LECTURE 4

Cross-Validation

It is clear that choice of the bandwidth will have an important effect on how good $f_n(x)$ is as an estimate of $f(x)$. The optimal asymptotic choice of bandwidth h does depend on the unknown function $f(x)$. This had led a number of people to suggest choices of h determined by the data itself. Some of these methods of choosing h are called cross-validation methods and we shall describe two of them.

The first is maximum likelihood cross-validation [Habbema, Hermans and Vanderbroek (1974)]. Let $X_1, \ldots, X_n$ be independent identically distributed random variables with unknown density function $f(x)$. A standard kernel density function estimate $f_n(x)$ based on the weight function w and bandwidth h is to be considered. To estimate one carries through the following procedure. Consider the estimate at X_i

$$_i f_n(X_i; h)$$

based on all the observations except for X_i with weight function w and bandwidth h. Look at the product

$$\prod_{i-1}^{n} {}_i f_n(X_i; h) = L_n(h)$$

and determine the value of h maximizing this product. Take this value h as the bandwidth in one's estimate of the density function. Chow, Geman and Wu (1983) have shown that if f is a density with compact support and w a continuous kernel positive at 0 and of compact support, that $f_n(x)$ using this cross-validated bandwidth converges in mean to f almost surely.

If the density f is not of compact support and the tail decreases at a sufficiently slow rate, the bandwidth $\hat{h}_n$ obtained by maximum likelihood cross-validation will not lead to a consistent density estimate when w is bounded and of compact support. The boundary between consistency and inconsistency appears to be given by the exponential distribution. This was pointed out by Schuster and Gregory (1981) and we give part of their argument. Let w be a kernel with support in $[-1, 1]$ that is bounded by M. The

argument is given in terms of the left tail of the density. Let F be the distribution function of the population and $g(u) = u/f(F^{-1}(u))$. Assume that f is continuous and $\liminf_{u \to 0} g(u) = \bar{g} > 0$. One can say then that f has a long left tail. We show that if $\bar{g} = \infty$, then $\hat{h}_n \to \infty$ and the kernel density estimate is inconsistent. A slightly more elaborate argument shows that if $\bar{g} > 0$ but finite, then one still has inconsistency [see Schuster and Gregory (1981) for this discussion]. Let $X_{1n} < X_{2n} < \cdots < X_{nn}$ be the order statistics of the sample. Then if $\hat{h}_n$ is well defined, we have $X_{1n} - X_{2n} \leq \hat{h}_n$ since otherwise ${}_{1n} f_n(X_{1n}; h) \equiv 0$ and then $L_n(h) \equiv 0$. If $u_{in} = F(X_{in})$, then

$$X_{2n} - X_{1n} = F^{-1}(u_{2n}) - F^{-1}(u_{1n}) = g(u_n^*)(u_{2n} - u_{1n})/u_n^*,$$

where $u_{1n} \leq u_n^* \leq u_{2n}$. Therefore

$$g(u_n^*)(u_{2n} - u_{1n})/u_{2n} \leq \hat{h}_n.$$

$(u_{2n} - u_{1n})/u_{2n}$ has a uniform distribution and one can see that

$$P(\hat{h}_n < b\varepsilon) \leq \varepsilon + P(h(u_n^*) < b)$$

if $b, \varepsilon > 0$. If $\bar{g} > 0$ by taking $0 < b < \bar{g}$, one can see that $\hat{h}_n$ cannot approach 0 in probability. If $\bar{g} = \infty$, it is clear that $\hat{h}_n \to \infty$ in probability. If $\hat{h}_n \to \infty$ in probability, then $\sup_x f_n(x) \to 0$ in probability and one cannot have consistency. Notice that in the case of the exponential distribution $A(x) = e^x$, $-\infty < x < 0$, $\bar{g}$ is positive and finite. That implies that if one has a density f such that

$$e^x/f(x) \to 0$$

as $x \to -\infty$, the corresponding $\bar{g} = \infty$ and one does not have consistency of the kernel estimates $f_n(x)$ with $\hat{h}_n$ determined by maximum likelihood cross-validation.

An alternative way of choosing a bandwidth in terms of the observations is called least squares cross-validation. It is motivated by the discussion given earlier where an optimal choice of h (not depending on the location x) is obtained by minimizing

$$E\int |f_n(x; h) - f(x)|^2 \, dx \tag{4.1}$$

as a function of h. Since the answer depends on the unknown f, it is useless. However, a plausible alternative might be minimizing

$$\int |f_n(x; h) - f(x)|^2 \, dx \tag{4.2}$$

as a function of h. In fact, it looks more plausible than minimizing (4.1) because it is in terms of an expression determined by the data themselves. Minimizing (4.2) is equivalent to minimizing

$$\int f_n(x; h)^2 \, dx - 2\int f_n(x; h) f(x) \, dx. \tag{4.3}$$

The second integral in (4.3) still depends on the unknown f and one would like to replace it by a good estimate given completely in terms of the observations. Now

$$E\int f_n(x,h)\,f(x)\,dx = \int f(x)\,E\omega\left(\frac{x-X}{h}\right)\frac{1}{h}\,dx$$

$$= \int f(x)\,f(y)\,\frac{1}{h}\omega\left(\frac{x-y}{h}\right)dx\,dy$$

$$= E\left[\frac{1}{n(n-1)}\sum_{i\neq j}\frac{1}{h}\omega\left(\frac{X_i-X_j}{h}\right)\right].$$

The expression

$$\frac{1}{n(n-1)}\sum_{i\neq j}\frac{1}{h}\omega\left(\frac{X_i-X_j}{h}\right) = \frac{1}{h}\sum_{j=1}^{n}{}_j f_n(X_j;h) \tag{4.4}$$

has the same expectation as $\int f_n(x,h)f(x)\,dx$ and one notion is to replace the second integral by this expression in (4.3). A small further modification would be to replace the normalization $n(n-1)$ in (4.4) by n^2.

In an earlier discussion we considered the value of $h_0 = h_0(n)$ that would minimize

$$E(h) = \int E\{f_n(x;h) - f(x)\}^2\,dx$$

asymptotically as $n \to \infty$ under appropriate assumptions. This choice of h_0 clearly depended on the unknown density function f. As already noted it had been suggested that one actually ought to try minimizing

$$I(h) = \int \{f_n(x;h) - f(x)\}^2\,dx$$

instead of (4.1). Let us call the minimizing value here $\hat{h}_0$. It is clear that $I(h_0) \geq I(\hat{h}_0)$. We have already noted some of the limitations of maximum likelihood cross-validation and noted that minimizing

$$C(h) = \int f_n^2(x;h) - 2n^{-1}\sum_{j=1}^{n}{}_j f_n(X_j;h) \tag{4.5}$$

is a plausible alternative to that of minimizing $I(h)$ in view of the lack of knowledge of f. Let us call the h value minimizing (4.5) $\hat{h}_c$.

Hall and Marron (1987) have compared the asymptotic behavior of the h-values, h_0, $\hat{h}_0$ and $\hat{h}_c$ under the following assumptions. They assume that $\omega(\cdot)$ is a symmetric function with finite support and Hölder-continuous derivative ω'. Further

$$\int \omega = 1, \qquad \int u^2\omega(u)\,du = 2c \neq 0.$$

Also f is assumed twice continuously differentiable with f and its derivatives bounded, f' and f'' integrable and f'' uniformly continuous.

Let $D = I(h) - E(h)$. Also it is clear that $C(h) = I(h) + \delta - \int f^2$, where

$$\tfrac{1}{2}\delta = \int f f_n - n^{-1} \sum_{j-1}^{n} {}_j f_n(X_j).$$

From an earlier discussion we have seen that

$$E(h) = c_1(nh)^{-1} + c_2 h^4 + o\{(nh)^{-1} + h^4\}$$

with $c_1 = \int \omega^2$ and $c_2 = c^2 \int (f'')^2$. On differentiating twice one finds

$$E''(h) = 2c_1(nh^3)^{-1} + 12c_2 h^2 + o\{(nh^3)^{-1} + h^2\}$$

as $h \to 0$ and $n \to \infty$. This implies that $h_0 \simeq c_0 n^{-1/5}$ with $c_0 = (c_1/4c_2)^{1/5}$ and $E''(h_0) \simeq c_3 n^{-2/5}$ with $c_3 = 2c_1 c_0^{-3} + 12c_2 c_0^2$.

First they prove a limit theorem for $\hat{h}_0 - h_0$. It is clear that

$$\begin{aligned} 0 = I'(\hat{h}_0) &= E'(\hat{h}_0) + D'(\hat{h}_0) \\ &= (\hat{h}_0 - h_0)E''(h^*) + D'(\hat{h}_0) \end{aligned}$$

with h^* between h_0 and $\hat{h}_0$. A succession of estimates allows them to show that

$$\hat{h}_0 = h_0 + O(n^{-1/5 - \varepsilon})$$

for some $\varepsilon > 0$ and that

$$D'(\hat{h}_0) = D'(h_0) + o(n^{-7/10}).$$

An argument like that used to prove asymptotic normality for $I(h)$ is used to show that

$$n^{7/10} D'(h_0) \to N(0, \sigma_0^2) \tag{4.6}$$

in distribution, where

$$\sigma_0^2 = (2/\sigma_0)^3 \left(\int f^2\right) \int \left[\int \omega(y+z)\{\omega(z) - L(z)\}\, dz\right]^2 dy$$

$$+ (4cc_0)^2 \left\{\int (f'')^2 f - \left(\int f'' f\right)^2\right\}$$

and

$$L(z) = -z\omega'(z).$$

Now $h^*/h_0 \to_p 1$ and so it follows that $E''(h^*) = c_3 n^{-2/5} + o(n^{-2/5})$. Also from (4.6) one sees that

$$h^{7/10} D'(\hat{h}_0) \to N(0, \sigma_0^2)$$

in distribution. This implies that

$$n^{3/10}(\hat{h}_c - h_0) \to N(0, \sigma_c^2 c_3^{-2})$$

in distribution. A more elaborate argument of a similar character shows that

$$n^{3/10}(\hat{h}_c - \hat{h}_0) \to N(0, \sigma_c^2 c_3^{-2}),$$

where

$$\sigma_0^2 = (2/c_0)^3\left(\int f^2\right)\left(\int L^2\right) + (4cc_0)^2\left\{\int (f'')^2 f - \left(\int f'' f\right)^2\right\}.$$

Under the additional condition that ω has a second derivative that is Hölder continuous, it is shown that

$$n\{I(h_0) - I(\hat{h}_0)\} \to \tfrac{1}{2}\sigma_0^2 c_3^{-1}\chi_1^2$$

and

$$n\{I(\hat{h}_c) - I(\hat{h}_0)\} \to \tfrac{1}{2}\sigma_0^2 c_3^{-1}\chi_1^2$$

in distribution where χ_1^2 is a chi-square variable with one degree of freedom.

In this lecture we have discussed a few versions of cross-validation. Rice (1984) has considered modifications of cross-validation and their usefulness in determining bandwidth in regression estimation. However, one should note that there are earlier and perhaps more immediately intuitive procedures that relate to data driven choices of bandwidth on either a local or global basis. It is clear from formula (2.3) that an asymptotically optimal local choice of bandwidth in density estimation would require initial estimates of the density function and its second derivatives. Woodroofe (1970) suggested a two-step procedure involving such initial estimates to implement the choice of an asymptotically optimal bandwidth sequence. Cross-validation is usually presented as a global procedure though it can be modified to obtain a more localized version. In many situations it is clear that a locally optimal procedure would have advantages. Adaptations of ideas centering about a data driven choice of bandwidth to the case of regression estimation can be found in Müller (1985) and Mack and Müller (1987).