Practicum 3: Least squares and data fitting

Efstratios Gavves and Leo Dorst

November 20, 2010

1 Introduction to the least squares method

In the section 5.4 of the book the least squares method is discussed. To provide some context, the least squares method was first described by the 24-years old Carl Friedrich Gauss around 1794. He used least squares to predict the orbit of the asteroid *Ceres*, which was observed by an Italian monk, Giuseppe Piazzi. Gauss's least squares approach made him worldwide famous for solving a problem that was considered to be extremely difficult, if not unsolvable. Following Gauss's paradigm, we will study least squares in the context of astronomy, trying to predict planetary orbits.

For this assignment, you have to know the basics of the least squares method. The theorem 5.4.6 and 5.4.7 from the book are a good start. Also, the examples 3 to 6, p.225, are pretty illustrative and will help you with this assignment.

You have to download the folder *practicum3* from blackboard(unzip the file *practicum3.tar.gz*). The folder contains the matlab files *test_models.m* and *planets.m*. The file *planets.m* is just a script, which instructs you how you could write down the commands so as it is easier for you to run the functions you implemented.

2 Using the least squares method

The basic idea between linear least squares data fitting is shaping the linear equations in the matrix format, $A\vec{x} = \vec{b}$. This will be the target of the first exercise. We first have to assume some model for the data. Without diving into the specifics of the rest of the assignment, let's assume that our model consists of the equation $f(x) = c_1 + c_2x + c_3x^2$. For such a model, we typically have pairs of values $(x, f(x))$ and we want to calculate what are the best(for the current assumptions) estimates $\hat{c_1}, \hat{c_2}, \hat{c_3}$. The model is not linear with respect to x, as it contains c_3x^2. However, the model is linear with respect to the parameters c_1, c_2, c_3, which we are anyways looking for. Therefore, whenever we perform some linear least squares fitting, we imply that the model is linear **with respect to the unknown parameters**.

Back to our problem, we assume that Hubble captured a signal from a different solar system. This signal contains evidence of a similar sun, around which an earth-like planet, the Pandora, revolves. Hubble was able to obtain 20 measurements of the (x, y, t) coordinates, where t where the time instances of the measurements and x, y Pandora's coordinates. The measurements though were noisy, because of the electrical storms that flipped some bits in the transmission. These data can be found in the mat file *observations.mat*.

Having the (x, y) coordinates and being an exceptional computer scientist, you have a handful of different models that could possibly explain the data despite the noise. These models are then

- $f_1(t) = c_1 + c_2t + c_3t^2$
- $f_2(t) = c_1 + c_2t + c_3t^2 + c_4t^3$
- $f_3(t) = c_1 + c_2sin(\frac{2\pi t}{360}) + c_3cos(\frac{2\pi t}{360})$

Exercise 1 *Generate the matrices A, $\vec{b}$, which contain all the observations according to each model. Because you have one independent variable, t and two dependent variables x and y that depend on t, you need two matrices A, one for each variable. Therefore, in the end of this first exercise we will have 6 matrices A and 6 matrices b, 2 for each one of the models. For A the matrices will be the $A_{x_1}, A_{y_1}, A_{x_2}, A_{y_2}, A_{x_3}, A_{y_3}$ and for b the matrices will be $b_{x_1}, b_{y_1}, b_{x_2}, b_{y_2}, b_{x_3}, b_{y_3}$. (1pt)*

Now that we have the observation matrices, we can proceed to the least square data fitting. This fitting will get as input the observations and will output the best fit coefficients, according to the model used.

Take care that you follow the specifications for the `leastSquares` function, that is name, input and output arguments. For one thing, you are CS/AI students, which means that you might not always work on software that is completely written by you and therefore if you deviate by the specifications things might just not work. For another, the code that is provided might assume these specifications, therefore if you do something else, maybe it won't run.

Exercise 2 *Write a function called* `leastSquares`. *The function should take **2 input arguments**, A and $\vec{b}$, and **1 output argument**, $\vec{c}$. The input arguments A and b are the equivalents of the matrices in Definition 5.4.4. The function should solve the linear system and return the parameters according to least squares fitting. (1pt)*

After you write the function `leastSquares`, apply the different matrices A and b from the different models that you found in exercise 1. After using the various matrices you will notice different behaviors depending on the method.

To test the various models, you should use the function `test_models`, which is already provided. `test_models` takes 4 input arguments. The first two input arguments declare the assumed models for the x-variable and for the y-variable. These two arguments can take one of the possible values `'f1', 'f2', 'f3'`. The last two arguments contain the coefficient for the models declared in the first two arguments, in a vector format. For example, for the model `'f1'` the coefficient vector would be $[c_1, c_2, c_3]^T$. An example call of that function would be

```
test_models('f1', 'f3', [c1_f1; c2_f1; c3_f1], [c1_f3; c2_f3; c3_f3])
```

`test_models` plots Pandora revolving around its sun. Also, it plots the observation data and the fitted model. Using that, you can visually check how well is the fitting done and argue on that. The estimated orbit is plotted with a green line with crosses. The true orbit is the blue dotted line and the observations of Pandora's positions are marked with the red stars.

Exercise 3 *(2pt) Use matrices A, b, from exercise 1 with* `leastSquares` *to estimate the model coefficients.*

1. *Report which model performs best in which case, that is what is better for the x variable and what is best for the y variable.*
2. *Determine the cumulative error $\|\vec{b} - A\hat{c}\|^2$ for each one of the models*
3. *Explain **why** polynomial models or sinusoidal models would perform better for our model.*
4. *Reason your explanations numerically (using the estimated errors) and visually (using the generated plots).*

As you can see from the plots, the orbit is more or less as expected. Having such an orbit, we know that variables x, y might depend on each other based on some pattern behind their motion. So far we have been able to successfully explain the orbit based on our data, is it though the best we can do? Are there any properties that would bond x, y together and allow us build an even better model of Pandora's orbit?

Exercise 4 *(Bonus 1pt) Consider what quantities have we been minimizing so far.*

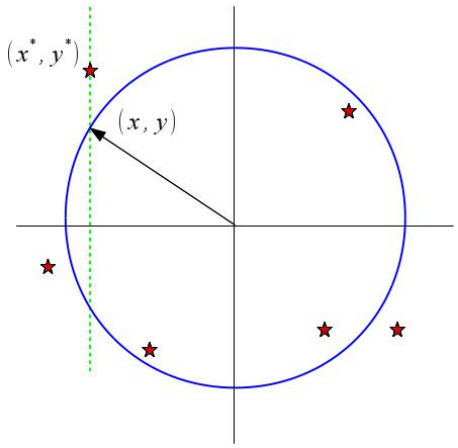

Figure 1:

1. *Get the plot of Pandora's orbit and show graphically what are we minimizing by regressing time t on x and on y. An example would be the plot of figure 1.*

2. *By now, we can assume that we know that Pandora's orbit is circular. Since Pandora follows a circular orbit, x and y are somehow related. What is that relation? To say it alternatively, since we know that Pandora should follow a* ***circular*** *orbit (think of the unit circle), like in figure 1, how else can we express Pandora's coordinates, combining x and y.*

3. *Having this a-priori knowledge of Pandora's orbit, we can use it for our own benefit to improve or simplify our model, which currently does regression on both x and y variables. Perform the linear regression on the new coordinate format. What is one of the advantages of the new coordinate systems? To answer that question, think of what remains constant on the points around a circle and how we can use that to simplify our initial model.*

3 What is required?

Please send an email to `egavves@uva.nl` with the m-files containing the functions you have implemented. Submit a report in pdf answering *ALL* the questions in the exercises. You can work in couples, putting both names in the final report.