A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation

S. J. Sheather, M. C. Jones

Journal of the Royal Statistical Society. Series B (Methodological), Volume 53, Issue 3 (1991), 683-690.

Journal of the Royal Statistical Society. Series B (Methodological)

J. R. Statist. Soc. B (1991)
53, *No.* 3, *pp.* 683–690

A Reliable Data-based Bandwidth Selection Method for Kernel Density Estimation

By S. J. SHEATHER and M. C. JONES†

University of New South Wales, Sydney, Australia *IBM Research Division, Yorktown Heights, USA*

[Received August 1989. Final revision July 1990]

SUMMARY

We present a new method for data-based selection of the bandwidth in kernel density estimation which has excellent properties. It improves on a recent procedure of Park and Marron (which itself is a good method) in various ways. First, the new method has superior theoretical performance; second, it also has a computational advantage; third, the new method has reliably good performance for smooth densities in simulations, performance that is second to none in the existing literature. These methods are based on choosing the bandwidth to (approximately) minimize good quality estimates of the mean integrated squared error. The key to the success of the current procedure is the reintroduction of a non-stochastic term which was previously omitted together with use of the bandwidth to reduce bias in estimation without inflating variance.

Keywords: ADAPTIVE CHOICE; BIAS REDUCTION; FUNCTIONAL ESTIMATION; SMOOTHING; SQUARED ERROR LOSS FUNCTIONS

1. INTRODUCTION

There is currently much interest in the problem of providing good data-based procedures for selecting the smoothing parameter—which controls the degree of smoothing applied to the data—employed in statistical curve estimation techniques. This is particularly so for nonparametric probability density function estimation by the kernel method, which is described in Silverman (1986), for example. In this context, we call the smoothing parameter the bandwidth and denote it by h.

An important recent paper in this area is Park and Marron (1990). Park and Marron studied an estimator, the performance of which—in both theory and simulations—proved to be clearly superior to other methods currently popular in the literature, such as the well-known 'least squares cross-validation', for estimating smooth densities. The work of the current paper makes further advances to the methodology over and above those already made by Park and Marron (1990).

The improvements that we make are on several levels. First, on the theoretical side, we improve the asymptotic rate of convergence of the estimated bandwidth to its theoretical (but practically unavailable) optimum value. Also, and this point is important because the rate improvement is only slight, we note that the constant coefficient of the leading term in the expansion of the performance measure we use is considerably reduced. Both Park and Marron's (1990) bandwidth selection procedure and our favoured one require the numerical solution of an equation; it appears that

†*Address for correspondence*: Department of Statistics, The Open University, Walton Hall, Milton Keynes, MK7 6AA, UK.

 0035-9246/91/53683 $2.00

our procedure is easier to compute because the function that we need to zero is the better behaved of the two. Finally, and most importantly, it must be stressed that the bandwidth estimator that we recommend has a practical performance second to none in the existing literature on the subject. Although the simulation study presented here is rather small, we can be confident of our claim of excellent practical results because of extensive recent, but as yet unpublished, simulations of J. S. Marron which confirm the reliably good performance of our proposed procedure.

2. PARK AND MARRON'S h SELECTOR

The usual kernel density estimate $\hat{f}_h$ of a univariate density f based on a random sample $X_1, \ldots, X_n$ of size n is

$$\hat{f}_h(x) = n^{-1} \sum_{i=1}^{n} h^{-1} K\{h^{-1}(x - X_i)\}. \tag{1}$$

The bandwidth h has already been introduced in Section 1; the function K is the kernel function which we take to be a symmetric probability density. All the data-based h selection procedures discussed in this paper are based on choosing h to (at least approximately) minimize a kernel-based estimate of mean integrated squared error (MISE) via the first two terms of its usual asymptotic expansion (AMISE) valid as $n \to \infty$ and $h = h(n) \to 0$:

$$\text{AMISE}(h) = (nh)^{-1} R(K) + \tfrac{1}{4} h^4 \sigma_K^4 R(f'') \tag{2}$$

(e.g. Silverman (1986), section 3.3). Here, the notation follows the convention $R(g) = \int g^2(x)\, dx$ and $\sigma_g^2 = \int x^2 g(x)\, dx$ for appropriate functions g, and the quantities involved are assumed to exist and be finite. Each objective function is thus of the form

$$\psi(h) = (nh)^{-1} R(K) + \tfrac{1}{4} h^4 \sigma_K^4 \hat{S}(\alpha) \tag{3}$$

where $\hat{S}(\alpha)$ is a kernel-based estimate of $R(f'')$, using some appropriate bandwidth α; $\hat{S}(\alpha)$ is discussed below. Note that if α did not depend on h, the minimization of ψ could be performed analytically to give

$$\tilde{h} = [R(K)/\{\sigma_K^4 \hat{S}(\alpha)\}]^{1/5} n^{-1/5}, \tag{4}$$

an estimate of the usual expression for the asymptotically optimal bandwith, h_*, say. In its turn, h_* is an approximation to h_0, the exact minimizer of MISE(h).

Differing in a negligible way from Park and Marron (1990), their h estimator takes $\hat{S}(\alpha)$ to be

$$\hat{S}_{\text{ND}}(\alpha) = \{n(n-1)\}^{-1} \alpha^{-5} \sum\sum_{i \neq j} L^{\text{iv}}\{\alpha^{-1}(X_i - X_j)\} \tag{5}$$

(this derives from the estimate $n^{-1} \Sigma \hat{f}_\alpha^{\text{iv}}(X_i)$ of $R(f'')$ and the subscript ND, standing for 'no diagonals', refers to the fact that the double sum does not include $i = j$ terms). Notice that, importantly, α is another bandwidth differing from h, and also that L is another symmetric density not necessarily K. Taking α to be an estimate of the asymptotically optimal bandwidth α_1, say, for estimating $R(f'')$ leads to Park and

Marron's (1990) 'plug-in' estimator of h_0, which develops ideas in Hall (1980) and Sheather (1983, 1986). From Hall and Marron (1987),

$$\begin{aligned} \alpha_1 &= C_1(L)\, C_2(f) n^{-2/13} \\ &= C_3(L)\, C_4(f) h_*^{10/13} \end{aligned} \tag{6}$$

where

$$C_1(L) = \{18\, R(L^{\mathrm{iv}})/\sigma_L^4\}^{1/13} = \{R(K)/\sigma_K^4\}^{2/13}\, C_3(L)$$

and

$$C_2(f) = \{R(f)/R^2(f''')\}^{1/13} = R^{-2/13}(f'')\, C_4(f).$$

Now, α_1 in turn depends on an unknown functional of f. However, at this second stage, it turns out to be sufficiently good to estimate this functional less well, in fact using a scale model for f, i.e. write $f = g_{\hat{\lambda}}$, say, where $g_\lambda(x) = \lambda^{-1} g_1(\lambda^{-1}x)$, g_1 is a fixed density, such as the (suitably standardized) normal, and the scale parameter is estimated robustly, giving $\hat{\lambda}$, say. Park and Marron's (1990) published algorithm, with which we compare the estimators to follow, is completed by considering equation (6) to give a general relationship between α and h, namely $\alpha = \alpha(h) = C_3(L)\, C_4(g_{\hat{\lambda}}) h^{10/13}$, i.e. α is taken to depend on h; then $\hat{h}_1$, say, is given by that h, found numerically, that solves equation (4) with $\tilde{h} = h$, $\hat{S} = \hat{S}_{\mathrm{ND}}$ and $\alpha = \alpha(h)$.

Theorem 3.3 of Park and Marron (1990) describes the asymptotic performance of $\hat{h}_1$. They show that, for sufficiently smooth f,

$$\hat{h}_1/h_0 = 1 + O_p(n^{-4/13}). \tag{7}$$

Moreover, the mean squared relative error (MSRE) of $\hat{h}_1$ is given by

$$\begin{aligned} \mathscr{E}(\hat{h}_1/h_0 - 1)^2 \simeq\ & 100^{-1}\, R^{-2}(f'')\{18\, R(L^{\mathrm{iv}})\, R(f)\}^{4/13} \\ & \times \{\sigma_L^2\, R(f''')\}^{18/13}(Q^4 + 4Q^{-9}/9) n^{-8/13} \end{aligned} \tag{8}$$

where $Q = C_4(g_\lambda)/C_4(f)$ (essentially as in Park and Marron (1990)). Note that the above rate of convergence is the best exhibited by Park and Marron and that, by and large, this superior performance carries over to (small sample) simulation results as well. Indeed, it is difficult to find another example of an automatic bandwidth selection procedure with as consistently good a simulation performance as $\hat{h}_1$ in the literature.

3. AN IMPROVED h SELECTOR

Given that the estimates of $R(f'')$ studied in Jones and Sheather (1991) improve, theoretically, on those of Hall and Marron (1987) on which Park and Marron's h selector is based, it is now natural to replace $\hat{S}_{\mathrm{ND}}$ by Jones and Sheather's $\hat{S}_{\mathrm{D}}$ in the above. The latter estimate of $R(f'')$ is given by equation (5) with $i = j$ terms added in (the subscript D means 'diagonals in'). The difference between $\hat{S}_{\mathrm{D}}$ and $\hat{S}_{\mathrm{ND}}$ is a non-stochastic term which contributes a positive amount to the bias in estimating $R(f'')$; the trick is then to recognize that the bias due to the smoothing is negative and to use the bandwidth α to (approximately) cancel the 'diagonal' term with the leading smoothing bias term. The value of α that this leads to is α_2, say, which, from equation (4) of Jones and Sheather (1991), is

$$\alpha_2 = D_1(L)\, R^{-1/7}(f''')n^{-1/7} \tag{9}$$

where

$$D_1(L) = \{2\, L^{\mathrm{iv}}(0)/\sigma_L^2\}^{1/7}.$$

By analogy with Park and Marron's (1990) algorithm, our first $\hat{h}$, denoted by $\hat{h}_{2S}$, say, solves the equation

$$h = [R(K)/\{\sigma_K^4 \hat{S}_D(\alpha_2(h))\}]^{1/5} n^{-1/5},$$

where, noting that $\alpha_2 = c_1 h_*^{5/7}$ for appropriate c_1, we have written $\alpha_2 = \alpha_2(h) = \hat{c}_1 h^{5/7}$, where $\hat{c}_1$ estimates c_1. It is tempting simply to use a scale model to estimate c_1 but this is not quite good enough here. The inconsistency of such an estimate (unless g_λ is fortuitously the true f) means that the leading bias terms do not cancel out sufficiently well and a non-negligible bias term is reintroduced. To obtain the necessary sufficiently small bias, it turns out that we must estimate $R(f''')$ by some $\hat{T}$ such that $\hat{T} = R(f''') + o_p(n^{-1/14})$ (Jones and Sheather, 1991). Thus, any of the consistent estimators of $R(f''')$ discussed by Hall and Marron (1987) or Jones and Sheather (1991) will suffice.

There are, however, various alternative options, each of which is investigated in the simulation study of Section 4. First, note that a computationally simpler and essentially asymptotically equivalent alternative estimator of h_0 arises by leaving $\alpha_2 = c_2 n^{-1/7}$, for appropriate c_2 as in equation (9), estimating c_2 and using equation (4) to give a direct formula for $\tilde{h}$. Call this $\hat{h}_{2P}$, say. A third approach uses essentially the same device as does $\hat{h}_{2S}$ in that we write α_2 in terms of h, but uses it in the minimization of expression (3) rather than the solution of equation (4); this gives $\hat{h}_{2M}$. Asymptotic analysis (not given) actually yields a slightly different optimal α in the minimization case which we utilize in practice. Although it is not theoretically necessary to do so, we find it best in practice to use optimal bandwidths in diagonals-in formulae for estimating functionals at the second stage, and estimate only the scale in these, using the same scale model ideas as in Section 2. We give full implementation details only for what proves to be the most successful of these options in Section 5. At times, we refer to any of the above estimators as $\hat{h}_2$.

At the (slight) expense of introducing a third stage in the above estimation procedure, we have obtained estimators $\hat{h}_2$ of h_0 with better theoretical properties than $\hat{h}_1$. This was proved by Jones and Sheather (1991). Their result 2 gives that, for f little smoother than that necessary for equations (7) and (8),

$$\hat{h}_2/h_0 = 1 + O_p(n^{-5/14}) \tag{10}$$

and that the MSRE of $\hat{h}_2$ is

$$\mathscr{E}(\hat{h}_2/h_0 - 1)^2 \simeq 25^{-1} \times 2^{-2/7} R^{-2}(f'')\, R(L^{\mathrm{iv}}) \times R(f)\{\sigma_L^2\, R(f''')/L^{\mathrm{iv}}(0)\}^{9/7} n^{-5/7} \tag{11}$$

for $\hat{h}_{2S}$ and $\hat{h}_{2P}$, while $\hat{h}_{2M}$ also has convergence rate (10) but proves to have an inferior constant coefficient of MSRE to that in approximation (11) (not given). Of course although the $O_p(n^{-5/14})$ term appearing in equation (10) guarantees $\hat{h}_2$'s superiority over $\hat{h}_1$—which has a corresponding $O_p(n^{-4/13})$ term in equation (7)—*for sufficiently large n*, it is not so clear what the small sample repercussions are. However, the very similarity of these rates makes it meaningful to compare constant multipliers in

MSREs. Denoting these by η_i, $i=1, 2$, corresponding to $\hat{h}_i$, $i=1, 2$, respectively ($\hat{h}_2 = \hat{h}_{2P}$ or $\hat{h}_2 = \hat{h}_{2S}$ here only), numerical calculation in the special case f is normal (and also taking K and L to be normal) gives $\eta_2 \simeq 0.503\eta_1$ in the case $Q=1$. (Getting g 'wrong' increases η_1, although discarding the scale model trick for a further kernel estimation step would retain η_1 as the appropriate constant). Such an improvement in the constant augurs well for $\hat{h}_2$'s performance, too.

Reverting to diagonals-in estimates $\hat{S}_D$ instead of $\hat{S}_{ND}$ was originally motivated by algorithmic considerations in numerically solving equations like equation (4) for Park and Marron's (1990) h selector. When using $\hat{S}_{ND}$, with a normal L, it was typical for $\hat{S}_{ND} < 0$ for some small values of h. When $\hat{S}_{ND}$ changes sign, the function in the equation being zeroed has a discontinuity and changes sign. This is particularly troublesome to many root finding procedures and can mislead them into reporting this discontinuity as a solution. Such problems do not exist if we use $\hat{S}_D$, since it has the great advantage that it is necessarily always positive. Plots of the relevant functions based on $\hat{S}_{ND}$ (full curve) and $\hat{S}_D$ (broken curve) are given for a typical sample of size $n=100$ from the standard normal distribution in Fig. 1; the figure illustrates this computational advantage of our proposal well. Moreover, although the $\hat{h}_{2S}$ function appears always to have the same shape as this one, in many cases the Park–Marron function is even more badly behaved than this, with more (discontinuous) zero crossings.

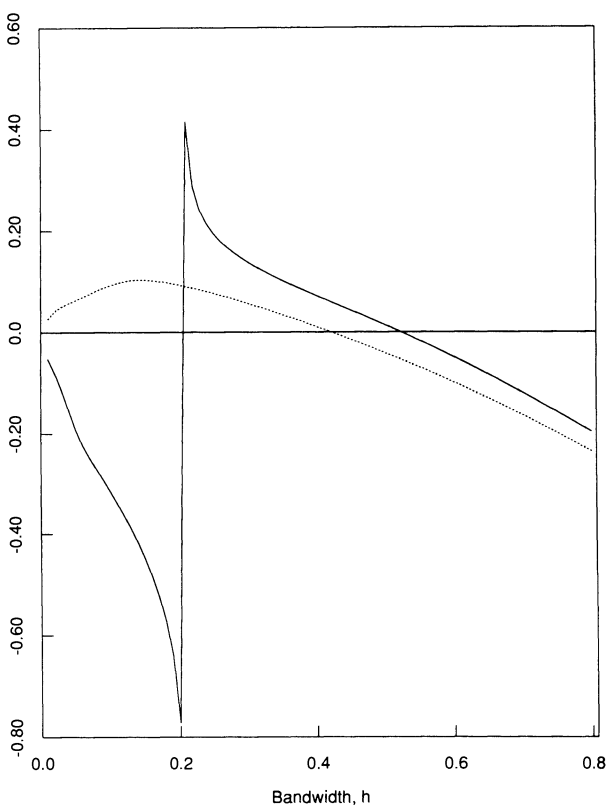

Fig. 1. Functions that must be zeroed: ———, for Park and Marron's (1990) method; --------, for $\hat{h}_{2S}$

4. SIMULATION RESULTS

The simulation study reported here is rather brief. We compare all three $\hat{h}_2$ proposals of Section 3 with each other and with Park and Marron's (1990) procedure, $\hat{h}_1$. We took the options of setting both K and L to be ϕ, the standard normal density, of using the interquartile range as $\hat{\lambda}$ wherever needed and of setting $g_1(x) = 1.349 \times \phi(1.349x)$. For those $\hat{h}$s not given by a direct formula, we used the Newton–Raphson method to solve the required equations numerically.

We generated $W = 500$ realizations of data sets of size $n = 50$ and $n = 100$ from each of four test densities. These densities are ϕ, the normal mean mixture $f_1(x) = \frac{1}{2}\phi(x+\frac{3}{2}) + \frac{1}{2}\phi(x-\frac{3}{2})$ and the normal variance mixtures $f_2(x) = \frac{1}{2}\phi(x) + \frac{1}{2}\sqrt{10}\,\phi(\sqrt{10}\,x)$ and $f_3(x) = \frac{1}{2}\phi(x) + 5\,\phi(10x)$.

In Fig. 2 are shown approximate confidence intervals for the quantities $R_M = 100\,\mathscr{E}\{\text{MISE}(\hat{h})/\text{MISE}(h_0) - 1\}$, where $\hat{h}$ stands for any of the bandwidths of interest. MISE($\hat{h}$) denotes plugging the value of $\hat{h}$ into the MISE formula for fixed h and its use is best justified by observing that R_M is essentially proportional to $\mathscr{E}(\hat{h}/h_0 - 1)^2$. It is fairly widely accepted that assessing the worth of any $\hat{f}_{\hat{h}}$ is sensibly done by comparing $\hat{h}$ with its 'target' h_0 (see Jones (1991) for reasons). R_M, however, gives us a feel for the practical worthwhileness of any bandwidth differences. Our approximate confidence intervals are precisely the pivoted confidence intervals of Park and Marron (1990) which we denote by $(\hat{R}_M/(1+V), \hat{R}_M/(1-V))$. Here, $\hat{R}_M$ is the obvious estimate of R_M obtained from the simulations and $V = 1.96(2/W)^{1/2} = 0.124$.

There are some interesting conclusions to be drawn from Fig. 2. Recall first that $\hat{h}_1$ is considered to exhibit good performance by the current standards of the literature. Away from the normal distribution, however, $\hat{h}_1$ is often considerably worse than the $\hat{h}_2$s. Of the $\hat{h}_2$s, $\hat{h}_{2S}$s is never dominated by $\hat{h}_{2M}$; $\hat{h}_{2P}$ does better than $\hat{h}_{2S}$ for the

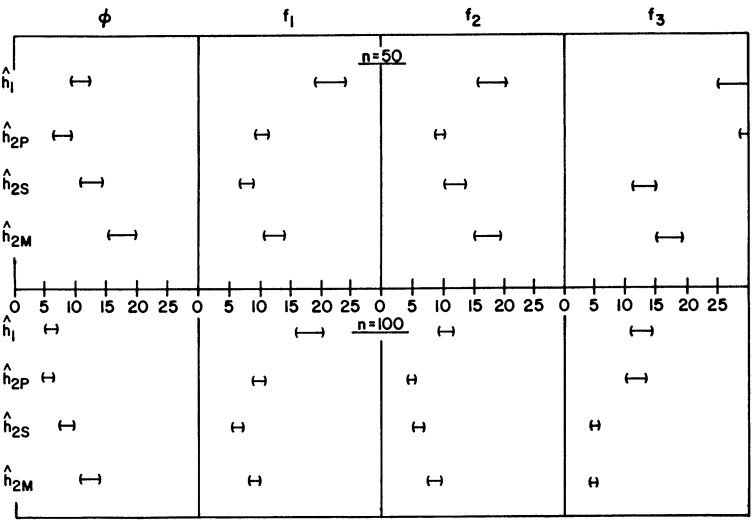

Fig. 2. Approximate 95% confidence intervals for R_M for each of four data-based bandwidths, for two sample sizes from each of four different underlying distributions (definitions of all these quantities are in the text)

normal and (normal-like) f_2 distributions, worse for f_1 and is considerably inferior to $\hat{h}_{2S}$ for data from f_3. On this evidence, the consistently good performance of $\hat{h}_{2S}$ suggests it as the method of choice. Further, on J. S. Marron's extensive simulation evidence, $\hat{h}_{2S}$ continues to perform very well, and better than other methods tried, over a wide range of smooth density shapes.

5. THE $\hat{h}$ OF CHOICE

In Section 4, we saw that the bandwidth selection procedure resulting in $\hat{h}_{2S}$ has much to recommend it. For the reader's convenience, here we give full details of the formulae for this bandwidth selector.

The bandwidth $\hat{h}_{2S}$, for use in equation (1), is the solution to the equation

$$[R(K)/\{\sigma_K^4 \hat{S}_D(\hat{\alpha}_2(h))\}]^{1/5} n^{-1/5} - h = 0 \tag{12}$$

where

$$\hat{S}_D(\alpha) = \{n(n-1)\}^{-1}\alpha^{-5} \sum_{i=1}^{n} \sum_{j=1}^{n} \phi^{iv}\{\alpha^{-1}(X_i - X_j)\}.$$

From manipulation of equation (9), we have

$$\hat{\alpha}_2(h) = 1.357\{\hat{S}_D(a)/\hat{T}_D(b)\}^{1/7} h^{5/7}.$$

Here, the constant is $D_1(\phi)/R^{1/7}(\phi)$ and the term in brackets is the second stage estimate of $R(f'')/R(f''')$;

$$\hat{T}_D(b) = -\{n(n-1)\}^{-1} b^{-7} \sum_{i=1}^{n} \sum_{j=1}^{n} \phi^{vi}\{b^{-1}(X_i - X_j)\}.$$

For this estimate the bandwidths a and b are given by a normal scale model estimate of equation (9) and of the corresponding formula for estimating $R(f''')$ in Jones and Sheather (1991) respectively to be

$$a = 0.920\hat{\lambda} n^{-1/7} \text{ and } b = 0.912\hat{\lambda} n^{-1/9},$$

where $\hat{\lambda}$ is the sample interquartile range.

We successfully use the Newton-Raphson method to solve equation (12). A Fortran subroutine is available on request from the first author.

ACKNOWLEDGEMENTS

The authors are very grateful to Peter Hall and especially Steve Marron for many helpful comments and discussions. We also acknowledge Rob Hyndman for programming assistance in an earlier version of this work. The editorial process has been greatly beneficial in improving the standard of presentation of this material. Some of M. C. Jones's work was supported by a Mathematical Sciences Research Centre Visiting Fellowship at the Australian National University, Canberra, Australia.

REFERENCES

Hall, P. (1980) Objective methods for the estimation of window size in the nonparametric estimation of a density. Unpublished.

Hall, P. and Marron, J. S. (1987) Estimation of integrated squared density derivatives. *Statist. Probab. Lett.*, **6**, 109–115.

Jones, M. C. (1991) The roles of ISE and MISE in density estimation. *Statist. Probab. Lett.*, to be published.

Jones, M. C. and Sheather, S. J. (1991) Using non-stochastic terms to advantage in kernel-based estimation of integrated squared density derivatives. *Statist. Probab. Lett.*, to be published.

Park, B. U. and Marron, J. S. (1990) Comparison of data-driven bandwidth selectors. *J. Am. Statist. Ass.*, **85**, 66–72.

Sheather, S. J. (1983) A data-based algorithm for choosing the window width when estimating the density at a point. *Comput. Statist. Data Anal.*, **1**, 229–238.

——(1986) An improved data-based algorithm for choosing the window width when estimating the density at a point. *Comput. Statist. Data Anal.*, **4**, 61–65.

Silverman, B. W. (1986) *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.