Assignment 1
Examining multivariate data

Kurskod och namn:	732A97 Multivariate Statistical Methods
Delmomentsansvarig:	Krzysztof Bartoszek
Instruktioner:	This assignment is part of the examination for the Multivariate Statistical Methods course You will work in groups of 2–4. Submit your report as a **.PDF** file Be concise and do not include unnecessary printouts and figures produced by the software and not required in the assignments. **All code (R) should be included as an appendix into your report.** A typical report should contain 2–4 pages of text plus some amount of figures plus appendix with codes. In the report reference **ALL** consulted sources and disclose **ALL** collaborations. The report should be handed in via LISAM (or alternatively in case of problems e–mailed to krzysztof.bartoszek@liu.se), by **23:59 24 November 2017** at latest. **Late submission may result in an additional penalty assignment.** The report can be written in English or Swedish. Notice there is a final deadline of **23:59 4 February 2018** after which no submissions nor corrections will be considered and you will have to redo the missing labs next year.

Assignment developed by Ann–Charlotte Hallberg and Bertil Wegmann.

Learning objectives

After reading the recommended text and doing the assignment the student shall be able to:

- give a brief description of a multivariate data set
- study the structure/relationships between the variables
- examine the possible extreme values or even outliers
- understand and use different multivariate residuals

Recommended reading

Chapters 1–2 in *Johnson, Wichern*
Chapters 1–2 in *Everitt, Hothorn*
For basic R code: the Little Book of R for Multivariate Analysis (`https://little-book-of-r-for-multivariate-analysis.readthedocs.io/en/latest/`).

The first step in any data analysis is an examination of the raw data. With multivariate data it is more involved then with univariate. But because of the added complexities even more important.

Question 1: Describing individual variables

Consider the data set in the `T1-9.dat` file, National track records for women. For 54 different countries we have the national records for 7 variables (100, 200, 400, 800, 1500, 3000m and marathon). Use R to do the following analyses.

a) Describe the 7 variables with mean values, standard deviations e.t.c.

b) Illustrate the variables with different graphs (explore what plotting possibilities R has). Make sure that the graphs look attractive (it is absolutely necessary to look at the labels, font sizes, point types). Are there any apparent extreme values? Do the variables seem normally distributed? Plot the best fitting (match the mean and standard deviation, i.e. method of moments) Gaussian density curve on the data's histogram. For the last part you may be interested in the `hist()` and `density()` functions.

Question 2: Relationships between the variables

a) Compute the covariance and correlation matrices for the 7 variables. Is there any apparent structure in them? Save these matrices for future use.

b) Generate and study the scatterplots between each pair of variables. Any extreme values?

c) Explore what other plotting possibilities R offers for multivariate data. Present other (at least two) graphs that you find interesting with respect to this data set.

Question 3: Examining for extreme values

a) Look at the plots (esp. scatterplots) generated in the previous question. Which 3–4 countries appear most extreme? Why do you consider them extreme?

One approach to measuring "extremism" is to look at the distance (needs to be defined!) between an observation and the sample mean vector, i.e. we look how far one is from the average. Such a distance can be called an *multivariate residual* for the given observation.

b) The most common residual is the Euclidean distance between the observation and sample mean vector, i.e.

$$d(\vec{x}, \bar{x}) = \sqrt{(\vec{x} - \bar{x})^T (\vec{x} - \bar{x})}.$$

This distance can be immediately generalized to the L^r, $r > 0$ distance as

$$d_{L^r}(\vec{x}, \bar{x}) = \left(\sum_{i=1}^{p} |\vec{x}_i - \bar{x}_i|^r \right)^{1/r},$$

where p is the dimension of the observation (here $p = 7$).

Compute the squared Euclidean distance (i.e. $r = 2$) of the observation from the sample mean for all 54 countries using R's matrix operations. First center the raw data by the means to get $\vec{x} - \bar{x}$ for each country. Then do a calculation with matrices that will result in a matrix that has on its diagonal the requested squared distance for each country. Copy this diagonal to a vector and report on the five most extreme countries. In this questions you **MAY NOT** use any loops.

c) The different variables have different scales so it is possible that the distances can be dominated by some few variables. To avoid this we can use the squared distance

$$d^2_{\mathbf{V}}(\vec{x}, \bar{x}) = (\vec{x} - \bar{x})^T \mathbf{V}^{-1} (\vec{x} - \bar{x}),$$

where $\mathbf{V}$ is a diagonal matrix with variances of the appropriate variables on the diagonal. The effect, is that for each variable the squared distance is divided by its variance and we have a scaled independent distance.

It is simple to compute this measure by standardizing the raw data with both means (centring) and standard deviations (scaling), and then compute the Euclidean distance for the normalized data. Carry out these computations and conclude which countries are the most extreme ones. How do your conclusions compare with the unnormalized ones?

d) The most common statistical distance is the *Mahalanobis distance*

$$d^2_{\mathrm{M}}(\vec{x}, \bar{x}) = (\vec{x} - \bar{x})^T \mathbf{C}^{-1} (\vec{x} - \bar{x}),$$

where $\mathbf{C}$ is the sample covariance matrix calculated from the data. With this measure we also use the relationships (covariances) between the variables (and not only the marginal variances as $d_{\mathbf{V}}(\cdot, \cdot)$ does). Compute the Mahalanobis distance, which countries are most extreme now?

e) Compare the results in b)–d). Some of the countries are in the upper end with all the measures and perhaps they can be classified as extreme. Discuss this. But also notice the different measures give rather different results (how does Sweden behave?). Summarize this graphically.