EXERCISE 15

WEIYU LI

1. Consider the following regression model

$$Y = f(X) + \epsilon, f(X) = \frac{\sin(12(X+0.2))}{X+0.2},$$

with $X \sim U(0,1), \epsilon \sim N(0,1)$. Generate $N = 100$ observations (x_i, y_i) at random.
(1) Fit the data via smooth spline, and choose the best tuning parameter via cross-validation.
(2) Plot the fitting curves with different degrees of freedom $df = 5, 9, 15$ and real curve. Then draw the pointwise confidence band.

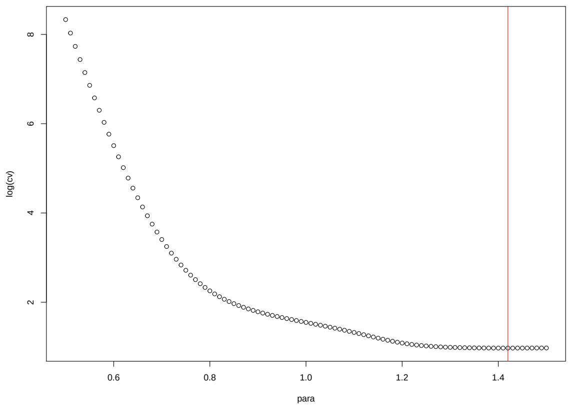
Solve. In the lecture, we don't specify how to estimate the confidence band. However, we can use what we have learned before, *e.g.* bootstrap method. My exampled code is based on `http://www.stat.cmu.edu/~cshalizi/402/lectures/11-splines/lecture-11.pdf`.

```
set.seed(0)
f <- function(x) sin(12 * (x + 0.2)) / (x + 0.2)
N <- 100
x <- runif(N, 0, 1)
e <- rnorm(N, 0, 1)
y <- f(x) + e

### (1)
para <- seq(0.5, 1.5, by = .01) # these are candidates of the parameter
cv <- para # these are the CV scores corresponding to each parameter
for (i in 1:length(para)){
  cv[i] <- smooth.spline(x, y, spar = para[i], cv = TRUE, all.knots = TRUE)$cv.crit
  # note that the problem asks for CV
  # otherwise setting cv = False gives the GCV scores
  # the two scores give similar but not same choice of parameter spar
}
plot(para, log(cv))
abline(v = para[which.min(cv)], col = 2)
cat("The best bandwidth via CV is spar = ", para[which.min(cv)])
# Output -> The best bandwidth via CV is spar =  1.42
## remark: in this example, the best choice of spar is 1.42
## one can also grid search the parameter lambda instead
## when spar is 1.42, the corresponding default lambda is 0.326.
```

The output image is at the top of the next page.

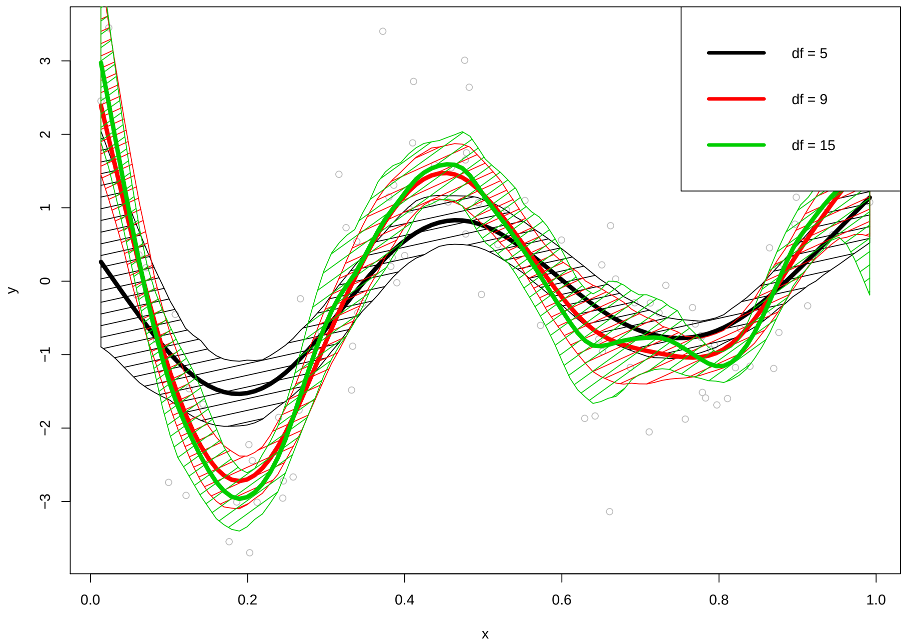
Date: 2019/12/02.
liweiyu@mail.ustc.edu.cn.




```
### (2) we draw the fitted curves
### as well as their corresponding bootstrap confidence bands
sp.boot.estimator <- function(x, y, xx, df) {
  # predictors of points at xx,
  # given training data (x,y), with degrees of freedom df
  index <- sample(1:N, size = N, replace = TRUE)
  fit <- smooth.spline(x = x[index], y = y[index], df = df,
                       cv = TRUE, all.knots = TRUE)
  return(predict(fit, x = xx)$y)
}
sp.spline.cis <- function(B, alpha, xx, df, fhat) {
  # draw B bootstrap samples, fit the spline to each,
  # then get the bootstrap confidence bounds
  spline.boots <- replicate(B, sp.boot.estimator(x, y, xx, df))
  cis.lower <- 2 * fhat - apply(spline.boots, 1, quantile, probs = 1 - alpha / 2)
  cis.upper <- 2 * fhat - apply(spline.boots, 1, quantile, probs = alpha / 2)
  return(c(cis.lower, rev(cis.upper)))
}

df=c(5, 9, 15)
xx <- seq(min(x), max(x), length=101)
plot(x, y, col = "gray")
for (i in 1:3) {
  fhat <- predict(smooth.spline(x, y, df = df[i]), xx)$y
  lines(xx, fhat, lwd = 5, col = i, lty = 1)
  sp.cis <- sp.spline.cis(B = 200, alpha = 0.05, xx, df = df[i], fhat)
  polygon(x = c(xx, rev(xx)), y = sp.cis,
          col = i, density = 20, angle = 10 * i)
}
legend("topright", legend = c("df = 5", "df = 9", "df = 15"),
       lwd = rep(4,3), lty = rep(1,3), col = 1:3)
```

□



2. Solve the following optimization problem

$$\min_f RSS(f,\lambda) = \sum_{i=1}^n w_i(y_i - f(x_i))^2 + \lambda \int \{f''(t)\}^2 dt,$$

where $w_i \geq 0$ are weights. Using the solution to study optimization problem of smoothing splines in the case that there exist ties in the observation points, *i.e.*, $x_i = x_j$ **for some** $i \neq j$**.**

Solve. From the theorem in Page 48 of *Lec15.pdf*, natural cubic splines always have smaller cost. Thus, with basis $e_1, \dots, e_n$ for the natural cubic spline with knots $x_1, \dots, x_n$, the minimizor is in the form of $\sum_{j=1}^n \beta_j^* e_j$. Equivalently,

$$\begin{aligned}\beta^* &= \arg\min_{\beta \in \mathbb{R}^n} \sum_{i=1}^n w_i(y_i - \sum_{j=1}^n \beta_j e_j(x_i))^2 + \lambda \int \{\sum_{j=1}^n \beta_j e_j''(t)\}^2 dt \\ &= \arg\min_{\beta \in \mathbb{R}^n} (y - N\beta)' W (y - N\beta) + \lambda \beta' \Omega \beta,\end{aligned}$$

where $N_{ij} = e_j(x_i)$, $W = diag(w_1, \dots, w_n)$ and $\Omega_{ij} = \int \{e_i''(t)\}^2 \{e_j''(t)\}^2 dt$ (c.f., Page 53 of *Lec15.pdf*). Then the minimizer is

$$\beta^* = (N'WN + \lambda\Omega)^{-1} N'Wy.$$

Therefore, if there exist ties, suppose that the knots are $x_1, \dots, x_k$ with $n_i \geq 1$ observations at knot x_i, then $n = \sum_{i=1}^k n_i$. In this case, we fit the natural spline with knots $x_1, \dots, x_k$ and weights $w_i = n_i$. Thus the fitted smoothing spline is $\sum_{j=1}^n \beta_j^* e_j$. □