BSCB 694: Functional Data analysis

Homework 4

Due: Tuesday, May 1

Answer at least 3 questions. Submit any Matlab code you used (you may print and submit an HTML demonstration, but do not need to).

1. Leave-one-curve-out CV

   (a) Recall the ordinary cross validation identity

$$\sum_{i=1}^{n}(y_i - \hat{f}^{-i}(t_i))^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{f}(t_i))^2}{(1 - S_{ii})^2}$$

   where $f^{-i}(t)$ is the prediction at t when the ith point has been left out and $\hat{f}$ results from a linear smoother with smoothing matrix S. Now suppose that we instead wish to leave out multiple observations at a time. Specifically let the numbers $1, \ldots, n$ be partitioned into sets $I_1, \ldots, I_k$, find an expression for

$$\sum_{j=1}^{k} \left(y_{I_j} - f^{-I_j}(t_{I_j})\right)^2$$

   that does not require re-calculating the regression k times.

   (b) Apply your answer above to leave-one-curve-out cross validation in functional linear regression. Provide specific formulae for the discounting terms.

   (c) Consider a multivariate response model that results from a linear smooth in the sense that

$$\text{vec}(\hat{f}) = S\text{vec}(y)$$

   where vec(y) stacks the n k-dimensional responses $y_i = f_i + \epsilon_i$ into one long vector. Further, assume that S preserves the true function $f = Sf$. Let the ϵ_i be iid vectors with covariance matrix Σ^2. Derive an unbiassed estimate for this matrix.

   Bonus: Show that GCV is invariant to linear transformations of the y, while OCV is not. Can you derive an equivalent for the multivariate response model?

2. Representor Theorems for Functional Models

   Recall the penalized functional linear model objective:

$$\sum_{i=1}^{n}(y_i - \int \xi(t)x_i(t)dt)^2 + \lambda \int (L\xi(t))^2 dt$$

and the penalized functional Principle Components Analysis model

$$\frac{\text{var} \int \xi(t)x_i(t)dt}{\int \xi(t)^2 dt + \lambda \int (L\xi(t))^2 dt}$$

Assume that each x_i and ξ are in W^m and that is is an mth-order linear differential operator.

(a) Show that in both cases, ξ is may be characterized by a finite linear combination of functions.

(b) Demonstrate that the basis-approximation to ξ used in the book can be given to any desired accuracy.

(c) Generalize these results to an objective function of the form

$$\sum N_i(\int \xi_1(t)x_{i1}(t), \dots, \int \xi_k(t)x_{ik}(t)dt) + \sum_{j=1}^{k} \int (L_j\xi_j(t))^2 dt$$

for N_i nonlinear functions of k arguments and $x_{i1}(t), ldots, x_{ik}(t)$ are co-variate functions for each observation i.

Bonus: derive explicit means to calculate each of the functions in the linear combination above.

3. Functional Linear Regression, Smoothing and Bias.

Part of the motivation for iterating principle differential analysis is that it helps to remove bias due to smoothing. Here we investigate a simple situation

(a) Consider the smoothing spline minimizing

$$\sum (y_i - \hat{f}(t_i))^2 + \lambda(L\hat{f}(t))^2 dt$$

with $y_i = f(t_i) + \epsilon_i$ where the ϵ_i are iid. Show that

$$\sum (f(t_i) - E\hat{f}(t_i))^2 \propto \int (Lf(t))^2 dt$$

(b) Suppose that we are interested in relating $Ly_i(t)$ to $x_i(t)$ in a concurrent linear model. Derive the bias for $x_i(t)$ in terms of the bias for $y_i(t)$.

(c) Show that each step in iterated principle differential analysis reduces the objective function

$$\sum_i \sum_j (y_{ij} - \hat{f}_i(t_j))^2 \quad + \sum_i \lambda \int \left(D^m \hat{f}_i(t) + \beta_m(t) D^{m-1} \hat{f}_i(t) + \ldots + \beta_0(t) \right)^2 dt$$
$$+ \sum_k \tau_k \int (L_k \beta_k(t))^2 dt$$

4. Smoothing mean functions

   Suppose that observations $y_{ij} = f_i(t_j) + \epsilon_{ij}$ of k functions are available where the observation times t_j are common across functions. We wish to estimate the average function $\sum f_i(t)/k$.

   (a) Show that smoothing the average values $\sum_i y_{ij}/k$ is equivalent to smoothing the k curves individually and then averaging.
   (b) How do the smoothing parameters for each relate?
   (c) Demonstrate the above in terms of the equivalent GCV formulae.

5. Computational inference.

   The `slider.mat` data file on the class website contains results from an experiment undertaken in the Psychology department of McGill University. Subjects were played a series of pitches and were asked to represent the pitch on a linear scale by use of a slider. The position of the slider was then recorded at about 150hz.

   The data files in `slider.mat` are

slider	the position of a slider relative to its value before each change in pitch
pchange	the change in pitch corresponding to each run.

   Additionally, `slider.m` provides an example principle differential analysis for modeling responses in relation to changes in pitch.

   The concern here is that variance functions given by the `fRegress` function are only estimated point-wise and require the curves to be distributed as approximately gaussian.

   (a) Design and implement a resampling scheme to provide estimates of the true variance in such a way as to account for
      - the discretisation errors apparent in the data
      - the process of smoothing the data
      - the distribution of curves about a response model
   (b) Design and implement a test for the effect of pitch change within the first-order principle differential analysis model that takes into account the concerns above.

   For the sake of computational efficiency, you may use cross-validated values for smoothing parameters that are chosen outside the resampling schemes.