Principal Components Analysis (PCA)

**Introduction:** Principal component analysis (PCA) is concerned with explaining the variance structure of a set of variables through a few linear combinations of the original variables. Its general objective is dimension reduction. Although components are required to reproduce the total system variability, often much of this variability can be accounted for by a small number of the principal components (. If so, there is (almost) as much information in the components as there is in the original variables. The principal components can then replace the initial variables for further analysis, for example, in developing principal component regression (PCR) models. The original data set, consisting of measurements on variables, is reduced to a data set consisting of measurements of principal components.

The first principal component has the greatest variability. The second component has the maximum variability among all linear combinations that are orthogonal to the first. The third principal component is orthogonal to both the first and second, and so on for further components. Variables with the greatest variance will be emphasized typically in first few principal components, and thus dominate the analysis.

**Why Principal Components?**

More often principal components are obtained for use as input to another analysis. For example, two situations in regression where principal components may be useful are (1) if the number of independent variables is large relative to the number of observations, a test may be ineffective or even impossible, and (2) if the independent variables are highly correlated (termed as multicollinearity), the estimates of repression coefficients may be unstable by having large standard error resulting in imprecise parameter estimations. In such cases, the independent variables can be reduced to a smaller number of principal components that will yield a better test or more stable estimates of the regression coefficients.

**Example**:In the Places Rated Almanac, Boyer and Savageau rated 329 communities according to the following nine criteria:

1. Climate and Terrain
2. Housing
3. Health Care and Environment
4. Crime
5. Transportation
6. Education
7. Arts
8. Recreation
9. Economics

Note: Within the dataset, except for housing and crime, the higher the score the better. For housing and crime, the lower the score the better. While some communities might rate better in the arts, other communities might rate better in other areas such as having a lower crime rate and good educational opportunities. The corresponding csv data named “places” is available in the Datasets module in Canvas course shell.

**Objective:** There are 329 observations representing the 329 communities in our dataset and 9 variables stated above. With a large number of variables, there would be too many pairwise covariances or correlations between the variables to consider. For , there would be such covariances or correlations or 36 two-dimensional scatterplots to be studied! With such a large number of correlations or plots, it is difficult to interpret the results and graphs.

To interpret the data in a more meaningful form, it is therefore necessary to reduce the number of variables to a few, interpretable linear combinations of the original variables. Each linear combination corresponds to a **principal component**.

**Learning objectives & outcomes:** Upon completion of this lesson, you should be able to do the following:

* Carry out a principal component analysis using R;
* Assess how many principal components should be considered in an analysis;
* Determine when a principal component analysis may be based on the variance-covariance matrix, and when the correlation matrix should be used.

**Principal Component Analysis (PCA) Procedure:** Algebraically, principal components are particular linear combinations of the random variables . Suppose that we have a random vector of variables . That is,

with population variance-covariance matrix .

It is to be noted that nothing in the formulation described below require to assume that has a multivariate normal distribution.

Consider the linear combinations

…

Each of these can be thought of as a linear regression, predicting each *Yi* from . There is no intercept term. Each are weights and known as *loadings*. These loadings form the vector

These loadings can be viewed as regression coefficients and they give the contribution of each variable to the principal component. We choose the loadings so that has the maximum variance and is mutually uncorrelated with other principal components. Note that are all scalars. Also, is a function of random variables , and therefore is a random variable.

Each has a population variance given by

where , and becomes variance when .

The population covariance between and is given by

Now let us define the principal components.

***First Principal Component* ():** The *first principal component* is the linear combination of -variables that has maximum variance among all linear combinations , so it accounts for most of the variation in the data. Specifically, we will define coefficients for in such a way that is maximized.

Note that this variance can be made arbitrarily large by multiplying by a large scalar.To avoid this ambiguity, we restrict the loadings to have unit length, that is,(that is, is a unit vector).

More formally, select that maximizes

subject to the constraint that

***Second Principal Component* ():** The *second principal component*  is the linear combination of -variables that accounts for as much of the remaining variation as possible, with the constraint that the first and second principal components are uncorrelated, that is, correlation and covariance between the first and second principal component is zero.

We select that maximizes the variance of this new component

subject to the constraint that the sums of squared coefficients add up to one,

along with the additional constraint that these two components will be uncorrelated with each other.

The goal is to make the variances as large as possible and all the covariances equal to zero.

All subsequent principal components have this same property – they are linear combinations of original - variables that account for as much of the remaining variation as possible and they are uncorrelated with other principal components.

We will do this in the same way with each additional component. For instance:

***i*th *Principal Component* ():**

We select that maximizes

subject to the constraint that the sum of squared coefficients adds up to one along with the additional constraint that this new component will be uncorrelated with all the previously defined components.

.....

Therefore, all principal components are uncorrelated with one another.

It is to be noted that the formulation presented below does not require any multivariate normality assumption for variables

**Question: How do we find the coefficients** **for a principal component?**

The solution involves the eigenvalues and corresponding eigenvectors of the covariance matrix . Note that the covariance matrix is a positive definite matrix and therefore, all the eigenvalues will be positive real numbers.

We are going to let through denote the eigenvalues of the covariance matrix . These are ordered so that is the largest eigenvalue and is the smallest, .

We are also going to let the vectors through denote the corresponding eigenvectors of . The vectors through are all unit vectors and mutually orthogonal.

Theoretical results show that the elements for these eigenvectors will be the coefficients of the principal components and the variance for the *i*th principal component is equal to the *i*th eigenvalue .

**Spectral Decomposition Theorem**

The covariance matrix can be written as the sum over the *p* eigenvalues, multiplied by the product of the corresponding eigenvector times its transpose:

The second expression is a useful approximation if are small. Therefore, we can approximate by .

Note that the total variation of is defined as the sum of variances of each component in . That is, the total variation of is the trace of the covariance matrix . It turns out that (Result 8.2 on page 432, Johnson)

This above relation enables us to interpret each principal component in terms of the amount of the full variation explained by each component. The proportion of variation explained by the *i*th principal component is equal to the eigenvalue associated with that component divided by the sum of the all eigenvalues. In other words, the *i*th principal component explains the following proportion of the total variation:

A related quantity is the proportion of variation explained by the first *k* principal components. This would be the sum of the first *k* eigenvalues divided by the total variation.

Naturally, if the proportion of variation explained by the first *k* principal components is large, then not much information is lost by considering only the first *k* principal components.

**Estimation:** All of the above is defined in terms of the population covariance matrix which remains to be unknown in most practical situations. In that case, we estimate by the sample covariance matrix which is given by the formula:

**Procedure:** We compute the eigenvalues of the sample covariance matrix , and the corresponding eigenvectors .

The estimated principle components using the eigenvectors as coefficients are defined as:

+

+

….

+

…..

+

Generally, we only retain the first *k* principal components. In doing so, we must balance two conflicting desires:

1. To obtain the simplest possible interpretation, we want *k* to be as small as possible. If we can explain most of the variation just by two principal components then this would give us a much simpler description of the data.

2. To avoid loss of information, we want the proportion of variation explained by the first *k* principal components to be large. Ideally as close to 1 as possible; i.e., we want

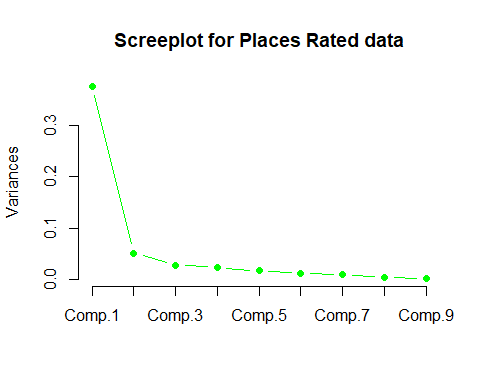
Note that **,**

To show that , we begin by  **(**using the equation )

Pre-multiplying both sides by , we get

(since andare orthogonal)

**How to determine the number of Principal Components to retain:** There is always a question of how many components to retain. There is no definite answer to this question. However, the following guidelines have been proposed:

1. Retain sufficient components to account for a specified percentage of the total variance, say 80%.
2. Retain the components whose eigenvalues are greater than the average of the eigenvalues.
3. A useful visual aid to determine an appropriate number of principal components is a **scree plot** (Scree is the rock debris at the bottom of a cliff). With the eigenvalues ordered from largest to the smallest, a scree plot is a plot of estimate of versus  *-* the magnitude of an estimated eigenvalue versus its rank. To determine the number of components to retain, we look for an elbow (bend) in the scree plot. The number of components is taken to be the point at which the remaining eigenvalues are relatively small and all about the same size. The recommendation is to retain those eigenvalues in the steep curve before the first one on the straight line. Below figure shows the scree plot for the Places Rated data with 9 principal components. An elbow occurs in the plot at . That is, the eigenvalues after are all relatively small and about the same size. In this case, it appears that two or perhaps three sample principal components effectively summarize the total sample variance. Later we find that the first three principal components collectively explains 87% of the total variation in the data, which is reasonably a large percentage.

**Notes:** Principal component analysis using the covariance matrix should only be considered if all of the variables have the same units of measurement. Variables measured on different scales are often standardized, and the eigenvalues and corresponding eigenvectors are calculated based on the sample correlation matrix .

**Standardized variables:** If the variables either have different units of measurement (i.e., pounds, feet, gallons, etc.), or if we wish each variable to receive equal weight in the analysis, then the variables should be standardized before a principal components analysis is carried out. Standardize the variables by subtracting its mean from that variable and dividing it by its standard deviation:

where = measurement for ith variable, = sample mean for ith variable, = sample standard deviation for ith variable.

After this transformation, we perform the principal component analysis using the standardized data.

**Principal Component Analysis Procedure using correlation matrix,**

The principal components are first calculated by obtaining the eigenvalues for the correlation matrix, :

and the corresponding eigenvectors

Then the estimated principle components are calculated using formulas similar to before, but instead of using the raw data we will use the standardized data in the formulas below:

+

+

…..

+

Rest of the procedure and the interpretations are same as before.

We now list some general comparisons of principal components from with those from :

**1.** The percent of variance accounted for by the components of will differ from the percent for .

**2.** The coefficients of the principal components from differ from those obtained from .

The magnitudes of the coefficients give the contributions of each variable to that component. Because the data have been standardized, they do not depend on the variances of the corresponding variables.

One of the problems with this analysis is that the analysis is not as 'clean' as one would like with all of the numbers involved. A variable may be significant in two different principal components which may lead to ambiguous interpretation.

**Benefits of PCs**

In summary:

1. PCA helps in dimensionality reduction. Through PC analysis, we convert set of correlated variables to non-correlated variables.
2. It finds a sequence of linear combinations of variables.
3. PCA also serves as a tool for better data visualization of high dimensional data. We can create a heat map to show the correlation between each component.
4. It is often used to help in dealing with multicollinearity before a model is developed.
5. Sometimes in regression settings, you might have a very large number of potential explanatory variables and you may not have much of an idea as to which ones you might think are important. You might perform a principal components analysis first and then perform a regression predicting the variables from the principal components themselves. The nice thing about this analysis is that the regression coefficients will be independent of one another because the components are independent of one another. In this case, you actually say how much of the variation in the variable of interest is explained by each of the individual components. This is something that you can not normally do in multiple regression.

R program: PCA, PCA\_AirPollution