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 $X_1, \dots, X_n$ 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