Mathematics Learning
Support Centre

Statistics: 3.2 Principal Components Analysis

Rosie Cornish. 2007.

1 Introduction

This handout is designed to provide only a brief introduction to principal components analysis and how it is done. Books giving further details are listed at the end.

Principal components analysis is a multivariate method used for data reduction purposes. The basic idea is to represent a set of variables by a smaller number of variables called **principal components**. These are chosen in such a way that they are uncorrelated (and are therefore measuring different, unrelated aspects, or dimensions, of the data).

2 Assumptions

Principal components analysis, like factor analysis, is designed for interval data, although it can also be used for ordinal data (e.g. scores assigned to Likert scales). The variables should be linearly related to each other. This can be checked by looking at scatterplots of pairs of variables. Obviously the variables must also be at least moderately correlated to each other, otherwise the number of principal components will be almost the same as the number of original variables, which means that carrying out a principal components analysis would be pointless.

3 What principal components analysis does

If you have p variables $X_1, X_2, \ldots, X_p$ measured on a sample of n subjects, then the i^{th} principal component, Z_i can be written as a linear combination of the original variables. Thus,

$$Z_i = a_{i1}X_1 + a_{i2}X_2 + \ldots + a_{ip}X_p$$

The principal components are chosen such that the first one, $Z_1 = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p$ accounts for as much of the variation in the data (i.e. in the original variables) as possible subject to the constraint that

$$a_{11}^2 + a_{12}^2 + \ldots + a_{1p}^2 = 1$$

Then the second principal component $Z_2 = a_{21}X_1 + a_{22}X_2 + \ldots + a_{2p}X_p$ is chosen such that its variance is as high as possible. A similar constraint applies — namely, that

$$a_{21}^2 + a_{22}^2 + \ldots + a_{2p}^2 = 1$$

Another constraint is that the second component is chosen such that it is uncorrelated with the first component. The remaining principal components are chosen in the same way.

When you do a principal components analysis you get what are called eigenvalues. It is not necessary to understand what eigenvalues are in order to understand the principles behind principal components analysis. However, it is useful to know that the eigenvalues are the variances of the principal components. In other words, the first eigenvalue is the variance of the first principal component, the second eigenvalue is the variance of the second principal component, and so on. Thus, because of the way the principal components are selected, the first eigenvalue will be the largest, the second the next largest, etc. There will be p eigenvalues altogether but some may be equal to zero.

Once the principal components have been calculated you will need to decide how many to keep. Essentially any principal components that account for only a small proportion of the variation in the data (i.e. those with small eigenvalues) are discarded. Different methods are used to decide which principal components to retain:

- Choose sufficient principal components to account for a particular percentage (e.g. 75%) of the total variability in the data.
- Choose only those principal components with eigenvalues over 1 (if using the correlation matrix).
- Use the scree plot of the eigenvalues. This will indicate whether there is an obvious cut-off between large and small eigenvalues.

4 Carrying out principal components analysis in SPSS

Note that SPSS will not give you the actual principal components. However, these can be calculated from the output provided.

– **Analyze**
– **Data Reduction**
– **Factor**
– Select the variables you want the factor analysis to be based on and move them into the **Variable(s)** box.
– In the **Extraction** window, select **Principal components**. Under **Analyze** ensure that **Correlation Matrix** is selected (this is the default). The default is also to extract eigenvalues over 1. You can either keep it like this or specify the number of factors to be equal to the number of original variables (later on you can decide which principal components to keep and which to discard). Click on **Continue**.
– In the **Rotation** window, select **None** under **Method**. Click on **Continue**.
– In the **Scores** window you can specify whether you want SPSS to save the values of the 'principal components' (as mentioned above, these are not actually the principal components but can be used to calculate them) for each observation (this will save them as new variables in the data set). Under **Method** choose **Regression**. Click on **Continue**.
– **OK**

IMPORTANT NOTE
Note that what SPSS gives you are NOT the principal components. However, these can be calculated quite easily:

1. To get the coefficients of the principal components (the as):
SPSS will give you a table entitled the **Component Matrix**. The components are listed as columns in this table; the variables are listed as rows. To get the values of the as you must **DIVIDE** the values in the table by the SQUARE ROOT of the corresponding eigenvalue. For example, to get $a_{11}, a_{12}, a_{13}, \ldots$ you should divide EACH number in the first column by the square root of the first (largest) eigenvalue (i.e. the eigenvalue corresponding to component 1). Similarly, the second column should be divided by the square root of the eigenvalue corresponding to component 2, and so on.

2. To get the values of the principal components as new variables in the data set:
SPSS will have saved variables called **FAC1_1, FAC2_1**, and so on. These are NOT the values of the principal components. To get the principal components you have to **MULTIPLY** these factor scores by the square root of the corresponding eigenvalue. For example, if the eigenvalue for the first principal component was 3.65, you would compute the first principal component in SPSS as follows:

– **Transform**
– **Compute**
– Under **Target variable** write PC1 (or something similar — to stand for first principal component) and under **Numeric Expression** type in FAC1_1 * sqrt(3.65).
– Click on **OK**

Use a similar process to compute the values of the second principal component (calling this one PC2, using FAC2_1 and replacing the 3.65 by whatever the second eigenvalue is).

5 References

- Manly, B.F.J. (2005), **Multivariate Statistical Methods: A primer**, Third edition, Chapman and Hall.
- Rencher, A.C. (2002), **Methods of Multivariate Analysis**, Second edition, Wiley.