5.9

Derive the Reinsch form $\mathbf{S}_\lambda = (\mathbf{I} + \lambda\mathbf{K})^{-1}$ for the smoothing spline.

We have

$$\begin{aligned}\mathbf{S} &= \mathbf{N}(\mathbf{N}^T\mathbf{N} + \lambda\mathbf{\Omega}_N)^{-1}\mathbf{N}^T \\ &= \mathbf{N}(\mathbf{N}^T(\mathbf{I} + \lambda(\mathbf{N}^T)^{-1}\mathbf{\Omega}_N\mathbf{N}^{-1})\mathbf{N})^{-1}\mathbf{N}^T \\ &= (\mathbf{I} + \lambda\mathbf{K})^{-1}\end{aligned}$$

where $\mathbf{K} = (\mathbf{N}^T)^{-1}\mathbf{\Omega}_N\mathbf{N}^{-1}$.

5.13

You have fitted a smoothing spline $\hat{f}_\lambda$ to a sample of N pairs (x_i, y_i). Suppose you augment your original sample with the pair $x_0, \hat{f}_\lambda(x_0)$, and refit; describe the result. Use this to derive the N-fold cross-validation formula (5.26).

Let $\hat{f}_\lambda^{(-i)}(x_i)$ denote the predicted value for the $i-$th case when $\{x_i, y_i\}$ is left out of the data doing the fitting. We claim that

$$\hat{f}_\lambda^{(-i)}(x_i) = \frac{1}{1 - S_\lambda(i,i)} \sum_{j \neq i} S_\lambda(i,j) y_j. \quad (1)$$

Starting from (1), we multiply $(1 - S_\lambda(i,i))$ on both sides and move one term from left side to right side, we have

$$\hat{f}_\lambda^{(-i)}(x_i) = \sum_{j \neq i} S_\lambda(i,j) y_j + S_\lambda(i,i)\hat{f}_\lambda^{(-i)}(x_i).$$

Recall that

$$\hat{f}_\lambda(x_i) = \sum_{j=1}^{n} S_\lambda(i,j) y_j,$$

we have

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

thus

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

It remains to prove (1). Intuitively, any reasonable smoother is constant preserving, which means $S_\lambda\mathbf{1} = \mathbf{1}$. Therefore, the rows of S_λ sum to one. Thus if we want to use the same smoother with the i-th row and column deleted, we must re-normalize the rows to sum to one, that gives (1). For a rigorous proof, please see Ex. 7.3 (a).

5.15

This exercise derives some of the results quoted in Section 5.8.1. Suppose $K(x, y)$ satisfying the conditions (5.45) and let $f(x) \in \mathcal{H}_K$. Show that

(a) $\langle K(\cdot, x_i), f \rangle_{\mathcal{H}_K} = f(x_i)$.

(b) $\langle K(\cdot, x_i), K(\cdot, x_j) \rangle_{\mathcal{H}_K} = K(x_i, x_j)$.

(c) If $g(x) = \sum_{i=1}^{N} \alpha_i K(x, x_i)$, then

$$J(g) = \sum_{i=1}^{N} \sum_{i=1}^{N} K(x_i, x_j) \alpha_i \alpha_j.$$

Suppose that $\tilde{g}(x) = g(x) + \rho(x)$, with $\rho(x) \in \mathcal{H}_K$, and orthogonal in $\mathcal{H}_K$ to each of $K(x, x_i)$, i=1,...,N. Show that

(d)

$$\sum_{i=1}^{N} L(y_i, \tilde{g}(x_i)) + \lambda J(\tilde{g}) \geq \sum_{i=1}^{N} L(y_i, g(x_i)) + \lambda J(g)$$

with equality iff $\rho(x) = 0$.

(a) Note that by (5.47) in text, the inner product $\mathcal{H}_K$ is

$$\left\langle \sum_{j\in J} a_j\phi_i(x), \sum_{j\in J} b_j\phi_j(x) \right\rangle_{\mathcal{H}_K} = \sum_{j\in J} \frac{a_j b_j}{\lambda_j}.$$

Therefore, by definition of K we have

$$\begin{aligned} \langle K(\cdot,y), f\rangle_{\mathcal{H}_K} &= \left\langle \sum_{i=1}^{\infty}(\gamma_i\phi_i(x))\phi_i(y), \sum_{i=1}^{\infty} c_i\phi_i(x) \right\rangle \\ &= \sum_{i=1}^{\infty} \frac{c_i\lambda_i\phi_i(y)}{\lambda_i} \\ &= f(y). \end{aligned}$$

(b) It follows from (a) by letting $f(\cdot) = K(\cdot, x_j)$.

(c) From (b) we have

$$\begin{aligned} J(g) &= \left\langle \sum_{i=1}^{N} \alpha_i K(x,x_i), \sum_{i=1}^{N} \alpha_i K(x,x_i) \right\rangle \\ &= \sum_{i=1}^{N}\sum_{i=1}^{N} K(x_i,x_j)\alpha_i\alpha_j. \end{aligned}$$

(d) Since ρ is orthogonal to each $K(x,x_i)$ for $i = 1,\ldots,N$, we have

$$\lambda J(\tilde{g}) = \lambda J(g) + \lambda\|\rho\|^2_{\mathcal{H}_K} \geq \lambda J(g).$$

Moreover, from (a), we have

$$\begin{aligned} \tilde{g}(x_i) &= \langle K(\cdot,x_i), \tilde{g}\rangle_{\mathcal{H}_K} \\ &= \langle K(\cdot,x_i), g+\rho\rangle_{\mathcal{H}_K} \\ &= \langle K(\cdot,x_i), g\rangle_{\mathcal{H}_K}, \end{aligned}$$

so that

$$L(y_i, \tilde{g}(x_i)) = L(y_i, g(x_i)),$$

that is, the loss only depends on the data space.

The proof is now complete.

5.16

Consider the ridge regression problem (5.53), and assume $M \geq N$. Assume you have a kernel K that computes the inner product $K(x, y) = \sum_{m=1}^{M} h_m(x) h_m(y)$.

(a)

Derive (5.62) on page 171 in the text. How would you compute the matrices $\mathbf{V}$ and $\mathbf{D}_\gamma$, given K? Hence show that (5.63) is equivalent to (5.53).

(b)

Show that

$$\begin{aligned}\hat{\mathbf{f}} &= \mathbf{H}\hat{\beta} \\ &= \mathbf{K}(\mathbf{K} + \lambda \mathbf{I})^{-1}\mathbf{y},\end{aligned}$$

where $\mathbf{H}$ is the $N \times M$ matrix of evaluations $h_m(x_i)$, and $\mathbf{K} = \mathbf{H}\mathbf{H}^T$ the $N \times N$ matrix of inner-product $h(x_i)^T h(x_j)$.

(c)

Show that

$$\begin{aligned}\hat{f}(x) &= h(x)^T \hat{\beta} \\ &= \sum_{i=1}^{N} K(x, x_i)\hat{\alpha}_i\end{aligned}$$

and $\hat{\alpha} = (\mathbf{K} + \lambda \mathbf{I})^{-1}\mathbf{y}$.

(d)

How would you modify your solution if $M < N$?

(a)

By definition of the kernel K, we have

$$K(x,y) = \sum_{m=1}^{M} h_m(x)h_m(y) = \sum_{i=1}^{\infty} \gamma_i \phi_i(x)\phi_i(y).$$

Multiply each summand above by $\phi_k(x)$ and calculate $\langle K(x,y), \phi_k(x) \rangle$,

$$\sum_{m=1}^{M} \langle h_m(x), \phi_k(x) \rangle h_m(y) = \sum_{i=1}^{\infty} \langle \phi_i(x), \phi_k(x) \rangle \phi_i(y). \quad (1)$$

Since $\{\phi_i, i = 1, \ldots, \infty\}$ are orthogonal, we have

$$\langle \phi_i(x), \phi_k(x) \rangle = \begin{cases} 1 & \text{if } i = k, \\ 0 & \text{otherwise.} \end{cases}$$

Thus, (1) becomes

$$\sum_{m=1}^{M} \langle h_m(x), \phi_k(x) \rangle h_m(y) = \gamma_k \phi_k(y).$$

Let $g_{km} = \langle h_m(x), \phi_k(x) \rangle$ and calculate $\langle K(x,y), \phi_l(y) \rangle$, we get

$$\begin{aligned} \sum_{m=1}^{M} g_{km} h_m(y) &= \gamma_k \phi_k(y), \\ \sum_{m=1}^{M} g_{km} \langle h_m(y), \phi_l(y) \rangle &= \gamma_k \langle \phi_k(y), \phi_l(y) \rangle, \\ \sum_{m=1}^{M} g_{km} g_{lm} &= \gamma_k \delta_{k,l} \end{aligned}$$

where $\delta_{k,l} = 1$ if $k = l$ and 0 otherwise.

Let $\mathbf{G}_M = \{g_{nm}\} \in \mathbb{R}^{M\times N}$, we have

$$\mathbf{G}_M\mathbf{G}_M^T = \mathrm{diag}\{\gamma_1, \gamma_2, \ldots, \gamma_M\} = \mathbf{D}_\gamma.$$

Let $\mathbf{V}^T = \mathbf{D}_\gamma^{-\frac{1}{2}}\mathbf{G}_M$, we have

$$\mathbf{V}\mathbf{V}^T\mathbf{G}_M^T = \mathbf{G}_M^T\mathbf{D}_\gamma^{-1}\mathbf{G}_M = \mathbf{I}_N.$$

Let $h(x) = (h_1(x), h_2(x), \ldots, h_M(x))^T$ and $\phi(x) = (\phi_1(x), \phi_2(x), \ldots, \phi_M(x))^T$, then the three equations above can be rewritten as

$$\begin{aligned}
\mathbf{G}_M h(x) &= \mathbf{D}_\gamma\phi(x)\\
\mathbf{V}\mathbf{D}_\gamma^{-\frac{1}{2}}\mathbf{G}_M h(x) &= \mathbf{V}\mathbf{D}_\gamma^{-\frac{1}{2}}\mathbf{D}_\gamma\phi(x)\\
h(x) &= \mathbf{V}\mathbf{D}_\gamma^{\frac{1}{2}}.
\end{aligned}$$

To show that (5.63) is equivalent to (5.53) in the text, we start with (5.63). Let $\beta = (\beta_1, \beta_2, \ldots, \beta_m)^T$ and $c = \mathbf{D}_\gamma^{\frac{1}{2}}\mathbf{V}^T\beta$,

$$\begin{aligned}
&\min_{\{\beta_m\}_1^M}\sum_{i=1}^N\left(y_i - \sum_{m=1}^M\beta_m h_m(x_i)\right)^2 + \lambda\sum_{m=1}^M\beta_m^2\\
=&\min_\beta\sum_{i=1}^N(y_i - \beta^T h(x_i))^2 + \lambda\beta^T\beta\\
=&\min_\beta\sum_{i=1}^N(y_i - \beta^T\mathbf{V}\mathbf{D}_\gamma^{\frac{1}{2}}\phi(x_i))^2 + \lambda\beta^T\beta\\
=&\min_c\sum_{i=1}^N(y_i - c^T\phi(x_i))^2 + \lambda(\mathbf{V}\mathbf{D}_\gamma^{\frac{1}{2}}c)^T\mathbf{V}\mathbf{D}_\gamma^{\frac{1}{2}}c\\
=&\min_c\sum_{i=1}^N(y_i - c^T\phi(x_i))^2 + \lambda c^T c\mathbf{D}_\gamma^{-1}\\
=&\min_{\{c_j\}_1^\infty}\sum_{i=1}^N\left(y_i - \sum_{j=1}^\infty c_j\phi_j(x_i)\right)^2 + \lambda\sum_{j=1}^\infty\frac{c_j^2}{\gamma_j},
\end{aligned}$$

which is (5.53) in the text.

（b）

Recall that in (a) we have

$$\min_{\beta} \sum_{i=1}^{N} (y_i - \beta^T h(x_i))^2 + \lambda \beta^T \beta.$$

Taking derivative w.r.t β and setting it to be zero yields

$$-\mathbf{H}^T(\mathbf{y} - \mathbf{H}\hat{\beta}) + \lambda\hat{\beta} = 0.$$

Thus we have

$$\hat{\beta} = (\mathbf{H}^T\mathbf{H} + \lambda\mathbf{I})^{-1}\mathbf{H}^T\mathbf{y}$$

and

$$\hat{\mathbf{f}} = \mathbf{H}(\mathbf{H}^T\mathbf{H} + \lambda\mathbf{I})^{-1}\mathbf{H}^T\mathbf{y}.$$

By Woodbury matrix identity, we have

$$(\mathbf{H}^T\mathbf{H} + \lambda\mathbf{I})^{-1} = \frac{1}{\lambda}\mathbf{I} - \frac{1}{\lambda}\mathbf{I}\mathbf{H}^T\left(\mathbf{I} + \frac{1}{\lambda}\mathbf{H}\mathbf{H}^T\right)^{-1}\mathbf{H}\cdot\frac{1}{\lambda}\mathbf{I}.$$

Therefore, we have

$$\begin{aligned}
\hat{\mathbf{f}} &= \frac{1}{\lambda}\mathbf{H}\mathbf{H}^T\mathbf{y} - \frac{1}{\lambda}\mathbf{H}\mathbf{H}^T\left(\lambda\mathbf{I} + \mathbf{H}\mathbf{H}^T\right)^{-1}\mathbf{H}\mathbf{H}^T\mathbf{y} \\
&= \frac{1}{\lambda}\mathbf{H}\mathbf{H}^T\left[\mathbf{I} - (\lambda\mathbf{I} + \mathbf{H}\mathbf{H}^T)^{-1}\mathbf{H}\mathbf{H}^T\right]\mathbf{y} \\
&= \frac{1}{\lambda}\mathbf{H}\mathbf{H}^T\left[(\lambda\mathbf{I} + \mathbf{H}\mathbf{H}^T)^{-1}(\lambda\mathbf{I} + \mathbf{H}\mathbf{H}^T) - (\lambda\mathbf{I} + \mathbf{H}\mathbf{H}^T)^{-1}\mathbf{H}\mathbf{H}^T\right]\mathbf{y} \\
&= \frac{1}{\lambda}\mathbf{H}\mathbf{H}^T\left[(\lambda\mathbf{I} + \mathbf{H}\mathbf{H}^T)^{-1}\lambda\mathbf{I}\right]\mathbf{y} \\
&= \mathbf{H}\mathbf{H}^T(\lambda\mathbf{I} + \mathbf{H}\mathbf{H}^T)^{-1}\mathbf{y} \\
&= \mathbf{K}(\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{y}.
\end{aligned}$$

（c）

This is directly derived from (b).

（d）

The solution remains the same as $\mathbf{K} + \lambda\mathbf{I}$ is invertible as long as $\lambda \neq 0$. When $\lambda = 0$ however, we have

$$\hat{\mathbf{f}} = \mathbf{H}\hat{\beta} = \mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{y} = \mathbf{y}.$$