CHAPTER 1

An Overview of Principal Components

OBJECTIVE

This chapter provides a high-level overview of Principal Component Analysis (PCA). It is a statistical technique used for analyzing multivariate, correlated data by transforming the original variables into new uncorrelated indicators called Principal Component Scores (PCS). You will learn that these indicators are ordered based on the proportion of total variation they capture, allowing the first few components to represent the majority of the data's variability. PCA effectively reduces dimensionality by replacing multiple correlated variables with fewer composite scores, making it easier to interpret complex datasets. You will also see through a geometric approach that PCA rotates the coordinate system to align the data with the principal components, facilitating a simplified representation while retaining critical information about the dataset.

Contents

1.1 *Introduction* 3
1.2 *Illustrating Composite Score Variables* 5
1.3 *Geometric Framework for Principal Component Analysis* 11
1.4 *Concluding Remarks* 14

1.1 Introduction

Principal Component Analysis (*PCA*) is recommended for analyzing multivariate, correlated data. When each unit of analysis has a large number of quantitative attributes, it becomes nearly impossible to extract meaningful insights from the dataset. Key questions arise: Which units are similar? Are some attributes more relevant than others? Are there outliers in the dataset that should be discarded? These are some of the critical questions an analyst seeks to answer.

If the units of analysis are to be examined using a single variable, then you are dealing with a one-dimensional problem. With two attributes, the problem becomes two-dimensional. In one-dimensional problems, you can conveniently rank all units along that dimension to get an initial understanding of your data and gain insights about the units of analysis. For two-dimensional problems, you can conduct basic exploratory data analysis using a scatter plot of the two variables, providing a practical and effective visual representation of your data. However, beyond two variables, the problem becomes three-dimensional, making the data structure difficult to interpret. In real-world scenarios, you will often encounter multi-dimensional problems involving dozens of variables. Extracting a coherent story from the data becomes a challenging task. This is where PCA proves valuable.

PCA is a statistical technique that involves taking a specific number of variables and replacing them with an equivalent number

of new indicators called *Principal Component Scores* (PCS). These scores act as surrogates for the original attributes and possess the following properties:

- The new indicators can be ranked in descending order according to the percentage of total variation in the data they account for.[1] The first few indicators are expected to account for most of the variation in the dataset.
- The new indicators are uncorrelated, meaning there is no redundancy in the information each indicator provides. In other words, each indicator reveals a unique aspect of the dataset.
- Collectively, the new indicators must account for 100% of the total information contained in the original dataset. However, the concentration of information in the first few indicators allows for a reduction in the problem's dimensionality.

Various terms are used in the literature to refer to Principal Component Scores. They are often called "Composite Scores," "Composite Score Variables," or "Composite Score Indicators." Some texts may mistakenly refer to them as "Principal Components." In reality, the term "Principal Component" refers to something else, which I will discuss in detail later in this chapter. Personally, I prefer the terms "Composite Score Variable" or "Composite Score Indicator," as any value taken by these variables is a composite score.

[1] Note that the amount of information contained in a dataset is measured by the variation that exists across the different units of analysis.

1.2 Illustrating Composite Score Variables

To gain an understanding of what Principal Component Analysis (PCA) does, let's consider the measurements of two variables, X and Y, taken on 12 units, as shown in Table 1.1. The correlation coefficient[2] between X and Y is 0.76, indicating a strong positive linear relationship. Thus, the information carried by these two variables is largely redundant.

If you are only interested in X, you could rank the 12 units based solely on their X values and divide them into more homogeneous groups. However, if both variables are of interest, the problem becomes two-dimensional. The data can be represented by a scatter plot, as shown in Figure 1.1. From the plot, it is evident that as X increases, Y tends to increase as well, confirming the positive correlation.

This raises a few important questions:

> Can you still group the 12 units into more homogeneous clusters, where homogeneity is defined by both X and Y? Is it possible to rank the units based on their X and Y values simultaneously? If so, what metric would be appropriate for this ranking?

In the rest of this section, I will explore how composite scores appear in this simple scenario and how they can be effectively used. Additionally, I will discuss the technical challenges that may arise when calculating these composite scores.

[2] A correlation coefficient quantifies the degree of linear relationship between two variables.

Table 1.1: Measurements of 2 variables X and Y taken on 12 units[a]

Unit	X	Y
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]Note that these are the first 12 observations of a larger dataset of 50 units. The larger dataset can be downloaded in CSV format using the following link: https://agreestat.com/books/knn/datasets/xydata.csv

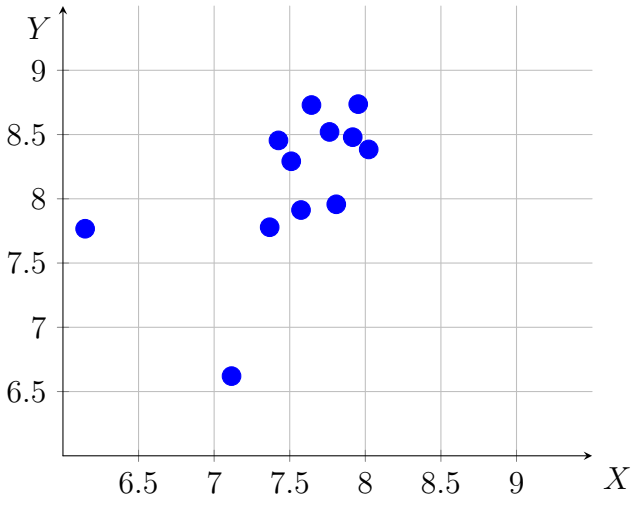

Figure 1.1: Scatter plot of the 2 variables X and Y of Table 1.1

If the X and Y values represent grades assigned by a teacher for two exams, you could calculate the average of both grades for each student to obtain a composite score, which can then be used to rank the students. Alternatively, if the grades are for a midterm and a final exam, you might assign a higher weight to the final exam, resulting in a weighted composite score for each student. The choice of how to weight the scores is entirely up to you. Regardless of the method used, the goal remains the same: to combine two scores into a single value for easier analysis. This is a simple example of dimensionality reduction. By creating a composite score, you reduce a two-dimensional problem to a more manageable one-dimensional format.

When selecting a dimensionality-reduction method, you may not always have the luxury of choosing your approach. For instance, consider an economist trying to identify key determinants of economic growth, with countries as the unit of analysis. Suppose the two variables of interest are the inflation rate and the female fertility rate (the average number of children per woman). In this scenario, it's impractical to weight these variables arbitrarily, as they represent distinct factors.

A solution is to compute a linear combination of these variables that maximizes their ability to distinguish between countries. Principal Component Analysis (PCA) addresses this need by transforming the original variables into two composite variables: the first captures most of the variation in the data, while the second captures the remaining marginal variation. By retaining the first composite variable, you can effectively summarize the

dataset with minimal loss of information.

I conducted a Principal Component Analysis (PCA) on the data presented in Table 1.1, resulting in two composite score variables, PC_X and PC_Y, as shown in Table 1.2. Each composite score variable is a linear combination of the original variables X and Y, calculated as follows:

$$\begin{aligned} \text{PC}_X &= 0.60643 \times X + 0.79514 \times Y \\ \text{PC}_Y &= 0.79514 \times X - 0.60643 \times Y \end{aligned} \tag{1.2.1}$$

For each individual unit, PC_X is computed by multiplying the X value by 0.60643 and the Y value by 0.79514, then summing the two products. Similarly, PC_Y is calculated using the same procedure as outlined in Equation 1.2.1. Note that PC_X represents the linear combination with the highest variance, while PC_Y captures the second highest variance.

One of the main challenges in conducting Principal Component Analysis (PCA) is determining the linear coefficients in Equation 1.2.1. In Chapter 2, I will demonstrate how to compute these coefficients using Excel. However, a thorough understanding of the computational procedure for deriving these coefficients typically requires knowledge in linear algebra, matrix algebra, and calculus, including concepts such as Lagrange multipliers. Claims of performing this analysis without the use of algebra or calculus should be approached with skepticism, as they may lack the necessary rigor.

If you are serious about gaining a thorough understanding of

the mathematics underlying PCA[3], I highly recommend starting with Singh (2014). Titled "Linear Algebra: Step by Step," this book is one of the most valuable resources on linear algebra that I have encountered. For a discussion on applying linear algebra and calculus to PCA with practical Excel implementation, refer to Gwet (2020). Additionally, Cohen (2022) offers a gentle introduction to linear algebra specifically tailored for data science.

In Section 1.3, I will provide a brief overview of the geometric interpretation of PCA to offer further insights into the topic. For a detailed discussion on how the geometric problem translates into linear algebra and calculus, consult Gwet (2020).

Table 1.2 demonstrates that the two composite variables are uncorrelated, indicating that each conveys unique information. Notably, the first composite variable, PC_X, explains 77.6% of the total variation in the data. As a result, analysis of all 12 units can be effectively reduced to a single dimension by relying solely on PC_X. In contrast, the original variables, X and Y, are highly correlated, meaning they provide redundant information. Consequently, selecting either X or Y individually would not yield meaningful insights, as their information cannot be usefully disentangled.

Figure 1.2 presents a scatter plot of the two composite variables. Aside from the two data points labeled 1 and 2, the dataset shows minimal variation along the PC_Y axis. Most of the variation is concentrated along the PC_X axis.

[3]This depth of understanding is not necessary for most practitioners who are focused on applying PCA to real-world problems.

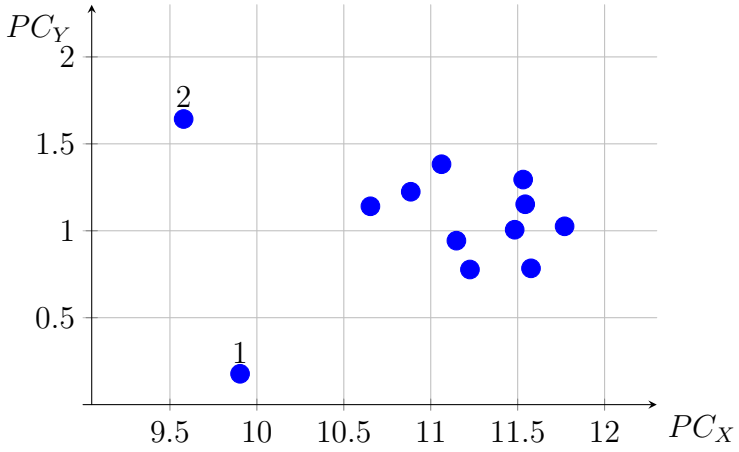

Figure 1.2: Scatterplot of the composite variables PC_X and PC_Y of Table 1.2

Table 1.2: Measurements of 2 variables X and Y taken on 12 units

Unit	X	Y	PC_X	PC_Y
1	6.147	7.767	9.903	0.177
2	7.115	6.620	9.579	1.643
3	7.367	7.779	10.653	1.141
4	7.425	8.454	11.225	0.777
5	7.510	8.292	11.148	0.943
6	7.575	7.913	10.885	1.224
7	7.644	8.730	11.577	0.784
8	7.763	8.520	11.483	1.006
9	7.808	7.957	11.062	1.383
10	7.917	8.480	11.543	1.153
11	7.953	8.737	11.770	1.026
12	8.022	8.384	11.532	1.294
Variance	0.258	0.346	0.469	0.136
Total Variance		0.605		0.605
% Total Variance	42.7%	57.3%	77.6%	22.4%
Correlation		0.972		0

1.3 Geometric Framework for Principal Component Analysis

Consider, for example, the data presented in Figure 1.3. Each data point is represented by its coordinates in the original coordinate system, defined by the natural basis $(\vec{\imath}, \vec{\jmath})$. It is important to note that these data points seem to vary along the horizontal $\vec{\imath}$ axis almost as much as they do along the vertical $\vec{\jmath}$ axis. As a result, both axes are essential for a thorough analysis of the dataset.

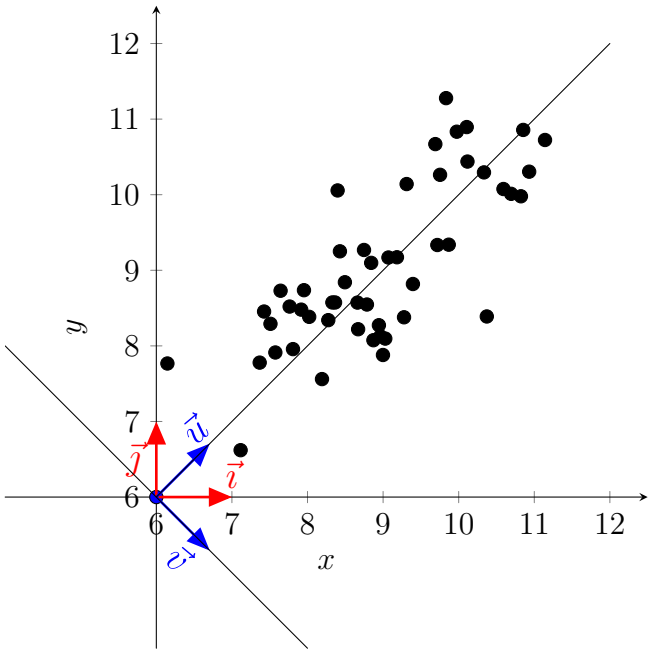

Figure 1.3: Scatter plot of a two-variable dataset in the initial $(\vec{\imath}, \vec{\jmath})$ coordinate system

Now, consider a new coordinate system defined by a basis of two vectors, $\vec{u}$ and $\vec{v}$. From Figure 1.3, it is evident that the data points vary much more along the $\vec{u}$ axis than along the $\vec{v}$ axis. Con-

sequently, if you wish to examine the data in a one-dimensional space using a single axis, you could focus solely on the $\vec{u}$ axis, which captures a significant portion of the data's variation. In doing so, you effectively reduce the initial two-dimensional analysis problem to a one-dimensional one. This is the fundamental idea behind dimensionality reduction techniques.

> If the basis of the new coordinate system maximizes the variation of the data along one axis, then the vectors $\vec{u}$ and $\vec{v}$ will be referred to as *principal components*. The coordinates of your data points in this new system, defined by the principal components, are known as the principal component scores, or composite scores.

As shown in Figure 1.3, if you project your data points onto the $\vec{u}$ axis, you will obtain the first principal component (PC) scores. Similarly, projecting onto the $\vec{v}$ axis will yield the second PC scores.

Next, you can display your data points using their coordinates in the new coordinate system defined by the principal components, as illustrated in Figure 1.4. This figure is produced by performing an approximately 30° clockwise rotation of the initial $(\vec{u}, \vec{v})$ axes.

Let me clarify a common source of confusion regarding the term "Principal Component" in the literature. A coordinate system is defined by its basis vectors. When basis vectors, such as $(\vec{u}, \vec{v})$, meet certain criteria (e.g. orthogonality, maximal data variation in one direction,...), they are referred to as principal components. The coordinates of your data points in this new coordinate system

are known as "Principal Component Scores."

However, the term "Principal Component" is often used to refer to both the basis vectors and the principal component scores, which can lead to ambiguity. In most cases, the context will clarify the correct meaning.

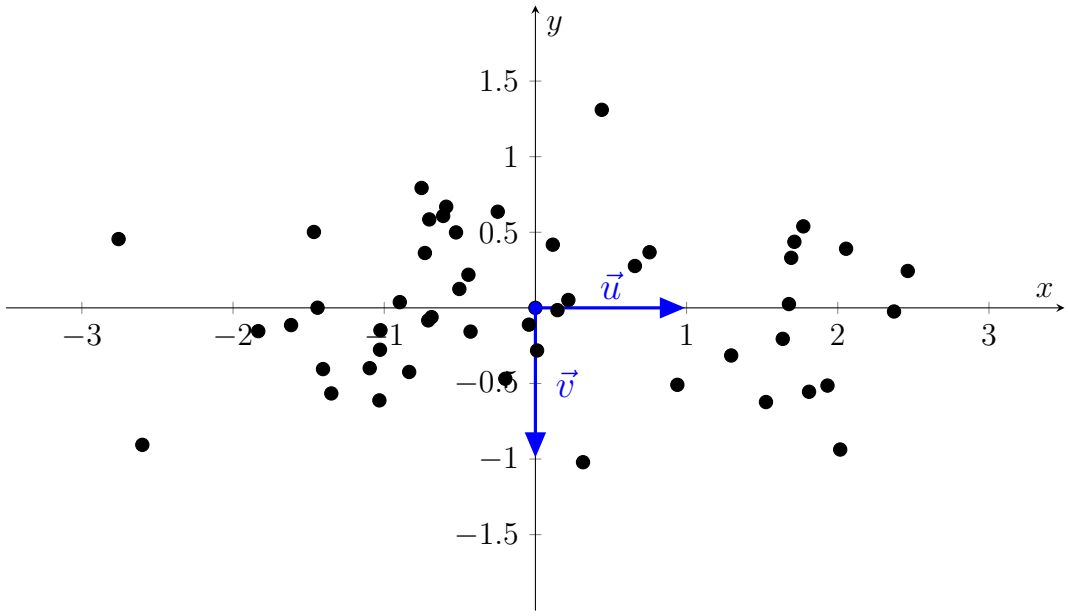

Figure 1.4: Scatterplot of a two-variable dataset in the alternative $(\vec{u}, \vec{v})$ coordinate system of principal components.

In general, the PCA of an n-dimensional problem yields a new coordinate system consisting of n principal components. These principal components are ordered by the proportion of total variation in the data that each explains, in descending order. The first principal component captures the largest portion of the variation, followed by the second principal component, which captures the second largest portion, and so on.

At this point, you can decide whether to retain only the first component, the first two components, or more, depending on the amount of total variation you are willing to ignore. This selection process is key to the dimensionality reduction technique that PCA provides, making it highly effective for simplifying complex datasets.

1.4 Concluding Remarks

Principal Component Analysis (PCA) serves as a powerful dimensionality reduction tool for multi-dimensional data analysis. By transforming the original correlated variables into a new set of uncorrelated components, PCA enables the representation of data in fewer dimensions without significant loss of information. The principal components are ordered by the amount of variation they capture, allowing analysts to focus on those that explain the most significant portion of the data's variability. This transformation simplifies complex datasets into manageable forms while preserving the most critical aspects of the information.

The ability to reduce dimensionality while retaining essential variation makes PCA particularly useful in real-world applications where large datasets often involve dozens or even hundreds of variables. By choosing an appropriate number of components to retain, one can strike a balance between reducing the complexity of the data and maintaining accuracy. Ultimately, PCA enhances data interpretability and enables more effective decision-making by identifying the key dimensions that contribute to the overall

structure and variation in the dataset.

Geometrically, PCA can be understood as rotating the original coordinate system to align the axes with the directions of maximum data variance. In this new coordinate system, each principal component axis captures as much variation as possible, with the first component explaining the greatest variance. By projecting the data onto the first few principal component axes, one can effectively reduce the dimensionality of the data, retaining most of the original structure. This geometric interpretation provides intuitive insight into how PCA condenses the dataset's complexity while highlighting the most meaningful patterns and relationships.