University of Washington

STAT / BIOSTAT 527 NONPARAMETRIC REGRESSION AND CLASSIFICATION

# Homework 3

**Issued:** Thursday, April 18, 2013 **Due:** Thursday, April 25, 2013

---

1. Please include your code with your solution.
2. Please hand in your 1-page project proposal (hard copy) in class on Thursday April 25. You should use the proposal format suggested on the course website, in the section of "Project Proposal" under "Project".

---

## Problem 3.1

Consider the truncated power series representation for cubic splines with $K$ interior knots. Let

$$f(X) = \sum_{j=0}^{3} \beta_j X^j + \sum_{k=1}^{K} \theta_k (X - \xi_k)_+^3. \tag{1}$$

Prove that the natural boundary conditions for natural cubic splines (Section 5.2.1 in the book "Elements of Statistical Learning") imply the following linear constraints on the coefficients:

$$\begin{aligned} \beta_2 &= 0, \qquad \sum_{k=1}^{K} \theta_k = 0, \\ \beta_3 &= 0, \qquad \sum_{k=1}^{K} \xi_k \theta_k = 0. \end{aligned}$$

Hence derive the basis for natural cubic spline:

$$N_1(X) = 1, \quad N_2(X) = X, \quad N_{k+2}(X) = d_k(X) - d_{K-1}(X), \tag{2}$$

where

$$d_k(X) = \frac{(X - \xi_k)_+^3 - (X - \xi_K)_+^3}{\xi_K - \xi_k}. \tag{3}$$

Each of these basis functions can be seen to have zero second and third derivative for $X \geq \xi_K$.

## Problem 3.2

In this question you will fit various spline models to the fossil data (in "fossil.csv") of Chaudhuri and Marron (1999). These data consist of 106 measurements of ratios of strontium isotopes found in fossil shells and their ages. Make a scatterplot of the data, superimposed with the fitted curves by (a),(b),(c) and part (i) of (d). Compare and discuss the fitted models.

(a) Fit a natural cubic spline, using ordinary cross-validation for choice of smoothing parameter (this model has $n$ knots). Report the selected smoothing parameter and effective degrees of freedom.
*Hint: The function* `sm.spline(,cv=TRUE)` *from the* `pspline` *package in* `R` *might be helpful.*

(b) Fit a natural cubic spline, using generalized cross-validation for choice of smoothing parameter (this model has $n$ knots). Report the selected smoothing parameter and effective degrees of freedom.
*Hint: The function* `sm.spline(,cv=FALSE)` *from the* `pspline` *package in* `R` *might be helpful.*

(c) Fit a penalized cubic regression spline with $K = 20$ equally-spaced knots, using ordinary cross-validation for choice of smoothing parameter. Report the selected smoothing parameter and effective degrees of freedom.
*Hint: The function* `gam(, knots=··· , sp=··· )` *from the* `mgcv` *package in* `R` *might be helpful. But it doesn't output the OCV score. Please write your own code to calculate OCV from the hat matrix.*

(d) Fit a penalized cubic regression spline with $K = 20$ knots, using generalized cross-validation for choice of smoothing parameter. Report the selected smoothing parameter and effective degrees of freedom.

(i) Fit the model with $K = 20$ equally-spaced knots in the range of the observed covariate.

(ii) Fit the model with $K = 20$ random knots in the range of the observed covariate. Repeat fitting this model 10 times (each with a set of 20 random knots). Plot the 10 fitted curves and the data points. Does the location of 20 knots change your fitted curve?

*Hint: The function* `gam(, method="GCV.Cp", knots=··· )` *from the* `mgcv` *package in* `R` *might be helpful.*

For the models in (c) and (d) part (i), give $\widehat{f}(x)$ and an asymptotic 95% confidence interval, for the (smoothed) function, at $x = 95$ and $x = 115$ years. For the models in (a) and (b), only report $\widehat{f}(x)$ at $x = 95$ and $x = 115$ years.