MATRIX THEORY: HOMEWORK 2 (v1), 13 October 2020

This homework focuses on least squares problems and low-rank approximations of matrices. This topic is partially covered in the chapters 2 and 3 of the lecture notes of the course. You may use any tools and results from these chapters.

Least squares problems and low-rank approximations are useful in many fields of engineering, for instance, in *system identification*, as we will see in this homework.

Exercise A: Least square problems

A *least square problem* is a problem of the form:

$$\min_{x\in\mathbb{R}^n} \|Ax-b\|_2$$

where $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ and $m \geq n$. In the following, we assume that A has full column rank.

(A1) Show that the solution of the least square problem is given by

$$A^\top A x = A^\top b,$$

and there exists a unique solution when A has full column rank. In this case, the solution can be expressed as $x = (A^\top A)^{-1}A^\top b$.

(A2) Suppose the QR decomposition of A is given by $Q\left[\begin{smallmatrix} R \\ 0 \end{smallmatrix}\right]$, where $Q \in \mathbb{R}^{m\times m}$ is unitary (i.e., $Q^\top Q = I$) and $R \in \mathbb{R}^{n\times n}$ is upper triangular. Express the solution of $A^\top Ax = A^\top b$ using the QR decomposition of A. Show that the computation of the solution can be reduced to the resolution of a single triangular system of linear equations (which can be solved efficiently using backward substitution).

Exercise B: Low-rank approximation

(B1) Define the singular values of a matrix $A \in \mathbb{R}^{m\times n}$. Are they unique?

Show that the number of nonzero singular values of a matrix is equal to its rank.

(B2) Let $X \in \mathbb{R}^{m\times n}$ be such that $|X_{ij}| \leq \varepsilon$ for all $i \in \{1,\ldots,m\}$ and $j \in \{1,\ldots,n\}$. Let $\|X\|_2$ be the 2-norm (aka. *spectral* norm) of X and let $\|X\|_F$ be its Frobenius norm.

Show that $\|X\|_2 \leq \|X\|_F \leq \sqrt{mn}\varepsilon$. Provide an example where the inequalities are tight (i.e., are equalities). Justify your answers.

(B3) Let $A \in \mathbb{R}^{m\times n}$ have rank r $(\leq \min(m,n))$ and let $X \in \mathbb{R}^{m\times n}$ be a perturbation matrix such that $|X_{ij}| \leq \varepsilon$ for all $i \in \{1,\ldots,m\}$ and $j \in \{1,\ldots,n\}$. Finally, let $B = A + X$. Using B1 and B2, what can you say about the smallest $\min(m,n) - r$ singular values of B?

Propose a method to estimate the rank of A if only B is known and assuming that $\sigma_r > 2\sqrt{mn}\varepsilon$, where σ_r is the smallest nonzero singular value of A.

Hint: Results from the lecture notes related to variational problems may be helpful.

Exercise C: Realization theory

A classical problem in control theory is to determine the model of a dynamical system when the only available data are the values of the output at different times. These systems are called *black-box* or *data-driven* systems. For the implementation part of this homework, you will be asked to build an autoregressive (AR) model from a sequence of outputs $\{y(0), y(1), \ldots, y(N)\}$ generated from a *black-box* system. Therefore, you will have to use the tools studied in the previous subsections.

As for the data, we consider the total number of hospitalizations from COVID-19 in Belgium available at Covidata.be. The data set is attached to the homework. For the implementation, you can use any coding language (e.g. Matlab, Julia, Python,...) you want.

An AR model is given by the following recurrence equation:

$$y(t) = \alpha_0 + \sum_{i=1}^{p} \alpha_i y(t-i)$$

where p is the *order* of the AR model and α_i are the *model parameters*.

Assuming that you have $N+1$ outputs $(N > p)$, explain how you can reformulate the problem of estimating the model parameters from the set of outputs as a system of linear equations. For a given sequence of outputs $\{\hat{y}(0), ..., \hat{y}(N)\}$ and a given model order p, we want to find the parameters that minimize the squared error $\|y - \hat{y}\|_2$. Show that this can be reexpressed as a least square problem for a system of linear equations.

To solve this least square problem, you can use the built-in QR function of the language you have chosen to get the triangular problem. Then, you can solve this problem by using backward substitution.

We consider the data from March 15th to July 31th. The confinement terminated at May 31th. We consider the period from March 15th to May 31th as the "confinement mode" and the period from June 1st to July 31th as the "social distancing mode". Due to different policies, people's behaviors were quite different. So, the underlying systems are also different. You are required to identify the system for each mode separately. During the identification, each data set has to be divided into a training set (70%) and a validation set (30%).

The squared error of the of the training and the validation are denoted by $\mathrm{ERR}_p^{\text{train}}$ and $\mathrm{ERR}_p^{\text{validation}}$ respectively. We are interested in the effect of p on these two errors. Precisely, we ask you to solve the least square problem for several values of the parameter $p \in [1, 30]$. For each value of p, plot $\mathrm{ERR}_p^{\text{train}}$ and $\mathrm{ERR}_p^{\text{validation}}$. Intuitively, $\mathrm{ERR}_p^{\text{train}}$ becomes smaller as p increases. Discuss what you observe.

Bonus question Assume now that not all the data from the system are available immediately, but arrive at regular time intervals. We want to maintain in real time a model that approximates the dynamics of our system and improve it when new data are provided. Explain in one sentence why your approach for solving the least square problem here is probably sub-optimal.

Practical information

The homework solution should be written in English.
Please, send it by email to zheming.wang@uclouvain.be, julien.calbert@uclouvain.be and guillaume.berger@uclouvain.be with as object: `[LINMA2380] - Homework 2 - Group XX`, and with the same name for the pdf file containing your solution *(failure to respect these guidelines may induce negative points)*.

Deadline for turning in the homework: Monday 02 November 2020 (11:59pm).

It is expected that each group makes the homework individually. If your group has problems or questions, you are welcome to contact the teaching assistants (see emails above).