More on Splines

Recall the basis

$$N_1(x) = 1, \quad N_2(x) = x$$

and

$$N_{2+l}(x) = \frac{(x-\xi_l)^3_+ - (x-\xi_K)^3_+}{\xi_K - \xi_l} - \frac{(x-\xi_{K-1})^3_+ - (x-\xi_K)^3_+}{\xi_K - \xi_{K-1}}$$

for $l = 1, \ldots, K-2$ for natural cubic splines. Observe that $N_1''(x) = N_2''(x) = 0$ and

$$N_{2+l}''(x) = \begin{cases} 6\frac{x-\xi_l}{\xi_K-\xi_l} & x \in (\xi_l, \xi_{K-1}] \\ 6\frac{(\xi_{K-1}-\xi_l)(\xi_K-x)}{(\xi_K-\xi_l)(\xi_K-\xi_{K-1})} & x \in (\xi_{K-1}, \xi_K) \\ 0 & x \leq \xi_l \text{ and } x \geq \xi_K \end{cases}$$

Assuming that $\xi_1 < \ldots < \xi_K$ the functions $N_3'', \ldots, N_K''$ are linearly independent.

Regularity of the Spline Smoother

If $x_1, \ldots, x_N$ are all different, $N_1, \ldots, N_N$ is the basis for the n.c.s. with knots $x_1, \ldots, x_N$ and $f = \sum_{i=1}^N \theta_i N_i$ we have

$$\theta^T \Omega_N \theta = \int_a^b (f''(x))^2 \mathrm{d}x = 0$$

if and only if $f''(x) = 0$ for all $x \in [a, b]$. Hence

$$\theta_3 = \ldots = \theta_N = 0.$$

If also $\theta^T \mathbf{N}^T \mathbf{N} \theta = 0$ then

$$(\theta_1 \ \theta_2) \begin{pmatrix} N & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} = 0,$$

which implies that $\theta_1 = \theta_2 = 0$ if $N \geq 2$. The in general positive semidefinite matrix

$$\mathbf{N}^T \mathbf{N} + \lambda \Omega_N$$

is thus positive definite for $\lambda > 0$.

The Reinsch Form

Let

$$\mathbf{S}_\lambda = \mathbf{N}(\mathbf{N}^T\mathbf{N} + \lambda\Omega_N)^{-1}\mathbf{N}^T$$

be the spline smoother and $\mathbf{N} = UDV^T$ the singular value decomposition of $\mathbf{N}$. Since $\mathbf{N}$ is square $N \times N$, U is orthogonal hence invertible with $U^{-1} = U^T$, and D is invertible since $\mathbf{N}$ has full rank N. Then

$$\begin{aligned}
\mathbf{S}_\lambda &= UDV^T(VD^2V^T + \lambda\Omega_N)^{-1}VDU^T \\
&= U(D^{-1}V^TVD^2V^TVD^{-1} + \lambda D^{-1}V^T\Omega_N VD^{-1})^{-1}U^T \\
&= U(I + \lambda D^{-1}V^T\Omega_N VD^{-1})^{-1}U^T \\
&= (U^TU + \lambda U^TD^{-1}V^T\Omega_N VD^{-1}U)^{-1} \\
&= (I + \lambda \underbrace{U^TD^{-1}V^T\Omega_N VD^{-1}U}_{\mathbf{K}})^{-1} \\
&= (I + \lambda\mathbf{K})^{-1}
\end{aligned}$$

The Demmler-Reinsch Basis

The matrix $\mathbf{K}$ is positive semidefinite and we write

$$\mathbf{K} = \bar{U} D \bar{U}^T$$

where $D = \text{diag}(d_1, \ldots, d_N)$ with $0 = d_1 = d_2 < d_3 \leq \ldots \leq d_N$ and $\bar{U}$ is orthogonal.

The columns in $\bar{U}$, denoted $\bar{u}_1, \ldots, \bar{u}_N$, are known as the Demmler-Reinsch basis.

The Demmler-Reinsch basis is a (the) orthonormal basis of $\mathbb{R}^N$ with the property that the smoother $\mathbf{S}_\lambda$ is diagonal in this basis:

$$\mathbf{S}_\lambda = \bar{U}(I + \lambda D)^{-1} \bar{U}^T$$

The eigenvalues are in decreasing order

$$\rho_k(\lambda) = \frac{1}{1 + \lambda d_k}$$

for $k = 1, \ldots, N$ – and $\rho_1(\lambda) = \rho_2(\lambda) = 1$.

The Demmler-Reinsch Basis

We may also observe that

$$\mathbf{S}_\lambda \bar{u}_k = \rho_k(\lambda)\bar{u}_k.$$

We think of and visualize $\bar{u}_k$ as a function evaluated in the points $x_1, \ldots, x_N$.

One important consequence of these derivations is that the Demmler-Reinsch basis does not depend upon λ and we can clearly see the effect of λ through the eigenvalues $\rho_k(\lambda)$ that work as shrinkage coefficients multiplied on the basis vectors.

A Bias-Variance Decomposition

Assume that conditionally on $\mathbf{X}$ the Y_i's are uncorrelated with common variance σ^2. Then with $\mathbf{f} = E(\mathbf{Y}|\mathbf{X}) = E(\mathbf{Y}^{\text{new}}|\mathbf{X})$ and $\mathbf{Y}^{\text{new}}$ independent of $\mathbf{Y}$

$$\begin{aligned}
E(||\mathbf{Y}^{\text{new}} - \hat{\mathbf{f}}||^2|\mathbf{X}) &= E(||\mathbf{Y}^{\text{new}} - \mathbf{S}_\lambda \mathbf{Y}||^2|\mathbf{X}) \\
&= E(||\mathbf{Y}^{\text{new}} - \mathbf{f}||^2|\mathbf{X}) + ||\mathbf{f} - \mathbf{S}_\lambda \mathbf{f}||^2 \\
&\quad + E(||\mathbf{S}_\lambda(\mathbf{f} - \mathbf{Y})||^2|\mathbf{X}) \\
&= N\sigma^2 + \underbrace{||(I - \mathbf{S}_\lambda)\mathbf{f}||^2}_{\text{Bias}(\lambda)^2} + \sigma^2 \text{trace}(\mathbf{S}_\lambda^2) \\
&= \sigma^2(N + \text{trace}(\mathbf{S}_\lambda^2)) + \text{Bias}(\lambda)^2
\end{aligned}$$

where we use that $E(\hat{\mathbf{f}}|\mathbf{X}) = E(\mathbf{S}_\lambda \mathbf{Y}|\mathbf{X}) = \mathbf{S}_\lambda \mathbf{f}$. We can also write

$$\text{Bias}(\lambda)^2 = \text{trace}((I - \mathbf{S}_\lambda)^2 \mathbf{f}\mathbf{f}^T).$$

Estimation of σ^2 using low bias estimates

It seems that

$$\mathrm{RSS}(\hat{\mathbf{f}}) = \sum_{i=1}^{N} (y_i - \hat{\mathbf{f}}_i)^2$$

is a natural estimator of $E(||\mathbf{Y} - \hat{\mathbf{f}}||^2|\mathbf{X})$, and its mean is computed as

$$\sigma^2(N - (\mathrm{trace}(2\mathbf{S}_\lambda - \mathbf{S}_\lambda^2)) + \mathrm{Bias}(\lambda)^2.$$

Choosing a low-bias – that is small λ – model we expect $\mathrm{Bias}(\lambda)^2$ to be negligible and we estimate σ^2 as

$$\hat{\sigma}^2 = \frac{1}{N - \mathrm{trace}(2\mathbf{S}_\lambda - \mathbf{S}_\lambda^2)} \mathrm{RSS}(\hat{\mathbf{f}}).$$

From this point of view it seems that

$$\mathrm{trace}(2\mathbf{S}_\lambda - \mathbf{S}_\lambda^2)$$

can also be justified as the effective degrees of freedom.

Reproducing Kernel Hilbert Spaces

On any space Ω, not necessarily a subset of $\mathbb{R}^p$, a kernel is a function

$$K : \Omega \times \Omega \to \mathbb{R}$$

with the property that if $x_1, \ldots, x_N \in \Omega$ then the $N \times N$ matrix

$$\mathbf{K} = \{K(x_i, x_j)\}_{i,j}$$

is positive semidefinite. We will only kernels that are positive definite.

The inner product space

$$\mathcal{H}_K^{\text{pre}} = \left\{ \sum_m \alpha_m K(\cdot, y_m) \right\}$$

with inner product

$$\left\langle \sum_m \alpha_m K(\cdot, y_m), \sum_n \alpha'_n K(\cdot, y'_n) \right\rangle = \sum_{m,n} \alpha'_n \alpha_m K(y'_n, y_m)$$

can be abstractly completed.

Reproducing Kernel Hilbert Spaces

The existence of the completion $\mathcal{H}_K$, which is a Hilbert space with reproducing kernel K is known as the Moore-Aronszajn theorem. If $f \in \mathcal{H}_K$ then

$$\langle f, K(\cdot, x) \rangle = f(x).$$

If $\Omega \subseteq \mathbb{R}^p$ then under additional regularity conditions there are orthogonal functions ϕ_i such that

$$K(x, y) = \sum_i \gamma_i \phi_i(x) \phi_i(y)$$

where $\gamma_i \geq 0$ and $\sum_i \gamma_i^2 < \infty$. This is known as Mercer's theorem. Then $\mathcal{H}_K$ becomes concrete as

$$f = \sum_i c_i \phi_i$$

with $\sum_i \frac{c_i^2}{\gamma_i} < \infty$.

The Finite-Dimensional Optimization Problem

Considering the abstract problem

$$\min_{f \in \mathcal{H}_K} \sum_{i=1}^{N} (y_i - f(x_i))^2 + \lambda ||f||_K^2$$

a solution is then of the form $\sum_{i=1}^{N} \alpha_i K(\cdot, x_i)$. We need to solve

$$\min_{\alpha \in \mathbb{R}^N} (\mathbf{y} - \mathbf{K}\alpha)^T (\mathbf{y} - \mathbf{K}\alpha) + \lambda \alpha^T \mathbf{K} \alpha.$$

The solution (unique when **K** is positive definite) is

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

and the predicted values are

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

Data acquisition – and interpretations

In this course we consider observational data. Roughly we have

- Observational data; Both X and Y are sampled from an (imaginary) population.
- Non-observational; e.g. a designed experiment where we fix X by the design and sample Y.

For observational data how should we interpret $Y|X$?

Example

In toxicology we are interested in measuring the effect of a (toxic) compound on the plant, say.

Consider a naturally occurring compound A and a plant Z.

- Full observational study: On N randomly selected fields we measure Y = the amount of plant Z and X = the amount of compound A.
- Semi-observational study: On each of N randomly selected fields we plant R plants Z. After T days we measure Y = the amount of plant Z and X = the amount of compound A.
- Designed experiment: On each of N identical fields we plant R plants Z. We add according to a design scheme the amount X_i of compound A to field i. After T days we measure Y = the amount of plant Z.

Causality

In toxicology – as in most parts of science – the basic question is causal relations.

Is the compound A toxic? Does it actually kill plant Z?

The pragmatic farmer; Can I grow plant Z on my soil?

The former question can only be answered by the designed experiment. The latter may be answered by prediction of the yield based on a measurement of compound A.

The latter prediction is not justified by causality – only by correlation.

Probability Models and Causality

Probability theory is completely blind to causation!

From a technical point of view the regression of Y on X is carried out precisely in the same manner whether the data are observational or from a designed experiment. The probability conditional model is the same.

For the ideal designed experiment we control X and all systematic variation in Y can only be ascribed to X.

For the observational study we observed the pair (X, Y) Systematic variations in Y can be due to X but there is no evidence of causality.

Interventions

Many, many studies are observational and many, many conclusions are causal.

- If the children in Gentofte get higher grades compared to Copenhagen, should I put my child in one of their schools?
- If the children in large schools get higher grades compared to children in small schools, should we build larger schools?
- If people on night-shifts get more ill than those with a regular job, is it then dangerous to take night-shifts? Should I not take a night-shift job?
- If smokers more frequently get lung cancer is that because they smoke? Should I stop smoking?

All four final questions are phrased as interventions. Data from an observational study does not alone provide information on the result of an intervention.

What if $Y|X$ then?

For observational data we must think of $Y|X$ as an observational conditional distribution meaning that (X, Y) must be sampled exactly the same way as $(x_1, y_1), \ldots, (x_N, y_1)$ were.

Then if $X = x$ but Y has not been disclosed to us, $Y|X = x$ is a sensible conditional distribution of Y.

If we remember to gather data using the same principles as when we later want to use $Y|X$ for predictions, we can expect that $Y|X$ is useful for predictions – even if there is no alternative evidence of causation.

Violations of a consistent sampling scheme is the Achilles heel of predictions based on observational data. And we can not trust predictions if we make interventions.