CEE 690: Health and Environmental Data Science, Spring 2019

Homework 5: 65 Total Points

Due: April 19th by 11:59 PM

Submissions required electronically through Sakai

Problem 1: Exploring a Time Series (20 points)

In the homework repository, there is a .txt file corresponding to the Southern Oscillation Index (SOI), which you can read more about at https://www.ncdc.noaa.gov/teleconnections/enso/indicators/soi/.

In Python, you can load this into a vector of measurements and yearly time indexes with:

```
data=pd.read_csv('data/soi.txt',sep='  ',header=2,index_col=0)
data=data[0:67]
x=np.double(data.values.reshape([-1]))
t=np.arange(0,67,1./12.)
```

Or you can use the template.

This returns time indexes t with respect to the starting year, and measurements x in a vector.

Calculate, document, and ***comment*** on the following:

a) The power spectral density
b) The empirical autocorrelation function
c) The empirical partial autocorrelation function

Given what you have observed from the autocorrelation and partial autocorrelation functions:

d) If you were using an autoregressive (AR) model, how many lags would you use?
e) If you were using a moving average (MA) model, how many lags would you use?

Problem 2: Time Series Prediction (20 points)

Because the SOI is highly correlated to meteorological events, it's of interest to be able to predict future patterns. Set up a time series cross-validation structure, and estimate how well you can predict the time series from.

a) An autoregressive model where you determine the number of lags by cross-validation. Is the suggested order from the autocorrelation or partial autocorrelation reasonable? How do the two values compare?
b) A nearest neighbor model. Determine what number of neighbors and what number of lags you should use by cross-validation.

Problem 3: Autoregressive models and Autocorrelation (20 Points)

Consider the autoregressive-1 model, given by the updates:

$$x_t = \beta x_{t-1} + \epsilon_t, \quad \epsilon_t \sim N(0,1).$$

An alternative way of describing this model is to state

$$x_t | x_{t-1} \sim N(\beta x_{t-1}, 1).$$

We want to understand what the autocovariance sequence looks like. First, we want to know what the *stationary* distribution is, which simply means what the distribution of a single sample would look like if we did not know its neighbors, which is for $\beta < 1$:

$$x_t \sim N\left(0, \left(\frac{1}{1-\beta^2}\right)\right).$$

Using these properties, determine the following:

a) Find the lag-1 autocovariance:

$$E[x_t x_{t-1}]$$

b) Find the function that defines the lag-τ autocovariance:

$$E[x_t x_{t-\tau}]$$

c) What is the analytical autocorrelation function as a function of t?

Problem 4: (5 points)

How many hours did it take you to complete this assignment?

Affirm that you adhered to the Duke Honor Code.