APTS Statistical Computing: Assessment 2011

The work provided here is intended to take students up to half a week to complete. Students should talk to their supervisors to find out whether or not their department requires this work as part of any formal accreditation process (APTS itself has no resources to assess or certify students). It is anticipated that departments will decide on the appropriate *level* of assessment locally, and may choose to drop some (or indeed all) of the parts, accordingly. So make sure that your supervisor or local organizer of APTS assessment has looked at the assignment before you start, and has told you which parts of it to do. In order to avoid undermining institutions' local assessment procedures the module lecturer will not respond to enquiries from students about this assignment.

This assignment is about "smoothing" with piecewise linear functions. Consider an increasing sequence of K x-values, $k_1, k_2 \dots k_K$. A piecewise linear function on $[k_1, k_K]$ can be defined as

$$f(x) = \sum_{i=1}^{K} \beta_i b_i(x)$$

where β_i is a coefficient (interpretable as the value of $f(k_i)$), and $b_i(x)$ is a *basis function*,

$$b_i(x) = \begin{cases} (k_{i+1} - x)/(k_{i+1} - k_i) & k_i \le x < k_{i+1}, \quad i < K \\ (x - k_{i-1})/(k_i - k_{i-1}) & k_{i-1} \le x < k_i, \quad i > 1 \\ 0 & \text{otherwise.} \end{cases}$$

The higher K is, the more flexible $f(x)$ will be.

1. Write an R function which will take a vector of x values at which to evaluate f, and a vector of k values, and will return the matrix $\mathbf{X}$ such that $X_{ij} = b_j(x_i)$. (For a bonus mark, extend your function so that $f(x)$ simply extends the first or last linear segment for x values outside the range $[k_1, k_K]$).

2. Examine the `mcycle` data from R package `MASS`. Use your function from part 1 and the `lm` command in R to fit piecewise linear models of the form

   $$\texttt{accel}_i = f(\texttt{times}_i) + \epsilon_i$$

   to the data. Produce a plot comparing models with $K = 10, 20, 30$ and 40, and equally spaced k_i values. i.e. produce a scatterplot of `accel` against `time`, and overlay the curves showing $\hat{f}(\texttt{times})$ against `times` for the different K values.

3. Controlling the complexity of the piecewise linear model by varying K is not very satisfactory. An alternative is to control the model's complexity by using a relatively large K, but penalizing the 'wiggliness' of the model as part of fitting. For example, if y is the response variable, the model can be fitted by minimizing a penalized version of the normal least squares objective,

   $$Q = \sum_{i=1}^{n} (y_i - f(t_i))^2 + \lambda \sum_{j=2}^{K-1} (\beta_{j-1} - 2\beta_j + \beta_{j+1})^2$$

   w.r.t. $\boldsymbol{\beta}$. λ is a smoothing parameter, controlling the tradeoff between smoothness of f and close fit to the data y_i. It can readily be verified that if $\mathbf{P}$ is a matrix of coefficients such as that produced by `diff(diff(diag(K)))`, in R, and if $\mathbf{X}$ is the matrix with entries $X_{ij} = b_j(t_i)$, then

   $$Q = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \boldsymbol{\beta}^\mathsf{T} \mathbf{P}^\mathsf{T} \mathbf{P} \boldsymbol{\beta}.$$

   Show that, for a given λ, the formal expression for the 'hat matrix' $\mathbf{A}$ such that $\hat{\boldsymbol{\mu}} = \mathbf{A}\mathbf{y}$ ($\boldsymbol{\mu} = \mathbb{E}(\mathbf{y})$), is given by $\mathbf{A} = \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda \mathbf{P}^\mathsf{T}\mathbf{P})^{-1}\mathbf{X}^\mathsf{T}$.

4. Produce a plot like the one from part 2, but now illustrating the smooth fits given by several alternative λ values (choose these to give results varying from very smooth to very wiggly, and include something in the middle that looks reasonably sensible for the data). Don't worry about stable or efficient computation for the moment, and use $K = 40$ with equally spaced k_i values.

5. One objective method for choosing the smoothing parameter λ is Generalized Cross Validation (GCV). This selects the λ value that minimizes

$$\mathcal{V}(\lambda) = \frac{n\|\mathbf{y} - \mathbf{A}\mathbf{y}\|^2}{(n - \text{tr}(\mathbf{A}))^2}$$

(the dependence of the r.h.s. on λ is via $\mathbf{A}$). In terms of n (the length of $\mathbf{y}$), what order would the computational cost of evaluating $\mathcal{V}$ be, for any trial value of λ, if the expression for $\mathbf{A}$ given in part 3 was used naively?

6. Let $\mathbf{X} = \mathbf{QR}$ be the QR decomposition of $\mathbf{X}$, and consider the singular value decomposition

$$\begin{pmatrix} \mathbf{PR}^{-1} \\ \mathbf{0} \end{pmatrix} = \mathbf{UDV}^\mathsf{T}$$

where $\mathbf{0}$ is a 2 row matrix of zeroes. Show that $\mathbf{A} = \mathbf{QV}(\mathbf{I} + \lambda\mathbf{D}^2)^{-1}\mathbf{V}^\mathsf{T}\mathbf{Q}^\mathsf{T}$. Hence show that $\text{tr}(\mathbf{A}) = \text{tr}((\mathbf{I} + \lambda\mathbf{D}^2)^{-1})$. Note that $\text{tr}(\mathbf{A})$ can be interpreted as the effective degrees of freedom of a penalized fit. (Hint: first substitute the QR decomposition of $\mathbf{X}$ into the expression for $\mathbf{A}$ and simplify. Then express $\mathbf{R}^{-\mathsf{T}}\mathbf{P}^\mathsf{T}\mathbf{PR}^{-1}$ in terms of the components of the given singular value decomposition, substitute and simplify further.)

7. Write an R function with arguments `y` and `x` containing the response variable to smooth, and the covariate to smooth with respect to (`accel` and `times` in the `mcycle` data example). The function should also take an argument `K` giving the number of basis functions to use.

   The function should smooth the x, y data using a penalized piecewise linear smoother of the sort introduced above. It should choose the smoothing parameter by GCV, by evaluating $\mathcal{V}$ for each $\log\lambda$ values generated by `seq(-20,20,length=100)`, and selecting the smoothing parameter yielding the smallest $\mathcal{V}$. The function should have $O(nK^2)$ computational cost overall, but you should ensure that the cost of each $\mathcal{V}$ evaluation in the $\log\lambda$ selection search is only $O(K)$. To do this use the results from part 6, and the fact that

$$\|\mathbf{y} - \hat{\boldsymbol{\mu}}\|^2 = \|\mathbf{V}^\mathsf{T}\mathbf{Q}^\mathsf{T}\mathbf{y} - (\mathbf{I} + \lambda\mathbf{D}^2)^{-1}\mathbf{V}^\mathsf{T}\mathbf{Q}^\mathsf{T}\mathbf{y}\|^2 + \|\mathbf{y}\|^2 - \|\mathbf{Q}^\mathsf{T}\mathbf{y}\|^2.$$

   The function should return a list containing 3 vectors: the fitted values $\hat{\boldsymbol{\mu}}$ corresponding to the GCV optimal smooth model, a 100-vector of $\log\lambda$ values, and a corresponding vector of GCV scores $\mathcal{V}$.

8. After testing your routine, produce a scatterplot of `accel` vs. `times` with the GCV optimal smooth estimates overlaid, alongside a plot of the GCV score against effective degrees of freedom.