Department of Statistics
Lund University

Data Mining
Fall 2013

# Project 1: Bootstrap and splines

*Introduction to R, Bootstrap, Splines*

*Perform all requested tasks. In principle, you should be able to do within the timeframe of two hours of lab. If for any reason you will not be able to conclude your work in the lab, it is expected that you deliver a written report showing that you have completed the work outside of the lab session. In such a case, on a separate paper give answers to the questions, attach a printout of the code you have used and all graphs that you deem important for this lab session. Your grade of the lab will be based on this material. Some useful R-code that can help in completing Project 1 can be found* **here**.

## Part One – Resampling from pairs of data

Health services and health insurance companies are interested in determining what kind of medical examinations and diagnostic procedures should be administered to a newborn child. In one approach, there is a score system based on which it is determined when a child is healthy and does not require any special attention or when he/she is not in which case a series additional medical tests are performed.

A random sample of records for 736 recently born children (singleton and not prematurely born) has been considered from hospital across a certain region. The records contain a large variety of information but extraction of weight and height data are given in the file:

**WeightHeight.txt**.

1. Read in the data using function `read.table()`.
2. Using the data estimate the mean, the covariance for the length and the weight of children and correlation coefficient.
3. Perform a bootstrap study of the accuracy of the correlation estimate.
4. Provide 95% bootstrap confidence interval for the coefficient.

**Part Two – Smooth spline fitting with fixed number of knots**

1. Download and activate the package `"splines"`. Use R-help feature to get information about the function `bs()`. Use this function to plot the cubic B-spline basis on interval [0,1] with the midpoint as the knot point (include also the endpoints).

2. Simulate data from the following model

$$y = x(1-x) + \epsilon$$

   where $\epsilon$ is distributed according to $N(0, 0.0625)$ (the second parameter is the variance so for **R** one needs to use the square root of it, i.e. 0.25). Consider regularly spaced predictor values $x_i = i/400$, $i = 1, ..., 400$. Plot the obtained data.

3. Define the matrix **H** as defined in the lecture, for $H_j$ being the cubic B-splines. What dimension does this matrix have?

4. Find the coefficient of the least squares fit to the data based on the basis made of B-splines. Observe how matrix computation are performed in R.

5. Plot the spline fit to the data and compare with the true value. Do you think that the smooth spline method worked well in this case?

6. For the remaining part of this problem. Perform non-parametric bootstrap from the data to obtain and graph several spline curves that would represent the bootstrap confidence region around the original spline fit. What type of resampling scheme did you use for this task?

7. Comment about the accuracy of the fit as indicated by the bootstrap analysis.

## Part Three – Examples of microarray data

This part of the project presents different microarray data sets that are publicly available for analysis. In this project no particular analysis of the data is required but you are asked to download the data and make yourself familiar with the format of the data. The following is a short description of four data sets that came from actual microarray experiments:

**ALL dataset** The dataset we will use comes from a study on acute lymphoblastic leukemia presented in

Chiaretti, S., Li, X., Gentleman, R., Vitale, A., Vignetti, M., Mandelli, F., Ritz, J., and Foa, R. (2004). Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. *Blood*, **103**(7), 2771-2778.

Li, X. (2009). ALL: A data package. R package version 1.4.7.

The data consists of microarray samples from 128 individuals with this type of disease. Actually, there are two different types of tumors among these samples: T-cell ALL (33 samples) and B-cell ALL (95 samples).

For B-cell ALL samples, one can distinguish different types of mutations. Namely, ALL1/AF4, BCR/ABL, E2A/PBX1, p15/p16 and also individuals with no cytogenetic abnormalities.

The ALL dataset is part of the *bioconductor* set of packages that was developed for **R** for analysis biomedical data and, in particular, microarray data. To use it, one needs to install at least a set of basic packages from *bioconductor*, see **http://www.bioconductor.org/**. To install a set of basic bioconductor packages and the ALL dataset, it is required that a computer has an active Internet connection. The sequence of instructions is available in the **code** prepared for this lab.

An example of analysis of this data set can be found in Chapter 5 of **Data Mining with R. Learning with Case Studies**, by Luis Torgo, CRC Press 2011.

**NCI60 dataset** NCI microarray data. The data contains expression levels on 6830 genes from 64 cancer cell lines. Cancer type is also recorded. This data set is described in the introductions to both our textbooks and is downloaded as part of the package ISRL.

**Ch10Ex11 dataset** On the first textbook website, **www.StatLearning.com**, there is a gene expression data set (Ch10Ex11.csv) that consists of 40 tissue samples with measurements on 1000 genes. The first 20 samples are from healthy patients, while the second 20 are from a diseased group.

**Ch18.3** In the second textbook, there are discussed data from another microarray experiment. The data are divided into a training set of 144 patients with 14 different types of cancer, and a test set of 54 patients. Gene expression measurements were available for 16, 063 genes. Cancer classes are labelled as follows:

1. breast
2. prostate
3. lung
4. collerectal
5. lymphoma
6. bladder

7. melanoma
8. uterus
9. leukemia
10. renal
11. pancreas
12. ovary
13. meso
14. cns

Reference:

S. Ramaswamy and P. Tamayo and R. Rifkin and S. Mukherjee and C.H. Yeang and M. Angelo and C. Ladd and M. Reich and E. Latulippe and J.P. Mesirov and T. Poggio and W. Gerald and M. Loda and E.S. Lander and T.R. Golub (2001) Multiclass Cancer Diagnosis Using Tumor Gene Expression Signatures.*Proc. Natl. Acad. Sci.*,**98**, pp 15149-15154.

1. Please, visit the webpage of the bioconductor package to get idea about this advanced project that was built for data mining techniques in biosciences.
2. Download the package and the data set ALL following the instruction in the **code**.
3. Use `help()` to obtain some information about the data set.
4. Load the package associated with the first textbook and check if the data set NCI60 is available.
5. Load in the `Ch10Ex11.csv` data using `read.csv()`. You will need to select header=F. Check the size of the data set.
6. Visit the webpage for the **second textbook**. Download the data set that is last in our list, i.e. the one described as Ch18.3. Load it into **R**.

## Part Four – Monte Carlo/bootstrap in Assignment 1

This part is complementary to the bootstrapping example discussed in Assignment 1, where two different parameter estimators were investigated and a possibility of modeling certain data with Poisson distribution was considered.

1. Perform the Monte Carlo study of two estimates for Poisson distribution discussed in Problem 1 of Assignment 1. Formulate the conclusions following from the study.

2. Perform a bootstrap study of goodness-of-fit of Poisson distribution to the data given in Problem 1 of Assignment 1. Is Poisson distribution a reasonable model for the data?

## Part Five – Microarray bootstrap study from Assignment 1

This part is complementary to the example of bootstrap study of microarray inaccuracies discussed in Assignment 1, where different sources of inaccuracies were discussed. Here we assume that we have five sextuples of experiments, each performed on the same tissue sample (the same spray of genetic material within each sextuple). The data are available in **here**.

1. Estimate parameters that describe the variability of the microarray technology.
2. Perform a bootstrap study of accuracy of these estimates.