Optimality of smoothing splines

Let us prove, in the first place, the following proposition.

Proposition. A set of $n \geq 2$ distinct points (x_i, y_i) can be interpolated using a natural cubic spline with the data points $x_1 < \dots < x_m$ as knots. The interpolating natural cubic spline is unique.

The result stated in this proposition means that there exist a unique $f(x)$ such that

$$y_i = f(x_i), \qquad i = 1, \dots, m,$$

among all possible cubic splines. Heuristically, the reason behind this proposition is due to the fact that

$$f(x) = \sum_{j=1}^{m} N_j(x)\beta_j, \quad \text{using } x_1, \dots, x_m \text{ as knots.}$$

There are m "free parameters" and "m data points", thus we can choose β such that

$$N\beta = y \implies \hat{\beta} = N^{-1}y.$$ It is now invertible

Note that $\hat{\beta} = N^{-1}y = \underbrace{(N^TN)^{-1}N^T}_{\text{Pseudo-inverse Moore-Penrose}} y =$ "Least square estimate".

The main result about the optimality of splines is provided below.

Theorem. Suppose $n \geq 2$ and that f is the natural cubic spline <u>interpolant</u> to the values (x_i, y_i) satisfying $a < x_1 < \dots < x_m < b$. Let $\tilde{f}$ be any function in $S_2[a,b]$ for which $\tilde{f}(x_i) = y_i$, for $i = 1, \dots, m$. The

$$\int_a^b \tilde{f}''(t)^2 dt \geq \int_a^b f''(t)^2 dt.$$

Proof. Let us define $h(t) = \tilde{f}(t) - f(t)$, which is also in $S_2[a,b]$. Then

$$\int_a^b \tilde{f}''(t)^2 dt = \int_a^b [f''(t) + h''(t)]^2 dt = \int_a^b f''(t)^2 dt + \underbrace{\int_a^b h''(t)^2 dt}_{\geq 0} + \cancel{2\int_a^b f''(t) h''(t) dt}_{=0}$$

$$\geq \int_a^b f''(t)^2 dt$$

In fact:

$$\int_a^b f''(t)h''(t)dt \underset{\text{Integration by parts}}{=} f''(t)h'(t)\Big|_a^b - \int_a^b \underbrace{f'''(t)}_{\text{it exists, because of the properties of natural cubic splines.}} h'(t)dt$$

$$= \underbrace{\cancel{f''(b)}}_{=0} h'(b) - \underbrace{\cancel{f''(a)}}_{=0} h'(a) - \int_a^b f'''(t)h'(t)dt$$

Because of the continuity constraints $f''(a) = f''(b)$

$$= -\int_a^b f'''(t)h'(t) = -\sum_{j=0}^{m} \int_{x_j}^{x_{j+1}} \underbrace{f'''(t)}_{\text{This is a constant!}} h'(t)dt$$

$$= -\sum_{j=1}^{m-1} f'''(x_j^+) \int_{x_j}^{x_{j+1}} \underbrace{h'(t)}_{\text{integrable!}} dt$$

having set $x_0 = a$, $x_{m+1} = b$. Summarizing, so far we proved that

$$\int_a^b f''(t)h''(t)dt = -\sum_{j=1}^{m-1} f'''(x_j^+)[h(x_{j+1}) - h(x_j)]$$

Note in addition that $h(x_i) = \tilde{f}(x_i) - f(x_i) = y_i - y_i = 0$. Hence

$$\int_a^b f''(t)h''(t)dt = -\sum_{j=1}^{m-1} f'''(x_j^+)[\underbrace{h(x_{j+1})}_{=0} - \underbrace{h(x_j)}_{=0}] = 0.$$

In principle we can have $\int_a^b h''(t)^2 dt = 0$, which means $h(t) = a + bt$, i.e. is a linear function. However, we also need $h(x_1) = \dots = h(x_m) = 0 \Rightarrow a, b = 0$.

Hence $\int_a^b h''(t)^2 dt = 0 \iff \tilde{f}(t) = f(t) \ \forall t \in [a,b]$.

From these results, we can finally conclude the following final Theorem.

Theorem Let $m \geq 2$ and consider a collection of distinct values $x_1, \dots, x_m$, associated with the responses $y_1, \dots, y_m$. Then there exist a unique minimizer

$$\hat{f}(x) = \underset{f \in S_2(a,b)}{\arg\min} \mathcal{L}(f) = \underset{f \in S_2(a,b)}{\arg\min} \sum_{i=1}^m \{y_i - f(x)\}^2 + \lambda \int_a^b f''(t)^2 dt,$$

The minimizer $\hat{f}(x)$ is a natural cubic spline with knots at $x_1, \dots, x_m$.

Proof. The theorem is just a consequence of the previous results. Consider a generic function $\tilde{f} \in S_2(a,b)$, whose predictions are $\tilde{f}(x_i)$. It is always possible to find a natural cubic spline with knots at $x_1, \dots, x_m$ called $f(x_i)$ such that

$$z_i = \tilde{f}(x_i) = f(x_i), \quad i = 1, \dots, m \quad \text{(Property of interpolation)},$$

which implies that

$$\sum_{i=1}^m \{y_i - \tilde{f}(x_i)\}^2 = \sum_{i=1}^m \{y_i - f(x_i)\}^2.$$

However, because of the previous Theorem (using (x_i, z_i) as data points)

$$\int_a^b \tilde{f}''(t)^2 dt \geq \int_a^b f''(t)^2 dt. \qquad \blacksquare$$

which concludes the proof, because $\mathcal{L}(\tilde{f}) \geq \mathcal{L}(\hat{f})$.

Basically the same proof of ridge regression.

From here, it follows that $\hat{f}(x) = N\hat{\beta}$ and

$$\hat{\beta} = \underset{\beta}{\arg\min} \sum_{i=1}^m \{y_i - N\beta\}^2 + \lambda \beta^T \Omega \beta \Rightarrow \hat{\beta} = (N^T N + \lambda \Omega)^{-1} N^T y.$$