Statistics & Probability Letters 6 (1988) 311–314
North-Holland

April 1988

ON THE MINIMIZATION OF ABSOLUTE DISTANCE IN KERNEL DENSITY ESTIMATION

Peter HALL and Matthew P. WAND
Australian National University, Canberra, Australia

Received May 1987
Revised August 1987

Abstract: We given an algorithm for determining the window-size which minimizes mean absolute distance in kernel density estimation and discuss the practical implications of our results.

Keywords: kernel estimator, L_1 distance, mean absolute error, nonparametric density estimation, optimal window-size.

1. Introduction

Suppose we wish to estimate the probability density f at some point x using a kernel estimator, $\hat{f}(x \mid h)$ where h is the window-size parameter. The choice of h is very important since it controls the "trade-off" between the bias and variance of the estimate. The theoretically optimal choice of this parameter is usually taken to be that h which minimizes an asymptotic formula for the mean squared distance between $\hat{f}(x \mid h)$ and $f(x)$, often referred to as the mean squared error (MSE). Under certain regularity conditions on $f(x)$ an exact expression for the asymptotically optimal value of h is readily derived (see e.g. Parzen (1962)). An alternative measure of loss is the mean absolute distance between $\hat{f}(x \mid h)$ and $f(x)$, which we shall call the mean absolute error (MAE). Specifically,

$$\mathrm{MAE}\{\hat{f}(x \mid h)\} = E \mid \hat{f}(x \mid h) - f(x) \mid,$$

which is the "local" analogue of the L_1 distance between f and $\hat{f}$. Bounds to the window-size which asymptotically minimize L_1 loss were derived by Devroye and Györfi (1985). An algorithm for finding the exact asymptotically optimal window-size with respect to this measure of loss was developed by Hall and Wand (1987). In this paper we modify the algorithm to compute the choice of h which asymptotically minimizes the MAE (Section 2). Numerical values of these optimal values are presented for certain densities and comparisons are made between these and the corresponding values for minimization of MSE (Section 3).

Also in Section 3, the practical implications of our results are discussed. We argue that there is hardly any difference between density estimates optimized for L_1 or L_2 distance, either locally or globally. Thus it would seem that the emphasis in recent work on the difference between L_1 and L_2 metrics, has little bearing on the actual estimation of density functions.

In Section 4 we point out that our techniques can be generalized to other L_q metrics for integer q.

2. Minimization of MAE

In this section unqualified integrals are assumed to be over the whole real line. Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables from a population with univariate density f. Consider the

0167-7152/88/$3.50

kernel estimator

$$\hat{f}(x \mid h) = (nh)^{-1} \sum_{i=1}^{n} K\{(x - X_i)/h\},$$

where K is a pth order kernel for some integer $p \geqslant 1$. This means that $\int |z^p K(z)| \, dz < \infty$, K is bounded, $\int K = 1$ and

$$\int z^j K(z) \, dz = \begin{cases} 0, & j = 1, \ldots, p-1, \\ (-1)^p \kappa_1 \neq 0, & j = p. \end{cases}$$

Consider some point x on the real line. Assume that at this point $f^{(p)}$ exists, is nonzero and continuous and that $f(x) > 0$. It is well-known that the bias and standard deviation of $\hat{f}(x \mid h)$ are asymptotic to $(\kappa_1/p!)h^p f^{(p)}(x)$ and $\kappa_2 (nh)^{-1/2} \times f(x)^{1/2}$ respectively, where $\kappa_2 = (\int K^2)^{1/2}$ (see e.g. Parzen (1962)). From these we obtain

$$\begin{aligned} &\hat{f}(x \mid h) - f(x) \\ &\quad = (\kappa_1/p!)h^p f^{(p)}(x) \\ &\qquad + \kappa_2 (nh)^{-1/2} f(x)^{1/2} Z(x) + o(h^p) \end{aligned} \tag{2.1}$$

as $n \to \infty$, $h = h(n) \to 0$ and $nh \to \infty$, where $Z = Z(x)$ is asymptotically a standard normal random variable.

To "balance" bias and standard deviation we must choose h so that each of these quantities are of the same order of magnitude. This involves taking $h = u^2 n^{-1/(2p+1)}$ for some positive constant u not depending on n. Let b_x and σ_x stand for $(\kappa_1/p!)f^{(p)}(x)$ and $\kappa_2 f(x)^{1/2}$ respectively. Then from (2.1) we see that $\mathrm{MAE}\{\hat{f}(x \mid h)\}$ is asymptotic to $n^{-p/(2p+1)} \lambda_x(u)$, where

$$\lambda_x(u) = \int_{-\infty}^{\infty} |u^{2p} b_x - u^{-1} \sigma_x z| \phi(z) \, dz, \tag{2.2}$$

and ϕ denotes the standard normal density function. Consequently, the asymptotically optimal value of u is that value which minimizes $\lambda_x(u)$.

Observe that (2.2) may be rewritten as

$$\lambda_x(u) = 2\sigma_x u^{-1} \int_{-\infty}^{u^{2p+1} b_x/\sigma_x} \Phi(z) \, dz - u^{2p} b_x$$

where $\Phi(y) = \int_{-\infty}^{y} \phi(z) \, dz$. Noting that $\int_{-\infty}^{y} z\phi(z) \, dz = -\phi(y)$ we obtain

$$\tfrac{1}{2}\lambda_x'(u) = u^{-2} \Lambda_x(u^{2p+1}),$$

where

$$\Lambda_x(v) = 2pvb_x\{\Phi(vb_x/\sigma_x) - \tfrac{1}{2}\} - \sigma_x \phi(vb_x/\sigma_x).$$

The value of u, u^* say, which minimizes $\lambda_x(u)$ is equal to $(v^*)^{1/(2p+1)}$ where v^* is that value of v for which $\Lambda_x(v) = 0$ $(v > 0)$. Now,

$$\begin{aligned} \Lambda_x'(v) &= 2pb_x\{\Phi(vb_x/\sigma_x) - \tfrac{1}{2}\} \\ &\quad + (2p+1)b_x^2 \sigma_x^{-1} v \phi(vb_x/\sigma_x), \end{aligned}$$

which is positive for all $v > 0$. Also $\lim_{v \to \infty} \Lambda_x(v) = \infty$, while for $b_x \neq 0$,

$$\lim_{v \downarrow 0} \Lambda_x(v) = -\sigma_x \phi(0) < 0,$$

proving that v^*, and therefore u^*, exists and is unique.

In practice we can locate v^* by iteration using Newton's method by forming the sequence $v_1, v_2, \ldots$ where $v_{i+1} = v_i - \Lambda_x(v_i)/\Lambda_x'(v_i)$. The limit of this sequence is v^*. The functions involved in the iteration are readily computable and the convergence of the sequence is very rapid. The value of h which minimizes MAE at x is asymptotic to $(v^*)^{2/(2p+1)} n^{-1/(2p+1)}$.

3. Numerical results

The algorithm derived in Section 2 computes the "optimal coefficient" for h, that is the coefficient of $n^{-1/(2p+1)}$ in the formula for the MAE optimal value of h when estimating the density at a point x. We shall let $c_1(x)$ be this quantity. The optimal coefficient for minimizing MSE at x will be denoted by $c_2(x)$ and is given by

$$c_2(x) = \left[\frac{\kappa_2^2 (p!)^2 f(x)}{2p\kappa_1^2 \{f^{(p)}(x)\}^2} \right]^{1/(2p+1)}$$

(see e.g. Parzen (1962)). The numerical values of $c_1(x)$ and $c_2(x)$ presented in this section are all based on the Gaussian kernel, $K(z) = (2\pi)^{-1/2} e^{-z^2/2}$, so we have $p = 2$, $\kappa_1 = 1$ and $\kappa_2 = (2\pi^{1/2})^{-1/2}$.

Table 3.1 lists values of $c_1(x)$ and $c_2(x)$ for the estimation of (a) the standard normal density, (b) the extreme value density $(f(x) = e^x e^{-e^x})$, (c) the equal-proportion, two-component normal mixture

Table 3.1.
Values of c_1 and c_2

(a) Standard normal density

x	0.0	0.5	1.0	1.5	2.0	2.5	3.0
$c_1(x)$	0.92	1.06	∞	1.05	0.88	0.88	0.98
$c_2(x)$	0.93	1.07	∞	1.07	0.90	0.90	1.00

(b) Extreme value density

x	−4.0	−3.0	−2.0	−1.0	0.0	1.0	2.0
$c_1(x)$	1.75	1.50	1.43	4.00	0.93	1.93	0.55
$c_2(x)$	1.77	1.52	1.45	4.06	0.95	1.96	0.56

(c) Normal mixture density

x	0.0	0.5	1.0	1.5	2.0	2.5	3.0
$c_1(x)$	0.44	1.26	0.53	19.9	0.51	0.56	0.88
$c_2(x)$	0.45	1.28	0.54	20.2	0.52	0.57	0.90

(d) Standard Cauchy density

x	0.0	0.5	1.0	1.5	2.0	2.5	3.0
$c_1(x)$	0.73	1.59	1.10	1.18	1.40	1.67	1.98
$c_2(x)$	0.74	1.61	1.12	1.19	1.42	1.70	2.01

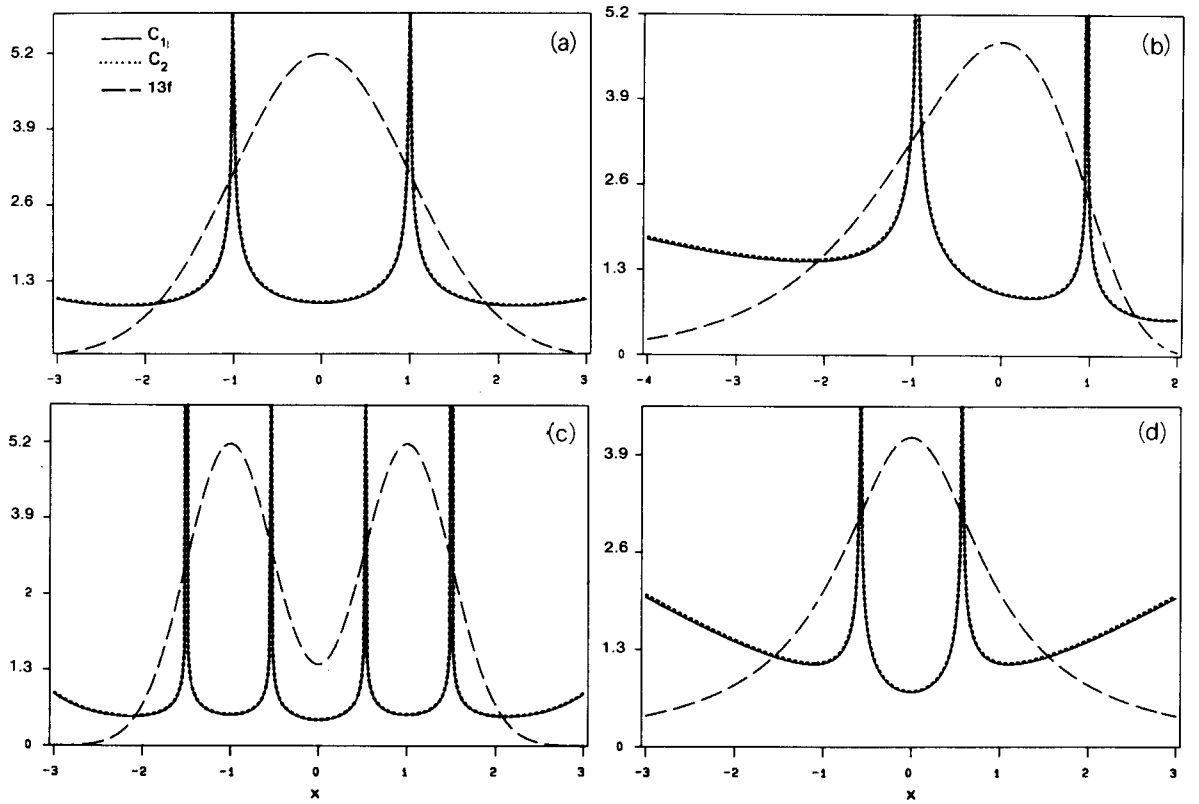

Fig. 3.1. Graphs of c_1, c_2 and $13 \times f$ for (a) $N(0, 1)$, (b) extreme value, (c) normal mixture, (d) Cauchy distributions.

with means $(-1, 1)$ and variances $(\frac{1}{4}, \frac{1}{4})$ and (d) the standard Cauchy density. We see that the values are very close in every case. The graphs in Figure 3.1 depict the functions $c_1(x)$, $c_2(x)$ and $f(x)$ ($f(x)$ is scaled up by a factor of 13) for the same four densities. In each case the curves for c_1 and c_2 are virtually indistinguishable. The values of $c_2(x)$ will always be greater than that of $c_1(x)$. Indeed, if in the case of the Gaussian kernel we put $v_2^* = c_2(x)^{5/2}$, we obtain

$$\begin{aligned}\Lambda_x(v_2^*) &= (2\pi^{1/2})^{-1/2}\left[2\left\{\Phi\left(\tfrac{1}{2}\right) - \tfrac{1}{2}\right\} - \phi\left(\tfrac{1}{2}\right)\right] f(x)^{1/2} \\ &= (0.0016\ldots) f(x)^{1/2},\end{aligned}$$

showing that v_2^* marginally over-estimates v^*. Taking v_2^* as a starting guess for Newton's method means that a very accurate estimate of v^* is usually obtained after only two or three iterations.

These examples, and many others of the same type, indicate that there is little difference between minimizing MAE and MSE in kernel density estimation. Figure 3.2 confirms that MAE is quite insensitive to the difference between c_1 and c_2: there is hardly any distinguishable difference in MAE if MAE-optimal or MSE-optimal windows are used. The same phenomenon may be observed for global L_1 loss: in the great majority of examples, the window-size which asymptotically minimizes L_1 distance is within a few percent of the window-size which asymptotically minimizes L_2 distance. These results have important implications for practical density estimation, since they show that minimizing L_1 distance (either pointwise or globally) is virtually equivalent to minimizing L_2 distance. In this sense,t he emphasis which several writers have placed recently on the difference between L_1 and L_2 metrics, is almost solely a technical matter. It has relatively little bearing on the actual estimation of density functions.

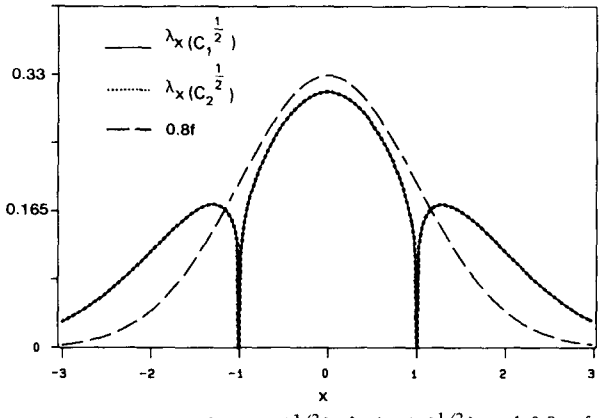

Fig. 3.2. Graphs of $\lambda_x\{c_1(x)^{1/2}\}$, $\lambda_x\{c_2(x)^{1/2}\}$ and $0.8 \times f$ for $N(0, 1)$ distribution.

4. Other L_q metrics

The theory discussed in Section 2 can be extended to the minimization of L_q loss, both globally and locally, for integers $q \geqslant 1$. In the case of local estimation of $f(x)$ the optimal coefficient, $c_q(x)$, for minimizing the error criterion $E|\hat{f}(x|h) - f(x)|^q$ is given by $(u^*)^2$ where u^* is the value of u which minimizes

$$\lambda_{x,q}(u) = \int_{-\infty}^{\infty} \left|u^{2p} b_x - u^{-1}\sigma_x z\right|^q \phi(z)\, dz.$$

For integral values of q the expression on the right-hand side can be expanded and then minimized by differentiation. The global minimization of L_q distance has a similar treatment.

Acknowledgement

We are grateful to Professor W. Schucany for a helpful comment.

References

Devroye, L. and L. Györfi (1985), *Nonparametric Density Estimation: The L_1 View* (Wiley, New York).

Hall, P. and M.P. Wand (1988), Minimizing L_1 distance in nonparametric density estimation, *J. Multivariate Anal.*, to appear.

Parzen, E. (1962), On the estimation of a probability density function and the mode, *Ann. Math. Statist.* **40**, 1065–1076.