AMATH 301 – Autumn 2019
Homework 6

Due: 7:00pm, November 20, 2019.

Instructions for submitting:

- Scorelator problems are submitted as a MATLAB script (`.m` file). You should **NOT** upload your `.dat` files. The `.dat` files should be created by your script. You have 5 attempts on Scorelator.
- Writeup problems are submitted to Gradescope as a single `.pdf` file that contains text and plots. Put the problems in order and label each writeup problem. When you submit, **you must indicate which problem is which on Gradescope**. **All code you used for this part of the assignment should be included either at the end of the problem or at the end of your .pdf file.**

Scorelator problems

1. Download the file `salmon_data.mat` included with the homework. Make sure this file is in the same folder as your .m file. This file contains the annual Chinook salmon counts taken at Bonneville on the Columbia river from the years 1938 to 2017 (*www.cbr.washington.edu*). In this problem we will fit this data with a curve to predict salmon populations.

   (a) Load `salmon_data.mat` into your workspace. **You do NOT need to upload this file to Scorelator. Your code will be tested using the counts for another species of salmon.**

   (b) In the video lectures, you learned that the following matrix equations could be used to determine the coefficients of a linear best fit:

   $$\begin{pmatrix} \sum_{k=1}^{N} t_k^2 & \sum_{k=1}^{N} t_k \\ \sum_{k=1}^{N} t_k & \sum_{k=1}^{N} 1 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^{N} t_k y_k \\ \sum_{k=1}^{N} y_k \end{pmatrix}.$$

   Here t_k is the k-th year and y_k is number of salmon at year t_k. The solution provides c_1 and c_2 such that $y = c_1 t + c_2$ is the RMS best-fit line. The above problem can be written as $A\boldsymbol{x} = \boldsymbol{b}$. To construct A and $\boldsymbol{b}$, you may want to use the `sum` command in MATLAB.

   Construct the matrix A and save it to `A1.dat`. Construct the vector $\boldsymbol{b}$ and save it to `A2.dat`.

(c) Solve the above system for c_1 and c_2. Save c_1 and c_2 as a 2×1 column vector with c_1 in the first component and c_2 in the second component. Save this vector to `A3.dat`.

(d) Use `polyfit` to find the best-fit polynomials of degree 3, 5, and 8. Save the coefficients for these polynomials in `A4.dat`, `A5.dat`, and `A6.dat` respectively.

(e) You now have four models for predicting the number of salmon in a given year: a degree-1 polynomial, a degree-3 polynomial, a degree-5 polynomial, and a degree-8 polynomial. Call these three polynomials p_1, p_3, p_5, and p_8 respectively. You will plot these in Gradescope problem 1.

The number of salmon in 2018 was 336030. For each of the four polynomials, find the absolute error between what your polynomial predicts is the number of salmon in 2018 and the true number of salmon. In other words, calculate

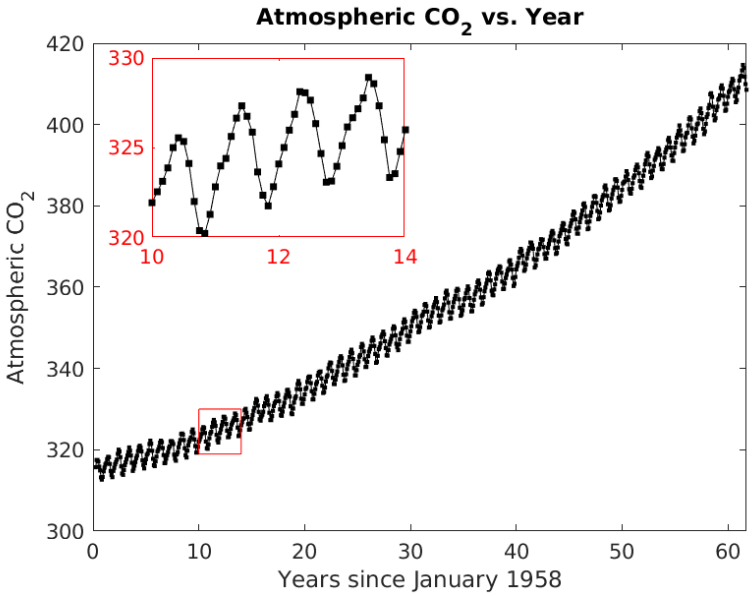
$$\begin{aligned} \text{err}_1 &= |p_1(2018) - 336030|, \\ \text{err}_2 &= |p_3(2018) - 336030|, \\ \text{err}_3 &= |p_5(2018) - 336030|, \\ \text{err}_4 &= |p_8(2018) - 336030|. \end{aligned}$$

Create a 1×4 row vector with the components

$$[\text{err}_1,\ \text{err}_2,\ \text{err}_3,\ \text{err}_4].$$

Save this vector to `A7.dat`.

2. The amount of CO_2 in the atmosphere is regularly measured at the Mauna Loa observatory in Hawaii. The file `CO2_data.mat`, which is included with the homework, contains the monthly averages since 1958. A plot of the data is shown below. The data has an overall upward trend as well as seasonal oscillations.



Data from:

1. `https://www.esrl.noaa.gov/gmd/ccgg/trends/data.html`

(a) Load this data into MATLAB using the `load` command. Be sure that the file `CO2_data.mat` has been downloaded into the same directory as your script file. **You do NOT need to upload this file to Scorelator. Scorelator has its own copy.** If the `load` command is successful, you will have two new vectors in your workspace, `t` and `CO2`. The values of the vector `t` are the number of years since 1958 corresponding to each month from March 1958 to October 2019. So the first value is `t(1) = 3/12` because March 1958 is the third month (out of 12) since the beginning of 1958. The vector `CO2` has the corresponding CO_2 levels.

(b) It looks like the overall trend of the data might be captured well by using an exponential fit. Recall from Video 13 that this can be done by data linearization. However, you can do a better job by adding a constant to the exponential. We seek a model of the form

$$y = ae^{rt} + b.$$

The data linearization trick does not work for a function of this form. Instead, you must create a MATLAB function that takes the values of a, r, and b as inputs and calculates the sum of squared errors as output. Then you can use `fminsearch` to find the values of a, r, and b that minimize the sum of squared errors (and therefore minimize the root-mean squared error). Use this method to find the best fit curve of the form $y = ae^{rt} + b$. For `fminsearch`, use initial guesses of $a = 30$, $r = 0.03$, and $b = 300$. Make a 1×3 row vector with the optimal parameters `a; r; b`, and save this vector in `A8.dat`. You will plot this curve in Gradescope problem 2.

(c) This best fit curve matches the overall trend well (again, I recommend plotting it!), but it still does not capture the seasonal oscillations. In order to capture the oscillations, we can find a best fit curve of the form

$$y = ae^{rt} + b + c\sin(d(t - e)).$$

Create an error function and use `fminsearch` to find the values of the parameters that minimize the sum of squared errors. For your initial guess, use the values of a, r, and b that you found in part (b) and use $c = -5$, $d = 4$, and $e = 0$. Make a 1×6 row vector with the optimal parameters `[a; r; b; c; d; e]`, and save this vector in `A9.dat`. You will plot this curve in Gradescope problem 2.

3. Which of the models, p_1, p_3, p_5 or p_8, gives the most reasonable estimate for the number of salmon in the year 2045? Here "reasonable estimate" means that the estimate is physically possible and/or does not give a number of salmon more than twice as big as the largest number of salmon in our data set. Save your answer to `A10.dat`.

A p_1

B p_3

C p_5

D p_8

Gradescope problems

1. This problem mirrors Scorelator problem 1. You need to turn in the plot in part (a), the responses from part(b)-(d), and all of the code used for this problem.

   (a) Create a plot of the salmon data from Scorelator problem 1 along with the four functions created in that problem, p_1, p_3, p_5 and p_8. Your plot should have the following features:

      i. The data should be plotted as black circles.

      ii. Your plot should be from 1930 to 2020 on the x-axis and from 150,000 to 1,500,000 on the y-axis.

      iii. p_1 should be in blue, p_3 in red, p_5 in green, and p_8 in magenta. These lines should have linewidth 2.

      iv. Your plot should have a legend.

      v. Label the x-axis with "Year" and the y-axis with "Salmon".

   (b) What is the real-world interpretation of the slope of the line of best fit for the Salmon data? What does it tell us about how the population is changing?

   (c) The coefficients of p_8 are very large. In particular, the y-intercept of this polynomial is very large. Describe why this is the case in a few sentences.

   (d) Of the different polynomials you tried in Problem 1, which gave the most accurate prediction of the 2018 salmon population and which gave the least accurate? If you had to predict the population in 2050, which polynomial fit would you trust the most? Justify your answer.

2. This problem mirrors Scorelator problem 2. You need to turn in the plot in part(a), the answers to parts (b)-(d), and all of the code used for this problem.

   (a) Create a plot that contains the atmospheric CO_2 data and your two curves of best fit from Scorelator problem 2. Make sure to plot with enough data points to capture the oscillations.

      i. The data should be plotted as black dots connected by black lines by using the line specification `'-k.'`.

      ii. Your plot should show from $t = 0$ to $t = 65$.

      iii. The exponential fit from Scorelator problem 2 b should be plotted as a red curve and it should have line thickness 2.

      iv. The exponential + sinusoidal fit should be plotted as a blue curve and it should have line thickness 2.

      v. Label the x-axis with "Years since January 1958."

      vi. Label the y-axis with "Atmospheric CO_2."

      vii. Include a legend.

   (b) In Scorelator problem 2 you calculated the root-mean square error for the two curves. Record the error for each of the methods and compare the error. Which of the two methods give smaller error?

(c) Which model is better for predicting atmospheric CO_2 for November 2019?

(d) Which model would you prefer for predicting average CO_2 levels in 2040? Is one significantly better than the other for this application?