Cogs 109: Modeling and Data Analysis

Homework 6

Due **Tuesday 11/20**

1. A student in a Modeling and Data Analysis course downloaded monthly San Diego temperature data covering the previous 2 years (n=24 data points). For each of the following methods, comment on whether it could potentially be an appropriate way of modeling the data. Explain your answer
   a. Polynomial regression with polynomial order k=2?
   b. Polynomial regression with order k=24?
   c. Polynomial regression with order k=48?
   d. Spline regression with 35 knots?
2. **Principal component (PC) regression.** Using the anesthesia dataset from last week, we will perform PCA regression.
   a. (3 points) Recall that the predictors are the power levels in each of a set of 103 frequency bands. Let X_i be the i'th predictor; since the predictor is actually a vector of values for each each time point, we can write this as a matrix: X_{ij} = the power of the i'th frequency band at the j'th time point. Compute the standard deviation (s.d.) of each of the predictors and make a plot showing them. Which frequency band has the largest s.d.?
   b. (2 points) The first step of our analysis will be to standardize the predictors by dividing each predictor by its standard deviation. This ensures that all predictors have the same variance. Using these values, define a new set of standardized predictors $S_{ij} = X_{ij}/std(X_i)$. Check that you have done this correctly by verifying that the s.d. of each of the standardized predictors is 1.
   c. (3 points) Compute the top 20 principal components (PCs) of the predictors. Make a plot showing the value of PC1 as a function of time throughout the experiment. Make similar plots showing the values of PC2 and PC3 vs. time.
   d. (4 points) Run PC regression by fitting a series of linear models using the top k=1, 2, 3, …, 20 principal components. Use 10-fold cross-validation to estimate the test set error for each linear model. Make a plot showing the MSE for training and test data as a function of k.
   e. (1 point) Based on these results, what value of k would you select to provide the best predictive accuracy?
   f. (1 point) Conceptual question (no coding required): Consider a linear model that includes just the top PC (k=1). Which frequencies contribute to this model's prediction? Explain.
3. **Polynomial and spline regression.** In this problem we will try to fit smooth curves to the timecourse of the behavioral response variable in the Anesthesia dataset.
   a. (1 point) Plot BehaviorResponse vs. time.

b. (2 points) Fit a linear model of the form: $y = \beta_0 + \beta_1 x$, where y is BehaviorResponse and x is time. Plot the resulting prediction, $\hat{y}$, as a function of time and overlay it on top of the true data. Is this model too simple or too complex for this dataset?

c. (2 points) Fit polynomials of order 2 (quadratic), 3 (cubic), 5, 10 and 20. For each polynomial, plot $\hat{y}$ vs. time.

d. (3 points) Using 10-fold cross-validation, determine the testing MSE for all polynomials of order 1 up to 20. Plot the training and test MSE vs. model order.
(NOTE: You may not see signs of overfitting in this case, even with model order=20)

e. (1 point) Based on these results, what model order would you recommend for fitting these data?

f. (Extra credit 3 points) Fit a cubic spline smoothing regression [NOTE: When fitting a smoothing spline you will not choose the knots]. Plot test MSE as a function of lambda. (Note in MATLAB the function uses a parameter p instead of lambda -- you can plot your result as a function of p). What lambda (or p) value gives the best predictive model? (NOTE: In MATLAB, the function csaps has a parameter p which plays the role of lambda. When p=0, lambda=∞, and when p=1 , lambda=0. For this problem you may plot your result as a function of p or lambda, but make sure to clearly label the axes.)

Code hints for MATLAB:

```
pca, std, polyfit, polyval, csaps (requires the curve fitting toolbox)
```

Python:

```
scipy.interpolate.UnivariateSpline
```