CHAPTER 1

Basics of Principal Components

OBJECTIVE

This chapter provides a high-level presentation of the "true" nature of Principal Component Analysis (PCA). The objective is to allow you to develop a global understanding of what PCA does and what you can use it for. You will see what it means to reduce your data dimensionality without getting into technical details. A special emphasis will be put on carefully defining new terms, which will be used throughout this book.

Contents

1.1 Introduction

Principal Component Analysis (PCA) is a popular dimensionality reduction technique, which allows you to reduce the analysis of a large number of variables to only a few key indicators without losing much information. Sometimes, analyzing a large number of variables would be impossible without the use of principal components. Therefore, developing a basic understanding of principal components is essential in this era of big data.

You cannot master PCA in one sitting. It is because, the material is technical and some experience/practice is required to become good at it. Just like operating a car is clearly not rocket science. However, nobody expects you to become an excellent driver the first day you put your hands on the car steering wheel. Researchers studying PCA techniques may aim at one of the following 2 objectives:

- You want to be able to implement PCA and interpret the results. After presenting a clear and general overview of the very nature of PCA, this book will help you achieve this goal using the R statistical software.

- Alternatively, you want to develop an in-depth understanding of why PCA works. This objective requires a closer look at the algorithms that PCA is based upon. These algorithms are based on complex linear/matrix algebra. If you fall into this category, then you could read my other book entitled "Beginner's Guide to Principal Components / Applications with Microsoft Excel" (see Gwet, 2020). In that book, I present the algebra underlying PCA, and attempt to show step by step, how the calculations are carried out with the help of Excel.

I assume in this book that you have never been exposed to R nor to RStudio's Integrated Development Environment (IDE) for R. In chapter 2, you will acquire the minimum knowledge necessary to perform PCA with R using RStudio's IDE. If you are already familiar with R, feel free to skip chapter 2, or refer to it only when necessary.

The good news about the R software is that it is free, and available in different operating systems including MS Windows, Linux, MacOS. Moreover, R offers a cloud version, which requires that you remain connected to the Internet at all times, but does not require any installation.

1.2 A Glimpse into Principal Component Analysis

Consider the small dataset of Table 1.1. It shows information about 12 human subjects on 2 dimensions named X_1 and X_2. This two-dimensional dataset is depicted in the scatterplot of Figure 1.1, where you can see a seemingly linear structure in the relationship between both variables.

Suppose that you want to analyze the 12 subjects on a single dimension, perhaps because you want to be able to rank them by the magnitude of that dimension. What are you going to do? One option would be to simply ignore one of the 2 variables (e.g. X_2) and use X_1 as the sole dimension for your analysis. This option is bad. Figure 1.1 shows substantial variation in the data along both dimensions X_1 and X_2. Ignoring one of these 2 dimensions will result in a substantial loss of information. Variation along one dimension occurs for a reason, which is to be understood.

The better approach to this dimensionality reduction problem requires the use of principal components. The data points in Figure 1.1 are displayed in the Cartesian coordinate system. The 2 vectors $\vec{\imath} = (1, 0)$ and $\vec{\jmath} = (0, 1)$ represent the basis of this coordinate system, whereas the 2 data series X_1 and X_2 of Table 1.1 are the coordinates of the 12 subjects in the $(\vec{\imath}, \vec{\jmath})$ basis. The Principal components (PC) represent a pair of vectors $(\vec{u}, \vec{v})$, which form the basis of a new and special coordinate system shown in Figure 1.2. This is the coordinate system of "Principal Components" as opposed to the more traditional Cartesian coordinate system.

What is it that makes the new coordinate system of principal components so special? Note in Figure 1.2 that most of the variation in your dataset occurs along the axis associated with the $\vec{u}$ principal component. The variation along the $\vec{v}$ axis is smaller. If you are going to retain a single dimension for your dataset, it must be the $\vec{u}$ principal component axis, as it minimizes the loss of information.

To successfully conduct a PCA, you need to overcome the following 2 challenges:

- Find the coordinates of the 2 principal components[1] $\vec{u}$ and $\vec{v}$, expressed in the Cartesian coordinate system, the first principal component representing the axis with maximum variation. For Table 1.1 data, the first

[1]The number of principal components equals the number of dimensions in your dataset. In our example, there are 2 dimensions X_1 and X_2 and therefore 2 principal components.

PC can be written as $\vec{u} = 0.707\vec{\imath} + 0.707\vec{\jmath} = (0.7070, 0.707)$, whereas the second PC would be written as $\vec{v} = 0.707\vec{\imath} - 0.707\vec{\jmath} = (0.7070, -0.707)$. You will learn in chapter 3 how these coordinates are calculated.

- Translate the original data points from the Cartesian coordinate system into the new coordinate system of principal components. These coordinates will be referred to as the *Principal Component Scores* or the *Scores*. Then conduct your analysis using PC scores as opposed to original data. In Table 1.1 for example, subject 1's coordinates in the PC coordinate system are $(9.8384, -1.1457) = 9.8384\vec{u} - 1.1457\vec{v}$. Typically, only the scores associated with the first PC $\vec{u}$ would be used in the analysis as a way of reducing the data dimensions from 2 to 1. Again, you will see in chapter 3 how these coordinates are calculated.

Table 1.1: Two Measurements X_1 and X_2 taken on 12 human subjects[a]

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

[a]This small table is an extract of a larger dataset of 50 subjects, which is accessible via the link https://bit.ly/3keldt5

A close look at Figure 1.2 makes the basis $(\vec{u}, \vec{v})$ of the coordinate system of principal components look as if it was obtained by rotating the Cartesian basis $(\vec{\imath}, \vec{\jmath})$, counterclockwise. In this sense, the word "rotation" is often used in R outputs, which are related to PCA.

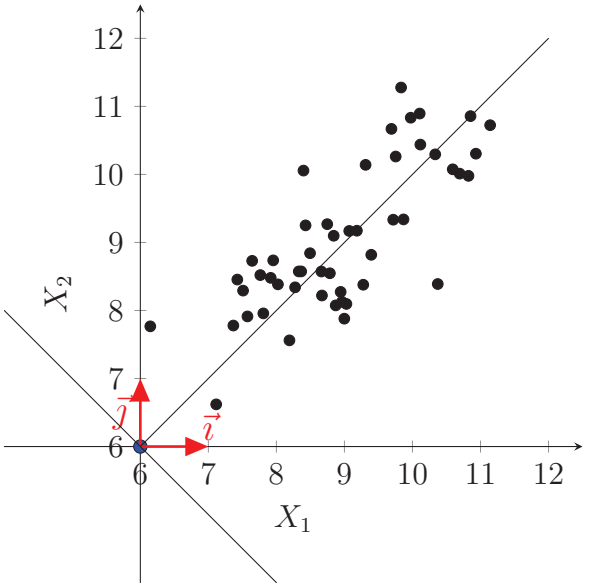

Figure 1.1: Scatter plot of the two-variable dataset of Table 1.1 in the initial $(\vec{\imath}, \vec{\jmath})$ cartesian coordinate system

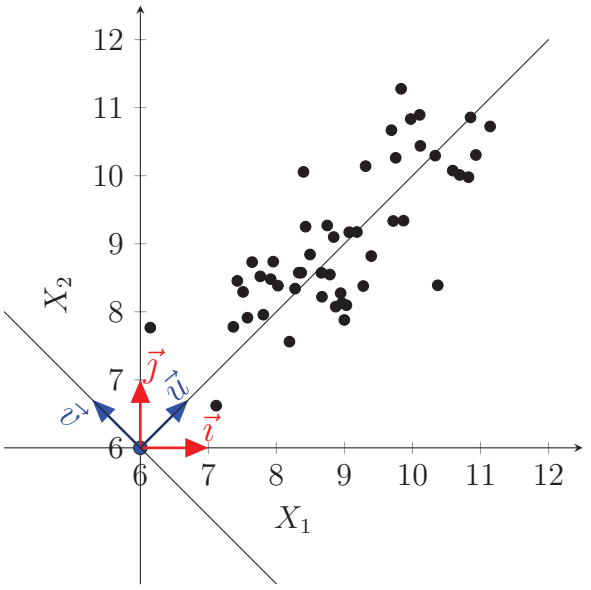

Figure 1.2: Scatter plot of Table 1.1 data in the Cartesian system $(\vec{\imath}, \vec{\jmath})$, and in the system of principal components $(\vec{u},\vec{v})$

1.3 Principal Components and Associated Scores

The 2 principal components associated with the 2-dimensional dataset of Table 1.1 are given in Table 1.2. In many statistical packages, principal components are labeled as PC1 and PC2 as opposed to $\vec{u}$ and $\vec{v}$. The numbers in this table represent the coordinates of the 2 components in the Cartesian coordinate system, and were used to display $\vec{u}$ and $\vec{v}$ in blue in Figure 1.2. These numbers are often referred to as Principal Component Coefficients (or PC coefficients). Once again, you can see that the $\vec{u}$ axis carries more information (i.e. more variation) than the $\vec{v}$ axis. Therefore, you can focus your analysis on the first principal component $\vec{u}$ and drop the second one $\vec{v}$, without losing too much information. It is how you reduce a two-dimensional problem to a one-dimensional problem.

Table 1.2: Principal Components $\vec{u}$ and $\vec{v}$ associated with 1.1 data

	$\vec{u}$	$\vec{v}$
X_1	0.70711	0.70711
X_2	0.70711	-0.70711

Now, you must translate all of your data points of Table 1.1 from the Cartesian system to the new and more useful coordinate system of principal components. These new coordinates, also known as Principal Component "Scores" (PCS or PC scores), or simply scores, are given in Table 1.3. From now on, your data analysis should be based on this new dataset. If you want to perform a one-dimensional analysis based on a single variable, then PCS_1 is it.

An interesting question to be asked is whether there is any link between Table 1.3 data and the initial Table 1.1 data. The answer is yes, there is. Consider Subject 1's coordinates $(6.15, 7.77)$ in the Cartesian coordinate system. These coordinates can be written as,

$$\begin{aligned} \texttt{Subject1} = (6.15, 7.77) &= (6.15, 0) + (0, 7.77), \\ &= 6.15\,(1, 0) + 7.77\,(0, 1), \\ &= \mathbf{6.15}\,\vec{\imath} + \mathbf{7.77}\,\vec{\jmath}, \end{aligned}$$

where $\vec{\imath} = (1, 0)$ and $\vec{\jmath} = (0, 1)$ form the basis of the Cartesian coordinate

system. These same coordinates can be rewritten as follows:

$$\begin{aligned}\texttt{Subject1} = (6.15, 7.77) &= (6.957 - 0.810, 6.957 + 810),\\ &= (6.957, 6.957) + (-0.810, 810),\\ &= 9.8384\,(0.7071, 0.7071) - 1.1457\,(0.7071, -0.7071),\\ &= \mathbf{9.8384}\,\vec{u} - \mathbf{1.1457}\,\vec{v},\end{aligned}$$

where $\vec{u} = (0.7071, 0.7071)$ and $\vec{v} = (0.7071, -0.7071)$ are the principal components that form the basis of the new coordinate system. You can now see the 2 coordinates 9.8384 and -1.1457 of subject 1, also shown in the first row of Table 1.3.

Therefore, principal component analysis does not change subject 1. Instead, it changes the way you look at it, by using a different system of coordinates.

Table 1.3: Two Series of Principal Component Scores PCS_1 and PCS_2 Associated with Variables X_1 and X_2 of Table 1.1[a]

Subject	PCS_1	PCS_2
1	9.8384	-1.1457
2	9.7123	0.3500
3	10.7099	-0.2909
4	11.2288	-0.7276
5	11.1737	-0.5529
6	10.9511	-0.2390
7	11.5777	-0.7678
8	11.5143	-0.5352
9	11.1476	-0.1057
10	11.5939	-0.3980
11	11.8015	-0.5541
12	11.6011	-0.2558

[a]The complete series of 50 principal component scores is accessible via the link http://bit.ly/3XIrv1R

You can now display your data points again on the new coordinate system of principal components as shown in Figure 1.3.

You will often see the term "Principal Component" used in the literature to refer to both the basis vectors and the principal component scores, making

you wonder what exactly the principal component is. The basis vectors $(\vec{u}, \vec{v})$ of a new coordinate system are the principal components. The coordinates of your data points in the new coordinate system are the "Principal Component Scores." Generally, the context will often tell you what the correct term should be. This confusion should not be a problem as long you understand how things work.

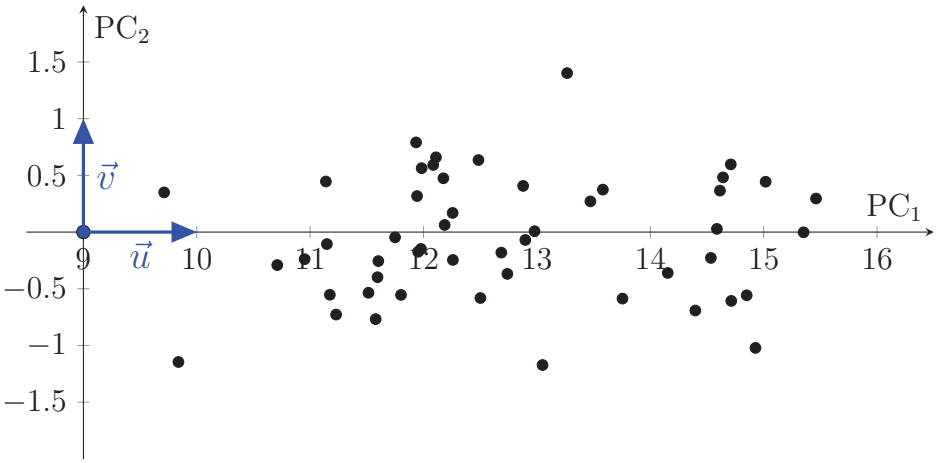

Figure 1.3: Scatterplot of a two-variable dataset in the alternative $(\vec{u}, \vec{v})$ coordinate system of principal components.

1.4 A More Realistic PCA Problem

In this section, I will dig a little deeper into the notion of principal component, and further explore the different aspects of Principal Component Analysis, and its importance in the broader field of exploratory data analysis. However, the more technical task of actually computing the computing the principal components and associated scores using R, is addressed in chapter 3.

Consider Table 1.4, which contains a dataset with 8 variables that describe different aspects of 11 of the 50 States that make up the United States of America. Here are the 8 variables:

- `Population:` population estimate as of July 1, 1975.
- `Income:` per capita income (1974).