MS-E2112 Multivariate Statistical Analysis Lietzén, Mellin, Ilmonen

Exercise 3

Problem 1: Principal Component Analysis

Continue the exercise from session 2. Download the file DECATHLON.txt into your R-workspace. The file contains the results of 48 decathletes from 1973. Familiarize yourself with the data and perform the correlation matrix based PCA transformation.

a) How much of the variation of the original data is explained by k principal components, where $k = 1, 2, ..., 10$.

b) Choose a sufficient amount of principal components and try to interpret them. Are the interpretations same as last week?

c) Calculate the sample mean and covariance matrix from the transformed data.

d) Add one clear outlier into the data set. Use PCA and try to detect the outlier.

e) Now download the data DECATHLON2.txt into your workspace. The supplementary variable at column 13 indicates the type of competition that the corresponding points were achieved. Use the first 23 athletes and all the variables except for rank and points in your analysis.

f) Can you find a combination of principal components that splits the data set into two groups? The package *ade4* has some nice tools for visualizing the data.

g) Use the principal components calculated from the first 23 athletes to predict the group of the last 4 athletes.

Problem 2: Affine equivariance

a) Show that the sample mean $T()$ is affine equivariant. In other words, if you transform your data $X \to Y$ such that

$$y_i = Ax_i + b,$$

then

$$T(Y) = AT(X) + b,$$

for all nonsingular $p \times p$ matrices A and for all p-vectors b.

b) Show that the sample covariance matrix $S()$ is affine equivariant. In other words, if you transform your data $X \to Y$ such that

$$y_i = Ax_i + b,$$

then

$$S(Y) = AS(X)A^T$$

for all nonsingular $p \times p$ matrices A and for all p-vectors b.

Home Exercise 3: Maximizing Variance

Let x denote a p variate random vector with a finite mean vector μ and a finite covariance matrix Σ and let y_k denote the kth principal component of x. Let $b \in \mathbb{R}^p$, $b^T b = 1$. Assume that $b^T x$ is uncorrelated with first $k - 1$ components of x. Show that $\mathrm{var}(y_k) \geq \mathrm{var}(b^T x)$.