Computational Statistics
Summer 2023
Exercise Sheet 4

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

May 25 2023

This week's exercise focuses on regularization and active set methods.

Problem 17 - Choose Wisely
State the appropriate variant of the Ordinary Least Squares (OLS) method for each of the following scenarios. Justify your answer.

A A proteomics (analysis of proteins in a cell) study investigates the influence of different proteins on the growth of bacteria. Quantitative data is available for 500 proteins. We expect to see a positive/negative effect for very few proteins, based on prior knowledge.

B Quality control requires knowledge of the exact composition of a drug. Multiple measurements are taken to detect the absolute contributions of the ingredients.

C A genetic study aims to estimate the influence of each gene on a target variable. During the application of OLS, researchers are facing numerical problems due to the high correlation of multiple genes.

Problem 18 - Ridge Regression
With a regression problem, the weight of students should be predicted from their height given in both $cm(x_1)$ and $m(x_2)$. For the sake of convenience we regard the following two observations:
$y = (50 \quad 100)', x_1 = (150 \quad 200)', x_2 = (1.5 \quad 2)'$

A Why is it not possible to obtain a ordinary least squares solution for this data set? Identify the specific point when the computation fails.

B Run a Ridge Regression with a suitable value of λ and give the solution vector $\hat{\beta}_{\text{Ridge}}$.

C What is the interpretation of the coefficients $\hat{\beta}_{\text{Ridge}}$? Comment on their goodness and any potential problem?

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 20 - Mass Spectrometry
Mass Spectrometry is an analytical technique that measures the mass-to-charge ratio of ions. One of its important applications is the accurate mass determination and characterization of proteins. On the Moodle course page you will find two files of such a real world proteomics data set. First, load both files into R (recommended command: load) or Python. The spectrum file consists of an excerpt of a mass spectrum of a protein. The first column gives the mass/charge position (m/z), the second column the measured intensity. The model file consists of a matrix of possible protein fragments. Each column corresponds to a single model, which represents the expected intensity of different protein fragments for all mass positions of the spectrum. The overall task in this problem is to find the coefficients for all the models that explain the sum of the measured spectrum. Note that the measured spectrum not only contain biological information, but also noise that cannot be explained by the models.

A Visualize the spectrum data (x-axes: m/z, y-axes: intensity). What you see is the raw output of the measurement.

B Visualize the models 210, 215, and 225 and shortly describe what you observe. Each model corresponds to one protein fragment that might be in the spectrum. E.g. to explain the whole spectrum the different models need to be considered together. Hint: Visualize a model by plotting the spectrum m/z and the model coefficients.

C Run an ordinary least squares regression in order to explain the intensities of the spectrum by all models. Argue whether it is appropriate to include an intercept and how an intercept could be interpreted.

D Visualize the coefficients of the regression model. What is the interpretation of each coefficient? Comment on the results (magnitude, sign, distribution of coefficients).

E Now, compute a lasso solution (recommended library and command: lars) with intercept. From biological knowledge, we can assume that this excerpt of the spectrum does not contain more than 15 protein fragments. Give the selected variables. Compare the results to the OLS solution and comment on the difference in the coefficients.

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 5, 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.