Demo Problem 1: Principal Component Analysis

Upload 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. Conduct the analysis without the variables: points, height and weight.

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? Visualize the observations with respect to the first two principal components.

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

Demo Problem 2: Affine Equivariance

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

$$y_i = Ax_i + b,$$

then

$$S(Y) = AS(X)A^\top$$

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

Homework Problem 1: Maximizing Variance

Let x denote a p-variate random vector with a finite mean vector μ and a finite full-rank covariance matrix Σ. Let $y_k = \gamma_k^\top (x - \mu)$ denote the kth principal component of x. Let $b \in \mathbb{R}^p$ such that $b^\top b = 1$. Assume that $b^\top x$ is uncorrelated with first $k-1$ principal components of x. Read lecture slides 2 carefully and give detailed proofs for the following.

(a) Let $b = d_1\gamma_1 + \ldots + d_p\gamma_p$. Show that $d_i = 0$, when $i < k$.

(b) Show that $\mathrm{var}(y_k) \geq \mathrm{var}(b^\top x)$.

Be careful with your notation and note that $y_k \neq \gamma_k$.