THE INSTITUTE OF STATISTICS

THE CONSOLIDATED UNIVERSITY OF NORTH CAROLINA

ON THE MODIFIED LIKELIHOOD FOR DENSITY ESTIMATION

by

J. S. Marron
University of North Carolina, Chapel Hill

and

R. L. Taylor
University of Georgia

Mimeo Series #1560

1984

DEPARTMENT OF STATISTICS
Chapel Hill, North Carolina

On the Modified Likelihood for Density Estimation

by

J. S. Marron[1]
University of North Carolina, Chapel Hill

and

R. L. Taylor
University of Georgia

1 Research partially supported by ONR contract N00014-81-K-0373

Summary

It is shown that some recent results of Wong (1983) concerning his version of the modified likelihood criterion for smoothing parameter selection in kernel density estimation can be very misleading, because the wrong mode of convergence is established. An example is given to demonstrate that the results are false when a more reasonable mode of convergence is used. Slightly stronger conditions are added and valid proofs for the correct version of these results are indicated.

I. Introduction

Let $\underset{\sim}{X} = \{X_1,\ldots,X_n\}$ be a random sample from a density f. The kernel estimator of f is

$$f_\lambda(x,\underset{\sim}{X}) = n^{-1}\sum_{i=1}^{n} K_\lambda(x-X_i), \tag{1.1}$$

where

$$K_\lambda(x) = \frac{1}{\lambda} K(\frac{x}{\lambda}).$$

The central problem in the field of density estimation is the choice of the smoothing parameter or bandwidth, λ. As noted in Wong (1983), several authors have proposed selecting λ by maximizing the cross-validation function

$$\hat{R}^{CV}(\lambda) = n^{-1}\sum_{i=1}^{n} \log f_\lambda(X_i,\underset{\sim}{X}_{(i)}) \tag{1.2}$$

where $\underset{\sim}{X}_{(i)}$ denotes the "leave one out" sample:

$$\underset{\sim}{X}\backslash\{X_i\}.$$

Wong (1983) studies the behavior of (1.2) by looking for conditions under which (letting * denote convolution)

$$\hat{R}^{CV}(\lambda) \xrightarrow{a.e.} \int \log(f*K_\lambda)\,dF. \tag{1.3}$$

This relationship is then used to establish an asymptotic equivalence of (1.2) with a Jackknife selector of λ given by

$$\hat{R}^{Jack}(\lambda) = n^{-1}\sum_{i=1}^{n}\left[\log f_\lambda(X_i,\underset{\sim}{X}) - \log f_\lambda(X_i,\underset{\sim}{X}_{(i)}) + n^{-1}\sum_{j=1}^{n}\log f_\lambda(X_i,\underset{\sim}{X}_{(j)})\right].$$

In the present paper it will be shown that Wong's results need very cautious interpretation for two reasons.

First, the proof of the pivotal Theorem 1 in Wong (1983) is based on a Law of Large Numbers for a sequence of independent, identically distributed random variables in Banach space, which forces the smoothing parameter, λ, to be fixed as $n \to \infty$. Allowing λ to vary with n necessitates consideration of arrays of random variables in Banach spaces where each row is an i.i.d. sequence (see Taylor (1982) and Taylor (1983) for results of this type). This type of asymptotics provides a poor model for studying kernel density estimation because it is well known that, as $n \to \infty$, one needs $\lambda \to 0$ to even have consistency of f_λ (i.e.: convergence to f). In section 2, an example is presented which demonstrates that this issue is vital to understanding when (1.3) holds and is not a minor technical detail.

Second, the theorems of Wong (1983) which study the asymptotic behavior of the functions $R^{CV}(\lambda)$ and $R^{Jack}(\lambda)$ are established only pointwise (in λ). But what is really of interest here is properties of the maximizers of these functions, and to make such inferences requires theorems which are uniform in λ. The example of Section 2 demonstrates that uniformity over all $\lambda > 0$ is impossible. However, in section 3 it is shown that, under stronger assumptions, (1.3) and the asymptotic equivalence of $R^{CV}(\lambda)$ and $R^{Jack}(\lambda)$ are true uniformly over a very reasonable λ range.

2. Counterexample

In Marron (1984) it is seen that if the cumulative distribution function of f is

$$F(x) = e^{-1/x} \qquad \text{for } x > 0,$$

if K is compactly supported, and if λ_n tends to 0 fast enough that

$$(\log n)^2 \lambda_n \to 0, \tag{2.1}$$

then

(2.2) $$\hat{R}^{CV}(\lambda_n) \to -\infty \qquad \text{in probability .}$$

To see the implications of this on the maximizer of $\hat{R}^{CV}$, note that the usual (see, for example, Rosenblatt (1971)) optimal bandwidth of $\lambda_n \sim n^{-1/5}$ easily satisfies (2.1). So the maximizer of $\hat{R}^{CV}$ will be (asymptotically) very far from optimal, or in other words cross-validation fails in this setting. The maximizer of $\hat{R}^{Jack}$ can be shown to suffer similar difficulties.

To relate (2.2) to the results of Wong (1983), note that if K is a probability density which is bounded and positive on a neighborhood of the origin, then an easy analytic argument shows that there is a constant $M > 0$ so that

$$-M < \int \log(f*K_\lambda)dF < M,$$

for $\lambda \in (0,1)$. Thus

$$\hat{R}^{CV}(\lambda_n)/\int \log(f*K_{\lambda n})dF \to -\infty$$

is probability and so (1.3) no longer holds. This does not contradict Wong's Theorem 1 because there (1.3) is established <u>pointwise</u> in $\lambda > 0$, but it does show one can not have (1.3) uniformly over $\lambda > 0$ or even for $\lambda \to 0$ (as with the optimal $\lambda_n \sim n^{-1/5}$), under these assumptions. Hence, the issues raised in section 1 are crucial to understanding the behavior of $\hat{R}^{CV}$ and $\hat{R}^{Jack}$, and not just technical details.

3. Positive Results

The pathology of the previous section is caused by the fact that f(x) is "close to 0 on a set of large measure." To avoid this (and thus have it possible for the maximizer of $\hat{R}^{CV}$ or $\hat{R}^{Jack}$ to have some optimality

properties) it will be assumed that f is supported and bounded away from 0 on an interval [a,b]. Gasser and Müller (1979) (working in the very similar setting of regression estimation) indicated that the estimator (1.1) does not perform well near a and b. To avoid technical details which would obscure the issues under discussion here, the "boundary effect" will be ignored. In Marron (1984) it is seen how to modify $\hat{R}^{CV}$ to overcome this difficulty.

Two technical points in the proof of Theorem 1 of Wong (1983) relate to the Banach space of bounded continuous functions, $C(-\infty,\infty)$. First, continuity of the estimator may be a problem without additional conditions on K (for example, the shifted histogram of Rosenblatt (1956) is discontinuous). However, the choice of K is of secondary importance in the kernel estimate, and most selections of K are continuous. Here it will be assumed that K is a probability density which is positive at the origin and is Lipschitz continuous in the sense that there are positive constants C_1 and α so that

$$|K(x) - K(y)| \leq C_1|x-y|^{\alpha}, \tag{3.1}$$

for all x,y. Most commonly used kernel functions satisfy these assumptions. The second technical point relates to separability, which is required by the referenced Strong Law of Large Numbers of Revesz (1968), but which does not hold for $C(-\infty,\infty)$. However, since the range of the i.i.d. random variables $X_1, X_2, \ldots$ (namely $\mathbb{R}$) is separable, condition (3.1) assures that the kernel estimators reside in a separable subspace of $C(-\infty,\infty)$.

To see the range of λ's to be considered, let β be a small positive constant and define

$$\underline{\lambda}_n = n^{-\beta} \quad , \quad \bar{\lambda}_n = n^{-\frac{1}{2}+\beta} \quad . \tag{3.2}$$

Note that for β sufficiently small,

$$\underline{\lambda} < n^{-1/5} < \overline{\lambda} ,$$

where the dependence of $\underline{\lambda}$ and $\overline{\lambda}$ on n is suppressed for notational convenience.

The positive results of this paper are:

Theorem 1: Under the above assumptions,

$$\sup_{\lambda \in [\underline{\lambda}, \overline{\lambda}]} \left| \hat{R}^{CV}(\lambda) - \int \log(f * K_\lambda) dF \right| \to 0 \qquad \text{a.s.}$$

Theorem 2: Under the above assumptions,

$$\sup_{\lambda \in [\underline{\lambda}, \overline{\lambda}]} \left| \hat{R}^{CV}(\lambda) - \hat{R}^{Jack}(\lambda) \right| \to 0 \qquad \text{a.s.} \quad .$$

The proofs of these theorems follow the general outline of that of Wong (1983) but are unfortunately quite technical in nature, in part because of the "uniform over λ" improvement. The details which establish the present version of (9) in Wong's argument are in the appendix.

Appendix: Proof of Theorems

Following Wong (1983), the key step is to prove

$$\text{(A.1)} \qquad \sup_{\substack{x\in[a,b]\\ \lambda\in[\underline{\lambda},\overline{\lambda}]}} \left| \frac{f_\lambda(x,\underset{\sim}{X})}{f^*K_\lambda(x)} - 1 \right| \to 0 \qquad \text{a.s.}$$

The consideration of the random variables in a separable subspace of $C(-\infty,\infty)$ with the sup norm neatly answers measurability questions concerning the sequence of random variables in (A.1). Consider only $n > n_0$ so that

$$f^*K_\lambda(x) \geq C_2 > 0$$

for all $x \in [a,b]$ and all $\lambda\in[\underline{\lambda},\overline{\lambda}]$. Choose partitions $x_1^n,\ldots,x_{Jn}^n \in [a,b]$ and $\lambda_1^n,\ldots,\lambda_{Ln}^n \in [\underline{\lambda},\overline{\lambda}]$ so that

$$x_1^n = a,\ x_{Jn}^n = b,\ \lambda_1^n = \underline{\lambda}_n,\ \lambda_{Ln}^n = \overline{\lambda}_n\ ,$$

and so that there is a constant C_3 such that

$$\text{(A.2)} \qquad Jn \leq C_3 n^{1+1/\alpha}\ , \qquad Ln \leq C_3 n^{2\alpha+1}\ ,$$

and so that, for $j = 2, \ldots, J_n$, $\ell = 2, \ldots, L_n$,

$$\text{(A.3)} \qquad \left|x_j^n - x_{j-1}^n\right| \leq n^{-(1+1/\alpha)},\quad \left|(\lambda_{\ell-1}^n)^{-1} - (\lambda_\ell^n)^{-1}\right| \leq n^{-1/2\alpha}\ ,$$

where α is the constant in (3.1) From here on, the dependence of the above quantities on n will be suppressed for notational convenience. Moreover, $C_1,\ldots,C_{12}$ will refer to constants (independent of n).

Expand (A.1) into

$$\text{(A.4)} \qquad \sup_{\substack{x\in[a,b]\\ \lambda\in[\underline{\lambda},\overline{\lambda}]}} \left| \frac{f_\lambda(x,\underset{\sim}{X})}{f^*K_\lambda(x)} - 1 \right| \leq \max_{\substack{j=1,\ldots,J\\ \ell=1,\ldots,L}} \left| \frac{f_{\lambda\ell}(x_j,\underset{\sim}{X})}{f^*K_{\lambda\ell}(x_j)} - 1 \right| +$$

$$+ \max_{\substack{j=2,\ldots,J \\ \ell=2,\ldots,L}} \sup_{\substack{x\in[x_{j-1},x_j] \\ \lambda\in[\lambda_{\ell-1},\lambda_\ell]}} \left| \frac{f_\lambda(x,\underset{\sim}{X})}{f^*K_\lambda(x)} - \frac{f_{\lambda\ell}(x_j,\underset{\sim}{X})}{f^*K_{\lambda\ell}(x_j)} \right| .$$

The second term on the right hand side of (A.4) can be bounded by writing, for $j=2, \ldots, J$; $\ell=2, \ldots, L$; $x\in[x_{j-1},x_j]$; $\lambda\in[\lambda_{\ell-1},\lambda_\ell]$

$$\left| \frac{f_\lambda(x,\underset{\sim}{X})}{f^*K_\lambda(x)} - \frac{f_{\lambda\ell}(x_j,\underset{\sim}{X})}{f^*K_{\lambda\ell}(x_j)} \right| \le n^{-1} \sum_{i=1}^{n} \left| \frac{K_\lambda(x-X_i)}{f^*K_\lambda(x)} - \frac{K_{\lambda\ell}(x_j-X_i)}{f^*K_{\lambda\ell}(x_j)} \right| .$$

But, for $i=1, \ldots, n$,

$$\text{(A.5)} \quad \left| \frac{K_\lambda(x-X_i)}{f^*K_\lambda(x)} - \frac{K_{\lambda\ell}(x_j-X_i)}{f^*K_{\lambda\ell}(x_j)} \right| \le |K_\lambda(x-X_i)| \left| \frac{f^*K_{\lambda\ell}(x_j)-f^*K_\lambda(x)}{f^*K_\lambda(x)f^*K_{\lambda\ell}(x_j)} \right| + \left| \frac{K_\lambda(x-X_i)-K_{\lambda\ell}(x_j-X_i)}{f^*K_{\lambda\ell}(x_j)} \right| .$$

But, since K is bounded and f is bounded away from 0 on [a,b],

$$|K_\lambda(x-X_i)| \left| \frac{f^*K_{\lambda\ell}(x_j)-f^*K_\lambda(x)}{f^*K_\lambda(x)f^*K_{\lambda\ell}(x_j)} \right| \le C_4\lambda^{-1} \left| f^*K_{\lambda\ell}(x_j)-f^*K_\lambda(x) \right| \le$$

$$\le C_4\lambda^{-1} \left| f^*(K_{\lambda\ell}(x_j)-K_{\lambda\ell}(x)) \right| + C_4\lambda^{-1} \left| f^*(K_{\lambda\ell}(x)-K_\lambda(x)) \right| .$$

And by integration by substitution, (3.1), (3.2) and (A.3)

$$\lambda^{-1} \left| f^*(K_{\lambda\ell}(x_j)-K_{\lambda\ell}(x)) \right| \le \lambda^{-1} \left| \int \left(\frac{1}{\lambda_\ell} K\left(\frac{y-x_j}{\lambda_\ell}\right) - \frac{1}{\lambda_\ell} K\left(\frac{y-x}{\lambda_\ell}\right) \right) f(y)\,dy \right| \le$$

$$\le \lambda^{-1} \int \left| K\left(u+\frac{x-x_j}{\lambda_\ell}\right) - K(u) \right| f(x+\lambda_\ell u)\,du \le$$

$$\le C_1 n^{\frac{1}{2}-\beta} |x-x_j|^\alpha \lambda_\ell^{-1-\alpha} \le C_5 n^{\frac{1}{2}-\beta} n^{-1-\alpha} n^{\frac{1}{2}+\alpha/2} = C_5 n^{-\beta-\alpha/2} .$$

In a similar spirit, using also the boundedness and compact support of f,

$$\lambda^{-1}\left|f^*(K_{\lambda\ell}(x)-K_\lambda(x))\right| = \lambda^{-1}\left|\int\left[\frac{1}{\lambda_\ell}K(\frac{y-x}{\lambda_\ell})-\frac{1}{\lambda}K(\frac{y-x}{\lambda})\right]f(y)dy\right| \leq$$

$$\leq \lambda^{-1}\int\left|K(u) - \frac{\lambda_\ell}{\lambda}K(\frac{\lambda_\ell}{\lambda}u)\right|f(x+\lambda_\ell u)du \leq$$

$$\leq \lambda^{-1}\left|1 - \frac{\lambda_\ell}{\lambda}\right|\int K(u)f(x+\lambda_\ell u)du + \frac{\lambda_\ell}{\lambda}\int\left|K(u)-K(\frac{\lambda_\ell}{\lambda}u)\right|f(x+\lambda_\ell u)du \leq$$

$$\leq C_6 n^{-1/2\alpha} + C_1\left|1 - \frac{\lambda_\ell}{\lambda}\right|^\alpha \leq C_7 n^{-1/2} .$$

It follows from the above that the first term on the right hand side of (A.5) tends to 0. To check the second term, note that by the boundedness of f above 0,

$$\left|\frac{K_\lambda(x-X_i)-K_{\lambda\ell}(x_j-X_i)}{f^*K_{\lambda\ell}(x_j)}\right| \leq C_8\left[\left|\frac{1}{\lambda} - \frac{1}{\lambda_\ell}\right|\left|K(\frac{x-X_i}{\lambda})\right| + \frac{1}{\lambda_\ell}\left|K(\frac{x-X_i}{\lambda})-K(\frac{x_j-X_i}{\lambda_\ell})\right|\right]$$

But, by the boundedness of K and (A.3),

$$\left|\lambda^{-1} - \lambda_\ell^{-1}\right|\left|K(\frac{x-X_i}{\lambda_\ell})\right| \leq C_4 n^{-1/2\alpha} .$$

And by the compactness of the support of f and (3.1),

$$\lambda_\ell^{-1}\left|K(\frac{x-X_i}{\lambda}) - K(\frac{x_j-X_i}{\lambda_\ell})\right| \leq C_9 n^{\frac{1}{2}-\beta}\left\{\left|\lambda^{-1} - \lambda_\ell^{-1}\right| + \lambda_\ell^{-1}\left|x-x_j\right|\right\}^\alpha \leq$$

$$\leq C_{10} n^{-\frac{1}{2}\beta}\ (n^{-\frac{1}{2}\alpha}\beta^\alpha = C_{10} n^{-\beta} .$$

It follows from the above that the second term on the right hand side of (A.4) tends to 0.

To check the almost sure convergence of the first term of (A.4),

for $\varepsilon > 0$ consider

$$P\left[\max_{j,\ell}\left|\frac{f_{\lambda\ell}(x_j,\underset{\sim}{X})}{f^*K_{\lambda\ell}(x_j)} - 1\right| > \varepsilon\right] \leq$$

(A.6)

$$\leq J_n \cdot L_n \cdot P[|f_{\lambda\ell}(x_j,\underset{\sim}{X}) - f^*K_{\lambda\ell}(x_j)| > \varepsilon\, C_2] \ .$$

Next pick q, an even positive integer, so that

(A.7) $$2 + \frac{1}{\alpha} + 2\alpha - \beta q < -1.$$

Then for each j and ℓ, using a Marcinkiewicz-Zygmund inequality (see, for example, Theorem 2 of Section 10.3 of Chow and Teicher (1978))

(A.8) $$P[|f_{\lambda\ell}(x_j,\underset{\sim}{X}) - f^*K_{\lambda\ell}(x_j)| > \varepsilon C_2] =$$

$$= P\left[\left|\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\lambda_\ell}(K(\frac{x_j-X_i}{\lambda_\ell}) - EK(\frac{x_j-X_i}{\lambda_\ell}))\right| > \varepsilon\, C_2\right] \leq$$

$$\leq E\left|\frac{i}{n\lambda_\ell}\sum_{i=1}^{n}(K(\frac{x_j-X_i}{\lambda_\ell}) - EK(\frac{x_j-X_i}{\lambda_\ell}))\right|^q / (\varepsilon C_2)^q \leq$$

$$\leq C_{11}(\varepsilon C_2)^{-q}(n\lambda_\ell)^{-q}E\left[(\sum_{i=1}^{n}(K(\frac{x_j-X_i}{\lambda_\ell}) - EK(\frac{x_j-X_i}{\lambda_\ell}))^2)^{q/2}\right] \leq$$

$$\leq C_{11}(\varepsilon C_2)^{-q}(n\lambda_\ell)^{-q}\, n^{q/2}\, E(K(\frac{x_j-X_1}{\lambda_\ell}) - EK(\frac{x_j-X_1}{\lambda_\ell}))^q \leq$$

$$\leq C_{12}\, n^{-q}\, n^{q/2}(\lambda_\ell)^{-q} \leq C_{12}(n\underline{\lambda}^2)^{-q/2}$$

for each n. From (A.6), (A.8), (A.2) and (3.2)

$$\sum_{n=1}^{\infty} P\left[\max_{j,\ell}\left|\frac{f_{\lambda\ell}(x_j,\underset{\sim}{X})}{f^*K_{\lambda\ell}(x_j)} - 1\right| > \varepsilon\right]$$

$$\leq n_0 + \sum_{n=n_0+1}^{\infty} C_3\, n^{1+\frac{1}{\alpha}}\, C_3 n^{2\alpha+1} C_{12}(n(n^{-\frac{1}{2}+\beta})^2)^{-q/2}$$

$$\leq n_0 + C_3^2 C_{12} \sum_{n=n_0+1}^{\infty} n^{2+\frac{1}{\alpha}+2\alpha-\beta q} < \infty,$$

by (A.7). Hence,

$$\max_{j,\ell}\left|\frac{f_{\lambda\ell}(x_j,\underset{\sim}{X})}{f^*K_{\lambda\ell}(x_j)} - 1\right|$$

converges completely and consequently a.s. to 0.

The rest of the proofs follow as in Wong (1983), except that the step

$$\sup_{\lambda} n^{-1} \sum_{i=1}^{n} \log \frac{f_\lambda(X_i, \underset{\sim}{X}_{(i)})}{f^*K_\lambda(X_i)} \to 0 \qquad \text{a.s.}$$

takes more work to verify. However, since f and f^*K_λ are bounded away from 0, this may be easily done using an argument which is similar to (but simpler than) the preceding arguments.

References

Chow, Y.S. and Teicher, H. (1978), Probability Theory, New York: Springer Verlag.

Gasser, T. and Müller, H.G. (1979), "Kernel estimation of regression functions", Smoothing Techniques for Curve Estimation, Lecture Notes in Mathematics, 757, 23-68.

Marron, J.S. (1984). "An asymptotically efficient solution to the bandwidth problem of kernel density estimation", North Carolina Institute of Statistics, Mimeo Series #1545.

Rosenblatt, M. (1956), "Remarks on some nonparametric estimates of a density function", Ann. Math. Statist., 27, 832-837.

Rosenblatt, M. (1971), "Curve estimates", Ann. Math. Statist. 42, 1815-1842.

Taylor, R.L. (1982), "Convergence of weighted sums of arrays of random elements in type p spaces with application to density estimation," Sankhyā, 44, 341-351.

Taylor, R.L. (1983), "Complete convergence for weighted sums of arrays of random elements", Internat. J. Math. and Math. Sci., 6, 69-79.

Wong, W.H. (1983), "A note on the modified likelihood for density estimation", J. Amer. Statist. Assoc., 78, 461-463.