Principal component analysis (PCA)

Linear transformations. The purpose is to demonstrate how PCA works. Generate n observations from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

Center your data (you can denote it by $\boldsymbol{\chi}_c$, for example) and calculate the sample covariance matrix $\mathbf{S}$.

1) Find the spectral decomposition of $\mathbf{S}$. Use the matrix of eigenvectors to transform the data ($\boldsymbol{\chi}_{tr} = \boldsymbol{\chi}_c\mathbf{V}$). Calculate the variances of the new variables and compare with the eigenvalues. Perform now PCA with function 'prcomp' in R.

2) Consider now an arbitrary linear transformation, that is $\boldsymbol{\chi}_{atr} = \boldsymbol{\chi}_c\mathbf{A}$, where $\mathbf{A}$ is orthogonal matrix. Calculate the variances of the new variables and the correlation between them.

Decathlon data (results of Olympic Decathlon from 1988). **Aim**: perform PCA and study whether the first principal component will rank the athletes in the same way as the score variable.

1) Explore the data set: study the number of observations and variables, measurement units and direction. Do all the variables have the same measurement direction, that is, a larger value corresponds to a better result? If not, change the scale direction. Are there any outliers? If there are observations that could bring us to wrong conclusions, perform the analysis without these.

2) Study the means and standard deviations of the sports events variables and also the correlation matrix.

3) Perform now PCA (function 'prcomp' in R). Observe that 'prcomp' uses the covariance matrix by default. If we want to perform the analysis using the correlation matrix, we have to specify *scale=T*.

- Write down the formula/expression for the first principal component. What does the first principal component measure in this example?

- What events dominate the second principal component? What about the third PC?
- What is the variance of the first and second PC? How much of the total variation is described by these two components?
- Study the covariance matrix of the principal components. Compare the variances of the principal components with the eigenvalues (which eigenvalues?).
- Compute the correlation between "Score" and PC1 and PC2. Make also the scatter plots.

Biplot. The plot is showing:

1) the score of each case (athlete) on the first two principal components,

2) the loading (weight) of each variable (i.e., each sporting event) on the first two principal components. The left and bottom axes are showing [normalized] principal component scores, the top and right axes are showing the loadings. In general it assumes that two components explain a sufficient amount of the variance to provide a meaningful visual representation of the structure of cases and variables. You can look to see which events are close together in the space. Where this applies, this may suggest that athletes who are good at one event are likely also to be good at the other proximal events. Alternatively you can use the plot to see which events are distant.

Swiss bank notes data. Is it possible to distinguish between genuine and counterfeit bank notes? Recall the explanatory data analysis for this data set. What variable(s) seemed most suitable for distinguishing between the two groups?

1) Apply now PCA to the bank data set. At first we do not standardize the data (why?).

- Find the vector of eigenvalues and the matrix of eigenvectors of $\mathbf{S}$.
- Write down the formula for the first and second PC. What variables dominate the first two PCs? Can you interpret them? Think like this: the first eigenvector gives the weights used in the linear combination of the original data in the first PC.
- How much of the variation do the first 3 PCs explain?

- Find the correlations of the original variables with PC1 and PC2. Plot the correlations in a scatter plot.

2) Apply now PCA to the standardized variables. Now, each original variable has the same weight in the analysis and the results are independent of the scale of each variable. Compare the results with the previous analysis, that is, with the analysis based on the covariance matrix. Compare for example: eigenvalues, eigenvectors, scatter plots of the PCs, correlations of the original variables with PCs.