CHAPTER 3

Using Principal Components

OBJECTIVE

After learning how to compute principal components in chapter 2, this chapter will show you what to do with them. You will learn to determine the minimum number of principal components that will define the few dimensions in which your data will be analyzed. You will also learn what the principal components represent with respect to the original variables, and how they can be used in subsequent analyses of your dataset.

Contents

"It is necessary, first of all, to find a correct logical starting point, one which can lead us to a natural and sound interpretation of the empirical facts."

Ernst Cassirer, An Essay on Man: An Introduction to a Philosophy of Human Culture

3.1 Introduction

An abstract discussion on the right way principal components should be interpreted would be inefficient. Instead, I decided to use 2 actual datasets that will guide the discussions throughout this chapter. In chapters 1 and 2, I presented the logic and the mathematics of principal components. In this chapter, I will use that background to analyze and interpret 2 actual real-life datasets. The first dataset contains the 1979 relative distribution of employed professionals by 9 major industry sectors in 26 European countries, while the second contains State government expenditures and debt by State, in the United States for the fiscal year of 2007. Among other things, I will discuss how to determine the correct number of principal components to retain for your analysis and how to interpret them with respect to the original variables. You will learn how a few principal components are used in a convenient way for identifying the most important variables and for detecting outlying observations. I will close this chapter by briefly discussing how principal components can be rotated to improve their interpretability and by stressing out some key limitations of the concept of rotation.

The employment dataset of 26 European countries to be analyzed in this chapter is listed in Table C.3 of section C.4 in appendix C, and the variables are defined in Table C.2 of the same section. This dataset can also be downloaded as a text file at the following web address:

`agreestat.com/books/pca/datasets/euro79employment.csv`

Interested readers could also download the principal component analysis of that dataset along with the rotation of the entire factor pattern matrix (i.e. the 9 principal component loadings) at the following web address:

`agreestat.com/books/pca/datasets/euro79employmentpca.csv`

The analysis contained in the above text file will be discussed in details later in this chapter.

The second dataset used for illustration purposes in this section contains State government expenditures and debt by State, in the United States for the fiscal year of 2007. For each of the 50 States, the dataset contains 18 expenditure variables listed in Table 3.1.

Table 3.1: List of variables in the 2007 States' expenditures dataset

VarId	Description	VarId	Description
1	Intergovernmental[a]	10	Housing & community development
2	Education	11	Solid waste management
3	Welfare	12	Governmental adminstration
4	Health	13	Interest on general debt
5	Hospitals	14	Other direct expenditures
6	Highways	15	Utility and liquor store[b]
7	Protection	16	Insurance trust[c]
8	Corrections	17	Cash and security holdings
9	Parks & Recreation	18	Total debt outstanding

[a]Intergovernmental expenditures are amounts paid to other governments as fiscal aid in the form of shared revenues and grants-in-aid, as reimbursements for performance of general government activities and for specific services for the paying government, or in lieu of taxes

[b]These expenditures are related to the establishment and operation of liquor stores, the purchase and construction of utility facilities among others

[c]These expenditures include social insurance payments to beneficiaries among others

Interested readers may get more refined definitions of these expenditures by visiting the relevant US Census Bureau's webpage,

`https://www.census.gov/govs/www/classexpdef.html`

You may also download this dataset from its original source on the US Census Bureau's website using the following URL:
`http://www2.census.gov/library/publications/2010/compendia/statab/130ed/tables/11s0452.xls`.

Alternatively, it can be downloaded using one of the following 2 options:

- MS Excel users can download the Excel workbook `states2007.xlsx` using the following link:

  `agreestat.com/books/pca/datasets/states2007.xlsx`

  This Excel workbook contains 3 worksheets named "2007 data," "Input data" and "PCA." The worksheet "2007 data" contains the original data

as reported on the US Census Bureau's website. "Input data" contains the input dataset as used in the principal component analysis (some of the original variables were excluded from the analysis) and "PCA" contains the principal component analysis output.

- Alternatively, you can download the text file `states2007data.csv`, which only requires a text editor and can be downloaded using the following URL:

  `agreestat.com/books/pca/datasets/states2007data.csv`

3.2 Interpreting Principal Components

I performed a principal component analysis on the `states2007data.csv` dataset using the `pca.xlsm` Excel macro[1] and obtained the 18 eigenvalues and associated 18 principal components as well as the principal component scores. Each of the 18 principal components is a vector of length 18, and represents a column in Table C.1 of section C.3 in appendix C. This table also shows the eigenvalues and the percent of total variance they explain individually as well as cumulatively. The complete output can be downloaded at the following address:

`agreestat.com/books/pca/datasets/states2007pca.csv`

Once the PCA output is obtained, you need to first decide how many principal components you are going to retain in your analysis. The number of principal components is expected to be small as it represents the smaller number of dimensions along which your analysis will be conducted. In the next step, you will use the associated principal component scores to extract useful information from your dataset.

3.2.1 *Optimal Number of Principal Components*

To determine how many principal components to use, you must carefully examine the magnitude of the eigenvalues, which by construction, are ranked from the highest value to the lowest. Since each eigenvalue represents the variance of the associated series of principal component scores, you want to retain the smallest number of principal components which explain the desired

[1] You may download this Excel template at the following web address: `https://agreestat.com/books/princomp/pca.xlsm`