CHAPTER 2

Computing Principal Components

OBJECTIVE

In this chapter, I will demonstrate how to use the *Multivariate.xlsm* Excel template to perform Principal Component Analysis (PCA). This template is compatible exclusively with the Windows version of Excel and has been tested specifically on Microsoft 365. Unfortunately, it is not supported by the Mac version of Excel.

Contents

2.1 Introduction

Throughout this chapter, I will use the *Iris* flower dataset to demonstrate how Principal Component Analysis (PCA) can be performed in Excel. This dataset is well-known and widely used in the field of machine learning for testing purposes (see Fisher, 1988). It consists of 150 iris flower samples, spanning three species: *Iris setosa*, *Iris virginica*, and *Iris versicolor*. Each sample is characterized by four variables, representing the length and width (in centimeters) of two botanical parts, namely the sepal and petal.

For simplicity, I randomly selected 20 records from the original dataset to create a smaller dataset, referred to as *Iris20*. This reduced dataset is shown in Table 2.1. The *iris* dataset can be downloaded from the link `https://bit.ly/4dmWAAK`, while the *iris20* dataset can be obtained from `https://bit.ly/3WMxTGZ`. Both datasets are in Excel format.

Before performing PCA, it is common practice to standardize the dataset. Standardization involves dividing each variable by its non-zero standard deviation, resulting in a dataset where each variable has a standard deviation of 1.

The primary purpose of standardization is to eliminate potential biases in the PCA process due to one variable having naturally high variation. For instance, income tends to be more variable than height, and without standardization, income might disproportionately influence the computation of principal component scores, even if it doesn't provide the greatest discriminatory power.

Another essential procedure is data centering. Centering involves subtracting the mean of each variable from its values, giving all variables a baseline of 0. This procedure often improves the robustness of certain machine learning algorithms. Centering is typically performed before standardization.

If a variable in the dataset maintains a constant value across all observations (i.e., has a standard deviation of zero), it should be removed. Such a variable does not contribute useful information for distinguishing between different units of analysis and offers no value to the PCA process. Remember, the goal of PCA is to identify new composite variables with maximum variation.

The *iris20* dataset shown in Table 2.1 consists of 20 samples, representing all three iris species, and contains four numeric variables: *S.Length*, *S.Width*, *P.Length*, and *P.Width*. PCA will focus on these four variables. The variable *Species* is merely an attribute of the flower and cannot differentiate between two flowers of the same species, so it will not be used in the analysis. Only variables that uniquely define each flower are relevant for this analysis.

Table 2.2 presents the input data after they have been centered and standardized. Centering shifts the mean of each variable to 0 by subtracting the mean from each value. Standardization (also called scaling) further transforms each variable by dividing the values by their standard deviation, resulting in standardized variables with a standard deviation of 1.

Centering and scaling are recommended for computational reasons. In both statistical computing and machine learning, algorithms tend to perform better with centered and scaled data. If large-scale and small-scale data are used in the same algorithm, they may not be treated uniformly, leading to unpredictable outcomes.

Table 2.1: An extract of 20 randomly chosen records from the iris flower dataset[a]

		Sepal		*Petal*	
Case#	*Species*	*S.Length*	*S.Width*	*P.Length*	*P.Width*
55	versicolor	6.5	2.8	4.6	1.5
136	virginica	7.7	3.0	6.1	2.3
116	virginica	6.4	3.2	5.3	2.3
68	versicolor	5.8	2.7	4.1	1.0
100	versicolor	5.7	2.8	4.1	1.3
24	setosa	5.1	3.3	1.7	0.5
109	virginica	6.7	2.5	5.8	1.8
21	setosa	5.4	3.4	1.7	0.2
137	virginica	6.3	3.4	5.6	2.4
50	setosa	5.0	3.3	1.4	0.2
122	virginica	5.6	2.8	4.9	2.0
18	setosa	5.1	3.5	1.4	0.3
129	virginica	6.4	2.8	5.6	2.1
146	virginica	6.7	3.0	5.2	2.3
93	versicolor	5.8	2.6	4.0	1.2
112	virginica	6.4	2.7	5.3	1.9
36	setosa	5.0	3.2	1.2	0.2
94	versicolor	5.0	2.3	3.3	1.0
15	setosa	5.8	4.0	1.2	0.2
58	versicolor	4.9	2.4	3.3	1.0

[a]This dataset is a subset of the larger and widely-used iris dataset provided by Fisher (1988)

When variables are centered, they all share the same baseline value of 0. After standardization, they also have the same range or spread, ensuring that no single variable disproportionately influences the analysis due to large natural variations.

Table 2.2: Standardized Iris flower dataset of Table 2.1

		Sepal		Petal	
Case#	Species	S.Length	S.Width	P.Length	P.Width
55	versicolor	0.83893	-0.43570	0.46100	0.26250
136	virginica	2.42431	0.03533	1.31470	1.23927
116	virginica	0.70682	0.50635	0.85939	1.23927
68	versicolor	-0.08587	-0.67121	0.17643	-0.34797
100	versicolor	-0.21799	-0.43570	0.17643	0.01831
24	setosa	-1.01068	0.74187	-1.18949	-0.95845
109	virginica	1.10316	-1.14224	1.14396	0.62879
21	setosa	-0.61434	0.97738	-1.18949	-1.32473
137	virginica	0.57470	0.97738	1.03013	1.36136
50	setosa	-1.14280	0.74187	-1.36023	-1.32473
122	virginica	-0.35011	-0.43570	0.63174	0.87298
18	setosa	-1.01068	1.21289	-1.36023	-1.20264
129	virginica	0.70682	-0.43570	1.03013	0.99508
146	virginica	1.10316	0.03533	0.80248	1.23927
93	versicolor	-0.08587	-0.90672	0.11952	-0.10378
112	virginica	0.70682	-0.67121	0.85939	0.75089
36	setosa	-1.14280	0.50635	-1.47405	-1.32473
94	versicolor	-1.14280	-1.61326	-0.27887	-0.34797
15	setosa	-0.08587	2.39046	-1.47405	-1.32473
58	versicolor	-1.27491	-1.37775	-0.27887	-0.34797