ME 7960 Scientific Machine Learning

Midterm project

Assigned:
Due: Friday February 28th, 6:00pm

Turn in all your codes in *.m or *.py or *.ipynb format. Also, turn in a pdf of all your solutions (plots, answers, etc.). Please also include your codes (e.g., a snapshot) in the pdf document. You are allowed to use the lecture codes (see your textbook's website for all codes/data also in Matlab format) but you will need to modify and customize as needed.

The exam is designed such that every student's submission will be very different from others.

IMPORTANT instruction for Problem 1: You should replace the 2D aneurysm dataset with a "spatiotemporal" dataset (scalar or vector) from your own research or lab. Make sure you have at least 20 time-steps (ideally at least 40). If you have 3D data you can slice through it to make it 2D. You can use the aneurysm data explained below and uploaded mat file to see how to structure your data (similar to the data matrix in lectures). If you do not have any spatiotemporal data in your research group (justify why not!) then you can use the aneurysm data.

1. We are given data matrix for blood flow velocity in a 2D aneurysm. The data is stored in *2D_velocity_vector_aneu.mat*. The velocity array is our typical spatiotemporal data matrix and has 3943 points and 49 time-steps (Note that this is **vector data** so the number of rows is 2*3943=7886). Each row first contains the x component of the velocity vector then the y component of velocity and then goes to the x and y component of the next spatial point. The coordinates array shows you the x and y coordinate position of these spatial points in the same order (useful for plotting your data). You can plot the data (e.g., velocity magnitude) using scatter plot (make sure you play with the range of colorbar to see features inside the aneurysm where the flow is very slow).

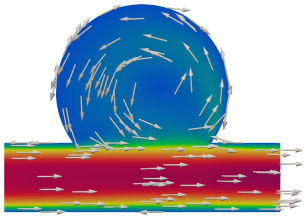

Figure 1: 2D blood flow in an aneurysm (sudden enlargement of a vessel).

Perform PCA (POD) on this data. Note that you can either work in vector format (no need to change the data matrix) or if you are just interested in velocity magnitude you can first pre-process the data to calculate velocity magnitude at each of the 3943 points, reduce the data matrix, and then perform PCA.

(a) Plot the first three dominant modes (velocity magnitude or vector). Set the min and max of the colorbar so you see flow features inside the aneurysm. Note that in practice for visualization, we plot over an unstructured grid not scatter points, but here we will plot

on scatter points to make the process easier. However, please do investigate and see how you can improve your simple visualization based on what you find and your data.

(b) Corrupt your data in two different approaches: 1- Adding noise randomly to 80% of your data. 2- Adding noise to a local specific region (use coordinates to know the location). When you add noise you generate a random number with zero mean and standard deviation 1 from a Gaussian distribution, then you modify each data entry as: $x_{noisy} = x+\zeta*rand*x$, where $rand$ is the random number you generated and ζ is the level of noise. Play with different noise levels (0.05, 0.1, 0.2, etc.). As an example, $\zeta = 0.1$ is adding 10% noise (we disturb the data by around 10% of its value with some randomness).

Similar to the lecture 9 plots, for each noise level and pattern you investigated plot the singular values sorted versus mode as well as the cumulative energy (for each mode this is the sum of all singular values up to that mode normalized by the sum of all singular values). Comment on what you learn.

(c) Perform RPCA and repeat the above to see how RPCA improves your results under noise (you might need to play with the hyperparameter λ in RPCA). Report on RPCA's performance based on noise level and pattern (random vs. localized).

2. Assume we have time-series data that we can represent as a Fourier series. We can define such a signal as:

$$u(t) = a_0 + \sum_{i=1}^{N} a_i cos(i\omega t) + \sum_{i=1}^{N} b_i sin(i\omega t)$$

Define such a signal with N=3. You may arbitrary select the coefficients as well as the frequency ω. Generate the signal for $0 < t < T$ and ensure you have a high enough resolution. Select T such that your data contains a few of the periods. This is your ground-truth data that you can use for calculating your data reconstruction accuracy below.

(a) Now imagine you are an experimentalist collecting u(t) data but you cannot measure u(t) at high resolution. You can only collect p points with p sensors where p is not a very large number (if you just plot this data with these p measurements your curve will look terrible compared to ground truth). Use **compressed sensing** (maybe discrete cosine transform as your sparsifying transform) to reconstruct high resolution data (same resolution as ground-truth) based on your p measurements. How low can you make p and still get good results? How does this depend on the frequency of your signal (if you make frequency higher do you need more sensors?) Please show all of your results and discuss your findings.

(b) Now consider the data with just one period (define T to be equal to one period). Does your method still work?

(c) Now add noise to your signal in part (a) similar to last example. How robust are your results based on the level of noise? Implement the optimization formulation we discussed in class that is more robust to noise to see how that improves your results.

3. **Traditional machine learning!** Consider the traditional machine learning topics we discussed in lecture 3 and 4 (regression and classification). Search on google scholar to find a good paper **"closely" related to your MS/PhD research area** that is using traditional machine learning (not SciML). Read the paper.

(a) Provide a citation for the paper and summarize the paper in 1 paragraph with a focus on the machine learning aspects. **Explain how this relates to your research**.

(b) List two items that you find interesting in this paper.

(c) List two criticisms you have related to the "machine learning aspects" of the paper.

(d) List at least one direction that the paper could be extended based on the SciML topics we have discussed so far.