Sobolev Take 2

Given a function h, the Sobolev penalty for h is

$$P(h) = \int (h^{(r)}(x))^2 dx$$

Consider the class of smoothing splines

$$\left\{ \hat{g}(\cdot|\lambda) = \arg\min_{g \in \mathcal{G}} \frac{1}{2}\|y - g\|_T^2 + \lambda P(g) : \lambda \in \Lambda \right\}$$

Ever function $\hat{g}(\cdot|\lambda)$ is a spline that can be expressed as the weighted sum of normalized B-splines of degree $r+1$ for a given set of k knots:

$$\hat{g}(x|\lambda) = \sum_{i=1}^{k} \theta_i N_{i,r+1}(x)$$

Note that the normalized B-splines have the property that they sum up to one at all points within the boundary of the knots. Also recall that B-splines are non-negative. (I might be wrong regarding the relationship between the number of knots and the number of basis functions, but the number of basis functions should be linear in the number of knots)

Therefore we can re-express the class of smoothing splines as a set of function parameters

$$\left\{ \hat{\theta}_\lambda = \arg\min_{\theta} \frac{1}{2}\|y - N_T\theta\|^2 + \lambda P(\theta) : \lambda \in \Lambda \right\}$$

where N_T is the normalized B-spline basis for the given set of knots evaluated at the points in the training set. $P(\theta)$ is the Sobolev penalty and can be written as $\theta^T \Omega \theta$ for an appropriate penalty matrix Ω. We will not need to express anything in terms of Ω so the penalty will be just written as $P(\theta)$.

Instead of considering the original smoothing spline problem with the roughness penalty, we will add a ridge penalty on the function parameters

$$\left\{ \hat{\theta}_\lambda = \arg\min_{\theta} \frac{1}{2}\|y - N_T\theta\|^2 + \lambda \left(P(\theta) + \frac{w}{2}\|\theta\|_2^2 \right) : \lambda \in \Lambda \right\}$$

Univariate "Sobolev" norm

Suppose we have the set of function parameters

$$\left\{ \hat{\theta}_\lambda = \arg\min_{\theta} \frac{1}{2}\|y - N_T\theta\|^2 + \lambda \left(P(\theta) + \frac{w}{2}\|\theta\|_2^2 \right) : \lambda \in \Lambda \right\}$$

Let k be the number of normalized B-spline basis functions.

Suppose that we restrict the function parameters such that$\sup_{\theta \in \Theta} \|\theta\|_2 \leq G$. (**Is this a reasonable assumption?**)

For any $\lambda^{(1)}, \lambda^{(2)} \in \Lambda$ such that

$$\|\lambda^{(1)} - \lambda^{(2)}\| \leq \frac{d}{2}\lambda_{min} w \left(\left(\frac{1}{\lambda_{min}} (2G + \|\epsilon\|_T) + wG \right) + wG \right)^{-1}$$

we can show that

$$\|\hat{\theta}_{\lambda^{(1)}} - \hat{\theta}_{\lambda^{(2)}}\|_2 \leq d$$

Hence

$$\|\hat{g}(\cdot|\lambda^{(1)}) - \hat{g}(\cdot|\lambda^{(2)})\|_\infty = \sup_x \sum_{i=1}^{k} \left(\hat{\theta}_{i,\lambda^{(1)}} - \hat{\theta}_{i,\lambda^{(2)}} \right) N_i(x) \leq kd$$

Proof

Let $\beta = \hat{\theta}_{\lambda^{(1)}} - \hat{\theta}_{\lambda^{(2)}}$. Suppose $\|\beta\|_2 > d$ for contradiction.

Consider the optimization problem

$$\hat{m}_\beta(\lambda) = \arg\min_m \frac{1}{2}\|y - N_T(\hat{\theta}_{\lambda^{(1)}} + m\beta)\|^2 + \lambda\left[P\left(\hat{\theta}_{\lambda^{(1)}} + m\beta\right) + \frac{w}{2}\|\hat{\theta}_{\lambda^{(1)}} + m\beta\|_2^2\right]$$

The KKT conditions give us

$$-\langle N_T\beta, y - N_T(\hat{\theta}_{\lambda^{(1)}} + \hat{m}_\beta(\lambda)\beta)\rangle + \lambda\left[\nabla_m P\left(\hat{\theta}_{\lambda^{(1)}} + m\beta\right)\Big|_{m=\hat{m}_\beta(\lambda)} + w\langle\beta, \hat{\theta}_{\lambda^{(1)}} + \hat{m}_\beta(\lambda)\beta\rangle\right] = 0$$

Implicit differentiation wrt λ gives us

$$\begin{aligned}\frac{\partial}{\partial\lambda}\hat{m}_\beta(\lambda) &= -\left(\|N_T\beta\|^2 + \lambda\left[\nabla_m^2 P\left(\hat{\theta}_{\lambda^{(1)}} + m\beta\right)\Big|_{m=\hat{m}_\beta(\lambda)} + w\|\beta\|_2^2\right]\right)^{-1} \\ &\quad \left(\nabla_m P\left(\hat{\theta}_{\lambda^{(1)}} + m\beta\right)\Big|_{m=\hat{m}_\beta(\lambda)} + w\langle\beta, \hat{\theta}_{\lambda^{(1)}} + \hat{m}_\beta(\lambda)\beta\rangle\right)\end{aligned}$$

To bound the first multiplicand, we use the convexity of the Sobolev penalty to get

$$\left(\|N_T\beta\|^2 + \lambda\left[\nabla_m^2 P\left(\hat{\theta}_{\lambda^{(1)}} + m\beta\right)\Big|_{m=\hat{m}_\beta(\lambda)} + w\|\beta\|_2^2\right]\right)^{-1} \leq \left(\lambda w\|\beta\|_2^2\right)^{-1}$$

To bound the second multiplicand, the KKT conditions give us

$$\begin{aligned}\left\|\nabla_m P\left(\hat{\theta}_{\lambda^{(1)}} + m\beta\right)\Big|_{m=\hat{m}_\beta(\lambda)}\right\| &= \left\|\frac{1}{\lambda}\langle N_T\beta, y - N_T(\hat{\theta}_{\lambda^{(1)}} + \hat{m}_\beta(\lambda)\beta)\rangle - w\langle\beta, \hat{\theta}_{\lambda^{(1)}} + \hat{m}_\beta(\lambda)\beta\rangle\right\| \\ &\leq \|\beta\|\left(\frac{1}{\lambda}\left\|N_T^T\left(y - N_T(\hat{\theta}_{\lambda^{(1)}} + \hat{m}_\beta(\lambda)\beta)\right)\right\| + w\left\|\hat{\theta}_{\lambda^{(1)}} + \hat{m}_\beta(\lambda)\beta\right\|\right)\end{aligned}$$

We can bound the maximum eigenvalue of N_T^T using the matrix entries. Since each row of N_T^T is the value of each of the normalized B-spline basis functions at a given point, each row of N_T^T sums to one. By the Gershgorin circle theorem (*), the maximum eigenvalue of a non-negative matrix is bounded by its maximum row sum, which in the case of N_T^T is 1. Hence

$$\begin{aligned}\left\|\nabla_m P\left(\hat{\theta}_{\lambda^{(1)}} + m\beta\right)\Big|_{m=\hat{m}_\beta(\lambda)}\right\| &\leq \|\beta\|\left(\frac{1}{\lambda}\left\|y - N_T(\hat{\theta}_{\lambda^{(1)}} + \hat{m}_\beta(\lambda)\beta)\right\| + w\left\|\hat{\theta}_{\lambda^{(1)}} + \hat{m}_\beta(\lambda)\beta\right\|\right) \\ &\leq \|\beta\|\left(\frac{1}{\lambda_{min}}\left(2kG + \|\epsilon\|_T\right) + wG\right)\end{aligned}$$

The second inequality follows from the fact that for any function parameters θ , we have that

$$\|N_T\theta\| \leq k\|\theta\| \leq kG$$

since the maximum eigenvalue of N_T is bounded by the maximum row sum of N_T, which is at most k.

Then

$$\begin{aligned}\left\|\frac{\partial}{\partial\lambda}\hat{m}_\beta(\lambda)\right\| &\leq \left(\lambda w\|\beta\|_2^2\right)^{-1}\left(\|\beta\|\left(\frac{1}{\lambda_{min}}\left(2kG + \|\epsilon\|_T\right) + wG\right) + w\|\beta\|\|\hat{\theta}_{\lambda^{(1)}} + \hat{m}_\beta(\lambda)\beta\|\right) \\ &\leq \left(\lambda_{min} w\|\beta\|_2\right)^{-1}\left(\left(\frac{1}{\lambda_{min}}\left(2kG + \|\epsilon\|_T\right) + wG\right) + wG\right)\end{aligned}$$

By the MVT, there is some $\alpha \in (0,1)$ such that

$$\begin{aligned}
\left|\hat{m}_\beta(\lambda^{(2)}) - \hat{m}_\beta(\lambda^{(1)})\right| &= \left|\lambda^{(2)} - \lambda^{(1)}\right| \left(\frac{\partial}{\partial \lambda} \hat{m}_\beta(\lambda) \Big|_{\lambda = \alpha\lambda^{(1)} + (1-\alpha)\lambda^{(2)}} \right) \\
&\leq \left|\lambda^{(2)} - \lambda^{(1)}\right| (\lambda_{min} w \|\beta\|_2)^{-1} \left(\left(\frac{1}{\lambda_{min}} (2kG + \|\epsilon\|_T) + wG \right) + wG \right) \\
&\leq \left|\lambda^{(2)} - \lambda^{(1)}\right| (\lambda_{min} w d)^{-1} \left(\left(\frac{1}{\lambda_{min}} (2kG + \|\epsilon\|_T) + wG \right) + wG \right)
\end{aligned}$$

By our choice of $\lambda^{(2)}, \lambda^{(1)}$, we have that

$$\left|\hat{m}_\beta(\lambda^{(2)}) - \hat{m}_\beta(\lambda^{(1)})\right| \leq 1/2$$

However this is a contradiction since we know that $\hat{m}_\beta(\lambda^{(2)}) = 1$ and $\hat{m}_\beta(\lambda^{(1)}) = 0$.

Footnotes

(*) https://en.wikipedia.org/wiki/Perron%E2%80%93Frobenius_theorem#Inequalities_for_Perron.E2.80.93Frobenius_e