J. R. Statist. Soc. B (1985)
47, *No.* 1, *pp.* 1–52

Some Aspects of the Spline Smoothing Approach to Non-parametric Regression Curve Fitting

By B. W. SILVERMAN

University of Bath

[*Read before the* Royal Statistical Society *at a meeting organized by the* Research Section *on Wednesday, October 10th, 1984*, Professor J. B. Copas *in the Chair*]

SUMMARY

Non-parametric regression using cubic splines is an attractive, flexible and widely-applicable approach to curve estimation. Although the basic idea was formulated many years ago, the method is not as widely known or adopted as perhaps it should be. The topics and examples discussed in this paper are intended to promote the understanding and extend the practicability of the spline smoothing methodology. Particular subjects covered include the basic principles of the method; the relation with moving average and other smoothing methods; the automatic choice of the amount of smoothing; and the use of residuals for diagnostic checking and model adaptation. The question of providing inference regions for curves—and for relevant properties of curves—is approached via a finite-dimensional Bayesian formulation.

Keywords: ROUGHNESS PENALTY; SMOOTHING; WEIGHT FUNCTION; VARIABLE KERNEL; CROSS-VALIDATION; AUTOMATIC SMOOTHING; RESIDUALS; REGRESSION DIAGNOSTICS; LOCAL REWEIGHTING; CHANGE POINT; MODEL CHOICE; BAYESIAN INFERENCE; EMPIRICAL BAYES; B-SPLINES; GROWTH CURVES; FUNCTIONALS OF CURVES; ROBUST SMOOTHING; GENERALIZED SMOOTHING; SURFACE ESTIMATION

1. INTRODUCTION

Consider the regression problem where we have *observations* Y_i at *design points* t_i, $i = 1, \ldots, n$ and the observations are assumed to satisfy

$$Y_i = g(t_i) + \epsilon_i. \tag{1.1}$$

In this paper the non-parametric estimation of the function g will be discussed. It will be assumed that the design points satisfy $t_1 \leqslant t_2 \leqslant \ldots \leqslant t_n$ and that the errors ϵ_i are uncorrelated with zero mean. At first the variances of the ϵ_i will be assumed to be equal, but later this assumption will be relaxed.

1.1. *Motivation*

Before embarking on any technical details, it is important to consider the reasons why we might be interested in the estimation of the curve g, or, indeed, in any regression technique. Regression, of whatever kind, has two main purposes. Firstly, it provides a way of exploring and presenting the relationship between the design variable and the response variable; secondly, it gives predictions of observations yet to be made. A method for estimating a curve g will also be used for a third purpose, to give estimates of interesting properties of g. For example, in the study of growth curves the maximum rate of growth is an important quantity and so the maximum of the derivative of g will be of interest. This example will be discussed further in Section 7 below.

Especially for the first and third of these purposes, a non-parametric method of estimation is

Present address: School of Mathematics, University of Bath, Bath, BA2 7AY.

 0035–9246/85/47001 $2.00

Drs W. Greblicki and M. Pawlak (Institute of Engineering Cybernetics, Wroclaw, Poland): We would like to point out one difference between various nonparametric estimates and we shall do it in the context of density estimation. For an unknown density f having p derivatives, Wahba (1975) suggested using a kernel K such that

$$\int y^i K(y)\,dy = 0, \quad i = 1, \ldots, p-1 \text{ and } \int y^p \mid K(y) \mid dy < \infty.$$

Then, for the kernel estimate $\hat{f}(x)$ of $f(x)$

$$E\{\hat{f}(x) - f(x)\}^2 \leqslant c_1 h^{2p-1} + c_2/nh, \tag{1}$$

c_1 and c_2 are positive and h is the smoothing parameter. Hence, for $h(n) \sim n^{-1/2p}$

$$E\{\hat{f}(x) - f(x)\}^2 = O(n^{-(2p-1)/2p}). \tag{2}$$

In turn, for the estimate $\bar{f}$ using the cosine orthonormal series

$$E\{\bar{f}(x) - f(x)\}^2 \leqslant c_3 N^{-(2p-1)} + c_4 N/n, \tag{3}$$

where N is the number of orthonormal functions in the estimate—see also Wahba (1975). For $N(n) \sim n^{1/2p}$, the rate of the cosine series estimate equals that in (2). Thus, generally speaking, the rates achieved for the kernel and the cosine series are the same. The same conclusion is true for mean integrated square error—see Rosenblatt (1971) for the kernel estimate and Greblicki and Pawlak (1984) for the orthogonal series.

Now let the density be analytic, i.e., let it have all derivatives (which is often the case) and let the kernel be selected in the way suggested by Wahba (1975) (now p is a number arbitrarily chosen by a statistician). Then, the right side in (1) and consequently the rate in (2) remain unchanged. For the cosine series estimate, however,

$$E\{\bar{f}(x) - f(x)\}^2 \leqslant c_3 N^{-1/\delta} + c_4 N/n \tag{4}$$

$\delta > 0$. Now, taking e.g. $N(n) \sim n^{\epsilon}$, $\epsilon > 0$, we get

$$E\{\bar{f}(x) - f(x)\}^2 = O(n^{-1+\epsilon}).$$

By selecting ϵ sufficiently small, one can obtain a rate better than that achieved for the kernel estimate. A similar property can be also observed for estimates employing other orthogonal series (see e.g. Greblicki and Pawlak, 1984).

In view of this it seems that, for analytic densities, orthogonal series estimates behave better than the kernel estimate. This difference between these two estimates is caused by the fact that the kernel is selected according to p, i.e. the number of existing derivatives of the unknown density, while in the orthogonal series estimate the kernel is uniquely determined by the

Christoffel-Darboux formula and gives a good fit of the estimate to smooth densities. Similar properties of both estimates can also be observed while estimating regression functions according to the model examined by Silverman in his paper. One problem, however, arises for analytic regressions (and densities): do spline estimates behave as the kernel or the orthogonal series estimate?