Applied nonparametric regression

Wolfgang Härdle
Fakultät Rechts- und Staatswissenschaften
Wirtschaftstheoretische Abteilung II
Adenauerallee 24-26
Rheinische-Friedrich-Wilhelms Universität
D-5300 Bonn
15.6. 1989

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, UK http: //www.cup.cam.ac.uk
40 West 20th Street, New York, NY 10011-4211, USA http: //www.cup.org
10 Stamford Road, Oakleigh, Melbourne 3166, Australia

First published 1990
First paperback edition 1991
Reprinted 1992, 1993, 1995, 1997, 1999

A catalogue record for this book is available from the British Library

Library of Congress Cataloguing-in-Publication data is available

ISBN 0 521 38248 3 hardback
ISBN 0 521 42950 1 paperback

Transferred to digital printing 2002

Contents

CHAPTER 1

Introduction

As regards problems of specification, these are entirely a matter for the practical statistician, for those cases where the qualitative nature of the hypothetical population is known do not involve any problems of this type.

Sir R. A. Fisher (1922)

A regression curve describes a general relationship between an explanatory variable X and a response variable Y. Having observed X, the average value of Y is given by the regression function. It is of great interest to have some knowledge about this relation. The form of the regression function may tell us where higher Y-observations are to be expected for certain values of X or whether a special sort of dependence between the two variables is indicated. Interesting special features are, for instance, monotonicity or unimodality. Other characteristics include the location of zeros or the size of extrema. Also, quite often the regression curve itself is not the target of interest but rather derivatives of it or other functionals.

If n data points $\{(X_i, Y_i)\}_{i=1}^n$ have been collected, the regression relationship can be modeled as

$$Y_i = m(X_i) + \varepsilon_i, \qquad i = 1, \ldots, n,$$

with the unknown regression function m and observation errors ε_i. A look at a scatter plot of X_i versus Y_i does not always suffice to establish an interpretable regression relationship. The eye is sometimes distracted by extreme points or fuzzy structures. An example is given in Figure 1.1, a scatter plot of X_i = rescaled net income versus Y_i = expenditure for potatoes from the Family Expenditure Survey (1968–1983). The scatter of points is presented in the form of a sunflower plot (see Cleveland and McGill, 1984, for construction of sunflower plots).

In this particular situation one is interested in estimating the mean expenditure as a function of income. The main body of the data covers only a quarter of the diagram with a bad "signal to ink ratio" (Tufte 1983): it seems therefore to be difficult to determine the average expenditure for given income X. The aim of a regression analysis is to produce a reasonable approximation to the unknown response function m. By

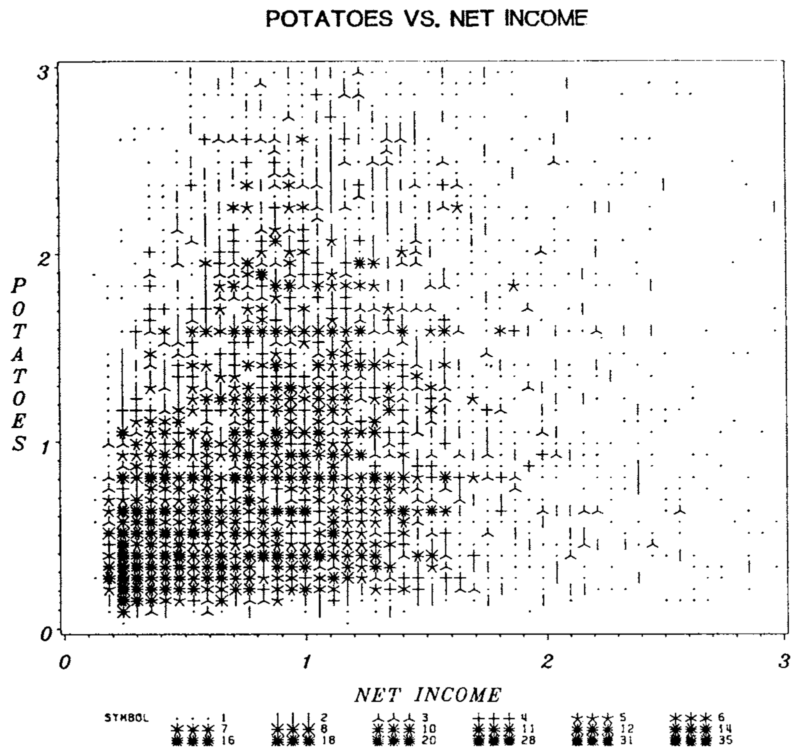

Figure 1.1. Potatoes versus net income. Sunflower plot of Y = expenditure for potatoes versus X = net income of British households for year 1973, n = 7125. Units are multiples of mean income and mean expenditure, respectively. The number of petals of the sunflower indicates the frequency of observations falling in the cell covered by the sunflower. Family Expenditure Survey (1968–1983).

reducing the observational errors it allows interpretation to concentrate on important details of the mean dependence of Y on X. This curve approximation procedure is commonly called "smoothing."

This task of approximating the mean function can be done essentially in two ways. The quite often used *parametric* approach is to assume that the mean curve m has some prespecified functional form, for example, a line with unknown slope and intercept. As an alternative one could try to estimate m *nonparametrically* without reference to a specific form. The first approach to analyze a regression relationship is called parametric since it is assumed that the functional form is fully described by a finite set of parameters. A typical example of a parametric model is a polynomial regression equation where the parameters are the coefficients of the independent variables. A tacit assumption of the parametric approach though is that the curve can be represented in terms of the parametric model or that, at least, it is believed that the approximation bias of the best parametric fit is a negligible quantity. By contrast, nonparametric modeling of a regression relationship does not project the observed data into a Procrustean bed of a fixed parametrization, for example, fit a line to the potato data. A preselected parametric model might be too restricted or too low-dimensional to fit unexpected features, whereas the nonparametric smoothing approach offers a flexible tool in analyzing unknown regression relationships.

The term *nonparametric* thus refers to the flexible functional form of the regression curve. There are other notions of "nonparametric statistics" which refer mostly to distribution-free methods. In the present context, however, neither the error distribution nor the functional form of the mean function is prespecified.

The question of which approach should be taken in data analysis was a key issue in a bitter fight between Pearson and Fisher in the twenties. Fisher pointed out that the nonparametric approach gave generally poor efficiency whereas Pearson was more concerned about the specification question. Tapia and Thompson (1978) summarize this discussion in the related setting of density estimation.

Fisher neatly side-stepped the question of what to do in case one did not know the functional form of the unknown density. He did this by separating the problem of determining the form of the unknown density (in Fisher's terminology, the problem of "specification") from the problem of determining the parameters which characterize a specified density (in Fisher's terminology, the problem of "estimation").

Both viewpoints are interesting in their own right. Pearson pointed out that the price we have to pay for pure parametric fitting is the possibility of gross misspecification resulting in too high a model bias. On the other hand, Fisher was concerned about a too pure consideration of parameter-free models which may result in more variable estimates, especially for small sample size n.

An example for these two different approaches is given in Figure 1.2, where the straight line indicates a linear parametric fit (Leser 1963, eq. 2a) and the other curve is a nonparametric smoothing estimate. Both curves model the market demand for potatoes as a function of income from the point cloud presented in Figure 1.1. The linear parametric model is unable to represent a decreasing demand for potatoes as a function of increasing income. The nonparametric smoothing ap-

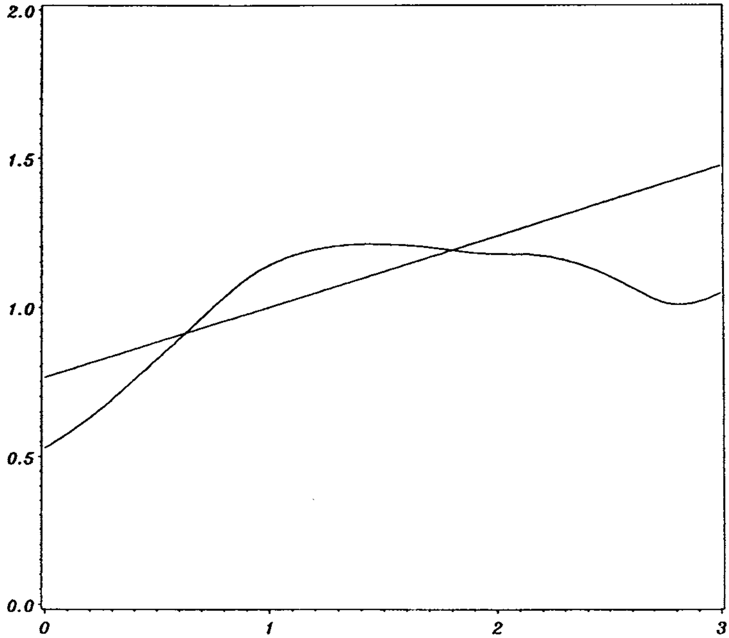

Figure 1.2. Potatoes versus net income. A linear parametric fit of Y = expenditure for potatoes versus X = net income (straight line) and a nonparametric kernel smoother (bandwidth = 0.4) for the same variables, year 1973, n = 7125. Units are multiples of mean income and mean expenditure, respectively. Family Expenditure Survey (1968–1983).

proach suggests here rather an approximate U-shaped regression relation between income and expenditure for potatoes. Of course, to make this graphical way of assessing features more precise we need to know how much variability we have to expect when using the nonparametric approach. This is discussed in Chapter 4. Another approach could be to combine the advantages of both methods in a semiparametric mixture. This line of thought is discussed in Chapters 9 and 10.

1.1 Motivation

The nonparametric approach to estimating a regression curve has four main purposes. First, it provides a versatile method of exploring a general relationship between two variables. Second, it gives predictions of observations yet to be made without reference to a fixed parametric model. Third, it provides a tool for finding spurious observations by studying the influence of isolated points. Fourth, it constitutes a flexible method of substituting for missing values or interpolating between adjacent X-values.

The *flexibility* of the method is extremely helpful in a preliminary and exploratory statistical analysis of a data set. If no a priori model information about the regression curve is available, the nonparametric analysis could help in suggesting simple parametric formulations of the regression relationship. An example is depicted in Figure 1.3. In that study of human longitudinal growth curves the target of interest was the first (respectively, second) derivative of the regression function (Gasser et al. 1984; Jørgensen et al. 1985).

The nonparametric regression smoothing method revealed an extra peak in the first derivative, the so-called mid-growth spurt at the age of about eight years. Other approaches based on ad hoc parametric modeling made it extremely difficult to detect this extra peak (dashed line Figure 1.3).

An analogous situation in the related field of density estimation was reported by Hildenbrand (1986) for the income density income of British households. It is important in economic theory, especially in demand and equilibrium theory, to have good approximations to income distributions. A traditional parametric fit – the Singh–Madalla model – resulted in Figure 1.4.

The parametric model class of Singh–Madalla densities can only produce unimodal densities per se. By contrast, the more flexible nonparametric smoothing method produced Figure 1.5. The nonparametric approach makes it possible to estimate functions of greater complexity and suggests instead a bimodal income distribution. This bimodality is present over the thirteen years from 1968 to 1981 and changes its shape: More people enter the "lower income range" and the "middle class" peak becomes less dominant.

An example which once more underlines this flexibility of modeling regression curves is presented in Engle et al. (1986). They consider a nonlinear relationship between electricity sales and temperature using a parametric–nonparametric estimation procedure. Figure 1.6 shows the result of a spline smoothing procedure that nicely models a kink in the electricity sales.

Another example arises in modeling alcohol concentration curves. A commonly used practice in forensic medicine is to approximate ethanol reduction curves with parametric models. More specifically, a linear regression model is used which in a simple way gives the so-called β_{60} value, the ethanol reduction rate per hour. In practice, of course, this

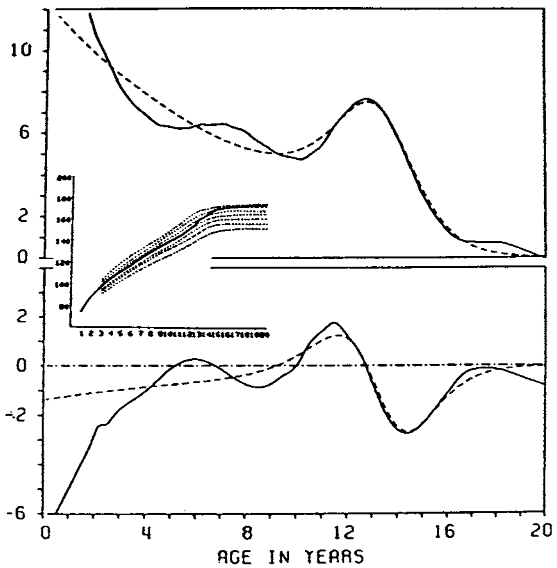

Figure 1.3. Human height growth versus age. The small graph gives raw data of height connected by straight lines (solid line) with cross-sectional sample quantiles (dashed lines). Velocity of height growth of a girl (above) and acceleration (below) modeled by a nonparametric smoother (solid line) and a parametric fit (dashed line). Units are cm (for height), cm/year (for velocity) and cm/year^2 (for acceleration). From Gasser and Müller (1984 figure 1) with the permission of the *Scandinavian Journal of Statistics*.

model can be used only in a very limited time interval; an extension into the "late ethanol reduction region" would not be possible. A nonparametric analysis based on splines suggested a mixture of a linear and exponential reduction curve. (Mattern et al. 1983).

The *prediction* of new observations is of particular interest in time series analysis. It has been observed by a number of people that in certain applications classical parametric models are too restrictive to give reasonable explanations of observed phenomena. The nonparametric prediction of times series has been investigated by Robinson (1983) and Doukhan and Ghindes (1985). Ullah (1987) applies kernel smoothing to a time series of stock market prices and estimates certain risk indexes. Deaton (1988) uses smoothing methods to examine demand patterns in Thailand and investigates how knowledge of those patterns affects the assessment of pricing policies. Yakowitz (1985b) applies smoothing techniques for one-day-ahead prediction of river flow. Figure 1.7 be-

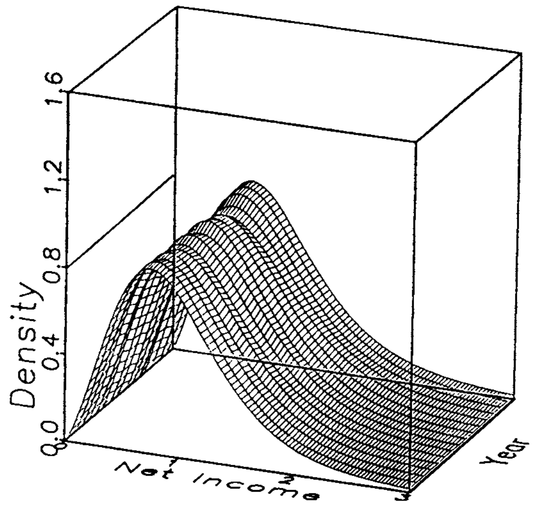

Figure 1.4. Net income densities over time. A Singh–Madalla fit to the densities of X = net income from 1969 to 1983. Units are mean income for each year. Family Expenditure Survey (1968–1983).

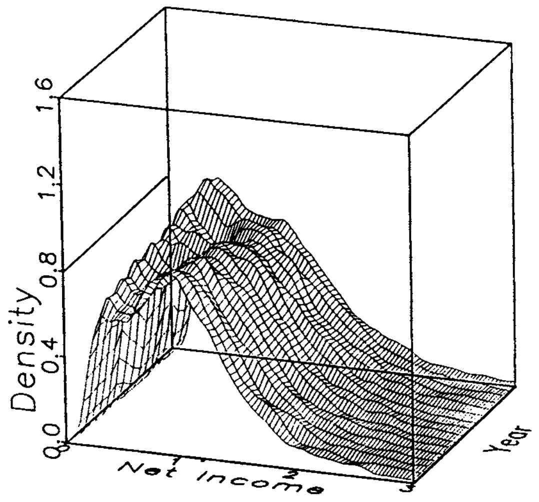

Figure 1.5. Net income densities over time. A nonparametric kernel fit (bandwidth $h = 0.2$) to the densities of X = net income from 1969 to 1981. Units are mean income for each year. Family Expenditure Survey (1968–1983).

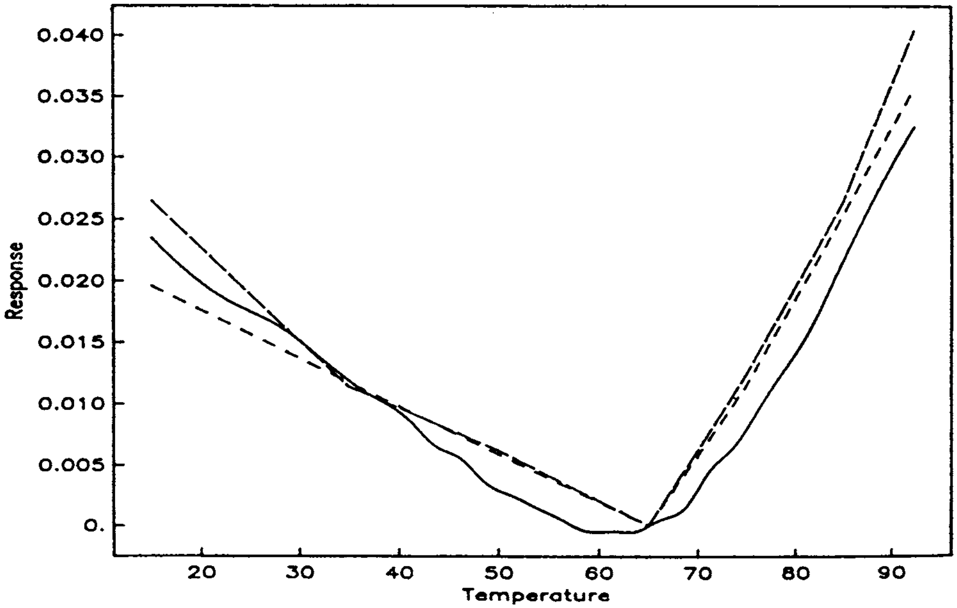

Figure 1.6. Temperature response function for Georgia. The non-parametric estimate is given by the solid curve and two parametric estimates by the dashed curves. From Engle et al. (1986) with the permission of the American Statistical Association.

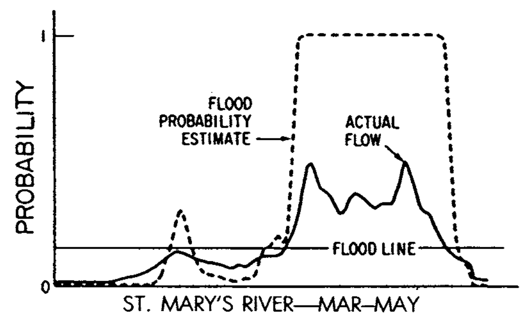

Figure 1.7. Nonparametric flow probability for the St. Mary's river. From Yakowitz (1985b) with permission of the Water Resources Research.

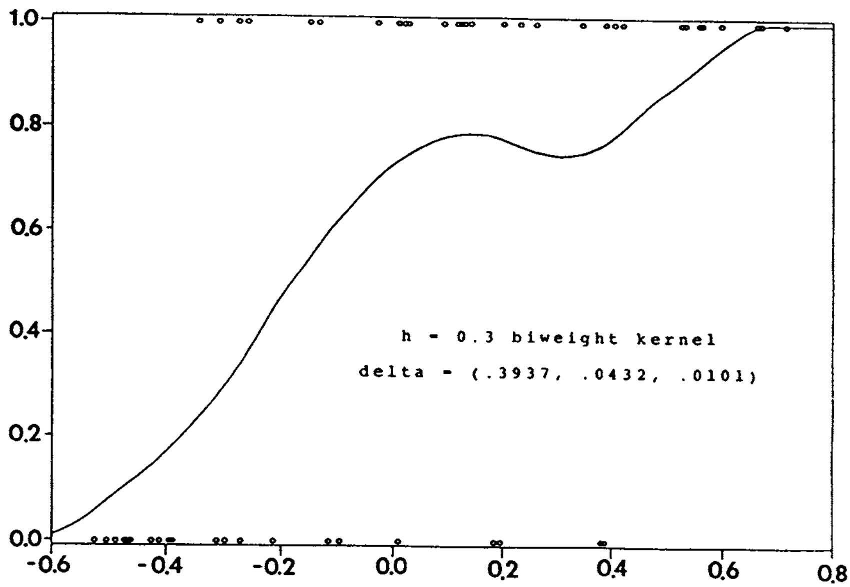

Figure 1.8. Indicators of fatal injury ($Y = 1$) as a function of an injury stress index together with an estimate of the regression curve. From Härdle and Scott (1988).

low shows a nonparametric estimate of the flow probability for the St. Mary's river.

A treatment of *outliers* is an important step in highlighting features of a data set. Extreme points affect the scale of plots so that the structure of the main body of the data can become invisible. There is a rich literature on robust parametric methods in which different kinds of outlier influence are discussed. There are a number of diagnostic techniques for parametric models which can usually cope with outliers. However, with some parametric models one may not even be able to diagnose an implausible value since the parameters could be completely distorted by the outliers. This is true in particular for isolated (leverage) points in the predictor variable X. An example is given in Rousseouw and Yohai (1984) in which a linear regression line fitted a few outliers but missed the main body of the data. Nonparametric smoothing provides a versatile pre-screening method for outliers in the x-direction without reference to a specific parametric model. Figure 1.8 shows a nonparametric smoother applied to analysis of simulated side impact studies. The curve shown is an approximation to the probability of a fatal injury as a function of anthropometric and biokinetic parameters. The Y-ordinates are binary in this case ($Y = 1$ denoting fatal injury). The curve shows visually what could also be derived from an influence analysis: it makes a dip at the isolated x-points in the far right. The points could be identified as observations from young persons which had a rather unnormal reaction behavior in these experiments; see Kallieris and Mattern (1984). This example is discussed in more detail in Section 10.4.

Missing data is a problem quite often encountered in practice. Some response variables may not have been recorded since an instrument broke down or a certain entry on an inquiry form was not answered. Nonparametric smoothing bridges the gap of missing data by interpolating between adjacent data points, whereas parametric models would involve all the observations in the interpolation. An approach in spatial statistics is to interpolate points by the "kriging" method. This method is used by statisticians in hydrology, mining, and petroleum engineering and is related to predicting values of noisy data in a nonparametric fashion; see Yakowitz and Szidarovszky (1986). Schmerling and Peil (1985) use local polynomial interpolation in anatomy to extrapolate missing data.

1.2 Scope of this book

This book takes the viewpoint of an applied statistician who is interested in a flexible regression analysis of exploratory character. In this spirit, I shall concentrate on simple smoothing techniques and analyze problems that typically arise in applications. Important practical questions are:

> *What is the right amount of smoothing?*
>
> *How close is the estimated curve to the underlying curve?*
>
> *How can we effectively estimate curves in dimensions higher than three?*

One of the simplest smoothing techniques is kernel estimation. It is straightforward to implement without further mathematical knowledge and it is understandable on an intuitive level. It is argued in Chapter 2 that kernel smoothing is a suitable tool in many ways. A variety of alternative smoothing techniques such as splines are discussed as well. In Chapter 3 it is seen that they are, in an asymptotic sense, equivalent to kernel smoothing.

The decision about the right amount of smoothing is crucial. Every smoothing method has to be tuned by some smoothing parameter which balances the degree of fidelity to the data against the smoothness of the estimated curve. A choice of the smoothing parameter has to be made in practice and controls the performance of the estimators. This *smoothing parameter selection problem* will be discussed in great detail and will be a centerpiece of this book (Chapters 4 and 5). The user of a nonparametric smoothing technique should be aware that the final decision about an estimated regression curve is partly subjective since even asymptotically optimal smoothers contain a considerable amount of noise that leaves space for subjective judgment. It is therefore of great importance to make such a decision in interaction with the data, which means that ideally one should have computer resources with some sort of interactive graphical display. Bearing this in mind, a great deal of the discussion will be devoted to algorithmic aspects of nonparametric smoothing.

In Chapters 6 and 7 I discuss smoothing in the presence of outliers and correlation, respectively. In Chapter 8 smoothing under qualitative constraints, such as monotonicity or more general piecewise monotonicity, is presented. Smoothing in dimensions higher than three creates problems on the computational and on the statistical side of the estimator. It takes longer to compute the estimators and the accuracy decreases exponentially as the dimension grows. Chapter 9 presents some semiparametric approaches to incorporate parametric components into nonparametric smoothing. Chapter 10 discusses additive models and gives some heuristics as to why these models achieve better accuracy and in this sense reduce the dimension problem.

The great flexibility of nonparametric curve estimation makes a precise theoretical description of the accuracy of the smoothers for finite sample sizes extremely difficult. It is therefore necessary to achieve some sort of simplification. This is done here in two ways. First, the mathematical arguments are of asymptotic character, that is, the accuracy of the nonparametric smoothing method will be evaluated as the sample size n tends to infinity. Second, the class of smoothers that is mainly considered here is of very simple structure (kernel estimators).

The reader interested in the applied aspects should not be too disappointed about the asymptotic mathematical results. I have tried to present them in the spirit aptly described by Murray Rosenblatt:

The arguments ... have been of an asymptotic character and it is a mistake to take them too literally from a finite sample point of view. But even asymptotic arguments if used and interpreted with care can yield meaningful ideas.

Technical details of the mathematical theory are kept simple or else deferred to exercises and complements. I believe that each chapter provides stimulation to work out some of the mathematical arguments. Some practically oriented readers might find themselves encouraged to try the methods in practice. This can be done, for instance, with graphically oriented computing environments and systems such as GAUSS (1987), ISP (1987), S (1988) or XploRe (1989).