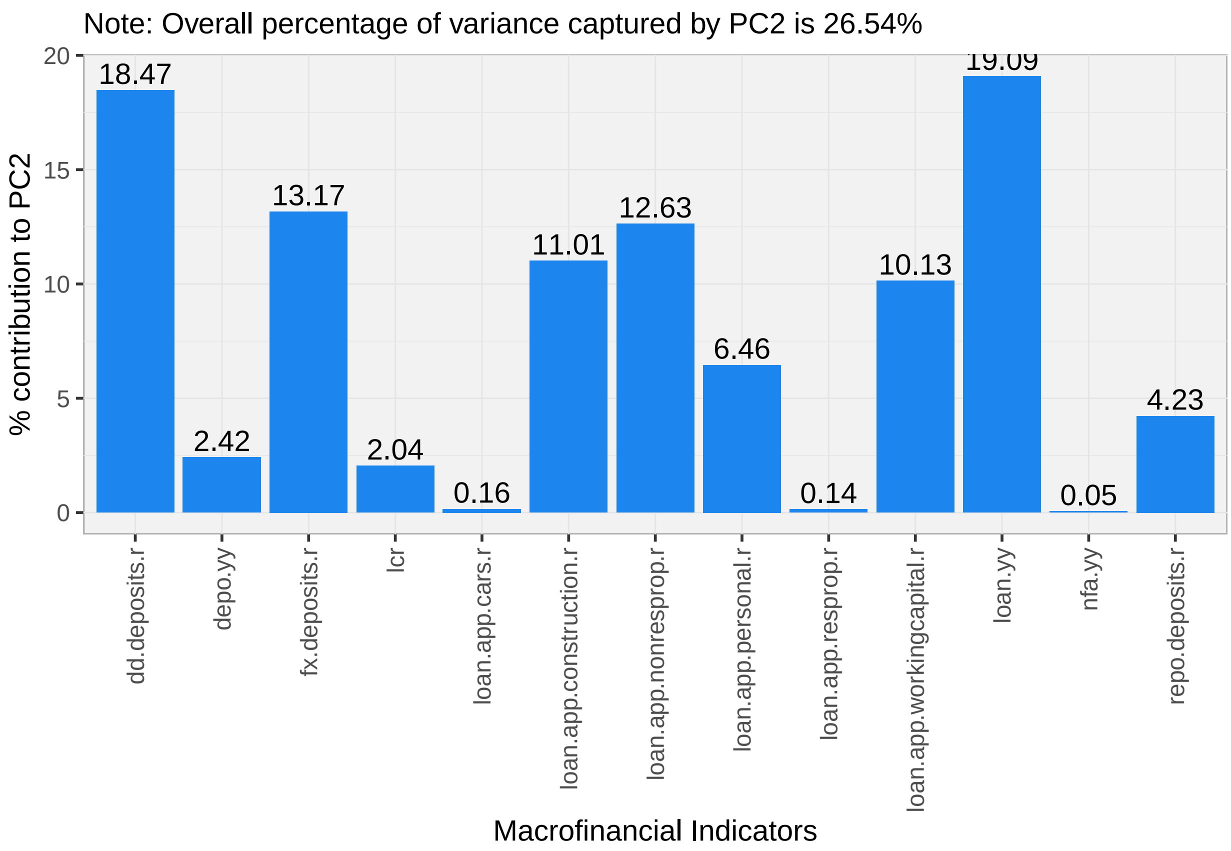

Figure 8: Influence of each macrofinancial indicator on PC2

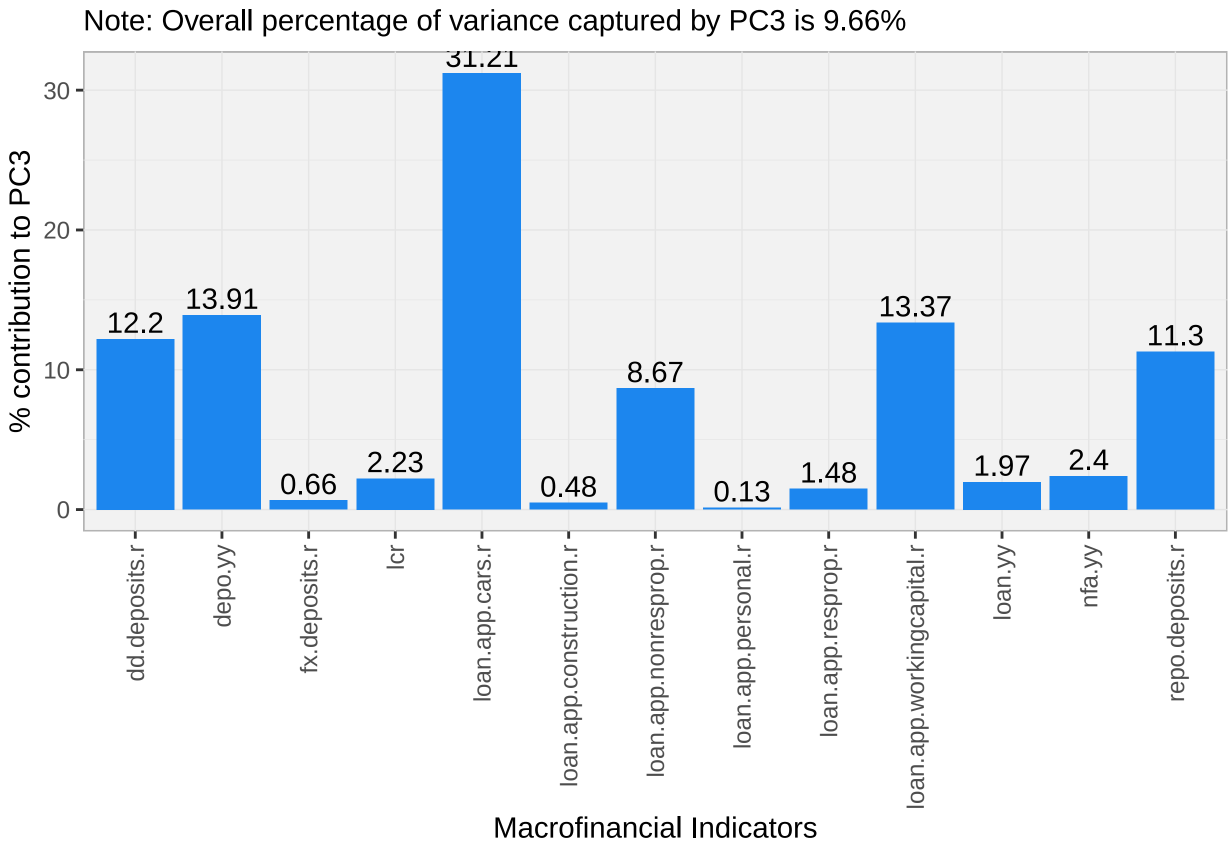

Figure 9: Influence of each macrofinancial indicator on PC3

Results

Plotting principal component scores

We use the principal component scores derived from each observation as coordinates to plot the objects in a scatterplot. Also a colour gradient is applied to each PC score to visualize the impact of moving towards certain regions in the PC1-PC2 plot on NPL ratios.

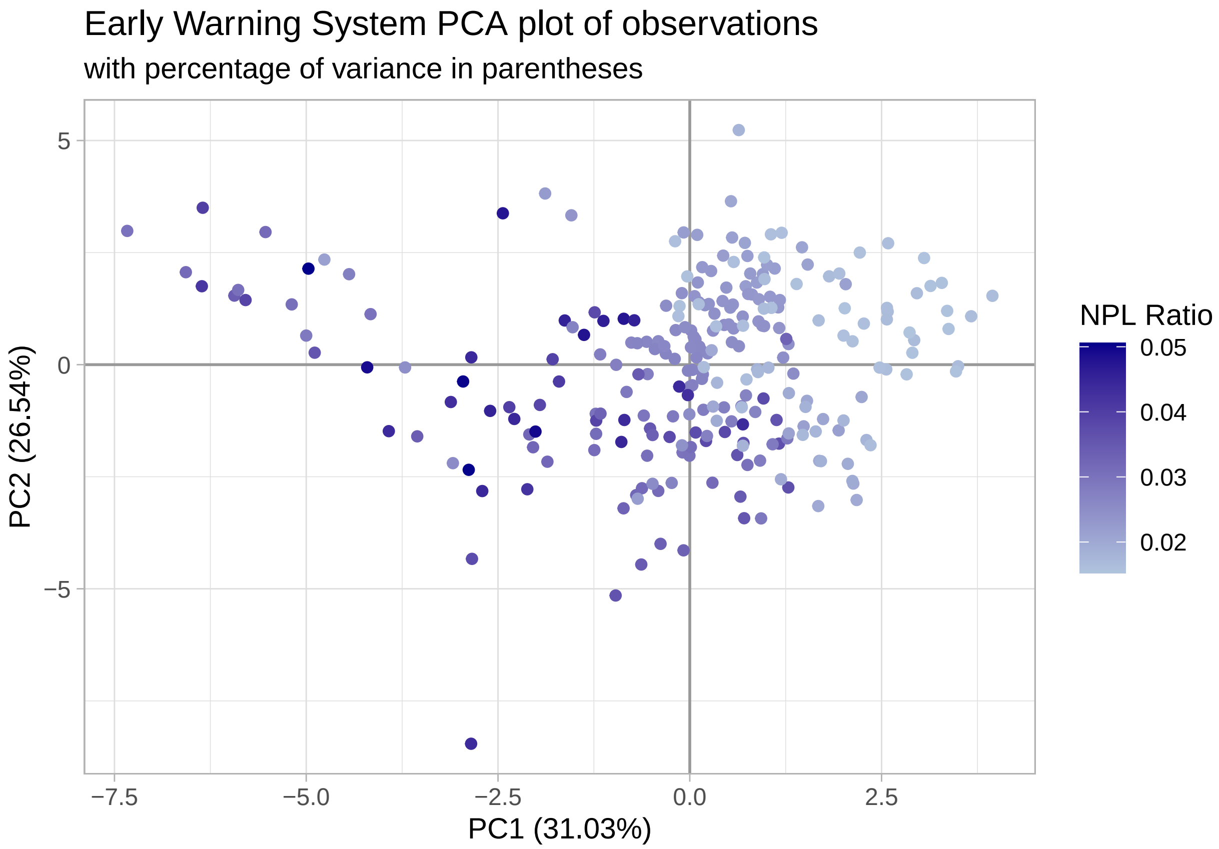

Figure 10: Scatterplot of PC scores on first two PC axes, applied colour gradient to each point to scale NPL ratio

It can be observed that moving into the negative region of PC1 generally implies higher nonperforming loans (NPL) ratios. This gives the PC1 axis a strong predictive power on the general outlook of banks' loan portfolio performance.

Conclusion

In this project, a regularized imputed PCA algorithm to impute missing values in historical datasets has been carried out. The new transformed and uniform dataset was used for a second order PCA to extract the factors and sensitivities that drives the performance of the analyzed banking system stemming from the performance of various macrofinancial indicators.

This project shows how extremely complex, dynamic and interlinked is the banking sector. Furthermore, it is shown how small gyrations in this sector can potentially impact the wider economy and vice versa and how this cluster of indicators can be used to detect potential warning signs in the wider banking system. In other words, we obtained a better understanding on how a cluster of factors can contribute to the overall macro effect, not simply through their individual mechanical effects, but also through their inter-linked micro effects.

The basic limitation of the PCA method is that is focused on finding orthogonal projections of the dataset that contains the highest variance possible in order to 'find hidden LINEAR correlations' between variables of the dataset. This means that if you have some of the variables in your dataset that are linearly correlated, PCA can find directions that represents your data. But if the data is not linearly correlated, then PCA would not be enough:

Another limitation could be the careless selection of the number of Principal Components. Although Principal Components try to cover maximum variance among the features in a dataset, if we don't select the number of Principal Components with care, it may miss some information as compared to the original list of features.

As future work towards enhancing the impact of this project, it could be the analysis of the wider banking system on country level, or it could also be applied on continental level. A bank should always be examined combined with the country´s financial dynamic and perspective, in order to acquire a realistic predictive power for its indicators.