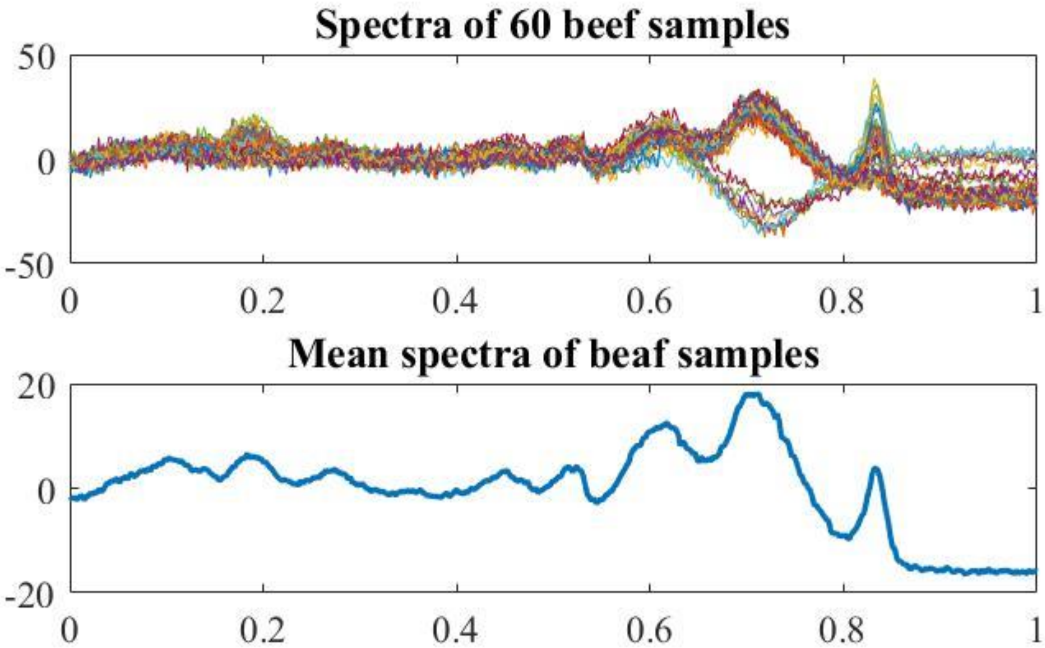
ISyE 8803 –Topics on High Dimensional Data Analytics

Exam I

- You are not allowed to discuss the exam content with your fellow students nor receive aid on this exam.
- You are expected to observe the Georgia Tech Honor Code throughout the exam.
- Exam is due on October 24, 11:59pm (U.S. Eastern Time). Late submission is NOT accepted. Submit your solutions via Canvas.

Question 1: (25 points)

Food spectrographs are used in chemometrics to classify food types, a task that has applications in food safety and quality assurance. The beef dataset consists of five classes of beef spectrograms, from pure to adulterated beef with varying degrees of offal. “Question1.csv” contains 60 beef samples. The last column represents classes. Use the first 30 data samples for training, and the rest for the test.



Part 1. The data samples are contaminated by noise. In this part, our goal is to use kernel $f(z)$ to smooth out the data by the following steps:

$$f(z) = \begin{cases} \frac{3}{4}(1 - z^2) & \text{if} \quad |z| \leq 1 \\ 0 & \text{else} \end{cases}$$

a) Compute the mean spectra of beef samples (see the lower panel of the figure above) and use 5-fold cross validation to determine the optimal kernel bandwidth.

- Report the optimal bandwidth and the cross validation MSE.

- Plot the mean spectra of beat samples and your smoothed curve.

b) Use the kernel bandwidth you found in Part 1a and smooth out the entire data set.

- Plot the smoothed data samples.

Part 2.

a) Use cubic B-spline with 28 knots to estimate the smoothed data you obtained from Part 1b.

b) Perform functional PCA. How many FPC-scores are required to explain 99% of variations?

b) Build a multi-class support vector machine (SVM). Evaluate the performance of your classifier on the test set.

- Report the accuracy and confusion matrix on the test set.

Question 2: (25 points)

"Question2.jpg" is contaminated with both additive Gaussian noise and salt & paper noise. In this question, we use two noise reduction techniques to suppress the noise. Apply each technique to the noisy image and display the output.

1- **Trimmed Mean Filter**: Suppose that we delete the $\frac{d}{2}$ lowest and the $\frac{d}{2}$ highest intensity values of $g(r,c)$ in the neighborhood S_{xy} to trim away outlier pixels. Let $g_R(r,c)$ represent the remaining $mn-d$ pixels in S_{xy}. A filter formed by averaging these remaining pixels is called a trimmed mean filter. The form of this filter is

$$\hat{f}(x,y) = \frac{1}{mn-d} \sum_{(r,c)\in S_{xy}} g_R(r,c)$$

Where m and n is the size of the filter. Consider filter size 7 × 7 and $d = 8$.

2- **Noise Suppression Filter**:

a) Estimate the standard deviation of the noise (σ_n) using the following formula:

$$\sigma_n = \sqrt{\frac{\pi}{2}} \frac{1}{6M\,N} \sum_{x=1}^{M} \sum_{y=1}^{N} |\mathrm{Image}(x,y) * W|$$

where M and N are the size of the image, $|\cdot|$ is the absolute value, '*' is the convolution operator, and $W = \begin{bmatrix} 1 & -2 & 1 \\ -2 & 4 & -2 \\ 1 & -2 & 1 \end{bmatrix}$. Report your estimated standard deviation of the noise (σ_n).

b) Our filter is to operate on a neighborhood, S_{xy} centered on coordinates (x,y). The response of the filter at (x,y) is to be based on the following quantities: $g(x,y)$ the value of the noisy image at (x,y); σ_n^2 the variance of the noise estimated in part a; $\bar{z}_{S_{xy}}$ the local average intensity of the pixels in S_{xy}; and $\sigma_{S_{xy}}^2$ the local variance of the intensities of pixels in S_{xy}. The filtered image $\hat{f}(x,y)$ is obtained by

$$\hat{f}(x,y) = \begin{cases} g(x,y) - \frac{\sigma_n^2}{\sigma_{S_{xy}}^2}\left(g(x,y) - \bar{z}_{S_{xy}}\right) & if \ \frac{\sigma_n^2}{\sigma_{S_{xy}}^2} \leq 1 \\ \bar{z}_{S_{xy}} & if \ \frac{\sigma_n^2}{\sigma_{S_{xy}}^2} > 1 \end{cases}$$

Consider the filter size 7 × 7.

Question 3: (25 points)

1) Electroencephalography (EEG) is an electrophysiological monitoring method to record electrical activity on the scalp that has been shown to represent the macroscopic activity of the surface layer of the brain underneath. “Question3.zip” contains the EEG data. There are 99 .csv files. Form tensor $\mathcal{X} \in \mathbb{R}^{64\times 640\times 99}$ so that each frontal slice is one csv file. Consider the following optimization problem:

$$\underset{a_r,b_r,c_r}{\operatorname{argmin}} \left\| \mathcal{X} - \sum_{r=1}^{R} a_r \circ b_r \circ c_r \right\|_F^2$$

Part 1: Write a pseudo code for finding the factor matrices A, B, and C using alternating least squares (ALS) method.

Part 2: Implement/code your algorithm (part 1) to find the factor matrices A, B, and C. Consider rank $R = 4$ and maximum iterations 50. Plot the loading matrices (A, B, C).

Question 4: (25 points)

Consider the likelihood function:

$$L(\alpha,\beta|y_1,\dots,y_n) = \prod_{i=1}^{n} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, y_i^{\alpha-1}\,(1-y_i)^{\beta-1}$$

Where $\Gamma(\cdot)$ is the gamma function; and $\alpha, \beta > 0$; and y_i can be found in ‘Question4.csv’.

a) Write down the log-likelihood function.

b) Write down the corresponding maximum likelihood formulation.

c) Derive the gradient and the Hessian of the log-likelihood function.

d) The data file (‘Question4.csv’) contains 1000 observations. Report your estimated $\hat{\alpha}_{MLE}$ and $\hat{\beta}_{MLE}$. Plot the value of the parameters $\hat{\alpha}_{MLE}$ and $\hat{\beta}_{MLE}$versus the number of iterations. Plot the value of the log-likelihood function versus the number of iterations. For computing digamma $\psi(x) = \frac{\Gamma'(x)}{\Gamma(x)}$ and trigamma $\psi'(x)$, you can use built-in functions.

- Accelerated Gradient descent
- Newton’s method