## Iteration II, Data Transformation, Principal Component Analysis

In the exploratory data analysis we have seen that there are several highly correlated variables. In the second iteration in order to reduce the dimensionality of the given dataset we are using the technic of principal component analysis (PCA). PCA’s central idea is to reduce the number of variables of a dataset while retaining as much variation (and by that also the information content) as possible. The PCA is applied as the last data pre-processing step, because its outcome, completely new set of uncorrelated variables with high variation, is not easily interpretable from the business point of view (cf. Jolliffe, 2002).

The variance of a dataset is a measure of the information content of a dataset. In the PCA a correlation matrix of the previously normalized (and standardized) dataset is created. The eigenvectors of the (symmetric) correlation matrix are the uncorrelated (orthogonal) principal components of the dataset. To find the principle components of the numeric variables of the dataset the r function “princomp(x, ...)“ has been used. The sum of the eigenvalues represent the total variance I the dataset, the corresponding eigenvalues of the eigenvectors represent the proportion of variance explained by each eigenvector of the correlation matrix (principle component). The plot in the Figure 1 pictures the amount of variance explained by the first 10 components.



Figure : Variance explained by the first 10 ecomponents

For the further analysis only the components which have the squared standard deviation (eigenvalues) higher then one will be retained, meaning that this components explain at least the same amount of variance as the original variables. By applying the PCA, the 87 numeric variables (“Customer\_ID” excluded) could be reduced down to 16.

Sources:

Jolliffe, I. (2002). *Principal component analysis*. New York: Springer.