Alma Mater Studiorum

Department of Computer Science and Engineering

Low-rank Approximation for Data Analysis: PCA on Pizza dataset

PhD Student
Marco Ferrati
1097861

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Principal Component Analysis

Principal Component Analysis (PCA) is an unsupervised dimension-reducing technique applicable to datasets (seen as matrices) with n features. This technique allows for the reduction of variables in a dataset while retaining as much information as possible (it is expected that features will lose their physical meaning).

- decrease computation time when working on the "new" dataset;
- decrease storage space consumption because the "new" dataset is smaller;
- increase the simplicity in making plot.

PCA allows you to find up to n Principal Components (PCs). These are new axes with the characteristic that from PC_1 to PC_n the quantity of information decreases. There are two techniques to find PCs:

1. **Covariance matrix**: if we want to keep some hierarchy of features. It means that some features are more important than others;
2. **Correlation matrix**: if all the features have the same importance.

The number of PCs that we want to find is a parameter to choose. There are several ways to find which number is the best, and using a combination of them is usually a good choice. With the idea to plot the new matrix, two or three PCs are the best choice. These techniques are:

- cumulative variance: choose k PCs, with $k : t_k > t^*$ with t^* the minimum value of cumulative variance that we want to have. Usually

> 70%;

- Kaiser rule: select the number of PCs comparing the eigenvalues of each PC with the mean of every PCs. Only the PC with eigenvalues greater than the mean are taken;

- scree plot: plot the eigenvalues for each component number and choose the number of PCs where a elbow is visible in the plot;

- LEV plot: it is a logarithmic transformation of the scree plot (LEV means Log EigenValues). It is used when the differences are small.

After choosing how many PCs to consider, the last step is to interpret the results. The idea is to understand how much any PCs is correlated to the original variables. This can be visualized with the help of the correlation circle if we chose to keep two PCs or with the help of a heat map of the correlation matrix between features and PCs.

1.2 Dataset

The dataset is from data.world, specifically the dataset ”Principal Component Analysis - Pizza Dataset” by Dhilip Subramanian has been chosen. The dataset has 300 rows and 9 columns.

The columns are:

- **brand:** pizza brand (class label);

- **id:** number of sample analysed;

- **mois:** amount of water per 100 grams in the sample;

- **prot:** amount of protein per 100 grams in the sample;

- **fat:** amount of fat per 100 grams in the sample;

- **ash:** amount of ash per 100 grams in the sample;

- **sodium:** amount of sodium per 100 grams in the sample;

- **carb:** amount of carbohydrates per 100 grams in the sample;

- **cal:** amount of calories per 100 grams in the sample.

1.3 Tools

Python3 has been chosen to perform the PCA algorithm with the help of other libraries:

- **scikit-learn:** Python package that provides a method to perform the PCA algorithm;
- **pandas:** Python package used to read and visualize the dataset;
- **matplotlib and seaborn:** Python packages used to make charts;
- **jupyter-lab:** used to write and execute the Python code in a ”notebook”.

Chapter 2

Methodology

2.1 Covariance or correlation matrix

To choose which matrix to use to compute the PCs we can look at the correlation matrix of the variables in the dataset and their standard deviations. The factors to consider are:

- important difference in standard deviation;
- high correlations between variables.

2.2 Standardization

Standardizing the input is an important task. In this way, we can address the problem that can occur if there are large differences between the ranges of initial variables; those variables with larger ranges will dominate over those with smaller ranges, which will lead to biased results. So, transforming the data to comparable scales can prevent this problem. Mathematically, this can be done by subtracting the mean and dividing by the standard deviation for each value of each variable.

$$z = \frac{value - mean}{\sigma} \tag{2.1}$$

2.3 Number of Principal Components

The scree plot and the cumulative variance can be used to choose the number of PCs to keep. The scree plot is a graphical technique where a person has to identify the elbow in the plot. For the cumulative variance, we set the threshold at 95%.

2.4 Understand the results

To better understand how the computed PCs are correlated with the features of the dataset, the correlation matrix can be computed, and to help its visualization, it can be plotted as a heat map.

2.5 Plot the Principal Components

After the execution of the PCA algorithm, we can plot the new matrix. It is interesting to reassign the class label to each sample. In this way, we can better visualize how the different samples are scattered on the plot.

Chapter 3

Results

3.1 Covariance or correlation matrix

Table 3.1 shows the correlation between the variables of the dataset and their standard deviations. A few cases have a high correlation (e.g., fat and sodium; ash and carbohydrates; ash and proteins; ash and sodium). Regarding the standard deviations, there are important differences between them ($\sigma_{carb} = 18.029722$, $\sigma_{sodium} = 0.370358$).

3.2 Number of Principal Component

From the scree plot in figure 3.1 the elbow is clearly visible at PC number three. Also, the cumulative variance (in table 3.2) reinforces this choice.

3.3 Understand the results

Figure 3.2 shows the correlation matrix between the PCs and the features of the dataset. It shows a high correlation between:

- PC1 with ash, carbohydrates, fat, and sodium;
- PC2 with calories and moisture;
- PC3 with moisture, proteins, and sodium.

	ash	**cal**	**carb**	**fat**	**mois**	**prot**	**sodium**
ash	1						
cal	0.326468	1					
carb	-0.898988	-0.023485	1				
fat	0.791634	0.764567	-0.640238	1			
mois	0.265556	-0.764441	-0.591802	-0.171318	1		
prot	0.823844	0.070258	-0.853542	0.498002	0.360248	1	
sodium	0.808122	0.671958	-0.620176	0.933325	-0.102279	0.429130	1
Standard Deviations	1.269724	0.620034	18.029722	8.975658	9.552987	6.434392	0.370358

Table 3.1: Correlation matrix between variables of the dataset and their standard deviation.

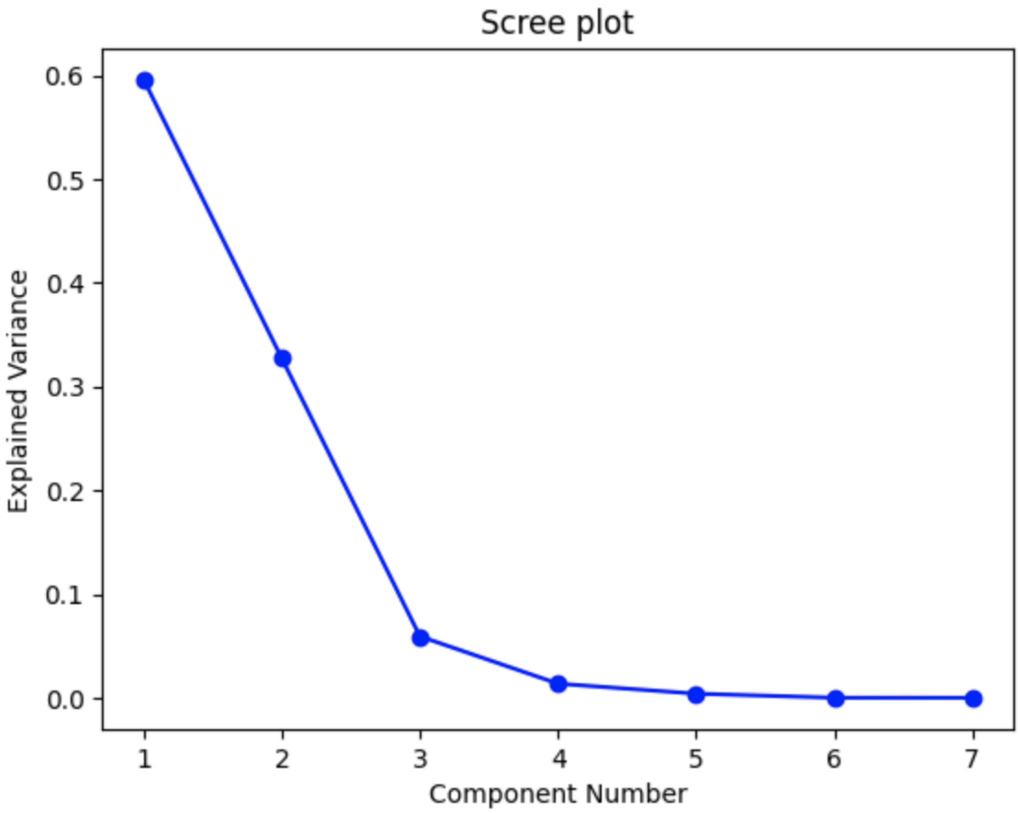

Figure 3.1: Scree plot

PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**
0.59596884	0.92317704	0.98240023	0.99599655	0.99995041	0.99999864	1

Table 3.2: Cumulative variances

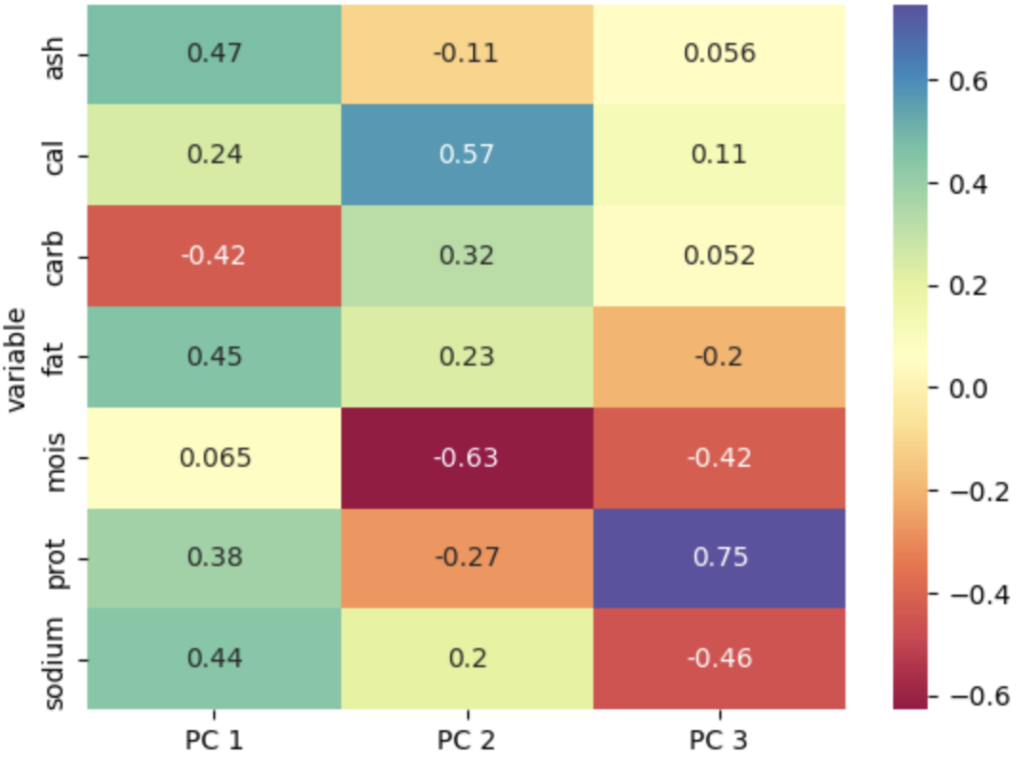

Figure 3.2: Correlation matrix between PCs and dataset's feature

3.4 Plot the Principal Components

To visualize the results of PCA, a 3D scatter plot has been used. The result (figure 3.3) shows in different colors the ten different brands in the dataset. Five groups can be easily recognized from the image:

- red;
- blue;
- orange;
- yellow;
- others colors.

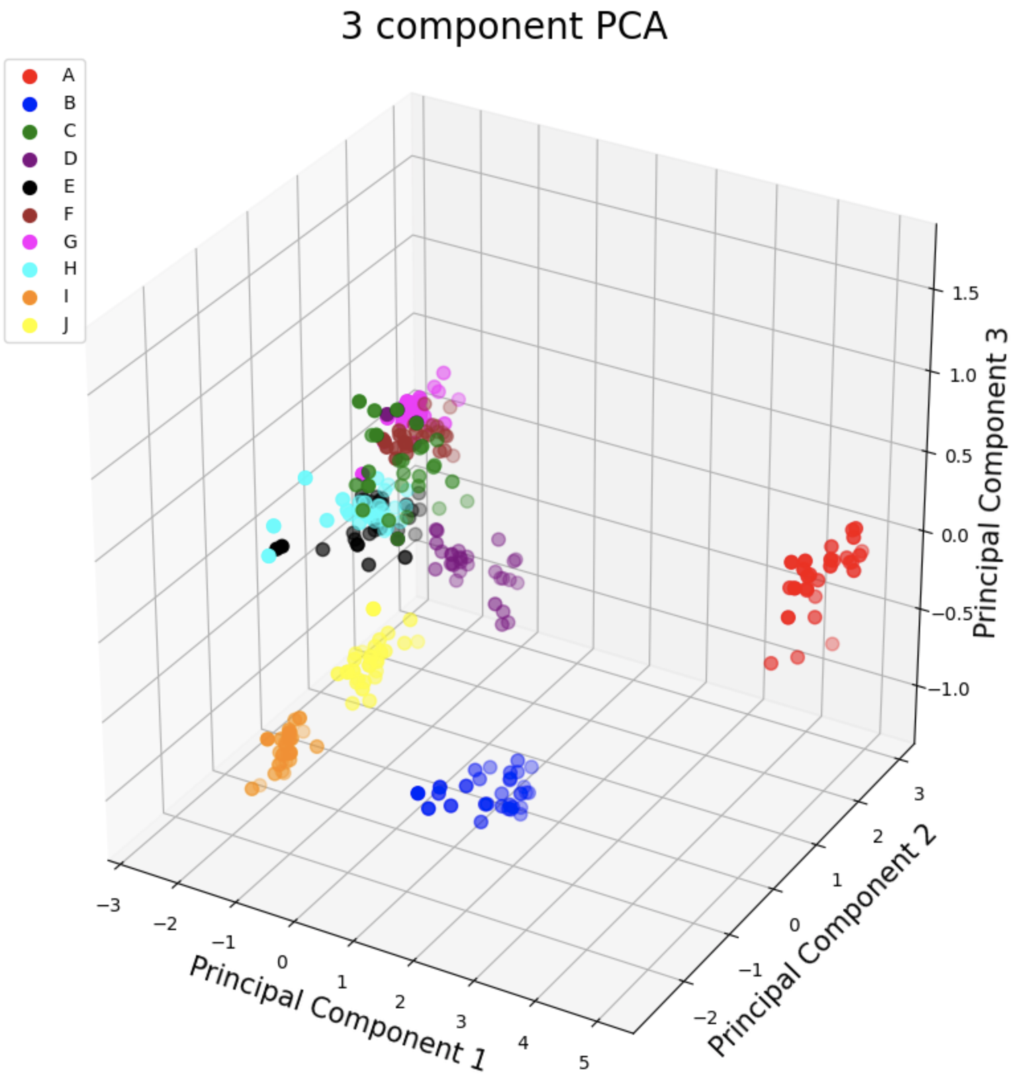

Figure 3.3: 3D Scatter plot of the first three PCs

Chapter 4

Conclusions

From the application of the PCA algorithm to the dataset, and in particular from the 3D scatter plot of the three PCs, we can say that among the ten brands of pizza, the A brand is the most different. Also, the brands B, I, and J are each part of his group. In contrast, brands C, D, E, F, G, and H are very similar. Thanks to the correlation matrix (figure 3.2) we can also add more information about how these clusters differ.

- **A**: PC1 and PC2 are both high. PC3 varies between 0 and -1;
- **B**: PC1 and PC2 are both around 0. PC3 is low;
- **I**: PC1 and PC3 are both low. PC2 is around 2;
- **J**: PC1 and PC3 are both low. PC2 is around 0. This brand is similar to the brand I with PC1 in mind;
- **CDEFGH**: PC1 is low. PC2 is high. PC3 is around 0.

Thanks to this analysis, customers can choose a brand based on their tastes. For example, any pizza from brands C, D, E, F, G, and H should taste very similar. Otherwise, if they would like to try something different, the A brand is the one more distant from the others. Finally, brands B, I, and J are different from brands A, C, D, E, F, G, and H, but they are close to each other.

Bibliography

[1] Flavia Esposito. *Slide course "Low-rank Approximation for Data Analysis: models, numerical methods and applications"*.

[2] Ian Jolliffe. “Principal component analysis”. In: *Encyclopedia of statistics in behavioral science* (2005).