CEE 154/254 Data Analytics for Physical Systems
Autumn 2020
Assignment 2

Assignment 2 is due on 10/12/2020 at the beginning of class

Goal: Understand and apply component analysis and statistical modeling concepts to real-world data. This assignment covers Lectures 5-7.

For this assignment we will use the Foshan data from Assignment 1.

First, load the data file *hw2_1.mat*
This file contains a data table that includes time, location (longitude, latitude), and PM2.5 concentration data collected in Foshan, China from 09/28/2018 to 10/07/2018.

Part I [30 points]: **Seasonal Trend Decomposition – Additive Modeling**
Create the hourly averaged PM2.5 data using the original data table. This data table is a 240 row by 4 column matrix, and each row contains the hourly timestamp, hourly averaged PM2.5 and GPS positions.

To create the hourly averaged data, one way is to use the MATLAB command `grpstats(__)`.Specifically, you can add two new columns for the dates and hours of datapoints as shown below:

```
date = datestr(data.time,'mm/dd/yyyy');
hour = datestr(data.time,'HH');
data = addvars(data, date, hour,'Before','time','NewVariableNames',{'date',
'hour'});
```

Further, you can obtain the hourly averaged PM2.5 using the `grpstats` function:

```
hourly_data = grpstats(data,{'date','hour'},{'mean'},'DataVars',{'pm2d5'});
```

Then, create a 4 by 1 window subplot.

a) In the first subplot, plot the hourly averaged PM2.5 data with the x-axis representing the time period.
b) Please select the period of data for seasonal trend analysis and explain why you decide to select your specific periodicity (T). Find the data trend across the entire 10-day collection period using a (T+1)-term moving average whose window is [1/2T, 1/T, 1/T, …, 1/T, 1/T, 1/2T]. Also, the moving average window should be symmetric, and its summation is one. Plot the trend in the second subplot with the same x-axis as a). What do you think would

happen if you reduce T to half and why do you think so? What if you increase it to 2T and why?

c) Find the seasonal trend of the PM2.5 data using the data periodicity T that was defined in the previous question. Plot this seasonal trend in the third subplot. What pattern do you observe? Explain why you think such pattern is observed. What do you think would happen if you reduce T to half and why do you think so? What if you increase it to 2T and why?

d) Find the residual (noise) in the PM2.5 data by subtracting the trend and seasonality from the hourly moving average PM2.5 data. Plot this residual in the fourth subplot. Do you think all the patterns have been extracted in the trend and seasonality based on your residual plot? Explain why you think so.

e) Discuss any interesting observations from each of the plots. How can this decomposition be useful to study PM2.5 data patterns?

Part II [40 points]: **Singular Value Decomposition**.

Create a 10 row by 1440 column matrix where each row is normalized one day PM2.5 measures (0-1 range) with the sampling rate of 1 minute. Then answer the following questions.

a) Plot each day of PM2.5 data together in one plot window, similar to the figure shown in Lecture note 5.15. The x-axis should represent minutes in each day (from 1 to 1440). Mark the y-axis with numbers from 1-10 representing the 10 days.

Perform a Singular Value Decomposition (SVD) on the PM2.5 data.

The MATLAB command for SVD is: `svd(A)`

b) Plot the first 3 columns of the "U" matrix from the decomposition.

c) Plot the singular values (S). Describe what this figure represents and what pattern you observe.

d) Plot the first three rows of "V^T." Describe your observation of the "V^T" matrix. How is this matrix related with the original data?

e) Reconstruct the first 3 days' PM2.5 data using the SVD decomposition.

   i. Reconstruct with at least 90% explained ratio, using the minimum number of singular values, and plot the results. How many singular values are required for this reconstruction?

   ii. Reconstruct with at least 95% explained ratio, using the minimum number of singular values, and plot the results. How many singular values are required for this reconstruction?

   iii. Reconstruct with at least 99% explained ratio, using the minimum number of singular values, and plot the results. How many singular values are required for this reconstruction?

   iv. Plot the residual from iii). Describe any interesting observations.

Part III [30 points]: **Data Statistical Analysis.** Consider the time series data for Foshan PM2.5.

a) Estimate the data distribution (*data.pm2d5*) using a kernel smoothing function.

   The MATLAB command for kernel density estimation is: `ksdensity(x)`

   i. Determine the kernel density estimation using a normal Gaussian kernel and bandwidth 1. Plot the distribution.

   ii. Determine the kernel density estimation using a Box kernel and bandwidth 1. Plot the distribution.

   iii. Describe the similarities and differences between the distributions estimated in i) and ii) above. Which do you think better represents the data distribution and why?

b) In one plot, overlay the kernel density estimation using a Gaussian kernel with bandwidth of i) 1, ii) 10, and ii) 0.1. Describe the differences between each distribution estimation. How does the bandwidth influence the curve?

c) Calculate basic statistics for the PM2.5 data. i) mean, ii) median, iii) standard deviation, iv) skewness, v) kurtosis. Describe any interesting observations from these values.