Scand J Statist 9: 65–78, 1982

Empirical Choice of Histograms and Kernel Density Estimators

MATS RUDEMO

The Royal Veterinary and Agricultural University, Copenhagen

Received May 1980, in final form July 1981

ABSTRACT. Methods of choosing histogram width and the smoothing parameter of kernel density estimators by use of data are studied. They are based on estimators of risk functions corresponding to mean integrated squared error and the Kullback-Leibler information measure. Two closely related risk function estimators are given, one of which is derived from cross-validation. In examples with simulated and real data the methods are applied to estimation of probability densities and the rate function of a time-dependent Poisson process.

Key words: empirical smoothing, histogram, kernel estimator, time-dependent Poisson process, non-parametric density estimation, cross-validation, series density estimator

1. Introduction

Suppose that $X_1, \ldots, X_n$ are independent, identically distributed, real-valued random variables with probability density f. We shall consider estimators $\hat{f}$ of f. First let $\mathbf{I}=(I_k)$ be a partition of the real line into disjoint intervals. Let h_k denote the length of I_k and let $N_k=\#\{i: X_i \in I_k,\ 1 \leqslant i \leqslant n\}$ be the number of observations in I_k. Put $\mathbf{X}=(X_1, \ldots, X_n)$ and

$$\hat{f}(x) = \hat{f}_{\mathbf{I}}(x, \mathbf{X}) = N_k/(nh_k), x \in I_k. \tag{1.1}$$

Then $\hat{f}=\hat{f}_{\mathbf{I}}$ is the histogram corresponding to $\mathbf{I}$ and $\mathbf{X}$. Typically we shall assume that all the intervals I_k have equal length h.

Let us say that a non-negative real function K is a kernel if $\int K(x)dx=1$. All unspecified integrals are taken over the entire line. Suppose that $h>0$ and put $\alpha=(K, h)$. The kernel estimator $\hat{f}=\hat{f}_\alpha$ is defined by

$$\hat{f}(x) = \hat{f}_\alpha(x, \mathbf{X}) = \frac{1}{nh}\sum_{i=1}^{n} K((x-X_i)/h). \tag{1.2}$$

Suppose more generally that we have a family of density estimators $(\hat{f}_\alpha)_{\alpha \in \mathcal{A}}$. The problem discussed in this paper is the choice of α by use of the observations. In particular we shall discuss the choice of h, i.e. the degree of smoothing, for the histogram and the kernel estimator.

For an estimator $\hat{f}$ of f we can use several risk functions such as

$$R^p(\hat{f}, f) = E\int |\hat{f}(x)-f(x)|^p \varrho(x)\, dx, \tag{1.3}$$

where $p \geqslant 1$ and ϱ is a non-negative weight function, or the Kullback-Leibler risk function

$$KL(\hat{f}, f) = -E\int f(x) \log \hat{f}(x)\, dx. \tag{1.4}$$

With f fixed $KL(g, f)$ is minimized for $g=f$, Kullback (1959, p. 15). The Kullback-Leibler risk function is not suitable with the histogram estimator (1.1) as $KL(\hat{f}, f)$ is then infinite. Note that $\hat{f}(x)$ in (1.1) is zero with positive probability unless $P(N_k=n)=1$. The risk function (1.4) will be used in the examples below together with (1.3) with $p=2$ and $\varrho(x) \equiv 1$. This risk function, often called MISE (mean integrated squared error), seems to be the most common global measure in studies of probability density estimators, including the pioneering paper Rosenblatt (1956).

Surveys of non-parametric density estimation can be found in Rosenblatt (1971), Wegman (1972), Tarter & Kronmal (1976), Fryer (1977) and in the recent books Tapia & Thompson (1978) and Wertz (1978). An extensive bibliography is given in Wertz & Schneider (1979). A general result for kernel

estimators is that as n increases h should decrease essentially as $n^{-1/5}$, cf. Rosenblatt (1971) and Silverman (1978). For the MISE risk function an explicit formula for the asymptotically optimal h-value can be found. With $a=\int (K(x))^2\,dx$ and $b=\int x^2K(x)\,dx$ it is

$$h_{asy}=(a/b^2)^{1/5}\left[\int (f''(x))^2\,dx\right]^{-1/5} n^{-1/5}, \tag{1.5}$$

provided that f'' is bounded, continuous and square integrable and that K is symmetric and bounded with a finite second moment, see Rosenblatt (1956, 1971) and Nadaraya (1974). For the histogram estimator with interval length h a general result is that h should decrease as $n^{-1/3}$, cf. Révész (1968, p. 163). For the MISE risk function Scott (1979) obtains the explicit asymptotic formula

$$h_{asy}=\left[6/\int (f'(x))^2\,dx\right]^{1/3} n^{-1/3}. \tag{1.6}$$

Most methods for choosing histograms and kernel estimators depend to some extent on subjective evaluation, as e.g. Silvermans (1978) test graph method, or involve the unknown density, see for instance the asymptotic formulas (1.5) and (1.6). An iterative method for estimation of the asymptotically optimal h-value in (1.5) is described in Tapia & Thompson (1978, pp. 67–68). The h-value h_i after i iterations is used to estimate the density and $\int (f''(x))^2\,dx$, from which h_{i+1} is obtained by (1.5). Assuming that the unknown density is close to some specified family of densities, such as the normal, we can use the formulas (1.5) and (1.6) as described by Scott (1979) for histograms.

Orthogonal series estimators of probability densities were introduced by Čencov (1962). Suppose that $(\varphi_k)_{k\in\mathcal{K}}$ is a system of orthonormal real functions with respect to a weight function ϱ. Let $\mathcal{A}$ be a set of subsets of $\mathcal{K}$. For each $\alpha\in\mathcal{A}$ we put

$$\hat{f}_\alpha(x,\mathbf{X})=\sum_{k\in\alpha}\hat{a}_k\varphi_k(x), \tag{1.7}$$

where

$$\hat{a}_k=\frac{1}{n}\sum_{i=1}^{n}\varphi_k(X_i)\varrho(X_i). \tag{1.8}$$

For trigonometric systems Kronmal & Tarter (1968) gave a stopping rule based on estimation of the contribution to a quadratic risk function when extra terms were included. Let us consider the orthonormal system of functions on (0, 1) consisting of $\varphi_0(x)\equiv 1$, $\varphi_{2k-1}(x)=\sqrt{2}\sin 2\pi kx$, $\varphi_{2k}(x)=\sqrt{2}\cos 2\pi kx$, $k\geq 1$, with the weight function $\varrho(x)\equiv 1$. Kronmal & Tarter's stopping rule then consists of including the terms corresponding to $\sin 2\pi kx$ and $\cos 2\pi kx$ if and only if

$$(1/n^2)\sum_{i,j}\cos 2\pi k(X_i-X_j)>2/(n+1). \tag{1.9}$$

In Habbema, Hermans & Van den Broek (1974), see also Habbema & Hermans (1977) and Duin (1976), a modified likelihood technique is described. Let $\mathbf{X}_i$ denote the observations with X_i excluded, i.e.

$$\mathbf{X}_i=(X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_n). \tag{1.10}$$

The technique consists of maximization of

$$L(\alpha)=\prod_{i=1}^{n}\hat{f}_\alpha(X_i,\mathbf{X}_i) \tag{1.11}$$

with $\hat{f}_\alpha$ given by say (1.2). This method is also discussed in Fryer (1977) on the basis of simulations. A related technique is used in Schuster & Gregory (1978), where the sample $\mathbf{X}$ is divided at random into two parts $\mathbf{X}'$ and $\mathbf{X}''$ with $m=n/2$ observations. After that,

$$L_1(\alpha)=\prod_{i=1}^{m}\hat{f}_\alpha(X_i',\mathbf{X}'')$$

is maximized, for $\alpha=\alpha'$ say. With the roles of $\mathbf{X}'$ and $\mathbf{X}''$ exchanged we get maximum for $\alpha=\alpha''$. The density estimator suggested is $\hat{f}=(\hat{f}_{\alpha'}+\hat{f}_{\alpha''})/2$.

Similar techniques have been studied by Rogers & Wagner (1978) and Devroye & Wagner (1979) in connection with discrimination problems. Here (X,θ), $(X_1,\theta_1),\ldots,(X_n,\theta_n)$ are supposed to be independent and identically distributed random pairs, $\theta\in\{1,\ldots,M\}$, and on the basis of observations of $(X_1,\theta_1),\ldots,(X_n,\theta_n)$ and X we shall guess θ. The risk function L_n is simply the probability of a wrong guess. For this risk function two estimators $\hat{L}_n$ are

studied; one, called the deleted estimator, is derived from a splitting of $\mathbf{X}$ into X_i and $\mathbf{X}_i$ as in (1.11), and another, called the hold out estimator, is based on division of the observations into two parts with $n/2$ observations in each part. Bounds of the type $E(\hat{L}_n - L_n)^2 \leq A/n$ are obtained for these risk function estimators.

As will be further discussed in Section 3, the last mentioned procedures are closely related to cross-validation, see Stone (1974). Cross-validation has been used by Wahba in a series of papers to estimate a smooth function from noisy data, see Wahba & Wold (1975) and Craven & Wahba (1979). In Wahba (1977) the problem of density estimation is transformed into a problem of recovering a smooth curve from noisy data and the smoothing parameter of an orthogonal series density estimator is chosen by cross-validation.

2. Estimation of a quadratic risk function

Suppose that our density estimator $\hat{f}$ can be written on the form

$$\hat{f}(x, \mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} g(x, X_i), \tag{2.1}$$

a type of density estimator considered already by Whittle (1958). Then $\hat{f}$ coincides with (1.1) if $g(x, y) = 1/h_k$ when x and y both belong to I_k, while $g(x, y) = 0$ when x and y belong to different intervals. Further, if $g(x, y) = (1/h)K((x-y)/h)$, then $\hat{f}$ coincides with (1.2). The orthogonal series density estimator (1.7) can also be written on the form (2.1) with

$$g(x, y) = \sum_{k \in \alpha} \varphi_k(x)\, \varphi_k(y)\, \varrho(y). \tag{2.2}$$

Other examples, showing the generality of (2.1), together with an analysis of the asymptotic properties can be found in Walter & Blum (1979).

As risk function we shall use

$$Q(\hat{f}, f) = \int \varrho(x)\, E(\hat{f}(x, \mathbf{X}))^2\, dx - 2\int f(x)\, \varrho(x)\, E\hat{f}(x, \mathbf{X})\, dx \tag{2.3}$$

for a known weight function ϱ. Note that $Q(\hat{f}, f) = R^2(\hat{f}, f) - \int f^2 \varrho\, dx$ and thus minimization of $Q(\hat{f}, f)$ is equivalent to minimization of $R^2(\hat{f}, f)$ for fixed f.

To assure that (2.3) is well-defined, we have to make some assumptions. Let g_1 and g_2 be defined by $g_1(x) = Eg(x, X)$ and $g_2(x) = E(g(x, X))^2$, where X is supposed to have density f. Assume that the functions $f\varrho g_1$, ϱg_2 and ϱg_1^2 are integrable. It follows that

$$Q(\hat{f}, f) = \frac{1}{n}\int \varrho g_2\, dx + \frac{n-1}{n}\int \varrho g_1^2\, dx - 2\int f\varrho g_1\, dx, \tag{2.4}$$

and that $Q(\hat{f}, f)$ tends to

$$Q_\infty(\hat{f}, f) = \int \varrho g_1^2\, dx - 2\int f\varrho g_1\, dx$$

as $n \to \infty$. Thus with $\hat{f}$ given by (2.1) for a fixed function g, the risk function $R^2(\hat{f}, f)$ has the limit $\int (f - g_1)^2 \varrho\, dx$ as $n \to \infty$. In general this limit is positive.

As an estimator of $Q(\hat{f}, f)$ we shall use

$$\hat{Q}(\hat{f}) = \int \varrho(x)(\hat{f}(x, \mathbf{X}))^2\, dx - \frac{2}{n(n-1)}\sum_{i \neq j} \varrho(X_i)\, g(X_i, X_j), \tag{2.5}$$

where the summation is performed over all the $n(n-1)$ pairs for which $i \neq j$. By direct computation we get, cf. (2.3),

$$E\hat{Q}(\hat{f}) = Q(\hat{f}, f), \tag{2.6}$$

i.e. $\hat{Q}(\hat{f})$ is an unbiased estimator of $Q(\hat{f}, f)$.

To assure that the variance of $\hat{Q}(\hat{f})$ is finite, we assume that $\varrho(X_1)\, g(X_1, X_2)$ and $\int \varrho(x)(g(x, X_1))^2 dx$ have finite second moments. Using Theorem 5.2 of Hoeffding (1948), we can then show that

$$\operatorname{Var} \hat{Q}(\hat{f}) = O\left(\frac{1}{n}\right) \tag{2.7}$$

if $n \to \infty$ with g fixed.

The estimator $\hat{Q}(\hat{f})$ is related to Schweder's estimator

$$\hat{\theta}_n = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} g(X_i, X_j)$$

of $\theta=\int f^2(x)\,dx$, see Section 3 of Schweder (1975), where an asymptotic formula for Var $\hat{\theta}_n$ is derived and the asymptotic normality of $\hat{\theta}_n$ is shown.

Let us now give some examples of evaluation of $\hat{Q}(\hat{f})$ for various choices of g.

Example 1. The histogram estimator. Suppose that $\varrho(x)\equiv 1$ and that $g(x,y)=1/h_k$ if x and y belong to the same interval I_k, while $g(x,y)=0$ if x and y belong to different intervals of a partition $\mathbf{I}=(I_k)$. We find that

$$\hat{Q}(\hat{f}_{\mathbf{I}})=\frac{2}{n(n-1)}\sum_k (N_k/h_k)-\frac{n+1}{n^2(n-1)}\sum_k (N_k^2/h_k).$$

In particular if $h_k=h$ for all k we get

$$\hat{Q}(\hat{f}_{\mathbf{I}})=\frac{2}{(n-1)h}-\frac{n+1}{n^2(n-1)h}\sum_k N_k^2. \tag{2.8}$$

Example 2. The normal kernel estimator. Put $K(x)=(1/\sqrt{2\pi})\exp(-x^2/2)$ and put $g(x,y)=(1/h)K((x-y)/h)$. Suppose further that $\varrho(x)\equiv 1$. A simple computation shows that

$$\hat{Q}(\hat{f})=A+B\sum_{i<j}[\exp(-\Delta_{ij}^2/4)-C\exp(-\Delta_{ij}^2/2)], \tag{2.9}$$

where

$$A=(2nh\sqrt{\pi})^{-1},\ B=(n^2h\sqrt{\pi})^{-1},$$
$$C=2\sqrt{2}\,n/(n-1) \text{ and } \Delta_{ij}=(X_i-X_j)/h.$$

Example 3. Orthogonal series estimators. Suppose that $\hat{f}_\alpha$ and g are given by (1.7) and (2.2). Substitution in (2.5) gives

$$\hat{Q}(\hat{f}_\alpha)=\sum_{k\in\alpha}\left[-\frac{n+1}{n-1}\hat{a}_k^2+\frac{2}{n(n-1)}\sum_{i=1}^n \varphi_k(X_i)^2\varrho(X_i)^2\right]. \tag{2.10}$$

In particular, for (φ_k) equal to the system of trigonometric functions specified before (1.9), $\alpha=\{0,1,\ldots,2k\}$, $\beta=\{0,1,\ldots,2k-2\}$ and $\varrho(x)\equiv 1$ we find that

$$\hat{Q}(\hat{f}_\alpha)-\hat{Q}(\hat{f}_\beta)=-\frac{2(n+1)}{(n-1)n^2}\sum_{i,j}\cos 2\pi k(X_i-X_j)+\frac{4}{n-1},$$

and we see that (1.9) is equivalent to $\hat{Q}(\hat{f}_\alpha)-\hat{Q}(\hat{f}_\beta)<0$.

For choice of histograms, kernel estimators or more generally estimators of type (2.1) the following method is suggested. Start with specification of a class $(\hat{f}_\alpha)_{\alpha\in\mathcal{A}}$ of possible estimators. Compute $\hat{Q}(\hat{f}_\alpha)$, $\alpha\in\mathcal{A}$, and either choose directly an $\hat{f}_\alpha$ which minimizes $\hat{Q}(\hat{f}_\alpha)$ or display $\hat{Q}(\hat{f}_\alpha)$ as a function of a suitable parameter such as h for (2.8) or (2.9) and choose an $\hat{f}_\alpha$ after inspection of this graph.

From (2.7) it follows that if we have a fixed finite number of possible estimators, then the probability that we choose one which minimizes $R^2(\hat{f}_\alpha,f)$ goes to one as $n\to\infty$. However, this is not a very satisfactory convergence theorem. Such a theorem should allow an infinite number of possible estimators or allow the number of elements in $\mathcal{A}$ to increase with n at a suitable rate.

Anyway, the usefulness of the suggested procedure must to a large extent be judged on its performance in examples. The studies in sections 5, 6 and 7 below give information in this respect.

3. Choice based on cross-validation

Let $(\hat{f}_\alpha)_{\alpha\in\mathcal{A}}$ be a family of density estimators. Suppose that we have a risk function that can be written on the form

$$L_n(\hat{f}_\alpha,f)=El_\alpha(X,\mathbf{X}), \tag{3.1}$$

where $X, X_1,\ldots,X_n$ are independent and identically distributed with density f. The expectation in (3.1) is taken with respect to the joint distribution of X and $\mathbf{X}=(X_1,\ldots,X_n)$. Further, $l_\alpha(x,\mathbf{x})$ is a real-valued function depending on $\alpha\in\mathcal{A}$, $x\in R$ and $\mathbf{x}\in R^n$. In particular $l_\alpha(x,\mathbf{x})$ shall not depend on the unknown density f. For instance, $l_\alpha(x,\mathbf{x})=-\log\hat{f}_\alpha(x,\mathbf{x})$ gives $L_n(\hat{f}_\alpha,f)=KL(\hat{f}_\alpha,f)$, see (1.4); and for a known weight function ϱ

$$l_\alpha(x,\mathbf{x})=\int(\hat{f}_\alpha(y,\mathbf{x}))^2\varrho(y)\,dy-2\hat{f}_\alpha(x,\mathbf{x})\varrho(x) \tag{3.2}$$

gives the risk function $Q(\hat{f}_\alpha,f)$ in (2.3). In (3.1) we can think of $l_\alpha(X,\mathbf{X})$ as the loss corresponding to

the use of $\hat{f}_\alpha$, based on $\mathbf{X}$, at prediction of a future observation X.

Let us consider a cross-validation estimator, cf. Stone (1974) and Geisser (1975), of $L_n(\hat{f}_\alpha, f)$. Let $\mathbf{X}_i$ be given by (1.10) and put

$$\hat{L}(\hat{f}_\alpha) = \frac{1}{n}\sum_{i=1}^{n} l_\alpha(X_i, \mathbf{X}_i). \tag{3.3}$$

It follows from (3.1) that $\hat{L}(\hat{f}_\alpha)$ is an unbiased estimator of $L_{n-1}(\hat{f}_\alpha, f)$.

To choose $\hat{f}_\alpha$ among $(\hat{f}_\alpha)_{\alpha\in\mathcal{A}}$ we can use $\hat{L}(\hat{f}_\alpha)$ in the same way as $\hat{Q}(\hat{f}_\alpha)$ in the previous section, i.e. we minimize $\hat{L}(\hat{f}_\alpha)$. For $l_\alpha(x, \mathbf{x}) = -\log \hat{f}_\alpha(x, \mathbf{x})$ corresponding to the Kullback-Leibler risk function we then get a method equivalent to maximization of (1.11).

Let us now suppose that $\hat{f}_\alpha$ is of the form (2.1) and that the risk function is (2.3). Then both the method of the previous and the method of the present section are applicable. Inserting l_α from (3.2) in (3.3) we find after a short computation an expression for $\hat{L}(\hat{f}_\alpha)$ which is very close to (2.5). In fact, if $\int \varrho(x)\, g(x, x_1)\, g(x, x_2)\, dx$ is bounded as a function of x_1 and x_2, the difference between $\hat{Q}(\hat{f}_\alpha)$ and $\hat{L}(\hat{f}_\alpha)$ is $O(1/n^2)$. Corresponding to a histogram estimator with constant interval length h we get for instance

$$\hat{L}(\hat{f}_1) = \frac{2n-1}{(n-1)^2 h} - \frac{1}{(n-1)^2 h}\sum_k N_k^2, \tag{3.4}$$

which should be compared with $\hat{Q}(\hat{f}_1)$ in (2.8). Suppose generally that we have a family $(\hat{f}_\alpha)$ of density estimators of type (2.1) and that we want to choose α by minimization of an estimator of the risk function (2.3). Though the difference between $\hat{Q}(\hat{f}_\alpha)$ and $\hat{L}(\hat{f}_\alpha)$ depends on α it turns out that the difference between the resulting density estimators is neglible for the examples in Sections 5, 6 and 7. On the other hand we shall see that use of the Kullback-Leibler risk function instead of the quadratic risk function can give quite large differences.

4. Kernel estimators for distributions on subintervals of the real line

Let us first suppose that we have a continuous distribution on the interval (l, ∞). As $K((x-X_i)/h)$ may be positive for $x<l$, the estimator (1.2) may place probability less than one on (l, ∞) and thus it has to be adjusted. Though several modifications are possible, let us for simplicity limit ourselves to

$$\hat{f}(x, l, \mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} [K_h(x-X_i) + K_h(x+X_i-2l)],\ x>l, \tag{4.1}$$

where $K_h(x) = h^{-1}K(x/h)$, which corresponds to reflection of the observations at l. Such reflection, which is discussed in more detail in Boneva, Kendall & Stefanov (1971), seems reasonable if $f(x)$ is approximately constant in the neighbourhood of $x=l$. It is easily verified that $\int K(x)\, dx = 1$ implies that the integral of (4.1) over (l, ∞) is one.

Consider the normal kernel K in Example 2 and the quadratic risk function estimator $\hat{Q}(\hat{f})$ in (2.5). Let $\hat{Q}_0$ denote the right member of (2.9). After some computations we find for $\hat{f}$ in (4.1)

$$\hat{Q}(\hat{f}) = \hat{Q}_0 + \hat{Q}_L \tag{4.2}$$

with

$$\hat{Q}_L = (A/n)\sum_{i=1}^{n} \exp\left(-P_i^2\right) + B\sum_{i<j} [\exp\left(-S_{ij}^2/4\right) - C\exp(-S_{ij}^2/2)],$$

where A, B and C are identical with A, B and C in (2.9), $P_i = (X_i - l)/\mathrm{h}$ and $S_{ij} = (X_i + X_j - 2l)/h$.

Similarly, if we have a continuous distribution on a finite interval (l, r), we can use the estimator

$$\hat{f}(x, l, r, \mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} [K_h(x-X_i) + K_h(x+X_i-2l) + K_h(x+X_i-2r)], \tag{4.3}$$

$l<x<r$, provided that $K(x)$ is negligible for $|x|>r-l$. For the normal kernel we get, provided that say $r-l>3h$, the approximation

$$\hat{Q}(\hat{f}) \approx \hat{Q}_0 + \hat{Q}_L + \hat{Q}_R \tag{4.4}$$

with $\hat{Q}_0$ and $\hat{Q}_L$ defined as before and

$$\hat{Q}_R = (A/n)\sum_{i=1}^{n} \exp\left(-R_i^2\right) + B\sum_{i<j} [\exp\left(-T_{ij}^2/4\right) - C\exp(-T_{ij}^2/2)],$$

where $R_i = (X_i - r)/h$ and $T_{ij} = (X_i + X_j - 2r)/h$.

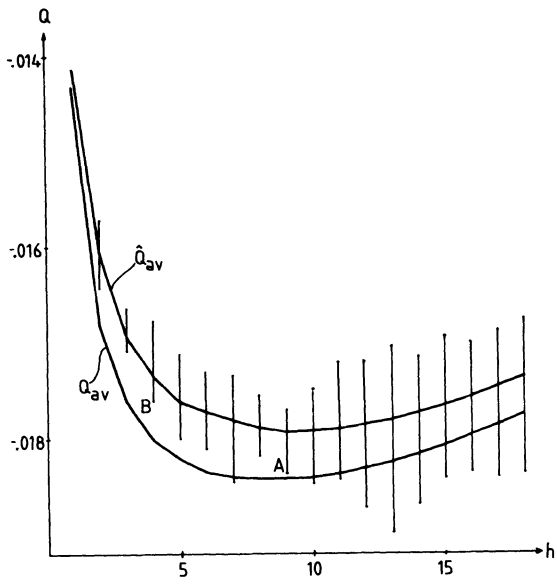

Fig. 1. Averages $\hat{Q}_{av}(h)$ of the risk function estimates $\hat{Q}(\hat{f}_{\mathbf{I}})$ for histograms, see (2.8), plotted as a function of h, $h=1, 2, \ldots, 18$, for Set no. 1. For each h-value the maximal and the minimal $\hat{Q}$-values are also shown, connected with a vertical line. The lower curve gives the averages $Q_{av}(h)$ of the true risk function (5.3) for $h=1,2,\ldots,18$. The histogram corresponding to the point marked A is shown in Fig 2 and the histogram corresponding to B can be found in Fig. 4. The h-axis meets the Q-axis at (5.4).

5. Simulations with a lognormal distribution

To study the methods of choosing smoothing degree described in the previous sections, pseudo-random variables with the density

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma x}\exp\left[-\frac{1}{2\sigma^2}\left(\log\frac{x}{\mu}\right)^2\right], \tag{5.1}$$

with $\mu=50$ and $\sigma=0.3$, were generated by means of a computer routine. Fifteen independent sets, each consisting of 200 independent random variables, were generated and the results of the computations on these sets are shown in Table 1. For Set no. 1 more detailed results are given in Figs. 1–4 and for Set no. 2 some of the corresponding graphs are shown in Figs. 5 and 6.

To determine a histogram or equivalently a partition $\mathbf{I}=(I_k)$, cf. (1.1), with constant interval length h, the following procedure was used. A partition is determined by two real numbers b and h, e.g. for $k=0, \pm1, \ldots,$ we can put $I_k=(b+(k-1)h,\ b+kh]$. It was decided to consider only pairs (b, h) of the form

$$(b, h)=(b_0+ih_0, jh_0), \tag{5.2}$$

where i and j are integers, $j>0$, and b_0 and h_0 are fixed real numbers, $h_0>0$. In the present case $b_0=0$ and $h_0=1$ were used. For a given $h=jh_0$ there are j different partitions determined by $b=b_0+ih_0$, $i=0, 1, \ldots, j-1$. Let $\hat{Q}_{av}(h)$, $\hat{Q}_{min}(h)$ and $\hat{Q}_{max}(h)$ denote the average, the minimum and the maximum of the $\hat{Q}$-values computed from (2.8) for these j partitions. For Set no. 1 the points $(h,\ \hat{Q}_{av}(h))$ for $h=1,2,\ldots,18$ are shown in Fig. 1 joined with straight-line segments. The curve has a minimum for $\hat{h}=9$. The corresponding $\hat{h}$-values for all the fifteen data sets are shown in the second column of Table 1. The points $(h,\ \hat{Q}_{min}(h))$ and $(h,\ \hat{Q}_{max}(h))$ joined with vertical lines are also shown in Fig. 1. A histogram is now chosen as follows. We choose $\hat{h}$ as the h-value which minimizes $\hat{Q}_{av}(h)$. For this h-value, choose that partition which gives the smallest $\hat{Q}$-value. For Set no. 1 we then get the partition corresponding to the point marked A in Fig. 1 and the corresponding histogram is shown in Fig. 2, where also the true density is drawn. A more rugged histogram corresponding to the point B in Fig. 1 is shown in Fig. 4 together with the true density and a kernel estimate. As regards terminology, the convention we use here is to write estimate for observed values of estimators also with density estimators and risk function estimators, for which an estimate thus is a function.

In Fig. 1 is further shown the graph of $Q_{av}(h)$ determined in the following way. From (2.4) one finds

$$Q(\hat{f}, f)=\frac{1}{nh}\left[1-(n+1)\sum_k\left(\int_{I_k} f\,dx\right)^2\right]. \tag{5.3}$$

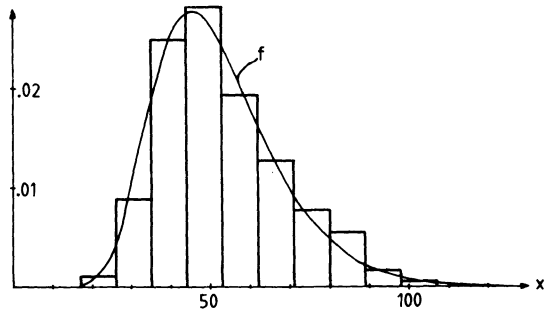

Fig. 2. Histogram estimate $\hat{f}_{\mathbf{I}}$ for Set no. 1 with width 9 corresponding to the point A in Fig. 1 and the true density f from (5.1).

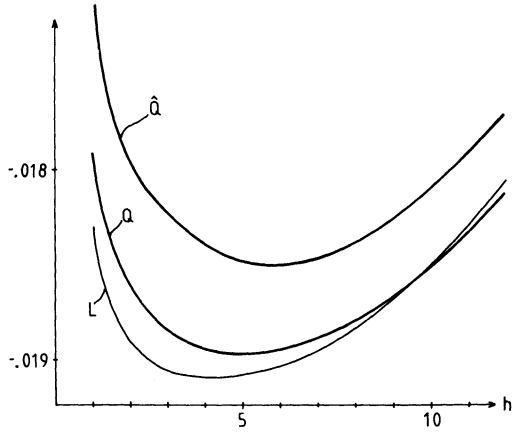

Fig. 3. Risk function estimate $\hat{Q}$ from (2.9) and loss function L from (5.6) for Set no. 1 as functions of h. The figure also shows the true risk function Q computed from (5.5). The h-axis meets the Q-axis at (5.4).

Then $Q_{av}(h)$ is determined as an average of Q-values from (5.3) for different partitions with the same h, in the same way as $\hat{Q}_{av}(h)$ was determined by $\hat{Q}$-values from (2.8). In Fig. 1 the h-axis cuts the Q-axis at

$$Q = -\int f^2\,dx = -0.019234. \tag{5.4}$$

Thus the vertical distance between the curve of Q_{av} and the h-axis corresponds to the risk function in (1.3) with $p=2$ and $\varrho\equiv 1$. It can be seen from Fig. 1 that the $\hat{h}$-values chosen for histograms in Table 1 all give an average risk close to the minimum.

To choose a kernel estimate the normal kernel of Example 2 was used. Fig. 3 shows the graph of the risk function estimate $\hat{Q}$ determined from (2.9) as a function of h. The minimum occurs for $\hat{h}=5.85$. The corresponding kernel estimate is shown in Fig. 4. The $\hat{h}$-values thus obtained for the fifteen data sets are shown in column 3 of Table 1. The curve marked Q in Fig. 3 is the risk function $Q(\hat{f},f)$ of (2.3) as a function of h. This function was evaluated by numerical integration as follows. Put $A(x,h)=\exp(-(x/h)^2)$ and

$$B(x,y) = \left(1-\frac{1}{n}\right)A(x-y,2h)-2\sqrt{2}\,A(x-y,\sqrt{2}\,h).$$

For the normal kernel we then get from (2.4)

$$Q(\hat{f},f) = \frac{1}{2h\sqrt{\pi}}\left\{\frac{1}{n}+\int\int f(x)f(y)B(x,y)\,dx\,dy\right\}. \tag{5.5}$$

The double integral in (5.5) was evaluated numerically by a computer routine which used bicubic spline approximation.

The curve marked L in Fig. 3 is a true loss function defined by

$$L(h) = \int(\hat{f}-f)^2\,dx - \int f^2\,dx. \tag{5.6}$$

As in Fig. 1 the h-axis meets the Q-axis at (5.4). The fourth column of Table 1 shows, for each data set, the h-value h_{opt} which minimizes (5.6). The fifth column gives

$$L^2(\hat{h}) = 1\,000\int(\hat{f}-f)^2\,dx, \tag{5.7}$$

with $\hat{h}$ from the third column used in $\hat{f}$. The sixth and seventh columns show the h-value $\hat{h}$ which minimizes $\hat{L}(\hat{f})$ in (3.3) with $l(x,\mathbf{x})=-\log\hat{f}(x,\mathbf{x})$ and the h-value h_{opt} which minimizes

$$L(\hat{f}) = -\int f(x)\log\hat{f}(x)\,dx. \tag{5.8}$$

As before, $\hat{f}$ is the kernel estimate with a normal kernel. The graphs of $\hat{L}(\hat{f})$ and $L(\hat{f})$ as functions of h were both similar to the curves in Fig. 3.

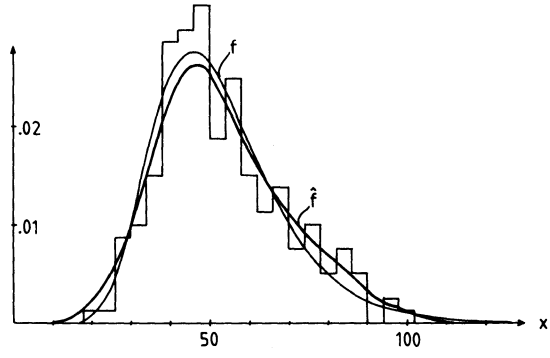

Fig. 4. Kernel density estimate $\hat{f}$ for Set no. 1 with the normal kernel and $h=5.85$ corresponding to the minimum of the curve $\hat{Q}$ in Fig. 3. The diagram also shows the true density f, see (5.1), and the histogram with width 4 corresponding to B in Fig. 1.

Table 1. *Results from computations on 15 synthetic data sets from a lognormal distribution*
Each set consists of $n=200$ observations. The variables are explained in the text

Set no.	Histogram choice $\hat{h}$	Kernel choice: Quadratic risk $\hat{h}$	Quadratic risk h_{opt}	Quadratic risk $L^2(\hat{h})$	Kullback-Leibler $\hat{h}$	Kullback-Leibler h_{opt}	Δ
1	9	5.85	4.37	.18	5.22	7.06	3
2	7	3.80	6.02	.40	5.06	5.74	9
3	10	6.12	4.78	.41	5.83	6.24	9
4	9	5.80	4.48	.30	5.47	5.77	11
5	8	4.96	5.29	.09	5.18	6.01	4
6	6	3.03	5.85	.58	7.73	5.30	28
7	7	5.12	6.33	.15	4.02	7.71	3
8	9	6.12	3.71	.23	5.40	5.97	9
9	9	4.99	5.40	.28	9.15	5.48	40
10	6	3.02	6.99	.99	5.28	5.74	15
11	8	3.85	5.02	.21	7.14	5.89	27
12	9	5.12	4.85	.21	7.56	5.35	25
13	9	4.99	4.70	.19	4.59	5.61	2
14	9	5.45	4.78	.28	4.36	5.79	2
15	7	3.39	5.27	.21	3.01	7.34	0
Median	9	4.99	5.02	.23	5.28	5.79	
Mean	8.13	4.77	5.19	.31	5.67	6.07	
s	1.25	1.09	0.84	.22	1.60	0.73	

Let us see how the $\hat{h}$-values given in Table 1 compare with h-values, that minimize the corresponding risk functions and the asymptotic h-values from (1.5) and (1.6). For histogram choice, the $\hat{h}$-values in Table 1 have mean 8.1 and standard deviation 1.2. With h restricted to integers, the risk function Q_{av}, cf. Fig. 1, has minimum at $h=8$ with $h=9$ giving a slightly larger risk. From (1.6) we find $h_{asy}=8.25$. Further, for a normal distribution (1.6) gives

$$h_{asy} = (24\sqrt{\pi})^{1/3} n^{-1/3} \sigma. \tag{5.9}$$

Assuming that the density that we estimate is close to a normal density, we can replace σ in (5.9) with the empirical standard deviation s. For the date sets 1–15 we then get h_{asy}-values ranging from 8.8 to 10.9, i.e. slightly too large width estimates.

For kernel choice with a quadratic risk function, the $\hat{h}$-values in Table 1 have mean 4.8 and standard deviation $s=1.1$. The true risk function $Q(\hat{f},f)$, see curve Q in Fig. 3, has minimum at $h=5.11$, while (1.5) gives $h_{asy}=4.78$. For a normal density, we obtain from (1.5)

$$h_{asy} = (4/3)^{1/5} n^{-1/5} \sigma. \tag{5.10}$$

With σ replaced by s, we get for the 15 data sets h_{asy}-values ranging from 5.41 to 6.69.

For kernel choice with the Kullback-Leibler risk function, the $\hat{h}$-values have mean 5.7 and standard deviation $s=1.6$. The theoretical risk function (1.4) can be estimated as the mean of the fifteen $L(\hat{f})$-functions computed from (5.8). To correct for bias and, in particular, to get an accuracy estimate, jackknifing was used in the estimation of h_{KL}, the minimizing h-value. In the jack-knife procedure fifteen h-values were computed, each time with exclusion of one of the data sets. The resulting h_{KL}-estimate is 6.19 with standard error 0.21. We see that about the same result is obtained if we use the fifteen h_{opt}-values in column 7 of Table 1. The corresponding mean is 6.07 with standard error $0.73/\sqrt{15}=0.19$.

Let us now compare the different $\hat{h}$-values in Table 1. It is seen that those data sets, that have small $\hat{h}$-values for histogram choice, also tend to have small $\hat{h}$-values for choice of a kernel with a quadratic risk function, while no similar connection can be seen with the $\hat{h}$-values in column 6. To find out what affects the $\hat{h}$ variations for the Kullback-Leibler risk function the difference between the largest and the next largest observation was computed. This difference Δ, rounded to the nearest integer, is shown in the last column of Table 1. For

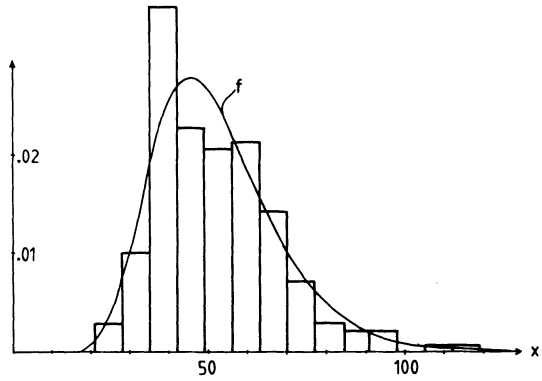

Fig. 5. Histogram estimate $\hat{f}_I$ for Set no. 2 with interval length 7 and the true density f from (5.1). The histogram estimate is determined in the same way as the histogram in Fig. 2.

the present distribution an isolated observation will typically mean a large Δ-value, and it is seen that it is precisely those data sets that give a large $\hat{h}$-value in column 6. Thus we get that choice of $\hat{h}$ by use of the Kullback-Leibler risk function estimator seems to be largely affected by isolated observations.

One further conspicuous feature of Table 1 is that the variations in columns 3 and 4, and similarly the variations in columns 6 and 7, are negatively correlated. Thus, for Set no. 10 we get with a quadratic risk function a small $\hat{h}$-value, while the optimal h-value h_{opt} for this set is large. At first this effect seems paradoxical, but it may be explained heuristically in the following way. If we, for a smooth distribution, by chance get an irregular data set, a large h-value is needed to smooth the empirical

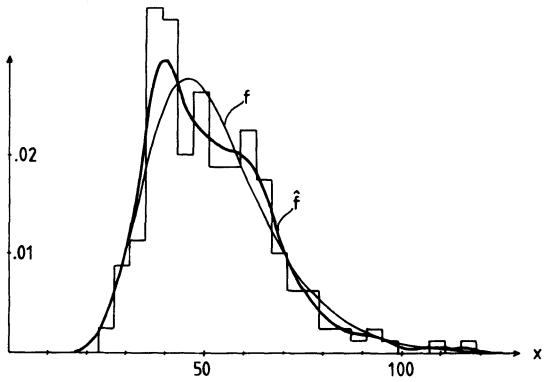

Fig. 6. Kernel density estimate $\hat{f}$ for Set no. 2 with the normal kernel and $h=3.80$. The diagram also shows the true density f from (5.1) and a histogram with width 4.

distribution sufficiently. On the other hand, the irregularity of the observed distribution suggests that the underlying distribution has a density with numerically large second derivative values, a feature that calls for a small h-value, cf. (1.5).

Let us finally remark that for Set no. 1, used in Figs. 1–4, the density estimates are comparatively good, which can be seen from column 5. Measured by $L^2(\hat{h})$, Set no. 1 gives the third best fit. On the other hand, Set no. 2 used in Figs. 5 and 6 has rank 12 among the 15 sets.

6. Simulations with a mixture of lognormal distributions

To study choice of smoothing degree by use of risk function estimation for a more complicated density than (5.1), pseudo-random variables with the density

$$f(x)=\sum_{i=1}^{3}\frac{p_i}{\sqrt{2\pi}\,\sigma_i x}\exp\left[-\frac{1}{2\sigma_i^2}\left(\log\frac{x}{\mu_i}\right)^2\right], \qquad (6.1)$$

with $p_1=0.80$, $p_2=p_3=0.10$, $\mu_1=50$, $\mu_2=65$, $\mu_3=80$, $\sigma_1=0.3$ and $\sigma_2=\sigma_3=0.04$, were generated. Ten independent sets, called Set no. 16–Set no. 25, each consisting of 400 independent random variables, were generated and results of computations on these sets are shown in Table 2. For Set no. 16 graphs similar to Figs. 1–4 are shown in Figs. 7–10. According to the measure in the fifth column of Table 2 Set no. 16 gives the second best fit among the 10 data sets. The h-axis in Figs. 7 and 9 meets the Q-axis at $Q=-\int f^2\,dx=-0.01732$.

Let us make an analysis of Table 2 similar to the analysis of Table 1 in Section 5. For histogram choice the $\hat{h}$-values in the second column of Table 2, determined as in Section 5 except that $h_0=0.5$, conform well with the true risk function average Q_{av}, which has minimum for $h=5$, cf. Fig. 7. The asymptotic h-value from (1.6), $h_{asy}=4.66$, is slightly smaller. Similarly, the $\hat{h}$-values in the third column of Table 2 are consistent with the true risk function Q, which has minimum for $h=2.38$, cf. Fig. 9. The asymptotic h-value from (1.5) is $h_{asy}=1.94$.

The true distribution (6.1) is far from a normal distribution, and hence we cannot expect (5.9) and (5.10) with σ replaced by s to give reasonable estimates of h. In fact, if we use that procedure we get too large h-values: for data sets 16–25 we find h-

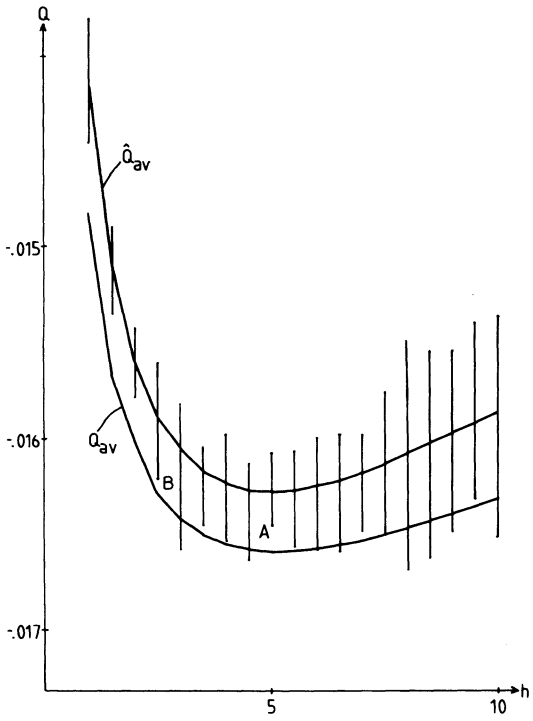

Fig. 7. Averages $\hat{Q}_{av}(h)$ of the risk function estimates $\hat{Q}(\hat{f}_1)$ for histograms, see (2.8), plotted as a function of h, $h=1.0, 1.5, 2.0, \ldots, 10$, for Set no. 16. For each h-value the maximal and the minimal $\hat{Q}$-values are also shown, connected with a vertical line. The figure further shows the averages Q_{av} of the risk function (5.3) for the same h-values. The histograms corresponding to the points A and B are drawn in Figs. 8 and 10.

values between 7.69 and 8.53 for histograms and h-values between 5.19 and 5.75 for kernel estimates.

Several of the variables in Table 1 showed clear co- and contravariation. Similar tendencies are indicated in Table 2 but they are less pronounced.

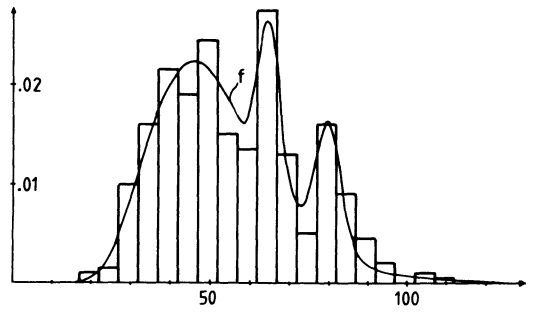

Fig. 8. Histogram estimate $\hat{f}_1$ for Set no. 16 with width 5 corresponding to the point A in Fig. 7 and the true density f from (6.1).

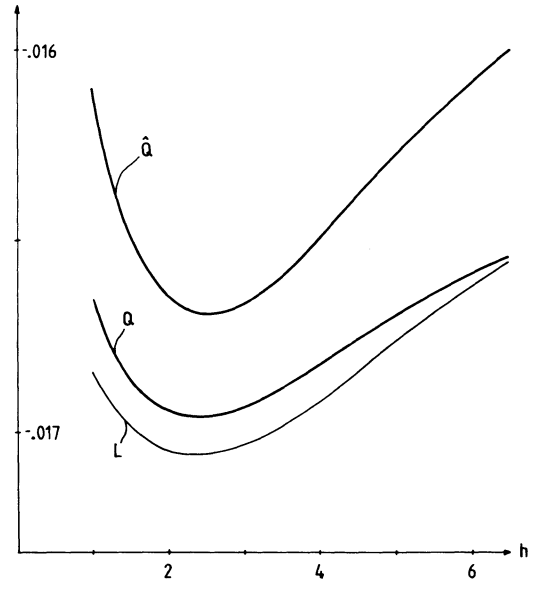

Fig. 9. Risk function estimate $\hat{Q}$ from (2.9) and loss function L from (5.6) for Set no. 16 as functions of h. The figure also shows the true risk function Q computed from (5.5).

The results from the simulations described in Table 1, Table 2 and Figs. 1–10 suggest that choice of smoothing degree by risk function estimation is a practical and reasonably effective method. From Figs. 1, 3, 7 and 9 it is seen that the risk function estimates can have rather large level errors, but that the form and the location of minima are fairly well estimated, and that is the essential point in estimation of smoothing degree. In the next section we shall see how the methods perform in an example with real data.

7. Analysis of the coal-mining diasters point process

Maguire, Pearson & Wynn (1952) give data on 109 intervals in days between successive coal-mining diasters in Great Britain for the period 1875–1951, see also Cox & Lewis (1966, p. 4). An extended and corrected data set is given in Jarrett (1979), which the author became aware of after the computations described below were made.

Consider generally a time-dependent Poisson process $(N_t)_{t\geq 0}$, which has increments of size one at $T_1<T_2<\ldots$, and with rate function λ defined by $EN_t = \int_0^t \lambda(s)\,ds$. To estimate λ, a natural procedure would be to fix a time T and observe N_t, $0<t<T$. Dividing the interval $(0, T)$ into subintervals of

Table 2. *Results from computations on 10 synthetic data sets from a mixture of lognormal distributions*
Each set consists of $n=400$ observations. The variables are explained in the text

Set no.	Histogram choice $\hat{h}$	Kernel choice: Quadratic risk $\hat{h}$	Quadratic risk h_{opt}	Quadratic risk $L^2(\hat{h})$	Kullback-Leibler $\hat{h}$	Kullback-Leibler h_{opt}	Δ
16	5	2.52	2.41	.25	2.48	4.65	2
17	4.5	2.21	2.10	.13	2.02	5.02	1
18	5	2.47	2.51	.29	3.64	4.08	14
19	5.5	2.77	2.46	.36	3.65	3.91	8
20	5.5	2.71	2.92	.63	2.40	4.78	3
21	6.5	3.29	2.12	.43	4.51	4.86	13
22	6	2.82	2.41	.39	3.99	3.66	14
23	3.5	1.34	2.51	.45	4.83	3.04	23
24	7.5	5.00	2.18	.72	7.73	6.28	41
25	5	2.05	2.34	.46	3.20	4.81	5
Median	5.25	2.62	2.41	.41	3.64	4.72	
Mean	5.40	2.72	2.40	.41	3.85	4.51	
s	1.10	0.96	0.24	.17	1.64	0.89	

length h we can use the histogram estimator $\hat{\lambda}(t)=(N_{kh}-N_{(k-1)h})/h$, $(k-1)h<t<kh$, or we can use a kernel estimator modified as in (4.3) and put

$$\hat{\lambda}(t)=(1/N_T)\sum_{i=1}^{N_T}[K_h(t-T_i)+K_h(t+T_i)+K_h(t+T_i-2T)], \tag{7.1}$$

where $K_h(t)=(1/h)K(t/h)$. For the present data we shall use a slightly different procedure and condition on T_{109} as in Barnard (1953).

Regard $T_1,\dots,T_n$, $n=108$, as ordered observations of a sample from a distribution on $(0, T)$,

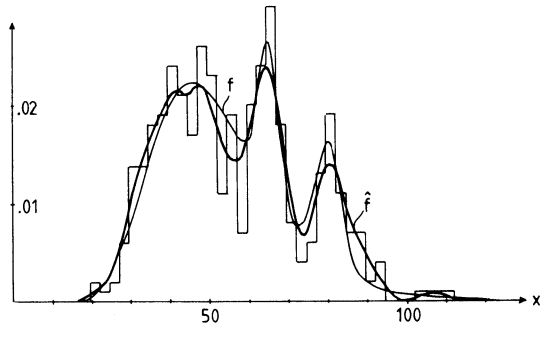

Fig. 10. Kernel density estimate $\hat{f}$ for Set no. 16 with the normal kernel and $h=2.52$ corresponding to the minimum of the curve $\hat{Q}$ in Fig. 9. The diagram also shows the true density f from (6.1) and the histogram with width 2.5 corresponding to B in Fig. 9.

where $T=T_{109}=26263$ days. We shall estimate the probability density

$$f(t)=\lambda(t)/\int_0^T \lambda(s)\,ds,\ 0<t<T.$$

From an estimate $\hat{f}$ of f we can obtain an estimate $\hat{\lambda}$ of λ. One could try to take account of the actual observation scheme used, but for simplicity we shall just put

$$\hat{\lambda}(t)=109\,\hat{f}(t),\ 0<t<T. \tag{7.2}$$

Let us first consider histogram estimators corresponding to partitions $\mathbf{I}=\mathbf{I}_m$ of $(0, T)$ into m subintervals of equal length $h=T/m$ for integer $m\geq 1$. The risk function estimate $\hat{Q}(\hat{f}_{\mathbf{I}})$ of (2.8) is shown in Fig. 11 as a function of m. The histogram with the smallest risk function estimate (A in Fig. 11) is shown in the upper diagram of Fig. 12 and the histogram with the next smallest risk function estimate (B in Fig. 11) is shown in the lower diagram of Fig. 12. The message of the upper histogram with five subintervals is clear: the disaster rate for the first fifth of the period is considerably larger than the rate for the remaining part. The lower histogram with nine subintervals is more difficult to interpret and indicates some oscillations of the disaster rate in addition to the decrease from a high initial rate.

For the kernel estimator (4.3) with $l=0$ and

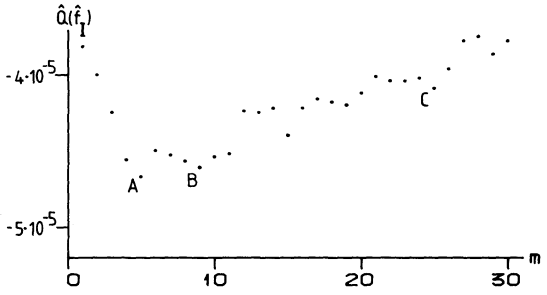

Fig. 11. Risk function estimates $\hat{Q}(\hat{f}_I)$ from (2.8) for the coal-mining disasters point process plotted against the number m of subintervals of **I**. The histograms corresponding to A and B are shown in Fig. 12 and the histogram corresponding to C can be found in Fig. 14.

$r=26263$ the quadratic risk function estimate $\hat{Q}(\hat{f})$ from (4.4) is shown in Fig. 13 as a function of h marked QR. The Kullback-Leibler risk function estimate $\hat{L}(\hat{f})$ from (3.3) for the same kernel estimator is shown as the curve marked KL. The two curves have minima for $h=2028$ and $h=1804$. The corresponding two kernel estimates are very close and only that one with $h=2028$ is shown in Fig. 14. There we also see the histogram estimate with 25 subintervals corresponding to C in Fig. 11. A comparison of the kernel estimate and the histogram indicates that the kernel estimate gives a suitable smoothing of the original data.

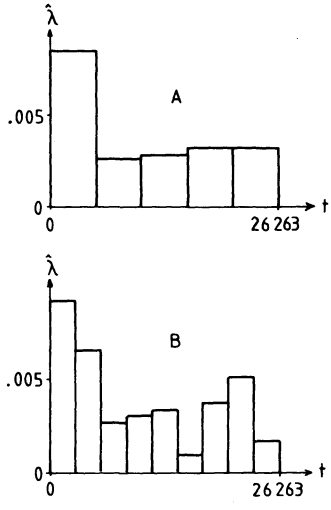

Fig. 12. Histogram estimates $\hat{\lambda}=109\hat{f}_I$ for the rate function of the coal-mining disasters point process. The histograms correspond to the points marked A (upper diagram) and B (lower diagram) in Fig. 11.

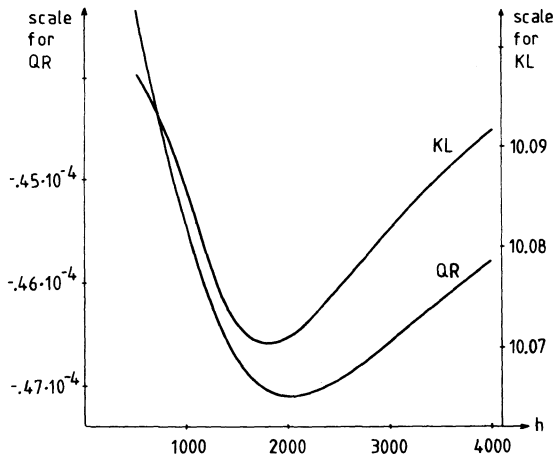

Fig. 13. Risk function estimates $\hat{Q}(\hat{f})$ from (4.4) for the quadratic risk function, marked QR, and $\hat{L}(f)$ from (3.3) for the Kullback-Leibler risk function, marked KL, for the coal-mining disasters point process. The risk function estimates are shown as functions of h.

For comparison a different type of rate function estimate is also shown in Fig. 14,

$$\hat{\lambda}(t)=\exp(\hat{\alpha}+\hat{\beta}t+\hat{\gamma}t^2),$$

with $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\gamma}$ determined by the method described in Cox & Lewis (1966, pp. 39–43). Excluding interval No. 57 with length 871, the remaining 108 intervals were added in groups of four. Let $Y_1,\ldots,Y_{27}$ denote the logarithms of these sums and let $t_1,\ldots,t_{27}$ denote the mid-points of the corresponding time intervals. A quadratic polynomial in t was fitted to the Y-values by the usual least squares method. Via Eq. (2.2.15) in Cox & Lewis (1966) an estimate $\hat{\lambda}(t)$ of $\lambda(t)$, $0<t<T$, where $T=26263$, was obtained and this rate function estimate is shown in Fig. 14 as the curve marked E.

8. Discussion and related topics

The risk function estimators described in the present paper can be used in two ways. Either as a tool suggesting which histograms or kernel estimates that are worth a closer look, or to construct a fully specified procedure for choosing a density estimate from a given set of estimates. In the last case the straightforward procedure is to choose the density estimate which minimizes the risk function estimate. Such a procedure has at least in principle the advantage over more subjective methods that it can

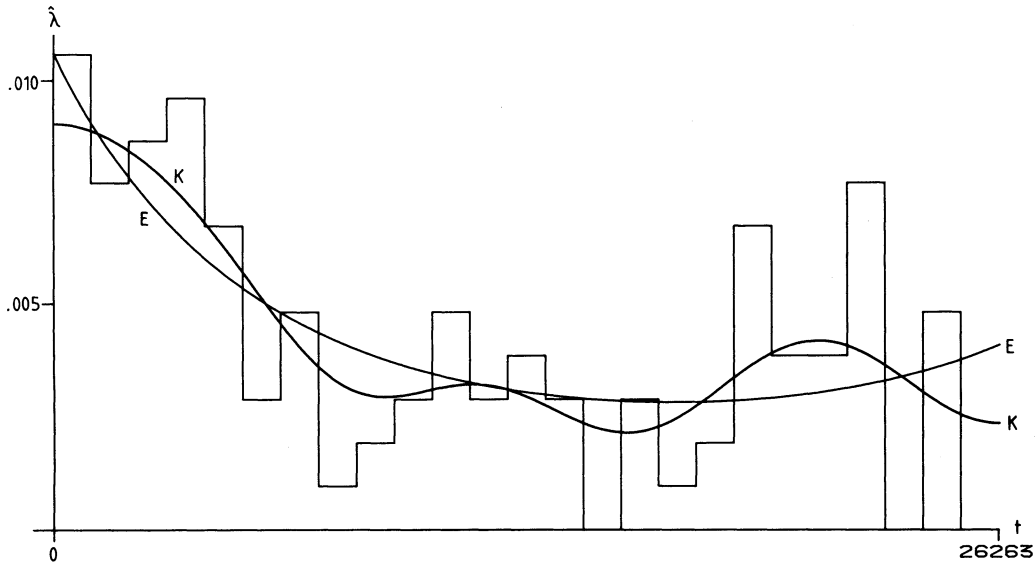

Fig. 14. Kernel estimate (marked *K*) of the rate function of the coal-mining disasters point process with $h=2\,028$ corresponding to the minimum of the risk function estimate *QR* in Fig. 13. The diagram also shows the histogram corresponding to *C* in Fig. 11 with 25 subintervals and an estimate (marked *E*) of the form $\hat{\lambda}(t)=\exp(\hat{\alpha}+\hat{\beta}t+\hat{\gamma}t^2)$.

be studied by simulation and by theoretical analysis of its asymptotic properties.

To plot the risk function estimates, cf. the $\hat{Q}$-curves in Figs. 3 and 9, and the *QR*- and *KL*-curves in Fig. 13, it is usually sufficient to compute the risk function estimates for a set of equidistant h-values over a suitable range. The spacings $\Delta h=0.5$, $\Delta h=0.25$ and $\Delta h=250$ were used in Figs. 3, 9 and 13, respectively. Suppose that among the computed risk function estimates, the minimum occurs for $h=h_0$. Then the minimizing h-value can be estimated by quadratic interpolation, using the risk function estimates for $h=h_0$ and $h=h_0\pm\Delta h$, or some further risk function estimates in the neighbourhood of h_0 can be computed. A procedure, which starts with a set of equidistant h-values over reasonably large range, should give a fair protection against local minima. For the data sets described in the present paper, no local minima were found for the risk function estimates corresponding to kernel estimators, but some were found for $\hat{Q}_{av}$, used to choose histogram estimators, among Sets 1–25.

It may be remarked that for a set of potato yield data, consisting of 240 observations in the range 109 to 176 with mean 134 and standard deviation 12, the author found global minima of $\hat{Q}$ at $h=4.9$ and of the Kullback-Leibler risk function estimate $\hat{L}$ at $h=3.8$ for normal kernel etimators. In addition local minima of both risk function estimates were found in the neighbourhood of $h=1.5$. In such cases it might be useful to plot the kernel estimates both for the global and the local minima. It may also be remarked that if we have rounded data, as in the example just mentioned, where the observations were rounded to integers, we may get an irregular behaviour of the risk function estimates for very small h-values due to the rounding. In the example referred to, that effect occurred for h of the order 0.1.

The treatment in this paper has been confined to distributions on the real line. It should be possible to generalize, in a relatively straightforward way, the methods both for histograms and kernel estimators to distributions in several dimensions.

In Section 7 estimation of the rate function of a time-dependent Poisson process is discussed. A natural generalization is to study estimation of the functions (α_i) for a multivariate point process with a multiplicative conditional intensity

$$\Lambda_i(t) = \alpha_i(t)\, Y_i(t), \tag{8.1}$$

where (Y_i) are observed stochastic processes, see Aalen (1978). In particular, one could consider kernel smoothing of Aalens non-parametric estimator of the integrals of (α_i). The model (8.1) can be used for hazard rate estimation in life tests with a general

type of censoring and also for estimation of the transition probabilities of time-dependent Markov chains, cf. Aalen & Johansen (1978), where a different type of smoothing is discussed.

Acknowledgement

This work was supported in part by the Danish Agricultural and Veterinary Research Council. I am further indebted to the referees for valuable references and suggestions.

References

Aalen, O. (1978). Nonparametric inference for a family of counting processes. *Ann. Statist.* **6,** 701–726.

Aalen, O. & Johansen, S. (1978). An empirical transition matrix for non-homogeneous Markov chains based on censored observations. *Scand. J. Statist.* **5,** 141–150.

Barnard, G. A. (1953). Time intervals between accidents—a note on Maguire, Pearson and Wynn's paper. *Biometrika* **40,** 212–213.

Boneva, L. I., Kendall, D. G. & Stefanov, I. (1971). Spline transformations: Three new diagnostic aids for the statistical analyst (with discussion). *J. Roy. Statist. Soc. Ser. B* **33,** 1–70.

Čencov, N. N. (1962). Evaluation of an unknown distribution density from observations. *Soviet Math.* **3,** 1559–1562.

Cox, D. R. & Lewis, P. A. W. (1966). *The statistical analysis of series of events*. Methuen, London.

Craven, P. & Wahba, G. (1979). Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by generalized cross-validation. *Numer. Math.* **31,** 377–403.

Devroye, L. P. & Wagner, T. J. (1979). Distribution-free inequalities for the deleted and hold out error estimates. *IEEE Trans. Information Theory* **IT-25,** 202–207.

Duin, R. P. W. (1976). On the choice of smoothing parameters for Parzen estimators of probability density functions. *IEEE Trans. Computers* **C-25,** 1174–1179.

Fryer, M. J. (1977). A review of some non-parametric methods of density estimation. *J. Inst. Math. Appl.* **20,** 335–354.

Geisser, S. (1975). The predictive sample reuse method with applications. *J. Amer. Statist. Assoc.* **70,** 320–328.

Habbema, J. D. F., Hermans, J. & Van den Broek, K. (1974). A stepwise discriminant analysis program using density estimation. *Compstat 1974, Proceedings in computational statistics,* pp. 101–110. Physica Verlag, Wien.

Habbema, J. D. F. & Hermans, J. (1977). Selection of variables in discriminant analysis by *F*-statistic and error rate. *Technometrics* **19,** 487–493.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Ann. Math. Statist.* **19,** 293–325.

Jarret, R. G. (1979). A note on the intervals between coalmining disasters. *Biometrika* **66,** 191–193.

Kronmal, R. A. & Tarter, M. (1968). The estimation of probability densities and cumulatives by Fourier series methods. *J. Amer. Statist. Assoc.* **63,** 925–952.

Kullback, S. (1959). *Information theory and statistics*. Wiley, New York.

Maguire, B. A., Pearson, E. S. & Wynn, A. H. A. (1952). The time intervals between industrial accidents. *Biometrika* **39,** 168–180.

Nadaraya, E. A. (1974). On the integral mean square error of some nonparametric estimates for the density function. *Theor. Probability Appl.* **19,** 133–141.

Révész, P. (1968). *The laws of large numbers*. Academic Press, New York.

Rogers, W. H. & Wagner, T. J. (1978). A finite sample distribution-free performance bound for local discrimination rules. *Ann. Statist.* **6,** 506–514.

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.

Schuster, E. F. & Gregory, G. G. (1978). Choosing the shape factor(s) when estimating a density. *Inst. Math. Statist. Bull.* **7,** 292.

Schweder, T. (1975). Window estimation of the asymptotic variance of rank estimators of location. *Scand. J. Statist.* **2,** 113–126.

Scott, D. W. (1979). On optimal and data-based histograms. *Biometrika* **66,** 605–610.

Silverman, B. W. (1978). Choosing the window width when estimating a density. *Biometrika* **65,** 1–11.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *J. Roy. Statist. Soc. Ser. B* **36,** 111–147.

Tapia, R. A. & Thompson, J. R. (1978). *Nonparametric probability density estimation*. Johns Hopkins Univ. Press, Baltimore.

Tarter, M. E. & Kronmal, R. A. (1976). An introduction to the implementation and theory of nonparametric density estimation. *Amer. Statistician* **30,** 105–112.

Wahba, G. (1977). Optimal smoothing of density estimates. *Classification and clustering* (ed. J. Van Ryzin), pp. 423–458. Academic Press, New York.

Wahba, G. & Wold, S. (1975). A completely automatic French curve: Fitting spline functions by cross validation. *Comm. Statist.* **4,** 1–17.

Walter, G. G. & Blum, J. R. (1979). Probability estimation using delta sequences. *Ann. Statist.* **7,** 328–340.

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

Wertz, W. (1978). *Statistical density estimation: a survey*. Vandenhoeck & Ruprecht, Göttingen.

Wertz, W. & Schneider, B. (1979). Statistical density estimation: a bibliography. *Internat. Statist. Rev.* **47,** 155–175.

Whittle, P. (1958). On the smoothing of probability density functions. *J. Roy. Statist. Soc. Ser. B* **20,** 334–343.

Mats Rudemo
Department of Mathematics and Statistics
The Royal Veterinary and Agricultural University
Thorvaldsensvej 40
1871 Copenhagen V
Denmark