THE INSTITUTE OF STATISTICS

THE CONSOLIDATED UNIVERSITY
OF NORTH CAROLINA

ON JACKKNIFING KERNEL REGRESSION FUNCTION ESTIMATORS

by

Wolfgang Härdle

Universität Heidelberg
Sonderforschungsbereich 123
Im Neuenheimer Feld 293
D-6900 Heidelberg

University of North Carolina
Department of Statistics
321 Phillips Hall 039A
Chapel Hill, NC 27514

Institute of Statistics Mimeo Series #1526

May 1983

DEPARTMENT OF STATISTICS
Chapel Hill, North Carolina

ON JACKKNIFING KERNEL REGRESSION FUNCTION ESTIMATORS*

Wolfgang Härdle

Universitat Heidelberg
Sonderforschungsbereich 123
Im Neuenheimer Feld 293
D-6900 Heidelberg

University of North Carolina
Department of Statistics
321 Phillips Hall 039A
Chapel Hill, North Carolina 27514

Abstract

Estimation of the value of a regression function at a point of continuity using a kernel-type estimator is discussed and improvements of the technique by a generalized jackknife estimator are presented. It is shown that the generalized jackknife technique produces estimators with faster bias rates. In a small example it is investigated, if the generalized jackknife method works for all choices of bandwidths. It turns out that an improper choice of this parameter may inflate the mean square error of the generalized jackknife estimator.

AMS Subject Classifications: Primary 62G05, Secondary 62J02.

Keywords and Phrases: Nonparametric regression function estimation, jackknife, bias reduction.

*Research partially supported by the Deutsche Forschungsgeneinschaft Sonderforschungsbereich 123 "Stochastische Mathematische Modelle."
Research also partially supported by the Air Force of Scientific Research Contract AFOSR-F49620 82 C 0009.

1. Introduction and Background

Let (x_1,Y_1), (x_2,Y_2),... be independent bivariate data following the mechanism

$$(1.1) \qquad Y_i = m(x_i) + \varepsilon_i \qquad i=1,2,\ldots$$

where x_i are fixed design points, the ε_i are zero mean random variables (rv) and $m(x)$ is the unknown regression function. Let us assume that n observations have been made with $0 \le x_1 \le \ldots \le x_n = 1$ and that Y_i is distributed as a rv with cumulative distribution function $F(y;\, x_i)$ and with probability density function

$$f(y;\, x_i) \in \mathcal{F} = \{f(y;\, x):\ 0 \le x \le 1\} .$$

The nonparametric regression function estimation problem is to estimate

$$m(x) = \int y\, f(y;\, x)\, dy$$

given n observations $(x_1,Y_1),\ldots,(x_n,Y_n)$.

Many estimators of $m(\cdot)$ in this "fixed design sampling " model (1.1) have been considered. (Priestley and Chao, 1972; Reinsch, 1967; Wold, 1974). Most of these authors assume a "i.i.d. error structure", i.e. $f(y;\, x) = f_0(y-m(x))$ for some fixed density f_0. We consider here kernel-type estimators

$$(1.2) \qquad \overline{m}_n(x) = \sum_{i=1}^{n} \alpha_i(x) Y_i \quad , \quad 0<x<1$$

where the sequence of (concentrating) weights $\alpha_i = \alpha_i^{(n)}$ are derived from a kernel function to be defined later. Nonparametric regression function estimators of this kind were introduced by Priestley and Chao (1972) and further discussed by Benedetti (1977). As can be seen from these early papers and more recently from Gasser and Müller (1979), estimators of the kind, defined in (1.2), are usually biased. Using a suggestion of Bartlett (1963) for the bias reduction of density estimators, Gasser and Müller gave an approximate expression for the bias and variance of such estimators (see Lemma 1).

The purpose of this paper is first to define a generalized jackknife estimator of m(x), as introduced by Schucany and Sommers (1977) for kernel-type density estimators. Then it is shown that the generalized jackknife, which is a method of forming linear combinations of kernel regression function estimators, reduces the bias of $\bar{m}_n(x)$. The generalized jackknife technique, applied to kernel regression function estimators, exhibits thus the same properties as in density estimation: Schucany and Sommers (1977) show that a jackknifed kernel estimator of a density f(x) reduces (asymptotically) bias in the way of Bartlett's (1963) original suggestion.

We will also show that for finite sample size in the regression setting considered here, the jackknifed estimator of $\bar{m}_n(x)$ may have a larger mean square error (MSE) than the original estimator when the weights $\alpha_i(x)$ are improperly chosen. A proper choice of this parameter is usually impossible in practical situations, so the experimenter is always confronted with the risk of selecting a "wrong" sequence of weights $\alpha_i^{(n)}$. Since the MSE is widely accepted as a criterion for measuring the accuracy of nonparametric estimators (Epanechnikov (1969), Rosenblatt (1971), among others) this (negative) result shows that in certain situations the generalized jackknife method, applied to kernel regression function estimators, may actually fail to improve the accuracy of the estimator.

We now present some of the choices of the weight functions $\alpha_i(\cdot)$ in (1.2). For instance, Priestley and Chao (1972) suggested to use

$$\alpha_i(x) = h^{-1}K((x-x_i)/h)[x_i-x_{i-1}] \tag{1.3}$$

with $x_0=0$, $h=h(n)$ a sequence of bandwidth tending to zero and a kernel function $K(\cdot)$, to be defined in (1.5). Gasser and Müller (1979) considered the following weights

$$\alpha_i(x) = h^{-1}\int_{s_{i-1}}^{s_i} K((x-u)/h)du \tag{1.4}$$

with a sequence of interpolating points $x_i \le s_i \le x_{i+1}$, $s_0=0$, $s_n=1$.

Cheng and Lin (1981) showed that with $x_i = s_i$ the weights defined through (1.4) or (1.3) give asymptotically the same consistency rate. We therefore restrict for convenience our attention to weights as defined in (1.4). We consider only even kernel functions $K(\cdot)$ which vanish outside $[-A,A]$ are continuous and satisfy for some integer r

$$\int_{-A}^{A} u^j K(u)du = 0 \qquad j=1,\ldots, r-1$$

(1.5)

$$\int_{-A}^{A} u^r K(u)du = r!\Lambda(K,r) < \infty .$$

These conditions on K are assumed to hold by some previous authors in the density estimation setting or in the regression function estimation case (Wegman, 1972a; Gasser and Müller, 1979). The assumption of finite support of K is not stringent. By reading through the proofs, it will be clear that the results also hold for kernels with infinite support. We are restricting ourselves on kernels with finite support only for computational convenience. In practical applications every kernel will be of finite support, due to lower bounds on machine precisions.

We further assume for the remainder of the paper that x_i are asymptotically equispaced: $\sup_{1\le j\le n} |s_j - s_{j-1}| = 0(n^{-1})$. We then have from Gasser and Müller (1979) the following result on the mean square error.

Lemma 1

Let $m^{(p)}(\cdot)$ be uniformly bounded with p=2t, t>0 and let K be a kernel function satisfying (1.5) with $r \le p$. Assume that $\sigma^2(x) = \int [y-m(x)]^2 f(y;\, x)dy$ is uniformly continuous for 0<x<1, then the leading term of the MSE of $\bar{m}_n(x)$, 0<x<1 is

$$(nh)^{-1}\beta_K\sigma^2(x) + [\sum_{s=1}^{t} h^{2s} m^{(2s)}\Lambda(K,2s)]^2 \quad ,$$

where $\beta_k = \int K^2(u)du$.

A similar MSE decomposition in variance and bias parts holds for density estimators $f_n(x) = (nh)^{-1}\sum_{i=1}^n K((x-X_i)/h)$ (Parzen, 1962) but with derivatives of $f(\cdot)$, the density, in the bias instead of derivatives of $m(\cdot)$. The first part of the following section is therefore quite analogous to Schucany and Sommers (1977).

2. Does the jackknifed estimate improve the kernel estimate?

We shall construct here combinations of kernel estimators, using the generalized jackknife method of Schucany and Sommers (1977). Note that, in the context of the generalized jackknife, the "leave-out" techniques, subscribed to the ordinary jackknife, will not be employed. We first state a lemma giving the convergence rate of the bias of the jackknifed estimator and discuss furthermore an example for which the jackknife estimator fails to improve the MSE of the original kernel estimate.

Define for $\ell=1,2$ $\bar{m}_{\ell,k}(x) = \sum_{i=1}^n \alpha_{\ell,i}(x)Y_i$ where

$$\alpha_{\ell,i}(x) = h_\ell^{-1} \int_{s_{i-1}}^{s_i} K_\ell((t-u)/h_\ell)du \quad , \quad h_1=h_1(n),\ h_2=h_2(n)$$

and h_ℓ denote different sequences of bandwidths and K_1, K_2 kernel functions with $r_1=r_2=2$ respectively. The generalized jackknife estimate of $\bar{m}_n(x)$ is then defined as

$$G[\bar{m}_{1,n}(x),\ \bar{m}_{2,n}(x)] = (1-R)^{-1}[\bar{m}_{1,n}(x)-R\bar{m}_{2,n}(x)] \quad , \quad R\neq 1 . \tag{2.1}$$

It is assumed for the remainder of the paper that $\sigma^2(x)$, as defined in Lemma 1, is uniformly continuous for $0<x<1$. The proof of the following lemma is evident in view of Lemma 1.

Lemma 2

Suppose that $m^{(p)}(\cdot)$ is uniformly bounded with $p=2t$, $t\geq 2$ and let K_1, K_2 be kernels satisfying (1.5) with $r_1, r_2 \leq p$. Then the leading bias term of $G[\bar{m}_{1,n}(x), \bar{m}_{2,n}(x)]$ is

$$(2.2) \qquad (1-R)^{-1} \sum_{s=1}^{t} [h_1^{2s}\Lambda(K_1,2s) - Rh_2^{2s}\Lambda(K_2,2s)]m^{(2s)}(x) .$$

The reduction of bias in (2.2) is now made possible by a suitable choice of the balancing constant R.

If we set $R=R_n=(h_1^2/h_2^2)\Lambda(K_1,2)/\Lambda(K_2,2)$,

then

$$h_1^2\Lambda(K_1,2) - Rh_2^2\Lambda(K_2,2) = 0 ,$$

and the first bias term containing $m^{(2)}(x)$, is eliminated. We thus have indeed an estimator with a faster bias rate and moreover we could have produced $G[\bar{m}_{1,n}(x), \bar{m}_{2,n}(x)]$ with the single kernel

$$K^*(u) = [K_1(u) - \nu c_n^3 K_2(c_n u)]/[1 - \nu c_n^2]$$

where $\nu = \Lambda(K_1,2)/\Lambda(K_2,2)$ and $c=c_n=h_1(n)/h_2(n)$. Note that, in contrast to K_1,K_2, the kernel $K^*(\cdot)$ may still depend on n, but satisfies (1.5) for all n with r=4 as is shown in Schucany and Sommers (1977), p. 421. Note also that the calculations in that paper, showing that K^* belongs to the class of kernel functions (see (1.5)) with r=4, do not depend on the density estimation setting.

As is shown in Lemma 1 and Lemma 2, the bias terms of both $\bar{m}_n$ and $G[\bar{m}_{1n},\bar{m}_{2n}]$ still depend on $m(\cdot)$ through the derivatives of the regression function. An optimal choice of h with respect to the MSE would thus involve the knowledge of the derivatives of the regression function. A conservative strategy of the experimenter could therefore be to subscribe a small amount of smoothness to $m(\cdot)$. He could assume, for instance, that the second derivative of $m(\cdot)$ exists and is continuous. On the other hand, he could do better if even the fourth derivative exists by choosing a kernel function K satisfying (1.5) with r=4. This justifies the use of a generalized jackknife kernel estimator, as defined in (2.1). If, in fact, m(•) is smoother than we expected, let's say we started with a kernel K with r=2, the generalized jackknife estimator would give us the faster vanishing bias term with kernel functions K_1 and K_2 with r=2 in the class defined in (1.5).

We now investigate the properties of $G[\bar{m}_{1n},\bar{m}_{2n}]$ in a small example. For this define $K_1=K_2=K_E$, where

$$K_E(u) = 3/4(1-u^2) \qquad |u|\le 1$$
$$= 0 \qquad |u|>1$$

is the Epanechnikov (1969) kernel. For optimality questions of this particular kernel, we refer to Rosenblatt (1971). This kernel function obviously satisfies (1.5) with r=2. The following calculations for the variance remain valid also in the density estimation setting, since β_K or β_{K*} occur also as variance factors there (Parzen, 1962). By straightforward computations it is easy to obtain:

$$\beta_{K_E} = \int K_E^2(u)\,du = 3/5$$

$$\Lambda(K_E,2) = 1/10$$

$$\Lambda(K_E,4) = 1/280 .$$

(2.3) $$\beta_{K*} = \int K^{*2}(u)\,du$$
$$= [1-c^2]^{-2}\{\int K_E^2(u)d + c^5\int K_E^2(u)\,du - 2c^3\int K_E(u)K_E(cu)\,du\}$$
$$= [1-c^2]^{-2}\{3/5 + 9/10\, c^5 - 3/2\, c^3\}$$
$$= 9/10[c^3 + 2c^2 + (4/3)c + 2/3]/[c+1]^2 .$$

Note that $\beta_{K*} \approx 9/8$ as $c \approx 1$ which is considerably higher than $\beta_{K_E} = 3/5$. This behavior of K* can also be drawn from table 1 in Schucany and Sommers (1977) for a normal density kernel and R=.99. It is therefore apparent that some caution must be exercised in selecting c, which is the same as choosing the balancing factor R or h_1 and h_2. To compensate on the trade-off between bias and variance, the faster rate of the bias of G_n suggests to choose $h_1 > h$. The calculation of the variance factors β_{K_E} and β_{K*} in (2.3) suggests the choice of $h_1 \approx [\beta_{K*}/\beta_{K_E}]h \approx 15h/8$ to balance the variance part of the MSE. How does this choice of bandwidths now affect the MSE of $\bar{m}_n(x)$ and $G[\bar{m}_{1,n}(x), \bar{m}_{2,n}(x)]$, given that $m^{(4)}(x)$ exists and is uniformly bounded. The leading term of the MSE of $\bar{m}_n(x)$ is given by

$$(2.4) \qquad (nh)^{-1}\, \beta_{K_E} + \{h^2 m^{(2)}(x)/10 + h^4 m^{(4)}(x)/280\}^2 ,$$

whereas the principal term of the MSE of $G[\bar{m}_{1,n}(x), \bar{m}_{2,n}(x)]$ is

$$(2.5) \qquad (nh_1)^{-1}\, \beta_{K*} + \{-c^{-2} h_1^4 m^{(4)}(x)/280\}^2 .$$

Assume now that $\alpha = m^{(2)}(x)/10$ and $\beta = m^{(4)}(x)/280$ are positive and $c \approx 1$, and assume in addition that by choosing $h_1 \approx 15h/8$, the variance parts of (2.4) and (2.5) are approximately the same. With this selection of h_1, the bias terms now read

$$\mathrm{bias}^2(\bar{m}_n(x)) = \{h^2\alpha + h^4\beta\}^2$$

$$\mathrm{bias}^2(G[\bar{m}_{1n}, \bar{m}_{2n}]) \approx \{-50625\beta h^4/4096\}^2$$

$$\approx 152.76\beta^2 h^8 .$$

Comparing now these bias terms, only depending on h now, shows that with the "wrong" choice of h_1 and h_2, the generalized jackknife estimator may fail to improve $\bar{m}_n(x)$ in MSE. More precise computations yield that, if h is chosen to fulfill

$$|(\beta/\alpha)h^2 - .00658| > .0814 ,$$

then

$$\mathrm{MSE}\{G[\bar{m}_{1n}, \bar{m}_{2n}]\} \gtrapprox \mathrm{MSE}\{\bar{m}_n(x)\} .$$

Some additional remarks should be made. The example seems to be somewhat artificially constructed since we restrict our attention to one particular kernel function. This is due to the computations of Rosenblatt (1971), Table 1, p. 1821, showing the relative insensitivity of the MSE to different kernels. (See also Wegm n (1972b)). Interesting is the fact that, if $m^{(2)}(x_0)$ happens to be zero for some x_0, the new jackknife estimator drastically loses MSE accuracy, provided that h_1 was chosen in such a way, that the variance parts of both $\bar{m}_n(x)$ and $G[\bar{m}_{1n},\bar{m}_{2n}]$ are approximately equal ($c \approx 1$). A proper choice of R and h_1 is in practice not obtainable, since we have no knowledge about the derivatives of the regression function. It is also not possible in general to compute the regions of bandwidths, where $G[\bar{m}_{1n},\bar{m}_{2n}]$ actually improves $\bar{m}_n(x)$, since these require the constants α and β.

3. Conclusion

Under the proper conditions, several of the original type of kernel regression estimators proposed by Priestley and Chao (1972) can be combined to form generalized jackknife estimators which have a faster bias rate. The new estimators produce an improved rate of the MSE due to cancellations of bias terms. The generalized jackknife estimator, as defined here for the regression function estimation setting, achieves thus the same properties as a similar estimator, introduced by Schucany and Sommers for density function estimation. In a small example, it is investigated how the new estimator performs when compared to ordinary kernel regression function estimates. It is shown there, than an improper choice of R, the balancing factor between $\bar{m}_{1n}$ and $\bar{m}_{2n}$ may introduce an inflation effect on the MSE of $G[\bar{m}_{1n},\bar{m}_{2n}]$. A jackknifing technique of kernel estimators of regression functions should therefore be cautiously performed with a proper inspection of the involved parameters.

REFERENCES

Bartlett, M.S. (1963). Statistical estimation of density functions, Sankhyā 25A, 145-54.

Benedetti, J.K. (1977). On the nonparametric estimation of regression functions. J. Roy. Statist. Soc. B 39, 248-253.

Cheng, K.F. and Lin, P. (1981). Nonparametric estimation of a regression function. Zeit. Wahrsch. verw. Geb. 57, 223-233.

Epanechnikov, V.A. (1969). Nonparametric estimation of a multivariate probability density. Theor. Prob. Appl. 14, 153-158.

Gasser, T. and Müller, H.G. (1979). Kernel estimation of regression functions in "Smoothing Techniques for Curve Estimation" ed. T. Gasser and M. Rosenblatt. Lecture Notes 757, Springer Verlag, Heidelberg.

Parzen, E. (1962). On estimation of a probability density function and its mode. Ann. Math. Stat. 33, 1065-1076.

Priestley, M.B. and Chao, M.T. (1972). Non-parametric function fitting. J. Roy. Statist. Soc. B 34, 385-392.

Reinsch, C. (1967). Smoothing by spline functions, Numer. Math. 10, 177-183.

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

Schucany, W.R. and Sommers, J.P. (1977). Improvement of kernel type density estimators. JASA 72, 420-423.

Wegman, E.J. (1972a). Nonparametric probability density estimation: I. A summary of available methods. Technometrics 14, 533-546.

Wegman, E.J. (1972b). Nonparametric probability density estimation: II. A comparison of density estimation methods. J. Statist. Computation and simulation 1, 225-245.

Wold, S. (1974). Spline functions in data analysis. Technometrics 16, 1-11.