Question 1: (20 points)

Compute the natural cubic spline S which interpolates f at the knots 1, 2, and 5.

$$f(x) = \frac{1}{x^2}$$

$$S(x) = \begin{cases} a_1 + b_1(x-1) + c_1(x-1)^2 + d_1(x-1)^3 & x \in [1,2] \\ a_2 + b_2(x-2) + c_2(x-2)^2 + d_2(x-2)^3 & x \in [2,5] \end{cases}$$

Find $a_1, a_2, b_1, b_2, c_1, c_2, d_1, d_2$(show your work). Sketch f and S for $x \in [1,5]$.

Question 2: (10 points)

Consider the penalized least square problems:

$$\hat{f}_1 = \min_f \left[\sum_{i=1}^{N} \left(y_i - f(x_i)\right)^2 + \lambda \int_a^b \left(f^{(m)}(x)\right)^2 dx\right]$$

$$\hat{f}_2 = \min_f \left[\sum_{i=1}^{N} \left(y_i - f(x_i)\right)^2 + \lambda \int_a^b \left(f^{(m+k)}(x)\right)^2 dx\right]$$

where k is a positive integer.

Part 1: Consider $\hat{f}_1$in the following cases and determine the degree of the polynomial.

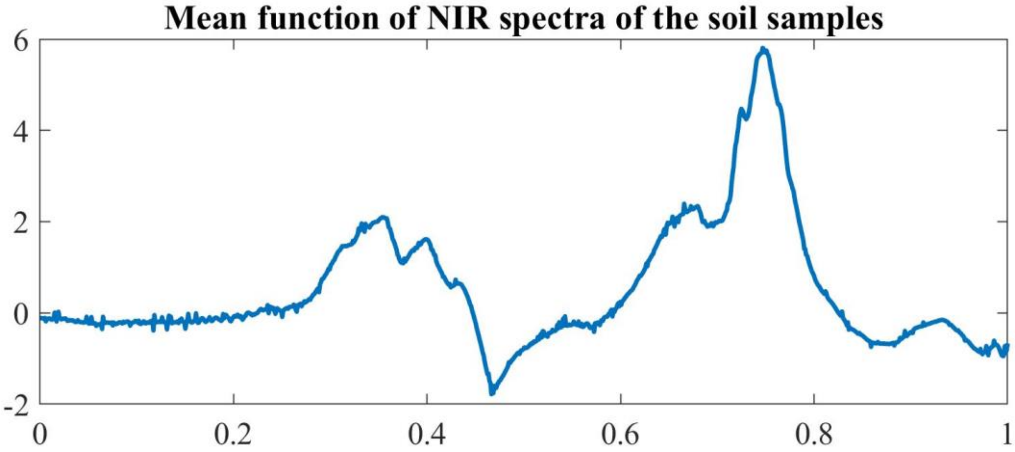
(1) $m = 0$ and $\lambda = \infty$
(2) $m = 1$ and $\lambda = \infty$
(3) $m = 2$ and $\lambda = \infty$
(4) $m = 3$ and $\lambda = 0$

Part2:

(1) Will $\hat{f}_1$ or $\hat{f}_2$ have the smaller training residual sum of squares as $\lambda \rightarrow \infty$.
(2) Will $\hat{f}_1$ or $\hat{f}_2$ have the smaller test residual sum of squares as $\lambda \rightarrow \infty$.
(3) Will $\hat{f}_1$ or $\hat{f}_2$ have the smaller training and test residual sum of squares for $\lambda = 0$.

Question 3: (40 points)

In precision farming and agriculture practices, soil fertility is an important factor which affects plant growth and maintaining soil health. To determine soil quality and fertility in a short time and without complicated sample preparations, near infrared reflectance spectroscopy (NIRS) is employed. “Question3.csv” contains NIRS of 40 bulk soil samples. Calculate the mean function of the data (figure below) and use the following methods to estimate it.

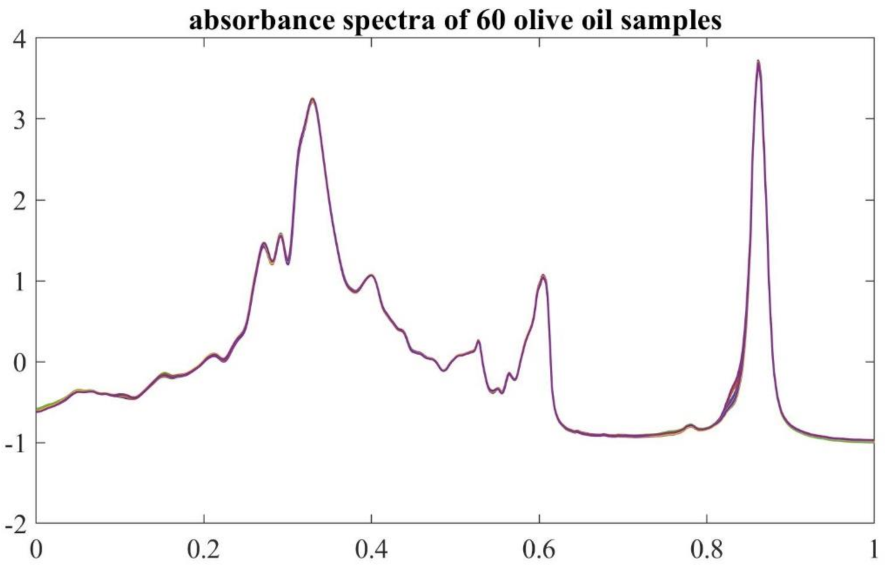


(a) Cubic Spline: vary the number of knots from 5 to 50 and use 5-fold cross validation to determine the optimal number of knots. Report the optimal number of knots and the cross validation MSE.
(b) Cubic B-splines: vary the number of knots from 5 to 50 and use 5-fold cross validation to determine the optimal number of knots. Report the optimal number of knots and the cross validation MSE.
(c) Smoothing Spline: Report the optimal lambda and MSE.
(d) Gaussian kernel: use 5-fold cross validation to determine the optimal kernel bandwidth. Report the optimal bandwidth and the cross validation MSE.

Plot your estimated curves along with the mean function of NIR spectra of the soil samples.

Question 4: (30 points)

Olive oil is cultivated extensively across the Mediterranean basin. It is known that the composition of olive oils varies with geographic origin due to a number of different factors: regional differences in climate, soil, and agricultural practice. Verifying the declared origin, or determining the origin of an unidentified olive oil, is, therefore, a challenging problem. In this question our goal is to distinguish extra virgin olive oils from different producing countries. “Question4.csv” dataset contains 60 authenticated samples of extra virgin olive oils, originating from four European producing countries: Greece, Italy, Portugal, and Spain. Use the first 30 data samples for training; and the rest to evaluate the accuracy of your models. The last column of the dataset indicates classes {1, 2, 3, 4}.



Group Designation	Country of Origin	Number of Samples
1	Greece	10
2	Italy	17
3	Portugal	8
4	Spain	25
Total		60

Part 1: In this part, we use cubic B-splines with 70 knots for dimension reduction and feature extraction. Use the extracted features (B-spline coefficients) to build a multi-class support vector machine (SVM). Evaluate the performance of your classifier on the test set.

- Report the accuracy and confusion matrix.

Part 2: In this part, we use functional PCA for dimension reduction and feature extraction. Build four multi-class SVMs using 2, 5, 8, and 10 harmonics. Evaluate the performance of your classifiers on the test set.

- Report the accuracy and confusion matrix of the four classifiers.