Homework 10

Due Friday, 11/20/20

1. Consider the data set "NIR" in the R package 'chemometrics', which contains the first derivatives (with respect to wavelength) of near infrared spectroscopy (NIR) absorbance values at $p = 235$ wavelengths between 1115-2285nm. The goal is to use these covariates to predict the glucose concentration in $n = 166$ alcoholic fermentation mashes of feedstock. The columns in the covariate matrix `NIR$xNIR` are arranged in order of increasing wavelength.

   (a) Concentration is typically right skewed, and often times must be transformed to meet linear modeling assumptions. Use ordinary least squares to regress glucose concentration onto ten randomly chosen predictors. Using the results from this regression, do you think a transformation is warrented? Explain. (**Hint**: To make results as interpretable as possible, it is usually best to avoid complex transformations if possible.)

   (b) Let y_i be the glucose concentration in fermentation mash i. Can ordinary least squares be used to estimate the parameters in the model

cannot because
main reason is p > n
=> non full rank

$$y_i = \sum_{j=1}^{p} \beta_j A_j + \epsilon_i, \quad i = 1, \ldots, n \tag{1}$$

   where A_j is the first derivative of the absorbance spectrum at wavelength j? If not, can you suggest four other methods that we've looked at in class that might be used to estimate $\beta_1, \ldots, \beta_p$ in this model?

   In the following questions, permute the observations by using the seed "1968", and then use the first 126 values for training and the last 40 values for testing.

   (c) Use the training set and principal component regression, using 9-fold cross validation to estimate the number of components, to estimate $\boldsymbol{\beta}$ from Model (1). That is, choose the number of components $\hat{K}$ to be

$$\mathcal{L}(k) = \frac{1}{126} \sum_{f=1}^{9} \sum_{i \in \text{fold } f} (y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_{(-f)}^{(k)})^2$$
$$\hat{K} = \underset{k \in \{0, \ldots, \min(n-1, p)\}}{\arg\min} \mathcal{L}(k),$$

   where $\boldsymbol{x}_i$ is the ith row of the 126×235 covariate matrix and $\hat{\boldsymbol{\beta}}_{(-f)}^{(k)}$ is PCR's estimate for $\boldsymbol{\beta}$ with k components using data from folds $1, \ldots, f-1, f+1, \ldots, \#\text{folds} = 9$. (Remember that you need to account for the intercept!)

   (i) Plot $\mathcal{L}(k)$ as a function of k. What is $\hat{K}$?

   (ii) Plot your estimate for $\boldsymbol{\beta}$ as a function of wavelength λ. What do you conclude?

   (iii) Repeat part (i) using leave one out cross validation instead of 9-fold cross validation. How does the loss compare to part (i)?

(d) Repeat (c), but with partial least squares.

(e) Now use LASSO to estimate $\boldsymbol{\beta}$ with λ chosen with 9-fold cross validation. Plot $\mathcal{L}(\lambda)$ as a function of $\log(\lambda)$.

(f) Use the test data to evaluate PCR's, PLS's, and LASSO's predictive performance on this dataset. Comment on $\mathcal{L}$'s ability to estimate the testing error.