Worksheet 5, Oct 8, 2020

Least squares and infections disease

Let us assume an infectious disease with the following reported new infections I_i on each day t_i, for $i = 1, \ldots, 10$. Using least squares fitting, we would like to understand the nature of this

Table 1: Number of new infections I_i on days t_i.

t_i:	1	2	3	4	5	6	7	8	9	10
I_i:	14	20	21	24	15	45	67	150	422	987

growth. We consider two models to describe the connection between time (i.e., days) t and the number of new infections, both with 3 unknown parameters (a, b, c):

$$I(t) = a + bt + ct^2 \qquad \text{(polynomial model)}$$

$$I(t) = a + bt + c\exp(t) \qquad \text{(exponential model)}$$

Our goal is to figure out which model describes the progression of the infections better, and we use least squares fitting to figure that out. Note that if a model would fit the data perfectly, $I(t_i) = I_i$ for all i. In general, you will not be able to find parameters that satisfy this, and thus have to use least squares fitting (sometimes this is also called *regression*).

1a. Formulate, assuming the polynomial model, the least squares problem for the parameters $\boldsymbol{x} = [a, b, c]^T$ by specifying the matrices A and the vector $\boldsymbol{b}$:

$$\min_{\boldsymbol{x} \in \mathbb{R}^3} \|A\boldsymbol{x} - \boldsymbol{b}\|_2^2$$

1b. Same as above, but for the exponential model.

1c. Use a QR-factorization in MATLAB or Python to solve these problems and plot the data as points, as well as the model as a line. Repeat using the normal equations $A^T A\boldsymbol{x} = A^T\boldsymbol{b}$.

1d. To decide which model describes the data better, we need to compute the distance between the model and the data points. Take a look at the proof from class for how the QR factorization can be used to solve least squares problems. In particular, we found that:

$$\|A\boldsymbol{x} - \boldsymbol{b}\|_2^2 \geq \|\boldsymbol{b}_2\|_2^2,$$

where $\boldsymbol{b}_2 = \hat{\hat{Q}}^T\boldsymbol{b}$. We also found that this inequality is an equality if $\boldsymbol{x}$ solves the least squares problem. Thus, the norm or $\boldsymbol{b}_2$ is a measure of how well the model fits the data. Use this to decide which of the two models above describes the data better.

2. We would like to reflect vectors in $\mathbb{R}^3$ on the plane orthogonal to $\boldsymbol{v} = [1, 1, 1]^T$. Compute the corresponding Householder reflection matrix $H = H(\boldsymbol{v})$, and use it to compute the reflection of the 3rd unit vector $\boldsymbol{e}_3$ on the plane normal to $\boldsymbol{v}$.