CHAPTER 2

Computing Principal Components

OBJECTIVE

The purpose of this chapter is to provide a technical review of the mathematics of principal components. The objective is not to provide a comprehensive review of all techniques that have been proposed for computing principal components. It is rather to present a detailed account of one possible approach that has been successfully implemented in practice. You will learn all the computation steps from raw data leading up to the principal components.

I will also discuss the preliminary work that must be done on your data before you start computing the principal components. The two issues discussed will be the centering and the standardization. While chapter 2 focuses on computational methods, chapter 3 is devoted to the interpretation and applications of principal components.

Contents

"Don't say the old lady screamed. Bring her on and let her scream."
Mark Twain

2.1 Introduction

Principal Component Analysis (PCA) is recommended if your dataset contains several correlated variables that were measured on each of many subjects and you are having difficulties extracting useful information from it. One such dataset is shown in Table 2.1. For example, credit card companies or credit bureaus often collect lots of data on individuals (e.g. annual income, length of credit history, current loans, FICO score, ...) considered to be key determinants of credit card default. An economist may gather country data (e.g. gross domestic product, male and female educational attainment, inflation rate, or female fertility rate) to study the determinants of economic growth. A key problem researchers often face when using multidimensional data is that many variables are often correlated, making it difficult to understand which variable has a meaningful impact on the characteristic being investigated.

When dealing with an initial large set of correlated variables, the solution is often to narrow it down to a smaller set of uncorrelated derived variables that retain most of the information contained in the original variables. Even if the interpretation of the derived variables is not straightforward, you can always study their correlation with the original variables to understand their structure. The original variable that has the highest correlation with a derived variable will contribute to its definition in a meaningful way. These correlations between original and derived variables are referred to in the PCA literature as the principal component loadings.

It is essential at this stage to realize that the covariance matrix[1] or the correlation matrix[2] is used to quantify the amount of information contained in a dataset. The covariance matrix contains the variance of each variable on the diagonal whereas its off-diagonal elements are the covariances of the different pairs of variables. Imagine a dataset where each variable has a zero variance. It means each variable takes a unique value for all subjects and does not carry any useful information about these subjects that can be exploited in any in-

[1]For p variables, this $p \times p$ matrix is created by calculating the covariance between all pairs of variables.

[2]The correlation matrix is the matrix of all pairwise correlation coefficients. All diagonal elements of this matrix equal 1.

vestigation. A zero variance also leads to a 0 covariance. You can see why the covariance matrix plays a pivotal in PCA. I will explain in section 2.3.2 why I always prefer the correlation matrix over the covariance matrix even though both matrices produce similar results in most situations. As an example, consider Table 2.2 showing the correlation matrix associated with the original variables of Table 2.1.

Table 2.1: Measurements of 7 attributes taken on 12 subjects.

Subject	V1	V2	V3	V4	V5	V6	V7
1	6.147	7.767	7.830	7.024	7.284	7.393	6.267
2	7.115	6.620	7.479	7.386	6.926	7.485	6.561
3	7.367	7.779	7.150	8.425	7.816	8.612	7.872
4	7.425	8.454	10.160	8.424	8.902	8.671	8.621
5	7.510	8.292	8.895	10.031	8.459	9.972	8.761
6	7.575	7.913	7.958	7.329	7.339	8.582	8.356
7	7.644	8.730	8.846	9.024	8.896	9.276	9.400
8	7.763	8.520	8.881	8.059	6.770	9.191	7.718
9	7.808	7.957	7.939	7.823	7.598	7.637	7.892
10	7.917	8.480	7.840	7.513	8.539	8.413	9.133
11	7.953	8.737	7.671	9.320	8.340	8.576	8.463
12	8.022	8.384	8.380	6.891	7.536	8.199	9.183

You can look at the original variables of Table 2.1 as coordinates of your data points in the standard coordinate system where each of the 7 basis vectors points in a different orthogonal direction. The 7×7 identity matrix describes this standard coordinate system, and each column of this matrix is a basis vector. Calculating the principal components amounts to finding an alternative coordinate system, represented by a special orthogonal matrix $\mathbf{V}$ whose columns are vectors called the *Principal Components* and which represent the 7 basis vectors of a new coordinate system. What is peculiar about the new coordinate system is the direction of each basis vector. The first principal component points in a special direction along which the data projections have the largest variance possible. The second principal component points in the direction of the second largest variance, and so on. This is how you can decide to retain only the first 2 principal components for example if their cumulative percent of the total variance explained is deemed sufficiently large. Table 2.3 contains the 7 principal components associated with the correlation matrix of

Table 2.2. This is the orthogonal matrix **V** that I mentioned in the previous paragraph, and which forms the new coordinate system within which the data will be analyzed. Each of the 7 principal components is an eigenvector of the correlation matrix of Table 2.2 with the associated eigenvalues given in Table 2.4. You can verify[3] that multiplying matrix **Σ** by a column of **V** will yield a vector that is a multiple of that same column scaled by a factor determined by the corresponding eigenvalue. I will discuss in section 2.4 how these principal components and associated eigenvalues are calculated.

Table 2.2: Correlation matrix **Σ** of the 7 variables of Table 2.1

Attributes	V1	V2	V3	V4	V5	V6	V7
V1	1	0.537	0.139	0.246	0.276	0.411	0.771
V2	0.537	1	0.503	0.437	0.598	0.606	0.752
V3	0.139	0.503	1	0.308	0.438	0.505	0.408
V4	0.246	0.437	0.308	1	0.607	0.762	0.382
V5	0.276	0.598	0.438	0.607	1	0.474	0.701
V6	0.411	0.606	0.505	0.762	0.474	1	0.626
V7	0.771	0.752	0.408	0.382	0.701	0.626	1

Table 2.3: The principal components represented by the columns of matrix **V**

PC_1	PC_2	PC_3	PC_4	PC_5	PC_6	PC_7
0.316	0.663	0.164	0.240	0.325	0.414	0.314
0.420	0.137	-0.221	-0.031	-0.842	0.214	0.017
0.300	-0.346	-0.733	0.296	0.331	0.227	-0.059
0.349	-0.447	0.566	0.038	0.058	0.451	-0.386
0.387	-0.182	-0.023	-0.765	0.197	-0.038	0.437
0.412	-0.235	0.247	0.490	-0.053	-0.569	0.384
0.439	0.365	-0.067	-0.161	0.174	-0.450	-0.641

Table 2.4 shows the 7 eigenvalues $\lambda_1, \cdots, \lambda_7$ associated with the eigenvectors of Table 2.3 along with the percent of total variance each explains individually and cumulatively. You can see that the first 2 principal components explain

[3]Tables 2.1, 2.2, 2.3, and 2.5 are included in the following downloadable text file: `https://agreestat.com/books/pca/datasets/table2x1pca.csv`

73.4% of the total variance. You could then decide to narrow your analysis from the initial 7 dimensions down to the 2 dimensions represented by the first 2 principal components PC_1 and PC_2. Consequently, you will need to translate your initial data points of Table 2.1 into the new coordinate system represented by the first 2 principal components. Table 2.5 contains the new coordinates of the initial data points in the coordinate system represented by all 7 principal components.

Table 2.4: The correlation matrix eigenvalues and their contribution to total variance

	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7
Eigenvalues (λ)	4.062	1.075	0.764	0.581	0.307	0.195	0.016
% Variance[a]	0.580	0.154	0.109	0.083	0.044	0.028	0.002
% Var. Cum.[b]	0.580	0.734	0.843	0.926	0.970	0.998	1

[a]Percentage of total variance explained individually
[b]Percentage of total variance explained cumulatively

Table 2.5: Coordinates[a] of Table 2.1 data points in the new coordinate system of principal components

Subject	PCS_1	PCS_2	PCS_3	PCS_4	PCS_5	PCS_6	PCS_7
1	-3.414	-1.431	-0.767	-0.619	-1.002	-0.139	-0.058
2	-3.633	-0.285	0.519	0.166	1.091	0.119	0.033
3	-0.742	-0.106	1.309	-0.261	-0.087	-0.347	0.075
4	1.802	-1.142	-1.664	-0.394	0.607	0.333	0.101
5	2.389	-1.518	0.907	0.534	0.300	-0.273	-0.095
6	-0.667	0.666	-0.057	0.471	0.074	-0.589	-0.026
7	2.541	-0.415	-0.039	-0.485	-0.044	-0.279	-0.104
8	0.230	0.047	-0.361	1.971	-0.566	0.244	0.139
9	-0.900	0.817	0.023	-0.209	0.234	0.745	-0.097
10	0.854	1.254	-0.093	-0.907	-0.087	-0.340	0.255
11	1.333	0.353	1.129	-0.473	-0.574	0.767	-0.019
12	0.208	1.760	-0.906	0.206	0.054	-0.241	-0.205

[a]These coordinates are referred to as the Principal Component Scores (PCS). PCS_1 refers to the scores associated with the first principal component.

You may then retain only the first 2 series of principal component scores for your data analysis. This is how your problem is changed from 7 dimensions down to only 2 dimensions. I will show in section 2.2 how these factors are calculated.

> Table 2.5 columns are often referred to in the literature as the PC factors. Not to be confused with the principal components. The principal components are the basis vectors defining a new coordinate system within which your data points will be analyzed. The coordinates of your data points within that new system are the PC factors or PC scores.
>
> You will always need PC scores to map your data points in the new coordinate system of principal components. We will see later that multiplying the correlation matrix of your original data by matrix of principal components always yields the PC scores.

Figure 2.1 depicts the projection of your data points on the two-dimensional space defined by the first 2 principal components. This allows you to have a graphical representation of your data reflecting over 73% of total variance. With this graph, you can perform some visual exploratory analysis and be able to detect possible outliers. This would have been impossible if you had to deal with all 7 dimensions.

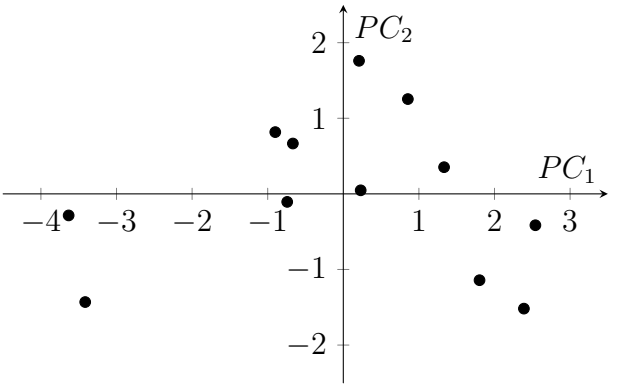

Figure 2.1: Scatterplot of Table 2.1 observations with respect to the first 2 principal components