

## AM261 – Recent Applications of Probability and Statistics

### Project 6: Cross-validated smoothing for kernel density estimators

Given a kernel function  $k(x)$ ,  $x \in R^1$  ( $k$  is non-negative and integrates to one), and a sample  $X_1, \dots, X_N$  from an (unknown) density function  $f$ , the kernel estimator for  $f$  is

$$\hat{f}_{X_1^N}^\sigma(x) = \frac{1}{N} \sum_{i=1}^N \frac{1}{\sigma} k\left(\frac{x - X_i}{\sigma}\right)$$

The cross-validated smoothing parameter,  $\sigma_N$ , comes from the cross-validated log-likelihood function:

$$\begin{aligned} CV(\sigma) &= \frac{1}{N} \sum_{i=1}^N \log\left\{\frac{1}{N-1} \sum_{j=1, j \neq i}^N \frac{1}{\sigma} k\left(\frac{X_i - X_j}{\sigma}\right)\right\} \\ \sigma_N &= \arg \max_{\sigma} CV(\sigma) \end{aligned}$$

and the cross-validated density estimator is then  $\hat{f}_{X_1^N}^\sigma$  with  $\sigma = \sigma_N$ .

1. (Distribution with a heavy tail approximated by kernels with light tails – the tail wags the dog.) Let  $f(x) = e^{-x}$ ,  $x \geq 0$ , and

$$k(x) = 30(x - .5)^2(x + .5)^2 \quad x \in [-.5, .5]$$

For each of  $N = 100, 500$ , and  $1000$ , draw  $N$  random samples from  $f$  and find  $\sigma_N$ , the cross-validated smoothing parameter. You can do this by “brute force” by computing the cross-validated log likelihood,  $CV(\sigma)$ , for each of  $\sigma = .05, .10, .15, \dots, 10$  and selecting the maximizing value of  $\sigma$ . (But you might as well confine your search to values of  $\sigma$  for which  $CV(\sigma) > -\infty$ : Let  $d_i$  be the distance from  $x_i$  to its nearest neighbor. Then  $CV(\sigma) > -\infty \Rightarrow \sigma > 2 \cdot \max_i d_i$ .) For each  $N$  plot  $CV(\sigma)$  against  $\sigma$ , and, on a second figure, plot three densities: (i)  $f(x)$ , (ii)  $\hat{f}_{X_1^N}^\sigma$  at  $\sigma = \sigma_N$ , and (iii)  $\hat{f}_{X_1^N}^\sigma$  with a “hand-crafted”  $\sigma$ . For the hand-crafted  $\sigma$ , cheat, by choosing  $\sigma$  to best match  $\hat{f}_{X_1^N}^\sigma$  to  $f$ , either by minimizing some criterion (e.g. the integral over  $x$  of the absolute difference) or, more simply, by eye after some trial-and-error.

Is it your sense that cross-validation is over-smoothing, under-smoothing, or getting it about right? Can you explain the performance in terms of the relationship between  $f$  and  $k$ ?

2. (Distribution and kernel well matched.) Use the same kernel as in part (1), but this time experiment with the density

$$f(x) = \frac{5}{81} * \{x^2(x-3)^2 1_{x \leq 3} + (x-2)^2(x-5)^2 1_{x \geq 2}\} \quad x \in [0, 5]$$

Proceed exactly as in part (1), but using sample sizes  $N = 10, 100,$  and  $1000$ . (An easy way to draw samples from  $f$  is by “acceptancerejection”: Choose a vector  $(X, Y)$  from the uniform distribution on  $[0, 5] \times [0, .35]$ , and then accept  $X$  if  $Y \leq f(X)$ .)

Answer again the question about the relationship between  $f$  and  $k$  and the degree of smoothing.

3. (Distribution with light tails approximated by kernels with heavy tails.) Let  $f$  be the uniform density on  $[0, 1]$  and let  $k$  be the Cauchy density:

$$k(x) = \frac{1}{\pi(1+x^2)} \quad x \in (-\infty, \infty)$$

Repeat again the experiment, with sample sizes  $N = 10, 100,$  and  $1000$ , but with  $\sigma$  running from  $.001$  to  $.2$  in steps of size  $.001$ . Be careful not to exclude, *a priori*, any values of  $\sigma$ : the Cauchy distribution has support on the entire real line and hence the cross-validated log likelihood is never negative infinity.

In this experiment, is the cross-validated estimator providing an appropriate degree of smoothing? Why?