On the Nonparametric Estimation of Regression Functions

By Jacqueline K. Benedetti

University of California at Los Angeles

[Received October 1976. Final revision March 1977]

Summary

We consider a nonparametric technique proposed by Priestley and Chao (1972) for estimating an unknown regression function. Conditions for strong convergence and asymptotic normality are discussed. Special consideration is given to the optimal choice of a weighting function.

Keywords: NONPARAMETRIC REGRESSION; KERNEL ESTIMATES; WEIGHTING FUNCTIONS; CURVE FITTING

1. Introduction

An important statistical problem is the estimation of a regression function $g(x) = E(y|x)$. Typically, $g(x)$ has a specified functional form and parameter estimates are obtained according to certain desirable criteria, such as least squares. Assuming normal errors the investigator can test the appropriateness of the hypothesized model. One may wish, however, to have an estimation technique applicable for an arbitrary $g(x)$.

In a recent paper, Priestley and Chao (1972) considered the problem of estimating an unknown regression function $g(x)$ given observations at a fixed set of points. Their estimate, referred to here as the Priestley–Chao (PC) estimate, is nonparametric in the sense that $g(x)$ is restricted only by certain smoothing requirements. It can be viewed as a moving average of sample Y's whose weights are based on a class of kernels suggested by Rosenblatt (1956) and Parzen (1962). These weights are similar to those used in nonparametric density estimation.

In this paper, the results of Priestley and Chao will be reviewed and further properties of their estimate considered.

2. The Priestley–Chao Estimate

Let $Y_1, ..., Y_n$ be n observations at fixed $x_1, ..., x_n$ according to the model

$$Y_i = g(x_i) + \varepsilon_i,$$

where g is an unknown function defined on x contained in the interval $[0, 1]$, and the errors are i.i.d. random variables with zero mean and finite variance σ^2. Without loss of generality we assume $0 \leqslant x_1 \leqslant ... \leqslant x_n \leqslant 1$. The Priestley–Chao estimate of $g(x)$ is of the form

$$g_n(x) = \sum_{i=1}^{n} Y_i \frac{x_i - x_{i-1}}{h_n} K\left(\frac{x - x_i}{h_n}\right),$$

where K is a weight function, satisfying

$$\text{(a) } K(u) \geqslant 0, \text{ for all } u; \text{ (b) } \int_{-\infty}^{\infty} K(u)\,du = 1; \text{ (c) } \int_{-\infty}^{\infty} K^2(u)\,du < \infty; \quad \text{(d) } K(u) = K(-u) \quad (1)$$

and $\{h_n\}$ is a sequence of positive real numbers converging to zero in such a way that $nh_n \to \infty$ as $n \to \infty$. Commonly used weight functions, or kernels, are given in Table 1.

Priestley and Chao (1970) establish consistency of the estimate through the following theorem.

TABLE 1

Functions and values associated with various kernels

Kernel	$K(u)$	$\kappa(u) = \int_{-\infty}^{\infty} e^{iux} K(x)\, dx$	r	$L = \int_{-\infty}^{\infty} K^2(u)\, du$	$M = \int_{-\infty}^{\infty} u^2 K(u)\, du$	$R = LM^{\frac{1}{2}}$
Uniform	$\begin{cases} \frac{1}{2}, & \lvert u \rvert \leqslant 1 \\ 0, & \lvert u \rvert > 1 \end{cases}$	$\dfrac{\sin(u)}{u}$	2	$\frac{1}{2}$	$\frac{1}{3}$	0·2886
Triangular	$\begin{cases} 1-\lvert u \rvert, & \lvert u \rvert \leqslant 1 \\ 0, & \lvert u \rvert > 1 \end{cases}$	$\left\{\dfrac{\sin(u/2)}{u/2}\right\}^2$	2	$\frac{2}{3}$	$\frac{1}{6}$	0·2721
Normal	$\{\surd(2\pi)\}^{-1} e^{-u^2/2}$	$e^{-u^2/2}$	2	$(2\surd\pi)^{-1}$	1	0·2821
Double exponential	$\frac{1}{2}\exp(-\lvert u \rvert)$	$(1+u^2)^{-1}$	2	$3/(5\surd 5)$	1	0·2682
Quadratic	$\begin{cases} (3/4\surd 5)(1-u^2/5), & \lvert u \rvert \leqslant \surd 5 \\ 0, & \lvert u \rvert > \surd 5 \end{cases}$	$\dfrac{3(\sin u\surd 5 - u\surd 5 \cos u\surd 5)}{(5\surd 5)\, u^3}$	2	$\frac{1}{4}$	2	0·3535
Fejer	$(2\pi)^{-1}\left\{\dfrac{\sin(u/2)}{u/2}\right\}^2$	$\begin{cases} 1-\lvert u \rvert, & \lvert u \rvert \leqslant 1 \\ 0, & \lvert u \rvert > 1 \end{cases}$	1	$(3\pi)^{-1}$	—	—
Cauchy	$\{\pi(1+u^2)\}^{-1}$	$e^{-\lvert u \rvert}$	1	π^{-1}	—	—

Theorem 1. Let $g(x)$ and $K(u)$ satisfy Lipschitz conditions of orders α and β respectively. Let $\delta_n = \max_i(x_i - x_{i-1}) = O(1/n)$ and $h_n = n^{-\gamma}$ for $\gamma > 0$. If $\gamma < \min(\alpha, \beta/1+\beta)$, then $g_n(x) \to g(x)$ in probability for $x \in (0,1)$.

Note that the restrictions on γ require it to be $< \frac{1}{2}$.

In the next two sections, asymptotic normality and a stronger form of convergence are demonstrated. The final section considers choices of the kernel and the bandwidth parameter.

3. Asymptotic Normality

To demonstrate asymptotic normality, we require the following lemma, which is stated without proof. (The details are a straightforward application of the dominated convergence theorem, and may be found in Benedetti, 1974.)

Lemma 1. Suppose K is continuous, and is such that $K(u)$ is nonincreasing for $u > 0$, and nondecreasing for $u < 0$, and suppose $\int_{-\infty}^{\infty} K^r(u)\,du < \infty$. If there exists some $\bar{\Delta}$ such that $\bar{\Delta}/n \geqslant \max(x_i - x_{i-1})$ for all n, where $x_0 = 0$, $x_{n+1} = 1$, and if $nh_n \to \infty$, then for $x \in (0,1)$

$$h_n^{-1} \sum (x_i - x_{i-1})\, K^r\left(\frac{x - x_i}{h_n}\right) \to \int_{-\infty}^{\infty} K^r(u)\,du.$$

Most commonly used kernels satisfy the requirements of Lemma 1. One exception is the Fejer kernel

$$K(y) = (2\pi)^{-1}\left\{\frac{\sin(y/2)}{y/2}\right\}^2,$$

which does not display the required monotonicity. However, since it is bounded above by $2/(\pi y^2)$, an appropriate dominating function may be constructed to carry forth the desired convergence.

Theorem 2. If K satisfies the conditions of Lemma 1 for $r = 3$, and if there exists $\bar{\Delta}$ and $\hat{\Delta}$ such that $0 < \hat{\Delta}/n \leqslant (x_i - x_{i-1}) \leqslant \bar{\Delta}/n$ for $i = 1, 2, \ldots, n$, and if the third moment of the ε's is finite, then

$$[g_n(x) - E\{g_n(x)\}]/[\operatorname{var}\{g_n(x)\}]^{\frac{1}{2}}$$

is asymptotically $N(0,1)$ for all $x \in (0,1)$.

Proof. Theorem 2 follows from the Liapounoff Central Limit Theorem and Lemma 1.

Let F_n be the distribution function of

$$\frac{g_n(x) - E\{g_n(x)\}}{[\operatorname{var}\{g_n(x)\}]^{\frac{1}{2}}},$$

$\Phi(x)$ be the standard normal distribution, and $M = \max_x K(x)$. Then $|F_n(x) - \Phi(x)|$ can be approximately bounded by

$$33E|\varepsilon|^3 M\bar{\Delta} \Big/ 4\hat{\Delta}\sigma^3 \left\{nh_n \int_{-\infty}^{\infty} K^2(u)\,du\right\}^{\frac{1}{2}}$$

which is obtained using the Berry–Essen theorem (Rosenblatt, 1971) and Lemma 1.

4. Convergence of $g_n(x)$

In this section we establish a stronger form of convergence for $g_n(x)$ than previously demonstrated by Priestley and Chao. We apply the following lemma to achieve our results.

Lemma 2. Let $\{y_n\}$ be a sequence of symmetric i.i.d. random variables with zero mean and finite fourth moment. If $\{a_{nk}\}$ is a sequence of constants such that

$$\sum_{n=1}^{\infty}\left(\sum_{k=1}^{n} a_{nk}^2\right)^2 < \infty,$$

then $y_n' \to 0$ almost surely, where $y_n' = \sum a_{nk} y_k$, the summation being from 1 to n.

Proof. The fourth moment μ_4' of y_n' is such that $\mu_4' \leqslant \mu_4(\sum a_{nk}^2)^2$, where μ_4 is the fourth moment of y_n.

The Chebyshev inequality implies

$$\sum P(|y_n'| > c) \leqslant \frac{\mu_4}{c^4}\sum_n \left|\sum_k a_{nk}^2\right|^2.$$

Hence it follows that $y_n' \to 0$ almost surely by the Borel Cantelli lemma.

Theorem 3. Under the conditions of Theorem 1, and if $E(\varepsilon)^4$ exists, and Lemma 1 holds for $r = 2$, $g_n(x) \to g(x)$ with probability one.

Proof. The result follows from Lemma 2 with

$$a_{ni} = \left(\frac{x_i - x_{i-1}}{h_n}\right) K\left(\frac{x - x_i}{h_n}\right).$$

5. Choices of Kernel

Let us assume now that the x_i's are equally spaced, i.e. $x_i - x_{i-1} = \delta$. Using a method of Parzen (1962) which will be outlined below, we find an asymptotic expression for the mean square error, which, under certain conditions, is found to be identical to the expression derived by Priestley and Chao. Given this expression, optimality of the choice of kernel will be considered.

Let

$$\kappa(t) = \int_{-\infty}^{\infty} e^{itx} K(x)\, dx$$

be the Fourier transform of the kernel K. Then

$$E\{g_n(x)\} = (2\pi)^{-1}\int_{-\infty}^{\infty}\left\{\sum_{j=1}^{n} g(x_j)\,\delta\, e^{izx_j}\right\} e^{-izx}\,\kappa(h_n z)\, dz.$$

Similarly, the bias term may be expressed as

$$E\{g_n(x) - g(x)\} = (2\pi)^{-1}\int_{-\infty}^{\infty}\kappa(h_n z)\, e^{-izx}\left\{\sum_{j=1}^{n} g(x_j)\,\delta e^{izx_j} - \phi_g(z)\right\} dz + (2\pi)^{-1}\int_{-\infty}^{\infty} e^{-izx} \times \phi_g(z)\{\kappa(h_n z) - 1\}\, dz,$$

where $\phi_g(z) = \int_0^1 g(x)\, e^{izx}\, dx$.

Now, if g is bounded and continuous, then by the dominated convergence theorem,

$$\sum_{j=1}^{n} g(x_j)\,\delta e^{izx_j} \to \phi_g(z),$$

and hence the first term of the bias expression converges to zero.

If there exists a positive r such that $k_r = \lim_{u\to 0}[\{1 - \kappa(u)\}/|u|^r]$ is nonzero, then r is called the characteristic exponent of κ, and k_r is the characteristic coefficient. Thus, if we divide the second term in the bias by h_n^r we obtain

$$(2\pi)^{-1}\int_{-\infty}^{\infty} e^{-izx}\,\phi_g(z)\frac{\kappa(h_n z) - 1}{h_n^r}\, dz = (2\pi)^{-1}\int_{-\infty}^{\infty} e^{-izx}\,\phi_g(z)\frac{\kappa(h_n z) - 1}{|h_n z|^r}|z|^r\, dz. \tag{2}$$

Since $\kappa(u)$ is the Fourier transform of K, $|1-\kappa(u)|\leqslant M|u|^r$ for some M, and for $|u|<\delta$. Hence, equation (2) is bounded by

$$\frac{M}{2\pi}\int_{-\delta/h_n}^{\delta/h_n}|u|^r|\phi_g(u)|\,du+\frac{Q}{2\pi\delta^r}\int_{|u|>\delta/h_n}|u|^r|\phi_g(u)|\,du,$$

where Q is a bound on $\kappa(u)-1$ which necessarily exists and is positive.

This equation is strictly less than $\{(P+1)/2\pi\}\int_{-\infty}^{\infty}|u|^r|\phi(u)|\,du$, where $P=\max(M,Q/\delta^r)$. Hence, by the dominated convergence theorem

$$(2\pi)^{-1}\int_{-\infty}^{\infty}\frac{e^{-izx}\phi_g(z)\{\kappa(h_n z)-1\}}{h_n^r}\,dz\to k_r g^r(x),$$

where

$$g^r(x)=-(2\pi)^{-1}\int_{-\infty}^{\infty}e^{-iux}|u|^r\phi_g(u)\,du,$$

and it is assumed that the integral converges absolutely. Now,

$$\operatorname{var}(g_n(x))=\frac{\delta^2}{h_n^2}\sigma^2\sum_{i=1}^{n}K^2\{(x-x_i)/h_n\}\sim\frac{\delta\sigma^2}{h_n}\int_{-\infty}^{\infty}K^2(u)\,du$$

by Lemma 1. Hence, an asymptotic expression for the mean square error is

$$(\sigma^2/h_n)\,\delta\int_{-\infty}^{\infty}K^2(u)\,du+h_n^{2r}|k_r g^r(x)|^2.$$

This is minimized with respect to h_n by taking

$$h_n=\left\{\sigma^2\,\delta\int K(u)\,du\bigg/2r|k_r g^r(x)|^2\right\}^{1/(1+2r)}.$$

Hence the mean square error becomes

$$(2r+1)\left\{\sigma^2\int K^2(u)\,du\right\}^{2r/(1+2r)}(2r/\delta)^{-2r/(1+2r)}\{k_r g^r(x)\}^{2/(1+2r)}\tag{3}$$

which tends to zero as $n^{-2r/1+2r}$. In the case of $r=2$ (see Table 1) this implies the estimates have order of consistency $n^{\frac{4}{5}}$, i.e. $n^{\frac{4}{5}}E(g_n(x)-g(x))^2\to C<\infty$ as $n\to\infty$. It should also be noted that this mean square error is precisely the result (2.3) derived by Priestley and Chao.

Suppose now, among the class of kernels with $r=2$, we would like to find the kernel which minimizes (3). This is the same as finding the K which minimizes

$$R_u=\left\{\int_{-\infty}^{\infty}K^2(u)\,du\right\}^2\left\{\int_{-\infty}^{\infty}u^2K(u)\,du\right\}$$

subject to the restraints $\int_{-\infty}^{\infty}K(u)=1$, $K(u)\geqslant 0$, $K(u)=K(-u)$. The next result shows that there is not a unique minimization of R, but rather a whole family of kernels $K(u)$ which will be acceptable.

Lemma 3. For any random variable U whose density function is an acceptable kernel, any scale transformation $X=cU$ leaves R unchanged.

Proof. Let U have variance σ^2, and range (a,b) where $\pm\infty$ are possible limits. Let $X=cU$. Then

$$f(x)=c^{-1}K(x/c)\text{ for }\quad ca\leqslant x\leqslant cb,\text{ if }c>0;\quad ca\geqslant x\geqslant cb,\text{ if }c<0;$$

hence

$$\int_{ca}^{cb}f^2(x)\,dx=c^{-2}\int_{ca}^{cb}K^2(x/c)\,dx=c^{-1}\int_a^b K^2(u)\,du$$

therefore

$$R_x = \left\{\int_{ca}^{cb} f^2(x)\,dx\right\}^2 \{\text{var}(x)\} = c^{-2}\left\{\int_a^b K^2(u)\,du\right\}^2 c^2\,\text{var}(u) = R_u.$$

To find one member of the family of "optimal" kernels, let us set the variance equal to 1. Thus we need only to consider minimizing

$$\int_{-\infty}^{\infty} K^2(u)\,du$$

subject to constraints (1) and $\int u^2 K(u)\,du = 1$. This minimization problem is precisely one considered by Epanechnikov (1969). The solution is the quadratic kernel.

$$K_0(u) = (3/4\sqrt{5})(1-u^2/5), \quad |u| \leqslant \sqrt{5},$$

which is just a scale transformation of the kernel

$$K(x) = \tfrac{3}{4}(1-x^2), \quad |x| \leqslant 1,$$

that Priestley and Chao conjectured to be optimal.

The above argument provides an optimal family of kernels in the sense of optimizing an asymptotic expression for the mean square error. There is no reason to believe, however, for finite samples that the same kernel should be best regardless of the underlying form of the function $g(x)$.

Table 1 suggests that, at least asymptotically, the choice of kernel is not critical. As has previously been noted for density estimation and the PC estimate, the properties of $g_n(x)$ tend to depend more critically on the choice of the bandwidth parameter. Guidelines for choosing an appropriate h_n are needed before such estimates can be employed in practical situations.

Acknowledgements

This research represents a portion of the author's doctoral dissertation, and was supported in part by the NIH Biometry Training Grant No. 5–T01–GM–1269–10. Thanks are due to Dr Richard Kronmal for suggesting the topic and guiding the research.

References

BENEDETTI, J. (1974). Kernel estimation of regression functions. Doctoral Dissertation, University of Washington.

EPANECHNIKOV, V. A. (1969). Nonparametric estimates of a multivariate probability density. *Theor. Probability Appl.*, **14**, 153–158.

PARZEN, E. (1962). On estimation of a probability density and mode. *Ann. Math. Statist.*, **33**, 1065–1076.

PRIESTLEY, M. B. and CHAO, M. T. (1970). Nonparametric function fitting. Department of Mathematics Technical Report No. 21, University of Manchester.

—— (1972). Non-parametric function fitting. *J. R. Statist. Soc.* B, **34**, 385–392.

ROSENBLATT, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.*, **27**, 832–837.

—— (1971). *Markov Processes: Structure and Asymptotic Behavior.* Die Grundlehren der Mathematischen, Wissenschaften und Einzeldarstellungen. New York: Springer-Verlag.