# Numerical Analysis 1 – Class 10

Friday, March 24th, 2023

## *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

This is a easy problem which tells a cautionary tale about extrapolating a model. Consider the following model for the world population growth over the last few centuries:

$$\frac{dp}{dt} = K p^2 \tag{1}$$

where $p$ is the world population in millions, and $K$ is a constant. The idea is that the population grows super-exponentially, perhaps due to technological progress or increased longevity or some other effect. Equation (1) has a solution of the form

$$p = \frac{1}{K}\frac{1}{t_0 - t} \tag{2}$$

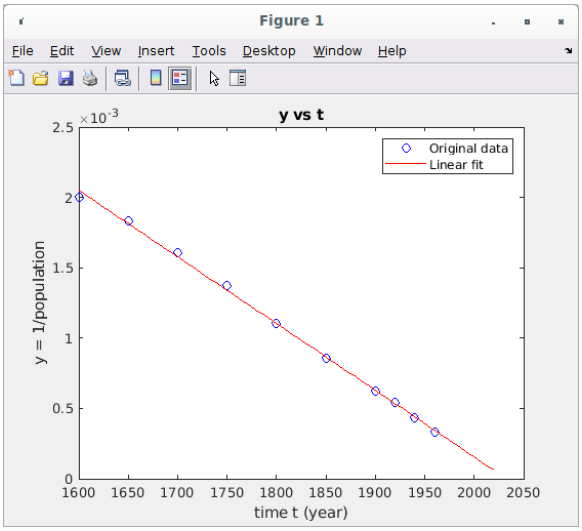
for some constant $t_0$ .

Here is some historical data for the world population:

| **t (year)** | 1600 | 1650 | 1700 | 1750 | 1800 | 1850 | 1900 | 1920 | 1940 | 1960 |
|---|---|---|---|---|---|---|---|---|---|---|
| **p (millions)** | 500 | 545 | 623 | 728 | 906 | 1171 | 1608 | 1834 | 2295 | 3003 |

Please do the following:

- Using pencil and paper, please verify that (2) is a valid solution of (1). Turn in your derivation.
- Consider change of variables $y = 1/p$ . Write down the expression for $y$ vs. $t$ . Make a plot of $y$ vs. $t$ . Your plot should evidence a linear relationship between $y$ and $t$ .
- Since $y$ depends linearly upon $t$ you can fit a line and extract $K$ and $t_0$ . Please use the normal equations to do the linear fit. Please add your fit to the plot and report your extracted values. My plot is shown below.




- Please also make a plot of the original variable $p$ vs. $t$ as well as your fitted function.
- Now consider the world population over the last forty years. Here is the data:

| **t (year)** | 1980 | 1990 | 2000 | 2010 | 2020 |
|---|---|---|---|---|---|
| **p (millions)** | 4458 | 5327 | 6143 | 6956 | 7794 |

Please add this data to your plot of $p$ vs. $t$ . Did the model accurately predict the world population after 1960? What (if anything) is wrong with the model?

- Regarding testing, for a regression program the best way to test is to first compute the RMS error of your model when evaluating the input data. Then take the fit parameters and perturb them and compute the RMS error again. The RMS error computed when using the perturbed parameters should be larger than when using the actual fit parameters – if your fit is correct.

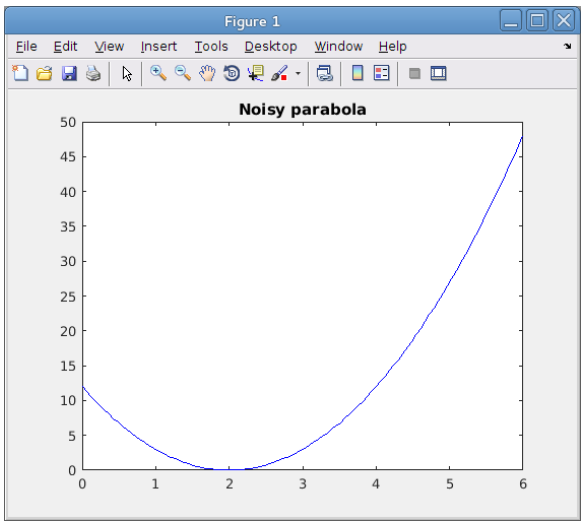
### Problem 2

Taking derivatives of real data is an important application of the Savitzky-Golay filter. A major reason for this is that the simplistic finite difference formulas for derivatives do not work well when used on real, noisy data. This problem explores the topic of taking derivatives of data.

Consider taking the second derivative of the data shown in the figure below. The data is a simple parabola corrupted by additive white Gaussian noise. Note that although the parabola in the figure doesn't look noisy, it is noisy enough that computing derivatives is problematic, as we will see.

The equation for this parabola is

$$y(t)=3(t-2)^2$$

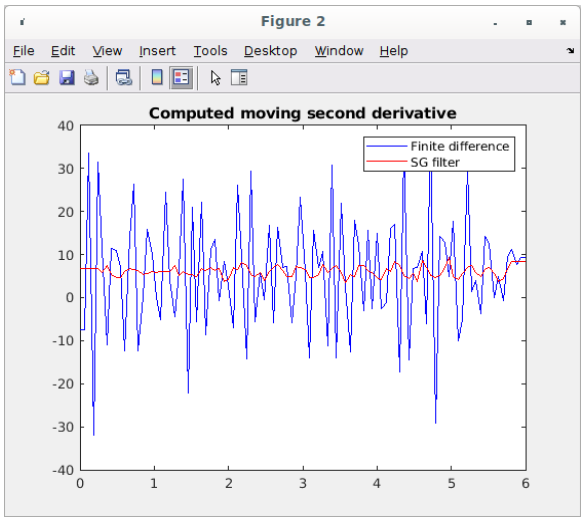



We will compute a moving second derivative of this parabola using a 7-point, 2nd order Savitzky-Golay filter, and then compare the result to a second derivative obtained using the usual finite-difference formula

$$(y'')_n=\frac{y_{n+1}-2y_n+y_{n-1}}{(\Delta t)^2}$$

Please do the following:

- Using the method shown in class, derive the coefficients for a 2nd derivative Savitzky-Golay filter for a polynomial of degree 2 fitting 7 points . Assume equally-spaced points.
- Write a program which returns the filter coefficients (i.e. the kernel). You can check your result against any of the online tabulations of the Savitzky-Golay filter coefficients.
- Now write a program which reads the parabola data out of the CSV file on Canvas, computes the moving second derivative using your Savitzky-Golay program, and plots the result. The CSV file holds the data as [t, y] pairs on each row.
- Now compute the second derivative using the finite difference formula (3). Plot your result in the same plot as above. My result is shown at right.



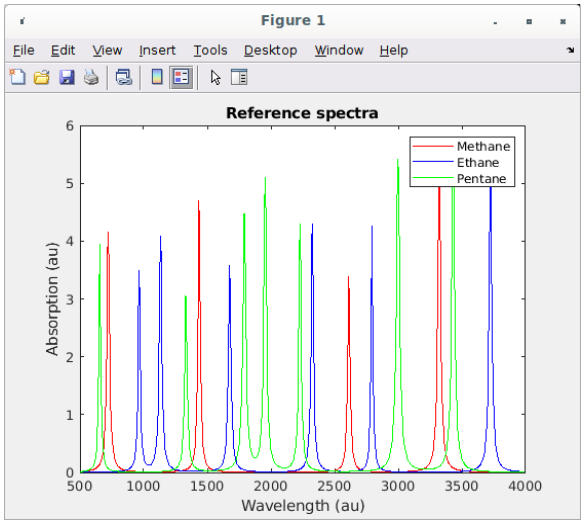

- To test your result, use the formula for the original parabola (above), extract its second derivative, and compare the analytic second derivative against the second derivative computed by your program.

Obviously, if you have noisy data, the second derivative computed using the Savitzky-Golay filter is preferable to that computed using the finite difference formula (3).

## Problem 3

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.

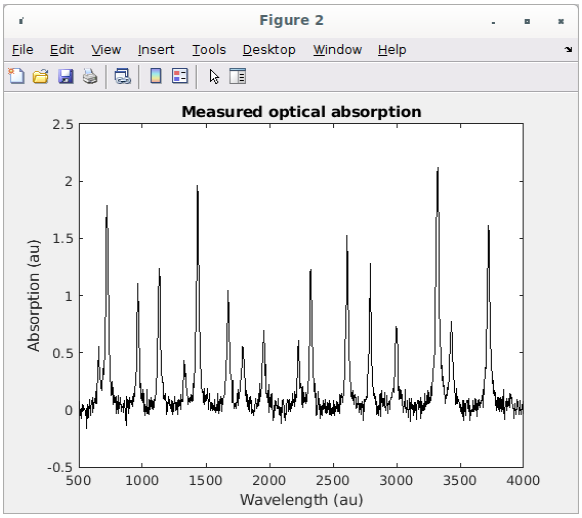
Your sample gas is assumed to contain only three optically active species: Methane, Ethane, and Propane.



You know the absorption spectrum of each gas (at 100% concentration). A plot of the three reference spectra is shown at right above.

Each component has a wavelength-dependent spectrum denoted by $R_i(\nu)$, where i indexes the three different species, and ν 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(ν) 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 functions, and concentrations $c_i$ are the fit coefficients.

I have placed a zip file on Canvas 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.