Chapter Five Homework

The use of cross-correlation methods is common in science, but most introductory level courses do not discuss it at all. As a result, many students who only have an introduction science fields, even up to the BS level, have never seen how the method works in practice. That seems a shame. Given the importance of the method in data analyses, students of science (even those who are not majoring in science) should have some idea of how data and theory are compared. In this section we will look at some simple examples that will illustrate how the method of cross-correlation can be used to make a determination of how strongly a set of measured data matches some model-generated data.

You can use any software you like for these exercises. A spreadsheet will work, but so will other packages or computer languages.

Homework Question 13

For this exercise, use the data located in the file DATASET1.csv. This file has two columns/rows of numbers. The first is time, in seconds. The other is some other quantity that depends on the time. The quantity could be anything, it just depends what kind of measurements were being made in the experiment that acquired the data. Your task is to find a mathematical model that will best fit these data. A general procedure you can follow to produce a model is as follows.

1. Inspect the data. Usually plotting one variable against the other is a good way to do this.
2. Look for trends in the data that can be fit with a familiar mathematical function. Are the data linear? Quadratic? Logarithmic? Exponential? Periodic?
3. Decide on a function to try to fit, and determine the number of parameters you will need to describe that function.
4. Write a computer code to compute a model for different parameters and compare the computed models to the data.
5. Determine the "best fit" model, usually by choosing the model with the minimal difference from the ensemble of data points. You must decide what is a reasonable criterion for determining what you mean by "minimal difference." You might use the sum of the differences between the model and the data for each point, the mean of the difference, the sum of the squares of the differences or their mean. Whichever method you use, think about what its strengths are, and its weaknesses.

When you have decided what your best fit model is, write a short description of it. Give the model parameters that work best and explain the criterion you used to make that decision.

Homework Question 14

In the previous exercise you had to fit a dataset using a mathematical model, but there was no physical motivation behind it. The dataset we provided you was just a generic time series with no physical interpretation given for it. This exercise will be different.
Use DATASET2.csv for this. It again lists time series data, but in this case the columns accompanying the time are the x-y positions of some object over the times provided. In other words, the data describe the trajectory of a particle. Since you have an idea of what the data are and how they might have been acquired, you can not only fit a mathematical model to them, but also give a physical interpretation for the data.
Use the same methodology as in the previous exercise, but this time you should include some physical explanation of the system along with your best fit model and parameters.

These exercises have given you some basic experience fitting models to data. Sometimes models can be physically motivated, as in the second of the two exercises. In other cases, there is no physical motivation at all. A model based only upon fits to data can still be used to make predictions, and that alone can be useful even if it doesn't necessarily further basic understanding of the system in question. These sorts of models are called "empirical" models. In the first exercise you produced one of them. In the second model you were given more information: you knew that the data provided positions over time. This allowed you to make a better interpretation of the data and could, in principle, even allow you to test some physical theory. If you did the exercise you have an idea of what theory these data could test.

Homework Question 15

Fetch DATASET3.csv from the Moodle website. It is another CSV file which can be easily read into a spreadsheet. For this exercise, multiply corresponding cells (from the same column) of the spreadsheet together and place the product of each pair into a third row. You should place this third row well below the first two, perhaps at row 20 or so. So, for example, multiply the first cell of rows one and two together, and place their product into the first cell of row 25. Continue this for the entire dataset, placing the product of each cell pair into the same column, but in row 25. When you have finished, compute the sum of all the cells in row 25. This sum is called the cross-correlation coefficient. You can put it at the end of the data in row 25 if you like.

Now make a copy of the second row, but shift all the cells right by one. Place this copy in row three and pad the empty first column with a zero. Delete any extra cells on the right that extend past the data in the first two rows. Repeat the procedure above for this new shifted version of row two. To clarify, you should duplicate row two with a version of itself in which every cell has been shifted to the right by one cell position while keeping the total length of the row constant. You then compute the cross-correlation coefficient using this duplicate and put the new products and their sum in the row below the first example - row 26 in the example we have used. Do this several times, shifting row two rightward by an additional cell each time and placing the new copy in the empty row immediately below the previous copy. Insert zeros on the left while deleting overflow in the right. Similarly, create a new row of products and compute a new cross-correlation coefficient for the shifted data. What do you notice happening? If it's not obvious, you have not done enough shifts and correlation coefficients. Do a few more.

Explain how this method can be used to fit model-generated data to observed data from an experiment. Or put another way, how does this technique relate to the previous two exercises?

If you prefer, you can use a computer code to complete this exercise rather than using a spreadsheet. The shift-multiply-add operations are straightforward to implement using a computer program. Certainly more so than doing them with a spreadsheet.

In the third exercise there was no model to fit, per se. You just had to match two signals using a cross-correlation method. This is similar to the way LIGO does template-matching.