Splines

- **knots**: a sequence of increasing numbers $\xi_1 < \xi_2 < \cdots < \xi_m$ on some interval $[a, b]$.

- A function $g$ defined on $[a, b]$ is a cubic spline w.r.t knots $\{\xi_i\}_{i=1}^m$ if the following two conditions are satisfied: (1) $g$ is a cubic polynomial on each of the $m+1$ intervals, that is,
  \[
g(x) = d_i x^3 + c_i x^2 + b_i x + a_i, \quad x \in [\xi_i, \xi_{i+1}]
  \]
  where $i = 0, 1, \ldots, m$, $\xi_0 = a$ and $\xi_{m+1} = b$; (2) $g$ is continuous up to the 2nd derivative, that is,
  \[
g^{(0,1,2)}(\xi^+_i) = g^{(0,1,2)}(\xi^-_i), \quad i = 1, \ldots, m.
  \]

All the cubic splines (w.r.t knots $\{\xi_i\}_{i=1}^m$) form a linear space of functions with $(m+4)$ degree of freedom. Below is a set of basis functions for that space:

\[
\begin{align*}
    h_1(x) &= 1; \quad h_2(x) = x; \quad h_3(x) = x^2; \quad h_4(x) = x^3; \\
    h_i+4(x) &= (x - \xi_i)^3, \quad i = 1, 2, \ldots, m.
\end{align*}
\]

- A cubic spline on $[a, b]$ is said to be a natural cubic spline (NCS) if its second and third derivatives are zero at $a$ and $b$, that is, it is linear on the two extreme intervals $[a, \xi_1]$ and $[\xi_m, b]$. Because of the four additional constraints $d_0 = c_0 = d_m = c_m = 0$, the degree of freedom of NCS’s with $m$ knots is $m$. One version of the basis functions can be found in the textbook (section 5.2.1).

Smoothing Splines

- **Roughness Penalty Approach** – Let $S[a, b]$ be the space of all “smooth” functions $g$ on $[a, b]$ that have two continuous derivatives. Among all the functions in $S[a, b]$, we are looking for the minimizer of the following penalized residual sum of squares
  \[
  \text{RSS}(g, \lambda) = \sum_{i=1}^n [y_i - g(x_i)]^2 + \lambda \int_a^b [g''(x)]^2 dx, \tag{1}
  \]
  where the first term quantifies the goodness-of-fit to the data, the second term measures the roughness, and $\lambda$ is a positive smoothing parameter.
parameter. The criterion (1) is defined on an infinite-dimensional space $S[a, b]$ (in fact, a Sobolev space). However, remarkably, it can be shown (by the following theorem) that when $n > 1$ the minimizer $\hat{g}$ lies in a finite dimensional space, the space of natural cubic splines with knots at the $n$ data points $x_1, \ldots, x_n$.

**THEOREM 1**: Let $g$ be any differentiable function on $[a, b]$ for which $g(x_i) = z_i$ for $i = 1, \ldots, n$. Suppose $n \geq 2$, and that $\tilde{g}$ is the natural cubic spline interpolant to the values $z_1, \ldots, z_n$ at points $x_1, \ldots, x_n$ with $a < x_1 < \cdots < x_n < b$. Then $\int g''^2 \geq \int \tilde{g}''^2$ with equality only if $\tilde{g} \equiv g$.

**PROOF**: Let $h(x) = g(x) - \tilde{g}(x)$. So $h(x_i) = 0$ for $i = 1, \ldots, n$. Want to show that
\[
\int \tilde{g}''(x)^2 dx \leq \int g''(x)^2 dx.
\]
It is because
\[
(1) \int g''^2 = \int (\tilde{g}'' + h'')^2
\]
\[
= \int \tilde{g}''^2 + 2 \int \tilde{g}'' h'' + \int h''^2
\]
\[
(2) \int_a^b \tilde{g}'' h'' dx = \tilde{g}''(x)h'(x)|_a^b - \int_a^b h'(x)\tilde{g}^{(3)}(x)dx
\]
\[
= -\sum_{i=1}^{n-1} \tilde{g}^{(3)}(x_j^+) \int_{x_j}^{x_{j+1}} h'(x)dx
\]
\[
= -\sum_{i=1}^{n-1} \tilde{g}^{(3)}(x_j^+)(h(x_{j+1}) - h(x_j))
\]
\[
= 0.
\]

Note that the result above will still hold true even if we change the square loss, the first term in (1), to other loss functions.

- **Generalized Ridge Regression** – Since the minimizer is a NCS, we can write it as
\[
g(x) = \sum_{i=1}^n \theta_i N_i(x),
\]
where \( N_i(x) \)'s are basis functions for NCS with knots at \( x_1, \ldots, x_n \). The criterion (1) can be written as
\[
\text{RSS}(\theta, \lambda) = (y - N\theta)^T(y - N\theta) + \lambda \theta^T \Omega \theta,
\]
where \( N \) and \( \Omega \) are \( n \times n \) matrices with \( (N)_{ij} = N_j(x_i) \) and \( (\Omega)_{ij} = \int N_i(x)^\prime N_j(x)^\prime \prime dx \). The solution is
\[
\hat{\theta} = \text{argmin} \text{RSS}(\theta, \lambda) = (N^T N + \lambda \Omega)^{-1} N^T y
\]
\[
\hat{y} = N\hat{\theta} = N(N^T N + \lambda \Omega)^{-1} N^T y = S_\lambda y.
\]
We will refer to \( S_\lambda \) as smoother matrix.

Demmler & Reinsch (1975) constructed a basis with the so-called double orthogonality property, i.e.
\[
N^T N = I, \quad \Omega = \text{diag}(d_i),
\]
where \( d_i \)'s are a non-negative increasing sequence and \( d_1 = d_2 = 0 \) (Why?). Using this basis, we can write \( \hat{y} \) as
\[
\hat{y} = N \text{diag}\left( \frac{1}{1 + \lambda d_i} \right) N^T y = \sum_{i=1}^{n} \frac{1}{1 + \lambda d_i} (u_i^T y) u_i
\]
where \( u_i \) is the \( i \)-th column of \( N \). Now we can easily see the connection between the smoothing spline and the ridge regression.

Similar to the ridge regression, we define the effective degree of freedom of a smoothing spline to be
\[
df(\lambda) = \text{trace} S_\lambda = \text{trace} N^T N(N^T N + \lambda \Omega)^{-1} = \sum_{i=1}^{n} \frac{1}{1 + \lambda d_i}.
\]

• How to Choose the Smoothing Parameter – Let us pretend that we do not observe the \( i \)-th observation and instead use the remaining \( n - 1 \) observations to fit a smoothing spline (w.r.t a smoothing parameter \( \lambda \)). Denote the estimated curve by \( \hat{g}^{(-i)} \) (in contrast to \( \hat{g} \), the smoothing spline calculated from the full data). By definition, \( \hat{g}^{(-i)}(x) \) minimizes
\[
\sum_{j \neq i} [(y_i - g(x_i))^2 + \lambda \int_a^b [g''(x)]^2 dx].
\]
The quality of $\hat{g}^{(-i)}$ as a predictor on a new observation can be judged by how well the value $\hat{g}^{(-i)}(x_i)$ predicts $y_i$. The idea of cross-validation is to choose $\lambda$ that minimizes the following score function

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

(4)

At first sight, it seems necessary to solve $n$ separate smoothing problems in order to calculate $CV(\lambda)$. However, we will show that the “deleted residual” $y_i - \hat{g}^{(-i)}(x_i)$ can be expressed in terms of $y_i - \hat{g}(x_i)$ and the $i$th diagonal element of the smoother matrix.

**THEOREM 2**: The cross-validation score (4) satisfies

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

(5)

where $S_\lambda(i, j)$ denotes the $(i, j)$-th entry of the smoother matrix $S_\lambda$.

**PROOF**: We will show that the following equality holds true,

$$
y_i - \hat{g}^{(-i)}(x_i) = \left( \frac{y_i - \hat{g}(x_i)}{1 - S_\lambda(i, i)} \right), \quad i = 1, \ldots, n.
$$

(6)

It suffices to show that it holds true when $i = 1$.

Define a vector $y^*$: $y^*_j = y_j$ for $j = 2, 3, \ldots, n$ and $y^*_1 = \hat{g}^{(-1)}(x_1)$. We claim that $\hat{g}^{(-1)}$ is the estimated smoothing spline for the $n$ pairs of new observations $(x_i, y^*_i)_{i=1}^n$, by showing that $\hat{g}^{(-1)}$ minimizes

$$
\sum_{i=1}^{n} [y^*_i - g(x_i)]^2 + \lambda \int [g''(x)]^2 dx.
$$

The summation above is bigger than or equal to

$$
\sum_{i=2}^{n} [y^*_i - g(x_i)]^2 + \lambda \int [g''(x)]^2 dx,
$$

$$
\geq \sum_{i=2}^{n} [y^*_i - \hat{g}^{(-1)}(x_i)]^2 + \lambda \int [\hat{g}^{(-1)''}(x)]^2 dx
$$

$$
= \sum_{i=1}^{n} [y^*_i - \hat{g}^{(-1)}(x_i)]^2 + \lambda \int [\hat{g}^{(-1)''}(x)]^2 dx.
$$
Therefore

\[
y_1^* = \sum_{j=1}^{n} S_{\lambda}(1, j)y_j^*
\]

\[
= S_{\lambda}(1, 1)y_1^* + \sum_{j=2}^{n} S_{\lambda}(1, j)y_j
\]

\[
= S_{\lambda}(1, 1)(y_1^* - y_1) + \sum_{j=1}^{n} S_{\lambda}(1, j)y_j
\]

\[
= S_{\lambda}(1, 1)(y_1^* - y_1) + \hat{g}(x_1).
\]

Substituting \(y_1^* = \hat{g}(-1)(x_1)\), we obtain (6).

**Generalized cross-validation** is a modified form of cross-validation. The basic idea is to replace the individual factors \((1 - S_{\lambda}(i, i))\) by their average value, i.e.,

\[
\text{GCV}(\lambda) = \frac{1}{n} \frac{1}{(1 - n^{-1} \text{trace} S_{\lambda})^2} \sum_{i=1}^{n} \left( y_i - \hat{g}(x_i) \right)^2.
\]