Biometrika (1987), **74**, 4, *pp.* 743–9
Printed in Great Britain

Bandwidth choice and confidence intervals for derivatives of noisy data

BY HANS-GEORG MÜLLER
Institut für Medizinisch-Biologische Statistik, Universität Marburg, Ernst-Giller-Strasse 20, D-3550 Marburg, Federal Republic of Germany

U. STADTMÜLLER
Abteilung für Mathematik I, Universität Ulm, Oberer Eselsberg, D-7900 Ulm, Federal Republic of Germany

AND THOMAS SCHMITT
Institut für Medizinisch-Biologische Statistik, Universität Marburg, Ernst-Giller-Strasse 20, D-3550 Marburg, Federal Republic of Germany

SUMMARY

We propose a method for automatic bandwidth selection for kernel estimators of derivatives of a regression function. The finite sample behaviour of this new method is compared with that of other methods in a Monte Carlo Study. The automatic estimation of derivatives can be employed for the construction of asymptotic local confidence intervals for the nonparametric estimate of the regression function and its first derivatives.

Some key words: Bandwidth selection; Cross-validation; Difference quotient; Estimation of derivatives; Factor method; Kernel estimator.

1. INTRODUCTION

Nonparametric regression and curve fitting methods for the estimation of derivatives of regression functions are useful when the dynamics underlying a measured time course are of interest. If these dynamics are unknown, as is often the case in biomedical applications, usually there is no satisfactory parametric model available which could be fitted to the data. There are several curve-fitting methods available for the nonparametric estimation of a regression function and its derivatives. All of them require the choice of a smoothing parameter that balances the trade-off between bias and variance of the curve estimate. The choice of this parameter is crucial to the quality of the estimated curve, and even more so for derivatives of the curve. Since the theoretically optimal smoothing parameter with respect to some error criterion depends on the unknown variance of the noise and the regression function itself, the choice of the smoothing parameter constitutes a difficult problem in practice.

To be more precise, we consider the model:

$$y_i = g(x_i) + \varepsilon_i \quad (1 \leqslant i \leqslant n), \tag{1}$$

where the known x_i, for $0 \leqslant x_1 \leqslant \ldots \leqslant x_n = 1$, represent time or other points of measurements y_i for the unknown regression function g, which is assumed to be 'smooth'. For simplicity we suppose that $x_i = i/n$ and g is infinitely differentiable on $[0, 1]$. The

measurements y_i are contaminated with noise variables ε_i, which are assumed to be independently and identically distributed with mean 0 and variance $\sigma^2 < \infty$. Our aim is to estimate the derivatives $g^{(\nu)}$ $(\nu = 0, 1, \ldots, m)$ for some positive integer m, by means of kernel estimators, for which the smoothing parameter is the so-called bandwidth. For details see § 2.

In the case $\nu = 0$, two well-known methods to estimate the bandwidth are cross-validation (Wahba & Wold, 1975) and a criterion suggested by Rice (1984a). The idea behind these methods is to minimize a data-dependent quantity which is asymptotically equivalent to the integrated mean squared error. In § 3·1, we propose a generalization of cross-validation for the estimation of derivatives, replacing the data y_i by appropriate difference quotients, in analogy of the generalization of the criterion by Rice (1984a), given by Rice (1986). Since difference quotients have large variances these methods may not be very stable. In § 3·2, we suggest a third method which uses the optimal bandwidths for the case $\nu = 0$, and a factor, depending on the kernels used, connecting the asymptotically optimal bandwidths for the case $\nu = 0$ with those for $\nu > 0$. The corresponding asymptotic considerations are summarized in § 2. It should be noted that this factor method can also be employed for bandwidth choice for kernel estimates of derivatives of density and spectral density functions.

Numerical computations and simulations were used to compare the three approaches. The results are summarized in § 4. As an application of the estimation of $g^{(\nu)}$, we consider in § 5 a possible construction of asymptotic local confidence intervals. Since boundary effects and their treatment were discussed by Rice (1984b) and Gasser & Müller (1984), we avoid a technical discussion of them, considering curves and confidence bands in a closed subinterval $I \subseteq (0, 1)$ only.

2. DIFFERENTIATION WITH THE KERNEL METHOD

Kernel estimators in nonparametric regression were introduced by Priestley & Chao (1972). We consider a modified version proposed by Gasser & Müller (1984). Define $s_0, \ldots, s_n$ by $s_0 = 0$, $s_n = 1$, $s_i = \frac{1}{2}(x_i + x_{i+1})$ $(i = 1, \ldots, n-1)$. Then the kernel estimate for $g^{(\nu)}(x)$, $x \in I$, is

$$\hat{g}^{(\nu)}(x) = \sum_{j=1}^{n} \omega_j^{(\nu)}(x) y_j, \tag{2}$$

where

$$\omega_j^{(\nu)}(x) = \frac{1}{b^{\nu+1}} \int_{s_{j-1}}^{s_j} K_\nu^{(\nu)}\left(\frac{x-v}{b}\right) dv,$$

where $b = b(n)$ is the smoothing parameter, called the bandwidth, and K_ν is a kernel function with support $[-1, 1]$. The kernel K_ν is assumed to be $(\nu-1)$-times continuously differentiable on $\mathbb{R}$ and infinitely often on its support and to satisfy the following moment conditions for some integer $k > \nu$ to be fixed in advance:

$$\int_{-1}^{1} K_\nu(v) v^j dv = \begin{cases} 1 & (j = 0), \\ 0 & (0 < j < k - \nu), \\ \neq 0 & (j = k - \nu). \end{cases} \tag{3}$$

Hence the kernel is of order $k - \nu$. Define

$$V_{k,\nu} = \int_{-1}^{1} K_\nu^{(\nu)}(v)^2 \, dv, \quad B_{k,\nu} = (-1)^k \int_{-1}^{1} K_\nu^{(\nu)}(v) v^k \, dv / k!,$$

and assume that $b = b(n)$ satisfies $b \to 0$, $nb \to 0$, as $n \to \infty$.

A routine argument (Gasser & Müller, 1984) shows that, as $n \to \infty$,

$$\text{var}\{\hat{g}^{(\nu)}(x)\} = (nb^{2\nu+1})^{-1}\{\sigma^2 V_{k,\nu} + o(1)\}, \tag{4}$$

$$E\{\hat{g}^{(\nu)}(x) - g^{(\nu)}(x)\} = b^{k-\nu} g^{(k)}(x) B_{k,\nu} + o(b^{k-\nu}). \tag{5}$$

Minimizing the leading terms of the integrated mean squared error over I, derived from (4) and (5), gives the asymptotically optimal bandwidth as

$$b^*_{\nu,k} = \left\{ \frac{2\nu+1}{2(k-\nu)} \frac{\sigma^2 V_{k,\nu} |I|}{B^2_{k,\nu} \int g^{(k)}(v)^2 \, dv} \frac{1}{n} \right\}^{1/(2k+1)}, \tag{6}$$

where the integral is over the range I of length $|I|$. So far, formula (6) is not helpful, since it contains the unknown quantities σ^2 and $\int g^{(k)}(v)^2 \, dv$.

By applying the Lindeberg condition, it is shown by Gasser & Müller (1984, Th. 3) that under the present assumptions the limiting distribution of

$$(nb^{2\nu+1})^{\frac{1}{2}}\{\hat{g}^{(\nu)}(x) - g^{(\nu)}(x)\} \tag{7}$$

is normal for any $x \in I$. In particular if $nb^{2k+1} \to d$, as $n \to \infty$, with some constant $d \geqslant 0$, we conclude from (4) and (5) that as $n \to \infty$:

$$(nb^{2\nu+1})^{\frac{1}{2}}\{\hat{g}^{(\nu)}(x) - g^{(\nu)}(x)\} \to N\{d^{\frac{1}{2}} g^{(k)}(x) B_{k,\nu}, \sigma^2 V_{k,\nu}\} \tag{8}$$

in distribution.

3. Bandwidth choice for derivatives

3·1. *Cross-validation and the Rice criterion*

Methods for bandwidth choice for kernel estimators of the regression function g itself are reviewed by Rice (1983) and Härdle & Marron (1985). In the following we assume $I = [0, 1]$ for simplicity. We define x- and y-coordinates of difference quotient operators iteratively by $x_i^{(0)} = x_i$, $\Delta_i^{(0)} = y_i$, for $i = 1, \ldots, n$, and

$$x_i^{(\nu)} = \tfrac{1}{2}(x_{i+1}^{(\nu-1)} + x_i^{(\nu-1)}), \quad \Delta_i^{(\nu)} = \frac{\Delta_{i+1}^{(\nu-1)} - \Delta_i^{(\nu-1)}}{x_{i+1}^{(\nu-1)} - x_i^{(\nu-1)}} \quad (i = 1, \ldots, n-\nu;\ \nu \geqslant 1).$$

We propose to generalize cross-validation for the estimation of the νth derivative ($\nu \geqslant 0$) of the regression function by selecting the bandwidth b which minimizes

$$\text{CV}^{(\nu)}(b) = \frac{1}{n-\nu} \sum_{j=1}^{n-\nu} [\Delta_j^{(\nu)} - \hat{g}^{(\nu)}_{-(j,j+\nu)}\{x_j^{(\nu)}\}]^2, \tag{9}$$

where $\hat{g}^{(\nu)}_{-(j,j+\nu)}\{x_j^{(\nu)}\}$ is the leave-$(\nu+1)$-out kernel estimator of type (2) based on the data $(x_1, y_1), \ldots, (x_{j-1}, y_{j-1}), (x_{j+\nu+1}, y_{j+\nu+1}), \ldots, (x_n, y_n)$.

With

$$\text{RSS}^{(\nu)}(b) = \frac{1}{n-\nu} \sum_{j=1}^{n-\nu} [\Delta_j^{(\nu)} - \hat{g}^{(\nu)}\{x_j^{(\nu)}\}]^2,$$

$$R^{(\nu)}(b) = \frac{1}{n} \sum_{j=1}^{n} E\{\hat{g}^{(\nu)}(x_j) - g^{(\nu)}(x_j)\}^2,$$

Rice (1986) proposed as a bandwidth estimate for derivatives $g^{(\nu)}$ the minimizer of

$$\hat{R}^{(\nu)}(b) = \text{RSS}^{(\nu)}(b) + (-1)^{\nu} 2K_{\nu}^{(2\nu)}(0)\hat{\sigma}^2/(nb^{2\nu+1}) - \binom{2\nu}{\nu}\hat{\sigma}^2, \tag{10}$$

where

$$\hat{\sigma}^2 = \frac{1}{2(n-1)} \sum_{j=1}^{n-1} (y_{j+1} - y_j)^2.$$

The methods (9) and (10) are motivated by the asymptotic equivalences

$$E\{\hat{R}^{(\nu)}(b)\} \doteq \text{IMSE}(b), \quad E\{\text{CV}^{(\nu)}(b)\} \doteq \text{IMSE}(b) + \binom{2\nu}{\nu}\sigma^2,$$

where IMSE(b) denotes the integrated mean squared error as a function of the bandwidth b.

3·2. *Factor method*

Stochastic difference quotients have been used in applications, for example by James & Conyers (1985), but are unstable since $\text{var}\{\Delta_i^{(\nu)}\} \doteq n^{2\nu}$ if $x_{i+1} - x_i = 1/n$. Therefore an alternative to the methods (9), (10) may be useful. We suggest the so-called factor method, which requires only a good method of bandwidth selection for $\nu = 0$ and an appropriate choice of kernels. Observe that if we choose a kernel K_0 of order k and a kernel K_ν of order $k - \nu$, with the same integer k, formula (6) gives the following connection between the asymptotically optimal bandwidths for $\hat{g}$ and $\hat{g}^{(\nu)}$:

$$\frac{b_{\nu,k}^*}{b_{0,k}^*} = \left\{ \frac{(2\nu+1)k}{k-\nu} \frac{V_{k,\nu} B_{k,0}^2}{V_{k,0} B_{k,\nu}^2} \right\}^{1/(2k+1)} = d_{\nu,k}, \tag{11}$$

say, and this constant only depends on the kernels K_0 and K_ν, but not on g and σ^2. Hence an estimate for the optimal bandwidth for $\hat{g}^{(\nu)}$ is given by

$$\hat{b}_{\nu,k} = \hat{b}_{0,k} d_{\nu,k}, \tag{12}$$

where $\hat{b}_{0,k}$ is the minimizer of (9) or (10) for $\nu = 0$ or a related criterion.

Formula (12) is partly of an asymptotic nature and partly data dependent. Obviously consistency and the rate of convergence of $\hat{b}_{0,k}$ will be inherited by $\hat{b}_{\nu,k}$. Under the assumptions of Theorem 3 of Rice (1984a) we obtain $(\hat{b}_{\nu,k}/b_{\nu,k}^* - 1) = O_p\{(b_{\nu,k}^*)^{\frac{1}{2}}\}$.

For good quality estimates it is important to use symmetric kernels $K_\nu^{(\nu)}$ if ν is even, and antisymmetric kernels if ν is odd. A certain class of optimal kernels for ν and k both even or both odd was discussed by Müller (1984). Since the factor method requires also kernels for $\nu = 0$ with odd k we constructed kernels for this case, too. All these kernels are polynomials, have bounded support $[-1, 1]$, satisfy the moment conditions (3), are continuous but not differentiable at the endpoints, and are smooth elsewhere. In addition they showed good finite sample properties. The kernels used are listed by Müller (1984), see the case $\mu = 1$, or are given by:

$$K(x) = \tfrac{15}{32}(3 - 3x - 10x^2 + 10x^3 + 7x^4 - 7x^5) \quad (\nu = 0, k = 3),$$

$$K(x) = \tfrac{35}{256}(15 - 105x^2 + 189x^4 - 99x^6) - 0{\cdot}2099(-5x + 35x^3 - 63x^5 + 33x^7) \quad (\nu = 0, k = 5).$$

The corresponding factors $d_{\nu,k}$ for these kernels are given in Table 1. The factor method yields consistent estimators of the asymptotically optimal bandwidth if $\hat{b}_{0,k}$ in (12) is chosen by criteria (9) or (10), applied for $\nu = 0$ respectively.

Table 1. *Factors $d_{\nu,k}$ for bandwidth choice for the νth derivative* (11)

Kernel $K_\nu^{(\nu)}$ for estimation of νth derivative		Corresponding kernel K_0		Factor $d_{\nu,k}$ (11)
ν	k	ν	k	
1	3	0	3	0·7083
1	5	0	5	0·7396
2	4	0	4	0·8919
2	6	0	6	0·9507
3	5	0	5	0·6788

4. Monte Carlo study

The methods of bandwidth choice for derivatives discussed so far, have been motivated by asymptotic considerations. To compare their finite sample behaviour, we carried out a simulation study for derivatives up to the third. We chose the interior interval $I = [0{\cdot}25, 0{\cdot}75]$ and used the kernels $K_\nu^{(\nu)}$, where $K^{(\nu)}$ of order 2 from § 3·2. The following procedures were investigated, the abbreviations in brackets identifying the respective method will be used in the following.

Method 1 (CV): Minimization of $\mathrm{CV}^{(\nu)}(b)$ using (9).

Method 2 (R): Minimization of $\hat{R}^{(\nu)}(b)$ using (10).

Method 3 (FAC-CV): Formula (12), choosing $\hat{b}_{0,k}$ as minimizer of (9) for $\nu = 0$.

Method 4 (FAC-R): Formula (12), choosing $\hat{b}_{0,k}$ as minimizer of (10) for $\nu = 0$.

Method 5 (CV−0): Minimization of (9) for $\nu = 0$.

Method 5 was used earlier lacking a better approach; the idea behind it is that the derivatives of a good curve estimate will be a good estimate of the derivative. From (6) we see that this belief is asymptotically wrong; nevertheless the finite sample behaviour of this procedure may still be interesting.

For various curves, sample sizes and residual variances the following observations on the averages obtained from 100 Monte Carlo runs per case were made. For $\nu = 0$, the differences between CV and R are only minor in terms of average integrated mean squared error and tolerance regions, i.e. percentage of cases where the integrated mean squared error is above $(1+\alpha)$ times the finite optimum for various $\alpha > 0$, although R tends to choose smaller bandwidths. Bad performance with respect to tolerance regions can be due to either a wrong average bandwidth with a small variance or to a large variance of the chosen bandwidths. The first cause seems to apply to CV-0 which always is too small on the average, has very bad tolerance region behaviour especially for $\nu = 2, 3$, and is therefore obsolete, even though the average integrated mean squared error is sometimes better than for CV or FAC-CV. The second cause seems to apply to CV which typically chooses bandwidths with the highest variance among all methods considered and shows therefore inferior tolerance regions. On the other hand, R suffers from the disadvantage that the average bandwidths usually are too small which leads to inferior average integrated mean squared errors. In part this might be due to some cases where we observed severe underestimation of σ by $\hat{\sigma}$. In nearly all cases, FAC-CV and FAC-R chose bandwidths closest to the optimal one with the smallest variances.

Summarizing, the following ranking of the methods emerges. In the present evaluation scheme, FAC-R is the best method. If tolerance regions are valued more than average integrated mean squared error, FAC-CV is second best; otherwise R; CV and CV-0 showed the worst performance.

5. LOCAL CONFIDENCE INTERVALS

Confidence bounds for curve estimates are often requested in order to assess the reliability of the estimated curves. Global confidence bands have been discussed by Knafl, Sacks & Ylvisaker (1985) and Stadtmüller (1986); a Bayesian approach was developed by Wahba (1983) and Silverman (1985).

We present a simple approach for the construction of local asymptotic confidence intervals employing the estimation of derivatives for the assessment of bias. Choice of optimal bandwidths implies that bias squared and variance of the local limiting distribution are of the same order of magnitude. Since the exact bias is hard to assess, already Clark (1980) proposed to base confidence intervals on the distribution $N\{0, \hat{S}\}$, where $\hat{S}$ is a consistent estimate of the local mean squared error, S, of the estimator $\hat{g}^{(\nu)}(x)$ in the sense that $\hat{S}/S \to 1$ in probability, as $n \to \infty$. By (8), $1-\alpha$ confidence intervals derived from this distribution will be conservative whenever, denoting the bias and variance of $\hat{g}^{(\nu)}(x)$ by β and v,

$$\Phi^{-1}(1-\tfrac{1}{2}\alpha) \geqslant \beta v^{-\frac{1}{2}}/\{(1+\beta^2 v^{-1})^{\frac{1}{2}}-1\},$$

where Φ^{-1} denotes the inverse of the normal distribution function.

As a consistent estimator of the mean squared error we use

$$\hat{S} = \hat{\sigma}^2 \sum_{i=1}^{n} \left(\int_{s_{i-1}}^{s_i} b^{-\nu-1} K_\nu^{(\nu)}\left(\frac{x-v}{b}\right) dv \right)^2 + \{b^{k-\nu}\hat{g}^{(k)}(x) B_{k,\nu}\}^2,$$

where $\hat{g}^{(k)}$ is estimated using the factor method for bandwidth selection for the kth derivative. The confidence bounds become

$$\hat{g}^{(\nu)}(x) \pm \Phi^{-1}(1-\tfrac{1}{2}\alpha)\hat{S}^{\frac{1}{2}}.$$

In an application to human movement data they were seen to be quite reasonable for the kernel estimates of the regression function and its first derivative.

ACKNOWLEDGEMENTS

The methods were tried out on data which were kindly provided by Dr Virgil Stokes, Karolinska Institute, Stockholm. We thank A. Reifenschneider and Chr. Lohrengel for help with the computations. This research was supported by Deutsche Forschungsgemeinschaft.

REFERENCES

CLARK, R. M. (1980). Calibration, cross-validation and Carbon-14, II. *J. R. Statist. Soc.* A **143**, 177–94.

GASSER, T. & MÜLLER, H. G. (1984). Estimating regression functions and their derivatives by the kernel method. *Scand. J. Statist.* **11**, 171–85.

HÄRDLE, W. & MARRON, J. (1985). Asymptotic nonequivalence of some bandwidth selectors in nonparametric regression. *Biometrika* **72**, 481–4.

JAMES, A. T. & CONYERS, R. A. J. (1985). Estimation of a derivative by a difference quotient: Its application to hepatocyte lactate metabolism. *Biometrics* **41**, 467–76.

KNAFL, G., SACKS, J. & YLVISAKER, D. (1985). Confidence bands for regression functions. *J. Am. Statist. Assoc.* **80**, 683–91.

MÜLLER, H. G. (1984). Smooth optimum kernel estimators of regression curves, densities and modes. *Ann. Statist.* **12**, 766–74.

PRIESTLEY, M. B. & CHAO, M. T. (1972). Nonparametric function fitting. *J. R. Statist. Soc.* B **34**, 385-92.
RICE, J. A. (1983). Methods for bandwidth choice in nonparametric kernel regression. In *Computer Science and Statistics: The Interface*, Ed. J. E. Gentle, pp. 186-90. Amsterdam: North-Holland.
RICE, J. A. (1984a). Bandwidth choice for nonparametric kernel regression. *Ann. Statist.* **12**, 1215-30.
RICE, J. A. (1984b). Boundary modification for kernel regression. *Comm. Statist.* A **13**, 893-900.
RICE, J. A. (1986). Bandwidth choice for differentiation. *J. Mult. Anal.* **19**, 251-64.
SILVERMAN, B. W. (1985). Some aspects of the spline-smoothing approach. *J. R. Statist. Soc.* B **47**, 1-21.
STADTMÜLLER, U. (1986). Asymptotic properties of curve estimates. *Period. Math. Hung.* **17**, 83-108.
WAHBA, G. (1983). Bayesian confidence intervals for the cross-validated smoothing spline. *J. R. Statist. Soc.* B **45**, 133-50.
WAHBA, G. & WOLD, S. (1975). A completely automatic French curve: fitting spline functions by cross-validation. *Comm. Statist.* A **4**, 1-17.

[*Received October* 1986. *Revised April* 1987]