Homework 2 10/36-702 Due Friday March 3 at 3:00 pm

1 Splines, kernels, and wavelets

In the next several parts, you'll do some work on the different ways to represent the smoothing spline estimator. (Some of the details can be found in the lecture notes and in that case you can just recapitulate the arguments rigorously here.)

(a) Let \hat{f} denote the cubic smoothing spline fit on the pairs $(x_i, y_i) \in \mathbb{R} \times \mathbb{R}$, i = 1, ..., n, and S denote its smoother matrix, i.e., so that $\hat{\mu} = (\hat{f}(x_1), ..., \hat{f}(x_n)) \in \mathbb{R}^n$ satisfies

$$\widehat{\mu} = S y.$$

Let g_1, \ldots, g_n be a basis for the set of natural cubic splines with knots at x_1, \ldots, x_n . You may assume (as proved in the lecture notes) that the smoothing spline solution \hat{f} lies in the span of g_1, \ldots, g_n . Show that the vector of fitted values can be written as $\hat{\mu} = N\hat{\beta}$, where

$$\widehat{\beta} = \operatorname*{argmin}_{\beta \in \mathbb{R}^n} \| y - N\beta \|_2^2 + \lambda \beta^T \Omega\beta,$$

where $\lambda \ge 0$ is the smoothing spline tuning parameter, and hence

$$S = N(N^T N + \lambda \Omega)^{-1} N^T$$

for a suitable matrices $N, \Omega \in \mathbb{R}^{n \times n}$ defined in terms of g_1, \dots, g_n . (Some details for this can be found in the lecture notes and you can just recapitulate the arguments rigorously here.)

(b) Recall that the leave-one-out cross-validation (CV) error is defined as

$$\operatorname{CV}(\widehat{f}) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \widehat{f}^{-i}(x_i) \right)^2.$$

where \hat{f}^{-i} is the estimator trained on all but the *i*th pair (x_i, y_i) , for i = 1, ..., n. The lecture notes gave the following magical shortcut computation for the leave-one-out CV error:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}^{-i}(x_i))^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{f}(x_i)}{1 - S_{ii}}\right)^2,\tag{1}$$

and here you will prove it for the smoothing spline estimator.

(i) First fix some arbitrary *i*. Define $z \in \mathbb{R}^n$ so that $z_j = y_j$ for all $j \neq i$, and $z_i = \hat{f}^{-i}(x_i)$. Prove that $\hat{f}^{-i}(x_i) = \sum_{j=1}^n S_{ij} z_j$, and hence

$$\widehat{f}^{-i}(x_i) = \frac{1}{1 - S_{ii}} \left(\widehat{f}(x_i) - S_{ii} y_i \right).$$
(2)

(Hint: there are different ways to prove this; one way involves recalling the Sherman-Morrison update formula, and another involves considering the smoothing spline criterion with response z, when we include and exclude the *i*th sample point.)

(ii) Establish using (2) that

$$\widehat{f}(x_i) - \widehat{f}^{-i}(x_i) = S_{ii} \left(y_i - \widehat{f}^{-i}(x_i) \right),$$

and hence

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

Square both sides, and sum over i = 1, ... n, to give the result in (1).

- (c) Verify the leave-one-out CV formula in (1) empirically. Simulate some data (or find some real data), perform leave-one-out CV according to the original definition, according to the shortcut formula, and show that these match over a discrete of tuning parameter values λ . Then plot the smoothing spline fit, on top of the data set, at the value of λ that minimizes the CV error. For this part and part (d) you may use any available package you prefer (e.g. smooth.spline and KernSmooth in R).
- (d) Prove that the property (2) and so the leave-one-out shortcut (1) also holds for kernel smoothing. Verify that this leave-one-out CV formula holds empirically, for an example data set with two-dimensional inputs points. As you did with smoothing splines, plot the data and the kernel smoothing estimate (now a 3d plot, or a contour plot, or a heat map, etc.), at the value of the bandwidth *h* that minimizes the CV error.
- (e) Let $W \in \mathbb{R}^{n \times n}$ be an orthogonal basis matrix (so, like the matrix N above but with orthonormal columns). Consider the wavelet smoothing estimator defined as $\hat{\mu} = W \hat{\beta}$, and

$$\widehat{\beta} = \operatorname*{argmin}_{\beta \in \mathbb{R}^n} \frac{1}{2} \|y - W\beta\|_2^2 + \lambda \|\beta\|_1,$$

where $\lambda \ge 0$ is a tuning parameter. Show that

$$\widehat{\beta}_i = T_{\lambda}((W^T y)_i), \quad i = 1, \dots, n,$$

where T_{λ} is the soft-thresholding operator with threshold level λ , i.e.,

$$T_{\lambda}(x) = \begin{cases} x - \lambda & x > \lambda \\ 0 & |x| \le \lambda \\ x + \lambda & x < -\lambda. \end{cases}$$

Conclude that wavelet smoothing is not a linear smoother, i.e., $\hat{\mu} = \hat{\mu}(y)$ is not a linear function of *y*. (For this, you must only demonstrate that $\hat{\mu}(ay+b) \neq a\hat{\mu}(y) + b$ for some $y, b \in \mathbb{R}^n$, $a \in \mathbb{R}$.)

(f) Assume the normal data model

$$y \sim N(\mu, \sigma^2 I),$$

~

for some mean vector $\mu \in \mathbb{R}^n$ and variance parameter $\sigma^2 > 0$. Prove that the risk of the wavelet smoothing estimator with soft-thresholding function, for an arbitrary choice of tuning parameter λ , satisfies

$$\frac{1}{n} \mathbb{E} \| \widehat{\mu} - \mu \|_2^2 = \frac{1}{n} \sum_{i=1}^n r((W^T \mu)_i, \lambda),$$

where

$$r(\theta,\lambda) = \theta^2 \int_{\frac{-\lambda-\theta}{\sigma}}^{\frac{\lambda-\theta}{\sigma}} \phi(z) dz + \int_{\frac{\lambda-\theta}{\sigma}}^{\infty} (\sigma z - \lambda)^2 \phi(z) dz + \int_{-\infty}^{\frac{-\lambda-\theta}{\sigma}} (\sigma z + \lambda)^2 \phi(z) dz,$$

and ϕ denotes the standard (univariate) normal density function.

(g) Prove, following from the last part, that for $\lambda = \sigma \sqrt{2 \log n}$, the wavelet smoothing smoother estimator with soft-thresholding function has risk satisfying

$$\frac{1}{n} \mathbb{E} \| \widehat{\mu} - \mu \|_2^2 \le \frac{\sigma^2}{n} + \frac{2\log n + 1}{n} \sum_{i=1}^n \min\left\{ (W^T \mu)_i^2, \sigma^2 \right\}$$

(Hint: start with $\sigma^2 = 1$ for simplicity. Prove that, for any $\lambda, \theta \ge 0$, it holds that $0 \le \partial r(\theta, \lambda)/\partial \theta \le 2\theta$; from this, argue that $r(\theta, \lambda)$ is monotone increasing in θ , and further

$$r(\theta, \lambda) \le r(0, \lambda) + \min\{\theta^2, r(\infty, \lambda)\}.$$

Then, derive upper bounds on $r(0,\lambda), r(\infty,\lambda)$ (for the former you can use Mills' ratio, and for the latter use direct arguments), to give

$$r(\theta,\lambda) \le e^{-\lambda^2/2} + \min\{\theta^2, 1+\lambda^2\}$$

Plug in $\lambda = \sqrt{2\log n}$ to give

$$r(\theta, \lambda) \le \frac{1}{n} + \min\{\theta^2, 1 + 2\log n\} \le \frac{1}{n} + (2\log n + 1)\min\{\theta^2, 1\}.$$

Finally, argue that an analogous bound holds for general σ^2 , and average the right-hand side above over $\theta = (W^T \mu)_i$, i = 1, ..., n, to give the result.)

In what situations (i.e., for what configurations of the mean μ) will this risk bound be small?

(Bonus) Consider a "poor man's" version of smoothing splines, whose fitted values are given by $\hat{\mu} = W \hat{\beta}$, and

$$\widehat{\beta} = \operatorname*{argmin}_{\beta \in \mathbb{R}^n} \|y - W\beta\|_2^2 + \lambda \|\beta\|_2^2,$$

where $W \in \mathbb{R}^{n \times n}$ is an orthogonal basis matrix. Show that

$$\widehat{\beta}_i = \frac{(W^T y)_i}{1+\lambda}, \ i = 1, \dots, n$$

Using the normal data model we considered above for wavelet smoothing, give an upper bound on the risk of this (poor man's) version of smoothing splines. Compare it to our previous upper bound for the risk of wavelet smoothing, across different settings for μ , when both use the same orthogonal transformation W, and both are properly tuned. When does wavelet smoothing have better risk? When does it have worse risk?

2 Density estimation

Let \hat{p}_h be the kernel density estimator (in one dimension) with bandwidth $h = h_n$ satisfying $h_n \to 0$ and $nh_n \to \infty$. Let $s_n^2(x) = \operatorname{var}(\hat{p}_h(x))$. Assume that the density function is bounded, and the kernel function is bounded with a compact support.

(a) Show that

$$\frac{\widehat{p}_h(x) - p_h(x)}{s_n(x)} \rightsquigarrow N(0, 1)$$

where $p_h(x) = \mathbb{E}[\hat{p}_h(x)]$.

Hint: The Lyapunov central limit theorem says the following: Suppose that Y_1, Y_2, \ldots are independent. Let $\mu_i = \mathbb{E}[Y_i]$ and $\sigma_i^2 = \operatorname{Var}(Y_i)$. Let $s_n^2 = \sum_{i=1}^n \sigma_i^2$. If

$$\lim_{n \to \infty} \frac{1}{s_n^{2+\delta}} \sum_{i=1}^n \mathbb{E}[|Y_i - \mu_i|^{2+\delta}] = 0$$

for some $\delta > 0$. Then $s_n^{-1} \sum_i (Y_i - \mu_i) \rightsquigarrow N(0, 1)$.

(b) Suppose that the bandwidth h_n is chosen optimally. Assume that the density function has a bounded continuous second derivative. Show that $bias^2(x)/s_n^2(x)$ does not necessarily tend to 0 as $n \to \infty$. To reduce the bias, one can use a simple trick called "twicing." Define

$$p^{\dagger}(x) = 2\widehat{p}_{h}(x) - \widehat{p}_{\sqrt{2}h}(x).$$

Show that, for this estimator, $bias^2(x)/s_n^2(x) \rightarrow 0$.

Comment: this implies that $\frac{p_h^{\dagger}(x)-p(x)}{s_n(x)} \rightsquigarrow N(0,1)$. The twicing estimator is equivalent to using the modified kernel $\widetilde{K} = 2K - K \star K$. Here, \star denotes convolution.

3 Clustering

Suppose that $Y_1, \ldots, Y_n \sim p$ where $Y_i \in \mathbb{R}^d$. Suppose that

$$p = \sum_{j=1}^k \pi_j p_j$$

where $\pi_j > 0$, $\sum_j \pi_j = 1$ and p_j is a density function supported on a compact, connected set A_j . Let

$$\delta_{j\ell} = \inf_{x \in A_j} \inf_{y \in A_\ell} ||x - y||$$

for $j \neq \ell$. Assume that $\min_{j,\ell} \delta_{j\ell} \ge \Delta > 0$. Let $\gamma = \max_j \operatorname{diameter}(A_j)$. Finally, we will also assume that p_j is uniform on A_j .

- (a) Show that, if Δ is sufficiently large, then k-means is consistent. That is, each A_j is contained in a separate Voronoi cell, with probability tending to 1. Show that if Δ is not large enough, consistency can fail.
- (b) Assume there exists $\epsilon_0, c > 0$ that for all $x \in \bigcup_{j=1}^k A_j$ and $\epsilon \in (0, \epsilon_0)$, $p(\mathscr{B}(x, \epsilon)) \ge c\omega\epsilon^d$, where $\mathscr{B}(x, \epsilon) := \{y \in \mathbb{R}^d : \|x y\| < \epsilon\}$ and ω being a volume of a unit ball. Show that for sufficiently small h > 0 (but not necessarily going to 0), level set clustering using the kernel density estimator with the boxcar kernel $K(x) = \frac{1}{\omega}I(x \in \mathscr{B}(0, 1))$ is consistent in the sense that, with probability tending to 1, there exist disjoint density level set clusters B_1, \ldots, B_k such that $A_j \subset B_j$ for each j. You can assume that conditions in Theorem 15 of Density Estimation notes are satisfied.