High Dimensional Data Analysis - Group Assignment 18

Dewi Amaliah, Aarathy Babu, Rahul Bharadwaj & Priya Dingorkar

9th September 2021

Contents

Introduction

The UCLA-LoPucki Bankruptcy Research Database (BRD) is a UCLA School of Law data gathering, data linking, and data distribution initiative. The goal of the BRD is to encourage bankruptcy research by making bankruptcy data available to academic investigators worldwide. All of the data was gathered when the companies declared bankruptcy. In this study, we will use a variety of high-dimensional analysis approaches like Multidimensional Scaling, Principle Component Analysis and Clustering to extract useful insights from the data.

Acknowledgement

Our sincere gratitude goes out to Ruben Loaiza-Maya and the our tutor Ari Handayani for their guidance and support. Their culminating efforts have placed us in a situation where we can produce this report collectively while showcasing our competence to employ diverse solutions that can be used with high dimensional data.

Data Description

This report is based on data from US businesses that declared bankruptcy between 1980 and 2000. The information is coming from the UCLA-LoPucki Brankruptcy Research Database. Let's take a closer look at these variables and what they mean. Let's go further into the dataset to see if we can find any examples of data cleaning or wrangling.

The dataset has 436 observations and 7 variables with their description explained below.

- Name: Name of the firm
- Assets: Total assets (in millions of dollars)
- CityFiled: City where filing took place
- CPI U.S CPI at the time of filing
- DaysIn: Length of bankruptcy process
- DENYOther: CityFiled, categorized as Wilmington (DE), New York (NY) or all other cities (OT)
- Ebit: Earnings (operating income) at time of filing (in millions of dollars)
- Employees: Number of employees before bankruptcy
- EmplUnion: Number of union employees before bankruptcy
- FilingRate: Total number of other bankrupcy filings in the year of this filing
- FirmEnd: Short description of the event that ended the firm's existence
- GDP: Gross Domestic Product for the Quarter in which the case was filed
- HeadCityPop: The population of the firms headquarters city
- HeadCourtCityToDE: The distance in miles from the firms headquarters city to the city in which the case was filed
- HeadStAtFiling: The state in which firms headquarters is located
- Liab: Total amount of money owed (in millions of dollars)
- MonthFiled: Categorical variable where numbers from 1 to 12 correspond to months from Jan to Dec
- PrimeFiling: Prime rate of interest on the bankruptcy filing date
- Sales: Sales before bankruptcy (in dollars)
- SICMajGroup: Standard industrial clasification code
- YearFiled: Year bankruptcy was filed

Let us further examine the bankruptcy statistics. We will undertake some preliminary data analysis. We will also go through the numerous approaches used for this high-dimensional data in detail later.

Princple Component Analysis (PCA)

- Now that we've seen how to input this high-dimensional data into Multidimensional scaling (MDS) to obtain a low (typically 2) dimensional representation. Let us now perform a Principal Component Analysis (PCA), which is a dimensional-reduction method that is frequently used to reduce the dimensional of large data sets by transforming a large set of variables into a smaller one that still contains the majority of the information in the large set.
- On performing a PCA on the clean dataset during the data investigation process we found that, just like MDS sees Texaco Inc. which comes under Petroleum SIC and Refining And Related Industries as the outlier we get similar results when we feed the complete dataset to carry out PCA.
- Keeping in mind, the word limit we have constrained to show only the PCA using the data without any outliers. Note the complete analysis of PCA for the data can be found in the Appendix section at the end of this report.
- Let's carry out PCA on our bankruptcy data. Lets investigate if our data is a good fit for PCA. Let's us further investigate how the variables in our data are correlated and how many PC's explain the variation of our data.
- Before we apply this principle to our data, it is very important that we standardized our variables as this ensures that results are not sensitive to the units of measurement. Thus giving us more accurate analysis.

PCA without outlier

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     1.7801 1.6335 1.1244 1.02884 1.00066 0.92352 0.86648
## Proportion of Variance 0.2641 0.2224 0.1054 0.08821 0.08344 0.07107 0.06257
## Cumulative Proportion  0.2641 0.4864 0.5918 0.67997 0.76342 0.83449 0.89706
##                            PC8     PC9    PC10   PC11    PC12
## Standard deviation     0.79510 0.56550 0.48833 0.1766 0.11704
## Proportion of Variance 0.05268 0.02665 0.01987 0.0026 0.00114
## Cumulative Proportion  0.94974 0.97639 0.99626 0.9989 1.00000
```

- Using the `summary` function we can infer the following:
  - Proportion of variance explained by the first four PCs together is now 67.99%
  - Proportion of variance explained by the first and second PC alone is 26.41% and 22.24% respectively
  - Using kaisers rule, choose those PC's whose variance and standard deviation greater than 1, we will choose 5 PC's that explains the most variation in our data.
- Let us now use the scree plot to find the total number of PC's that best explain our data.

- Interestingly the scree plot and the kaiser's rule do not agree with each other. But since the scree plot is a more accurate measure in helping us confirm that most of the variations is captured by the first three principal components from our dataset.
- As we can see from the summary of our PCA that our third PC explains approx 11% of variation in our data but we cannot visualize this with a biplot. Nevertheless we still plot the biplot for our first two PC's as they explain around 50.00% of variation in our data.

- Looking at the figure 2 we can infer the following:

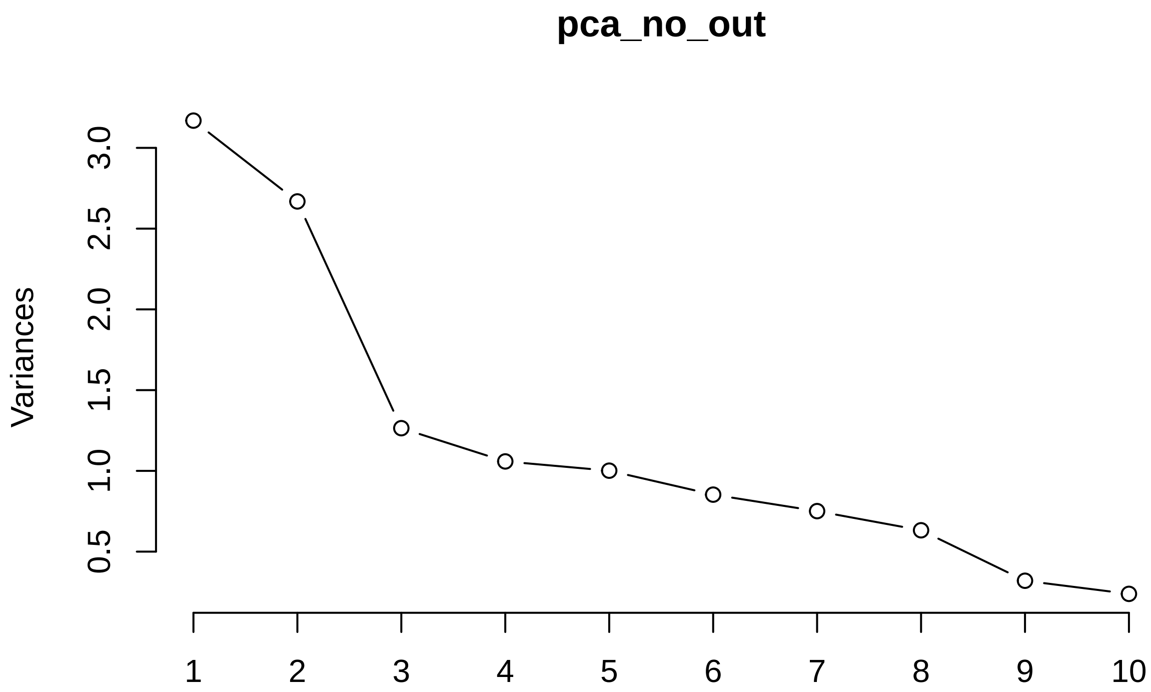

Figure 1: Screeplot

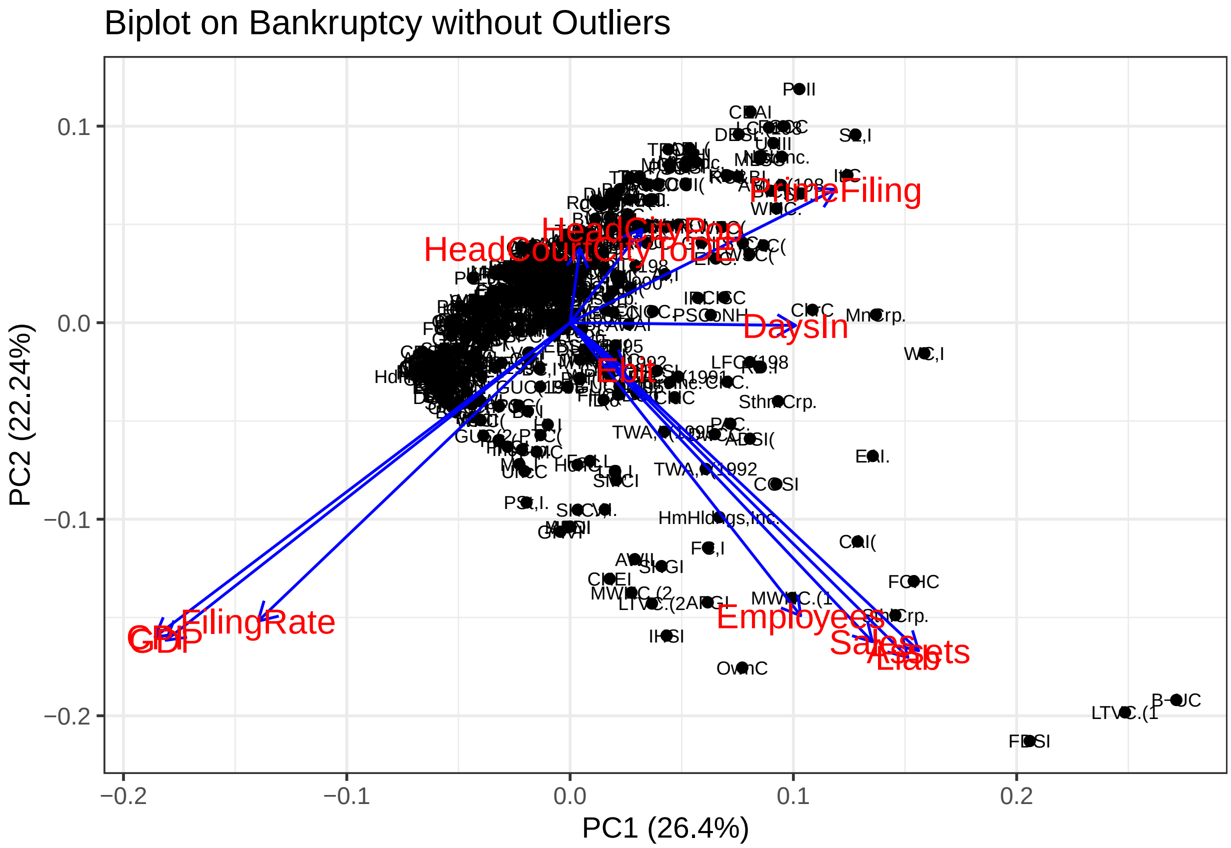

Figure 2: PCA Biplot without Outliers

- We can see the various spread of the different companies while filing for bankruptcy. We can see how these companies are been spread out in different direction showing different properties with the surrounding variables. For instance two companies that show similar characteristics are FDSI and LDVC and the ones that show very dissimilar characteristics to each other are FDSI and DESL.
- The further away these vectors or variables are from a PC origin, the more influence they have on that PC. For instance taking a closer a look at the left quadrant, we infer that CPI has more influence, followed by GDP and FilingRate whereas EBit has the least influence among all the variables.
- We also know that variables at an angle of 90° indicates no correlation between them, In our data we can infer that Employees with GDP/CPI shows an angle of almost 90° thus showing no correlation between these variables.
- We also infer that variables with 180° angle indicates negative correlation. In our case we can say that PrimeFiling is negative correlated with GDP, CPI and FilingRate making an almost 180° angle.
- Similarly,variables with angle close to 0° indicates positive correlation. In our dataset we can see that CPI and GDP are highly positively correlated. Assest, Liab, Ebit and Sales are all also positively correlated as the angle between all them is nearly zero.

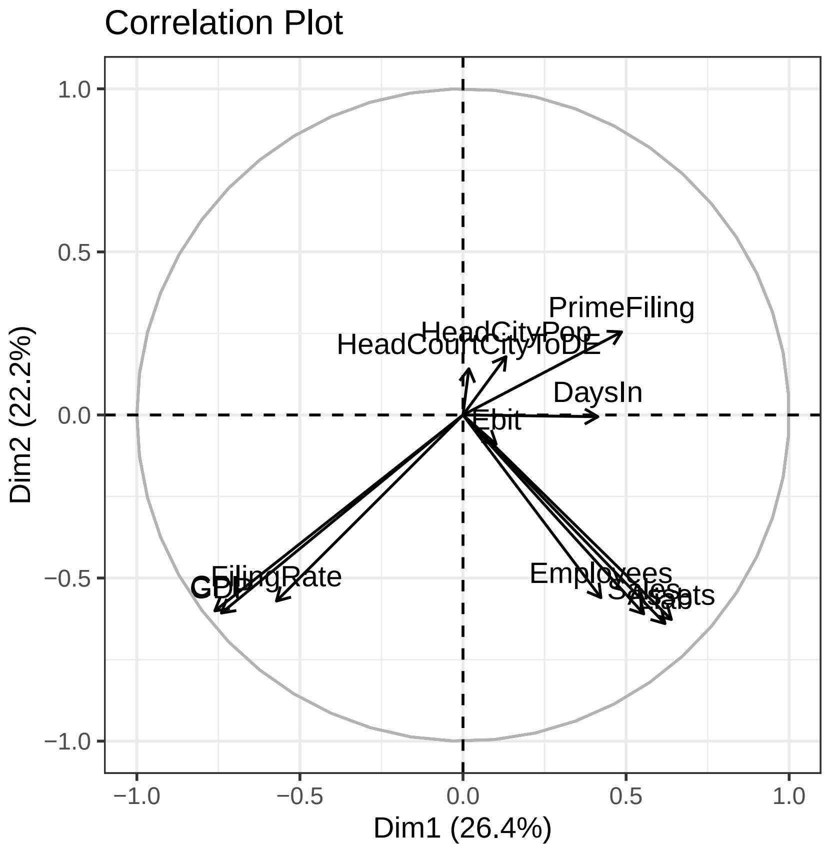

Figure 3: PCA Correlation Plot

- We can also refer to figure 3 for more clear visuals of the angles between variables making it clearly for the audience to distinguish between the variables that are positively, negatively or not correlated at all.

Limitation

- One of the most important limitation in our PCA analysis is that, as seen in the screeplot at figure we inferred that our data was most explained by the the first three PCs. But due to the limitations of biplot we have visualized our data using the first and the second PCs only. This mean they might be certain patterns in the data and to visualize third PCs we will further need to work on the structure of the data.

Appendix

PCA

- PCA performed on the complete data including outliers

```
## Importance of components:
##                          PC1    PC2    PC3     PC4    PC5     PC6     PC7
## Standard deviation     1.900 1.6842 1.1354 1.02794 0.9392 0.85634 0.82987
## Proportion of Variance 0.301 0.2364 0.1074 0.08806 0.0735 0.06111 0.05739
## Cumulative Proportion  0.301 0.5374 0.6448 0.73283 0.8063 0.86745 0.92484
##                            PC8     PC9    PC10    PC11   PC12
## Standard deviation     0.68629 0.49184 0.39448 0.14244 0.1148
## Proportion of Variance 0.03925 0.02016 0.01297 0.00169 0.0011
## Cumulative Proportion  0.96409 0.98424 0.99721 0.99890 1.0000
```

- Using the `summary` function we can infer the following:
  - Proportion of variance explained by the first four PCs together is 73.28%
  - Proportion of variance explained by the first and second PC alone is 30.10% and 23.64% respectively
  - Using kaisers rule, we choose those PC's whose variance and standard deviation is greater than 1, in our bankruptcy data we will choose 4 PC's.
- Let us now plot use the scree plot to find the total number of PC's that best explain our data.

- Using the scree plot we infer that our bankruptcy data is explained by the first three PC's. Also note this is different to kaisers rule.
- Let us now plot a biplot, that will help us infer intersting features about our data. A PCA biplot shows both PC scores of samples (dots) and loadings of variables(vectors).

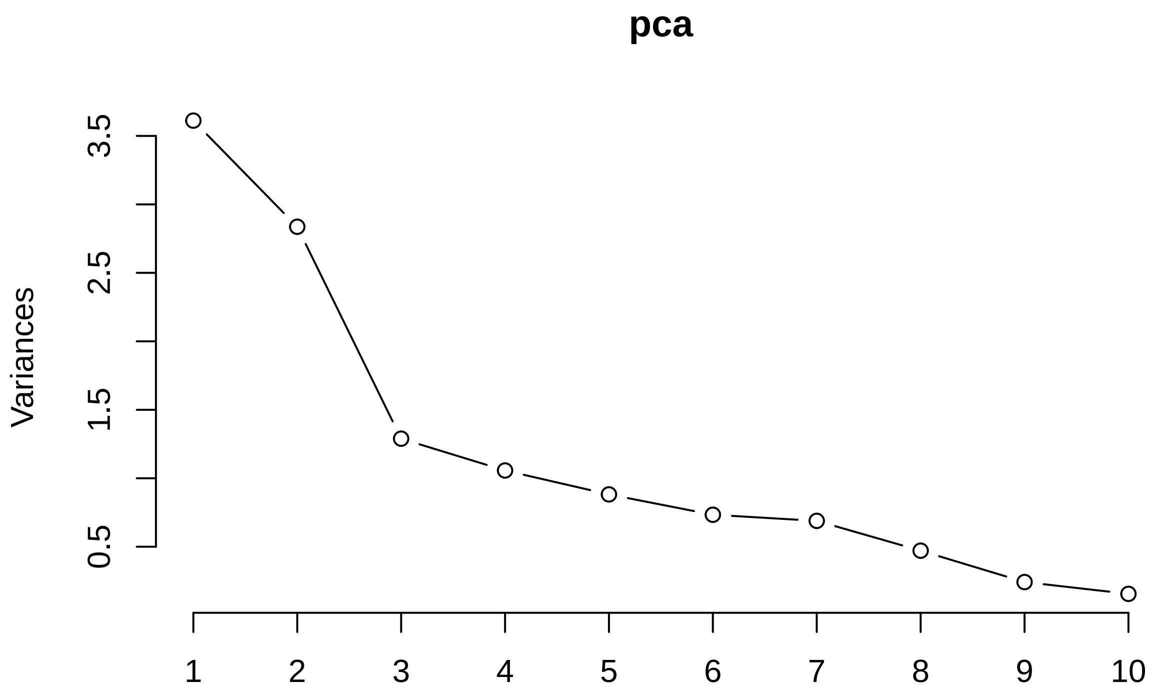

Figure 4: Scree Plot

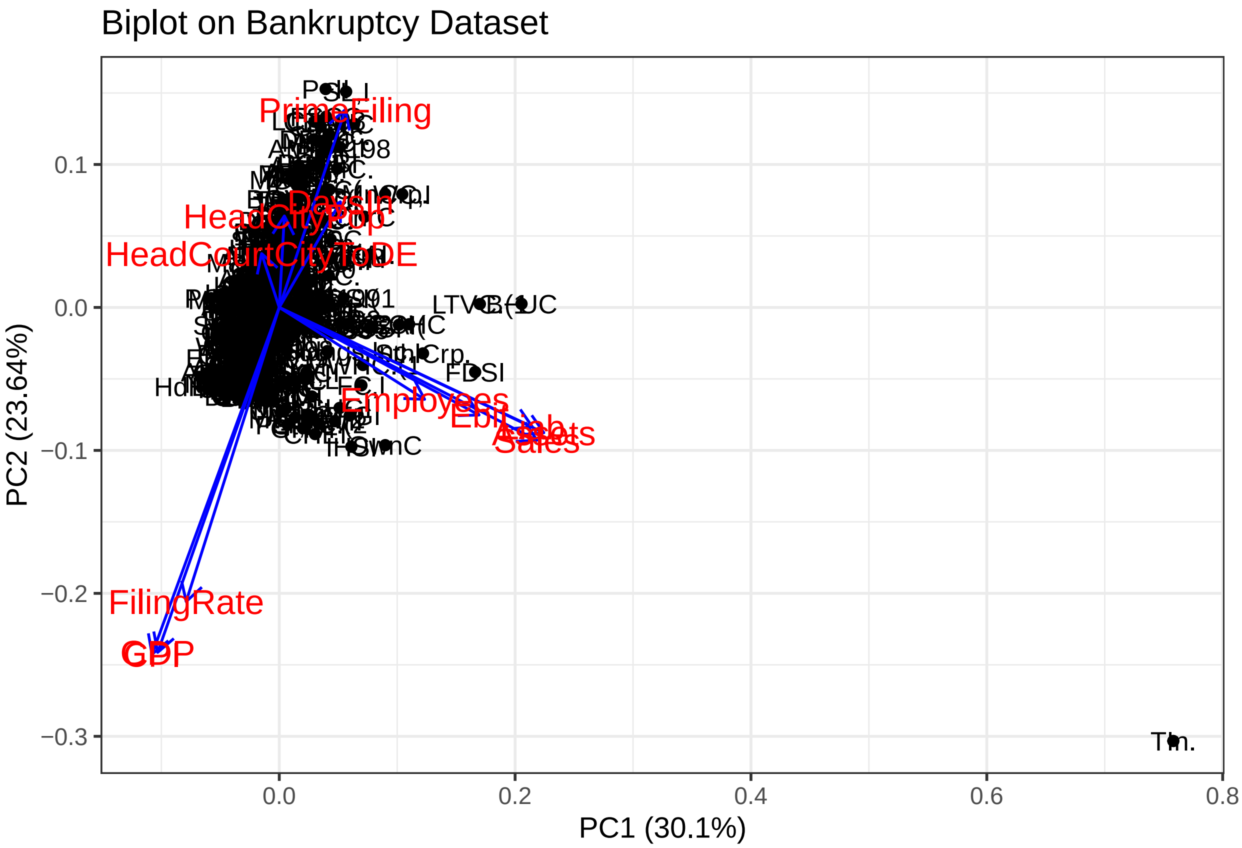

Figure 5: PCA Biplot