Homework 15

Liu Huihang
SA18017026
QQ: 184050473
MAIL: huihang@mail.ustc.edu.cn

Problem 1 (15.1) *Consider the following regression model*

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

Let $X \sim U(0, 1), \epsilon \sim N(0, 1)$. *Randomly generate* $N = 100$ *samples* (x_i, y_i),

(1) Use smooth spline to fit the data set, and use CV to select the best turing parameters.

(2) Draw the fitted curve and real curve under different degrees of freedom $df(5, 9, 15)$, *with point-by-point confidence band.*

Problem 2 (15.1) *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 weighted value to observations. Use this conclusion to investigate the solution of the optimization problem of smooth splines when there are ties in the observation points (that is, there are duplicates in the data).*

Solution

We can choose a basis $\eta_1, \ldots, \eta_n$ for the set of kth-order natural splines with knots over $x_1, \ldots, x_n$, and reparametrize the problem into a finite-dimensional problem. Let $(\hat{f}(x_1), \ldots, \hat{f}(x_n))^T = N\hat{\beta}$, where basis matrix N from the B-spline basis with $N_{ij} = \eta_j(x_i)$. Then we can rewrite the equation as following

$$\min_f RSS(f, \lambda) = (y - N\beta)^T W(y - N\beta) + \lambda \beta^T \Omega \beta,$$

where $W = diag(w_1, \ldots, w_n)$, $y = (y_1, \ldots, y_n)$, Ω is penalty matrices with $\Omega_{ij} = \int \eta_i''(x)\eta_j''(x)dx$.

$$\frac{\partial RSS}{\partial \beta} = -2N^T W(y - N\beta) + 2\lambda\Omega\beta$$

Let $\partial RSS/\partial\beta = 0$, we can obtain

$$\hat{\beta} = (N^T W N + \lambda\Omega)^{-1} N^T W y.$$

If there are ties within the data. Assume that there are n_0 distinct observations, and each observation x_i appears n_i times.

Then, we can rewrite the optimization problem

$$\begin{aligned}
\min_f RSS(f,\lambda) &= \sum_{i=1}^{n} w_i(y_i - f(x_i))^2 + \lambda \int \{f''(t)\}^2 dt \\
&= \sum_{i=1}^{n_0}\sum_{j=1}^{n_i} w_i(y_{ij} - f(x_i))^2 + \lambda \int \{f''(t)\}^2 dt \\
&= \sum_{i=1}^{n_0}\sum_{j=1}^{n_i} w_i(y_{ij}^2 - 2y_{ij}f(x_i) + f(x_i)^2) + \lambda \int \{f''(t)\}^2 dt \\
&= \sum_{i=1}^{n_0} n_i w_i \left(\overline{y_{i\cdot}^2} - 2\overline{y_{i\cdot}} f(x_i) + f(x_i)^2\right) + \lambda \int \{f''(t)\}^2 dt. \\
&= \sum_{i=1}^{n_0} n_i w_i \left(\overline{y_{i\cdot}}^2 - 2\overline{y_{i\cdot}} f(x_i) + f(x_i)^2\right) - \sum_{i=1}^{n_0}\sum_{j\neq k} y_{ij}y_{ik} + \lambda \int \{f''(t)\}^2 dt. \\
&\quad \text{(the second term is ignored)} \\
&= \sum_{i=1}^{n_0} n_i w_i \left(\overline{y_{i\cdot}}^2 - 2\overline{y_{i\cdot}} f(x_i) + f(x_i)^2\right) + \lambda \int \{f''(t)\}^2 dt.
\end{aligned}$$

Further, let $y_{new} = (\overline{y_{1\cdot}}, \cdots, \overline{y_{n_0\cdot}})^T$, $W_{new} = diag(n_1w_1, \ldots, n_{n_0}w_{n_0})$, $N_{ij} = \eta_j(x_i)$ for $i = 1, \ldots, n_0$ and we can obtain

$$\hat{\beta}_{ties} = (N^T W_{new} N + \lambda\Omega)^{-1} N^T W_{new} y_{new}.$$

■

Problem 3 (16.1) *Consider the following regression model*

$$y_i = f(x_i) + e_i, i = 1, \ldots, n.$$

Let $N_j(x), j = 1, \ldots, n$ *be the basis of cubic nature spline,* $f(x) = \sum_{j=1}^{n} N_j(x)\beta_j$ *and* $\Omega = (\Omega_{jk}), \Omega_{jk} = \int N_j'' N_k''(t)dt$.

(1) Obtain the estimation of $\beta = (\beta_1, \ldots, \beta_n)$ *using smooth spline method.*

(2)Let $S_\lambda = N(N^T N + \lambda\Omega)^{-1} N^T$*, prove*

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}^{(-i)}(x_i)\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{y_i - \hat{f}(x_i)}{1 - S_\lambda(i,i)}\right]^2.$$

where $\hat{f}^{(-i)}(x_i)$ *represents the estimation of* f *at* x_i *with removing* i*-th sample.*

Solution

(1) Let $\hat{\beta}$ be the minimizer of the OLS, there are n different observations and n variables,

$$\frac{1}{n}\|y - N\beta\|^2$$

Then we have $\hat{\beta} = (N^T N)^{-1} N^T y$.

(2) Let $\hat{f}^{(-i)}$ be the minimizer of the PLS based on all observations except (x_i, y_i)

$$\frac{1}{n}\sum_{j\neq i}(y_j - \mathcal{L}_j f)^2 + \lambda \|P_1 f\|^2$$

For any fixed i, $\hat{\beta}^{-i}$ is the minimizer of

$$\frac{1}{n}\left(\mathcal{L}_i \hat{f}^{-(i)} - \mathcal{L}_i f\right)^2 + \frac{1}{n}\sum_{j\neq i}(y_j - \mathcal{L}_j f)^2 + \lambda \|P_1 f\|^2$$

because for any function f, we have

$$\begin{aligned}
&\frac{1}{n}\left(\mathcal{L}_i \hat{f}^{-(i)} - \mathcal{L}_i f\right)^2 + \frac{1}{n}\sum_{j\neq i}(y_j - \mathcal{L}_i f)^2 + \lambda \|P_1 f\|^2 \\
&\geq \frac{1}{n}\sum_{j\neq i}(y_j - \mathcal{L}_i f)^2 + \lambda \|P_1 f\|^2 \\
&\geq \frac{1}{n}\sum_{j\neq i}\left(y_j - \mathcal{L}_j \hat{f}^{-(i)}\right)^2 + \lambda \left\|P_1 \hat{f}^{-(i)}\right\|^2 \\
&= \frac{1}{n}\left(\mathcal{L}_i \hat{f}^{-(i)} - \mathcal{L}_i \hat{f}^{-(i)}\right)^2 + \frac{1}{n}\sum_{j\neq i}\left(y_j - \mathcal{L}_j \hat{f}^{-(i)}\right)^2 + \lambda \left\|P_1 \hat{f}^{-(i)}\right\|^2
\end{aligned}$$

It indicates that the solution to the PLS without the ith observation, $\hat{f}^{(-i)}$, is also the solution to the OLS with the ith observation (x_i, y_i) being replaced by the fitted value $\mathcal{L}_i \hat{f}^{-(i)}$.

Note that the hat matrix S_λ depends on the model space and operators $\mathcal{L}_i$ only. It does not depend on observations of the dependent variable. Therefore, $\hat{f} = S_\lambda y$ and $\hat{f}^{(-i)} = S_\lambda y^{-i}$.

$$\begin{aligned}
\mathcal{L}_i \hat{f} &= \sum_{j=1}^{n} S_\lambda(i,j) y_j \\
\mathcal{L}_i \hat{f}^{-(i)} &= \sum_{j\neq i} S_\lambda(i,j) y_j + S_\lambda(i,i)\mathcal{L}_i \hat{f}^{-(i)}
\end{aligned}$$

Solving for $\mathcal{L}_i \hat{f}^{(-i)}$, we have

$$\mathcal{L}_i \hat{f}^{-(i)} = \frac{\mathcal{L}_i \hat{f} - S_\lambda(i,i) y_i}{1 - S_\lambda(i,i)}$$

Then

$$y_i - \hat{f}^{(-i)}(x_i) = \frac{y_i - \hat{f}(x_i)}{1 - S_\lambda(i,i)}$$

and

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}^{(-i)}(x_i)\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{y_i - \hat{f}(x_i)}{1 - S_\lambda(i,i)}\right]^2.$$

■

I don't understand the last question, please help me.