THE AMERICAN UNIVERSITY IN CAIRO | SCHOOL OF SCIENCES AND ENGINEERING

DEPARTMENT OF MATHEMATICS AND ACTUARIAL SCIENCE

Principal Component Analysis

Multivariate Analysis (MACT4233-02)

Dr Ali Hadi

Spring 2023

Laila El Saadawi

900201723

Masa Tantawy

900201312

Table of Contents

Introduction 2

Data Description 3

Variables 3

Data Preparation 4

Non-Robust Principal Component Analysis 6

Robust Principal Component Analysis 7

Conclusion 8

Introduction

India’s monsoon season, typically lasting from June to September, is its wettest time of year and is an integral element of India's agriculture and economy. The rain provides water for irrigation, helps recharge groundwater supplies, moderates the climate, and reduces the risk of droughts. However, the monsoon rains can also be destructive. Heavy rains can cause flooding, landslides, crop damage, and widespread ruin. Therefore, analyzing the patterns of rainfall throughout the year using multivariate analysis techniques could facilitate the identification of relationships existing within the data, allowing for better and more accurate predictions on future rainfall events. This will in turn allow for the mitigation of the monsoon season’s risks and the harnessing of its benefits.

This project aims to apply the Principal Component Analysis technique on data about this very topic, so that an improved analysis of the variations present between the data’s multiple variables can be made through the reduction of its dimensionality. The overall objective is to uncover the relationships between the variables available in the chosen data so that correlated features can be removed, all while preserving the integrity of the data and the information it provides.

Data Description

Data Source: Kaggle

https://www.kaggle.com/datasets/rajkumarpandey02/rainfall-in-all-india-dataset-1901-2016

The dataset provides information on rainfall measurements in all India from 1901 to 2016 for the Monsoon season, which are the months June to September, and its percentage departure from normal rainfall.

Number of Observations and variables: 116 observations and 11 variables

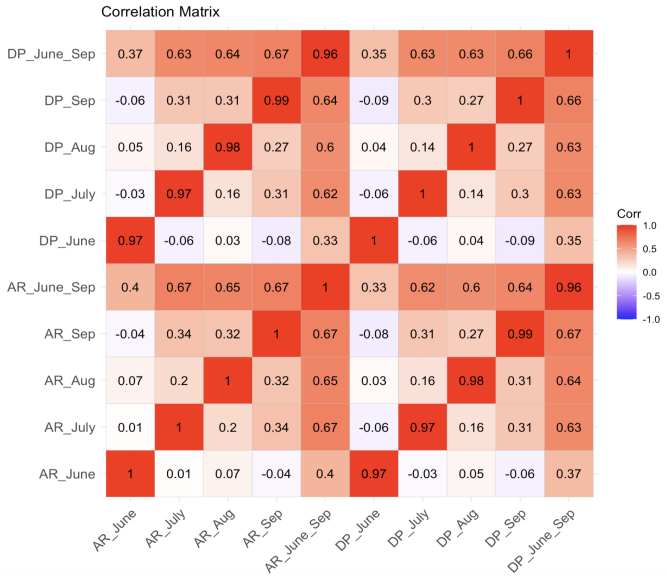
Variables

	Name	Description	Type	Units of Measurement
1.	Year	Observation Year	Categorical	(Unitless)
2.	Actual.Rainfall..JUN	Amount of rainfall during the month of June.	Quantitative	Millimeters (mm)
3.	Actual.Rainfall..JUL	Amount of rainfall during the month of July.	Quantitative	Millimeters (mm)
4.	Actual.Rainfall..AUG	Amount of rainfall during the month of August.	Quantitative	Millimeters (mm)
5.	Actual.Rainfall..SEPT	Amount of rainfall during the month of September.	Quantitative	Millimeters (mm)
6.	Actual.Rainfall..JUNE.SEPT	Total amount of rainfall from June to September.	Quantitative	Millimeters (mm)
7.	Departure.Percentage..JUN	Percentage departure of rainfall from normal during June.	Quantitative	Millimeters (mm)

8.	Departure.Percentage ..JUL	Percentage departure of rainfall from normal during July.	Quantitative	Millimeters (mm)
9.	Departure.Percentage ..AUG	Percentage departure of rainfall from normal during August.	Quantitative	Millimeters (mm)
10.	Departure.Percentage ..SEP	Percentage departure of rainfall from normal during September.	Quantitative	Millimeters (mm)
11.	Departure.Percentage ..JUNE.SEPT	Percentage departure of total rainfall from normal from June to September.	Quantitative	Millimeters (mm)

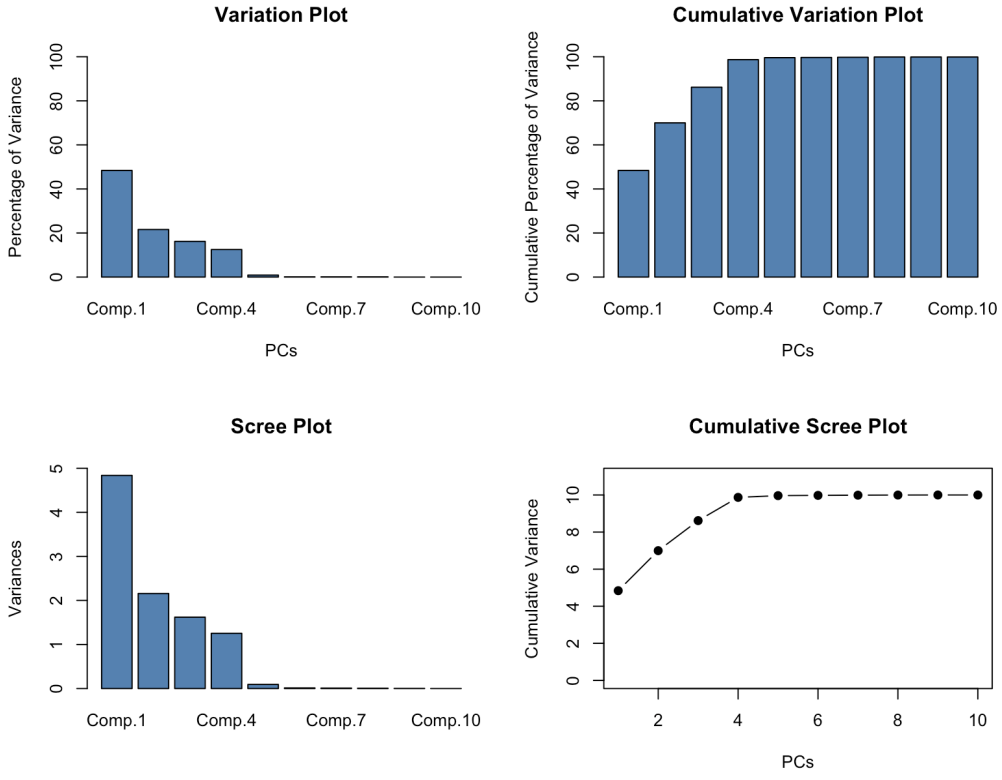
Data Preparation

The primary step is to prepare the dataset to be used. Since the aim of this project is the implementation of Principle Component Analysis, only numeric columns can be used. Hence, the column *Year* was dropped from the dataset since it is a categorical variable, and the columns were renamed into shorter names for convenience. The dataset was checked for the existence of any missing values, reaching the conclusion that the data contained no missing entries. As for outliers, 7 observations were identified as outliers by the blocked adaptive computationally efficient outlier nominators (BACON) algorithm proposed by Billor, Hadi, and Velleman (2000). These outliers will be eliminated in the Robust Principal Component Analysis section of this report. Next, the numeric columns of the dataset were scaled and the dependency structure was studied using the correlation matrix shown below. It appeared that the dataset contained high collinearity since multiple columns had high correlation, illustrated by red cells in the correlation matrix. For example, the columns actual rainfall for all the months had a strong positive correlation with the departure percentage of that month, such as *AR_June* and *DP_June*. One of the issues that PCA solves is this dependency of variables; it will be clarified that after performing PCA, the components will be orthogonal to each other thus linearly independent.



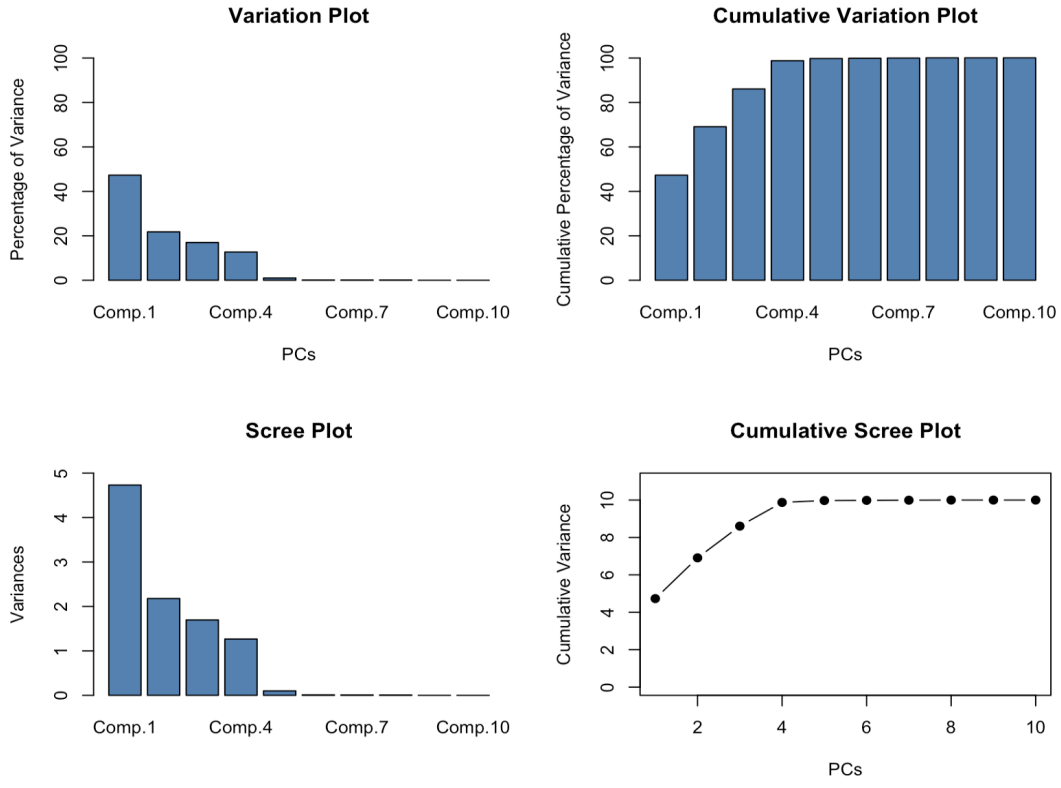
Non-Robust Principal Component Analysis

First, principal component analysis was performed on the 10 numeric columns in the dataset using all the observations in the dataset. The 4 plots below show the results of the non-robust PCA: a variation plot illustrating the percentage variation explained by each of the 10 components, a cumulative variation plot, a scree plot with the variance of each component, and a cumulative scree plot. The general rule is that the number of PCs is selected using the average lambda (lambda bar), which was identified as 1 using the covariance matrix above. Therefore, only the first 4 components are necessary to capture the maximum variance. This is because the variances of components 5 to 10 are less than 1, which indicates that their contribution to the data as individual variables is minimal. Likewise, the scree plot and the other 3 plots indicate that only the first 4 principle components should be kept, as they capture 98.7% of the data's variance, while the others should be eliminated.



Robust Principal Component Analysis

Next, robust principal component analysis was performed on the 10 numeric columns in the dataset using only the observations that were not identified as outliers by BACON. The 4 plots below are the results of the robust PCA after removing the 7 outliers from the dataset. The output did not vary compared to the non-robust PCA. According to the rule where the number of PCs is chosen using the average lambda (lambda bar), which is 1 according to the correlation matrix, the first 4 components are sufficient for capturing the maximum variance. This conclusion is illustrated by the scree and variance plots. These 4 components will explain 98.7% of the variation in the dataset.



Conclusion

Overall, PCA is a powerful tool that can be used to improve the understanding and analysis of data through the elimination of variables with little to no variance so that the data's dimensionality can be reduced and its observations better visualized. By applying PCA on the Indian Monsoon dataset, the dimensions of the data were reduced from 10 variables to only 4. These four components capture the vast majority of the data's information due to their high variance and low collinearity. They were thus made to be orthogonal to each other and were made to represent the data's principal components. The remaining six variables had very high collinearity (low variance) which meant that they were too dependent on one another. The information that they could capture from the dataset as individual variables was rendered as of little use, which justifies their consequent elimination.