STAT 260 - Nonparametric Statistics

Ying Sun
Statistics (STAT) Program
Computer, Electrical and Mathematical Sciences & Engineering (CEMSE) Division
King Abdullah University of Science and Technology (KAUST)

HOMEWORK V

Henrique Aparecido Laureano

Spring Semester 2018

Contents

Problem 1

Read section 5.3.1 in the textbook. Answer the questions for cubic regression spline (5.3).

(a)

Show that the second derivative of the spline can be expressed as

$$f''(x) = \sum_{i=2}^{k-2} \delta_i d_i(x)$$

where

$$d_i(x) = \begin{cases} (x - x_{i-1})/h_{i-1}, & x_{i-1} \le x \le x_i, \\ (x_{i+1} - x)/h_i, & x_i \le x \le x_{i+1}, \\ 0, & \text{otherwise.} \end{cases}$$

Solution:

The spline can be written as

$$f(x) = \frac{x_{i+1} - x}{h_i}\beta_i + \frac{x - x_i}{h_i}\beta_{i+1} + \frac{(x_{i+1} - x)^3/h_i - h_i(x_{i+1} - x)}{6}\delta_i + \frac{(x - x_i)^3/h_i - h_i(x - x_i)}{6}\delta_{i+1},$$

if $x_i \le x \le x_{i+1}$.

Differentiating we have,

$$f'(x) = \frac{1}{6}\frac{3(x_{i+1} - x)^2}{h_i}\delta_i + \frac{1}{6}\frac{3(x - x_i)^2}{h_i}\delta_{i+1}, \quad \text{if} \quad x_i \le x \le x_{i+1}.$$

Differentiating again we have,

$$f''(x) = \frac{x_{i+1} - x}{h_i}\delta_i + \frac{x - x_i}{h_i}\delta_{i+1}, \quad \text{if} \quad x_i \le x \le x_{i+1}.$$

What can be re-written simply as

$$f''(x) = \sum_{i=2}^{k-2} \delta_i d_i(x) \quad \text{where} \quad d_i(x) = \begin{cases} (x - x_{i-1})/h_{i-1}, & x_{i-1} \le x \le x_i, \\ (x_{i+1} - x)/h_i, & x_i \le x \le x_{i+1}, \\ 0, & \text{otherwise.} \end{cases}$$

□

(b)

Hence show that, in the notation of section 5.3.1,

$$\int f''(x)^2 dx = \delta^{-\top} B \delta^-$$

Solution:

$$f''(x) = \frac{x_{i+1} - x}{h_i}\delta_i + \frac{x - x_i}{h_i}\delta_{i+1} = \sum_{i=2}^{k-2} \delta_i d_i(x) \quad \text{where} \quad d_i(x) = \begin{cases} (x - x_{i-1})/h_{i-1}, & x_{i-1} \le x \le x_i, \\ (x_{i+1} - x)/h_i, & x_i \le x \le x_{i+1}, \\ 0, & \text{otherwise.} \end{cases}$$

and the matrix B used to define the cubic regression spline

$$B_{i,i} = \frac{h_i + h_{i+1}}{3}, \quad i = 1, \ldots, k-2, \qquad B_{i,i+1} = \frac{h_{i+1}}{6} \quad \text{and} \quad B_{i+1,i} = \frac{h_{i+1}}{6}, \quad i = 1, \ldots, k-3.$$

Also $h_i = x_{i+1} - x_i$.

We can re-written

$$\int f''(x)^2 dx = \delta^{-\top} \int d(x) d(x)^\top dx \delta^-,$$

with $d(x)$ being a vector with i-th element $d_{i+1}(x)$ and with the first and last elements having coefficients zero.

Each $d_i(x)$ is non-zero over only 2 intervals, is easy to see that $\int d(x)d(x)^\top dx$ is tri-diagonal and symmetric. The $i-1$-th leading diagonal element is given by

$$\begin{aligned} \int_{x_{i-1}}^{x_{i+1}} d_i(x)^2 dx &= \int_{x_{i-1}}^{x_i} \frac{(x - x_{i-1})^2}{h_{i-1}^2} dx - \int_{x_i}^{x_{i+1}} \frac{(x_{i+1} - x)^2}{h_i^2} dx = \frac{(x - x_{i-1})^3}{3h_{i-1}^2}\Big|_{x_{i-1}}^{x_i} - \frac{(x_{i+1} - x)^3}{3h_i^2}\Big|_{x_i}^{x_{i+1}} \\ &= \frac{(x_i - x_{i-1})^3}{3h_{i-1}^2} + \frac{(x_{i+1} - x_i)^3}{3h_i^2} \\ &= \frac{h_{i-1}^3}{3h_{i-1}^2} + \frac{h_i^3}{3h_i^2} = \frac{h_{i-1}}{3} + \frac{h_i}{3} = B_{i,i}, \quad i = 2, \ldots, k-1. \end{aligned}$$

Following the same reasoning the off-diagonal elements $(i-1, i)$ and $(i, i-1)$ are given by

$$\int_{x_{i-1}}^{x_i} d_i(x) d_{i-1}(x) dx = \int_{x_{i-1}}^{x_i} \frac{x - x_{i-1}}{h_{i-1}} \frac{x_i - x}{h_{i-1}} dx = \frac{h_{i-1}}{6} = B_{i-1,i} \quad \text{and} \quad B_{i,i-1}, \quad i = 3, \ldots, k-1.$$

In this way we see that $\int d(x)d(x)^\top dx = B$, and therefore,

$$\int f''(x)^2 dx = \delta^{-\top} B \delta^-.$$

□

(c)

Finally show that

$$\int f''(x)^2 dx = \beta^\top D^\top B^{-1} D\beta$$

Solution:

$$\int f''(x)^2 dx = \delta^{-\top} B \delta^{-}.$$

From Equation (5.4) we know that

$$B\delta^- = D\beta \quad \Rightarrow \quad \delta^- = B^{-1}D\beta.$$

Then,

$$\delta^{-\top} = (B^{-1}D\beta)^\top = \beta^\top D^\top (B^{-1})^\top,$$

and therefore,

$$\int f''(x)^2 dx = \beta^\top D^\top (B^{-1})^\top B B^{-1} D\beta = \int f''(x)^2 dx = \beta^\top D^\top B^{-1} D\beta.$$

Given the symmetry of B, $(B^{-1})^\top B = (B^{-1})^\top B^\top = I$.

□

Problem 2

Read section 5.4.2 in the textbook. The natural parameterization is particularly useful for understanding the way in which penalization causes bias in estimates, and this question explores this issue.

(a)

Find an expression for the bias in a parameter estimator $\hat{\beta}_i''$ in the natural parameterization (bias being defined as $\mathbb{E}\{\hat{\beta}_i'' - \beta_i''\}$). What does this tell you about the bias in components of the model which are unpenalized, or only very weakly penalized, and in components for which the 'true value' of the corresponding parameter is zero or nearly zero?

Solution:

We know that

$$\mathbb{E}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^\top\mathbf{X} + \boldsymbol{\lambda}\mathbf{S})^{-1}\mathbf{X}^\top\mathbb{E}(\mathbf{y}) = (\mathbf{X}^\top\mathbf{X} + \boldsymbol{\lambda}\mathbf{S})^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta},$$

and that in the natural parametrization

$$\mathbb{E}(\hat{\boldsymbol{\beta}}'') = (\mathbf{I} + \boldsymbol{\lambda}\mathbf{D})^{-1}\boldsymbol{\beta}''.$$

Then,

$$\text{bias}(\hat{\beta}_i'') = \mathbb{E}\{\hat{\beta}_i'' - \beta_i''\} = \frac{\beta_i'' - (1 + \lambda D_{ii})\beta_i''}{(1 + \lambda D_{ii})} = \frac{-\beta_i''\lambda D_{ii}}{(1 + \lambda D_{ii})}.$$

If $\beta_i'' = 0$ or $\lambda D_{ii} = 0$, then the estimator is unbiased.

The bias will be small for small 'true parameter value' or weakly penalization. Just moderate or strongly penalizations of substantial magnitude that are subject to substantial bias.

□

(b)

The mean square error in a parameter estimator (MSE) is defined as $\mathbb{E}\{(\hat{\beta}_i - \beta_i)^2\}$ (dropping the primes for notational convenience). Show that the MSE of the estimator is in fact the estimator variance plus the square of the estimator bias.

Solution:

$$\begin{aligned}
\mathbb{E}\{(\hat{\beta}_i - \beta_i)^2\} &= \mathbb{E}\{(\hat{\beta}_i - \mathbb{E}(\hat{\beta}_i) + \mathbb{E}(\hat{\beta}_i) - \beta_i)^2\} \\
&= \mathbb{E}\{(\hat{\beta}_i - \mathbb{E}(\hat{\beta}_i))^2\} + \mathbb{E}\{(\mathbb{E}(\hat{\beta}_i) - \beta_i)^2\} + \mathbb{E}\{(\hat{\beta}_i - \mathbb{E}(\hat{\beta}_i))(\mathbb{E}(\hat{\beta}_i) - \beta_i)\} \\
&= \mathbb{V}(\hat{\beta}_i) + \text{bias}(\hat{\beta}_i)^2 + 0 \\
&= \mathbb{V}(\hat{\beta}_i) + \text{bias}(\hat{\beta}_i)^2.
\end{aligned}$$

□

(c)

Find an expression for the mean square error of the i-th parameter of a smooth in the natural parameterization.

Solution:

The variance expression is given in page 212.

$$\text{Mean Square Error}: \quad \begin{aligned} \text{MSE} &= \mathbb{V}(\hat{\beta}_i) + \text{bias}(\hat{\beta}_i)^2 \\ &= \frac{\sigma^2}{(1+\lambda D_{ii})^2} + \frac{(\beta_i \lambda D_{ii})^2}{(1+\lambda D_{ii})^2} \\ &= \frac{\sigma^2 + (\beta_i \lambda D_{ii})^2}{(1+\lambda D_{ii})^2}. \end{aligned}$$

□

(d)

Show that the lowest achievable MSE, for any natural parameter, is bounded above by σ^2, implying that penalization always has the potential to reduce the MSE of a parameter *if the right smoothing parameter value is chosen*. Comment on the proportion of the minimum achievable MSE that is contributed by the squared bias term, for different magnitudes of parameter value.

Solution:

We can write

$$\frac{\text{MSE}}{\sigma^2} = \frac{1 + (\beta_i^2/\sigma^2)\lambda^2 D_{ii}^2}{(1+\lambda D_{ii})^2}.$$

Minimizing in λ we have

$$\begin{aligned} \frac{\partial \text{MSE}/\sigma^2}{\partial \lambda} = 0 \quad \Rightarrow \quad 2\lambda \frac{\beta_i^2}{\sigma^2} D_{ii}^2 (1+\lambda D_{ii})^2 &= 2(1 + \frac{\beta_i^2}{\sigma^2}\lambda^2 D_{ii}^2) D_{ii}(1+\lambda D_{ii}) \\ \lambda \frac{\beta_i^2}{\sigma^2} D_{ii}(1+\lambda D_{ii}) &= 1 + \frac{\beta_i^2}{\sigma^2}\lambda^2 D_{ii}^2 \\ \lambda \frac{\beta_i^2}{\sigma^2} D_{ii} + \frac{\beta_i^2}{\sigma^2}\lambda^2 D_{ii}^2 &= 1 + \frac{\beta_i^2}{\sigma^2}\lambda^2 D_{ii}^2 \\ \lambda \frac{\beta_i^2}{\sigma^2} D_{ii} &= 1 \\ \lambda^* &= \frac{\sigma^2}{D_{ii}\beta_i^2}. \end{aligned}$$

Putting this λ^* in MSE we obtain

$$\begin{aligned} \text{MSE} = \frac{\sigma^2 + \beta_i^2 (\lambda^*)^2 D_{ii}^2}{(1+\lambda^* D_{ii})^2} = \frac{\sigma^2 + \beta_i^2 \frac{\sigma^4}{D_{ii}^2 \beta_i^4} D_{ii}^2}{\left(1 + \frac{\sigma^2}{D_{ii}\beta_i^2} D_{ii}\right)^2} = \frac{\sigma^2 + \frac{\sigma^4}{\beta_i^2}}{\left(1 + \frac{\sigma^2}{\beta_i^2}\right)^2} = \frac{\frac{\sigma^2 \beta_i^2 + \sigma^4}{\beta_i^2}}{\left(\frac{\beta_i^2 + \sigma^2}{\beta_i^2}\right)^2} = \frac{\sigma^2 \beta_i^2 + \sigma^4}{\beta_i^2} \frac{\beta_i^4}{(\beta_i^2 + \sigma^2)^2} \\ = \sigma^2 (\beta_i^2 + \sigma^2) \frac{\beta_i^2}{(\beta_i^2 + \sigma^2)^2} = \frac{\sigma^2 \beta_i^2}{\beta_i^2 + \sigma^2}. \end{aligned}$$

So the lowest achievable MSE is $\sigma^2\beta_i^2/(\beta_i^2+\sigma^2)$. Comparing with σ^2 we see that

$$\frac{\sigma^2\beta_i^2}{\beta_i^2+\sigma^2}=\sigma^2 \quad \Rightarrow \quad \sigma^2\beta_i^2=\sigma^2(\beta_i^2+\sigma^2) \quad \Rightarrow \quad \beta_i^2 \leq \beta_i^2+\sigma^2.$$

In the natural parameterization the unpenalized estimator variance and unpenalized MSE is σ^2. $\sigma^2\beta_i^2/(\beta_i^2+\sigma^2)$ is always smaller than σ^2.

If λ could be chosen to minimize the MSE for a given parameter, then from $\sigma^2\beta_i^2/(\beta_i^2+\sigma^2)$ is clear that small magnitude β_i's would lead to high penalization of MSE dominated by the bias term, while large magnitude β_i's would lead to low penalization of MSE dominated by the variance.

□

Problem 3

Your R function for fitting a penalized regression spline (slides Topic 9 page 38)

```
# <r code> ============================================================================ #
# data: it is often claimed, at least by people with little actual knowledge of
#       engines, that a car engine with a larger cylinder capacity will wear out
#       less quickly than a smaller capacity engine.

# the data were collected from 19 Volvo engines.

          # reading the data and scaling the engine capacity data to lie in [0, 1]
size <- c(1.42, 1.58, 1.78, 1.99, 1.99,
          1.99, 2.13, 2.13, 2.13, 2.32,
          2.32, 2.32, 2.32, 2.32, 2.43,
          2.43, 2.78, 2.98, 2.98)

wear <- c(4.0, 4.2, 2.5, 2.6, 2.8,
          2.4, 3.2, 2.4, 2.6, 4.8,
          2.9, 3.8, 3.0, 2.7, 3.1,
          3.3, 3.0, 2.8, 1.7)

x <- size - min(size) ; x <- x / max(x)

par(mar = c(5, 4, 2, 2) + .1)                                  # graphical definition
plot(x, wear, xlab = "Scaled engine size", ylab = "Wear index", pch = 16)
# </r code> ============================================================================ #
```

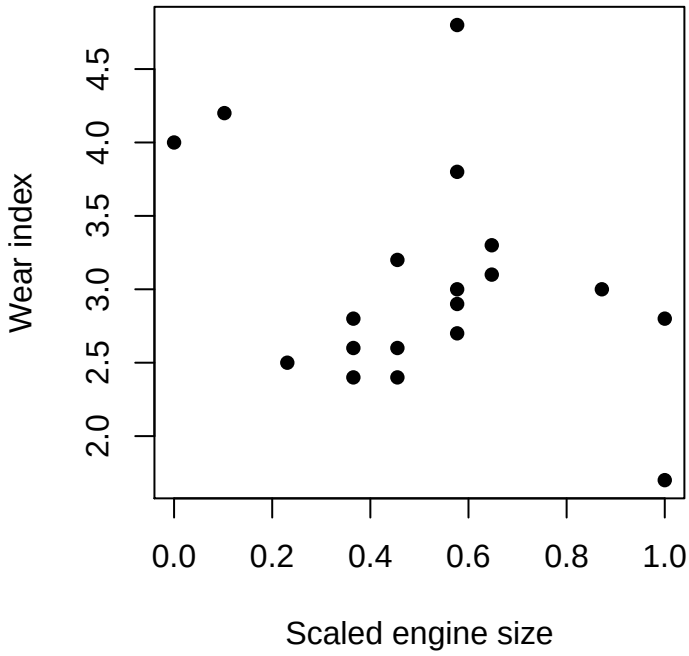

Figure 1: Data scatter plot.

```
# <r code> ============================================================================ #
                                                    # stablishing basis function
rk <- function(x, z){                          # R(x, z) for cubic spline on [0, 1]
  ((z - .5)**2 - 1/12) * ((x - .5)**2 - 1/12) / 4 -
    ( (abs(x - z) - .5)**4 - (abs(x - z) - .5)**2 / 2 + 7 / 240 ) / 24}
# taking a sequence of knots and an array of x values to produce a design matrix X
#             for the spline (setting up model matrix for cubic regression spline)
spl.X <- function(x, xk){                 # x the data vector, x_{k} the knot vector
  q = length(xk)                                                  # number of knots
  p = q + 2                                                  # number of parameters
  n = length(x)                                                    # number of data
  X = matrix(1, n, p)                                     # initialized model matrix
  X[ , 2] = x                                               # set second column to x
  X[ , 3:p] = outer(x, xk, FUN = rk)                 # and remaining to R(x, x_{k})
  X}
                    # setting up the penalized regression spline penalty matrix S,
                    #                                    given knot sequence x_{k}
spl.S <- function(xk){
  q = length(xk)
  p = q + 2
  S = matrix(0, p, p)                                      # initialize matrix to 0
  S[3:p, 3:p] = outer(xk, xk, FUN = rk)                     # fill in non-zero part
  S}
                           # fitting a penalized regression spline to x, y data,
                           # with knots x_{k}, given smoothing parameter, lambda
```

```r
prs.fit <- function(y, x, xk, lambda){
  X = spl.X(x, xk)                                  # computing design matrix X
  S = spl.S(xk)                                    # computing penalty matrix S
  inv = solve(t(X) %*% X + lambda * S)                      # computing inverse
  beta = inv %*% t(X) %*% y                          # computing the \beta's
  hat = X %*% inv %*% t(X)                         # computing the hat matrix
  hat.y = hat %*% y                                        # computing \hat{Y}
  return(list(coef = beta, fitted = hat.y))}       # returning \beta's and \hat{Y}
prs <- prs.fit(y = wear, x = x, xk = 1:7 / 8, lambda = 1e-4)              # fitting

par(mfrow = c(1, 2), mar = c(5, 4, 2, 2) + .1)             # graphical definitions
plot(x, wear, xlab = "Scaled engine size", ylab = "Wear index", pch = 16)
lines(x, prs$fitted, lwd = 2, col = "#0080ff")                # plotting \hat{Y}

plot(x, wear, xlab = "Scaled engine size", ylab = "Wear index", pch = 16)
   # building a prediction matrix and plotting the fitted spline for this new data
lines(0:100/100, spl.X(0:100/100, 1:7 / 8) %*% prs$coef, lwd = 2, col = "#0080ff")
# </r code> ========================================================================== #
```

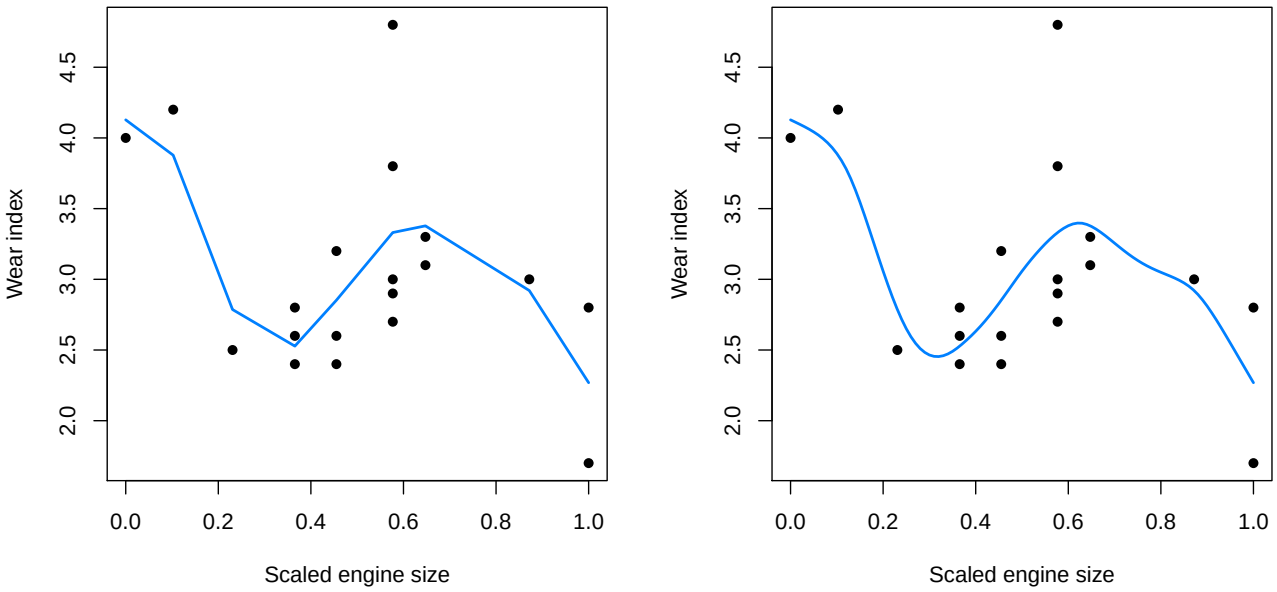

Figure 2: Scatter plots with fitted splines in blue. In the left we have $\widehat{Y} = A(\lambda)Y$, but the curve is sharp, since we have only 19 data points. In the right we have the fitted spline for a grid of 100 points, as consequence the curve is much more smooth.

Here we used $q = 7$ knots, evenly spread over $[0, 1]$, and a $\lambda = 0.0001$ (best fit in slides Topic 9 page 39).

■