Numerical Analysis 1 – Class 10

Thursday, March 25th, 2021

Subjects covered

- Regression – least squares fitting 1D – minimizing fit error and the normal equations
- Least squares fitting in ND – Normal equations and projection operators
- Regresssion using QR.
- Polynomial regression.
- Application area: Chemometrics.

Reading

- “Multiple Linear Regression Analysis: A Matrix Approach with Matlab”, Scott Brown. (Linked on Canvas.)
- "The Singular Value Decomposition and the Pseudoinverse", G. Gregorcic. (Linked on Canvas)

Problems

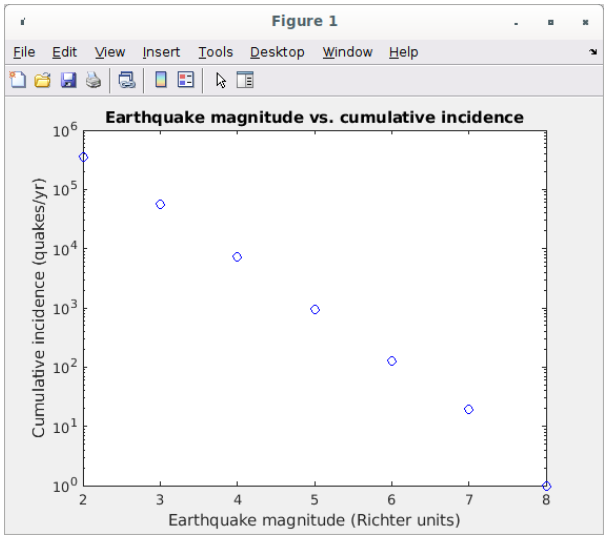
Most of the following problems require you to write a program. For each program you write, please make sure you also write a test which validates your program. Please use Canvas to upload your submissions under the “Assignments” link for this problem set.

Problem 1

The Gutenberg-Richter law relates the magnitude of earthquakes to the frequency at which they happen (i.e. the number of earthquakes in a given time period). The law may be expressed as

$$N(M)=10^{a-bM}$$

where M is the magnitude of the quake in Richter units and N is the number of earthquakes per time period of magnitude M and larger. The parameters a and b are fitting coefficients. The b parameter is of particular interest to scientists. Real-world values of b generally lie close to 1.0 but can vary depending upon geography, the geology of the area (i.e. type of underlying rock), and other effects.



I have placed a file of real earthquake data onto Canvas called “RawData.csv”. Your job is to write a program which will read the datafile (perhaps after some cleaning) and then do linear regression to find the b coefficient. Please use the normal equations to do the fit (i.e. please don’t use any Matlab built-in functions for regression). Note that the data in the datafile is “binned”, so you need to first create cumulative values for N.

Regarding testing, I recommend you validate your program’s correctness by creating some synthetic data with known a and b coefficients and verify your regression program returns the known values.

Problem 2

In class I derived the projection operator $P_A = A\,A^T$, with $A = \begin{pmatrix} \vdots & \vdots & \\ u_1 & u_2 & \cdots \\ \vdots & \vdots & \end{pmatrix}$, which maps an arbitrary vector v onto the subspace subtended by span(A). This expression for P_A is valid only when the vectors u_i form an orthonormal set.

In class, I also asserted that when the u_i are not orthonormal, the projection operator is given by

$$P_A = A\left(A^T A\right)^{-1} A^T$$

In this case, the stuff inside parenthesis $\left(A^T A\right)^{-1}$ acts like a normalization factor, and the stuff outside parenthesis $A(\cdots)A^T$ corresponds to the projection operator above. Please prove this assertion. That is, show that the full expression for the projection operator in the case where the u_i do not form an orthonormal set is

$$P_A = A\left(A^T A\right)^{-1} A^T$$

Hint: Consider a QR factorization of *A*. You may assume *A* is non-singular.

Problem 3

Recall the normal equations which are encountered as part of solving the linear least squares problem. They may be written as $A^T(Ax-b)=0$. Please prove the following theorem:

- If x satisfies $A^T(Ax-b)=0$, then x solves the least squares problem, i.e. x minimizes the 2-norm of the residual, $\|b-Ax\|_2$.

Hint: consider the 2-norm $\|b-A(x+y)\|_2$ for arbitrary vector y and show that

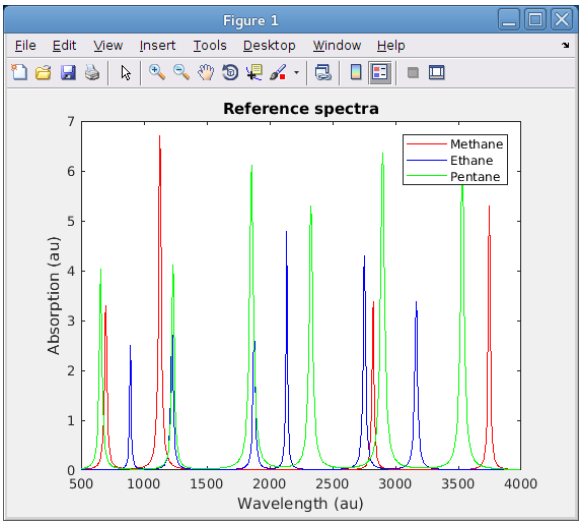
$$\|b-A(x+y)\|_2 \geq \|b-A\,x\|_2$$

for all y.

Problem 4

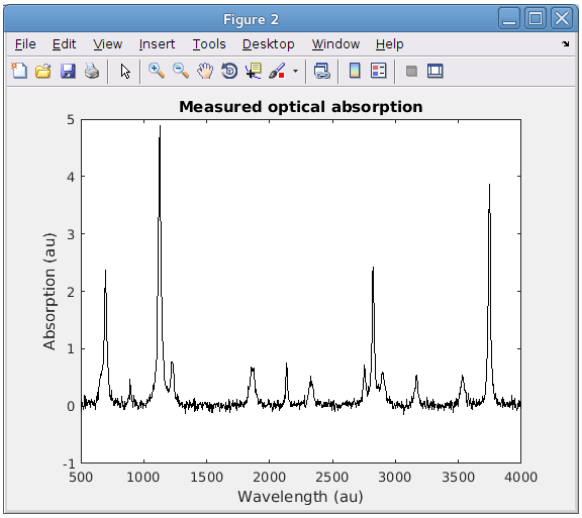
As you learned in the lecture, a big application area for regression techniques lies at the intersection of Chemistry and Spectroscopy. This sub-discipline is called “Chemometrics”, and is involves using light absorption to measure the concentration of chemicals in a sample. In this problem, you will write a program implementing multivariate linear regression to compute the concentration of some gasses given a measured absorption spectrum.

Your sample gas is assumed to contain only three optically active species: Methane, Ethane, and Propane. You know the absorption spectrum of each gas (at 100% concentration). A plot of the three reference spectra is shown below.



Each component has a wavelength-dependent spectrum denoted by $R_i(\nu)$, where i indexes the three different species, and v is related to the light's wavelength.

You are given an optical absorption measurement of a sample of unknown gas. The measured spectrum of the mixture is shown below.



You know a priori that the unknown mixture is composed of a weighted combination of the three known reference gasses and an inert gas which is not optically active (nitrogen). That is, you know the measured spectrum S(v) may be written as

$$S(\nu)=c_{meth}R_{meth}(\nu)+c_{eth}R_{eth}(\nu)+c_{pro}R_{pro}(\nu)$$

where the c coefficients are the concentrations of each species. Because they are concentrations, the c coefficients each obey $0\le c_i\le 1$. Also, since the total concentration of all gasses cannot be larger than 1, we have $c_{meth}+c_{eth}+c_{pro}\le 1$.

Note that the expression for $S(\nu)$ means the problem of finding the concentrations c_i from a measured spectrum $S(\nu)$ is a linear regression problem. The reference spectra $R_i(\nu)$ are basis func-

tions, and concentrations c_i are the fit coefficients.

I have placed a zip file on Blackboard with the reference spectra and an unknown sample spectrum. Please write a program which performs linear regression on the unknown spectrum and reports the concentrations (c coefficients) of the three different gasses present in the mixture.