Non-parametric Function Fitting

By M. B. PRIESTLEY and M. T. CHAO†

University of Manchester Institute of Science and Technology

[Received August 1971. Revised February 1972]

SUMMARY

In this note we consider the problem of fitting a general functional relationship between two variables. We require only that the function to be fitted is, in some sense, "smooth", and do not assume that it has a known mathematical form involving only a finite number of unknown parameters.

Keywords: CURVE FITTING; REGRESSION; DENSITY FUNCTION ESTIMATION; SMOOTHING WINDOWS; MEAN-SQUARE ERRORS

1. INTRODUCTION

IN a recent paper Tischendorf and Chao (1970) discussed the problem of estimating the indefinite integral of an arbitrary function, given only observations on the function at a discrete set of points. This problem arose in connection with a method of estimating a phase spectrum, $\phi(\omega)$, given m observations on the derivative, $\phi'(\omega)$, at a discrete set of frequencies, $\omega_1, \ldots, \omega_m$. Here, the points $\{\omega_i\}$ are equally spaced, and since the frequency bandwidth may be assumed to be finite it is convenient to take the domain of the function ϕ to be the interval $(0, 1)$.

Let τ_i denote the observed value of $\phi'(\omega_i)$. The above authors note that a crude estimate of $\phi(\omega)$ is simply

$$\hat{\phi}(\omega) = m^{-1} \sum_{i=1}^{j} \tau_i, \quad \omega_j \leqslant \omega < \omega_{j+1}, \tag{1.1}$$

but they suggest that an improved estimate of $\phi(\omega)$ may be obtained by first constructing an estimate of $\phi'(\omega)$ for *all* ω in $(0, 1)$, using a suitable weighting scheme on the observations $\{\tau_i\}$. Specifically, they propose the estimate

$$\hat{\phi}'(\omega) = \sum_{i=1}^{m} \binom{m}{i} \tau_i \, \omega^i (1-\omega)^{m-i}, \quad (0 \leqslant \omega \leqslant 1) \tag{1.2}$$

(corresponding to binomial weights)
in which case $\phi(\omega)$ is then estimated by

$$\hat{\phi}(\omega) = \int_0^{\omega} \hat{\phi}'(u) \, \mathrm{d}u. \tag{1.3}$$

Thus, the problem reduces essentially to that of estimating a function (which may be assumed to be in some sense "smooth" but is otherwise quite arbitrary), given only observations at a discrete set of points. The conventional approach to this problem via regression analysis assumes that the required function has a known mathematical form which involves only a finite number of unknown parameters. We now describe below an alternative approach which is "non-parametric" in the sense that it requires only that the function be "smooth"—in a sense to be made more precise in Section 6.

† Present address: Bell Telephone Laboratories, Holmdel, New Jersey, U.S.A.

2. STATEMENT OF THE GENERAL PROBLEM

We now restate the problem in more general terms, and accordingly change the notation somewhat from that used in Section 1.

Suppose that we are given m observations, $y_1, y_2, \ldots, y_m$, which are described by the model,

$$y_i = f(x_i) + \epsilon_1, \quad i = 1, 2, \ldots, m, \quad 0 \leqslant x_1 \leqslant x_2 \leqslant \ldots \leqslant x_m \leqslant 1, \tag{2.1}$$

where the $\{x_i\}$ have known values, $f(x)$ is an unknown function defined for $0 \leqslant x \leqslant 1$ and the $\{\epsilon_i\}$ are uncorrelated random variables with zero mean and constant variance σ_ϵ^2, i.e.

$$E[\epsilon_i] = 0, \quad E[\epsilon_i^2] = \sigma_\epsilon^2, \quad i = 1, \ldots, m.$$

Initially, we will assume further that the observations are taken at equally spaced intervals, so that

$$x_{i+1} - x_i = \delta, \text{ say}, \quad i = 1, \ldots, (m-1).$$

(Later, we remove this restriction.) We require estimates of $f(x)$ and $F(x) = \int_0^x f(u)\,du$, for all $x \in (0, 1)$.

We may start by considering the simple "step-function" type estimate of $f(x)$, namely,

$$\hat{f}_1(x) = \begin{cases} y_1, & 0 < x \leqslant x_1, \\ y_i, & x_{i-1} < x \leqslant x_i, \quad i = 1, \ldots, (m-1), \\ y_m, & x_m < x \leqslant 1. \end{cases} \tag{2.2}$$

Alternatively, we may consider a "piece-wise linear" estimate obtained simply by joining the points (x_i, y_i) by straight lines, namely,

$$\hat{f}_2(x) = \begin{cases} \dfrac{x}{x_1} y_1, & 0 \leqslant x < x, \\ \dfrac{(x - x_1) y_2 + (x_2 - x) y_1}{(x_2 - x_1)}, & x_1 \leqslant x < x_2 \\ \vdots & \\ (x - x_{m-1}) y_m + (x_m - x) y_{m-1}, & x_{m-1} \leqslant x < x_m, \\ \left(\dfrac{1 - x}{1 - x_m}\right) y_m, & x_m \leqslant x \leqslant 1. \end{cases} \tag{2.3}$$

Of course, neither $\hat{f}_1(x)$ nor $\hat{f}_2(x)$ would, in general be a satisfactory estimate of $f(x)$ since, for example, $\hat{f}_1(x)$ is discontinuous at $x = x_i$, while $\hat{f}_2(x)$ has a discontinuous derivative at $x = x_i$ $(i = 1, \ldots, m)$. We note, however, that $\hat{f}_1(x)$ and $\hat{f}_2(x)$ may be re-written in the forms:

$$\hat{f}_1(x) = \delta \sum_{i=1}^{m} y_i \left\{ \delta^{-1} W^{(1)}\left(\frac{x - x_i}{\delta}\right) \right\}, \tag{2.4}$$

where

$$W^{(1)}(x) = \begin{cases} 1, & -1 < x \leqslant 0, \\ 0, & \text{otherwise} \end{cases} \tag{2.5}$$

and

$$\hat{f}_2(x) = \delta \sum_{i=1}^{m} y_i \left\{ \delta^{-1} W^{(2)}\left(\frac{x - x_i}{\delta}\right) \right\}, \tag{2.6}$$

where

$$W^{(2)}(x) = \begin{cases} 1 - |x|, & 0 < |x| \leqslant 1, \\ 0, & \text{otherwise.} \end{cases} \tag{2.7}$$

Equations (2.4) and (2.6) suggest that we may now consider a general class of estimates of $f(x)$, of the form,

$$\hat{f}(x) = \delta \sum_{i=1}^{m} y_i W_k(x - x_i) \tag{2.8}$$

where, for each k,

$$W_k(x) = k^{-1} W_0(x/k) \tag{2.9}$$

and $W_0(x)$ is a general type of "weight function" defined for $-\infty < x < \infty$, and satisfying

(a) $$W_0(x) \geqslant 0, \text{ all } x,$$

(b) $$\int_{-\infty}^{\infty} W_0(x)\,\mathrm{d}x = 1 \text{ and}$$

(c) $$\int_{-\infty}^{\infty} \{W_0(x)\}^2\,\mathrm{d}x < \infty.$$

Note that the quantity k plays the role of a "bandwidth" parameter, i.e. by varying the value of k we may vary the "width" of the function $W_k(x)$. Note further that both $\hat{f}_1(x)$ and $\hat{f}_2(x)$ are special cases of the general form (2.8) in which

$$W_0(x) = W^{(1)}(x) \quad \text{and} \quad k = \delta, \quad \text{for } \hat{f}_1(x)$$

and

$$W_0(x) = W^{(2)}(x) \quad \text{and} \quad k = \delta, \quad \text{for } \hat{f}_2(x).$$

Thus, $\hat{f}_1(x)$ corresponds to the *rectangular* weight function given by (2.5), and $\hat{f}_2(x)$ corresponds to the *triangular* weight function given by (2.7). In each case, the bandwidth of the weight function is exactly equal to the interval spacing of the points $\{x_i\}$.

Using (2.8) to estimate $f(x)$, we may then estimate $F(x)$ by

$$\hat{F}(x) = \int_0^x \hat{f}(u)\,\mathrm{d}u \tag{2.10}$$

3. ANALOGY WITH DENSITY FUNCTION ESTIMATION

The use of weight functions of the form (2.9) is well known in the context of estimating both probability and spectral density functions (see, e.g. Parzen, 1961, 1962). There is, in fact, a superficial resemblance between the present problem and that of probability density function estimation which may be observed by writing the general estimate (2.8) in the form

$$\hat{f}(x) = \int_0^1 W_k(x-u)\, \mathrm{d}\hat{F}_0(u), \tag{3.1}$$

where $\hat{F}_0(x)$ is the crude step-function estimate of $F(x)$ analogous to (1.1). Specifically,

$$\hat{F}_0(x) = \begin{cases} 0, & x < x_1 \\ \delta \sum_{i=1}^{j} y_i, & x_j \leqslant x < x_{j+1} \quad (j = 1, \ldots, m) \\ \delta \sum_{i=1}^{m} y_i, & x_m \leqslant x \leqslant 1. \end{cases}$$

(Note that, with equal spacing, $\delta \sim m^{-1}$.) Thus, we may regard $\hat{f}(x)$ as a "smoothed" version of "$\{\mathrm{d}\hat{F}_0(x)\}/\mathrm{d}x$". However, we would point out that the connection between the two problems is only superficial, and the sampling properties of $\hat{f}(x)$ are certainly not the same as those of a probability density function estimate. (Recall, in particular, that the observations y_i are *not* frequency counts.)

4. Approximate Mean and Variance of $\hat{f}(x)$

We now derive approximate expressions for the asymptotic mean and variance of $\hat{f}(x)$ as m (the number of observations) $\to \infty$. The approach presented here is heuristic, but illustrates the essential features of the sampling properties of $\hat{f}(x)$. (A more rigorous derivation of these results is presented in Section 6.)

We have, from (2.1) and (2.8),

$$\begin{aligned} E\{\hat{f}(x)\} &= \delta \sum_{i=1}^{m} f(x_i)\, W_k(x-x_i) \\ &\sim \int_0^1 f(y)\, W_k(x-y)\, \mathrm{d}y, \quad \text{as } \delta \to 0 \quad (\text{i.e. } m \to \infty) \\ &\qquad (\text{recall that, with equal spacing, } \delta \sim m^{-1} \to 0 \text{ as } m \to \infty) \\ &= \int_0^1 f(y) \left\{ k^{-1} W_0\left(\frac{x-y}{k}\right) \right\} \mathrm{d}y \\ &= \int_{-(1-x)/k}^{x/k} f(x-kv)\, W_0(v)\, \mathrm{d}v \qquad (4.1) \\ &\sim f(x) \left\{ \int_{-\infty}^{\infty} W_0(v)\, \mathrm{d}v \right\}, \quad \text{as } k \to 0 \\ &= f(x), \quad \text{by condition (a) on } W_0(x). \end{aligned}$$

Also, from (2.8),

$$\begin{aligned}
\operatorname{var}\{\hat{f}(x)] &= \delta^2 \sigma_\epsilon^2 \sum_{i=1}^{m} W_k^2(x-x_i) \\
&\sim \delta\sigma_\epsilon^2 \int_0^1 W_k^2(x-y)\,\mathrm{d}y, \quad \text{as } \delta\to 0, \\
&= \frac{\delta\sigma_\epsilon^2}{k}\int_0^1 k^{-1}\left\{W_0\left(\frac{x-y}{k}\right)\right\}^2 \mathrm{d}y \\
&\sim \frac{\delta}{k}\sigma_\epsilon^2 \int_{-\infty}^{\infty} \{W_0(v)\}^2\,\mathrm{d}v, \quad \text{as } k\to 0. \\
&= O(\delta/k), \quad \text{since } \int_{-\infty}^{\infty} W_0^2(v)\,\mathrm{d}v<\infty \text{ by condition (b).}
\end{aligned} \tag{4.2}$$

Thus, provided that, as $m\to\infty$, $\delta\to 0$ and $k\to 0$ in such a way that $(\delta/k)\to 0$, then $\hat{f}(x)$ will be asymptotically unbiased and $\operatorname{var}\{\hat{f}(x)\}\to 0$. We may expect, therefore, that in this case $\hat{f}(x)$ will be a consistent estimate of $f(x)$.

Note that the condition $\delta = O(k)$ (as $m\to\infty$) implies that, for consistency, the "bandwidth" of the weight function $W_k(x)$ must extend over more than *one* interval. If $k = \delta$ then, essentially, the "smoothing effects" of $W_k(x)$ covers only one interval and $\hat{f}(x)$ is no longer a consistent estimate of $f(x)$.

Example

Take $\delta = m^{-1}$ and choose $k = m^{-\alpha}$, $\alpha<1$. Then, as $m\to\infty$, $\delta\to 0$, $k\to 0$, $\delta/k\to 0$, so that

$$E\{\hat{f}(x)\}\to f(x), \quad \text{as } m\to\infty$$

$$\operatorname{var}\{\hat{f}(x)\}\to 0, \quad \text{as } m\to\infty.$$

The above results show that as we decrease the value of k the bias of $\hat{f}(x)$ decreases but the variance increases. Conversely, as we increase k the bias increases but the variance decreases. In practice, the "optimum" value of k will depend on which criterion we use to measure the overall precision of $f(x)$. For example, if we use the mean-square error criterion then, in principle, we could choose k so as to minimize

$$M^2\{\hat{f}(x)\} = E\{\hat{f}(x)-f(x)\}^2 = \operatorname{var}\{\hat{f}(x)\}+[E\{\hat{f}(x)\}-f(x)]^2. \tag{4.3}$$

This is similar to the problem of choosing the optimum bandwidth of a spectral window when estimating power spectra (see, e.g., Priestley, 1965), and presents the same difficulty, namely, that the magnitude of the bias term in the above expression depends on the "smoothness' 'of $f(x)$ relative to $W_k(x)$, and cannot be evaluated explicitly unless we have *a priori* information on the behaviour of the derivatives of $f(x)$. More specifically, consider the case where $f(x)$ is a polynomial of degree K (say), and $W_0(x)$ is an even function of x whose moments up to order K are all finite. Then, expanding $f(x-kv)$ in (4.1) we obtain, for small k,

$$E\{\hat{f}(x)\} \sim f(x)+(k^2/2)f''(x)\int_{-\infty}^{\infty} v^2 W_0(v)\,\mathrm{d}v+o(k^2). \tag{4.4}$$

Thus, we find from (4.2) to (4.4) that

$$M^2\{\hat{f}(x)\} \sim \tfrac{1}{4}k^4\{f''(x)\}^2\,\sigma_W^4 + (\delta^2/k)\,\sigma_\epsilon^4 I_W^2, \tag{4.5}$$

where

$$\sigma_W^2 = \int_{-\infty}^{\infty} v^2\, W_0(v)\,\mathrm{d}v$$

and

$$I_W = \int_{-\infty}^{\infty} \{W_0(v)\}^2\,\mathrm{d}v.$$

Minimizing (4.5) with respect to k gives

$$k = \left[\frac{2\delta^2\,\sigma_\epsilon^4 I_W^2}{\{f''(x)\}^2\,\sigma_W^4}\right]^{1/6} \tag{4.6}$$

and substituting (4.6) in (4.5) gives

$$M_{\min}^2[\hat{f}(x)] \sim \left(\frac{3}{2^{4/3}}\right)[\{f''(x)\}^{\frac{1}{2}}\,\sigma_\epsilon^2\,\delta\sigma_W\, I_W]^{4/3}. \tag{4.7}$$

Equation (4.6) indicates that the optimum value of k is $0(\delta^{\frac{1}{3}})$, i.e. $O(m^{-\frac{1}{3}})$, but the precise value of (4.6) depends, of course, on the value of $f''(x)$.

In practice, one might possibly obtain a numerical value for k by replacing $f''(x)$ in (4.6) by $\max_x\{f''(x)\}$ (assumed >0)—in cases where this quantity is known, *a priori*. The reasoning underlying this approach is based on the observation that the magnitude of the bias term will be largest in the region where $f(x)$ has its "sharpest" peak. However, it may be noted that, in contrast with the case of *density* function estimation, the prior information required for the evaluation of the bias term cannot be expressed physically purely in terms of a "bandwidth" parameter. In the problem of density function estimation the variance term is proportional to $f^2(x)$ and the natural criterion to minimize is the *relative* mean-square error. The corresponding quantity which then arises in evaluating the "bias" term is $\min_x\{f(x)/f''(x)\}$; this latter quantity being related to the width of the narrowest peak in $f(x)$—see Priestley (1965).

5. CHOICE OF WEIGHT FUNCTION

As in the case of density function estimation the choice of $W_0(x)$ is, to a large extent, arbitrary, and one may expect that the properties of $\hat{f}(x)$ will depend more critically on the chosen value of the parameter k rather than on the mathematical form of $W_0(x)$. In fact, the simple weight functions $W^{(1)}(x)$ and $W^{(2)}(x)$ (as given by (2.5) and (2.7)) would be suitable choices for $W_0(x)$. [The unsatisfactory nature of $\hat{f}_1(x)$ and $\hat{f}_2(x)$ is due to the fact that, for these estimates, $k = \delta$.] However, it is interesting to note that $W^{(2)}(x)$ corresponds to the convolution of $W^{(1)}(x)$ with itself, so that $W^{(2)}(x)$ corresponds to the probability density function of the sum of two independent rectangular variables. We may now consider a third weight function, $W^{(3)}(x)$ which corresponds to the probability density function of the sum of three independent rectangular variables, and so on. In the limit we are led to considering a Gaussian weight function of the form

$$W^{(\infty)}(x) = 1/\sqrt{(2\pi)}\,e^{-\frac{1}{2}x^2}. \tag{5.1}$$

A rough comparison of different weight functions may be made on the basis of the approximate expression for $M^2_{\min}\{\hat{f}(x)\}$, as given by (4.7). We note that the dependence of the right-hand side of (4.7) on the form of the function $W_0(x)$ arises only in terms of the product, $(\sigma_W I_W)$. In Table 1 we tabulate the values of $(\sigma_W I_W)$

TABLE 1

Values of $(\sigma_W I_W)$

	Rectangular	*Triangular*	*Gaussian*	*Quadratic*
σ_w	$1/\sqrt{12}$	$1/\sqrt{6}$	1	$1/\sqrt{5}$
I_w	1	$\frac{2}{3}$	$1/2\sqrt{\pi}$	$\frac{3}{5}$
$\sigma_w I_w$	0·2886	0·2721	0·2821	0·2682

for the rectangular, triangular and Gaussian weight functions (corresponding to $W^{(1)}(x)$, $W^{(2)}(x)$ and $W^{(\infty)}(x)$, respectively), and also for the quadratic weight function,

$$W^{(4)}(x) = \begin{cases} \frac{3}{4}(1-x^2), & |x| \leqslant 1, \\ 0, & |x| > 1. \end{cases} \tag{5.2}$$

(The rectangular weight function, $W^{(1)}(x)$, as defined by (2.5), is not of course, an even function, but can be replaced by

$$W^{(1)}(x) = \begin{cases} 1, & -\frac{1}{2} \leqslant x \leqslant \frac{1}{2}, \\ 0, & \text{otherwise}, \end{cases}$$

by a trivial modification to $\hat{f}_1(x)$.)

Thus, it would appear that the quadratic weight function is to be preferred to the Gaussian form, but it should be remembered that the expression (4.7) for $M^2_{\min}[\hat{f}(x)]$ is only an approximation, and was obtained by neglecting in particular, terms of $o(k^2)$ in expression (4.4) for the bias term. Previous studies of the density function estimation problem (Priestley, 1962; Bartlett, 1963) have indicated that, with the above approximation, the quadratic weight function produces the smallest relative mean-square error, and Table 1, while not conclusive, is certainly consistent with this assertion. However, perhaps the more important feature of the above calculations is the fact that there is relatively little variation between the values of $(\sigma_W I_W)$ for the four weight functions considered. This would tend to confirm that, as previously remarked, one would expect the sampling properties of $\hat{f}(x)$ to depend more critically on the value of the parameter k rather than on the mathematical form of $W_0(x)$.

6. CONVERGENCE OF $f(x)$

As noted in Section 4, one would expect $\hat{f}(x)$ to be a consistent estimate of $f(x)$ provided $f(x)$ satisfies certain "smoothness" conditions, and $\delta/k \to 0$ as $m \to \infty$. We now state this result in a more precise form in the following theorem in which we remove the restriction that the $\{x_i\}$ are equally spaced over (0, 1), and consider a slightly more general form of (2.8), namely,

$$\hat{f}_{k,m}(x) = \sum_{i=1}^{m} y_i(x_i - x_{i-1})\, k^{-1} W_0\left(\frac{x-x_i}{k}\right). \tag{6.1}$$

(The proof of the theorem is straightforward and is omitted; details may be found in Priestley and Chao, 1971.)

Theorem. Let $f(x), W_0(x)$ satisfy Lipschitz conditions of orders α, β, respectively. Let $\delta_m = \max_i(x_i - x_{i-1}) = O(m^{-1})$ and $k = m^{-\gamma}$. If

$$\gamma < \min\left(\alpha, \frac{\beta}{1+\beta}\right),$$

then the estimate

$$\hat{f}_{k,m}(x) = \sum_{i=1}^{m} y_i(x_i - x_{i-1})\, k^{-1} W_0\left(\frac{x-x_i}{k}\right)$$

converges to $f(x)$, in probability, for all $x \in (0,1)$.

REFERENCES

BARTLETT, M. S. (1963). Statistical estimation of density functions. *Sankhyā*, A, **25**, 245.
PARZEN, E. (1961). Mathematical considerations in the estimation of spectra. *Technometrics*, **3**, 167–190.
—— (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.*, **33**, 1065–1076.
PRIESTLEY, M. B. (1962). Basic considerations in the estimation of power spectra. *Technometrics*, **4**, 551–564.
—— (1965). The role of bandwidth in spectral analysis. *Appl. Statist.*, **14**, 33–47.
PRIESTLEY, M. B. and CHAO, M. T. (1971). Non-parametric function fitting. Internal Bell Telephone Laboratories memo.
TISCHENDORF, J. A. and CHAO, M. T. (1970). Statistical properties of estimates of an integral. Internal Bell Telephone Laboratories memo.