Homework 5

Problem 1: Salmon Population

The file `salmon.mat` (included with the homework) contains the annual Chinook salmon counts taken at Bonneville on the Columbia river from the years 1938 to 2017. (You can find more information at `www.cbr.washington.edu`. The data is from the DART adult passage annual summary.) You can load this file into matlab with the command `load('salmon.mat')`. **Do not upload this file to scorelator. Scorelator has its own copy of salmon.mat, with counts from a different species.** If the load command is successful, you will have two new vectors in your workspace: `t` and `salmon`. The vector `t` represents years since 1938, so $t = 0$ is the year 1938 and $t = 79$ is the year 2017. Each entry in `salmon` represents the number of Chinook salmon that passed through Bonneville in the corresponding year. In this homework, we will try many methods to predict how many Chinook salmon will pass through Bonneville this year ($t = 80$). (At the time of this writing, two salmon have already made it through.)

You should plot the data alongside every curve that we fit to it. This will not only give you some insight into salmon population dynamics, but also provides a good sanity check as to whether or not your code is working correctly.

(a) Find the best fit line for this data. That is, find a line $y = mt + b$, where t is the time in years and y is the salmon population. Your line should be "best" in the sense of minimizing root-mean-square error. Save the slope of this line in `A1.dat`. Now use this formula to predict the population of salmon in 2018 and save the predicted population in `A2.dat`. Find the root-mean-square error of your prediction and save it in `A3.dat`.

Things to think about: Is the slope positive or negative? What does this tell you about the salmon population dynamics? Has the salmon population increased or decreased over the last three years? Is that trend reflected in the best fit line?

(b) Now find the best fit quadratic for this data (i.e., the best fit polynomial of order 2, where "best" is in the root-mean-square sense). Use this curve to predict the salmon population in 2018 and save this prediction in `A4.dat`. Repeat this process for the best fit polynomials of orders 5 and 20. Save your predictions in `A5.dat` and `A6.dat`, respectively.

Things to think about: What are the root-mean-square errors for each of these fits? Which would you most trust to predict this years population? To see what Matlab means by the warning "polynomial is badly conditioned", try this problem with a 25th order polynomial (or even higher). Some of what you are seeing is the polynomial wiggle phenomenon that we discussed in class, but most of it is (unavoidable) rounding error.

(c) It is often reasonable to model population growth with an exponential equation of the form $y = N_0 e^{rt}$. Here, N_0 is the population at time 0 and r is the population growth rate. Transform this data by taking the natural log of the population values (as discussed in class and the videos) and then use a linear fit of the transformed data to calculate N_0 and r. Store these two values as a 1×2 row vector $[N_0,\ r]$ and save this row vector in `A7.dat`. Use the exponential model to predict the salmon population in 2018 and save your prediction in `A8.dat`.

Things to think about: Do you get the same parameters if you use `fminsearch` to find N_0 and r directly? Why do you think this is/is not? What did you get for a root-mean-square error? What does this tell you?

(d) Suppose that a biologist proposes the following function as a model for salmon population growth:

$$y = e^{at^2+bt+c},$$

where a, b and c are parameters. Write a function to compute the root-mean-square error between this curve and the population data `salmon` for any given values of a, b and c. Use `fminsearch` to find the parameters that minimize this error (using an initial guess of $a = 0.001$, $b = -0.01$ and $c = 10$), then use this optimal set of parameters to predict the chinook population in 2018. Save your prediction in `A9.dat`.

Things to think about: What was the root-mean-square error? What does that tell you about the fit? How does this error compare to that from part (c)? How are these two models related? Do you think this is a biologically plausible model?

(e) It is not unreasonable to think that the salmon populations over the last few years are much more relevant to our prediction than the population from 1938. One way we could focus on more recent data is to use a cubic spline. Use a spline of the data to predict the salmon population in 2018. Save your prediction in `A10.dat`.

Things to think about: Do you think it is wise to use a spline for extrapolation? Can you think of a better extrapolation method that primarily uses the more recent data?