Computational Statistics
Summer 2023
Exercise Sheet 5

Fabio Miranda - Fabio.MalcherMiranda@hpi.de
Hendrik Raetz - Hendrik.Raetz@hpi.de
Pauline Hiort - Pauline.Hiort@hpi.de

June 12th 2023

This week's exercise focuses on NNLS, bootstrapping and RANSAC.

Problem 19 (pen & paper) - NNLS
Please compute the NNLS solution for the following data: $y = \binom{2}{1}$ $X = \left(\begin{smallmatrix} 3 & 0 \\ 3 & 1 \end{smallmatrix}\right)$
Use the following R results:

```
>lm(y~0+x)$coefficients
0.6666667   -1.0
>lm(y~0+x[,1])$coefficients
0.5
>lm(y~0+x[,2])$coefficients
1
```

Problem 21 - Bootstrapping
A spoiled investment banker only flies to his favorite private island in the Caribean when the water temperature is predominantly within his preferred range, which is defined as between 29 and 32 degrees Celsius. The weather forecast, which even money cannot influence to be without a random error, gives the following temperatures as well as random resampling (bootstrap samples) results of these temperatures (with replacement). Compute 80% empirical confidence intervals (CI) for the mean and median temperature and give advice if the vacation should be canceled. Hint: Use the percentiles to compute the CI.

day	1	2	3	4	5
forecast	29	31	27	30	34
bootstrap sample 1	29	27	31	29	31
bootstrap sample 2	31	30	30	27	27
bootstrap sample 3	27	31	29	27	34
bootstrap sample 4	30	27	27	27	34
bootstrap sample 5	29	27	29	31	27
bootstrap sample 6	27	27	31	29	30
bootstrap sample 7	29	30	27	29	30
bootstrap sample 8	34	29	30	34	27
bootstrap sample 9	31	30	29	29	34
bootstrap sample 10	34	30	30	31	34

Problem 22 - RANSAC
This task focuses on the application of the RANSAC (Random Sample Consensus) algorithm. We will use the famous (NOT $infamous$) $mtcars$ dataset, which you can load into R using `library(datasets)` or via the Moodle course page, if you prefer using Python.

A Plot the *displacement* values against *horsepower* and look for outliers in the data set

B Fit a linear regression of *horsepower* (dependent variable) onto *displacement* (independent variable) and plot the fitted line.

C Implement the RANSAC algorithm for a linear regression model from B as a function `RANSAC<-function(n, m, t)` with $n = number\ of\ iterations$, $m = minimum\ consensus\ set\ size$ and $t = maximum\ distance\ of\ a\ data\ point\ from\ the\ model$ that returns the best model as well as the set of inliers and outliers for this model.

D Run your implemented function for $n = 50, 100, 500, 1000$; $m = 10, 15, 20$ and use for t the 80th percentile of the residual errors from the simple linear fit (fitting all points). Plot your solutions and state which is the best fitted model.

Problem 23 - Normalization
Argue for the following two scenarios whether a normalization should be undertaken:

A A farmer wants to figure out whether worms prefer one of the two species of apples he grows. Thus, he counts the number of apples with a worm for the two species of apple trees daily. He has 50 trees for apple species A in one field. Apple species B is of gold color and he can sell a lot more of these. Therefore, he has 100 trees in a different field.

B Santa Claus has a new sleigh and wants to compare its speed to the old one. He undertook overall 100 test runs using the same set of reindeer and recorded the times. Randomly switched sleighs during the tests and did 50 test runs with each sleigh.

Problem 24 (pen & paper) - Normalization
Compute by hand the sum total, median and quantile normalization for the following dataset:
A: 2, 4, 3
B: 4, 2, 5

Problem 25 - Normalization

Let's apply what we have learned to biology! We have created a large dataset by creating technical replicates (we analyzed the 'same' sample four times) and then applied different techniques to normalize the data. This is performed to aid in quantifying RNA in biological samples[1]. The data represents the different counts for specific RNA molecules (rows) across the replicates (columns).
Optional: A specialized bioinformatics R package for this type of analysis is DESeq. You will find a short summary on the normalization online, if you are interested[2].

A Read the fly data into a dataframe and make appropriate quality control plots to visualize the four different count distributions (recommended R commands: `read.table`, `boxplot`). Given the nature of the technical replicates, what would you expect to see in the boxplot? Describe the data and comment on the need for normalization. *Hint:* Choose a sensible plotting range and filter the data for instances where molecules counts were observed.

B Apply the sum total normalization, make appropriate plots and comment on the quality and appropriateness.

C Apply the median normalization, make appropriate plots and comment on the quality and appropriateness.

D Apply the quantile normalization (recommended R library: `preprocessCore` from bioconductor, recommended R commands: `as.matrix`, `normalize.quantiles`), make appropriate plots and comment on the quality and appropriateness.

E *Optional*: Apply the DESeq normalization (recommended R library: `DESeq2` from bioconductor, recommended R commands: `DESeqDataSetFromMatrix`, `estimateSizeFactors`, `counts`, note that you can set the condition for all repeats to 'untreated'), make appropriate plots and comment on the quality and appropriateness.

F Compare the results from the previous steps. Which method appears most appropriate?

[1] `https://en.wikipedia.org/wiki/RNA-Seq`

[2] 'DESeq2-normalized counts: Median of ratios method': `https://hbctraining.github.io/DGE_workshop/lessons/02_DGE_count_normalization.html`

These exciting real world problems are not graded, however, we believe these will enrich your experience in the Computational Statistics course and therefore, it is strongly recommended that you attempt to complete these problems. We will discuss the problems on Monday, **June 19, 2023**. Each student is *required* to present at least one solution during the semester. To ensure an interactive exercise experience, solutions will be presented each week by either volunteers or randomly generated volunteers via R.