CHAPTER 6

Statistical Analysis Based on Principal Components

OBJECTIVE

You learned to compute principal components in chapter 3, to visualize them in chapter 4, and to apply them to basic real-word problems in chapter 5. In this chapter, you will learn how some known statistical techniques can be improved with the use of principal components. In particular, I will cover outlier detection techniques, regression analysis, and linear discriminant analysis.

Contents

6.1 Introduction

In this chapter, I will discuss 4 ways of using principal components (PC) to tackle statistical problems.

(i) *Principal components have often been used to identify the most relevant of a long list of variables.* For this application of PCs, synthetic variables are used primarily for identifying the most important variables. In section 6.2, I will review the two most popular methods for selecting these relevant variables, which are the "inclusion method" and the "elimination method." The inclusion method automatically selects all variables associated with loadings that exceed a predetermined threshold. The elimination method on the other hand, you proceed by elimination where all variables with the highest loadings on the least dominant synthetic variables are eliminated.

(ii) *Practitioners have used PCs to identify outlying observations in a large and complex dataset.* Identifying aberrant values in a multidimensional dataset can be challenging. There is no effective visualization technique you can use to display your data point. PCA allows you to explore your data points in 2 or 3 dimensions, making it possible to identify aberrant values that stray away from the bulk of the data. These aberrant values may require further examination before a decision is made on whether or not they should be kept before analysis can begin. This problem is discussed in section 6.3.

(iii) *Analysts commonly use regression analysis to investigate the relationship between a response and a set of explanatory variables.* The use of this well-known statistical technique in practice presents considerable challenges, which can be resolved by using Principal Component Regression (PCR) instead. This technique is discussed extensively in section 6.4. A well-documented challenge you need to overcome when conducting regression analysis is multicollinearity[1]. This problem will be eliminated with the use of synthetic variables instead.

(iv) *Principal Components can improve standard classification methods based on Linear Discriminant Analysis.* Linear Discriminant Analysis (LDA) is a well-known statistical technique often used by analysts to classify objects into predefined categories. A classification model must be trained on a training sample, before the trained model can be used to predict category membership

[1] In regression analysis, multicollinearity occurs when explanatory variables have a high correlation with one another. This issue makes it difficult to determine the individual effect that each independent variable has on the response variable.

of the testing sample objects. In section 6.5, you will learn how using a few synthetic variables (e.g. 2 or 3) from the PCA you can improve the correct classification rate of LDA-based models.

6.2 Identifying Relevant Variables with Principal Components

The traditional use of PCA consists of deriving the synthetic variables and using a few of the most dominant as surrogates for the original variables, in your data analysis task. However, some analysts prefer to use some selected original variables with the most influence over the retained synthetic variables[2]. In other words, dimensionality reduction is not achieved by replacing a large number of original variables with a few principal components. Instead, it is achieved by using the principal components to select a few original variables that are representative of the entire set of original variables. In this section, I will discuss 2 procedures for selecting the most relevant original variables.

If several variables are highly correlated, then it is logical to consider using only a handful of them to represent the others. For example, women fertility rate and their educational attainment are expected to have a high negative correlation. Therefore, using only one of these 2 variables will be more appropriate in some applications. Moreover, principal component scores are not always easily interpretable, as you do not know what they represent. Any analysis based on them could be difficult to interpret. McCabe (1984) and Jolliffe (1986) discuss various strategies for selecting the most relevant variables. I am going to confine myself to 2 such strategies, which are based on PC loadings.

As a reminder, PC loadings represent the correlation coefficients between original and synthetic variables. The original variables with the highest correlation coefficients in absolute value, play a bigger role and are the most influential in the composition of the synthetic variables they are correlated to. Here are the 2 strategies for deciding about the most relevant variables that are considered in this section:

- **Inclusion Method**

  Jolliffe's rule - see Jolliffe (2002) - regarding the number of PCs to retain, recommends the retention of all PCs with a standard deviation that ex-

[2] Using original as opposed to synthetic variables could make your analysis results easier to interpret.

ceeds 0.70. The inclusion method consists of reviewing all retained PCs from the most dominant[3] to the least dominant and to select the variable with the highest loading as long as it has not already been selected to represent a more dominant principal component. In this case, you will select the variable with the second highest loading.

- **Elimination Method**

  Still based on Jolliffe's rule, the second variable selection strategy proceeds by elimination. For each discarded PC going from the least dominant and proceeding towards the more dominant, you will eliminate the variable associated with the highest loading among all variables that have not already been eliminated. If the variable with the highest loading has already been eliminated, then you should move to the next variable with the second highest loading until you find a variable to eliminate. *The rationale behind this procedure is based on the fact that the most important variable in the least important principal component must not really be that important anyway.*

Consider the R built-in dataset `mtcars` that I previously discussed. I performed the PCA on the 9 quantitative variables `mpg, cyl, disp, hp, drat, wt, qsec, gear`, and `carb`. Table 6.1 contains all 9 principal components, and Table 6.2 contains the associated loadings that I will use for implementing the variable selection procedures of this section.

As previously discussed, there are 2 ways of computing PC loadings. You can either compute all 81 pairwise correlation coefficients, or alternatively you can multiply each column of PC coefficients given in Table 6.1 by the associated standard deviation[4]. Thus, the `PC1` column of Table 6.2 is obtained by multiplying the corresponding column in Table 6.1 by the associated standard deviation of 2.3782. You can repeat the same calculation for the remaining 8 columns.

The R script of Program[5] 6.1 was executed to create the following outputs among others:

[3] PC dominance is determined by the magnitude of the associated synthetic variable's standard deviation.

[4] This PCA produced the following 9 standard deviations for the synthetic variables: 2.3782, 1.4429, 0.7101, 0.5148, 0.428, 0.3518, 0.3241, 0.2419, 0.149.

[5] You can download this script file using the link `https://bit.ly/3sDXx5X`.

- The 9 synthetic variable standard deviations 2.3782, 1.4429, 0.7101, 0.5148, 0.428, 0.3518, 0.3241, 0.2419, 0.149.
- Tables 6.1 and 6.2, which show the principal components and associated loadings for the `MTcars` data frame.
- Tables 6.3 and 6.4, which show how the inclusion and elimination methods for selecting relevant variables work. These 2 tables are discussed later in this section.

Lines `#01-#13` for the part of the script where PCA is performed on the quantitative measurements of `MTcars` data frame selected in line `#06`. This is where the synthetic variable standard deviations, the principal components (c.f. Table 6.1) and the loadings (c.f. Table 6.2) are calculated and displayed.

Lines `#15-#22` implement the inclusion method and creates Table 6.3, whereas lines `#24-#31` implement the elimination method, which leads to Table 6.4.

Program 6.1. R script for implementing the inclusion and elimination variable selection methods, on the `MTcars` data frame

```
01  #
02  #-- Implementing inclusion and elimination algorithms for
03  #   selecting the most relevant variables.
04  #
05  pacman::p_load(tidyverse,xlsx)
06  mtcars.df <- mtcars[,c(1:7,10,11)]
07  mtcars.pca <- prcomp(mtcars.df, center = TRUE, scale = TRUE)
08  lambda <- mtcars.pca$sdev
09  loadings <- mtcars.pca$rotation%*%diag(lambda)
10  colnames(loadings) <- colnames(mtcars.pca$rotation)
11  cat("\nStandard dev:",round(lambda,4),"\n\n") #Standard dev.
12  print(round(mtcars.pca$rotation,4)) #PC coefficients
13  print(round(loadings,4)) #- PC loadings
14
15  #-- Inclusion method ---
16
17  pc.coeffs.inc <- mtcars.pca$rotation[,1:3]
18  lambda.gt07 <- mtcars.pca$sdev[(mtcars.pca$sdev>0.7)]
19  loadings.inc <- pc.coeffs.inc %*% diag(lambda.gt07)
20  round(lambda.gt07,4)
21  round(pc.coeffs.inc,4)
```

```
round(loadings.inc,4)

#-- Elimination method ---

pc.coeffs.elim <- mtcars.pca$rotation[,-c(1:3)]
lambda.le07 <- mtcars.pca$sdev[(mtcars.pca$sdev<=0.7)]
loadings.elim <- pc.coeffs.elim %*% diag(lambda.le07)
round(lambda.le07,4)
round(pc.coeffs.elim,4)
round(loadings.elim,4)
```

End of Script

Table 6.1: Principal components associated to the 9 quantitative variables of the `mtcars` dataset

```
         PC1     PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9
mpg  -0.3931  0.0275 -0.2212 -0.0061 -0.3208  0.7202 -0.3814 -0.1247  0.1149
cyl   0.4026  0.0157 -0.2523  0.0407  0.1171  0.2243 -0.1589  0.8103  0.1627
disp  0.3974 -0.0889 -0.0783  0.3395 -0.4868 -0.0197 -0.1823 -0.0642 -0.6619
hp    0.3671  0.2694 -0.0172  0.0683 -0.2947  0.3539  0.6962 -0.1657  0.2518
drat -0.3118  0.3417  0.1500  0.8457  0.1619 -0.0154  0.0477  0.1351  0.0381
wt    0.3735 -0.1719  0.4537  0.1913 -0.1875 -0.0838 -0.4278 -0.1984  0.5692
qsec -0.2244 -0.4840  0.6281 -0.0303 -0.1482  0.2575  0.2762  0.3561 -0.1687
gear -0.2095  0.5508  0.2066 -0.2824 -0.5625 -0.3230 -0.0856  0.3164  0.0472
carb  0.2446  0.4843  0.4641 -0.2145  0.3998  0.3571 -0.2060 -0.1083 -0.3205
```

Table 6.2: PC loadings associated to the 9 principal components of Table 6.1

```
         PC1     PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9
mpg  -0.9350  0.0397 -0.1571 -0.0032 -0.1373  0.2534 -0.1236 -0.0302  0.0171
cyl   0.9574  0.0227 -0.1792  0.0210  0.0501  0.0789 -0.0515  0.1960  0.0242
disp  0.9450 -0.1283 -0.0556  0.1748 -0.2083 -0.0069 -0.0591 -0.0155 -0.0986
hp    0.8730  0.3888 -0.0122  0.0352 -0.1261  0.1245  0.2257 -0.0401  0.0375
drat -0.7416  0.4930  0.1065  0.4354  0.0693 -0.0054  0.0155  0.0327  0.0057
wt    0.8882 -0.2481  0.3222  0.0985 -0.0802 -0.0295 -0.1387 -0.0480  0.0848
qsec -0.5336 -0.6985  0.4460 -0.0156 -0.0634  0.0906  0.0895  0.0861 -0.0251
gear -0.4982  0.7948  0.1467 -0.1454 -0.2407 -0.1136 -0.0277  0.0765  0.0070
carb  0.5817  0.6988  0.3296 -0.1104  0.1711  0.1256 -0.0668 -0.0262 -0.0477
```

Using Table 6.2, I am going to determine the most relevant variables based on each of the 2 selection methods previously described.

Inclusion Method

Table 6.3 shows the PC loadings associated with the 3 PCs that have a standard deviation exceeding the 0.70 threshold[6]. The highest loading in columns `PC1`, `PC2`, and `PC3` are associated with the 3 variables `cyl`, `qsec`, and `gear` as seen in Table 6.3.

A close examination of Table 6.3 reveals that the first principal component `PC1` can be well represented by a few other variables such as `disp` or `mpg`, all of which have high loadings. In fact, all these variables are different measures of the car engine power and may well be used interchangeably.

Table 6.3: Loadings associated with the first 3 and most dominant principal components[a]

Variable	PC1	PC2	PC3
mpg	-0.9350	0.0397	-0.1571
cyl	0.9574	0.0227	-0.1792
disp	0.9450	-0.1283	-0.0556
hp	0.8730	0.3888	-0.0122
drat	-0.7416	0.4930	0.1065
wt	0.8882	-0.2481	0.3222
qsec	-0.5336	-0.6985	0.4460
gear	-0.4982	0.7948	0.1467
carb	0.5817	0.6988	0.3296

[a] The highlighted numbers are associated with the 3 variables that must be retained.

Elimination Method

For the `MTcars` dataset, implementing the elimination method requires that you start with the 6 least dominant principal components associated with a

[6] The standard deviation of each standardized original variable is 1. You will want to retain PCs with a standard deviation exceeding 1. Due to sampling errors potentially underestimating the "true" standard deviation, this threshold is reduced to 0.70, as recommended by Jolliffe (2002).

standard deviation, which is smaller than or equal 0.70. These PCs and associated loadings are shown in Table 6.4.

Unlike the "inclusion" method, where the table of PCs must be examined from left to right, the "elimination" method requires that you analyze the table of least dominant PCs from right to left. It follows from Table 6.4 that the 6 variables that are eliminated are `disp`, `cyl`, `hp`, `mpg`, `gear` and `drat`. Therefore, the 3 variables that will be retained are `wt` (the car weight), `qsec` (the car's 1/4 mile time[7]) and `carb` (the number of carburetors in the car).

Table 6.4: Loadings associated with the 6 least dominant principal components[a]

Variable	PC4	PC5	PC6	PC7	PC8	PC9
mpg	-0.0032	-0.1373	0.2534	-0.1236	-0.0302	0.0171
cyl	0.0210	0.0501	0.0789	-0.0515	0.1960	0.0242
disp	0.1748	-0.2083	-0.0069	-0.0591	-0.0155	-0.0986
hp	0.0352	-0.1261	0.1245	0.2257	-0.0401	0.0375
drat	0.4354	0.0693	-0.0054	0.0155	0.0327	0.0057
wt	0.0985	-0.0802	-0.0295	-0.1387	-0.0480	0.0848
qsec	-0.0156	-0.0634	0.0906	0.0895	0.0861	-0.0251
gear	-0.1454	-0.2407	-0.1136	-0.0277	0.0765	0.0070
carb	-0.1104	0.1711	0.1256	-0.0668	-0.0262	-0.0477

[a] Highlighted numbers are associated with the variables that must be eliminated. The remaining 3 variables (`wt, qsec, carb`) must be retained.

Note that the 3 variables selected by the "inclusion" method (`cyl, qsec, gear`) are different from those retained by the "elimination" method (`wt, qsec, carb`). Nevertheless, the 2 sets of 3 variables have the following 2 things in common:

- The `qsec` variable is common to both sets.
- For the other 2 variables that are different, `cyl` and `wt` are both highly correlated with the first principal component `PC1`, whereas `gear` and `carb`

[7] This is the time that a car takes to travel 1/4 mile.

are highly correlated with the second principal component PC2. Consequently, these 2 sets of variables may well be used interchangeably.

The science behind these 2 variable selection methods is not well developed. By selecting a few variables to represent the most dominant PCs, you cannot quantify what your gain and loss are. Nevertheless, if you cannot use the synthetic variables in your analysis, you may want to be mindful in your use of these variable selection methods. Those interested, may learn more about this in Jolliffe (1986).

6.3 Identifying Outliers with Principal Components

Outlier detection is a major field of application for principal components. If a large number of quantitative measurements are collected on each unit of analysis, then it becomes nearly impossible to identify units that stray away from the bulk of data, with a visual inspection of the original variables. Identifying outlying data is essential since they will potentially have an unduly high influence on your analysis. You must first identify potential outliers for further inspection, before deciding whether or not they are due to errors and must be discarded, or whether they are legitimate data points that carry useful information about the underlying data distribution.

A key advantage for using principal components to identify outlying observations is their orthogonality. That is, a pairwise analysis of the synthetic variables can be an effective way for identifying outliers due to the absence of correlation. A simple scatterplot depicting 2 synthetic variables PC1 and PC2 can show you what observation is outlying and potentially influential. The same analysis can be performed with different pairs of synthetic variables. All these analyses can be interpreted independently, due to the orthogonality of principal components.

To further investigate this issue, I am going to consider a dataset named `country.data137` that contains 15 measurements on each of 137 selected countries[8]. These measurements include military spending, financial and demographic data. Here is a high-level description of these variables:

[8] You may download this dataset in CSV format with the link `https://bit.ly/3N0qf81`, or in Excel format with the link `https://bit.ly/47sdQBX`.

```
LABEL          Country label (sequential number starting from 1).
COUNTRY        Country name.
CONTINENT      The continent the country belongs to.
MIL.SPENDING   Estimated 2023 military spending in millions USD.
POP2023        Country estimated 2023 population in millions.
GDP.MILLIONS   Country Gross Domestic Product in millions of USD.
DENSITY.POP    Country 2023 population density.
LAND.AREA      Country land area.
FERTILITY.RATE Country fertility rate.
AGE.MEDIAN     Country 2023 median age.
URBAN.POP      Country 2023 urban population.
DEBT.MILLIONS  Country national debt in millions of USD (2019-2021).
DEBT.GDP       Country ratio of debt over GDP (2019-2021).
DEBT.CAPITA    Country debt per capita (2019-2021).
IMPORTS        Annual imports in millions of USD (2019-2021).
XPORTS         Annual exports in millions of USD (2018-2021).
```

Note that GDP data refer to different years. For most countries however, the data refer to the years 2020-2022.

This dataset will be used in this section for outlier detection, as well as in section 6.4 devoted to the important topic of principal component regression.

Internal Structure of the country.data137 Dataset

Before I start manipulating the `country.data137` dataset, I want to explore its internal structure. You can accomplish this objective in R by running the command `str(co.data134)` on R console. This will produce the following output:

```
> str(co.data134)
'data.frame': 134 obs. of  16 variables:
 $ label        : int  1 2 3 4 5 6 7 8 9 10 ...
 $ country      : chr  "Afghanistan" "Albania" "Algeria" ...
 $ continent    : chr  "Asia" "Europe" "Africa" "Africa" ...
 $ mil.spending : num  12000 250 13000 7000 4200 ...
 $ pop2023      : num  42.24 2.83 45.61 36.68 45.77 ...
 $ gdp.millions : num  14583 18882 191913 106714 632770 ...
 $ density.pop  : int  60 105 18 26 17 104 3 109 123 2239 ...
 $ land.area    : num  652860 27400 2381740 1246700 2736690 ...
```