ISYE 8803 Exam 1 Problem 1

Nick DiNapoli, ndinapoli6@gatech.edu

June 26, 2022

1 Problem 1

In this problem, I analyze the near infrared (NIR) spectra from 166 fermentation mashes (absorbance as a function of wavelength). These spectra serve as feature data where each sample has a corresponding ethanol concentration target value. I aim to predict the ethanol concentration using two different regression models. Each model was built using a different dimensionality reduction technique, the first being b-splines and the second, functional principal component analysis (FPCA). I use the first 100 samples as training data and the remaining 66 as the test set. Figure 1 shows the mean of all of the data to represent the rough shape of the sample signals.

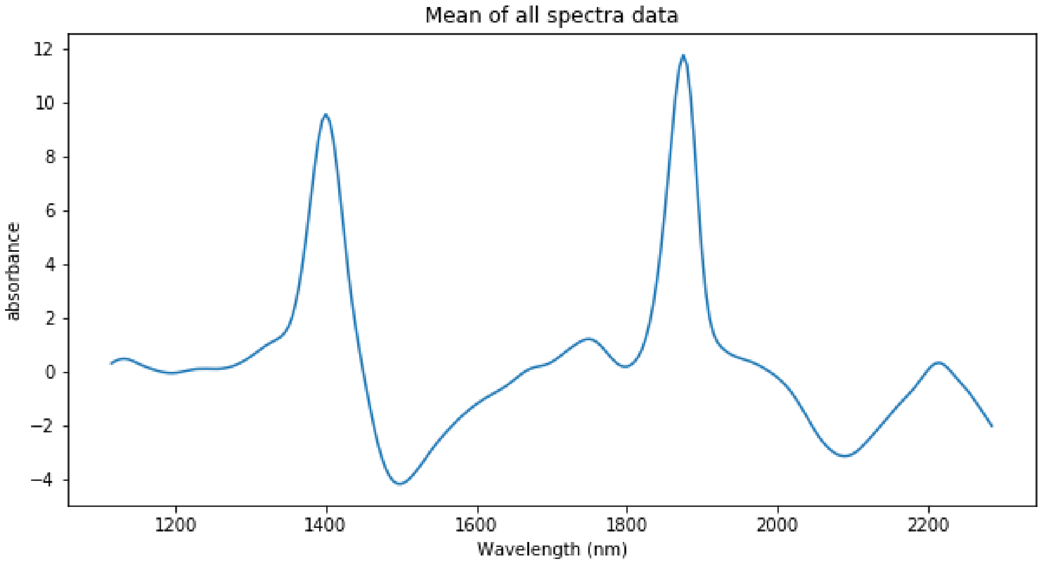

Figure 1: Average signal across all samples.

1.1 B-splines

First, I implement cubic b-splines for dimensionality reduction and feature extraction. Specifically, I varied the number of knots from 5 to 50 and completed 5-fold cross-validation to determine the optimal number of knots. For this task, I averaged the mean squared error (MSE) of the test set across each fold and compared the average MSE for each knot value to determine the optimal number of knots. The cubic b-splines estimate and MSE for each different knot value can be seen in Figures 2 & 3. The optimal number of knots was found to be 12. The b-splines estimate of the training data using the optimal number of knots is superimposed on the mean of the true training data values in Figure 4.

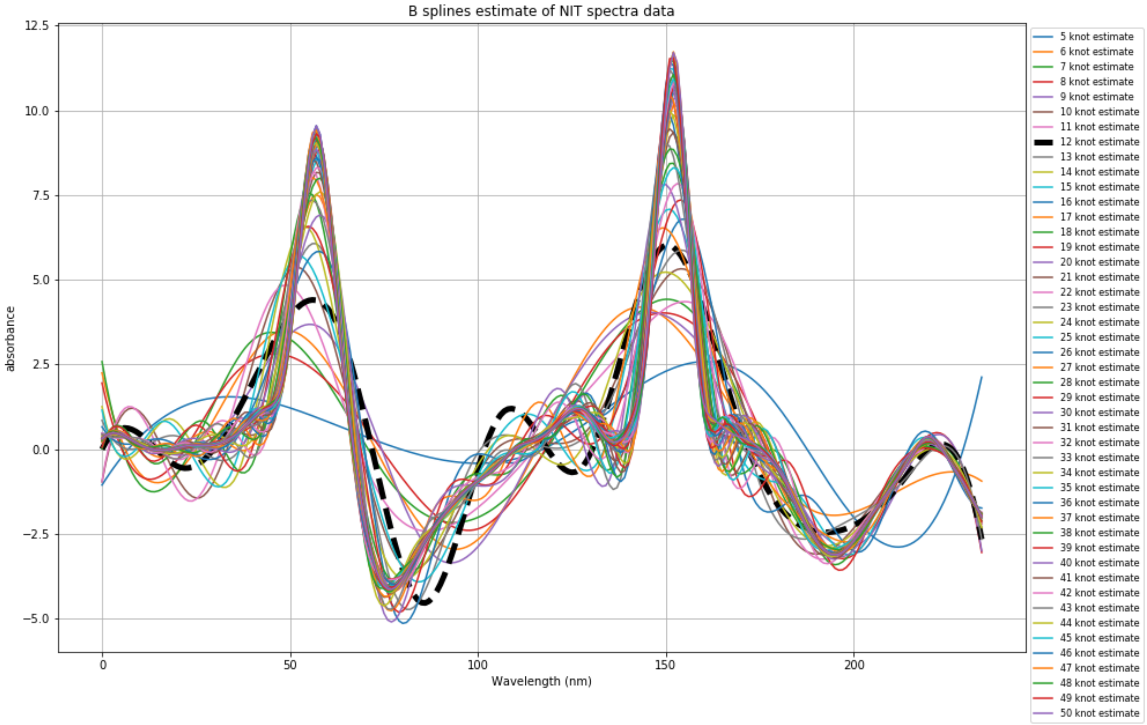

Figure 2: B-splines estimate for each knot value. The estimate using the optimal value of 12 is plotted as a black dotted line.

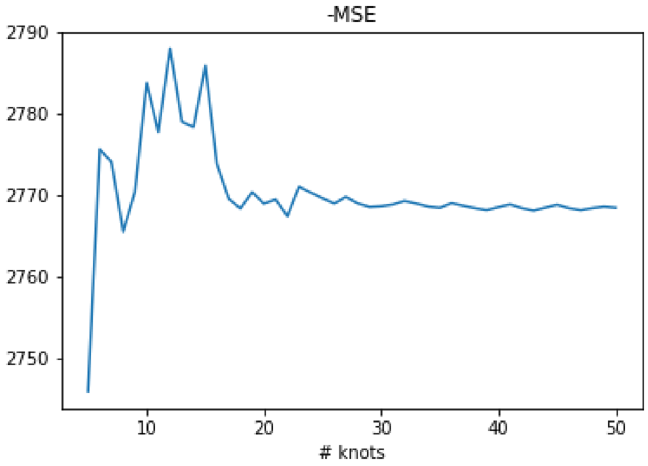

Figure 3: Negative average MSE using 5-fold cross-validation as the number of knots is varied.

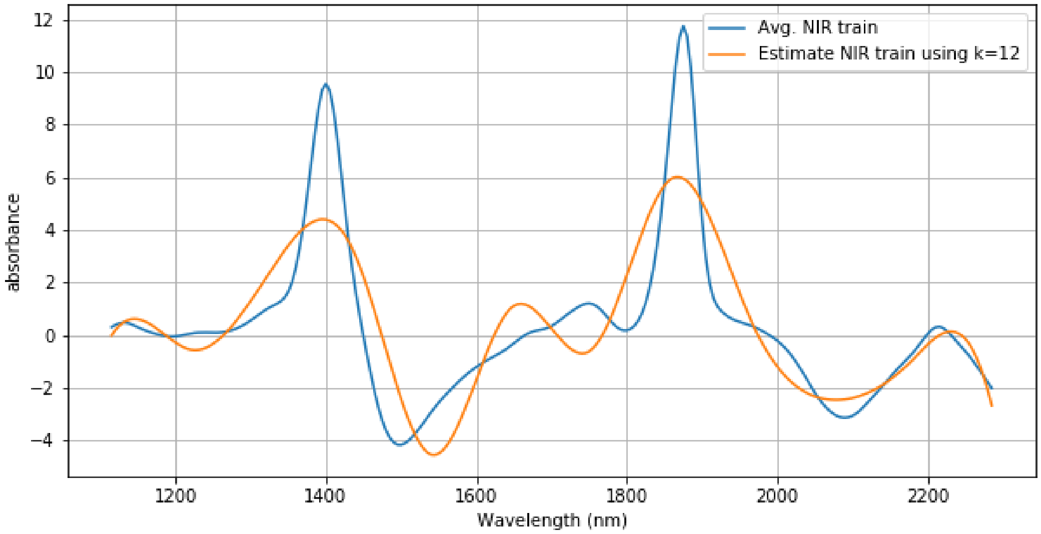

Figure 4: B-splines estimate when the number of knots is 12 on top of the true mean of the training data.

I then performed feature extraction on the training and test sets using cubic b-splines and 12 knots and built a linear regression model to make predictions on the ethanol concentration of the test set. After completing this task, I found the residual sum of squares (RSS) for this model on the in sample data to be 23.16 and the MSE of the test set to be 17.58.

1.2 FPCA

I completed a similar implementation but used FPCA for dimensionality reduction. Using the optimal number of knots from the previous section, I show the cubic b-splines estimates of the entire dataset (train and test) in Figure 5. I discovered that 5 FPC-scores are necessary to explain 99% of the variations in the dataset. In order to draw direct comparisons between the two methods, I built another linear regression model based on the scores on the principal components of the training data and made predictions on the training and test sets once again. Using FPCA yields an RSS of 1869.14 and a MSE on the test set of 83.32.

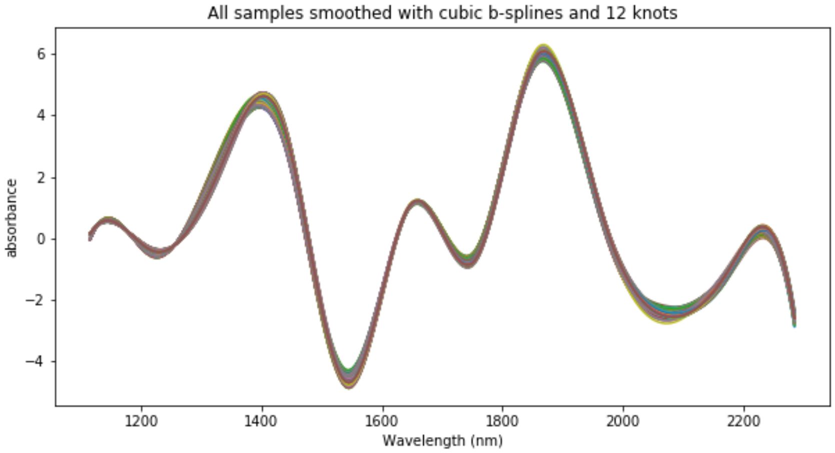

Figure 5: B-splines estimate of each signal on the entire dataset.