Aim: Principal Component Analysis-Finding Principal Components, Variance and Standard Deviation calculations of principal components.(Using R)

Principal Component Analysis

In simple words, principal component analysis is a method of extracting important variables (in form of components) from a large set of variables available in a data set. It extracts low dimensional set of features from a high dimensional data set with a motive to capture as much information as possible. With fewer variables, visualization also becomes much more meaningful. PCA is more useful when dealing with 3 or higher dimensional data.

It is always performed on a symmetric correlation or covariance matrix. This means the matrix should be numeric and have standardized data.

Normalization:

The principal components are supplied with normalized version of original predictors. This is because, the original predictors may have different scales. For example: Imagine a data set with variables' measuring units as gallons, kilometers, light years etc. It is definite that the scale of variances in these variables will be large.

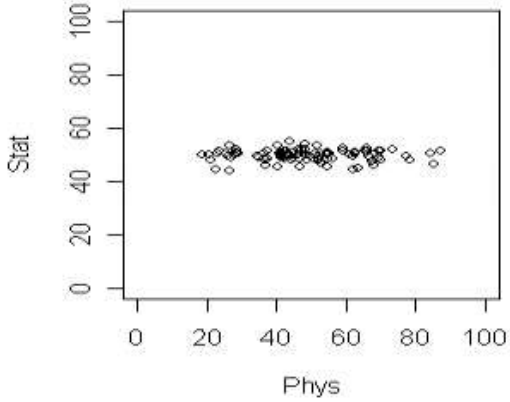
Performing PCA on un-normalized variables will lead to insanely large loadings for variables with high variance. In turn, this will lead to dependence of a principal component on the variable with high variance. This is undesirable.

PCA highlights [2]

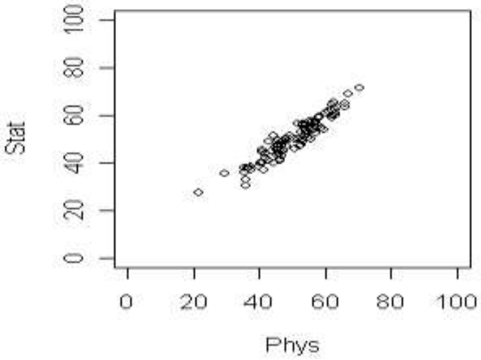
1. PCA is used to overcome features redundancy in a data set.
2. These features are low dimensional in nature.
3. These features a.k.a components are a resultant of normalized linear combination of original predictor variables.
4. These components aim to capture as much information as possible with high explained variance.
5. The first component has the highest variance followed by second, third and so on.
6. The components must be uncorrelated (remember orthogonal direction ?). See above.
7. Normalizing data becomes extremely important when the predictors are measured in different units.
8. PCA works best on data set having 3 or higher dimensions. Because, with higher dimensions, it becomes increasingly difficult to make interpretations from the resultant cloud of data.
9. PCA is applied on a data set with numeric variables.
10. PCA is a tool which helps to produce better visualizations of high dimensional data.

Example:

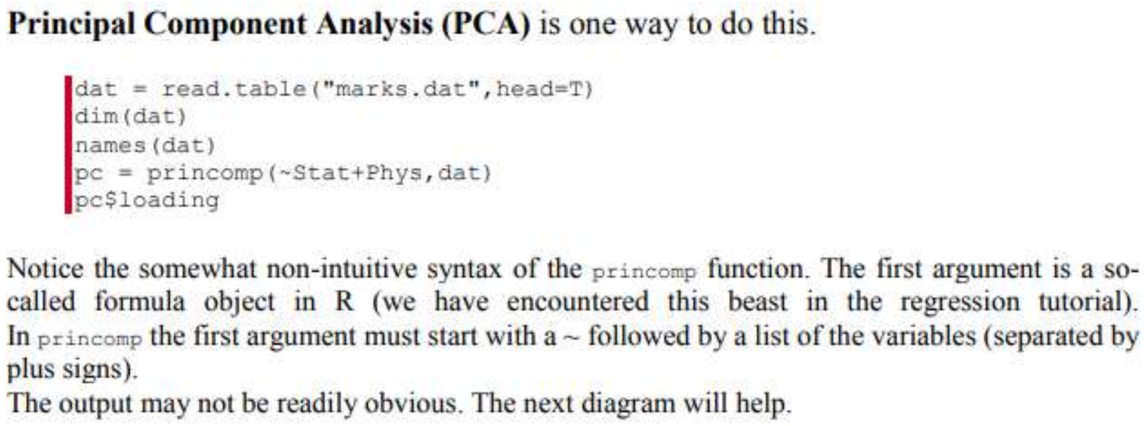
Consider 100 students with Physics and Statistics grades shown in the diagram below. The data set is in `marks.dat`.



If we want to compare among the students which grade should be a better discriminating factor? Physics or Statistics? Surely Physics, since the variation is larger there. This is a common situation in data analysis where the direction along which the data *varies the most* is of special importance. Now suppose that the plot looks like the following. What is the best way to compare the students now?



Here the direction of maximum variation is like a slanted straight line. This means we should take linear combination of the two grades to get the best result. In this simple data set the direction of

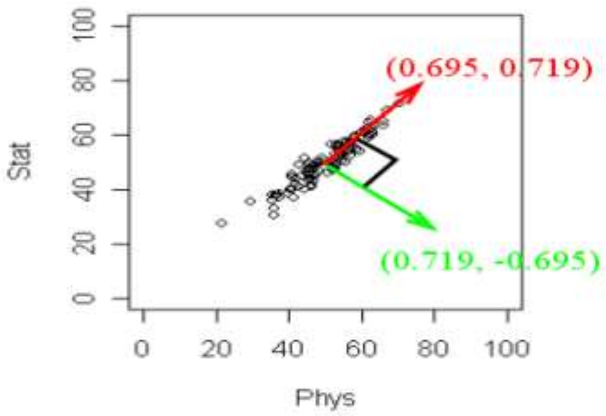


Principal Component Analysis (PCA) is one way to do this.

```
dat = read.table("marks.dat",head=T)
dim(dat)
names(dat)
pc = princomp(~Stat+Phys,dat)
pc$loading
```

Notice the somewhat non-intuitive syntax of the `princomp` function. The first argument is a so-called formula object in R (we have encountered this beast in the regression tutorial). In `princomp` the first argument must start with a ~ followed by a list of the variables (separated by plus signs).

The output may not be readily obvious. The next diagram will help.



R has returned two principal components. (Two because we have two variables). These are a unit vector at right angles to each other. You may think of PCA as choosing a new coordinate system for the data, the principal components being the unit vectors along the axes. The first principal component gives the direction of the maximum spread of the data. The second gives the direction of maximum spread perpendicular to the first direction. These two directions are packed inside the matrix pc$loadings. Each column gives a direction. The direction of maximum spread (the first principal component) is in the first column, the next principal component in the second and so on.

Conclusion:

Thus we have found principal components, calculated the mean and variance of principal components.