ROB 101: Computational Linear Algebra Project Regression: Precipitation in Alaska (A True Story)

Kira Biener, Madhav Achar, Tribhi Kathuria, Maani Ghaffari, and Jessy Grizzle University of Michigan Robotics Institute

This problem set counts for 20% of your course grade. You are encouraged to work as a group and divide non-trivial code or solution steps amongst yourselves.

Submission Instructions

Post in GitHub and submit link - just like we do for your Julia Homework.

Notebooks vs. PDF

The notebooks are meant to be self-explanatory, and this PDF should be looked at as a supplement, pro- viding extra information that is not in the notebook, and might help better understand the problem.

Objective

The objective of this course is to get you to deal with huge amounts of data and manipulate it in a meaningful manner using Matrix algebra. In this project, you will get a chance to see how the none-obvious problem of surface fitting can be solved with linear algebra

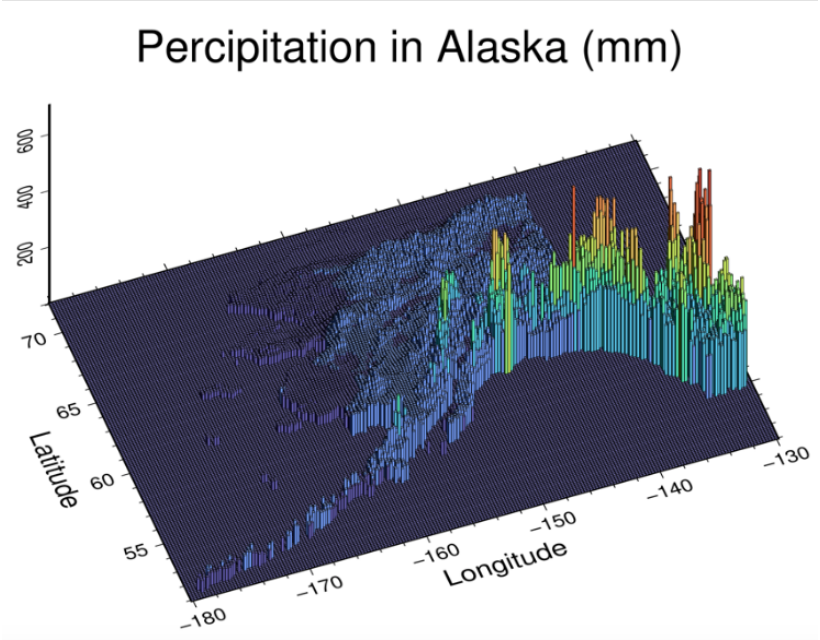

Figure 1: Precipitation in Alaska during July 2020.

1 Problem Statement

Linear Regression is a powerful tool employed in fields ranging from business to arts to explain data and understand trends. There are two main categories of tasks that can be set up as regression problems:

1. Factor Analysis: The objective of such studies to find the correlation between different data points in available data sets. Example Problem: Suppose you have data for different cars in a particular area of the US; you must analyze the data and find out what factors are more important to people when they are buying cars and build a pricing model. So a Car-selling company will hire you to look at the data and find these factors for them. Relation to Regression: How do you set this up as a regression problem then? Well, we have been able to look at our data x and predict y = f (x) setting up the system of equation as $\hat{Y} = \Phi * \alpha$. Simplifying the further analysis, we can naively say that if from our plots we could accurately predict the car price using a combination of certain data factors (x), then those factors should be important in our pricing strategy. So you might find things like people care about the colors of their cars only if they are high end, and so on.

2. Prediction: Prediction on the same car-pricing model problem would involve drawing inference on your model. You are given a new car and asked to predict its optimal price, given the value of factors like brand, type, color, age, etc. Relation to Regression: You will solve the problem of predicting the Precipitation in Alaska in this Project, shown in Figure 1, and see the relation for yourself!

If you find this problem interesting, you should try and compete in challenges at Kaggle. We picked this example problem right from there; you will find many more there. Although you should realize that Linear Regression is only a foot in the door, machine learning is a vast subject in itself, and if you find yourself interested in it, you will greatly benefit from learning Linear Regression well! For this project, you will use linear regression methods that you have learned in previous assignments and apply it to a much larger dataset. The dataset we will be using is from the U.S. National Oceanic and Atmospheric Administration (NOAA) [1] and contains information about the amount of precipitation that occurred during July 2020 in Alaska. An important takeaway in this project is that only how the basis functions [2] are combined must be linear in a linear regression problem. The basis functions themselves can be nonlinear. At the end of this project, you will have learned how to take a specific nonlinear basis function called the radial basis function and apply it towards surface regression.

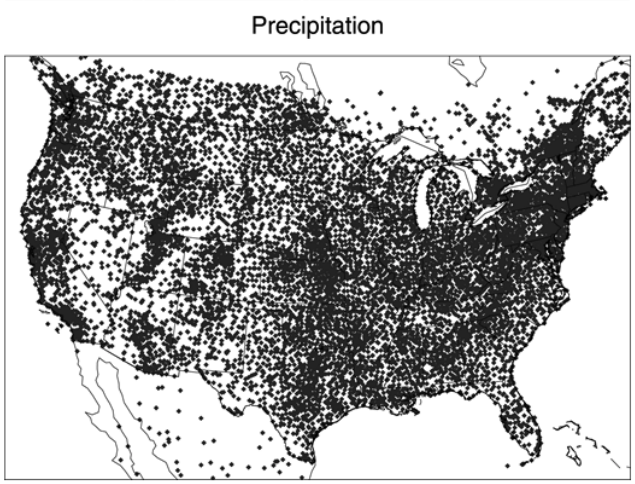

Figure 2: Map of the locations of the weather stations reporting precipitations data.

1.1 Project 2

NOAA's full dataset holds precipitation information in Alaska and the continental US at 5km increments both in the longitudinal and lateral direction. In reality, there is not a weather station located in the US 5km a part in every direction. The data presented is itself the result of regression efforts to estimate the precipitation across the US accurately. See the paper summarizing the efforts for more details [3]. Figure 2 shows the location of the weather stations reporting precipitation. We can see that they are not evenly spaced out. That, along with stations located at a variety of altitudes, geographical barriers, and other fac- tors reduce the quality and accuracy of a precipitation estimate based purely on the average precipitation reporting of neighboring weather stations. A lot of critical research is based on accurate climate maps, which motivates using regression techniques to create high-quality models. In the project, we will try to mimic the efforts done to produce the map. We will sample a small portion of the precipitation measure- ments in the original dataset and build a model that can produce a precipitation estimate at any queried longitude/latitude coordinates within Alaska.

[1]Vose, Russell S., Applequist, Scott, Squires, Mike, Durre, Imke, Menne, Matthew J., Williams, Claude N. Jr., Fenimore, Chris, Gleason, Karin, and Arndt, Derek (2014): Gridded 5km GHCN-Daily Temperature and Precipitation Dataset (nCLIMGRID), Ver- sion 1. 202007.prcp.alaska.pnt. NOAA National Centers for Environmental Information. DOI:10.7289/V5SX6B56 Aug 1, 2020.

[2]In HW, you used the monomials as basis functions. You sought to express a value y as $\hat{y} = \alpha_0 + \alpha_1 x + \cdots + \alpha_m x^m$. In this case, the functions x^k , $0 \le k \le m$ are the basis functions. And note, they are written as a linear combination!

[3] https://journals.ametsoc.org/jamc/article/53/5/1232/13722/Improved-Historical-Temperature-and-Precipitation

2 Your Tasks

This project has two separate notebooks: p1_basis_functions.ipynb and p2_alaska.ipynb. The first notebook, p1_basis_functions.ipynb, serves as an introduction to the radial basis function. In the notebook, you will follow along with an example showing the limitations of using only monomials as your basis functions and how the radial basis function addresses some limitations. Then, in the second notebook, you will use your understanding of the radial basis and the least squares pipeline that we provide to perform a surface regression. For each task, make sure you read/run/edit all of the notebook cells until the task before starting it as there are information and definitions in the notebook not introduced in here.

Task 1: Monomial Fitting (4 points) Prerequisites

• Review section 8.3 and 9.7 in the textbook.

In this task, you will create regressor matrix Φ that will be used to find the optimal coefficients for a monomial-based model (Eq. (1)).

$$\hat{y} = a_1 + a_2x + a_3x^2 + a_4x^3 \quad (1)$$

Once the matrix is constructed, we can use our least squares pipeline to solve our model's optimal coefficients. Equation (2) below is the exact problem definition. For examples of how to construct the matrix, refer to sections 8.3 and 9.7 in the textbook. To pass this task, run the auto-graded cell directly beneath the cell you modify in the notebook. If there is no error displayed, you have correctly built the matrix.

$$\alpha^* = \arg\min_{\alpha} \|Y - \Phi\alpha\|^2 \quad (2)$$

Task 2: Fitting with Radial Basis Functions (2+2 points)

Part A: Hand-tuned parameters

In this part, you will try to manually fit the initial function in the first notebook by using a model composed of just two Radial Basis Functions (RBF) (Eq. (3)). Figure 3 shows an image of one part of the initial function fit with a single RBF.

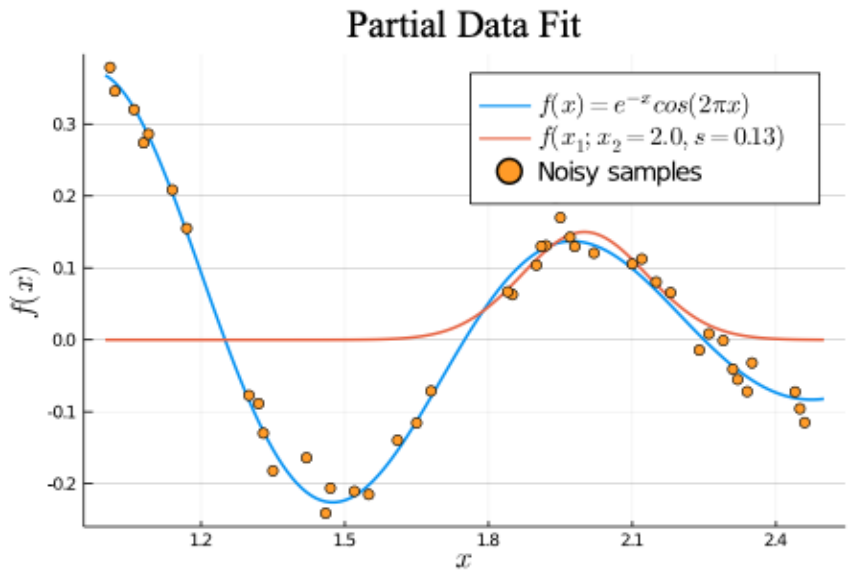

Figure 3: Example process fitted with a single radial basis function from p1_basis_functions.ipynb.

This task aims to help you develop an intuition for adjusting the different parameters of the RBF that affect the overall model. As you tune the parameters, observe the overall fitting error displayed in the cell and how it compares to the fitting error of the initial monomial model. To pass this task, you will need to produce an overall fitting error with the two RBF models less than 0.5.

$$\hat{y} = a_1 f(x; x_{c_1}, s) + a_2 f(x; x_{c_2}, s) \quad (3)$$

Part B: Fitting using Least Squares and RBFs

In this part, you will take the knowledge you gained in Part A and use it to define a larger RBF-based model. The model in (4) will have a constant offset term and a variable number of RBFs based on a parameter M that you will set. You will also need to set the parameter s in the same cell that defines all RBFs' scale in the model. To pass this task, you will need to fit a model that produces an overall fitting error lower than the initial monomial fitting error.

$$\hat{y} = a_1 + a_2 f(x; x_{c_1}, s) + a_3 f(x; x_{c_2}, s) + \ldots + a_{M+1} f(x; x_{c_M}, s) \quad (4)$$

Warning: It is very tempting to say, hey, I will use M = 200 basis functions! That will fit wonderfully! Go ahead and try it. What you will find is that the columns of your regressor matrix Φ become "nearly" linearly dependent, and hence $\Phi^T\Phi$ is not invertible. When you do the QR Factorization, you will find that R has tiny numbers on its diagonal. You should check that out! Here is some code (don't miss the dot after the abs command!):

```
diagMin=minimum(abs.(diag(R)))
```

After your back substitution algorithm blows up, because you are almost dividing by zero, you'll come back here and thank us for this warning!

Task 3: Build a Model to predict the precipitation in Alaska (4+3+2+3 points)

Part A: Modify the Helper Functions

With the context of the actual problem set in the cells in the second notebook preceding Task 3, we now begin on our main regression problem. Take a moment to scan through all of the cells in the section. You will notice that there are 5 functions, all of them introduced in the previous notebook: rbf, backwardsub, calc_phi_row, regressor_matrix and least_squares_qr.

The functions are implemented as they were in the first notebook and are provided in a compiling state. However, as the beginning of the second notebook explained, our data has grown in dimension. The points that we evaluate with our model belong to R^2 instead of R. Specifically, we will send in (longitude, latitude) coordinate pairs to our model, and it will provide the estimated precipitation at that coordinate. Thus, you must identify and make the necessary adjustments in the code to handle the new data dimension. The part of the code you will need to modify has been left out for you to fill in. Out of the 5 functions only two of them calc_phi_row and regressor_matrix, will need modifications. In Part B, when you build the pipeline, you can test and see if all the necessary changes were made correctly. A good indication of that will be if Julia does not display any dimension mismatch errors. Run the autogrded cell in at the end of the cells to see if you passed.

Part B: Calculate the Model Weights

In the second part of this task, you will utilize the helper functions provided to solve for the optimal set of coefficients, a_star = $[a_1,...,a_{M+1}]^T$, for our model in (5). The basis centers and the radial basis scale s are already defined for you. Do not change these. From our investigation at the beginning of the notebook, we saw that using 1% of the precipitation data provided an even coverage across the state of Alaska, so we use those coordinates as our radial basis centers.

$$\hat{y} = a_1 + a_2 f(x; x_{c_1}, s) + a_3 f(x; x_{c_2}, s) + \cdots + a_{M+1} f(x; x_{c_M}, s) \quad (5)$$

To pass this, task you can verify that the model coefficients you solved for are correct with the auto-graded cell at the very end of the task.

Part C: Inference

In this part of the task, you are given the latitude and longitude of Juneau (a place in Alaska) and asked to predict the precipitation using your model. Check to see if your model predicted 317.4 mm of precipitation.

Part D: Plot the Surface

For the final task, you will use the model we defined in (5) and the coefficients you just obtained for esti- mating the precipitation across Alaska. We will hold this information in a variable called precipitation. At this point, all variables in the model, i.e., $a_1,...,x_{c_1},...$, s, are defined and x is the (lon,lat) pair or query point.

Fill in the code in the indicated cell to calculate the precipitation at a defined (lon, lat) pair and write the value to precipitation[i,j] in the matrix. Pay attention to the double for loop structure already

defined. You only need to write the code for estimating the precipitation for a single coordinate, and the looping structure will take care of making sure that we visit and evaluate all the points in the grid. Once the code is completed, run the final cells in the notebook and visually see the results. It should look similar to Figure 4.

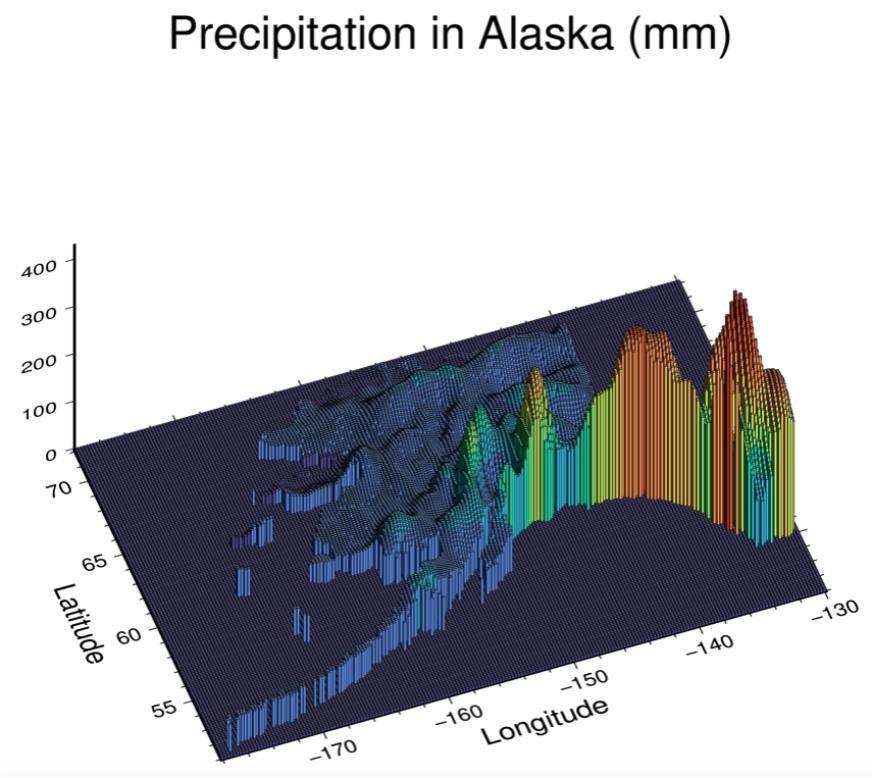

Figure 4: Result of surface regression

Compare the results to Figure 1. which is the regression fit performed by the team who published the dataset. You may notice that your fit seems a little bumpier than their fit, but that is okay. They use a slightly different technique for data interpolation and smoothing known as Thin Plate Splines: https: //en.wikipedia.org/wiki/Thin_plate_spline. If you follow the link, you'll notice though that there is a relation to the radial basis functions that we used ourselves!