CHAPTER 1

Principal Components and Linear Algebra

OBJECTIVE

This chapter provides a short review of linear and matrix algebra concepts that are relevant in the study of principal components. I do believe that without an in-depth understanding of the notions of coordinate systems and change of basis, it is near impossible to have a good grasp of principal components. Therefore, part of this chapter is devoted to showing how a dataset can be graphically represented in different coordinate systems. For researchers who like to understand how principal components are actually calculated, I introduce some key concepts in matrix algebra, such as matrix similarity, characteristic polynomials, eigenvalues, eigenvectors and the method of Lagrange multipliers. You are not required to have any specific prerequisite to find this chapter useful since I start with the simplest examples before moving to the more complex ones. However, any prior knowledge of linear and matrix algebra will help.

Contents

"The man who grasps principles can successfully select his own methods. The man who tries methods, ignoring principles, is sure to have trouble."
Ralph Waldo Emmerson (May 25, 1803 - April 27, 1882)

1.1 Introduction

Principal Component Analysis (PCA) is recommended if you want to analyze multivariate correlated data. If a single variable is measured, then you can rank all subjects according to that variable, get a first look at your data and learn something about your subjects. If 2 variables are measured, then a basic exploratory data analysis based on a scatter plot can provide you with a practical and effective visual description of your data. Beyond 2 variables, our problem becomes three-dimensional and our view of the data structure blurry. Extracting a useful story from our data becomes a challenging task. PCA is a statistical technique that consists of taking a large set of variables of interest and narrow it down to a smaller and conceptually more coherent set of variables called the "Principal Components."

Depending upon your interest and the extent to which you want to understand the underlying mathematics, the study of principal components can be more or less time-consuming. I decided to start with the simplest case involving only 2 variables, where PCA is not needed, before moving on to the more complex situation where its use becomes essential. The rationale behind this approach is that the mathematics necessary for 2 variables are simple and can be used as a stepping stone for the study of multivariate data, which requires some linear and matrix algebra.

To have a glimpse into what PCA does, consider the measurements associated with 2 variables X_1 and X_2 taken on 12 subjects and shown in Table 1.1. If X_1 is the only variable of interest to you, then you can rank all 12 subjects by X_1 or compute summary statistics such as the mean, the median, the standard deviation and others. These indicators will give you a good first look at your dataset. If both variables are of interest, then your data can be depicted as shown in figure[1] 1.1. You can see a positive correlation between the 2 variables indicating that as variable X_1 increases, so is variable X_2. The rate at which

[1] Note that figure 1.1 is based on a larger dataset of 50 subjects from which Table 1.1 was extracted. The larger data can be downloaded in CSV format using the following link: https://agreestat.com/books/princomp/datasets/x1x2data.csv

one of the 2 variables increases with respect to the other can be approximated with basic linear regression techniques.

Table 1.1: Measurements of 2 variables X_1 and X_2 taken on 12 subjects

Subject	X_1	X_2
1	6.15	7.77
2	7.12	6.62
3	7.37	7.78
4	7.43	8.45
5	7.51	8.29
6	7.57	7.91
7	7.64	8.73
8	7.76	8.52
9	7.81	7.96
10	7.92	8.48
11	7.95	8.74
12	8.02	8.38

Now, what do you do if you want to characterize each subject with respect to both variables X_1 and X_2? If these scores were grades that a teacher assigned to students in 2 exams, then you could average both grades associated with each student and obtain a composite grade for the class before using it to rank the students. Alternatively, if the grades are for an interim and a final exams, then the final could receive a higher weight than the interim, which will result in a weighted composite score for each student. How you weight both series of scores in this situation is entirely up to you. Any technique you use has the same objective, which is to transform 2 scores into a single score to simplify analysis. This is one of the simplest examples of dimensionality reduction. By creating a single composite score, you reduce a two-dimensional problem to a one-dimensional problem.

You will not always have the option of selecting your own dimensionality-reduction method. Suppose you are an economist attempting to discover the key determinants of economic growth, that the subject is a country and the 2 variables are the country inflation rate and its female fertility rate. Then you can no longer arbitrarily weight both variables. Instead, you need to compute a linear combination of the 2 variables that will discriminate the different countries the most. This is precisely what Principal Components Analysis does. It takes both variables and derive a single variable (the principal component) that best captures the variability in your data. There will generally be a smaller lost of variability (hopefully negligible) captured by a second principal component that we will be able to safely ignored.

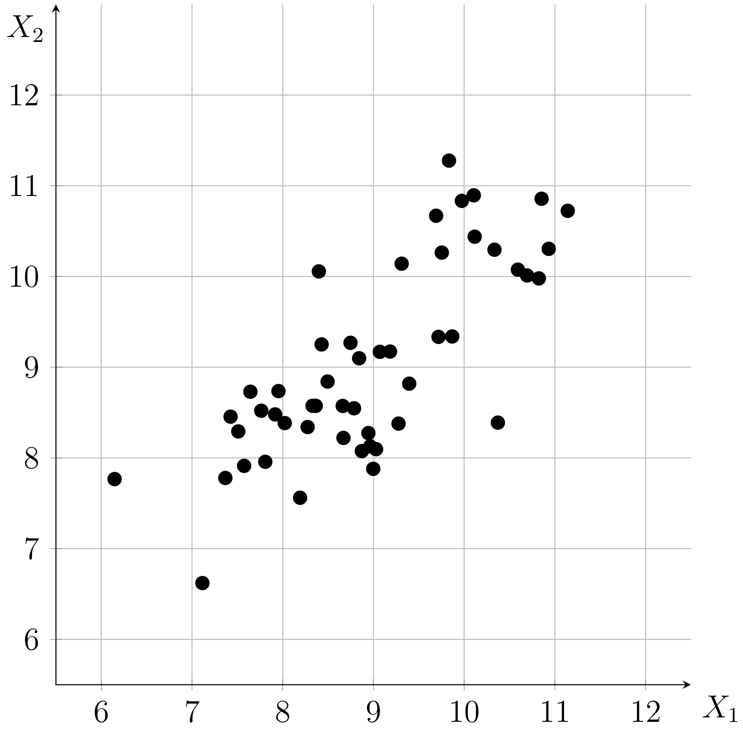

Figure 1.1: Scatter plot of 2 variables X_1 and X_2 based on 50 observations

Consider Figure 1.2 where the same 50 subjects of Figure 1.1 are plotted on a different coordinate system defined by the 2 principal components PC_1 and PC_2 (I will later discuss how these principal components are calculated). It follows that the PC_1 axis alone approximates the set of observations better than any one of the initial axes X_1 or X_2. That is, the orthogonal projection of all observations onto the line along PC_1 (see Figure 1.3) produces more variation than the orthogonal projection of the same observations onto any other axis conceivable. Consequently, your data analysis along the PC_1 axis alone will still tell you more of what you need to know about your data than any other one-dimensional coordinate system.

The coordinates of the 50 data points in the new coordinate system are on a different scale when compared to their representation in the standard coordinate system of Figure 1.1. The data location is different as well. However, you can always shift that location at will by adding a constant value to the coordinates. But it is not recommended to rescale these coordinates as such an operation will jeopardize the statistical properties of the principal components. I will further discuss about these issues in subsequent chapters.

The number of principal components is always the same as the number of variables. The main advantage of principal components lies in that they are uncorrelated and they are ranked in descending order of the proportion of total variance each of them can explain. Therefore, the bigger proportion of total variance is concentrated on the first few principal components, making it possible to built a sound analysis on a small set of uncorrelated variables that are easier to interpret.

Many researchers want to have some understanding of the way principal components are constructed. My intend is to provide a detailed description of the building blocks in this process. However, this will require some introduction to matrix algebra. Matrix algebra is not as important when dealing with 2 variables, as it is when dealing with a large number of variables. This introductory note to matrix algebra is provided in section 1.2, where I review the important link between matrix algebra and the resolution of systems of linear equations. The important method of Lagrange multipliers for finding solutions to optimization problems will be reviewed in section 1.3.

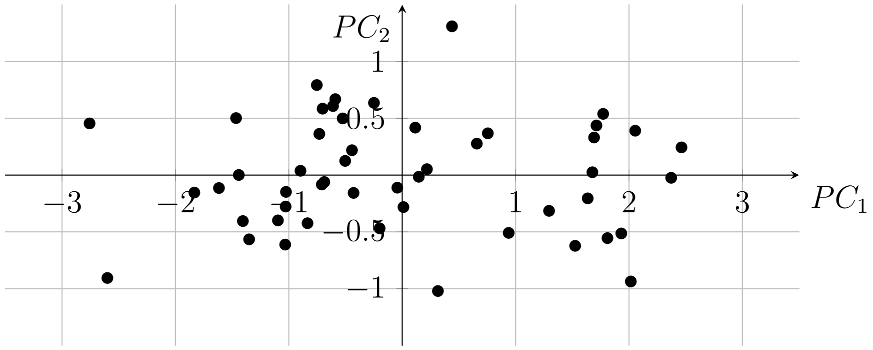

Figure 1.2: Scatter plot of the 2 principal components PC_1 and PC_2 based on the same 50 observations of figure 1.1

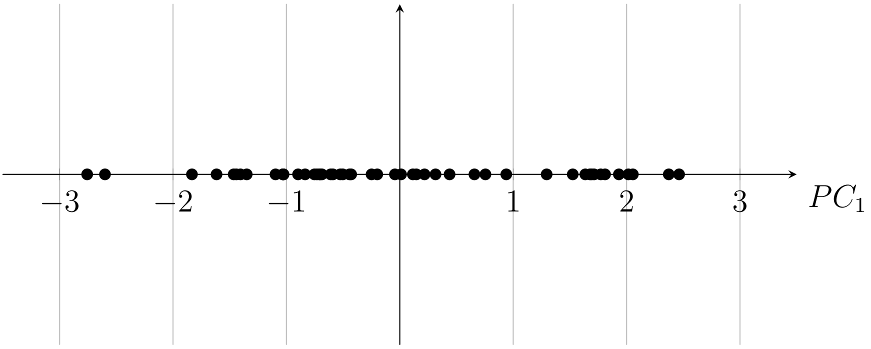

Figure 1.3: Plot of the first principal component PC_1 based on the same 50 observations of figure 1.1

1.2 Basic Matrix Algebra

The resolution of systems of linear equations is essential in the method of Lagrange multipliers, which is needed to compute principal components. Before I tackle the method of Lagrange multipliers in section 1.3, I am going to provide a brief introduction to systems of linear equations, matrices and their use in linear transformations and in defining new coordinate systems. Even if you have previously been exposed to these notions in a linear algebra class, note that the emphasis here is less on the underlying mathematics and more on their interpretation in the context of multivariate data analysis.

Systems of linear equations occur in various contexts in the real world. Consider the following problem:

> The admission fee at a small fair is \$1.50 for children and \$4.00 for adults. On a given day 2,200 people enter the fair and \$5,050 is collected. *How many children* and *how many adults attended*?

Let me define 2 variables x = “Number of children” and y = “Number of adults.” These 2 variables are suggested by the very question at the end of the problem statement. No need to look far for what the 2 unknowns should be. The fact that 2,200 people have entered the fair translates mathematically into the equation $x + y = 2{,}200$. Since the total revenue is \$5,500 and has been generated by the children and the adults, I can formulate this mathematically as