Principal Component Analysis

Cheng Peng

Contents

1 Introduction

In this note, we use factor analysis to aggregate the information of the 12-item self-compassion instrument and 6-item gratitude instrument. The principal factors extracted from the two survey instruments will be used the regression analysis that will reflect the association between self-compassion and gratitude scores.

Before we summarize aggregate the information in the individual items on the survey instruments. provide a non-technical description of the idea of the principal component analysis (similar ideas in factor analysis) and how it works.

2 Principle Component Analysis (PCA)

The method of principal component analysis (PCA) is one of the important dimension reduction methods in the areas of data science and machine learning. Depending on which optimization is used, you can call it a statistical dimension reduction method.

Here we an intuitive example that explains the idea, logic, and some terms related to the principal component analysis (PCA) and the factor analysis (FA).

2.1 Non-numerical example: fish size

Suppose that 50 fish were measured, the following plot might be obtained.

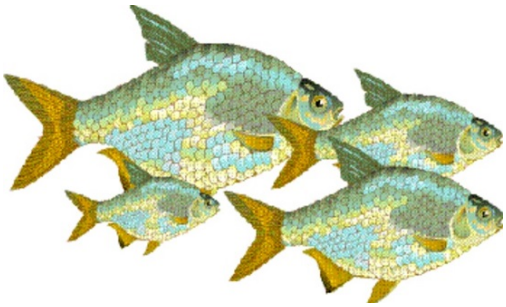

Figure 1: A school of fish with different sizes

There is an obvious relationship between length and breadth, longer fish tend to be broader.

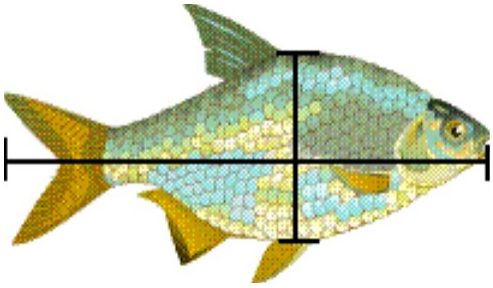

Figure 2: Breadth and length of Fish

Assume we make a scatter plot based on the 50 data points, we have the following plot. It is not surprising that the breadth and length are linearly dependent. This means that the two variables have shared information - one variable contains some amount of information about the other variable.

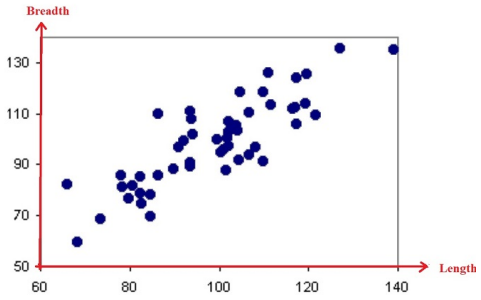

Figure 3: Scatter Plot: Breadth vs Length

Since breadth and length have some information in common, we should figure out an effective away in using the two variables to describe the "size" of fish. The question is how to "effectively" combine the information in different variables that measure a similar type of information?

- **Simply add the values of the breadth and length?** In this particular case, it seems okay to add up the two variables to define a variable measuring the size of fish? However, you weight variable in the data, you really cannot add breadth, length, and weigh together.
- **Pick one of them because they are highly correlated?** This method seems not to be appropriate since dropping one variable will lose information unless they are not 100% correlated (i.e. correlation coefficient is not equal to 1).

- **Because the involved variables may have completely different units (hence different magnitudes)**, we use the correlation coefficient between variables with different magnitudes since the correlation coefficient is a kind of index with is not dependent on the magnitude of the individual variables.

Two statistical methods, principal component analysis (PCA) and factor analysis (FA) can be used to combine variables with possibly different magnitudes based on the correlation between variables.

2.2 The geometric description of PCA

The objective in this example is to define a new variable, size, using the information in breadth and length. Let x_1 = the standardized breadth and x_2 = standardized length. We transform the coordinate system of (x_1, x_2) to a new system (P_1, P_2) in such a way that P_1 axis assumes the maximum variances (this implies P_1 axis has the smallest variance).

The following three figures give the geometric steps to transform the two coordinate systems.

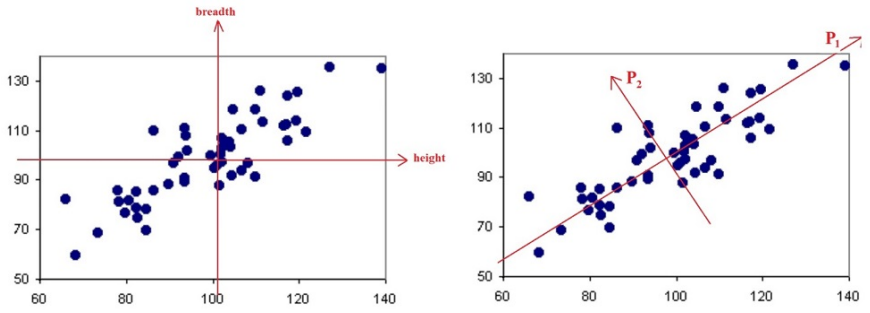

Figure 4: Geometric illustration of the linear transformation of the two coordinate system: (breadth, length) = (P1, P2)

The mathematical expression od the linear transformation between the coordinate system (x_1, x_2) and coordinate system (P_1, P_2) is given by

$$\begin{aligned} P_1 = &\quad a_{11}x_1 + a_{12}x_2; \\ P_2 = &\quad a_{21}x_1 + a_{22}x_2. \end{aligned}$$

where P_1 and P_2 are called the first and the second principal components respectively. From the above figure, we can see that the first principal component (P_1) take the most variation of the data cloud and the second principal component (P_2) take the next maximum variation of the data.

The coefficients of the system of the linear transformation a_{11}, a_{12}, a_{21}, and a_{22} are called **factor loadings**. The magnitudes and signs of these factor loadings reflect the importance and impact of the corresponding standardized factors.

In statistics, the "information" of a standardized variable is primarily reflected in the variation of the variable. From Figure 3, we can see that P_1 contains most of the variation in the data cloud. P_2 contains a small amount of information. This implies we aggregate the information of the original variables to P_1 that contains the majority of the information contained in the original two variables. So we can ignore P_2 in the practical applications to make the analysis simpler. This why we call PCA a dimension reduction method.

If you look at the steps we describe above, we did not use any distribution of variables in the dataset. The only statistical term used in the description is the **variance and covariance** of the variables in the data set. Another relevant method is called **factor analysis (FA)** which assumes parametric distributions of the variables in the data set. We will not go to detail of the FA.

2.3 How to use the principal components?

In some cases, the first principal component can explain up to 80% of the total variation. In applications, we can use the only principal component for analyses such as linear and non-linear association analysis.

The principal component can also be used as predictive models. You can think about how to obtain an IQ score of a person through an IQ test.

Let's still use the **fish** example to illustrate how to use the PCA as predictive model. Assume we use the 50 data records to find the estimated factor loadnings $(\hat{a}_{11}, \hat{a}_{12}, \hat{a}_{21}, \hat{a}_{22},)$. Then the predictive PCA model is explicitly given by

$$P_1 = \hat{a}_{11}x_1 + \hat{a}_{12}x_2;$$
$$P_2 = \hat{a}_{21}x_1 + \hat{a}_{22}x_2.$$

Let assume that we caught a new fish and measured the breadth and length of the fish, we then plug in the measurements to the above system of equations to get the predicted principal component scores (i.e., the type of "size" of the fish). This is analogous to a person who took an IQ test and plugged in the scores earned in each individual test items to the IQ score equation to predict the person's IQ score.

3 A numerical example based on Iris data

3.1 Introduction

The Iris Dataset contains four features (length and width of sepals and petals) of 50 samples of three species of Iris (Iris setosa, Iris virginica and Iris versicolor). These measures were used to create a linear discriminant model to classify the species. The dataset is often used in data mining, classification and clustering examples and to test algorithms.

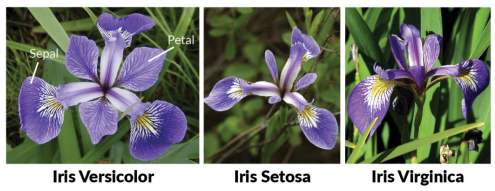

Figure 5: Iris data set: variables

This 100 years old data set has been included in the R base package. The first few records of the data set are displayed in the following table.

```
kable(head(iris))
```

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa

3.2 Fitting PCA model to Iris data

We want to PCA method to reduce the dimensions from 4 (numerical variables) to a smaller number. The R function **prcomp()** to the factor loadings associated the four numerical variables.

```
log.iris = log(iris[,-5])    # drop the categorical variable in the original
                             # data set and transform all numerical to the
                             # log-scale
ir.pca <- prcomp(log.iris, center = TRUE, scale = TRUE)
# summary(ir.pca)[6]    # use the command to explore the possible information
                        # available in the output of the summary.
```

In the above R function, three arguments are explained in the following

```
log.iris = log of the four variables
cater = TRUE, this means the variables are cetered,basically you move the origin of the original coordi
scale = TRUE, divide the difference between the value of each variable and its mean by the standard devi
```

Next, we find the factor loading of the above fitted PCA. We can write an explicit system of linear transformation by using the loadings.

```
kable(round(ir.pca$rotation, 2), caption="Factor loadings of the PCA")
```

Table 2: Factor loadings of the PCA

	PC1	PC2	PC3	PC4
Sepal.Length	0.50	-0.45	0.71	0.19
Sepal.Width	-0.30	-0.89	-0.33	-0.09
Petal.Length	0.58	-0.03	-0.22	-0.79
Petal.Width	0.57	-0.04	-0.58	0.58

The explicit expression of the predictive system of PC is given by

$$PC_1 = 0.50Sepal.Length - 0.30Sepal.Width + 0.58Petal.Length + 0.57Petal.Width$$
$$PC_2 = -0.45Sepal.Length - 0.89Sepal.Width - 0.03Petal.Length - 0.04Petal.Width$$
$$PC_3 = 0.71Sepal.Length - 0.33Sepal.Width - 0.22Petal.Length - 0.58Petal.Width$$
$$PC_4 = 0.19Sepal.Length - 0.09Sepal.Width - 0.79Petal.Length + 0.58Petal.Width$$

The importance of the principal components is given by

```
kable(summary(ir.pca)$importance, caption="The importance of each principal component")
```

Table 3: The importance of each principal component

	PC1	PC2	PC3	PC4
Standard deviation	1.712458	0.9523797	0.3647029	0.165684
Proportion of Variance	0.733130	0.2267600	0.0332500	0.006860
Cumulative Proportion	0.733130	0.9598900	0.9931400	1.000000

From the above table, we can see that the first PC explains about 73.33% of the variation. But we first two principal components explain about 96% of the total variation. In the data analysis, you only need to use the first two PCs that loose about 4% of the information.

3.3 Extracting PC scores

The predictive principle scores are values of the new transformed variables. We can choose first few principal components to use as response variables to do relevant modeling.

The following code extracts the PC scores from the PCA procedure. These scores are the values of the new transformed variables. They can be used as response or predictor variables in statistical models.

```
kable(ir.pca$x[1:15,], caption = "The first 15 PC scores transformed from the original variable. In the
```

Table 4: The first 15 PC scores transformed from the original variable. In the analysis, you want to either the first PC or the first two PCs.

PC1	PC2	PC3	PC4
-2.406639	-0.3969554	0.1939647	0.0047795
-2.223539	0.6901804	0.3500015	0.0488684
-2.581105	0.4275418	0.0188976	0.0499095
-2.450869	0.6860074	-0.0687460	-0.1496465
-2.536853	-0.5082516	0.0293226	-0.0400482
-1.841495	-1.2899381	-0.2527683	0.1638906
-2.479490	0.1011323	-0.4974043	0.1227590
-2.348593	-0.1569003	0.1360186	-0.0954982
-2.535948	1.2477681	-0.1118812	-0.0754686
-2.625580	0.5074073	0.6594737	-0.4732591
-2.252707	-0.9305324	0.3266434	-0.0450709
-2.431184	-0.0290417	-0.0929138	-0.2368403
-2.697278	0.7816300	0.6575035	-0.3883884
-3.325521	1.1499290	0.1948436	-0.2162823
-2.380613	-1.6326568	0.5878310	0.2993832

4 Your Turn

4.1 Read in data file

I have converted the original Excel file to a csv file. You can read in data file directly from my S3 bucket on the AWS or download the csv file to your local drive then read in the file from your local computer. My WCU machine's firewall setting does not allow reading the data file directly from the AWS where I host the course web page.

```
survey = read.csv("https://raw.githubusercontent.com/pengdsci/STA490/main/w09-SurveyDataCsvFinal.csv", h
dim(survey)
```

```
## [1] 104  33
```

Here are my suggestions for the rest of the analysis;

4.2 Data cleaning

As I mentioned in the consulting meeting, you need to manage the original data. You are expected to combine the categories of many demographic variables. I would expect the number of categories should not be more than three since the sample size is relatively small.

There are also some missing values found in some of the items in the survey instruments. You can use simple imputation methods like replacing the missing values with the mean or median of the corresponding variables.

Once you complete the data cleaning, you can continue the analysis based on the following steps.

4.3 Subsetting data sets

Create two data sets:

Self-Compassion contains only the data associated with the 12 items in the survey instrument. The original data file, the 12 variables are named as $Q2_1, Q2_2, ..., Q2_{12}$.

Gratitude Questionnaire contains only the variables associated with gratitude questions. The variables used in the original data file are $Q3_1, Q3_2, ..., Q3_6$.

4.4 Applying the principal component analysis to two data sets

Using a similar code on the two data set and then use one or two PCs in each data sets as the response variables to do further analysis. For example, you can rename the PCs in the self-compassion data using SC1 (= PC1 in self-compassion) and SC2 (= PC2) to denote the self-compassion index. Then you can combine the transformed response variables with other demographic data to do regression analysis.

You can similarly do the same thing with the gratitude data and extract one or two PCs for further analysis.

4.5 Statistical analysis for addressing the research questions

In your final **analytic data set** with transformed self-compassion index variable(s) and gratitude index variable(s) and modified demographic variables.

4.6 Project ideas based on clients' requests

You can think about statistical methods such as regression modeling to address the research questions. The principal components extracted from the survey instruments can be used as either response or predictor variables in the regression models. These ideas with be the topics to discuss next week.