(1989) Härdle, W. Discussion of papers by Eagleson, Tibshirani and Rice

Proceedings of the International Statistical Institute (ISI) Meeting, Paris, September (1989)

Paper not retrievable

Resampling for Inference from Curves

W.Härdle
Wirtschaftstheorie II, Adenauerallee 24-26
Rheinische-Friedrich-Wilhelms Universität
D-5300 Bonn

Abstract

Nonparametric curve estimation resampling methods have a long tradition. Cross-Validation is used for instance to optimize the smoothing parameter. In this paper a resampling method is studied that is helpful in drawing inferences from curves. More specifically a variant of the Bootstrap is proposed to construct errorbars and to compare with parametric curves. This so-called Wild Bootstrap is easy to implement and does not require complicated plug-in estimation.

RESUME

Rééchantillonner pour faire de l'estimation non paramétrique, est une technique ancienne.

Les méthodes de validation croisée sont utilisées, par exemple, pour optimiser le paramètre de lissage.

Dans ce papier, on étudie une méthode de rééchantillonnage utile pour faire de l'inférence sur les courbes de régression.

Plus spécialement, une variante du Bootstrap (facile à mettre en oeuvre), appelée Wild-Bootstrap est proposée pour construire des intervalles de confiance et est comparée avec les techniques paramétriques.

1. The need for computer assistance

A typical task for a statistician is that of model construction and comparison with known or traditional models. In curve estimation a common approach to this task is to start with a nonparametric curve estimate and then to analyze its qualitative features. Certain shape characteristics (e.g. the location of peaks) guide and help in proposing and constructing a suitable (parametric) model.

A good example for this approach is the human growth curve study by Gasser et al. (1984). They compared a nonparametric regression growth curve with a traditional parametric model and found that the parametric model did not model a pre-pubertal growth spurt. In the field of density estimation Marron and Schmitz (1989) describe the evolution of income distributions over time. They found that, in contrast to more traditional log normal density estimates, the nonparametric curve shows two distinct modes that were changing height and location.

A typical scenario in these studies was the interactive graphical comparison of the curve estimates. Curves have been compared for example with parametric fits or among each other when a smooothing parameter varied. Of course this is only a method of "graphical inference" but it helps in developing a sense for the real shape of a curve. Usual methods for the inference in curve estimation include error bars or measures of distance between curves. For sensible inference these error bars should be constructed with simultaneous coverage probability.

Both approaches for inference in curve estimation have been done theoretically, see e.g. Konakov and Piterbarg (1984); Härdle and Mammen (1989). A drawback of this theoretical approach is that its use in practice requires "plug-in" estimation of complicated functionals of the data distribution. The purpose of this paper is to show how resampling techniques help in finding asymptotically correct error bars or the distribution of a test statistic for comparing nonparametric with parametric regression models. These models are completely automatic resampling methods and require no knowledge about the functionals entering the asymptotic distributions of the test statistics.

In section 2 I describe the Bootstrap in the setting of curve estimation. It is called the *Wild Bootstrap* since resampling is done from *one* single residual to infer the conditional distribution. Section 3 is devoted to the problem of comparison between nonparametric and parametric regression models. The construction of simulated simultaneous error bars with the wild bootstrap is described in Section 4.

2. The Wild Bootstrap in Curve Estimation

Stochastic design regression is based on iid. observations $\{(X_i, Y_i)\}_{i=1}^n \in \mathbb{R}^{d+1}$ and the goal is to estimate $m(x) = E(Y|X = x) : \mathbb{R}^d \to \mathbb{R}$. The form of the nonparametric kernel regression estimator, developed by Nadaraya (1964) and Watson (1964) is

$$\hat{m}_h(x) = n^{-1} \sum_{i=1}^{n} K_h(x - X_i) Y_i / \hat{f}_h(x) \tag{2.1}$$

where

$$\hat{f}_h(x) = n^{-1} \sum_{i=1}^{n} K_h(x - X_i) \tag{2.2}$$

and where $K_h(u) = h^{-d} K(u/h)$ is a kernel weight function with bandwidth h. All results of this paper are given in terms of this estimator. The essential ideas though carry over to other regression estimators.

Resampling methods in nonparametric regression are used for a variety of purposes. Asymptotically optimal bandwith sequences for example are found by cross-validation, see Rice (1984); Härdle and Marron (1985). I concentrate on the Bootstrap resampling method here. Bootstrap techniques are well known tools for assessing variability. In the present context a little care has to be taken to properly account for smoothing bias. In particular the so-called *naive Bootstrap*

Resample from the pairs $\{(X_i, Y_i)\}_{i=1}^n$

is not appropiate for the questions I deal with here. The naive bootstrap does not reflect the bias correctly. Indeed

$$E^* \hat{m}_h^*(x) \quad - \quad \hat{m}_h(x) \quad = \quad 0,$$

where E^* denotes expectation under the bootstrap distribution and $\hat{m}_h^*(x)$ denotes the above (2.1) computed from the bootstrap observations $\{(X_i, Y_i)\}_{i=1}^n$.

My approach to this problem is to resample on the basis of the residuals

$$\hat{\varepsilon}_i = Y_i - \hat{m}_h(X_i)$$

Note that at each X_i we may have a different conditional distribution G_i of $(Y_i|X_i)$, so we should not resample from the whole set of residuals as in Härdle and Bowman (1988).

In order to retain the characteristics of G_i I will use the Wild Bootstrap which is a resampling method based on the single residuals $\hat{\varepsilon}_i$. More precisely I define a two point destribution $\hat{G}_i$ which has mean zero, variance equal to the square of the residuals, and third moment equal to the cube of the residual. Some algebra reveals that if $\hat{G}_i = \gamma\delta_a + (1-\gamma)\delta_b$, then $a = \hat{\varepsilon}_i(1-\sqrt{5})/2$, $b = \hat{\varepsilon}_i(1+\sqrt{5})/2$ and $\gamma = (5+\sqrt{5})/10$. These parameters ensure that if $\varepsilon_i^* \sim \hat{G}_i$, then $E\varepsilon^* = 0$, $E\varepsilon^{*2} = \hat{\varepsilon}_i^2$, $E\varepsilon^{*3} = \hat{\varepsilon}_i^3$. In a certain sense the resampling distribution can be thought of as attempting to reconstruct the distribution of each residual through the use of one single observation. Therefore it is called the Wild Bootstrap. More formally we have the following

Wild Bootstrap Algorithm

1. Define at each X_i the two point distribution $\hat{G}_i$.
2. Generate Bootstrap errors $\varepsilon_i^* \sim \hat{G}_i$.
3. Define Boootstrap observations

$$Y_i^* = \hat{m}_g(X_i) + \varepsilon_i^*;$$

for error bar computation, respectively

$$Y_i^* = m_{\hat{\theta}}(X_i) + \varepsilon_i^*$$

for comparison with a parametric model $\{m_\theta : \theta \in \Theta\}$.

Here $\hat{m}_g$ denotes the kernel smoother (2.1) with bandwith $g >> h$ and $m_{\hat{\theta}}$ the least squares estimator for a parametric model $\{m_\theta : \theta \in \Theta\}$ of the regression curve. A related resampling technique was considered by Wu (1986).

3. Resampling for comparison with a parametric model

Suppose that one has computed a nonparametric curve estimator together with a parametric fit from a model $\{m_\theta : \theta \in \Theta\}$ to explain the regression of Y on X. A possible way to compare these two curves is to compute the integrated squared differences between $\hat{m}_h$ and a parametric fit $m_{\hat{\theta}}$. Since

$$E(\hat{m}_h(x)|\underset{\sim}{X}) = K_{h,n}m(x)$$

where $\underset{\sim}{X} = \{X_i\}_{i=1}^n$ and

$$K_{h,n}p(\bullet) = n^{-1}\sum_{i=1}^{n} K_h(\bullet - X_i)p(X_i)/\hat{f}_h(\bullet)$$

it makes more sense to compare $\hat{m}_h(\bullet)$ with $K_{h,n}m_{\hat{\theta}}(\bullet)$. Therefore I propose to consider

$$T_n = nh^{d/2}\int(\hat{m}_h(x) - K_{h,n}m_{\hat{\theta}}(x))^2 w(x)dx \tag{3.1}$$

as a teststatistic to test the parametric hypothesis $m \in \{m_\theta : \theta \in \Theta\}$. Here w denotes a weight function.

This test statistic will be small under the hypothesis and can be interpretated as a smoothed variant of the χ^2-statistic. For linear regression models,

$$m(x) = \sum_{j=1}^{k}\theta_j p_j(x) =< \theta, p(x) >,$$

it is easy to see that the Least Squares estimate $m_{\hat{\theta}}$ can be expanded as

$$m_{\hat{\theta}}(x) = m_{\theta_0}(x) + n^{-1}\sum_{i=1}^{n} < p(x), q(X_i) > \varepsilon_i + o_p((n \quad log \quad n)^{-1/2}) \tag{3.2}$$

with bounded functions p, q. Assume now that the kernel K is a bounded symmetric probability density function with compact support and that $h \sim n^{-1/(4+d)}$, the optimal rate for estimating m nonparametrically (Stone, 1982). Under smoothness conditions on $m, f, \sigma^2(x) = var(Y|X = x)$ and moment assumptions on ε_i, Härdle and Mammen (1989) have shown

Theorem 1. *Under the hypothesis "$m \in \{m_\theta : \theta \in \Theta\}$" and validity of expansion (3.2)*

$$d(\mathcal{L}(T_n), N(b_h, V)) \to 0,$$

where

$$b_h = h^{-d/2} K^{(2)}(0) \int \frac{\sigma^2(x) w(x)}{f(x)} dx,$$

$$V = 2K^{(4)}(0) \int \frac{\sigma^4(x) w(x)^2}{f^2(x)} dx,$$

$d(\bullet, \bullet)$ denotes the Mallows metric and $K_h^{(j)}$ denotes the j-times convolution product of K_h.

For a practically oriented statistician an application of this theorem might be a nightmare: He has to estimate all those rather complicated constants in Theorem 1 and then to plug them in into the asymptotic distribution to obtain level α sets for hypothesis testing. A way to avoid such obstacles is an automatic resampling method which yields the desired rejection regions. In a first attempt we could try to simulate the distribution of T_n by using the naive bootstrap. Unfortunately this method fails in approximating the $N(b_h, V)$ distribution of T_n, see Theorem 1. The reason lies in the fact that the regression function is not the conditional expectation of the observation under the bootstrap distribution. Therefore the bias is not correctly reflected.

The Wild Bootstrap works though: The statistic T_n^* computed from simulated data as described in the above wild bootstrap algorithm has the correct asymptotic normal distribution. More formally we can write for a resampling scheme over B replications:

```
FOR b=1 TO B DO BEGIN
1.  Generate Wild Bootstrap observations $(X_i, Y_i^*)$.
2.  Create $T_n^*$ like $T_n$.
END
From $\mathcal{L}^*(T_n^*)$ define the $(1-\alpha)$ quantile $\hat{t}_\alpha^*$
  and REJECT, IF $T_n > \hat{t}_\alpha^*$.
```

A proof for the correctness of this procedure can be found in Härdle and Mammen (1989). How well the bootstrap distribution $\mathcal{L}^*(T_n^*)$ approximates $\mathcal{L}(T_n)$ in seen from the plot below. It shows from $M = 1000$ Monte Carlo runs and $B = 100$ resampling steps four distributions approximating $\mathcal{L}(T_n)$.

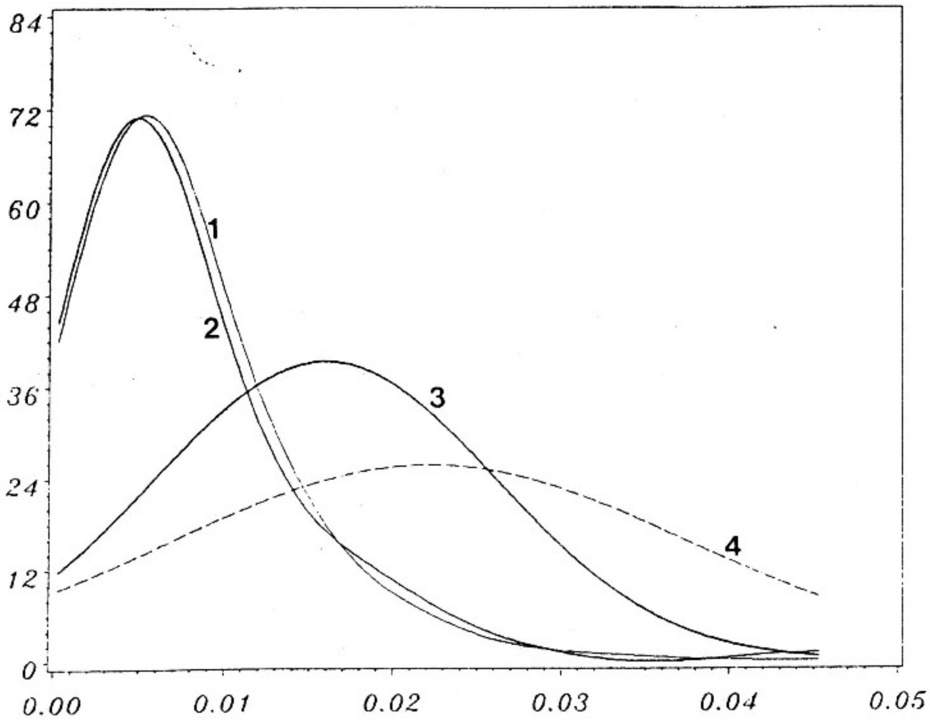


Figure 1. Four densities of T_n. The line with label 1 denotes the (kernel) density of the distribution of T_n over ALL Monte Carlo runs (M = 1000). The line with label 2 is the (kernel) density of T_n from ONE Monte Carlo run using the Wild Bootstrap method from B = 100 bootstrap curves. The curve labelled 3 is the Normal theory density from Theorem 1 with the KNOWN constants b_h and V. The curve labelled 4 is the Normal theory density from Theorem 1 with ESTIMATED constants $\hat{b}_h$ and $\hat{V}$. From Härdle and Mammen (1988).

The thin line (label 1) denotes the Monte Carlo kernel density estimate of the T_n-distance from the M runs. The medium thin line (label 2) is the kernel density of one bootstrap sample out of the M runs (taken at random). The thick line corresponds to the Normal theory density as given in Theorem 1 based on the true b_h and V (label 3). The dashed line finally shows the Normal theory density based on estimated b_h and V (label 4). In all four cases the bootstrap estimates the distribution of the distance quite well. The normal approximations are totally misleading. Power estimates and an application of this technique to the determination of the functional form of demand curves are presented in Härdle and Mammen (1989).

4. Simulated Simultaneous Error Bars.

Simultaneous error bars are intervals $I(x_j), j = 1, \ldots, N$ such that at distinct design points $x_1, \ldots, x_N$ with probability at least $(1-\alpha)$

$$\hat{m}_h(x_j) - m(x_j) \in I(x_j), j = 1, \ldots, N.$$

A quite common approach to this problem is to work with the limiting distribution of $\sqrt{nh^d}[\hat{m}_h(x_j) - m(x_j)]$ at the gridpoints $\underset{\sim}{x} = \{x_j\}_{j=1}^N$. From the limiting Normal distribution one can obtain via the "plug-in" method quantiles at the gridpoints and can correct for the level via the Bonferroni method.

The essential drawback of this approach is that it requires estimation of asymptotic bias and variance. In particular the bias of the Nadaraya-Watson estimator (2.1) is a rather complicated functional of the joint distribution of (X, Y), see Collomb (1981). A computer assisted automatic resampling method for finding error bars may resolve these practical problems. Consider the Wild Bootstrap again. Given a bootstrap sample $\{(X_i, Y_i^*)\}_{i=1}^n$ one can compute a kernel smoother $\hat{m}_h^*(x)$ from (2.1). The hope is now that a number of replications of $\hat{m}_h^*(x)$ can be used for approximating the distribution of

$$\sqrt{nh^d}[\hat{m}_h(\underset{\sim}{x}) - m(\underset{\sim}{x})].$$

This hope can be fulfilled if $\hat{m}_h^*(x)$ is correctly centered as the following theorem shows.

Theorem 2. *Given the assumptions of Theorem 1 (except (3.2), we have along almost all sample sequences and for all $\underset{\sim}{z} \in \mathbb{R}^N$*

$$\begin{aligned} sup_h sup_g |P^{Y|X}\{\sqrt{nh^d}[\hat{m}_h(\underset{\sim}{x}) - m(\underset{\sim}{x})] < \underset{\sim}{z}\} \\ - P^*\{\sqrt{nh^d}[\hat{m}_h^*(\underset{\sim}{x}) - \hat{m}_g(\underset{\sim}{x})] < \underset{\sim}{z}\}| \to 0, \quad n \to \infty. \end{aligned}$$

Here h and g run over sets

$$H_n = [\underline{c}n^{-1/(4+d)}, \bar{c}n^{-1/(4+d)}], \quad 0 < \underline{c} < \bar{c} < \infty.$$

$$G_n = [n^{-1/(4+d)+\delta}, n^{-\delta}], \delta > 0.$$

respectively.

For an intuitive understanding of why the bandwidth g used in the construction of the bootstrap residuals should be oversmoothed, consider the means of $\hat{m}_h(x) - m(x)$ under the $Y|X-$ distribution and $\hat{m}_h^*(x) - \hat{m}_g(x)$ under the $*-$ distribution in the simple situation when the marginal density $f(x)$ is constant in a neighborhood of x. Asymptotic analysis as in Rosenblatt (1969) shows that

$$E^{Y|X}(\hat{m}_h(x) - m(x)) \approx h^2(\int u^2 K/2)m^{''}(x),$$

$$E^*(\hat{m}_h^*(x) - \hat{m}_g(x)) \approx h^2(\int u^2 K/2)\hat{m}_g^{''}(x).$$

Hence for these two distributions to have the same bias we need $\hat{m}_g^{''}(x) \rightarrow m^{''}(x)$. This requires choosing g tending to zero at a rate slower than the *optimal bandwidth* h for estimating $m(x)$, see Gasser and Müller (1984). A data-driven method for chosing g is also reported in Härdle and Marron (1989).

As a practical method for finding the actual pointwise levels β_j at each x_j I suggest the following "halving" approach. In particular, motivated from the Bonferroni method, first try $\beta = \alpha/2M$, and calculate α_β. If the result is more than α/M, then try $\beta = \alpha/4M$, otherwise next try $\beta = 3\alpha/4M$. Continue this halving approach until neighboring (since only finitely many bootstrap replications are made, there is only a finite grid of possible β's available) values β_* and β^* are found so that $\alpha_{\beta_*} < \alpha/M < \alpha_{\beta^*}$. Finally take a weighted average of the β_* and the β^* intervals where the weights are $(\alpha_{\beta^*} - \alpha/M)/(\alpha_{\beta^*} - \alpha_{\beta_*})$ and $(\alpha/M - \alpha_{\beta_*})/(\alpha_{\beta^*} - \alpha_{\beta_*})$ respectively.

More formally we can write the algorithm for finding error bars as follows.

```
FOR b=1 TO B DO BEGIN
1.  Generate Wild Bootstrap observations $\{(X_i, Y_i^*)\}_{i=1}^n$,
                    $Y_i^* = \hat{m}_g(X_i) + \varepsilon_i^*, g >> h.$
2.  Create $\hat{m}_h^*(\underset{\sim}{x})$ like $\hat{m}_h(\underset{\sim}{x})$.
END
From $\mathcal{L}^*\{\sqrt{nh^d}[\hat{m}_h^*(\underset{\sim}{x}) - \hat{m}_g(\underset{\sim}{x})]\}$ define the error bars using the above
halving approach.
```

For an illustration of these ideas, consider Figure 2. Figure 2a shows a scatter plot of the expenditure for potatoes as a function of income for the year

1973, from the Family Expenditure Survey (1968-1983). Figure 2b shows a nonparametric regression estimate which was obtained by smoothing the point cloud, using the kernel algorithm. As a means of understanding the variability in the kernel smooth, Figure 2b also shows some error bars, constructed by the boostrap method proposed here. These bars are estimated simultaneous 80 % confidence intervals. Note that the error bars are longer on the right hand side, which reflects the fact that there are fewer observations there, and hence more uncertainty in the curve estimate. Note also that the error bars are asymmetric since the bootstrap method correctly corrects for the bias of the Nadaraya-Watson kernel smoother.

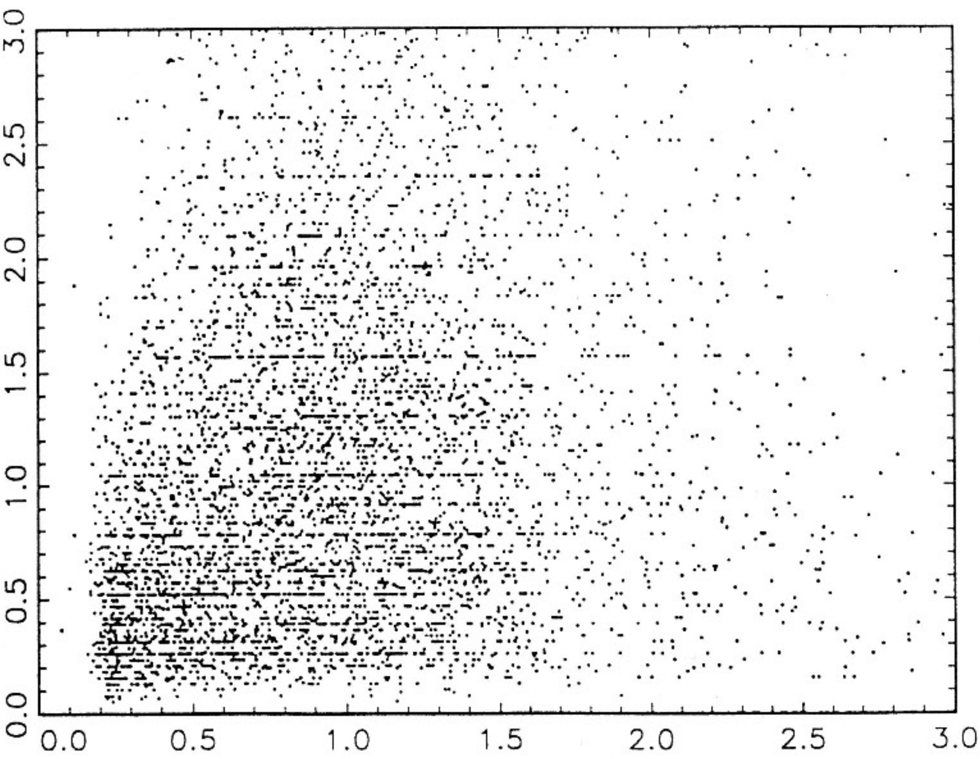


Figure 2b. Potato expenditure vs. income (a) Scatter plot (b) Regression smooth and error bars

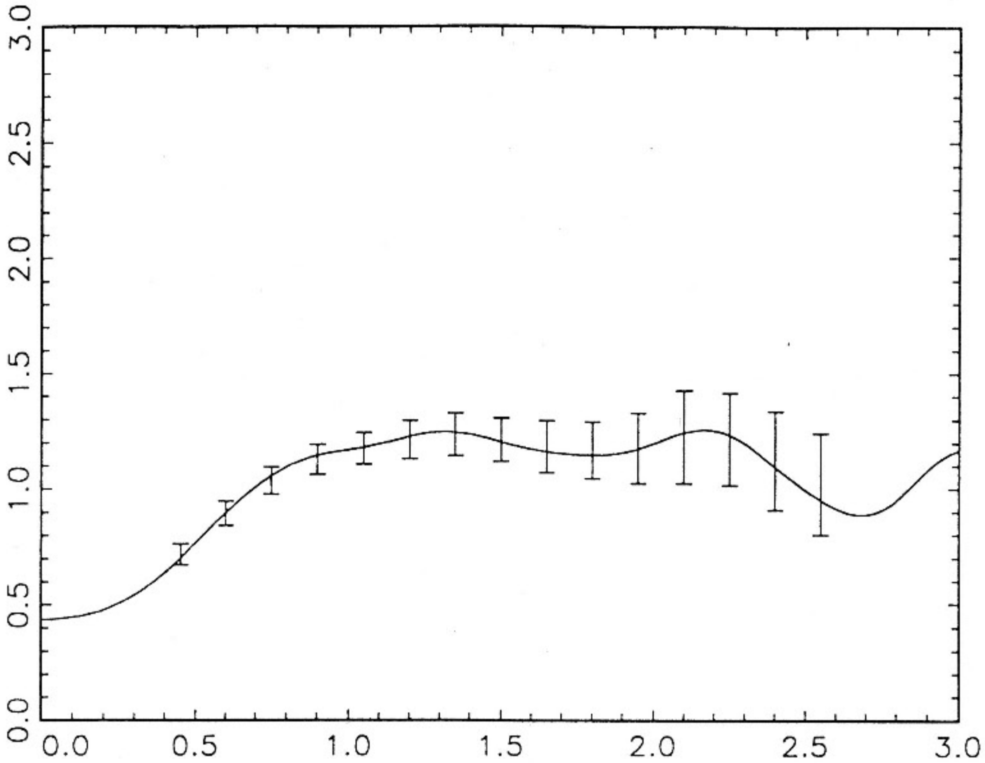


Figure 2a. Potato expenditure vs. income (a) Scatter plot (b) Regression smooth and error bars

References

Collomb, G. (1981). Estimation Non-paramétrique de la Régression: Revue Bibliographique. *International Statistical Review, 49, 75–93.*

Family Expenditure Survey, Annual Base Tapes (1968 – 1983) Department of Employment, Statistics Division, Her Majesty's Stationery Office, London 1968–1983. *The data utilized in this paper were made available by the ESRC Data Archive at the University of Essex.*

Gasser, T., Müller, H.G., Köhler, W., Molinari, L. and Prader, A. (1984). Nonparametric Regression Analysis of Growth Curves. *Annals of Statistics, 12, 210–229.*

Gasser, T. and Müller, H.G. (1984). Estimating Regression Functions and their Derivatives by the Kernel Method. *Scandanavian Journal of Statistics, 11, 171–185.*

Härdle, W. and Bowman, A. (1988). Bootstrapping in Nonparametric Regression: Local Adaptive Smoothing and Confidence Bands. *Journal of the American Statistical Association, 83, 102-110.*

Härdle, W. and Marron, J.S. (1985). Optimal Bandwidth Selection in Nonparametric Regression Function Estimation. *Annals of Statistics, 13, 1465-1481.*

Härdle, W. and Marron, J.S. (1989). Bootstrap Simultaneous Error Bars for Nonparametric Regression. *Annals of Statistics, submitted.*

Härdle, W. and Mammen, E. (1989). Comparing Nonparametric versus Parametric Regression Fits. *submitted to Econometrica*

Konakov, V.D. and Piterbarg, V.I. (1984). On the Convergence Rate of Maximal Deviation for Kernel Regression Estimates. *Journal of Multivariate Analysis, 15, 279-294.*

Marron, J.S. and Schmitz, H.P. (1989). Aggregation over Individuals and Size Distributions of Income. *SFB 303 Discussionpaper, A-186 Universität Bonn.*

Nadaraya, E.A. (1964). On Estimating Regression. *Theory Prob. Appl. 10, 186-190.*

Rice,J.A. (1984a). Bandwidth Choice for Nonparametric Regression. *Annals of Statistics, 12, 1215-30.*

Rosenblatt, M. (1969). Conditional Probability Density and Regression Estimators. *in: Multivariate Analysis II, 25-31, New York: Academic Press.*

Stone, C.J. (1982). Optimal Global Rates of Convergence for Nonparametric Regression. *Annals of Statistics, 10, 1040-1053.*

Watson, G.S. (1964). Smooth Regression Analysis. *Sankhyā, Series A, 26, 359-372.*

Wu, C.F.J. (1986). Jacknknife, bootstrap and other resampling methods in regression analysis. *Annals of Statistics, 14, 1261-1343.*

Proceedings of the International Symposium on Data Analysis, Learning symbolic and numeric knowledge, Antibes, September (1989)

THE INTERPLAY BETWEEN STATISTICS AND COMPUTING IN DATA ANALYSIS

Dedicated to Friedhelm Eicker

W.Härdle
Wirtschaftstheorie II
Adenauerallee 24-26
Rheinische-Friedrich-Wilhelms Universität
D-5300 Bonn

Abstract

The personal computer is becoming a widely used tool for data analysis. I describe some minimum requirements for a data analysis machine and discuss the theoretical objects (machine data structures) in order to efficiently organize the interplay between statistics and computing. The ideas are shown to be implementable on a desktop system like an IBM AT or PS/2. The particular implementation discussed here is called *XploRe* and is written in TURBO PASCAL 5.0.

Detecting structure is a typical task for a statistician interested in model construction and analysis of relations between variables. ”Structure” can have many meanings: outliers at specific locations (or indices), clusters of points, linear dependence between variables, etc. Before we can see any kind of structure we have to ask how to detect it and what tools do we need to provide in order to decide whether something is really a structure. Very simple structural elements of a data set may be detected just by looking at the list of values, e.g. a column with small numbers with the exception of one large one will indicate an outlier. On a finer level summary statistics for the variables and the use of graphical techniques like Boxplots help us to see data structures like skewness, for example. Suppose one has detected by these techniques that certain variables have a right skew distribution. A very natural next question is whether the points causing this skewness create other structural features in the other variables as well. Can we efficiently perform this task with one of the ”classical packages” ?

Most statistical packages now in use, such as SAS or MINITAB, were designed to operate in a batch-like mode. The user enters commands from the keyboard and the system scrolls subsequent stages of the analysis and displays on explicit request values or plots. Therefore the data analytic task, ”investigate the behavior of other variables given the detection of skewness in a particular variable” cannot be done instantaneously. Instead we have to write a separate program. This program has to first identify the indices of the interesting subset and then make several statistics and plots from other variables on the basis of this subset.

The development of hardware in the last years has made it possible to carry out this task and similar ones within a desktop system. It should be emphasized though that the hardware, in principle, gives the ”possiblity” to perform such operations efficently at a high ”power/price” ratio but in many cases the computing environment (the software) is not capable of doing what a data analyst wants to do. It is the object of this paper to propose some standards for an ideal computing environment and to demonstrate a first approximation to these ideals in an existing computing environment based on a widely used IBM desktop system (IBM AT, IBM PS/2).

The desktop concept allows us to analyze various facets of the data from different viewpoints (windows) at the same time. Stuetzle (1987) has made this evident, stating:

Each window and in particular each plot, corresponds to a sheet of paper on the desktop.

Before computers entered our work we shuffled paper on the desk to compare and evaluate different structural elements of the data. Now the paper shuffling is replaced by window clicking and logical links between the sheets are not only in the statistician's mind but also between defined computer objects inside the machine. These computer objects and their implementation are described below. It is argued in Section 2 that instead of expensive hardware it is a matter of software to bring data analysis into life, especially data analysis in high dimensions. In Section 3 I describe a current approximation to a data analysis computing environment. Section 4 is devoted to an example session with the system *XploRe*.

2. The Corvette versus the Chevette: Do four cylinders suffice ?

In analysing data we are programming, in effect, our view towards the data and the structures we want to and can see. By selecting certain scales we can concentrate our eyes dynamically on local or global features, see McDonald and Pederson (1986). For example, with a single column of numbers, by using LOG–transformations we can sharpen our eye for peaks or other local structures. As a consequence we need a flexible programming and computing environment that allows us to sharpen our eyes if necessary for interesting details on the macro or micro scale. Assisting tools in this task are four "cylinders" (minimum requirements):

⊙ The capability of showing 3D rotation motion.

⊙ The capability of a multi-window system.

⊙ A high resolution display (at least 600 × 400 pixels).

- ⊙ A graphical input system like a mouse or a tracker ball.

Even today, high resolution display and the mouse are available at low cost. The important remaining cylinders are the capability of a multi-window system and 3D motion. Having the possibility of a multi-window view on the data enhances our capability of finding structure and the 3D motion allows us to see in an exploratory way features in 3D space. These two cylinders can be driven and supported from different levels of the machine's hard- and software. By adding more cylinders (e.g. hardware), the multi-window technique, for example, can be realized on the operating system level or close to it. More cylinders mean in this case that window operations like opening, closing, moving, reshaping, etc. are available on a (fast) low level. On the other hand we need more than elementary window handling techniques.

Let me give a simple example. Suppose we are looking at similarly scaled 2D data repeatedly by studying a sequence of scatter plots. Our eyes would "detect" artifacts due to different scales unless the scale of the 2D scatterplot pictures are the same for all plots. It is therefore highly desirable to have the possibility of a STATIC2DPICTURE object (possibly of window type) which stays unchanged in subsequent calls. Practically speaking this means that a picture shown in a window system "has to know what scales it had in earlier calls". In a more modern language, we would say that the picture should have the capability of inheriting properties of earlier stages of an analysis.

This capability of inheritance in an Object Oriented Programming System (OOPS) creates extra difficulties in software construction. If we try to realize this concept of a statistical STATIC2DPICTURE object in a eight cylinder vehicle (e.g. the machine with a built in window system) we face extra difficulties since we have to artificially attach statistical meanings to windows constructed for a different purpose. Thus the construction of a "statistical data window" can bring us down to ordinary programming speed. With other words more cylinders can be more expensive and may even slow us down. For this reason some people have called such vehicles "Corvette data analysis machines".

The Corvettes are very powerful, not very frequent among the data analysis vehicles and sometimes have to drive the speed of a Chevette.

A Chevette is a data analysis machine that operates from a lower powered level but has the (dis)advantage of free object programming. Free object programming means that we can define our programming objects by ourselves in one horizontal level without time consuming "vertical diffusions" to other levels. Of course this is an advantage from a pure theoretical point of view but a disadvantage on the programmer's side. The other advantages, such as price and portability, make it clear that the "Chevette data analysis machine" is the preferable choice. Note that portability has two meanings here. First we can export the code more easily since it is written in one level although possibly simulating deeper levels. Second we can implement the code on portable machines (laptops) to show data analysis live to other statisticians ! Summarizing these thoughts leads me to the conclusion that a four cylinder personal computer like an IBM AT or PS/2 is acceptable for interactive data analysis if we tune it correctly e.g. program it correctly. In the following I describe a four cylinder Chevette called *XploRe*.

3. XploRing data

XploRe is an interactive, graphically oriented computing environment designed to analyse various kinds of relations between data and to apply and compare different smoothing methods. *XploRe* is suitable for investigating high dimensional data. It supports the user with the above four cylinders and other sophisticated data management tools such as masking, brushing, labeling and rotation of data. In addition, a wide variety of additive models are available, among them ACE (Alternating Conditional Expectations), Breiman and Friedman (1985), ADE (Average Derivative Estimation), Härdle and Stoker (1989), PPR (Projection Pursuit Regression), Friedman and Stuetzle (1981).

XploRe is designed as an open system. It is basically a framework awaiting more software in the form of user written submodules. The user can write his or her own programs and add them into *XploRe* via an object oriented user interface. The construction of *XploRe* has been influenced by similar systems like S (Becker, Chambers and Wilks, 1988) and ISP (Interactive Scientific Processor). It differs from both systems by the fully graphically oriented user interface. The hard- and software background of *XploRe* consists of: an IBM Personal Computer AT, XT, PS/2 or a closely compatible machine, runnning DOS Version 2.0 or later.

XploRe is an OBJECT oriented system. An object can be of one of four types:

- VECTOR
- WORKUNIT
- PICTURE
- TEXT

A *VECTOR* object is a data vector as a logical unit with which to work. This vector may contain strings or real numbers and can be of variable length. The simplest form of a *WORKUNIT* object is an ordered collection of data vectors. However, a workunit can also include display attributes and a mask vector. Display attributes concern the layout of scatter plots such as data marking symbols, linestyle, line pattern, line thickness etc. A *TEXT* object is necessary to show information as text on the display. This can be data you are analyzing, documentation you want to pin down or system output. A *PICTURE* object contains the viewport characteristics of certain views that you have on data, e.g. the name of the picture and the axes and the scaling of the axes. For 3D rotation the rotation angles, the initial distance from the point cloud, location of the origin and zooming increments are kept in this picture object.

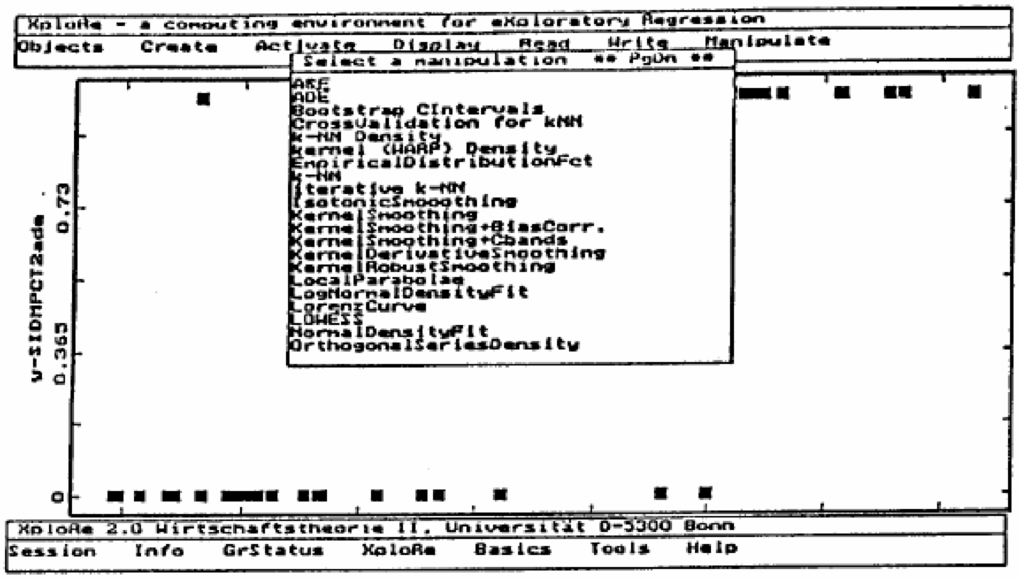


Fig.3.1: The two menu bars of the XploRe main menu after pressing "M". Underlying is a STATIC2DPICTURE scatterplot.

XploRe has a menu structure. Two menu bars will appear on your display (see Fig. 3.1). A third menu bar appears, if you press and hold down the <ALT> key. You can choose an option by typing the capitalized letter of the corresponding menu entry. Whenever a pull down menu appears at the next step, it is in most cases possible to get *QUICK HELP* by pressing **<ALT>+F1**. For example, if you are not familiar with the ACE algorithm, you press this key sequence and a help file explaining the algorithm pops up. The main keys are explained below.

OBJECTS (o,O) After clicking *Objects* you can see an overview of the existing objects with their current names. In addition the type of object (vector, workunit, ...) is indicated.

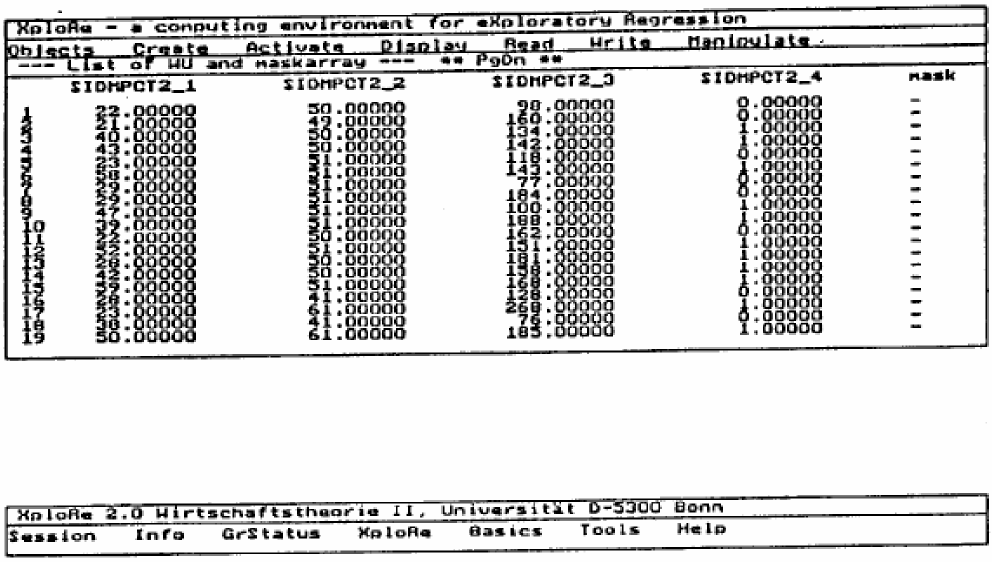
CREATE (c,C) You are asked which object you would like to create. Clicking *WORKUNIT* will show you a window "select a vector number: 1 or ESC" containing all vectors of the active workunit. In this way you can create other workunits from existing vector objects. You can select as many vectors you want. Clicking *VECTOR* will allow you to make a new vector from an existing vector, e.g. by taking LOGs. After clicking *PICTURE*, you are asked for a desired picture type. By choosing the option *STATIC2DGRAPHICS* you can create a picture object for 2 dimensional static graphics, while the option *DYNAMIC3DGRAPHICS* defines an object for dynamic 3D graphics. The menu entry *DRAFTMANSPLOT* will create up to 25 2-dimensional scatterplots by plotting each of up to five selected vectors (which are kept in the same workunit) against each other of them. After clicking *TEXT* a window "create text" appears on the screen. This means that you have invoked the editor of *XploRe* and are able to write ASCII texts. It is now possible to write datasets or comments on data without leaving *XploRe*.

ACTIVATE (a,A) You are asked which object type you would like to activate. By clicking the object of a certain type (in a window "activation") you activate the object. This means that this object becomes the default data set or picture for following operations.

DISPLAY (d,D) This feature allows you to display any existing object of *XploRe*. Again you will be asked by a window to select an object type to display. Suppose you have created some workunits and you want to display one of them. First you will be asked to choose a corresponding picture object by showing a window. The picture object contains the characteristics of the viewport (axes and origin) of the graphical display. There are three different kinds of display styles available. The option *STATIC2DGRAPHICS* will show you a two-dimensional picture display of two selected vectors of your dataset, whereas *DYNAMIC3DGRAPHICS* shows you a three-dimensional picture display of accordingly three vectors of the workunit. If you would like to display a dataset with more than three dimensions, you have the possibility to display two-dimensional scatterplots of up to five vectors against each other of them. In this case you must choose the option *DRAFTMANSPLOT*. If you want to display a text object (the data vectors of a workunit or the corresponding explanations) you first have to read the workunit or the corresponding text file as a *TEXT* object. After this action you can display the vectors or the explanatory help file.

READ (r,R) You are asked what kind of object you want to read. If you want to display data as text it is neccessary to choose the *TEXT* option. In a next step you have to select the subdirectory of your file . You can choose a standard DOS wildcard mask. After clicking a file *XploRe* creates the reads the desired object.

WORKUNIT INFORMATION (i,I) By clicking the menu option *Info* you first have to select a workunit. After it a window (Fig. 3.2) will be shown.



```
XploRe - a computing environment for eXploratory Regression
Objects   Create   Activate   Display   Read   Write   Manipulate
--- List of WU and maskarray ---   ** PgOn **
        SIDMPCT2_1       SIDMPCT2_2       SIDMPCT2_3       SIDMPCT2_4     mask
 1      22.00000         50.00000         98.00000         0.00000        -
 2      21.00000         49.00000        160.00000         0.00000        -
 3      40.00000         50.00000        134.00000         1.00000        -
 4      43.00000         50.00000        142.00000         1.00000        -
 5      23.00000         51.00000        118.00000         0.00000        -
 6      58.00000         51.00000        143.00000         1.00000        -
 7      29.00000         51.00000         77.00000         0.00000        -
 8      29.00000         51.00000        184.00000         0.00000        -
 9      47.00000         51.00000        100.00000         1.00000        -
10      39.00000         51.00000        188.00000         1.00000        -
11      22.00000         50.00000        162.00000         0.00000        -
12      52.00000         51.00000        151.00000         1.00000        -
13      28.00000         50.00000        181.00000         1.00000        -
14      42.00000         50.00000        158.00000         1.00000        -
15      59.00000         51.00000        168.00000         1.00000        -
16      28.00000         41.00000        128.00000         0.00000        -
17      53.00000         61.00000        268.00000         1.00000        -
18      38.00000         41.00000         76.00000         0.00000        -
19      50.00000         61.00000        185.00000         1.00000        -

XploRe 2.0 Wirtschaftstheorie II, Universität D-5300 Bonn
Session   Info   GrStatus   XploRe   Basics   Tools   Help
```

Fig.3.2: Workunit Information of a workunit containing the sideimpact data, see section 4.

MANIPULATE (m,M) With this operation you invoke the manipulation part of *XploRe*, see Figure 3.1. At first it is necessary to activate the corresponding object you would like to use by the *Manipulate* option (see *Activate*). Then you can select an operation under the menu entry *Manipulate*. At the next step you have to select an *XploRe* object which should be the "input" of the selected operation. The calculation time depends on the complexity of the procedure being used. The result of this operation is stored in a new workunit which will be the top most entry in the window if you click the menu option *Objects*.

SESSION INFORMATION (s,S) This menu option shows a window containing all important information on the actual *XploRe* session. These are the active objects, time and the remaining memory (number of available bytes).

GRAPHICSTATUS (g,G) This option allows you to alter the graphics driver of your screen display. You have the possibility to select of 8 drivers. These are

- CGA
- MCGA
- EGA64
- EGAMONO
- HERCMONO
- ATT400
- VGA
- PC3270

BASIC STATISTICS (b,B) By this menu option you can select one of the following four basic statistics. The menu option *Boxplot* shows you a parallel boxplot of all vectors which correspond to the workunit you have to select. In the box on the left side of your display you can see a scale (extends from minimum to maximum of all data) and a legend which contains the marker symbols of median, mean, inner- and outer fence. The option *Stem and Leaf Plot* shows you the corresponding display of all vectors after selecting the desired workunit. The option *Data Summary* gives a summary of a selected workunit by showing minimum, maximium, range, mean, median, variance and upper and lower quartile of all vectors. Finally the last option *Correlation Matrix* shows a matrix containing the correlation between all vectors of the workunit.

TOOLS (t,T) You can *Edit workunit display attributes* and *Edit picture display attributes*. If you have created many new objects during an *XploRe* session and you don't want to write all objects separately, it is useful to click the option *Save all* in order to write all existing objects held in memory at this moment. To read all objects of an *XploRe* session back into memory you have to click the option *Load all*.

HELP (h,H) This option informs you about some important keys to get help or to leave *XploRe*.

CLEARSCREEN (<ALT>+c,C) Clears the screen, but leaves the current objects active.

EXIT (<ALT>+x,X) With this option you can leave *XploRe* and return to DOS. Don't forget to save the *XploRe* objects which you want to keep on disk for later use.

OS SHELL (<ALT>+o,O) If you want to use the DOS shell during an *XploRe* session type one of the above keys. To go back to *XploRe* type EXIT and the main menu appears again.

DELETE (<ALT>+d,D) You are asked which object type you would like to delete. By clicking the object type and the object name a deletion of the object is performed.

ENVIRONMENT (<ALT>+e,E) This option shows the content of the pascal source file `typedef.pas` and contains the declarations of the variables types used by the *XploRe* pascal modules.

INVERTSCREEN (<ALT>+i,I) Inverts the actual screen display.

4. An example of an XploRe session

In this section I present an analysis using *XploRe* with the side impact data set given in Table 3, Appendix 2 of Härdle (1989). These data have been gathered by simulating side impacts with Post Mortal Test Objects (PMTO). The response variable is $Y \in \{0,1\}$ a binary variable denoting fatal injury ($Y = 1$) or non–fatal injury ($Y = 0$). The predictor variables are $X_1 = AGE$, the age of the PMTO, $X_2 = VEL$, the measured speed (in km/h) and $X_3 = T12RM$, the measured accelaration (in g) at the 12th rib. The aim of the analysis is to devise a model for predicting the probability of fatal injury given a certain x.

Figure 4.1a shows a scatterplot of the variables AGE and Y in this side impact data set. One can immediately see that there are a few high AGE variables which seem to fall out of the response pattern for the other observations. How can we investigate this further ? We invoke the cursor and move it to the points we are interested in. In Figure 4.1b this action is shown with the rightmost AGE value. By studying the values of the other variables shown in the window on the left top of the display we see that this observations is from an experiment with a velocity of $45km/h$ and received an input of $T12RM = 103g$.

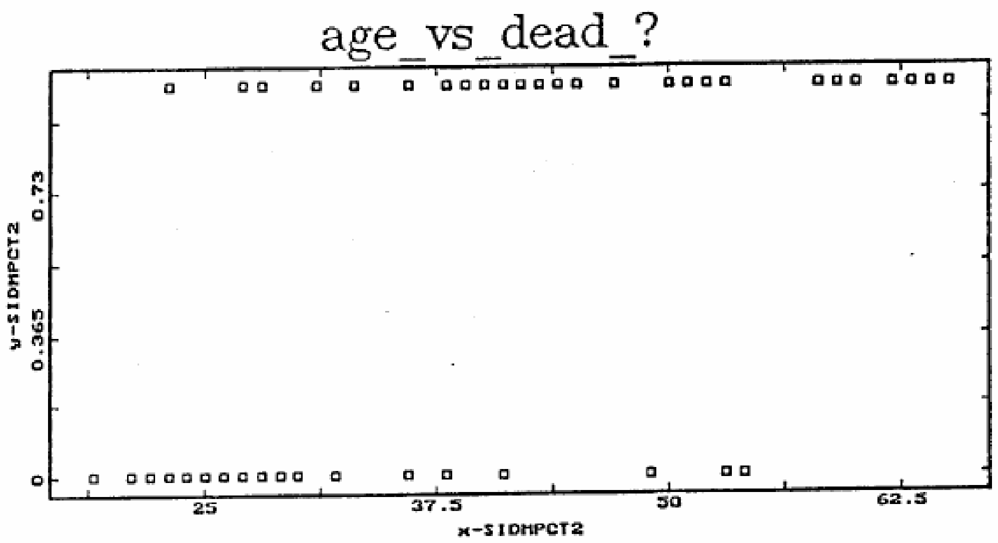
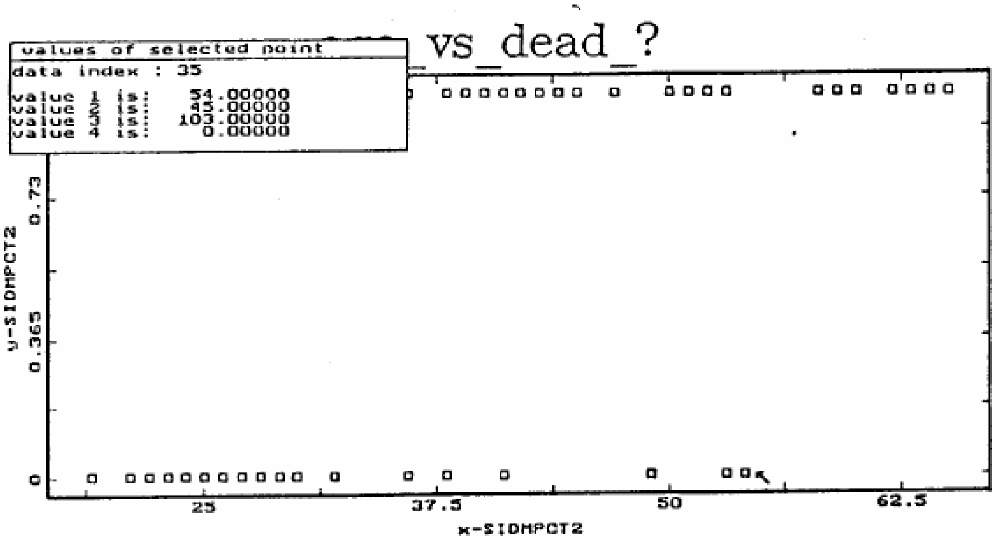


Figure 4.1 a. A scatterplot of the variables *AGE* and *Y*.



```
values of selected point
data index : 35
value 1 is:    54.00000
value 2 is:    45.00000
value 3 is:   103.00000
value 4 is:     0.00000
```

Figure 4.1 b. A scatterplot of the variables *AGE* and *Y* with activated cursor for control of values. The cursor is at $(x,y)=(53,0)$.

We have now a hint this specific data value is an outlier in this marginal scatterplot. But how can we see more ? We display the same data set in a DRAFTMAN'S plot, a system of all pairwise scatterplots, see Figure 4.2. By using a brush (in the $Y vs. X_2$ scatter) and by highlighting the values with $VEL = V0 = 45 km/h$ we can see that those observations correspond to a nearly uniform AGE distribution in the scatter plot AGE vs.VEL. Brushing is a conditioning technique: by conditioning on the experiments with $VEL = 45 km/h$ we see the distribution of the other marginal variables. By moving the brush to the right we can see that the pattern of the highlighted points in the Y vs. AGE plot changes from many "0"s to many "1"s. So we can expect on an intuitive level a positive effect of this influental variable on $m(x) = P(Y = 1|X = x), x \in \mathbb{R}^d$.

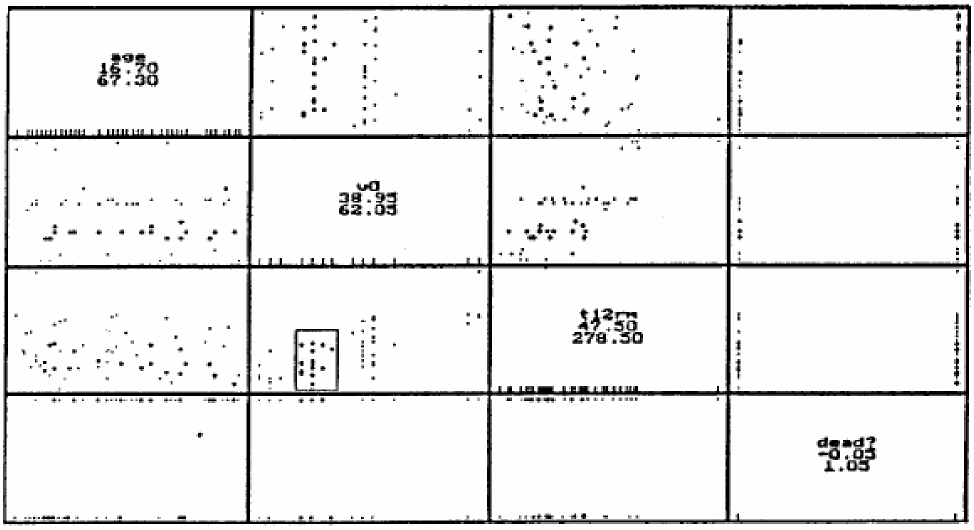


Figure 4.2. A draftman's plot with a brush at the scatterplot $Y vs. X_2$. The brush is at the $45 km/h$ group.

To investigate this point more deeply we apply the ADE method. The ADE method is a technique for finding the average derivative

$$\delta = E_X[m'(X)],$$

here $m'(\bullet) \in \mathbb{R}^d$ denotes the gradient vector of $m(\bullet)$. For Generalized Linear

Models with unknown link function, i.e.,

$$m(x) = g(x^T\beta),$$

the average derivate is proportional to β, $\delta = \gamma\beta$ with a scalar γ. Thus, computing γ gives an estimate of β (up to scale) without knowledge of the link function. For this purpose we press "M", for manipulation and have a menu of a variety of smoothing techniques (see Figure 3.1). After pressing the ENTER key when the cursor is on the "ADE" line we are asked to enter the bandwidth h. Using a bandwidth of $h = 1$ we obtain a value of

$$\delta = (0.423, 0.061, 0.137)$$

for the standardized regressors.

The corresponding projection $\hat{\delta}^T X$ vs. Y is shown in Figure 4.3 together with a kernel smooth and confidence bands. One clearly sees the asymmetry of the response function due to a cluster of five points (high AGE, but low $T12RM$). One may now compare this nonparametric link function with more traditional approaches such as logistic regression analysis using the GLIM module of *XploRe* .

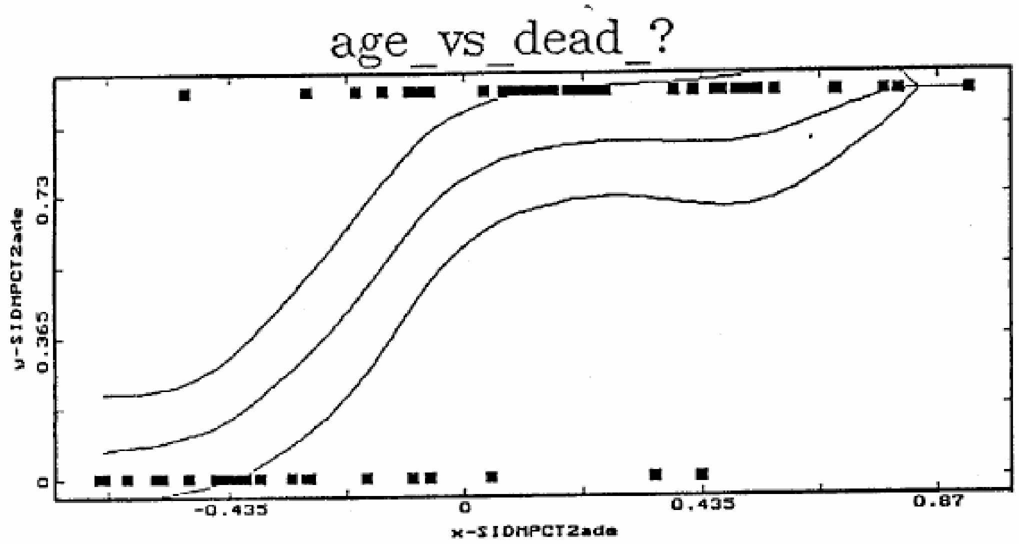


Figure 4.3. The projected data, a kernel smooth plus confidence bands in a STATIC2DGRAPHICS Picture.

Acknowledgements

I would like to thank Ray Carroll and David Scott for numerous fruitful discussions on the subject of statistical computing. They helped sharpening and programming my thoughts on the interplay between computing and statistics.

References

Becker, R.A., Chambers, J.C. and Wilks, A.R. (1988). The New S language: a programming environment for data analysis and graphics. *Wadsworth and Brooks/Cole Advanced Books and Software, Pacific Grove, California.*

Breiman, L. and Friedman, J. (1985). Estimating Optimal Transformations for Multiple Regression and Correlation. *Journal of the American Statistical Association, 80, 580–619.*

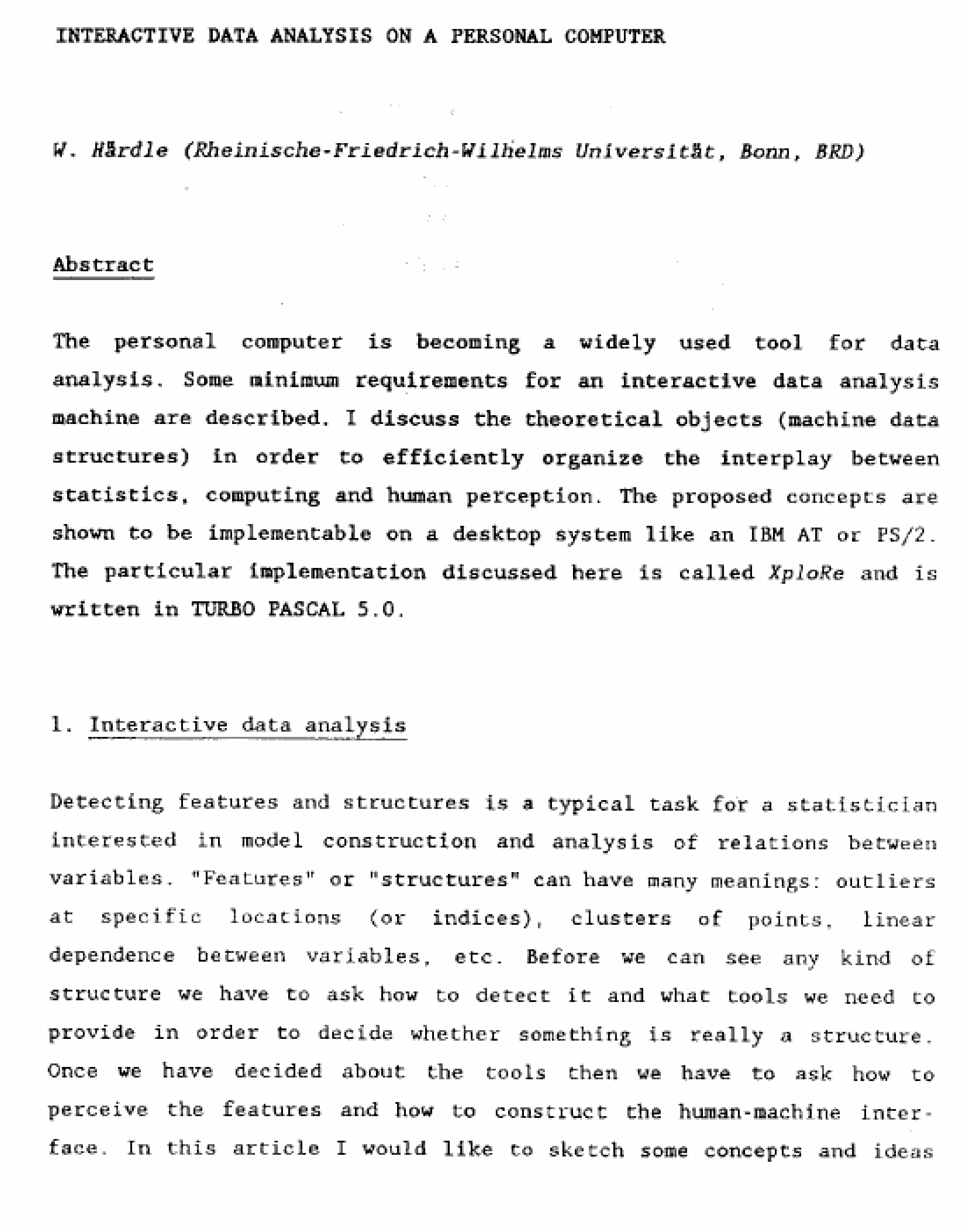
Friedman, J. and Stuetzle, W. (1981). Projection Pursuit Regression. *Journal of the American Statistical Association, 76, 817–823.*

Härdle, W. and Stoker, T. (1989). Investigating Smooth Multiple Regression by the Method of Average Derivatives. *Journal of the American Statistical Association, to appear.*

Härdle, W. (1989). Applied Nonparametric Regression. *Econometric Society Monograph Series, Cambridge University Press, in print.*

McDonald, J.A. and Pederson, J. (1986). Computing Environments for Data Analysis: Part 3, Programming Environments. *Laboratory for Computational Statistics, Stanford, Technical Report, 24.*

Stuetzle, W. (1987). Plot Windows. *Journal of the American Statistical Association, 82, 466–475.*



INTERACTIVE DATA ANALYSIS ON A PERSONAL COMPUTER

W. Härdle (Rheinische-Friedrich-Wilhelms Universität, Bonn, BRD)

Abstract

The personal computer is becoming a widely used tool for data analysis. Some minimum requirements for an interactive data analysis machine are described. I discuss the theoretical objects (machine data structures) in order to efficiently organize the interplay between statistics, computing and human perception. The proposed concepts are shown to be implementable on a desktop system like an IBM AT or PS/2. The particular implementation discussed here is called *XploRe* and is written in TURBO PASCAL 5.0.

1. Interactive data analysis

Detecting features and structures is a typical task for a statistician interested in model construction and analysis of relations between variables. "Features" or "structures" can have many meanings: outliers at specific locations (or indices), clusters of points, linear dependence between variables, etc. Before we can see any kind of structure we have to ask how to detect it and what tools we need to provide in order to decide whether something is really a structure. Once we have decided about the tools then we have to ask how to perceive the features and how to construct the human-machine interface. In this article I would like to sketch some concepts and ideas of how this task can be done in an interactive data analysis computing environment.

Very simple structural elements of a data set may be detected just by looking at the list of values, e.g. a column with small numbers with the exception of one large one will indicate an outlier. On a finer level summary statistics for the variables and the use of graphical techniques like Boxplots or residual plots help us to see data structures. A Boxplot for example is a simple graphical technique to diagnose skewness. Suppose one has detected by the techniques that variable 1 has a right skew distribution. A very natural next question is whether the points causing this skewness in variable 1 create other structural features in variables 2, 3, 4 ... as well. Can we efficiently perform this task with one of the "classical packages"?

Most statistical packages now in use, such as SAS or MINITAB, were designed to operate in a batch-like mode. The user enters commands from the keyboard and the system scrolls subsequent stages of the analysis and displays on explicit request values or plots. Therefore the data analytic task, "investigate the behavior of other variables given the detection of skewness in a particular variable" cannot be done instantaneously. Instead we have to write a separate program. This program has to first identify the indices of the interesting subset and then make several statistics and plots from other variables on the basis of this subset.

The development of hardware in the last years has made it possible to carry out this task and similar ones within a desktop system. It should be emphasized though that the hardware, in principle, gives the "possibility" to perform such operations efficiently at a high "power/price" ratio but in many cases the computing environment (the software) is not capable of doing what a data analyst wants to do. We will see below a first approximation of a statistical desktop system based on a widely used IBM desktop system (IBM AT, IBM PS/2).

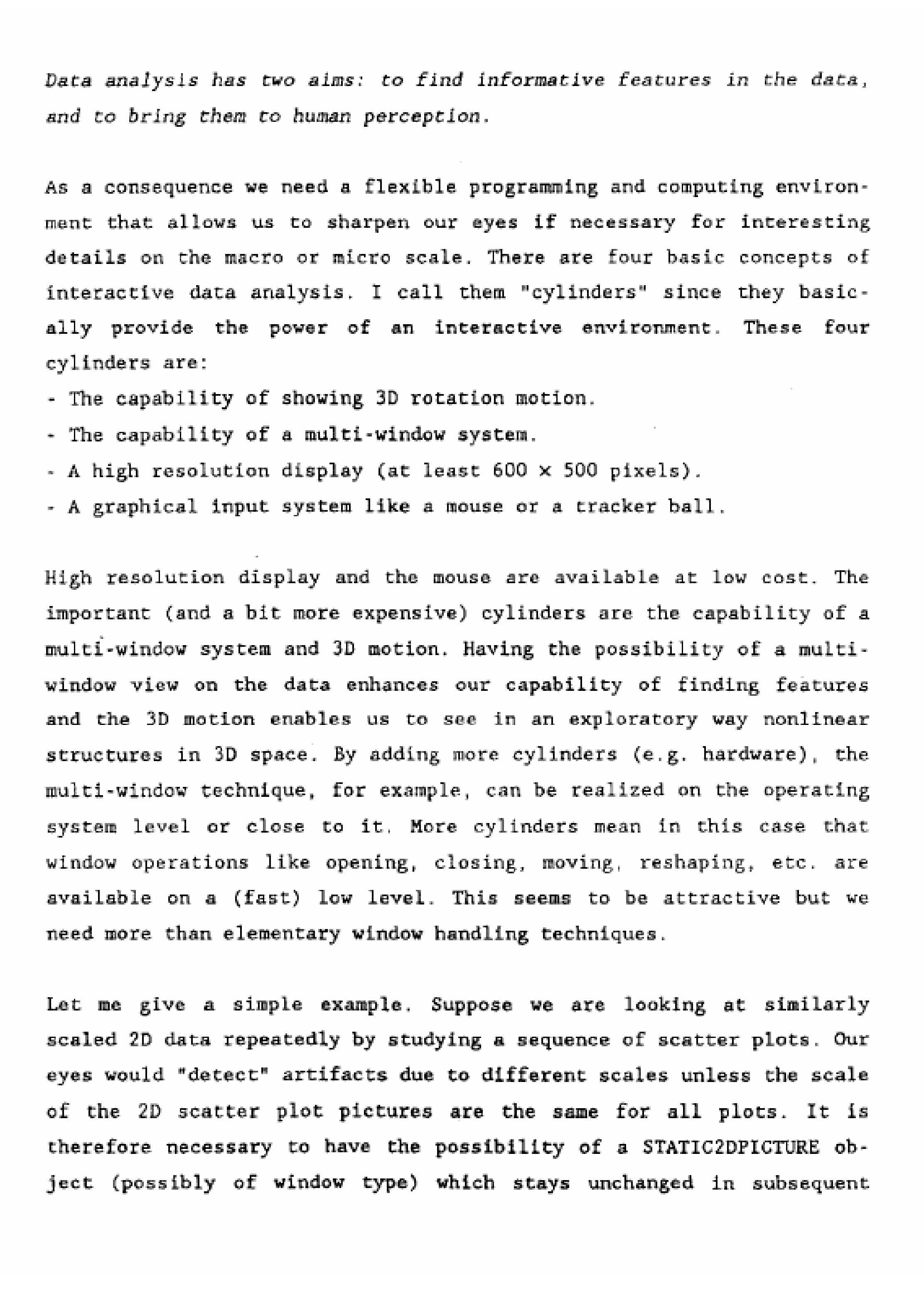
The desktop concept allows us to analyze various aspects of the data from different viewpoints (windows) at the same time. Stuetzle (1987) has made this evident, stating:

Each window and in particular each plot, corresponds to a sheet of paper on the desktop.

Before computers entered our work we shuffled paper on the desk to compare and evaluate different structural elements of the data. Now the paper shuffling is replaced by window clicking and logical links between the sheets are not only in the statistician's mind but also between defined computer objects inside the machine. These computer objects and their implementation are described below. It is argued in Section 2 that instead of expensive hardware it is a matter of software to bring interactive data analysis into life, especially data analysis in high dimensions. In Sections 3 I describe a current approximation to a data analysis computing environment. Section 4 is devoted to an example session with the system *XploRe*.

2. Four concepts of interactive data analysis

In analysing data, we are programming, in effect, our view towards the data and the structures we want to and can see. By selecting certain scales we can concentrate our eyes dynamically on local or global features, see McDonald and Pederson (1986). For example, with a single column of numbers, by using LOG-transformations we can sharpen our eye for peaks or other local structures. Günter Sawitzki (1989) has described this aptly as follows.



Data analysis has two aims: to find informative features in the data, and to bring them to human perception.

As a consequence we need a flexible programming and computing environment that allows us to sharpen our eyes if necessary for interesting details on the macro or micro scale. There are four basic concepts of interactive data analysis. I call them "cylinders" since they basically provide the power of an interactive environment. These four cylinders are:

- The capability of showing 3D rotation motion.
- The capability of a multi-window system.
- A high resolution display (at least 600 x 500 pixels).
- A graphical input system like a mouse or a tracker ball.

High resolution display and the mouse are available at low cost. The important (and a bit more expensive) cylinders are the capability of a multi-window system and 3D motion. Having the possibility of a multi-window view on the data enhances our capability of finding features and the 3D motion enables us to see in an exploratory way nonlinear structures in 3D space. By adding more cylinders (e.g. hardware), the multi-window technique, for example, can be realized on the operating system level or close to it. More cylinders mean in this case that window operations like opening, closing, moving, reshaping, etc. are available on a (fast) low level. This seems to be attractive but we need more than elementary window handling techniques.

Let me give a simple example. Suppose we are looking at similarly scaled 2D data repeatedly by studying a sequence of scatter plots. Our eyes would "detect" artifacts due to different scales unless the scale of the 2D scatter plot pictures are the same for all plots. It is therefore necessary to have the possibility of a STATIC2DPICTURE object (possibly of window type) which stays unchanged in subsequent calls. Practically speaking this means that a picture shown in a window system "has to know what scales it had in earlier calls". In a more modern language, we would say that the picture should have the capability or inheriting properties of earlier stages of an analysis.

This capability of inheritance in an Object Oriented Programming System (OOPS) creates difficulties in software construction. If we try to realize this concept of a statistical STATIC2DPICTURE object in a eight cylinder vehicle (e.g. the machine with a built-in window system) we face extra difficulties since we have to artificially attach statistical meanings to windows constructed for a different purpose. By forcing the construction of a "statistical data window" in a window system not designed for data analysis we can slow down to ordinary programming speed. With other words: more cylinders can be more expensive and contra-productive. For this reason some people have called such vehicles "Corvette data analysis machines".

The Corvettes are very powerful, not very frequent among the data analysis vehicles, and sometimes have to drive the speed of a Chevette.

A Chevette is a data analysis machine that operates from a lower powered level but has the (dis)advantage of free object programming. Free object programming means that we can define our programming objects (e.g. statistical windows) by ourselves in one horizontal level without time consuming "vertical diffusions" to other levels. Of course this is an advantage from a pure theoretical point of view but a disadvantage on the programmer's side. The other advantages, such as price and portability, make it clear that the "Chevette data analysis machine" is the preferable choice. Note that portability has two meanings here. First we can export the code more easily since it is written in one level although possibly simulating deeper levels. Second we can implement the code on portable machines (laptops) to show data analysis live to other statisticians! Summarizing these thoughts leads me to the conclusion that a four cylinder personal computer like an IBM AT or PS/2 is acceptable for interactive data analysis if we tune it correctly e.g. program it correctly. In the following I describe a four cylinder Chevette called *XploRe*.

3. XploRing data

XploRe is an interactive, graphically oriented computing environment designed to analyse various kinds of relations between data and to apply and compare different smoothing methods. *XploRe* is suitable for investigating high dimensional data. It supports the user with the above four cylinders and other sophisticated data management tools such as masking, brushing, labeling and rotation of data. In addition, a wide variety of additive models are available, among them:

- ACE (Alternating Conditional Expectations), Breiman and Friedman (1985),
- ADE (Average Derivative Estimation), Härdle and Stoker (1989),
- PPR (Projection Pursuit Regression), Friedman and Stuetzle (1981).

XploRe is designed as an open system. It is basically a framework awaiting more software in the form of user written submodules. The user can write his or her own programs and add them into *XploRe* via an object oriented user interface. User supplied help files can be attached so that the data analyst in a strict sense is able to design his own computing environment.

The construction of *XploRe* has been influenced by similar systems like S (Becker, Chambers and Wilks, 1988) and ISP (Interactive Scientific Processor). It differs from both systems by the fully graphically oriented user interface. The hard- and software background of *XploRe*

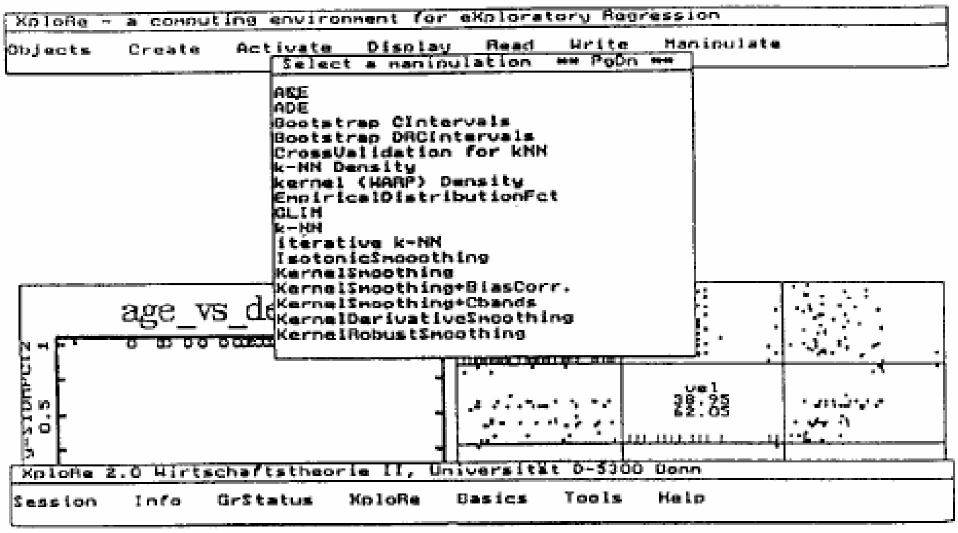
consists of: an IBM Personal Computer AT, XT, PS/2 or a closely compatible machine, running DOS version 2.0 or later.
XPloRe is an OBJECT oriented system. An object can be of one of four types:

- VECTOR
- WORK UNIT
- PICTURE
- TEXT

A *VECTOR* object is a data vector as a logical unit with which to work. This vector may contain strings or real numbers and can be of variable length. The simplest from of a *WORK UNIT* object is an ordered collection of data vectors. However, a work unit can also include display attributes and a mask vector. Display attributes concern the layout of scatter plots such as data marking symbols, linestyle, line pattern, line thickness, etc. A *TEXT* object is necessary to show information as text on the display. This can be data you are analyzing, documentation you want to pin down, or system output. A *PICTURE* object contains the viewpoint characteristics of certain views that you have on data, e.g. the name of the picture and the axes and the scaling of the axes. For 3D rotation the rotation angles, the initial distance from the point cloud, location of the origin and zooming increments are kept in this picture object.

XploRe has a menu structure. Two menu bars will appear on your display (see figure 3.1). A third menu bar appears, if you press and hold down the **<ALT>** key. You can choose an option by typing the capitalized letter of the corresponding menu entry. Whenever a pull down menu appears at the next step, it is in most cases possible to get *QUICK HELP* by pressing **<ALT>**+F1. For example, if you are not familiar with the ACE algorithm, you press this key sequence and a help file explaining the algorithm pops up. The main keys are explained below.

Figure 3.1. The two menu bars of the XploRe main menu after pressing "M". Underlying is a STATIC2DPICTURE scatter plot of AGE vs. Y in the lower left corner and a draftman's plot of AGE, VEL, T12RM in the lower right corner



OBJECTS (o,O) After clicking *Objects* you can see an overview of the existing objects with their current names. In addition the type of object (vector, work unit, ...) is indicated.

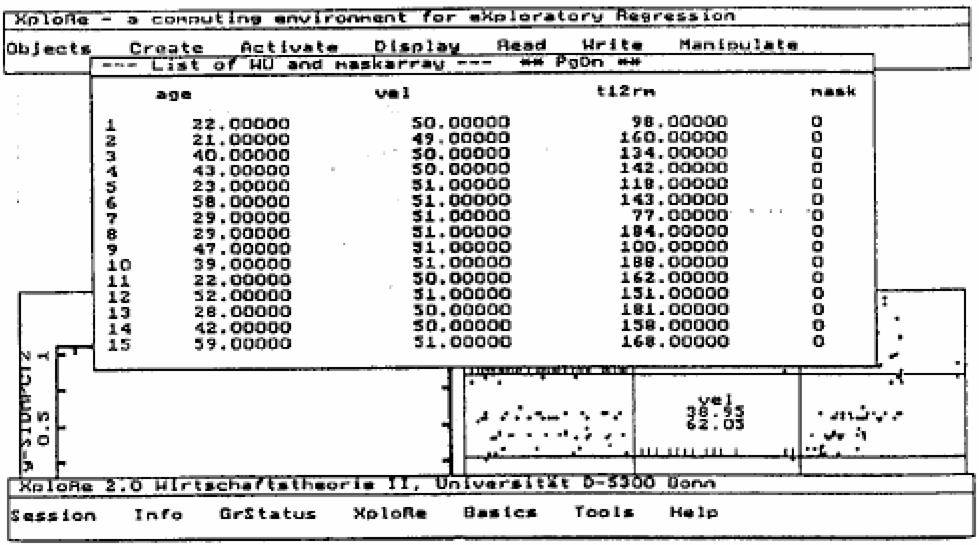
CREATE (c,C) You are asked which object you would like to create. Clicking *WORK UNIT* will show you a window "select a vector number: 1 or ESC" containing all vectors of the active work unit. In this way you can create other work units from existing vector objects. You can select as many vectors you want. Clicking *VECTOR* will allow you to make a new vector from an existing vector, e.g. by taking LOGs. After clicking *PICTURE*, you are asked for a desired picture type. By choosing the option *STATIC2DGRAPHICS* you can create a picture object for 2 dimensional static graphics, while the option *DYNAMIC3DGRAPHICS* defines an object for dynamic 3D graphics. The menu entry *DRAFTMANSPLOT* will create up to 25 2-dimensional scatter plots by plotting each of up to five selected vectors (which are kept in the same work unit) against each other of them. After clicking *TEXT* a window "create text" appears on the screen. This means that you have invoked the editor of *XploRe* and are able to write ASCII texts. It is now possible to write data sets or comments on data without leaving *XploRe*.

ACTIVATE (a,A) You are asked which object type you would like to activate. By clicking the object of a certain type (in a window "activation") you activate the object. This means that this object becomes the default data set or picture for following operations.

DISPLAY (d,D) This feature allows you to display any existing object of *XploRe*. Again you will be asked by a window to select an object type to display. Suppose you have created some work units and you want to display one of them. First you will be asked to choose a corresponding picture object by showing a window. The picture object contains the characteristics of the viewport (axes and origin) of the graphical display. There are three different kinds of display styles available. The option *STATIC2DGRAPHICS* will show you a two-dimensional picture display of two selected vectors of your data set, whereas *DYNAMIC3DGRAPHICS* shows you a three-dimensional picture display of accordingly three vectors of the work unit. If you would like to display a data set with more than three dimensions, you have the possibility to display two-dimensional scatter plots of up to five vectors against each other. In this case you must choose the option *DRAFTMANSPLOT*. If you want to display a text object (the data vectors of a work unit of the corresponding explanations) you first have to read the work unit or the corresponding text file as a *TEXT* object. After this action you can display the vectors or the explanatory help file.

READ (r,R) You are asked what kind of object you want to read. If you want to display data as text it is necessary to choose the *TEXT* option. In a next step you have to select the subdirectory of your file. You can choose a standard DOS wildcard mask. After clicking a file *XploRe* creates and reads the desired object.

Figure 3.2. Work unit Information of a work unit containing the side impact data, see section 4


XploRe - a computing environment for eXploratory Regression
Objects
Create
Activate
Display
Read
Write
Manipulate
vel
38.95
62.03
XploRe 2.0 Wirtschaftstheorie II, Universität D-5300 Bonn
Session
Info
GrStatus
XploRe
Basics
Tools
Help

```
--- List of WU and maskarray ---    ** PgDn **

          age              vel              t12rm           mask

1        22.00000         50.00000          98.00000        0
2        21.00000         49.00000         160.00000        0
3        40.00000         50.00000         134.00000        0
4        43.00000         50.00000         142.00000        0
5        23.00000         51.00000         118.00000        0
6        58.00000         51.00000         143.00000        0
7        29.00000         51.00000          77.00000        0
8        29.00000         51.00000         184.00000        0
9        47.00000         51.00000         100.00000        0
10       39.00000         51.00000         188.00000        0
11       22.00000         50.00000         162.00000        0
12       52.00000         51.00000         151.00000        0
13       28.00000         50.00000         181.00000        0
14       42.00000         50.00000         158.00000        0
15       59.00000         51.00000         168.00000        0
```

WORK UNIT INFORMATION (I,i) By clicking the menu option *Info* you first have to select a work unit. After that a window (figure 3.2) will be shown.

MANIPULATE (m,M) With this operation you invoke the manipulation part of *XploRe*, see figure 3.1. At first it is necessary to activate the corresponding object you would like to use by the *Manipulate* option (see *Activate*). Then you can select an operation under the menu entry *Manipulate*. At the next step you have to select an *XploRe* object which should be the "input" of the selected operation. The calculation time depends on the complexity of the procedure being used. The result of this operation is stored in a new work unit which will be the topmost entry in the window if you click the menu option *Objects*.

SESSION INFORMATION (s,S) This menu option shows a window containing all important information on the actual *XploRe* session. These are the active objects, time and the remaining memory (number of available bytes).

GRAPHICSTATUS (g,G) This option allows you to alter the graphics driver of your screen display. You have the possibility to select 8 drivers. These are:
- CGA
- MCGA
- EGA64
- EGAMONO
- HERCMONO
- ATT400
- VGA
- PC3270

BASIC STATISTICS (b,B) By this menu option you can select one of the following four basic statistics. The menu option *Boxplot* shows you a parallel boxplot of all vectors which correspond to the work unit you have to select. In the box on the left side of your display you can see a scale (extends from minimum to maximum of all data) and a legend which contains the marker symbols of median, mean, inner- and outer fence. The option *Stem and Leaf Plot* shows you the corresponding display of all vectors after selecting the desired work unit. The option *Data Summary* gives a summary of a selected work unit by showing minimum, maximum, range, mean, median, variance and upper and lower quartile of all vectors. Finally the last option *Correlation Matrix* shows a matrix containing the correlation between all vectors of the work-unit.

TOOLS (t,T) You can *Edit work unit display attributes* and *Edit picture display attributes*. If you have created many new objects during an *XploRe* session and you don't want to write all objects separately, it is useful to click the option *Save all* in order to write all existing objects held in memory at this moment. To read all objects of an *XploRe* session back into memory you have to click the option *Load all*.

HELP (h,H) This option informs you about some important keys to get help or to leave *XploRe*.

CLEARSCREEN (<ALT>+c,C) Clears the screen, but leaves the current objects active.

EDIT (<ALT>+x,X) With this option you can leave *XploRe* and return to DOS. Don't forget to save the *XploRe* objects which you want to keep on disk for later use.

OS SHELL (<ALT>+o,O) If you want to use the DOS shell during an *XploRe* session type one of the above keys. To go back to *XploRe* type *EXIT* and the main menu appears again.

DELETE (<ALT>+d,D) You are asked which object you would like to delete. By clicking the object type and the object name a deletion of the object is performed.

ENVIRONMENT (<ALT>+e,E) This option shows the content of the pascal source file *typedef.pas* and contains the declarations of the variables types used by the *XploRe* pascal modules.

INVERTSCREEN (<ALT>+i,I) Inverts the actual screend display.

4. An example of an XploRe session

In this section I present an analysis using *XploRe* with the side impact data set given in Table 3, Appendix 2 of Härdle (1989). These data have been gathered by simulating side impacts with Post Mortal Test Objects (PMTO). The response variable is $Y \in \{0,1\}$ a binary variable denoting fatal injury ($Y = 1$) or non-fatal injury ($Y = 0$). The predictor variables are $X_1 = AGE$, the age of the PMTO, $X_2 = VEL$, the measured speed (in km/h) and $X_3 = T12RM$, the measured acceleration (in g) at the 12th rib. The aim of the analysis is to devise a model for predicting the probability of fatal injury given a certain x.

Figure 4.1a shows a scatter plot of the variables *AGE* and Y in this side impact data set. One can immediately see that there are a few high *AGE* variables which seem to fall out of the response pattern for the other observations. How can we investigate this further? We invoke the cursor and move it to the points we are interested in. In Figure

Figure 4.1a. A scatter plot of the variables *AGE* and *Y*. The display style of the work unit has been set to circles

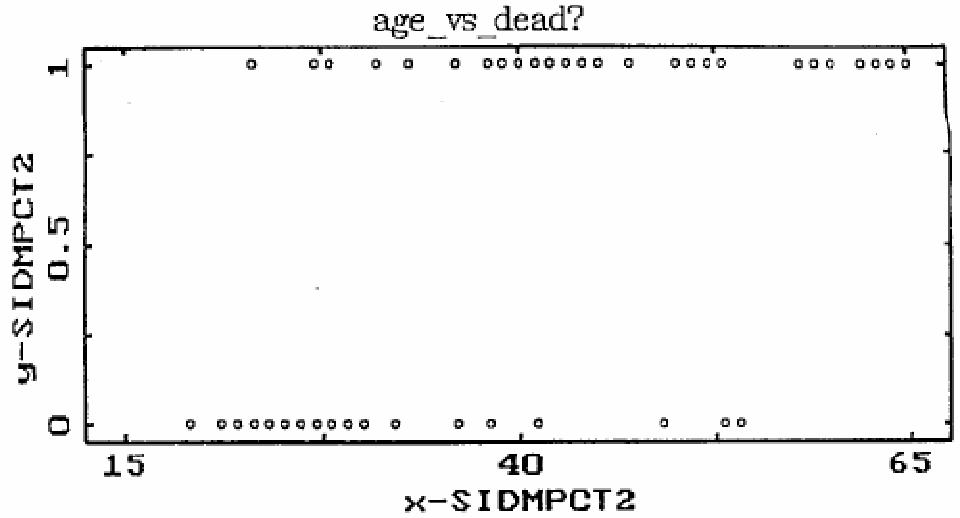
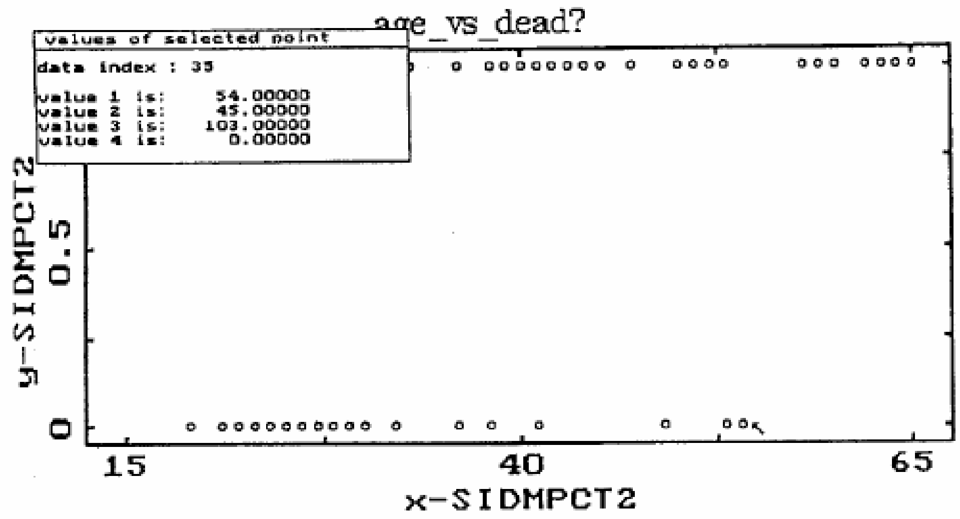


Figure 4.1b. A scatter plot of the variables *AGE* and *Y* with activated cursor for control of values. The cursor is at (*x*,*y*)=(54,0)



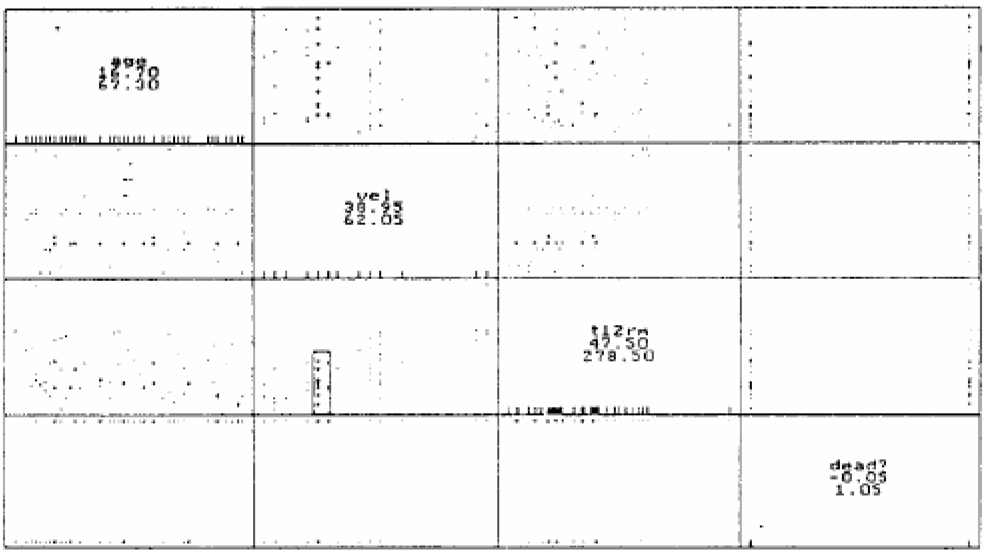
```
values of selected point
data index : 35

value 1 is:      54.00000
value 2 is:      45.00000
value 3 is:     103.00000
value 4 is:       0.00000
```

4.1b this action is shown with the right most *AGE* value. By studying the values of the other variables shown in the window on the left top of the display we see that this observation is from an experiment with a velocity of 45*km/h* and received an input of *T12RM* = 103*g*.

We have now a hint that this specific data value is an outlier in this marginal scatter plot. But how can we see more? We display the same data set in a DRAFTMAN'S plot, a system of all pairwise scatter plots, see Figure 4.2a. By using a brush (in the $Yvs.X_2$ scatter) and by highlighting the values with *VEL* = *VO* = 45*km/h* we can see that those observations correspond to a nearly uniform AGE distribution in the scatter plot *AGE* vs. *VEL*. Brushing is a conditioning technique: by conditioning on the experiments with *VEL* = 45*km/h* we see the distribution of the other marginal variables. By moving the brush to the

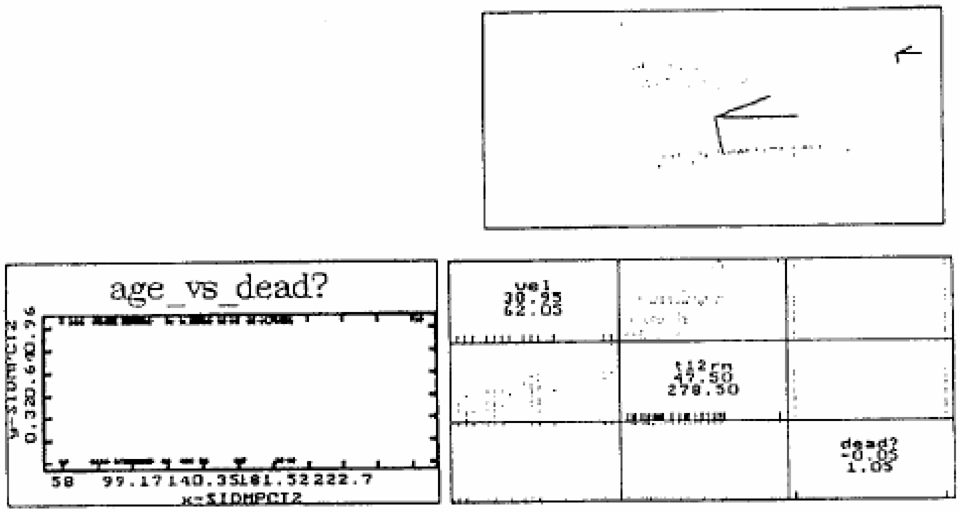
Figure 4.2a. A draftman's plot with a brush at the scatter plot $Yvs.X_2$. The brush is at the 45*km/h* group



right we can see that the pattern of the highlighted points in the Y vs. *AGE* plot changes from many "0"s to many "1"s. So we can expect on an intuitive level a positive effect of this influental variable on $m(x) = P(Y = 1|X = x), x \in \mathbb{R}^d$.

Using the multi window technique, an important cylinder in our data analysis chevette, we can enhance our view even more. Figure 4.2b shows a combination of three display techniques. In the upper right corner we see a DYNAMIC3DPICTURE with the variables *AGE*, *VEL* and Y. Below a draftman's plot of the three predictor variables. Finally in the lower left the STATIC2DPICTURE that we saw already in Figure 3.1. By using the brush now in one of the plots we can link all these pictures and highlight the respective points in each of the different windows. In an explanatory way we can see a positive effect of increasing the predictor variables.

Figure 4.2b. A multi window view on the side impact data set



To investigate this point more deeply we apply the ADE method. The ADE method is a technique for finding the average derivative

$$\delta = E_X[m'(X)],$$

here $m'(o) \in \mathbb{R}^d$ denotes the gradient vector of $m(o)$. For Generalized Linear Models with unknown link function, i.e.,

$$m(x) = g(x^T\beta),$$

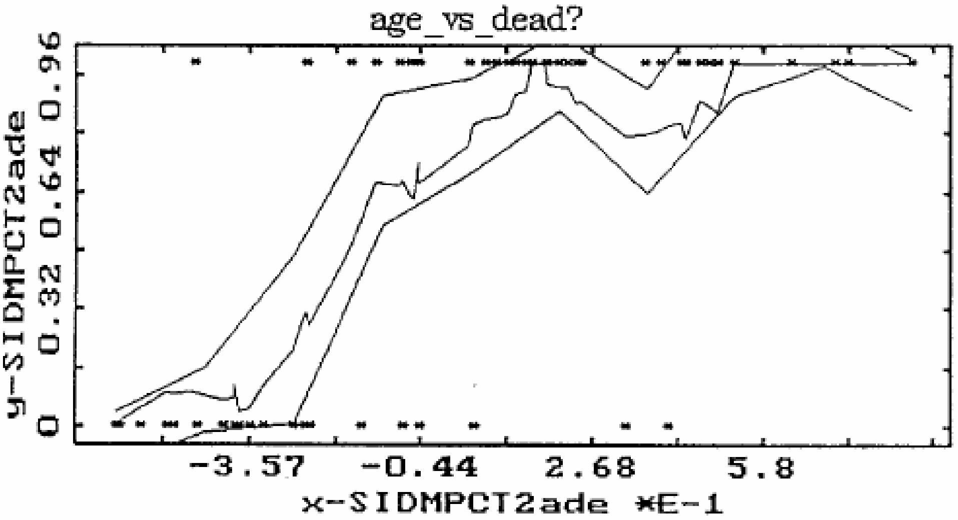
the average derivate is proportional to β, $\delta = \gamma\beta$ with a scalar γ. Thus computing γ gives an estimate of β (up to scale) without knowledge of the link function. For this purpose we press "M", for manipulation and have a menu of a variety of smoothing techniques (see Figure 3.1). After pressing the ENTER key when the cursor is on the "ADE" line we are asked to enter the bandwidth h. Using a bandwidth of $h = 1$ we obtain a value of

$$\delta = (0.423, 0.061, 0.137)$$

for the standardized regressors.

The corresponding projection $\hat{\delta}^T X$ vs. Y is shown in Figure 4.3 together with a kernel smooth and confidence bands. One clearly sees the asymmetry of the response function due to a cluster of five points (high *AGE*, but low *T12RM*). One may now compare this nonparametric link function with more traditional approaches such as logistic regression analysis using the GLIM module of *XploRe*.

Figure 4.3. The projected data, a *k-N N* smooth bootstrap confidence bands in a STATIC2DGRAPHICS picture



The confidence bands are computed using the Wild Bootstrap technique of Härdle and Marron (1989).

I would like to thank Ray Carroll, Richard Gill and David Scott for numerous fruitful discussions on the subject of statistical computing. They helped sharpening and programming my thoughts on the interplay between computing and statistics.

References

Becker, R.A., J.C. Chambers and A.R. Wilks, 1988, The New S language: a programming environment for data analysis and graphics. Wadsworth and Brooks/Cole Advanced Books and Software, Pacific Grove, California.

Breiman, L. and J. Friedman, 1985, Estimating Optimal Transformations for Multiple Regression and Correlation. Journal of the American Statistical Association, 80, pp. 580-619.

Friedman, J. and W. Stuetzle, 1981, Projection Prusuit Regression. Journal of the American Statistical Association, 76, pp. 817-823.

Härdle, W. and J.S. Marron, 1989, Bootstrap Error Bars for Nonparametric Regression. Annals of Statistics, submitted.

Härdle, W. and T. Stoker, 1989, Investigating Smooth Multiple Regression by the Method of Average Derivatives. Journal of the American Statistical Association, to appear.

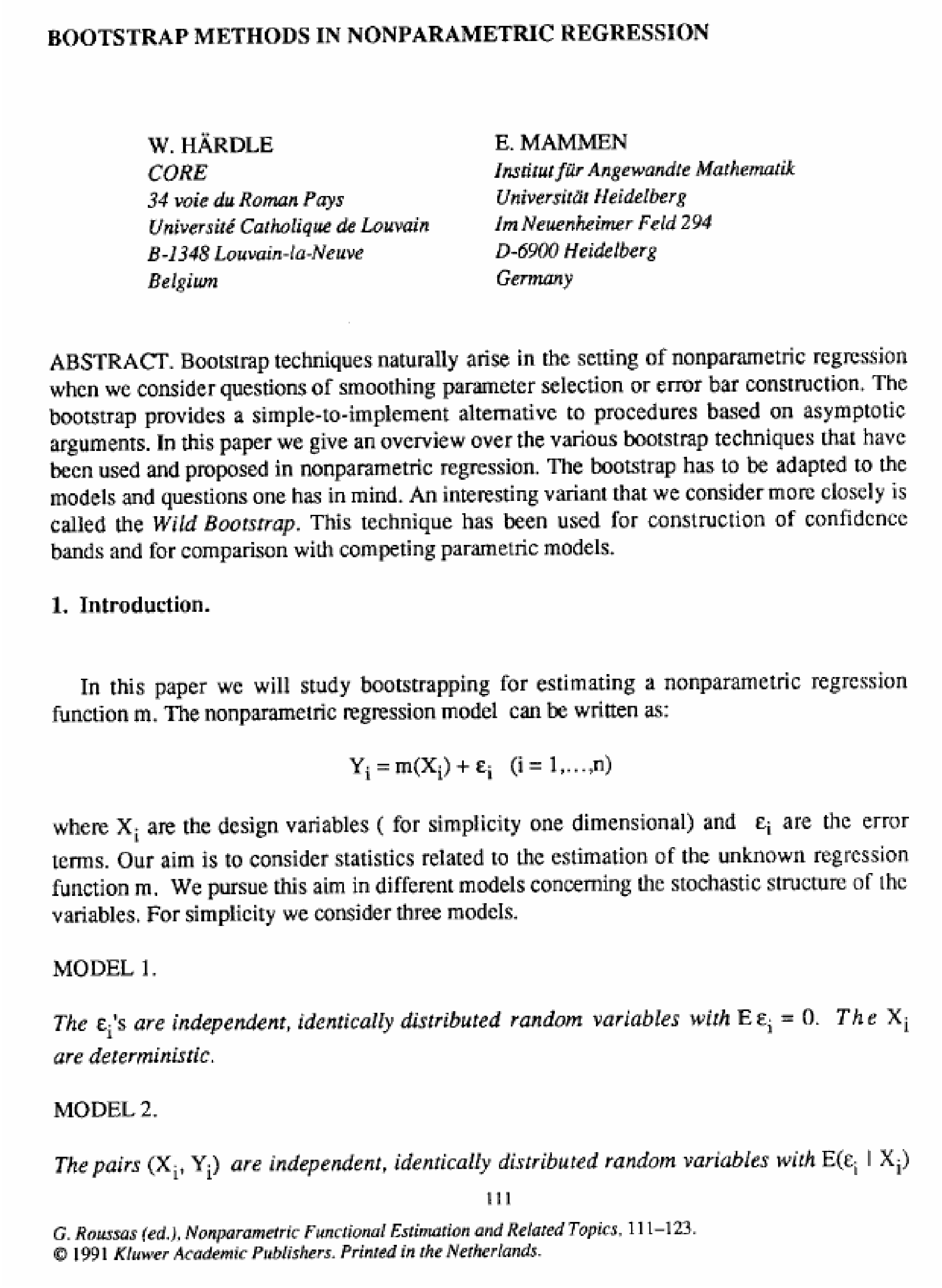
Härdle, W., 1989, Applied Nonparametric Regression. Econometric Society Monograph Series, Cambridge University Press, in print.

McDonald, J.A. and J. Pederson, 1986, Computing Environments for Data Analysis: Part 3, Programming Environments. Laboratory for Computational Statistics, Stanford, Technical Report, 24.

Sawitzki, G., 1989, Tools and Concepts in Data Analysis. Softstat 1989, F. Faulbaum et al. eds.

Stuetzle, W., 1987, Plot Windows. Journal of the American Statistical Association, 82, pp. 466-475.

Nonparametric Functional estimation and Related Topics, Edited by G.Roussas, Kluwer Publishing company, Series C: Mathematical and Physical Sciences, 335, 111-124



BOOTSTRAP METHODS IN NONPARAMETRIC REGRESSION

W. HÄRDLE
CORE
34 voie du Roman Pays
Université Catholique de Louvain
B-1348 Louvain-la-Neuve
Belgium

E. MAMMEN
Institut für Angewandte Mathematik
Universität Heidelberg
Im Neuenheimer Feld 294
D-6900 Heidelberg
Germany

ABSTRACT. Bootstrap techniques naturally arise in the setting of nonparametric regression when we consider questions of smoothing parameter selection or error bar construction. The bootstrap provides a simple-to-implement alternative to procedures based on asymptotic arguments. In this paper we give an overview over the various bootstrap techniques that have been used and proposed in nonparametric regression. The bootstrap has to be adapted to the models and questions one has in mind. An interesting variant that we consider more closely is called the *Wild Bootstrap*. This technique has been used for construction of confidence bands and for comparison with competing parametric models.

1. Introduction.

In this paper we will study bootstrapping for estimating a nonparametric regression function m. The nonparametric regression model can be written as:

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

where X_i are the design variables (for simplicity one dimensional) and ε_i are the error terms. Our aim is to consider statistics related to the estimation of the unknown regression function m. We pursue this aim in different models concerning the stochastic structure of the variables. For simplicity we consider three models.

MODEL 1.

The ε_i*'s are independent, identically distributed random variables with* $E\,\varepsilon_i = 0$. *The* X_i *are deterministic.*

MODEL 2.

The pairs (X_i, Y_i) *are independent, identically distributed random variables with* $E(\varepsilon_i \mid X_i)$

G. Roussas (ed.), Nonparametric Functional Estimation and Related Topics, 111–123.

= 0. *Then* $m(x) = E(Y_i \mid X_i = x)$.

MODEL 3.

The ε_i's *are independent random variables with* $E\ \varepsilon_i = 0$. *The* X_i *are deterministic. The distribution of the errors may depend on the design variable.*

The model that has been most dominantly investigated is MODEL 1, the socalled 'fixed design' model, see Eubank (1988) and Härdle (1990). Note that MODEL 3 covers the class of models of type 1. Also if we condition on the design variables then MODEL 2 is contained in MODEL 3. Of course there is a wide range of possibilities between these above model classes. For example, with the variance function $\sigma^2(x)$ =var($\varepsilon \mid X = x$), one can assume that in MODEL 3 the additional assumption holds that $\varepsilon_i / \sigma(X_i)$ are i.i.d. . One such approach could be to parametrize the variance function as in Carroll (1982) and estimate the parameters before entering a bootstrap step.

Each of these models suggests a different resampling procedure. Furthermore one may use a resampling procedure motivated by a larger model (for instance MODEL 3) in a smaller model (for instance MODEL 1). This makes sense if one wants to safeguard oneself in MODEL 1 against deviations from MODEL 1. But clearly if one is more interested in efficiency than model robustness one should prefer resampling methods motivated by the assumed model. In the context of resampling procedures for linear models this point has also been made in Liu and Singh (1989). To our knowledge the first bootstrap study for making inference about m can be found in a film by McDonald(1982). He assumed MODEL 2 and resampled from the pairs of observations (X_i, Y_i). A recent bootstrap overview has been given in Mammen (1990c).

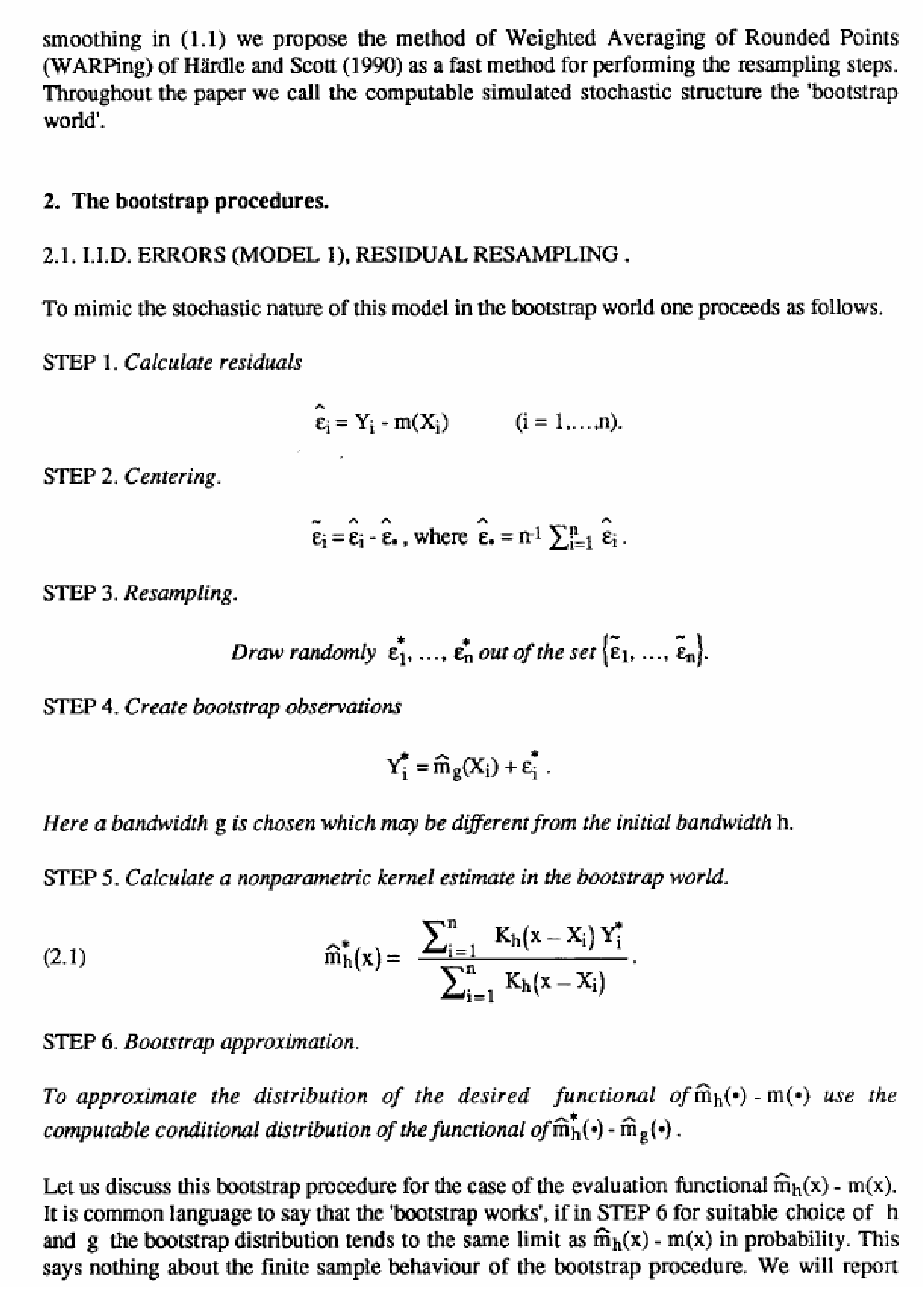
Throughout our paper we use the kernel estimator $\hat{m}_h$ with bandwidth $h = h_n$ and kernel K (Nadaraya, 1964; Watson, 1964)

$$(1.1) \qquad \hat{m}_h(x) = \frac{\sum_{i=1}^{n} K_h(x - X_i)\, Y_i}{\sum_{i=1}^{n} K_h(x - X_i)} ,$$

$$K_h(\bullet) = h^{-1} K(\bullet / h) .$$

For simplification of notation the dependence of h on n will be suppressed. Generalizations of the results presented here to higher dimensional design variables are straightforward. In the fixed design model with equidistant $X_i = i/n$ the denominator in (1.1) is often conveniently replaced by n, see Müller (1988).

We are mainly interested in the distribution of functionals of $\hat{m}_h(\bullet) - m(\bullet)$, for instance the L_2 norm of this function or the evaluations of this function at a set of points. But we have also other functionals such as shape parameters in mind (see Mammen, 1990a). We discuss the bootstrap procedures which have been proposed in the literature for MODEL 1 - 3 in the next section. We do not address in this paper the problem of computational feasibility of bootstrap in this context. To avoid the computer intensive direct computation of the



smoothing in (1.1) we propose the method of Weighted Averaging of Rounded Points (WARPing) of Härdle and Scott (1990) as a fast method for performing the resampling steps. Throughout the paper we call the computable simulated stochastic structure the 'bootstrap world'.

2. The bootstrap procedures.

2.1. I.I.D. ERRORS (MODEL 1), RESIDUAL RESAMPLING .

To mimic the stochastic nature of this model in the bootstrap world one proceeds as follows.

STEP 1. *Calculate residuals*

$$\hat{\varepsilon}_i = Y_i - m(X_i) \qquad (i = 1,\ldots,n).$$

STEP 2. *Centering.*

$$\tilde{\varepsilon}_i = \hat{\varepsilon}_i - \hat{\varepsilon}_\bullet\text{, where } \hat{\varepsilon}_\bullet = n^{-1} \textstyle\sum_{i=1}^{n} \hat{\varepsilon}_i\,.$$

STEP 3. *Resampling.*

Draw randomly $\varepsilon_1^*, \ldots, \varepsilon_n^*$ *out of the set* $\{\tilde{\varepsilon}_1, \ldots, \tilde{\varepsilon}_n\}$.

STEP 4. *Create bootstrap observations*

$$Y_i^* = \hat{m}_g(X_i) + \varepsilon_i^*\,.$$

Here a bandwidth g *is chosen which may be different from the initial bandwidth* h.

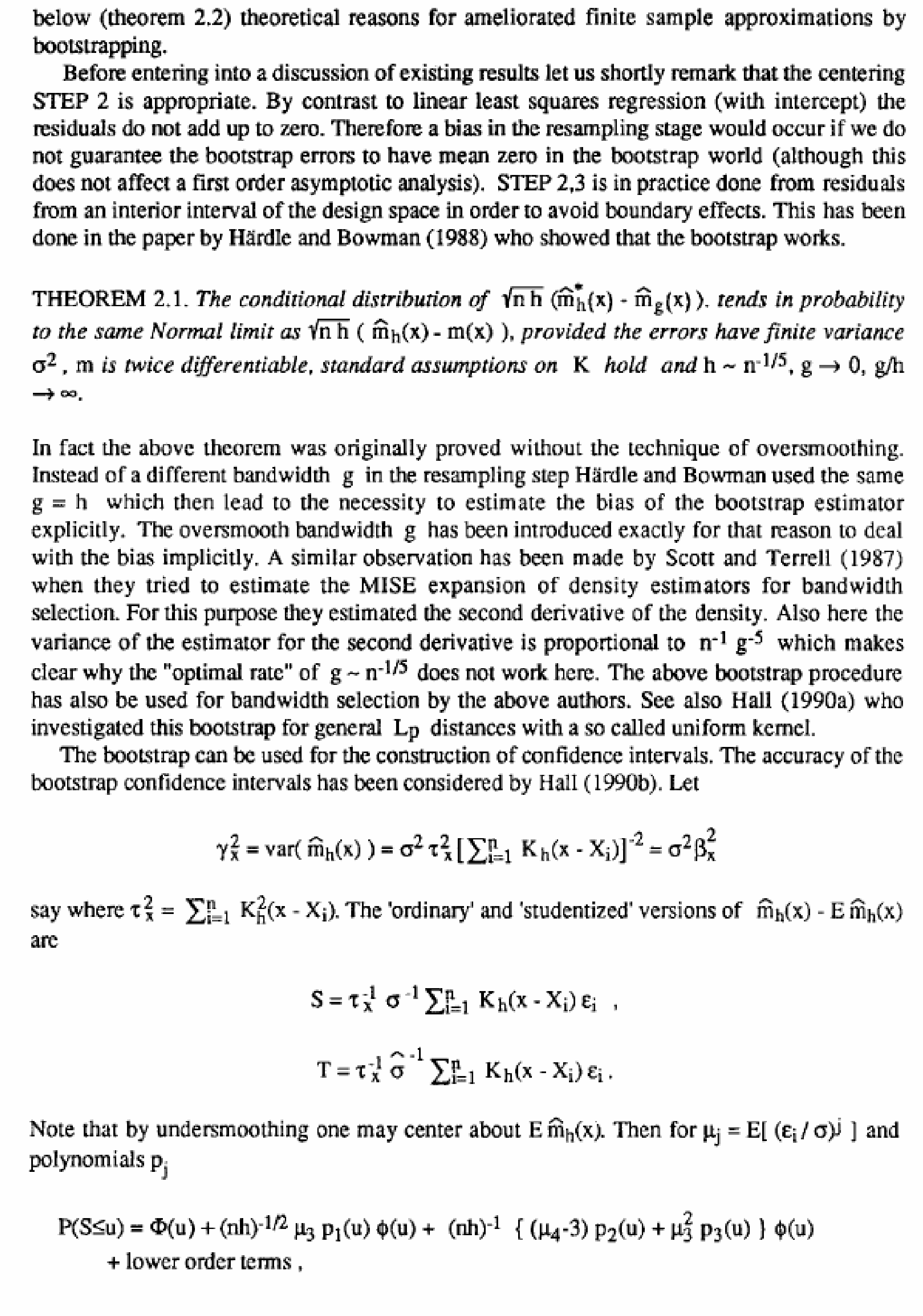
STEP 5. *Calculate a nonparametric kernel estimate in the bootstrap world.*

$$\hat{m}_h^*(x) = \frac{\sum_{i=1}^{n} K_h(x - X_i)\, Y_i^*}{\sum_{i=1}^{n} K_h(x - X_i)}\,. \tag{2.1}$$

STEP 6. *Bootstrap approximation.*

To approximate the distribution of the desired functional of $\hat{m}_h(\bullet) - m(\bullet)$ *use the computable conditional distribution of the functional of* $\hat{m}_h^*(\bullet) - \hat{m}_g(\bullet)$.

Let us discuss this bootstrap procedure for the case of the evaluation functional $\hat{m}_h(x) - m(x)$. It is common language to say that the 'bootstrap works', if in STEP 6 for suitable choice of h and g the bootstrap distribution tends to the same limit as $\hat{m}_h(x) - m(x)$ in probability. This says nothing about the finite sample behaviour of the bootstrap procedure. We will report



below (theorem 2.2) theoretical reasons for ameliorated finite sample approximations by bootstrapping.

Before entering into a discussion of existing results let us shortly remark that the centering STEP 2 is appropriate. By contrast to linear least squares regression (with intercept) the residuals do not add up to zero. Therefore a bias in the resampling stage would occur if we do not guarantee the bootstrap errors to have mean zero in the bootstrap world (although this does not affect a first order asymptotic analysis). STEP 2,3 is in practice done from residuals from an interior interval of the design space in order to avoid boundary effects. This has been done in the paper by Härdle and Bowman (1988) who showed that the bootstrap works.

THEOREM 2.1. *The conditional distribution of* $\sqrt{n\,h}\,(\hat{m}^*_h(x) - \hat{m}_g(x))$. *tends in probability to the same Normal limit as* $\sqrt{n\,h}\,(\hat{m}_h(x) - m(x))$, *provided the errors have finite variance* σ^2, m *is twice differentiable, standard assumptions on* K *hold and* $h \sim n^{-1/5}$, $g \to 0$, $g/h \to \infty$.

In fact the above theorem was originally proved without the technique of oversmoothing. Instead of a different bandwidth g in the resampling step Härdle and Bowman used the same $g = h$ which then lead to the necessity to estimate the bias of the bootstrap estimator explicitly. The oversmooth bandwidth g has been introduced exactly for that reason to deal with the bias implicitly. A similar observation has been made by Scott and Terrell (1987) when they tried to estimate the MISE expansion of density estimators for bandwidth selection. For this purpose they estimated the second derivative of the density. Also here the variance of the estimator for the second derivative is proportional to $n^{-1}\,g^{-5}$ which makes clear why the "optimal rate" of $g \sim n^{-1/5}$ does not work here. The above bootstrap procedure has also be used for bandwidth selection by the above authors. See also Hall (1990a) who investigated this bootstrap for general L_p distances with a so called uniform kernel.

The bootstrap can be used for the construction of confidence intervals. The accuracy of the bootstrap confidence intervals has been considered by Hall (1990b). Let

$$\gamma_x^2 = \mathrm{var}(\,\hat{m}_h(x)\,) = \sigma^2\,\tau_x^2\,[\textstyle\sum_{i=1}^n K_h(x - X_i)]^{-2} = \sigma^2\beta_x^2$$

say where $\tau_x^2 = \sum_{i=1}^n K_h^2(x - X_i)$. The 'ordinary' and 'studentized' versions of $\hat{m}_h(x) - E\,\hat{m}_h(x)$ are

$$S = \tau_x^{-1}\,\sigma^{-1}\textstyle\sum_{i=1}^n K_h(x - X_i)\,\varepsilon_i \quad ,$$

$$T = \tau_x^{-1}\,\hat{\sigma}^{-1}\textstyle\sum_{i=1}^n K_h(x - X_i)\,\varepsilon_i\,.$$

Note that by undersmoothing one may center about $E\,\hat{m}_h(x)$. Then for $\mu_j = E[\,(\varepsilon_i/\sigma)^j\,]$ and polynomials p_j

$$P(S \le u) = \Phi(u) + (nh)^{-1/2}\,\mu_3\,p_1(u)\,\phi(u) + (nh)^{-1}\,\{\,(\mu_4 - 3)\,p_2(u) + \mu_3^2\,p_3(u)\,\}\,\phi(u)$$
$$+ \text{ lower order terms },$$

$$P(T\le u) = \Phi(u) + (nh)^{-1/2}\,\mu_3\, p_1(u)\,\phi(u) + (nh)^{-1}\,\{ (\mu_4\text{-}3)\, p_2(u) + \mu_3^2\, p_3(u) \}\,\phi(u)$$
$$+ (h/n)^{1/2}\,\mu_3\, p_4(u)\,\phi(u) + \text{lower order terms.}$$

In the bootstrap world we have the complete analogue

$$P^*(S^*\le u) = \Phi(u) + (nh)^{-1/2}\,\widehat{\mu}_3\, p_1(u)\,\phi(u) + (nh)^{-1}\,\{ (\widehat{\mu}_4\text{-}3)\, p_2(u) + \widehat{\mu}_3^2\, p_3(u) \}\,\phi(u)$$
$$+ \text{lower order terms,}$$

$$P^*(T^*\le u) = \Phi(u) + (nh)^{-1/2}\,\widehat{\mu}_3\, p_1(u)\,\phi(u) + (nh)^{-1}\,\{ (\widehat{\mu}_4\text{-}3)\, p_2(u) + \widehat{\mu}_3^2\, p_3(u) \}\,\phi(u)$$
$$+ (h/n)^{1/2}\,\widehat{\mu}_3\, p_4(u)\,\phi(u) + \text{lower order terms.}$$

Since the constants $\widehat{\mu}_j$ converge to μ_j at the rate $n^{-1/2}$ we have the following Theorem.

THEOREM 2.2. *Under some regularity conditions the following expansions hold*

$$P^*(S^*\le u) - P(S\le u) = O_p(n^{-1}\, h^{-1/2}\,),$$

$$P^*(T^*\le u) - P(T\le u) = O_p(n^{-1}\, h^{-1/2}\,).$$

Note that this rate does not hold for the nonstudentized statistic

$$U = \beta_x^{-1}\,(\widehat{m}_h - E\,\widehat{m}_h)(x)\,.$$

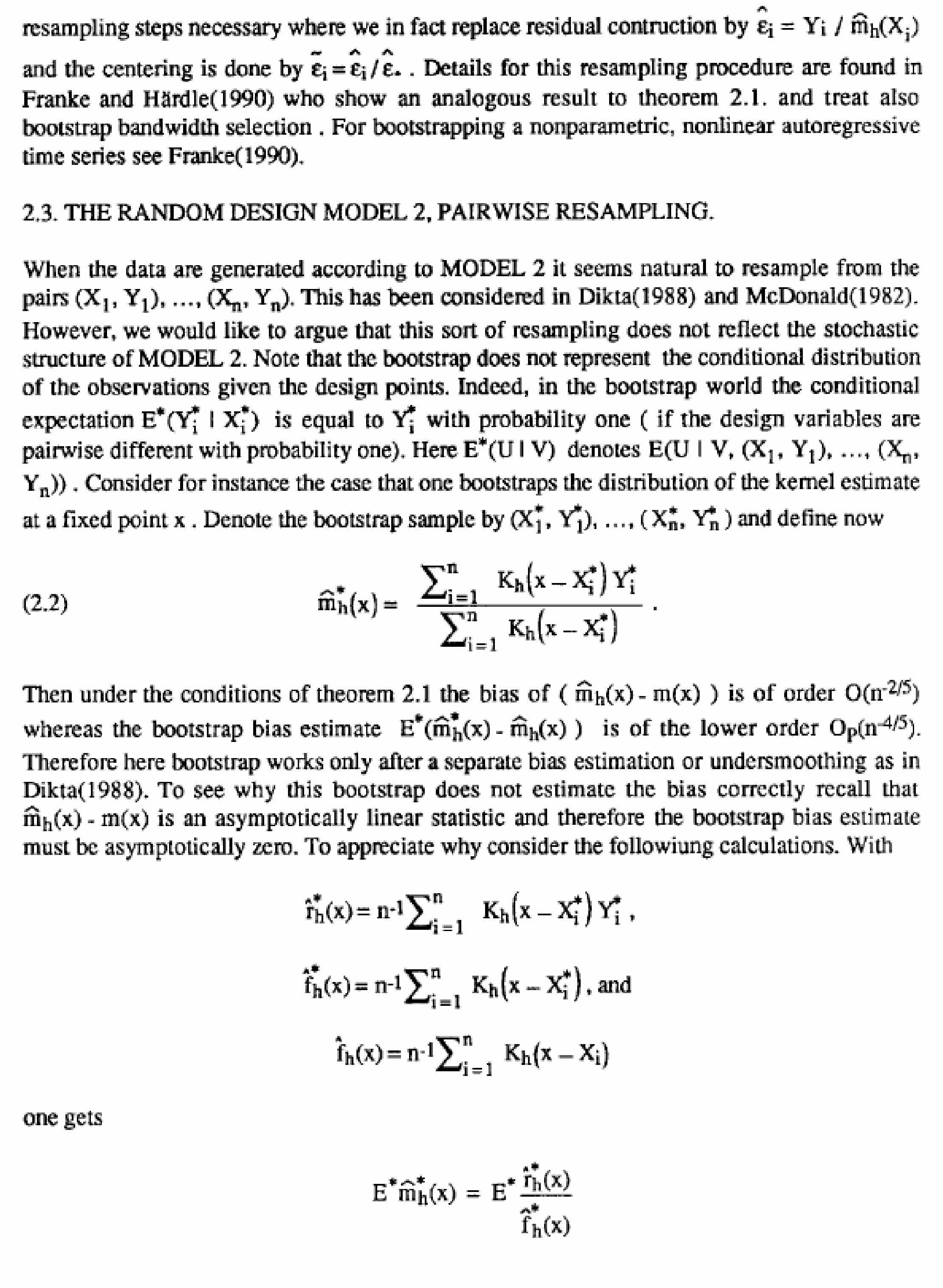
Then we observe a significantly weaker approximation

$$P^*(U^*\le u) - P(U\le u) = \Phi(u/\widehat{\sigma}) - \Phi(u/\sigma) + \text{lower order terms,}$$

since this difference is only of the order $\widehat{\sigma} - \sigma = O_p(n^{-1/2})$, see eg. Gasser et al.(1986). Also in other contexts (see for instants Hall (1988) and Mammen (1990b)) it is well known that studentizing gives a considerable improvement of coverage error.

2.2. AN APPLICATION WITH ALMOST I.I.D. ERRORS.

A direct application of the above bootstrap to kernel sprectral density estimates is not straightforward since the periodogram values become only "asymptotically independent" and also the error structure for this regression problem is of multiplicative nature. Consider a strictly stationary real valued process and let Y_i be the periodogram values at discrete equidistant frequencies in $[-\pi, \pi]$. The kernel spectral density estimate is of the form (1.1) (with denominator equal to $2\pi n$ since the design is equidistant on $[-\pi, \pi]$). The regression equation is mulitplicative, i.e. $Y_i = m(X_i)\,\varepsilon_i$. This makes a slight modification of the above



resampling steps necessary where we in fact replace residual contruction by $\hat{\varepsilon}_i = Y_i / \hat{m}_h(X_i)$ and the centering is done by $\tilde{\varepsilon}_i = \hat{\varepsilon}_i / \hat{\varepsilon}_\cdot$. Details for this resampling procedure are found in Franke and Härdle(1990) who show an analogous result to theorem 2.1. and treat also bootstrap bandwidth selection . For bootstrapping a nonparametric, nonlinear autoregressive time series see Franke(1990).

2.3. THE RANDOM DESIGN MODEL 2, PAIRWISE RESAMPLING.

When the data are generated according to MODEL 2 it seems natural to resample from the pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$. This has been considered in Dikta(1988) and McDonald(1982). However, we would like to argue that this sort of resampling does not reflect the stochastic structure of MODEL 2. Note that the bootstrap does not represent the conditional distribution of the observations given the design points. Indeed, in the bootstrap world the conditional expectation $E^*(Y_i^* \mid X_i^*)$ is equal to Y_i^* with probability one (if the design variables are pairwise different with probability one). Here $E^*(U \mid V)$ denotes $E(U \mid V, (X_1, Y_1), \ldots, (X_n, Y_n))$. Consider for instance the case that one bootstraps the distribution of the kernel estimate at a fixed point x . Denote the bootstrap sample by $(X_1^*, Y_1^*), \ldots, (X_n^*, Y_n^*)$ and define now

$$(2.2) \qquad \hat{m}_h^*(x) = \frac{\sum_{i=1}^n K_h\left(x - X_i^*\right) Y_i^*}{\sum_{i=1}^n K_h\left(x - X_i^*\right)} \; .$$

Then under the conditions of theorem 2.1 the bias of ($\hat{m}_h(x) - m(x)$) is of order $O(n^{-2/5})$ whereas the bootstrap bias estimate $E^*(\hat{m}_h^*(x) - \hat{m}_h(x))$ is of the lower order $O_p(n^{-4/5})$. Therefore here bootstrap works only after a separate bias estimation or undersmoothing as in Dikta(1988). To see why this bootstrap does not estimate the bias correctly recall that $\hat{m}_h(x) - m(x)$ is an asymptotically linear statistic and therefore the bootstrap bias estimate must be asymptotically zero. To appreciate why consider the followiung calculations. With

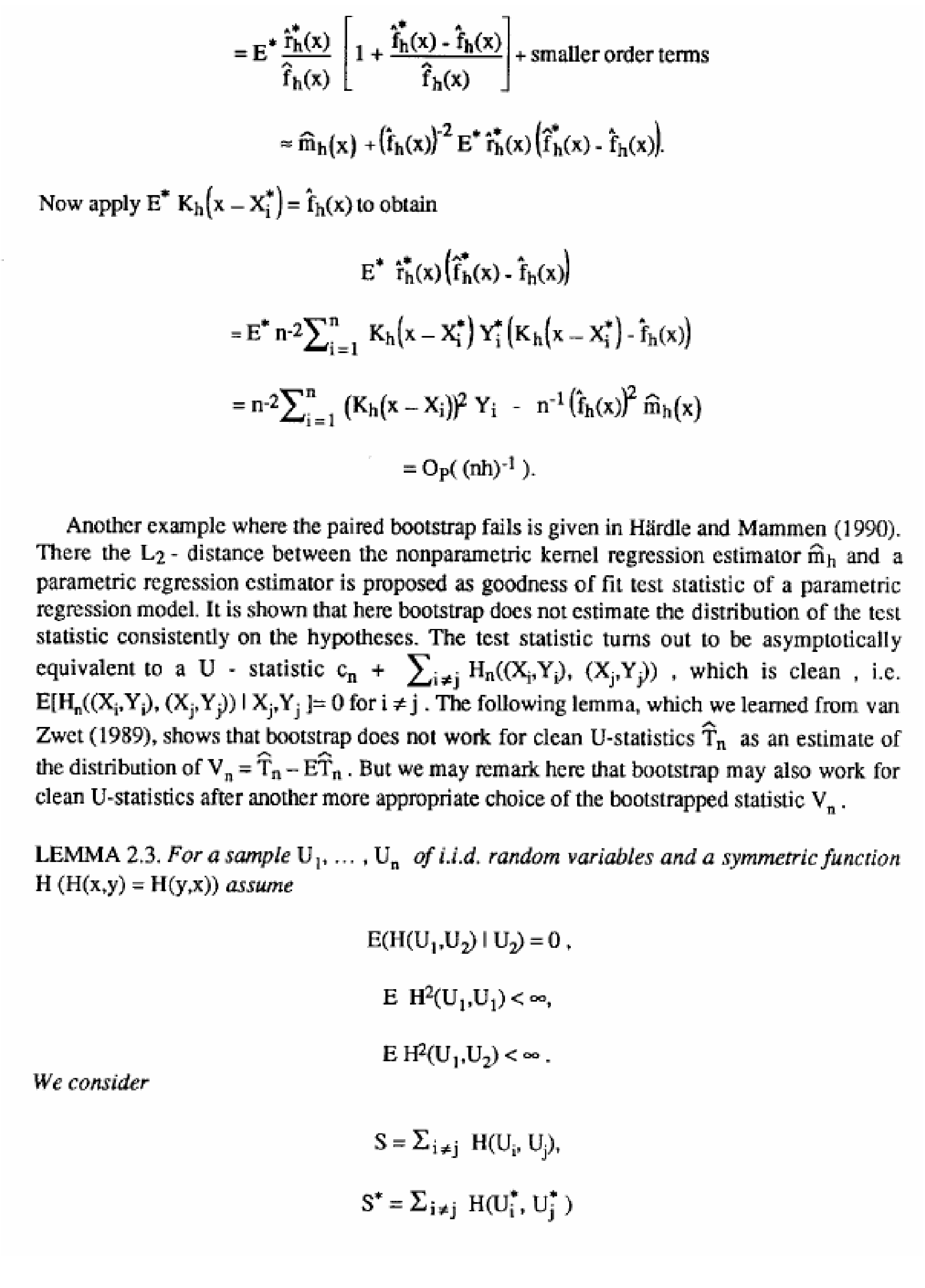
$$\hat{r}_h^*(x) = n^{-1} \sum_{i=1}^n K_h\left(x - X_i^*\right) Y_i^* ,$$

$$\hat{f}_h^*(x) = n^{-1} \sum_{i=1}^n K_h\left(x - X_i^*\right), \text{ and}$$

$$\hat{f}_h(x) = n^{-1} \sum_{i=1}^n K_h(x - X_i)$$

one gets

$$E^* \hat{m}_h^*(x) = E^* \frac{\hat{r}_h^*(x)}{\hat{f}_h^*(x)}$$



$$= E^* \frac{\hat{r}^*_h(x)}{\hat{f}_h(x)} \left[1 + \frac{\hat{f}^*_h(x) - \hat{f}_h(x)}{\hat{f}_h(x)} \right] + \text{smaller order terms}$$

$$\approx \hat{m}_h(x) + \left(\hat{f}_h(x)\right)^{-2} E^* \hat{r}^*_h(x) \left(\hat{f}^*_h(x) - \hat{f}_h(x)\right).$$

Now apply $E^* K_h\left(x - X^*_i\right) = \hat{f}_h(x)$ to obtain

$$E^* \; \hat{r}^*_h(x) \left(\hat{f}^*_h(x) - \hat{f}_h(x)\right)$$

$$= E^* n^{-2} \sum_{i=1}^{n} K_h\left(x - X^*_i\right) Y^*_i \left(K_h\left(x - X^*_i\right) - \hat{f}_h(x)\right)$$

$$= n^{-2} \sum_{i=1}^{n} \left(K_h(x - X_i)\right)^2 Y_i \; - \; n^{-1} \left(\hat{f}_h(x)\right)^2 \hat{m}_h(x)$$

$$= O_P(\,(nh)^{-1}\,).$$

Another example where the paired bootstrap fails is given in Härdle and Mammen (1990). There the L_2 - distance between the nonparametric kernel regression estimator $\hat{m}_h$ and a parametric regression estimator is proposed as goodness of fit test statistic of a parametric regression model. It is shown that here bootstrap does not estimate the distribution of the test statistic consistently on the hypotheses. The test statistic turns out to be asymptotically equivalent to a U - statistic $c_n + \sum_{i \neq j} H_n((X_i,Y_i), (X_j,Y_j))$, which is clean , i.e. $E[H_n((X_i,Y_i), (X_j,Y_j)) \mid X_j,Y_j]= 0$ for $i \neq j$. The following lemma, which we learned from van Zwet (1989), shows that bootstrap does not work for clean U-statistics $\hat{T}_n$ as an estimate of the distribution of $V_n = \hat{T}_n - E\hat{T}_n$. But we may remark here that bootstrap may also work for clean U-statistics after another more appropriate choice of the bootstrapped statistic V_n .

LEMMA 2.3. *For a sample* $U_1, \ldots , U_n$ *of i.i.d. random variables and a symmetric function* H $(H(x,y) = H(y,x))$ *assume*

$$E(H(U_1,U_2) \mid U_2) = 0\,,$$

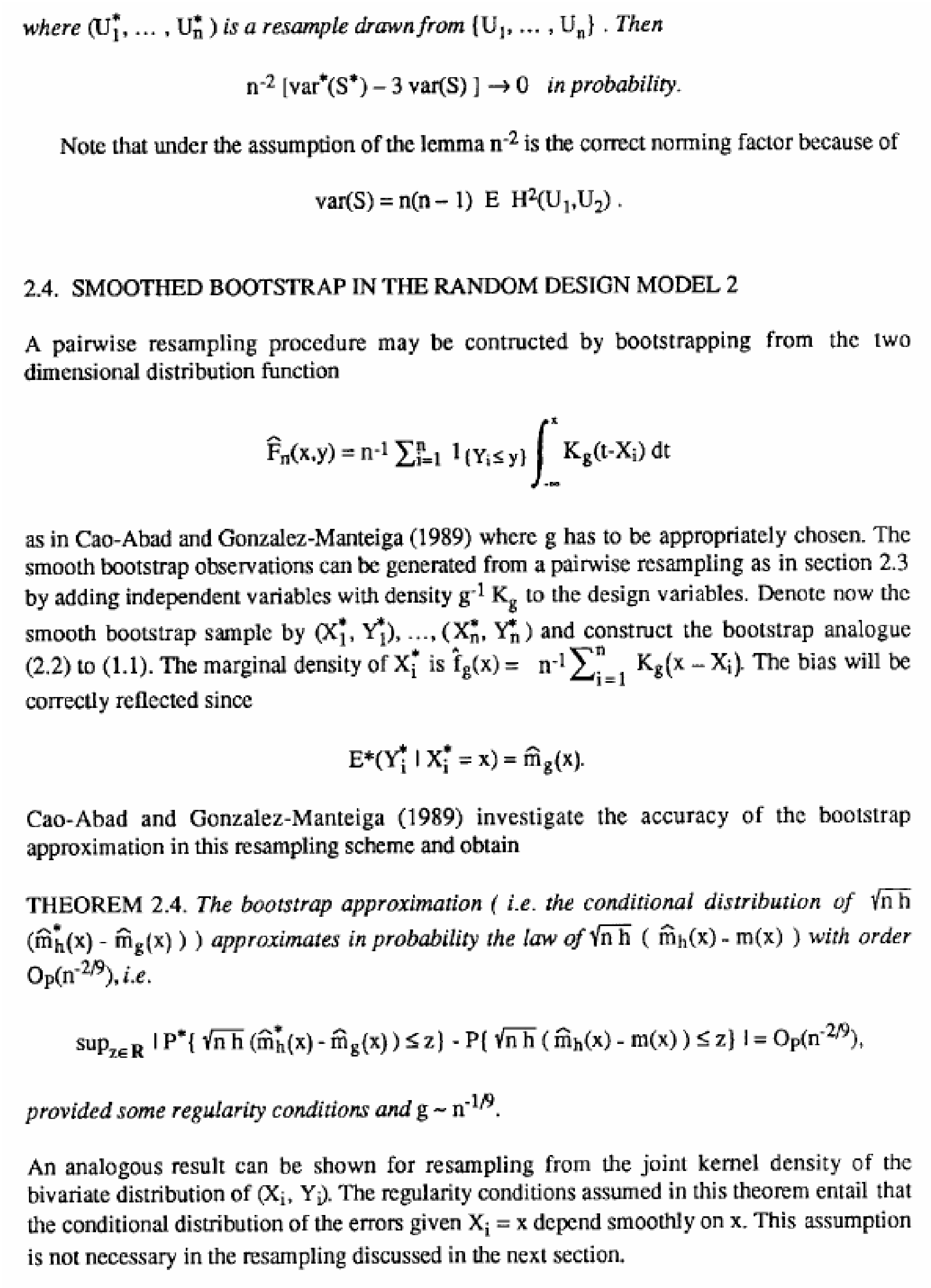
$$E \;\; H^2(U_1,U_1) < \infty,$$

$$E\, H^2(U_1,U_2) < \infty\,.$$

We consider

$$S = \Sigma_{i \neq j} \;\; H(U_i,\, U_j),$$

$$S^* = \Sigma_{i \neq j} \;\; H(U^*_i,\, U^*_j\,)$$



where $(U_1^*, \dots, U_n^*)$ *is a resample drawn from* $\{U_1, \dots, U_n\}$ *. Then*

$$n^{-2}\,[\mathrm{var}^*(S^*) - 3\,\mathrm{var}(S)\,] \to 0 \quad \textit{in probability.}$$

Note that under the assumption of the lemma n^{-2} is the correct norming factor because of

$$\mathrm{var}(S) = n(n-1)\ \ E\ \ H^2(U_1,U_2)\,.$$

2.4. SMOOTHED BOOTSTRAP IN THE RANDOM DESIGN MODEL 2

A pairwise resampling procedure may be contructed by bootstrapping from the two dimensional distribution function

$$\hat{F}_n(x,y) = n^{-1} \sum_{i=1}^{n} 1_{\{Y_i \le y\}} \int_{-\infty}^{x} K_g(t-X_i)\,dt$$

as in Cao-Abad and Gonzalez-Manteiga (1989) where g has to be appropriately chosen. The smooth bootstrap observations can be generated from a pairwise resampling as in section 2.3 by adding independent variables with density $g^{-1} K_g$ to the design variables. Denote now the smooth bootstrap sample by $(X_1^*, Y_1^*), \dots, (X_n^*, Y_n^*)$ and construct the bootstrap analogue (2.2) to (1.1). The marginal density of X_i^* is $\hat{f}_g(x) = \ n^{-1}\sum_{i=1}^{n} K_g(x - X_i)$. The bias will be correctly reflected since

$$E^*(Y_i^* \mid X_i^* = x) = \hat{m}_g(x).$$

Cao-Abad and Gonzalez-Manteiga (1989) investigate the accuracy of the bootstrap approximation in this resampling scheme and obtain

THEOREM 2.4. *The bootstrap approximation (i.e. the conditional distribution of* $\sqrt{n\,h}$ $(\hat{m}_h^*(x) - \hat{m}_g(x)\,)$ *) approximates in probability the law of* $\sqrt{n\,h}$ $(\ \hat{m}_h(x) - m(x)\ )$ *with order* $O_P(n^{-2/9})$, *i.e.*

$$\sup_{z\in\mathbf{R}} \mid P^*\{\ \sqrt{n\,h}\ (\hat{m}_h^*(x) - \hat{m}_g(x)\,) \le z\} - P\{\ \sqrt{n\,h}\ (\ \hat{m}_h(x) - m(x)\ ) \le z\} \mid = O_P(n^{-2/9}),$$

provided some regularity conditions and $g \sim n^{-1/9}$.

An analogous result can be shown for resampling from the joint kernel density of the bivariate distribution of (X_i, Y_i). The regularity conditions assumed in this theorem entail that the conditional distribution of the errors given $X_i = x$ depend smoothly on x. This assumption is not necessary in the resampling discussed in the next section.

2.5. NOT IDENTICALLY DISTRIBUTED ERRORS (MODEL 3), WILD BOOTSTRAPPING.

For MODEL 3 the *wild bootstrap* procedure has been introduced in Härdle and Mammen (1990). The wild bootstrap is related to proposals in linear regression models of Wu (1986) (see Beran (1986), Liu (1988), Liu and Singh(1988), Mammen (1989)). Since in this approach one is going to use one *single residual* $\hat{\varepsilon}_i$ to estimate the 'conditional' distribution $\mathcal{L}(Y_i - m(X_i) \mid X_i)$ by an estimate $\hat{F}_i$ we are calling it the *wild bootstrap*. More precisely $\hat{F}_i$ is defined as an arbitrary distribution which fulfills

$$E_{\hat{F}_i} Z = 0,$$

$$E_{\hat{F}_i} Z^2 = (\hat{\varepsilon}_i)^2 .$$

$$E_{\hat{F}_i} Z^3 = (\hat{\varepsilon}_i)^3 .$$

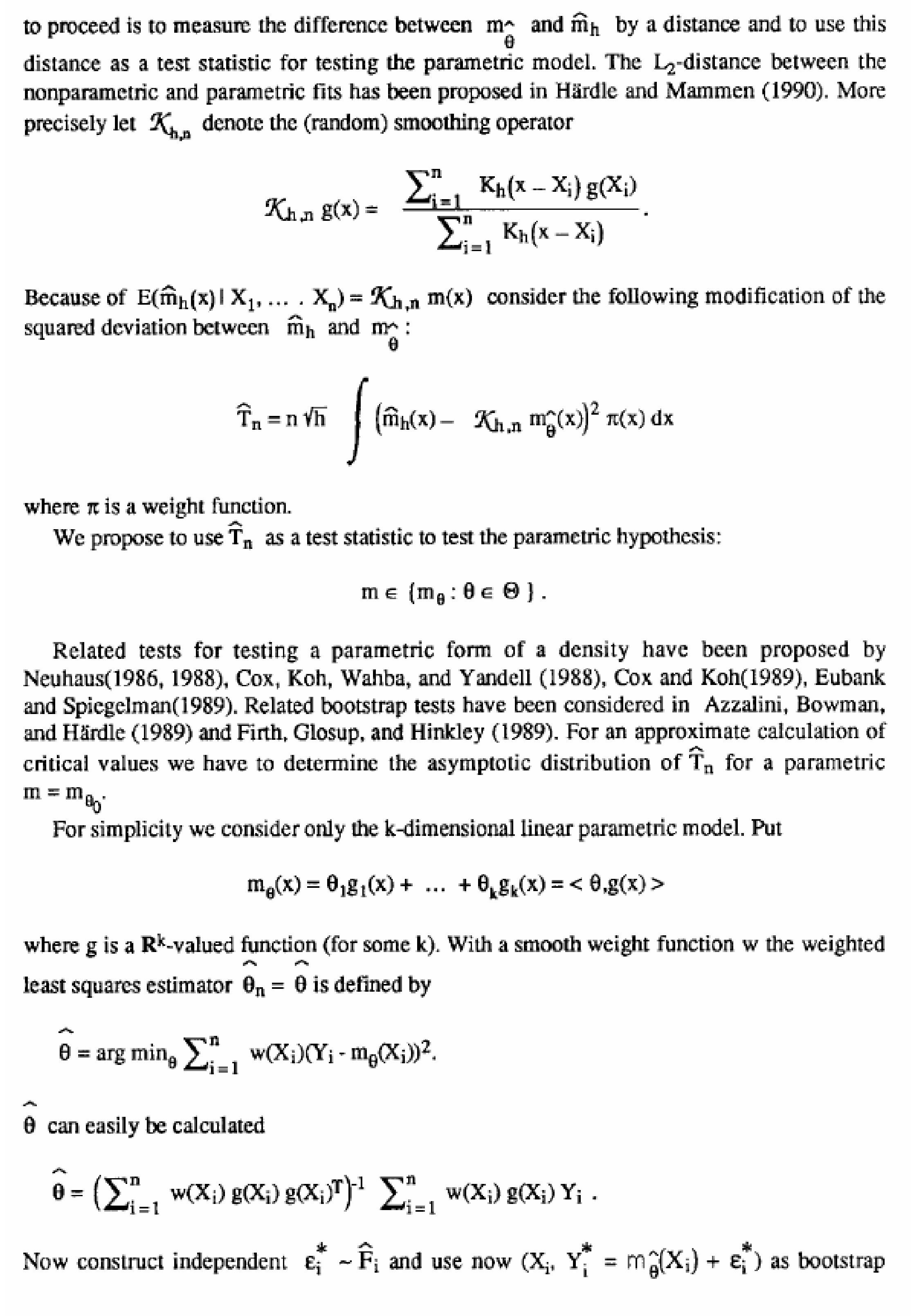
For instance one may use a two point distribution which is uniquely determined by these requirements. Then $\hat{F}_i = \mathcal{L}(Z_i)$ where $Z_i = -(\sqrt{5}-1)\,\hat{\varepsilon}_i / 2$ with probability $(\sqrt{5}+1)/(2\sqrt{5})$ and $Z_i = (\sqrt{5}+1)\,\hat{\varepsilon}_i / 2$ with probability $1 - (\sqrt{5}+1)/(2\sqrt{5})$. Or in another construction one may put $Z_i = \hat{\varepsilon}_i U_i$ where $\mathcal{L}(U_i)$ does not depend on i and where $EU_i = 0$, $E\,U_i^2 = E\,U_i^3 = 1$, e.g. $U_i = V_i/\sqrt{2} + (V_i^2 - 1)/2$ where the V_i's are independent N(0,1) - distributed variables.

For the construction of the bootstrap observations one generates independent $\varepsilon_i^* \sim \hat{F}_i$. Note that STEP 2 of section 2.1 is not necessary since the ε_i^* have automatically mean zero by construction. STEP 4,5 though are identical, one uses $(X_i, Y_i^* = \hat{m}_g(X_i) + \varepsilon_i^*)$ as bootstrap observations (for an appropriate choice of the bandwidth g , see section 2.1). Then one creates $\hat{m}_h^*(\cdot)$ according to (2.1).

To avoid technical regularity conditions on the 'conditional' error distributions let us consider the asymptotic performance of wild bootstrap in the random design MODEL 2. For the distribution of $\hat{m}_h(x) - m(x)$ at a finite number of points x wild bootstrap has been studied in Härdle and Marron(1990) see section 2.6 below.

Let us consider wild bootstrap for the test statistic of Härdle and Mammen(1990) which we have mentioned in the last section. This test statistic tests the hypothesis of a parametric regression model and it is based on the distance between the parametric and the nonparametric kernel regression estimate $\hat{m}_h$.

For the regression function $m(\cdot) = E(Y_i \mid X_i = \cdot)$ a parametric model $\{m_\theta : \theta \in \Theta\}$ may be given. The parametric approach shall be compared with the nonparametric analysis which are only based on the assumption that $m(\cdot)$ is a 'smooth' function. To this account a parametric regression estimator $m_{\hat{\theta}}$ may be plotted against a kernel estimator $\hat{m}_h$. The question arises if visible differences between $m_{\hat{\theta}}$ and $\hat{m}_h$ can be explained by stochastic fluctuations or if they suggest to use nonparametric instead of parametric methods. One way



to proceed is to measure the difference between $m_{\hat\theta}$ and $\hat{m}_h$ by a distance and to use this distance as a test statistic for testing the parametric model. The L_2-distance between the nonparametric and parametric fits has been proposed in Härdle and Mammen (1990). More precisely let $\mathcal{K}_{h,n}$ denote the (random) smoothing operator

$$\mathcal{K}_{h,n}\, g(x) = \frac{\sum_{i=1}^{n} K_h(x - X_i)\, g(X_i)}{\sum_{i=1}^{n} K_h(x - X_i)}.$$

Because of $E(\hat{m}_h(x) \mid X_1, \dots . X_n) = \mathcal{K}_{h,n}\, m(x)$ consider the following modification of the squared deviation between $\hat{m}_h$ and $m_{\hat\theta}$:

$$\hat{T}_n = n \sqrt{h} \int \left(\hat{m}_h(x) - \mathcal{K}_{h,n}\, m_{\hat\theta}(x)\right)^2 \pi(x)\, dx$$

where π is a weight function.

We propose to use $\hat{T}_n$ as a test statistic to test the parametric hypothesis:

$$m \in \{m_\theta : \theta \in \Theta\}\,.$$

Related tests for testing a parametric form of a density have been proposed by Neuhaus(1986, 1988), Cox, Koh, Wahba, and Yandell (1988), Cox and Koh(1989), Eubank and Spiegelman(1989). Related bootstrap tests have been considered in Azzalini, Bowman, and Härdle (1989) and Firth, Glosup, and Hinkley (1989). For an approximate calculation of critical values we have to determine the asymptotic distribution of $\hat{T}_n$ for a parametric $m = m_{\theta_0}$.

For simplicity we consider only the k-dimensional linear parametric model. Put

$$m_\theta(x) = \theta_1 g_1(x) + \ \dots\ + \theta_k g_k(x) = <\theta, g(x)>$$

where g is a $\mathbf{R}^k$-valued function (for some k). With a smooth weight function w the weighted least squares estimator $\hat\theta_n = \hat\theta$ is defined by

$$\hat\theta = \arg\min_\theta \sum_{i=1}^{n} w(X_i)(Y_i - m_\theta(X_i))^2.$$

$\hat\theta$ can easily be calculated

$$\hat\theta = \left(\sum_{i=1}^{n} w(X_i)\, g(X_i)\, g(X_i)^T\right)^{-1} \sum_{i=1}^{n} w(X_i)\, g(X_i)\, Y_i\ .$$

Now construct independent $\varepsilon_i^* \sim \hat{F}_i$ and use now $(X_i,\ Y_i^* = m_{\hat\theta}(X_i) + \varepsilon_i^*)$ as bootstrap observations. Then create $\widehat{T}_n^*$ like $\widehat{T}_n$ by the squared deviance between the parametric fit and the nonparametric fit. From the Monte Carlo approximation of $\mathcal{L}^*(\widehat{T}_n^*)$ construct the $(1-\alpha)$ quantile $\hat{t}_\alpha^W$ and reject the parametric hypothesis if $\widehat{T}_n > \hat{t}_\alpha^W$. The following Theorem has been shown in Härdle and Mammen(1990).

THEOREM 2.5. *Assume, that m lies in the parametric hypotheses* $\{m_\theta\colon \theta \in \Theta\}$. *Then under some regularity conditions for a deterministic sequence* c_n *the conditional distribution of* $\widehat{T}_n^* - c_n$ *converges weakly to the same limit as the distribution of* $\widehat{T}_n - c_n$ *(in probability).*

2.6 SIMULTANEOUS ERROR BARS

Under the condition of theorem 2.1. the conditional distribution of $\sqrt{nh}\,(\widehat{m}_h^*(x) - \widehat{m}_g(x))$ tends in probability to the same Normal limit as $\sqrt{nh}\,(\widehat{m}_h(x) - m(x))$. This convergence holds in fact uniformly over a grid of points, so it can be used for the construction of error bars with simultaneous coverage probability. The accuracy of these confidence intervals has been investigated by Cao - Abad (1990) and Härdle, Huet and Jolivet(1990).

The main advantage of bootstrapping in this context lies in the fact that the simulated distribution of $\sqrt{nh}\,(\widehat{m}_h^*(\cdot) - \widehat{m}_g(\cdot))$ at a finite number of points can be easily used to contruct confidence intervals with simultaneuos coverage probability. A conservative way of constructing confidence intervals at a finite number M of design points is the Bonferroni method where we use M pointwise confidence intervals each with coverage probability $1 - \alpha/M$. A more accurate method is to construct first pointwise confidence intervals with coverage probability $1 - \beta$, say, such that the uniform coverage is $1-\alpha$. We suggest the following halving approach. First try individual (i.e. at each design point) coverage probabilty $1 - \beta = 1 - \alpha/(2M)$ and calculate by simulation the resulting simultaneuos coverage $1 - \alpha_\beta$. If the result is more than α then try $\beta = \alpha/(4M)$ otherwise next try $\beta = 3\alpha/(4M)$. After stopping this halving approach find neighboring values β_* and β^* so that $\alpha_{\beta_*} < \alpha < \alpha_{\beta^*}$. Finally take the weighted average of the β_* and β^* intervals. For an application in an econometric context see Härdle and Marron (1990).

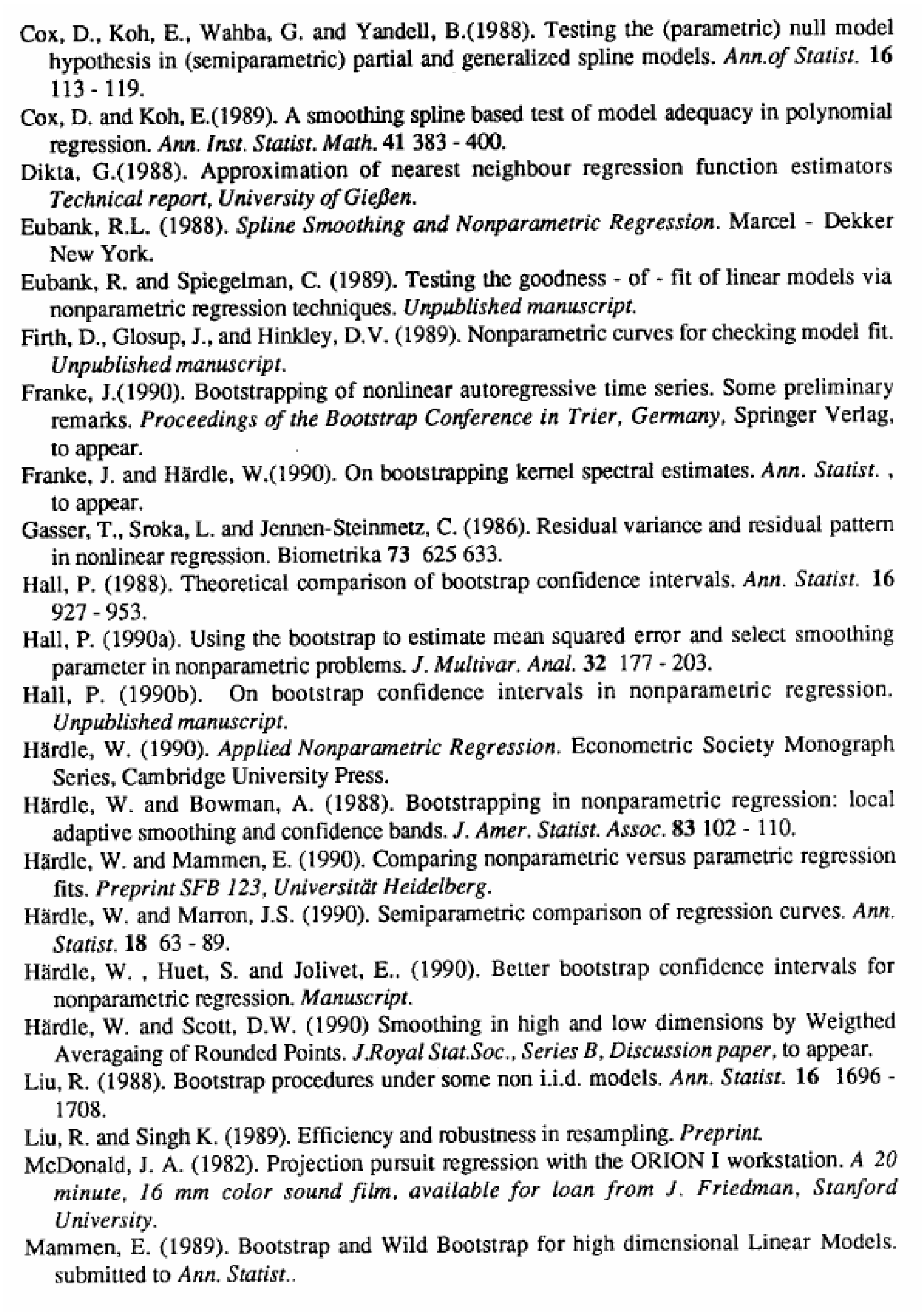
References

Azzalini, A., Bowman, A.W. and Härdle, W. (1989). On the use of nonparametric regression for model checking. *Biometrika*. **76** 1 - 11.

Beran,R.(1986). Discussion to Wu,C.F.J.: Jackknife, bootstrap and other resampling methods in regression analysis. *Ann. Statist.* **14** 1295-1298.

Cao - Abad, R. (1990). Rate of convergence for the wild bootstrap in nonparametric regression. *Ann. Statist.*, to appear.

Carroll, R.J. (1982). Adapting for heteroscedasticity in linear models. *Ann. Statist.* **10** 1224 - 1233.



Cox, D., Koh, E., Wahba, G. and Yandell, B.(1988). Testing the (parametric) null model hypothesis in (semiparametric) partial and generalized spline models. *Ann.of Statist.* **16** 113 - 119.

Cox, D. and Koh, E.(1989). A smoothing spline based test of model adequacy in polynomial regression. *Ann. Inst. Statist. Math.* **41** 383 - 400.

Dikta, G.(1988). Approximation of nearest neighbour regression function estimators *Technical report, University of Gießen.*

Eubank, R.L. (1988). *Spline Smoothing and Nonparametric Regression*. Marcel - Dekker New York.

Eubank, R. and Spiegelman, C. (1989). Testing the goodness - of - fit of linear models via nonparametric regression techniques. *Unpublished manuscript.*

Firth, D., Glosup, J., and Hinkley, D.V. (1989). Nonparametric curves for checking model fit. *Unpublished manuscript.*

Franke, J.(1990). Bootstrapping of nonlinear autoregressive time series. Some preliminary remarks. *Proceedings of the Bootstrap Conference in Trier, Germany*, Springer Verlag, to appear.

Franke, J. and Härdle, W.(1990). On bootstrapping kernel spectral estimates. *Ann. Statist.* , to appear.

Gasser, T., Sroka, L. and Jennen-Steinmetz, C. (1986). Residual variance and residual pattern in nonlinear regression. Biometrika **73** 625 633.

Hall, P. (1988). Theoretical comparison of bootstrap confidence intervals. *Ann. Statist.* **16** 927 - 953.

Hall, P. (1990a). Using the bootstrap to estimate mean squared error and select smoothing parameter in nonparametric problems. *J. Multivar. Anal.* **32** 177 - 203.

Hall, P. (1990b). On bootstrap confidence intervals in nonparametric regression. *Unpublished manuscript.*

Härdle, W. (1990). *Applied Nonparametric Regression*. Econometric Society Monograph Series, Cambridge University Press.

Härdle, W. and Bowman, A. (1988). Bootstrapping in nonparametric regression: local adaptive smoothing and confidence bands. *J. Amer. Statist. Assoc.* **83** 102 - 110.

Härdle, W. and Mammen, E. (1990). Comparing nonparametric versus parametric regression fits. *Preprint SFB 123, Universität Heidelberg.*

Härdle, W. and Marron, J.S. (1990). Semiparametric comparison of regression curves. *Ann. Statist.* **18** 63 - 89.

Härdle, W. , Huet, S. and Jolivet, E.. (1990). Better bootstrap confidence intervals for nonparametric regression. *Manuscript.*

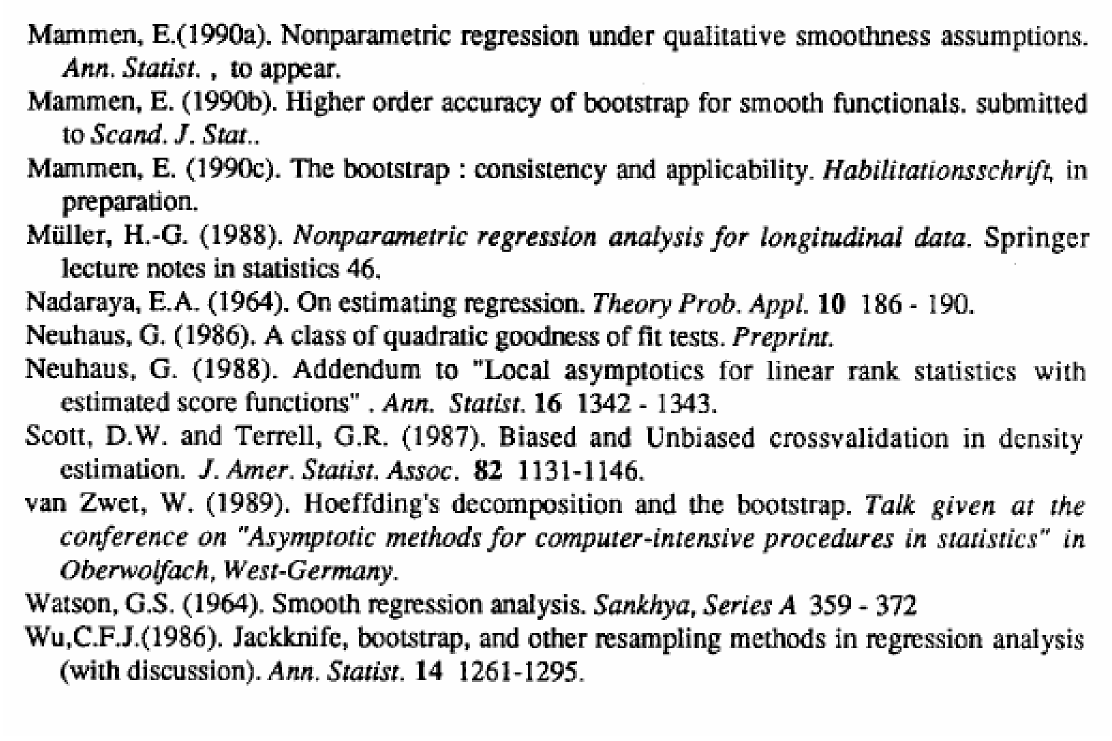
Härdle, W. and Scott, D.W. (1990) Smoothing in high and low dimensions by Weigthed Averagaing of Rounded Points. *J.Royal Stat.Soc., Series B, Discussion paper,* to appear.

Liu, R. (1988). Bootstrap procedures under some non i.i.d. models. *Ann. Statist.* **16** 1696 - 1708.

Liu, R. and Singh K. (1989). Efficiency and robustness in resampling. *Preprint.*

McDonald, J. A. (1982). Projection pursuit regression with the ORION I workstation. *A 20 minute, 16 mm color sound film, available for loan from J. Friedman, Stanford University.*

Mammen, E. (1989). Bootstrap and Wild Bootstrap for high dimensional Linear Models. submitted to *Ann. Statist.*.



Mammen, E.(1990a). Nonparametric regression under qualitative smoothness assumptions. *Ann. Statist.* , to appear.

Mammen, E. (1990b). Higher order accuracy of bootstrap for smooth functionals. submitted to *Scand. J. Stat.*.

Mammen, E. (1990c). The bootstrap : consistency and applicability. *Habilitationsschrift*, in preparation.

Müller, H.-G. (1988). *Nonparametric regression analysis for longitudinal data*. Springer lecture notes in statistics 46.

Nadaraya, E.A. (1964). On estimating regression. *Theory Prob. Appl.* **10** 186 - 190.

Neuhaus, G. (1986). A class of quadratic goodness of fit tests. *Preprint*.

Neuhaus, G. (1988). Addendum to "Local asymptotics for linear rank statistics with estimated score functions" . *Ann. Statist.* **16** 1342 - 1343.

Scott, D.W. and Terrell, G.R. (1987). Biased and Unbiased crossvalidation in density estimation. *J. Amer. Statist. Assoc.* **82** 1131-1146.

van Zwet, W. (1989). Hoeffding's decomposition and the bootstrap. *Talk given at the conference on "Asymptotic methods for computer-intensive procedures in statistics" in Oberwolfach, West-Germany.*

Watson, G.S. (1964). Smooth regression analysis. *Sankhya, Series A* 359 - 372

Wu,C.F.J.(1986). Jackknife, bootstrap, and other resampling methods in regression analysis (with discussion). *Ann. Statist.* **14** 1261-1295.

meaningful to perform an asymptotic analysis at places where there are hardly any data. It may seem a bit ironical that Chu and Marron make the same assumption "bounded from below" in the same paper (assumption A.4 of Section 3). In an interesting paper, Fan (1990) concludes independently about the Nadaraya-Watson estimator (remark 2, Section 3) "... hence its asymptotic minimax efficiency is arbitrary small."

CONCLUSIONS

Our conclusion is that the convolution weights are clearly superior to evaluation weights for fixed design, since we have the same variance for both methods but a nasty bias for evaluation weights. For random design, the problem seems to us more open: There is a minimax argument, and we would like to repeat a general argument, which is not well quoted by Chu and Marron (Section 3): "The latter authors [Gasser and colleagues] in particular seem to feel that variability is not a major issue, apparently basing their feelings on the premise that it is always easy to gather simply more data." What we said when discussing the structural bias of the evaluation weights was the following (Gasser and Engel, 1990): "These bias problems are particularly accentuated in the scientific process of many empirical sciences: studies are usually replicated by sticking to the design of the previously published study. In this way, qualitatively misleading phenomena as obtained by the Nadaraya-Watson estimator will be attributed even more confidence."

OUTLOOK

One way out of this problem has been opened by Fan (1990), who showed that for random design local polynomials have the same bias as convolution weights and the same variance as evaluation weights (the equivalence of local polynomials to convolution type kernel estimators for fixed design had been shown by Müller, 1987). A further possibility for improving the variance properties of convolution weights has been described by Chu and Marron in Section 6.

Comment

Birgit Grund and Wolfgang Härdle

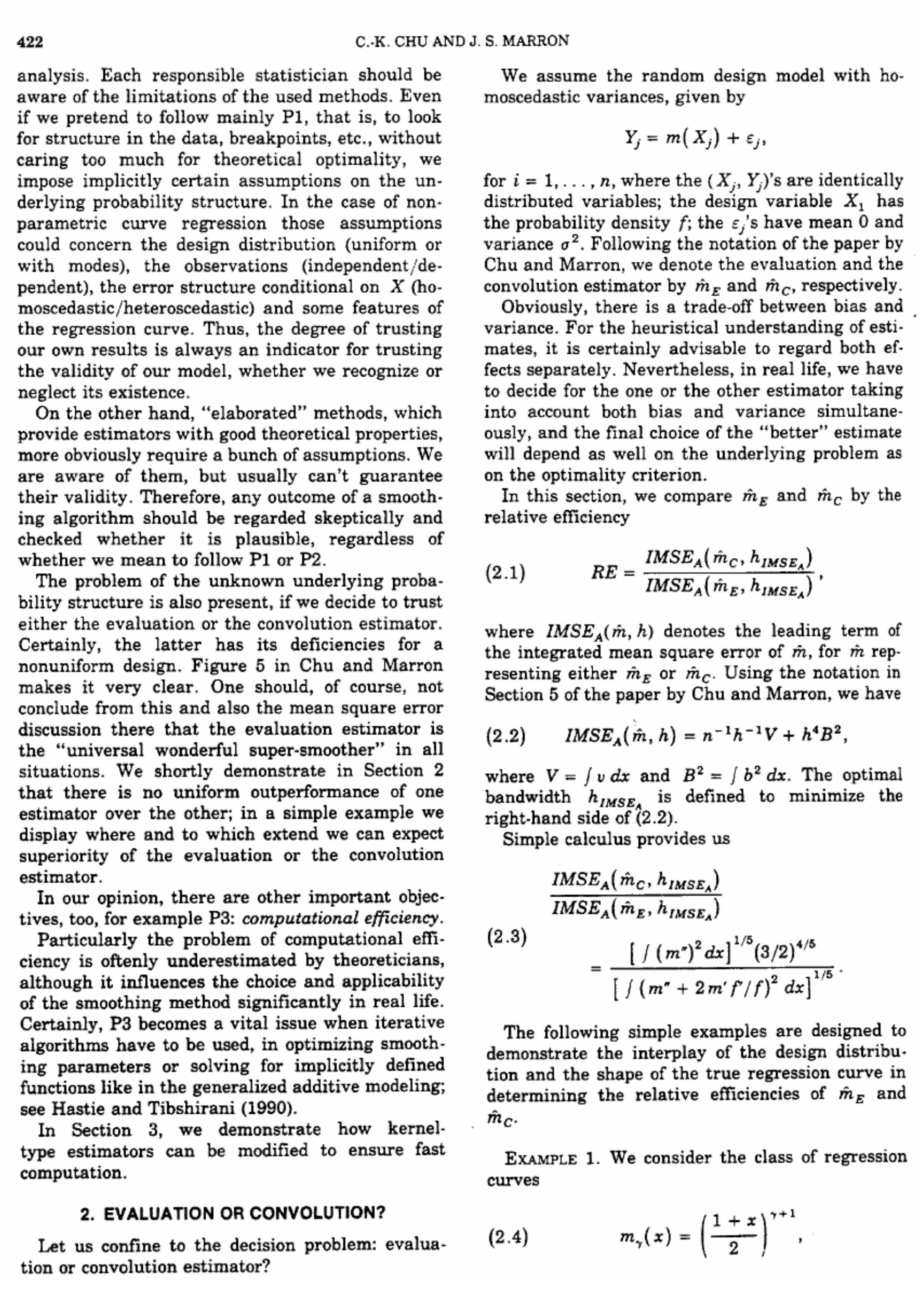
1. OBJECTIVES OF SMOOTHING

Smoothing has become a standard data analytic tool. A good indicator of this is the increased offer of smoothing procedures in a variety of standard statistical software packages. It is therefore high time to provide background information that enables statisticians and users to critically evaluate the—in the meantime—rich basket of smoothing tools. The paper by Chu and Marron meets this demand for information and compares two different kernel regression estimators on an easy, understandable level. The authors combine successfully careful mathematical discussion with heuristic arguments in a well-done exposition. Cleverly chosen striking examples provide an easy access to not immediately apparent problems in smoothing for data analysis. We congratulate the authors to this valuable contribution.

Among the many objectives of smoothing, there are certainly the two perhaps most discussed. These are P1: to find structure; and P2: to construct estimators from a probability distribution.

We agree that the interplay of these two objectives is vital for an honest parameter-free data

Birgit Grund is Assistant Professor, School of Statistics, University of Minnesota, 352 Classroom-Office Building, 1994 Bufford Avenue, St. Paul, Minnesota 55108, and CORE Research Fellow, Université Catholique de Louvain, Belgium. Wolfgang Härdle is Professor of Statistics, CORE, Université Catholique de Louvain, 34 Voie du Roman Pays, B01348 Louvain-la-Neuve, Belgium. He is currently visiting CentER, Tilburg University, The Netherlands.



analysis. Each responsible statistician should be aware of the limitations of the used methods. Even if we pretend to follow mainly P1, that is, to look for structure in the data, breakpoints, etc., without caring too much for theoretical optimality, we impose implicitly certain assumptions on the underlying probability structure. In the case of nonparametric curve regression those assumptions could concern the design distribution (uniform or with modes), the observations (independent/dependent), the error structure conditional on X (homoscedastic/heteroscedastic) and some features of the regression curve. Thus, the degree of trusting our own results is always an indicator for trusting the validity of our model, whether we recognize or neglect its existence.

On the other hand, "elaborated" methods, which provide estimators with good theoretical properties, more obviously require a bunch of assumptions. We are aware of them, but usually can't guarantee their validity. Therefore, any outcome of a smoothing algorithm should be regarded skeptically and checked whether it is plausible, regardless of whether we mean to follow P1 or P2.

The problem of the unknown underlying probability structure is also present, if we decide to trust either the evaluation or the convolution estimator. Certainly, the latter has its deficiencies for a nonuniform design. Figure 5 in Chu and Marron makes it very clear. One should, of course, not conclude from this and also the mean square error discussion there that the evaluation estimator is the "universal wonderful super-smoother" in all situations. We shortly demonstrate in Section 2 that there is no uniform outperformance of one estimator over the other; in a simple example we display where and to which extend we can expect superiority of the evaluation or the convolution estimator.

In our opinion, there are other important objectives, too, for example P3: *computational efficiency*.

Particularly the problem of computational efficiency is oftenly underestimated by theoreticians, although it influences the choice and applicability of the smoothing method significantly in real life. Certainly, P3 becomes a vital issue when iterative algorithms have to be used, in optimizing smoothing parameters or solving for implicitly defined functions like in the generalized additive modeling; see Hastie and Tibshirani (1990).

In Section 3, we demonstrate how kernel-type estimators can be modified to ensure fast computation.

2. EVALUATION OR CONVOLUTION?

Let us confine to the decision problem: evaluation or convolution estimator?

We assume the random design model with homoscedastic variances, given by

$$Y_j = m(X_j) + \varepsilon_j,$$

for $i = 1, \ldots, n$, where the (X_j, Y_j)'s are identically distributed variables; the design variable X_1 has the probability density f; the ε_j's have mean 0 and variance σ^2. Following the notation of the paper by Chu and Marron, we denote the evaluation and the convolution estimator by $\hat{m}_E$ and $\hat{m}_C$, respectively.

Obviously, there is a trade-off between bias and variance. For the heuristical understanding of estimates, it is certainly advisable to regard both effects separately. Nevertheless, in real life, we have to decide for the one or the other estimator taking into account both bias and variance simultaneously, and the final choice of the "better" estimate will depend as well on the underlying problem as on the optimality criterion.

In this section, we compare $\hat{m}_E$ and $\hat{m}_C$ by the relative efficiency

$$RE = \frac{IMSE_A(\hat{m}_C, h_{IMSE_A})}{IMSE_A(\hat{m}_E, h_{IMSE_A})}, \tag{2.1}$$

where $IMSE_A(\hat{m}, h)$ denotes the leading term of the integrated mean square error of $\hat{m}$, for $\hat{m}$ representing either $\hat{m}_E$ or $\hat{m}_C$. Using the notation in Section 5 of the paper by Chu and Marron, we have

$$IMSE_A(\hat{m}, h) = n^{-1}h^{-1}V + h^4B^2, \tag{2.2}$$

where $V = \int v\,dx$ and $B^2 = \int b^2\,dx$. The optimal bandwidth h_{IMSE_A} is defined to minimize the right-hand side of (2.2).

Simple calculus provides us

$$\begin{aligned}\frac{IMSE_A(\hat{m}_C, h_{IMSE_A})}{IMSE_A(\hat{m}_E, h_{IMSE_A})} \\ = \frac{\left[\int (m'')^2\,dx\right]^{1/5}(3/2)^{4/5}}{\left[\int (m'' + 2m'f'/f)^2\,dx\right]^{1/5}}.\end{aligned} \tag{2.3}$$

The following simple examples are designed to demonstrate the interplay of the design distribution and the shape of the true regression curve in determining the relative efficiencies of $\hat{m}_E$ and $\hat{m}_C$.

Example 1. We consider the class of regression curves

$$m_\gamma(x) = \left(\frac{1+x}{2}\right)^{\gamma+1}, \tag{2.4}$$

for $x \in [-1, 1]$ and for $\gamma \geq 0$, and the family of random designs with densities

$$f_\omega(x) = (1-\omega)(1/2) + \omega\varphi_{[-1,1]}(x), \tag{2.5}$$

for $\omega \in [0, 1]$, where $\varphi_{[-1,1]}$ denotes the density of the standard normal distribution, truncated to $[-1, 1]$. For $\omega = 0$, the design variable X is uniformly distributed on $[-1, 1]$; the most concentrated design in this example is the truncated standard normal distribution ($\omega = 1$).

Some representatives of the regarded regression curves are shown in Figure 1. The parameter $\gamma = 0$ corresponds to a linear function, and for growing γ the curves deviate more and more from the straight line.

Figure 2 below displays RE (see (2.1)), for all combinations of designs f_ω and true regression curves m_γ. With respect to the γ-scale here, the curves in Figure 1 are equidistant and represent the whole range.

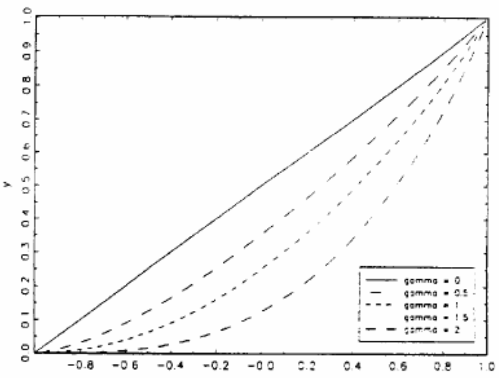


FIG. 1. *Regression curves $m_\gamma(x) = ((1+x)/2)^{1+\gamma}$ for different values of γ. With respect to the γ-scale of the Figures 2 and 4, the curves are equidistant and cover the whole range.*

We have chosen the above families for convenient control of bias and variance.

Under uniform design ($\omega = 0$) both $\hat{m}_E$ and $\hat{m}_C$ have the same bias, but $\hat{m}_C$ has a bigger variance. More precisely, $v_C = 3v_E/2$ for all $x \in [-1, 1]$, for all bandwidths and any regression curve. This setting causes

$$RE = (3/2)^{4/5} \quad \text{for all } \gamma \geq 0$$

and is reflected by the straight line on the right front side of the box.

The left front side of the box in Figure 2 corresponds to estimating a straight line under increasingly nonuniform design, the ideal background for the convolution estimator. We see that the trade-off between bias and variance begins to favor $\hat{m}_C$ at about $\omega \approx 1/3$.

The region where the convolution estimator is superior to the evaluation estimator ($RE \leq 1$) is rather small for this example. Even under the most nonuniform design ($\omega = 1$) the convolution estimator is better just for $\gamma < 0.5$ (see Figure 1).

EXAMPLE 2. We regard the same class of regression curves, but the class of random designs is now given by the densities

$$f_\omega(x) = \varphi_{[-1,1]}\left(\frac{x}{\omega}\right),$$

for $\omega > 0$, that is, the truncated $N(0, \omega^2)$ normal distributions. Figure 3 below shows some representatives. We see that $\omega = 2.3$ describes almost the uniform design.

The relative efficiency RE is displayed in Figure 4. Note, that the densities in Figure 3 are not

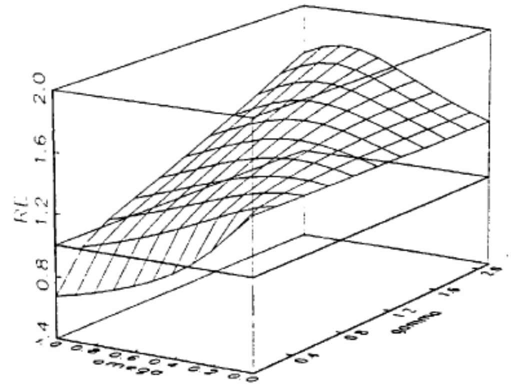


FIG. 2. *Relative efficiency RE of $\hat{m}_C$ and $\hat{m}_E$, in dependence from design density f_ω and regression curve m_γ; see Example 1.*

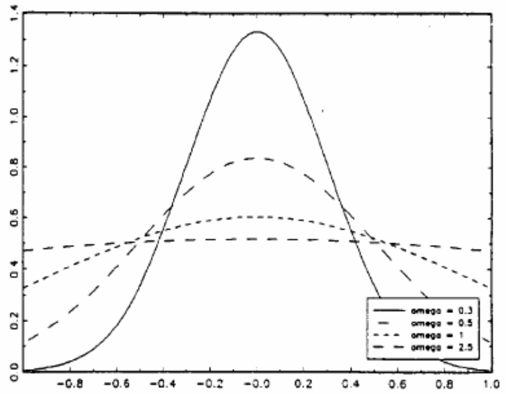


FIG. 3. *Truncated normal densities $f_\omega(x) = \varphi_{[-1,1]}(x/\omega)$ for different values of ω.*

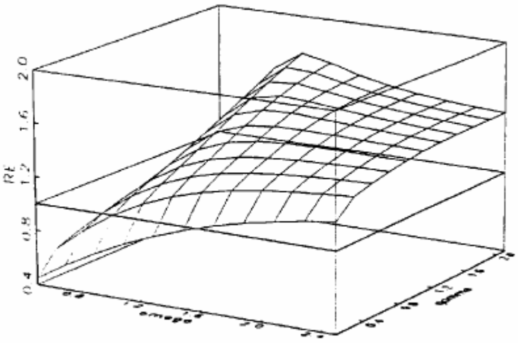


FIG. 4. *Relative efficiency RE of $\hat{m}_C$ and $\hat{m}_E$, in dependence from design density f_ω and regression curve m_γ; see Example 2.*

equidistant with respect to the ω-scale of Figure 4.

Figure 4 provides a similar impression about the relative behavior of the convolution and the evaluation estimator as we saw in Example 1. Again, the convolution estimator is preferable for regression curves with low slope (γ small) and rather concentrated design (ω small). The region where the convolution estimator is superior seems to be greater here. But note that values $\omega \leq 0.4$ correspond to a rather peaky design density.

3. COMPUTATIONALLY FAST ESTIMATORS

We have motivated the need for fast smoothing techniques. One possibility to speed up computation is to use weighted averaging of rounded points (WARPing). This technique is based on the following three steps: discretize the data, generate kernel weights and convolute the binned data with the kernel.

Regression data are discretized by counting the X observations that fall into bins $[(j-1/2)\delta, (j+1/2)\delta)$, where δ denotes the (small) bandwidth and j varies over all integers. Additionally, one records the sum of the response Y's in these bins and maintains a pointer structure to nonempty bins. We describe this technique here only for the evaluation estimator, it could of course be applied also to the $\hat{m}_C$ estimator.

WARPing is designed for kernels with compact support. Let us assume that K is supported on $[-1,1]$. Define the index function

$$(3.1)\qquad \iota(x) = j \Leftrightarrow x \in \left[(j-\tfrac{1}{2})\delta, (j+\tfrac{1}{2})\delta\right),$$

which returns the index of the small bin that x belongs to. Then the WARPing approximation to the evaluation estimator $\hat{m}_E$ is

$$\hat{m}_{M,K}(x) = \frac{\sum_{i=1}^{n} K((\iota(x)-\iota(X_i))/M)Y_i}{\sum_{i=1}^{n} K((\iota(x)-\iota(X_i))/M)};$$

here the integer $M \approx h\delta^{-1}$ stands for the bandwidth of the discretized kernel. An easy recalculation leads to

$$(3.2)\quad \hat{m}_{M,K}(x) = \frac{\sum_{l=1-M}^{M-1} K(l/M)Y_{\bullet\iota(x)+l}}{\sum_{l=1-M}^{M-1} K(l/M)n_{\iota(x)+l}},$$

where n_j and $Y_{\bullet j}$ denote the number of X's that fall into bin j and the sum of the corresponding Y's, respectively.

Formula (3.2) shows that essentially the problem of varying h depends now only on the number of bins, which is usually orders of magnitude smaller than the sample size n.

Thus, the above mentioned iterations and successive calls of kernel smoothing subroutines is performed much faster. Suppose we want to estimate m at N points. The evaluation kernel estimator requires $O(nN)$ operations for a kernel with noncompact support like the Gaussian kernel. For a kernel with compact support, this numerical effort is reduced slightly to $O(nhN)$. For the WARPing

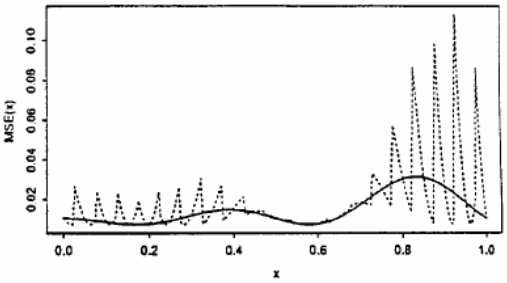


FIG. 5. *Leading term of the MSE of $\hat{m}_E$ (solid line) and of the WARPing step function $\hat{m}_{M,K}$. Underlying model: $m(x) = x\sin(2\pi x) + 3x$, uniform design, $\sigma^2 = 0.25$. Parameters: $h = 0.25$, $M = 5$, $n = 100$, quartic kernel.*

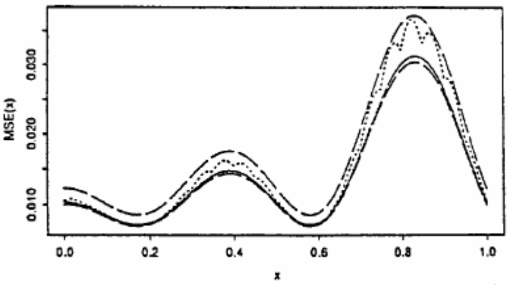
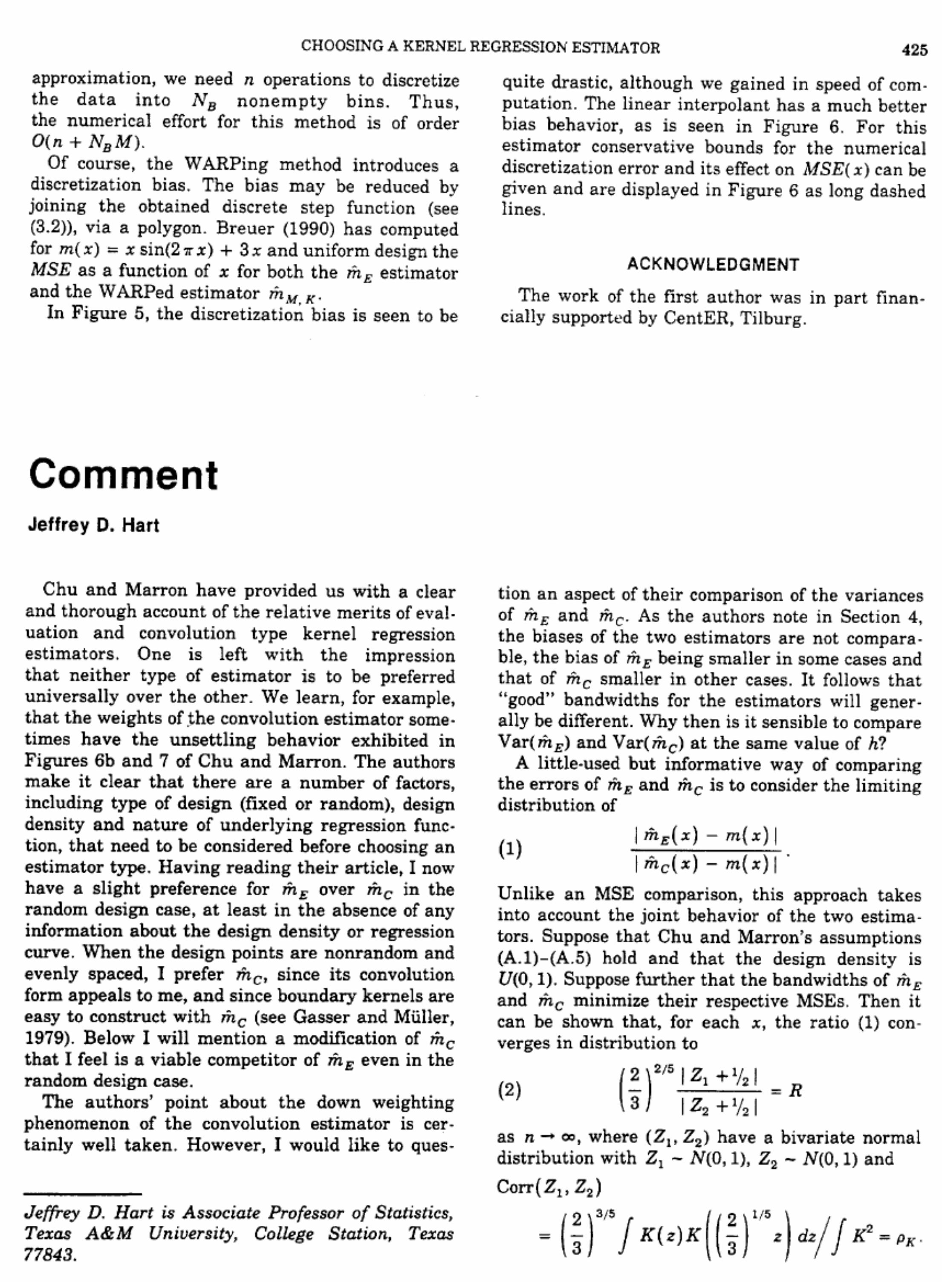


FIG. 6. *Leading term of the MSE of $\hat{m}_E$ (solid line) and of the corresponding WARPing polygon function (dotted line), with conservative bounds for the latter (dashed line). The same model and parameters as in Figure 5.*



approximation, we need n operations to discretize the data into N_B nonempty bins. Thus, the numerical effort for this method is of order $O(n + N_B M)$.

Of course, the WARPing method introduces a discretization bias. The bias may be reduced by joining the obtained discrete step function (see (3.2)), via a polygon. Breuer (1990) has computed for $m(x) = x \sin(2\pi x) + 3x$ and uniform design the MSE as a function of x for both the $\hat{m}_E$ estimator and the WARPed estimator $\hat{m}_{M,K}$.

In Figure 5, the discretization bias is seen to be quite drastic, although we gained in speed of computation. The linear interpolant has a much better bias behavior, as is seen in Figure 6. For this estimator conservative bounds for the numerical discretization error and its effect on $MSE(x)$ can be given and are displayed in Figure 6 as long dashed lines.

ACKNOWLEDGMENT

The work of the first author was in part financially supported by CentER, Tilburg.

Comment

Jeffrey D. Hart

Chu and Marron have provided us with a clear and thorough account of the relative merits of evaluation and convolution type kernel regression estimators. One is left with the impression that neither type of estimator is to be preferred universally over the other. We learn, for example, that the weights of the convolution estimator sometimes have the unsettling behavior exhibited in Figures 6b and 7 of Chu and Marron. The authors make it clear that there are a number of factors, including type of design (fixed or random), design density and nature of underlying regression function, that need to be considered before choosing an estimator type. Having reading their article, I now have a slight preference for $\hat{m}_E$ over $\hat{m}_C$ in the random design case, at least in the absence of any information about the design density or regression curve. When the design points are nonrandom and evenly spaced, I prefer $\hat{m}_C$, since its convolution form appeals to me, and since boundary kernels are easy to construct with $\hat{m}_C$ (see Gasser and Müller, 1979). Below I will mention a modification of $\hat{m}_C$ that I feel is a viable competitor of $\hat{m}_E$ even in the random design case.

The authors' point about the down weighting phenomenon of the convolution estimator is certainly well taken. However, I would like to question an aspect of their comparison of the variances of $\hat{m}_E$ and $\hat{m}_C$. As the authors note in Section 4, the biases of the two estimators are not comparable, the bias of $\hat{m}_E$ being smaller in some cases and that of $\hat{m}_C$ smaller in other cases. It follows that "good" bandwidths for the estimators will generally be different. Why then is it sensible to compare $\mathrm{Var}(\hat{m}_E)$ and $\mathrm{Var}(\hat{m}_C)$ at the same value of h?

A little-used but informative way of comparing the errors of $\hat{m}_E$ and $\hat{m}_C$ is to consider the limiting distribution of

$$(1) \qquad \frac{|\hat{m}_E(x) - m(x)|}{|\hat{m}_C(x) - m(x)|}.$$

Unlike an MSE comparison, this approach takes into account the joint behavior of the two estimators. Suppose that Chu and Marron's assumptions (A.1)–(A.5) hold and that the design density is $U(0, 1)$. Suppose further that the bandwidths of $\hat{m}_E$ and $\hat{m}_C$ minimize their respective MSEs. Then it can be shown that, for each x, the ratio (1) converges in distribution to

$$(2) \qquad \left(\frac{2}{3}\right)^{2/5} \frac{|Z_1 + 1/2|}{|Z_2 + 1/2|} = R$$

as $n \to \infty$, where (Z_1, Z_2) have a bivariate normal distribution with $Z_1 \sim N(0, 1)$, $Z_2 \sim N(0, 1)$ and

$$\mathrm{Corr}(Z_1, Z_2) = \left(\frac{2}{3}\right)^{3/5} \int K(z) K\left(\left(\frac{2}{3}\right)^{1/5} z\right) dz \bigg/ \int K^2 = \rho_K.$$

Jeffrey D. Hart is Associate Professor of Statistics, Texas A&M University, College Station, Texas 77843.

"Bootstrapping and Related Techniques", ed. Jöckel, K.H., Rothe, G. and Sendler, W., Springer Lecture Notes in Economics and Mathematical Systems, 376, 63-70

BOOTSTRAP CONFIDENCE BANDS

Wolfgang HÄRDLE
C.O.R.E.
Voie du Roman Pays 34, B-1348 Louvain-la-Neuve

Michael Nussbaum
Karl-Weierstrass-Institut für Mathematik
Mohrenstr. 39
D-1086 Berlin

Abstract

Bootstrap confidence bands are constructed for nonparametric regression. Resampling is based on a suitably estimated residual distribution. The procedure is called the *Wild Bootstrap*. The method is to construct first a fine grid of error bars with simultaneous coverage probability. Second the end-points of these error bars are joined via polygon pieces or parabolae using assumptions on the local curvature of the regression curve.

1. Motivation

Nonparametric regression smoothing is a flexible method for estimation of mean curves. Since this technique makes no structural assumptions on the underlying curve, it is very important to have a device for understanding when observed features are significant. An often asked question in this context is whether or not an observed peak or valley is actually a feature of the underlying regression function or is only an artifact of the observational noise. For such issues confidence bands should be used.

This paper proposes and analyzes a method of obtaining confidence bands based on simultaneous error bars at a grid of points. The method is simple to implement and relies on local smoothness of the regression curve. The construction is based on a residual resampling technique which models the conditional error distribution and also takes the bias properly into account.

For an understanding of these ideas, consider Figure 1. Figure 1a shows a scatter plot of the expenditure for potatoes as a function of income for the year 1973, from the Family Expenditure Survey (1968-1983). Figure 1b shows a nonparametric regression estimate which was obtained by smoothing the point cloud, using the kernel algorithm described in Section 2. As a means of understanding the variability in the kernel smooth, Figure 1b also shows error bars, constructed by the Wild Bootstrap method proposed in Härdle and Marron (1990). These bars are estimated simultaneous 80 % confidence intervals. Note that the error bars are longer on the right hand side, which reflects the fact that there are fewer observations there, and hence more uncertainty in the curve estimate. The error bars are also asymmetric in particular at points with high curvature which reflects the correct centering of the bars by a bias term. We propose a method of joining these error bars in order to obtain a bootstrap confidence band.

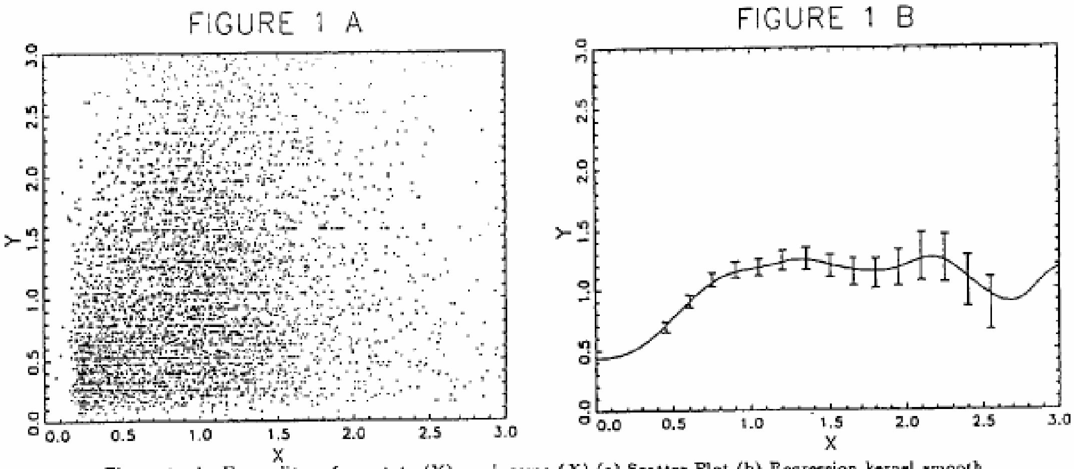


Figure 1 a,b. Expenditure for potato (Y) vs. income (X) (a) Scatter Plot (b) Regression kernel smooth (quartic kernel with band with h=0.3) and errors bars.

Clearly there is a need for confidence bands in all applications of nonparametric regression. Hall and Titterington (1986) constructed a confidence band for calibration of radio carbon dating assuming Normal errors. Knafl, Sacks and Ylvisaker (1984) derived uniform variability bands under the same error structure.

Our approach is based on resampling from estimated residuals. This form of bootstrapping preserves the error structure in the data and guarantees that the bootstrap observations have errors with mean zero. There are two main advantages to this approach. First it correctly accounts for the bias and hence does not require additional estimation of bias or the use of a sub-optimal (under smoothed) curve estimator. Second, no assumption of homoscedasticity is required, the method automatically adapts to different residual variances at different locations.

In Section 2 we give a technical introduction into simultaneous error bars constructed via the Wild Bootstrap. In Section 3 we consider bootstrap confidence bands. Proofs are given in the forthcoming paper by Härdle and Marron (1990).

2. Simultaneous error bars via the Wild Bootstrap

Stochastic design nonparametric regression is based on iid. observations $\{(X_i, Y_i)\}_{i=1}^n \in \mathbb{R}^{d+1}$ and the goal is to estimate $m(x) = E(Y|X = x) : \mathbb{R}^d \to \mathbb{R}$. The form of the kernel regression estimator we consider is

$$\hat{m}_h(x) = n^{-1}\sum_{i=1}^{n} K_h(x - X_i)Y_i/\hat{f}_h(x) \tag{2.1}$$

where

$$\hat{f}_h(x) = n^{-1}\sum_{i=1}^{n} K_h(x - X_i) \tag{2.2}$$

and where $K_h(u) = h^{-d}K(u/h)$ is a kernel weight function with bandwidth h. All results of this paper are stated in terms of this estimator, although the essential ideas clearly extend to other types of kernel estimators such as those of Gasser and Müller (1984) and also to other regression estimators, such as spline methods, as discussed in Eubank (1988).

One approach to the problem of finding simultaneous error bars would be to work with limiting normal distributions of the estimator at the grid points. However the joint distribution of the estimator at close gridpoints has substantial positive correlation, which makes the derivation of joint normal theory confidence intervals nontrivial. In fact, they essentially should be done by simulation methods. Since simulation methods are needed anyway, it seems more economical to use direct resampling.

While bootstrap methods are well known tools for assessing variability, more care must be taken to properly account for the type of bias encountered in nonparametric curve estimation. In particular, the naive bootstrap approach, of resampling from the pairs $\{(X_i, Y_i) : i = 1, ..., n\}$ is inappropriate because the bootstrap bias will be 0, see Härdle and Mammen (CORE DP 9049). Our approach to this problem is to first use the estimated residual

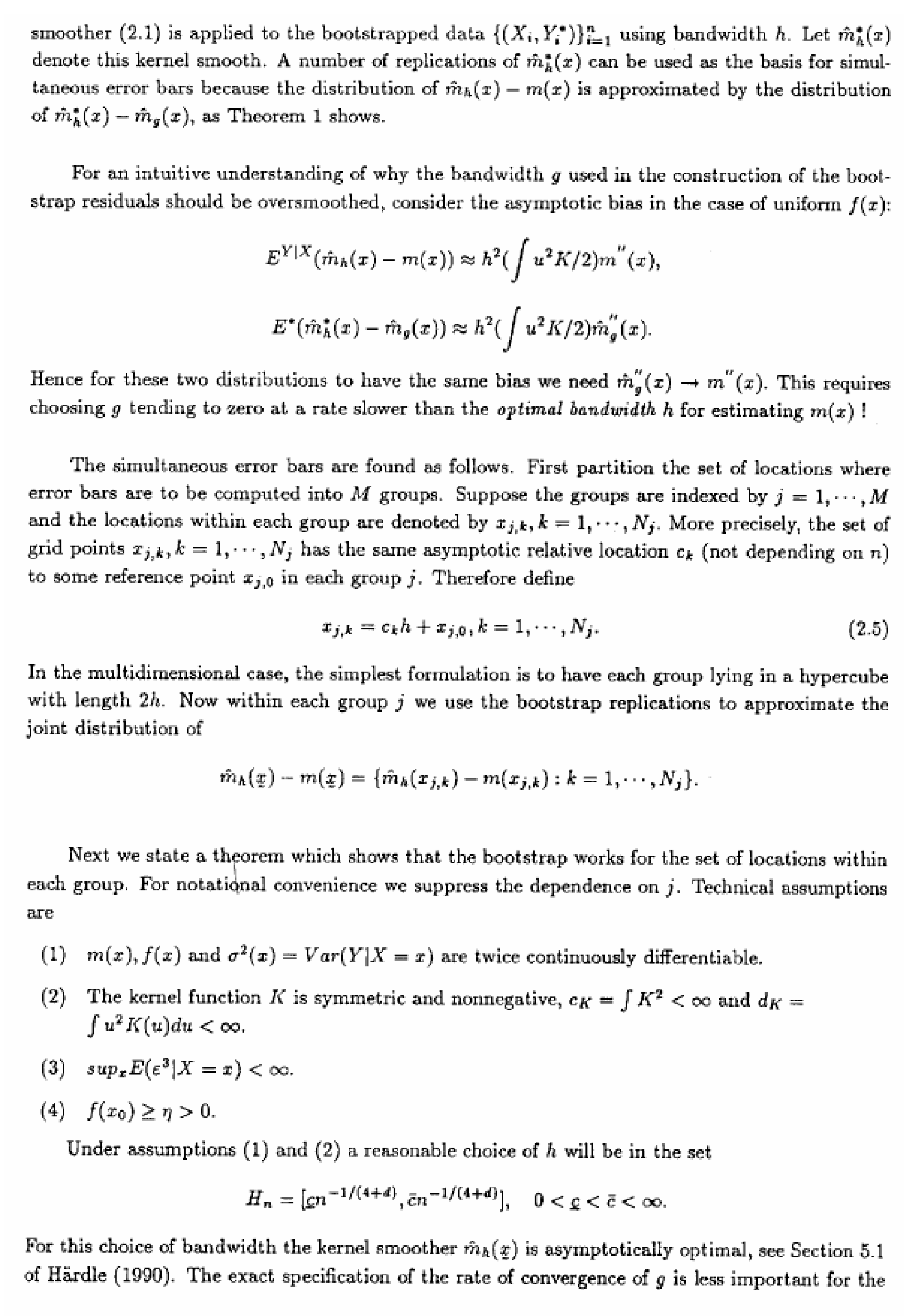
$$\hat{\varepsilon}_i = Y_i - \hat{m}_h(X_i). \tag{2.3}$$

To better retain the conditional distributional characteristics of the estimate, we do not resample from the entire set of residuals, as in Härdle and Bowman (1988). We use the idea of *wild bootstrapping*, as used in Härdle and Marron (1990) where each bootstrap residual is drawn from the two point distribution which has mean zero, variance equal to the square of the residual, and third moment equal to the cube of the residual. In particular define a new random variable ε_i^* having a two point distribution $\hat{G}_i$, where $\hat{G}_i = \gamma\delta_a + (1-\gamma)\delta_b$ is defined through the three parameters a, b, γ, and where δ_a, δ_b denote point measures at a, b respectively. Some algebra reveals that the parameters a, b, γ at each location X_i are given by $a = \hat{\varepsilon}_i(1-\sqrt{5})/2$, $b = \hat{\varepsilon}_i(1+\sqrt{5})/2$ and $\gamma = (5+\sqrt{5})/10$. These parameters ensure that $E\varepsilon^* = 0, E\varepsilon^{*2} = \hat{\varepsilon}_i^2$ and $E\varepsilon^{*3} = \hat{\varepsilon}_i^3$. The above formulation of the wild bootstrap, based on a two point distribution, is only one possible approach. Other distributions such as mixtures of normals could be considered as well.

After resampling, bootstrap data

$$Y_i^* = \hat{m}_g(X_i) + \varepsilon_i^* \tag{2.4}$$

are defined, where $\hat{m}_g(x)$ is a kernel estimator with bandwidth g taken to be larger than h (a heuristic explanation of why it is essential to oversmooth g is given below). Then the kernel



smoother (2.1) is applied to the bootstrapped data $\{(X_i, Y_i^*)\}_{i=1}^n$ using bandwidth h. Let $\hat{m}_h^*(x)$ denote this kernel smooth. A number of replications of $\hat{m}_h^*(x)$ can be used as the basis for simultaneous error bars because the distribution of $\hat{m}_h(x) - m(x)$ is approximated by the distribution of $\hat{m}_h^*(x) - \hat{m}_g(x)$, as Theorem 1 shows.

For an intuitive understanding of why the bandwidth g used in the construction of the bootstrap residuals should be oversmoothed, consider the asymptotic bias in the case of uniform $f(x)$:

$$E^{Y|X}(\hat{m}_h(x) - m(x)) \approx h^2(\int u^2 K/2) m''(x),$$

$$E^*(\hat{m}_h^*(x) - \hat{m}_g(x)) \approx h^2(\int u^2 K/2) \hat{m}_g''(x).$$

Hence for these two distributions to have the same bias we need $\hat{m}_g''(x) \to m''(x)$. This requires choosing g tending to zero at a rate slower than the *optimal bandwidth* h for estimating $m(x)$!

The simultaneous error bars are found as follows. First partition the set of locations where error bars are to be computed into M groups. Suppose the groups are indexed by $j = 1, \cdots, M$ and the locations within each group are denoted by $x_{j,k}, k = 1, \cdots, N_j$. More precisely, the set of grid points $x_{j,k}, k = 1, \cdots, N_j$ has the same asymptotic relative location c_k (not depending on n) to some reference point $x_{j,0}$ in each group j. Therefore define

$$x_{j,k} = c_k h + x_{j,0}, k = 1, \cdots, N_j. \tag{2.5}$$

In the multidimensional case, the simplest formulation is to have each group lying in a hypercube with length $2h$. Now within each group j we use the bootstrap replications to approximate the joint distribution of

$$\hat{m}_h(\underset{\sim}{x}) - m(\underset{\sim}{x}) = \{\hat{m}_h(x_{j,k}) - m(x_{j,k}) : k = 1, \cdots, N_j\}.$$

Next we state a theorem which shows that the bootstrap works for the set of locations within each group. For notational convenience we suppress the dependence on j. Technical assumptions are

(1) $m(x), f(x)$ and $\sigma^2(x) = Var(Y|X = x)$ are twice continuously differentiable.

(2) The kernel function K is symmetric and nonnegative, $c_K = \int K^2 < \infty$ and $d_K = \int u^2 K(u) du < \infty$.

(3) $sup_x E(\epsilon^3 | X = x) < \infty$.

(4) $f(x_0) \geq \eta > 0$.

Under assumptions (1) and (2) a reasonable choice of h will be in the set

$$H_n = [\underline{c} n^{-1/(4+d)}, \bar{c} n^{-1/(4+d)}], \quad 0 < \underline{c} < \bar{c} < \infty.$$

For this choice of bandwidth the kernel smoother $\hat{m}_h(\underset{\sim}{x})$ is asymptotically optimal, see Section 5.1 of Härdle (1990). The exact specification of the rate of convergence of g is less important for the validity of the following theorem, although it must tend to zero at a rate slower than h. Hence it is assumed that g is chosen from the set

$$\cdot G_n = [n^{-1/(4+d)+\delta}, n^{-\delta}], \delta > 0.$$

A fine tuning of the choice of bandwidth g is presented in Theorem 3 of Härdle and Marron (1990).

Theorem 1. *Given the assumptions above, we have along almost all sample sequences and for all $\underset{\sim}{z} \in I\!R^N$*

$$sup_{h\in H_n} sup_{g\in G_n} |P^{Y|X}\{\sqrt{nh^d}[\hat{m}_h(\underset{\sim}{x}) - m(\underset{\sim}{x})] < \underset{\sim}{z}\}$$
$$- P^*\{\sqrt{nh^d}[\hat{m}^*_h(\underset{\sim}{x}) - \hat{m}_g(\underset{\sim}{x})] < \underset{\sim}{z}\}| \to 0.$$

The reason that uniform convergence (in h and g) in the above result is important, is that it ensures that the result still holds when h or g are replaced by random data driven bandwidths. For each group j this joint distribution is used to obtain simultaneous $1 - \alpha/M$ error bars that are simultaneous over $k = 1, \cdots, N_j$ as follows. Let $\beta > 0$ denote a generic size for individual confidence intervals. Our goal is to choose β so that the resulting simultaneous size is $1 - \alpha/M$. For each $x_{j,k}, k = 1, \cdots, N_j$ define the interval $I_{j,k}(\beta)$ to have endpoints which are the $\beta/2$ and the $1 - \beta/2$ quantiles of the $(\hat{m}^*_h(x_{j,k}) - \hat{m}_g(x_{j,k}))$ distribution. Then define α_β to be the empirical *simultaneous* size of the β confidence intervals, i.e. the proportion of curves which lie outside at least one of the intervals in the group j. Next find the value of β , denoted by β_j, which makes $\alpha_{\beta_j} = \alpha/M$. The resulting β_j intervals within each group j will then have confidence coefficient $1 - \alpha/M$. Hence by the Bonferroni bound the entire collection of intervals $I_{j,k}(\beta_j), k = 1, \cdots, N_j, j = 1, \cdots, M$ will simultaneously contain at least $1 - \alpha$ of the distribution of $\hat{m}^*_h(x_{j,k})$ about $\hat{m}_g(x_{j,k})$. Thus the intervals $I_{j,k}(\beta_j) - \hat{m}_g(x_{j,k}) + \hat{m}_h(x_{j,k})$ will be simultaneous confidence intervals with confidence coefficient at least $1 - \alpha$. The result of this process is summarized as

Theorem 2. *Define M groups of locations $x_{j,k}, k = 1, \cdots, N_j, j = 1, \cdots, M$ where simultaneous error bars are to be established. Compute uniform confidence intervals for each group. Correct the significance level across groups by the Bonferroni method. Then the bootstrap error bars establish asymptotic simultaneous confidence intervals, i.e.*

$$lim_{n\to\infty} P\{m(x_{j,k}) \in I_{j,k}(\beta_j) - \hat{m}_g(x_{j,k}) + \hat{m}_h(x_{j,k}), \tag{2.6}$$

$$k = 1, \cdots, N_j, j = 1, \cdots, M\} \geq 1 - \alpha$$

As a practical method for finding β_j for each group j we suggest the following "halving" approach (also called a bisection search). In particular, first try $\beta = \alpha/2M$, and calculate α_β. If the result is more than α/M, then try $\beta = \alpha/4M$, otherwise next try $\beta = 3\alpha/4M$. Continue this halving approach until neighboring (since only finitely many bootstrap replications are made, there is only a finite grid of possible β's available) values β_* and β^* are found so that $\alpha_{\beta_*} < \alpha/M < \alpha_{\beta^*}$. Finally take a weighted average of the β_* and the β^* intervals where the weights are $(\alpha_{\beta^*} - \alpha/M)/(\alpha_{\beta^*} - \alpha_{\beta_*})$ and $(\alpha/M - \alpha_{\beta_*})/(\alpha_{\beta^*} - \alpha_{\beta_*})$ respectively.

All of the results in this paper have been stated in terms of the so-called stochastic design model where the regressors X are thought of as realizations of random variables. Since these results are all conditional on $X_1, \cdots, X_n$ our ideas carry over immediately to the case where the $X's$ are fixed and chosen by the experimenter.

In the case of binary regression (dose-response curves, Cox (1970, p.8)), where the response variable Y takes on only the values 0 or 1, there are more natural ways of obtaining bootstrap confidence intervals than those described here. A direct application of our method would give bootstrapped data Y^* which take on values different from 0 and 1. A seemingly more natural approach would be to bootstrap from a Bernoulli distribution with parameter $\hat{m}_g(X_i)$. It is interesting to know how fast the convergence in Theorem 1 takes place. This has been analysed via Berry-Esseen bounds as in Cao-Abad (1989). More precisely Cao-Abad shows that the convergence in Theorem 1 is of order $n^{-2/5}$. Härdle, Huet and Jolivet (1990) consider Edgeworth expansions of the studentized statistic $\sqrt{nh}(\hat{m}_h(x) - m(x))/\widehat{var}x$ and as in Hall (1990) find slightly better rates.

3. Bootstrap Confidence Bands

Once simultaneous error bars have been constructed on a grid of points the extension to uniform confidence bands $[\underline{c}(x), \overline{c}(x)]$ such that

$$P\{\underline{c}(x) \leq m(x) \leq \overline{c}(x) \text{ for all } x\} \approx 1 - \alpha$$

can be done in several ways. One approach is based on bounds on the first derivative $m'(x)$. Let us consider for simplicity just two points x_1 and x_2 of the set of grid points. From section 2 we know that with probablity $(1-\alpha)$ the true curve $m(x)$ lies at these points in $[\underline{c}(x_j), \overline{c}(x_j)]$, $j = 1, 2$. The exact for of $\underline{c}, \overline{c}$ is given in Theorem 2. By the mean value theorem we know that for some point $\xi \in [x_1, x_2]$: $m(x_2) - m(x_1) = m'(\xi)(x_2 - x_1)$ thus it is natural to use bounds on the first derivative. Suppose that

$$\underline{\delta} \leq m'(x) \leq \overline{\delta}, x_1 \leq x \leq x_2$$

then the two error bars can be joined with two line segments to ensure an overall upper bound between x_1 and x_2. These two lines segments are

$$x \rightarrow \overline{c}_1(x) = \overline{\delta}(x - x_1) + \overline{c}(x_1)$$
$$x \rightarrow \overline{c}_2(x) = \underline{\delta}(x - x_2) + \overline{c}(x_2).$$

In a similar way the lower bound can be constructed by

$$x \rightarrow -\underline{c}_1(x) = \underline{\delta}(x - x_1) + \underline{c}(x_1)$$
$$x \rightarrow -\underline{c}_2(x) = \overline{\delta}(x - x_2) + \underline{c}(x_2).$$

Thus the desired confidence band is

$$\underline{c}(x) = \underline{c}_1(x) \vee \underline{c}_2(x) \text{ and } \overline{c}(x) = \overline{c}_1(x) \wedge \overline{c}_2(x).$$

Obviously this set of four lines contains the true curve with probability $(1-\alpha)$. The construction can be extended to an arbitrary set of grid points provided we have constructed error bars of simultaneous coverage probablity. Thus the confidence band is a sequence of connected hexagons.

Another approach we propose is based on bounds on the second derivative. This will lead to parabolae joining the error bars as indicated in Figure 2.

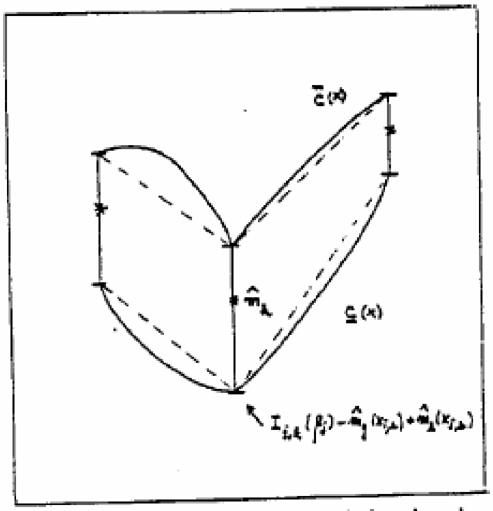


Figure 2. Bootstrap confidence bands based on bounds on the second derivative. The parabolae are different since the bound in the adjacent intervals may be different.

To get some insight into this construction consider for simplicity error bars in $x_1 = 0$, $x_2 = 1$. Let us construct the upper band. If the error bars have lengths $c(x_1)$ and $c(x_2)$ we can subtract the line joining the bars since it has second derivative equal to zero. So we may assume that the upper band $\bar{c}(x), 0 \le x \le 1$ passes zero at x_1 and x_2. Assume now that $|m''(x)| \le L$ in this interval and that for some z, $m(z) \ge t > 0$. By the mean value theorem we find $\xi_j, j = 1,2$ such that

$$m'(\xi_1) = \frac{t}{z}, m'(\xi_2) = -\frac{t}{1-z}.$$

By yet another application of the mean value theorem, there is a ξ_3 :

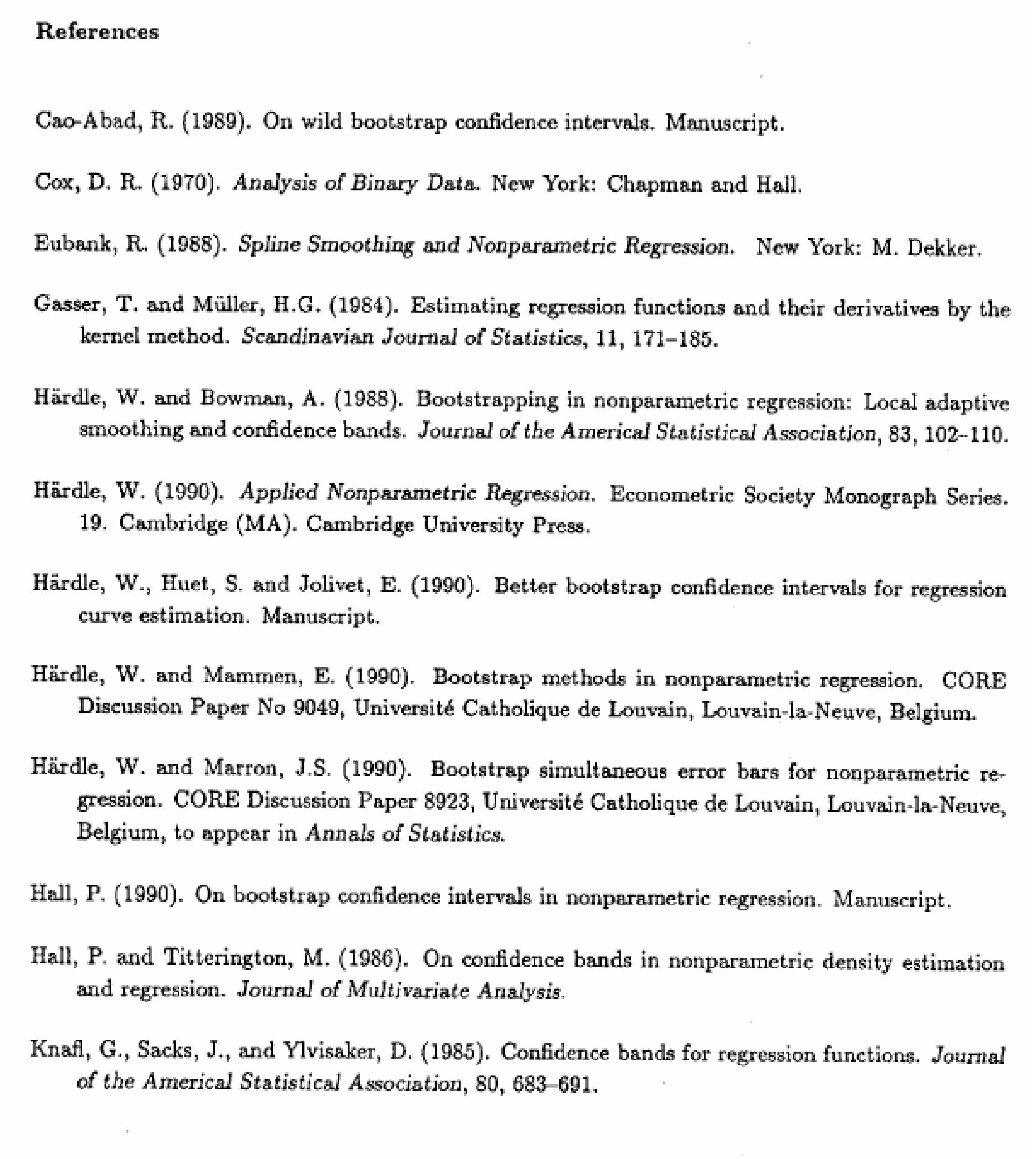
$$\begin{aligned} m''(\xi_3) &= \frac{\frac{t}{z} - (-\frac{t}{1-z})}{\xi_2 - \xi_1} \\ &\ge \frac{t}{z} + \frac{t}{1-z} \\ &= \frac{t}{z(1-z)}. \end{aligned}$$

Hence, using the bound on $m''(x)$ we have

$$t \le L(z(1-z)).$$

This means that $m(x)$ must be below the parabola $(1-x)\bar{c}(x_1) + x\bar{c}(x_2) + L(x(1-x))$.

A different approach could be based on global bounds of the arclength of the curve between adjacent error bars. The arclength between x_1 and x_2 is $\int_{x_1}^{x_2} \sqrt{1+(m'(x))^2}dx$, thus a bound on this quantity would give us another possibility to construct confidence bands. The geometric location of points with a fixed distance to two foci is an ellipse. The obtained confidence band would thus be a nice esthetic sequence of intersecting ellipses.



References

Cao-Abad, R. (1989). On wild bootstrap confidence intervals. Manuscript.

Cox, D. R. (1970). *Analysis of Binary Data.* New York: Chapman and Hall.

Eubank, R. (1988). *Spline Smoothing and Nonparametric Regression.* New York: M. Dekker.

Gasser, T. and Müller, H.G. (1984). Estimating regression functions and their derivatives by the kernel method. *Scandinavian Journal of Statistics*, 11, 171–185.

Härdle, W. and Bowman, A. (1988). Bootstrapping in nonparametric regression: Local adaptive smoothing and confidence bands. *Journal of the Americal Statistical Association*, 83, 102–110.

Härdle, W. (1990). *Applied Nonparametric Regression.* Econometric Society Monograph Series. 19. Cambridge (MA). Cambridge University Press.

Härdle, W., Huet, S. and Jolivet, E. (1990). Better bootstrap confidence intervals for regression curve estimation. Manuscript.

Härdle, W. and Mammen, E. (1990). Bootstrap methods in nonparametric regression. CORE Discussion Paper No 9049, Université Catholique de Louvain, Louvain-la-Neuve, Belgium.

Härdle, W. and Marron, J.S. (1990). Bootstrap simultaneous error bars for nonparametric regression. CORE Discussion Paper 8923, Université Catholique de Louvain, Louvain-la-Neuve, Belgium, to appear in *Annals of Statistics.*

Hall, P. (1990). On bootstrap confidence intervals in nonparametric regression. Manuscript.

Hall, P. and Titterington, M. (1986). On confidence bands in nonparametric density estimation and regression. *Journal of Multivariate Analysis.*

Knafl, G., Sacks, J., and Ylvisaker, D. (1985). Confidence bands for regression functions. *Journal of the Americal Statistical Association*, 80, 683–691.

Journal of the American Statistical Association, 86, 333-335

REMARKS ON SLICED INVERSE REGRESSION

W. Härdle
CORE, Université Catholique de Louvain

A.B. Tsybakov
Institute for Problems of Information Transmission
Academy of Sciences, U.S.S.R.

April 1990

Abstract

This is a discussion of sliced inverse regression.

ISSN-0771 3894

(1991) Härdle, W. and Tsybakov, A.B. Remarks on Sliced Inverse Regression. Discussion of a paper by Ker-Chau Li

Journal of the American Statistical Association, 86, 333-335

The paper by Professor Ker-Chau Li proposes a new and very useful approach to dimensionality reduction in multivariate nonparametric regression. The advantage of this approach as compared to others is the exceptional simplicity both of the idea and of the computational tools. We suppose that this would give rise to a wide implementation of slicing inverse regression (SIR).

As many simple ideas, of course also Sliced Inverse Regression will have its pitfalls in "nonsimple" situations. In particular, SIR depends very much on the probability structure of the $\mathbf{x}$-variables described by

For any b in $\mathbb{R}^p$, the conditional expectation $E(b\mathbf{x} \mid \beta_1\mathbf{x}, \ldots, \beta_K\mathbf{x})$ is linear in $\beta_1\mathbf{x}, \ldots, \beta_{\mathbf{x}}\mathbf{x}$; that is, for some constants $c_0, c_1, \ldots, c_K$,

$$E(b\mathbf{x} \mid \beta_1\mathbf{x}, \ldots, \beta_K\mathbf{x}) = c_0 + c_1\beta_1\mathbf{x} + \cdots + c_K\beta_K\mathbf{x}. \tag{3.1}$$

A "nonsimple" situation might be where the distribution of $\mathbf{x}$ is a mixture of two normal distributions or has a more complicated nonelliptical structure. In this case a nonparametric technique based on estimating the multivariate density of $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$ might be reasonable to check the assumption (3.1). We discuss later an approach based on this (more complicated) technique.

There are at least two questions that are important for a practioner: how to choose the number of principal directions K and how to choose the number of slices H? These questions are addressed to some extent but we feel that they deserve some more comments.

It is said that the root n consistency property in estimation of directions holds no matter how H is chosen and that it even holds when each slice contains only two observations. This is probably somewhat misleading. If H can be chosen arbitrarily, then it seems possible to use the simplest estimate, i.e., to put $H = 1$. But this is, of course, bizarre since in this case $p_h = 1$, and the estimate will be close to $m_h = E(Z) = 0$. When H increases the number of nontrivial eigenvectors of the matrix V will also increase. Although, it will not be evident for what H all the K principal eigenvectors are present. This could suggest that H should rather be chosen large to make sure that we catch all the principal directions. Thus one might incline to the other extreme, i.e., to choosing only two observations per slice. To understand this extreme, let us think of one observation per slice, then $\widehat{V} = \sum_{i=1}^{n} \tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^T$. Thus the principal directions are chosen from the covariance structure of $\mathbf{x}$ as in principal components analysis. Thus, between these two extremes of SIR, there is a lot of freedom which makes alternative approaches

(1991) Härdle, W. and Tsybakov, A.B. Remarks on Sliced Inverse Regression.
Discussion of a paper by Ker-Chau Li

interesting. One of them is based on a different identification method of the e.d.r. space, the second is based on average derivative estimation (ADE). Finally, we propose a nonparametric version of factor analysis.

Let us consider instead of V the matrix

$$B = E_Y[E(\mathbf{x} \mid y)E(\mathbf{x}^T \mid y)]$$

(assume here that $\mathbf{x}$ is already standardized). Elements of B can be expressed as

$$b_{jk} = \int m_j(y)m_k(y)F(dy)$$

where $m_j(y)$ is the regression function of y on the jth component of $\mathbf{x}$, and F is the marginal distribution of y. To estimate b_{jk}, replace F by the empirical distribution F_n, and m_j, m_k by the nonparametric regression estimates $\widehat{m}_j$, $\widehat{m}_k$. Thus

$$\widehat{b}_{jk} = \int \widehat{m}_j(y)\widehat{m}_k(y)F_n(dy) = \frac{1}{n}\sum_{i=1}^{n} \widehat{m}_j(y_i)\widehat{m}_k(y_i).$$

The functions $\widehat{m}_j$, $\widehat{m}_k$ may be kernel, orthogonal series or any other estimates. If $\widehat{m}$ is regressogram, then we get something very similar to SIR, namely,

$$\widehat{B} = \frac{1}{n}\sum_{i=1}^{n} \widehat{m}(y_i)\widehat{m}^T(y_i),$$

where

$$\widehat{m}(y) = \frac{1}{np_h}\sum_{h=1}^{H}\sum_{s=1}^{n} I\{y_s \in I_h, y \in I_h\}\tilde{\mathbf{x}}_s.$$

This estimate will of course have a bias decreasing with $H \to \infty$. Similar functionals like the average derivative have a variance proportional to $1/n$. We suspect therefore that a careful choice of H will yield a $\sqrt{n}$-convergence of $\widehat{B}$ to B.

All the eigenvectors of B that correspond to nonzero eigenvalues are contained in the e.d.r. space. In fact, it follows from Corollary 3.1 that

$$E(\mathbf{x} \mid y) = c_1(y)\beta_1 + \cdots + c_K(y)\beta_K$$

where $c_j(y)$ are some functions. Therefore $B = \sum_{j,m=1}^{K} \tilde{c}_{jm}\beta_j\beta_m^T$ where $\tilde{c}_{jm} = E(c_j(y)c_m(y))$. Thus if b is not in e.d.r. space, i.e., $b \perp \{\beta_1, \ldots, \beta_K\}$, then $Bb = 0$.

In the simplest case of $K = 1$ one gets

$$B = \tilde{c}_{11}\beta_1\beta_1^T, \qquad \tilde{c}_{11} = E(c_1^2(y)).$$

Assume that β_1 is normalized so that $||\beta_1|| = 1$. Then β_1 is the eigenvector of B corresponding to the maximal eigenvalue $\tilde{c}_{11}$:

$$B\beta_1 = \tilde{c}_{11}\beta_1; \quad \tilde{c}_{11} \geq b^T Bb, \quad \forall\, b : ||b|| = 1.$$

Another approach first developed for the case $K = 1$ is ADE, see Härdle and Stoker (1989), Härdle, Hart, Marron, and Tsybakov (1990). The *average derivative* is defined by

$$\int \nabla m(\mathbf{x}) f_X(\mathbf{x}) d\mathbf{x}$$

where $\nabla m(x)$ is the gradient of the unknown regression function $m(\mathbf{x}) = E(Y \mid X = \mathbf{x})$ and $f_X(\mathbf{x})$ is the marginal density of $\mathbf{x}$. The average derivative can be estimated $\sqrt{n}$-consistently. Although all the previous work on ADE was concerned with the case of $K = 1$, its extension to the more general model $y = m(\beta_1^T\mathbf{x}, \ldots, \beta_K^T\mathbf{x}, \varepsilon)$ is straightforward. In fact, the average derivative is then

$$\begin{aligned} \mathcal{AD} &= E\left(\nabla_{\mathbf{x}} m(\beta_1^T\mathbf{x}, \ldots, \beta_K^T\mathbf{x}, \varepsilon)\right) \\ &= c_1\beta_1 + \cdots + c_K\beta_K, \end{aligned}$$

where

$$c_j = E\left(\frac{\partial}{\partial t} m(\beta_1^T\mathbf{x}, \ldots, \beta_{j-1}^T\mathbf{x}, t, \beta_{j+1}^T\mathbf{x}, \ldots, \beta_K^T\mathbf{x}, \varepsilon)\Big|_{t=\beta_j^T\mathbf{x}}\right).$$

Define the matrix $B_1 = \mathcal{AD} \cdot \mathcal{AD}^T$. This matrix is an analogue of B defined earlier since all the eigenvectors of B that correspond to nonzero eigenvalues are in the e.d.r. space. Thus, in the same way as earlier, we can choose the estimates $\widehat{\beta}_1, \ldots, \widehat{\beta}_K$ of principal directions as the first K eigenvectors of

$$\widehat{B}_1 = \widehat{\mathcal{AD}} \quad \widehat{\mathcal{AD}}^T,$$

where $\widehat{\mathcal{AD}}$ is an average derivative estimator.

The choice of the number of principal directions K can be addressed in at least three different ways:

(ii) the candidates for principal directions are known and ordered; the first K directions are principal; K must be estimated;

(i) the candidates for principal directions are known; the number K of principal directions and their positions are unknown; these directions must be estimated;

(iii) the candidates for principal directions are unknown, their number is also unknown.

Professor Ker-Chau Li proposes an interesting way of treating the problem in case (iii) for normally distributed $\mathbf{x}$. His approach is based on the correlation structure of $\mathbf{x}$ only. This can be viewed as an analogue to sequential hypothesis testing techniques in linear regression. However, the extension to the case of non-Gaussian $\mathbf{x}$ seems to be somewhat difficult.

Note that (i) is solved if one has a solution of (ii). Under (ii) we can assume in general that possible candidates for principal directions are all the coordinate axes. For example, this assumption is quite reasonable if one thinks of a nonparametric version of factor analysis. Thus, the unknown regression function $m(\mathbf{x})$, $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_p) \in \mathbb{R}^p$, is of the form

$$m(\mathbf{x}) = \sum_{K=1}^{K} g_{j_k}(\mathbf{x}_{j_k}), \quad j_k \in \{1, \dots, p\},$$

where $K < p$ is some integer, $K \geq 1$. the problem is to estimate the set $J = \{j_1, \dots, j_K\}$. Given a sample $(\mathbf{x}_1, Y_1), \dots, (\mathbf{x}_n, Y_n)$ define

$$r_{nj}(\mathbf{x}_j) = \frac{1}{nh_n} \sum_{i=1}^{n} Y_i \tilde{K}\left(\frac{\mathbf{x}_{ij} - \mathbf{x}_j}{h_n}\right),$$

$$f_{nj}(\mathbf{x}_j) = \frac{1}{nh_n} \sum_{i=1}^{n} \tilde{K}\left(\frac{\mathbf{x}_{ij} - \mathbf{x}_j}{h_n}\right).$$

Here f_{nj} is the kernel estimate of the marginal density f_j of jth component, $\mathbf{x}_{ij}$'s are the components of vectors $\mathbf{x}_i = (\mathbf{x}_{i1}, \dots, \mathbf{x}_{ip})$, $\tilde{K}$ is a kernel and $h_n > 0$ is a bandwidth. Consider the following procedure of estimating J.

1) Calculate the quantities

$$S_{nj} = \frac{1}{n} \sum_{i=1}^{n} r_{nj}^2(\mathbf{x}_{ij}), \quad j = 1, \dots, p.$$

2) Arrange S_{nj} in the decreasing order:

$$S_n^{(1)} \geq S_n^{(2)} \geq \cdots \geq S_n^{(p)}.$$

Let $(1)_n$ be the integer that equalts to j with maximal value $S_{nj} = S_n^{(1)}$, $(2)_n$ be the integer that equals to j with $S_{nj} = S_n^{(2)}$, etc. Thus

$$(K)_n = j \in \{1, \ldots, p\} : S_{nj} = S_n^{(K)}.$$

Without loss of generality assume that all $S_n^{(k)}$ are different (thus $(K)_n$ is uniquely defined). In particular we have

$$S_n^{(K)} = \frac{1}{n} \sum_{i=1}^{n} r^2_{n(K)_n}(\mathbf{x}_{i(K)_n}).$$

3) Choose K_n as the minimizer of the following statistic

$$K_n = \left[\arg\min_{K \leq p}(S_n^{(p)} + K b_n)\right] - 1,$$

where b_n is a sequence that tends to zero as $n \to \infty$ and $nb_n^2 \to \infty$.

The estimate of the set $\{j_1, \ldots, j_K\}$ is defined as $J_n = \{(1)_n, \ldots, (K_n)_n\}$, and the corresponding estimate of the regression function is

$$m_n(\mathbf{x}) = \sum_{K \in J_n} g_{nj_k}(\mathbf{x}_{j_k}),$$

where

$$g_{nj}(x_j) = \frac{r_{nj}(\mathbf{x}_j)}{f_{nj}(\mathbf{x}_j)}.$$

It can be proved that under suitable assumptions $P\{J_n = J\} \to 1$, $n \to \infty$ (Härdle and Tsybakov (1990)). Moreover, the estimate $m_n(\mathbf{x})$ is pointwise asymptotically normal and converges to $m(\mathbf{x})$ with the rate that is achievable for the case of univariate regression function estimation.

This idea of estimating "principal components" can be viewed as a modification of AIC-BIC criteria with the additional reordering of components according to some stochastic criterion. Note that instead of S_{nj}'s we could take for reordering any other data-dependent quantities that are asymptotically nonzero for "principal components" and are zero for negligible components.

REFERENCES

Härdle, W., Hart, J., Marron, J.S., and A.B. Tsybakov (1989). Bandwidth choice for average derivative estimation. *Journal of the American Statistical Association*, submitted.

Härdle, W. and T. Stoker (1989). Investigating smooth multiple regression by the method of average derivatives. *Journal of the American Statistical Association*, 84, 986–995.

Härdle, W. and A.B. Tsybakov (1990) How many terms should be added into an additive model? *Annals of Statistics*, submitted.

Journal of the Royal Statistical Society, 54, 37-38

On an Efficient Smoothing Parameter Selector Proposed By Hall And Johnstone

Wolfgang Härdle

C.O.R.E., Universite Catholique de Louvain

Byeong U. Park

C.O.R.E., Universite Catholique de Louvain
and Seoul National University, Seoul, KOREA

September 16, 1991

Abstract

The difficulty in selecting smoothing parameters is discussed. A proposed selector by Hall and Johnstone avoids the negative correlation with the desired smoothing parameter. This selector is compared with cross-validation and extended to the case of regression with non uniform covariates.

(1992) Park, B. and Härdle, W. Discussion of the paper by Hall and Johnstone, (September 1991)

Journal of the Royal Statistical Society, 54, 37-38

1. Efficient smoothing parameter selection

The difficulty of assessing accurate smoothing parameter selectors has long been underestimated. This is mainly due to the fact that in the last few years a large toolbox of data driven selectors has been developed. The size of this toolbox created an overoptimistic approach in using automated smoothing procedures: All these methods are asymptotically optimal, some are even root-n convergent, so why care about a specific one? The paper by Hall and Johnstone has given us a very precise quantification of the difficulties inherent in smoothing parameter selection and has shown us relative merits of different methods. Moreover, the proposed efficient selector works much better than classical tools like, for example, cross-validation. For this and for the insight into the apparent circulus vitiosus, namely the negative correlation of data driven selectors and the desired optimal selector, we would like to thank the authors. They have combined brilliant mathematical analysis with important empirical and practical questions and have made a deep problem of nonparametric statistics accessible for a wide readership.

An implementation of the proposed efficient selector needs a fine tuned estimator of $J_{r+s/2}$, see Section 5. In fact, some accuracy is required in the stage of estimating this tuning constant, see Park and Marron(1991). One needs more than consistency,

$$\hat{J}_{r+s/2} - J_{r+s/2} = O_p(n^{-\alpha})$$

for some small positive α. This is because $g_*^{2r+2s+1}$ is chosen to cancel the two leading bias terms of $\hat{J}_r$, and a bit of mistuning for this constant yields some bias, so that one can not get the full advantage of Jones and Sheather's device. Hence just replacing m by $\hat{\lambda} m_0$ in the first tuning stage, as in the present paper, may not work too well when m is far different from m_0 in its shape. We suspect that the simulation gave good results since m_0 was used which is the same as m except for a scale factor.

2. An argument for cross-validation

The technical approach to describing the difficulties of data driven bandwidth choice is to assume that the object under study, the density or the regression function, has more than two derivatives. Thus, in a sense, one employs higher order smoothness to describe a problem typical for lower order

(1992) Park, B. and Härdle, W. Discussion of the paper by Hall and Johnstone, (September 1991) smooth functions. This is somewhat unsatisfactory since for the higher order smooth functions one knows how to construct better estimators based on higher order kernels. Is this another circulus vitiosus that we can not avoid?

Cross-validation does not face this problem since it can be applied independent of the knowledge of the smoothness class: It is asymptotically optimal for kernel density estimates provided the density f is bounded. In the mind of many statisticians cross-validation plays the role of "$\bar{x}$" for smoothing parameter selection. So one can imagine that a number of statisticians will still use cross-validation in the future.

3. An extension for regression with non uniform covariates

The theory developed in the present paper may be extended to the case of regression with non uniform covariates. Suppose we are given observations of independent identically distributed random variables $\{(X_i, Y_i)\}_{i=1}^{n}$, where X_i, $Y_i \in R$. For a simple extension to this case, let us assume that the marginal density function, $f(x)$, of the covariates is known to us. An estimator of the conditional expectation $m(x) = E(Y|X = x)$, similar to the one proposed by Nadaraya(1964) and Watson(1964), is given by

$$m_h(x) = r_h(x)/f(x)$$

where

$$\hat{r}_h(x) = (nh)^{-1} \sum_{i=1}^{n} K\{(x - X_i)/h\} Y_i \,.$$

Note that in contrast to the Nadaraya-Watson estimator we divide here by the *true and known* density and not by an estimate of the marginal density $f(x)$. It is known that these estimators have different variances, but we have not been able to extend the results of the present paper to this full generality.

An appropriate measure to assess the performance of $\hat{m}_h$ in this case is

$$\Delta(h) = \int \{\hat{m}_h(x) - m(x)\}^2 u(x) f(x) dx \,.$$

By paralleling arguments in the present paper, the same representation

$$\hat{h}_0 = A_1 + n^{1/5} A_2 J_1 + o_p(n^{-3/10}) \,,$$

may be established with slight changes in J_r, $\hat{J}_r$ and $I_1(h)$. In fact,

$$J_r = (-1)^r \int m(x)\{m(x)u(x)\}^{(2r)} f(x)dx\,,$$

$$\hat{J}_r = \{n(n-1)g^{2r+1}\}^{-1}(-1)^r \sum\sum_{i\neq j} Y_i Y_j u(X_j) f^{-1}(X_j) L^{(2r)}\{(X_i - X_j)/g\}\,,$$

$$I_1(h) = \int \hat{r}_h(x)\hat{a}_h(x)u(x)f^{-1}(x)dx\,.$$

Also the definition of $\hat{b}$ should be modified according to the changes in the asymptotic representation of $M''(h)$. Write $\hat{J}$ and $\hat{B}$ for consistent estimators of $J = \int r''^2 u f^{-1}$ and $B = \int m^2 u$ respectively, where $r(x) = m(x)f(x)$. Then $\hat{b}$ is defined by

$$\hat{b} = \check{h}_0\{2(n\check{h}_0^3)^{-1}(\hat{\sigma}^2 u_1 + \hat{B})k_{ns} + 3\check{h}_0^2 k_2^2 \hat{J}\}\,.$$

The modified forms of J_r and $\hat{J}_r$, for instance, come from equation (65). In our case

$$\begin{aligned} n^{-1}\sum_{i=1}^{n} Y_i v^{(4)}(X_i) &= E[Y v^{(4)}(X)] + O_p(n^{-1/2}) \\ &= E[m(X)v^{(4)}(X)] + O_p(n^{-1/2}) \end{aligned}$$

and, with modified $\hat{J}_r$ defined above,

$$(-1)^r n^{-1}\sum_{i=1}^{n} Y_i v^{(2r)}(X_i) = \frac{1}{2}(J_r + \hat{J}_r) + o_p(n^{-1/2})\,.$$

It is exactly at this point where we could not see how to extend it to the random estimate of the density. We hope that this will be solved in the future.

References

[1] Nadaraya, E.A.(1964). On estimating regression. *Theory of Probability and Its Application*, **10**, 186-90.

[2] Park, B.U. and Marron, J.S.(1991). On the use of pilot estimators in bandwidth selection, *Journal of Nonparametric Statistics*, to appear.

[3] Watson, G.S.(1964). Smooth regression analysis. *Sankhyā*, Series A, 26, 359-72.

Robust Locally Adaptive Nonparametric Regression

By

Wolfgang HÄRDLE
CORE
Université Catholique de Louvain

Alexander TSYBAKOV
Institute for Problems of Information Transmission
Academy of Sciences of the U.S.S.R.

April 1990

Abstract

The problem of robust nonparametric regression estimation is considered. We study pointwise asymptotic normality of variable bandwidth M-smoothers. A locally optimal bandwidth is derived, and the "plug-in" method is used for data-driven local bandwidth selection. Asymptotic optimality of local bandwidth selectors based on robust pilot estimators is proved. The work improves upon earlier contributions since we get the estimates that have smaller mean squared error under weaker assumptions on the error distribution and on the ψ-function of M-smoother.

AMS 1980 Subject classification 62G05.
Key words and phrases: Robust curve estimation, M-estimation, nonparametric regression, smoothing, locally adaptive bandwidth selection.

ISSN-0771 3894

1. Introduction

Let $(X_1, Y_1), \cdots, (X_n, Y_n)$ be i.i.d. bivariate random variables such that

$$Y_i = m(X_i) + \varepsilon_i$$

where m is an unknown regression function and ε_i are i.i.d. random errors.

Define a smoother $m_n(x)$ as a solution of the following optimization problem

$$m_n(x) = \underset{t \in I\!R}{\operatorname{argmin}} \sum_{i=1}^{n} \rho(Y_i - t)\ K(\frac{X_i - x}{h_n}) \tag{1.1}$$

where $\rho(\cdot)$ is a convex function, $K(\cdot)$ is a nonnegative function (a kernel) and $h_n > 0$ is the smoothing parameter or the bandwidth. Let ψ be the left-side derivative of ρ. If ψ is continuous then the smoother (1.1) satisfies the equation

$$\sum_{i=1}^{n} \psi(Y_i - m_n(x))\ K(\frac{X_i - x}{h_n}) = 0 \tag{1.2}$$

The estimates (1.1) are called M-smoothers. They are straightforward generalizations of kernel regression estimates based on the idea of M-estimation see Härdle and Tsybakov (1988). The so-called Nadaraya-Watson regression estimate is a special case of M-smoother for $\rho(u) = u^2$. Other possibilities are the median smoother $(\rho(u) = |u|)$ and Huber-type smoothers with

$$\rho(u) = \begin{cases} u^2/2, & |u| \le c, \\ c|u| - c^2/2, & |u| > c, \end{cases}$$

where c is some positive number.

The asymptotic properties of M-smoothers have been studied by several authors. Pointwise consistency and asymptotic normality are investigated in Tsybakov (1982a, b; 1983) and Härdle (1984a). The fixed x-design case is considered by Härdle and Gasser (1984). A recursive modification of M-smoother is introduced and analysed in Tsybakov (1983). Locally-polynomial M-smoothers has been considered by (Katkovnik (1985)). For asymptotic normality of locally polynomial M-smoothers, including the case of discontinuous ψ-functions, and for optimal bandwidth selection see Tsybakov (1986). Other possibilities of robust data smoothing is based on M-type splines (Huber (1979), Cox (1983)). Also nonparametric regression M-estimates on functional classes have been introduced (Nemirovskii, Polyak and Tsybakov (1983, 1985)). They have the advantage to be robust estimates that inherit qualitative behavior of the unknown function m (e.g. convexity, monotonicity, etc.). In neither of these papers the data-driven choice of smoothing parameter has been considered.

This paper is concerned with this problem. The asymptotically optimal bandwidth calculated in Tsybakov (1982b) depends on some a priori constants that in practice are unknown. This raises the point of data-driven bandwidth selection for M-smoothers. For the fixed design case Härdle (1984b) proposed to use the cross-validation technique and related global bandwidth selectors based on the residual sum of squares. More recently Hall and Jones (1989) proved the asymptotic optimality of cross-validation bandwidth selector for random design robust nonparametric regression with the Huber-type function ρ. They also considered the adaptive choice of tuning parameter c ocurring in the definition of ρ. Proofs of asymptotic optimality in Hall and Jones (1989) are based on the assumption that all moments of ε_i are finite. It is conjectured that cross-validation and related bandwidth selection criteria are not asymptotically optimal unless some higher moments of ε_i are finite. This comes from the fact that such criteria contain oscillating terms that are linear in Y_i. The higher moments assumption, however, is not reasonable in our view if one believes in gross errors.

Thus there exists the problem of finding adative bandwidth selectors that are asymptotically optimal in a sense to be defined here under milder assymptions on the error distribution. Another problem is data-driven bandwidth selection for a wide class of M-smoothers including the case of discontinuous ψ-functions (e.g. the median-smoother). These problems are addressed here. We study a variable bandwidth M-smoother and we introduce the "plug-in" technique to construct locally adaptive stochastic bandwidths. The use of variable bandwidth kernel smoothers is motivated by the simple observation that in asymptotics the mean integrated squared error (MISE) of the best variable bandwidth estimator is smaller than MISE of the best constant bandwidth estimator (see e.g. Müller and Stadtmüller (1987), Tsybakov (1987)). This property can be explained intuitively by the possibility to reduce the bandwidth and therefore the local mean squared error near peaks, and to increase the bandwidth in flat parts of the curve.

The "plug-in" technique i.e. the use of of estimated asymptotically optimal bandwidths goes back to Woodroofe (1970) who used it in density estimation (see

Devroye and Györfi (1985); chapter 6, for further references). Mack and Müller (1987) and Tsybakov (1987) proved the asymptotic optimality of "plug-in" bandwidth choice for the Nadaraya-Watson regression estimator.

In this paper we extend to robust M-smoothers the results of Tsybakov (1987) concerning data-driven local bandwidth selection. The class of M-smoothers studied here is rather broad, also smoothers with discontinuous ψ-functions such as the median smoother satisfy our assumptions.

2. Main Results

First consider the asymptotic normality of variable bandwidth M-smoothers at a fixed point x. Assume the following.

(A1) The kernel K is nonnegative, bounded, compactly supported, and

$$\int K(u)du = 1 \, , \int uK(u)du = 0.$$

(A2) The regression function m is twice continuously differentiable in some neighborhood of x.

(A3) The marginal density $f(\cdot)$ of X_1 is continuously differentiable in some neighborhood of $x, f(x) \neq 0$, and

$$b_1(x) = f'(x)\, m'(x)/f(x) + m''(x)/2 \neq 0.$$

(A4) The function ψ is nondecreasing.

(A5) The function $\varphi(u) = \int \psi(u+v)dF(v)$ where F is the distribution of errors ε_i is twice continuously differentiable in some neighborhood of the point $u = 0$,

$$\varphi(0) = \varphi''(0) = 0 \, , \varphi'(0) > 0.$$

(A6) The function $\varphi_2(u) = \int \psi^2(u+v)dF(v)$ is continuous and positive in some neighborhood of the point $u = 0$.

(A7) The bandwidth h_n is of the form $h_n = \beta(x)n^{-1/5}$ where $\beta(x) > 0$ is a constant.

In the following we use the notation $C_K = \int K^2(u)du$ and $d_K = \int u^2 K(u)du$.

Assumption (A7) is introduced since it guarantees the optimality of the rate of convergence for the class of regression functions m with bounded second derivative (Ibragimov, Hasminskii (1981), Stone (1980)).

THEOREM 1. Let (A1) - (A7) be satisfied. Then, as $n \to \infty$, the sequence $n^{2/5}(m_n(x) - m(x))$ is asymptotically normal with mean

$$b(x)\beta^2(x)b_1(x) \int u^2 K(u)du$$

and variance

$$\sigma^2(x) = \frac{\varphi_2(0)}{(\varphi'(0))^2} \frac{\int K^2(u)du}{\beta(x)f(x)}.$$

Remark 1: Note that the bias of M-smoother is the same as the bias of the Nadaraya-Watson regression estimate (Collomb (1977)). The variance differs in that we have now $\varphi_2(0)/(\varphi'(0))^2 = V(\psi, F)$ instead of conditional variance Var $(Y|X = x)$.

Remark 2: Theorem 1 is closely related to the earlier results by Tsybakov (1982b) and Härdle (1984a) although it is not the direct consequence of these.

For the following we need some more concepts. Denote $R(\beta(x), K, x) = b^2(x) + \sigma^2(x)$ the MSE of M-smoother at a fixed point x calculated from the asymptotic distribution. We call the M-smoother $m_n(x)$ with bandwidth $h_n = h_n(x)$ pointwise optimal if $h_n = \beta^*(x)n^{-1/5}$ where

$$\beta^*(x) = \underset{\beta>0}{\operatorname{argmin}} R(\beta, K, x) =$$

$$= (\frac{V(\psi, F)\ C_K}{4f(x)b_1(x)\ (d_K)^2})^{1/5}.$$

The MSE of pointwise optimal M-smoother is

$$R^*(K, x) = R(\beta^*(x), K, x) =$$

$$= (5/4^{4/5})\ V(\psi, F)^{4/5}\ b_1^{2/5}(x) f^{-4/5}(x)\ (d_K)^{2/5}\ (C_K)^{4/5}$$

Define the locally adaptive M-smoother as

$$\hat{m}_n(x) = \underset{t\in \mathbb{R}}{\operatorname{argmin}} \sum_{i=1}^{n} \rho(Y_i - t)K(\frac{X_i - x}{\hat{h}_n}) \tag{2.1}$$

where $\hat{h}_n = \hat{h}_n(X_1, \cdots, X_n, \cdots, Y_n, x)$ is a sequence of stochastic bandwidths such that

$$\frac{\hat{h}_n}{\beta(x)n^{-1/5}} \xrightarrow{P}, \; n \to \infty, \tag{2.2}$$

for every point x where $\beta(x) > 0$. The sequence $\hat{h}_n$ satisfying (2.1) can be constructed using consistent pilot estimators of b_1, f and V. Such estimators are presented in Section 3.

To prove asymptotic normality of locally adaptive M-smoother we need the following additional assumption.

(A8) The kernel K is continuous and there exist such $L > 0, \alpha \in (1,2]$, $\varepsilon_0 \in (0,1)$ that

$$\int (K(qu) - K(u))^2 du \le L|q-1|^\alpha \; , \; |q-1| < \varepsilon_0.$$

THEOREM 2. Let (A1) - (A8) and (2.2) hold. Then for any solution $\hat{m}_n(x)$ of (2.1) the sequence $n^{2/5}(\hat{m}_n(x) - m(x))$ is asymptotically normal with mean $b(x)$ and variance $\sigma^2(x)$.

Thus $\hat{m}_n$ is pointwise asymptotically equivalent to m_n provided (2.2) is true. In particular, the locally adaptive estimate $\hat{m}_n$ with $\hat{h}_n = \hat{\beta}_n(x)n^{-1/5}$ where $\hat{\beta}_n(x) \xrightarrow{P} \beta^*(x)$, $n \to \infty$, is asymptotically equivalent to the pointwise optimal M-smoother.

3. Pilot Estimators.

To estimate $\beta^*(x)$ consistently we need consistent estimators $f_n^{(0)}(x), f_n^{(1)}(x)$, $m_n^{(1)}(x), m_n^{(2)}(x), V_n$ of f, f', m', m'', V respectively. The estimates of f and f' are standard (see e.g. Devroye and Györfi (1985)). For example we can take

$$f_n^{(\ell)}(x) = \frac{1}{na_n^{\ell+1}} \sum_{i=1}^n K_\ell \left(\frac{X_i - x}{a_n}\right) , \; \ell = 0, 1,$$

where the kernel K_0 satisfies (A1), K_1 is such that

$$\int K_1(u)du = 0 \; , \; \int uK_1(u)du = 1,$$

and $a_n \to 0$ so that $na_n^3 \to \infty$.

The estimation of $m'(x)$ and $m''(x)$ is somewhat more sophisticated since robust estimators are to be used here. The standard kernel pilot estimates of m' and m'' as in Tsybakov (1987) are linear in Y_i and hence they are sensitive to outliers. A possible way of consistent robust derivatives estimation is the local approximation method (Katkovnik (1985), Tsybakov (1986)). However this method is rather involved from the computational point of view. Some simple estimates are preferable. For example, define

$$\begin{aligned} m_n^{(1)}(x) &= (m_n(x+a_n) - m_n(x))/a_n, \\ m_n^{(2)}(x) &= (m_n(x+a_n) + m_n(x-a_n) - 2m_n(x))/(2a_n^2) \end{aligned}$$

where $a_n \to 0$ and $m_n(x)$ is the robust estimate (1.1).

It follows from Härdle, Collomb (1986) that under mild conditions on the error distribution

$$\lim_n \sup P\{b_n \sup_{|z-x|\le\delta} |m_n(z) - m(z)| \ge \varepsilon\} = 0\ , \quad \forall\varepsilon > 0\ , \ \forall x, \tag{3.1}$$

where $b_n = O((n/\log n)^{1/3})$, $n \to \infty$, and $\delta > 0$ is small enough. Note that

$$\begin{aligned} &P\{\ |m_n^{(1)}(x) - m'(x)|\ \ge \varepsilon\}\ \le \\ &P\{\ |m_n^{(1)}(x)\ -\ \frac{m(x+a_n)-m(x)}{a_n}|\ \ge \frac{\varepsilon}{2}\} + . \\ &P\{\ |m'(x) - \frac{m(x+a_n)-m(x)}{a_n}|\ \ge \frac{\varepsilon}{2}\} \end{aligned}$$

The second probability in the RHS of this inequality vanishes if n is large enough. The first probability does not exceed

$$P\{2a_n^{-1} \sup_{|z-x|\le\delta} |m_n(z) - m(z)| \ge \varepsilon/2\}$$

for n such that $a_n < \delta$. If $a_n^{-1} = o((n/\log n)^{1/3})$ then this probability tends to 0 as $n \to \infty$. This proves consistency of $m_n^{(1)}(x)$. Consistency of $m_n^{(2)}(x)$ follows from the same arguement (here, however, we have to choose $a_n^{-1} = o((n/\log n)^{1/6})$.

As the estimate of variance V choose

$$V_n = \frac{\frac{1}{n}\sum_{i=1}^{n}\psi^2(Y_i - m_n(X_i))}{(\frac{1}{a_n}(\varphi_n(a_n) - \varphi_n(0)))^2}$$

where

$$\varphi_n(a) = \frac{1}{n}\sum_{i=1}^{n}\psi(a + Y_i - m_n(X_i)).$$

The estimate V_n converges in probability to V as $n \to \infty$ under some choice of a_n provided ψ is bounded and Lipschitz continuous and (3.1) holds. In fact under these conditions

$$| \frac{1}{n} \sum_{i=1}^{n} \psi^2(Y_i - m_n(X_i)) - \frac{1}{n} \sum_{i=1}^{n} \psi^2(\varepsilon_i)| \le$$

$$\le \frac{c_1}{n} \sum_{i=1}^{n} |m(X_i) - m_n(X_i)| \ = \ o_p(b_n), n \to \infty.$$

This entails convergence of the numerator of V_n to $\varphi_2(0)$. Similarly

$$\sup_a \ |\varphi_n(a) - \frac{1}{n} \sum_{i=1}^{n} \psi(\varepsilon_i + a)| \le$$

$$\le \frac{c_2}{n} \sum_{i=1}^{n} |m(X_i) - m_n(X_i)| \ = \ o_p(b_n), n \to \infty.$$

Here and in the sequel $c_i, i = 1, 2, \cdots$are positive constants. Convergence of the denominator of V_n to $(\varphi'(0))^2$ follows now from the same arguement as the convergence $m_n^{(1)}(x) \xrightarrow{P} m'(x)$ that has been just proved.

Note that consistent estimates of V can be constructed also for the case of discontinuous ψ. For example, if we want to use the median smoother (i.e. $\psi(u) =$ sign u) we have to estimate $V = (4p^2(0))^{-1}$ where $p(\cdot)$ is the density of ε_1. A possible estimator of V is now $V_n \ = \ (4p_n^2(0))^{-1}$ where

$$p_n(0) \ = \ (na_n)^{-1}\sum_{i=1}^{n} K(\frac{Y_i - m_n(X_i)}{a_n}).$$

Assuming that a_n tends to 0 slowly enough, imposing Lipshitz condition on K and using (3.1) we easily get consistency of $p_n(0)$.

4. Proofs

To simplify the notation set w.l.o.g. $m(x) = 0, \sigma(x) = \sigma, b(x) = b, \beta(x) = \beta, h_n(x) = h_n = \beta n^{-1/5}$. Since ψ is monotone and K is nonnegative we obtain

(4.1)
$$\{\hat{m}_n(x) < u\sigma n^{-2/5}\} \subseteq \hat{W}_n \cup \{\sum_{i=1}^{n} K(\frac{X_i - x}{\hat{h}_n}) = 0\},$$
$$\hat{W}_n \subseteq \{\hat{m}_n(x) \le u\sigma n^{-2/5}\}$$

where $u \in I\!R$ is arbitrary, and

$$\hat{W}_n = \{\sum_{i=1}^{n} \psi(Y_i - u\sigma n^{-2/5})K(\frac{X_i - x}{\hat{h}_n}) < 0\}.$$

Denote

$$p_{1n}(\varepsilon,\delta) = P\Big\{ \sup_{|t|\le\delta} |\frac{1}{\sqrt{nh_n}} \sum_{i=1}^{n} \psi(Y_i - u\sigma n^{-2/5})\{K(\frac{X_i - x}{h_n}) - K(\frac{X_i - x}{h_n(1+t)}))| \ge \varepsilon \Big\}$$

$$p_{2n} = P\{\ |\hat{h}_n/h_n - 1| > \delta\}.$$

where $\varepsilon > 0, 0 < \delta < 1$. It is clear that

$$(4.2) \qquad P(\hat{W}_n) \le P\{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \zeta_{ni} < \varepsilon\} + p_{1n}(\varepsilon,\delta) + p_{2n}$$

and

$$(4.3) \qquad P(\hat{W}_n) \ge P\{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \zeta_{ni} < -\varepsilon\} - p_{1n}(\varepsilon,\delta) - p_{2n}$$

where $\zeta_{ni} = h_n^{-1/2}\psi\ (Y_i - u\sigma n^{-2/5})K((X_i - x)/h_n)$.

Define

$$s_n^2 = \text{var}\{\zeta_{ni}\}, Y_{ni} = (\zeta_{ni} - E\{\zeta_{ni}\})/s_n, i = 1, \cdots, n.$$

Note that Y_{ni} are standardized - i.i.d. random variables.

If $\hat{h}_n$ is not random and $\hat{h}_n = h_n = \beta n^{-1/5}$ then (4.2) and (4.3) hold with $\varepsilon = 0, p_{1n} = p_{2n} = 0$. Theorems 1 and 2 follow from (4.1) - (4.3) and the next relations:

$$(4.4) \qquad \lim_n P\{\sum_{i=1}^{n} K(\frac{X_i - x}{\hat{h}_n}) = 0\} = 0,$$

$$(4.5) \qquad \limsup_{\delta\to 0}\ \limsup_n\ p_{1n}(\varepsilon,\delta) = 0,\ \forall\varepsilon > 0,$$

$$(4.6) \qquad \lim_n P\{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \zeta_{ni} < \varepsilon\} =$$

$$\Phi(u - \frac{b}{\sigma} + \frac{\varepsilon}{\varphi_2(0)f(x)\int K^2(v)dv}),\ \forall\varepsilon \in I\!R,$$

where Φ is the standard normal *cdf*. In fact, Theorem 1 is the consequence of (4.1) - (4.4), (4.6) with $\varepsilon = 0, \hat{h}_n = h_n = \beta n^{-1/5}$. To prove Theorem 2 note that by (2.2), (4.2), (4.5) and (4.6)

$$\limsup_n P(\hat{W}_n) \le \Phi(u - \frac{b}{\sigma} + \frac{\varepsilon}{\varphi_2(0) f(x) \int K^2(v)dv}) \tag{4.7}$$

for any $\varepsilon > 0$. Similarly,

$$\liminf_n P(\hat{W}_n) \ge \Phi(u - \frac{b}{\sigma} - \frac{\varepsilon}{\varphi_2(0) f(x) \int K^2(v)dv}) \tag{4.8}$$

Since $\varepsilon > 0$ is arbitrary we get

$$\lim_n P(\hat{W}_n) \;=\; \Phi(u - \frac{b}{\sigma}).$$

This together with (4.1) and (4.4) entails Theorem 2. Thus it remains to prove (4.4) to (4.6).

Proof of (4.4). Denote $\eta_1 \;=\; 2f(x)$. We have

$$\begin{aligned}
&(4.9)\; P\{\sum_{i=1}^n K(\frac{X_i - x}{\hat{h}_n}) \;=\; 0\} = \\
&= P\{\frac{1}{n}\sum_{i=1}^n (\frac{1}{\hat{h}_n} K(\frac{X_i - x}{\hat{h}_n}) - \frac{1}{h_n} K(\frac{X_i - x}{h_n})) \;= \\
&= -\frac{1}{nh_n}\sum_{i=1}^n K(\frac{X_i - x}{h_n})\} \le \\
&\le P\{-\frac{1}{nh_n}\sum_{i=1}^n K(\frac{X_i - x}{h_n}) < -\eta_1\} + \\
&+ P\left\{\sup_{|t|\le\delta} |\frac{1}{nh_n}\sum_{i=1}^n \left(\frac{1}{1+t} K(\frac{X_i - x}{h_n(1+t)}) - K(\frac{X_i - x}{h_n})\right)| \ge \eta_1\right\} + p_{2n}
\end{aligned}$$

The first probability in the RHS of (4.9) tends to 0 as $n \to \infty$ since

$$\frac{1}{nh_n}\sum_{i=1}^n K(\frac{X_i - x}{h_n}) \xrightarrow{P} f(x), n \to \infty, \tag{4.10}$$

(Parzen (1962)). The second probability tends to 0 by Lemma 1 of Tsybakov (1987) if $\delta > 0$ is chosen to be small enough. The third probability p_{2n} tends to 0 by (2.2).

If $\hat{h}_n = h_n$ then only the first probability in the RHS of (4.9) is nonzero, and (4.4) follows directly from (4.10). In this case condition (A8) is redundant.

Proof of (4.5). Use the following result of Prokhorov (1956). Let $f(t)$ be a continuous random process on $T = [-\delta, \delta]$ and

$$E(f(t+h) - f(t))^2 \leq \kappa h^\alpha \ , \ t, t+h \in T, \alpha \in (1,2] \ h > 0, \tag{4.11}$$

where $\kappa > 0$ is a constant. Then

$$P\{\sup_{t \in T} |f(t) - f(0)\ | \geq \varepsilon\} \leq c_3 \kappa \varepsilon^{-\alpha} \delta^{(\alpha-1)/4}, \varepsilon > 0. \tag{4.12}$$

Define

$$f(t) = \frac{1}{\sqrt{nh_n}} \sum_{i=1}^n \psi(Y_i - u\sigma n^{-2/5}) K(\frac{X_i - x}{h_n(1+t)}),$$

$$K_{1n}(z,h) = K(\frac{z-x}{h_n(1+t+h)}) - K(\frac{z-x}{h_n(1+t)}).$$

We prove now that the process $f(t)$ satisfies (4.11) for n large enough. This entails (4.5) since the LHS of (4.12) equals to $p_{1n}(\varepsilon, \delta)$ in our case. We have

$$\begin{aligned} &E(f(t+h) - f(t))^2 = \\ &= \frac{1}{nh_n} E(\sum_{i=1}^n \psi(Y_i - u\sigma n^{-2/5})\ K_{1n}(X_i, h))^2 = \\ &= \frac{1}{nh_n} [nE\{\psi^2(Y_1 - u\sigma n^{-2/5}) K_{1n}^2(X_1, h)\} + \\ &+ n(n-1)(E\{\psi(Y_1 - u\sigma n^{-2/5}) K_{1n}(X_1, h)\})^2]. \end{aligned} \tag{4.13}$$

Assume that $\delta \in (0,1)$ is small enough that $2\delta/(1-\delta) < \varepsilon_0$. This guarantees that $|\frac{1+t}{1+t+h} - 1| \ < \varepsilon_0$ for any $t, t+h \in T$. By (A8)

$$h_n^{-1} \int K_{1n}^2(z,h) dz \ \leq \ c_4 h^\alpha. \tag{4.14}$$

It follows easily from (A8) and from Cauchy inequality that

$$h_n^{-1} \int \ |K_{1n}(z,h)|\ dz \ \leq \ c_5 h^{\alpha/2}. \tag{4.15}$$

Using (4.15) and the fact that $\varphi_2(u)$ is bounded in some neighborhood of $u = 0$, and $f(z)$ is bounded in some neighborhood of $z = x$ we get

$$\begin{aligned} &\frac{1}{h_n} E\{\psi^2(Y_1 - u\sigma n^{-2/5})\ K_{1n}^2(X_1, h)\} = \\ &= \frac{1}{h_n} \int \varphi_2(m(z) - u\sigma n^{-2/5})\ f(z) K_{1n}^2(z,h) dz \leq c_6 h^\alpha. \end{aligned} \tag{4.16}$$

Moreover

$$E\{\psi(Y_1 - u\sigma n^{-2/5})K_{1n}(X_1,h)\} =$$

$$= \int \varphi(m(z) - u\sigma n^{-2/5})\ f(z)K_{1n}(z,h)dz\ =\ \sum_{j=1}^{4} I_j,$$

where

$$\begin{aligned}
I_1 &= \int(\varphi(m(z) - u\sigma n^{-2/5}) - \varphi'(0)(m(z) - u\sigma n^{-2/5}))\ f(x)K_{1n}(z,h)dz \\
I_2 &= \int \varphi(m(z) - u\sigma n^{-2/5})\ (f(x) - f(z))K_{1n}(z,h)dz, \\
I_3 &= \int f(x)\varphi'(0)\ (m'(x)(z-x)\ +\ \frac{m''(x)}{2}\ (z-x)^2 - u\sigma n^{-2/5})K_{1n}(z,h)dz, \\
I_4 &= \int f(x)\varphi'(0)(m(z) - m'(x)(z-x) - \frac{m''(x)}{2}\ (z-x)^2)K_{1n}(z,h)dz.
\end{aligned}$$

Note that (4.15), (A2) and the boundedness of supp K entail

$$|I_4|\ \leq c_7 n^{-2/5} \int |K_{1n}(z,h)|\ dz \leq c_8 n^{-3/5}h^{\alpha/2}.$$

Using (A1) we get

$$I_3\ =\ f(x)\varphi'(0)\ [\frac{m''(x)}{2}\ ((1+t+h)^3 - (1+t)^3)h_n^3 -$$

$$u\sigma n^{-2/5} \int K_{1n}(z,h)dz]$$

Hence

$$|I_3|\ \leq\ c_9 n^{-3/5}(h + h^{\alpha/2}).$$

The derivatives m', f' are finite and continuous in some neighborhood of x, and the derivative φ' is finite and continuous in some neighborhood of zero. This together with the condition $\varphi(0) = 0$ gives that

$$|I_2|\ \leq\ c_{10} n^{-2/5} \int\ |K_{1n}(z,h)\ |\ dz \leq c_{11} n^{-3/5}h^{\alpha/2}.$$

Finally, the condition $\varphi(0)\ =\ \varphi''(0)\ =\ 0$ and continuity of φ'' in some neighborhood of zero entail

$$\sup_{|z-x|\leq\mathcal{D}(1+|t|+|h|)h_n} |\varphi(m(z) - u\sigma n^{-2/5}) - \varphi'(0)(m(z) -$$

$$- u\sigma n^{-2/5})| \leq c_{12} \sup_{|z-x|\leq\mathcal{D}(1+|t|+|h|)h_n} |m(z) - u\sigma n^{-2/5}|^2 \leq$$

$$\leq c_{13} n^{-2/5}, \mathcal{D} = \max\{|z| : K(z) \neq 0\}.$$

Using this inequality and (4.15) we obtain

$$|I_1| \leq c_{14} n^{-3/5} h^{\alpha/2}.$$

Thus we proved that

$$(4.17) \qquad |E\{\psi(Y_1 - u\sigma n^{-2/5})K_{1n}(X_1,h)\}| \leq c_{15} n^{-3/5}(h + h^{\alpha/2})$$

It follows from (4.13), (4.16), (4.17) that

$$E(f(t+h) - f(t))^2 \leq c_{16}(h^2 + h^\alpha).$$

This yields (4.11) since by (A8) $\alpha \in (1,2]$.

Proof of (4.6). We have

$$P\{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \zeta_{ni} < \varepsilon\} =$$
$$= P\{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_{ni} < (-\sqrt{n}\; E\{\zeta_{n1}\} + \varepsilon)/s_n\}.$$

Here

$$(4.18) \qquad \sqrt{n}\; E\;\{\zeta_{n1}\} =$$
$$= \sqrt{n/h_n} \int \varphi(m(z) - u\sigma n^{-2/5}) f(z) K(\frac{z-x}{h_n}) dz =$$
$$= \sqrt{nh_n} \int K(v)\varphi(m(x+vh_n) - u\sigma n^{-2/5}) f(x+vh_n) dv.$$

Denote

$$H(t) \;=\; \varphi(m(x+t) - u\sigma n^{-2/5}).$$

Using (4.1) we obtain

$$(4.19) \qquad \int K(v)H(vh_n)f(x+vh_n)dv =$$
$$= \int K(v)(H(0) + H'(0)vh_n + \frac{1}{2}H''(\theta vh_n)(vh_n)^2)\times$$
$$\times\, (f(x) + f'(x+\theta_1 vh_n)vh_n)dv =$$
$$= H(0)f(x) + h_n^2 \int v^2 K(v)(H'(0)f'(x+\theta_1 vh_n) +$$
$$+ (1/2)H''(\theta vh_n)f(x))dv + \alpha_n$$

where $0 \le \theta, \theta_1 \le 1$ and $\alpha_n = O(h_n^3)$ by (A1) and by boundedness of $f'(x+t)$ and $H''(t)$ for small t.

Now

(4.20)
$$\begin{aligned} H(0) &= \varphi(-u\sigma n^{-2/5}) = \\ &= -u\sigma n^{-2/5}\varphi'(0) + O(n^{-4/5}), \\ H'(0) &= \varphi'(-u\sigma n^{-2/5})m'(x) = \\ &= \varphi'(0)m'(x)(1+o(1)), n \to \infty. \end{aligned}$$

Since K is bounded and compactly supported and f', H'' are continuous we get

(4.21)
$$\begin{aligned} &\int v^2K(v)f'(x+\theta_1 vh_n)dv = \\ &= f'(x)\int v^2K(v)dv + o(1), \\ &\int v^2K(v)H''(\theta vh_n)dv = \\ &= H''(0)\int v^2K(v)dv + o(1), \end{aligned}$$

where by continuity of φ' and φ''

(4.22)
$$\begin{aligned} &H''(0) = \varphi''(-u\sigma n^{-2/5})m'(x)+ \\ &+\varphi'(-u\sigma n^{-2/5})m''(x) = \varphi''(0)m'(x)+ \\ &+\varphi'(0)m''(x) + o(1) = \varphi'(0)m''(x) + o(1), \\ &n \to \infty. \end{aligned}$$

Combining (4.18) - (4.22) we find

(4.23)
$$\begin{aligned} &\lim_n \sqrt{n}\, E\{\zeta_{n1}\} = \sqrt{\beta}\,(-u\sigma\varphi'(0))f(x)+ \\ &+\beta^{5/2}\varphi'(0)\int v^2K(v)dv\,(g(x)f(x)). \end{aligned}$$

Next
$$s_n^2 = E\{\zeta_{n1}^2\} - (E\{\zeta_{n1}\})^2$$

where
$$\begin{aligned} E\{\zeta_{n1}^2\} &= \int \varphi_2(m(x+vh_n) - u\sigma n^{-2/5})K^2(v)f(x+vh_n)dv \\ &= \varphi_2(0)f(x)\int K^2(v)dv + o(1). \end{aligned}$$

This together with (4.23) entails

$$\lim_n s_n^2 \;=\; \varphi_2(0)f(x)\int K^2(v)dv. \tag{4.24}$$

From (4.23), (4.24) one obtains

$$\lim_n\,(-\sqrt{n}\;E\{\zeta_{n1}\}/s_n) \;=\; u - b/\sigma.$$

To prove (4.6) it remains to show that the distribution of $n^{-1/2}\sum_{i=1}^n Y_{ni}$ converges towards the standard normal. By the normal convergence criterion, as given in Loève (1960), p.295, it suffices to prove that

$$\lim_n E\{Y_{n1}^2 I(|Y_{n1}| \geq \sqrt{n}\varepsilon)\} \;=\; 0, \forall \varepsilon > 0. \tag{4.25}$$

Using (4.23), (4.24) we get

$$E\{Y_{n1}^2 I_n\} \;=\; s_n^{-2}E\{\zeta_{n1}^2 I_n\} + o(1)$$

where $I_n \;=\; I\{|Y_{n1}| \geq \sqrt{n}\varepsilon\} \;\leq\; I\{|\zeta_{n1}| \geq \delta_n\}, \delta_n \;=\; \sqrt{n}\;\varepsilon s_n - |E\{\zeta_{n1}\}|$. Thus (4.25) follows from

$$\lim_n\; E\{\zeta_{n1}^2 I_n\} \;=\; 0. \tag{4.26}$$

Now we prove (4.26). By (A1) we have $\mathcal{D} \;=\; \max\{|z| : K(z) \neq 0\} < \infty$. The monotonicity of ψ entails

$$\begin{aligned}\sup_{|z-x|\leq \mathcal{D}h_n} &|\psi(\xi + m(z) - u\sigma n^{-2/5})| \leq \\ &\leq \max\{|\psi(\xi + u')|, |\psi(\xi - u')|\} \triangleq w(\xi), \forall \in I\!R,\end{aligned} \tag{4.27}$$

where $0 < u' < \infty$ is chosen such that $u\sigma n^{-2/5} + \max_{|z-x|\leq \mathcal{D}h_n} |m(z) - m(x)| \leq u'$ for n large enough. Using (4.23), (4.24), (4.27) and the boundedness of K we find

$$\begin{aligned}|\zeta_{n1}| \;&\leq\; h_n^{-1/2}\; K(\frac{X_1 - x}{h_n})w(\xi_1), \\ |\zeta_{n1}|\delta_n^{-1} \;&\leq\; c_{17}n^{-2/5}w(\xi_1).\end{aligned}$$

Hence

$$\begin{aligned}&E\{\zeta_{n1}^2 I_n\} \leq \\ &\leq c_{18}\int w^2(\xi)I(w(\xi) \geq c_{17}^{-1}\; n^{2/5})dF(\xi).\end{aligned} \tag{4.28}$$

It follows from (A6) that the integral $\int w^2(\xi)dF(\xi)$ is finite for $u' > o$ enough. Thus (4.28) tends to 0 as $n \to \infty$, and therefore (4.26) is true. This concludes the proof of (4.6).

REFERENCES

Collomb, G. (1977), "Quelques propriétés de la méthode du noyau pour l'estimation non paramétrique de la regression en un point fixé.", *C.R. Acad. Sci. Paris*, série A, 285, 289-292.

Cox, D.D., (1983), "Asymptotics for M-type smoothing splines.", *Ann. Statist.*, 11, 530-551.

Devroye, L. and L. Györfi, (1985), *"Nonparametric density estimation : the L_1-view."* Wiley, New-York.

Hall, P. and M.C. Jones, (1989), "Adaptive M-estimation in nonparametric regression", Manuscript.

Härdle, W., (1984a), "Robust regression function estimation.", *J. Multivariate Anal.*, 14, 169-180.

Härdle, W., (1984b), "How to determine the bandwidth of nonlinear smoothers in practice ?", In *Robust and Nonlinear Time Series Analysis* (J. Franke, W. Härdle, and D. Martin, eds.), Springer, New-York, pp.163-184.

Härdle, W. and T. Gasser, (1984), "Robust nonparametric function fitting.", *J. Roy. Statist. Soc.*, Ser. B., 46, 42-51.

Härdle, W. and Collomb, G., (1986), "Strong uniform convergence rates in robust nonparametric time series analysis and prediction: kernel regression estimation from dependent observations.", *Stochastic Processes and their Applications*, 23, 77-89.

Härdle, W. and A.B. Tsybakov, (1988), "Robust nonparametric regression with simultaneous scale curve estimation.", *Ann. Statist.*, 16, 120-135.

Huber, P.J., (1964), "Robust estimation of a location parameter.", *Ann. Math. Statist.*, 35, 73-101.

Ibragimov, I.A. and R.Z. Hasminskii, (1981), *"Statistical Estimation Asymptotic Theory."*, Springer, Berlin e.a.

Katkovnik, V.Ya., (1985), *"Nonparametric identification and smoothing of data."*, Nauka, Moscow (in Russian).

Loève, M., (1960), "*Probability theory.*" (2nd ed.), Van Nostrand, New York.

Mack, Y.P. and H.G. Müller, (1987), "Adaptive nonparametric estimation of a multivariate regression function.", *J. Multivar. Anal.*, 23, n.2, 169-182.

Müller, H.G., (1985), "Empirical bandwidth choice for nonparamatric kernel regression by means of pilot estimators.", *Statist. and Decisions*, Suppl. Issue, n.2, 193-206.

Müller, H.G. and Stadtmüller, U., (1987), "Variable bandwidth kernel estimators of regression curves.", *Ann. Statist.*, 15, n.1, 182-201.

Nadaraya, E.A., (1964), "On estimating regression.", *Probab. Theory Appl.*, 9, 141-142.

Nemirovskii, A.S., Polyak, B.T. and A.B. Tsybakov, (1983), "Estimators of maximum likelihood type for nonparametric regression.", *Soviet Math.*, Dokl., 28, 788-792.

Nemirovskii, A.S., Polyak, B.T. and A.B. Tsybakov, (1985), "Rate of convergence of nonparametric estimates of maximum likelihood type.", *Problems of Information Transmission*, 21, n.4, 258-272.

Parzen, E., (1962), "On estimation of a probability density function and mode.", *Ann. Math. Statist.*, 31, 1065-1076.

Prokhorov, Yu.V., (1956), "Convergence of random processes and the limit theorems of probability theory.", *Probab. Theory Appl.*, 1, n.2, 177-238.

Stone, C.J., (1980), "Optimal rates of convergence for nonparametric estimators.", *Ann. Statist.*, 8, n.6, 1348-1360.

Tsybakov, A.B., (1982a), "Nonparametric signal estimation when there is incomplete information on the noise distribution.", *Problems of Information Transmission*, 18, 116-130.

Tsybakov, A.B., (1982b), "Robust estimates of a function.", *Problems Inform. Transmission*, 18, 190-201.

Tsybakov, A.B., (1983), "Convergence of nonparametric robust algorithms of reconstruction of functions.", *Automation and Remote Control*, 44, 1582-1591.

Tsybakov, A.B., (1986), "Robust reconstruction of functions by the local-approximation method.", *Problems of Information Transmission*, 22, n.2, 133-146.

Tsybakov, A.B., (1987), "On bandwidth choice for kernel nonparametric regression.", *Probab. Theory Appl.*, 32, n.1, 142-148.

Watson, G.S., (1964), "Smooth regression analysis.", *Sankhya*, 26, 359-372.

Woodroofe, M. (1970), "On choosing a delta sequence.", *Ann. Math. Statist.*, 41, 1665-1671.

"Advanced in GLIM and Statistical Modelling", Springer Lecture Notes in Statistics

Nonparametric Approaches to Generalized Linear Models

Wolfgang K. Härdle
Berwin A. Turlach
C.O.R.E. and Institut de Statistique
Université Catholique de Louvain
B-1348 Louvain-la-Neuve, Belgium

1. Introduction and Motivation

In this paper we consider classes of statistical models that are natural generalizations of *generalized linear models.* Generalized linear models cover a very broad class of classical statistical models including linear regression, ANOVA, logit, and probit models. An important element of generalized linear models is that they contain parametric components of which the influence has to be determined by the experimentator. Here we describe some lines of thought and research relaxing the parametric structure of these components.

In generalized linear models response variable and explanatory variables are related by predetermined functional forms, e.g., the logit model with a logistic link function and a linear form on the explanatory variables, see McCullagh and Nelder (1989). In this example the fixed parametric structures are the logistic distribution function and the (linear) form of the influence of the explanatory variables. Generalizing such a type of model means to abandon the form of either of these fixed components, i.e., the logistic (inverse) *link function* or the *linear predictor.* Generalizing the form of the link function means to allow for a flexible or parameter free form. Generalizing the form of the linear predictor means to allow for any unknown function of the explanatory variables.

Allowing for any functional form of influence for the predictor variables leads into well known dimensionality problems, commonly called the *curse of dimensionality* (Huber 1985). In order to avoid this curse of dimensionality Hastie and Tibshirani (1990) proposed to generalize the linear predictor by a sum of non-parametric univariate functions. This leads to so called *generalized additive models.* They contain generalized linear models as a special case when the link function is known and the univariate functions operating on the explanatory variables are linear.

Relaxing the form of the link function means to keep the linear predictor but to replace, in terms of our previous example, the logistic function by a non-parametric (preferably monotone) function. More generally several of these types of response models can be added, each using a different linear predictor and (non-parametric) link function. These models are known as projection pursuit regression (PPR) models due to an algorithm developped by Friedman and Stützle (1981).

(1992) Härdle, W. and Turlach, B. Nonparametric Approaches
to Generalized Linear Models.

If we take just on term, i.e., an unknown (inverse) link function operating on a linear combination of the explanatory variables, this is called a one term projection pursuit model, in econometrics also called a *single index model.* Stoker (1992a, pp. 17 20) from an entirely economic point of view, considers labor supply leading to such a single index model.

2. Nonparametric Approaches to generalized linear models

We have argued that natural generalizations of generalized linear models are weakening and relaxing the link function or the linear form of the explanatory variables. To fix ideas let $X \in \mathbb{R}^d$ denote the explanatory variable and $Y \in \mathbb{R}$ be the response variable. A generalized linear model (GLM) connects the mean μ of Y with the predictor $\eta = X^T\beta$ via a link function G, i.e., $\mu = G(\eta)$. As a running examples we shall use the case of binary response models, i.e., $Y \in \{0, 1\}$. The GLM then reads as $P[Y = 1|x^T\beta = x] = G(x^T\beta)$. The aim is to estimate β when the link function G is fixed. Here and in the following we use the term *link* where McCullagh and Nelder (1989) mean the inverse link. Since in our examples this link is monotone there is no problem of confusion.

Single Index Models

Single Index Models keep the linear component but generalize the link. In our running example this reads as

$$P[Y = 1|X = x] = g\,(x^T\beta) \tag{2.1}$$

with g an unknown univariate "smooth" function. Note that here some standardization of the parameter β is asked for, since as such, (2.1) does not identify β but rather the direction of β. The aim here is to estimate β and the unknown link. For illustration of statistical and numerical procedures to be described later we would like to introduce

example 1.

$$\begin{aligned} &X \sim \mathcal{N}_2(0, I_2), \quad \beta = (1,1)^T \\ &g(\eta) = L(\eta) + \rho\varphi'(\eta), \quad L(\eta) = \exp(\eta)/[1 + \exp(\eta)] \\ &P[Y = 1|X = x] = g(x^T\beta) \end{aligned} \tag{2.2}$$

This model is almost a Logit model, only the skew deviation term $\rho\varphi'(\eta)$ makes it different from a GLM. For $\rho = 0$ it falls into the class of GLMs. For later illustrations we have set $\rho = 0.6$

and have generated $n = 200$ datapoints (x_i, y_i) according to (2.2). A graphical inspectation of the data gives a taste of nonparametric structure. Figure 1 shows a three dimensional scatterplot of the data. If we project the X variables in the 45° line we obtain Figure 2. This picture shows the projected data $x_i^T\beta$ against y_i together with the link $g(\eta)$. All these graphics and future computations were done in XploRe (1992).

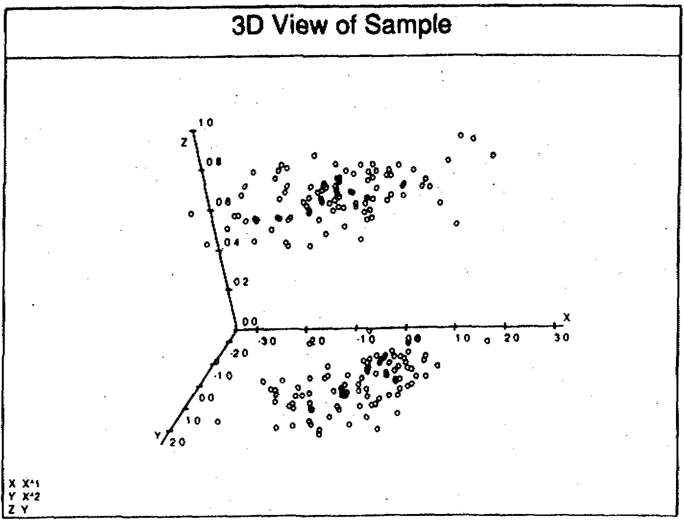


Figure 1: A three-dimensional scatterplot of the sample $\{(x_i^1, x_i^2, y_i)\}_{i=1}^{200}$ for example 1.

Generalized Additive Models

Generalized Additive Models keep the link but generalize the linear predictor to a sum of nonparametric functions. In our running example this reads as

$$P[Y = 1|X = x] = G\left(\alpha + \sum_{j=1}^{d} g_j(X^j)\right) \tag{2.3}$$

where X^j denotes the j^{th} component of the vector $X = (X^1, \dots, X^d)^T$ and the g_j are unknown univariate "smooth" functions. Again some standardization is necessary since the model as such does not identify the unknown $\{g_j\}_{j=1}^{d}$. The aim here is to estimate the nonparametric functions g_j. For illustrations of later techniques let us introduce

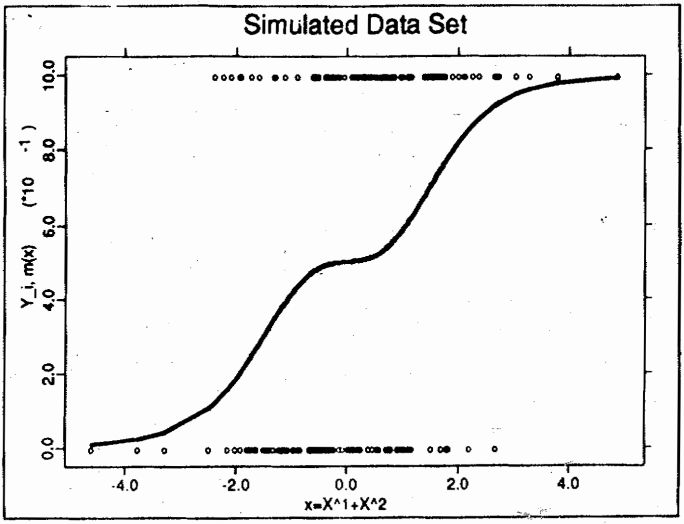


Figure 2: The observations y_i for example 1 plotted against $\eta = x^T\beta = x_i^1 + x_i^2$. The link $g(x) = L(\eta) + 0.6\varphi'(\eta)$ is shown as the solid line.

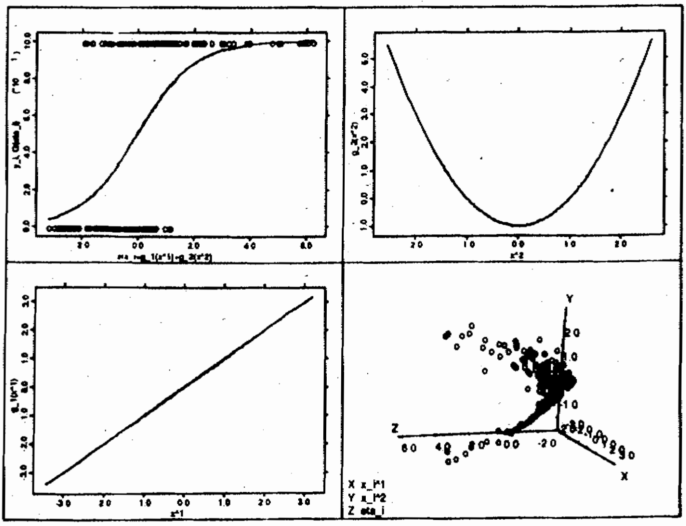


Figure 3: A four picture display with $\{(\eta_i, y_i)\}_{i=1}^{200}$ and $G(\eta_i)$ in the upper left. The nonparametric components are in the lower left and upper right. A rotated view of the surface $\{(x_i^1, x_i^2, \eta_i)\}_{i=1}^{200}$ is given in the lower right.

example 2.

$$\begin{aligned} &X \sim \mathcal{N}_2(0, I_2), \quad g_1(x^1) = x^1, \quad g_2(x^2) = \left[x^2\right]^2 - 1 \\ &G = L \\ &P[Y = 1|X = x] = G(g_1(x^1) + g_2(x^2)) \end{aligned} \tag{2.4}$$

This model is almost a Logit model, only the second predictor variable has a nonlinear influence on η. Figure 3 shows a four picture display with the data $\{(\eta_i, y_i)\}_{i=1}^{200}$ in the upper left corner together with the Logistic link. Note that the predictor is $\eta_i = g_1(x_1^1) + g_2(x_1^2)$. The "nonparametric" components g_j are shown in the lower left und the upper right. An impression of the nonlinear components can be gained by rotating the three dimensional surface $\{(x_i^1, x_i^2, \eta_i)\}_{i=1}^{200}$.

3. Single Index Models

Model (2.1) is called a single index model or a one term projection pursuit model. This terminology is due to Friedman and Stützle (1981) who considered the more general model:

$$P[Y = 1|X = x] = \sum_{j=1}^{K} g_j(x^T \beta_j)$$

where the $\beta_j \in \mathbb{R}^d$ are unknown parameters and the g_j's are unknown functions, satisfying some "smoothness" assumptions. In order to make the β_j's and the g_j's identifiable one has to impose restrictions on the scale, usually $\|\beta\| = 1$, or $\beta^1 = 1$.

Friedman and Stützle (1981) proposed to estimate K, β_j and g_j by the method of "Projection Pursuit Regression"(PPR) algorithm. This procedure estimates terms $g_j(X^T\beta_j)$ as long as the fraction of unexplained variance is below a userspecified treshhold. In each step that β_j is choosen which maximizes the fraction of unexplained variance given the previous terms (*projection pursuit*). The fitted model after convergence is

$$P[Y = 1|X = x] = \sum_{j=1}^{\hat{K}} \hat{g}_j(x_j^T \hat{\beta}_j).$$

From a mathematical point of view, a drawback of this method is, that it is not clear which value of K is to be chosen. Research has therefore focussed on one term projection pursuit models. In this line Hall (1989) constructs a root-n consistent estimator of β. A different method is that of Härdle and Stoker (1989) also called ADE for Average Derivative Estimation. It is based on the following idea. Define $m(x) := g(x^T\beta)$ and observe that for the average derivative δ, as defined below, we have

$$\delta := E_X[m'(X)] = E_X[\frac{dg}{d(x^T\beta)}(X^T\beta)]\beta. \tag{3.1}$$

Thus δ determines β up to scale. Let $f(x)$ denote the density of X and l its vector of the negative log-derivatives (partial), $l = -\frac{\partial \log f}{\partial x} = -\frac{f'}{f}$ (l is also called *score vector*). Under assumptions on f this enables us to write

$$\delta = E[m'(X)] = E[lY] \tag{3.2}$$

and to estimate δ by $\hat{\delta} = n^{-1}\sum_{i=1}^{n} \hat{l}_h(x_i)y_i$. Here $\hat{l}_h$ is an estimator of l based on a kernel density smoother with bandwidth h. For an easy access to kernel density smoothing see the book by Silverman (1986). With root-n estimates for δ precise estimates for the link can be obtained. The convergence rate for g is one dimensional, however in practice there remains the problem of selecting the bandwidth h. This was investigated in Härdle, Hart, Marron, and Tsybakov (1992) and for a weighted average derivative by Härdle and Tsybakov (1991). Stoker (1991) proposed alternative estimators for δ. A Monte Carlo comparison of these different methods was done by Stoker and Villas-Boas (1992b).

The estimation of the score vector l via a kernel density estimator involves a number of intensive calculations, especially when we optimize over h. Therefore discretization or WARPing ideas should be used (Turlach 1992). For our simulated example Figure 4 shows the result of this method. We calculated $\hat{\delta}$ and used the *Nadaraya-Watson* regression estimator to estimate $\hat{g}$. Note that the horizontal scale on this figure is different since (3.1) suggest that δ has different scale then β. In fact for ADE the scale of δ changes with g but it does not matter for the statistical interpretation of the link g that we are interested in.

The estimation of δ and its asymptotic covariance matrix $\hat{\Sigma}_\delta$ for example 1 was done with Program 1 in Section 5. Note that for this example we have $\delta = \binom{0.135}{0.135}$. The binning parameter d was chosen in such a way that maximal 20 bins were used in each coordinate, i.e., $d \approx \binom{0.33}{0.256}$. The estimate for the average derivative and the asymptotic covariance matrix was calculated using the three adjacent bins which equals a bandwidth $h \approx \binom{0.99}{0.78}$. As result we have

$$\hat{\delta} = \begin{pmatrix} 0.124 \\ 0.118 \end{pmatrix}, \qquad \hat{\Sigma}_\delta = \begin{pmatrix} 0.188 & 0.036 \\ 0.036 & 0.208 \end{pmatrix}.$$

These results allow us to test some hypothesis formally using a Wald statistic (see Stoker (1992a), pp. 53–54). In particular, to test the restriction $R\delta = r_0$, the Wald statistic

$$W = n(R\hat{\delta} - r_0)^T(R\hat{\Sigma}_\delta R^T)^{-1}(R\hat{\delta} - r_0)$$

is compared to a χ^2(rank R) critical value. Table 3.1 gives some examples for this technique.

Restriction	Value W	d.f.	$P[\chi^2(\text{d.f.}) > W]$
$\delta^1 = \delta^2 = 0$	25.25	2	0
$\delta^1 = \delta^2 = 0.135$	0.365	2	0.83
$\delta^1 = \delta^2$	0.126	1	0.72

Table 3.1: Wald Statistics for some restrictions on δ.

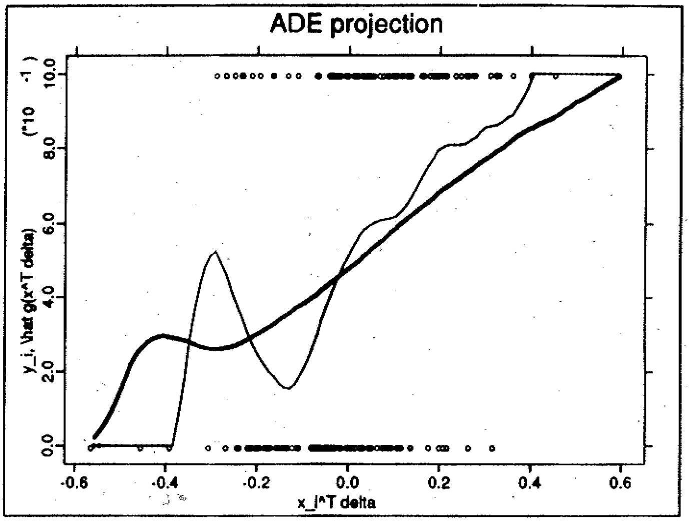


Figure 4: For the simulated data set of example 1 $x_i^T\hat{\delta}$ vs. y_i and two estimates of $\hat{g}(x_i^T\hat{\delta})$ are shown. The thick line shows the Nadaraya-Watson regression estimator for $\hat{g}$ with a bandwidth of $h = 0.3$, for the thin line $h = 0.1$ was chosen.

Another method to estimate g and β in (2.1) was proposed by Ichimura (1992). Let ε denote the error term inherent to the response variable. Observing that (β_0 denotes the true parameter):

(1) The variation in Y results from both the variation in $X^T\beta_0$ and the variation in ε.

(2) On the contour line $X^T\beta_0 = c$, where c is a given constant, the variability in Y results only from the variation in ε.

(3) Observation (2) does not necessarily hold on a contour line defined by $X^T\beta = c$ for $\beta \neq \beta_0$. Along this contour line, the value of $X^T\beta_0$ changes and therefore the variability in Y again results from the variation in both $X^T\beta_0$ and ε.

To identify β_0 Ichimura thus proposes to estimate the conditional variance

$$Var\left[Y|X^T\beta = c\right] = E\left[\left\{Y - E\left[Y|X^T\beta = c\right]\right\}^2\right] \tag{3.3}$$

by estimating $E\left[Y|X^T\beta = c\right]$ by a kernel estimator and to find than the vector β that minimizes (3.3). Härdle, Hall and Ichimura (1991) proposed a simple and effective crossvalidation method for this setting which yields a root-n consistent estimator of β_0 and an asymptotically optimal estimator of h_0, the bandwidth which should be used to calculate the kernel estimate of g.

A way of testing a GLM against this specific single index alternative has been given by Horowitz and Härdle (1992). They constructed a conditional moment test based on ideas of Bierens (1990) and Newey (1985). Another approach for such a test via Bootstrapping ideas was investigated by Rodríguez-Campos and Cao-Abad (1992).

4. Generalized Additive Models

A generalized additive model differs from a GLM in that an additive predictor replaces the linear predictor η. The estimation of this model is usually a highly iterative procedure. Estimation of α and $g_1, \ldots, g_d$ in (2.3) is accomplished by replacing the weighted linear regression in the adjusted dependent variable by an appropriate algorithm for fitting a weighted additive model (Hastie and Tibshirani 1990). This iterative fitting of a weighted additive model is known as *local scoring* since it generalizes the Fisher scoring procedure. Each estimation of a weighted additive model is done in an iterative process known as *backfitting*. In the backfitting step non-parametric estimates for $g_1, \ldots, g_d$ are calculated. The properties of the backfitting algorithm have been studied by Craig and Kohn (1991) or Härdle and Hall (1992) for example.

More specifically we have to estimate functions g_j in the model

$$P[Y = 1|X = x] = G\left(\alpha + \sum_{j=1}^{d} g_j(X^j)\right).$$

The explicit algorithm of finding the nonparametric components is given by (see Hastie and Tibshirani 1987):

Local Scoring Algorithm

```
Initialization  $\hat{f}_j^{(0)} \equiv 0$ for $j = 1, \ldots, d$, $\hat{\alpha}^{(0)}$ =logit$(\bar{y})$.
Loop            over outer interation counter $m$
                $\hat{\eta}^{(m)}(x_i) = \hat{\alpha}^{(m)} + \sum_{j=1}^{d} \hat{f}_j^{(m)}(x_i^j)$
                $\hat{p}_i$ =logit$^{-1}(\hat{\eta}^{(m)}(x_i))$
                $z_i = \hat{\eta}^{(m)}(x_i) + (y_i - \hat{p}_i)/[\hat{p}_i(1 - \hat{p}_i)]$
                $w_i = \hat{p}_i(1 - \hat{p}_i), \qquad i = 1, \ldots, n.$
                Obtain $\hat{\alpha}^{(m+1)}, \hat{f}_j^{(m+1)}, j = 1, \ldots, d$ by applying the backfitting
                algorithm to $z_i$ with explanatory variables $x_i$ and observation weights $w_i$.
until           the deviance $D(y, \hat{p}) = -2 \sum_i [y_i \log(\hat{p}_i) + (1 - y_i) \log(1 - \hat{p}_i)]$ converges.
```

Backfitting Algorithm

```
Initialization  $\hat{f}_j^{(0)} \equiv 0$ for $j = 1, \ldots, d$, $\hat{\alpha}^{(0)} = \bar{y}$.
Repeat          for $j = 1, \ldots, d$ repeat such cycles:
                $r_i = y_i - \hat{\alpha} - \sum_{k=1, k \neq j}^{d} \hat{f}_k(x_i^k) \quad i = 1, \ldots, n$
                $\hat{f}_j(x_i^j) = S(r|w, x_i^j) \quad i = 1, \ldots, n$
Until           $RSS = \sum_{i=1}^{n} \left( y_i - \hat{\alpha} - \sum_{j=1}^{d} \hat{f}_j(x_i^j) \right)^2$ converges.
```

Here $S(r|w, x_i^j)$ denotes the value of the function obtained by smoothing the scatterplot (r, x) with weights w at the point x_i.

Since non-parametric estimation methods are used in the backfitting step two main problems arise. The first problem is how to choose the smoothing parameter in this non-parametric fit regardless whether splines, kernel estimators or others are used, see Buja, Hastie and Tibshirani (1989). The second problem is, since the whole process is iterative, how to make the calculations of the non-parametric fits as fast as possible.

For kernel regression estimates this leads to WARPing (see Scott 1985, Härdle and Scott 1992, and Fan and Marron 1992). A third problem is how to incorporate the weights in the non-parametric smoothing step (see Hastie and Tibshirani 1990, pp. 72-74). Especially in logistic models, as we discuss them here, these weights can cause numeric problems. If the estimated probability $\hat{p}_i = P[Y_i = 1|X]$ is very close to 0 or 1 the weight for this observation in the backfitting step will be very small. But the adjusted dependent variable will be very big resulting in a big partial residual. This can result in a bad fit within the backfitting algorithm which leads in the next step of the local scoring to the same problem.

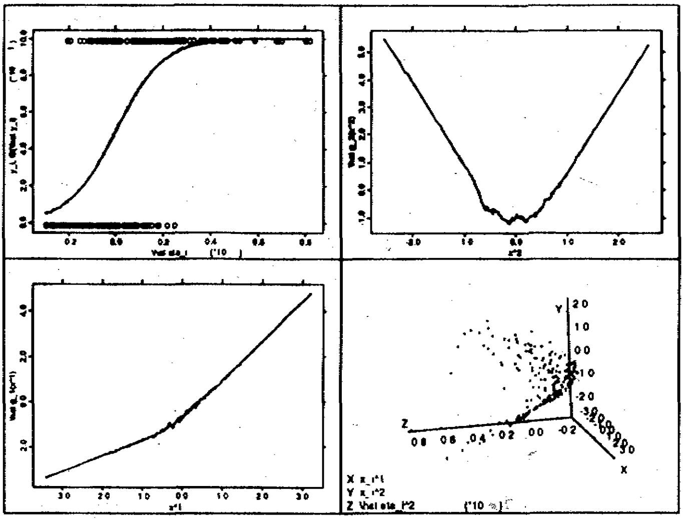


Figure 5: A four picture display with the results of the fitting procedure for the Generalized Additive Model. Legend is the same as for Figure 3 where η_i is replaced by $\hat{\eta}_i$.

Program 3 in Section 5 demonstrates how the Generalized Additive Model can be estimated in XploRe (1992). The result of this fitting is visualized in Figure 5. The backfitting algorithm provides estimates of the function g_j in the multiple additive regression model $E[y|x] = \alpha + \sum_j g_j(x^j)$ with $E[g_j(x^j)] = 0$ for $j = 1, \ldots, d$. It is easily seen that in example 2 given by (2.4) we have $E[g_j(x^j)] = 0$, $j = 1, 2$. Thus for our example we would expect that α is estimated as 0. In fact the result is $\hat{\alpha} = 0.25$.

5. The implementation in XploRe

The above calculations have been performed in the language XploRe (1992). In this section we give some programs that are useful in solving the iterative procedure for Generalized Additive Models for example or for ADE. The Single Index Model for example 1 has been estimated using the ADE technique with the following program.

```
library(smoother)                         ;load the necessary libraries
library(addmod)
randomize(0)
```

```
x = normal(200 2)                          ;generate the explanatory variable
rho = 0.6
beta = #(1 1)
eta = x*beta                               ;eta, notation as in (2.2)
g = 1./(1+exp(-eta)) - rho * eta.*pdfn(eta);calculate g(eta)
u = uniform (200 2)
y = u.<g                                   ;generate the response variable
d = (max(x)-min(x))/20                     ;choosing a binning parameter
(xb yb) = bindata(x d 0 y)                 ;binning the data
(del dvar) = adeind(xb yb d 3)
    ;estimate the average derivative and the asymptotic covariance matrix
est = (x*del)~y  ;calculate the projection
gh1 = regest(est 0.1)                      ;find estimates for g
gh2 = regest(est 0.3)
show(est gh1 gh2 s2d)                      ;show results (Picture 4)
```

Program 1: This program generates and estimates example 1

The commands of XploRe (1992) are similar to GAUSS but more fine tuned for smoothing and nonparametric methods in high dimensions. The Generalized Additive Model (GAM) of example 2 was created using the following code:

```
randomize(0)
x = normal(200 2)
g1 = x[,1]
g2 = x[,2].*x[,2]-1
eta = g1+g2
px = 1./(1+exp(-px))
u = uniform(200)
y = u.<px
createdisplay(pic3, 2 2, s2d s2d s2d d3d)
show(eta~y eta~px s2d1, x[,1]~g1 s2d2, x[,2]~g2 s2d3, x~eta d3d1)
```

Program 2: This program generates Picture 3

The estimation of the GAM was done by

```
proc(fx alpha dev)=lscore(x y)
  dim = cols(x)
  gx = matrix(rows(x) dim 0)                       ;initialize g_j
  xs = 1                                           ;used to store information
                                                   ;to sort the covariates
  ybar = mean(y)                                   ;initialize alpha
  alpha = ln(ybar/(1-ybar))
  loop = 1
  devold = 0
  dev = 100000;
```

```
    while( (abs(dev-devold) > 0.01) && (loop < 6) )
      eta = alpha + sumr(gx)
      p = 1./(1+exp(-eta))
      w = p.*(1-p)                                   ;calculate the weights
      z = eta + (y-p)./w                             ;calculate the adjusted
                                                     ;dependent variable
      (gx alpha xs)=backfit(x z xs w 0.4)            ;the backfitting step
      devold = dev
      dev = -2*sum(y.*ln(p)+(1-y).*ln(1-p))          ;calculate the deviance
      loop = loop+1
    endo
  endp
```

Program 3: This program implements the Local Scoring Algorithm

REFERENCES

Ansley, C.F. and Kohn R. (1991), "Convergence of the Backfitting Algorithm for Additive Models," *Working Paper 91-013*, Australian Graduate School of Management.

Bierens, H.J. (1990), "A consistent conditional moment test of functional form," *Econometrica*, **58**, 1443-1458.

Breiman, L. and Friedman, J.H. (1985), "Estimating optimal transformations for multiple regression and correlation (with discussion)," *Journal of the American Statistical Association*, **80**, 580-619.

Buja, A., Hastie, T.J. and Tibshirani, R.J. (1989), " Linear smoothers and additive models (with discussion)," *The Annals of Statistics*, **17**, 453-555

Fan, J. and Marron, J. S. (1992) "Fast implementations of nonparametric curve estimators", unpublished manuscript.

Friedman, J.H. and Stuetzle, W. (1981), "Projection Pursuit Regression," *Journal of the American Statistical Association*, **76**, 817-823.

Huber, P.J. (1985), "Projection Pursuit," *The Annals of Statistics*, **13**, 435 475.

Härdle, W., Hart J., Marron J.S., and Tsybakov, A.B. (1992), "Bandwidth Choice for Average Derivative Estimation," *Journal of the American Statistical Association*, **87**, 817 823.

Härdle, W. and Hall, P. (1992), "Simple formulae for steps and limits in the backfitting algorithm," *Statistica Neerlandica*, to appear.

Härdle, W., Hall, P. and Ichimura, H. (1991), "Optimal Smoothing in Single Index Models," *Core Discussion Paper N° 9107.*

Härdle, W. and Scott, D.W. (1992), "Smoothing by Weighted Averaging of Rounded Points," *Computational Statistics*, in print.

Härdle, W. and Stoker, T.M. (1989), "Investigating Smooth Multiple Regression by the Method of Average Derivatives," *Journal of the American Statistical Association*, **84**, 986 995.

Härdle, W. and Tsybakov, A.B. (1991), "How sensitive are average derivatives?," *CORE Discussion Paper N° 9144.*

Hall, P. (1989), "On Projection Pursuit Regression", *The Annals of Statistics*, **17**, 573 588

Hastie, T.J. and Tibshirani, R.J. (1987), "Non-parametric Logistic and Proportional Odds Regression," *Applied Statistics*, **36**, 260-276.

Hastie, T.J. and Tibshirani, R.J. (1990), *Generalized Additive Models*, Chapman and Hall, London.

Horowitz, J. and Härdle, W. (1992), " Testing a parametric model against a semiparametric alternative", *CentER Discussion Paper.*

Ichimura, H. (1992), "Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single-Index Models," *Journal of Econometrics*, special issue on "Nonparametric Approaches to Discrete Choice Models", ed. W. Härdle and C.F. Manski.

Newey, W.K. (1985), "Maximum likelihood specification testing and conditional moment test," *Econometrica*, **53**, 1047–1070.

McCullagh, P. and Nelder, J.A. (1989), *Generalized Linear Models, 2nd Edition*, Chapman and Hall, London.

Rodríguez-Campos, M.C. and Cao-Abad, R. (1992),"Nonparametric Bootstrap Confidence Intervals for Discrete Regression Functions," *Journal of Econometrics*, special issue on "Nonparametric Approaches to Discrete Choice Models", ed. W. Härdle and C.F. Manski.

Scott, D.W. (1985), "Averaged Shifted Histograms: Effective Nonparametric Density Estimators in Several Dimensions," *The Annals of Statistics*, **13**, 1024-1040

Silverman, B.W. (1986), '*Density Estimation for Statistical and Data Analysis*, Chapman and Hall, London.

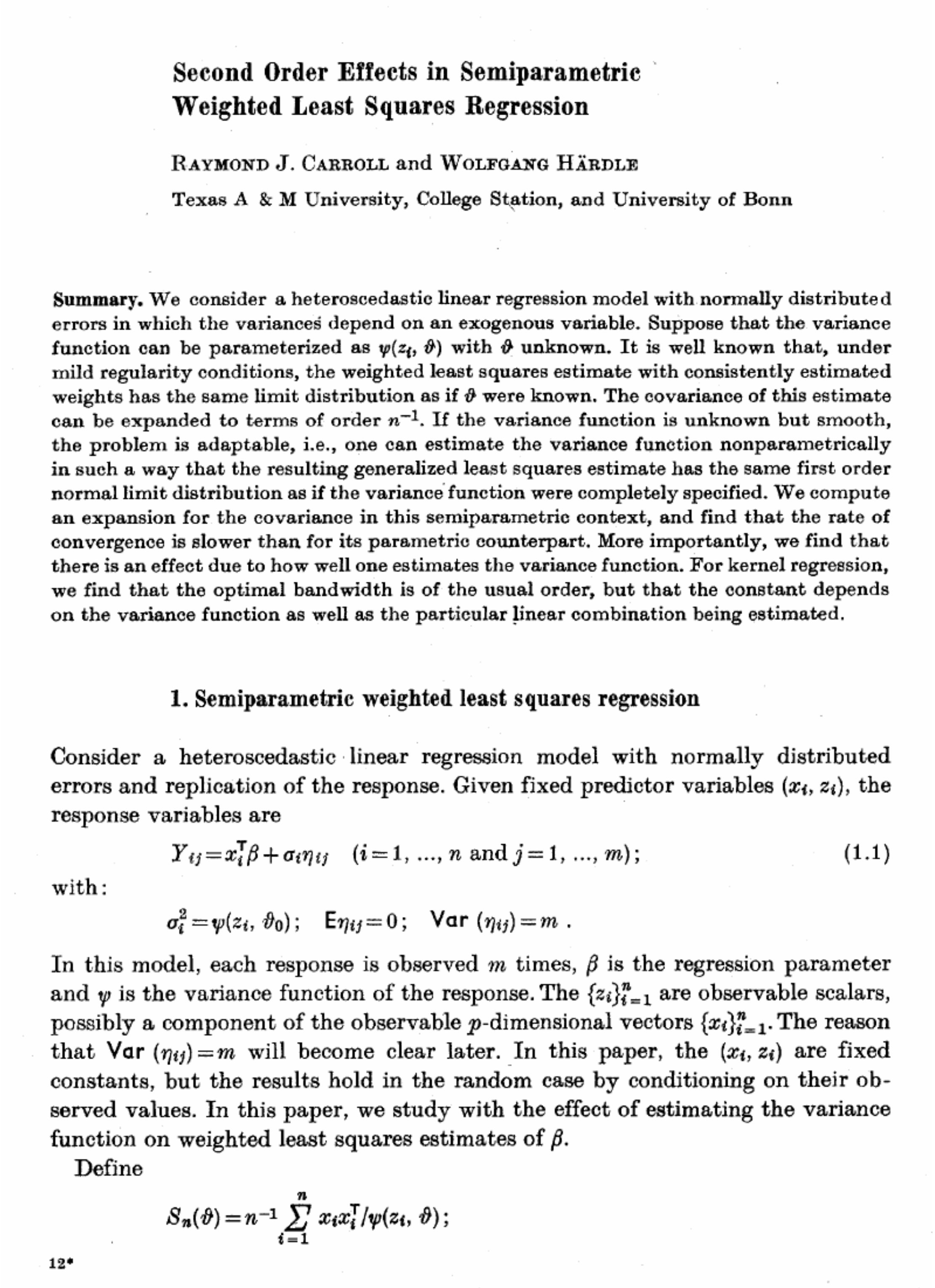
Stoker, T.M. (1991), "Equivalence of Direct, Indirect and Slope Estimators of Average Derivatives," in *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, Barnett, W.A., J.L. Powell and G.Tauchen, eds., Cambridge University Press.

Stoker, T.M. (1992a), *Lectures on Semiparametrics Econometrics*, CORE Lecture Series, Louvain-la-Neuve.

Stoker, T.M. and Villas-Boas, J.M. (1992b), "Monte Carlo Simulation of Average Derivative Estimators," *Discussion Paper*, Sloan School of Management, MIT.

Turlach, B. (1992), "Discretization Methods in high-dimensional smoothing," *CORE Discussion Paper.*

XploRe (1992), XploRe 3.0 — a computing environment for eXploratory Regression and data analysis. Available from XploRe Systems, C.O.R.E. Université Catholique de Louvain, Belgium.



Second Order Effects in Semiparametric Weighted Least Squares Regression

Raymond J. Carroll and Wolfgang Härdle

Texas A & M University, College Station, and University of Bonn

Summary. We consider a heteroscedastic linear regression model with normally distributed errors in which the variances depend on an exogenous variable. Suppose that the variance function can be parameterized as $\psi(z_i, \vartheta)$ with ϑ unknown. It is well known that, under mild regularity conditions, the weighted least squares estimate with consistently estimated weights has the same limit distribution as if ϑ were known. The covariance of this estimate can be expanded to terms of order n^{-1}. If the variance function is unknown but smooth, the problem is adaptable, i.e., one can estimate the variance function nonparametrically in such a way that the resulting generalized least squares estimate has the same first order normal limit distribution as if the variance function were completely specified. We compute an expansion for the covariance in this semiparametric context, and find that the rate of convergence is slower than for its parametric counterpart. More importantly, we find that there is an effect due to how well one estimates the variance function. For kernel regression, we find that the optimal bandwidth is of the usual order, but that the constant depends on the variance function as well as the particular linear combination being estimated.

1. Semiparametric weighted least squares regression

Consider a heteroscedastic linear regression model with normally distributed errors and replication of the response. Given fixed predictor variables (x_i, z_i), the response variables are

$$Y_{ij}=x_i^{\mathsf T}\beta+\sigma_i\eta_{ij} \quad (i=1, \ldots, n \text{ and } j=1, \ldots, m); \qquad (1.1)$$

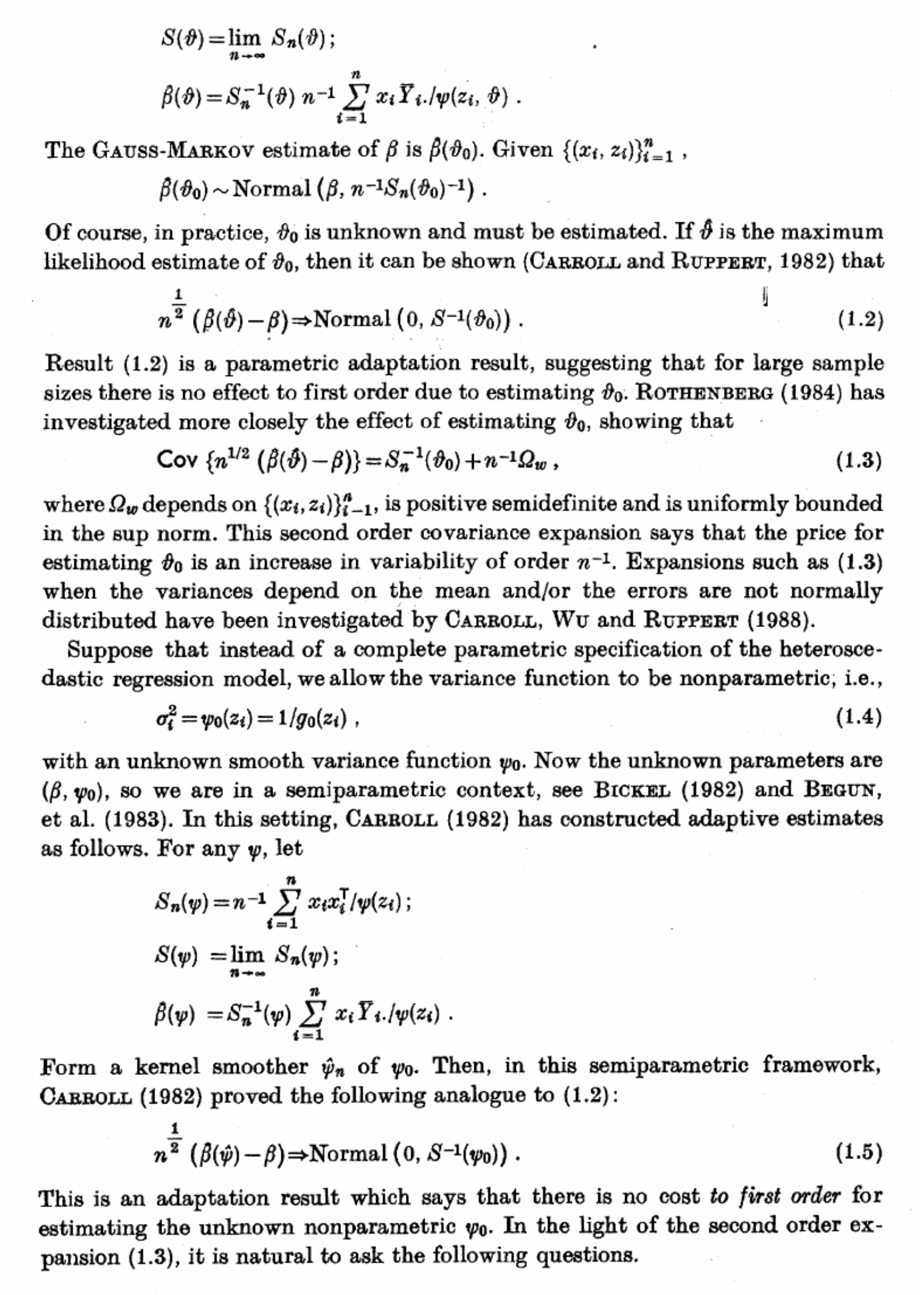
with:

$$\sigma_i^2=\psi(z_i, \vartheta_0); \quad \mathsf{E}\eta_{ij}=0; \quad \mathsf{Var}\,(\eta_{ij})=m\ .$$

In this model, each response is observed m times, β is the regression parameter and ψ is the variance function of the response. The $\{z_i\}_{i=1}^n$ are observable scalars, possibly a component of the observable p-dimensional vectors $\{x_i\}_{i=1}^n$. The reason that $\mathsf{Var}\,(\eta_{ij})=m$ will become clear later. In this paper, the (x_i, z_i) are fixed constants, but the results hold in the random case by conditioning on their observed values. In this paper, we study with the effect of estimating the variance function on weighted least squares estimates of β.

Define

$$S_n(\vartheta)=n^{-1}\sum_{i=1}^{n} x_i x_i^{\mathsf T}/\psi(z_i, \vartheta);$$



$$S(\vartheta) = \lim_{n\to\infty} S_n(\vartheta);$$

$$\beta(\vartheta) = S_n^{-1}(\vartheta)\, n^{-1} \sum_{i=1}^{n} x_i \bar{Y}_{i\cdot}/\psi(z_i, \vartheta)\ .$$

The Gauss-Markov estimate of β is $\hat{\beta}(\vartheta_0)$. Given $\{(x_i, z_i)\}_{i=1}^{n}$,

$$\hat{\beta}(\vartheta_0) \sim \text{Normal}\,(\beta,\, n^{-1}S_n(\vartheta_0)^{-1})\ .$$

Of course, in practice, ϑ_0 is unknown and must be estimated. If $\hat{\vartheta}$ is the maximum likelihood estimate of ϑ_0, then it can be shown (Carroll and Ruppert, 1982) that

$$n^{\frac{1}{2}}\,(\hat{\beta}(\hat{\vartheta}) - \beta) \Rightarrow \text{Normal}\,(0,\, S^{-1}(\vartheta_0))\ . \tag{1.2}$$

Result (1.2) is a parametric adaptation result, suggesting that for large sample sizes there is no effect to first order due to estimating ϑ_0. Rothenberg (1984) has investigated more closely the effect of estimating ϑ_0, showing that

$$\text{Cov}\,\{n^{1/2}\,(\hat{\beta}(\hat{\vartheta}) - \beta)\} = S_n^{-1}(\vartheta_0) + n^{-1}\Omega_w\ , \tag{1.3}$$

where Ω_w depends on $\{(x_i, z_i)\}_{i-1}^{n}$, is positive semidefinite and is uniformly bounded in the sup norm. This second order covariance expansion says that the price for estimating ϑ_0 is an increase in variability of order n^{-1}. Expansions such as (1.3) when the variances depend on the mean and/or the errors are not normally distributed have been investigated by Carroll, Wu and Ruppert (1988).

Suppose that instead of a complete parametric specification of the heteroscedastic regression model, we allow the variance function to be nonparametric, i.e.,

$$\sigma_i^2 = \psi_0(z_i) = 1/g_0(z_i)\ , \tag{1.4}$$

with an unknown smooth variance function ψ_0. Now the unknown parameters are (β, ψ_0), so we are in a semiparametric context, see Bickel (1982) and Begun, et al. (1983). In this setting, Carroll (1982) has constructed adaptive estimates as follows. For any ψ, let

$$S_n(\psi) = n^{-1} \sum_{i=1}^{n} x_i x_i^{\mathsf{T}}/\psi(z_i)\,;$$

$$S(\psi) \ = \lim_{n\to\infty} S_n(\psi);$$

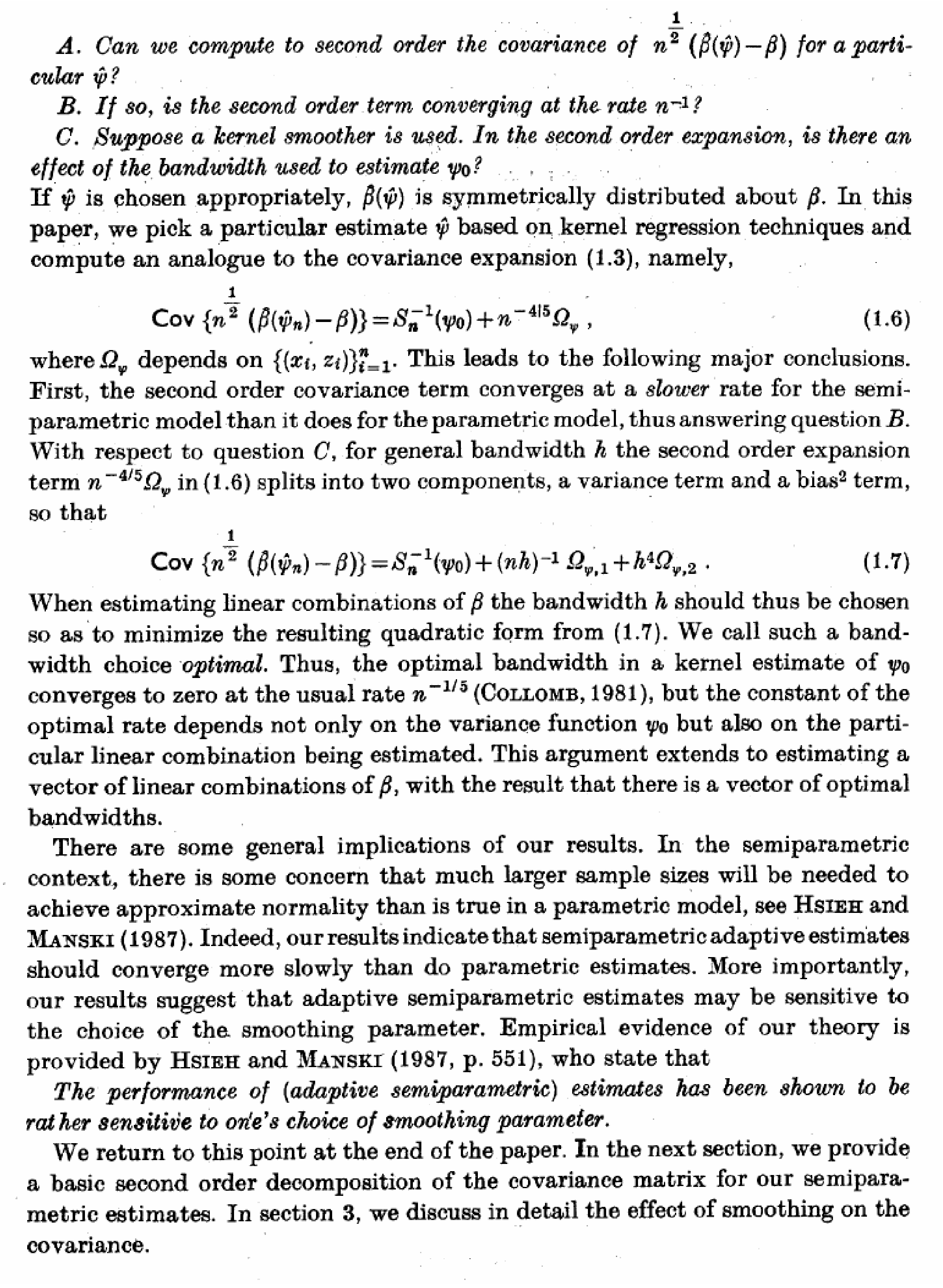
$$\hat{\beta}(\psi) \ = S_n^{-1}(\psi) \sum_{i=1}^{n} x_i \bar{Y}_{i\cdot}/\psi(z_i)\ .$$

Form a kernel smoother $\hat{\psi}_n$ of ψ_0. Then, in this semiparametric framework, Carroll (1982) proved the following analogue to (1.2):

$$n^{\frac{1}{2}}\,(\hat{\beta}(\hat{\psi}) - \beta) \Rightarrow \text{Normal}\,(0,\, S^{-1}(\psi_0))\ . \tag{1.5}$$

This is an adaptation result which says that there is no cost *to first order* for estimating the unknown nonparametric ψ_0. In the light of the second order expansion (1.3), it is natural to ask the following questions.

Carroll, R. and Härdle, W. (1989) A note on second order effects in a semiparametric context



A. Can we compute to second order the covariance of $n^{\frac{1}{2}}(\hat\beta(\hat\psi)-\beta)$ *for a particular* $\hat\psi$*?*

B. If so, is the second order term converging at the rate n^{-1}*?*

C. Suppose a kernel smoother is used. In the second order expansion, is there an effect of the bandwidth used to estimate ψ_0*?*

If $\hat\psi$ is chosen appropriately, $\hat\beta(\hat\psi)$ is symmetrically distributed about β. In this paper, we pick a particular estimate $\hat\psi$ based on kernel regression techniques and compute an analogue to the covariance expansion (1.3), namely,

$$\mathrm{Cov}\,\{n^{\frac{1}{2}}(\hat\beta(\hat\psi_n)-\beta)\}=S_n^{-1}(\psi_0)+n^{-4/5}\Omega_\psi\,, \tag{1.6}$$

where Ω_ψ depends on $\{(x_i, z_i)\}_{i=1}^{n}$. This leads to the following major conclusions. First, the second order covariance term converges at a *slower* rate for the semiparametric model than it does for the parametric model, thus answering question B. With respect to question C, for general bandwidth h the second order expansion term $n^{-4/5}\Omega_\psi$ in (1.6) splits into two components, a variance term and a bias^2 term, so that

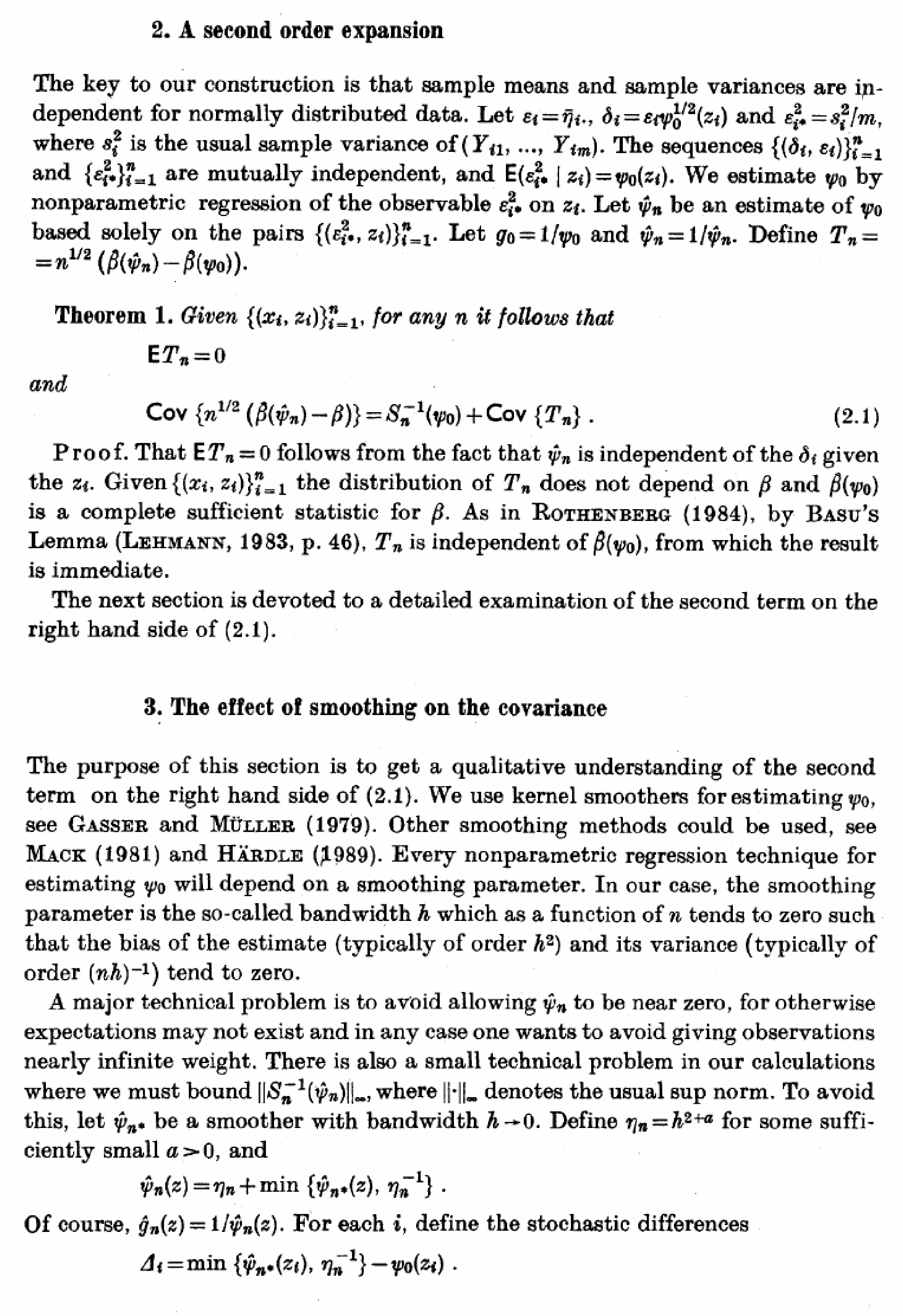
$$\mathrm{Cov}\,\{n^{\frac{1}{2}}(\hat\beta(\hat\psi_n)-\beta)\}=S_n^{-1}(\psi_0)+(nh)^{-1}\,\Omega_{\psi,1}+h^4\Omega_{\psi,2}\,. \tag{1.7}$$

When estimating linear combinations of β the bandwidth h should thus be chosen so as to minimize the resulting quadratic form from (1.7). We call such a bandwidth choice *optimal*. Thus, the optimal bandwidth in a kernel estimate of ψ_0 converges to zero at the usual rate $n^{-1/5}$ (Collomb, 1981), but the constant of the optimal rate depends not only on the variance function ψ_0 but also on the particular linear combination being estimated. This argument extends to estimating a vector of linear combinations of β, with the result that there is a vector of optimal bandwidths.

There are some general implications of our results. In the semiparametric context, there is some concern that much larger sample sizes will be needed to achieve approximate normality than is true in a parametric model, see Hsieh and Manski (1987). Indeed, our results indicate that semiparametric adaptive estimates should converge more slowly than do parametric estimates. More importantly, our results suggest that adaptive semiparametric estimates may be sensitive to the choice of the smoothing parameter. Empirical evidence of our theory is provided by Hsieh and Manski (1987, p. 551), who state that

The performance of (adaptive semiparametric) estimates has been shown to be rather sensitive to one's choice of smoothing parameter.

We return to this point at the end of the paper. In the next section, we provide a basic second order decomposition of the covariance matrix for our semiparametric estimates. In section 3, we discuss in detail the effect of smoothing on the covariance.



2. A second order expansion

The key to our construction is that sample means and sample variances are independent for normally distributed data. Let $\varepsilon_i = \bar{\eta}_{i\cdot}$, $\delta_i = \varepsilon_i \psi_0^{1/2}(z_i)$ and $\varepsilon_{i*}^2 = s_i^2/m$, where s_i^2 is the usual sample variance of $(Y_{i1}, \ldots, Y_{im})$. The sequences $\{(\delta_i, \varepsilon_i)\}_{i=1}^n$ and $\{\varepsilon_{i*}^2\}_{i=1}^n$ are mutually independent, and $\mathsf{E}(\varepsilon_{i*}^2 \mid z_i) = \psi_0(z_i)$. We estimate ψ_0 by nonparametric regression of the observable ε_{i*}^2 on z_i. Let $\hat{\psi}_n$ be an estimate of ψ_0 based solely on the pairs $\{(\varepsilon_{i*}^2, z_i)\}_{i=1}^n$. Let $g_0 = 1/\psi_0$ and $\hat{\psi}_n = 1/\hat{\psi}_n$. Define $T_n = = n^{1/2}(\hat{\beta}(\hat{\psi}_n) - \hat{\beta}(\psi_0))$.

Theorem 1. *Given $\{(x_i, z_i)\}_{i=1}^n$, for any n it follows that*

$$\mathsf{E}T_n = 0$$

and

$$\mathsf{Cov}\,\{n^{1/2}(\hat{\beta}(\hat{\psi}_n) - \beta)\} = S_n^{-1}(\psi_0) + \mathsf{Cov}\,\{T_n\}\,. \qquad (2.1)$$

Proof. That $\mathsf{E}T_n = 0$ follows from the fact that $\hat{\psi}_n$ is independent of the δ_i given the z_i. Given $\{(x_i, z_i)\}_{i=1}^n$ the distribution of T_n does not depend on β and $\hat{\beta}(\psi_0)$ is a complete sufficient statistic for β. As in ROTHENBERG (1984), by BASU's Lemma (LEHMANN, 1983, p. 46), T_n is independent of $\hat{\beta}(\psi_0)$, from which the result is immediate.

The next section is devoted to a detailed examination of the second term on the right hand side of (2.1).

3. The effect of smoothing on the covariance

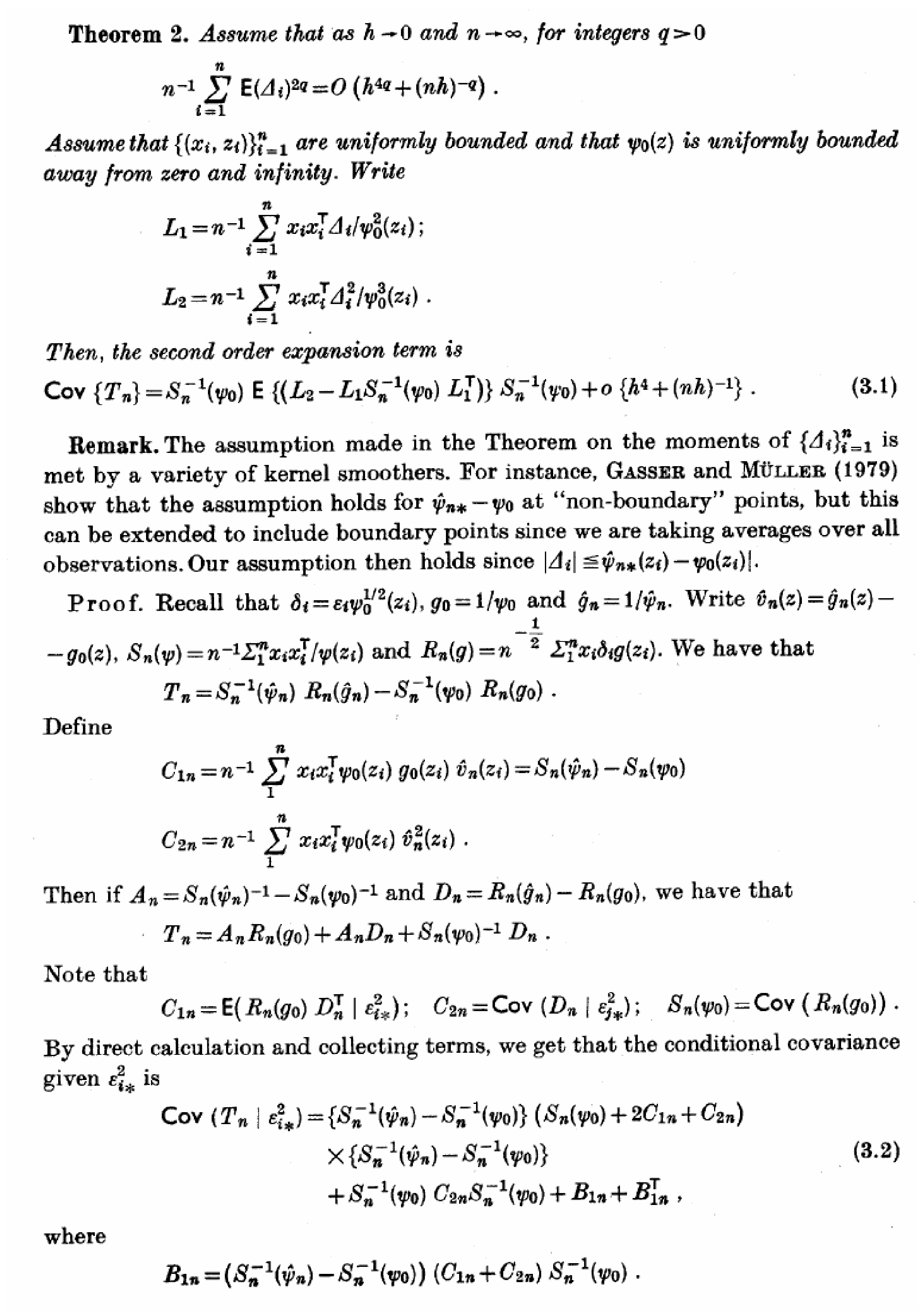
The purpose of this section is to get a qualitative understanding of the second term on the right hand side of (2.1). We use kernel smoothers for estimating ψ_0, see GASSER and MÜLLER (1979). Other smoothing methods could be used, see MACK (1981) and HÄRDLE (1989). Every nonparametric regression technique for estimating ψ_0 will depend on a smoothing parameter. In our case, the smoothing parameter is the so-called bandwidth h which as a function of n tends to zero such that the bias of the estimate (typically of order h^2) and its variance (typically of order $(nh)^{-1}$) tend to zero.

A major technical problem is to avoid allowing $\hat{\psi}_n$ to be near zero, for otherwise expectations may not exist and in any case one wants to avoid giving observations nearly infinite weight. There is also a small technical problem in our calculations where we must bound $\|S_n^{-1}(\hat{\psi}_n)\|_\infty$, where $\|\cdot\|_\infty$ denotes the usual sup norm. To avoid this, let $\hat{\psi}_{n*}$ be a smoother with bandwidth $h \to 0$. Define $\eta_n = h^{2+a}$ for some sufficiently small $a > 0$, and

$$\hat{\psi}_n(z) = \eta_n + \min\,\{\hat{\psi}_{n*}(z), \eta_n^{-1}\}\,.$$

Of course, $\hat{g}_n(z) = 1/\hat{\psi}_n(z)$. For each i, define the stochastic differences

$$\Delta_i = \min\,\{\hat{\psi}_{n*}(z_i), \eta_n^{-1}\} - \psi_0(z_i)\,.$$



Theorem 2. *Assume that as $h \to 0$ and $n \to \infty$, for integers $q > 0$*

$$n^{-1} \sum_{i=1}^{n} \mathsf{E}(\Delta_i)^{2q} = O\left(h^{4q} + (nh)^{-q}\right) .$$

Assume that $\{(x_i, z_i)\}_{i=1}^{n}$ are uniformly bounded and that $\psi_0(z)$ is uniformly bounded away from zero and infinity. Write

$$L_1 = n^{-1} \sum_{i=1}^{n} x_i x_i^{\mathsf{T}} \Delta_i / \psi_0^2(z_i) ;$$

$$L_2 = n^{-1} \sum_{i=1}^{n} x_i x_i^{\mathsf{T}} \Delta_i^2 / \psi_0^3(z_i) .$$

Then, the second order expansion term is

$$\mathsf{Cov}\,\{T_n\} = S_n^{-1}(\psi_0)\ \mathsf{E}\,\{(L_2 - L_1 S_n^{-1}(\psi_0)\ L_1^{\mathsf{T}})\}\ S_n^{-1}(\psi_0) + o\,\{h^4 + (nh)^{-1}\} . \tag{3.1}$$

Remark. The assumption made in the Theorem on the moments of $\{\Delta_i\}_{i=1}^{n}$ is met by a variety of kernel smoothers. For instance, GASSER and MÜLLER (1979) show that the assumption holds for $\hat{\psi}_{n*} - \psi_0$ at "non-boundary" points, but this can be extended to include boundary points since we are taking averages over all observations. Our assumption then holds since $|\Delta_i| \leq \hat{\psi}_{n*}(z_i) - \psi_0(z_i)|$.

Proof. Recall that $\delta_i = \varepsilon_i \psi_0^{1/2}(z_i)$, $g_0 = 1/\psi_0$ and $\hat{g}_n = 1/\hat{\psi}_n$. Write $\hat{v}_n(z) = \hat{g}_n(z) - - g_0(z)$, $S_n(\psi) = n^{-1} \Sigma_1^n x_i x_i^{\mathsf{T}} / \psi(z_i)$ and $R_n(g) = n^{-\frac{1}{2}}\ \Sigma_1^n x_i \delta_i g(z_i)$. We have that

$$T_n = S_n^{-1}(\hat{\psi}_n)\ R_n(\hat{g}_n) - S_n^{-1}(\psi_0)\ R_n(g_0) .$$

Define

$$C_{1n} = n^{-1} \sum_{1}^{n} x_i x_i^{\mathsf{T}} \psi_0(z_i)\ g_0(z_i)\ \hat{v}_n(z_i) = S_n(\hat{\psi}_n) - S_n(\psi_0)$$

$$C_{2n} = n^{-1} \sum_{1}^{n} x_i x_i^{\mathsf{T}} \psi_0(z_i)\ \hat{v}_n^2(z_i) .$$

Then if $A_n = S_n(\hat{\psi}_n)^{-1} - S_n(\psi_0)^{-1}$ and $D_n = R_n(\hat{g}_n) - R_n(g_0)$, we have that

$$T_n = A_n R_n(g_0) + A_n D_n + S_n(\psi_0)^{-1}\ D_n .$$

Note that

$$C_{1n} = \mathsf{E}(\,R_n(g_0)\ D_n^{\mathsf{T}} \mid \varepsilon_{i*}^2) ; \quad C_{2n} = \mathsf{Cov}\ (D_n \mid \varepsilon_{j*}^2) ; \quad S_n(\psi_0) = \mathsf{Cov}\ (\,R_n(g_0)) .$$

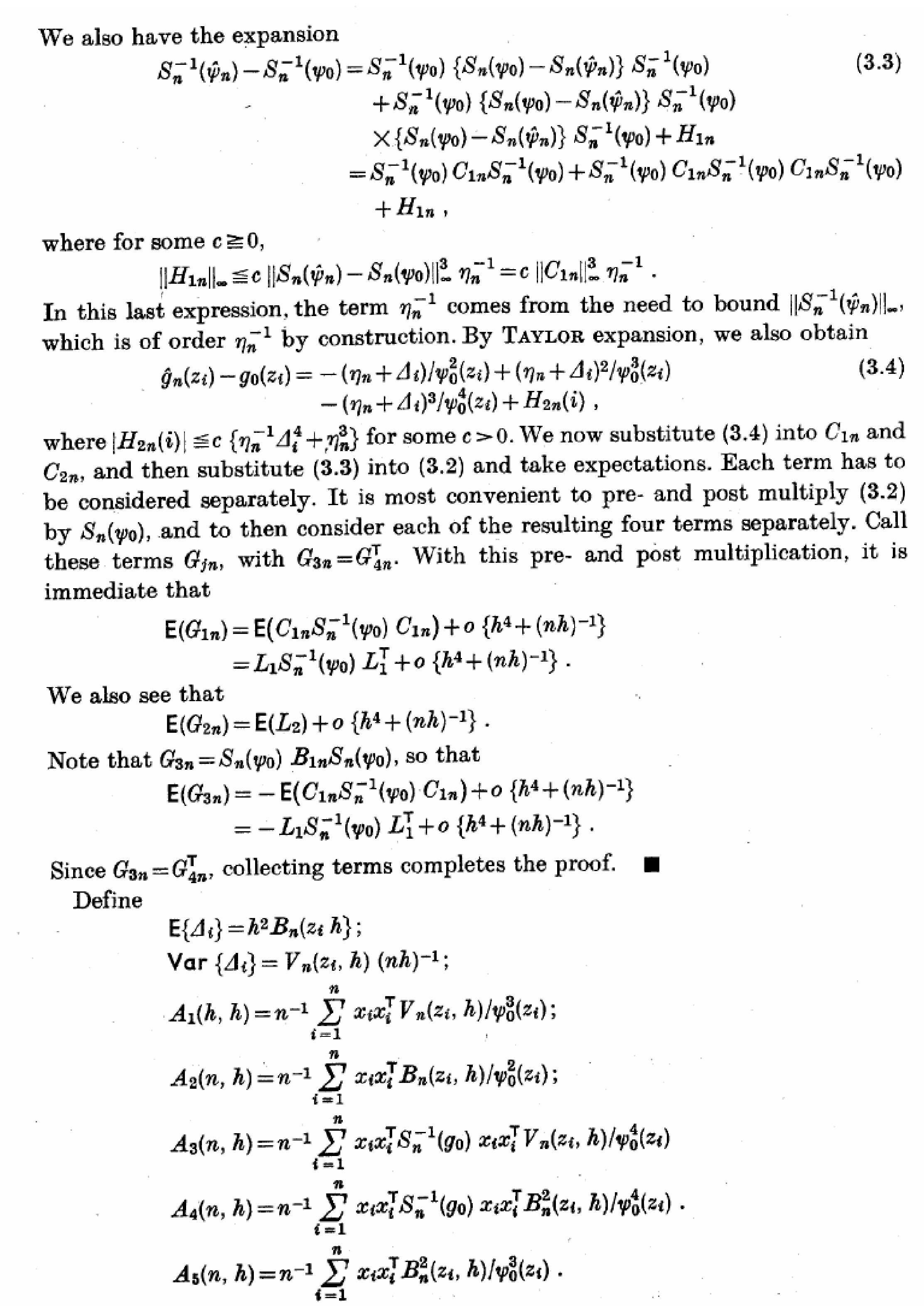
By direct calculation and collecting terms, we get that the conditional covariance given ε_{i*}^2 is

$$\begin{aligned} \mathsf{Cov}\ (T_n \mid \varepsilon_{i*}^2) = {} & \{S_n^{-1}(\hat{\psi}_n) - S_n^{-1}(\psi_0)\}\ (S_n(\psi_0) + 2C_{1n} + C_{2n}) \\ & \times \{S_n^{-1}(\hat{\psi}_n) - S_n^{-1}(\psi_0)\} \\ & + S_n^{-1}(\psi_0)\ C_{2n} S_n^{-1}(\psi_0) + B_{1n} + B_{1n}^{\mathsf{T}} , \end{aligned} \tag{3.2}$$

where

$$B_{1n} = (S_n^{-1}(\hat{\psi}_n) - S_n^{-1}(\psi_0))\ (C_{1n} + C_{2n})\ S_n^{-1}(\psi_0) .$$

Carroll, R. and Härdle, W. (1989) A note on second order effects in a semiparametric context



We also have the expansion

$$S_n^{-1}(\hat{\psi}_n)-S_n^{-1}(\psi_0)=S_n^{-1}(\psi_0)\,\{S_n(\psi_0)-S_n(\hat{\psi}_n)\}\,S_n^{-1}(\psi_0)$$
$$+S_n^{-1}(\psi_0)\,\{S_n(\psi_0)-S_n(\hat{\psi}_n)\}\,S_n^{-1}(\psi_0)$$
$$\times\{S_n(\psi_0)-S_n(\hat{\psi}_n)\}\,S_n^{-1}(\psi_0)+H_{1n}$$
$$=S_n^{-1}(\psi_0)\,C_{1n}S_n^{-1}(\psi_0)+S_n^{-1}(\psi_0)\,C_{1n}S_n^{-1}(\psi_0)\,C_{1n}S_n^{-1}(\psi_0)$$
$$+H_{1n}\,, \tag{3.3}$$

where for some $c \geqq 0$,

$$\|H_{1n}\|_\infty \leqq c\,\|S_n(\hat{\psi}_n)-S_n(\psi_0)\|_\infty^3\,\eta_n^{-1}=c\,\|C_{1n}\|_\infty^3\,\eta_n^{-1}\,.$$

In this last expression, the term η_n^{-1} comes from the need to bound $\|S_n^{-1}(\hat{\psi}_n)\|_\infty$, which is of order η_n^{-1} by construction. By TAYLOR expansion, we also obtain

$$\hat{g}_n(z_i)-g_0(z_i)=-(\eta_n+\Delta_i)/\psi_0^2(z_i)+(\eta_n+\Delta_i)^2/\psi_0^3(z_i)$$
$$-(\eta_n+\Delta_i)^3/\psi_0^4(z_i)+H_{2n}(i)\,, \tag{3.4}$$

where $|H_{2n}(i)| \leqq c\,\{\eta_n^{-1}\Delta_i^4+\eta_n^3\}$ for some $c>0$. We now substitute (3.4) into C_{1n} and C_{2n}, and then substitute (3.3) into (3.2) and take expectations. Each term has to be considered separately. It is most convenient to pre- and post multiply (3.2) by $S_n(\psi_0)$, and to then consider each of the resulting four terms separately. Call these terms G_{jn}, with $G_{3n}=G_{4n}^{\mathsf{T}}$. With this pre- and post multiplication, it is immediate that

$$\mathsf{E}(G_{1n})=\mathsf{E}(C_{1n}S_n^{-1}(\psi_0)\,C_{1n})+o\,\{h^4+(nh)^{-1}\}$$
$$=L_1S_n^{-1}(\psi_0)\,L_1^{\mathsf{T}}+o\,\{h^4+(nh)^{-1}\}\,.$$

We also see that

$$\mathsf{E}(G_{2n})=\mathsf{E}(L_2)+o\,\{h^4+(nh)^{-1}\}\,.$$

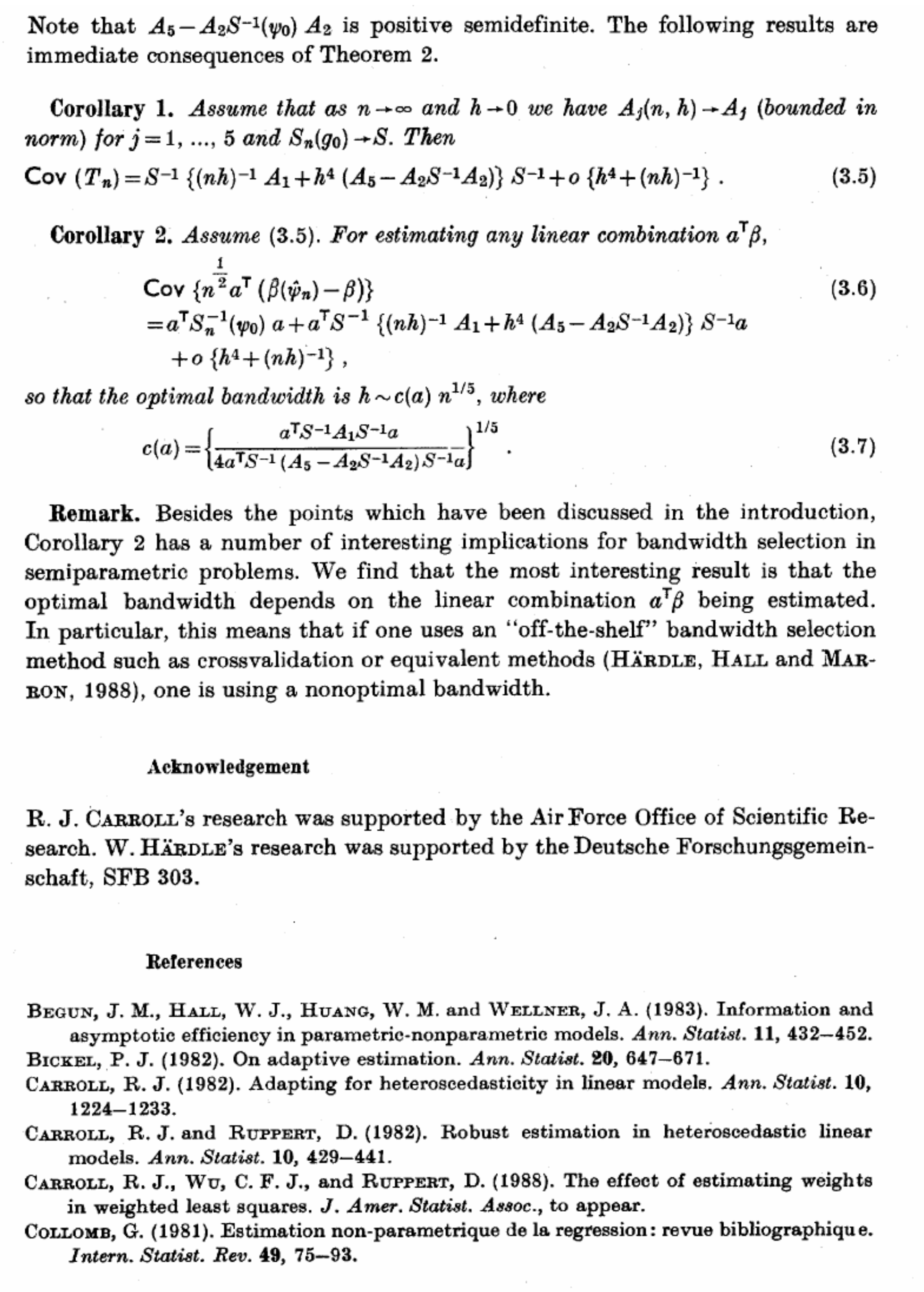
Note that $G_{3n}=S_n(\psi_0)\;B_{1n}S_n(\psi_0)$, so that

$$\mathsf{E}(G_{3n})=-\mathsf{E}(C_{1n}S_n^{-1}(\psi_0)\,C_{1n})+o\,\{h^4+(nh)^{-1}\}$$
$$=-L_1S_n^{-1}(\psi_0)\,L_1^{\mathsf{T}}+o\,\{h^4+(nh)^{-1}\}\,.$$

Since $G_{3n}=G_{4n}^{\mathsf{T}}$, collecting terms completes the proof. ■

Define

$$\mathsf{E}\{\Delta_i\}=h^2B_n(z_i\,h\}\,;$$
$$\mathsf{Var}\,\{\Delta_i\}=V_n(z_i,\,h)\,(nh)^{-1}\,;$$
$$A_1(h,\,h)=n^{-1}\sum_{i=1}^{n}x_ix_i^{\mathsf{T}}V_n(z_i,\,h)/\psi_0^3(z_i)\,;$$
$$A_2(n,\,h)=n^{-1}\sum_{i=1}^{n}x_ix_i^{\mathsf{T}}B_n(z_i,\,h)/\psi_0^2(z_i)\,;$$
$$A_3(n,\,h)=n^{-1}\sum_{i=1}^{n}x_ix_i^{\mathsf{T}}S_n^{-1}(g_0)\,x_ix_i^{\mathsf{T}}V_n(z_i,\,h)/\psi_0^4(z_i)$$
$$A_4(n,\,h)=n^{-1}\sum_{i=1}^{n}x_ix_i^{\mathsf{T}}S_n^{-1}(g_0)\,x_ix_i^{\mathsf{T}}B_n^2(z_i,\,h)/\psi_0^4(z_i)\,.$$
$$A_5(n,\,h)=n^{-1}\sum_{i=1}^{n}x_ix_i^{\mathsf{T}}B_n^2(z_i,\,h)/\psi_0^3(z_i)\,.$$



Note that $A_5 - A_2 S^{-1}(\psi_0)\, A_2$ is positive semidefinite. The following results are immediate consequences of Theorem 2.

Corollary 1. *Assume that as $n \to \infty$ and $h \to 0$ we have $A_j(n, h) \to A_j$ (bounded in norm) for $j = 1, \ldots, 5$ and $S_n(g_0) \to S$. Then*

$$\mathsf{Cov}\ (T_n) = S^{-1}\,\{(nh)^{-1}\, A_1 + h^4\, (A_5 - A_2 S^{-1} A_2)\}\, S^{-1} + o\,\{h^4 + (nh)^{-1}\}\ . \tag{3.5}$$

Corollary 2. *Assume* (3.5). *For estimating any linear combination $a^{\mathsf{T}}\beta$,*

$$\begin{aligned} &\mathsf{Cov}\ \{n^{\frac{1}{2}} a^{\mathsf{T}}\, (\beta(\hat{\psi}_n) - \beta)\} \\ &= a^{\mathsf{T}} S_n^{-1}(\psi_0)\, a + a^{\mathsf{T}} S^{-1}\, \{(nh)^{-1}\, A_1 + h^4\, (A_5 - A_2 S^{-1} A_2)\}\, S^{-1} a \\ &\quad + o\,\{h^4 + (nh)^{-1}\}\ , \end{aligned} \tag{3.6}$$

so that the optimal bandwidth is $h \sim c(a)\, n^{1/5}$, where

$$c(a) = \left\{\frac{a^{\mathsf{T}} S^{-1} A_1 S^{-1} a}{4 a^{\mathsf{T}} S^{-1}\, (A_5 - A_2 S^{-1} A_2)\, S^{-1} a}\right\}^{1/5}\ . \tag{3.7}$$

Remark. Besides the points which have been discussed in the introduction, Corollary 2 has a number of interesting implications for bandwidth selection in semiparametric problems. We find that the most interesting result is that the optimal bandwidth depends on the linear combination $a^{\mathsf{T}}\beta$ being estimated. In particular, this means that if one uses an "off-the-shelf" bandwidth selection method such as crossvalidation or equivalent methods (Härdle, Hall and Marron, 1988), one is using a nonoptimal bandwidth.

Acknowledgement

R. J. Carroll's research was supported by the Air Force Office of Scientific Research. W. Härdle's research was supported by the Deutsche Forschungsgemeinschaft, SFB 303.

References

Begun, J. M., Hall, W. J., Huang, W. M. and Wellner, J. A. (1983). Information and asymptotic efficiency in parametric-nonparametric models. *Ann. Statist.* **11,** 432–452.

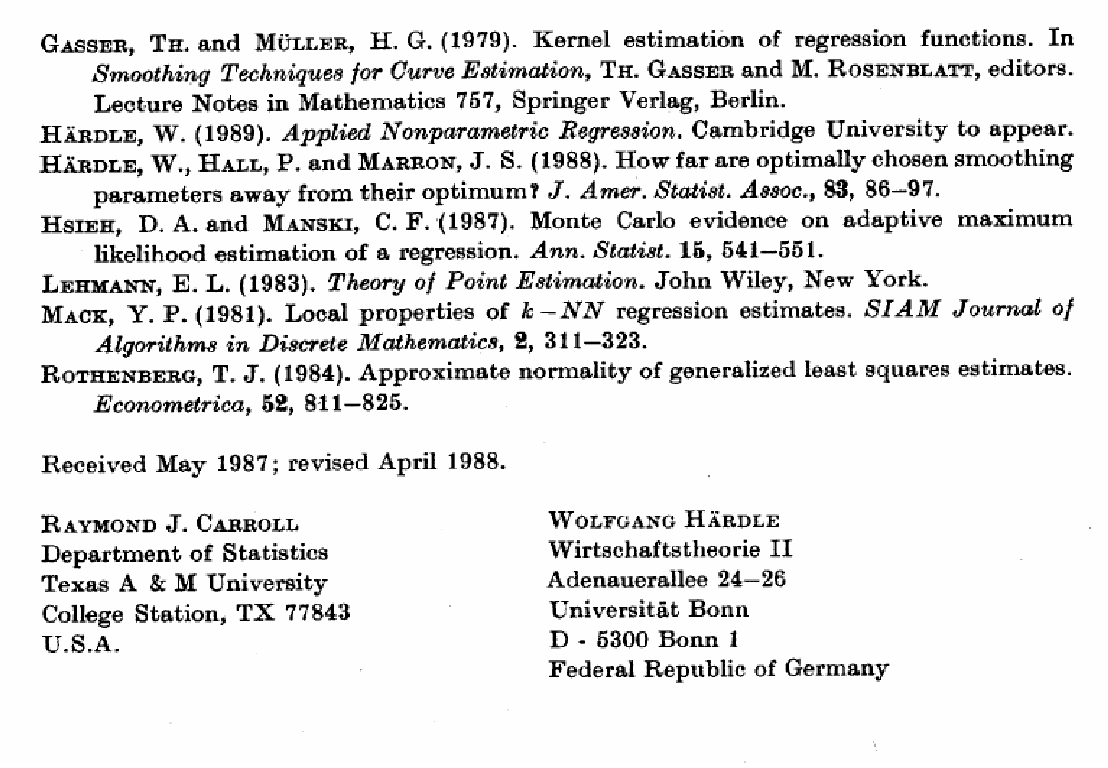
Bickel, P. J. (1982). On adaptive estimation. *Ann. Statist.* **20,** 647–671.

Carroll, R. J. (1982). Adapting for heteroscedasticity in linear models. *Ann. Statist.* **10,** 1224–1233.

Carroll, R. J. and Ruppert, D. (1982). Robust estimation in heteroscedastic linear models. *Ann. Statist.* **10,** 429–441.

Carroll, R. J., Wu, C. F. J., and Ruppert, D. (1988). The effect of estimating weights in weighted least squares. *J. Amer. Statist. Assoc.*, to appear.

Collomb, G. (1981). Estimation non-parametrique de la regression: revue bibliographique. *Intern. Statist. Rev.* **49,** 75–93.



Gasser, Th. and Müller, H. G. (1979). Kernel estimation of regression functions. In *Smoothing Techniques for Curve Estimation*, Th. Gasser and M. Rosenblatt, editors. Lecture Notes in Mathematics 757, Springer Verlag, Berlin.

Härdle, W. (1989). *Applied Nonparametric Regression*. Cambridge University to appear.

Härdle, W., Hall, P. and Marron, J. S. (1988). How far are optimally chosen smoothing parameters away from their optimum? *J. Amer. Statist. Assoc.*, **83**, 86–97.

Hsieh, D. A. and Manski, C. F. (1987). Monte Carlo evidence on adaptive maximum likelihood estimation of a regression. *Ann. Statist.* **15**, 541–551.

Lehmann, E. L. (1983). *Theory of Point Estimation*. John Wiley, New York.

Mack, Y. P. (1981). Local properties of $k-NN$ regression estimates. *SIAM Journal of Algorithms in Discrete Mathematics*, **2**, 311–323.

Rothenberg, T. J. (1984). Approximate normality of generalized least squares estimates. *Econometrica*, **52**, 811–825.

Received May 1987; revised April 1988.

Raymond J. Carroll
Department of Statistics
Texas A & M University
College Station, TX 77843
U.S.A.

Wolfgang Härdle
Wirtschaftstheorie II
Adenauerallee 24–26
Universität Bonn
D - 5300 Bonn 1
Federal Republic of Germany

Asymptotic Maximal Deviation of *M*-Smoothers*

WOLFGANG HÄRDLE

Universität Heidelberg, Heidelberg, West Germany and
University of North Carolina, Chapel Hill, North Carolina

Communicated by the Editors

Let $(X_1, Y_1),..., (X_n, Y_n)$ be i.i.d. rv's and let $m(x) = E(Y \mid X = x)$ be the regression curve of Y on X. A M-smoother $m_n(x)$ is a robust, nonlinear estimator of $m(x)$, defined in analogy to robust M-estimators of location. In this paper the asymptotic maximal deviation $\sup_{0 \leq t \leq 1} |m_n(t) - m(t)|$ is considered. The derived result allows the construction of a uniform confidence band for $m(x)$.

1. INTRODUCTION

Let (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n) be a sequence of independent identically distributed bivariate random variables with joint probability density function $f(x, y)$. Let $m(x) = E(Y \mid X = x)$ denote the regression curve of Y on X. Nadaraya [11] and Watson [18] independently proposed the estimator of $m(x)$,

$$m_n^*(x) = (nh_n)^{-1} \sum_{i=1}^{n} K((x - X_i)/Y_i/[(nh_n)^{-1} \sum_{j=1}^{n} K((x - X_j)/h_n)], \qquad (1.1)$$

where $K: \mathbb{R} \to \mathbb{R}$ denotes a positive kernel function and $h = h_n$ is a sequence of bandwidths tending to zero as n tends to infinity. The Nadaraya–Watson estimator, $m_n^*(x)$ can be considered as a local least-squares estimate, since $m_n^*(x)$ minimizes

$$H_n^*(\theta) = (nh_n)^{-1} \sum_{i=1}^{n} K((x - X_i)/h_n)(Y_i - \theta)^2$$

Received February 11, 1983; revised July 5, 1984.

AMS 1980 subject classifications: 62G05, 62J02.

Key words and phrases: Nonparametric regression; M-smoothing; uniform confidence bands.

* Research partially supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 123 and by Air Force Office of Scientific Research Contract AFOSR-F49620-82-C0009.

with respect to θ. Equivalently, $m_n^*(x)$ can be viewed as a local average of those Y-observations with corresponding X-observation in a neighborhood of x. The size of that neighborhood is regulated by the bandwidth sequence $\{h_n\}$.

It is well known that the sample mean is highly sensitive to outliers. It is therefore expected that $m_n^*(x)$, as a local average of the Y-observations, may give rise of misinterpretations when outliers are present. A huge outlier, for instance, may mimic peaks or bumps. Such outliers occur quite often in practice, see for instance Ruppert *et al.* [15, Fig. 2] or Bussian and Härdle [3].

In this paper we investigate so called M-smoothers, as considered by Härdle [5]. M-smoothers are nonlinear curve estimates and are implicitly defined as a zero (w.r.t. θ) of the function

$$G_n(\theta) = (nh_n)^{-1} \sum_{i=1}^{n} K((x - X_i)/h_n)\, \psi(Y_i - \theta). \tag{1.2}$$

Here $\psi\colon \mathbb{R} \to \mathbb{R}$ denotes a bounded monotone, antisymmetric function. We call the M-smoother $m_n(x)$. It is shown in this paper that

$$P\left\{(2\delta \log n)^{1/2} \left[\sup_{0 \leqslant t \leqslant 1} r(t)\, |(m_n(t) - m(t)|/\lambda(K)^{1/2} - d_n\right] < x\right\} \xrightarrow[n \to \infty]{} \exp(-2 \exp(-x)), \tag{1.3}$$

where δ, $r(t)$, $\lambda(K)$, d_n are suitable scaling parameters. This result allows the construction of (asymptotic) uniform confidence bands for $m(x)$. In a small Monte Carlo study (Section 3) the behavior of both $m_n^*(x)$ and $m_n(x)$ is investigated when the data contains outliers, generated by heavy tailed conditional distributions of $(Y \mid X = x)$.

The result (1.3) improves upon that of Johnston [9] in a number of ways. First, Johnston obtains results like (1.3), but for estimates different to the Nadaraya–Watson estimator (1.1); our result (1.3) applies to the Nadaraya–Watson estimator as a special case (set $\psi(u) = u$). Second, (1.3) holds for a much broader class of estimators. Finally, we obtain (1.3) under assumptions weaker than those needed by Johnston.

The function ψ entering into the definition of the M-smoother $m_n(x)$, can be chosen in various ways. For instance, the classical ψ-function

$$\psi(u) = \min\{c\colon \max\{x, -c\}\}, \qquad c > 0$$

can be used [8]. In this paper we do not emphasize the choice of a particular ψ-function; any of the ψ-functions to be specified below yields a robust estimate $m_n(x)$ of $m(x)$. The choice of a particular ψ-function depends on the kind of contamination model that is assumed to have generated the outliers. One possible contamination model and an adopted ψ-function thereof is described in Härdle [5].

As a footnote we would like to mention some related work. Stuetzle and Mittal [16] obtained bias and variance rates with $K(u)=\frac{1}{2}I_{[-1,1]}(u)$ and Härdle and Gasser [6] showed some asymptotic properties of $m_n(x)$ in a fixed design setting.

For the rest of the paper we will write h instead of h_n.

2. Results

The following assumptions will be convenient.

(A1) The kernel $K(\cdot)$ is positive has compact support $[-A, A]$ and is continuously differentiable;

(A2) $(nh)^{-1/2}(\log n)^{3/2} \to 0$, $(n \log n)^{1/2}\, h^{5/2} \to 0$, $(nh^3)^{-1}(\log n)^2 \leqslant M$, M a constant;

(A3) $h^{-3}(\log n)\int_{|y|>a_n} f_Y(y)\,dy = O(1)$, $f_Y(y)$ the marginal density of Y, $\{a_n\}_{n=1}^{\infty}$ a sequence of constants tending to infinity as $n\to\infty$;

(A4) $\inf_{0\leqslant t\leqslant 1} |q(t)| \geqslant q_0 > 0$, where $q(t)=E(\Psi'(Y-m(t)) \mid X=t)\, f_X(t)$, f_X the marginal density of X;

(A5) the regression function $m(x)$ is twice continuously differentiable, the conditional densities $f(y \mid x)$ are symmetric for all x; Ψ is piecewise twice continuously differentiable.

Define also

$$\sigma^2(t) = E(\Psi^2(Y-m(t)) \mid X=t)$$

$$H_n(t) = (nh)^{-1} \sum_{i=1}^{n} K((t-X_i)/h)\, \Psi(Y_i - m(t))$$

$$D_n(t) = (nh)^{-1} \sum_{i=1}^{n} K((t-X_i)/h)\, \Psi'(Y_i - m(t))$$

and assume that $\sigma^2(t)$ and $f_X(t)$ are differentiable.

Assumption (A1) on the compact support of the kernel could possibly be relaxed introducing a cutoff technique as Csörgő and Hall [4] for density estimators. Assumption (A2) has purely technical reasons: to keep the bias at a lower rate than the variance and to ensure the vanishing of some nonlinear remainder terms. Assumption (A3) appears in a somewhat modified form also in Johnston's paper [9]. When we want to apply the following theorem to the Nadaraya–Watson estimator $m_n^*(x)$ we have to restate (A2) as $h^{-3}(\log n)\int_{|y|>a_n} y^2 f_y(y)\,dy$ (which is assumption A1 in Johnston [9]). Assumption (A5) asking for the symmetry of the conditional densities is a common assumption in robust estimation [8]. It guarantees that the only solution of $\int \Psi(y-\cdot)f(y\mid x)\,dy=0$ is $m(x)=E(Y\mid X=x)$. If we had skew distributions then we would no longer estimate the conditional mean but rather some different conditional measure of location.

THEOREM. *Let* $h=n^{-\delta}$, $\frac{1}{5}<\delta<\frac{1}{3}$ and $\lambda(K)=\int_{-A}^{A} K^2(u)\,du$ *and*

$$d_n=(2\delta\log n)^{1/2}+(2\delta\log n)^{-1/2}\log(c_1(K)/\pi^{1/2})+\tfrac{1}{2}[\log\delta+\log\log n]\},$$

$$\text{if } c_1(K)=K^2(A)+K^2(-A)/[2\lambda(K)]>0$$

$$d_n=(2\delta\log n)^{1/2}+(2\delta\log n)^{-1/2}\{\log(c_2(K)/2\pi)\}$$

$$\text{otherwise with } c_2(K)=\int_{-A}^{A}[K'(u)]^2\,du/[2\lambda(K)].$$

Then (1.3) *holds with*

$$r(t)=(nh)^{1/2}q(t)[\sigma^2(t)f_X(t)]^{-1/2}.$$

This theorem can be used to construct uniform confidence intervals for the regression function as stated in the following corollary.

COROLLARY. *Under the assumptions of the theorem above, an approximate* $(1-\alpha)$ x 100% *confidence band over* [0, 1] *is*

$$m_n(t)\pm(nh)^{-1/2}[\hat{\sigma}^2(t)\hat{f}_X(t)\,\lambda(K)]^{1/2}$$
$$\times q^{-1}(t)[d_n+c(\alpha)(2\delta\log n)^{-1/2}]\cdot[\lambda(K)]^{1/2},$$

where $c(\alpha)=\log 2-\log|\log(1-\alpha)|$ and $\hat{\sigma}(t),\hat{f}_X(t)$ *are consistent estimates for* $\sigma(t), f_X(t)$.

The proof is essentially based on a linearization argument after a Taylor series expansion. The leading linear term will then be approximated in a similar way as in Johnston [9], Bickel and Rosenblatt [1]. The main idea behind the proof is a strong approximation of the empirical process of $\{(X_i, Y_i)\}_{i=1}^n$ by a sequence of Brownian bridges (with 2-dimensional time) as provided by Tusnady [17].

It follows by Taylor expansions applied to the defining equation (1.2) that

$$m_n(t)-m(t)=(H_n(t)-EH_n(t))/q(t)+R_n(t), \tag{2.1}$$

where $[H_n(t) - EH_n(t)]/q(t)$ is the leading linear term and

$$\begin{aligned} R_n(t) = {} & H_n(t)[q(t) - D_n(t)]/[D_n(t) \cdot q(t)] + EH_n(t)/q(t) \\ & + \tfrac{1}{2}(m_n(t) - m(t))^2 \cdot [D_n(t)]^{-1} \\ & \cdot (nh)^{-1} \sum_{i=1}^{n} K((x - X_i)/h)\, \Psi''(Y_i - m(t) + r_n^{(i)}(t)), \\ & |r_n^{(i)}(t)| < |m_n(t) - m(t)|. \end{aligned} \tag{2.2}$$

is the remainder term. In the third section it is shown (Lemma 3.1) that $\|R_n\| = \sup_{0 \leqslant t \leqslant 1} |R_n(t)| = o_p((nh \log n)^{-1/2})$.

Furthermore, the rescaled linear part

$$Y_n(t) = (nh)^{1/2}[\sigma^2(t) f_X(t)]^{-1/2}(H_n(t) - EH_n(t))$$

is approximated by a sequence of Gaussian processes, leading finally to the process

$$Y_{5,n}(t) = h^{-1/2} \int K((t-x)/h)\, dW(x),$$

as in Bickel and Rosenblatt [1].

We also need the Rosenblatt transformation [13],

$$T(x, y) = (F_{X|y}(x \mid y), F_Y(y)),$$

which transforms (X_i, Y_i) into $T(X_i, Y_i) = (X'_i, Y'_i)$ mutually independent uniform rv's. With the aid of this transformation, Theorem 1 of Tusnady [17] may be applied to obtain the following lemma.

LEMMA 2.1. *On a suitable probability space there exists a sequence of Brownian bridges B_n such that*

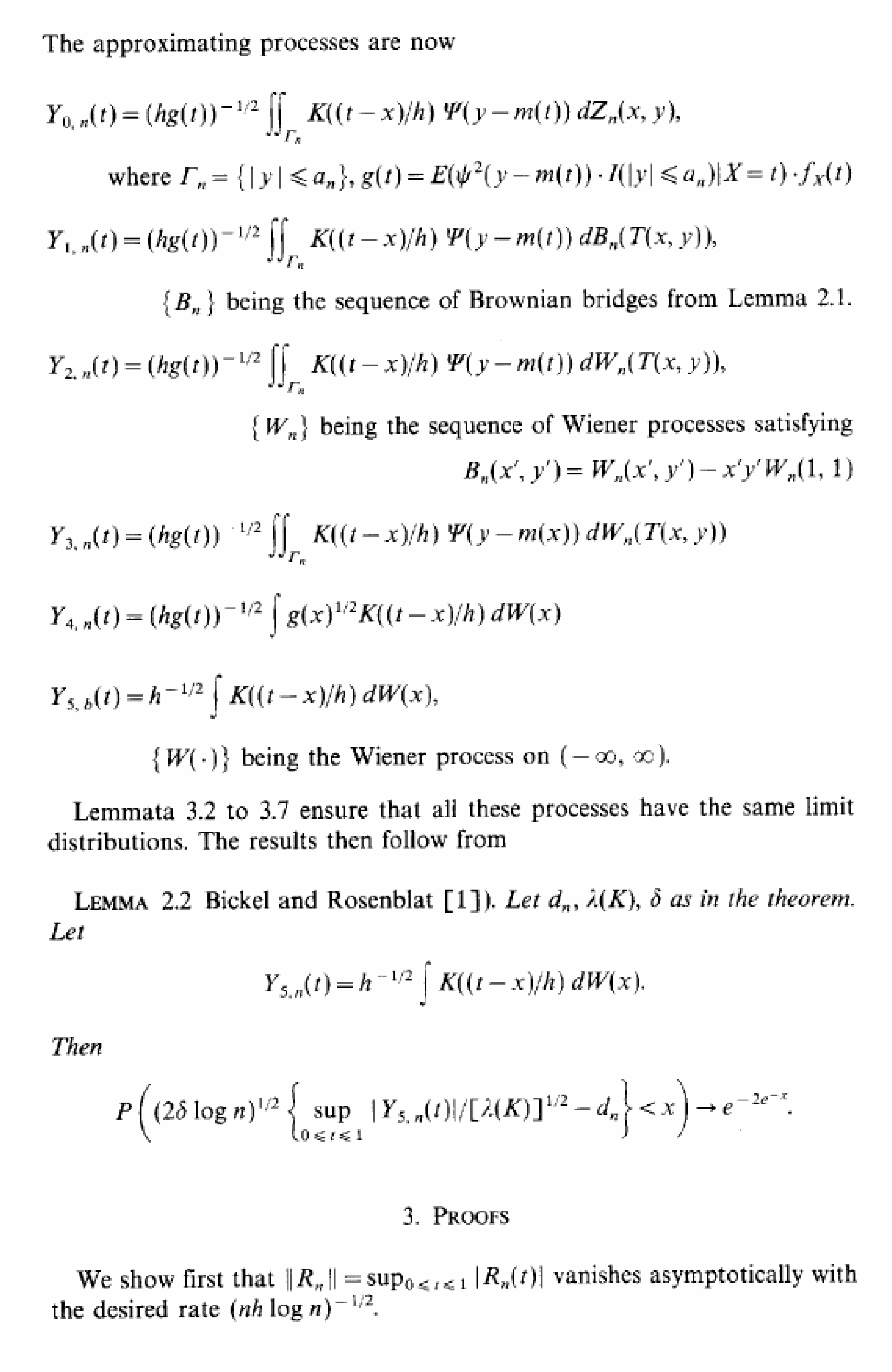
$$\sup_{x, y} |Z_n(x, y) - B_n(T(x, y))| = O(n^{-1/2}(\log n)^2) \qquad a.s.,$$

where $Z_n(x, y) = n^{1/2}[F_n(x, y) - F(x, y)]$ denotes the empirical process of $\{(X_i, Y_i)\}_{i=1}^{n}$.

Before we define the different approximating processes let us first rewrite $Y_n(t)$ as a stochastic integral with respect to the empirical process $Z_n(x, y)$,

$$Y_n(t) = h^{-1/2} g'(t)^{-1/2} \iint K((t-x)/h)\, \Psi(y - m(t))\, dZ_n(x, y),$$

$$g'(t) = \sigma^2(t) f_X(t).$$



The approximating processes are now

$$Y_{0,n}(t)=(hg(t))^{-1/2}\iint_{\Gamma_n} K((t-x)/h)\,\Psi(y-m(t))\,dZ_n(x,y),$$

where $\Gamma_n=\{|y|\leqslant a_n\}$, $g(t)=E(\psi^2(y-m(t))\cdot I(|y|\leqslant a_n)|X=t)\cdot f_X(t)$

$$Y_{1,n}(t)=(hg(t))^{-1/2}\iint_{\Gamma_n} K((t-x)/h)\,\Psi(y-m(t))\,dB_n(T(x,y)),$$

$\{B_n\}$ being the sequence of Brownian bridges from Lemma 2.1.

$$Y_{2,n}(t)=(hg(t))^{-1/2}\iint_{\Gamma_n} K((t-x)/h)\,\Psi(y-m(t))\,dW_n(T(x,y)),$$

$\{W_n\}$ being the sequence of Wiener processes satisfying

$$B_n(x',y')=W_n(x',y')-x'y'W_n(1,1)$$

$$Y_{3,n}(t)=(hg(t))^{-1/2}\iint_{\Gamma_n} K((t-x)/h)\,\Psi(y-m(x))\,dW_n(T(x,y))$$

$$Y_{4,n}(t)=(hg(t))^{-1/2}\int g(x)^{1/2}K((t-x)/h)\,dW(x)$$

$$Y_{5,b}(t)=h^{-1/2}\int K((t-x)/h)\,dW(x),$$

$\{W(\cdot)\}$ being the Wiener process on $(-\infty,\infty)$.

Lemmata 3.2 to 3.7 ensure that all these processes have the same limit distributions. The results then follow from

LEMMA 2.2 Bickel and Rosenblat [1]). *Let* d_n, $\lambda(K)$, δ *as in the theorem. Let*

$$Y_{5,n}(t)=h^{-1/2}\int K((t-x)/h)\,dW(x).$$

Then

$$P\left((2\delta\log n)^{1/2}\left\{\sup_{0\leqslant t\leqslant 1}|Y_{5,n}(t)|/[\lambda(K)]^{1/2}-d_n\right\}<x\right)\to e^{-2e^{-x}}.$$

3. PROOFS

We show first that $\|R_n\|=\sup_{0\leqslant t\leqslant 1}|R_n(t)|$ vanishes asymptotically with the desired rate $(nh\log n)^{-1/2}$.

LEMMA 3.1. *For the remainder term $R_n(t)$ defined in* (2.2) we have

$$\|R_n\| = o_p((bh \log n)^{-1/2}). \tag{3.1}$$

Proof. First we have by the positivity of the kernel K and $|\Psi''| < C_1$,

$$\|R_n\| \leqslant \left[\inf_{0 \leqslant t \leqslant 1} (|D_n(t)| \cdot q(t))\right]^{-1} \{\|H_n\| \cdot \|q - D_n\| + \|D_n\| \cdot \|EH_n\|\}$$

$$+ C_1 \cdot \|m_n - m\|^2 \cdot \left[\inf_{0 \leqslant t \leqslant 1} |D_n(t)|\right]^{-1} \cdot \|f_n\|,$$

where $f_n = (nh)^{-1} \sum_{i=1}^n K((x - X_i)/h)$.

The desired result (3.1) will then follow if we prove

$$\|H_n\| = o_p(n^{-1/2} h^{-1/4} \cdot (\log n)^{-1/2}) \tag{3.2}$$

$$\|q - D_n\| = o_p(n^{-1/4} h^{-1/4} (\log n)^{-1/2}) \tag{3.3}$$

$$\|EH_n\| = O(h^2) \tag{3.4}$$

$$\|m_n - m\|^2 = o_p((nh)^{-1/2} (\log n)^{-1/2}). \tag{3.5}$$

Define $U_n(t) = n^{1/4} h^{1/4} (\log n)^{1/2} [H_n(t) - EH_n(t)]$. We first show that $U_n(t) \to^p 0$ for all t. This follows from Markov's inequality since

$$U_n(t) = \sum_{i=1}^n U_{i,n}(t),$$

where $U_{i,n}(t) = n^{-3/4} h^{-3/4} (\log n)^{1/2} [K((t - X_i)/h) \ \Psi(Y_i - m(t)) - EK((t - X)/h) \cdot \Psi(y - m(t))]$, are i.i.d. rv's and thus

$$P(|U_n(t)| > \varepsilon) \leqslant \varepsilon^{-2} n^{-1/2} h^{-1/2} (\log n) \cdot h^{-1} EK^2((t - X)/h) \ \Psi^2(Y - m(t)).$$

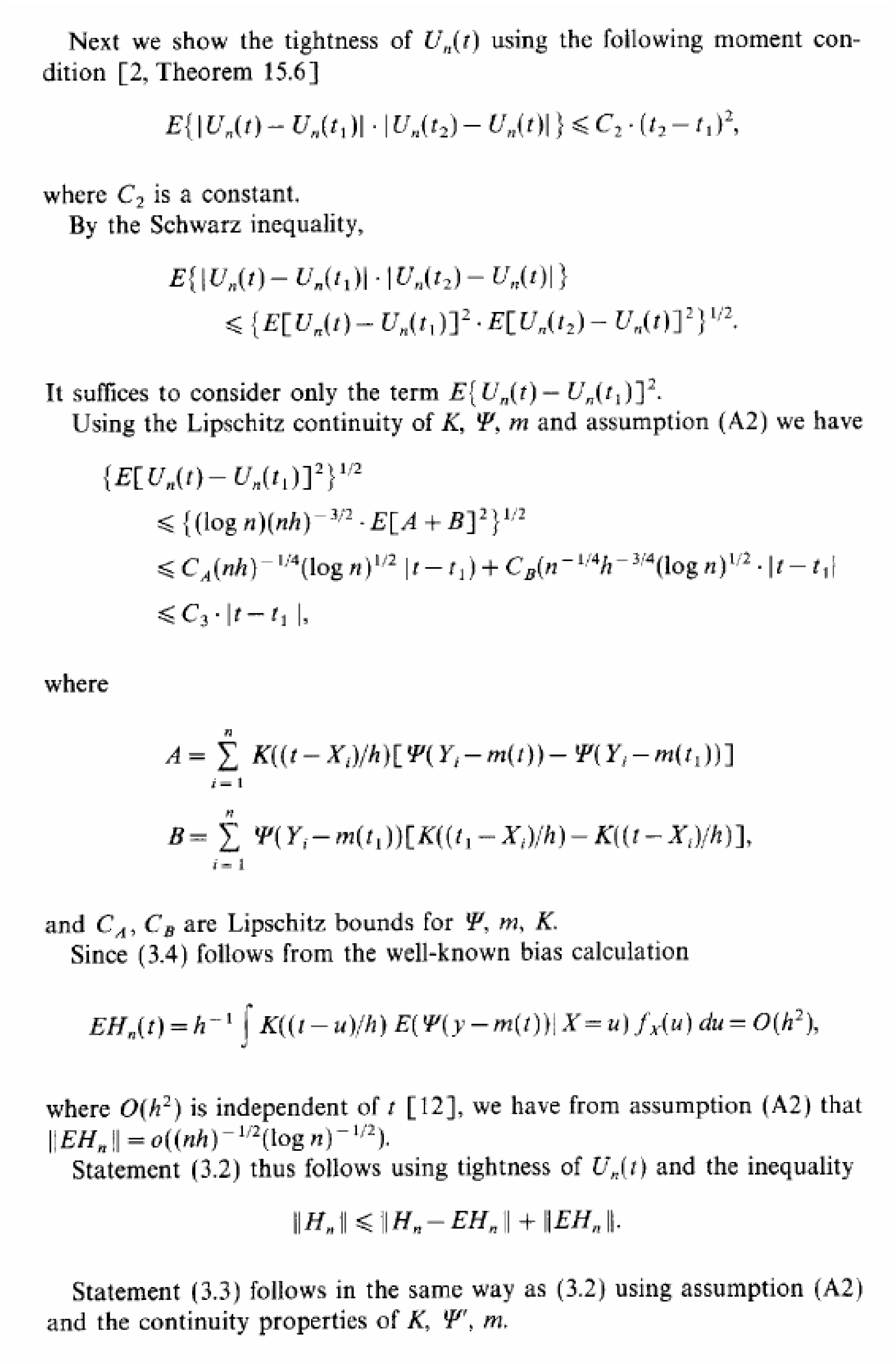
The RHS of this inequality tends to zero, since

$$h^{-1} EK^2((t - X)/h) \ \Psi^2(Y - m(t))$$

$$= h^{-1} \int K^2((t - u)/h) \ E(\Psi^2(Y - m(t)) | \ X = u) f_X(u) \, du$$

$$\sim \sigma^2(t) \cdot f_X(t) \cdot \int K^2(u) \, du$$

by continuity of $\sigma^2(t)$ and $f_X(t)$.



Next we show the tightness of $U_n(t)$ using the following moment condition [2, Theorem 15.6]

$$E\{|U_n(t)-U_n(t_1)|\cdot|U_n(t_2)-U_n(t)|\}\leqslant C_2\cdot(t_2-t_1)^2,$$

where C_2 is a constant.

By the Schwarz inequality,

$$\begin{aligned}&E\{|U_n(t)-U_n(t_1)|\cdot|U_n(t_2)-U_n(t)|\}\\&\qquad\leqslant\{E[U_n(t)-U_n(t_1)]^2\cdot E[U_n(t_2)-U_n(t)]^2\}^{1/2}.\end{aligned}$$

It suffices to consider only the term $E\{U_n(t)-U_n(t_1)]^2$.

Using the Lipschitz continuity of K, Ψ, m and assumption (A2) we have

$$\begin{aligned}&\{E[U_n(t)-U_n(t_1)]^2\}^{1/2}\\&\quad\leqslant\{(\log n)(nh)^{-3/2}\cdot E[A+B]^2\}^{1/2}\\&\quad\leqslant C_A(nh)^{-1/4}(\log n)^{1/2}\,|t-t_1)+C_B(n^{-1/4}h^{-3/4}(\log n)^{1/2}\cdot|t-t_1|\\&\quad\leqslant C_3\cdot|t-t_1|,\end{aligned}$$

where

$$A=\sum_{i=1}^{n}K((t-X_i)/h)[\Psi(Y_i-m(t))-\Psi(Y_i-m(t_1))]$$

$$B=\sum_{i=1}^{n}\Psi(Y_i-m(t_1))[K((t_1-X_i)/h)-K((t-X_i)/h)],$$

and C_A, C_B are Lipschitz bounds for Ψ, m, K.

Since (3.4) follows from the well-known bias calculation

$$EH_n(t)=h^{-1}\int K((t-u)/h)\,E(\Psi(y-m(t))|\,X=u)\,f_X(u)\,du=O(h^2),$$

where $O(h^2)$ is independent of t [12], we have from assumption (A2) that $\|EH_n\|=o((nh)^{-1/2}(\log n)^{-1/2})$.

Statement (3.2) thus follows using tightness of $U_n(t)$ and the inequality

$$\|H_n\|\leqslant\|H_n-EH_n\|+\|EH_n\|.$$

Statement (3.3) follows in the same way as (3.2) using assumption (A2) and the continuity properties of K, Ψ', m.

Finally from Härdle and Luckhaus [7], where uniform consistency of $m_n(t) - m(t)$ is shown, we have

$$\|m_n - m\| = O_p((nh)^{-1/2}(\log n)^{1/2}),$$

which implies (3.5).

Now the assertion of the lemma follows, since by tightness of $D_n(t)$, $\inf_{0 \leqslant t \leqslant 1} |D_n(t)| \to_p q_0$ and thus

$$\|R_n\| = o_p((nh)^{-1/2}(\log n)^{-1/2})(1 + \|f_n\|).$$

Finally, by Theorem 3.1 of Bickel and Rosenblatt [1], $\|f_n\| = O_p(1)$; thus the desired result $\|R_n\| = o_p((nh)^{-1/2}(\log n)^{-1/2})$ follows. In the non-robust case, i.e., $\Psi(u) = u$, the remainder term R_n reads

$$R_n = [m_n^* - m][f_X - f_n] f_X^{-1} + E(\hat{m}_n - m) f_n / f_X, \tag{3.6}$$

where $m_n(x) = (nh)^{-1} \sum_{i=1}^n K((x - X_i)/h)\, Y_i$.

Johnston [9] proved that $(\hat{m}_n - E\hat{m}_n)/f$ has the desired asymptotic distribution as stated in our theorem.

So if we apply the recent result of Mack and Silverman [10] or Härdle and Luckhaus [7] to $\|m_n^* - m\|$ and the well-known result from Bickel and Rosenblatt [1] to $\|f_X - f_n\|$, we may conclude that the first term on the RHS of (3.6) is $o_p((nh)^{-1/2}(\log n)^{-1/2})$. The second term in (3.6) is

$$\left[h^{-1} \int K((t-u)/h) \cdot m(u) f(u)\, du - m(t)\, h^{-1} \int K((t-u) f(u)\, du\right] \Big/ f_X(t)$$

which is by the same calculations as mentioned above [12] of the order $O(h^2)$. This shows that our result generalizes Johnston's paper. Our theorem says also that the confidence bounds are smaller. Johnston had $s^2(t) = E(Y^2 \mid X = t)$ as a factor for the asymptotic confidence bound, we have $\sigma^2(t) = \operatorname{var}(Y \mid X = t)$ which is in general smaller than $s^2(t)$. We now begin with the subsequent approximations of the processes $Y_{0,n}$ to $Y_{5,n}$.

LEMMA 3.2. $\|Y_{0,n} - Y_{1,n}\| = O((nh)^{-1/2}(\log n)^2)$ *a.s.*

Proof. Let t be fixed and put $L(y) = \Psi(y - m(t))$ still depending on t. Use integration by parts and obtain

$$\iint_{\Gamma_n} L(y)\, K((t-x)/h)\, dZ_n(x, y)$$

$$= \int_{u=-A}^{A} \int_{y=-a_n}^{a_n} L(y)\, K(u)\, dZ_n(t - h \cdot u, y)$$

$$= \int_{-A}^{A} \int_{-a_n}^{a_n} Z_n(t - h \cdot u, y)\, d[L(y)\, K(u)] + L(a_n) \int_{-A}^{A} Z_n(t \cdot u, a_n)\, dK(u)$$

$$- L(-a_n) \int_{-A}^{A} Z_n(t - h \cdot u, -a_n)\, dK(u)$$

$$+ K(A) \Bigg[\int_{-a_n}^{a} Z_n(t - h \cdot u, y)\, dL(y)$$

$$+ L(a_n)\, Z_{n_a}(t - h \cdot A, a_n) - L(-a_n)\, Z_n(t - h \cdot A, -a_n) \Bigg]$$

$$- K(-A) \Bigg[\int_{-a_n}^{n} L_n(t + h \cdot A, y)\, dL(y) + L(a_n)\, Z_n(t + h \cdot A, a_n)$$

$$- L(-a_n)\, Z_n(t + h \cdot A, -a_n) \Bigg].$$

If we apply the same operations to $Y_{1,n}$ with $B_n(T(x, y))$ instead of $Z_n(x, y)$ and use Lemma 2.1, we finally obtain

$$\sup_{0 \leqslant t \leqslant 1} h^{1/2} g(t)^{1/2} |\, Y_{0,n}^{(t)} - Y_{1,n}(t)| = O((nh)^{-1/2} (\log n)^2) \qquad \text{a.s.},$$

using the differentiability and boundedness of ψ.

Lemma 3.3. $\| Y_{1,n} - Y_{2,n} \| = O_p(h^{1/2})$.

Proof. Note that the Jacobi of $T(x, y)$ is $f(x, y)$ hence

$$|\, Y_{1,n}(t) - Y_{2,n}(t)|$$

$$= \left| (q(t)\, h)^{-1/2} \iint_{\Gamma_n} \psi(y - m(t)\, K((t - x)/h)\, f(x, y)\, dx\, dy \right| \cdot |W_n(1, 1)|\,.$$

It follows that

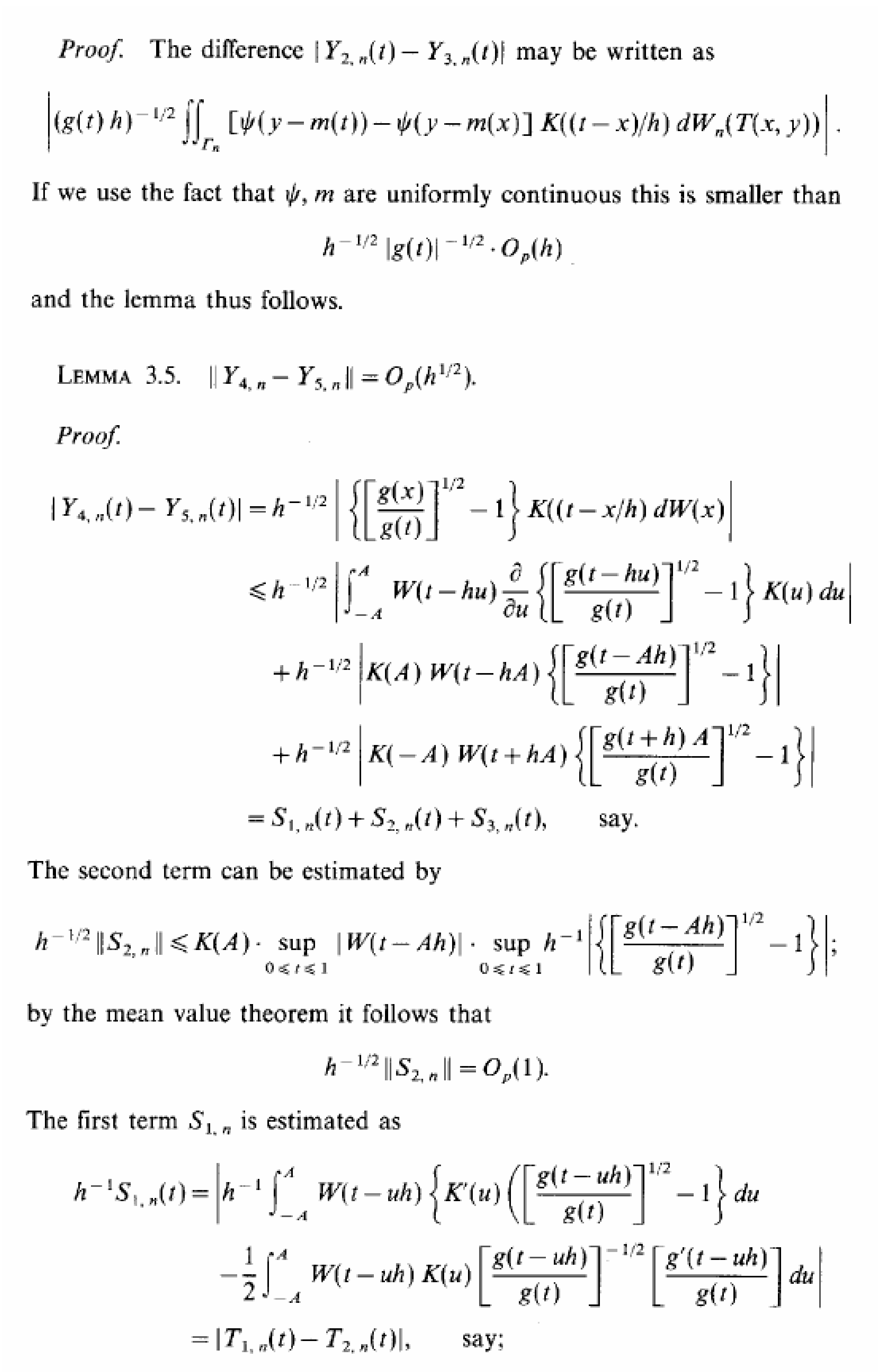
$$h^{-1/2} \| Y_{1,n} - Y_{2,n} \| \leqslant |W_n(1, 1)| \cdot \| g^{-1/2} \|$$

$$\cdot \sup_{0 \leqslant t \leqslant 1} h^{-1} \iint_{\Gamma_n} |\psi(y - m(t))\, K((t - x)/h)|\, f(x, y)\, dx\, dy.$$

Since $\| g^{-1/2} \|$ is bounded by assumption and ψ is bounded, we have

$$h^{-1/2} \| Y_{1,n} - Y_{2,n} \| \leqslant |W_n(1, 1)| \cdot C_4 \cdot h^{-1} \int (K((t - x)/h))\, dx = O_p(1).$$

Lemma 3.4. $\| Y_{2,n} - Y_{3,n} \| = O_p(h^{1/2})$.



Proof. The difference $|Y_{2,n}(t)-Y_{3,n}(t)|$ may be written as

$$\left|(g(t)\,h)^{-1/2}\iint_{\Gamma_n}[\psi(y-m(t))-\psi(y-m(x)]\,K((t-x)/h)\,dW_n(T(x,y))\right|.$$

If we use the fact that ψ, m are uniformly continuous this is smaller than

$$h^{-1/2}\,|g(t)|^{-1/2}\cdot O_p(h)$$

and the lemma thus follows.

LEMMA 3.5. $\|Y_{4,n}-Y_{5,n}\|=O_p(h^{1/2})$.

Proof.

$$\begin{aligned}|Y_{4,n}(t)-Y_{5,n}(t)|&=h^{-1/2}\left|\left\{\left[\frac{g(x)}{g(t)}\right]^{1/2}-1\right\}K((t-x/h)\,dW(x)\right|\\&\leqslant h^{-1/2}\left|\int_{-A}^{A}W(t-hu)\frac{\partial}{\partial u}\left\{\left[\frac{g(t-hu)}{g(t)}\right]^{1/2}-1\right\}K(u)\,du\right|\\&\quad+h^{-1/2}\left|K(A)\,W(t-hA)\left\{\left[\frac{g(t-Ah)}{g(t)}\right]^{1/2}-1\right\}\right|\\&\quad+h^{-1/2}\left|K(-A)\,W(t+hA)\left\{\left[\frac{g(t+h)\,A}{g(t)}\right]^{1/2}-1\right\}\right|\\&=S_{1,n}(t)+S_{2,n}(t)+S_{3,n}(t),\qquad\text{say.}\end{aligned}$$

The second term can be estimated by

$$h^{-1/2}\|S_{2,n}\|\leqslant K(A)\cdot\sup_{0\leqslant t\leqslant 1}|W(t-Ah)|\cdot\sup_{0\leqslant t\leqslant 1}h^{-1}\left|\left\{\left[\frac{g(t-Ah)}{g(t)}\right]^{1/2}-1\right\}\right|;$$

by the mean value theorem it follows that

$$h^{-1/2}\|S_{2,n}\|=O_p(1).$$

The first term $S_{1,n}$ is estimated as

$$\begin{aligned}h^{-1}S_{1,n}(t)=\Bigg|h^{-1}\int_{-A}^{A}W(t-uh)\left\{K'(u)\left(\left[\frac{g(t-uh)}{g(t)}\right]^{1/2}-1\right\}du\right.\\-\frac{1}{2}\int_{-A}^{A}W(t-uh)\,K(u)\left[\frac{g(t-uh)}{g(t)}\right]^{-1/2}\left[\frac{g'(t-uh)}{g(t)}\right]du\Bigg|\\=|T_{1,n}(t)-T_{2,n}(t)|,\qquad\text{say;}\end{aligned}$$

$\|T_{2,n}\| \leqslant C_5 \cdot \int_{-A}^{A} |W(t-hu)|\, du = O_p(1)$ by assumption on $g(t) = \sigma^2(t) \cdot f_X(t)$. To estimate $T_{1,n}$ we again use the mean value theorem to conclude that

$$\sup_{0 \leqslant t \leqslant 1} h^{-1} \left| \left[\frac{g(t-uh)}{g(t)} \right]^{1/2} - 1 \right| < C_6 \cdot |u|;$$

hence

$$\|T_{1,n}\| \leqslant C_6 \cdot \sup_{0 \leqslant t \leqslant 1} \int_{-A}^{A} |W(t-hu)\, K'(u)\, u/\, du = O_p(1).$$

Since $S_{3,n}(t)$ is estimated as $S_{2,n}(t)$, we finally obtain the desired result.

The next lemma shows that the truncation introduced through $\{a_n\}$ does not affect the limiting distribution.

LEMMA 3.6. $\|Y_n - Y_{0,n}\| = O_p((\log n)^{-1/2})$.

Proof. We shall only show that $g'(t)^{-1/2} h^{-1/2} \iint_{\mathbb{R} - \Gamma_n} \psi(y - m(t))\, K((t-x)/h)\, dZ_n(x, y)$ fulfills the lemma.

The replacement of $g'(t)$ by $g(t)$ may be proved as in Johnston [9]. The quantity above is less than $h^{-1/2} \|g^{-1/2}\| \cdot \|\iint_{\{|y| > a_n\}} \psi(y - m(\cdot))\, K((\cdot - x)/h)\, dZ(x, y)\|$. It remains to show that the last factor tends to zero at a rate $O_p((\log n)^{1/2})$. We show first that

$$V_n(t) = (\log n)^{1/2} h^{-1/2} \iint_{\{|y| > a_n\}} \psi(y - m(t))\, K((t-x)/h)\, dZ_n(x, y)$$

$$\xrightarrow{P} 0 \qquad \text{for all } t$$

and then we show tightness of $V_n(t)$, the result then follows:

$$V_n(t) = (\log n^{1/2} (nh)^{-1/2} \sum_{i=1}^{n} \{\psi(Y_i - m(t)\, I_{\{|y| > a_n\}}(Y_i)\, K((t - X_i)/h)$$

$$- E\psi(Y_i - m(t)) \cdot I_{\{|y| > a_n\}}(Y_i)\, K((t - X_i)/h)\}$$

$$= \sum_{i=1}^{n} X_{n,i}(t),$$

where $\{X_{n,i}(t)\}_{i=1}^{n}$ are i.i.d. for each n with $EX_{n,i}(t) = 0$ for all $t \in [0, 1]$. We have then

$$EX_{n,i}^2(t) \leqslant (\log n)(nh)^{-1}\, E\psi^2(Y_i - m(t))\, I_{\{|y| > a_n\}}(Y_i)\, K^2((t - X_i)/h)$$

$$\leqslant \sup_{-A \leqslant u \leqslant A} K^2(u) \cdot (\log n)(nh)^{-1} E\psi^2(Y_i - m(t))\, I_{\{|y| > a_n\}}(Y_i);$$

hence

$$\operatorname{var}\{V_n(t)\} = E\left(\sum_{i=1}^{n} X_{n,i}(t)\right)^2 = n \cdot EX_{n,i}^2(t)$$

$$\leqslant \sup_{-A \leqslant u \leqslant A} K^2(u)\, h^{-1}(\log n) \int_{\{|y| > a_n\}} f_y(y)\, dy \cdot M_\psi,$$

where M_ψ denotes an upper bound for ψ^2. This term tends to zero by assumption (A3). Thus by Markov's inequality we conclude that

$$V_n(t) \xrightarrow{p} 0 \qquad \text{for all } t \in [0, 1].$$

To prove tightness of $\{V_n(t)\}$ we refer again to the following moment condition as stated in Lemma 3.1:

$$E\{|V_n(t) - V_n(t_1) \cdot |V_n(t_2) - V_n(t)|\} \leqslant C' \cdot (t_2 - t_1)^2$$

$$C' \text{ denoting a constant,} \qquad t \in [t_1, t_2].$$

We again estimate the left-hand side by Schwarz's inequality and estimate each factor separately,

$$E[V_n(t) - V_n(t_1)]^2 = (\log n)(nh)^{-1}\, E\left\{\sum_{i=1}^{n} \Psi_n(t, t_1, X_i, Y_i) \cdot I_{\{|y| > a_n\}}(Y_i)\right.$$

$$\left. - E(\Psi_n(t, t_1, X_i, Y_i) \cdot I_{\{|y| > a_n\}}(Y_i))\right\}^2,$$

where $\Psi_n(t, t_1, X_i, Y_i) = \psi(Y_i - m(t))\, K((t - X_i)/h) - \psi(Y_i - m(t_1))$ $K((t_1 - X_1)/h)$. Since ψ, m, K are Lipschitz continuous, it follows that

$$\{E[V_n(t) - V_n(t_1)]^2\}^{1/2}$$

$$\leqslant C_7 \cdot (\log n^{1/2} h^{-3/2}\, |t - t_1| \cdot \left\{\int_{\{|y| > a_n\}} f_y(y)\, dy\right\}^{1/2}.$$

It we apply the same estimations to $V_n(t_2) - V_n(t_1)$ we finally have

$$E\{|V_n(t) - V_n(t_1)| \cdot V_n(t_2) - V_n(t)|\}$$

$$\leqslant C_7^2\, (\log n)\, h^{-3} |t - t_1|\, t_2 - t| \times \int_{\{|y| > a_n\}} f_y(y)\, dy$$

$$\leqslant C' \cdot |t_2 - t_1|^2 \text{ since } t \in [t_1, t_2] \qquad \text{by (A3).}$$

LEMMA 3.7. *Let $\lambda(K)=\int K^2(u)\,du$ and let $\{d_n\}$ be as in the theorem. Then*

$$(2\delta \log n)^{1/2}[\|Y_{3,n}\|/[\lambda(K)]^{1/2}-d_n]$$

has the same asymptotic distribution as

$$(2\delta \log n)^{1/2}[\|Y_{4,n}\|/[\lambda(K)]^{1/2}-d_n].$$

Proof. $Y_{3,n}(t)$ is a Gaussian process with

$$EY_{3,n}(t)=0$$

and covariance function

$$\begin{aligned} r_3(t_1, t_2) &= EY_{3,n}(t_1)\, Y_{3,n}(t_2) \\ &= [g(t_1)\, g(t_2)]^{-1/2}\, h^{-1} \iint_{\Gamma_n} \psi^2(y-m(x))\, K((t_1-x)/h) \\ &\quad \times K((t_2-x)/h)\, f(x, y)\, dx\, dy. \\ &= h^{-1}[g(t_1)\, g(t_2)]^{-1/2} \iint_{\Gamma_n} \psi^2(y-m(x))\, f(y \mid x)\, dy K((t_1-x)/h) \\ &\quad \times K((t_2-x)/h)\, f_X(x)\, dx \\ &= h^{-1}[g(t_1)\, g(t_2)]^{-1/2} \int g(x)\, K((t_1-x)/h)\, K((t_2-x)/h)\, dx \\ &= r_4(t_1, t_2) \end{aligned}$$

where $r_4(t_1, t_2)$ is the covariance function of the Gaussian process $Y_{4,n}(t)$, which proves the lemma.

4. A MONTE CARLO STUDY

In a small Monte Carlo study $m_n(x)$, together with its uniform confidence band, and $m_n^*(x)$, the (linear) Nadaraya–Watson estimator, were compared. The pseudo-random number generators GGUW for uniform rv's in $[0, 1]$ and GGNPM for normal rv's (both from the IMSL package) were used to generate bivariate data $\{(X_i, Y_i)\}_{i=1}^{n}$, $n=100$ with joint pdf

$$\begin{aligned} f(x, y) &= g(y-m(x))\, I_{[0,1]}(x) \\ g(u) &= \tfrac{9}{10}\varphi(y) + \tfrac{1}{90}\varphi(u/9). \end{aligned} \tag{4.1}$$

We took $m(x) = \sin(\pi x)$ and used the kernel

$$K(u) = \tfrac{3}{4}(1 - u^2), \qquad |u| \leqslant 1,$$
$$= 0, \qquad |u| > 1.$$

In Fig. 1 the raw data, together with the regression curve $m(x)$, is displayed. The random variables generated with probability $\frac{1}{10}$ from the longtailed pdf $\frac{1}{9}\varphi(u/9)$ (see (4.1)) are marked as squares whereas the standard normal rv's are shown as stars.

We then computed both $m_n^*(x)$ and $m_n(x)$ from the data. The bandwidth was set to $n^{-1/4} \approx \frac{1}{3}$ and Huber's ψ-function with a cutoff point of $c = 0.2$ was used. In Fig. 2 the two estimators together with the uniform confidence band (according to the corollary) with 95% coverage probability is shown. The true regression curve and the confidence band are shown as fine dotted lines, whereas the robust M-smoothers are shown as a solid line and the Nadaraya–Watson estimate is displayed as a broken line. The raw data is overlaid with the same conventions as for Fig. 1, but note that some of the outliers are clipped since Fig. 2 has a different scale. At first sight $m_n^*(x)$ has clearly more variation and has the expected sensitivity to outliers. A closer look reveals that $m_n^*(x)$ for $x \approx 0.55$ even leaves the confidence band. It may be surprising that this happens at $x \approx 0.55$ where no outlier is

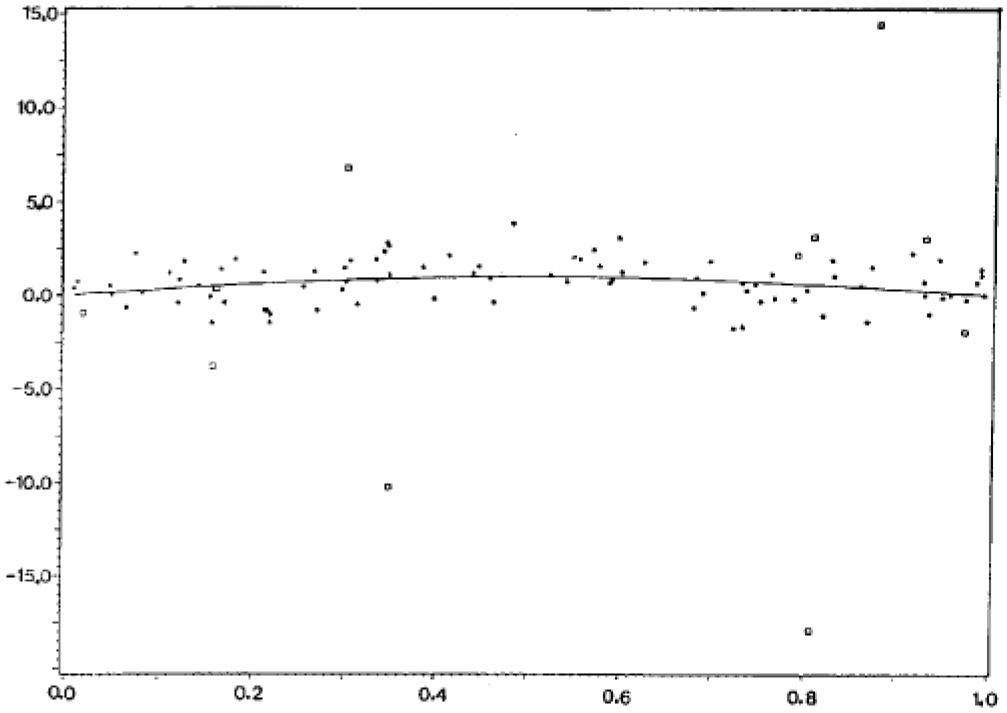


FIG. 1. Raw data with outliers. The regression curve $m(x) = \sin(\pi x)$ and the raw data points.

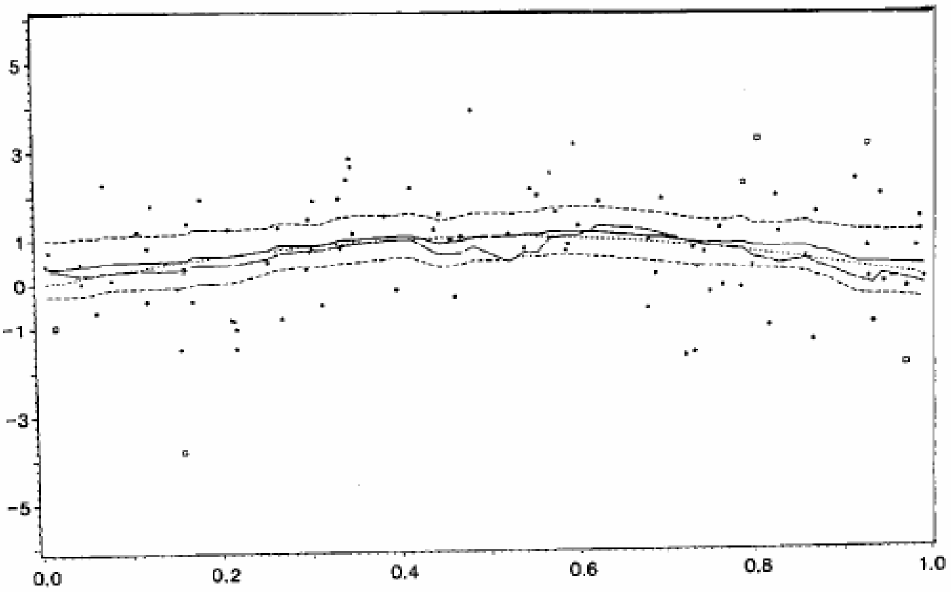


FIG. 2. Smoothed data with uniform confidence bands. The regression curve $m(x) = \sin(\pi x)$, the M-smoother $m_n(x)$, the Nadaraya–Watson estimator $m_n^*(x)$, and 95% confidence band.

placed, but a closer look at Fig. 1 shows that the large negative data value at $x \approx 0.8$ causes the trouble. This data value is inside the window $(h \approx \frac{1}{3})$ and therefore distorts $m_n^*(x)$ for $x \approx 0.55$, whereas the estimate $m_n^*(0.8)$ is not affected since the positive huge outlier at $x \approx 0.9$ balances the sensitivity effect (symmetry assumption). The M-smoother $m_n(x)$ (solid line) is unaffected and stays fairly close to the true reression curve $m(x)$.

ACKNOWLEDGMENTS

The author wishes to thank R. J. Carroll and an anonymous referee for helpful remarks and suggestions.

REFERENCES

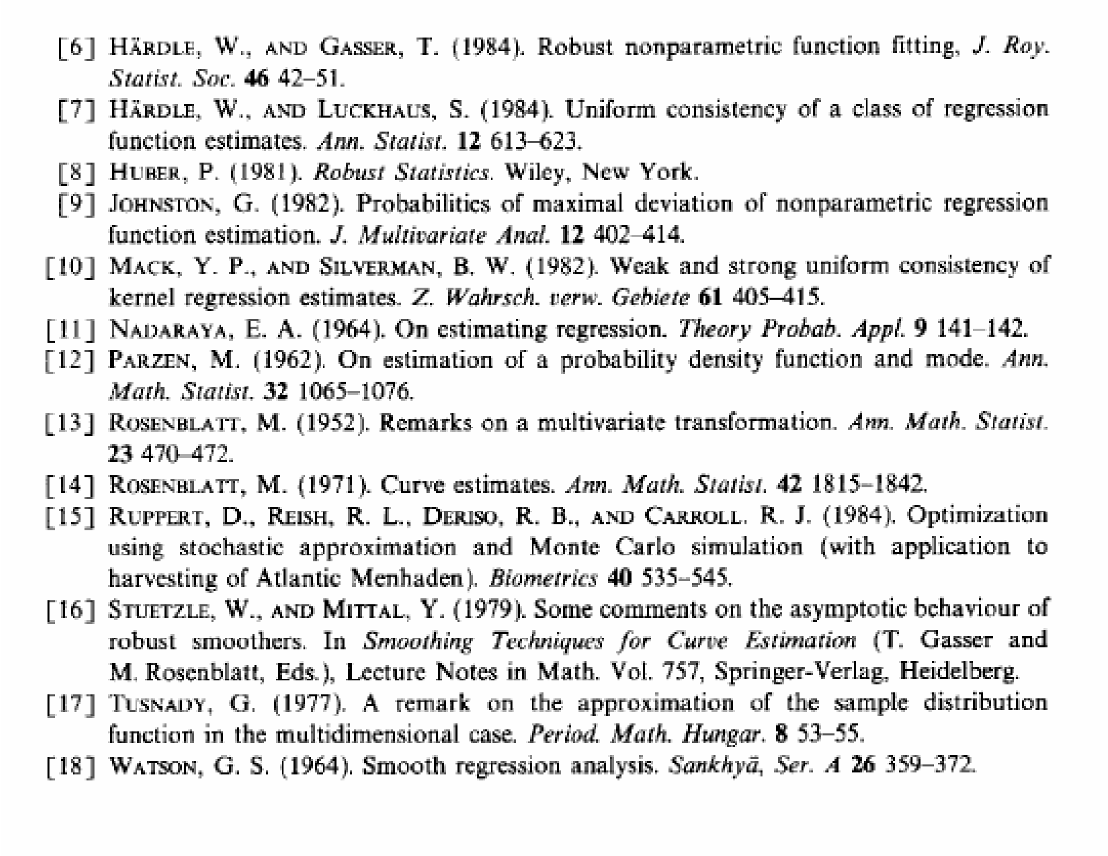
[1] BICKEL, P., AND ROSENBLATT, M. (1973). On some global measures of the deviation of density function estimators. *Ann. Statist.* **1** 1071–1095.

[2] BILLINGSLEY, P. (1968). *Convergence of Probability Measures.* Wiley, New York.

[3] BUSSIAN, B. M., AND HÄRDLE, (1984). Robust smoothing applied to white noise and single outlier contaminated Raman spectra. *Appl. Spectroscopy* **38** 309–313.

[4] CSÖRGŐ, S., AND HALL, P. (1982). Upper and lower classes for triangular arrays. *Z. Wahrscheinlichkeitstheorie und verw. Gebiete* **61** 207–222.

[5] HÄRDLE, W. (1982). Robust regression function estimation. *J. Multivariate Anal.* **14** 169–180.



[6] Härdle, W., and Gasser, T. (1984). Robust nonparametric function fitting, *J. Roy. Statist. Soc.* **46** 42–51.

[7] Härdle, W., and Luckhaus, S. (1984). Uniform consistency of a class of regression function estimates. *Ann. Statist.* **12** 613–623.

[8] Huber, P. (1981). *Robust Statistics.* Wiley, New York.

[9] Johnston, G. (1982). Probabilitics of maximal deviation of nonparametric regression function estimation. *J. Multivariate Anal.* **12** 402–414.

[10] Mack, Y. P., and Silverman, B. W. (1982). Weak and strong uniform consistency of kernel regression estimates. *Z. Wahrsch. verw. Gebiete* **61** 405–415.

[11] Nadaraya, E. A. (1964). On estimating regression. *Theory Probab. Appl.* **9** 141–142.

[12] Parzen, M. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.* **32** 1065–1076.

[13] Rosenblatt, M. (1952). Remarks on a multivariate transformation. *Ann. Math. Statist.* **23** 470–472.

[14] Rosenblatt, M. (1971). Curve estimates. *Ann. Math. Statist.* **42** 1815–1842.

[15] Ruppert, D., Reish, R. L., Deriso, R. B., and Carroll. R. J. (1984). Optimization using stochastic approximation and Monte Carlo simulation (with application to harvesting of Atlantic Menhaden). *Biometrics* **40** 535–545.

[16] Stuetzle, W., and Mittal, Y. (1979). Some comments on the asymptotic behaviour of robust smoothers. In *Smoothing Techniques for Curve Estimation* (T. Gasser and M. Rosenblatt, Eds.), Lecture Notes in Math. Vol. 757, Springer-Verlag, Heidelberg.

[17] Tusnady, G. (1977). A remark on the approximation of the sample distribution function in the multidimensional case. *Period. Math. Hungar.* **8** 53–55.

[18] Watson, G. S. (1964). Smooth regression analysis. *Sankhyā, Ser. A* **26** 359–372.

SYMMETRIZED NEAREST NEIGHBOR REGRESSION ESTIMATES

R.J. CARROLL *

Department of Statistics, Texas A & M University, College Station, TX 77843, USA.

W. HÄRDLE

Universität Bonn, Rechts- und Staatswissenschaftliche Fakultät, Wirtschaftstheoretische Abteilung II, Adenauerallee 24–26, D-5300 Bonn, FR Germany

Received December 1987
Revised March 1988

Abstract: We consider univariate nonparametric regression. Two standard nonparametric regression function estimates are kernel estimates and nearest neighbor estimates. Mack (1981) noted that both methods can be defined with respect to a kernel or weighting function, and that for a given kernel and a suitable choice of bandwidth, the optimal mean squared error is the same asymptotically for kernel and nearest neighbor estimates. Yang (1981) defined a new type of nearest neighbor regression estimate using the empirical distribution function of the predictors to define the window over which to average. This has the effect of forcing the number of neighbors to be the same both above and below the value of the predictor of interest; we call these symmetrized nearest neighbor estimates. The estimate is a kernel regression estimate with "predictors" given by the empirical distribution function of the true predictors. We show that for estimating the regression function at a point, the optimum mean squared error of this estimate differs from that of the optimum mean squared error for kernel and ordinary nearest neighbor estimates. No estimate dominates the others. They are asymptotically equivalent with respect to mean squared error if one is estimating the regression function at a mode of the predictor.

Keywords: nonparametric regression, kernel regression, nearest neighbor regression, bias, mean squared error.

We consider nonparametric regression with a random univariate predictor. Let (X, Y) be a bivariate random variable with joint distribution H, and denote the regression function of Y on X by $m(x) = E(Y \mid X = x)$. If it exists, let f_x denote the marginal density of X. A sample of size n is taken, (y_i, x_i) for $i = 1, \ldots, n$. Two common estimates of the regression function are the Nadaraya–Watson kernel estimate and the nearest neighbor estimate, see Nadaraya (1964), Watson (1964) and Stute (1984) for the former, and Mack (1981) for the latter. Fix x_0 and suppose we wish to estimate $m(x_0)$. The kernel and nearest neighbor estimates are defined as follows. Let K be a nonnegative even density function.

Kernel estimates, Let h_{ker} be a bandwidth depending on n. Then the kernel estimate is

$$\hat{m}_{\text{ker}}(x_0) = \frac{\sum_{i=1}^{n} y_i K\left(\frac{x_i - x_0}{h_{\text{ker}}}\right)}{\sum_{i=1}^{n} K\left(\frac{x_i - x_0}{h_{\text{ker}}}\right)}. \tag{1}$$

Nearest neighbor estimates. Let $k = k(n)$ be a sequence of positive integers, and let R_n be the Euclidean distance between x_0 and its kth nearest neighbor. Then the nearest neighbor estimate is

$$\hat{m}_{kNN}(x_0) = \frac{\sum_{i=1}^{n} y_i K\left(\frac{x_i - x_0}{R_n}\right)}{\sum_{i=1}^{n} K\left(\frac{x_i - x_0}{R_n}\right)}. \tag{2}$$

* Research supported by the Air Force Office of Scientific Research and by Sonderforschungsbereich 303, Universität Bonn.

0167-7152/89/$3.50 © 1989, Elsevier Science Publishers B.V. (North-Holland)

Under the differentiability conditions on the marginal density f_x, Mack has shown that the asymptotically optimal versions of the kernel and nearest neighbor estimates have the same behavior. Let $m^{(j)}$ and $f_x^{(j)}$ denote the jth derivative of m and f_x respectively. If $c_K = \int K^2(x)\,\mathrm{d}x$ and $d_K = \int x^2 K(x)\,\mathrm{d}x$, remembering that K is symmetric, the kernel estimate has bias

$$\begin{aligned}\text{bias}_{\text{ker}} = h_{\text{ker}}^2\, d_K &\\ \times \frac{m^{(2)}(x_0) f_x(x_0) + 2m^{(1)}(x_0) f_x^{(1)}(x_0)}{2 f_x(x_0)}&\\ + \mathrm{o}\left(h_{\text{ker}}^2\right)& \qquad (3)\end{aligned}$$

and variance

$$\begin{aligned}\text{var}_{\text{ker}} = c_K \operatorname{Var}(Y \mid X = x_0)/(nh_{\text{ker}} f_x(x_0))&\\ + \mathrm{o}\left((nh_{\text{ker}})^{-1}\right).& \qquad (4)\end{aligned}$$

Of course, (3) is not the exact bias of the kernel estimator but is instead an asymptotic bias based upon a linearization argument. There is obviously a bias versus variance trade-off here, so that if one wants to achieve the minimum mean squared error, the optimal bandwidth is $h_{\text{ker}} \sim n^{-1/5}$ and the optimal mean squared error is of order $\mathrm{O}(n^{-4/5})$. The formulae for bias and variance of the kth nearest neighbor estimate are the same as in (3) and (4) if one substitutes $2f_x(x_0)nh_{\text{ker}}$ for k.

Let F denote the distribution function of X, and let F_n denote the empirical distribution of the sample from X. Let h_{snn} be a bandwidth tending to zero. The estimate proposed by Yang (1981) and studied by Stute (1984) is

$$\begin{aligned}\hat{m}_{snn}(x_0) = (nh_{snn})^{-1}&\\ \times \sum_{i=1}^{n} y_i K\left(\frac{F_n(x_i) - F_n(x_0)}{h_{snn}}\right).& \qquad (5)\end{aligned}$$

The nearest neighbor estimate defines neighbors in terms of the Euclidean norm, which in this case is just absolute difference. The estimate (5) is also a nearest neighbor estimate, but now neighbors are defined in terms of distance based on empirical distribution function. This makes for computational efficiency if the uniform kernel is used. A direct application of (5) would result in $\mathrm{O}(n^2 h)$ operations, but using updating as the window moves over the span of the x's results in $\mathrm{O}(n)$ operations. Other smooth kernels can be computed efficiently by iterated smoothing, i.e., higher order convolution of the uniform kernel. Another possible device is the Fast Fourier transform (Härdle, 1987). Since the difference between (2) and (5) is that (5) picks its neighbors symmetrically, we call it a symmetrized nearest neighbor estimate. Note that $\hat{m}_{kNN}$ always averages over a symmetric neighborhood in the x-space, but may have an asymmetric distribution of x points in this neighborhood. By contrast, $\hat{m}_{snn}$ always averages over the same amount of points left and right of x_0, but may in effect average over an asymmetric neighborhood in the x-space. The esimate $\hat{m}_{snn}$ has an intriguing relationship with the k-NN estimator used by Friedman (1986). The variable span smoother proposed by Friedman uses the same type of neighborhood as does $\hat{m}_{snn}$ and is used as an elementary building block for ACE, see Breiman and Friedman (1985). The estimate (5) also looks appealingly like a kernel regression estimate of Y against not X but rather $F_n(X)$. Define

$$\begin{aligned}&\bar{m}_{snn}(x_0)\\ &\quad = h_{snn}^{-1} \int m(x) K\left(\frac{F(x) - F(x_0)}{h_{snn}}\right) F(\mathrm{d}x). \qquad (6)\end{aligned}$$

Then Stute shows that as as $n \to \infty$, $h_{snn} \to 0$ and $nh_{snn}^3 \to \infty$,

$$\begin{aligned}&(nh_{snn})^{1/2}\left(\hat{m}_{snn}(x_0) - \bar{m}_{snn}(x_0)\right)\\ &\quad \Rightarrow \quad \text{Normal}\left(0,\, c_K \operatorname{Var}(Y \mid X = x_0)\right). \qquad (7)\end{aligned}$$

This has the form (4) as long as $h_{snn} = h_{\text{ker}} f_x(x_0)$. However, the asymptotic bias term (see just after (4)) is not the same as the kernel estimator. If we define

$$\text{bias}_{snn} = \bar{m}_{snn}(x_0) - m(x_0),$$

then it follows from Stute (1984, p. 925) that

$$\begin{aligned}\text{bias}_{snn} &= h_{snn}^{-1} \int \left[m(x) - m(x_0)\right]\\ &\qquad \times K\left(\frac{F(x_0) - F(x)}{h_{snn}}\right) F(\mathrm{d}x)\\ &= \int \left[m \circ F^{-1}\left(F(x_0) - uh_{snn}\right)\right.\\ &\qquad \left. - m \circ F^{-1}\left(F(x_0)\right)\right] K(u)\,\mathrm{d}u,\end{aligned}$$

so that by a simple Taylor series expansion,

$$\text{bias}_{snn} = h^2_{snn}\, d_K \frac{m^{(2)}(x_0)f_x(x_0) - m^{(1)}(x_0)f_x^{(1)}(x_0)}{2f_x^3(x_0)} + \text{o}\left(h^2_{snn}\right). \tag{8}$$

Comparison of (3) and (8) shows that even when the variances of all three estimates are the same (the case $h_{snn} = h_{\text{ker}} f_x(x_0)$), the bias properties differ unless

$$m^{(1)}(x_0)f_x^{(1)}(x_0) = 0.$$

Otherwise, the optimal choice of bandwidth for the kernel and ordinary nearest neighbor estimates will lead to a different mean squared error than what obtains for the symmetrized nearest neighbor estimate.

The preceeding discussion presumed that we are interested in estimating the regression function only at the point x_0 and that bandwidth was chosen locally so as to minimize asymptotic mean squared error. In practice, one is usually interested in the regression curve over an interval, and the bandwidth is chosen globally, see for example Härdle, Hall and Marron (1988). Inspection of (3), (4) and (8) shows the usual tradeoff between kernel and nearest neighbor estimates; in the tails of the distribution of x, the former are more variable but less biased.

The symmetrized nearest neighbor estimate is a kernel estimate based on transforming the x data by F_n. Other transformations are possible, e.g., $\log(x)$. In general, if we transform by $w = G(x)$, if $m_*(w) = m(x)$ and w has density f_w, then the bias and variance properties of the resulting kernel estimate are given by (3)–(4) in m_* and f_w, the translation to f_x and m being immediate by the chain rule.

Example. For illustrative purposes we use a large data set ($n = 7125$) of the relationship of $Y =$ expenditure for potatoes versus $X =$ net income of British households (in tenth of a pence) in 1973. The data come from the *Family Expenditure Survey, Annual Base Tapes 1968–1983*, Department of Employment, Statistics Division, Her Majesty's Stationary Office, London, and were made available by the ESRC Data Archive at the University of Essex. See Härdle (1988, Chapter 1) for a discussion. A sunflower plot of the data is given in Figure 1. For these data, we used the quartic kernel

$$K(u) = \tfrac{15}{16}\left(1 - u^2\right)^2 I(|u| \leqslant 1).$$

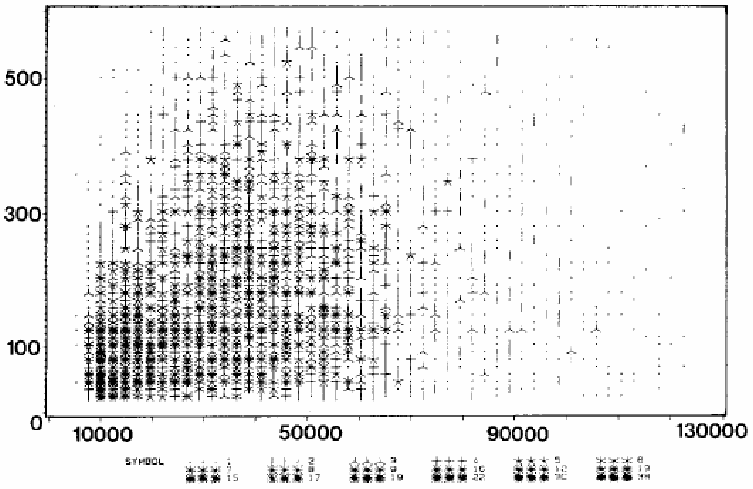


Fig. 1. Potatoes vs. Netincome. Sunflower Plot of $Y =$ expenditure for Potatoes versus $X =$ net income of British households (both reported in tenth of a Pence) for the Year 1973. $n = 7125$. The number of petals of the sunflower indicates the frequency of observations falling in the cell covered by the sunflower.

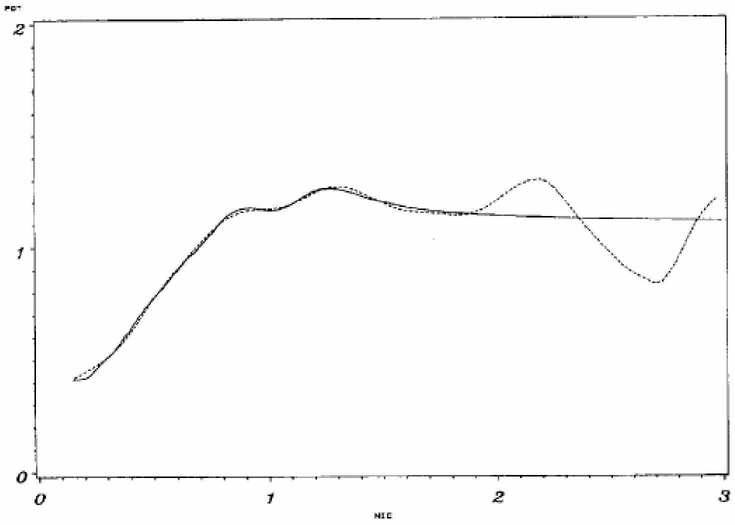


Fig. 2.

We computed the ordinary kernel estimate (1) and the symmetrized nearest neighbor estimate (5), the bandwidths being selected by crossvalidation, see Härdle and Marron (1985). The crossvalidated bandwidths were $h_{\text{ker}} = 0.25$ on the scale (0,3) of Figure 2 and $h_{snn} = 0.15$ on the F_n scale. The resulting regression curves are plotted in Figure 2. The two curves are similar for $x \leqslant 1$, which is where most of the data lie. There is a sharp discrepancy for larger values of x, the kernel estimate showing evidence of a bimodal relationship and the symmetrized neighbor estimate indicating either an asymptote or even a slight decrease as income rises. In the context, the latter seems to make more sense economically and looks quite similar to to curve in Hildebrand and Hildebrand (1986). Statistically, it is in this range of the data that the density f_x takes on small values, which is exactly when we expect the biggest differences in the estimates, i.e., the kernel estimate should be more variable but less biased.

References

Breiman, L. and J. Friedman (1985), Estimating optimal transformations for multiple regression and correlation, *Journal of the American Statistical Association* **80**, 580–619.

Friedman, J. (1986). A variable span smoother, Department of Statistics Technical Report LCS5, Stanford University.

Härdle, W. (1987), Resistant smoothing using the Fast Fourier transform, AS 222, *Applied Statistics* **36**, 104–111.

Härdle, W. (1988) *Applied Nonparametric Regression*, to appear.

Härdle, W., P. Hall and J.S. Marron (1988), How far are automatically chosen regression smoothing parameters away from their optimum? (with discussion), *Journal of the American Statistical Association*, to appear.

Härdle, W. and J.S. Marron (1985). Optimal bandwidth selection in nonparametric regression function estimation, *Annals of Statistics* **13**, 1465–1481.

Hildenbrand, K. and W. Hildenbrand (1986). On the mean income effect: a data analysis of the U.K. family expenditure survey, in: W. Hildebrand and A. Mas-Colell, eds., *Contributions to Mathematical Economics* (North-Holland, Amsterdam).

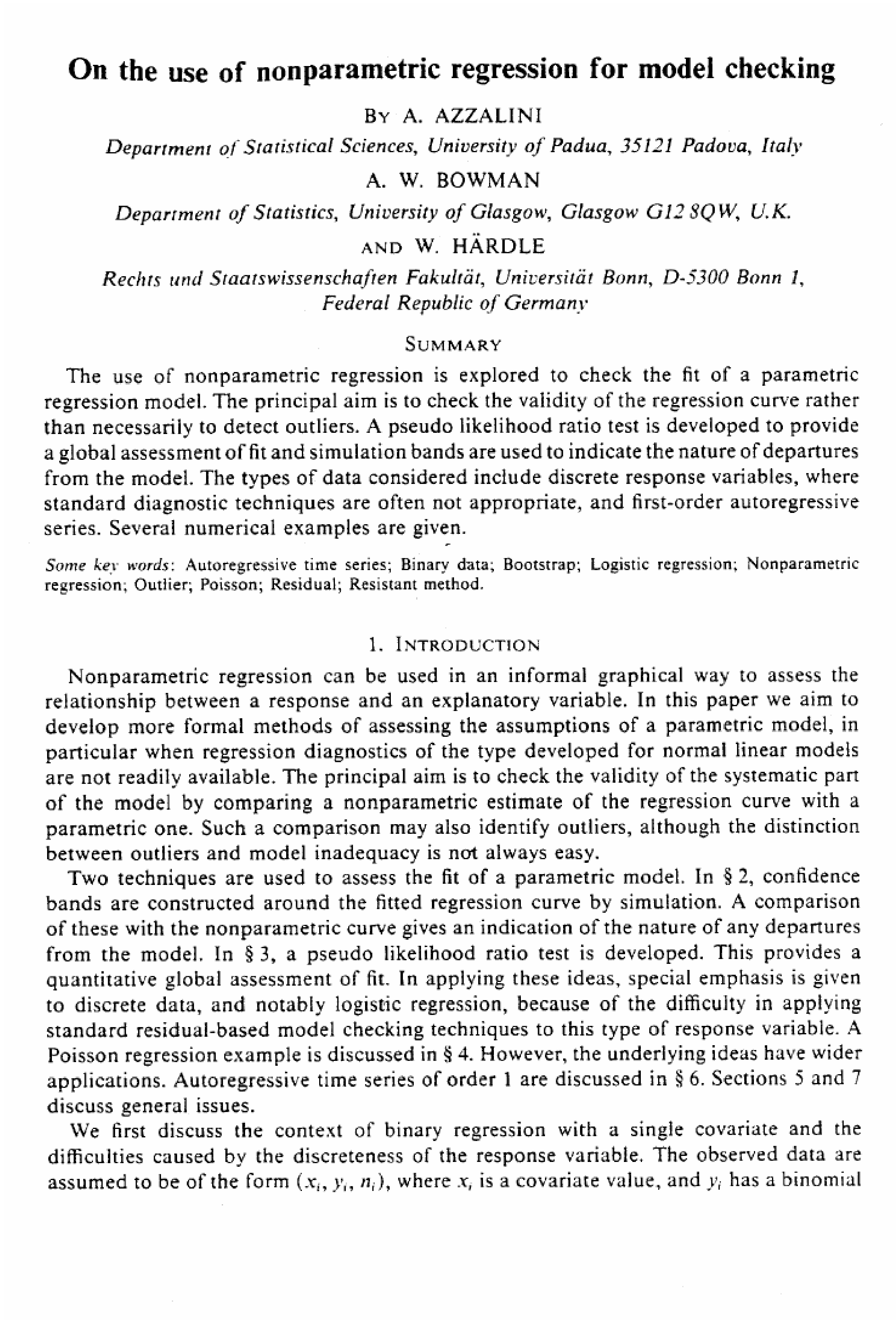
Mack, Y.P. (1981), Local properties of *k-NN* regression estimates, *SIAM J. Alg. Disc. Meth.* **2**, 311–323.

Nadaraya, E.A. (1964), On estimating regression, *Theory of Probability and its Applications* **9**, 141–142.

Stute, W. (1984), Asymptotic normality of nearest neighbor regression function estimates, *Annals of Statistics* **12**, 917–926.

Watson, G.S. (1964), Smooth regression analysis, *Sankhyā Series A* **26**, 359–372.

Yang, S.S. (1981), Linear functions of concomitants of order statistics with applications to nonparametric estimation of a regression function, *Journal of the American Statistical Association* **76**, 658–662.



On the use of nonparametric regression for model checking

By A. AZZALINI

Department of Statistical Sciences, University of Padua, 35121 Padova, Italy

A. W. BOWMAN

Department of Statistics, University of Glasgow, Glasgow G12 8QW, U.K.

AND W. HÄRDLE

Rechts und Staatswissenschaften Fakultät, Universität Bonn, D-5300 Bonn 1,
Federal Republic of Germany

Summary

The use of nonparametric regression is explored to check the fit of a parametric regression model. The principal aim is to check the validity of the regression curve rather than necessarily to detect outliers. A pseudo likelihood ratio test is developed to provide a global assessment of fit and simulation bands are used to indicate the nature of departures from the model. The types of data considered include discrete response variables, where standard diagnostic techniques are often not appropriate, and first-order autoregressive series. Several numerical examples are given.

Some key words: Autoregressive time series; Binary data; Bootstrap; Logistic regression; Nonparametric regression; Outlier; Poisson; Residual; Resistant method.

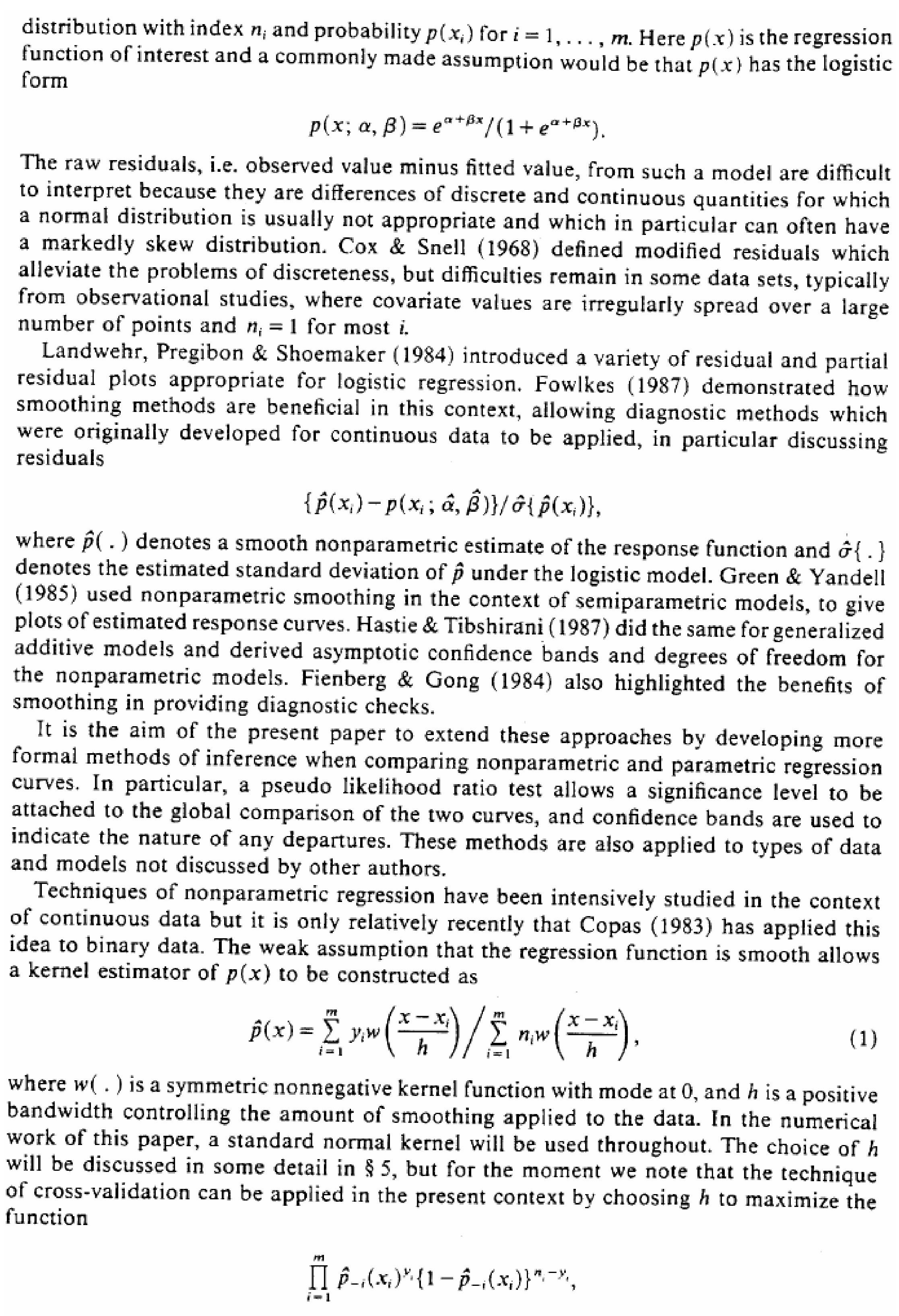
1. Introduction

Nonparametric regression can be used in an informal graphical way to assess the relationship between a response and an explanatory variable. In this paper we aim to develop more formal methods of assessing the assumptions of a parametric model, in particular when regression diagnostics of the type developed for normal linear models are not readily available. The principal aim is to check the validity of the systematic part of the model by comparing a nonparametric estimate of the regression curve with a parametric one. Such a comparison may also identify outliers, although the distinction between outliers and model inadequacy is not always easy.

Two techniques are used to assess the fit of a parametric model. In § 2, confidence bands are constructed around the fitted regression curve by simulation. A comparison of these with the nonparametric curve gives an indication of the nature of any departures from the model. In § 3, a pseudo likelihood ratio test is developed. This provides a quantitative global assessment of fit. In applying these ideas, special emphasis is given to discrete data, and notably logistic regression, because of the difficulty in applying standard residual-based model checking techniques to this type of response variable. A Poisson regression example is discussed in § 4. However, the underlying ideas have wider applications. Autoregressive time series of order 1 are discussed in § 6. Sections 5 and 7 discuss general issues.

We first discuss the context of binary regression with a single covariate and the difficulties caused by the discreteness of the response variable. The observed data are assumed to be of the form (x_i, y_i, n_i), where x_i is a covariate value, and y_i has a binomial

Azzalini, A., Bowman, A. and Härdle, W. (1989) On the use of nonparametric regression for model checking



distribution with index n_i and probability $p(x_i)$ for $i = 1, \ldots, m$. Here $p(x)$ is the regression function of interest and a commonly made assumption would be that $p(x)$ has the logistic form

$$p(x;\, \alpha, \beta) = e^{\alpha+\beta x}/(1+e^{\alpha+\beta x}).$$

The raw residuals, i.e. observed value minus fitted value, from such a model are difficult to interpret because they are differences of discrete and continuous quantities for which a normal distribution is usually not appropriate and which in particular can often have a markedly skew distribution. Cox & Snell (1968) defined modified residuals which alleviate the problems of discreteness, but difficulties remain in some data sets, typically from observational studies, where covariate values are irregularly spread over a large number of points and $n_i = 1$ for most i.

Landwehr, Pregibon & Shoemaker (1984) introduced a variety of residual and partial residual plots appropriate for logistic regression. Fowlkes (1987) demonstrated how smoothing methods are beneficial in this context, allowing diagnostic methods which were originally developed for continuous data to be applied, in particular discussing residuals

$$\{\hat{p}(x_i) - p(x_i;\, \hat{\alpha}, \hat{\beta})\}/\hat{\sigma}\{\hat{p}(x_i)\},$$

where $\hat{p}(\,.\,)$ denotes a smooth nonparametric estimate of the response function and $\hat{\sigma}\{\,.\,\}$ denotes the estimated standard deviation of $\hat{p}$ under the logistic model. Green & Yandell (1985) used nonparametric smoothing in the context of semiparametric models, to give plots of estimated response curves. Hastie & Tibshirani (1987) did the same for generalized additive models and derived asymptotic confidence bands and degrees of freedom for the nonparametric models. Fienberg & Gong (1984) also highlighted the benefits of smoothing in providing diagnostic checks.

It is the aim of the present paper to extend these approaches by developing more formal methods of inference when comparing nonparametric and parametric regression curves. In particular, a pseudo likelihood ratio test allows a significance level to be attached to the global comparison of the two curves, and confidence bands are used to indicate the nature of any departures. These methods are also applied to types of data and models not discussed by other authors.

Techniques of nonparametric regression have been intensively studied in the context of continuous data but it is only relatively recently that Copas (1983) has applied this idea to binary data. The weak assumption that the regression function is smooth allows a kernel estimator of $p(x)$ to be constructed as

$$\hat{p}(x) = \sum_{i=1}^{m} y_i w\left(\frac{x-x_i}{h}\right) \bigg/ \sum_{i=1}^{m} n_i w\left(\frac{x-x_i}{h}\right), \qquad (1)$$

where $w(\,.\,)$ is a symmetric nonnegative kernel function with mode at 0, and h is a positive bandwidth controlling the amount of smoothing applied to the data. In the numerical work of this paper, a standard normal kernel will be used throughout. The choice of h will be discussed in some detail in § 5, but for the moment we note that the technique of cross-validation can be applied in the present context by choosing h to maximize the function

$$\prod_{i=1}^{m} \hat{p}_{-i}(x_i)^{y_i}\{1-\hat{p}_{-i}(x_i)\}^{n_i-y_i},$$

where $p_{-i}(\,.\,)$ denotes the nonparametric estimator constructed from the data with the ith observation omitted. Although such an approach has not been studied in the present context, similar likelihood criteria have been extensively investigated in the related areas of regression with continuous data and in density estimation, where strong theoretical justification has been provided (Härdle & Marron, 1985).

2. Simulation bands

As an illustrative example of logistic regression, we use the data of Finney (1947) which has been analysed by several other authors. The data consist of 39 observations on the presence or absence of vasoconstriction in the skin of the digits at a variety of volumes and rates of air flow. In order to keep the development simple, we shall employ a single covariate, x equal to log (volume) plus log (rate), as implicitly suggested by the author. The case of two separate covariates will be discussed at a later stage. Since most of the n_i's are 1, it is appropriate to pool the information in neighbouring data using the estimator (1) under the assumption that the relationship between the mean of y and x is smooth.

The cross-validatory bandwidth in this case is $h = 0{\cdot}06$ and Fig. 1 displays the nonparametric regression estimate; two observations, numbers 4 and 18 in Finney's (1947) listing, cause a large peak in the estimate near $x = 0{\cdot}2$. This agrees with the analyses of other authors who found these observations to be outliers. In this case the departure of the regression function from the logistic shape is extreme but in general there is a problem in assessing whether observed differences indicate significant departures from the model.

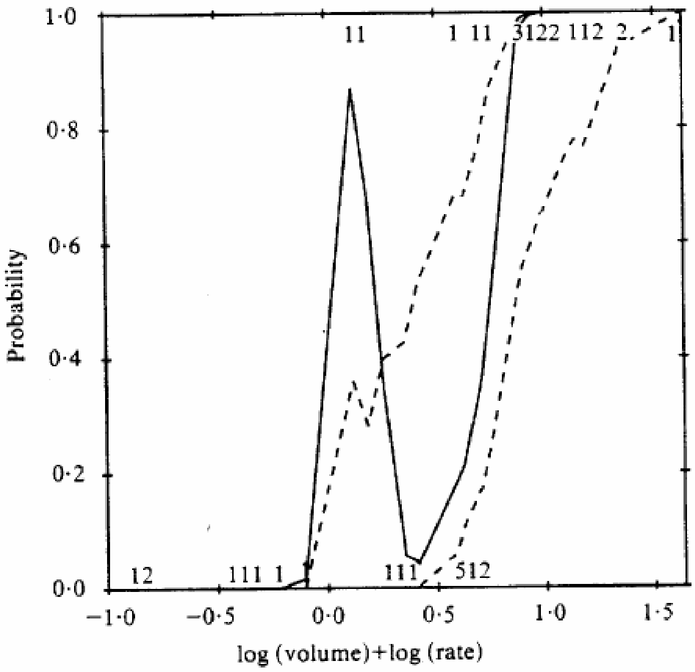


Fig. 1. Finney's data, with nonparametric estimate of regression function, shown by solid line, and approximate 95% confidence bands derived by simulation from logistic model, broken lines. Frequencies of zeros and ones indicated at top and bottom of graph.

First, the logistic model must be fitted. Since one of the purposes of model checking and diagnostics is to identify outliers, it is more appropriate that the unknown parameters are estimated by a resistant technique rather than by maximum likelihood. Pregibon (1982) describes a resistant technique for fitting binomial data which is approximately 95% efficient when the chosen model is correct, and this leads to the estimates $\hat{\alpha} = -5\cdot252$, $\hat{\beta} = 7\cdot719$ for Finney's data. Copas (1988) discusses general issues associated with resistant fitting and proposes a simple alternative model.

We now compare the fitted model with the data by constructing simulation bands for the nonparametric curve under the assumption that the logistic model is correct. It is straightforward to estimate the mean and variance of $\hat{p}$ under the logistic model, but the use of simulation removes the assumption of normality implicit in the use of ±2 standard deviations as a reference. Moreover, the same simulations will be used for an alternative technique to be described in § 3. Simulation was employed by Atkinson (1981) to produce an envelope on a probability plot of residuals from a regression model. Landwehr et al. (1984) used an analogous plot with raw residuals from a logistic fit. The pooling of neighbouring information involved in the estimator (1) has the attractive feature of making use of the smoothness of $p(x)$ and allows assessment of the adequacy of the model to be carried out on the probability scale, which is the natural one for exploratory purposes.

To construct pointwise simulation bands in the present context, a complete set of simulated responses $\{y_1^*, \dots, y_m^*\}$ is derived from the fitted model, that is y_i^* has a binomial $(n_i, p(x_i; \hat{\alpha}, \hat{\beta}))$ distribution, and a new nonparametric estimate p^* is produced, using

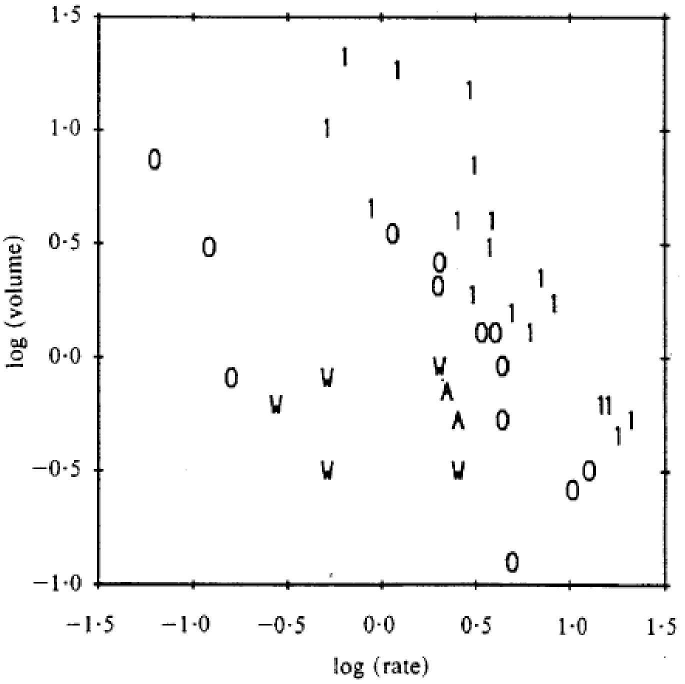
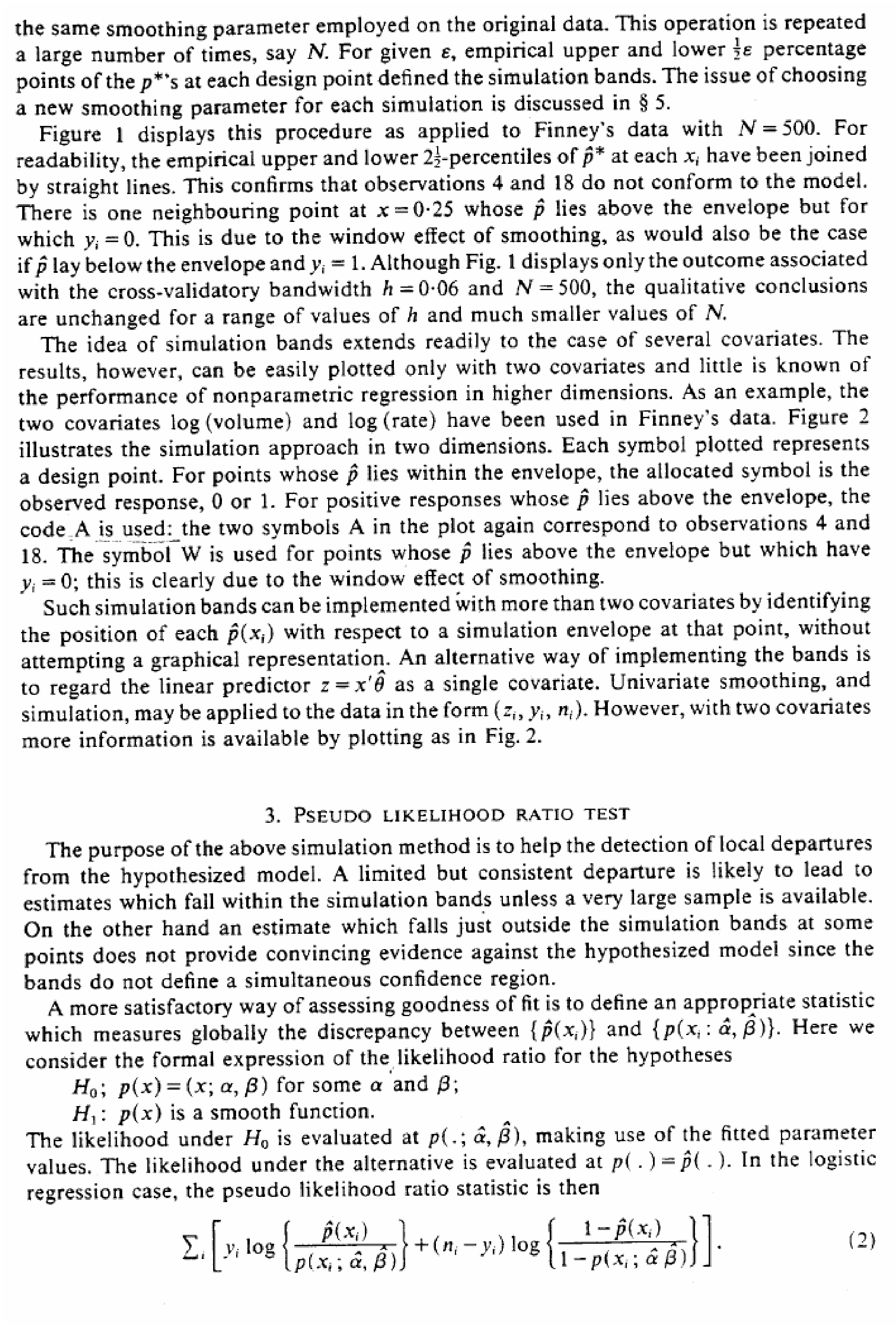


Fig. 2. Finney's data with two covariates and simulation envelope, 95% level, derived from logistic model; 0, 1, observed response whose $\hat{p}$ lies within envelope; A, positive response whose $\hat{p}$ lies above the envelope; W, response whose $\hat{p}$ lies outside the envelope, due only to window effect of smoothing.



the same smoothing parameter employed on the original data. This operation is repeated a large number of times, say N. For given ε, empirical upper and lower $\frac{1}{2}\varepsilon$ percentage points of the p^*'s at each design point defined the simulation bands. The issue of choosing a new smoothing parameter for each simulation is discussed in § 5.

Figure 1 displays this procedure as applied to Finney's data with $N = 500$. For readability, the empirical upper and lower $2\frac{1}{2}$-percentiles of $\hat{p}^*$ at each x_i have been joined by straight lines. This confirms that observations 4 and 18 do not conform to the model. There is one neighbouring point at $x = 0{\cdot}25$ whose $\hat{p}$ lies above the envelope but for which $y_i = 0$. This is due to the window effect of smoothing, as would also be the case if $\hat{p}$ lay below the envelope and $y_i = 1$. Although Fig. 1 displays only the outcome associated with the cross-validatory bandwidth $h = 0{\cdot}06$ and $N = 500$, the qualitative conclusions are unchanged for a range of values of h and much smaller values of N.

The idea of simulation bands extends readily to the case of several covariates. The results, however, can be easily plotted only with two covariates and little is known of the performance of nonparametric regression in higher dimensions. As an example, the two covariates log (volume) and log (rate) have been used in Finney's data. Figure 2 illustrates the simulation approach in two dimensions. Each symbol plotted represents a design point. For points whose $\hat{p}$ lies within the envelope, the allocated symbol is the observed response, 0 or 1. For positive responses whose $\hat{p}$ lies above the envelope, the code A is used: the two symbols A in the plot again correspond to observations 4 and 18. The symbol W is used for points whose $\hat{p}$ lies above the envelope but which have $y_i = 0$; this is clearly due to the window effect of smoothing.

Such simulation bands can be implemented with more than two covariates by identifying the position of each $\hat{p}(x_i)$ with respect to a simulation envelope at that point, without attempting a graphical representation. An alternative way of implementing the bands is to regard the linear predictor $z = x'\hat{\theta}$ as a single covariate. Univariate smoothing, and simulation, may be applied to the data in the form (z_i, y_i, n_i). However, with two covariates more information is available by plotting as in Fig. 2.

3. Pseudo likelihood ratio test

The purpose of the above simulation method is to help the detection of local departures from the hypothesized model. A limited but consistent departure is likely to lead to estimates which fall within the simulation bands unless a very large sample is available. On the other hand an estimate which falls just outside the simulation bands at some points does not provide convincing evidence against the hypothesized model since the bands do not define a simultaneous confidence region.

A more satisfactory way of assessing goodness of fit is to define an appropriate statistic which measures globally the discrepancy between $\{\hat{p}(x_i)\}$ and $\{p(x_i : \hat{\alpha}, \hat{\beta})\}$. Here we consider the formal expression of the likelihood ratio for the hypotheses

H_0; $p(x) = (x; \alpha, \beta)$ for some α and β;

H_1: $p(x)$ is a smooth function.

The likelihood under H_0 is evaluated at $p(.; \hat{\alpha}, \hat{\beta})$, making use of the fitted parameter values. The likelihood under the alternative is evaluated at $p(\,.\,) = \hat{p}(\,.\,)$. In the logistic regression case, the pseudo likelihood ratio statistic is then

$$\textstyle\sum_i \left[y_i \log \left\{ \frac{\hat{p}(x_i)}{p(x_i\,;\,\hat{\alpha}, \hat{\beta})} \right\} + (n_i - y_i) \log \left\{ \frac{1 - \hat{p}(x_i)}{1 - p(x_i\,;\,\hat{\alpha}\,\hat{\beta})} \right\} \right]. \qquad (2)$$

Now H_0 and H_1 are nested hypotheses. However, H_1 is not being fitted by maximum likelihood and this need not even be the case for H_0. The test statistic may therefore occasionally take negative values, although in nearly all cases it will be positive. With the usual interpretation of likelihood ratio statistics, the procedure may be viewed as constructing an estimate of the Kullback-Leibler distance between the two models. In this sense the test is consistent because as the sample size increases the normalized test statistics will converge to zero if H_0 is true and to some nonzero value if H_0 is false.

That the test statistic is derived from a likelihood ratio argument does not imply that its distribution is approximately chi-squared. We now examine the null hypothesis behaviour of the test statistic (2) by simulating data from the fitted model as described in § 2. Hastie & Tibshirani (1987) discuss similar statistics in the context of additive models and give an argument for an approximate number of degrees of freedom. In a forthcoming paper W. Härdle and E. Mammen show that the distribution of (2) is asymptotically normal.

The significance of the observed test statistic is calculated from its position in the ordered test statistics derived from the simulated data. With Finney's data the observed significance level is 0·4% with simulation size $N=500$. Notice that the smoothing parameter h and the parameter estimates $\hat\alpha$ and $\hat\beta$ are determined once from the original data and are subsequently kept fixed throughout the simulations of statistic (2). The effects of this on the significance level are discussed in § 5.

4. Poisson regression

The ideas of the previous section can be applied to the data of Bissell (1972) giving the lengths, x_i, of 32 pieces of cloth and the corresponding numbers of observed flaws, y_i. A natural model is a regression where y_i is assumed to have a Poisson distribution with mean βx_i. A resistant fit to the data, using the technique referred to in § 1, gives $\hat\beta = 0{\cdot}0143$. This is close to the maximum likelihood estimate, 0·0151.

Figure 3 displays the data and a nonparametric estimate of the regression line of the form (1) with $n_i = 1$, with a normal kernel, and with the smoothing parameter $h=100$ chosen by cross-validation. The pseudo likelihood ratio test statistic analogous to (2) is

$$\sum_{i=1}^{n}\left\{-\hat r(x_i)+\hat\beta x_i+y_i\log\frac{\hat r(x_i)}{\hat\beta x_i}\right\}, \tag{3}$$

where n denotes the sample size and the notation $\hat r(\,.\,)$ for the nonparametric regression reflects that we are no longer working on a probability scale. The observed significance level is 1·4%. This provides some evidence that the linear model is inadequate and simulation bands, $\varepsilon = 0{\cdot}05$, obtained as in § 3, can be used to help identify how the data differ from the model. These bands are also displayed on Fig. 3 which suggests an inadequacy of the model at high covariate values. The observed curve also strays outside the simulation envelope near $x=300$. This is caused by the presence of a very large observation at $x=371$, although a departure is not exhibited exactly at that point because of the balancing effect of a very small observation at $x=417$.

Bissell (1972) also reached the conclusion that the simple linear model is inadequate via different methods. An extended model was proposed where β has a gamma distribution, entailing larger variability in the data. This broader model can be assessed in a similar way to the Poisson regression. One effect is that the simulation bands are increased

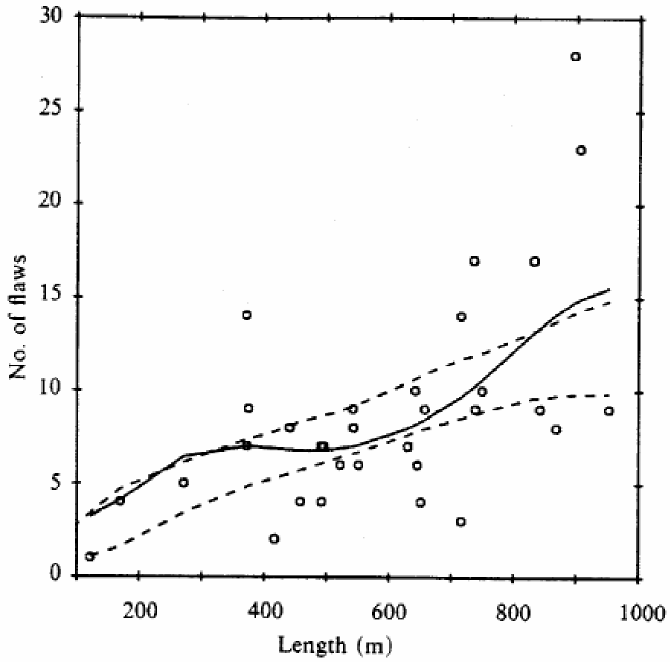


Fig. 3. Bissell's data with a nonparametric estimate of regression function, shown by solid line, and simulation bands, 95% level, derived from Poisson linear model, shown by broken lines.

in width and now contain the entire nonparametric curve. However, it is interesting that the discrepancies between the data and the linear model are predominantly negative between 400 and 600 and predominantly positive elsewhere.

5. Choice of smoothing and model parameters

In the above simulations, the smoothing parameter h of the nonparametric curve and the parameter estimates of the proposed models were determined once and for all from the original data. They were not recalculated on each new set of simulated data. The Poisson model offers a convenient framework within which the effects of this can be discussed, since the maximum likelihood estimate of the model parameter has the particularly simple form $\hat{\beta} = \Sigma\, y_i/\Sigma\, x_i$. For the remainder of this section, we therefore use this maximum likelihood estimator instead of the resistant version. As previously noted, there is little numerical difference between these with Bissell's data.

In the test statistic (3), the component corresponding to the linear model is based on a maximum likelihood fit to the data. If the distribution of this test statistic is to be accurately computed, then this source of variability should be incorporated into the simulations by re-estimating β on each of the simulated data sets. If this is done for Bissell's data, the significance of the observed test statistic drops from 1·4% to 0·2%. The reason is that, for each simulation, the likelihood component is maximized, and so the test statistic is minimized, by re-estimating β. The position of the observed test statistic is therefore made more extreme with respect to the simulated values. This effect holds in general, whenever the parameters of a proposed model are re-estimated on the simulated data. The strategy of estimating the model parameters once, and keeping them fixed throughout the subsequent simulations, is therefore conservative, and has the advantage of avoiding a large number of additional maximum likelihood calculations.

The choice of the smoothing parameter h in the nonparametric component of the test statistic (3) may also be viewed as playing the role of a fit to the observed data. There are two obvious ways in which this feature can be incorporated into the simulations. The first is to choose a new value of h by cross-validation on each simulated data set. This is not an attractive option since smoothing parameter choice is a rather imprecise operation and so a large amount of variability is added to the distribution of interest unless the sample size is large. An additional problem is the large amount of computational effort involved.

An alternative is to choose h in a way tailored to the proposed model rather than to the data. A natural approach would be to choose h to minimize the expected value of the pseudo likelihood ratio (3), where the expectation is taken under the Poisson linear model. However, a more tractable approach is to choose h to minimize

$$E\left[\sum_{i=1}^{n} \frac{\{y_i - \hat{r}(x_i)\}^2}{\beta x_i}\right].$$

This is a sum of standardized squared residuals, with expectation again taken under the Poisson linear model. This expression can be evaluated algebraically, and minimized numerically, using $\beta = \hat{\beta}$. This produces a smoothing parameter of 87 and an observed test statistic with an associated significance of 2%.

In summary, the parameters of the proposed model should be re-estimated on each simulated set of data if this is computationally feasible. Of the two possibilities for choice of smoothing parameter, the one tailored to the proposed model is more attractive since it avoids the computational effort and the extra variability incurred by cross-validation on each simulated set of data. There are, however, occasions when a model based choice of smoothing parameter is inappropriate. For example, a regression model with slope near zero leads to a very large smoothing parameter which obscures any nonlinearity in the data.

6. First-order autoregressive series

The parametric models discussed so far have been of generalized linear type, with discrete response variables. However, the basic ideas developed can be applied to a wider variety of models. As an illustration consider autoregressive time series of order 1. The model is

$$y_t = \rho y_{t-1} + \varepsilon_t, \qquad (4)$$

where $-1 < \rho < 1$ and $\{\varepsilon\}$ is a sequence of independent normal errors, with mean 0 and variance σ^2. There is now a scale parameter, which has not been the case in previous examples.

A simulated time series of size 200 is displayed in Fig. 4 as a plot of y_t against y_{t-1}, the form in which the linearity of the model (4) can be most easily examined. The slope parameter of the linear model is estimated by

$$\hat{\rho} = \sum_{t=2}^{n} y_t y_{t-1} \Big/ \sum_{t=1}^{n} y_t^2.$$

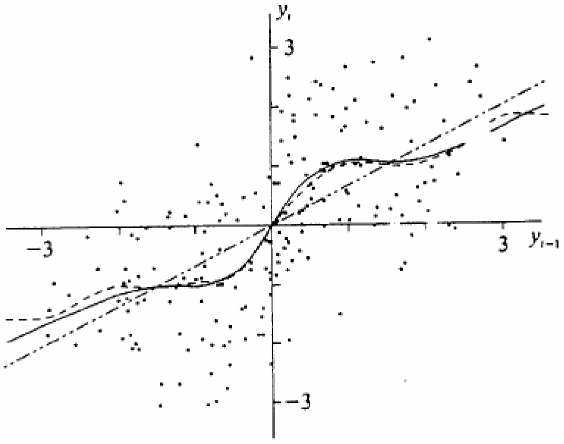


Fig. 4. Simulated time series data, y_t against y_{t-1}, with true regression function, shown by solid line, nonparametric estimate $\hat r$, dashed line, and fitted linear model, dotted and dashed line.

This form ensures that the estimate lies within the feasible region $(-1, 1)$. Our more general model assumes that $y_t = r(y_{t-1}) + \varepsilon_t$, where the function $r(\,.\,)$ is not necessarily linear. A nonparametric estimate of r is provided by smoothing the data of Fig. 4, namely

$$\hat r(y) = \sum_{t=2}^{n} y_t w\left(\frac{y-y_{t-1}}{h}\right) \Big/ \sum_{t=2}^{n} w\left(\frac{y-y_{t-1}}{h}\right).$$

For consistency and other asymptotic properties of a wide class of kernel estimates which includes $\hat r(y)$, see Robinson (1983). To take account of the bias introduced by smoothing, we adjust $\hat r$ so that the estimate is approximately unbiased when the linear model is correct. It is easy to show that

$$\hat r(y) + \hat\rho y \left\{1 + \frac{\hat\sigma^2}{h^2(1-\hat\rho^2)}\right\}^{-1}$$

is approximately unbiased. This bias-corrected curve is plotted in Fig. 4 and will be referred to by $\hat r$.

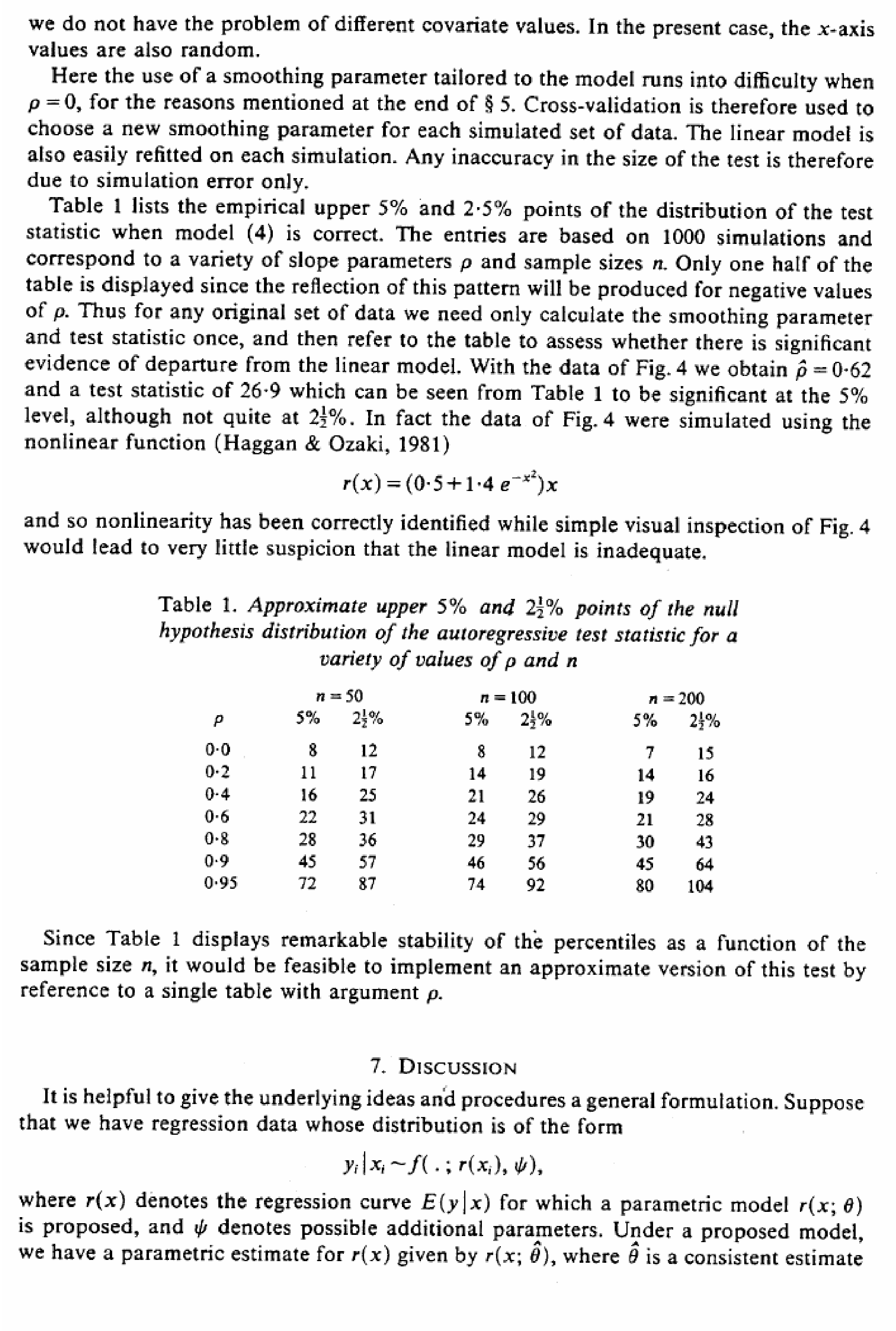
Cross-validation is used to select an appropriate smoothing parameter by minimizing

$$\sum_{t=2}^{n} \{y_t - \hat r_{-t}(y_{t-1})\}^2,$$

where $\hat r_{-t}$ denotes the estimator with the point (y_t, y_{t-1}) omitted. This yields the value 0·3 for the data of Fig. 4. If the linear model is correct, the contribution to the likelihood from the first observation is asymptotically negligible; ignoring this term, the pseudo likelihood ratio statistic is equivalent to

$$\frac{\sum\{y_t-\hat\rho y_{t-1}\}^2 - \sum\{y_t - \hat r(y_{t-1})\}^2}{\sum\{y_t-\hat r(y_{t-1})\}^2/n}. \qquad (5)$$

It is not hard to show that both $\hat\rho$ and the test statistic (5) are independent of the nuisance parameter σ^2. The particular structure of the present problem enables us to simulate the null hypothesis distribution once, and to refer to this for any data set of interest, because



we do not have the problem of different covariate values. In the present case, the x-axis values are also random.

Here the use of a smoothing parameter tailored to the model runs into difficulty when $\rho=0$, for the reasons mentioned at the end of § 5. Cross-validation is therefore used to choose a new smoothing parameter for each simulated set of data. The linear model is also easily refitted on each simulation. Any inaccuracy in the size of the test is therefore due to simulation error only.

Table 1 lists the empirical upper 5% and 2·5% points of the distribution of the test statistic when model (4) is correct. The entries are based on 1000 simulations and correspond to a variety of slope parameters ρ and sample sizes n. Only one half of the table is displayed since the reflection of this pattern will be produced for negative values of ρ. Thus for any original set of data we need only calculate the smoothing parameter and test statistic once, and then refer to the table to assess whether there is significant evidence of departure from the linear model. With the data of Fig. 4 we obtain $\hat{\rho}=0{\cdot}62$ and a test statistic of 26·9 which can be seen from Table 1 to be significant at the 5% level, although not quite at $2\frac{1}{2}$%. In fact the data of Fig. 4 were simulated using the nonlinear function (Haggan & Ozaki, 1981)

$$r(x)=(0{\cdot}5+1{\cdot}4\,e^{-x^2})x$$

and so nonlinearity has been correctly identified while simple visual inspection of Fig. 4 would lead to very little suspicion that the linear model is inadequate.

Table 1. *Approximate upper 5% and $2\frac{1}{2}$% points of the null hypothesis distribution of the autoregressive test statistic for a variety of values of ρ and n*

	$n=50$		$n=100$		$n=200$	
ρ	5%	$2\frac{1}{2}$%	5%	$2\frac{1}{2}$%	5%	$2\frac{1}{2}$%
0·0	8	12	8	12	7	15
0·2	11	17	14	19	14	16
0·4	16	25	21	26	19	24
0·6	22	31	24	29	21	28
0·8	28	36	29	37	30	43
0·9	45	57	46	56	45	64
0·95	72	87	74	92	80	104

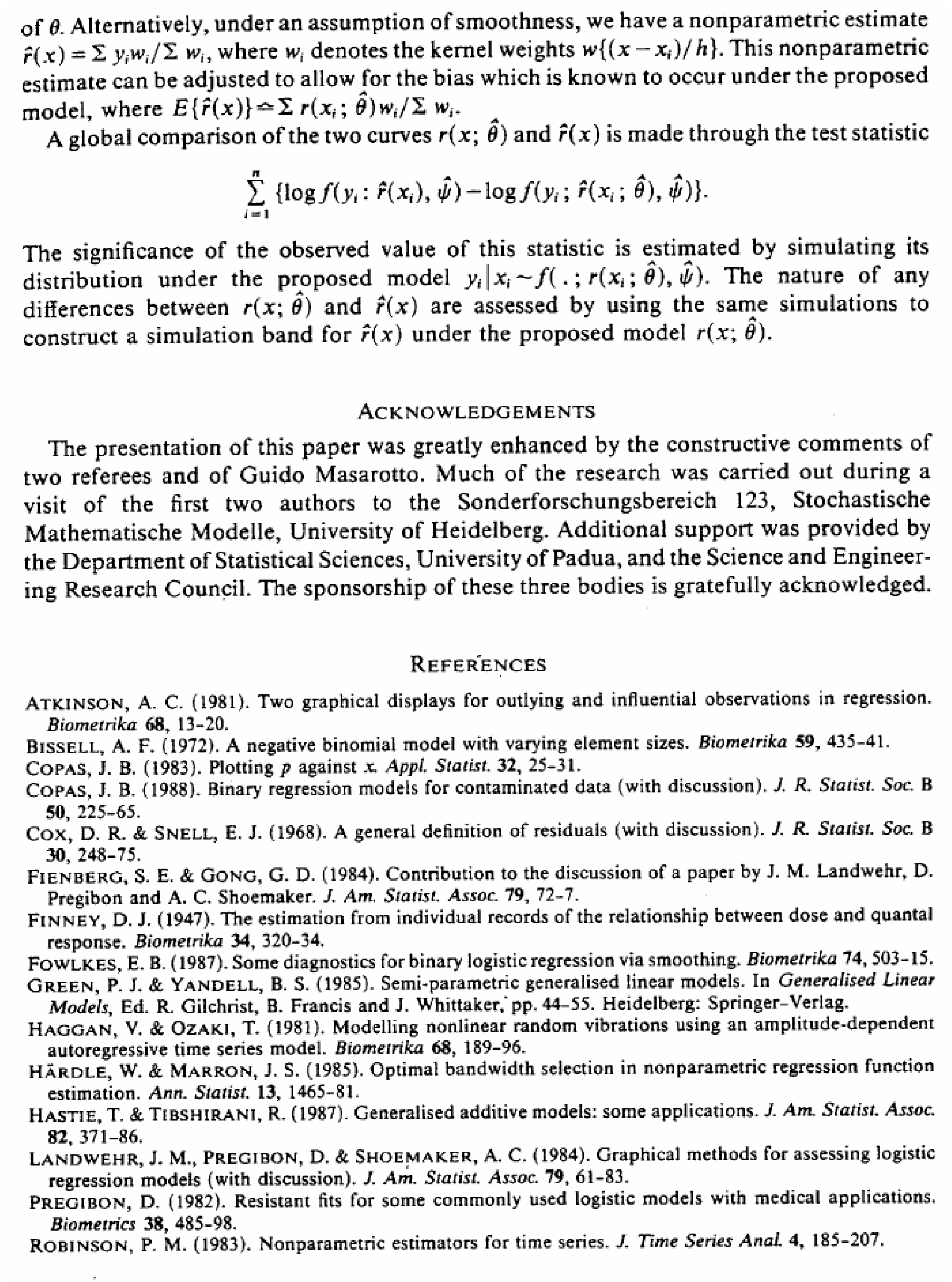
Since Table 1 displays remarkable stability of the percentiles as a function of the sample size n, it would be feasible to implement an approximate version of this test by reference to a single table with argument ρ.

7. Discussion

It is helpful to give the underlying ideas and procedures a general formulation. Suppose that we have regression data whose distribution is of the form

$$y_i\,|\,x_i \sim f(\,.\,;\,r(x_i),\,\psi),$$

where $r(x)$ denotes the regression curve $E(y\,|\,x)$ for which a parametric model $r(x;\,\theta)$ is proposed, and ψ denotes possible additional parameters. Under a proposed model, we have a parametric estimate for $r(x)$ given by $r(x;\,\hat{\theta})$, where $\hat{\theta}$ is a consistent estimate



of θ. Alternatively, under an assumption of smoothness, we have a nonparametric estimate $\hat{r}(x)=\Sigma\, y_i w_i/\Sigma\, w_i$, where w_i denotes the kernel weights $w\{(x-x_i)/h\}$. This nonparametric estimate can be adjusted to allow for the bias which is known to occur under the proposed model, where $E\{\hat{r}(x)\}\simeq \Sigma\, r(x_i;\,\hat{\theta})w_i/\Sigma\, w_i$.

A global comparison of the two curves $r(x;\,\hat{\theta})$ and $\hat{r}(x)$ is made through the test statistic

$$\sum_{i=1}^{n}\{\log f(y_i : \hat{r}(x_i),\, \hat{\psi})-\log f(y_i;\, \hat{r}(x_i;\, \hat{\theta}),\, \hat{\psi})\}.$$

The significance of the observed value of this statistic is estimated by simulating its distribution under the proposed model $y_i|x_i \sim f(\,.\,;\, r(x_i;\,\hat{\theta}),\, \hat{\psi})$. The nature of any differences between $r(x;\,\hat{\theta})$ and $\hat{r}(x)$ are assessed by using the same simulations to construct a simulation band for $\hat{r}(x)$ under the proposed model $r(x;\,\hat{\theta})$.

Acknowledgements

The presentation of this paper was greatly enhanced by the constructive comments of two referees and of Guido Masarotto. Much of the research was carried out during a visit of the first two authors to the Sonderforschungsbereich 123, Stochastische Mathematische Modelle, University of Heidelberg. Additional support was provided by the Department of Statistical Sciences, University of Padua, and the Science and Engineering Research Council. The sponsorship of these three bodies is gratefully acknowledged.

References

Atkinson, A. C. (1981). Two graphical displays for outlying and influential observations in regression. *Biometrika* **68**, 13–20.

Bissell, A. F. (1972). A negative binomial model with varying element sizes. *Biometrika* **59**, 435–41.

Copas, J. B. (1983). Plotting p against x. *Appl. Statist.* **32**, 25–31.

Copas, J. B. (1988). Binary regression models for contaminated data (with discussion). *J. R. Statist. Soc.* B **50**, 225–65.

Cox, D. R. & Snell, E. J. (1968). A general definition of residuals (with discussion). *J. R. Statist. Soc.* B **30**, 248–75.

Fienberg, S. E. & Gong, G. D. (1984). Contribution to the discussion of a paper by J. M. Landwehr, D. Pregibon and A. C. Shoemaker. *J. Am. Statist. Assoc.* **79**, 72–7.

Finney, D. J. (1947). The estimation from individual records of the relationship between dose and quantal response. *Biometrika* **34**, 320–34.

Fowlkes, E. B. (1987). Some diagnostics for binary logistic regression via smoothing. *Biometrika* **74**, 503–15.

Green, P. J. & Yandell, B. S. (1985). Semi-parametric generalised linear models. In *Generalised Linear Models*, Ed. R. Gilchrist, B. Francis and J. Whittaker, pp. 44–55. Heidelberg: Springer–Verlag.

Haggan, V. & Ozaki, T. (1981). Modelling nonlinear random vibrations using an amplitude-dependent autoregressive time series model. *Biometrika* **68**, 189–96.

Härdle, W. & Marron, J. S. (1985). Optimal bandwidth selection in nonparametric regression function estimation. *Ann. Statist.* **13**, 1465–81.

Hastie, T. & Tibshirani, R. (1987). Generalised additive models: some applications. *J. Am. Statist. Assoc.* **82**, 371–86.

Landwehr, J. M., Pregibon, D. & Shoemaker, A. C. (1984). Graphical methods for assessing logistic regression models (with discussion). *J. Am. Statist. Assoc.* **79**, 61–83.

Pregibon, D. (1982). Resistant fits for some commonly used logistic models with medical applications. *Biometrics* **38**, 485–98.

Robinson, P. M. (1983). Nonparametric estimators for time series. *J. Time Series Anal.* **4**, 185–207.

Bandwidth Choice for Density Derivatives

By WOLFGANG HÄRDLE, J. S. MARRON† and M. P. WAND

Universität Bonn, FRG

University of North Carolina at Chapel Hill, USA

Australian National University, Canberra, Australia

[Received February 1988. Final revision June 1989]

SUMMARY

An adaptation of least squares cross-validation is proposed for bandwidth choice in the kernel estimation of the derivatives of a probability density. The practicality of the method is demonstrated by an example and a simulation study. Theoretical justification is provided by an asymptotic optimality result.

Keywords: CROSS-VALIDATION DENSITY DERIVATIVES; KERNEL ESTIMATORS

1. INTRODUCTION

The kernel approach provides an attractive method for estimation of both probability densities and their derivatives. Such estimators have been successfully used in the exploration and presentation of data; see, for example, Silverman (1986). Density derivatives are of particular interest for the evaluation of modes and inflection points. They are also of theoretical importance, as they occur both directly and indirectly in asymptotic expansions of error criteria for density estimation. In addition, density derivatives are of practical importance for estimating scores in certain additive models; see Härdle and Stoker (1990). Another application is to the empirical verification of uniqueness of equilibria of market demand, where the estimation of derivatives of densities enters through so-called income effects; see Hildenbrand and Hildenbrand (1986).

As with any type of smoothing method, the performance of kernel estimators is heavily dependent on the choice of smoothing parameter. If the effective amount of local averaging is too small, the resulting curve estimate is subject to too much sample variability, which appears in the form of a curve which is too wiggly. In contrast, too much local averaging results in the introduction of an unacceptably large bias, in the sense that features of the true curve will be smoothed away.

A practical approach to the problem of smoothing parameter selection for density estimation is provided by least squares cross-validation, which was proposed by Rudemo (1982) and Bowman (1984). Strong theoretical justification has been provided by several asymptotic optimality results which demonstrate that the selected smoothing parameter is, in the limit, effectively the same as the squared error optimal choice; see Hall (1983), Stone (1984) and Burman (1985).

In this paper the cross-validation idea is extended to the estimation of density derivatives. The extension is motivated and made precise in Section 2. The practical

†*Address for correspondence*: Department of Statistics, University of North Carolina at Chapel Hill, CB 3260, Phillips Hall, Chapel Hill, NC 27599-3260, USA.

0035-9246/90/52223 $2.00

effectiveness of this method is demonstrated through an example and a simulation study in Section 3.

Section 4 provides theoretical underpinning for the cross-validation method. In particular, two types of asymptotic optimality results are established. It is shown that the cross-validated smoothing parameter is asymptotically the same as the squared error optimal choice, and also that the squared error performance is effectively the same. Section 5 contains the proofs of the theoretical results.

2. CROSS-VALIDATION FOR DENSITY DERIVATIVES

We consider here the estimation of the kth derivative $f^{(k)}(x)$ of a probability density $f(x)$ from a random sample $X_1, \ldots, X_n$. A kernel estimator of $f^{(k)}(x)$, motivated by taking the kth derivative of the kernel estimate of f, is given by

$$\hat{f}_h^{(k)}(x) = n^{-1} \sum_{i=1}^{n} h^{-k-1} K^{(k)}\{(x - X_i)/h\},$$

where h is called the bandwidth or smoothing parameter and K is the kernel function which is assumed to be a symmetric probability density. Gasser *et al.* (1985) have developed an interesting asymptotic theory for the optimal choice of K which shows that the best choice of the function $K^{(k)}$ is not necessarily the kth derivative of the optimal kernel for estimating f. However, the present form is used here because we prefer its intuitive content and are concerned about numerical instabilities (see Section 3 for more details on this, as well as a strong reason for not using the normal kernel, especially for large k). As already noted, the choice of the amount of smoothing, quantified here by h, is crucial to the performance of $\hat{f}_h^{(k)}(x)$.

The essential idea of least squares cross-validation, for the estimation of f (the special case $k = 0$ here), is to use the bandwidth which minimizes the function

$$\mathrm{CV}(h) = \int \hat{f}_h(x)^2 \, \mathrm{d}x - 2n^{-1} \sum_{i=1}^{n} \hat{f}_{h,i}(X_i),$$

where $\hat{f}_{h,i}$ denotes the leave-one-out kernel estimator (defined for general k later). This method of bandwidth selection can be motivated by observing that the function $\mathrm{CV}(h)$ provides a reasonable, and indeed unbiased, estimate of the first two terms in the expansion of the integrated square error,

$$\begin{aligned} d_{\mathrm{I}}(\hat{f}_h, f) &= \int \{\hat{f}_h(x) - f(x)\}^2 \, \mathrm{d}x \\ &= \int \hat{f}_h^2 - 2\int \hat{f}_h f + \int f^2. \end{aligned}$$

So, since the third term is independent of h, the minimizer of $\mathrm{CV}(h)$ may be expected to be reasonably close to the minimizer of d_{I}.

This idea can be extended to the estimation of derivatives of the density by observing that

$$\begin{aligned} d_{\mathrm{I}}(\hat{f}_h^{(k)}, f^{(k)}) &= \int \{\hat{f}_h^{(k)}(x) - f^{(k)}(x)\}^2 \, \mathrm{d}x \\ &= \int \hat{f}_h^{(k)2} - 2\int \hat{f}_h^{(k)} f^{(k)} + \int f^{(k)2}. \end{aligned}$$

As before, the last term is independent of h, so it has no effect on the location of the minimizer, and the first term is available to the experimenter. By integration by parts, the second term may be estimated by the second term of the cross-validation function,

$$\mathrm{CV}_k(h) = \int \hat{f}_h^{(k)}(x)^2 \,\mathrm{d}x - 2n^{-1}(-1)^k \sum_{i=1}^{n} \hat{f}_{h,i}^{(2k)}(X_i),$$

where

$$\hat{f}_{h,i}^{(2k)}(x) = (n-1)^{-1} \sum_{j \neq i} h^{-2k-1} K^{(2k)}\{(x - X_j)/h\}.$$

Hence the bandwidth $\hat{h}_k$, which minimizes CV_k, should be close to h_k^*, the minimizer of $d_1(\hat{f}_h^{(k)}, f^{(k)})$.

$\mathrm{CV}_k(h)$ can be simplified to give the computationally more straightforward version

$$\mathrm{CV}_k(h) = (-1)^k n^{-1} h^{-2k-1} \left[n^{-1} \sum_i \sum_j (K*K)^{(2k)}\{(X_i - X_j)/h\} \right.$$
$$\left. - 2(n-1)^{-1} \sum_{i \neq j}\sum K^{(2k)}\{(X_i - X_j)/h\} \right],$$

where here and throughout an asterisk denotes convolution. This can either be used directly, or easily adapted to give an efficient fast Fourier transform approximation, as described in Section 3.5 of Silverman (1986). For some kernels, the fact that $(K*K)^{(2k)} = K^{(k)}*K^{(k)}$ can also be useful.

3. EXAMPLE AND SIMULATIONS

We tested the bandwidth selection method described in Section 2 on several data sets for estimation of $f^{(1)}(x)$, the first derivative of $f(x)$. We had the best success when the kernel was standard normal. Piecewise polynomial kernels, with asymptotic optimality properties of the type described in Gasser *et al.* (1985), sometimes gave a numerically unstable derivative cross-validation function, especially for the smaller data sets. This seems to be caused by the fact that CV_k makes use of the 2*k*th derivative of K, which for $k = 1$ is discontinuous for some popular kernels. One approach to this problem would be to use piecewise polynomials that have an optimality property under smoothness constraints, as developed in Müller (1984), although we have not tried this.

The normal kernel is attractive also for larger data sets, as the function $\mathrm{CV}_k(h)$ may be efficiently calculated by a fast Fourier transform algorithm as mentioned at the end of Section 2. To see how much loss in efficiency could be expected from using the normal kernel, we calculated an analog of Table 3.1 of Silverman (1986). The analog of Silverman's $C(K)$ (although see Marron and Nolan (1989) for a more convincing derivation) for estimating the *k*th derivative is

$$C_k(K) = \{\textstyle\int (K^{(k)})^2\}^{4/(5+2k)} (\textstyle\int x^{2+k} K^{(k)})^{(4k+2)/(5+2k)}.$$

Table 1 shows the efficiency (in the sense of Silverman) of the normal kernel, with respect to some of the optimal kernels of Müller (1984) (indexed by the amount of 'smoothness' μ in Müller's notation), defined as

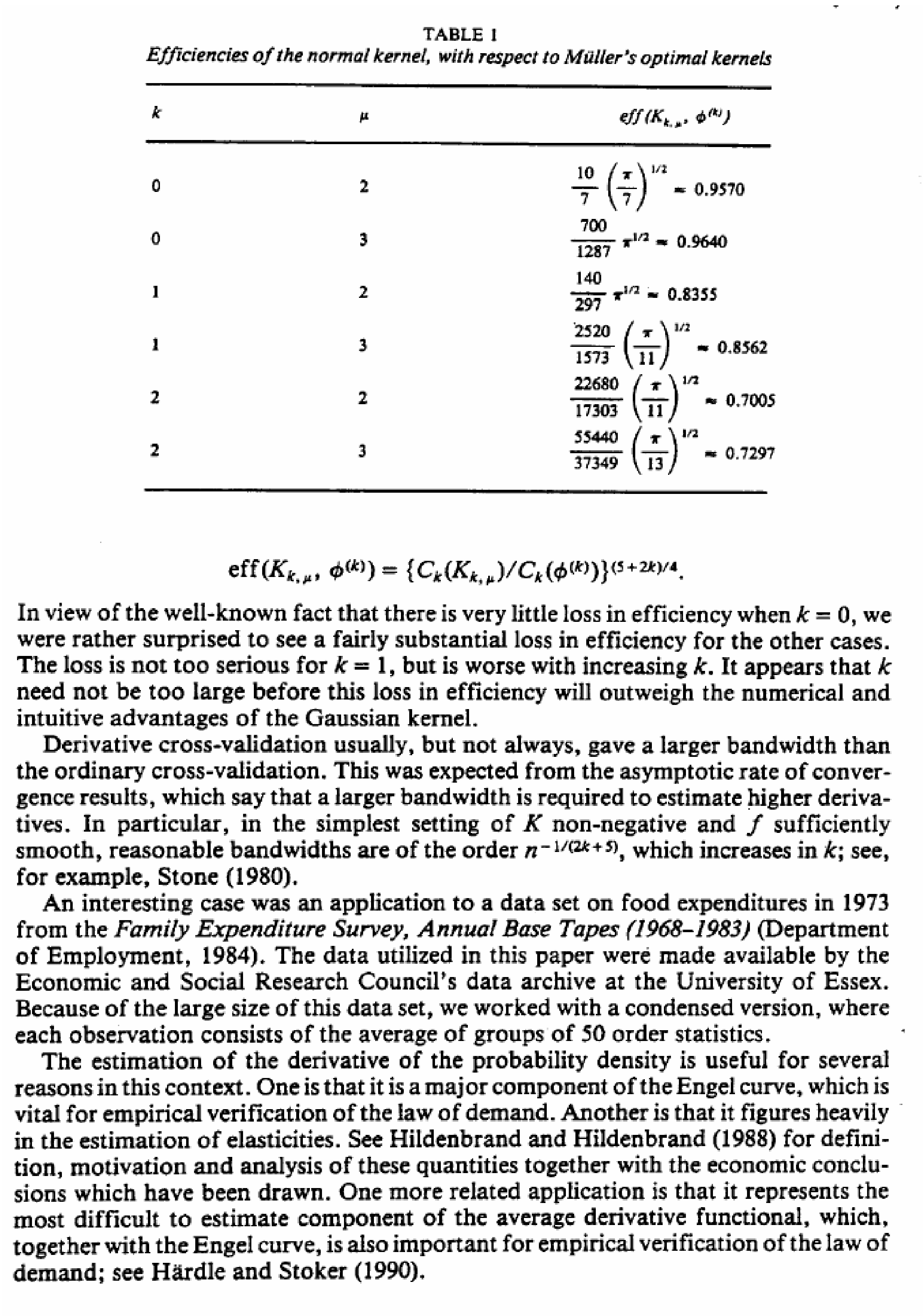


TABLE 1
Efficiencies of the normal kernel, with respect to Müller's optimal kernels

k	μ	$eff(K_{k,\mu}, \phi^{(k)})$
0	2	$\frac{10}{7}\left(\frac{\pi}{7}\right)^{1/2} \approx 0.9570$
0	3	$\frac{700}{1287}\pi^{1/2} \approx 0.9640$
1	2	$\frac{140}{297}\pi^{1/2} \approx 0.8355$
1	3	$\frac{2520}{1573}\left(\frac{\pi}{11}\right)^{1/2} \approx 0.8562$
2	2	$\frac{22680}{17303}\left(\frac{\pi}{11}\right)^{1/2} \approx 0.7005$
2	3	$\frac{55440}{37349}\left(\frac{\pi}{13}\right)^{1/2} \approx 0.7297$

$$\mathrm{eff}(K_{k,\mu}, \phi^{(k)}) = \{C_k(K_{k,\mu})/C_k(\phi^{(k)})\}^{(5+2k)/4}.$$

In view of the well-known fact that there is very little loss in efficiency when $k = 0$, we were rather surprised to see a fairly substantial loss in efficiency for the other cases. The loss is not too serious for $k = 1$, but is worse with increasing k. It appears that k need not be too large before this loss in efficiency will outweigh the numerical and intuitive advantages of the Gaussian kernel.

Derivative cross-validation usually, but not always, gave a larger bandwidth than the ordinary cross-validation. This was expected from the asymptotic rate of convergence results, which say that a larger bandwidth is required to estimate higher derivatives. In particular, in the simplest setting of K non-negative and f sufficiently smooth, reasonable bandwidths are of the order $n^{-1/(2k+5)}$, which increases in k; see, for example, Stone (1980).

An interesting case was an application to a data set on food expenditures in 1973 from the *Family Expenditure Survey, Annual Base Tapes (1968–1983)* (Department of Employment, 1984). The data utilized in this paper were made available by the Economic and Social Research Council's data archive at the University of Essex. Because of the large size of this data set, we worked with a condensed version, where each observation consists of the average of groups of 50 order statistics.

The estimation of the derivative of the probability density is useful for several reasons in this context. One is that it is a major component of the Engel curve, which is vital for empirical verification of the law of demand. Another is that it figures heavily in the estimation of elasticities. See Hildenbrand and Hildenbrand (1988) for definition, motivation and analysis of these quantities together with the economic conclusions which have been drawn. One more related application is that it represents the most difficult to estimate component of the average derivative functional, which, together with the Engel curve, is also important for empirical verification of the law of demand; see Härdle and Stoker (1990).

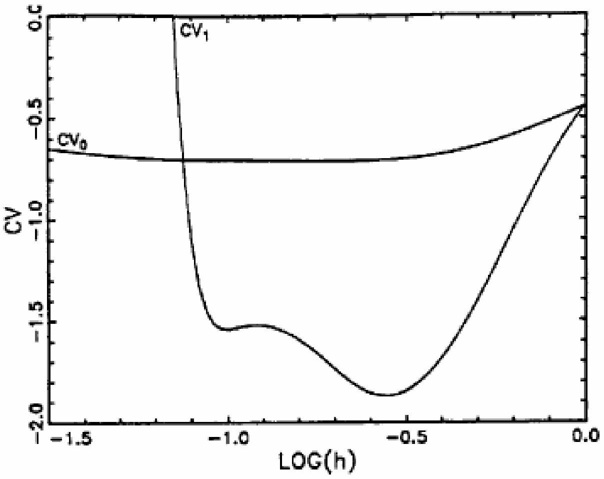


Fig. 1. Cross-validation functions CV_1 and CV_0 for food expenditure data, K standard normal

Fig. 1 shows a superimposition of the cross-validation functions $\mathrm{CV}_0(h)$ and $\mathrm{CV}_1(h)$. A feature that was typical of all the examples that we have considered is that the minimum is much better defined for the derivative. This appears related to the fact that bandwidths chosen by cross-validation have better stability properties in settings where curve estimation is more difficult, such as higher dimensional estimation. See Section 4 of Marron (1986) for a discussion of this seeming paradox.

An interesting feature of this data set, that we did not observe for any other, is that if we extend the range of hs, for which minimization is performed, to include some very small values then the function $\mathrm{CV}_1(h)$ has its global minimum at an unreasonably small value. This is not a practical problem for this data set, because the bandwidths in the extended range represent amounts of smoothing which give a far too wiggly curve to be seriously considered. However, it is worth noting because similar phenomena have been observed for ordinary density cross-validation; see Rudemo (1982) and Scott and Terrell (1987).

The bandwidth selection rule was also applied to some simulated data. 15 samples of size 750 of data having the extreme value density

$$f(x) = \mathrm{e}^x \mathrm{e}^{-\mathrm{e}^x}$$

were generated to assess the performance of $\hat{h}_1$ in the estimation of

$$f^{(1)}(x) = \mathrm{e}^x \mathrm{e}^{-\mathrm{e}^x}(1 - \mathrm{e}^x).$$

The selected bandwidths are listed in Table 2. With only 15 data sets, the results are far from conclusive, but they do give some insight.

Most of them are in reasonably close agreement with the bandwidth which minimizes the mean integrated square error, which was roughly 0.34 in this case. To give an idea about the performance of the resulting curve estimates, we chose the sample which gave the median value of $d_I(\hat{f}^{(1)}_{\hat{h}_1}, f^{(1)})$ among our 15 replications. Fig. 2 shows the resulting curve estimate $\hat{f}^{(1)}_{\hat{h}_1}$ as a broken line and the true underlying curve $f^{(1)}$ as the full curve.

TABLE 2
Values of $\hat{h}_I$ for 15 samples from the extreme value density with $n = 750$ using the standard normal kernel

0.16	0.44	0.20	0.45	0.33
0.44	0.30	0.46	0.43	0.28
0.18	0.34	0.34	0.47	0.16

4. THEORETICAL RESULTS

For ordinary density estimation, i.e. for $k = 0$, the effective asymptotic performance of the cross-validated bandwidth has been established by the optimality results of Hall (1983), Stone (1984) and Burman (1985). In this section, it is seen how these results may be extended to general k.

Assume that K is a compactly supported probability density with $2k$ bounded derivatives and that f has $2k+2$ continuous bounded derivatives. The assumption of compact support of K does not include the Gaussian kernel used in Section 3. The results proven here can be extended to this case by a straightforward truncation argument. This is not explicitly done because the increased technical complexity of the proof only detracts from the main points.

The bandwidths under consideration are assumed to come, for each n, from a set H_n so that $\sup_{h \in H_n} h \leqslant n^{-\delta}$, $\inf_{h \in H_n} h \geqslant n^{(-1+\delta)/(k+1)}$ and $\mathrm{card}(H_n) \leqslant n^{\rho}$ for some constants $\delta > 0$ and $\rho > 0$.

The cross-validated bandwidth $\hat{h}_k$ is asymptotically the same as the optimal bandwidth h_k^* (both chosen as minimizers over the set H_n) in the following sense.

Theorem 1. Under the above assumptions, as $n \to \infty$,

$$\hat{h}_k / h_k^* \to 1, \qquad \text{almost surely.}$$

The fact that this result means that the cross-validated bandwidth is useful for estimation is demonstrated by the following theorem.

Theorem 2. Under the above assumptions, as $n \to \infty$,

$$d_I(\hat{f}^{(k)}_{\hat{h}_k}, f^{(k)}) / d_I(\hat{f}^{(k)}_{h_k^*}, f^{(k)}) \to 1, \qquad \text{almost surely.}$$

Remark 1. Since d_I is random, i.e. changes for different data sets, one may prefer as an error criterion its expected value, the mean integrated square error,

$$d_M(\hat{f}^{(k)}_h, f^{(k)}) = E d_I(\hat{f}^{(k)}_h, f^{(k)}).$$

The proof of theorems 1 and 2 may be adapted in a straightforward fashion to give the d_M analogues of those results:

$$\hat{h}_k / h_k^{\cdot} \to 1, \qquad \text{almost surely,}$$

$$d_M(\hat{f}^{(k)}_{\hat{h}_k}, f^{(k)}) / d_M(\hat{f}^{(k)}_{h_k^{\cdot}}, f^{(k)}) \to 1, \qquad \text{almost surely,}$$

where $h_k^{\cdot}$ denotes the minimizer of $d_M(\hat{f}^{(k)}_h, f^{(k)})$. However, no statement is made here about $E d_I(\hat{f}^{(k)}_{\hat{h}_k}, f^{(k)})$.

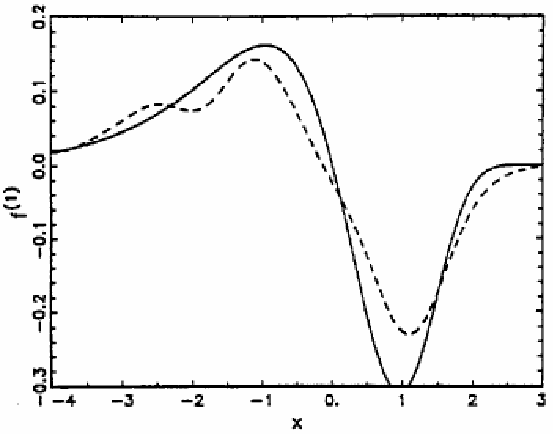


Fig. 2. Simulation target curve (——) and curve estimate (---) for the data set giving median performance, from the extreme value distribution, using the standard normal kernel

Remark 2. The assumption that f has $2k+2$ derivatives is substantially stronger than that made, for $k=0$, by most of the researchers cited at the beginning of this section. With more technical effort (and messy notation) than seems justified to us, this assumption can be weakened somewhat, but observe that the assumption of at least $2k$ derivatives appears to be essential for the establishment of asymptotic optimality for this method of cross-validation.

5. PROOFS

Theorems 1 and 2 are a consequence of the following two lemmas.

Lemma 1.

$$\sup_{h\in H_n}|\{d_{\mathrm{I}}(\hat{f}_h^{(k)}, f^{(k)})-A(h)\}/A(h)|\rightarrow 0, \qquad \text{almost surely,}$$

where

$$A(h)=\int(K^{(k)})^2 n^{-1}h^{-(2k+1)}+(\int u^2K/2)^2\int(f^{(k+2)})^2h^4.$$

Lemma 2.

$$\sup_{h,h'\in H_n}|B(h, h')|\rightarrow 0, \qquad \text{almost surely,}$$

where

$$B(h, h')=[\mathrm{CV}(h)-d_{\mathrm{I}}(\hat{f}_h^{(k)}, f^{(k)})-\{\mathrm{CV}(h')-d_{\mathrm{I}}(\hat{f}_{h'}^{(k)}, f^{(k)})\}]/\{A(h)+A(h')\}.$$

Calculations of the type leading to equation (3.20) of Silverman (1986), for example, show that $A(h)$ asymptotically approximates $d_{\mathrm{M}}(\hat{f}_h^{(k)}, f^{(k)})$ in the sense that

$$\sup_{h\in H_n}|\{d_{\mathrm{M}}(\hat{f}_h^{(k)}, f^{(k)})-A(h)\}/A(h)|\rightarrow 0.$$

This can be used to verify the claims made in remark 1.

The proof of lemma 1 follows very closely the proof of theorem 1 of Marron and

Härdle (1986). To see how to adapt the current set-up to the notation of that paper, define

$$g(x) := h^k f^{(k)}(x),$$
$$\hat{g}(x) := h^k \hat{f}_h^{(k)}(x),$$
$$\lambda := h^{-1},$$
$$\delta_\lambda(x, y) := h^{-1} K^{(k)}(x-y)/h,$$
$$w(x)\ \mathrm{d}F(x) := \mathrm{d}x.$$

The results of Marron and Härdle (1986) cannot be directly applied here because the target function $g(x)$ in that paper is not allowed to depend on h. However, an inspection of the proof in that paper shows that the result still holds, even in the current slightly more general context. This completes the proof of lemma 1.

Lemma 2 is a consequence of the following lemma.

Lemma 3.

$$\sup_{h \in H_n} \left| \left\{ n^{-1} \sum_{i=1}^{n} \hat{f}_{h,i}^{(2k)}(X_i) - \int \hat{f}^{(2k)} f - R \right\} \Big/ A(h) \right| \to 0, \qquad \text{almost surely,}$$

where

$$R = n^{-1} \sum_{i=1}^{n} f^{(2k)}(X_i) - \int f^{(2k)} f.$$

To prove lemma 3, define

$$U_{i,j} := h^{-2k-1} K^{(2k)}\{(X_i - X_j)/h\} - h^{-2k-1} \int K^{(2k)}\{(x - X_j)/h\}\ f(x)\ \mathrm{d}x$$
$$- f^{(2k)}(X_i) + \int f^{(2k)}(x)\ f(x)\ \mathrm{d}x.$$
$$V_i := E(U_{i,j} \,|\, X_i),$$
$$W_{i,j} := U_{i,j} - V_i.$$

To finish the proof it is sufficient to show that

$$\sup_{h \in H_n} \left| n^{-1} \sum_{i=1}^{n} V_i \Big/ A(h) \right| \to 0, \qquad \text{almost surely,} \tag{1}$$

and that

$$\sup_{h \in H_n} \left| n^{-2} \sum_{i \neq j} \sum W_{i,j} \Big/ A(h) \right| \to 0, \qquad \text{almost surely.} \tag{2}$$

To verify expression (1), by the Borel–Cantelli lemma it is sufficient to show that for $\epsilon > 0$

$$\sum_{n=1}^{\infty} \mathrm{card}(H_n)\ \sup_{h \in H_n} P \left\{ \left| n^{-1} \sum_{i=1}^{n} V_i \right| > \epsilon A(h) \right\} < \infty.$$

Hence by the Chebyshev inequality, it is sufficient to show that there is a constant $\gamma > 0$, so that for $m = 1, 2, \ldots$ there are constants C_m such that

$$\sup_{h \in H_n} E\left\{ n^{-1} \sum_{i=1}^{n} V_i \Big/ A(h) \right\}^{2m} \leqslant C_m n^{-\gamma m}. \tag{3}$$

To establish inequality (3), observe that $\{\Sigma_{i=1}^{n} V_i\}$ is a martingale with respect to the sequence of sigma fields generated by $\{X_1, \ldots, X_n\}$. An application of equation (21.5) of Burkholder (1973) (which is essentially Rosenthal's inequality), with $\Phi(x) = x^{2m}$, to the finitely (from $1, \ldots, n$) indexed martingale, gives

$$E\left(\sum_{i=1}^{n} V_i \right)^{2m} \leqslant C(n^m h^{4m} + n),$$

for some constant C. Inequality (3) follows from this and the definition of $A(h)$. This completes the proof of inequality (3), and hence also that of expression (1).

To verify expression (2), by a development similar to that leading to inequality (3), it is sufficient to show that

$$\sup_{h \in H_n} E\left\{ n^{-2} \sum_{i>j}\sum W_{i,j} \Big/ A(h) \right\}^{2m} \leqslant C_m n^{-\gamma m}. \tag{4}$$

Since

$$E(W_{i,j} \mid X_i) = E(W_{i,j} \mid X_j) = 0,$$

$\{\Sigma\Sigma_{i>j} W_{i,j}\}$ is a martingale with respect to the same sequence of sigma fields as before. Applying the same inequality to this finitely indexed martingale gives

$$E\left(\sum_{i>j}\sum W_{i,j} \right)^{2m} \leqslant C(n^{2m} h^{-(2k+1)m} + n^{m+1} h^{-(2k+1)2m}),$$

for another constant C. A consequence of this is inequality (4). This completes the proof of expression (2) and hence that of lemmas 2 and 3.

ACKNOWLEDGEMENTS

The research of the first two authors was supported by the Deutsche Forschungsgemeinschaft, SFB 303; that of the second author was also partially supported by the National Science Foundation.

REFERENCES

Bowman, A. (1984) An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, **71**, 353–360.

Burkholder, D. L. (1973) Distribution function inequalities for martingales. *Ann. Probab.*, **1**, 19–42.

Burman, P. (1985) A data dependent approach to density estimation. *Z. Wahrsch. Ver. Geb.*, **69**, 609–628.

Department of Employment (1984) *Family Expenditure Survey, Annual Base Tapes (1968–1983)*. London: Her Majesty's Stationery Office.

Gasser, T., Müller, H.-G. and Mammitzsch, V. (1985) Kernels for nonparametric curve estimation. *J. R. Statist. Soc.* B, **47**, 238–252.

Hall, P. (1983) Large sample optimality of least square cross-validation in density estimation. *Ann. Statist.*, **11**, 1156–1174.

Härdle, W. and Stoker, T. (1990) Investigating smooth multiple regression by the method of average derivatives. *J. Am. Statist. Ass.*, to be published.

Hildenbrand, K. and Hildenbrand, W. (1986) On the mean income effect: a data analysis of the U.K. family expenditure survey. In *Contributions to Mathematical Economics, in Honor of Gerard Debreu* (eds W. Hildenbrand and A. Mas-Colell), pp. 247–268. Amsterdam: North-Holland.

Marron, J. S. (1986) Will the art of smoothing ever become a science? *Contemp. Math.*, **9**, 169–178.

Marron, J. S. and Härdle, W. (1986) Random approximations to some measures of accuracy in nonparametric curve estimation. *J. Multiv. Anal.*, **20**, 91–113.

Marron, J. S. and Nolan, D. (1989) Canonical kernels for density estimation. *Statist. Probab. Lett.*, **7**, 195–199.

Müller, H. G. (1984) Smooth optimum kernel estimators of densities, regression curves and modes. *Ann. Statist.*, **12**, 766–774.

Rudemo, M. (1982) Empirical choice of histograms and kernel density estimators. *Scand. J. Statist.*, **9**, 65–78.

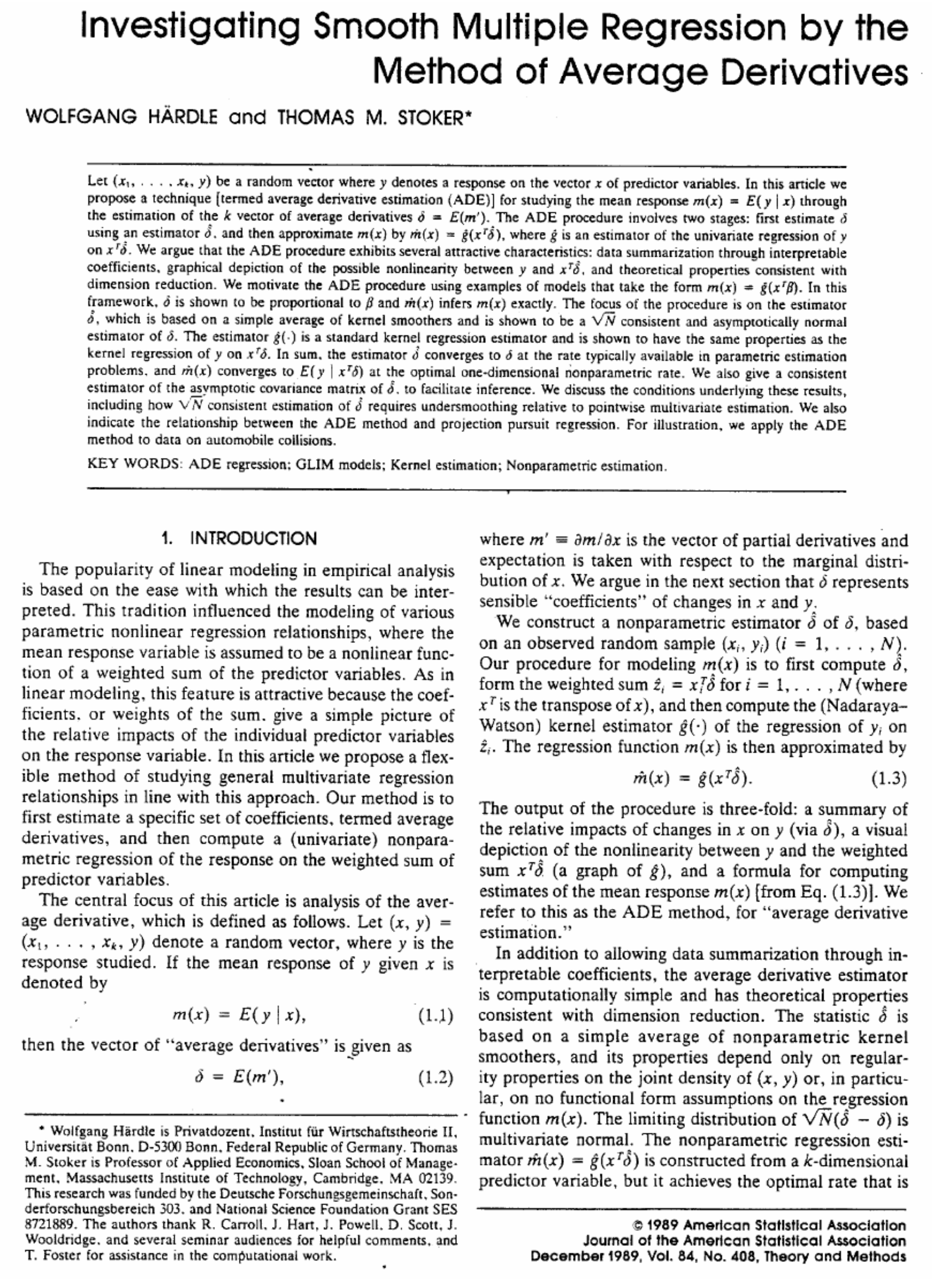
Scott, D. W. and Terrell, G. R. (1987) Biased and unbiased cross-validation in density estimation. *J. Am. Statist. Ass.*, **82**, 1131–1146.

Silverman, B. W. (1986) *Density Estimation for Statistics and Data Analysis*. New York: Chapman and Hall.

Stone, C. J. (1980) Optimal convergence rates for nonparametric estimators. *Ann. Statist.*, **8**, 1348–1360.

—— (1984) An asymptotically optimal window selection rule for kernel density estimates. *Ann. Statist.*, **12**, 1285–1297.

Journal of the American Statistical Assoication, 84, 986-995.



Investigating Smooth Multiple Regression by the Method of Average Derivatives

WOLFGANG HÄRDLE and THOMAS M. STOKER*

Let $(x_1, \ldots, x_k, y)$ be a random vector where y denotes a response on the vector x of predictor variables. In this article we propose a technique [termed average derivative estimation (ADE)] for studying the mean response $m(x) = E(y \mid x)$ through the estimation of the k vector of average derivatives $\delta = E(m')$. The ADE procedure involves two stages: first estimate δ using an estimator $\hat{\delta}$, and then approximate $m(x)$ by $\hat{m}(x) = \hat{g}(x^T\hat{\delta})$, where $\hat{g}$ is an estimator of the univariate regression of y on $x^T\hat{\delta}$. We argue that the ADE procedure exhibits several attractive characteristics: data summarization through interpretable coefficients, graphical depiction of the possible nonlinearity between y and $x^T\hat{\delta}$, and theoretical properties consistent with dimension reduction. We motivate the ADE procedure using examples of models that take the form $m(x) = \tilde{g}(x^T\beta)$. In this framework, δ is shown to be proportional to β and $\hat{m}(x)$ infers $m(x)$ exactly. The focus of the procedure is on the estimator $\hat{\delta}$, which is based on a simple average of kernel smoothers and is shown to be a $\sqrt{N}$ consistent and asymptotically normal estimator of δ. The estimator $\hat{g}(\cdot)$ is a standard kernel regression estimator and is shown to have the same properties as the kernel regression of y on $x^T\delta$. In sum, the estimator $\hat{\delta}$ converges to δ at the rate typically available in parametric estimation problems, and $\hat{m}(x)$ converges to $E(y \mid x^T\delta)$ at the optimal one-dimensional nonparametric rate. We also give a consistent estimator of the asymptotic covariance matrix of $\hat{\delta}$, to facilitate inference. We discuss the conditions underlying these results, including how $\sqrt{N}$ consistent estimation of $\hat{\delta}$ requires undersmoothing relative to pointwise multivariate estimation. We also indicate the relationship between the ADE method and projection pursuit regression. For illustration, we apply the ADE method to data on automobile collisions.

KEY WORDS: ADE regression; GLIM models; Kernel estimation; Nonparametric estimation.

1. INTRODUCTION

The popularity of linear modeling in empirical analysis is based on the ease with which the results can be interpreted. This tradition influenced the modeling of various parametric nonlinear regression relationships, where the mean response variable is assumed to be a nonlinear function of a weighted sum of the predictor variables. As in linear modeling, this feature is attractive because the coefficients, or weights of the sum, give a simple picture of the relative impacts of the individual predictor variables on the response variable. In this article we propose a flexible method of studying general multivariate regression relationships in line with this approach. Our method is to first estimate a specific set of coefficients, termed average derivatives, and then compute a (univariate) nonparametric regression of the response on the weighted sum of predictor variables.

The central focus of this article is analysis of the average derivative, which is defined as follows. Let $(x, y) = (x_1, \ldots, x_k, y)$ denote a random vector, where y is the response studied. If the mean response of y given x is denoted by

$$m(x) = E(y \mid x), \tag{1.1}$$

then the vector of "average derivatives" is given as

$$\delta = E(m'), \tag{1.2}$$

where $m' \equiv \partial m / \partial x$ is the vector of partial derivatives and expectation is taken with respect to the marginal distribution of x. We argue in the next section that δ represents sensible "coefficients" of changes in x and y.

We construct a nonparametric estimator $\hat{\delta}$ of δ, based on an observed random sample (x_i, y_i) $(i = 1, \ldots, N)$. Our procedure for modeling $m(x)$ is to first compute $\hat{\delta}$, form the weighted sum $\hat{z}_i = x_i^T\hat{\delta}$ for $i = 1, \ldots, N$ (where x^T is the transpose of x), and then compute the (Nadaraya–Watson) kernel estimator $\hat{g}(\cdot)$ of the regression of y_i on $\hat{z}_i$. The regression function $m(x)$ is then approximated by

$$\hat{m}(x) = \hat{g}(x^T\hat{\delta}). \tag{1.3}$$

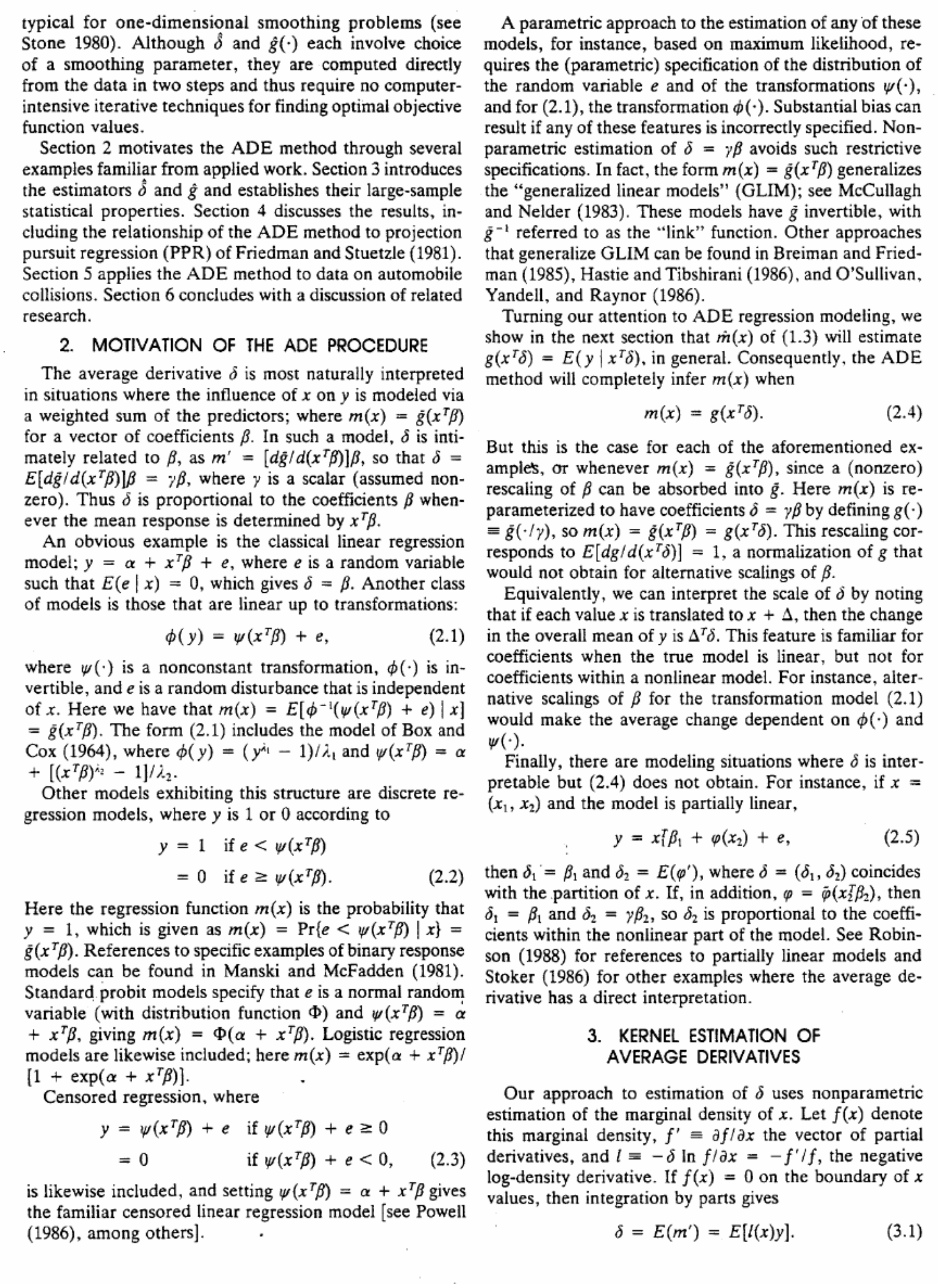
The output of the procedure is three-fold: a summary of the relative impacts of changes in x on y (via $\hat{\delta}$), a visual depiction of the nonlinearity between y and the weighted sum $x^T\hat{\delta}$ (a graph of $\hat{g}$), and a formula for computing estimates of the mean response $m(x)$ [from Eq. (1.3)]. We refer to this as the ADE method, for "average derivative estimation."

In addition to allowing data summarization through interpretable coefficients, the average derivative estimator is computationally simple and has theoretical properties consistent with dimension reduction. The statistic $\hat{\delta}$ is based on a simple average of nonparametric kernel smoothers, and its properties depend only on regularity properties on the joint density of (x, y) or, in particular, on no functional form assumptions on the regression function $m(x)$. The limiting distribution of $\sqrt{N}(\hat{\delta} - \delta)$ is multivariate normal. The nonparametric regression estimator $\hat{m}(x) = \hat{g}(x^T\hat{\delta})$ is constructed from a k-dimensional predictor variable, but it achieves the optimal rate that is

* Wolfgang Härdle is Privatdozent, Institut für Wirtschaftstheorie II, Universität Bonn, D-5300 Bonn, Federal Republic of Germany. Thomas M. Stoker is Professor of Applied Economics, Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA 02139. This research was funded by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 303, and National Science Foundation Grant SES 8721889. The authors thank R. Carroll, J. Hart, J. Powell, D. Scott, J. Wooldridge, and several seminar audiences for helpful comments, and T. Foster for assistance in the computational work.

Journal of the American Statistical Association
December 1989, Vol. 84, No. 408, Theory and Methods

Härdle, W. and Stoker, T. (1989) Investigating Smooth Multiple Regression by the Method of Average Derivatives.



typical for one-dimensional smoothing problems (see Stone 1980). Although $\hat{\delta}$ and $\hat{g}(\cdot)$ each involve choice of a smoothing parameter, they are computed directly from the data in two steps and thus require no computer-intensive iterative techniques for finding optimal objective function values.

Section 2 motivates the ADE method through several examples familiar from applied work. Section 3 introduces the estimators $\hat{\delta}$ and $\hat{g}$ and establishes their large-sample statistical properties. Section 4 discusses the results, including the relationship of the ADE method to projection pursuit regression (PPR) of Friedman and Stuetzle (1981). Section 5 applies the ADE method to data on automobile collisions. Section 6 concludes with a discussion of related research.

2. MOTIVATION OF THE ADE PROCEDURE

The average derivative δ is most naturally interpreted in situations where the influence of x on y is modeled via a weighted sum of the predictors; where $m(x) = \bar{g}(x^T\beta)$ for a vector of coefficients β. In such a model, δ is intimately related to β, as $m' = [d\bar{g}/d(x^T\beta)]\beta$, so that $\delta = E[d\bar{g}/d(x^T\beta)]\beta = \gamma\beta$, where γ is a scalar (assumed nonzero). Thus δ is proportional to the coefficients β whenever the mean response is determined by $x^T\beta$.

An obvious example is the classical linear regression model; $y = \alpha + x^T\beta + e$, where e is a random variable such that $E(e \mid x) = 0$, which gives $\delta = \beta$. Another class of models is those that are linear up to transformations:

$$\phi(y) = \psi(x^T\beta) + e, \tag{2.1}$$

where $\psi(\cdot)$ is a nonconstant transformation, $\phi(\cdot)$ is invertible, and e is a random disturbance that is independent of x. Here we have that $m(x) = E[\phi^{-1}(\psi(x^T\beta) + e) \mid x] = \bar{g}(x^T\beta)$. The form (2.1) includes the model of Box and Cox (1964), where $\phi(y) = (y^{\lambda_1} - 1)/\lambda_1$ and $\psi(x^T\beta) = \alpha + [(x^T\beta)^{\lambda_2} - 1]/\lambda_2$.

Other models exhibiting this structure are discrete regression models, where y is 1 or 0 according to

$$\begin{aligned} y &= 1 \quad \text{if } e < \psi(x^T\beta) \\ &= 0 \quad \text{if } e \geq \psi(x^T\beta). \end{aligned} \tag{2.2}$$

Here the regression function $m(x)$ is the probability that $y = 1$, which is given as $m(x) = \Pr\{e < \psi(x^T\beta) \mid x\} = \bar{g}(x^T\beta)$. References to specific examples of binary response models can be found in Manski and McFadden (1981). Standard probit models specify that e is a normal random variable (with distribution function Φ) and $\psi(x^T\beta) = \alpha + x^T\beta$, giving $m(x) = \Phi(\alpha + x^T\beta)$. Logistic regression models are likewise included; here $m(x) = \exp(\alpha + x^T\beta)/[1 + \exp(\alpha + x^T\beta)]$.

Censored regression, where

$$\begin{aligned} y &= \psi(x^T\beta) + e \quad &&\text{if } \psi(x^T\beta) + e \geq 0 \\ &= 0 &&\text{if } \psi(x^T\beta) + e < 0, \end{aligned} \tag{2.3}$$

is likewise included, and setting $\psi(x^T\beta) = \alpha + x^T\beta$ gives the familiar censored linear regression model [see Powell (1986), among others].

A parametric approach to the estimation of any of these models, for instance, based on maximum likelihood, requires the (parametric) specification of the distribution of the random variable e and of the transformations $\psi(\cdot)$, and for (2.1), the transformation $\phi(\cdot)$. Substantial bias can result if any of these features is incorrectly specified. Nonparametric estimation of $\delta = \gamma\beta$ avoids such restrictive specifications. In fact, the form $m(x) = \bar{g}(x^T\beta)$ generalizes the "generalized linear models" (GLIM); see McCullagh and Nelder (1983). These models have $\bar{g}$ invertible, with $\bar{g}^{-1}$ referred to as the "link" function. Other approaches that generalize GLIM can be found in Breiman and Friedman (1985), Hastie and Tibshirani (1986), and O'Sullivan, Yandell, and Raynor (1986).

Turning our attention to ADE regression modeling, we show in the next section that $\hat{m}(x)$ of (1.3) will estimate $g(x^T\delta) = E(y \mid x^T\delta)$, in general. Consequently, the ADE method will completely infer $m(x)$ when

$$m(x) = g(x^T\delta). \tag{2.4}$$

But this is the case for each of the aforementioned examples, or whenever $m(x) = \bar{g}(x^T\beta)$, since a (nonzero) rescaling of β can be absorbed into $\bar{g}$. Here $m(x)$ is reparameterized to have coefficients $\delta = \gamma\beta$ by defining $g(\cdot) \equiv \bar{g}(\cdot/\gamma)$, so $m(x) = \bar{g}(x^T\beta) = g(x^T\delta)$. This rescaling corresponds to $E[dg/d(x^T\delta)] = 1$, a normalization of g that would not obtain for alternative scalings of β.

Equivalently, we can interpret the scale of δ by noting that if each value x is translated to $x + \Delta$, then the change in the overall mean of y is $\Delta^T\delta$. This feature is familiar for coefficients when the true model is linear, but not for coefficients within a nonlinear model. For instance, alternative scalings of β for the transformation model (2.1) would make the average change dependent on $\phi(\cdot)$ and $\psi(\cdot)$.

Finally, there are modeling situations where δ is interpretable but (2.4) does not obtain. For instance, if $x = (x_1, x_2)$ and the model is partially linear,

$$y = x_1^T\beta_1 + \varphi(x_2) + e, \tag{2.5}$$

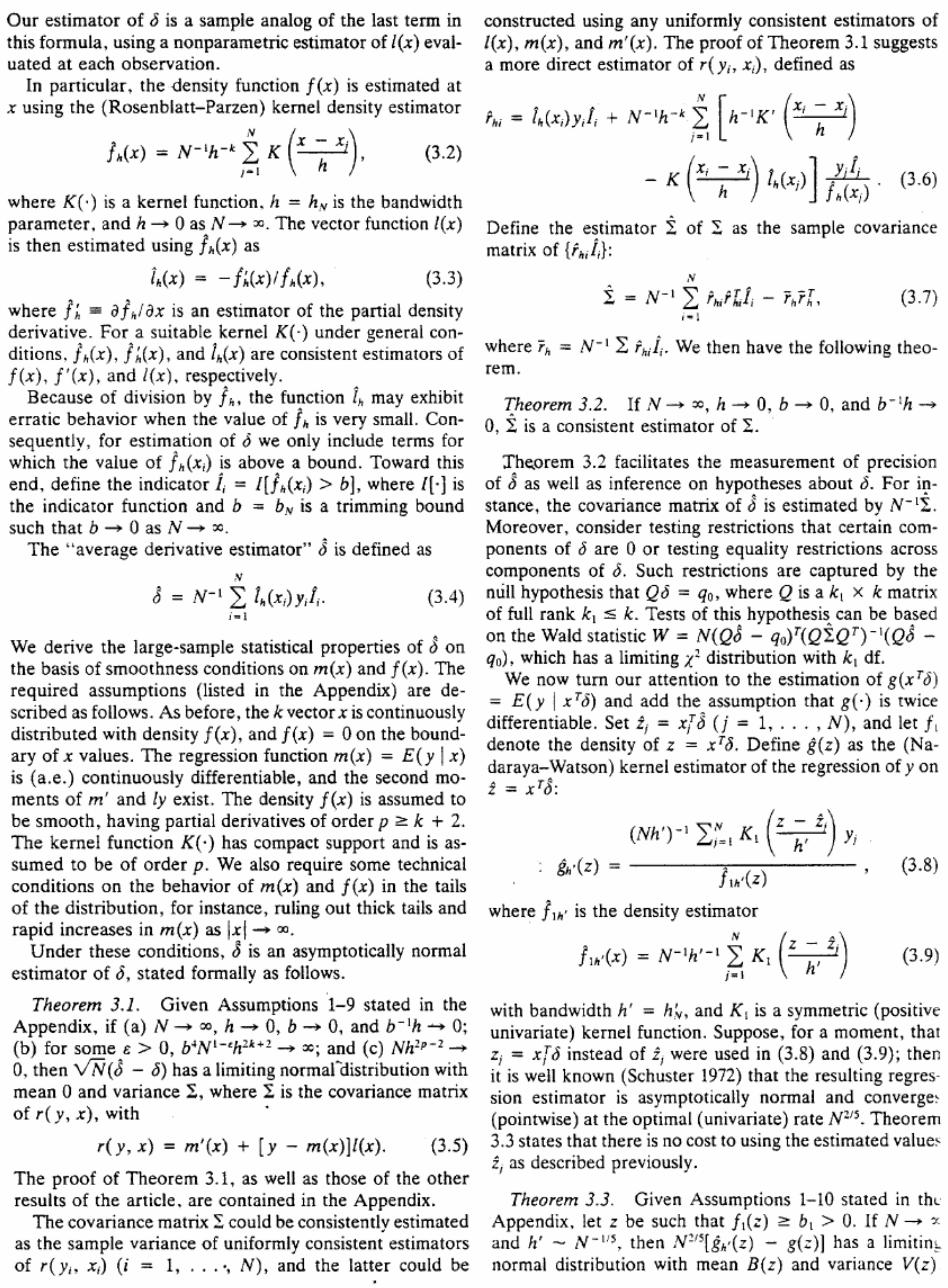
then $\delta_1 = \beta_1$ and $\delta_2 = E(\varphi')$, where $\delta = (\delta_1, \delta_2)$ coincides with the partition of x. If, in addition, $\varphi = \bar{\varphi}(x_2^T\beta_2)$, then $\delta_1 = \beta_1$ and $\delta_2 = \gamma\beta_2$, so δ_2 is proportional to the coefficients within the nonlinear part of the model. See Robinson (1988) for references to partially linear models and Stoker (1986) for other examples where the average derivative has a direct interpretation.

3. KERNEL ESTIMATION OF AVERAGE DERIVATIVES

Our approach to estimation of δ uses nonparametric estimation of the marginal density of x. Let $f(x)$ denote this marginal density, $f' \equiv \partial f/\partial x$ the vector of partial derivatives, and $l \equiv -\delta \ln f/\partial x = -f'/f$, the negative log-density derivative. If $f(x) = 0$ on the boundary of x values, then integration by parts gives

$$\delta = E(m') = E[l(x)y]. \tag{3.1}$$

Journal of the American Statistical Assoication, 84, 986-995.



Our estimator of δ is a sample analog of the last term in this formula, using a nonparametric estimator of $l(x)$ evaluated at each observation.

In particular, the density function $f(x)$ is estimated at x using the (Rosenblatt–Parzen) kernel density estimator

$$\hat{f}_h(x) = N^{-1}h^{-k} \sum_{j=1}^{N} K\left(\frac{x - x_j}{h}\right), \qquad (3.2)$$

where $K(\cdot)$ is a kernel function, $h = h_N$ is the bandwidth parameter, and $h \to 0$ as $N \to \infty$. The vector function $l(x)$ is then estimated using $\hat{f}_h(x)$ as

$$\hat{l}_h(x) = -\hat{f}_h'(x)/\hat{f}_h(x), \qquad (3.3)$$

where $\hat{f}_h' \equiv \partial\hat{f}_h/\partial x$ is an estimator of the partial density derivative. For a suitable kernel $K(\cdot)$ under general conditions, $\hat{f}_h(x)$, $\hat{f}_h'(x)$, and $\hat{l}_h(x)$ are consistent estimators of $f(x)$, $f'(x)$, and $l(x)$, respectively.

Because of division by $\hat{f}_h$, the function $\hat{l}_h$ may exhibit erratic behavior when the value of $\hat{f}_h$ is very small. Consequently, for estimation of δ we only include terms for which the value of $\hat{f}_h(x_i)$ is above a bound. Toward this end, define the indicator $\hat{I}_i = I[\hat{f}_h(x_i) > b]$, where $I[\cdot]$ is the indicator function and $b = b_N$ is a trimming bound such that $b \to 0$ as $N \to \infty$.

The "average derivative estimator" $\hat{\delta}$ is defined as

$$\hat{\delta} = N^{-1} \sum_{i=1}^{N} \hat{l}_h(x_i) y_i \hat{I}_i. \qquad (3.4)$$

We derive the large-sample statistical properties of $\hat{\delta}$ on the basis of smoothness conditions on $m(x)$ and $f(x)$. The required assumptions (listed in the Appendix) are described as follows. As before, the k vector x is continuously distributed with density $f(x)$, and $f(x) = 0$ on the boundary of x values. The regression function $m(x) = E(y \mid x)$ is (a.e.) continuously differentiable, and the second moments of m' and ly exist. The density $f(x)$ is assumed to be smooth, having partial derivatives of order $p \geq k + 2$. The kernel function $K(\cdot)$ has compact support and is assumed to be of order p. We also require some technical conditions on the behavior of $m(x)$ and $f(x)$ in the tails of the distribution, for instance, ruling out thick tails and rapid increases in $m(x)$ as $|x| \to \infty$.

Under these conditions, $\hat{\delta}$ is an asymptotically normal estimator of δ, stated formally as follows.

Theorem 3.1. Given Assumptions 1–9 stated in the Appendix, if (a) $N \to \infty$, $h \to 0$, $b \to 0$, and $b^{-1}h \to 0$; (b) for some $\varepsilon > 0$, $b^4N^{1-\varepsilon}h^{2k+2} \to \infty$; and (c) $Nh^{2p-2} \to 0$, then $\sqrt{N}(\hat{\delta} - \delta)$ has a limiting normal distribution with mean 0 and variance Σ, where Σ is the covariance matrix of $r(y, x)$, with

$$r(y, x) = m'(x) + [y - m(x)]l(x). \qquad (3.5)$$

The proof of Theorem 3.1, as well as those of the other results of the article, are contained in the Appendix.

The covariance matrix Σ could be consistently estimated as the sample variance of uniformly consistent estimators of $r(y_i, x_i)$ $(i = 1, \ldots, N)$, and the latter could be constructed using any uniformly consistent estimators of $l(x)$, $m(x)$, and $m'(x)$. The proof of Theorem 3.1 suggests a more direct estimator of $r(y_i, x_i)$, defined as

$$\hat{r}_{hi} = \hat{l}_h(x_i)y_i\hat{I}_i + N^{-1}h^{-k} \sum_{j=1}^{N} \left[h^{-1}K'\left(\frac{x_i - x_j}{h}\right) - K\left(\frac{x_i - x_j}{h}\right)\hat{l}_h(x_j)\right]\frac{y_j\hat{I}_j}{\hat{f}_h(x_j)}. \qquad (3.6)$$

Define the estimator $\hat{\Sigma}$ of Σ as the sample covariance matrix of $\{\hat{r}_{hi}\hat{I}_i\}$:

$$\hat{\Sigma} = N^{-1} \sum_{i=1}^{N} \hat{r}_{hi}\hat{r}_{hi}^T\hat{I}_i - \bar{r}_h\bar{r}_h^T, \qquad (3.7)$$

where $\bar{r}_h = N^{-1} \sum \hat{r}_{hi}\hat{I}_i$. We then have the following theorem.

Theorem 3.2. If $N \to \infty$, $h \to 0$, $b \to 0$, and $b^{-1}h \to 0$, $\hat{\Sigma}$ is a consistent estimator of Σ.

Theorem 3.2 facilitates the measurement of precision of $\hat{\delta}$ as well as inference on hypotheses about δ. For instance, the covariance matrix of $\hat{\delta}$ is estimated by $N^{-1}\hat{\Sigma}$. Moreover, consider testing restrictions that certain components of δ are 0 or testing equality restrictions across components of δ. Such restrictions are captured by the null hypothesis that $Q\delta = q_0$, where Q is a $k_1 \times k$ matrix of full rank $k_1 \leq k$. Tests of this hypothesis can be based on the Wald statistic $W = N(Q\hat{\delta} - q_0)^T(Q\hat{\Sigma}Q^T)^{-1}(Q\hat{\delta} - q_0)$, which has a limiting χ^2 distribution with k_1 df.

We now turn our attention to the estimation of $g(x^T\delta) = E(y \mid x^T\delta)$ and add the assumption that $g(\cdot)$ is twice differentiable. Set $\hat{z}_j = x_j^T\hat{\delta}$ $(j = 1, \ldots, N)$, and let f_1 denote the density of $z = x^T\delta$. Define $\hat{g}(z)$ as the (Nadaraya–Watson) kernel estimator of the regression of y on $\hat{z} = x^T\hat{\delta}$:

$$\hat{g}_{h'}(z) = \frac{(Nh')^{-1} \sum_{j=1}^{N} K_1\left(\frac{z - \hat{z}_j}{h'}\right) y_j}{\hat{f}_{1h'}(z)}, \qquad (3.8)$$

where $\hat{f}_{1h'}$ is the density estimator

$$\hat{f}_{1h'}(x) = N^{-1}h'^{-1} \sum_{j=1}^{N} K_1\left(\frac{z - \hat{z}_j}{h'}\right) \qquad (3.9)$$

with bandwidth $h' = h_N'$, and K_1 is a symmetric (positive univariate) kernel function. Suppose, for a moment, that $z_j = x_j^T\delta$ instead of $\hat{z}_j$ were used in (3.8) and (3.9); then it is well known (Schuster 1972) that the resulting regression estimator is asymptotically normal and converges (pointwise) at the optimal (univariate) rate $N^{2/5}$. Theorem 3.3 states that there is no cost to using the estimated values $\hat{z}_j$ as described previously.

Theorem 3.3. Given Assumptions 1–10 stated in the Appendix, let z be such that $f_1(z) \geq b_1 > 0$. If $N \to \infty$ and $h' \sim N^{-1/5}$, then $N^{2/5}[\hat{g}_{h'}(z) - g(z)]$ has a limiting normal distribution with mean $B(z)$ and variance $V(z)$

Härdle, W. and Stoker, T. (1989) Investigating Smooth Multiple Regression by the Method of Average Derivatives.

where

$$B(z) = [g''(z)/2 + g'(z)f_1'(z)/f_1(z)] \int u^2 K_1(u)\, du$$

$$V(z) = [\operatorname{var}(y \mid x'\delta = z)/f_1(z)] \int K_1(u)^2\, du. \qquad (3.10)$$

The bias and variance given in (3.10) can be estimated consistently for each z using y, $\hat{g}_h$, and $\hat{f}_{1h}$, and their derivatives, using standard methods. Therefore, asymptotic confidence intervals can be constructed for $\hat{g}_{h'}(z)$. It is clear that the same confidence intervals apply to $\hat{m}(x) = \hat{g}_{h'}(x^T\hat{\delta})$, for $z = x^T\hat{\delta}$.

4. REMARKS AND DISCUSSION

4.1 The Average Derivative Estimator

As indicated in the Introduction, the most interesting feature of Theorem 3.1 is that $\hat{\delta}$ converges to δ at rate $\sqrt{N}$. This is the rate typically available in parametric estimation problems and is the rate that would be attained if the values $l(x_i)$ $(i = 1, \ldots, N)$ were known and used in the average (3.4). The estimator $\hat{l}_h(x)$ converges pointwise to $l(x)$ at a slower rate, so Theorem 3.1 gives a situation where the average of nonparametric estimators converges more quickly than any of its individual components. This occurs because of the overlap between kernel densities at different evaluation points; for instance, if x_i and x_j are sufficiently close, the data used in the local average $\hat{f}_h(x_i)$ will overlap with that used in $\hat{f}_h(x_j)$. These overlaps lead to the approximation of $\hat{\delta}$ by U statistics with kernels depending on N. The asymptotic normality of $\hat{\delta}$ follows from results on the equivalence of such U statistics to (ordinary) sample average. In a similar spirit, Powell, Stock, and Stoker (in press) obtained $\sqrt{N}$ convergence rates for the estimation of "density weighted" average derivatives, and Carroll (1982), Robinson (1988), and Härdle and Marron (1987) showed how kernel densities can be used to obtain $\sqrt{N}$ convergence rates for certain parameters in specific semiparametric models. We also note that our method of trimming follows Bickel (1982), Manski (1984), and Robinson (1988).

For any given sample size, the bandwidth h and the trimming bound b can be set to any (positive) values, so their choice can be based entirely on the small-sample behavior of $\hat{\delta}$. Conditions (a)–(c) of Theorem 3.1 indicate how the initial bandwidth and trimming bound must be decreased as the sample size is increased. These conditions are certainly feasible; suppose that $h = h_0N^{-\zeta}$ and $b = b_0N^{-\eta}$, then (a)–(c) are equivalent to $0 < \eta \leq \zeta$ and $p/(2p - 2) < \zeta < (1 - 4\eta - \varepsilon)/(2k + 2)$. Since $p \geq k + 2$ and ε is arbitrarily small, η can be chosen small enough to fulfill the last condition.

The bandwidth conditions arise as follows. Condition (b) assures that the estimator $\hat{\delta}$ can be "linearized" to 1 without an estimated denominator and is a sufficient condition for asymptotic normality. Condition (c) assures that the bias of $\hat{\delta}$ vanishes at rate $\sqrt{N}$. Conditions (a)–(c) are one-sided in implying that the trimming bound b cannot converge too quickly to 0 as $N \to \infty$, but rather must converge slowly. The behavior of the bandwidth h as $N \to \infty$ is bounded both below and above by Conditions (b) and (c).

Condition (c) does imply that the pointwise convergence of $\hat{f}_h(x)$ to $f(x)$ must be suboptimal. Stone (1980) showed that the optimal pointwise rate of convergence under our conditions is $N^{p/(2p+k)}$, and Collomb and Härdle (1986) showed that this rate is achievable with kernel density estimators such as (3.2), for instance, by taking ƙ $= h_0N^{-1/(2p+k)}$. But we have that Nƙ$^{2p-2} \to \infty$, which violates Condition (c), so as $N \to \infty$, h must converge to 0 more quickly than ƙ.This occurs because (c) is a bias condition; as $N \to \infty$, the (pointwise) bias of $\hat{f}_h(x)$ must vanish at a faster rate than its (pointwise) variance, for the bias of $\hat{\delta}$ to be $o(N^{-1/2})$. In other words, for $\sqrt{N}$ consistent estimation of δ, one must "undersmooth" the nonparametric component $\hat{l}_h(x)$.

4.2 Modeling Multiple Regression

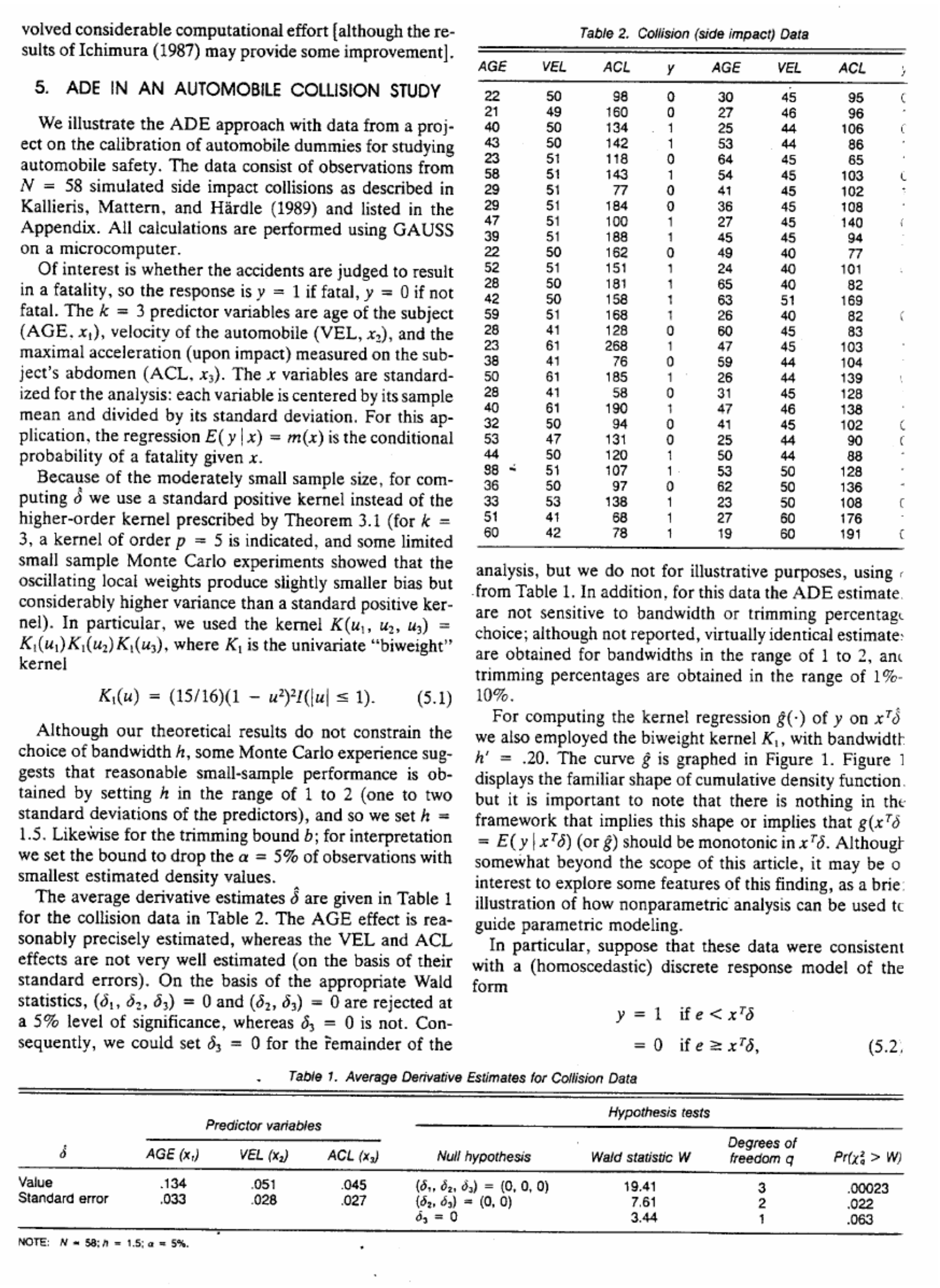
Theorem 3.3 shows that the optimal one-dimensional convergenee rate is achievable in the estimation of $g(x^T\delta) = E(y \mid x^T\delta)$, using $\hat{\delta}$ instead of δ. The requirement that $g(\cdot)$ is twice differentiable affixes the optimal rate at $N^{2/5}$, but otherwise plays no role: if $g(\cdot)$ is assumed differentiable of order q and $K_1(\cdot)$ is a kernel of order q, then it is easily shown that the optimal rate of $N^{q/(2q+1)}$ is attained. The attainment of optimal one-dimensional rates of convergence is possible for the ADE method because the additive structure of $g(x^T\delta)$ is sufficient to follow the "dimension reduction principle" of Stone (1986). Alternative uses of additive structure can be found in Breiman and Friedman (1985) and Hastie and Tibshirani (1986).

The ADE method can be regarded as a version of PPR of Friedman and Stuetzle (1981). The first step of PPR is to choose β (normalized as a direction) and $\bar{g}$ to minimize $s(\bar{g}, \beta) = \Sigma\, [y_i - \bar{g}(x_i^T\beta)]^2$, and any model of the form $m(x) = \bar{g}(x^T\beta)$ is inferred by the ADE estimator $\hat{m}(x) = \hat{g}_{h'}(x^T\hat{\delta})$ at the optimal one-dimensional rate of convergence. For a general regression function, however, $\hat{m}(x)$, $\hat{g}_{h'}$, and $\hat{\delta}$ will not necessarily minimize the sum of squares $s(\bar{g}, \beta)$: given $\bar{g}$, β is chosen such that $\{y_i - \bar{g}(x_i^T\beta)\}$ is orthogonal to $\{x_i\bar{g}'(x_i^T\beta)\}$, which does not imply that $\beta = \hat{\delta}/|\hat{\delta}|$.

Given $\hat{\delta}$, $\hat{g}_{h'}$ is a local least squares estimator; namely, $\Sigma\, K_1[(z - x_i^T\hat{\delta})/h'](y_i - t)^2$ is minimized by $t = \hat{g}_{h'}(z)$. Moreover, $\hat{\delta}$ is a type of least squares estimator, as follows. Set $\hat{\mathfrak{L}}_i = (S_l)^{-1}\hat{l}_h(x_i)\hat{I}_i$, where S_l is the sample moment $S_l = N^{-1}\,\Sigma\, \hat{l}_h(x_i)\hat{l}_h(x_i)^T\hat{I}_i$. Then $\hat{\delta}$ is the value of d that minimizes the sum of squares $\Sigma\, [y_i - \hat{\mathfrak{L}}_i^Td]^2$, or equivalently, $S_l^{-1}\hat{\delta}$ are the coordinates of $\{y_i\}$ projected onto the subspace spanned by $\{\hat{l}_h(x_i)\hat{I}_i\}$.

ADE and PPR thus represent different computational methods of inferring $m(x) = \bar{g}(x^T\beta)$. The possible advantages of ADE arise from reduced computational effort; (given h, b, and h') $\hat{m}(x) = \hat{g}(x^T\hat{\delta})$ is computed directly from the data, whereas minimizing $s(\bar{g}, \beta)$ (by checking all directions β and computing $\bar{g}$ for each β) typically in-

Journal of the American Statistical Assoication, 84, 986-995.



volved considerable computational effort [although the results of Ichimura (1987) may provide some improvement].

5. ADE IN AN AUTOMOBILE COLLISION STUDY

We illustrate the ADE approach with data from a project on the calibration of automobile dummies for studying automobile safety. The data consist of observations from $N = 58$ simulated side impact collisions as described in Kallieris, Mattern, and Härdle (1989) and listed in the Appendix. All calculations are performed using GAUSS on a microcomputer.

Of interest is whether the accidents are judged to result in a fatality, so the response is $y = 1$ if fatal, $y = 0$ if not fatal. The $k = 3$ predictor variables are age of the subject (AGE, x_1), velocity of the automobile (VEL, x_2), and the maximal acceleration (upon impact) measured on the subject's abdomen (ACL, x_3). The x variables are standardized for the analysis: each variable is centered by its sample mean and divided by its standard deviation. For this application, the regression $E(y \mid x) = m(x)$ is the conditional probability of a fatality given x.

Because of the moderately small sample size, for computing $\hat{\delta}$ we use a standard positive kernel instead of the higher-order kernel prescribed by Theorem 3.1 (for $k = 3$, a kernel of order $p = 5$ is indicated, and some limited small sample Monte Carlo experiments showed that the oscillating local weights produce slightly smaller bias but considerably higher variance than a standard positive kernel). In particular, we used the kernel $K(u_1, u_2, u_3) = K_1(u_1)K_1(u_2)K_1(u_3)$, where K_1 is the univariate "biweight" kernel

$$K_1(u) = (15/16)(1 - u^2)^2 I(|u| \le 1). \qquad (5.1)$$

Although our theoretical results do not constrain the choice of bandwidth h, some Monte Carlo experience suggests that reasonable small-sample performance is obtained by setting h in the range of 1 to 2 (one to two standard deviations of the predictors), and so we set $h = 1.5$. Likewise for the trimming bound b; for interpretation we set the bound to drop the $\alpha = 5\%$ of observations with smallest estimated density values.

The average derivative estimates $\hat{\delta}$ are given in Table 1 for the collision data in Table 2. The AGE effect is reasonably precisely estimated, whereas the VEL and ACL effects are not very well estimated (on the basis of their standard errors). On the basis of the appropriate Wald statistics, $(\delta_1, \delta_2, \delta_3) = 0$ and $(\delta_2, \delta_3) = 0$ are rejected at a 5% level of significance, whereas $\delta_3 = 0$ is not. Consequently, we could set $\delta_3 = 0$ for the remainder of the analysis, but we do not for illustrative purposes, using from Table 1. In addition, for this data the ADE estimate are not sensitive to bandwidth or trimming percentag choice; although not reported, virtually identical estimate are obtained for bandwidths in the range of 1 to 2, an trimming percentages are obtained in the range of 1%-10%.

Table 2. Collision (side impact) Data

AGE	VEL	ACL	y	AGE	VEL	ACL
22	50	98	0	30	45	95
21	49	160	0	27	46	96
40	50	134	1	25	44	106
43	50	142	1	53	44	86
23	51	118	0	64	45	65
58	51	143	1	54	45	103
29	51	77	0	41	45	102
29	51	184	0	36	45	108
47	51	100	1	27	45	140
39	51	188	1	45	45	94
22	50	162	0	49	40	77
52	51	151	1	24	40	101
28	50	181	1	65	40	82
42	50	158	1	63	51	169
59	51	168	1	26	40	82
28	41	128	0	60	45	83
23	61	268	1	47	45	103
38	41	76	0	59	44	104
50	61	185	1	26	44	139
28	41	58	0	31	45	128
40	61	190	1	47	46	138
32	50	94	0	41	45	102
53	47	131	0	25	44	90
44	50	120	1	50	44	88
98	51	107	1	53	50	128
36	50	97	0	62	50	136
33	53	138	1	23	50	108
51	41	68	1	27	60	176
60	42	78	1	19	60	191

For computing the kernel regression $\hat{g}(\cdot)$ of y on $x^T\hat{\delta}$ we also employed the biweight kernel K_1, with bandwidth $h' = .20$. The curve $\hat{g}$ is graphed in Figure 1. Figure 1 displays the familiar shape of cumulative density function, but it is important to note that there is nothing in the framework that implies this shape or implies that $g(x^T\delta = E(y \mid x^T\delta)$ (or $\hat{g}$) should be monotonic in $x^T\delta$. Althoug somewhat beyond the scope of this article, it may be o interest to explore some features of this finding, as a brie illustration of how nonparametric analysis can be used to guide parametric modeling.

In particular, suppose that these data were consistent with a (homoscedastic) discrete response model of the form

$$\begin{aligned} y &= 1 \quad \text{if } e < x^T\delta \\ &= 0 \quad \text{if } e \ge x^T\delta, \end{aligned} \qquad (5.2$$

Table 1. Average Derivative Estimates for Collision Data

$\hat{\delta}$	Predictor variables AGE (x_1)	 VEL (x_2)	 ACL (x_3)	Hypothesis tests Null hypothesis	 Wald statistic W	 Degrees of freedom q	 $Pr(\chi^2_q > W)$
Value	.134	.051	.045	$(\delta_1, \delta_2, \delta_3) = (0, 0, 0)$	19.41	3	.00023
Standard error	.033	.028	.027	$(\delta_2, \delta_3) = (0, 0)$	7.61	2	.022
				$\delta_3 = 0$	3.44	1	.063

NOTE: $N = 58$; $h = 1.5$; $\alpha = 5\%$.

Härdle, W. and Stoker, T. (1989) Investigating Smooth Multiple Regression by the Method of Average Derivatives.

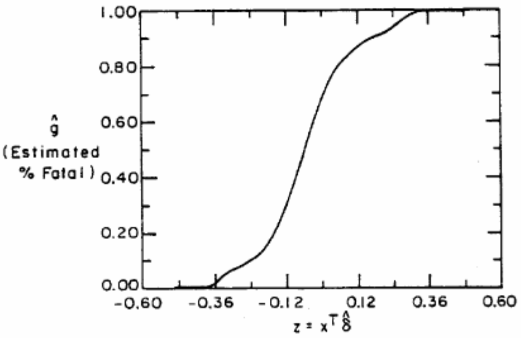


Figure 1. ADE Regression for Collision Data.

where e is distributed independently of x (possibly with nonzero mean). This formulation specializes (2.2) by setting $\psi(z) = z$ and normalizing β to δ. As discussed previously, in this model $E(y \mid x) = m(x) = g(y \mid x^T\delta) = \Pr(e < x^T\delta)$, so $g(z)$ is the cumulative distribution function of e. Moreover, under this model, $g' = dg/dz$ is the density function of e, which we can estimate by the kernel estimator of g' (the derivative of $\hat{g}$). The estimator is graphed in Figure 2. Its multimodal shape indicates the possibility of a "mixture" distribution for e, which is in contrast with standard parameterizations of binary response models (e.g., in a probit model e is assumed to be normally distributed). Although we do not pursue these issues further here, at minimum, these results suggest that one should test for the presence of a mixture, as well as look for additional distinctions in the data (or design discrimination rules) that can be built into the model so that a unimodal density for e is statistically appropriate.

The appearance of several modes of $\hat{g}'(\cdot)$ is not due to undersmoothing; it remains with a tripling of the bandwidth h' to .6. Moreover, it is not due to using the imprecise estimate $\hat{\delta}_3$; dropping ACL and reestimating gives a more pronounced multimodal shape of $\hat{g}'$.

There is one feature of the results, which appears in conflict with the framework, that merits further study. The average of $\hat{g}'(x_i^T\hat{\delta})$ over the data is 1.76, which contrasts with the normalization $E(g') = 1$. Although possibly due to sampling error or our particular choice of h' (doubling h' to .4 decreases the average to 1.17), this could signal underestimation of δ or, in particular, underestimation of the scale of δ. Although we could easily "correct" for this, our intention here is just to indicate the need for further study of scaling and/or normalization of $\hat{\delta}$. With regard to the preceding discussion, it is important to note that a rescaling of $\hat{\delta}$ would only relabel the horizontal axes of Figures 1 and 2. In particular, the scaling of $\hat{\delta}$ does not affect the substantive conclusions of Figures 1 and 2, nor does it affect the fitted values $\hat{m}(x)$ of the ADE model (1.3).

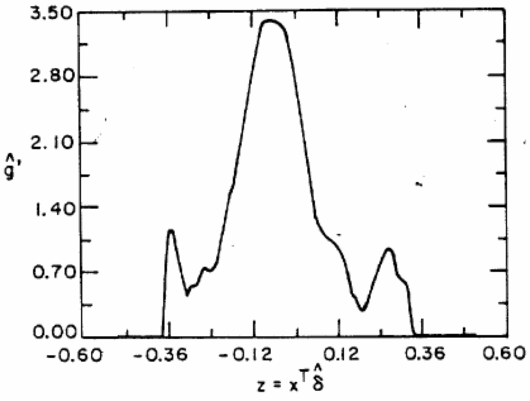


Figure 2. ADE Regression Derivative for Collision Data.

6. CONCLUDING REMARKS

In this article we have advanced the ADE method as a useful yet flexible tool for studying general regression relationships. At its center is the estimation of average derivatives, which we propose as sensible coefficients for measuring the relative impacts of separate predictor variables on the mean response. Although we have established attractive statistical properties for the estimators, it is important to stress that the real motivation for the ADE method is the economy it offers for nonparametric data summarization. Instead of attempting to interpret a fully flexible nonparametric regression, the ADE method permits the significance of individual predictor variables to be judged via simple hypothesis tests on the value of the average derivatives. Nonlinearity of the relationship is summarized by a graph of the function $\hat{g}$. As such, we regard the ADE method as a natural outgrowth of linear modeling, or "running (ordinary least squares) regressions," as a useful method of data summarization.

Although the results of our empirical illustration are encouraging [another application is given in Härdle, Hildenbrand, and Jerison (1988)], many questions can be posed regarding practical implementation of the ADE estimators. For instance, are there automatic methods for setting the bandwidth and trimming parameters that assure good small-sample performance of the estimators? Would small-sample performance be improved by normalizing the scale of $\hat{\delta}$ or using alternative methods of nonparametric approximation for the ingredients of $\hat{\delta}$? These sorts of issues need to be addressed as part of future research.

The ADE estimators are simple to compute, using standard software packages available for microcomputers. In addition, these procedures are being implemented as part of the exploratory data software package XploRe of Härdle (1988).

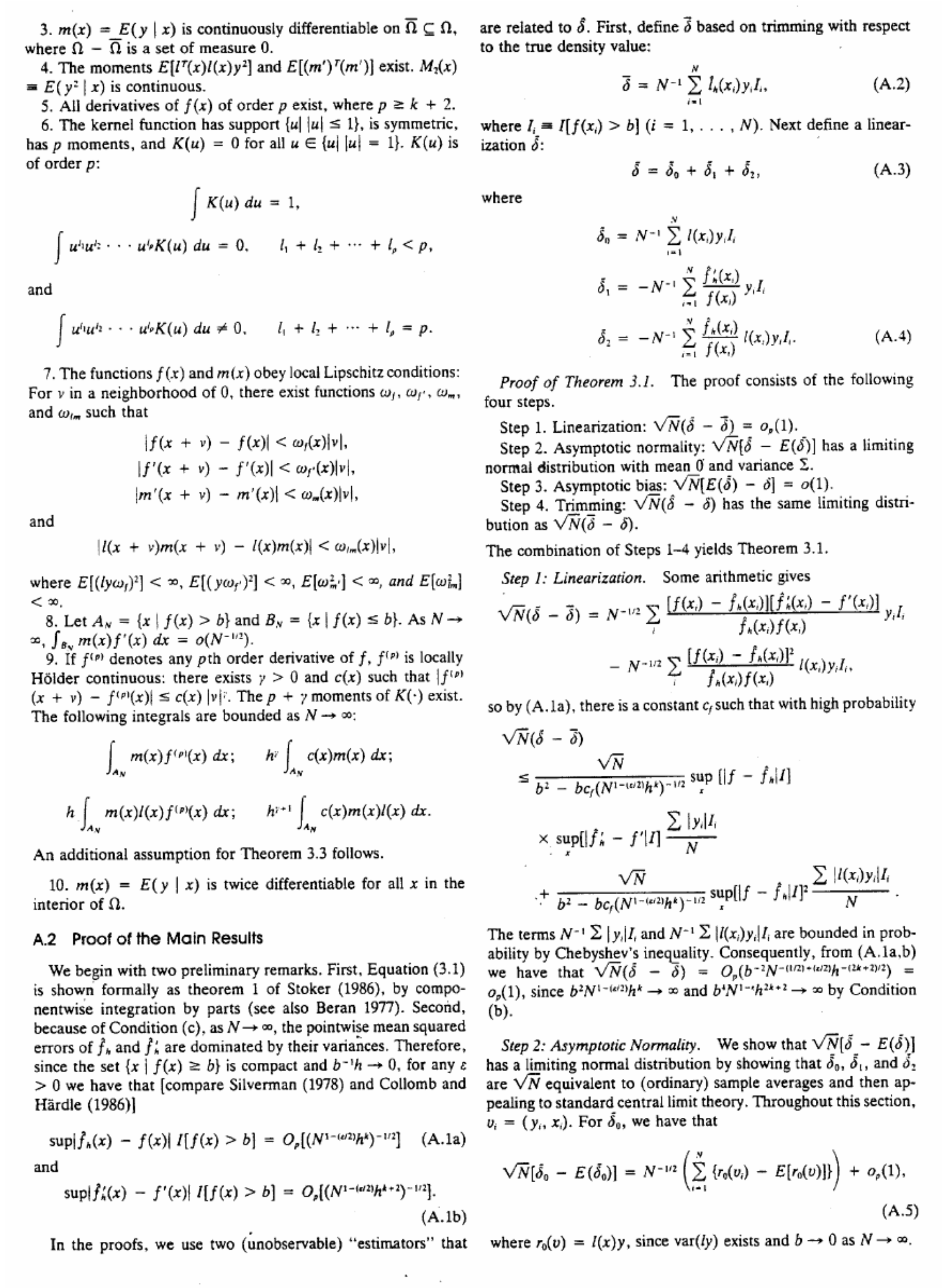
APPENDIX: ASSUMPTIONS, PROOFS OF THEOREMS, AND DATA

A.1 Assumptions for Theorems 3.1, 3.2, and 3.3

1. The support Ω of f is a convex, possibly unbounded subset of $\mathbf{R}^k$ with nonempty interior. The underlying measure of (y, x) can be written as $\mu_y \times \mu_x$, where μ_x is Lebesgue measure.

2. $f(x) = 0$ for all $x \in d\Omega$, where $d\Omega$ is the boundary of Ω.

Härdle, W. and Stoker, T. (1989) Investigating Smooth Multiple Regression by the Method of Average Derivatives.



3. $m(x) = E(y \mid x)$ is continuously differentiable on $\overline{\Omega} \subseteq \Omega$, where $\Omega - \overline{\Omega}$ is a set of measure 0.

4. The moments $E[l^T(x)l(x)y^2]$ and $E[(m')^T(m')]$ exist. $M_2(x) \equiv E(y^2 \mid x)$ is continuous.

5. All derivatives of $f(x)$ of order p exist, where $p \geq k + 2$.

6. The kernel function has support $\{u \mid |u| \leq 1\}$, is symmetric, has p moments, and $K(u) = 0$ for all $u \in \{u \mid |u| = 1\}$. $K(u)$ is of order p:

$$\int K(u)\,du = 1,$$

$$\int u^{l_1}u^{l_2}\cdots u^{l_p}K(u)\,du = 0, \qquad l_1 + l_2 + \cdots + l_p < p,$$

and

$$\int u^{l_1}u^{l_2}\cdots u^{l_p}K(u)\,du \neq 0, \qquad l_1 + l_2 + \cdots + l_p = p.$$

7. The functions $f(x)$ and $m(x)$ obey local Lipschitz conditions: For v in a neighborhood of 0, there exist functions ω_f, $\omega_{f'}$, ω_m, and ω_{lm} such that

$$|f(x + v) - f(x)| < \omega_f(x)|v|,$$
$$|f'(x + v) - f'(x)| < \omega_{f'}(x)|v|,$$
$$|m'(x + v) - m'(x)| < \omega_m(x)|v|,$$

and

$$|l(x + v)m(x + v) - l(x)m(x)| < \omega_{lm}(x)|v|,$$

where $E[(ly\omega_f)^2] < \infty$, $E[(y\omega_{f'})^2] < \infty$, $E[\omega^2_{m'}] < \infty$, *and* $E[\omega^2_{lm}] < \infty$.

8. Let $A_N = \{x \mid f(x) > b\}$ and $B_N = \{x \mid f(x) \leq b\}$. As $N \to \infty$, $\int_{B_N} m(x)f'(x)\,dx = o(N^{-1/2})$.

9. If $f^{(p)}$ denotes any pth order derivative of f, $f^{(p)}$ is locally Hölder continuous: there exists $\gamma > 0$ and $c(x)$ such that $|f^{(p)}(x + v) - f^{(p)}(x)| \leq c(x)\,|v|^{\gamma}$. The $p + \gamma$ moments of $K(\cdot)$ exist. The following integrals are bounded as $N \to \infty$:

$$\int_{A_N} m(x)f^{(p)}(x)\,dx; \qquad h^{\gamma}\int_{A_N} c(x)m(x)\,dx;$$

$$h\int_{A_N} m(x)l(x)f^{(p)}(x)\,dx; \qquad h^{\gamma+1}\int_{A_N} c(x)m(x)l(x)\,dx.$$

An additional assumption for Theorem 3.3 follows.

10. $m(x) = E(y \mid x)$ is twice differentiable for all x in the interior of Ω.

A.2 Proof of the Main Results

We begin with two preliminary remarks. First, Equation (3.1) is shown formally as theorem 1 of Stoker (1986), by componentwise integration by parts (see also Beran 1977). Second, because of Condition (c), as $N \to \infty$, the pointwise mean squared errors of $\hat{f}_h$ and $\hat{f}'_h$ are dominated by their variances. Therefore, since the set $\{x \mid f(x) \geq b\}$ is compact and $b^{-1}h \to 0$, for any $\varepsilon > 0$ we have that [compare Silverman (1978) and Collomb and Härdle (1986)]

$$\sup|\hat{f}_h(x) - f(x)|\, I[f(x) > b] = O_p[(N^{1-(\varepsilon/2)}h^k)^{-1/2}] \qquad \text{(A.1a)}$$

and

$$\sup|\hat{f}'_h(x) - f'(x)|\, I[f(x) > b] = O_p[(N^{1-(\varepsilon/2)}h^{k+2})^{-1/2}]. \qquad \text{(A.1b)}$$

In the proofs, we use two (unobservable) "estimators" that are related to $\hat{\delta}$. First, define $\tilde{\delta}$ based on trimming with respect to the true density value:

$$\tilde{\delta} = N^{-1}\sum_{i=1}^{N} l_h(x_i)y_iI_i, \qquad \text{(A.2)}$$

where $I_i \equiv I[f(x_i) > b]$ $(i = 1, \ldots, N)$. Next define a linearization $\bar{\delta}$:

$$\bar{\delta} = \bar{\delta}_0 + \bar{\delta}_1 + \bar{\delta}_2, \qquad \text{(A.3)}$$

where

$$\bar{\delta}_0 = N^{-1}\sum_{i=1}^{N} l(x_i)y_iI_i$$

$$\bar{\delta}_1 = -N^{-1}\sum_{i=1}^{N}\frac{\hat{f}'_h(x_i)}{f(x_i)}\,y_iI_i$$

$$\bar{\delta}_2 = -N^{-1}\sum_{i=1}^{N}\frac{\hat{f}_h(x_i)}{f(x_i)}\,l(x_i)y_iI_i. \qquad \text{(A.4)}$$

Proof of Theorem 3.1. The proof consists of the following four steps.

Step 1. Linearization: $\sqrt{N}(\hat{\delta} - \tilde{\delta}) = o_p(1)$.

Step 2. Asymptotic normality: $\sqrt{N}[\bar{\delta} - E(\bar{\delta})]$ has a limiting normal distribution with mean 0 and variance Σ.

Step 3. Asymptotic bias: $\sqrt{N}[E(\bar{\delta}) - \delta] = o(1)$.

Step 4. Trimming: $\sqrt{N}(\hat{\delta} - \delta)$ has the same limiting distribution as $\sqrt{N}(\tilde{\delta} - \delta)$.

The combination of Steps 1–4 yields Theorem 3.1.

Step 1: Linearization. Some arithmetic gives

$$\sqrt{N}(\hat{\delta} - \tilde{\delta}) = N^{-1/2}\sum_i \frac{[f(x_i) - \hat{f}_h(x_i)][\hat{f}'_h(x_i) - f'(x_i)]}{\hat{f}_h(x_i)f(x_i)}\,y_iI_i$$
$$- N^{-1/2}\sum_i \frac{[f(x_i) - \hat{f}_h(x_i)]^2}{\hat{f}_h(x_i)f(x_i)}\,l(x_i)y_iI_i,$$

so by (A.1a), there is a constant c_f such that with high probability

$$\sqrt{N}(\hat{\delta} - \tilde{\delta})$$
$$\leq \frac{\sqrt{N}}{b^2 - bc_f(N^{1-(\varepsilon/2)}h^k)^{-1/2}}\sup_x[|f - \hat{f}_h|I]$$
$$\times \sup_x[|\hat{f}'_h - f'|I]\,\frac{\sum |y_i|I_i}{N}$$
$$+ \frac{\sqrt{N}}{b^2 - bc_f(N^{1-(\varepsilon/2)}h^k)^{-1/2}}\sup_x[|f - \hat{f}_h|I]^2\,\frac{\sum |l(x_i)y_i|I_i}{N}.$$

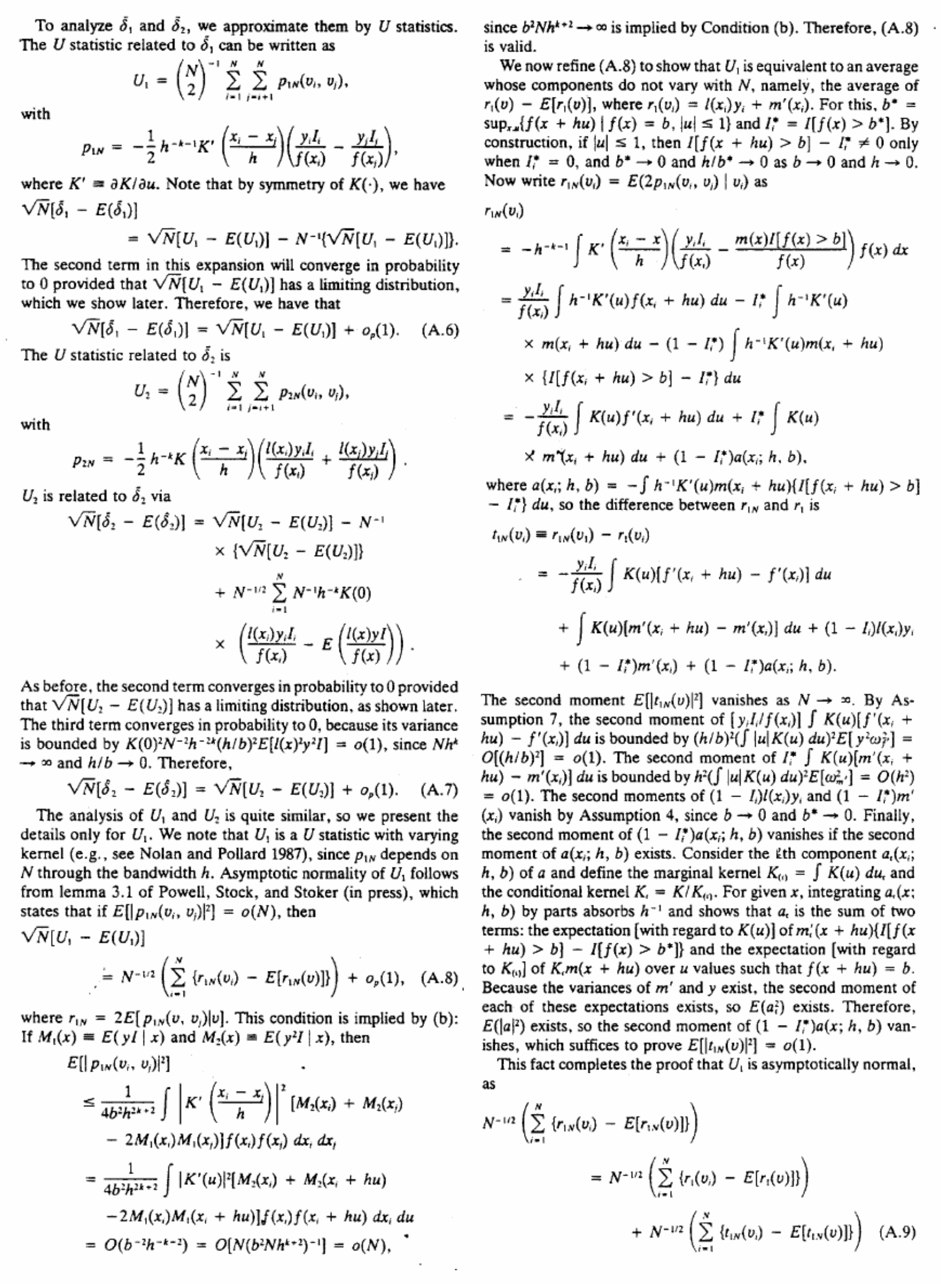
The terms $N^{-1}\sum |y_i|I_i$ and $N^{-1}\sum |l(x_i)y_i|I_i$ are bounded in probability by Chebyshev's inequality. Consequently, from (A.1a,b) we have that $\sqrt{N}(\hat{\delta} - \tilde{\delta}) = O_p(b^{-2}N^{-(1/2)+(\varepsilon/2)}h^{-(2k+2)/2}) = o_p(1)$, since $b^2N^{1-(\varepsilon/2)}h^k \to \infty$ and $b^4N^{1-\varepsilon}h^{2k+2} \to \infty$ by Condition (b).

Step 2: Asymptotic Normality. We show that $\sqrt{N}[\bar{\delta} - E(\bar{\delta})]$ has a limiting normal distribution by showing that $\bar{\delta}_0$, $\bar{\delta}_1$, and $\bar{\delta}_2$ are $\sqrt{N}$ equivalent to (ordinary) sample averages and then appealing to standard central limit theory. Throughout this section, $\upsilon_i = (y_i, x_i)$. For $\bar{\delta}_0$, we have that

$$\sqrt{N}[\bar{\delta}_0 - E(\bar{\delta}_0)] = N^{-1/2}\left(\sum_{i=1}^{N}\{r_0(\upsilon_i) - E[r_0(\upsilon)]\}\right) + o_p(1), \qquad \text{(A.5)}$$

where $r_0(\upsilon) = l(x)y$, since var(ly) exists and $b \to 0$ as $N \to \infty$.

Härdle, W. and Stoker, T. (1989) Investigating Smooth Multiple Regression by the Method of Average Derivatives.



To analyze $\bar{\delta}_1$ and $\bar{\delta}_2$, we approximate them by U statistics. The U statistic related to $\bar{\delta}_1$ can be written as

$$U_1 = \binom{N}{2}^{-1} \sum_{i=1}^{N} \sum_{j=i+1}^{N} p_{1N}(v_i, v_j),$$

with

$$p_{1N} = -\frac{1}{2} h^{-k-1} K' \left(\frac{x_i - x_j}{h}\right)\left(\frac{y_i I_i}{f(x_i)} - \frac{y_j I_j}{f(x_j)}\right),$$

where $K' \equiv \partial K / \partial u$. Note that by symmetry of $K(\cdot)$, we have

$$\sqrt{N}[\bar{\delta}_1 - E(\bar{\delta}_1)]$$
$$= \sqrt{N}[U_1 - E(U_1)] - N^{-1}\{\sqrt{N}[U_1 - E(U_1)]\}.$$

The second term in this expansion will converge in probability to 0 provided that $\sqrt{N}[U_1 - E(U_1)]$ has a limiting distribution, which we show later. Therefore, we have that

$$\sqrt{N}[\bar{\delta}_1 - E(\bar{\delta}_1)] = \sqrt{N}[U_1 - E(U_1)] + o_p(1). \quad \text{(A.6)}$$

The U statistic related to $\bar{\delta}_2$ is

$$U_2 = \binom{N}{2}^{-1} \sum_{i=1}^{N} \sum_{j=i+1}^{N} p_{2N}(v_i, v_j),$$

with

$$p_{2N} = -\frac{1}{2} h^{-k} K \left(\frac{x_i - x_j}{h}\right)\left(\frac{l(x_i) y_i I_i}{f(x_i)} + \frac{l(x_j) y_j I_j}{f(x_j)}\right).$$

U_2 is related to $\bar{\delta}_2$ via

$$\begin{aligned}\sqrt{N}[\bar{\delta}_2 - E(\hat{\delta}_2)] &= \sqrt{N}[U_2 - E(U_2)] - N^{-1} \\ &\quad \times \{\sqrt{N}[U_2 - E(U_2)]\} \\ &\quad + N^{-1/2} \sum_{i=1}^{N} N^{-1} h^{-k} K(0) \\ &\quad \times \left(\frac{l(x_i) y_i I_i}{f(x_i)} - E\left(\frac{l(x) y I}{f(x)}\right)\right).\end{aligned}$$

As before, the second term converges in probability to 0 provided that $\sqrt{N}[U_2 - E(U_2)]$ has a limiting distribution, as shown later. The third term converges in probability to 0, because its variance is bounded by $K(0)^2 N^{-2} h^{-2k} (h/b)^2 E[l(x)^2 y^2 I] = o(1)$, since $Nh^k \to \infty$ and $h/b \to 0$. Therefore,

$$\sqrt{N}[\hat{\delta}_2 - E(\hat{\delta}_2)] = \sqrt{N}[U_2 - E(U_2)] + o_p(1). \quad \text{(A.7)}$$

The analysis of U_1 and U_2 is quite similar, so we present the details only for U_1. We note that U_1 is a U statistic with varying kernel (e.g., see Nolan and Pollard 1987), since p_{1N} depends on N through the bandwidth h. Asymptotic normality of U_1 follows from lemma 3.1 of Powell, Stock, and Stoker (in press), which states that if $E[|p_{1N}(v_i, v_j)|^2] = o(N)$, then

$$\sqrt{N}[U_1 - E(U_1)]$$
$$= N^{-1/2}\left(\sum_{i=1}^{N} \{r_{1N}(v_i) - E[r_{1N}(v)]\}\right) + o_p(1), \quad \text{(A.8)}$$

where $r_{1N} = 2E[p_{1N}(v, v_j)|v]$. This condition is implied by (b): If $M_1(x) \equiv E(yI \mid x)$ and $M_2(x) \equiv E(y^2 I \mid x)$, then

$$\begin{aligned} &E[|p_{1N}(v_i, v_j)|^2] \\ &\le \frac{1}{4b^2 h^{2k+2}} \int \left| K'\left(\frac{x_i - x_j}{h}\right)\right|^2 [M_2(x_i) + M_2(x_j) \\ &\quad - 2M_1(x_i)M_1(x_j)] f(x_i) f(x_j)\, dx_i\, dx_j \\ &= \frac{1}{4b^2 h^{2k+2}} \int |K'(u)|^2 [M_2(x_i) + M_2(x_i + hu) \\ &\quad - 2M_1(x_i)M_1(x_i + hu)] f(x_i) f(x_i + hu)\, dx_i\, du \\ &= O(b^{-2} h^{-k-2}) = O[N(b^2 N h^{k+2})^{-1}] = o(N),\end{aligned}$$

since $b^2 N h^{k+2} \to \infty$ is implied by Condition (b). Therefore, (A.8) is valid.

We now refine (A.8) to show that U_1 is equivalent to an average whose components do not vary with N, namely, the average of $r_1(v) - E[r_1(v)]$, where $r_1(v_i) = l(x_i) y_i + m'(x_i)$. For this, $b^* = \sup_{x,u}\{f(x + hu) \mid f(x) = b, |u| \le 1\}$ and $I_i^* = I[f(x) > b^*]$. By construction, if $|u| \le 1$, then $I[f(x + hu) > b] - I_i^* \ne 0$ only when $I_i^* = 0$, and $b^* \to 0$ and $h/b^* \to 0$ as $b \to 0$ and $h \to 0$. Now write $r_{1N}(v_i) = E(2p_{1N}(v_i, v_j) \mid v_i)$ as

$$\begin{aligned} &r_{1N}(v_i) \\ &= -h^{-k-1} \int K'\left(\frac{x_i - x}{h}\right)\left(\frac{y_i I_i}{f(x_i)} - \frac{m(x) I[f(x) > b]}{f(x)}\right) f(x)\, dx \\ &= \frac{y_i I_i}{f(x_i)} \int h^{-1} K'(u) f(x_i + hu)\, du - I_i^* \int h^{-1} K'(u) \\ &\quad \times m(x_i + hu)\, du - (1 - I_i^*) \int h^{-1} K'(u) m(x_i + hu) \\ &\quad \times \{I[f(x_i + hu) > b] - I_i^*\}\, du \\ &= -\frac{y_i I_i}{f(x_i)} \int K(u) f'(x_i + hu)\, du + I_i^* \int K(u) \\ &\quad \times m'(x_i + hu)\, du + (1 - I_i^*) a(x_i; h, b),\end{aligned}$$

where $a(x_i; h, b) = -\int h^{-1} K'(u) m(x_i + hu)\{I[f(x_i + hu) > b] - I_i^*\}\, du$, so the difference between r_{1N} and r_1 is

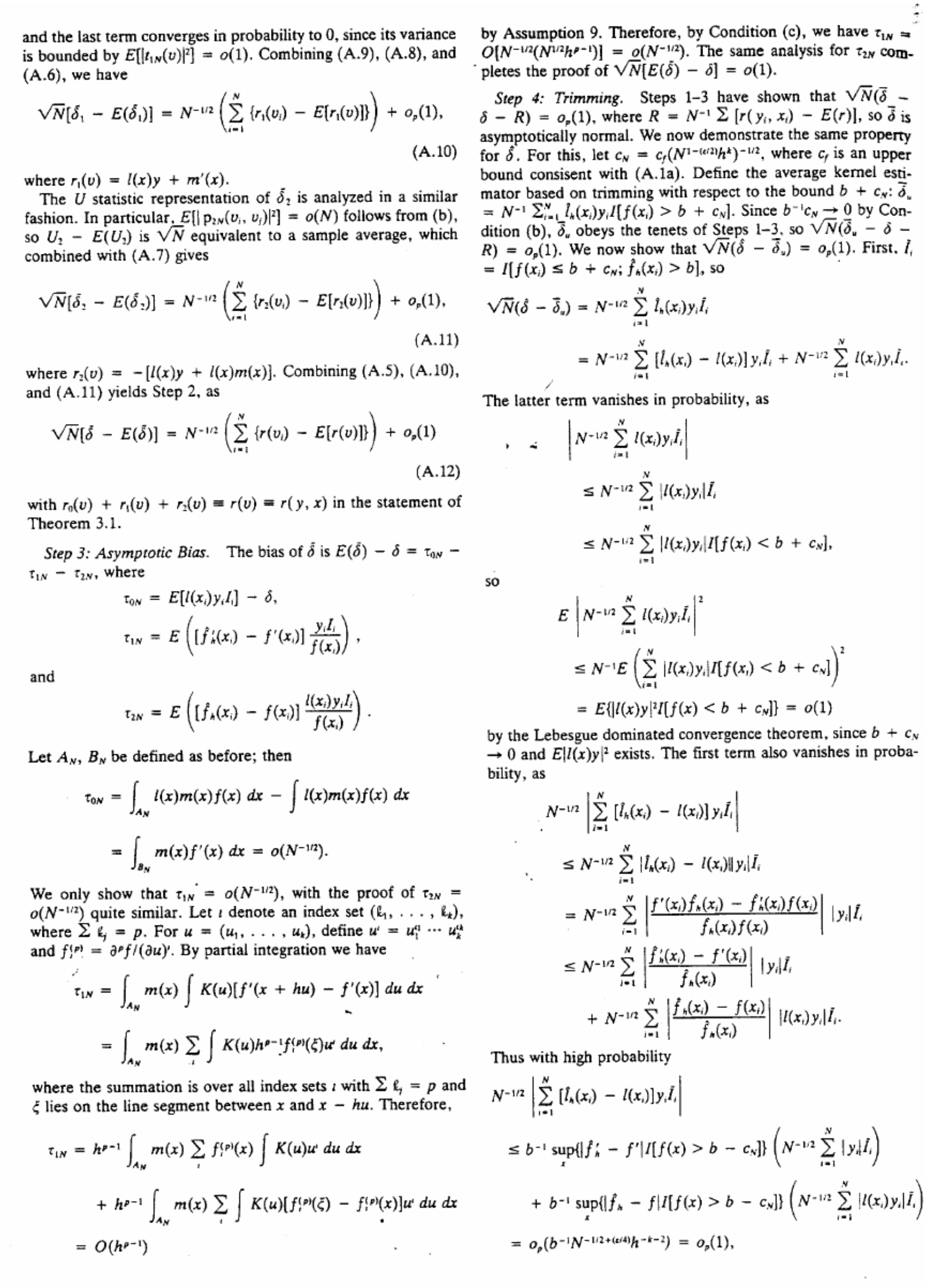
$$\begin{aligned} t_{1N}(v_i) &\equiv r_{1N}(v_1) - r_1(v_i) \\ &= -\frac{y_i I_i}{f(x_i)} \int K(u)[f'(x_i + hu) - f'(x_i)]\, du \\ &\quad + \int K(u)[m'(x_i + hu) - m'(x_i)]\, du + (1 - I_i) l(x_i) y_i \\ &\quad + (1 - I_i^*) m'(x_i) + (1 - I_i^*) a(x_i; h, b).\end{aligned}$$

The second moment $E[|t_{1N}(v)|^2]$ vanishes as $N \to \infty$. By Assumption 7, the second moment of $[y_i I_i / f(x_i)] \int K(u)[f'(x_i + hu) - f'(x_i)]\, du$ is bounded by $(h/b)^2 (\int |u| K(u)\, du)^2 E[y^2 \omega_{f'}^2] = O[(h/b)^2] = o(1)$. The second moment of $I_i^* \int K(u)[m'(x_i + hu) - m'(x_i)]\, du$ is bounded by $h^2 (\int |u| K(u)\, du)^2 E[\omega_{m'}^2] = O(h^2) = o(1)$. The second moments of $(1 - I_i) l(x_i) y_i$ and $(1 - I_i^*) m'(x_i)$ vanish by Assumption 4, since $b \to 0$ and $b^* \to 0$. Finally, the second moment of $(1 - I_i^*) a(x_i; h, b)$ vanishes if the second moment of $a(x_i; h, b)$ exists. Consider the ℓth component $a_\ell(x_i; h, b)$ of a and define the marginal kernel $K_{(\ell)} = \int K(u)\, du_\ell$ and the conditional kernel $K_\ell = K / K_{(\ell)}$. For given x, integrating $a_\ell(x; h, b)$ by parts absorbs h^{-1} and shows that a_ℓ is the sum of two terms: the expectation [with regard to $K(u)$] of $m_\ell'(x + hu)\{I[f(x + hu) > b] - I[f(x) > b^*]\}$ and the expectation [with regard to $K_{(\ell)}$] of $K_\ell m(x + hu)$ over u values such that $f(x + hu) = b$. Because the variances of m' and y exist, the second moment of each of these expectations exists, so $E(a_\ell^2)$ exists. Therefore, $E(|a|^2)$ exists, so the second moment of $(1 - I_i^*) a(x; h, b)$ vanishes, which suffices to prove $E[|t_{1N}(v)|^2] = o(1)$.

This fact completes the proof that U_1 is asymptotically normal, as

$$\begin{aligned} &N^{-1/2}\left(\sum_{i=1}^{N} \{r_{1N}(v_i) - E[r_{1N}(v)]\}\right) \\ &\qquad = N^{-1/2}\left(\sum_{i=1}^{N} \{r_1(v_i) - E[r_1(v)]\}\right) \\ &\qquad\quad + N^{-1/2}\left(\sum_{i=1}^{N} \{t_{1N}(v_i) - E[t_{1N}(v)]\}\right) \quad \text{(A.9)}\end{aligned}$$

Härdle, W. and Stoker, T. (1989) Investigating Smooth Multiple Regression by the Method of Average Derivatives.



and the last term converges in probability to 0, since its variance is bounded by $E[|t_{1N}(v)|^2] = o(1)$. Combining (A.9), (A.8), and (A.6), we have

$$\sqrt{N}[\tilde{\delta}_1 - E(\tilde{\delta}_1)] = N^{-1/2}\left(\sum_{i=1}^{N}\{r_1(v_i) - E[r_1(v)]\}\right) + o_p(1), \tag{A.10}$$

where $r_1(v) = l(x)y + m'(x)$.

The U statistic representation of $\tilde{\delta}_2$ is analyzed in a similar fashion. In particular, $E[|\,p_{2N}(v_i, v_j)|^2] = o(N)$ follows from (b), so $U_2 - E(U_2)$ is $\sqrt{N}$ equivalent to a sample average, which combined with (A.7) gives

$$\sqrt{N}[\tilde{\delta}_2 - E(\tilde{\delta}_2)] = N^{-1/2}\left(\sum_{i=1}^{N}\{r_2(v_i) - E[r_2(v)]\}\right) + o_p(1), \tag{A.11}$$

where $r_2(v) = -[l(x)y + l(x)m(x)]$. Combining (A.5), (A.10), and (A.11) yields Step 2, as

$$\sqrt{N}[\tilde{\delta} - E(\tilde{\delta})] = N^{-1/2}\left(\sum_{i=1}^{N}\{r(v_i) - E[r(v)]\}\right) + o_p(1) \tag{A.12}$$

with $r_0(v) + r_1(v) + r_2(v) \equiv r(v) \equiv r(y, x)$ in the statement of Theorem 3.1.

Step 3: Asymptotic Bias. The bias of $\tilde{\delta}$ is $E(\tilde{\delta}) - \delta = \tau_{0N} - \tau_{1N} - \tau_{2N}$, where

$$\tau_{0N} = E[l(x_i)y_iI_i] - \delta,$$

$$\tau_{1N} = E\left([\hat{f}_h'(x_i) - f'(x_i)]\,\frac{y_iI_i}{f(x_i)}\right),$$

and

$$\tau_{2N} = E\left([\hat{f}_h(x_i) - f(x_i)]\,\frac{l(x_i)y_iI_i}{f(x_i)}\right).$$

Let A_N, B_N be defined as before; then

$$\tau_{0N} = \int_{A_N} l(x)m(x)f(x)\,dx - \int l(x)m(x)f(x)\,dx$$

$$= \int_{B_N} m(x)f'(x)\,dx = o(N^{-1/2}).$$

We only show that $\tau_{1N} = o(N^{-1/2})$, with the proof of $\tau_{2N} = o(N^{-1/2})$ quite similar. Let ι denote an index set $(\ell_1, \ldots, \ell_k)$, where $\sum \ell_j = p$. For $u = (u_1, \ldots, u_k)$, define $u^\iota = u_1^{\ell_1} \cdots u_k^{\ell_k}$ and $f_\iota^{(p)} = \partial^p f/(\partial u)^\iota$. By partial integration we have

$$\tau_{1N} = \int_{A_N} m(x) \int K(u)[f'(x + hu) - f'(x)]\,du\,dx$$

$$= \int_{A_N} m(x) \sum_\iota \int K(u)h^{p-1}f_\iota^{(p)}(\xi)u^\iota\,du\,dx,$$

where the summation is over all index sets ι with $\sum \ell_j = p$ and ξ lies on the line segment between x and $x - hu$. Therefore,

$$\tau_{1N} = h^{p-1}\int_{A_N} m(x) \sum_\iota f_\iota^{(p)}(x) \int K(u)u^\iota\,du\,dx$$

$$+ h^{p-1}\int_{A_N} m(x) \sum_\iota \int K(u)[f_\iota^{(p)}(\xi) - f_\iota^{(p)}(x)]u^\iota\,du\,dx$$

$$= O(h^{p-1})$$

by Assumption 9. Therefore, by Condition (c), we have $\tau_{1N} = O[N^{-1/2}(N^{1/2}h^{p-1})] = o(N^{-1/2})$. The same analysis for τ_{2N} completes the proof of $\sqrt{N}[E(\tilde{\delta}) - \delta] = o(1)$.

Step 4: Trimming. Steps 1–3 have shown that $\sqrt{N}(\tilde{\delta} - \delta - R) = o_p(1)$, where $R = N^{-1}\sum[r(y_i, x_i) - E(r)]$, so $\tilde{\delta}$ is asymptotically normal. We now demonstrate the same property for $\hat{\delta}$. For this, let $c_N = c_f(N^{1-(\varepsilon/2)}h^k)^{-1/2}$, where c_f is an upper bound consisent with (A.1a). Define the average kernel estimator based on trimming with respect to the bound $b + c_N$: $\bar{\delta}_u = N^{-1}\sum_{i=1}^N \hat{l}_h(x_i)y_iI[f(x_i) > b + c_N]$. Since $b^{-1}c_N \to 0$ by Condition (b), $\bar{\delta}_u$ obeys the tenets of Steps 1–3, so $\sqrt{N}(\bar{\delta}_u - \delta - R) = o_p(1)$. We now show that $\sqrt{N}(\hat{\delta} - \bar{\delta}_u) = o_p(1)$. First, $\bar{I}_i = I[f(x_i) \le b + c_N;\ \hat{f}_h(x_i) > b]$, so

$$\sqrt{N}(\hat{\delta} - \bar{\delta}_u) = N^{-1/2}\sum_{i=1}^{N} \hat{l}_h(x_i)y_i\bar{I}_i$$

$$= N^{-1/2}\sum_{i=1}^{N}[\hat{l}_h(x_i) - l(x_i)]\,y_i\bar{I}_i + N^{-1/2}\sum_{i=1}^{N} l(x_i)y_i\bar{I}_i.$$

The latter term vanishes in probability, as

$$\left|N^{-1/2}\sum_{i=1}^{N} l(x_i)y_i\bar{I}_i\right|$$

$$\le N^{-1/2}\sum_{i=1}^{N}|l(x_i)y_i|\bar{I}_i$$

$$\le N^{-1/2}\sum_{i=1}^{N}|l(x_i)y_i|I[f(x_i) < b + c_N],$$

so

$$E\left|N^{-1/2}\sum_{i=1}^{N} l(x_i)y_i\bar{I}_i\right|^2$$

$$\le N^{-1}E\left(\sum_{i=1}^{N}|l(x_i)y_i|I[f(x_i) < b + c_N]\right)^2$$

$$= E\{|l(x)y|^2I[f(x) < b + c_N]\} = o(1)$$

by the Lebesgue dominated convergence theorem, since $b + c_N \to 0$ and $E|l(x)y|^2$ exists. The first term also vanishes in probability, as

$$N^{-1/2}\left|\sum_{i=1}^{N}[\hat{l}_h(x_i) - l(x_i)]\,y_i\bar{I}_i\right|$$

$$\le N^{-1/2}\sum_{i=1}^{N}|\hat{l}_h(x_i) - l(x_i)||y_i|\bar{I}_i$$

$$= N^{-1/2}\sum_{i=1}^{N}\left|\frac{f'(x_i)f_h(x_i) - f_h'(x_i)f(x_i)}{f_h(x_i)f(x_i)}\right|\,|y_i|\bar{I}_i$$

$$\le N^{-1/2}\sum_{i=1}^{N}\left|\frac{\hat{f}_h'(x_i) - f'(x_i)}{\hat{f}_h(x_i)}\right|\,|y_i|\bar{I}_i$$

$$+ N^{-1/2}\sum_{i=1}^{N}\left|\frac{\hat{f}_h(x_i) - f(x_i)}{\hat{f}_h(x_i)}\right|\,|l(x_i)y_i|\bar{I}_i.$$

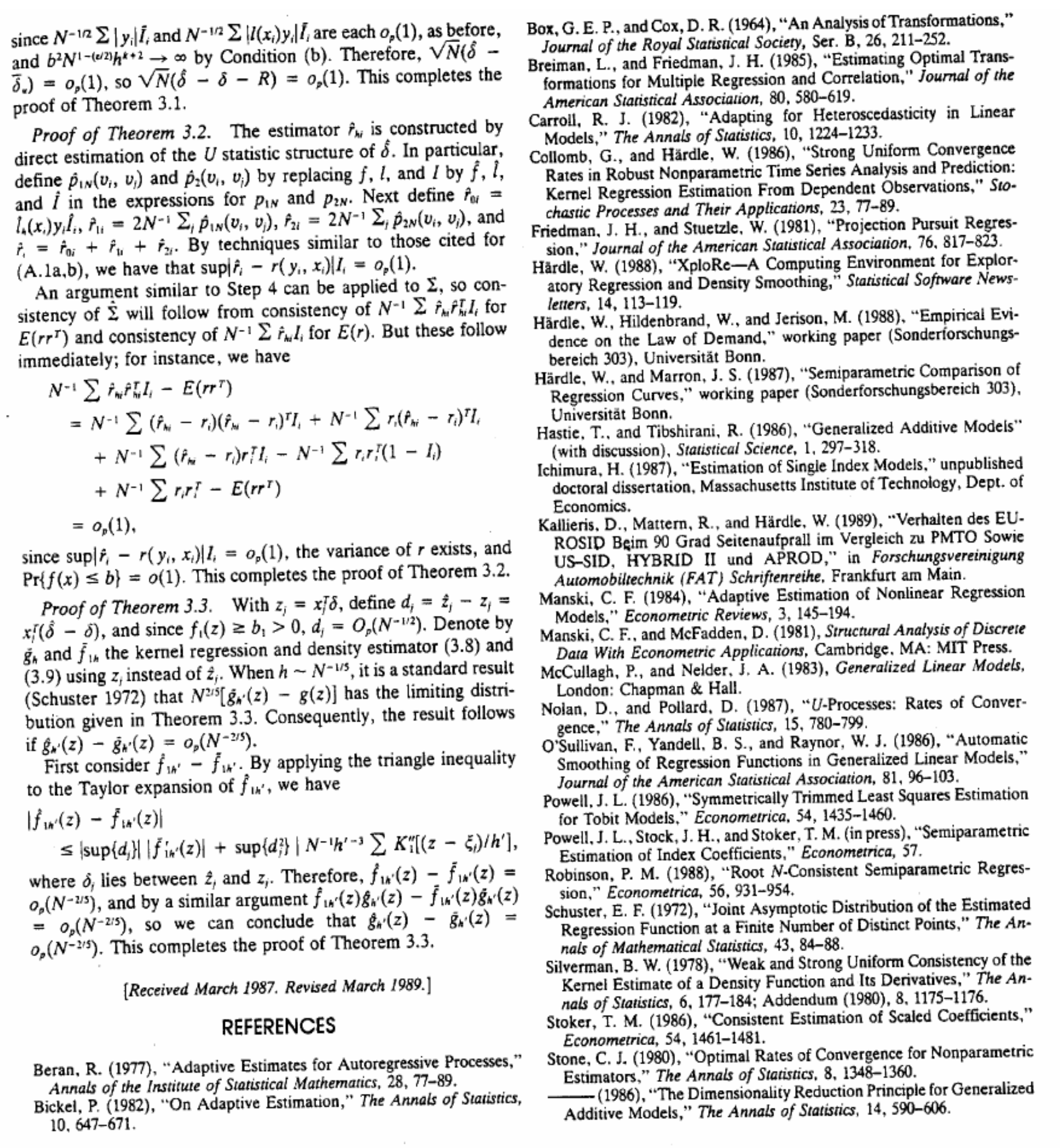
Thus with high probability

$$N^{-1/2}\left|\sum_{i=1}^{N}[\hat{l}_h(x_i) - l(x_i)]y_i\bar{I}_i\right|$$

$$\le b^{-1}\sup_x\{|\hat{f}_h' - f'|I[f(x) > b - c_N]\}\left(N^{-1/2}\sum_{i=1}^{N}|y_i|\bar{I}_i\right)$$

$$+ b^{-1}\sup_x\{|\hat{f}_h - f|I[f(x) > b - c_N]\}\left(N^{-1/2}\sum_{i=1}^{N}|l(x_i)y_i|\bar{I}_i\right)$$

$$= o_p(b^{-1}N^{-1/2+(\varepsilon/4)}h^{-k-2}) = o_p(1),$$



since $N^{-1/2}\sum|y_i|\bar{I}_i$ and $N^{-1/2}\sum|l(x_i)y_i|\bar{I}_i$ are each $o_p(1)$, as before, and $b^2N^{1-(\varepsilon/2)}h^{k+2} \to \infty$ by Condition (b). Therefore, $\sqrt{N}(\hat{\delta} - \bar{\delta}_*) = o_p(1)$, so $\sqrt{N}(\hat{\delta} - \delta - R) = o_p(1)$. This completes the proof of Theorem 3.1.

Proof of Theorem 3.2. The estimator $\hat{r}_{hi}$ is constructed by direct estimation of the U statistic structure of $\hat{\delta}$. In particular, define $\hat{p}_{1N}(v_i, v_j)$ and $\hat{p}_2(v_i, v_j)$ by replacing f, l, and I by $\hat{f}$, $\hat{l}$, and $\hat{I}$ in the expressions for p_{1N} and p_{2N}. Next define $\hat{r}_{0i} = \hat{l}_h(x_i)y_i\hat{I}_i$, $\hat{r}_{1i} = 2N^{-1}\sum_j \hat{p}_{1N}(v_i, v_j)$, $\hat{r}_{2i} = 2N^{-1}\sum_j \hat{p}_{2N}(v_i, v_j)$, and $\hat{r}_i = \hat{r}_{0i} + \hat{r}_{1i} + \hat{r}_{2i}$. By techniques similar to those cited for (A.1a,b), we have that $\sup|\hat{r}_i - r(y_i, x_i)|I_i = o_p(1)$.

An argument similar to Step 4 can be applied to Σ, so consistency of $\hat{\Sigma}$ will follow from consistency of $N^{-1}\sum \hat{r}_{hi}\hat{r}_{hi}^T I_i$ for $E(rr^T)$ and consistency of $N^{-1}\sum \hat{r}_{hi}I_i$ for $E(r)$. But these follow immediately; for instance, we have

$$
\begin{aligned}
&N^{-1}\sum \hat{r}_{hi}\hat{r}_{hi}^T I_i - E(rr^T)\\
&\quad = N^{-1}\sum(\hat{r}_{hi} - r_i)(\hat{r}_{hi} - r_i)^T I_i + N^{-1}\sum r_i(\hat{r}_{hi} - r_i)^T I_i\\
&\qquad + N^{-1}\sum(\hat{r}_{hi} - r_i)r_i^T I_i - N^{-1}\sum r_i r_i^T(1 - I_i)\\
&\qquad + N^{-1}\sum r_i r_i^T - E(rr^T)\\
&\quad = o_p(1),
\end{aligned}
$$

since $\sup|\hat{r}_i - r(y_i, x_i)|I_i = o_p(1)$, the variance of r exists, and $\Pr\{f(x) \le b\} = o(1)$. This completes the proof of Theorem 3.2.

Proof of Theorem 3.3. With $z_j = x_j^T\delta$, define $d_j = \hat{z}_j - z_j = x_j^T(\hat{\delta} - \delta)$, and since $f_1(z) \ge b_1 > 0$, $d_j = O_p(N^{-1/2})$. Denote by $\bar{g}_h$ and $\bar{f}_{1h}$ the kernel regression and density estimator (3.8) and (3.9) using z_j instead of $\hat{z}_j$. When $h \sim N^{-1/5}$, it is a standard result (Schuster 1972) that $N^{2/5}[\bar{g}_{h'}(z) - g(z)]$ has the limiting distribution given in Theorem 3.3. Consequently, the result follows if $\hat{g}_{h'}(z) - \bar{g}_{h'}(z) = o_p(N^{-2/5})$.

First consider $\hat{f}_{1h'} - \bar{f}_{1h'}$. By applying the triangle inequality to the Taylor expansion of $\hat{f}_{1h'}$, we have

$$
\begin{aligned}
&|\hat{f}_{1h'}(z) - \bar{f}_{1h'}(z)|\\
&\qquad \le |\sup\{d_j\}|\,|\bar{f}'_{1h'}(z)| + \sup\{d_j^2\}\,|\,N^{-1}h'^{-3}\sum K_1''[(z - \xi_j)/h'],
\end{aligned}
$$

where δ_j lies between $\hat{z}_j$ and z_j. Therefore, $\hat{f}_{1h'}(z) - \bar{f}_{1h'}(z) = o_p(N^{-2/5})$, and by a similar argument $\hat{f}_{1h'}(z)\hat{g}_{h'}(z) - \bar{f}_{1h'}(z)\bar{g}_{h'}(z) = o_p(N^{-2/5})$, so we can conclude that $\hat{g}_{h'}(z) - \bar{g}_{h'}(z) = o_p(N^{-2/5})$. This completes the proof of Theorem 3.3.

[*Received March 1987. Revised March 1989.*]

REFERENCES

Beran, R. (1977), "Adaptive Estimates for Autoregressive Processes," *Annals of the Institute of Statistical Mathematics*, 28, 77–89.

Bickel, P. (1982), "On Adaptive Estimation," *The Annals of Statistics*, 10, 647–671.

Box, G. E. P., and Cox, D. R. (1964), "An Analysis of Transformations," *Journal of the Royal Statistical Society*, Ser. B, 26, 211–252.

Breiman, L., and Friedman, J. H. (1985), "Estimating Optimal Transformations for Multiple Regression and Correlation," *Journal of the American Statistical Association*, 80, 580–619.

Carroll, R. J. (1982), "Adapting for Heteroscedasticity in Linear Models," *The Annals of Statistics*, 10, 1224–1233.

Collomb, G., and Härdle, W. (1986), "Strong Uniform Convergence Rates in Robust Nonparametric Time Series Analysis and Prediction: Kernel Regression Estimation From Dependent Observations," *Stochastic Processes and Their Applications*, 23, 77–89.

Friedman, J. H., and Stuetzle, W. (1981), "Projection Pursuit Regression," *Journal of the American Statistical Association*, 76, 817–823.

Härdle, W. (1988), "XploRe—A Computing Environment for Exploratory Regression and Density Smoothing," *Statistical Software Newsletters*, 14, 113–119.

Härdle, W., Hildenbrand, W., and Jerison, M. (1988), "Empirical Evidence on the Law of Demand," working paper (Sonderforschungsbereich 303), Universität Bonn.

Härdle, W., and Marron, J. S. (1987), "Semiparametric Comparison of Regression Curves," working paper (Sonderforschungsbereich 303), Universität Bonn.

Hastie, T., and Tibshirani, R. (1986), "Generalized Additive Models" (with discussion), *Statistical Science*, 1, 297–318.

Ichimura, H. (1987), "Estimation of Single Index Models," unpublished doctoral dissertation, Massachusetts Institute of Technology, Dept. of Economics.

Kallieris, D., Mattern, R., and Härdle, W. (1989), "Verhalten des EUROSID Bęim 90 Grad Seitenaufprall im Vergleich zu PMTO Sowie US–SID, HYBRID II und APROD," in *Forschungsvereinigung Automobiltechnik (FAT) Schriftenreihe*, Frankfurt am Main.

Manski, C. F. (1984), "Adaptive Estimation of Nonlinear Regression Models," *Econometric Reviews*, 3, 145–194.

Manski, C. F., and McFadden, D. (1981), *Structural Analysis of Discrete Data With Econometric Applications*, Cambridge, MA: MIT Press.

McCullagh, P., and Nelder, J. A. (1983), *Generalized Linear Models*, London: Chapman & Hall.

Nolan, D., and Pollard, D. (1987), "*U*-Processes: Rates of Convergence," *The Annals of Statistics*, 15, 780–799.

O'Sullivan, F., Yandell, B. S., and Raynor, W. J. (1986), "Automatic Smoothing of Regression Functions in Generalized Linear Models," *Journal of the American Statistical Association*, 81, 96–103.

Powell, J. L. (1986), "Symmetrically Trimmed Least Squares Estimation for Tobit Models," *Econometrica*, 54, 1435–1460.

Powell, J. L., Stock, J. H., and Stoker, T. M. (in press), "Semiparametric Estimation of Index Coefficients," *Econometrica*, 57.

Robinson, P. M. (1988), "Root *N*-Consistent Semiparametric Regression," *Econometrica*, 56, 931–954.

Schuster, E. F. (1972), "Joint Asymptotic Distribution of the Estimated Regression Function at a Finite Number of Distinct Points," *The Annals of Mathematical Statistics*, 43, 84–88.

Silverman, B. W. (1978), "Weak and Strong Uniform Consistency of the Kernel Estimate of a Density Function and Its Derivatives," *The Annals of Statistics*, 6, 177–184; Addendum (1980), 8, 1175–1176.

Stoker, T. M. (1986), "Consistent Estimation of Scaled Coefficients," *Econometrica*, 54, 1461–1481.

Stone, C. J. (1980), "Optimal Rates of Convergence for Nonparametric Estimators," *The Annals of Statistics*, 8, 1348–1360.

——— (1986), "The Dimensionality Reduction Principle for Generalized Additive Models," *The Annals of Statistics*, 14, 590–606.

SEMIPARAMETRIC COMPARISON OF REGRESSION CURVES

BY W. HÄRDLE[1] AND J. S. MARRON[1,2]

Universität Bonn and Universität Bonn and University of North Carolina at Chapel Hill

The comparison of nonparametric regression curves is considered. It is assumed that there are parametric (possibly nonlinear) transformations of the axes which map one curve into the other. Estimation and testing of the parameters in the transformations are studied. The rate of convergence is $n^{-1/2}$ although the nonparametric components of the model typically have a rate slower than that. A statistic is provided for testing the validity of a given completely parametric model.

1. Introduction. An important case of regression analysis is the comparison of regression curves from related samples. Even when there is no reasonable parametric model for each regression curve a way of quantifying differences across individual curves is often desirable. A well-known example is the study of child growth curves, where individual curves certainly seem to require nonparametric estimation techniques [Gasser, Müller, Köhler, Molinari and Prader (1984)] but may have a simple relationship between them. Another example appears in Figures 1(a) and 1(b), which show acceleration data from a study on automobile side impacts [Kallieris, Mattern and Härdle (1986)].

The curves give the impression that they are noisy versions of similar regression curves, where the main difference is that the time axis is shifted and there is a vertical rescaling. A parametric model that could be deduced from a physical or biomechanical theory is not available here; see Eppinger, Marcus and Morgan (1984), so a nonparametric smoothing technique seems to be a reasonable way to estimate the acceleration curves for inference regarding this data set. The problem of comparison of the two curves could be modeled parametrically because, to a large extent, the difference between them seems to be quantified by two parameters, horizontal shift and vertical scale. Hence, a comparison of nonparametric regression curves in a parametric framework is desirable for studying data sets of this type.

The main objective of this paper is the analysis of general semiparametric models where nonparametric curves are related in a parametric way. The case that is treated in detail is where there are two curves which are the same up to a transformation of the horizontal axis and a transformation of the vertical axis, and these transformations are indexed by some parameters. The techniques of this paper are adaptable to other semiparametric models such as multiplicative

Received January 1987; revised March 1989.

[1] Research supported by Deutsche Forschungsgemeinschaft, SFB 303.

[2] Research partially supported by NSF Grants DMS-8400602 and DMS-8701201.

AMS 1980 *subject classifications*. Primary 62G05; secondary 62G99.

Key words and phrases. Semiparametric regression, nonparametric smoothing, parametric comparison, kernel estimators.

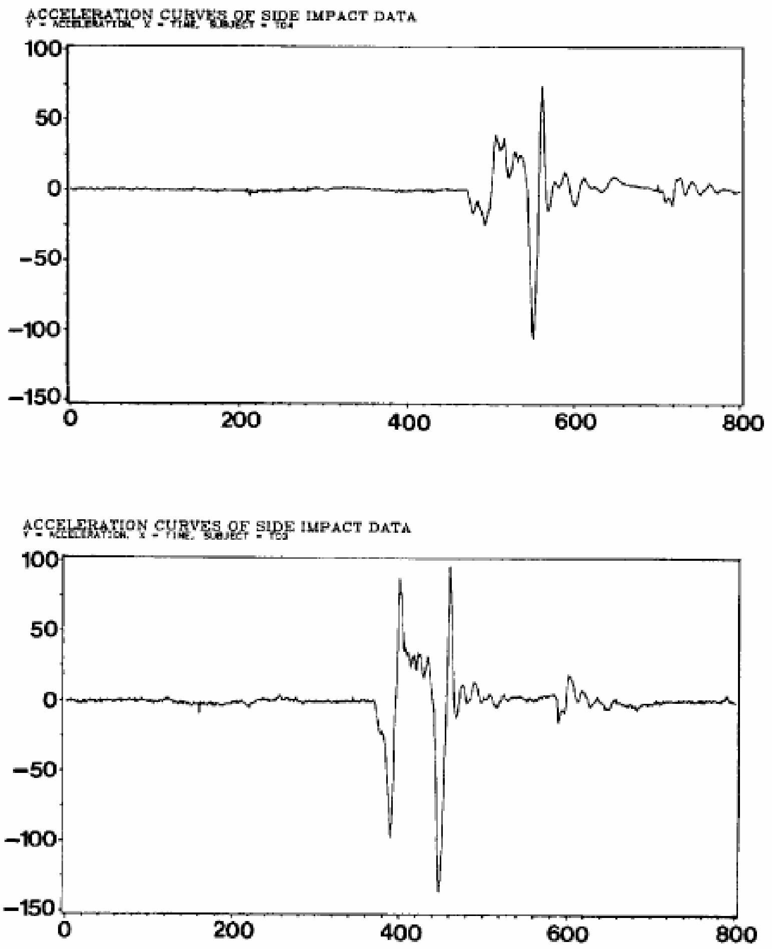
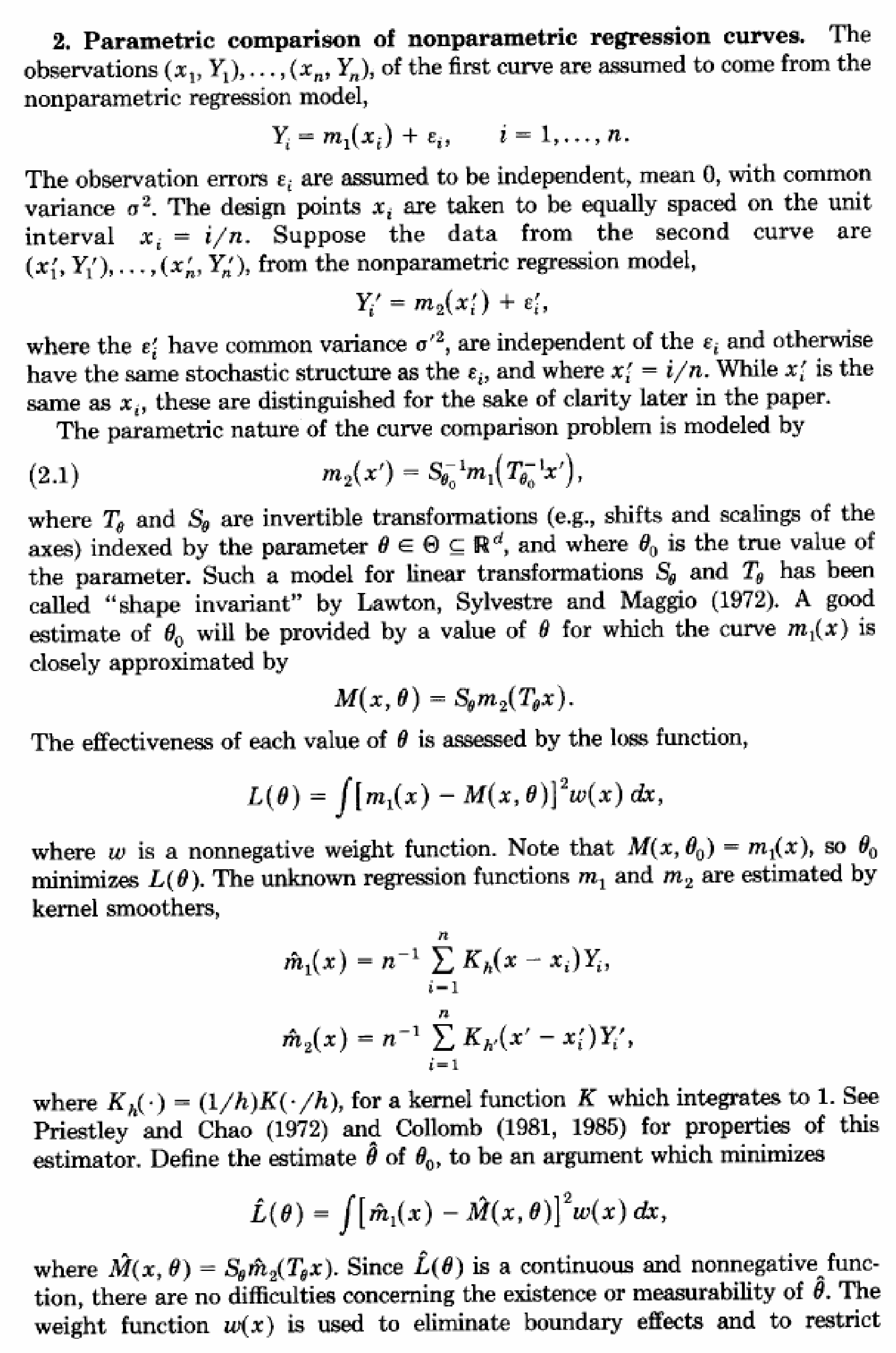


FIG. 1. *Two impact acceleration curves from Kallieris, Mattern and Härdle* (1986).

or additive combination of a nonparametric regression curve with a parametric "modulation" function. An additional benefit of the theory developed in this paper is that, with no extra work, a statistic is provided for testing the validity of a given completely parametric model. This test quantifies the idea of checking a parametric model by comparing the parametric fit to a nonparametric regression curve.

Section 2 contains a mathematical formulation of these ideas, together with a proposal for estimating the parameters. This parameter estimate is seen to be consistent under very mild conditions in Section 3. Asymptotic normality, with the rate of convergence typical to parametric problems, is established under somewhat stronger conditions in Section 4. Section 5 gives test statistics, together with their asymptotic null distributions, for testing whether some parameters can be eliminated from the model and also for testing whether a given semiparametric model is in fact appropriate.



2. Parametric comparison of nonparametric regression curves. The observations $(x_1, Y_1), \ldots, (x_n, Y_n)$, of the first curve are assumed to come from the nonparametric regression model,

$$Y_i = m_1(x_i) + \varepsilon_i, \qquad i = 1, \ldots, n.$$

The observation errors ε_i are assumed to be independent, mean 0, with common variance σ^2. The design points x_i are taken to be equally spaced on the unit interval $x_i = i/n$. Suppose the data from the second curve are $(x_1', Y_1'), \ldots, (x_n', Y_n')$, from the nonparametric regression model,

$$Y_i' = m_2(x_i') + \varepsilon_i',$$

where the ε_i' have common variance σ'^2, are independent of the ε_i and otherwise have the same stochastic structure as the ε_i, and where $x_i' = i/n$. While x_i' is the same as x_i, these are distinguished for the sake of clarity later in the paper.

The parametric nature of the curve comparison problem is modeled by

$$m_2(x') = S_{\theta_0}^{-1} m_1\left(T_{\theta_0}^{-1} x'\right), \tag{2.1}$$

where T_θ and S_θ are invertible transformations (e.g., shifts and scalings of the axes) indexed by the parameter $\theta \in \Theta \subseteq \mathbb{R}^d$, and where θ_0 is the true value of the parameter. Such a model for linear transformations S_θ and T_θ has been called "shape invariant" by Lawton, Sylvestre and Maggio (1972). A good estimate of θ_0 will be provided by a value of θ for which the curve $m_1(x)$ is closely approximated by

$$M(x, \theta) = S_\theta m_2(T_\theta x).$$

The effectiveness of each value of θ is assessed by the loss function,

$$L(\theta) = \int [m_1(x) - M(x, \theta)]^2 w(x)\, dx,$$

where w is a nonnegative weight function. Note that $M(x, \theta_0) = m_1(x)$, so θ_0 minimizes $L(\theta)$. The unknown regression functions m_1 and m_2 are estimated by kernel smoothers,

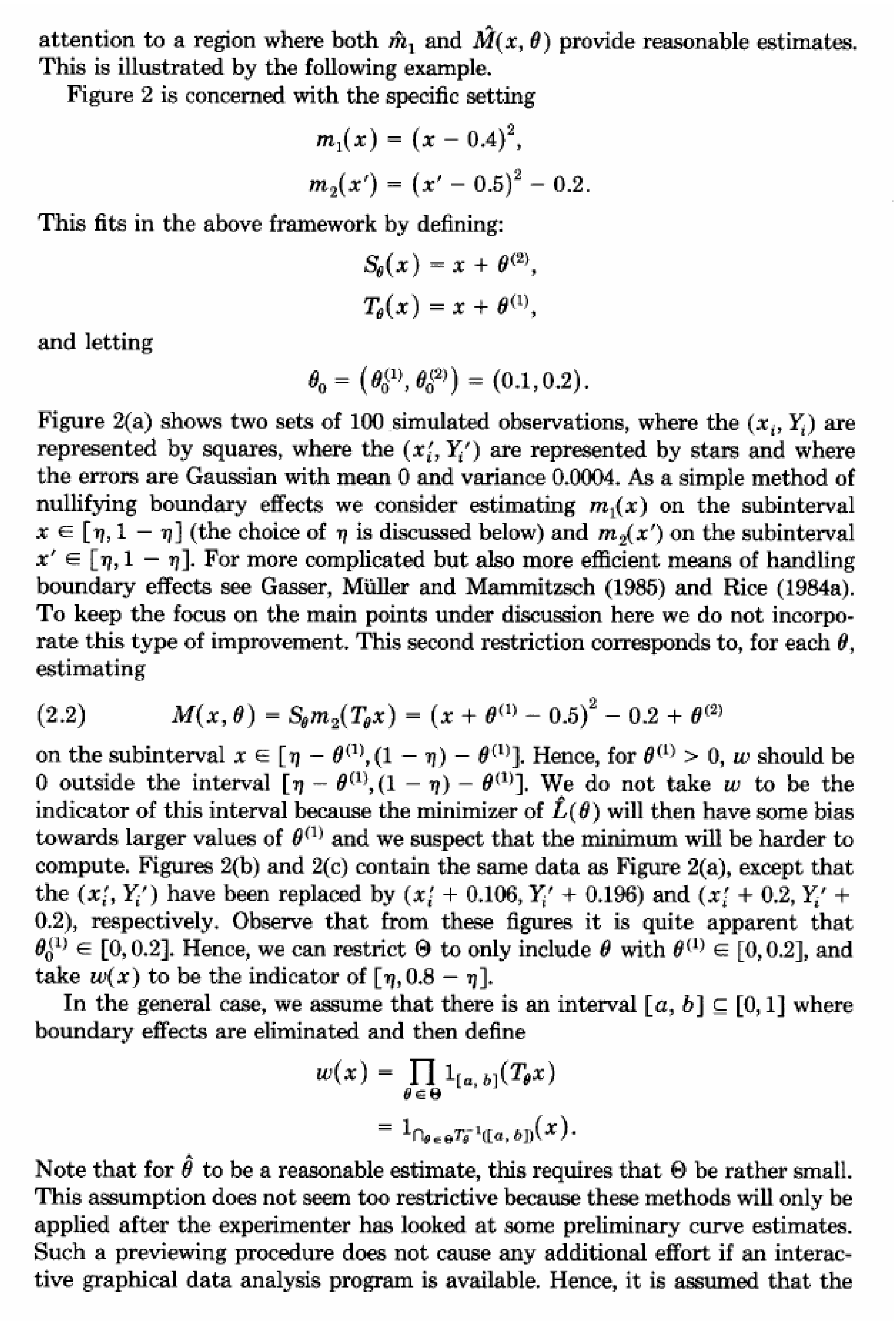
$$\hat{m}_1(x) = n^{-1} \sum_{i=1}^{n} K_h(x - x_i) Y_i,$$

$$\hat{m}_2(x) = n^{-1} \sum_{i=1}^{n} K_{h'}(x' - x_i') Y_i',$$

where $K_h(\cdot) = (1/h)K(\cdot/h)$, for a kernel function K which integrates to 1. See Priestley and Chao (1972) and Collomb (1981, 1985) for properties of this estimator. Define the estimate $\hat{\theta}$ of θ_0, to be an argument which minimizes

$$\hat{L}(\theta) = \int \left[\hat{m}_1(x) - \hat{M}(x, \theta)\right]^2 w(x)\, dx,$$

where $\hat{M}(x, \theta) = S_\theta \hat{m}_2(T_\theta x)$. Since $\hat{L}(\theta)$ is a continuous and nonnegative function, there are no difficulties concerning the existence or measurability of $\hat{\theta}$. The weight function $w(x)$ is used to eliminate boundary effects and to restrict



attention to a region where both $\hat{m}_1$ and $\hat{M}(x,\theta)$ provide reasonable estimates. This is illustrated by the following example.

Figure 2 is concerned with the specific setting

$$m_1(x) = (x - 0.4)^2,$$

$$m_2(x') = (x' - 0.5)^2 - 0.2.$$

This fits in the above framework by defining:

$$S_\theta(x) = x + \theta^{(2)},$$

$$T_\theta(x) = x + \theta^{(1)},$$

and letting

$$\theta_0 = \left(\theta_0^{(1)}, \theta_0^{(2)}\right) = (0.1, 0.2).$$

Figure 2(a) shows two sets of 100 simulated observations, where the (x_i, Y_i) are represented by squares, where the (x_i', Y_i') are represented by stars and where the errors are Gaussian with mean 0 and variance 0.0004. As a simple method of nullifying boundary effects we consider estimating $m_1(x)$ on the subinterval $x \in [\eta, 1 - \eta]$ (the choice of η is discussed below) and $m_2(x')$ on the subinterval $x' \in [\eta, 1 - \eta]$. For more complicated but also more efficient means of handling boundary effects see Gasser, Müller and Mammitzsch (1985) and Rice (1984a). To keep the focus on the main points under discussion here we do not incorporate this type of improvement. This second restriction corresponds to, for each θ, estimating

$$\text{(2.2)} \qquad M(x,\theta) = S_\theta m_2(T_\theta x) = \left(x + \theta^{(1)} - 0.5\right)^2 - 0.2 + \theta^{(2)}$$

on the subinterval $x \in [\eta - \theta^{(1)}, (1 - \eta) - \theta^{(1)}]$. Hence, for $\theta^{(1)} > 0$, w should be 0 outside the interval $[\eta - \theta^{(1)}, (1 - \eta) - \theta^{(1)}]$. We do not take w to be the indicator of this interval because the minimizer of $\hat{L}(\theta)$ will then have some bias towards larger values of $\theta^{(1)}$ and we suspect that the minimum will be harder to compute. Figures 2(b) and 2(c) contain the same data as Figure 2(a), except that the (x_i', Y_i') have been replaced by $(x_i' + 0.106, Y_i' + 0.196)$ and $(x_i' + 0.2, Y_i' + 0.2)$, respectively. Observe that from these figures it is quite apparent that $\theta_0^{(1)} \in [0, 0.2]$. Hence, we can restrict Θ to only include θ with $\theta^{(1)} \in [0, 0.2]$, and take $w(x)$ to be the indicator of $[\eta, 0.8 - \eta]$.

In the general case, we assume that there is an interval $[a, b] \subseteq [0,1]$ where boundary effects are eliminated and then define

$$w(x) = \prod_{\theta \in \Theta} 1_{[a,b]}(T_\theta x)$$

$$= 1_{\bigcap_{\theta \in \Theta} T_\theta^{-1}([a,b])}(x).$$

Note that for $\hat{\theta}$ to be a reasonable estimate, this requires that Θ be rather small. This assumption does not seem too restrictive because these methods will only be applied after the experimenter has looked at some preliminary curve estimates. Such a previewing procedure does not cause any additional effort if an interactive graphical data analysis program is available. Hence, it is assumed that the

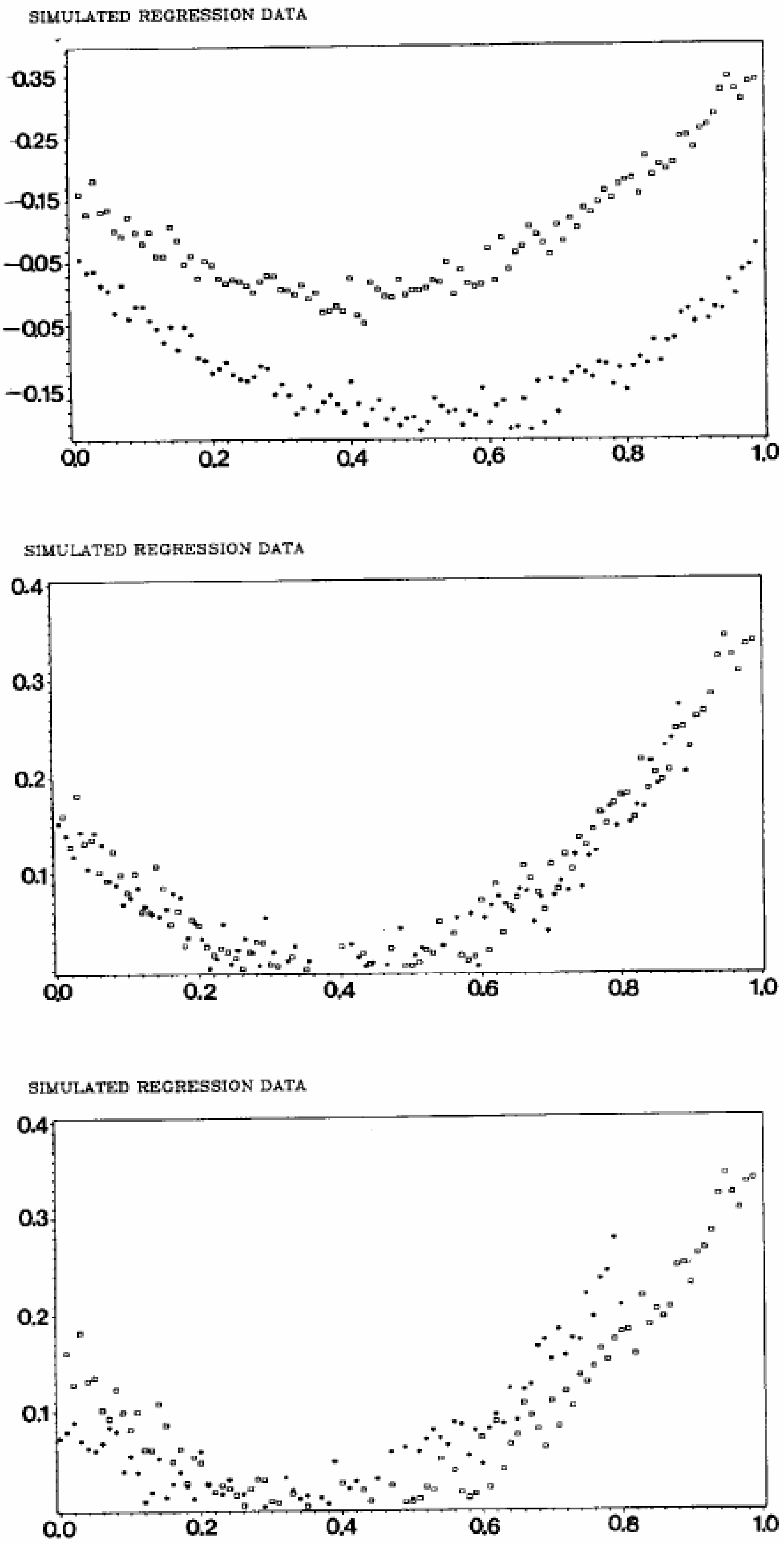


FIG. 2. *Simulated data.*

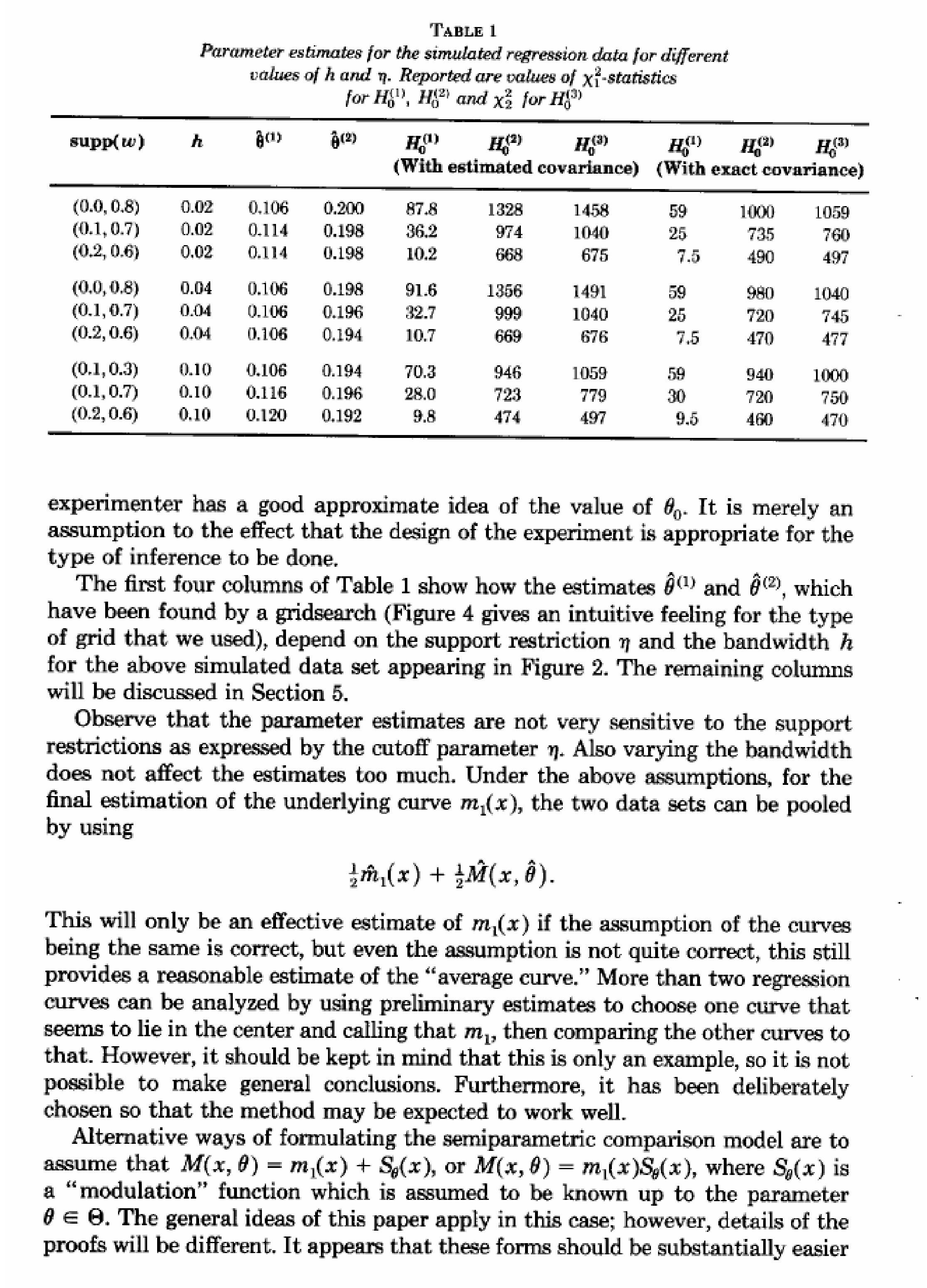


TABLE 1

Parameter estimates for the simulated regression data for different values of h and η. Reported are values of χ^2_1-statistics for $H_0^{(1)}$, $H_0^{(2)}$ and χ^2_2 for $H_0^{(3)}$

supp(w)	h	$\hat{\theta}^{(1)}$	$\hat{\theta}^{(2)}$	$H_0^{(1)}$	$H_0^{(2)}$	$H_0^{(3)}$	$H_0^{(1)}$	$H_0^{(2)}$	$H_0^{(3)}$
				(With estimated covariance)			(With exact covariance)		
(0.0, 0.8)	0.02	0.106	0.200	87.8	1328	1458	59	1000	1059
(0.1, 0.7)	0.02	0.114	0.198	36.2	974	1040	25	735	760
(0.2, 0.6)	0.02	0.114	0.198	10.2	668	675	7.5	490	497
(0.0, 0.8)	0.04	0.106	0.198	91.6	1356	1491	59	980	1040
(0.1, 0.7)	0.04	0.106	0.196	32.7	999	1040	25	720	745
(0.2, 0.6)	0.04	0.106	0.194	10.7	669	676	7.5	470	477
(0.1, 0.3)	0.10	0.106	0.194	70.3	946	1059	59	940	1000
(0.1, 0.7)	0.10	0.116	0.196	28.0	723	779	30	720	750
(0.2, 0.6)	0.10	0.120	0.192	9.8	474	497	9.5	460	470

experimenter has a good approximate idea of the value of θ_0. It is merely an assumption to the effect that the design of the experiment is appropriate for the type of inference to be done.

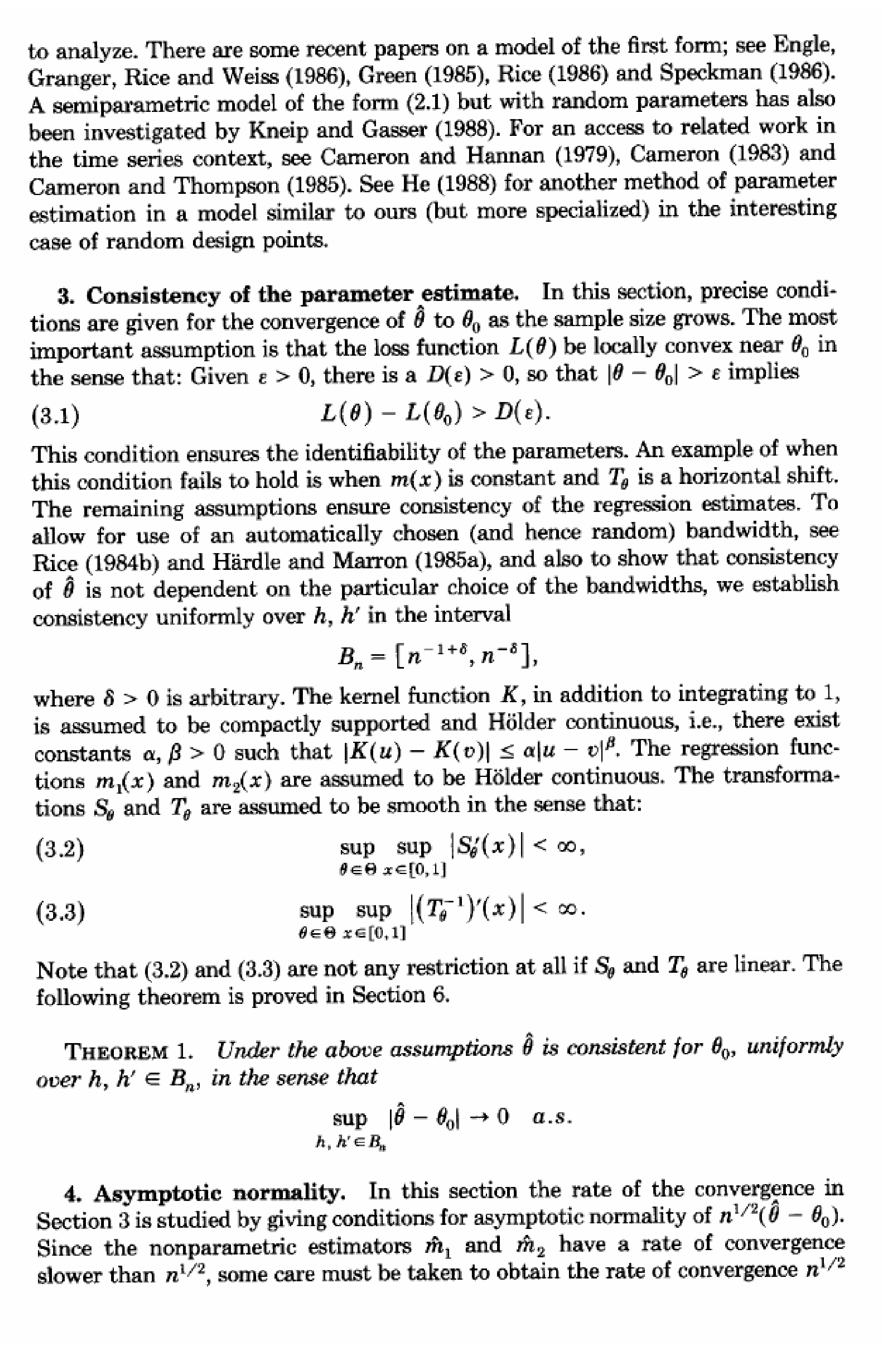
The first four columns of Table 1 show how the estimates $\hat{\theta}^{(1)}$ and $\hat{\theta}^{(2)}$, which have been found by a gridsearch (Figure 4 gives an intuitive feeling for the type of grid that we used), depend on the support restriction η and the bandwidth h for the above simulated data set appearing in Figure 2. The remaining columns will be discussed in Section 5.

Observe that the parameter estimates are not very sensitive to the support restrictions as expressed by the cutoff parameter η. Also varying the bandwidth does not affect the estimates too much. Under the above assumptions, for the final estimation of the underlying curve $m_1(x)$, the two data sets can be pooled by using

$$\tfrac{1}{2}\hat{m}_1(x) + \tfrac{1}{2}\hat{M}(x, \hat{\theta}).$$

This will only be an effective estimate of $m_1(x)$ if the assumption of the curves being the same is correct, but even the assumption is not quite correct, this still provides a reasonable estimate of the "average curve." More than two regression curves can be analyzed by using preliminary estimates to choose one curve that seems to lie in the center and calling that m_1, then comparing the other curves to that. However, it should be kept in mind that this is only an example, so it is not possible to make general conclusions. Furthermore, it has been deliberately chosen so that the method may be expected to work well.

Alternative ways of formulating the semiparametric comparison model are to assume that $M(x, \theta) = m_1(x) + S_\theta(x)$, or $M(x, \theta) = m_1(x)S_\theta(x)$, where $S_\theta(x)$ is a "modulation" function which is assumed to be known up to the parameter $\theta \in \Theta$. The general ideas of this paper apply in this case; however, details of the proofs will be different. It appears that these forms should be substantially easier



to analyze. There are some recent papers on a model of the first form; see Engle, Granger, Rice and Weiss (1986), Green (1985), Rice (1986) and Speckman (1986). A semiparametric model of the form (2.1) but with random parameters has also been investigated by Kneip and Gasser (1988). For an access to related work in the time series context, see Cameron and Hannan (1979), Cameron (1983) and Cameron and Thompson (1985). See He (1988) for another method of parameter estimation in a model similar to ours (but more specialized) in the interesting case of random design points.

3. Consistency of the parameter estimate. In this section, precise conditions are given for the convergence of $\hat{\theta}$ to θ_0 as the sample size grows. The most important assumption is that the loss function $L(\theta)$ be locally convex near θ_0 in the sense that: Given $\varepsilon > 0$, there is a $D(\varepsilon) > 0$, so that $|\theta - \theta_0| > \varepsilon$ implies

$$L(\theta) - L(\theta_0) > D(\varepsilon). \tag{3.1}$$

This condition ensures the identifiability of the parameters. An example of when this condition fails to hold is when $m(x)$ is constant and T_θ is a horizontal shift. The remaining assumptions ensure consistency of the regression estimates. To allow for use of an automatically chosen (and hence random) bandwidth, see Rice (1984b) and Härdle and Marron (1985a), and also to show that consistency of $\hat{\theta}$ is not dependent on the particular choice of the bandwidths, we establish consistency uniformly over h, h' in the interval

$$B_n = [n^{-1+\delta}, n^{-\delta}],$$

where $\delta > 0$ is arbitrary. The kernel function K, in addition to integrating to 1, is assumed to be compactly supported and Hölder continuous, i.e., there exist constants $\alpha, \beta > 0$ such that $|K(u) - K(v)| \le \alpha|u - v|^\beta$. The regression functions $m_1(x)$ and $m_2(x)$ are assumed to be Hölder continuous. The transformations S_θ and T_θ are assumed to be smooth in the sense that:

$$\sup_{\theta\in\Theta}\ \sup_{x\in[0,1]} |S_\theta'(x)| < \infty, \tag{3.2}$$

$$\sup_{\theta\in\Theta}\ \sup_{x\in[0,1]} |(T_\theta^{-1})'(x)| < \infty. \tag{3.3}$$

Note that (3.2) and (3.3) are not any restriction at all if S_θ and T_θ are linear. The following theorem is proved in Section 6.

THEOREM 1. *Under the above assumptions $\hat{\theta}$ is consistent for θ_0, uniformly over h, $h' \in B_n$, in the sense that*

$$\sup_{h,\,h'\in B_n} |\hat{\theta} - \theta_0| \to 0 \quad a.s.$$

4. Asymptotic normality. In this section the rate of the convergence in Section 3 is studied by giving conditions for asymptotic normality of $n^{1/2}(\hat{\theta} - \theta_0)$. Since the nonparametric estimators $\hat{m}_1$ and $\hat{m}_2$ have a rate of convergence slower than $n^{1/2}$, some care must be taken to obtain the rate of convergence $n^{1/2}$

for the $\hat{\theta} - \theta_0$ limiting distribution. To this end we assume that

$$T_\theta x = \theta^{(1)} + \theta^{(2)} x, \tag{4.1}$$

$$S_\theta \text{ only depends on } \theta^{(3)}, \dots, \theta^{(d)}, \tag{4.2}$$

and that $\hat{m}_1$ and $\hat{m}_2$ employ the same amount of smoothing in the sense that

$$h' = \theta^{(2)} h. \tag{4.3}$$

Assumption (4.3) seems quite restrictive at first glance; however, an inspection of the proofs reveals that it is in fact necessary for $n^{1/2}$ convergence of the parameter estimates. A simple way of implementing this in practice is to choose the bandwidth for only $\hat{m}_1$, say by cross validation, and then using a preliminary estimate of $\theta^{(2)}$ to get an improved $\hat{\theta}^{(2)}$ and iterating. More efficient methods would pool the information from the two curves, as discussed in Marron and Rudemo (1988) and Marron and Schmitz (1988). This is complicated in the present situation because the smoothing parameter selection is confounded with the estimation of θ, but a promising possibility to be investigated is to choose both h and $\hat{\theta}$ to be the joint minimizers of the sum of $\hat{L}(\theta)$ and the cross-validation score functions for the two curves.

As in Section 3, a critical assumption concerns the identifiability of θ_0. Assume that

$$H(\theta_0) \text{ is positive definite,} \tag{4.4}$$

where $H(\theta)$ is the $d \times d$ matrix whose l, l'th entry is

$$\int M_l(x,\theta) M_{l'}(x,\theta) w(x)\, dx,$$

using the notation $M_l(x,\theta) = (\partial/\partial\theta^{(l)}) M(x,\theta)$. Under the assumptions of this section, it can be shown that (4.4) implies (3.1). To gain some insight into this, consider the case $S_\theta(x) = \theta^{(3)} + \theta^{(4)} x$, where

$$M_1(x,\theta) = \theta^{(4)} m_2'(\theta^{(1)} + \theta^{(2)} x),$$

$$M_2(x,\theta) = \theta^{(4)} m_2'(\theta^{(1)} + \theta^{(2)} x) x,$$

$$M_3(x,\theta) = 1,$$

$$M_4(x,\theta) = m_2(\theta^{(1)} + \theta^{(2)} x).$$

Observe that (4.4) is then essentially requiring that the functions 1, $m(x)$, $m'(x)$ and $xm'(x)$ be linearly independent in $L^2(w)$. To facilitate Taylor expansion arguments, it is assumed that $S_\theta(x)$ is smooth in the sense that the following functions are uniformly continuous and bounded uniformly over $x \in \text{supp}(w)$

over $\theta \in \Theta$, and over $l, l' = 1, \ldots, d$:

$$
\begin{aligned}
S'_\theta(x) &= \frac{\partial}{\partial x} S_\theta(x), \\
S_{\theta,l}(x) &= \frac{\partial}{\partial \theta^{(l)}} S_\theta(x), \\
S_{\theta,l,l'}(x) &= \frac{\partial}{\partial \theta^{(l')}} S_{\theta,l}(x), \qquad (4.5)\\
S'_{\theta,l}(x) &= \frac{\partial}{\partial x} S_{\theta,l}(x), \\
S'_{\theta,l,l'}(x) &= \frac{\partial}{\partial x} S_{\theta,l,l'}(x).
\end{aligned}
$$

This assumption is trivial if S_θ is linear. Also to facilitate expansions, assume

(4.6) $\qquad m_1''(x)$ exists and is uniformly continuous.

A consequence of (4.5), (4.6) and the linearity of T_θ is that

$$M_{l,l'}(x,\theta) = \frac{\partial}{\partial \theta^{(l')}} M_l(x,\theta) \qquad (4.7)$$

is uniformly continuous and bounded uniformly over $x \in \operatorname{supp}(w)$, $\theta \in \Theta$, and $l, l' = 1, \ldots, d$. Also assume that K is a compactly supported probability density with Hölder continuous second derivative, and that $E\varepsilon_i^k < \infty$, for $k = 1, 2, \ldots$, uniformly over $i = 1, \ldots, n$. The final requirement is that the bandwidth h is taken to be an automatically selected bandwidth $\hat{h}$, as discussed in Rice (1984b), Härdle and Marron (1985a, b) and Härdle, Hall and Marron (1988) have shown that under the above assumptions

$$\hat{h} = h_0 + O_p(n^{-3/10}),$$

where $h_0 = c_0 n^{-1/5}$, for a constant c_0. Hence, if B_n^* is defined by

$$B_n^* = \left[h_0 - n^{-3/10+\alpha}, h_0 + n^{-3/10+\alpha}\right],$$

for some $\alpha \in (0, 1/10)$, then $P[\hat{h} \in B_n^*] \to 1$. Note that $\hat{h}$ is chosen only from the data $Y_1, \ldots, Y_n$. This allows assumption (4.3) to be satisfied in a simple fashion. See the discussion there for other possibilities.

THEOREM 2. *Under the above assumptions*

$$\sqrt{n}(\hat{\theta} - \theta_0) \to_{\mathscr{L}} N(0, H^{-1}(\theta_0) \Sigma H^{-1}(\theta_0)),$$

where the l, l'th entry of Σ is

$$4 \int \left[\sigma^2 + \sigma'^2\left(S'_{\theta_0}\left(m_2(T_{\theta_0}x)\right)\right)^2\right] M_l(x,\theta_0) M_{l'}(x,\theta_0) w(x)\, dx.$$

The proof of Theorem 2 is in Section 7. To add insight into this theorem, consider the special case of the example given in Section 2. Note that $T_\theta x =$

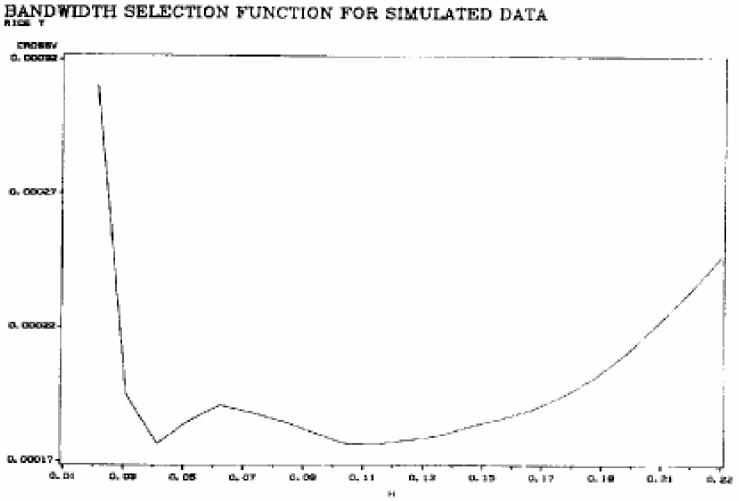


FIG. 3. *Bandwidth selection function based on the Epanečnikov kernel and weight function on* (0.1, 0.7).

$\theta^{(1)} + x$ and $S_\theta x = \theta^{(2)} + x$, so an obvious modification of the notation of this section will be made. In particular, from (2.2),

$$\begin{aligned} M_1(x, \theta) &= 2(x + \theta^{(1)} - 0.5), \\ M_2(x, \theta) &= 1, \\ S'_{\theta_0}(x) &= 1, \end{aligned}$$

and so

$$H(\theta_0) = \begin{pmatrix} \frac{8}{3}(0.4 - \eta)^3 & 0 \\ 0 & 2(0.4 - \eta) \end{pmatrix}.$$

Thus, $\sqrt{n}(\hat{\theta} - \theta)$ has asymptotic covariance matrix

$$\frac{\sigma^2}{(0.4 - \eta)^3} \begin{pmatrix} 3 & 0 \\ 0 & 4(0.4 - \eta)^2 \end{pmatrix}.$$

The bandwidth selection function computed for (x_i, Y_i), with w supported on $[0.1, 0.7]$ and the Epanečnikov kernel had a global minimum at $h = 0.04$ (Figure 3) but had a pronounced local minimum. In this simulated example we used

$$T(h) = n^{-1} \sum_{i=1}^{n} [Y_i - \hat{m}_1(x_i)]^2 w(x_i) / [1 - 2n^{-1}h^{-1}K(0)]$$

as a bandwidth selector. See Härdle, Hall and Marron (1988) for a more complete discussion of the issues of bandwidth selection. The negative loss function for this bandwidth is shown in Figure 4. Note that Figure 4 shows that the loss function is more sensitive to changes in $\theta^{(2)}$ than to changes in $\theta^{(1)}$. This is reflected intuitively by thinking about vertical and horizontal shifts in Figure 2(a).

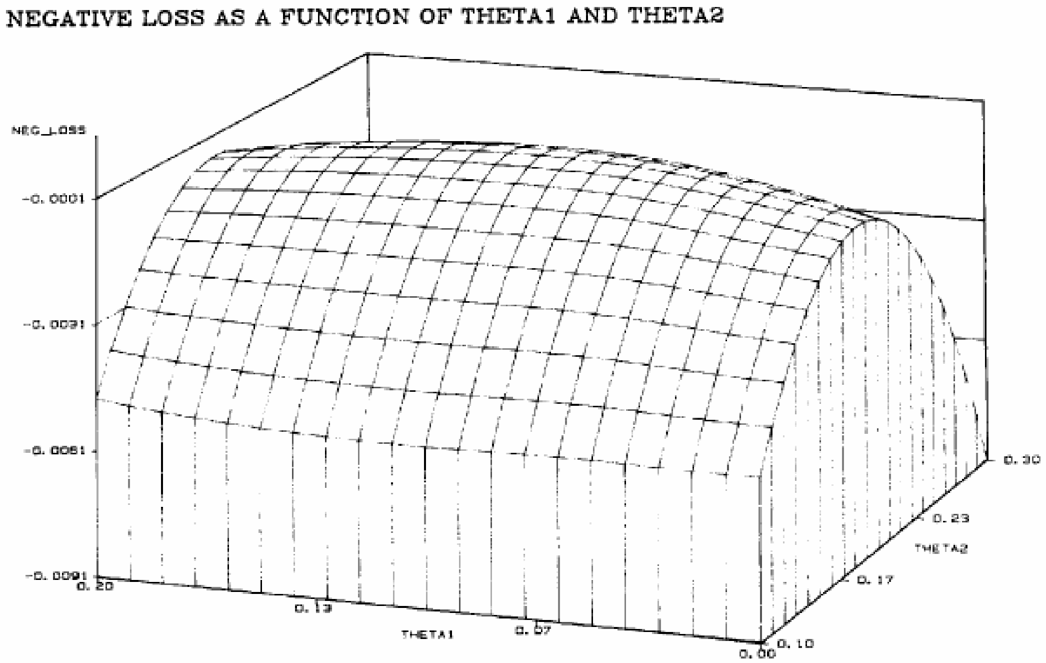


FIG. 4. *Negative loss as a function of* $\theta^{(1)}$ *and* $\theta^{(2)}$. $m(X) = (X - 0.4)^2$. *Errors* $N(0, 0.0004)$. *Weight function on* $(0.1, 0.7)$.

While the vertical shift is obvious, we find it much more difficult to justify a horizontal shift just by "eye inspection." Statistically, this can be quantified by $\mathrm{var}(\hat{\theta}^{(1)}) \approx 0.0444$, $\mathrm{var}(\hat{\theta}^{(2)}) \approx 0.0053$ (where these are the entries in the asymptotic covariance matrix given in Theorem 2). $L(\theta)$ is minimized at $\hat{\theta} = (\hat{\theta}^{(1)}, \hat{\theta}^{(2)}) = (0.106, 0.196)$ which is the shift used in the construction of Figure 2(b). An intuitive understanding of $\hat{\theta}$ can also be gained from Figure 5, which shows $\hat{m}_1(x)$ (solid line) and $\hat{M}(x, \hat{\theta})$ (dashed line). Note that either a horizontal or a vertical shift in the relative position of these curves will increase the integrated (over $[0.1, 0.7]$) squared difference between these.

A look at Figure 1 indicates that the shift-scale model, $T_\theta = \theta^{(1)} + x$, $S_\theta = \theta^{(4)}x$ (using notation consistent with this section) should be appropriate for the automobile side impact data. After transforming the X-values into the unit interval, the bandwidth $\hat{h} = 0.012$ was obtained by cross validation over the interval $[0.1, 0.7]$ for the data set shown in Figure 1(b), which we took to be $\{(x_i, Y_i)\}_{i=1}^n$ with $n = 800$. The negative loss function $\hat{L}(\theta)$ is shown in Figure 6, which for its form is called the "Sidney Opera.". As expected from a comparison of Figures 1(a) and 1(b), the choice of $\theta^{(1)}$ is more critical than that of $\theta^{(4)}$. The "side ridges" in the negative loss correspond to values of $\theta^{(1)}$, where there is a matching of "first peaks" to "second peaks." The loss function was minimized at $\hat{\theta} = (\hat{\theta}^{(1)}, \hat{\theta}^{(4)}) = (0.13, 1.45)$. Figure 7 shows how $\hat{m}_1(x)$ (solid curve) compares with $\hat{M}(x, \hat{\theta})$ (dashed curve).

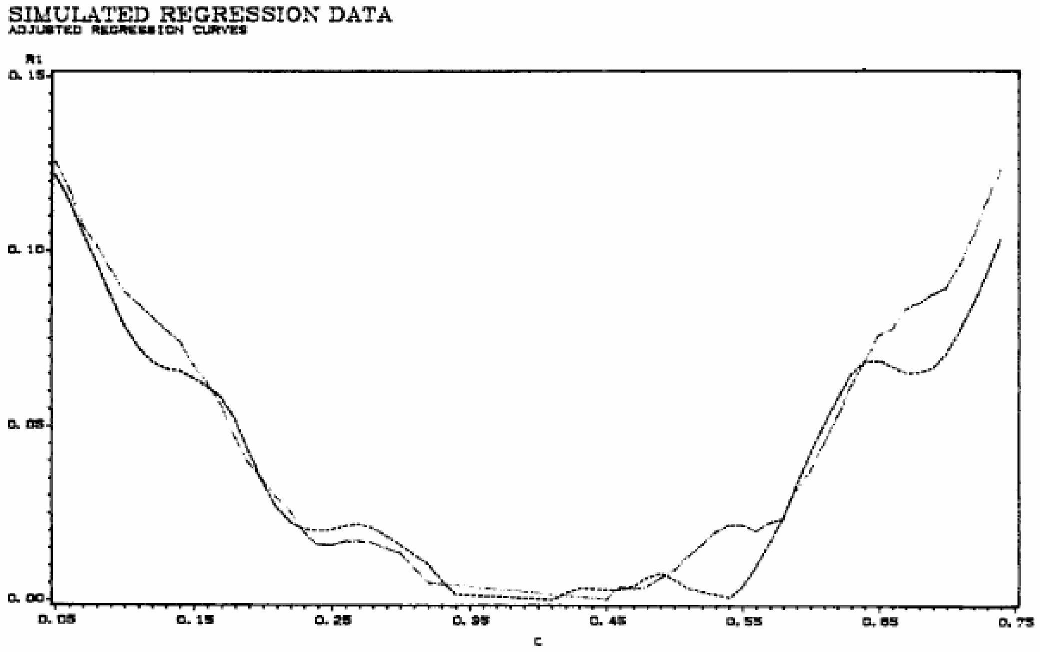


FIG. 5. *Adjusted regression curves for the simulated data.* $m(X) = (X - 0.4)^2$. *Errors* $N(0, 0.0004)$. $\theta^{(1)} = 0.106$; $\theta^{(2)} = 0.196$.

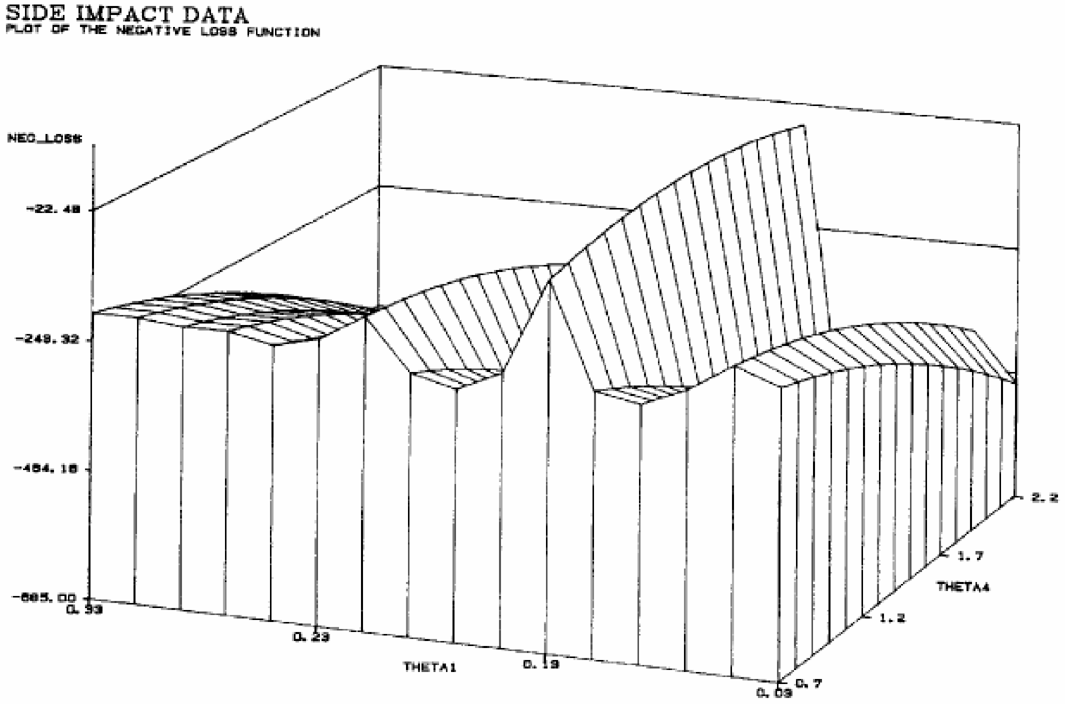


FIG. 6. *"Sidney Opera" negative loss function for the side impact data. Weight function on* (0.1, 0.7).

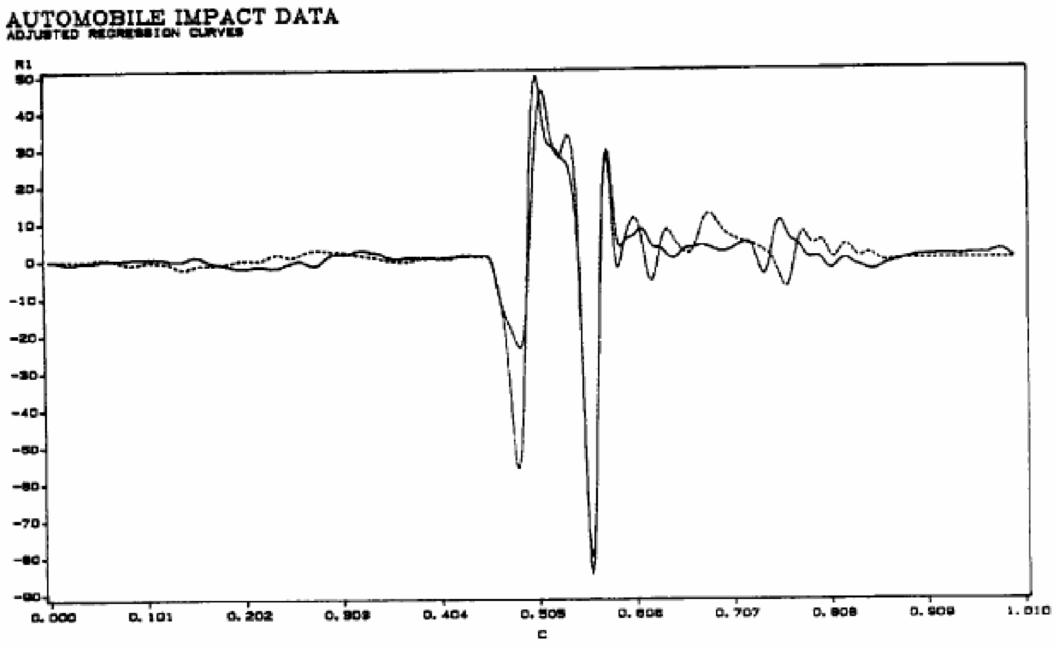


FIG. 7. *Adjusted regression curves for the automobile side impact data. Weight function on* $(0.1, 0.8)$. $\theta^{(1)} = 0.13$; $\theta^{(4)} = 1.45$.

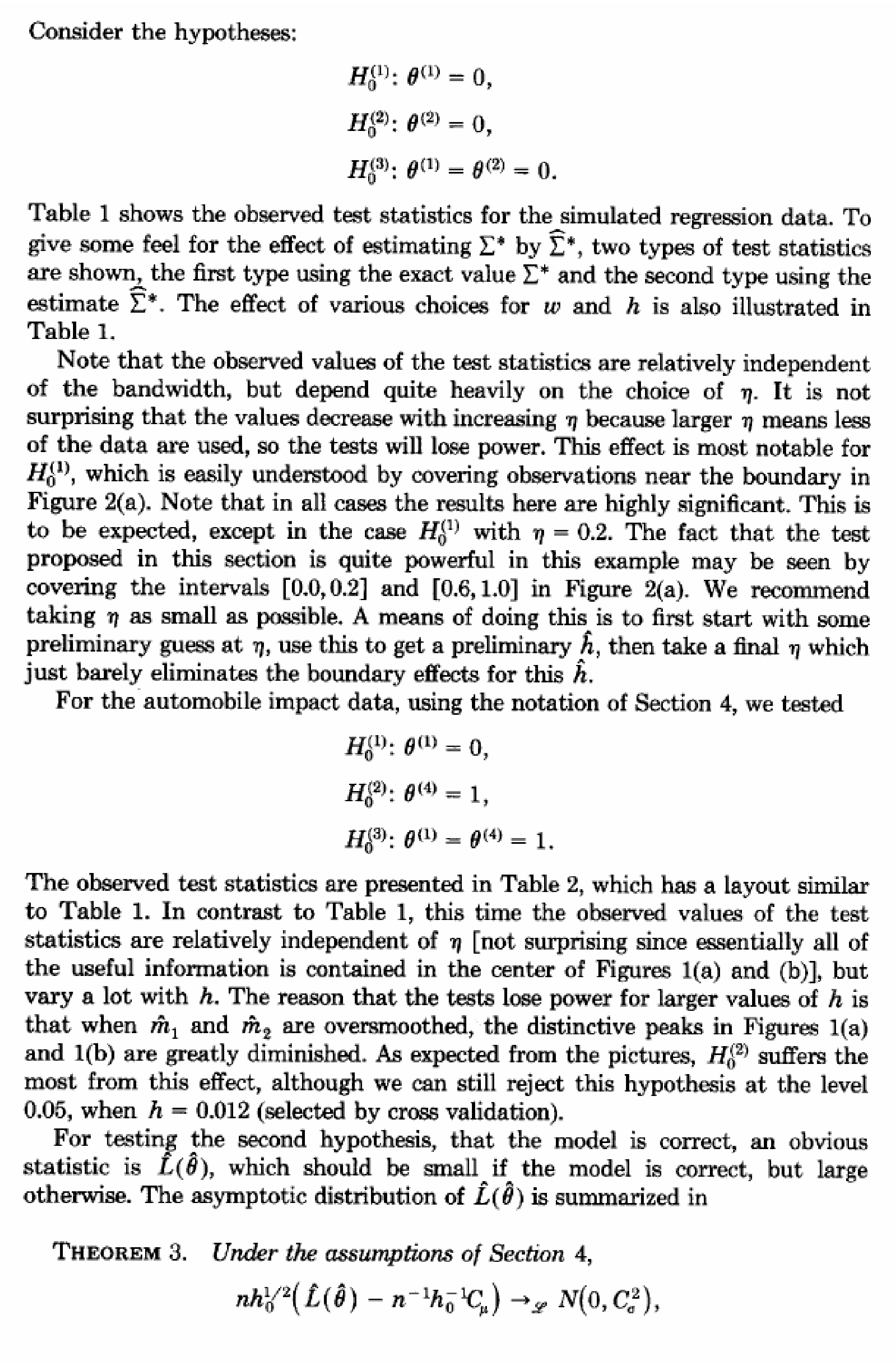
5. Hypothesis testing. There are two important hypotheses to test in this semiparametric model. First, can the parametric part of the model be reduced? (For example: Can a horizontal shift and scale be reasonably replaced by just a shift? Is an apparent vertical shift really significant?) Second, is the semiparametric model of this paper appropriate for a particular data set? [That is: Is $m_2(x)$ really a simple transformation of $m_1(x)$?] To formulate the first hypothesis, suppose there is a $\theta^* \in \Theta$ so that

$$m_1(x) = M(x, \theta^*).$$

For example, components of θ^* corresponding to the types of shifts discussed earlier are 0 and to the scaling are 1. A general way to formulate the hypothesis is

$$H_0\colon A(\theta_0 - \theta^*) = 0,$$

for an $r \times d$ matrix A of rank r. A reasonable basis for a hypothesis test is $A(\hat{\theta} - \theta^*)$, which has an asymptotic $N(0, \Sigma^*)$ distribution under H_0, where $\Sigma^* = AH(\theta_0)^{-1}\Sigma H(\theta_0)^{-1}A^T$. This suggests rejecting H_0 when $(\hat{\theta} - \theta^*)^T A^T \widehat{\Sigma}^{*-1} A(\hat{\theta} - \theta^*)$ is larger than the 95th percentile of the χ^2_r distribution, where $\widehat{\Sigma}^*$ is a consistent estimate of Σ^*. These ideas can be illustrated in the simulated data example of Section 2 which is depicted in Figure 2(a).



Consider the hypotheses:

$$H_0^{(1)}\colon \theta^{(1)} = 0,$$

$$H_0^{(2)}\colon \theta^{(2)} = 0,$$

$$H_0^{(3)}\colon \theta^{(1)} = \theta^{(2)} = 0.$$

Table 1 shows the observed test statistics for the simulated regression data. To give some feel for the effect of estimating Σ^* by $\widehat{\Sigma}^*$, two types of test statistics are shown, the first type using the exact value Σ^* and the second type using the estimate $\widehat{\Sigma}^*$. The effect of various choices for w and h is also illustrated in Table 1.

Note that the observed values of the test statistics are relatively independent of the bandwidth, but depend quite heavily on the choice of η. It is not surprising that the values decrease with increasing η because larger η means less of the data are used, so the tests will lose power. This effect is most notable for $H_0^{(1)}$, which is easily understood by covering observations near the boundary in Figure 2(a). Note that in all cases the results here are highly significant. This is to be expected, except in the case $H_0^{(1)}$ with $\eta = 0.2$. The fact that the test proposed in this section is quite powerful in this example may be seen by covering the intervals $[0.0, 0.2]$ and $[0.6, 1.0]$ in Figure 2(a). We recommend taking η as small as possible. A means of doing this is to first start with some preliminary guess at η, use this to get a preliminary $\hat{h}$, then take a final η which just barely eliminates the boundary effects for this $\hat{h}$.

For the automobile impact data, using the notation of Section 4, we tested

$$H_0^{(1)}\colon \theta^{(1)} = 0,$$

$$H_0^{(2)}\colon \theta^{(4)} = 1,$$

$$H_0^{(3)}\colon \theta^{(1)} = \theta^{(4)} = 1.$$

The observed test statistics are presented in Table 2, which has a layout similar to Table 1. In contrast to Table 1, this time the observed values of the test statistics are relatively independent of η [not surprising since essentially all of the useful information is contained in the center of Figures 1(a) and (b)], but vary a lot with h. The reason that the tests lose power for larger values of h is that when $\hat{m}_1$ and $\hat{m}_2$ are oversmoothed, the distinctive peaks in Figures 1(a) and 1(b) are greatly diminished. As expected from the pictures, $H_0^{(2)}$ suffers the most from this effect, although we can still reject this hypothesis at the level 0.05, when $h = 0.012$ (selected by cross validation).

For testing the second hypothesis, that the model is correct, an obvious statistic is $\hat{L}(\hat{\theta})$, which should be small if the model is correct, but large otherwise. The asymptotic distribution of $\hat{L}(\hat{\theta})$ is summarized in

THEOREM 3. *Under the assumptions of Section* 4,

$$nh_0^{1/2}\left(\hat{L}(\hat{\theta}) - n^{-1}h_0^{-1}C_\mu\right) \to_{\mathscr{L}} N\left(0, C_\sigma^2\right),$$

TABLE 2
χ^2-statistics for different bandwidths and support restrictions for impact data

supp(w)	h	$H_0^{(1)}$	$H_0^{(2)}$ (With estimated covariance)	$H_0^{(3)}$
(0.0, 0.8)	0.005	248	13.4	2180
(0.1, 0.7)	0.005	246	13.3	2150
(0.2, 0.6)	0.005	245	13.0	2130
(0.0, 0.8)	0.012	80.9	4.36	232
(0.1, 0.7)	0.012	80.3	4.32	229
(0.2, 0.6)	0.012	80.0	4.25	227
(0.0, 0.8)	0.040	41.9	2.26	62.3
(0.1, 0.7)	0.040	41.6	2.24	61.4
(0.2, 0.6)	0.040	41.4	2.20	60.9

where

$$C_\mu = \left(\int K^2\right)\left(\int\left[\frac{\sigma^2}{\theta_0^{(2)}} + \sigma'^2\left(S'_{\theta_0}(m_2(x))\right)^2\right]w\left(T_{\theta_0}^{-1}(x)\right)dx\right),$$

$$C_\sigma^2 = 2\theta_0^{(2)}\left(\int (K*K)^2\right)\left(\int\left[\frac{\sigma^2}{\theta_0^{(2)}} + \sigma'^2\left(S'_{\theta_0}(m_2(x))\right)^2\right]^2 w\left(T_{\theta_0}^{-1}(x)\right)dx\right).$$

The proof of Theorem 3 is in Section 8. It follows from Theorem 3 that a reasonable test, of the hypothesis that m_2 is indeed a parametric shift of m, will reject when

$$\hat{L}(\hat{\theta}) > (n\hat{h})^{-1}\hat{C}_\mu + n^{-1}\hat{h}^{-1/2}\hat{C}_\sigma z_{1-\alpha},$$

where z_{1-d} is the $(1-\alpha)$th quantile of the standard normal distribution, and where the estimates

$$\hat{C}_\mu = \left(\int K^2\right)\left(\int\left[\frac{\hat{\sigma}^2}{\hat{\theta}^{(2)}} + \hat{\sigma}'^2(S'_{\hat{\theta}}(\hat{m}_2(x)))^2\right]w\left(T_{\hat{\theta}}^{-1}(x)\right)dx\right),$$

$$\hat{C}_\sigma^2 = 2\hat{\theta}^{(2)}\left(\int (K*K)^2\right)\left(\int\left[\frac{\hat{\sigma}^2}{\hat{\theta}^{(2)}} + \hat{\sigma}'^2(S'_{\hat{\theta}}(\hat{m}_2(x)))\right]^2 w\left(T_{\hat{\theta}}^{-1}(x)\right)dx\right),$$

$$\hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}(Y_i - \hat{m}_1(x_i))^2, \qquad \hat{\sigma}'^2 = n^{-1}\sum_{i=1}^{n}(Y_i' - \hat{m}_2(x_i'))^2$$

have been used. The observed test statistics for the side impact data set are listed in Table 3 for a weight function concentrated on (0.1, 0.7). The shift-scale model that we proposed achieved a p-value of 0.02, whereas all the other studied submodels had p-values less than 0.001. Figure 7 provides an intuitive feeling for the power involved in this test. Note that while $\hat{\theta}$ clearly provides an informative choice of the parameters, it is also clear that the curves are certainly not the same. The fact that, at least in this example, the parameter estimation method

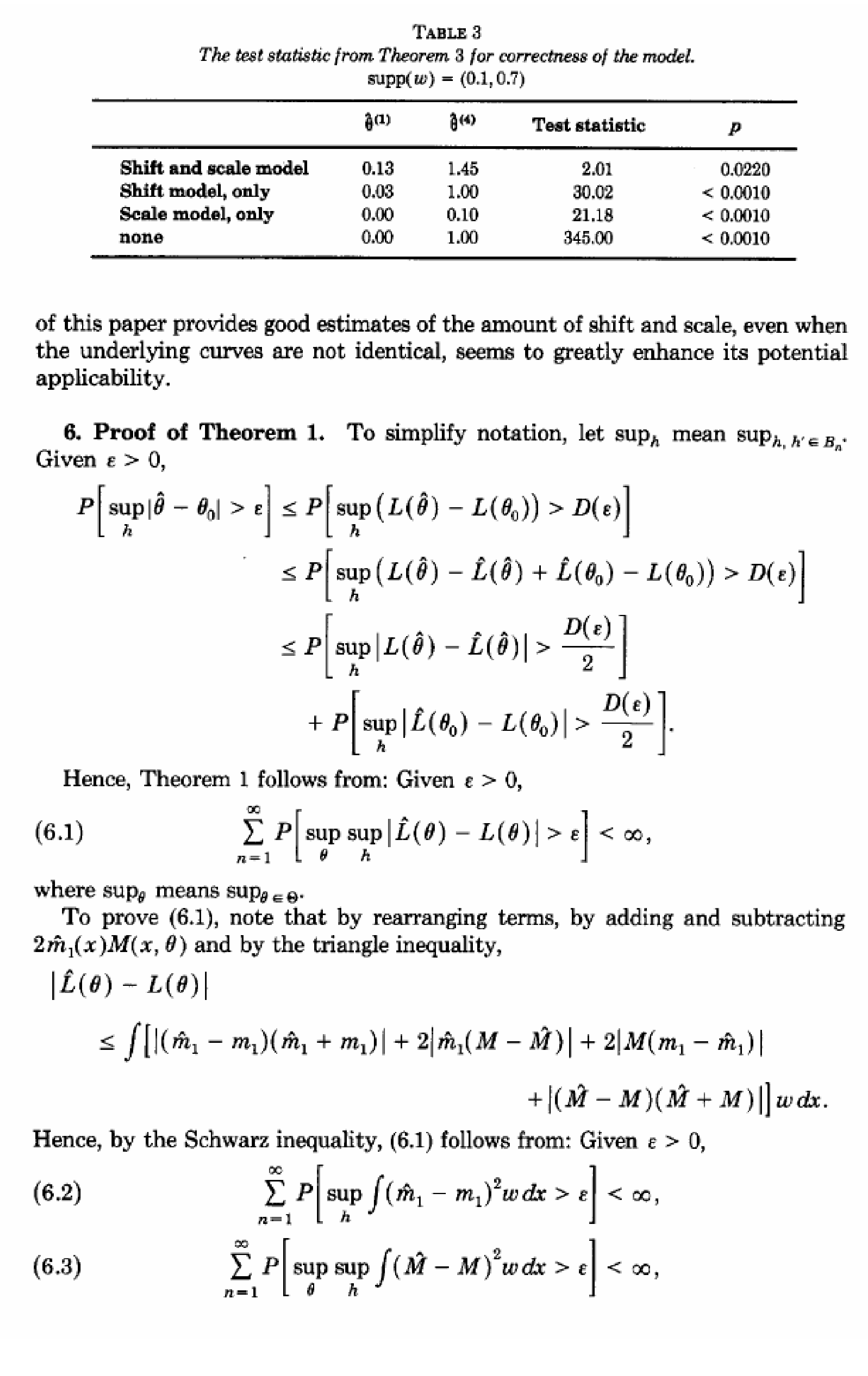


TABLE 3

The test statistic from Theorem 3 for correctness of the model.
supp$(w) = (0.1, 0.7)$

	$\hat\theta^{(1)}$	$\hat\theta^{(4)}$	Test statistic	p
Shift and scale model	0.13	1.45	2.01	0.0220
Shift model, only	0.03	1.00	30.02	< 0.0010
Scale model, only	0.00	0.10	21.18	< 0.0010
none	0.00	1.00	345.00	< 0.0010

of this paper provides good estimates of the amount of shift and scale, even when the underlying curves are not identical, seems to greatly enhance its potential applicability.

6. Proof of Theorem 1. To simplify notation, let $\sup_h$ mean $\sup_{h, h' \in B_n}$. Given $\varepsilon > 0$,

$$\begin{aligned} P\left[\sup_h |\hat\theta - \theta_0| > \varepsilon\right] &\le P\left[\sup_h \left(L(\hat\theta) - L(\theta_0)\right) > D(\varepsilon)\right] \\ &\le P\left[\sup_h \left(L(\hat\theta) - \hat L(\hat\theta) + \hat L(\theta_0) - L(\theta_0)\right) > D(\varepsilon)\right] \\ &\le P\left[\sup_h \left|L(\hat\theta) - \hat L(\hat\theta)\right| > \frac{D(\varepsilon)}{2}\right] \\ &\quad + P\left[\sup_h \left|\hat L(\theta_0) - L(\theta_0)\right| > \frac{D(\varepsilon)}{2}\right]. \end{aligned}$$

Hence, Theorem 1 follows from: Given $\varepsilon > 0$,

$$\sum_{n=1}^{\infty} P\left[\sup_\theta \sup_h \left|\hat L(\theta) - L(\theta)\right| > \varepsilon\right] < \infty, \tag{6.1}$$

where $\sup_\theta$ means $\sup_{\theta \in \Theta}$.

To prove (6.1), note that by rearranging terms, by adding and subtracting $2\hat m_1(x) M(x, \theta)$ and by the triangle inequality,

$$\begin{aligned} &\left|\hat L(\theta) - L(\theta)\right| \\ &\quad \le \int \Big[\left|(\hat m_1 - m_1)(\hat m_1 + m_1)\right| + 2\left|\hat m_1(M - \hat M)\right| + 2\left|M(m_1 - \hat m_1)\right| \\ &\qquad\qquad + \left|(\hat M - M)(\hat M + M)\right|\Big] w\,dx. \end{aligned}$$

Hence, by the Schwarz inequality, (6.1) follows from: Given $\varepsilon > 0$,

$$\sum_{n=1}^{\infty} P\left[\sup_h \int (\hat m_1 - m_1)^2 w\,dx > \varepsilon\right] < \infty, \tag{6.2}$$

$$\sum_{n=1}^{\infty} P\left[\sup_\theta \sup_h \int (\hat M - M)^2 w\,dx > \varepsilon\right] < \infty, \tag{6.3}$$

together with

$$\int m_1^2 w\,dx < \infty, \tag{6.4}$$

$$\sup_\theta \int M^2 w\,dx < \infty. \tag{6.5}$$

To prove (6.2), note that for $B_n' \subseteq B_n$,

$$\begin{aligned} P\left[\sup_h \int (\hat{m}_h - m_1)^2 w\,dx > \varepsilon\right] &\le P\left[\sup_{h\in B_n'} \int (\hat{m}_h - m_1)^2 w\,dx > \frac{\varepsilon}{2}\right] \\ &\quad + P\left[\sup_{h\in B_n} \inf_{h_1\in B_n'} \left|\int (\hat{m}_h - m_1)^2 w\,dx - \int (\hat{m}_{h_1} - m_1)^2 w\,dx\right| > \frac{\varepsilon}{2}\right]. \end{aligned}$$

By Hölder continuity of m_1 and K, B_n' can be chosen so that the second term is 0, for n sufficiently large, and so that $\#(B_n') \le n^\xi$, some $\xi > 0$. Hence, by Theorem 1 of Marron and Härdle (1986), (6.2) follows from

$$\sup_{h\in B_n'} E\int (\hat{m}_h - m_1)^2 w\,dx \to 0,$$

which is easily established by the methods of Rosenblatt (1971).

To prove (6.3), note that

$$\begin{aligned} &\int [\hat{M}(x,\theta) - M(x,\theta)]^2 w(x)\,dx \\ &\quad = \int [S_\theta \hat{m}_2(T_\theta x) - S_\theta m_2(T_\theta x)]^2 w(x)\,dx \\ &\quad = \int_0^1 [S_\theta'(\xi)(\hat{m}_2(u) - m_2(u))]^2 w(T_\theta^{-1}(u))(T_\theta^{-1})'(u)\,du. \end{aligned}$$

Hence, (6.3) follows from (3.2), (3.3) and the methods used to establish (6.2).

Note that (6.4) is a consequence of the Hölder continuity of $m_1(x)$. To prove (6.5), use Hölder continuity of $m_2(x)$ and an argument of the type used on (6.3). This completes the proof of Theorem 1. □

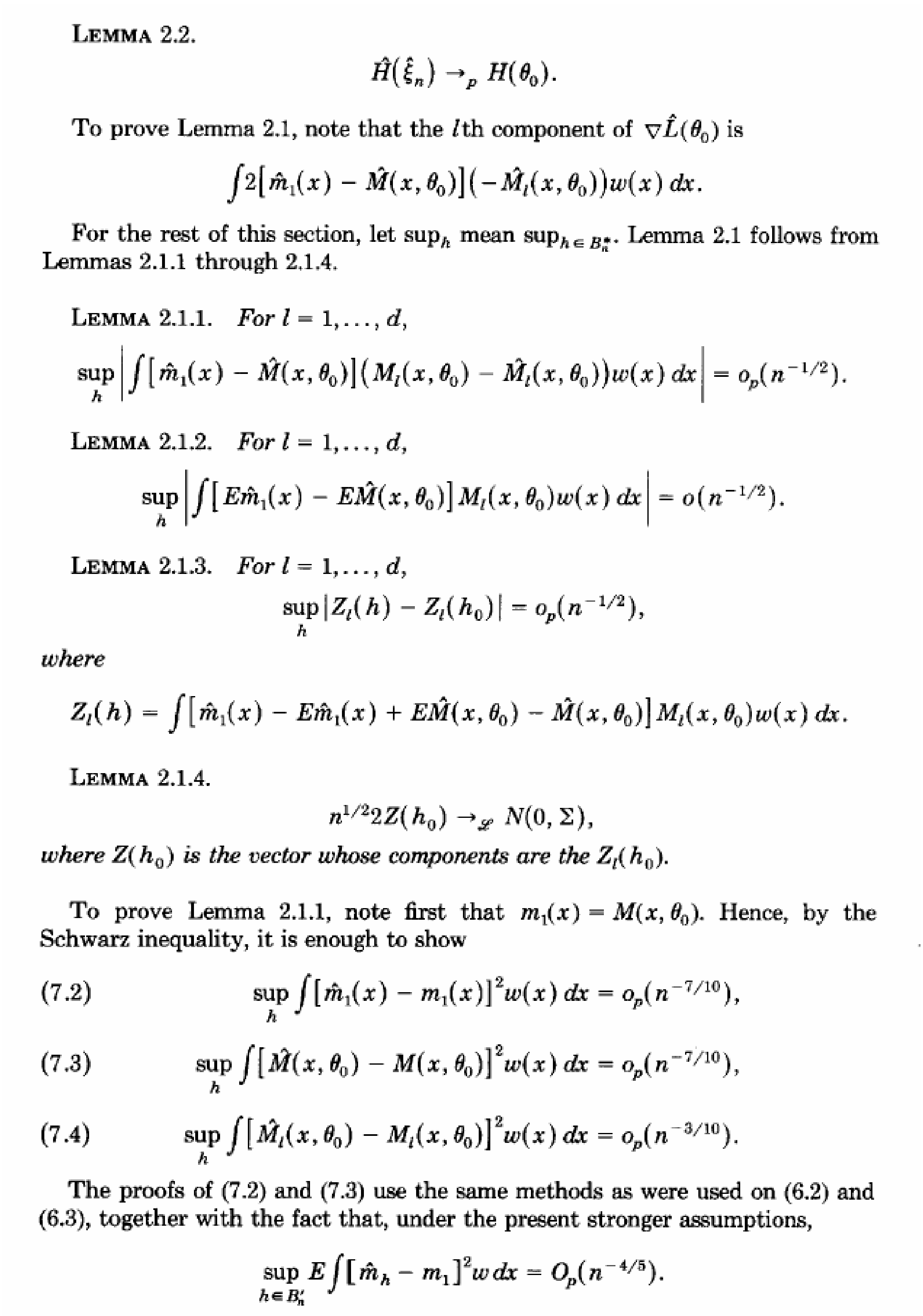
7. Proof of Theorem 2. Let $\nabla\hat{L}(\theta)$ denote the d-dimensional vector of partial derivatives, $\hat{L}_l(\theta) = (\partial/\partial\theta^{(l)})\hat{L}(\theta)$. Note that

$$0 = \nabla\hat{L}(\hat{\theta}) = \nabla\hat{L}(\theta_0) + \hat{H}(\hat{\xi}_n)(\hat{\theta} - \theta_0), \tag{7.1}$$

where $\hat{H}(\theta)$ is the Hessian matrix, whose components are $\hat{L}_{l,l'}(\theta) = (\partial/\partial\theta^{(l')})\hat{L}_l(\theta)$, and where $\hat{\xi}_n$ lies on a line segment connecting $\hat{\theta}$ and θ_0. Theorem 2 is a consequence of (7.1), together with the following lemmas.

LEMMA 2.1.

$$\sqrt{n}\,\nabla\hat{L}(\theta_0) \to_{\mathscr{L}} N(0,\Sigma).$$



LEMMA 2.2.

$$\hat{H}(\hat{\xi}_n) \to_p H(\theta_0).$$

To prove Lemma 2.1, note that the lth component of $\nabla \hat{L}(\theta_0)$ is

$$\int 2\left[\hat{m}_1(x) - \hat{M}(x, \theta_0)\right]\left(-\hat{M}_l(x, \theta_0)\right) w(x)\, dx.$$

For the rest of this section, let $\sup_h$ mean $\sup_{h \in B_n^*}$. Lemma 2.1 follows from Lemmas 2.1.1 through 2.1.4.

LEMMA 2.1.1. *For $l = 1, \dots, d$,*

$$\sup_h \left| \int [\hat{m}_1(x) - \hat{M}(x, \theta_0)]\left(M_l(x, \theta_0) - \hat{M}_l(x, \theta_0)\right) w(x)\, dx \right| = o_p(n^{-1/2}).$$

LEMMA 2.1.2. *For $l = 1, \dots, d$,*

$$\sup_h \left| \int [E\hat{m}_1(x) - E\hat{M}(x, \theta_0)]\, M_l(x, \theta_0) w(x)\, dx \right| = o(n^{-1/2}).$$

LEMMA 2.1.3. *For $l = 1, \dots, d$,*

$$\sup_h |Z_l(h) - Z_l(h_0)| = o_p(n^{-1/2}),$$

where

$$Z_l(h) = \int [\hat{m}_1(x) - E\hat{m}_1(x) + E\hat{M}(x, \theta_0) - \hat{M}(x, \theta_0)]\, M_l(x, \theta_0) w(x)\, dx.$$

LEMMA 2.1.4.

$$n^{1/2} 2 Z(h_0) \to_{\mathscr{L}} N(0, \Sigma),$$

where $Z(h_0)$ is the vector whose components are the $Z_l(h_0)$.

To prove Lemma 2.1.1, note first that $m_1(x) = M(x, \theta_0)$. Hence, by the Schwarz inequality, it is enough to show

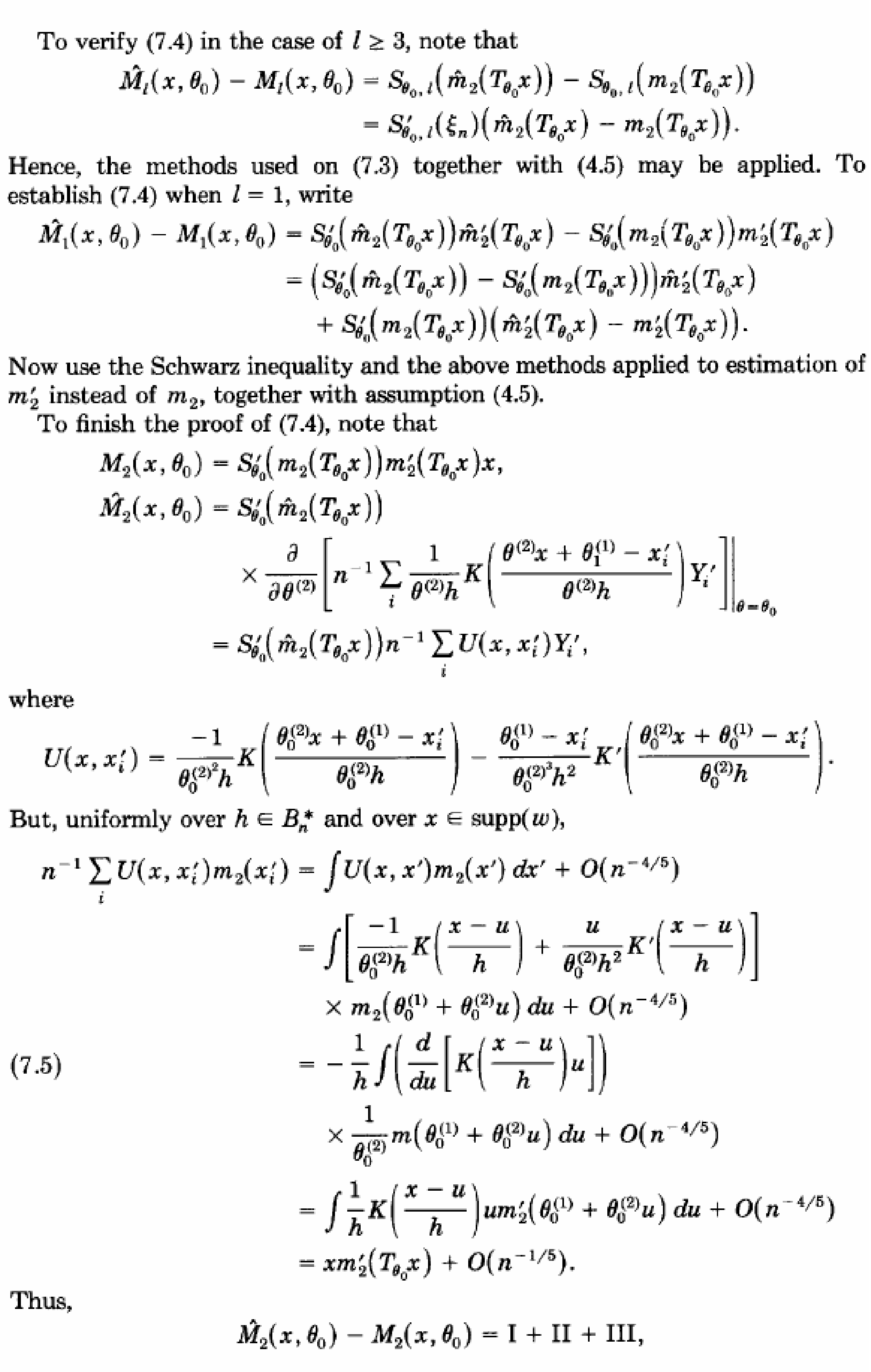
$$\sup_h \int [\hat{m}_1(x) - m_1(x)]^2 w(x)\, dx = o_p(n^{-7/10}), \tag{7.2}$$

$$\sup_h \int [\hat{M}(x, \theta_0) - M(x, \theta_0)]^2 w(x)\, dx = o_p(n^{-7/10}), \tag{7.3}$$

$$\sup_h \int [\hat{M}_l(x, \theta_0) - M_l(x, \theta_0)]^2 w(x)\, dx = o_p(n^{-3/10}). \tag{7.4}$$

The proofs of (7.2) and (7.3) use the same methods as were used on (6.2) and (6.3), together with the fact that, under the present stronger assumptions,

$$\sup_{h \in B_n'} E \int [\hat{m}_h - m_1]^2 w\, dx = O_p(n^{-4/5}).$$



To verify (7.4) in the case of $l \geq 3$, note that

$$\begin{aligned}\hat{M}_l(x,\theta_0) - M_l(x,\theta_0) &= S_{\theta_0,l}\big(\hat{m}_2(T_{\theta_0}x)\big) - S_{\theta_0,l}\big(m_2(T_{\theta_0}x)\big)\\ &= S'_{\theta_0,l}(\xi_n)\big(\hat{m}_2(T_{\theta_0}x) - m_2(T_{\theta_0}x)\big).\end{aligned}$$

Hence, the methods used on (7.3) together with (4.5) may be applied. To establish (7.4) when $l = 1$, write

$$\begin{aligned}\hat{M}_1(x,\theta_0) - M_1(x,\theta_0) &= S'_{\theta_0}\big(\hat{m}_2(T_{\theta_0}x)\big)\hat{m}_2'(T_{\theta_0}x) - S'_{\theta_0}\big(m_2(T_{\theta_0}x)\big)m_2'(T_{\theta_0}x)\\ &= \Big(S'_{\theta_0}\big(\hat{m}_2(T_{\theta_0}x)\big) - S'_{\theta_0}\big(m_2(T_{\theta_0}x)\big)\Big)\hat{m}_2'(T_{\theta_0}x)\\ &\quad + S'_{\theta_0}\big(m_2(T_{\theta_0}x)\big)\big(\hat{m}_2'(T_{\theta_0}x) - m_2'(T_{\theta_0}x)\big).\end{aligned}$$

Now use the Schwarz inequality and the above methods applied to estimation of m_2' instead of m_2, together with assumption (4.5).

To finish the proof of (7.4), note that

$$\begin{aligned}M_2(x,\theta_0) &= S'_{\theta_0}\big(m_2(T_{\theta_0}x)\big)m_2'(T_{\theta_0}x)x,\\ \hat{M}_2(x,\theta_0) &= S'_{\theta_0}\big(\hat{m}_2(T_{\theta_0}x)\big)\\ &\quad\times \frac{\partial}{\partial\theta^{(2)}}\left[n^{-1}\sum_i \frac{1}{\theta^{(2)}h}K\left(\frac{\theta^{(2)}x + \theta_1^{(1)} - x_i'}{\theta^{(2)}h}\right)Y_i'\right]\Bigg|_{\theta=\theta_0}\\ &= S'_{\theta_0}\big(\hat{m}_2(T_{\theta_0}x)\big)n^{-1}\sum_i U(x,x_i')Y_i',\end{aligned}$$

where

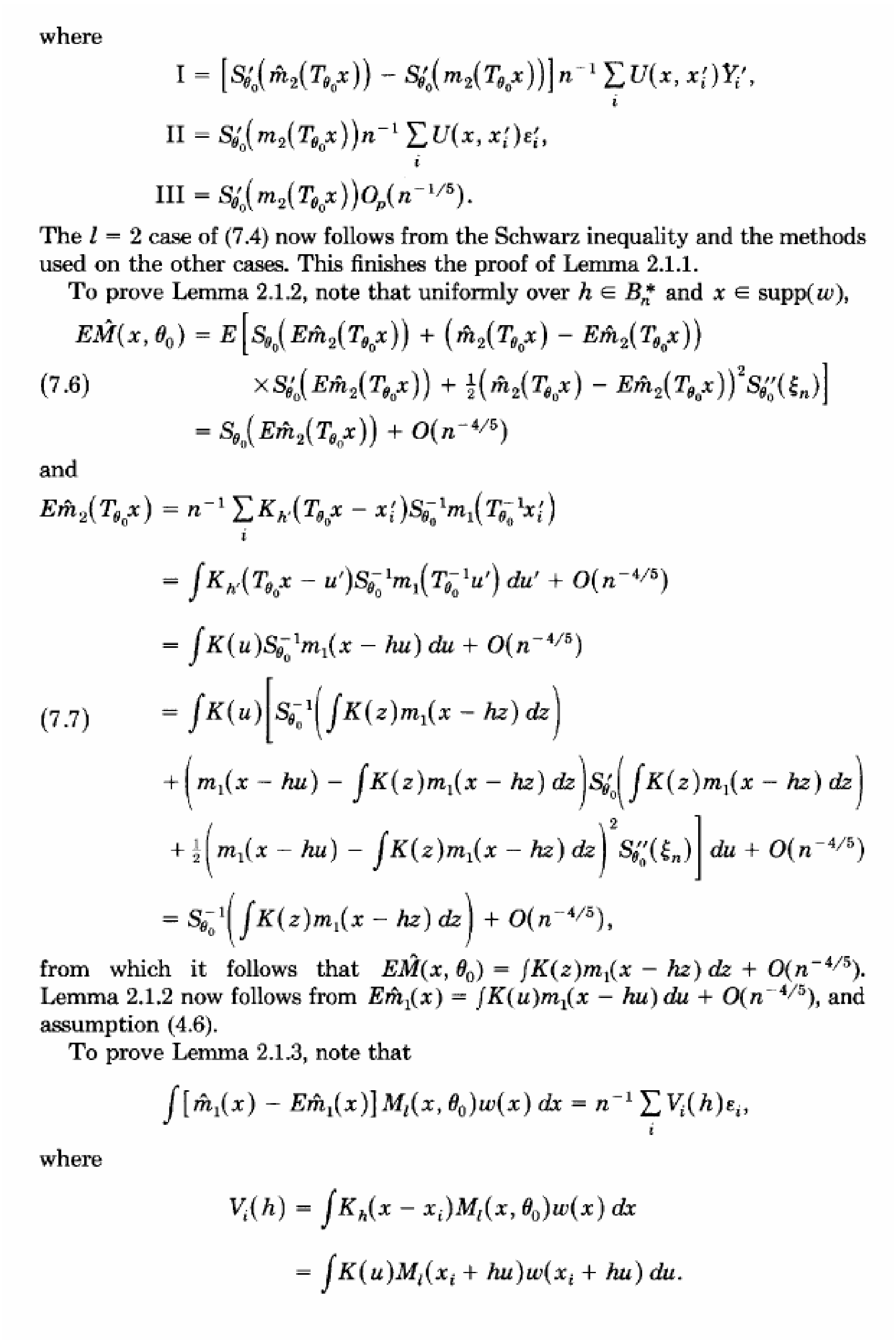
$$U(x,x_i') = \frac{-1}{\theta_0^{(2)^2}h}K\left(\frac{\theta_0^{(2)}x + \theta_0^{(1)} - x_i'}{\theta_0^{(2)}h}\right) - \frac{\theta_0^{(1)} - x_i'}{\theta_0^{(2)^3}h^2}K'\left(\frac{\theta_0^{(2)}x + \theta_0^{(1)} - x_i'}{\theta_0^{(2)}h}\right).$$

But, uniformly over $h \in B_n^*$ and over $x \in \operatorname{supp}(w)$,

$$\begin{aligned}n^{-1}\sum_i U(x,x_i')m_2(x_i') &= \int U(x,x')m_2(x')\,dx' + O(n^{-4/5})\\ &= \int\left[\frac{-1}{\theta_0^{(2)}h}K\left(\frac{x-u}{h}\right) + \frac{u}{\theta_0^{(2)}h^2}K'\left(\frac{x-u}{h}\right)\right]\\ &\quad\times m_2\big(\theta_0^{(1)} + \theta_0^{(2)}u\big)\,du + O(n^{-4/5})\\ &= -\frac{1}{h}\int\left(\frac{d}{du}\left[K\left(\frac{x-u}{h}\right)u\right]\right)\\ &\quad\times\frac{1}{\theta_0^{(2)}}m\big(\theta_0^{(1)} + \theta_0^{(2)}u\big)\,du + O(n^{-4/5})\\ &= \int\frac{1}{h}K\left(\frac{x-u}{h}\right)um_2'\big(\theta_0^{(1)} + \theta_0^{(2)}u\big)\,du + O(n^{-4/5})\\ &= xm_2'(T_{\theta_0}x) + O(n^{-1/5}).\end{aligned}\tag{7.5}$$

Thus,

$$\hat{M}_2(x,\theta_0) - M_2(x,\theta_0) = \mathrm{I} + \mathrm{II} + \mathrm{III},$$



where

$$\mathrm{I} = \left[S'_{\theta_0}\left(\hat{m}_2(T_{\theta_0}x)\right) - S'_{\theta_0}\left(m_2(T_{\theta_0}x)\right)\right] n^{-1} \sum_i U(x, x'_i) Y'_i,$$

$$\mathrm{II} = S'_{\theta_0}\left(m_2(T_{\theta_0}x)\right) n^{-1} \sum_i U(x, x'_i) \varepsilon'_i,$$

$$\mathrm{III} = S'_{\theta_0}\left(m_2(T_{\theta_0}x)\right) O_p(n^{-1/5}).$$

The $l = 2$ case of (7.4) now follows from the Schwarz inequality and the methods used on the other cases. This finishes the proof of Lemma 2.1.1.

To prove Lemma 2.1.2, note that uniformly over $h \in B_n^*$ and $x \in \operatorname{supp}(w)$,

$$\begin{aligned} E\hat{M}(x, \theta_0) &= E\Big[S_{\theta_0}\left(E\hat{m}_2(T_{\theta_0}x)\right) + \left(\hat{m}_2(T_{\theta_0}x) - E\hat{m}_2(T_{\theta_0}x)\right) \\ &\qquad \times S'_{\theta_0}\left(E\hat{m}_2(T_{\theta_0}x)\right) + \tfrac{1}{2}\left(\hat{m}_2(T_{\theta_0}x) - E\hat{m}_2(T_{\theta_0}x)\right)^2 S''_{\theta_0}(\xi_n)\Big] \\ &= S_{\theta_0}\left(E\hat{m}_2(T_{\theta_0}x)\right) + O(n^{-4/5}) \end{aligned} \tag{7.6}$$

and

$$\begin{aligned} E\hat{m}_2(T_{\theta_0}x) &= n^{-1} \sum_i K_{h'}(T_{\theta_0}x - x'_i) S^{-1}_{\theta_0} m_1\left(T^{-1}_{\theta_0} x'_i\right) \\ &= \int K_{h'}(T_{\theta_0}x - u') S^{-1}_{\theta_0} m_1\left(T^{-1}_{\theta_0} u'\right) du' + O(n^{-4/5}) \\ &= \int K(u) S^{-1}_{\theta_0} m_1(x - hu)\, du + O(n^{-4/5}) \\ &= \int K(u)\Big[S^{-1}_{\theta_0}\left(\int K(z) m_1(x - hz)\, dz\right) \\ &\quad + \left(m_1(x - hu) - \int K(z) m_1(x - hz)\, dz\right) S'_{\theta_0}\left(\int K(z) m_1(x - hz)\, dz\right) \\ &\quad + \tfrac{1}{2}\left(m_1(x - hu) - \int K(z) m_1(x - hz)\, dz\right)^2 S''_{\theta_0}(\xi_n)\Big]\, du + O(n^{-4/5}) \\ &= S^{-1}_{\theta_0}\left(\int K(z) m_1(x - hz)\, dz\right) + O(n^{-4/5}), \end{aligned} \tag{7.7}$$

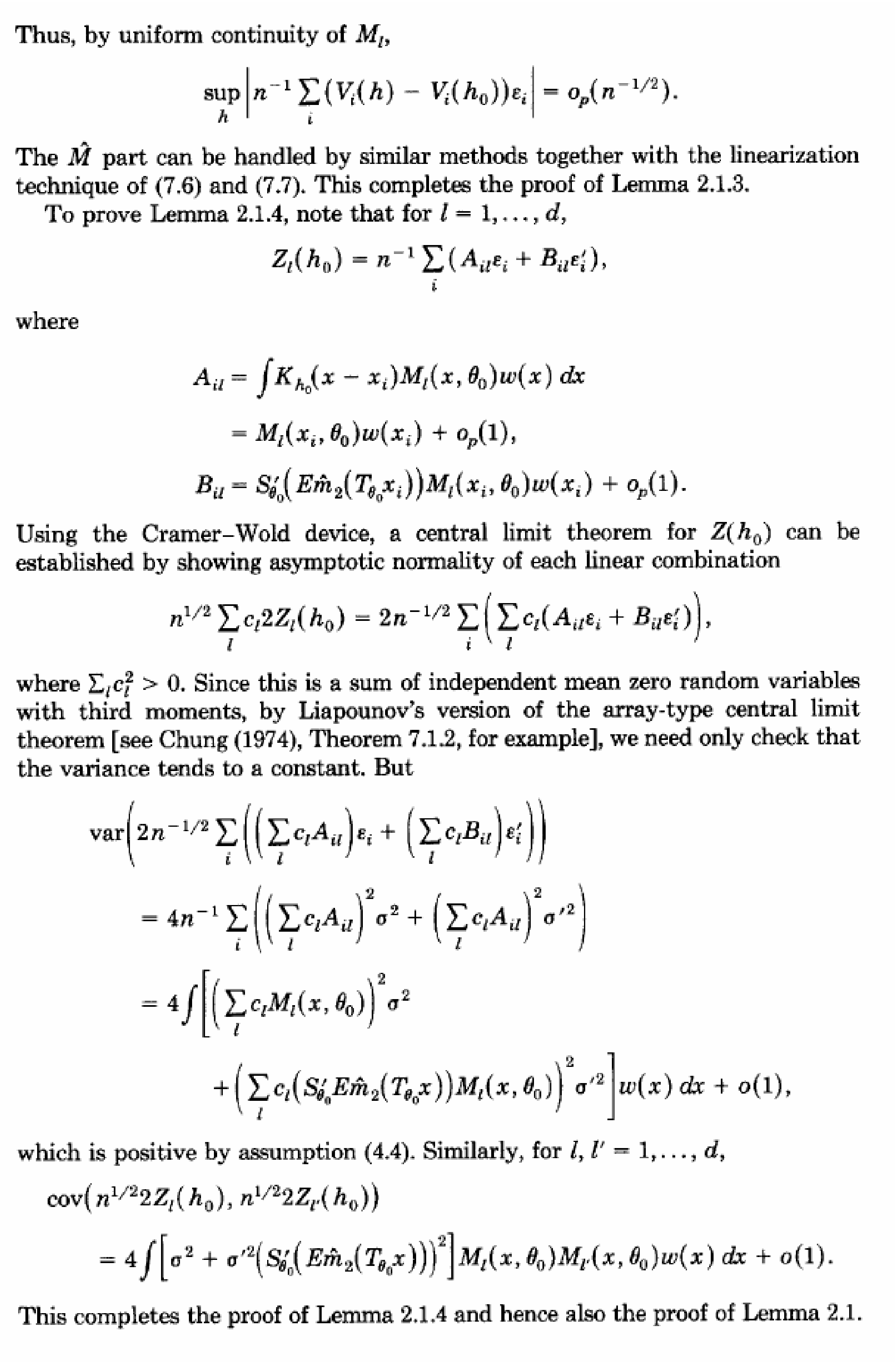
from which it follows that $E\hat{M}(x, \theta_0) = \int K(z) m_1(x - hz)\, dz + O(n^{-4/5})$. Lemma 2.1.2 now follows from $E\hat{m}_1(x) = \int K(u) m_1(x - hu)\, du + O(n^{-4/5})$, and assumption (4.6).

To prove Lemma 2.1.3, note that

$$\int [\hat{m}_1(x) - E\hat{m}_1(x)] M_l(x, \theta_0) w(x)\, dx = n^{-1} \sum_i V_i(h) \varepsilon_i,$$

where

$$\begin{aligned} V_i(h) &= \int K_h(x - x_i) M_l(x, \theta_0) w(x)\, dx \\ &= \int K(u) M_l(x_i + hu) w(x_i + hu)\, du. \end{aligned}$$



Thus, by uniform continuity of M_l,

$$\sup_h \left| n^{-1} \sum_i (V_i(h) - V_i(h_0))\varepsilon_i \right| = o_p(n^{-1/2}).$$

The $\hat{M}$ part can be handled by similar methods together with the linearization technique of (7.6) and (7.7). This completes the proof of Lemma 2.1.3.

To prove Lemma 2.1.4, note that for $l = 1, \ldots, d$,

$$Z_l(h_0) = n^{-1} \sum_i (A_{il}\varepsilon_i + B_{il}\varepsilon_i'),$$

where

$$\begin{aligned} A_{il} &= \int K_{h_0}(x - x_i) M_l(x, \theta_0) w(x)\, dx \\ &= M_l(x_i, \theta_0) w(x_i) + o_p(1), \\ B_{il} &= S'_{\theta_0}\big(E\hat{m}_2(T_{\theta_0} x_i)\big) M_l(x_i, \theta_0) w(x_i) + o_p(1). \end{aligned}$$

Using the Cramer–Wold device, a central limit theorem for $Z(h_0)$ can be established by showing asymptotic normality of each linear combination

$$n^{1/2} \sum_l c_l 2 Z_l(h_0) = 2n^{-1/2} \sum_i \left(\sum_l c_l (A_{il}\varepsilon_i + B_{il}\varepsilon_i') \right),$$

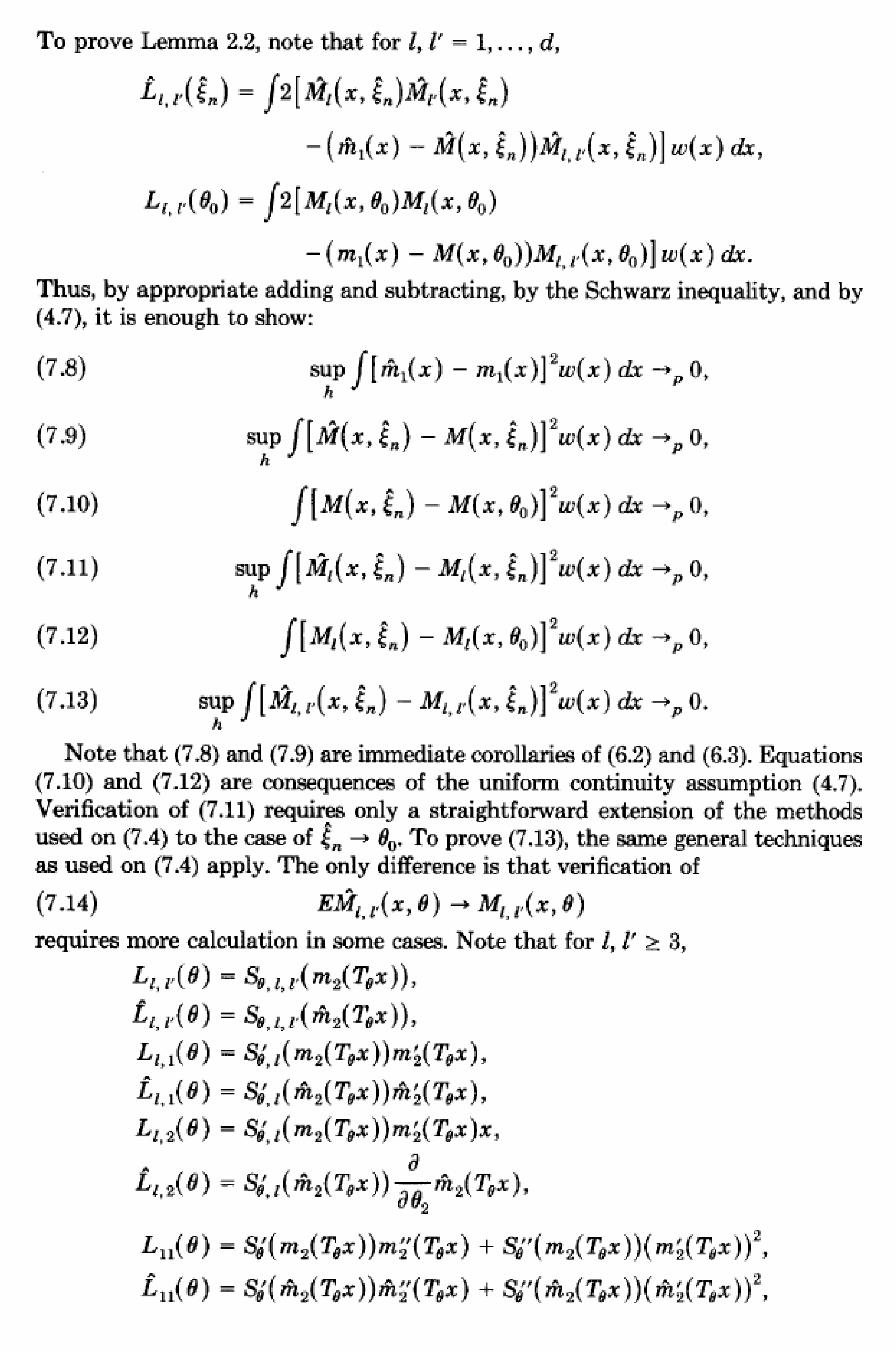
where $\sum_l c_l^2 > 0$. Since this is a sum of independent mean zero random variables with third moments, by Liapounov's version of the array-type central limit theorem [see Chung (1974), Theorem 7.1.2, for example], we need only check that the variance tends to a constant. But

$$\begin{aligned} &\operatorname{var}\left(2n^{-1/2} \sum_i \left(\left(\sum_l c_l A_{il} \right) \varepsilon_i + \left(\sum_l c_l B_{il} \right) \varepsilon_i' \right) \right) \\ &\quad = 4n^{-1} \sum_i \left(\left(\sum_l c_l A_{il} \right)^2 \sigma^2 + \left(\sum_l c_l A_{il} \right)^2 \sigma'^2 \right) \\ &\quad = 4 \int \Bigg[\left(\sum_l c_l M_l(x, \theta_0) \right)^2 \sigma^2 \\ &\qquad\qquad + \left(\sum_l c_l \big(S'_{\theta_0} E\hat{m}_2(T_{\theta_0} x) \big) M_l(x, \theta_0) \right)^2 \sigma'^2 \Bigg] w(x)\, dx + o(1), \end{aligned}$$

which is positive by assumption (4.4). Similarly, for $l, l' = 1, \ldots, d$,

$$\begin{aligned} &\operatorname{cov}\big(n^{1/2} 2 Z_l(h_0), n^{1/2} 2 Z_{l'}(h_0) \big) \\ &\quad = 4 \int \left[\sigma^2 + \sigma'^2 \big(S'_{\theta_0}\big(E\hat{m}_2(T_{\theta_0} x) \big) \big)^2 \right] M_l(x, \theta_0) M_{l'}(x, \theta_0) w(x)\, dx + o(1). \end{aligned}$$

This completes the proof of Lemma 2.1.4 and hence also the proof of Lemma 2.1.



To prove Lemma 2.2, note that for $l, l' = 1, \ldots, d$,

$$\hat{L}_{l,l'}(\hat{\xi}_n) = \int 2\big[\hat{M}_l(x,\hat{\xi}_n)\hat{M}_{l'}(x,\hat{\xi}_n)$$
$$\qquad - \big(\hat{m}_1(x) - \hat{M}(x,\hat{\xi}_n)\big)\hat{M}_{l,l'}(x,\hat{\xi}_n)\big]\, w(x)\, dx,$$
$$L_{l,l'}(\theta_0) = \int 2\big[M_l(x,\theta_0)M_l(x,\theta_0)$$
$$\qquad - \big(m_1(x) - M(x,\theta_0)\big)M_{l,l'}(x,\theta_0)\big]\, w(x)\, dx.$$

Thus, by appropriate adding and subtracting, by the Schwarz inequality, and by (4.7), it is enough to show:

$$\sup_h \int [\hat{m}_1(x) - m_1(x)]^2 w(x)\, dx \to_p 0, \tag{7.8}$$

$$\sup_h \int \big[\hat{M}(x,\hat{\xi}_n) - M(x,\hat{\xi}_n)\big]^2 w(x)\, dx \to_p 0, \tag{7.9}$$

$$\int \big[M(x,\hat{\xi}_n) - M(x,\theta_0)\big]^2 w(x)\, dx \to_p 0, \tag{7.10}$$

$$\sup_h \int \big[\hat{M}_l(x,\hat{\xi}_n) - M_l(x,\hat{\xi}_n)\big]^2 w(x)\, dx \to_p 0, \tag{7.11}$$

$$\int \big[M_l(x,\hat{\xi}_n) - M_l(x,\theta_0)\big]^2 w(x)\, dx \to_p 0, \tag{7.12}$$

$$\sup_h \int \big[\hat{M}_{l,l'}(x,\hat{\xi}_n) - M_{l,l'}(x,\hat{\xi}_n)\big]^2 w(x)\, dx \to_p 0. \tag{7.13}$$

Note that (7.8) and (7.9) are immediate corollaries of (6.2) and (6.3). Equations (7.10) and (7.12) are consequences of the uniform continuity assumption (4.7). Verification of (7.11) requires only a straightforward extension of the methods used on (7.4) to the case of $\hat{\xi}_n \to \theta_0$. To prove (7.13), the same general techniques as used on (7.4) apply. The only difference is that verification of

$$E\hat{M}_{l,l'}(x,\theta) \to M_{l,l'}(x,\theta) \tag{7.14}$$

requires more calculation in some cases. Note that for $l, l' \geq 3$,

$$L_{l,l'}(\theta) = S_{\theta,l,l'}(m_2(T_\theta x)),$$
$$\hat{L}_{l,l'}(\theta) = S_{\theta,l,l'}(\hat{m}_2(T_\theta x)),$$
$$L_{l,1}(\theta) = S'_{\theta,l}(m_2(T_\theta x))m'_2(T_\theta x),$$
$$\hat{L}_{l,1}(\theta) = S'_{\theta,l}(\hat{m}_2(T_\theta x))\hat{m}'_2(T_\theta x),$$
$$L_{l,2}(\theta) = S'_{\theta,l}(m_2(T_\theta x))m'_2(T_\theta x)x,$$
$$\hat{L}_{l,2}(\theta) = S'_{\theta,l}(\hat{m}_2(T_\theta x))\frac{\partial}{\partial\theta_2}\hat{m}_2(T_\theta x),$$
$$L_{11}(\theta) = S'_\theta(m_2(T_\theta x))m''_2(T_\theta x) + S''_\theta(m_2(T_\theta x))(m'_2(T_\theta x))^2,$$
$$\hat{L}_{11}(\theta) = S'_\theta(\hat{m}_2(T_\theta x))\hat{m}''_2(T_\theta x) + S''_\theta(\hat{m}_2(T_\theta x))(\hat{m}'_2(T_\theta x))^2,$$

so these cases can be handled as was done for (7.4). The hard cases arise because

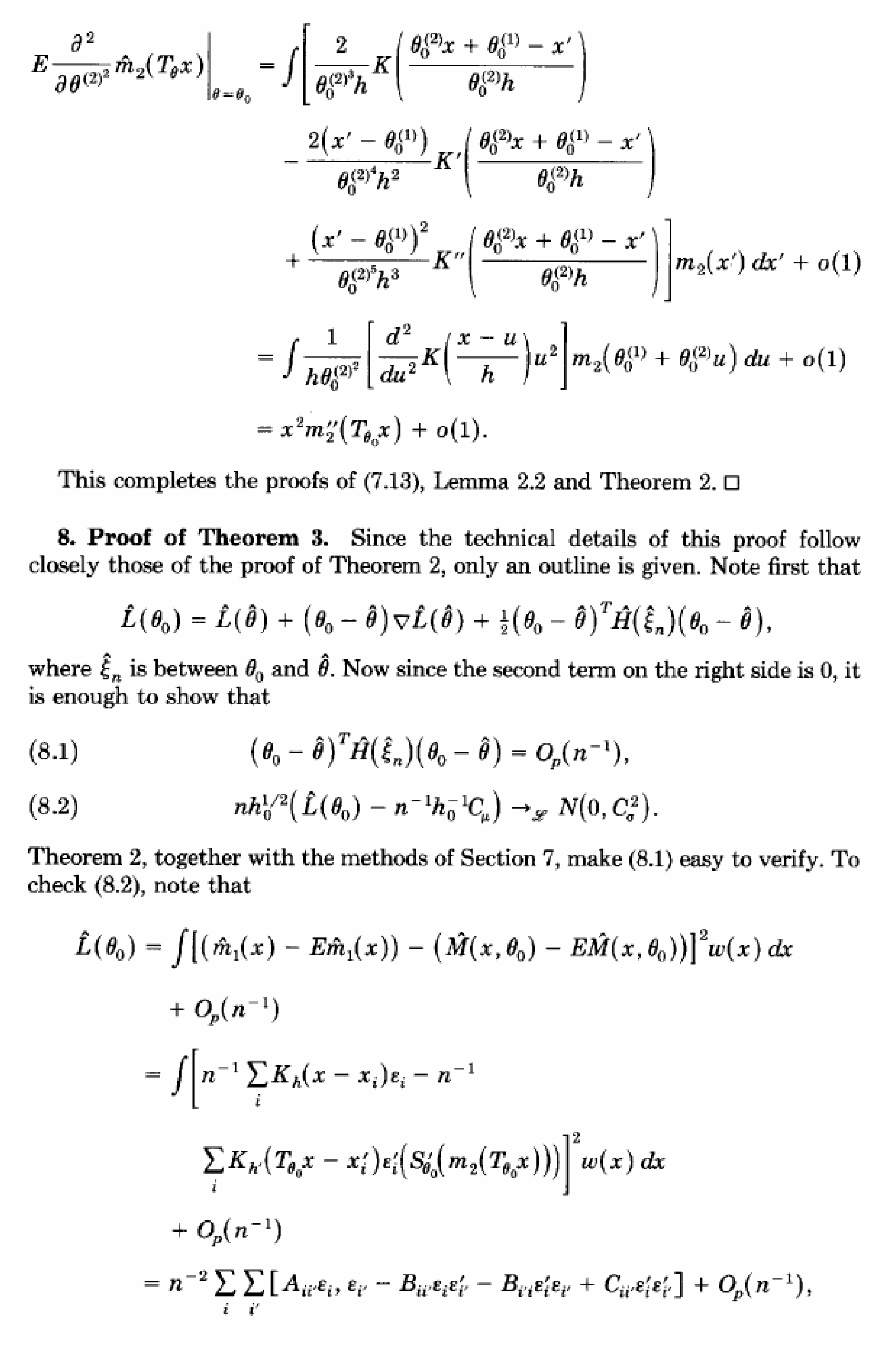
$$\begin{aligned}
L_{1,2}(\theta_0) &= S'_{\theta_0}\big(m_2(T_{\theta_0}x)\big)m_2''(T_{\theta_0}x)x + S''_{\theta_0}\big(m_2(T_{\theta_0}x)\big)\big(m_2'(T_{\theta_0}x)\big)^2 x,\\
\hat{L}_{1,2}(\theta_0) &= S'_{\theta_0}\big(\hat{m}_2(T_{\theta_0}x)\big)\left[\frac{\partial}{\partial\theta^{(2)}}\hat{m}_2'(T_\theta x)\right]_{\theta=\theta_0}\\
&\qquad + S''_{\theta_0}\big(\hat{m}_2(T_{\theta_0}x)\big)\left[\frac{\partial}{\partial\theta^{(2)}}\hat{m}_2(T_\theta x)\right]_{\theta=\theta_0}\hat{m}_2'(T_{\theta_0}x),\\
L_{2,2}(\theta_0) &= S'_{\theta_0}\big(m_2(T_{\theta_0}x)\big)\left[\frac{\partial^2}{\partial\theta^{(2)^2}}m_2(T_\theta x)\right]_{\theta=\theta_0}\\
&\qquad + S''_{\theta_0}\big(m_2(T_{\theta_0}x)\big)\left[\frac{\partial}{\partial\theta^{(2)}}m_2(T_\theta x)\right]_{\theta=\theta_0},\\
\hat{L}_{2,2}(\theta_0) &= S'_{\theta_0}\big(\hat{m}_2(T_{\theta_0}x)\big)\left[\frac{\partial^2}{\partial\theta^{(2)^2}}\hat{m}_2(T_\theta x)\right]_{\theta=\theta_0}\\
&\qquad + S''_{\theta_0}\big(\hat{m}_2(T_{\theta_0}x)\big)\left[\frac{\partial}{\partial\theta^{(2)}}\hat{m}_2(T_\theta x)\right]_{\theta=\theta_0}.
\end{aligned}$$

In view of the work done for (7.4), it remains to show that

$$\begin{aligned}
E\frac{\partial}{\partial\theta^{(2)}}\hat{m}_2(T_\theta x)\bigg|_{\theta=\theta_0} &\to \frac{\partial}{\partial\theta^{(2)}}m_2'(T_\theta x)\bigg|_{\theta=\theta_0} = m_2''(T_{\theta_0}x)x,\\
E\frac{\partial^2}{\partial\theta^{(2)^2}}\hat{m}_2(T_\theta x)\bigg|_{\theta=\theta_0} &\to \frac{\partial^2}{\partial\theta^{(2)^2}}m_2(T_\theta x)\bigg|_{\theta=\theta_0} = m_2''(T_{\theta_0}x)x^2.
\end{aligned}$$

To check these, observe that, as in (7.4),

$$\begin{aligned}
E\frac{\partial}{\partial\theta^{(2)}}\hat{m}_2'(T_\theta x)\bigg|_{\theta=\theta_0} &= \int\left[\frac{-2}{\theta_0^{(2)^3}h^2}K'\left(\frac{\theta_0^{(2)}x + \theta_0^{(1)} - x'}{\theta_0^{(2)}h}\right)\right.\\
&\qquad\left. + \frac{\big(x' - \theta_0^{(1)}\big)}{\theta_0^{(2)^4}h^3}K''\left(\frac{\theta_0^{(2)}x + \theta_0^{(1)} - x'}{\theta_0^{(2)}h}\right)\right]m_2(x')\,dx' + o(1)\\
&= \int\frac{1}{\theta_0^{(2)^2}h}\frac{d^2}{du^2}\left[K\left(\frac{x-u}{h}\right)u\right]m_2\big(\theta_0^{(1)} + \theta_0^{(2)}u\big)\,du + o(1)\\
&= \int K_h(x-u)um_2''\big(\theta_0^{(1)} + \theta_0^{(2)}u\big)\,du + o(1)\\
&= xm_2''(T_{\theta_0}x) + o(1),
\end{aligned}$$



$$E\frac{\partial^2}{\partial\theta^{(2)^2}}\hat{m}_2(T_\theta x)\bigg|_{\theta=\theta_0} = \int\Bigg[\frac{2}{\theta_0^{(2)^3}h}K\left(\frac{\theta_0^{(2)}x+\theta_0^{(1)}-x'}{\theta_0^{(2)}h}\right)$$

$$-\frac{2\left(x'-\theta_0^{(1)}\right)}{\theta_0^{(2)^4}h^2}K'\left(\frac{\theta_0^{(2)}x+\theta_0^{(1)}-x'}{\theta_0^{(2)}h}\right)$$

$$+\frac{\left(x'-\theta_0^{(1)}\right)^2}{\theta_0^{(2)^5}h^3}K''\left(\frac{\theta_0^{(2)}x+\theta_0^{(1)}-x'}{\theta_0^{(2)}h}\right)\Bigg]m_2(x')\,dx'+o(1)$$

$$=\int\frac{1}{h\theta_0^{(2)^2}}\left[\frac{d^2}{du^2}K\left(\frac{x-u}{h}\right)u^2\right]m_2\left(\theta_0^{(1)}+\theta_0^{(2)}u\right)du+o(1)$$

$$=x^2m_2''(T_{\theta_0}x)+o(1).$$

This completes the proofs of (7.13), Lemma 2.2 and Theorem 2. □

8. Proof of Theorem 3. Since the technical details of this proof follow closely those of the proof of Theorem 2, only an outline is given. Note first that

$$\hat{L}(\theta_0)=\hat{L}(\hat{\theta})+(\theta_0-\hat{\theta})\nabla\hat{L}(\hat{\theta})+\tfrac{1}{2}(\theta_0-\hat{\theta})^T\hat{H}(\hat{\xi}_n)(\theta_0-\hat{\theta}),$$

where $\hat{\xi}_n$ is between θ_0 and $\hat{\theta}$. Now since the second term on the right side is 0, it is enough to show that

$$(\theta_0-\hat{\theta})^T\hat{H}(\hat{\xi}_n)(\theta_0-\hat{\theta})=O_p(n^{-1}), \tag{8.1}$$

$$nh_0^{1/2}\left(\hat{L}(\theta_0)-n^{-1}h_0^{-1}C_\mu\right)\to_{\mathscr{L}}N(0,C_\sigma^2). \tag{8.2}$$

Theorem 2, together with the methods of Section 7, make (8.1) easy to verify. To check (8.2), note that

$$\hat{L}(\theta_0)=\int\left[(\hat{m}_1(x)-E\hat{m}_1(x))-\left(\hat{M}(x,\theta_0)-E\hat{M}(x,\theta_0)\right)\right]^2w(x)\,dx$$

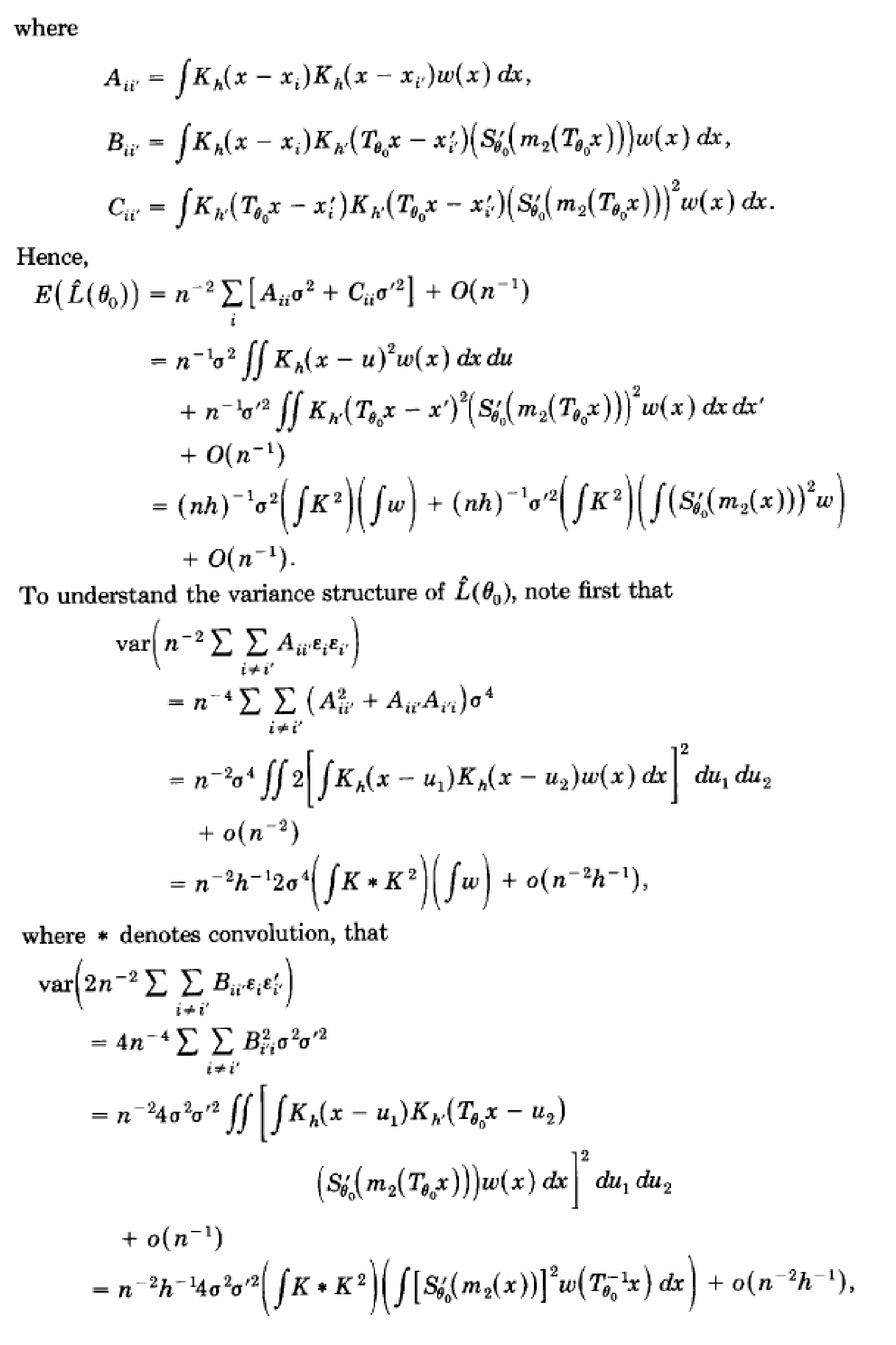
$$+O_p(n^{-1})$$

$$=\int\Bigg[n^{-1}\sum_i K_h(x-x_i)\varepsilon_i-n^{-1}$$

$$\sum_i K_{h'}(T_{\theta_0}x-x_i')\varepsilon_i'\left(S_{\theta_0}'\left(m_2(T_{\theta_0}x)\right)\right)\Bigg]^2w(x)\,dx$$

$$+O_p(n^{-1})$$

$$=n^{-2}\sum_i\sum_{i'}\left[A_{ii'}\varepsilon_i,\varepsilon_{i'}-B_{ii'}\varepsilon_i\varepsilon_{i'}'-B_{i'i}\varepsilon_i'\varepsilon_{i'}+C_{ii'}\varepsilon_i'\varepsilon_{i'}'\right]+O_p(n^{-1}),$$



where

$$A_{ii'} = \int K_h(x - x_i)K_h(x - x_{i'})w(x)\,dx,$$

$$B_{ii'} = \int K_h(x - x_i)K_{h'}(T_{\theta_0}x - x'_{i'})\big(S'_{\theta_0}\big(m_2(T_{\theta_0}x)\big)\big)w(x)\,dx,$$

$$C_{ii'} = \int K_{h'}(T_{\theta_0}x - x'_i)K_{h'}(T_{\theta_0}x - x'_{i'})\big(S'_{\theta_0}\big(m_2(T_{\theta_0}x)\big)\big)^2 w(x)\,dx.$$

Hence,

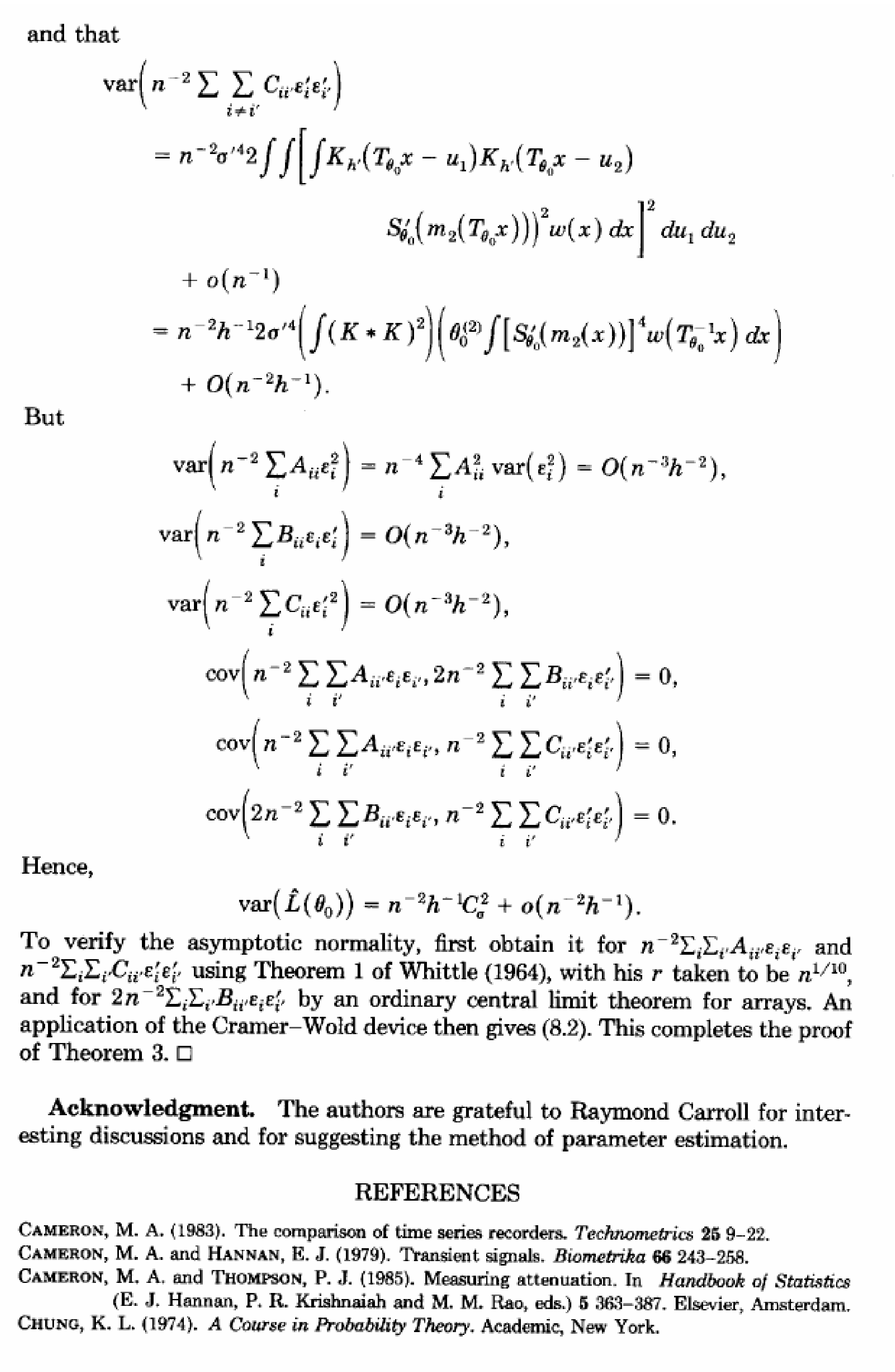
$$\begin{aligned}
E\big(\hat{L}(\theta_0)\big) &= n^{-2}\sum_i \big[A_{ii}\sigma^2 + C_{ii}\sigma'^2\big] + O(n^{-1}) \\
&= n^{-1}\sigma^2 \iint K_h(x-u)^2 w(x)\,dx\,du \\
&\quad + n^{-1}\sigma'^2 \iint K_{h'}(T_{\theta_0}x - x')^2\big(S'_{\theta_0}\big(m_2(T_{\theta_0}x)\big)\big)^2 w(x)\,dx\,dx' \\
&\quad + O(n^{-1}) \\
&= (nh)^{-1}\sigma^2\Big(\int K^2\Big)\Big(\int w\Big) + (nh)^{-1}\sigma'^2\Big(\int K^2\Big)\Big(\int \big(S'_{\theta_0}(m_2(x))\big)^2 w\Big) \\
&\quad + O(n^{-1}).
\end{aligned}$$

To understand the variance structure of $\hat{L}(\theta_0)$, note first that

$$\begin{aligned}
&\operatorname{var}\Big(n^{-2}\sum\sum_{i\neq i'} A_{ii'}\varepsilon_i\varepsilon_{i'}\Big) \\
&\quad = n^{-4}\sum\sum_{i\neq i'}\big(A_{ii'}^2 + A_{ii'}A_{i'i}\big)\sigma^4 \\
&\quad = n^{-2}\sigma^4 \iint 2\Big[\int K_h(x-u_1)K_h(x-u_2)w(x)\,dx\Big]^2 du_1\,du_2 \\
&\qquad + o(n^{-2}) \\
&\quad = n^{-2}h^{-1}2\sigma^4\Big(\int K * K^2\Big)\Big(\int w\Big) + o(n^{-2}h^{-1}),
\end{aligned}$$

where $*$ denotes convolution, that

$$\begin{aligned}
&\operatorname{var}\Big(2n^{-2}\sum\sum_{i\neq i'} B_{ii'}\varepsilon_i\varepsilon'_{i'}\Big) \\
&\quad = 4n^{-4}\sum\sum_{i\neq i'} B_{i'i}^2\sigma^2\sigma'^2 \\
&\quad = n^{-2}4\sigma^2\sigma'^2 \iint\Big[\int K_h(x-u_1)K_{h'}(T_{\theta_0}x - u_2) \\
&\qquad\qquad \big(S'_{\theta_0}\big(m_2(T_{\theta_0}x)\big)\big)w(x)\,dx\Big]^2 du_1\,du_2 \\
&\qquad + o(n^{-1}) \\
&\quad = n^{-2}h^{-1}4\sigma^2\sigma'^2\Big(\int K * K^2\Big)\Big(\int\big[S'_{\theta_0}(m_2(x))\big]^2 w\big(T_{\theta_0}^{-1}x\big)\,dx\Big) + o(n^{-2}h^{-1}),
\end{aligned}$$



and that

$$\operatorname{var}\left(n^{-2}\sum_{i \neq i'}\sum C_{ii'}\varepsilon_i'\varepsilon_{i'}'\right)$$

$$= n^{-2}\sigma'^4 2\int\int\left[\int K_{h'}(T_{\theta_0}x - u_1)K_{h'}(T_{\theta_0}x - u_2)\right.$$

$$\left.S_{\theta_0}'\big(m_2(T_{\theta_0}x)\big)^2 w(x)\,dx\right]^2 du_1\,du_2$$

$$+ o(n^{-1})$$

$$= n^{-2}h^{-1}2\sigma'^4\left(\int (K * K)^2\right)\left(\theta_0^{(2)}\int\left[S_{\theta_0}'(m_2(x))\right]^4 w\left(T_{\theta_0}^{-1}x\right)dx\right)$$

$$+ O(n^{-2}h^{-1}).$$

But

$$\operatorname{var}\left(n^{-2}\sum_i A_{ii}\varepsilon_i^2\right) = n^{-4}\sum_i A_{ii}^2 \operatorname{var}(\varepsilon_i^2) = O(n^{-3}h^{-2}),$$

$$\operatorname{var}\left(n^{-2}\sum_i B_{ii}\varepsilon_i\varepsilon_i'\right) = O(n^{-3}h^{-2}),$$

$$\operatorname{var}\left(n^{-2}\sum_i C_{ii}\varepsilon_i'^2\right) = O(n^{-3}h^{-2}),$$

$$\operatorname{cov}\left(n^{-2}\sum_i\sum_{i'}A_{ii'}\varepsilon_i\varepsilon_{i'}, 2n^{-2}\sum_i\sum_{i'}B_{ii'}\varepsilon_i\varepsilon_{i'}'\right) = 0,$$

$$\operatorname{cov}\left(n^{-2}\sum_i\sum_{i'}A_{ii'}\varepsilon_i\varepsilon_{i'}, n^{-2}\sum_i\sum_{i'}C_{ii'}\varepsilon_i'\varepsilon_{i'}'\right) = 0,$$

$$\operatorname{cov}\left(2n^{-2}\sum_i\sum_{i'}B_{ii'}\varepsilon_i\varepsilon_{i'}, n^{-2}\sum_i\sum_{i'}C_{ii'}\varepsilon_i'\varepsilon_{i'}'\right) = 0.$$

Hence,

$$\operatorname{var}\big(\hat{L}(\theta_0)\big) = n^{-2}h^{-1}C_\sigma^2 + o(n^{-2}h^{-1}).$$

To verify the asymptotic normality, first obtain it for $n^{-2}\Sigma_i\Sigma_{i'}A_{ii'}\varepsilon_i\varepsilon_{i'}$ and $n^{-2}\Sigma_i\Sigma_{i'}C_{ii'}\varepsilon_i'\varepsilon_{i'}'$ using Theorem 1 of Whittle (1964), with his r taken to be $n^{1/10}$, and for $2n^{-2}\Sigma_i\Sigma_{i'}B_{ii'}\varepsilon_i\varepsilon_{i'}'$ by an ordinary central limit theorem for arrays. An application of the Cramer–Wold device then gives (8.2). This completes the proof of Theorem 3. □

Acknowledgment. The authors are grateful to Raymond Carroll for interesting discussions and for suggesting the method of parameter estimation.

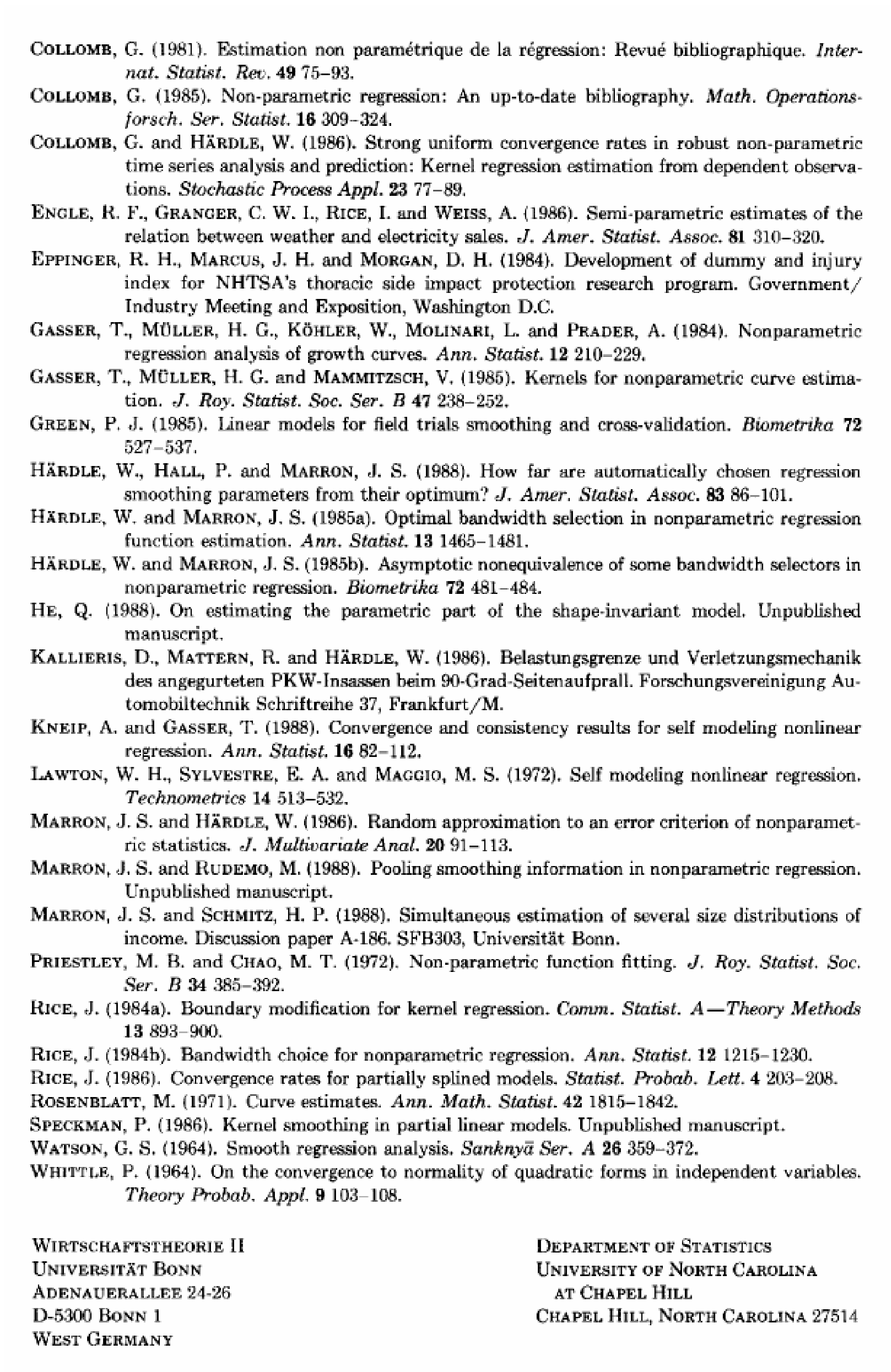
REFERENCES

CAMERON, M. A. (1983). The comparison of time series recorders. *Technometrics* **25** 9–22.

CAMERON, M. A. and HANNAN, E. J. (1979). Transient signals. *Biometrika* **66** 243–258.

CAMERON, M. A. and THOMPSON, P. J. (1985). Measuring attenuation. In *Handbook of Statistics* (E. J. Hannan, P. R. Krishnaiah and M. M. Rao, eds.) **5** 363–387. Elsevier, Amsterdam.

CHUNG, K. L. (1974). *A Course in Probability Theory*. Academic, New York.



COLLOMB, G. (1981). Estimation non paramétrique de la régression: Revué bibliographique. *Internat. Statist. Rev.* **49** 75–93.

COLLOMB, G. (1985). Non-parametric regression: An up-to-date bibliography. *Math. Operationsforsch. Ser. Statist.* **16** 309–324.

COLLOMB, G. and HÄRDLE, W. (1986). Strong uniform convergence rates in robust non-parametric time series analysis and prediction: Kernel regression estimation from dependent observations. *Stochastic Process Appl.* **23** 77–89.

ENGLE, R. F., GRANGER, C. W. I., RICE, I. and WEISS, A. (1986). Semi-parametric estimates of the relation between weather and electricity sales. *J. Amer. Statist. Assoc.* **81** 310–320.

EPPINGER, R. H., MARCUS, J. H. and MORGAN, D. H. (1984). Development of dummy and injury index for NHTSA's thoracic side impact protection research program. Government/Industry Meeting and Exposition, Washington D.C.

GASSER, T., MÜLLER, H. G., KÖHLER, W., MOLINARI, L. and PRADER, A. (1984). Nonparametric regression analysis of growth curves. *Ann. Statist.* **12** 210–229.

GASSER, T., MÜLLER, H. G. and MAMMITZSCH, V. (1985). Kernels for nonparametric curve estimation. *J. Roy. Statist. Soc. Ser. B* **47** 238–252.

GREEN, P. J. (1985). Linear models for field trials smoothing and cross-validation. *Biometrika* **72** 527–537.

HÄRDLE, W., HALL, P. and MARRON, J. S. (1988). How far are automatically chosen regression smoothing parameters from their optimum? *J. Amer. Statist. Assoc.* **83** 86–101.

HÄRDLE, W. and MARRON, J. S. (1985a). Optimal bandwidth selection in nonparametric regression function estimation. *Ann. Statist.* **13** 1465–1481.

HÄRDLE, W. and MARRON, J. S. (1985b). Asymptotic nonequivalence of some bandwidth selectors in nonparametric regression. *Biometrika* **72** 481–484.

HE, Q. (1988). On estimating the parametric part of the shape-invariant model. Unpublished manuscript.

KALLIERIS, D., MATTERN, R. and HÄRDLE, W. (1986). Belastungsgrenze und Verletzungsmechanik des angegurteten PKW-Insassen beim 90-Grad-Seitenaufprall. Forschungsvereinigung Automobiltechnik Schriftreihe 37, Frankfurt/M.

KNEIP, A. and GASSER, T. (1988). Convergence and consistency results for self modeling nonlinear regression. *Ann. Statist.* **16** 82–112.

LAWTON, W. H., SYLVESTRE, E. A. and MAGGIO, M. S. (1972). Self modeling nonlinear regression. *Technometrics* **14** 513–532.

MARRON, J. S. and HÄRDLE, W. (1986). Random approximation to an error criterion of nonparametric statistics. *J. Multivariate Anal.* **20** 91–113.

MARRON, J. S. and RUDEMO, M. (1988). Pooling smoothing information in nonparametric regression. Unpublished manuscript.

MARRON, J. S. and SCHMITZ, H. P. (1988). Simultaneous estimation of several size distributions of income. Discussion paper A-186. SFB303, Universität Bonn.

PRIESTLEY, M. B. and CHAO, M. T. (1972). Non-parametric function fitting. *J. Roy. Statist. Soc. Ser. B* **34** 385–392.

RICE, J. (1984a). Boundary modification for kernel regression. *Comm. Statist. A—Theory Methods* **13** 893–900.

RICE, J. (1984b). Bandwidth choice for nonparametric regression. *Ann. Statist.* **12** 1215–1230.

RICE, J. (1986). Convergence rates for partially splined models. *Statist. Probab. Lett.* **4** 203–208.

ROSENBLATT, M. (1971). Curve estimates. *Ann. Math. Statist.* **42** 1815–1842.

SPECKMAN, P. (1986). Kernel smoothing in partial linear models. Unpublished manuscript.

WATSON, G. S. (1964). Smooth regression analysis. *Sanknyā Ser. A* **26** 359–372.

WHITTLE, P. (1964). On the convergence to normality of quadratic forms in independent variables. *Theory Probab. Appl.* **9** 103–108.

WIRTSCHAFTSTHEORIE II
UNIVERSITÄT BONN
ADENAUERALLEE 24-26
D-5300 BONN 1
WEST GERMANY

DEPARTMENT OF STATISTICS
UNIVERSITY OF NORTH CAROLINA
AT CHAPEL HILL
CHAPEL HILL, NORTH CAROLINA 27514

Östereichische Zeitschrift für Statistik und Informatik, 20, 53-64.

Wolfgang Härdle
Raymond James Carroll

Biased Crossvalidation for a Kernel Regression Estimator and its Derivatives *)

Vortrag, gehalten im Rahmen des Arbeitskreises 'Statistische Kolloquien' am 19. April 1989.

For univariate nonparametric regression, we compute the mean squared error of a kernel regression estimator and its derivatives (GASSER and MÜLLER, 1984), extending slightly the conditions of applicability of this estimator. We show how to estimate this mean squared error and thus the best smoothing parameter by what SCOTT and TERRELL (1987) call biased crossvalidation, which is essentially a refined version of the "plug-in" method. This bandwidth estimator is shown to be asymptotically optimal in the sense of HÄRDLE and MARRON (1985).

1. Introduction

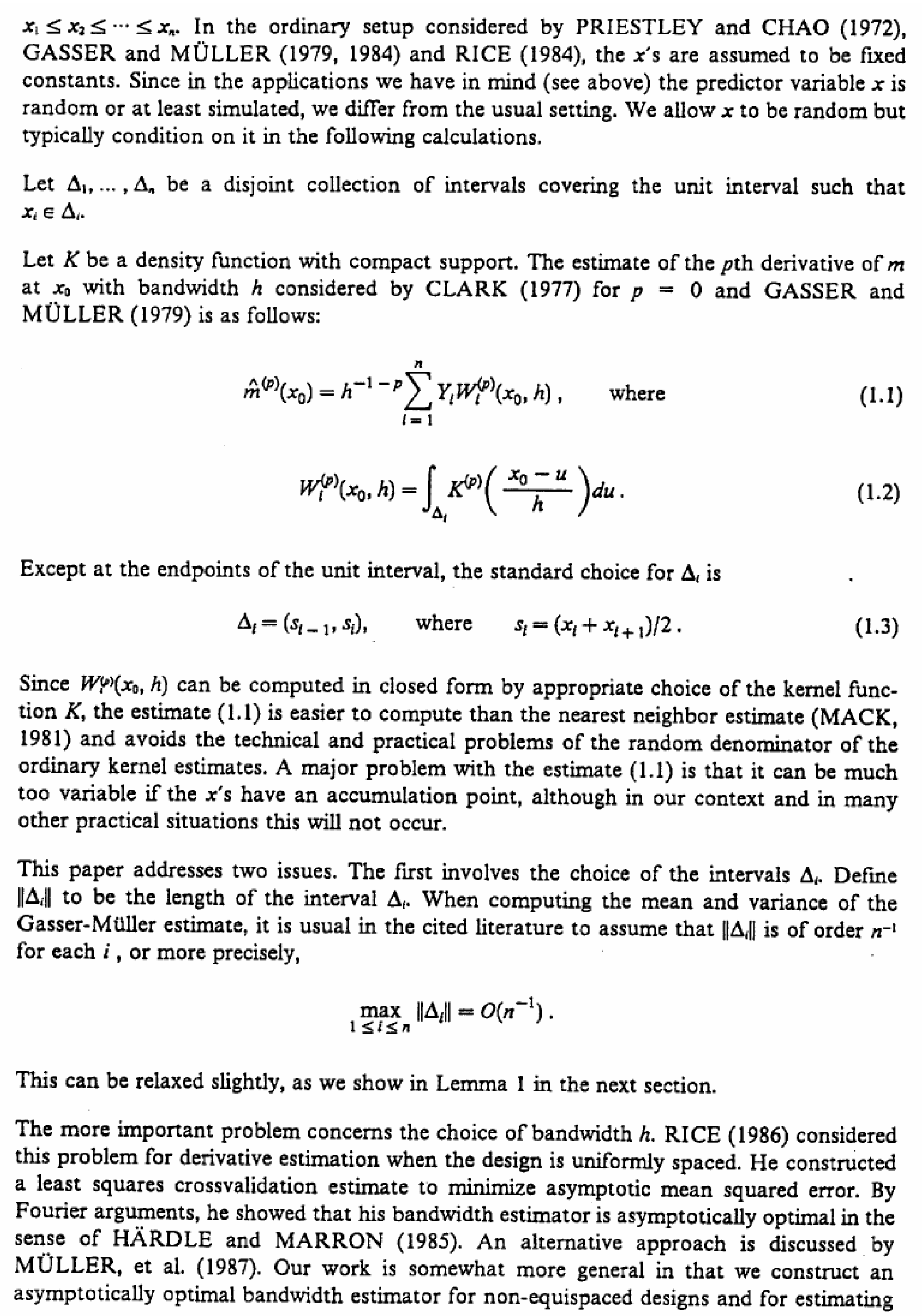
Suppose that given x, Y has regression function $E(Y \mid x) = m(x)$ and variance function $\sigma^2(x) = var(Y \mid x)$. We consider the estimation of $m(x)$ and its derivatives using a kernel regression estimate.

One motivation for our problem arises in chemistry, where estimation of the regression function and its first derivative is of interest, see LUCCHESE (1985). His experiments are based upon physical models for gas-surface scattering, with x representing parameters (input factors) in his model. The experimental method is a Monte-Carlo simulation based upon the underlying physical model. The input factors x are thus ours to control, and may be generated in a fashion chosen by the experimenter. Lucchese uses a uniform density over a specific range. The values x might not be generated uniformly or with equal spacing, but rather a random design might be appropriate.

Another motivation stems from the empirical verification of the law of demand, where derivatives of mean demand curves occur in the estimation procedure, see HILDENBRAND and HILDENBRAND (1986). Mean demand curves are regression functions of $Y =$ demand for some good against $X =$ income computable from a cross-section of an economy.

The setting for the estimate is as follows. Assume that the values of x are confined to the unit interval, and suppose we have n independent observation (Y_i, x_i), with

*) Research supported by the Air Force Office of Scientific Research, the National Institutes of Health and Sonderforschungsbereich 303, Universität Bonn.



$x_1 \le x_2 \le \cdots \le x_n$. In the ordinary setup considered by PRIESTLEY and CHAO (1972), GASSER and MÜLLER (1979, 1984) and RICE (1984), the x's are assumed to be fixed constants. Since in the applications we have in mind (see above) the predictor variable x is random or at least simulated, we differ from the usual setting. We allow x to be random but typically condition on it in the following calculations.

Let $\Delta_1, \dots, \Delta_n$ be a disjoint collection of intervals covering the unit interval such that $x_i \in \Delta_i$.

Let K be a density function with compact support. The estimate of the pth derivative of m at x_0 with bandwidth h considered by CLARK (1977) for $p = 0$ and GASSER and MÜLLER (1979) is as follows:

$$\hat{m}^{(p)}(x_0) = h^{-1-p} \sum_{i=1}^{n} Y_i W_i^{(p)}(x_0, h)\,, \qquad \text{where} \tag{1.1}$$

$$W_i^{(p)}(x_0, h) = \int_{\Delta_i} K^{(p)}\left(\frac{x_0 - u}{h}\right) du\,. \tag{1.2}$$

Except at the endpoints of the unit interval, the standard choice for Δ_i is

$$\Delta_i = (s_{i-1}, s_i), \qquad \text{where} \qquad s_i = (x_i + x_{i+1})/2\,. \tag{1.3}$$

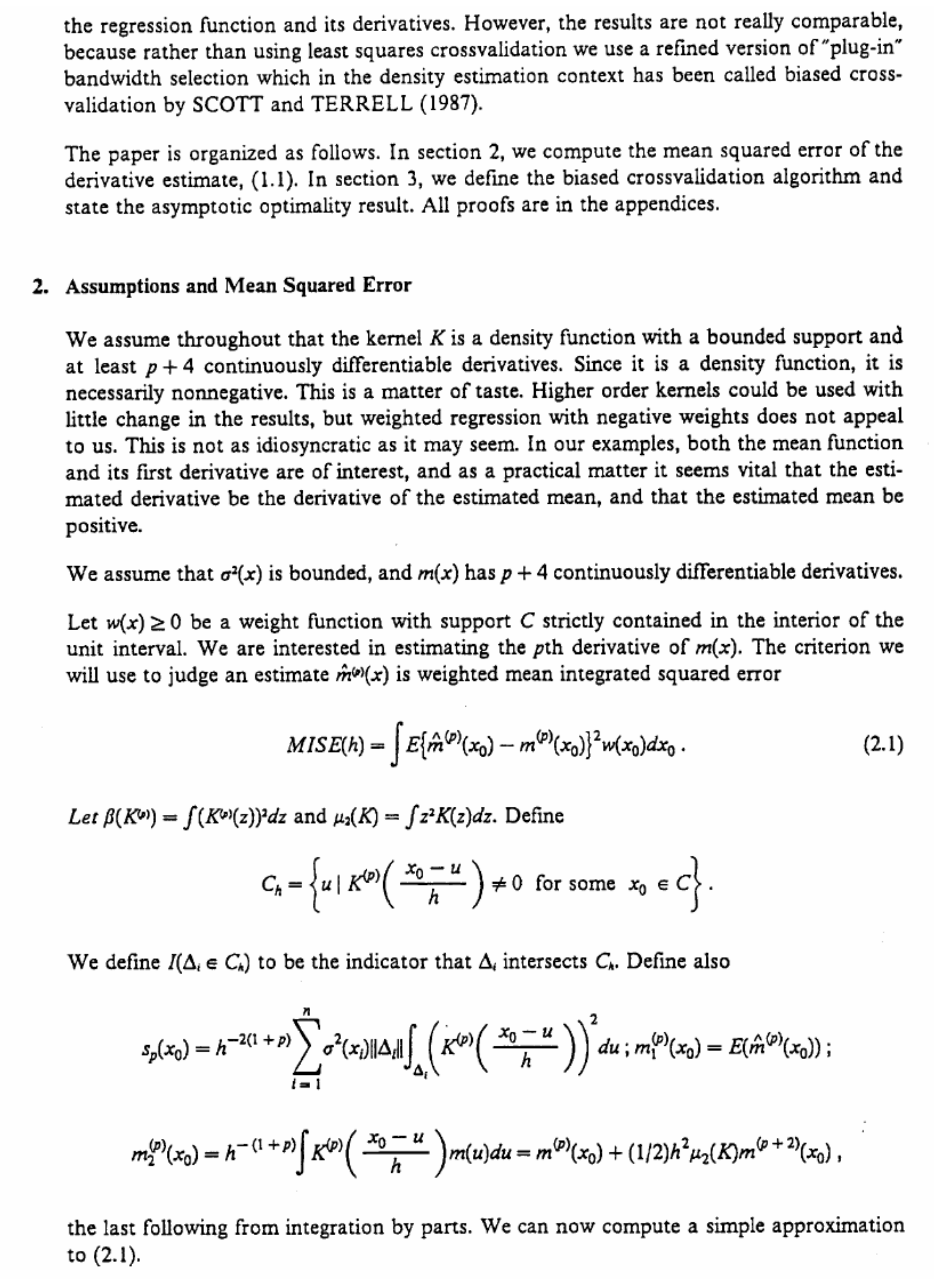
Since $W_i^{(p)}(x_0, h)$ can be computed in closed form by appropriate choice of the kernel function K, the estimate (1.1) is easier to compute than the nearest neighbor estimate (MACK, 1981) and avoids the technical and practical problems of the random denominator of the ordinary kernel estimates. A major problem with the estimate (1.1) is that it can be much too variable if the x's have an accumulation point, although in our context and in many other practical situations this will not occur.

This paper addresses two issues. The first involves the choice of the intervals Δ_i. Define $\|\Delta_i\|$ to be the length of the interval Δ_i. When computing the mean and variance of the Gasser-Müller estimate, it is usual in the cited literature to assume that $\|\Delta_i\|$ is of order n^{-1} for each i, or more precisely,

$$\max_{1 \le i \le n} \|\Delta_i\| = O(n^{-1})\,.$$

This can be relaxed slightly, as we show in Lemma 1 in the next section.

The more important problem concerns the choice of bandwidth h. RICE (1986) considered this problem for derivative estimation when the design is uniformly spaced. He constructed a least squares crossvalidation estimate to minimize asymptotic mean squared error. By Fourier arguments, he showed that his bandwidth estimator is asymptotically optimal in the sense of HÄRDLE and MARRON (1985). An alternative approach is discussed by MÜLLER, et al. (1987). Our work is somewhat more general in that we construct an asymptotically optimal bandwidth estimator for non-equispaced designs and for estimating



the regression function and its derivatives. However, the results are not really comparable, because rather than using least squares crossvalidation we use a refined version of "plug-in" bandwidth selection which in the density estimation context has been called biased crossvalidation by SCOTT and TERRELL (1987).

The paper is organized as follows. In section 2, we compute the mean squared error of the derivative estimate, (1.1). In section 3, we define the biased crossvalidation algorithm and state the asymptotic optimality result. All proofs are in the appendices.

2. Assumptions and Mean Squared Error

We assume throughout that the kernel K is a density function with a bounded support and at least $p+4$ continuously differentiable derivatives. Since it is a density function, it is necessarily nonnegative. This is a matter of taste. Higher order kernels could be used with little change in the results, but weighted regression with negative weights does not appeal to us. This is not as idiosyncratic as it may seem. In our examples, both the mean function and its first derivative are of interest, and as a practical matter it seems vital that the estimated derivative be the derivative of the estimated mean, and that the estimated mean be positive.

We assume that $\sigma^2(x)$ is bounded, and $m(x)$ has $p+4$ continuously differentiable derivatives.

Let $w(x) \geq 0$ be a weight function with support C strictly contained in the interior of the unit interval. We are interested in estimating the pth derivative of $m(x)$. The criterion we will use to judge an estimate $\hat{m}^{(p)}(x)$ is weighted mean integrated squared error

$$MISE(h) = \int E\{\hat{m}^{(p)}(x_0) - m^{(p)}(x_0)\}^2 w(x_0) dx_0 \,. \tag{2.1}$$

Let $\beta(K^{(p)}) = \int (K^{(p)}(z))^2 dz$ and $\mu_2(K) = \int z^2 K(z) dz$. Define

$$C_h = \left\{ u \mid K^{(p)}\left(\frac{x_0 - u}{h} \right) \neq 0 \text{ for some } x_0 \in C \right\}.$$

We define $I(\Delta_i \in C_h)$ to be the indicator that Δ_i intersects C_h. Define also

$$s_p(x_0) = h^{-2(1+p)} \sum_{i=1}^{n} \sigma^2(x_i) \|\Delta_i\| \int_{\Delta_i} \left(K^{(p)}\left(\frac{x_0 - u}{h} \right) \right)^2 du \,;\; m_1^{(p)}(x_0) = E(\hat{m}^{(p)}(x_0)) \,;$$

$$m_2^{(p)}(x_0) = h^{-(1+p)} \int K^{(p)}\left(\frac{x_0 - u}{h} \right) m(u) du = m^{(p)}(x_0) + (1/2) h^2 \mu_2(K) m^{(p+2)}(x_0) \,,$$

the last following from integration by parts. We can now compute a simple approximation to (2.1).

Lemma 1: Assume that for $k \leq 4$, we have that

$$\sum_{i=1}^{n} \|\Delta_i\|^k I(\Delta_i \in C_h) = O(n^{-k+1}) . \tag{2.2}$$

Then, except for terms of order $O(h^6 + (n^2h^{3+2p})^{-1} + h^2(n^2h^{1+2p})^{-1/2})$,

$$MISE(h) = \int s_p(x_0)w(x_0)dx_0 + ((1/2)\mu_2(K))^2h^4\int(m^{(p+2)}(x_0))^2w(x_0)dx_0 . \tag{2.3}$$

Assumption (2.2) is of course satisfied in the cases considered by GASSER and MÜLLER (1979, 1984) and RICE (1986). It can also be shown to hold almost surely if the intervals Δ_i satisfy (1.3) and if the x_i are the order statistics of a sample from a population whose density and its first two derivatives are bounded. However, (2.2) fails in the case that the design has an isolated mass point in C, the support of the weight function w.

We would like to minimize (2.3), but this requires knowledge of the conditional variance curve $\sigma^2(x)$ and the $(p+2)^{nd}$ derivative $m^{(p+2)}(x)$ of the regression curve. The plug-in method is to estimate these last two functions from the data and plug them into (2.3), which is then minimized. Suppose now that the plug-in bandwidth is chosen so that it balances bias² and variance of $MISE(h)$, which means that $h \sim n^{-1/(5+2p)}$. This method does not work since with this choice of bandwidth $m^{(p+2)}$ is not estimated consistently. A similar phenomenon was observed by SCOTT and TERRELL (1987) in the field of density estimation. We therefore modify the plug-in estimate in order to get a consistent estimate of (2.3). This is done in the next section.

3. Biased Crossvalidation

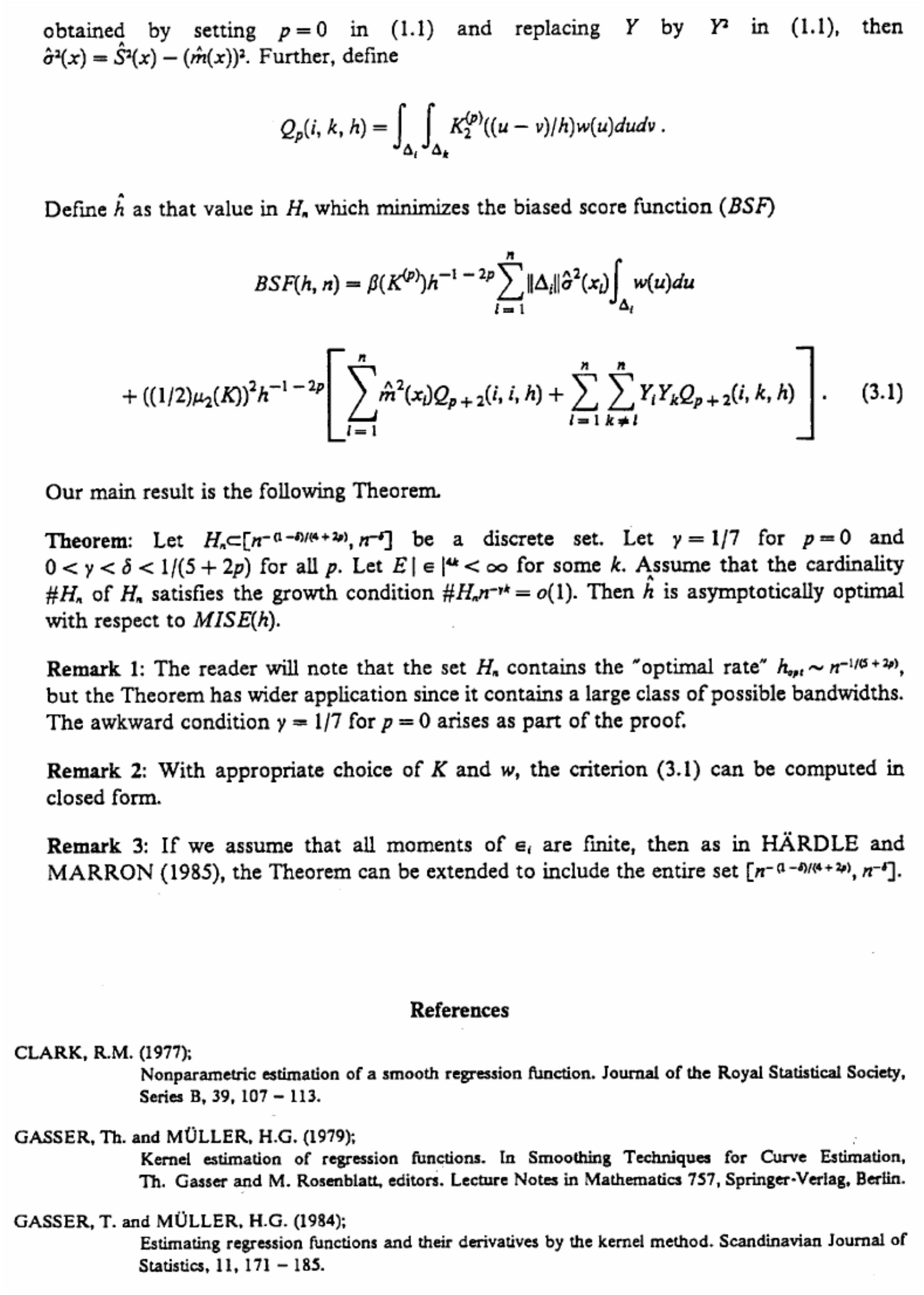
In this section, we define our estimate and h and state the main result. The derivation of the estimate is outlined in Appendix A, while the proof of the main result is given in Appendix B. Let H_n be a discrete set of h's; precise details are given in the statement of the Theorem. Let $h_{\min}$ minimize $MISE(h)$ (see 2.3) in H_n. Following HÄRDLE (1990), we say that a bandwidth selection rule $\hat{h}$ is asymptotically optimal with respect to $MISE(h)$ if

$$\frac{MISE(\hat{h})}{MISE(h_{\min})} \to 1 \text{ in probability}$$

Biased Crossvalidation Algorithm: Define

$$K_2^{(p)}(.) = \int K^{(p)}(z)K^{(p)}(z+.)dz$$

and let $\hat{\sigma}^2(x)$ be a kernel estimate of the variance function, defined through the formula $\sigma^2(x) = E(Y^2 \mid x) - \{E(Y \mid x)\}^2$. That is, if $S^2(x) = E(Y^2 \mid x)$ and $\hat{S}^2(x)$ is the estimate of $S^2(x)$



obtained by setting $p=0$ in (1.1) and replacing Y by Y^2 in (1.1), then $\hat{\sigma}^2(x) = \hat{S}^2(x) - (\hat{m}(x))^2$. Further, define

$$Q_p(i,k,h) = \int_{\Delta_i}\int_{\Delta_k} K_2^{(p)}((u-v)/h)w(u)dudv\,.$$

Define $\hat{h}$ as that value in H_n which minimizes the biased score function (BSF)

$$BSF(h,n) = \beta(K^{(p)})h^{-1-2p}\sum_{l=1}^{n}\|\Delta_l\|\hat{\sigma}^2(x_l)\int_{\Delta_l} w(u)du$$

$$+\,((1/2)\mu_2(K))^2 h^{-1-2p}\left[\sum_{l=1}^{n}\hat{m}^2(x_l)Q_{p+2}(i,i,h) + \sum_{l=1}^{n}\sum_{k\neq l}^{n} Y_l Y_k Q_{p+2}(i,k,h)\right]. \qquad (3.1)$$

Our main result is the following Theorem.

Theorem: Let $H_n \subset [n^{-(1-\delta)/(4+2p)}, n^{-\delta}]$ be a discrete set. Let $\gamma = 1/7$ for $p=0$ and $0<\gamma<\delta<1/(5+2p)$ for all p. Let $E|\in|^{4k} < \infty$ for some k. Assume that the cardinality $\#H_n$ of H_n satisfies the growth condition $\#H_n n^{-\gamma k} = o(1)$. Then $\hat{h}$ is asymptotically optimal with respect to $MISE(h)$.

Remark 1: The reader will note that the set H_n contains the "optimal rate" $h_{opt} \sim n^{-1/(5+2p)}$, but the Theorem has wider application since it contains a large class of possible bandwidths. The awkward condition $\gamma = 1/7$ for $p=0$ arises as part of the proof.

Remark 2: With appropriate choice of K and w, the criterion (3.1) can be computed in closed form.

Remark 3: If we assume that all moments of $\in_i$ are finite, then as in HÄRDLE and MARRON (1985), the Theorem can be extended to include the entire set $[n^{-(1-\delta)/(4+2p)}, n^{-\delta}]$.

References

CLARK, R.M. (1977);
Nonparametric estimation of a smooth regression function. Journal of the Royal Statistical Society, Series B, 39, 107 – 113.

GASSER, Th. and MÜLLER, H.G. (1979);
Kernel estimation of regression functions. In Smoothing Techniques for Curve Estimation, Th. Gasser and M. Rosenblatt, editors. Lecture Notes in Mathematics 757, Springer-Verlag, Berlin.

GASSER, T. and MÜLLER, H.G. (1984);
Estimating regression functions and their derivatives by the kernel method. Scandinavian Journal of Statistics, 11, 171 – 185.

HÄRDLE, W. (1990);
Applied Nonparametric Regression. Cambridge University Press.

HÄRDLE, W. and MARRON, J.S. (1985);
Optimal bandwidth selection in nonparametric regression function estimation. Annals of Statistics, 13, 1 465 – 1 481.

HILDENBRAND, K. and HILDENBRAND, W. (1986);
On the mean income effect: a data analysis of the U.K. Family Expenditure Survey. In Contributions to Mathematical Economics, W. Hildenbrand and A. Mas-Colell, eds. North Holland, Amsterdam.

LUCCHESE, R.R. (1985);
Stochastic sensitivity analysis applied to gas surface scattering. Journal of Chemical Physics, 83, 3 118 – 3 128.

MÜLLER, H.G., STADTMÜLLER, U. and SCHMITT, T. (1987);
Bandwidth choice and confidence intervals for derivatives of noisy data. Biometrika, 74, 743 – 750.

PRIESTLEY, M.B. and CHAO, M.T. (1972);
Nonparametric function fitting. Journal of the Royal Statistical Society, Series B, 34, 385 – 392.

RICE, J. (1986);
Bandwidth choice for differentiation. Journal of Multivariate Analysis, 16, 251 – 264.

SCOTT, D. and TERRELL, G.(1988);
Biased and unbiased crossvalidation in density estimation. Journal of the American Statistical Association, 82, 1 131 – 1 146.

WHITTLE, P. (1960);
Bounds for the moments of linear and quadratic forms in independent variables. Theory of Probability and its Applications, 5, 302 – 305.

Appendix A: Derivation of (3.1)

We now show how we arrived at the bandwidth selection criterion (3.1). Proofs are given at the end of the section. In estimating (2.3), we consider the two terms separately. The two parts of the second term of (3.1) are complex, resulting from estimating the second term in (2.3). We make a few preparatory definitions. Define

$$A(x_0, h, p) = h^{-2(1+p)} \sum_{i=1}^{n} \sum_{k \neq i}^{n} y_i y_k W_i^{(p)}(x_0, h) W_k^{(p)}(x_0, h)\,; \qquad A(h, p) = \int A(x_0, h, p) w(x_0) dx_0\,; \tag{A.1}$$

$$B(h, p) = h^{-1-2p} \sum_{i=1}^{n} m^2(x_i) \int_{\Delta_i} \int_{\Delta_i} K_2^{(p)}\Big(\frac{u-v}{h}\Big) w(u) du dv\,. \tag{A.2}$$

For the second term in (2.3), substituting (1.1) with $p^* = p + 2$ results in a double sum, the cross terms of which add an extra bias term and must be eliminated. If we eliminate these cross terms, then the resulting estimate of the second term in (2.3) is $A(h, p+2)$, see (A.1). The analysis of this term is summarized in the following Lemma.

Define

$$K_3^{(p)}(c, v, h) = \int K^{(p)}(u) K^{(p)}(u+c) w(v+hu) du\,,$$

$$D_h = \{(\Delta_i, \Delta_k) : K_3^{(p)}\Big(\frac{u-v}{h}, u, h\Big) \neq 0 \ \text{ for some } \ u \in \Delta_i, v \in \Delta_k\}\,.$$

In addition to the assumptions of Lemma 1, assume that

$$\sum_{i=1}^{n}\sum_{k\neq i}^{n}\|\Delta_i\|^2\|\Delta_k\|^2 I(\Delta_i \in C_h)I(\Delta_k \in C_h)I((\Delta_i, \Delta_k) \in D_h) = O(hn^{-2})\,. \tag{A.3}$$

Lemma 2: For $h \in H_n$, if $\xi_1(p, n, h) = h^2 + (nh^{1/2+p})^{-1} + h(n^2h^{1+2p})^{-1/4}$, then

$$A(h, p) = \int (m^{(p)}(x_0))^2 w(x_0)dx_0 - B(h, p) + U_{1n}(h) + O(h(nh^{1+2p})^{-1} + \xi_1(p, n, h))\,, \tag{A.4}$$

where for $p \geq 0$,

$$E(U_{1n}(h))^{2k_0}) = O\big(h(nh^{1+2p})^{-1} + h^{1+2p}(nh^{1+2p})^{-3}\big)^{k_0}\,. \tag{A.5}$$

Assumption (A.3) holds for the designs considered by GASSER and MÜLLER (1979, 1984) and by RICE (1986). It also holds almost surely if the x's are a sample from a distribution with a continuously differentiable density.

Since the term $U_{1n}(h) \to 0$ in probability uniformly in H_n, we see that $A(h, p+2)$ is a biased estimate for the integrated bias² term of $MISE(h)$ (see 2.3). In order to construct a consistent estimate, we have to estimate the term $B(h, p+2)$, see (A.2). The obvious method is plug in $\hat{m}(x)$ into (A.2). This substitution in summed up in the following result.

Lemma 3: Make the assumptions of Lemmas 1 and 2. Define $\delta_{i,p} = Q_p(i, i, h)$. Write

$$\xi_2(p, n, h) = h^p(nh^{1+2p})^{-2} + h^2(nh^{1+2p})^{-1} + (n^5h^{3+2p})^{-1}\,.$$

Then, for $h \in H_n$,

$$h^{-1-2p}\sum_{i=1}^{n}\hat{m}^2(x_i)\delta_{i,p} - B(h, p) = h^{-1-2p}\sum_{i=1}^{n}\hat{m}^2(x_i)\delta_{i,p} - h^{-1-2p}\sum_{i=1}^{n}m^2(x_i)\delta_{i,p} = \Omega_n\,,$$

where

$$\Omega_n = O(\xi_2(p, n, h)) + U_{2n}(h)\,; \qquad E(U_{2n}(h))^{2k_0} = O(n^3h^{6+4p})^{k_0}\,. \tag{A.6}$$

We now consider the first term in (3.1). Note that by (2.2),

$$\int s_p(x)w(x)dx = O((nh^{1+2p)-1})) = T(h, n) + O(h(nh^{1+2p})^{-1})\,, \quad \text{where}$$

$$T(h, n) = \beta(K^{(p)})h^{-1-2p}\sum_{i=1}^{n}\|\Delta_i\|\sigma^2(x_i)\int_{\Delta_i} w(u)du\,. \tag{A.7}$$

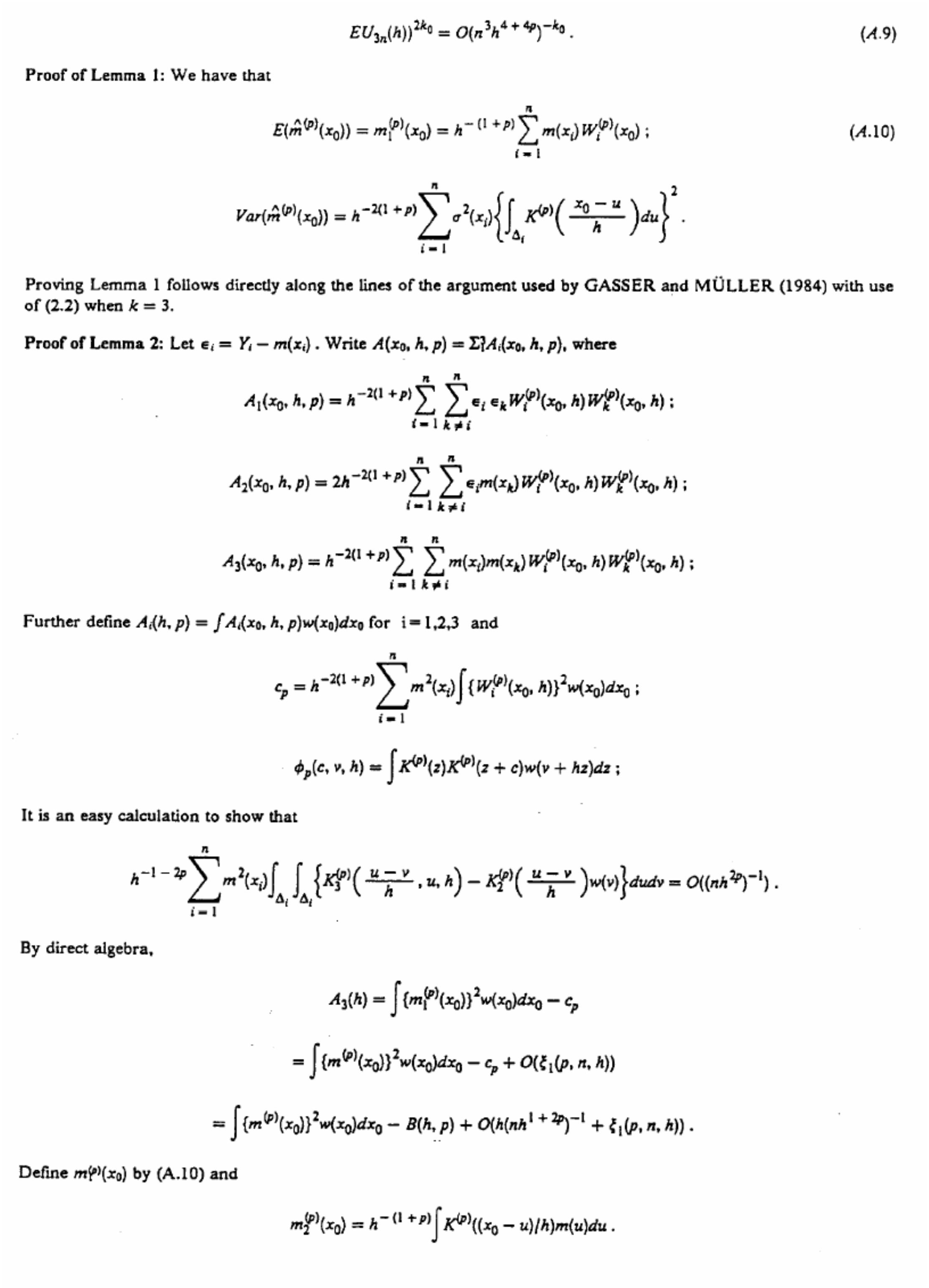
Thus the first term in (3.1) is just $T(h, n)$ with an estimated variance function. To see that it is consistent, we offer the following result.

Lemma 4: Make the assumptions of Lemmas 1 and 2. Let Ω_n be as in Lemma 3. Then, for $h \in H_n$,

$$\hat{T}(h, n) = \beta(K^{(p)})h^{-1-2p}\sum_{i=1}^{n}\|\Delta_i\|\sigma^2(x_i)\int_{\Delta_i} w(u)du$$

$$= \int s_p(x)w(x)dx + \Omega_n + U_{3n}(h) + O(h(nh^{1+2p})^{-1})\,, \tag{A.8}$$

where



$$EU_{3n}(h))^{2k_0} = O(n^3h^{4+4p})^{-k_0}\,. \tag{A.9}$$

Proof of Lemma 1: We have that

$$E(\hat{m}^{(p)}(x_0)) = m_1^{(p)}(x_0) = h^{-(1+p)}\sum_{i=1}^{n} m(x_i)\,W_i^{(p)}(x_0)\,; \tag{A.10}$$

$$Var(\hat{m}^{(p)}(x_0)) = h^{-2(1+p)}\sum_{i=1}^{n}\sigma^2(x_i)\left\{\int_{\Delta_i}K^{(p)}\left(\frac{x_0-u}{h}\right)du\right\}^2.$$

Proving Lemma 1 follows directly along the lines of the argument used by GASSER and MÜLLER (1984) with use of (2.2) when $k = 3$.

Proof of Lemma 2: Let $\epsilon_i = Y_i - m(x_i)$. Write $A(x_0, h, p) = \Sigma_1^3 A_i(x_0, h, p)$, where

$$A_1(x_0, h, p) = h^{-2(1+p)}\sum_{i=1}^{n}\sum_{k\neq i}^{n}\epsilon_i\,\epsilon_k W_i^{(p)}(x_0, h)\,W_k^{(p)}(x_0, h)\,;$$

$$A_2(x_0, h, p) = 2h^{-2(1+p)}\sum_{i=1}^{n}\sum_{k\neq i}^{n}\epsilon_i m(x_k)\,W_i^{(p)}(x_0, h)\,W_k^{(p)}(x_0, h)\,;$$

$$A_3(x_0, h, p) = h^{-2(1+p)}\sum_{i=1}^{n}\sum_{k\neq i}^{n}m(x_i)m(x_k)\,W_i^{(p)}(x_0, h)\,W_k^{(p)}(x_0, h)\,;$$

Further define $A_i(h, p) = \int A_i(x_0, h, p)w(x_0)dx_0$ for i = 1,2,3 and

$$c_p = h^{-2(1+p)}\sum_{i=1}^{n}m^2(x_i)\int\{W_i^{(p)}(x_0, h)\}^2 w(x_0)dx_0\,;$$

$$\phi_p(c, v, h) = \int K^{(p)}(z)K^{(p)}(z + c)w(v + hz)dz\,;$$

It is an easy calculation to show that

$$h^{-1-2p}\sum_{i=1}^{n}m^2(x_i)\int_{\Delta_i}\int_{\Delta_i}\left\{K_3^{(p)}\left(\frac{u-v}{h}, u, h\right) - K_2^{(p)}\left(\frac{u-v}{h}\right)w(v)\right\}dudv = O((nh^{2p})^{-1})\,.$$

By direct algebra,

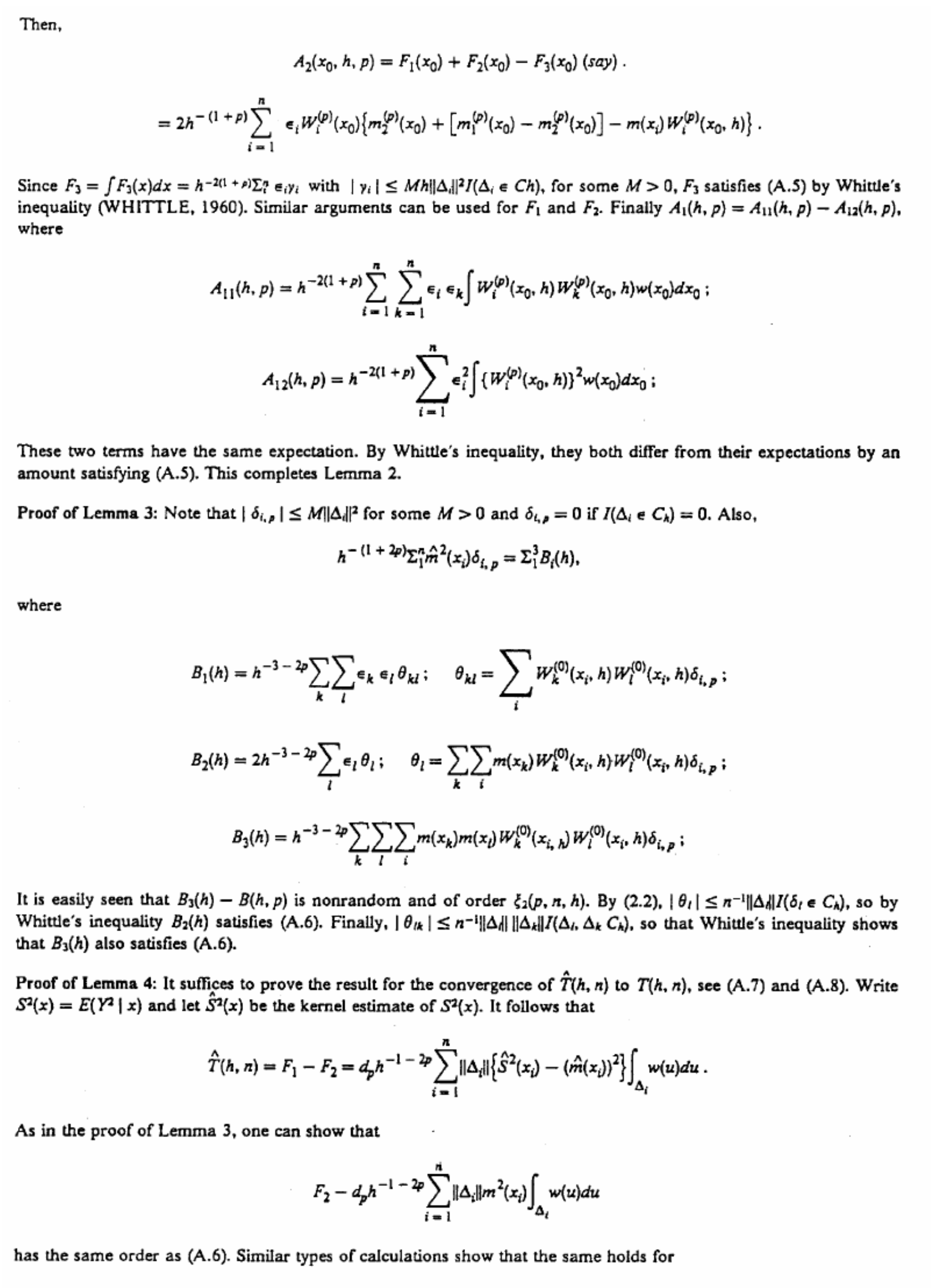
$$A_3(h) = \int\{m_1^{(p)}(x_0)\}^2 w(x_0)dx_0 - c_p$$

$$= \int\{m^{(p)}(x_0)\}^2 w(x_0)dx_0 - c_p + O(\xi_1(p, n, h))$$

$$= \int\{m^{(p)}(x_0)\}^2 w(x_0)dx_0 - B(h, p) + O(h(nh^{1+2p})^{-1} + \xi_1(p, n, h))\,.$$

Define $m^{(p)}(x_0)$ by (A.10) and

$$m_2^{(p)}(x_0) = h^{-(1+p)}\int K^{(p)}((x_0 - u)/h)m(u)du\,.$$



Then,

$$A_2(x_0, h, p) = F_1(x_0) + F_2(x_0) - F_3(x_0)\ (say)\ .$$

$$= 2h^{-(1+p)}\sum_{i=1}^{n} \epsilon_i W_i^{(p)}(x_0)\{m_2^{(p)}(x_0) + [m_1^{(p)}(x_0) - m_2^{(p)}(x_0)] - m(x_i)W_i^{(p)}(x_0, h)\}\ .$$

Since $F_3 = \int F_3(x)dx = h^{-2(1+p)}\Sigma_i^n \epsilon_i \gamma_i$ with $|\gamma_i| \le Mh\|\Delta_i\|^2 I(\Delta_i \in Ch)$, for some $M > 0$, F_3 satisfies (A.5) by Whittle's inequality (WHITTLE, 1960). Similar arguments can be used for F_1 and F_2. Finally $A_1(h, p) = A_{11}(h, p) - A_{12}(h, p)$, where

$$A_{11}(h, p) = h^{-2(1+p)}\sum_{i=1}^{n}\sum_{k=1}^{n}\epsilon_i\,\epsilon_k\int W_i^{(p)}(x_0, h)W_k^{(p)}(x_0, h)w(x_0)dx_0\ ;$$

$$A_{12}(h, p) = h^{-2(1+p)}\sum_{i=1}^{n}\epsilon_i^2\int\{W_i^{(p)}(x_0, h)\}^2 w(x_0)dx_0\ ;$$

These two terms have the same expectation. By Whittle's inequality, they both differ from their expectations by an amount satisfying (A.5). This completes Lemma 2.

Proof of Lemma 3: Note that $|\delta_{i,p}| \le M\|\Delta_i\|^2$ for some $M > 0$ and $\delta_{i,p} = 0$ if $I(\Delta_i \in C_h) = 0$. Also,

$$h^{-(1+2p)}\Sigma_1^n\hat{m}^2(x_i)\delta_{i,p} = \Sigma_1^3 B_i(h),$$

where

$$B_1(h) = h^{-3-2p}\sum_k\sum_l \epsilon_k\,\epsilon_l\,\theta_{kl}\ ; \quad \theta_{kl} = \sum_i W_k^{(0)}(x_i, h)W_l^{(0)}(x_i, h)\delta_{i,p}\ ;$$

$$B_2(h) = 2h^{-3-2p}\sum_l \epsilon_l\,\theta_l\ ; \quad \theta_l = \sum_k\sum_i m(x_k)W_k^{(0)}(x_i, h)W_l^{(0)}(x_i, h)\delta_{i,p}\ ;$$

$$B_3(h) = h^{-3-2p}\sum_k\sum_l\sum_i m(x_k)m(x_l)W_k^{(0)}(x_{i,}\,h)W_l^{(0)}(x_i, h)\delta_{i,p}\ ;$$

It is easily seen that $B_3(h) - B(h, p)$ is nonrandom and of order $\xi_2(p, n, h)$. By (2.2), $|\theta_l| \le n^{-1}\|\Delta_l\| I(\delta_l \in C_h)$, so by Whittle's inequality $B_2(h)$ satisfies (A.6). Finally, $|\theta_{lk}| \le n^{-1}\|\Delta_l\|\,\|\Delta_k\| I(\Delta_l, \Delta_k\ C_h)$, so that Whittle's inequality shows that $B_3(h)$ also satisfies (A.6).

Proof of Lemma 4: It suffices to prove the result for the convergence of $\hat{T}(h, n)$ to $T(h, n)$, see (A.7) and (A.8). Write $S^2(x) = E(Y^2 | x)$ and let $\hat{S}^2(x)$ be the kernel estimate of $S^2(x)$. It follows that

$$\hat{T}(h, n) = F_1 - F_2 = d_p h^{-1-2p}\sum_{i=1}^{n}\|\Delta_i\|\{\hat{S}^2(x_i) - (\hat{m}(x_i))^2\}\int_{\Delta_i} w(u)du\ .$$

As in the proof of Lemma 3, one can show that

$$F_2 - d_p h^{-1-2p}\sum_{i=1}^{n}\|\Delta_i\| m^2(x_i)\int_{\Delta_i} w(u)du$$

has the same order as (A.6). Similar types of calculations show that the same holds for

$$F_1 - d_p h^{-1-2p} \sum_{i=1}^{n} \|\Delta_i\| \beta(x_i) \int_{\Delta_i} w(u)du \,.$$

This completes the proof.

Appendix B: Proof of the Theorem

We will show that

$$\sup_{h,h' \in H_n} \left| \frac{MISE(h) - MISE(h') - (BSF(h) - BSF(h'))}{MISE(h) + MISE(h')} \right| \xrightarrow{P} 0, \text{ as } n \to \infty \qquad (B.1)$$

It follows from (B.1) that $\hat{h}$ is asymptotically optimal since with high probability for any $\eta > 0$,

$$\frac{MISE(\hat{h}) - MISE(h_0) - (BSF(\hat{h}) - BSF(h_0))}{MISE(\hat{h}) + MISE(h_0)} \le \eta \,,$$

where $h_0 = \operatorname{argmin}_{h \in H_n} MISE(h)$. Now from this

$$0 \ge BSF(\hat{h}) - BSF(h_0) \ge (1-\eta) MISE(\hat{h}) - (1+\eta) MISE(h_0)$$

which gives

$$1 \le \frac{MISE(\hat{h})}{MISE(h_0)} \le \frac{1+\eta}{1-\eta} \,.$$

Hence $\hat{h}$ is asymptotically optimal! To show (B.1) it suffices to show that

$$\sup_{h \in H_n} \left| \frac{MISE(h) - BSF(h)}{MISE(h)} \right| \xrightarrow{P} 0 \,. \qquad (B.2)$$

Consider now the different terms of the difference $MISE(h) - BSF(h)$. We have to show that they tend to zero "uniformly over h faster than $MISE(h)$, $h \in H_n$ itself tends to zero" in the sense that is made precise in formula (B.2). Let $T_n(h)$ denote one of those terms. The general idea is to apply Bonferroni's inequality and then use the inequalities given below. Let $\eta > 0$ be given. Then

$$P\{\sup_{h \in H_n} | T_n(h)/MISE(h) | > \eta\} \# H_n \sup_{h \in H_n} P\{ | T_n(h)/MISE(h) | > \eta\} \,.$$

The established Lemmas 1 − 4 will ensure that this last term will tend to zero and thus (B.2) is shown. First of all observe that we can work with the approximation to $MISE(h)$ as given in Lemma 1. Indeed

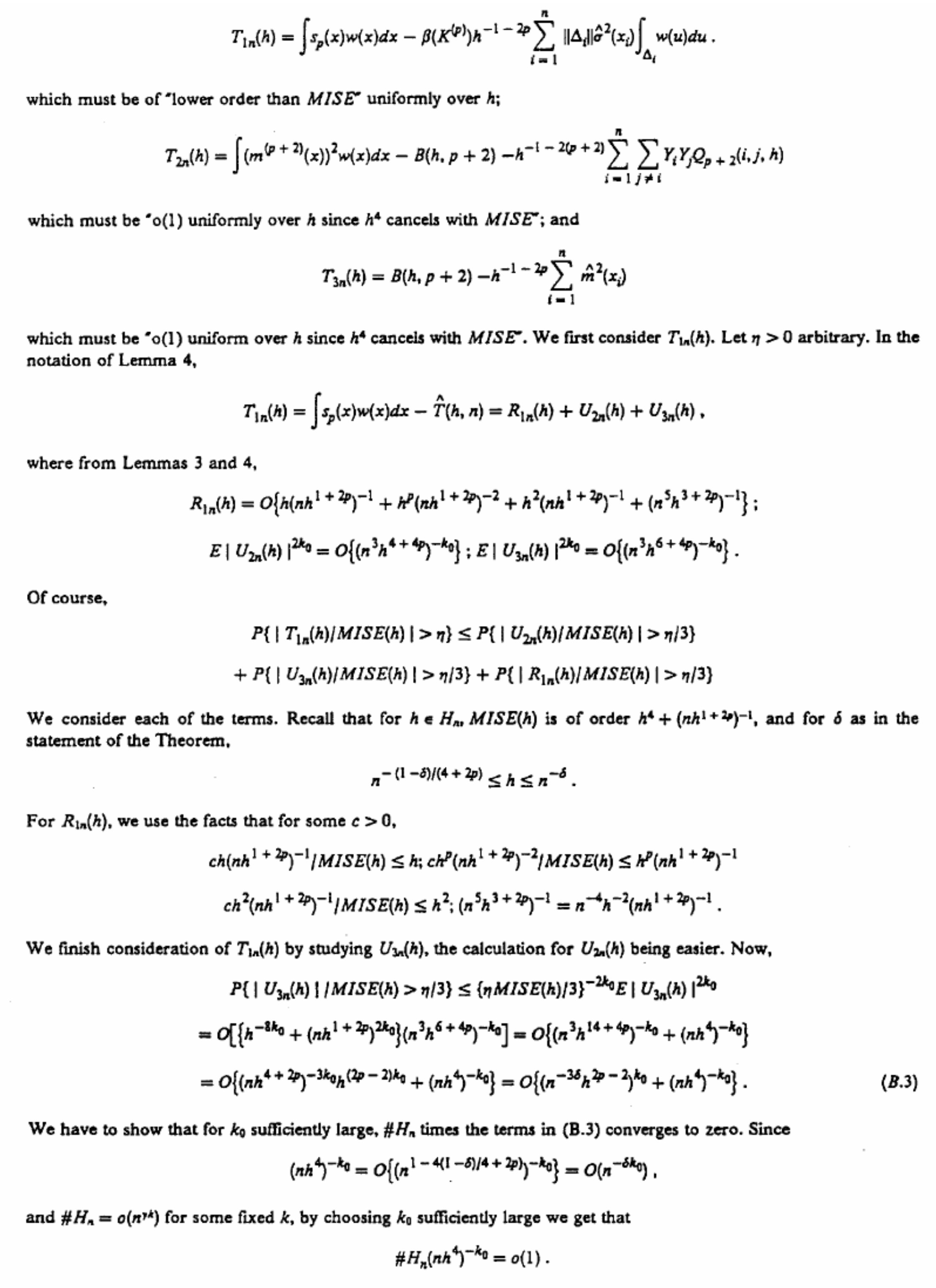
$$\sup_{h \in H_n} \left| \frac{MISE(h) - \int s_p(x) w(x) dx - ((1/2)\mu_2(K))^2 h^4 \int (m^{(p+2)}(x))^2 w(x) dx}{MISE(h)} \right|$$

$$\le \sup_{h \in H_n} | C_1 h^2 + C_2 (nh^2)^{-1} + C_3 n^{-1/2} | \xrightarrow{P} 0 \,.$$

To see this, first note that $\int s_p(x) w(x) dx = O_p(nh^{1+2p})^{-1}$, and then use

$$\frac{h^6}{h^4 + (nh^{1+2p})^{-1}} \le h^2 \,; \quad \frac{(nh^2)^{-1}(nh^{1+2p})^{-1}}{h^4 + (nh^{1+2p})^{-1}} \le (nh^2)^{-1} \quad \frac{n^{-1/2} h^2 (nh^{1+2p})^{-1/2}}{h^4 + [(nh^{1+2p})^{-1/2}]^2} \le n^{-1/2}$$

It remains to consider the terms



$$T_{1n}(h) = \int s_p(x)w(x)dx - \beta(K^{(p)})h^{-1-2p}\sum_{i=1}^{n} \|\Delta_i\|\hat{\sigma}^2(x_i)\int_{\Delta_i} w(u)du\,.$$

which must be of "lower order than $MISE$" uniformly over h;

$$T_{2n}(h) = \int (m^{(p+2)}(x))^2 w(x)dx - B(h, p+2) - h^{-1-2(p+2)}\sum_{i=1}^{n}\sum_{j\neq i} Y_i Y_j Q_{p+2}(i,j,h)$$

which must be "o(1) uniformly over h since h^4 cancels with $MISE$"; and

$$T_{3n}(h) = B(h, p+2) - h^{-1-2p}\sum_{i=1}^{n} \hat{m}^2(x_i)$$

which must be "o(1) uniform over h since h^4 cancels with $MISE$". We first consider $T_{1n}(h)$. Let $\eta > 0$ arbitrary. In the notation of Lemma 4,

$$T_{1n}(h) = \int s_p(x)w(x)dx - \hat{T}(h, n) = R_{1n}(h) + U_{2n}(h) + U_{3n}(h)\,,$$

where from Lemmas 3 and 4,

$$R_{1n}(h) = O\left\{h(nh^{1+2p})^{-1} + h^p(nh^{1+2p})^{-2} + h^2(nh^{1+2p})^{-1} + (n^5h^{3+2p})^{-1}\right\};$$

$$E\mid U_{2n}(h)\mid^{2k_0} = O\left\{(n^3h^{4+4p})^{-k_0}\right\};\; E\mid U_{3n}(h)\mid^{2k_0} = O\left\{(n^3h^{6+4p})^{-k_0}\right\}.$$

Of course,

$$P\{\mid T_{1n}(h)/MISE(h)\mid > \eta\} \leq P\{\mid U_{2n}(h)/MISE(h)\mid > \eta/3\}$$

$$+ P\{\mid U_{3n}(h)/MISE(h)\mid > \eta/3\} + P\{\mid R_{1n}(h)/MISE(h)\mid > \eta/3\}$$

We consider each of the terms. Recall that for $h \in H_n$, $MISE(h)$ is of order $h^4 + (nh^{1+2p})^{-1}$, and for δ as in the statement of the Theorem,

$$n^{-(1-\delta)/(4+2p)} \leq h \leq n^{-\delta}\,.$$

For $R_{1n}(h)$, we use the facts that for some $c > 0$,

$$ch(nh^{1+2p})^{-1}/MISE(h) \leq h;\; ch^p(nh^{1+2p})^{-2}/MISE(h) \leq h^p(nh^{1+2p})^{-1}$$

$$ch^2(nh^{1+2p})^{-1}/MISE(h) \leq h^2;\; (n^5h^{3+2p})^{-1} = n^{-4}h^{-2}(nh^{1+2p})^{-1}\,.$$

We finish consideration of $T_{1n}(h)$ by studying $U_{3n}(h)$, the calculation for $U_{2n}(h)$ being easier. Now,

$$P\{\mid U_{3n}(h)\mid /MISE(h) > \eta/3\} \leq \{\eta MISE(h)/3\}^{-2k_0} E\mid U_{3n}(h)\mid^{2k_0}$$

$$= O\left[\left\{h^{-8k_0} + (nh^{1+2p})^{2k_0}\right\}(n^3h^{6+4p})^{-k_0}\right] = O\left\{(n^3h^{14+4p})^{-k_0} + (nh^4)^{-k_0}\right\}$$

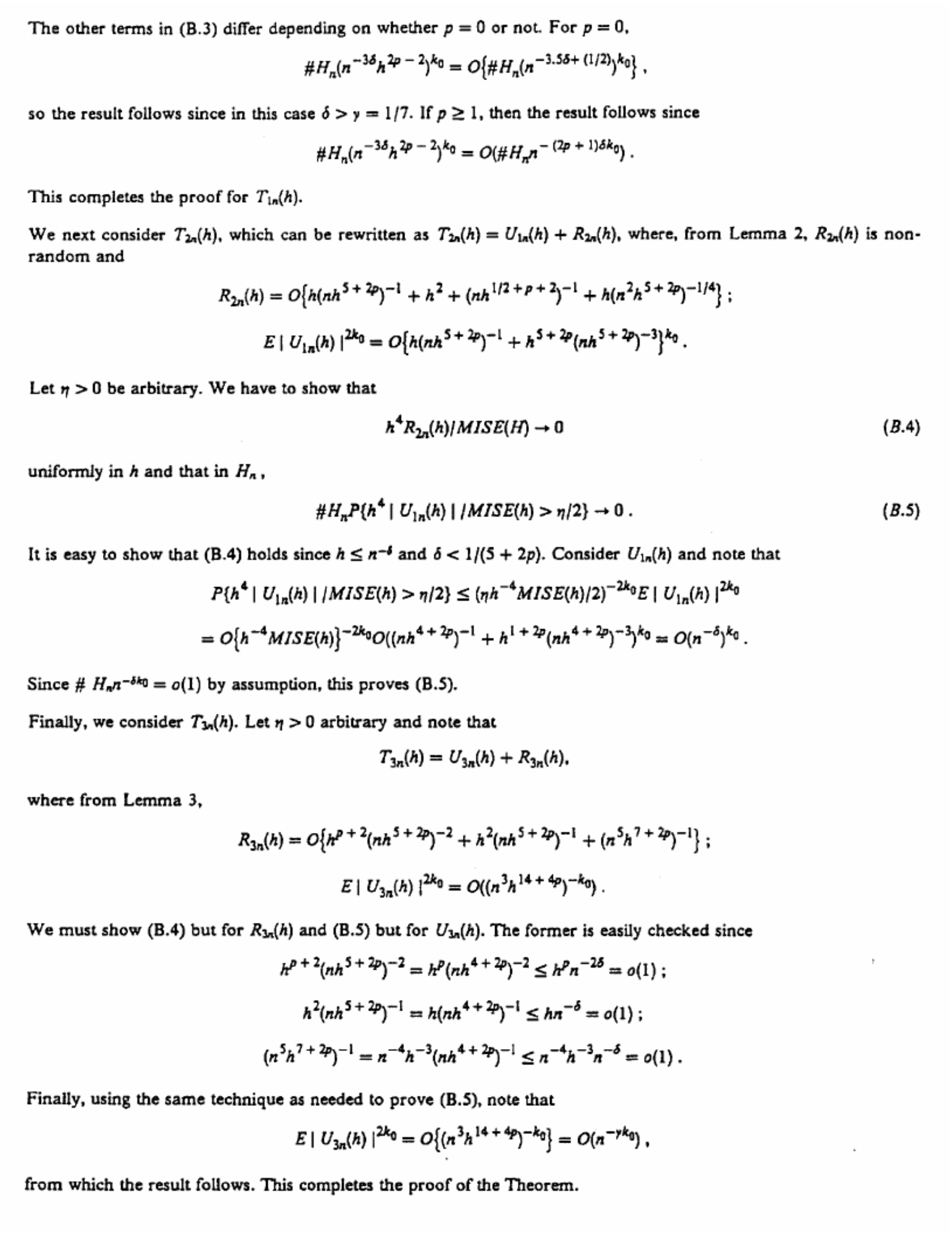
$$= O\left\{(nh^{4+2p})^{-3k_0}h^{(2p-2)k_0} + (nh^4)^{-k_0}\right\} = O\left\{(n^{-3\delta}h^{2p-2})^{k_0} + (nh^4)^{-k_0}\right\}. \qquad (B.3)$$

We have to show that for k_0 sufficiently large, $\#H_n$ times the terms in (B.3) converges to zero. Since

$$(nh^4)^{-k_0} = O\left\{(n^{1-4(1-\delta)/4+2p})^{-k_0}\right\} = O(n^{-\delta k_0})\,,$$

and $\#H_n = o(n^{\gamma k})$ for some fixed k, by choosing k_0 sufficiently large we get that

$$\#H_n(nh^4)^{-k_0} = o(1)\,.$$



The other terms in (B.3) differ depending on whether $p = 0$ or not. For $p = 0$,

$$\#H_n(n^{-3\delta}h^{2p-2})^{k_0} = O\{\#H_n(n^{-3.5\delta+(1/2)})^{k_0}\},$$

so the result follows since in this case $\delta > \gamma = 1/7$. If $p \geq 1$, then the result follows since

$$\#H_n(n^{-3\delta}h^{2p-2})^{k_0} = O(\#H_n n^{-(2p+1)\delta k_0}).$$

This completes the proof for $T_{1n}(h)$.

We next consider $T_{2n}(h)$, which can be rewritten as $T_{2n}(h) = U_{1n}(h) + R_{2n}(h)$, where, from Lemma 2, $R_{2n}(h)$ is nonrandom and

$$R_{2n}(h) = O\{h(nh^{5+2p})^{-1} + h^2 + (nh^{1/2+p+2})^{-1} + h(n^2h^{5+2p})^{-1/4}\};$$

$$E\,|\,U_{1n}(h)\,|^{2k_0} = O\{h(nh^{5+2p})^{-1} + h^{5+2p}(nh^{5+2p})^{-3}\}^{k_0}.$$

Let $\eta > 0$ be arbitrary. We have to show that

$$h^4R_{2n}(h)/MISE(H) \to 0 \qquad (B.4)$$

uniformly in h and that in H_n,

$$\#H_nP\{h^4\,|\,U_{1n}(h)\,|\,/MISE(h) > \eta/2\} \to 0\,. \qquad (B.5)$$

It is easy to show that (B.4) holds since $h \leq n^{-\delta}$ and $\delta < 1/(5+2p)$. Consider $U_{1n}(h)$ and note that

$$P\{h^4\,|\,U_{1n}(h)\,|\,/MISE(h) > \eta/2\} \leq (\eta h^{-4}MISE(h)/2)^{-2k_0}E\,|\,U_{1n}(h)\,|^{2k_0}$$

$$= O\{h^{-4}MISE(h)\}^{-2k_0}O((nh^{4+2p})^{-1} + h^{1+2p}(nh^{4+2p})^{-3})^{k_0} = O(n^{-\delta})^{k_0}.$$

Since $\#\,H_nn^{-\delta k_0} = o(1)$ by assumption, this proves (B.5).

Finally, we consider $T_{3n}(h)$. Let $\eta > 0$ arbitrary and note that

$$T_{3n}(h) = U_{3n}(h) + R_{3n}(h),$$

where from Lemma 3,

$$R_{3n}(h) = O\{h^{p+2}(nh^{5+2p})^{-2} + h^2(nh^{5+2p})^{-1} + (n^5h^{7+2p})^{-1}\};$$

$$E\,|\,U_{3n}(h)\,|^{2k_0} = O((n^3h^{14+4p})^{-k_0}).$$

We must show (B.4) but for $R_{3n}(h)$ and (B.5) but for $U_{3n}(h)$. The former is easily checked since

$$h^{p+2}(nh^{5+2p})^{-2} = h^p(nh^{4+2p})^{-2} \leq h^pn^{-2\delta} = o(1);$$

$$h^2(nh^{5+2p})^{-1} = h(nh^{4+2p})^{-1} \leq hn^{-\delta} = o(1);$$

$$(n^5h^{7+2p})^{-1} = n^{-4}h^{-3}(nh^{4+2p})^{-1} \leq n^{-4}h^{-3}n^{-\delta} = o(1).$$

Finally, using the same technique as needed to prove (B.5), note that

$$E\,|\,U_{3n}(h)\,|^{2k_0} = O\{(n^3h^{14+4p})^{-k_0}\} = O(n^{-\gamma k_0}),$$

from which the result follows. This completes the proof of the Theorem.

STATISTICAL METHODS FOR DEVELOPING AND DISTINGUISHING MULTINOMINAL RESPONSE MODELS IN THE TRAUMATOLOGICAL ANALYSIS OF SIMULATED AUTOMOBILE IMPACTS

W. Härdle
Rheinische Friedrich-Wilhelms-Universität Bonn
D-5300 Bonn, Federal Republic of Germany

D. Kallieris
Ruprecht-Karls-Universität Heidelberg
D-6900 Heidelberg, Federal Republik of Germany

R. Mattern
Johannes Gutenberg-Universität Mainz
D-6500 Mainz, Federal Republic of Germany

ABSTRACT

Simulated car-to-car side impacts, designed for the analysis of traumatological aspects, involve two sets of variables. Predictors include exogenous biomechanical factors as well as anthropometric variables, such as age. The response is measured a scale of injuy scores and is thus multinominal.

It is the aim of a statistical analysis of such data to devise a multinominal response model that explains possible patterns of injury as a function of a suitable set of predictor variables. Several approaches for modelling such a multinominal response relationship have been proposed in the literature, among them the Logistic and the Weibull regression models. Two major questions in applying such models are as follows: What model is appropriate and how should different models be compared. Another concern is how the quality of a given model should be presented for varying sets of predictors.

In this paper we discuss the first question by constructing a goodness-of-fit test based on bootstrapping flexible, non-parametric alternatives to a given parametric candidate model. Secondly, we present several graphical techniques that allow relatively simple comparisons of different models.

1. Modelling the influence of anthropometric and mechanical parameters on trauma indices:

The aim of the statistical analysis of simulated car impacts is to develop models that allow one to understand how the severity of impacts depend on observable input variables. Typically such input variables can be divided into two types. The first set of variables is describing

1

the test subject's physical characteristics, such as height or age. A second set is concerned with the actual experimental setting, and contains such parameters as velocity of the impact and acceleration measured at various places. These input variables determine jointly the response variable. The observed response variable is a trauma index usually scaled according to some injury scale, e.g. AIS (1980). The AIS trauma index, for example, is a discrete variable in $\{0,1,2,3,4,5,6\}$, with the lightest (or non) injury indexed by "0" and the severest injury indexed by "6". The input variables are mostly of continuous nature, i.e. they can possibly take each value in a certain interval.

Phrased in terms of statistical methodology we are given a discrete regression problem, i.e. discrete response variables (trauma index) are regressed on various kinds of predictor variables (possibly continuous or also discrete).(See Bickel and Doksum (1977), Neter and Wasserman (1974, Chapter 9)). The aim of this statistical problem is to construct suitable models for explaining the probability of a certain level of trauma index as a function of the given covariables. In this paper we denote by (X_i, Y_i), $i = 1,\ldots, n$, the data points from such an experiment; X standing for the vector of predictor variables (input) and Y denoting the discrete response variable (output vector). Since the response variable is multinominal (i.e. takes values in a discrete ordered set) it is reasonable to define the regression function as the probability that Y is bigger than some value c. Hence,we are dealing with a set of regression functions

$$p_c(x) = P(Y \geq c | X=x),$$

where c runs through the discrete set of possible response values (trauma indices). In determing such functions p one would like to have some basis requirements fulfilled that are direct consequences of the experimental setup. These are

(1.1) Monotonicity, i.e. if the input variables are ordered in some natural way then increasing the strength of impact or increasing age, the probability of having a trauma index greater than or equal to c should also increase.

(1.2) Consistency, i.e. $p_{c_1} \geq p_{c_2}$ for $c_1 \leq c_2$

Consistency means that the curves p_c should be so that the probability of having trauma index greater than c increases if c decreases.

In the next section we discuss several multinominal response models. In section 3 we show how nonparametric smoothing techniques help in selecting a suitable response model. In section 4 we discuss some graphical methods for enhancing the summary statistics of a given fit when the set of predictor variables is varied. In section 5 the application of these methods to the Heidelberg side impact data is presented. Section 6 is devoted to conclusions.

2. Multinominal Response Models

There are two different approaches to model the dependence of the conditional probability $p_c(x) = P(Y \geq c|X=x)$ as a function of the covariables x. The first approach is to assume that this function p_c is a member of a specific class of parameterized functions. The second approach is called non-parametric since the form of p_c is not restricted by any requirement except those of (1.1) and (1.2) above. The parametric approach has the advantage of easier interpretation of coefficients and also of numerical computations, whereas the non-parametric approach has the advantage of not being bound to any functional form. Both should serve each other as an alternative and should not be seen as mutually exclusive models. Well-known parametric models include the Logistic and the Probit regression models. The basic structural assumption for both approaches is the same; both are models based on linear combinations (projections) of the predictor variable x, i.e. the function p_c is modelled as

$$p_c(x) = G_c(\beta^T x).$$

with a link function G_c and parameter ß. The parametric approach consists of fixing the function $G_c(.) = G_c(\alpha_c + .)$ to a certain shape whereas the non-parametric approach does not prescribe the form of G_c. In the following we just write G to describe the general form of G_c.

In a Logit analysis one assumes that G is of the form of a logistic distribution function, i.e.

$$G(z) = \exp(z)/(1+\exp(z)).$$

The functions P_c are determined by the maximum likelihood method, i.e. one maximizes for each c

$$\Pi_{i=1}^{n} P(Y_i \geq c \mid X_i = x_i)$$

$$=\Pi_{i=1}^{n} G(\alpha_c + \beta^T x_i)^{Y_i^c}(1 - G(\alpha_c + \beta^T x_i)^{(1 - Y_i^c)}.$$

$$Y_i^c = I(Y_i \geq c).$$

subject to the consistency condition. In the same way other models like the Probit model with G equal to the standard normal distribution function can be adapted. Yet another shape function is the Weibull distribution function.

The non-parametric approach does not fix the shape function G, but rather lets it be any smooth function following the requirements (1.1) and (1.2). Given the parameter vector ß the link function G is determined by a non-parametric smoothing technique, such as spline or kernel, see Härdle (1988). The kernel smoother $\hat{G}_h(z)$ at the point

$$z = \beta^T x \text{ for data } (Z_i=\beta^T X_i,\ Y_i)$$

is defined by

$$\hat{G}_h(z) = n^{-1}\Sigma_{i=1}^n K_h(z-Z_i)Y_i/n^{-1}\Sigma_{i=1}^n K_h(z-Z_i)$$

where $K_h(u) = h^{-1}K(u/h)$ is a delta function sequence with bandwidth h and kernel K, where K is a continuous probability density. The kernel smoother is a consistent estimate of G if h → 0 as the sample size n tends to infinity. The parameter ß can be determined in various numerical ways, since the function G is not determined up to scale. One of the possibilities is to determine G and ß jointly by minimizing the Residual Sum of Squares (RSS) or other measures of accuracy. This amounts to finding G and ß such that

$$n^{-1}\Sigma_{i=1}^n (Y_i - G(\beta^T X_i))^2$$

is minimal. This minimization is done iteratively by searching over all possible directions ß, that is why this method is called Projection Pursuit Regression (PPR), see Friedman and Stuetzle (1981). Another method is called Average Derivative Estimation (ADE). In ADE estimates of ß are obtained in a direct way without involving the link function G. This estimate of ß is defined as

$$\hat{\beta} = n^{-1}\Sigma_{i=1}^n Y_i \hat{f}'(X_i)/\hat{f}(X_i)$$

where $\hat{f}$ denotes an estimate of the partial derivatives of f, the density of X. For details see Härdle and Stoker (1988).

3. Selecting a suitable model

The task finding a suitable model among the many possible parametric and non-parametric alternatives involves the statistical precision of the model as well as the numerical applicability. It is widely known that the Logistic regression model can be quite easily fitted numerically, SAS Supplementary User's Guide (1985). Other link functions G, for example the Probit curve have a similar shape (see Berkson, 1951) but require more computational effort. Also the non-parametric smoothing method requires a lot more on computations but has the advantage of not being restricted in its functional form. In particular the symmetry of the link function that is inherent to the Logit model is no restriction for the non-parametric approach. Indeed the response of the side impact experiments is somewhat asymmetric, as was pointed out by several people who tried a skewed Weibull distribution as a link function G. The price one has to pay though for this additional feature is that the number of parameters, and thus the numerical cost and precision of the algorithm, increase.

Since the non-parametric alternative allows fitting in a much wider class of functions it seems reasonable that it can be used in a formal test of goodness of fit of low dimensional parametric models. To simplify matters let us consider only a binominal response model of one dimensional X variables, i.e. Y takes the values 0 or 1. the proposed test is based on smoothing the response variables of a given parametric fit $p(x;\hat{ß})$. One defines the kernel smoother on data (X_i, Y_i) as

$$\hat{p}(X_j) = n^{-1}\Sigma_{i=1}^{n} K_h(X_j - X_i)Y_i / n^{-1}\Sigma_{i=1}^{n} K_h(X_j - X_i).$$

The smoothing parameter h can be determined by crossvalidation, see Härdle (1988). The test is described formally as follows.

1. Fit a candidate parametric model $(p\ (x;\hat{ß})$

2. Simulate new observations $(X^*_i\ Y^*_i)$ from this model by using a pseudo random number generator based on $p(x;\hat{ß})$ (bootstrapping).

3. Determine for each X_i that has been observed the empirical 5 % quantiles of a kernel smoother of the simulated data.

4. Center these 5 % bands around the assumed parametric candidate model.

5. Check whether the kernel smoother based on the original data lies in between these bands.

Figure 3.1

Another method is based on comparing the likelihood for different models with a bias correction for different number of parameters. This is related to ideas of Akaike (1977) and works as follows. One compares the Log-Likelihoods under both models, i.e.

$$n\ L_1(\dot{\beta}_1) - n\ L_2(\beta_2) - (\text{dim}(\text{model}\ {}_1) - \text{dim}(\text{model}\ {}_2)).$$

Based on the limiting chi-square distribution of twice the likelihood ratio statistic one cannont distinguish the two models if the magnitude of the above difference is less than 0.5.

4. Comparing similar models

If the above models are run for several types and sets of input variables it is important to compare the output of the different fits. In the study of the Heidelberg data we found the following, mostly graphically oriented tools very convenient.

Concomitant pairs

Concomitant pairs are defined through all pairs of observations with different response values. Now count all pairs of observations where the current model fit predicted a higher probability for the higher Y-value. Then compute the share of these pairs among all pairs with different Y-values. Certainly if this share of concomitant pairs is close to 1 the model fits quite well. The procedure LOGIST of the SAS system computes this number on request.

Prediction Table

The prediction table is simply a frequency table of the observed trauma indices versus the predicted trauma indices. The number of correctly predicted response variables is the classification rate. This number lies between O and 1. Certainly a number close to one is desirable. It is quite intuitive that the empirically determined classification rates are over optimistic since the data is used to determine the model as well as to judge it. An unbiased estimate of the classification rate can be obtained by, for example, cross validation. In this method the whole analysis is performed n times on n subsamples each of size n-1 (leave one out method). The left out observation is predicted by the model constructed from the rest of the observations. This leads to an unbiased estimate of the prediction error, as was shown by Stone (1974).

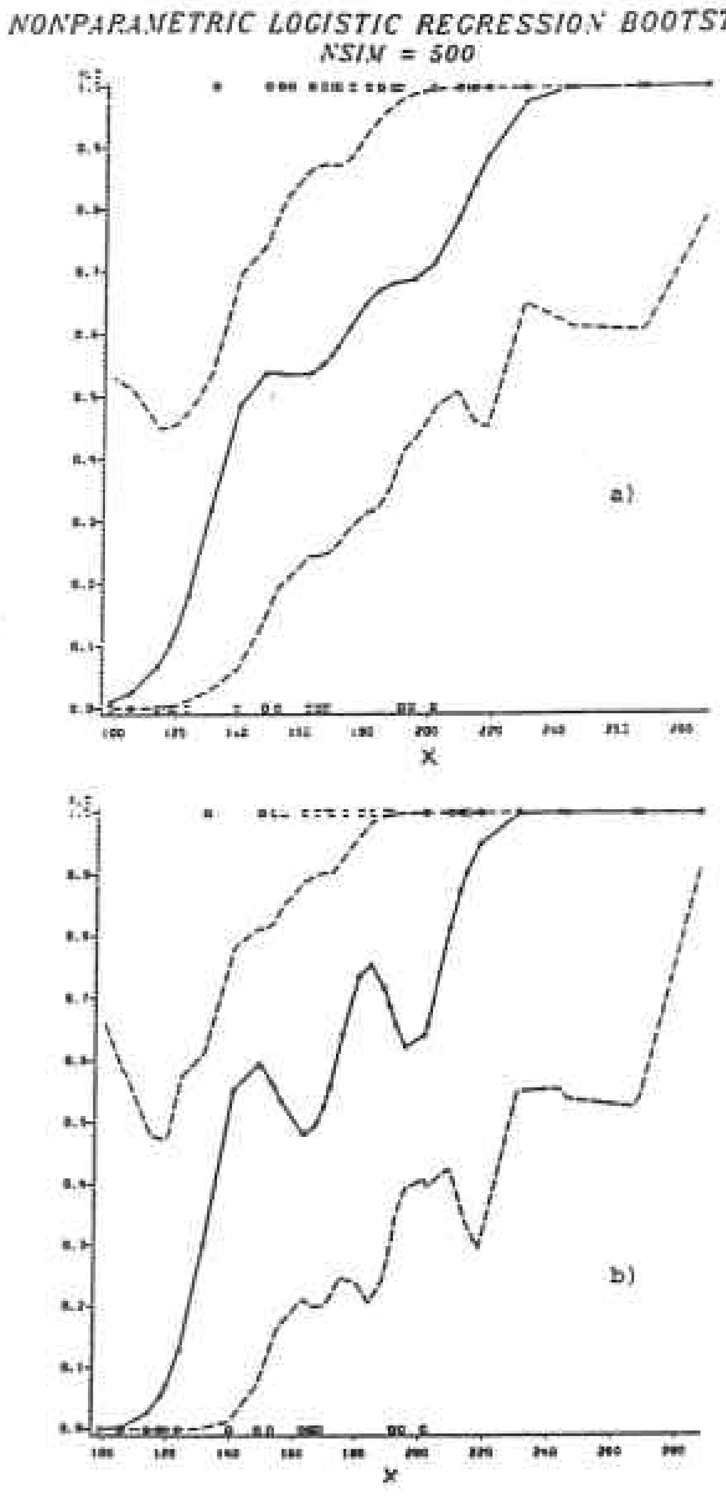


Fig. 3.1 Nonparametric logistic distribution of the injury severity (y = 1 for AIS ≥ 3 and Y = 0 for AIS < 3) over the TTI with 5 % confidencebands for 500 simulations according to the bootstrap method.
a) bandwidth h = 13
b) bandwidth h = 9

The enhanced histogram of prediction errors

This is a histogram of the observed differences between the observed trauma index and the predicted index where large indices are marked in a special way. The procedure is as follows. 1. Compute all the differences predicted response - observed response. 2. Index all large trauma values (for the AIS values (predicted or observed) greater than 4.
3. Draw a histogram of these differences where the big trauma indices get marked by using special symbol.

In figure 4.1 we show an enhanced histogram for the TTI (Eppinger et al., 1984) as a predictor variable for the TOAIS (thorax AIS).

Figure 4.1

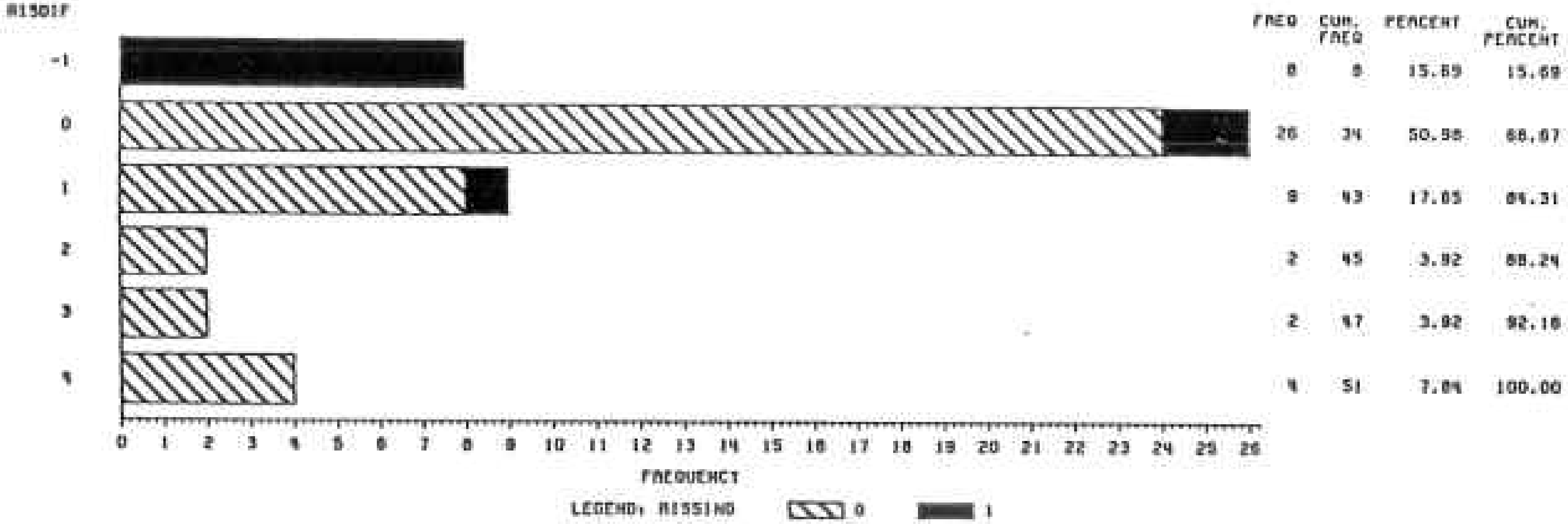
This Thoracic Trauma Index is defined through

$$TTI = 1.4 \text{ AGE} + 0.5 \text{ FORCE}.$$

One sees from this enhanced histogram of prediction erros that the TTI leans toward over estimating the true responses. Indeed, the histogram is skewed to the right. There are 11 observations involving the thorax AIS value of 4. Two of these eleven observations have prediction error zero.One observation has been predicted to have AIS value 4, but really had value 3 (prediction error 2 to the right in the histogram), and eight observations had AIS value 4 but were wrongly classified as 3. One should therefore search for a model that more faithfully predicts the high AIS values.

A distortion measure

As a measure of distortion of current fit we would like to propose two subintegrals of the above histogram. This pair of numbers tells first whether the fit is skew, i.e. has a bias towards over- or underestimating the true response value. Secondly the size of the subintegrals relative to the sample size immediately gives a goodness of fit criterion. The first subintegral just counts the number of positive exceedances (to the right of the column zero in figure 4.1). The second subintegral counts the number of negative exceedances, in this case -8. This together gives the distortion mesure (-8, 35) which describes in a very condensed form the skewness of the prediction and how much the true values are missed by the above model.



AISDIF	FREQ	CUM. FREQ	PERCENT	CUM. PERCENT
-1	8	8	15.69	15.69
0	26	34	50.98	66.67
1	9	43	17.65	84.31
2	2	45	3.92	88.24
3	2	47	3.92	92.16
4	4	51	7.84	100.00

Fig. 4.1 Logistic prediction of the body-AIS (TAAIS) from TTI. Enhanced AIS-difference-histogram between the observed predicted TAAIS.
Marked black: body-AIS-predictions, in which one TAAIS 5 is involved.

The Isoquants

The plot of isoquants is designed for two dimensional predictor variables and shows in a graphical way what trauma indices are to be expected given all possible combinations of covariables. In figure 4.2 we show the predicted thorax AIS classes as a function of AGE and FORCE, as defined in Kallieris, Mattern and Härdle (1986).

Figure 4.2

The region indicated by the letter A would be the region of (AGE, FORCE) combinations where AIS = 0 would be predicted. The region with AIS = 3 is shown by D and the highest AIS value of 4 is marked by an E. Overlaid in this plane are the original data values (0,1,2,3,4). This plot allows simple comparison of different fits by simply studying the regions that determine the AIS values. Given for instance the age of 30 one can easily determine by raising the values of FORCE at what points of FORCE the prediction to higher AIS classes would happen. (FORCE level 140 jumg to predicted AIS 3, FORCE level 250 jump to predicted AIS 4).

5. Application to the Heidelberg data

Only a few research onsets are suited to determine the connection beween mechanical influence and injury severity when measured in AIS degrees. There are real accident analyses on one hand and crash tests with post mortem human subjects (PMHS) on the other hand. Both research onsets are not ideal. The advantage of crash tests with PMHS is, e.g., that by defined conditions of the accident severity, loads acting on the body can be measured in physical magnitudes like acceleration at ribs, sternum, vertebral bodies and head. This is not possible in the real accident analyses. Differences of the injury limits against the living human beings are criticized as a disadvantage of the crash tests with PMHS. The load values measured on the bodies of the PMHS however, are indispensable basis data for the construction of dummies, if these dummies should be qualified for the injury prediction in crash tests.

At the Institute for Legal Medicine of the University of Heidelberg crash tests were conducted with PMHS and dummies for many years to investigate this research concept. As follows, the investigation of lateral collisions should represent which connections exist between loading parameters at the body of the PMHS, anthropometric data and injury severity and how these connections can be used for injury prediction by utilization of the statistical methods described above. Basis of the connection analyses are 58 90-degree lateral collisions. In these collisions PMHS have been loaded in near side position in the impacted/standing vehicle.

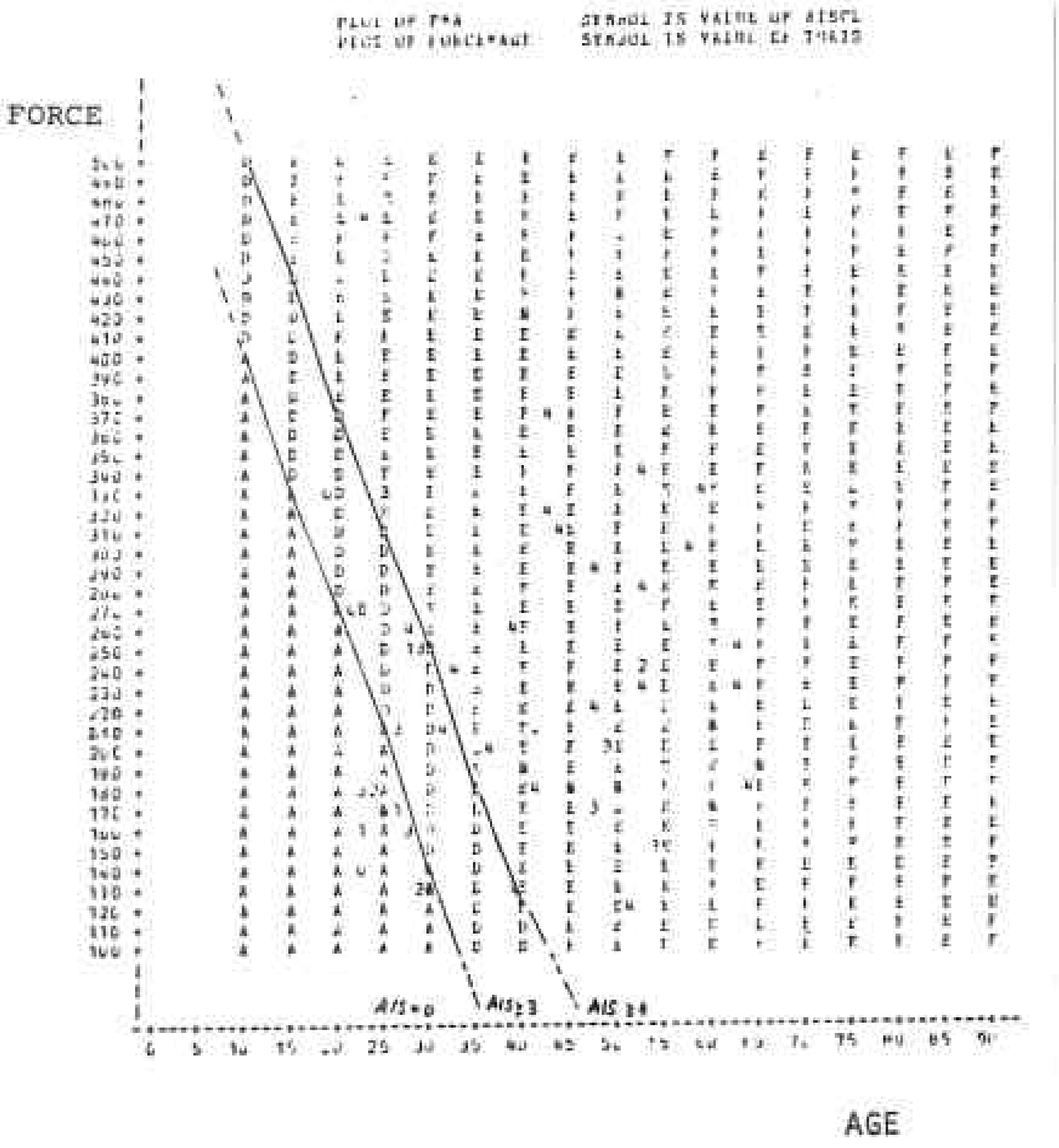


Fig. 4.2 Isoquantplot for the illustration of the prediction results of the logistic regression from AGE and FORCE.
Zone A: prediction of TOAIS = 0
Zone B: prediction of TOAIS = 3
Zone E: prediction of TOAIS = 4
Numbers in the zones: observed thorax-injury degrees
FORCE = 1/2 (accel.max. 4th rib impacted side + max. result. accel. Th 12) x bodymass / 75

The crash tests have been conducted at impact velocities of 40, 45, 50 and 60 km/h (Kallieris et al., 1987). In the PMHS 22 acceleration values at head, thorax, spinal column and pelvis have been recorded for each test. The injuries of the PMHS have been scaled according to AIS 80. It was seen in the statistical analyses that the injury levels could be most effectively predicted by the method of logistic regression. In the 90 degree lateral collisions the body injury severity (TAAIS) was generally leading and determined the maximum injury severity (MAIS). Therefore, the prediction of the body injury severity for right side lateral collisions is presented here as an example. Among the 22 as maximum and 3 ms values recorded accelerations the following proved to be the best predictors:

1. Acceleration (3 ms value) in x-direction at lower sternum (BUX3) (g);
2. acceleration (3 ms value) at the 12th thoracic vertebra in y-direction (T12Y3) (g);

The further improvement of the injury prediction has been reached in considering the Body Mass (BMASS) (kg) as covariable. With these covariable combination, the logistic model estimated the following parameters for the injury index Z:

$$Z = 0.15\ \text{BMASS} + 0.08\ \text{T12Y3} + 0.06\ \text{BUX3}.$$

The probability curves for TAAIS rankings 0,4 and 5 are shown in figure 5.1, for impacts from the right. The three tests with TAAIS 2 and 3 in the test series were not considered.

Figure 5.1

Below a Z value of 18.3, the envelope of the AIS probability curves indicates a high probability to be uninjured (the highest probability is below Z = 18). Between Z = 18.3 and Z = 20, a TAAIS of 4 is largely to be expected and above Z = 20 the probability for TAAIS 5 of about 45 % increases continuously to 100 % (at Z = 25). The enhanced TAAIS difference histogram (see section 4) in figure 5.2 shows that the above mentioned covariable combination as correctly predicts 59 % of the cases. The model predicts the TAAIS in 19 % too high and in 15 % a level too low; each one time, the model underestimated the observed injury for two and 4 AIS degrees.

Figure 5.2

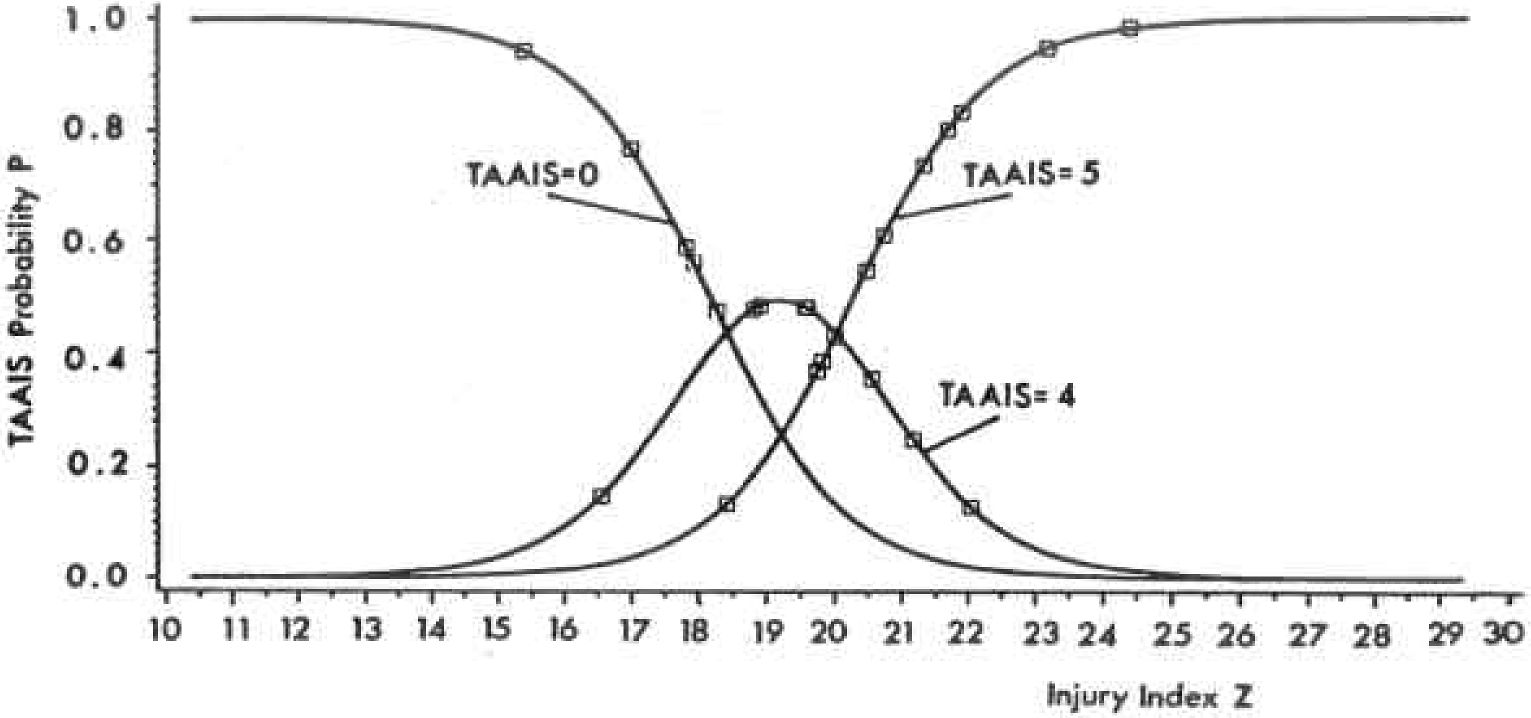
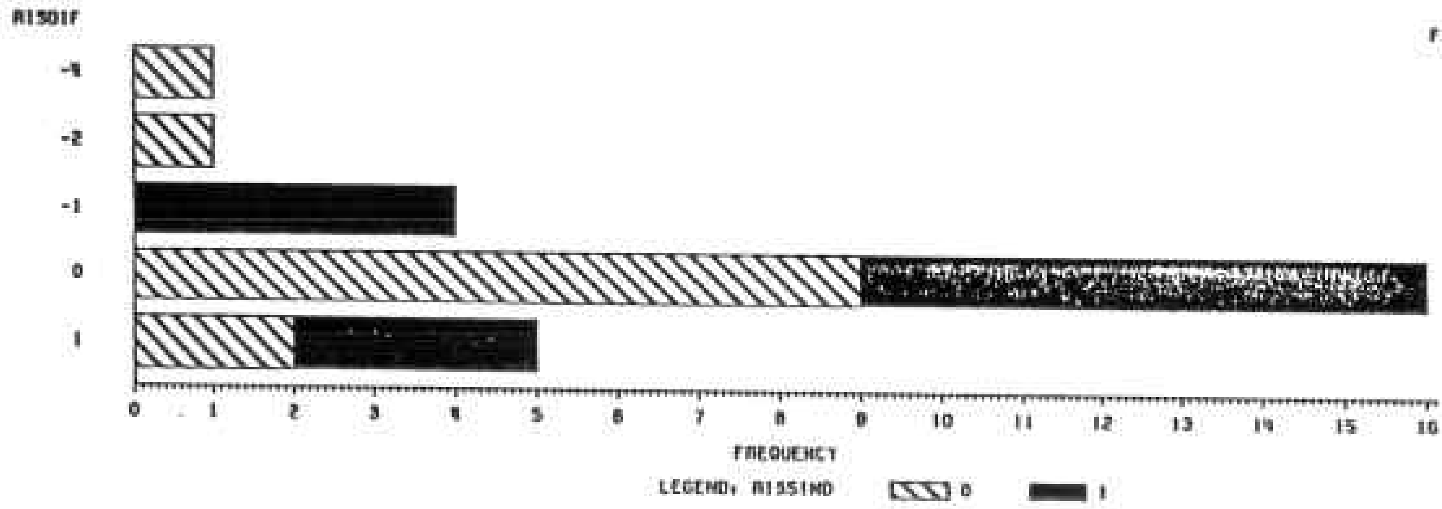


Fig. 5.1 Predicted probability of torso injury (TAAIS) for impacts from the right

□ Observed degree of injury



AISDIF	FREQ	CUM. FREQ	PERCENT	CUM. PERCENT
-4	1	1	3.70	3.70
-2	1	2	3.70	7.41
-1	4	6	14.81	22.22
0	16	22	59.26	81.48
1	5	27	18.52	100.00

Fig. 5.2 Enhanced AIS-difference-histogram for the logistic prediction of the body-AIS (TAAIS) grom Z = 0,15 AGE + 0,08 accel. Y-direct. Th 12 + 0,06 3 ms accel. X-direct. upper sternum

6. Conclusions

We have presented several multinominal response models of parametric and non-parametric nature. A way of comparing these models and deciding which one is more appropriate than others is given by considering non-parametric alternatives in the construction of a simulation band. This simulation band technique (section 3) lead for the Heidelberg data to the conclusion tht the Logistic response model is appropriate for the analysis of car-to-car side impacts. Comparing the Likelihoods of the Logistic and the Weibull link functions we found no better fit for the Weibull model, see Kallieris, Mattern and Härdle (1986). We furthermore presented a variety of graphical techniques which are of great assistance when looking for suitable predictor variables X, see section 4. Using these techniques we found for example that the Logistic model using the trauma index

$$Z = 0.15\ \mathrm{BMASS} + 0.08\ \mathrm{T12Y3} + 0.06\ \mathrm{BUX3}$$

had good prediction properties for the TAAIS, see section 5.

REFERENCES

AIS (1980) States JD, Huelke DF, Baker SP, Bryant RW et. al. The Abbreviated Injury Scale, 1980 Revision.

Akaike H (1977) On Entropy Maximization Principle. In Applications of Statistics, Ed.P.R. Krishaniah. Amsterdam, North Holland

Berkson J (1951) Why I prefer Logits to Probits. Biometrics 7: 327-339

Bickel P, Doksum K (1977) Mathematical Statistics. Holden-Day Inc., San Fransisco

Eppinger RH, Marcus JH, Morgan RM (1984) Development of Dummy and Injury Index for NHTSA's thoracic side impact protection research program, SAE technical paper series 840885, Government/Industry Meeting and Exposition Washington D.C.

Friedman J, Stuetzle W (1981) Projection Pursuit Regression. J.Amer.Statist.Assoc. 76: 817-823

Härdle W, Stoker T (1988) Investigating smooth multiple regression models by the method of Average Derivatives. J. Amer.Statist.Assoc., to appear

Härdle W (1988) Applied Nonparametric Regression. Book to appear

Kallieris D, Mattern R, Härdle W (1986) Belastbarkeitsgrenzen und Verletzungsmechanik des angegurteten Pkw-Insassen beim Seitaufprall. Phase II: Ansätze für Verletzungsprädiktionen. Schriftenreihe der Forschungsvereinigung Automobiltechnik e.V. (FAT) Nr. 60, Frankfurt/Main 17

Kallieris D, Schmidt Gg, Mattern R (1987) Vertebral Column Injuries in 90 degree Collisions - A study with Post Mortem Human Subjects. Proc. of Intern. IRCOBI Conf. on the Biomechanics of Impacts, Birmingham, 189-192

Neter J, Wasserman W (1974) Applied Linear Statistical Models. Richad D. Irwin, Inc. Homewood Illinois

SAS - Statistical Analysis System, Cary North Carolina

SAS - Supplementary User's Guide, Cary North Carolina

Stone M (1974) Crossvalidatory choice and assessment of statistical predictions (with discussion). Journal of the Royal Statistical Society, Series B, 36, 111-147

This paper is kindly dedicated to Prof. Dr. med. Georg Schmidt, Heidelberg, on the occasion of his 65th birthday.

The Annals of Statistics
1991, Vol. 19, No. 2, 778–796

BOOTSTRAP SIMULTANEOUS ERROR BARS FOR NONPARAMETRIC REGRESSION[1]

BY W. HÄRDLE AND J. S. MARRON

Université Catholique de Louvain and Universität Bonn and Universität Bonn

Simultaneous error bars are constructed for nonparametric kernel estimates of regression functions. The method is based on the bootstrap, where resampling is done from a suitably estimated residual distribution. The error bars are seen to give asymptotically correct coverage probabilities uniformly over any number of gridpoints. Applications to an economic problem are given and comparison to both pointwise and Bonferroni-type bars is presented through a simulation study.

1. Motivation. Regression smoothing is an effective method for estimation of mean curves in a flexible nonparametric way. Since this technique makes no structural assumptions on the underlying curve, it is very important to have a device for understanding when observed features are significant. A question often asked in this context is whether or not an observed peak or valley is actually a feature of the underlying regression function or is only an artifact of the observational noise. For such issues, confidence intervals should be used that are simultaneous (i.e., uniform over location) in nature. This paper proposes and analyzes a method of obtaining any number of simultaneous error bars at a grid of points. The method is simple to implement and does not rely on the evaluation of quantities which appear in asymptotic distributions. The construction is based on a residual resampling technique which models the conditional error distribution and also takes the bias properly into account (at least asymptotically).

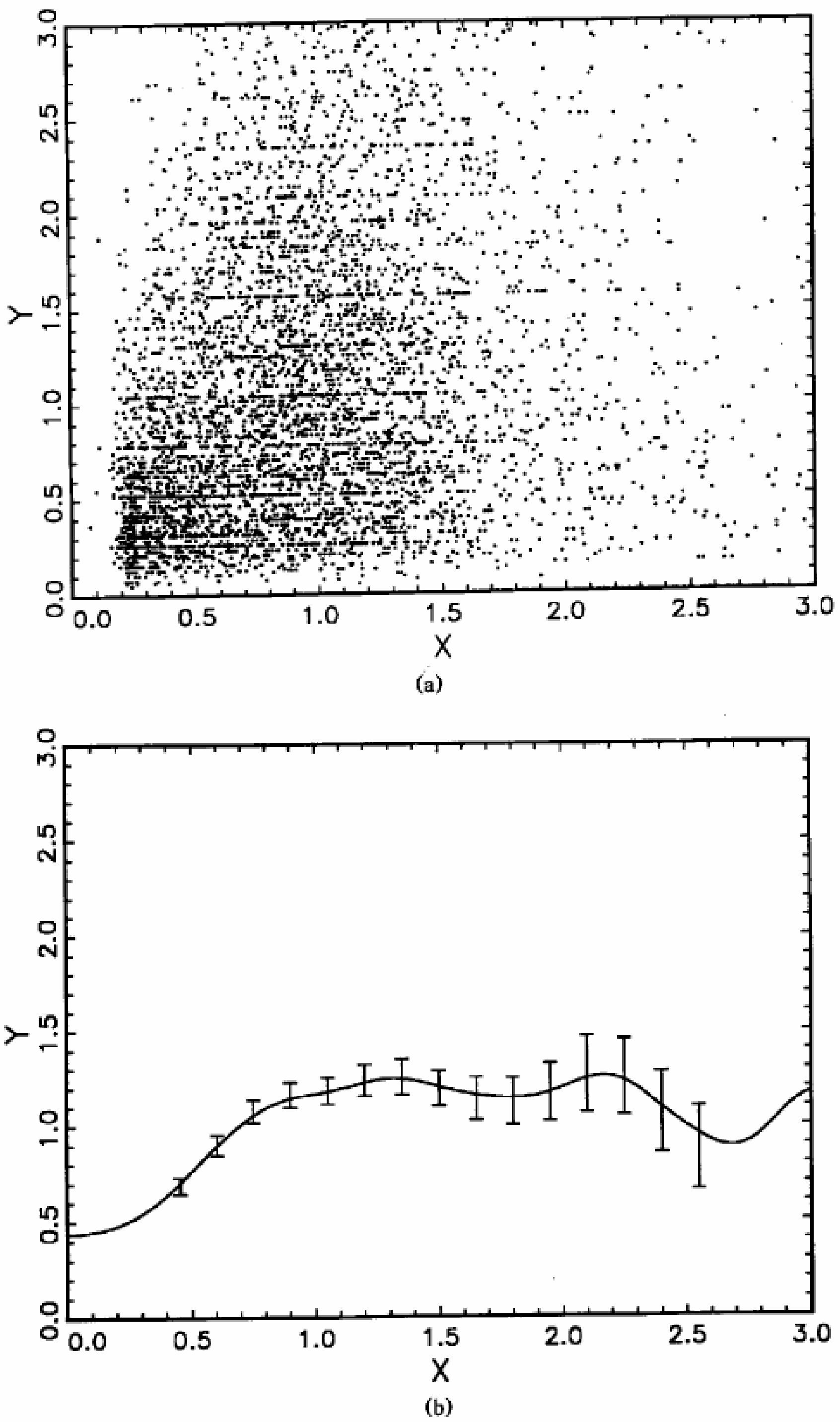
For an understanding of these ideas, consider Figure 1. Figure 1a shows a scatter plot of the expenditure for potatoes as a function of income for the year 1973, from the Family Expenditure Survey (1968–1983). Figure 1b shows a nonparametric regression estimate which was obtained by smoothing the point cloud, using the kernel algorithm described in Section 2. As a means of understanding the variability in the kernel smooth, Figure 1b also shows error bars, i.e., vertical confidence intervals constructed by the bootstrap method proposed in Section 2. These bars are estimated simultaneous 80% confidence intervals. Note that the error bars are longer on the right-hand side, which reflects the fact that there are fewer observations there and hence more uncertainty in the curve estimate. The error bars are asymmetric in particular

Received March 1989; revised March 1990.

[1] Research supported by Deutsche Forschungsgemeinschaft, SFB 303.

AMS 1980 *subject classifications*. Primary 62G05; secondary 62G99.

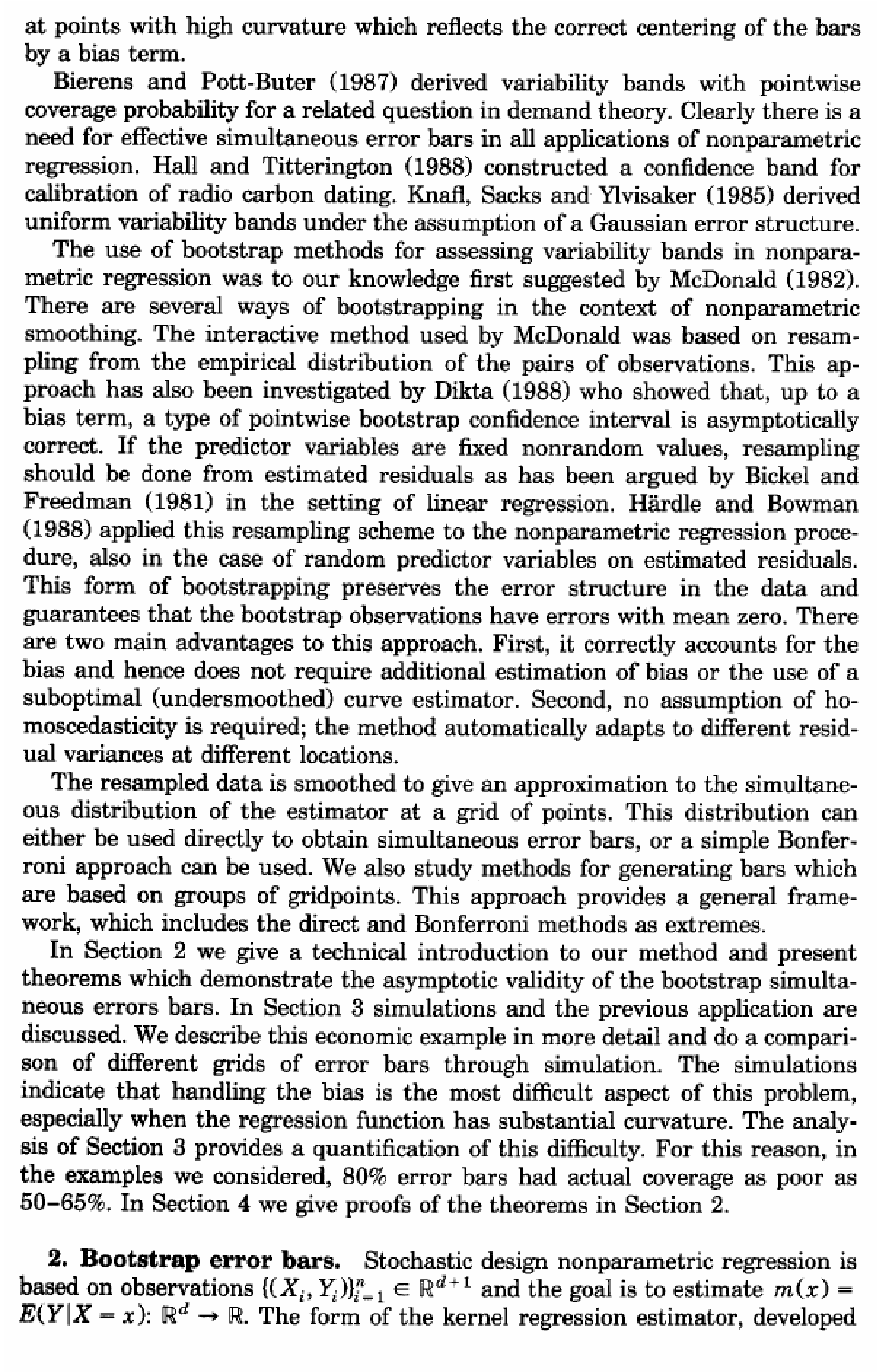
Key words and phrases. Bootstrap, error bars, kernel smoothing, nonparametric regression, variability bound.



(a)

(b)

FIG. 1. *Expenditure for potato vs. income* (a) *scatter plot* (b) *regression kernel smooth* (*quartic kernel with band with h* = 0.3) *and errors bars.*



at points with high curvature which reflects the correct centering of the bars by a bias term.

Bierens and Pott-Buter (1987) derived variability bands with pointwise coverage probability for a related question in demand theory. Clearly there is a need for effective simultaneous error bars in all applications of nonparametric regression. Hall and Titterington (1988) constructed a confidence band for calibration of radio carbon dating. Knafl, Sacks and Ylvisaker (1985) derived uniform variability bands under the assumption of a Gaussian error structure.

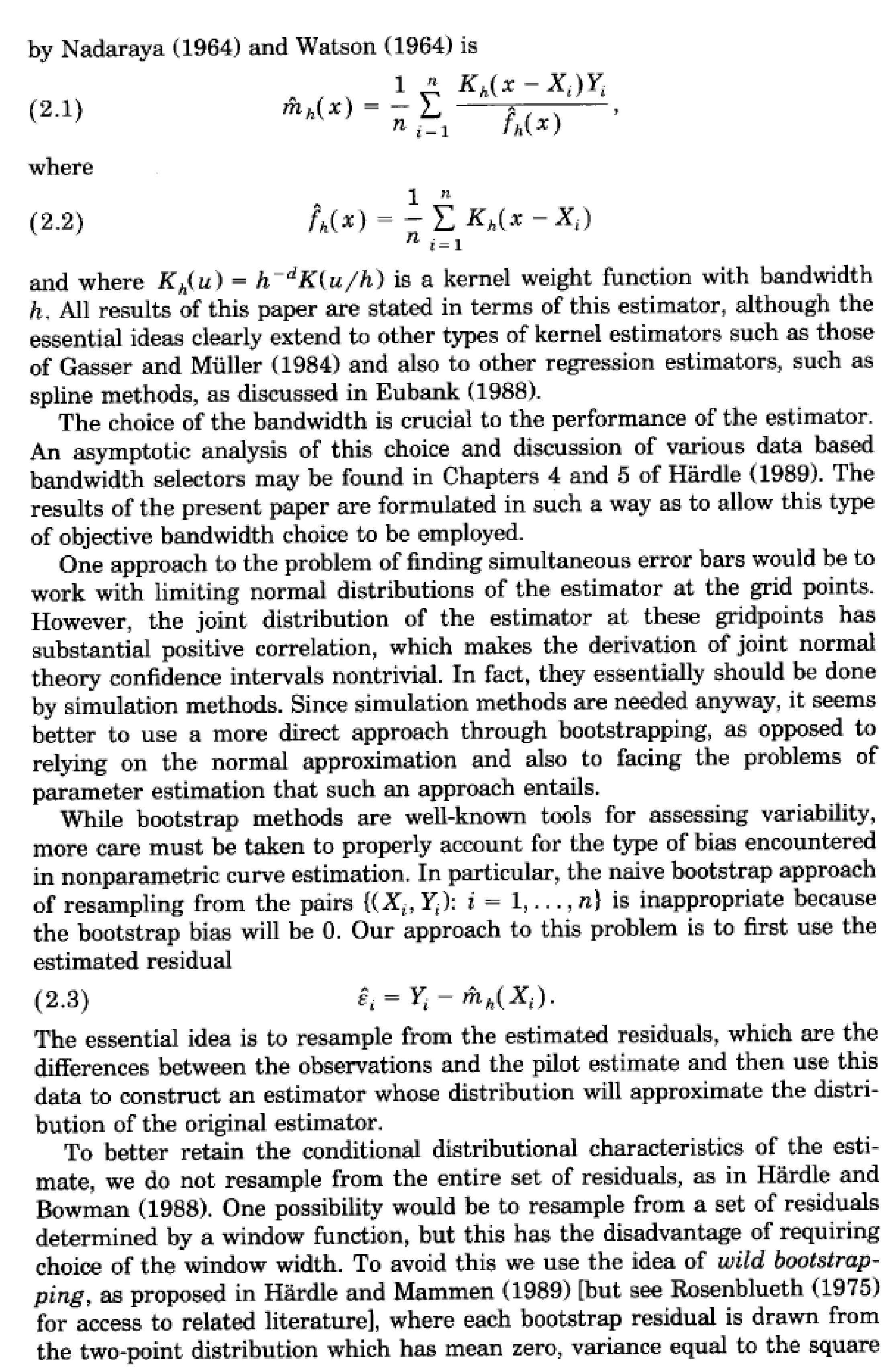
The use of bootstrap methods for assessing variability bands in nonparametric regression was to our knowledge first suggested by McDonald (1982). There are several ways of bootstrapping in the context of nonparametric smoothing. The interactive method used by McDonald was based on resampling from the empirical distribution of the pairs of observations. This approach has also been investigated by Dikta (1988) who showed that, up to a bias term, a type of pointwise bootstrap confidence interval is asymptotically correct. If the predictor variables are fixed nonrandom values, resampling should be done from estimated residuals as has been argued by Bickel and Freedman (1981) in the setting of linear regression. Härdle and Bowman (1988) applied this resampling scheme to the nonparametric regression procedure, also in the case of random predictor variables on estimated residuals. This form of bootstrapping preserves the error structure in the data and guarantees that the bootstrap observations have errors with mean zero. There are two main advantages to this approach. First, it correctly accounts for the bias and hence does not require additional estimation of bias or the use of a suboptimal (undersmoothed) curve estimator. Second, no assumption of homoscedasticity is required; the method automatically adapts to different residual variances at different locations.

The resampled data is smoothed to give an approximation to the simultaneous distribution of the estimator at a grid of points. This distribution can either be used directly to obtain simultaneous error bars, or a simple Bonferroni approach can be used. We also study methods for generating bars which are based on groups of gridpoints. This approach provides a general framework, which includes the direct and Bonferroni methods as extremes.

In Section 2 we give a technical introduction to our method and present theorems which demonstrate the asymptotic validity of the bootstrap simultaneous errors bars. In Section 3 simulations and the previous application are discussed. We describe this economic example in more detail and do a comparison of different grids of error bars through simulation. The simulations indicate that handling the bias is the most difficult aspect of this problem, especially when the regression function has substantial curvature. The analysis of Section 3 provides a quantification of this difficulty. For this reason, in the examples we considered, 80% error bars had actual coverage as poor as 50–65%. In Section 4 we give proofs of the theorems in Section 2.

2. Bootstrap error bars.

Stochastic design nonparametric regression is based on observations $\{(X_i, Y_i)\}_{i=1}^n \in \mathbb{R}^{d+1}$ and the goal is to estimate $m(x) = E(Y|X = x)$: $\mathbb{R}^d \to \mathbb{R}$. The form of the kernel regression estimator, developed



by Nadaraya (1964) and Watson (1964) is

$$\hat{m}_h(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{K_h(x - X_i) Y_i}{\hat{f}_h(x)}, \tag{2.1}$$

where

$$\hat{f}_h(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i) \tag{2.2}$$

and where $K_h(u) = h^{-d} K(u/h)$ is a kernel weight function with bandwidth h. All results of this paper are stated in terms of this estimator, although the essential ideas clearly extend to other types of kernel estimators such as those of Gasser and Müller (1984) and also to other regression estimators, such as spline methods, as discussed in Eubank (1988).

The choice of the bandwidth is crucial to the performance of the estimator. An asymptotic analysis of this choice and discussion of various data based bandwidth selectors may be found in Chapters 4 and 5 of Härdle (1989). The results of the present paper are formulated in such a way as to allow this type of objective bandwidth choice to be employed.

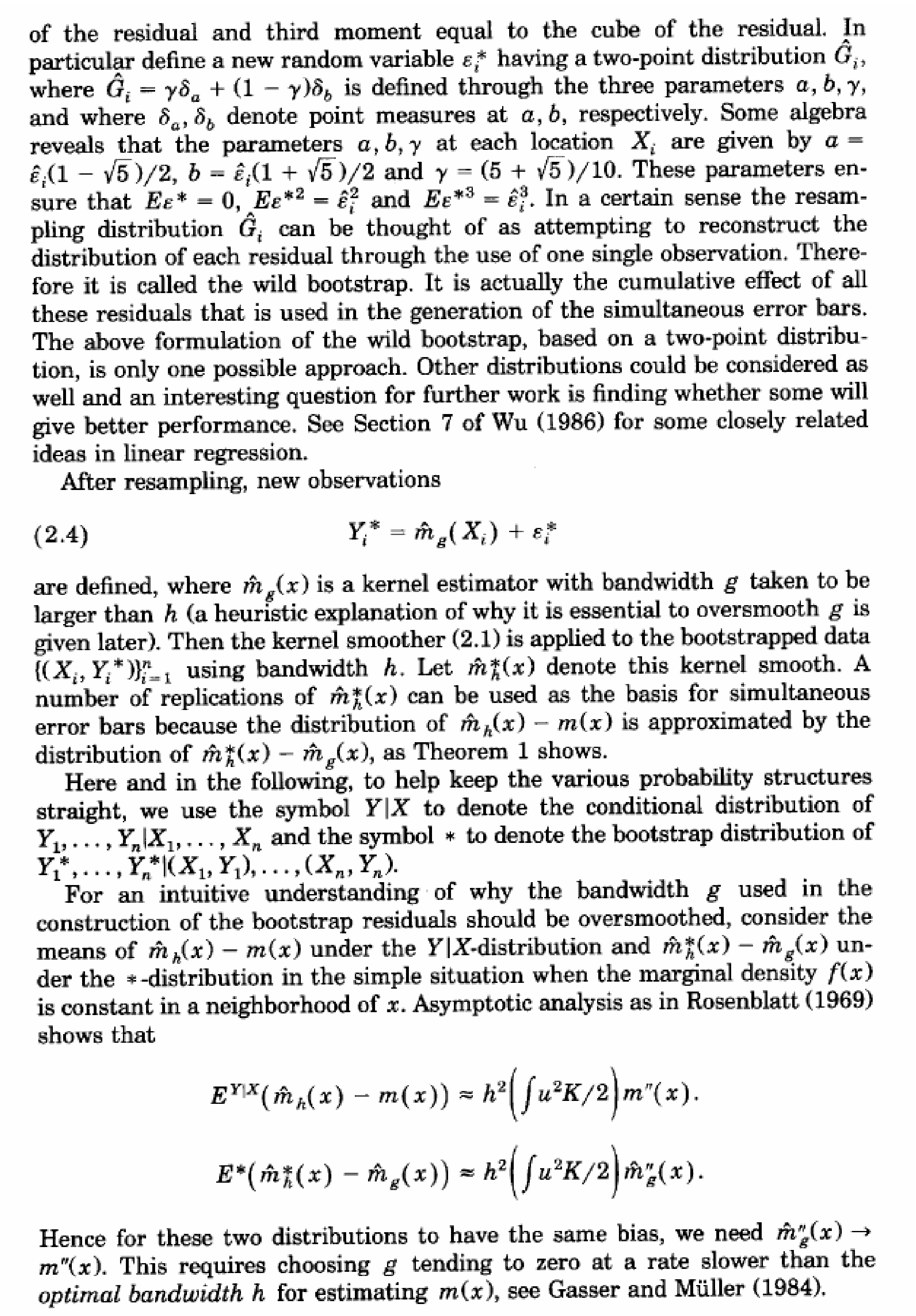
One approach to the problem of finding simultaneous error bars would be to work with limiting normal distributions of the estimator at the grid points. However, the joint distribution of the estimator at these gridpoints has substantial positive correlation, which makes the derivation of joint normal theory confidence intervals nontrivial. In fact, they essentially should be done by simulation methods. Since simulation methods are needed anyway, it seems better to use a more direct approach through bootstrapping, as opposed to relying on the normal approximation and also to facing the problems of parameter estimation that such an approach entails.

While bootstrap methods are well-known tools for assessing variability, more care must be taken to properly account for the type of bias encountered in nonparametric curve estimation. In particular, the naive bootstrap approach of resampling from the pairs $\{(X_i, Y_i): i = 1, \dots, n\}$ is inappropriate because the bootstrap bias will be 0. Our approach to this problem is to first use the estimated residual

$$\hat{\varepsilon}_i = Y_i - \hat{m}_h(X_i). \tag{2.3}$$

The essential idea is to resample from the estimated residuals, which are the differences between the observations and the pilot estimate and then use this data to construct an estimator whose distribution will approximate the distribution of the original estimator.

To better retain the conditional distributional characteristics of the estimate, we do not resample from the entire set of residuals, as in Härdle and Bowman (1988). One possibility would be to resample from a set of residuals determined by a window function, but this has the disadvantage of requiring choice of the window width. To avoid this we use the idea of *wild bootstrapping*, as proposed in Härdle and Mammen (1989) [but see Rosenblueth (1975) for access to related literature], where each bootstrap residual is drawn from the two-point distribution which has mean zero, variance equal to the square



of the residual and third moment equal to the cube of the residual. In particular define a new random variable ε_i^* having a two-point distribution $\hat{G}_i$, where $\hat{G}_i = \gamma\delta_a + (1-\gamma)\delta_b$ is defined through the three parameters a, b, γ, and where δ_a, δ_b denote point measures at a, b, respectively. Some algebra reveals that the parameters a, b, γ at each location X_i are given by $a = \hat{\varepsilon}_i(1-\sqrt{5})/2$, $b = \hat{\varepsilon}_i(1+\sqrt{5})/2$ and $\gamma = (5+\sqrt{5})/10$. These parameters ensure that $E\varepsilon^* = 0$, $E\varepsilon^{*2} = \hat{\varepsilon}_i^2$ and $E\varepsilon^{*3} = \hat{\varepsilon}_i^3$. In a certain sense the resampling distribution $\hat{G}_i$ can be thought of as attempting to reconstruct the distribution of each residual through the use of one single observation. Therefore it is called the wild bootstrap. It is actually the cumulative effect of all these residuals that is used in the generation of the simultaneous error bars. The above formulation of the wild bootstrap, based on a two-point distribution, is only one possible approach. Other distributions could be considered as well and an interesting question for further work is finding whether some will give better performance. See Section 7 of Wu (1986) for some closely related ideas in linear regression.

After resampling, new observations

$$Y_i^* = \hat{m}_g(X_i) + \varepsilon_i^* \tag{2.4}$$

are defined, where $\hat{m}_g(x)$ is a kernel estimator with bandwidth g taken to be larger than h (a heuristic explanation of why it is essential to oversmooth g is given later). Then the kernel smoother (2.1) is applied to the bootstrapped data $\{(X_i, Y_i^*)\}_{i=1}^n$ using bandwidth h. Let $\hat{m}_h^*(x)$ denote this kernel smooth. A number of replications of $\hat{m}_h^*(x)$ can be used as the basis for simultaneous error bars because the distribution of $\hat{m}_h(x) - m(x)$ is approximated by the distribution of $\hat{m}_h^*(x) - \hat{m}_g(x)$, as Theorem 1 shows.

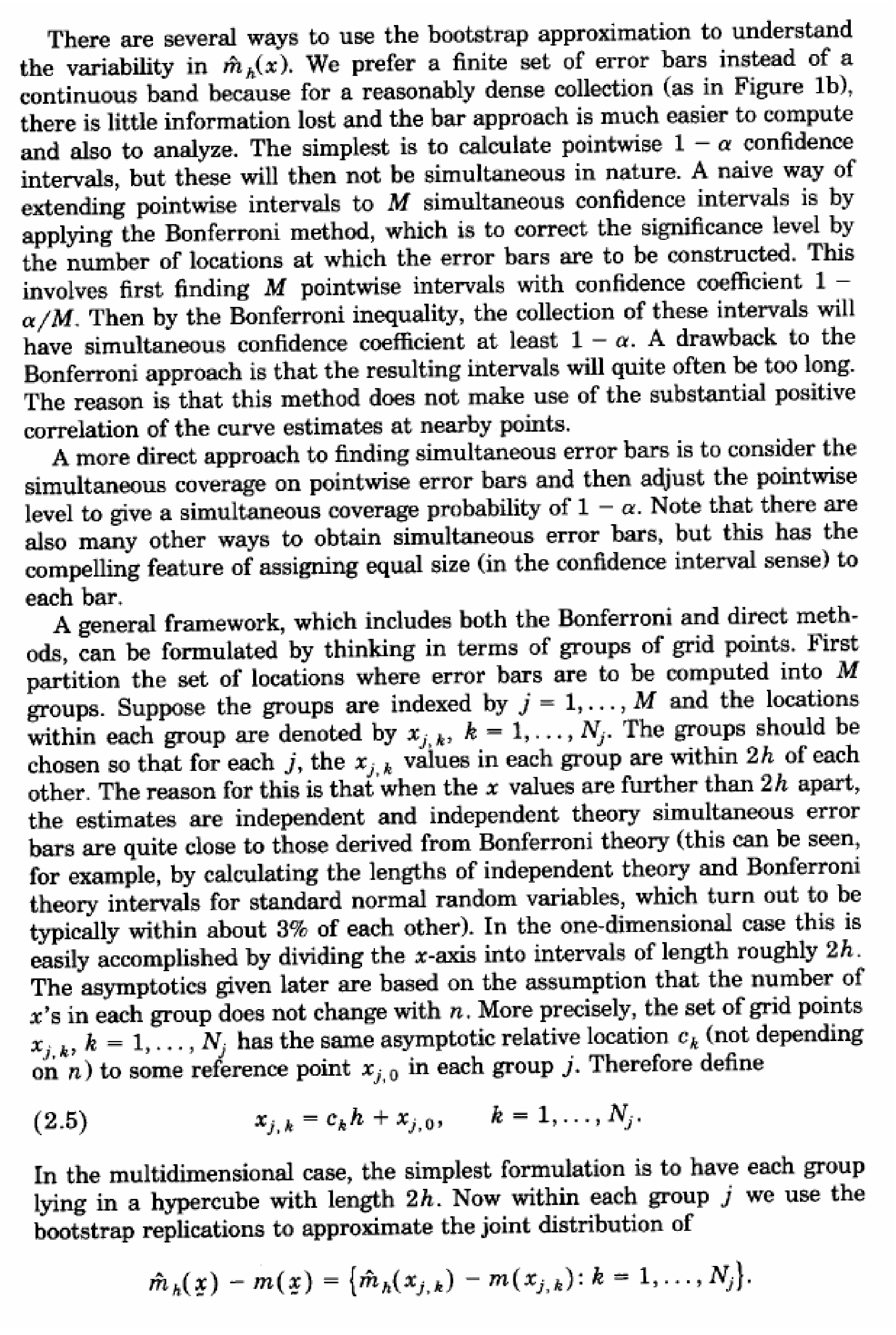
Here and in the following, to help keep the various probability structures straight, we use the symbol $Y|X$ to denote the conditional distribution of $Y_1, \ldots, Y_n | X_1, \ldots, X_n$ and the symbol $*$ to denote the bootstrap distribution of $Y_1^*, \ldots, Y_n^* | (X_1, Y_1), \ldots, (X_n, Y_n)$.

For an intuitive understanding of why the bandwidth g used in the construction of the bootstrap residuals should be oversmoothed, consider the means of $\hat{m}_h(x) - m(x)$ under the $Y|X$-distribution and $\hat{m}_h^*(x) - \hat{m}_g(x)$ under the $*$-distribution in the simple situation when the marginal density $f(x)$ is constant in a neighborhood of x. Asymptotic analysis as in Rosenblatt (1969) shows that

$$E^{Y|X}(\hat{m}_h(x) - m(x)) \approx h^2\left(\int u^2K/2\right)m''(x).$$

$$E^*(\hat{m}_h^*(x) - \hat{m}_g(x)) \approx h^2\left(\int u^2K/2\right)\hat{m}_g''(x).$$

Hence for these two distributions to have the same bias, we need $\hat{m}_g''(x) \to m''(x)$. This requires choosing g tending to zero at a rate slower than the *optimal bandwidth* h for estimating $m(x)$, see Gasser and Müller (1984).



There are several ways to use the bootstrap approximation to understand the variability in $\hat{m}_h(x)$. We prefer a finite set of error bars instead of a continuous band because for a reasonably dense collection (as in Figure 1b), there is little information lost and the bar approach is much easier to compute and also to analyze. The simplest is to calculate pointwise $1-\alpha$ confidence intervals, but these will then not be simultaneous in nature. A naive way of extending pointwise intervals to M simultaneous confidence intervals is by applying the Bonferroni method, which is to correct the significance level by the number of locations at which the error bars are to be constructed. This involves first finding M pointwise intervals with confidence coefficient $1-\alpha/M$. Then by the Bonferroni inequality, the collection of these intervals will have simultaneous confidence coefficient at least $1-\alpha$. A drawback to the Bonferroni approach is that the resulting intervals will quite often be too long. The reason is that this method does not make use of the substantial positive correlation of the curve estimates at nearby points.

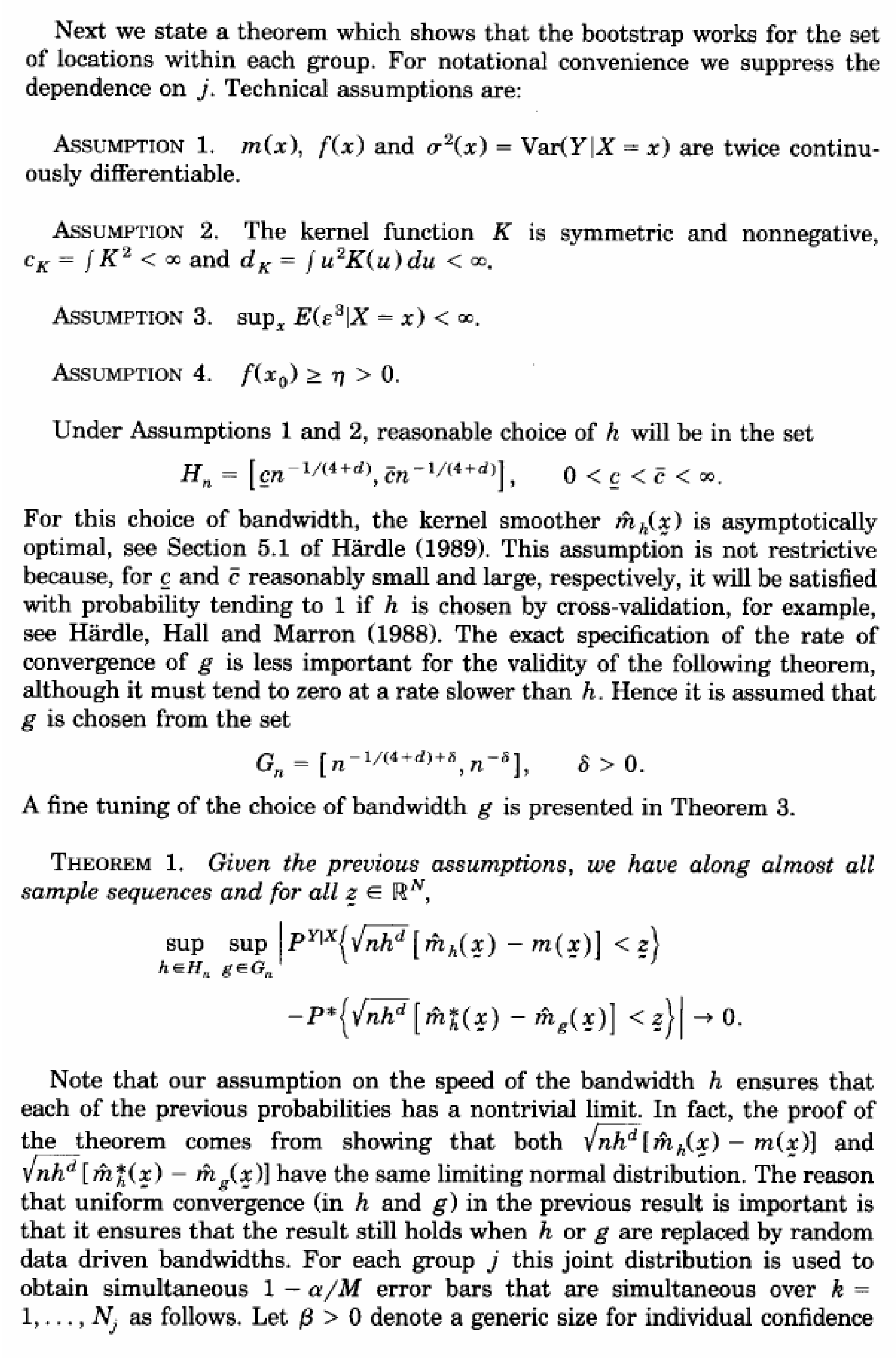
A more direct approach to finding simultaneous error bars is to consider the simultaneous coverage on pointwise error bars and then adjust the pointwise level to give a simultaneous coverage probability of $1-\alpha$. Note that there are also many other ways to obtain simultaneous error bars, but this has the compelling feature of assigning equal size (in the confidence interval sense) to each bar.

A general framework, which includes both the Bonferroni and direct methods, can be formulated by thinking in terms of groups of grid points. First partition the set of locations where error bars are to be computed into M groups. Suppose the groups are indexed by $j=1,\dots,M$ and the locations within each group are denoted by $x_{j,k}$, $k=1,\dots,N_j$. The groups should be chosen so that for each j, the $x_{j,k}$ values in each group are within $2h$ of each other. The reason for this is that when the x values are further than $2h$ apart, the estimates are independent and independent theory simultaneous error bars are quite close to those derived from Bonferroni theory (this can be seen, for example, by calculating the lengths of independent theory and Bonferroni theory intervals for standard normal random variables, which turn out to be typically within about 3% of each other). In the one-dimensional case this is easily accomplished by dividing the x-axis into intervals of length roughly $2h$. The asymptotics given later are based on the assumption that the number of x's in each group does not change with n. More precisely, the set of grid points $x_{j,k}$, $k=1,\dots,N_j$ has the same asymptotic relative location c_k (not depending on n) to some reference point $x_{j,0}$ in each group j. Therefore define

$$x_{j,k}=c_k h+x_{j,0},\qquad k=1,\dots,N_j. \tag{2.5}$$

In the multidimensional case, the simplest formulation is to have each group lying in a hypercube with length $2h$. Now within each group j we use the bootstrap replications to approximate the joint distribution of

$$\hat{m}_h(\underset{\sim}{x})-m(\underset{\sim}{x})=\left\{\hat{m}_h(x_{j,k})-m(x_{j,k}):k=1,\dots,N_j\right\}.$$



Next we state a theorem which shows that the bootstrap works for the set of locations within each group. For notational convenience we suppress the dependence on j. Technical assumptions are:

Assumption 1. $m(x)$, $f(x)$ and $\sigma^2(x) = \mathrm{Var}(Y|X = x)$ are twice continuously differentiable.

Assumption 2. The kernel function K is symmetric and nonnegative, $c_K = \int K^2 < \infty$ and $d_K = \int u^2K(u)\,du < \infty$.

Assumption 3. $\sup_x E(\varepsilon^3|X = x) < \infty$.

Assumption 4. $f(x_0) \geq \eta > 0$.

Under Assumptions 1 and 2, reasonable choice of h will be in the set

$$H_n = \left[\underline{c}n^{-1/(4+d)}, \bar{c}n^{-1/(4+d)}\right], \qquad 0 < \underline{c} < \bar{c} < \infty.$$

For this choice of bandwidth, the kernel smoother $\hat{m}_h(\underset{\sim}{x})$ is asymptotically optimal, see Section 5.1 of Härdle (1989). This assumption is not restrictive because, for $\underline{c}$ and $\bar{c}$ reasonably small and large, respectively, it will be satisfied with probability tending to 1 if h is chosen by cross-validation, for example, see Härdle, Hall and Marron (1988). The exact specification of the rate of convergence of g is less important for the validity of the following theorem, although it must tend to zero at a rate slower than h. Hence it is assumed that g is chosen from the set

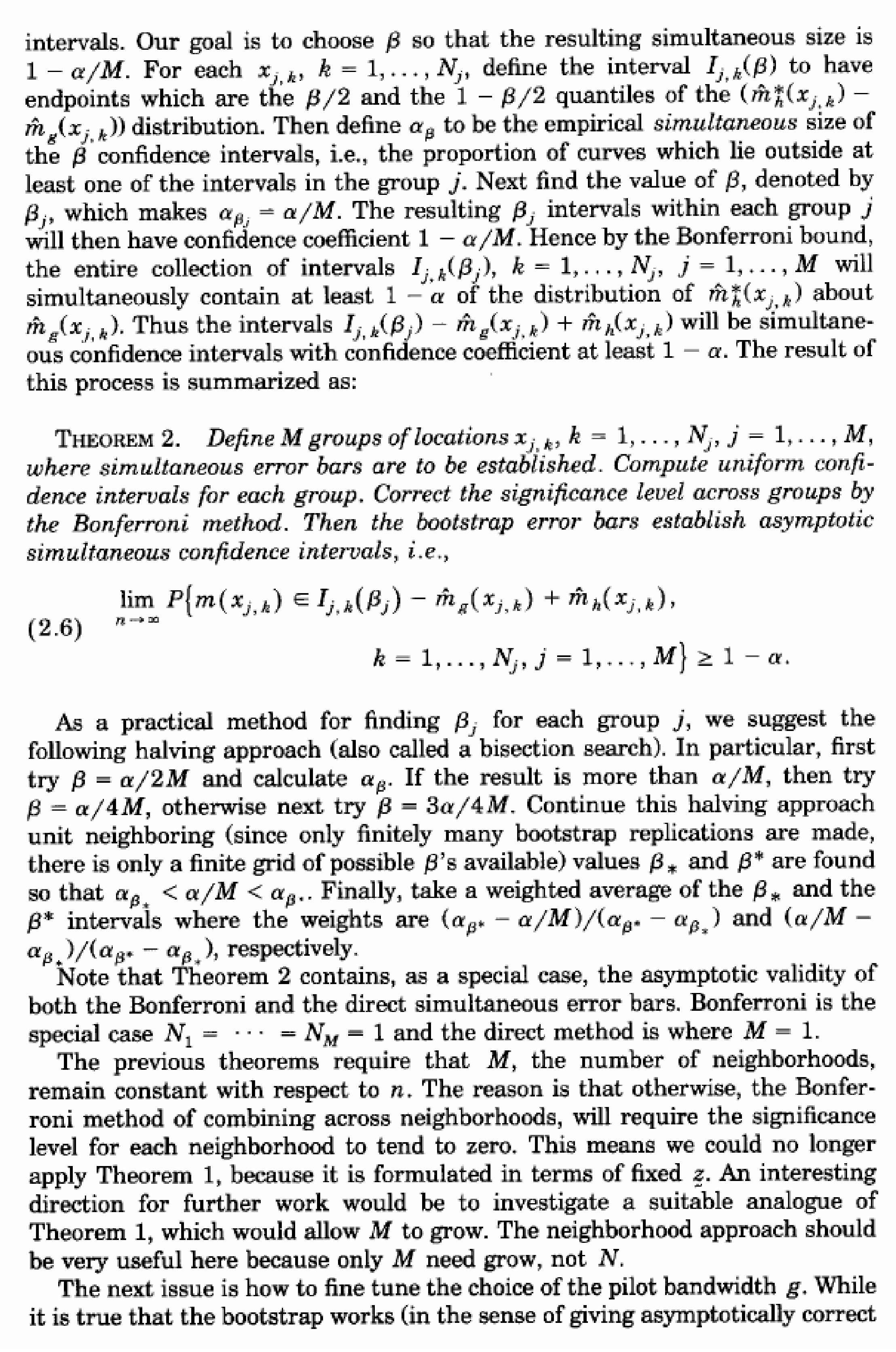
$$G_n = [n^{-1/(4+d)+\delta}, n^{-\delta}], \qquad \delta > 0.$$

A fine tuning of the choice of bandwidth g is presented in Theorem 3.

Theorem 1. *Given the previous assumptions, we have along almost all sample sequences and for all $\underset{\sim}{z} \in \mathbb{R}^N$,*

$$\sup_{h\in H_n}\ \sup_{g\in G_n}\left|P^{Y|X}\left\{\sqrt{nh^d}\left[\hat{m}_h(\underset{\sim}{x}) - m(\underset{\sim}{x})\right] < \underset{\sim}{z}\right\} - P^*\left\{\sqrt{nh^d}\left[\hat{m}_h^*(\underset{\sim}{x}) - \hat{m}_g(\underset{\sim}{x})\right] < \underset{\sim}{z}\right\}\right| \to 0.$$

Note that our assumption on the speed of the bandwidth h ensures that each of the previous probabilities has a nontrivial limit. In fact, the proof of the theorem comes from showing that both $\sqrt{nh^d}[\hat{m}_h(\underset{\sim}{x}) - m(\underset{\sim}{x})]$ and $\sqrt{nh^d}[\hat{m}_h^*(\underset{\sim}{x}) - \hat{m}_g(\underset{\sim}{x})]$ have the same limiting normal distribution. The reason that uniform convergence (in h and g) in the previous result is important is that it ensures that the result still holds when h or g are replaced by random data driven bandwidths. For each group j this joint distribution is used to obtain simultaneous $1 - \alpha/M$ error bars that are simultaneous over $k = 1, \ldots, N_j$ as follows. Let $\beta > 0$ denote a generic size for individual confidence



intervals. Our goal is to choose β so that the resulting simultaneous size is $1-\alpha/M$. For each $x_{j,k}$, $k=1,\dots,N_j$, define the interval $I_{j,k}(\beta)$ to have endpoints which are the $\beta/2$ and the $1-\beta/2$ quantiles of the $(\hat{m}_h^*(x_{j,k})-\hat{m}_g(x_{j,k}))$ distribution. Then define α_β to be the empirical *simultaneous* size of the β confidence intervals, i.e., the proportion of curves which lie outside at least one of the intervals in the group j. Next find the value of β, denoted by β_j, which makes $\alpha_{\beta_j}=\alpha/M$. The resulting β_j intervals within each group j will then have confidence coefficient $1-\alpha/M$. Hence by the Bonferroni bound, the entire collection of intervals $I_{j,k}(\beta_j)$, $k=1,\dots,N_j$, $j=1,\dots,M$ will simultaneously contain at least $1-\alpha$ of the distribution of $\hat{m}_h^*(x_{j,k})$ about $\hat{m}_g(x_{j,k})$. Thus the intervals $I_{j,k}(\beta_j)-\hat{m}_g(x_{j,k})+\hat{m}_h(x_{j,k})$ will be simultaneous confidence intervals with confidence coefficient at least $1-\alpha$. The result of this process is summarized as:

THEOREM 2. *Define M groups of locations $x_{j,k}$, $k=1,\dots,N_j$, $j=1,\dots,M$, where simultaneous error bars are to be established. Compute uniform confidence intervals for each group. Correct the significance level across groups by the Bonferroni method. Then the bootstrap error bars establish asymptotic simultaneous confidence intervals, i.e.*,

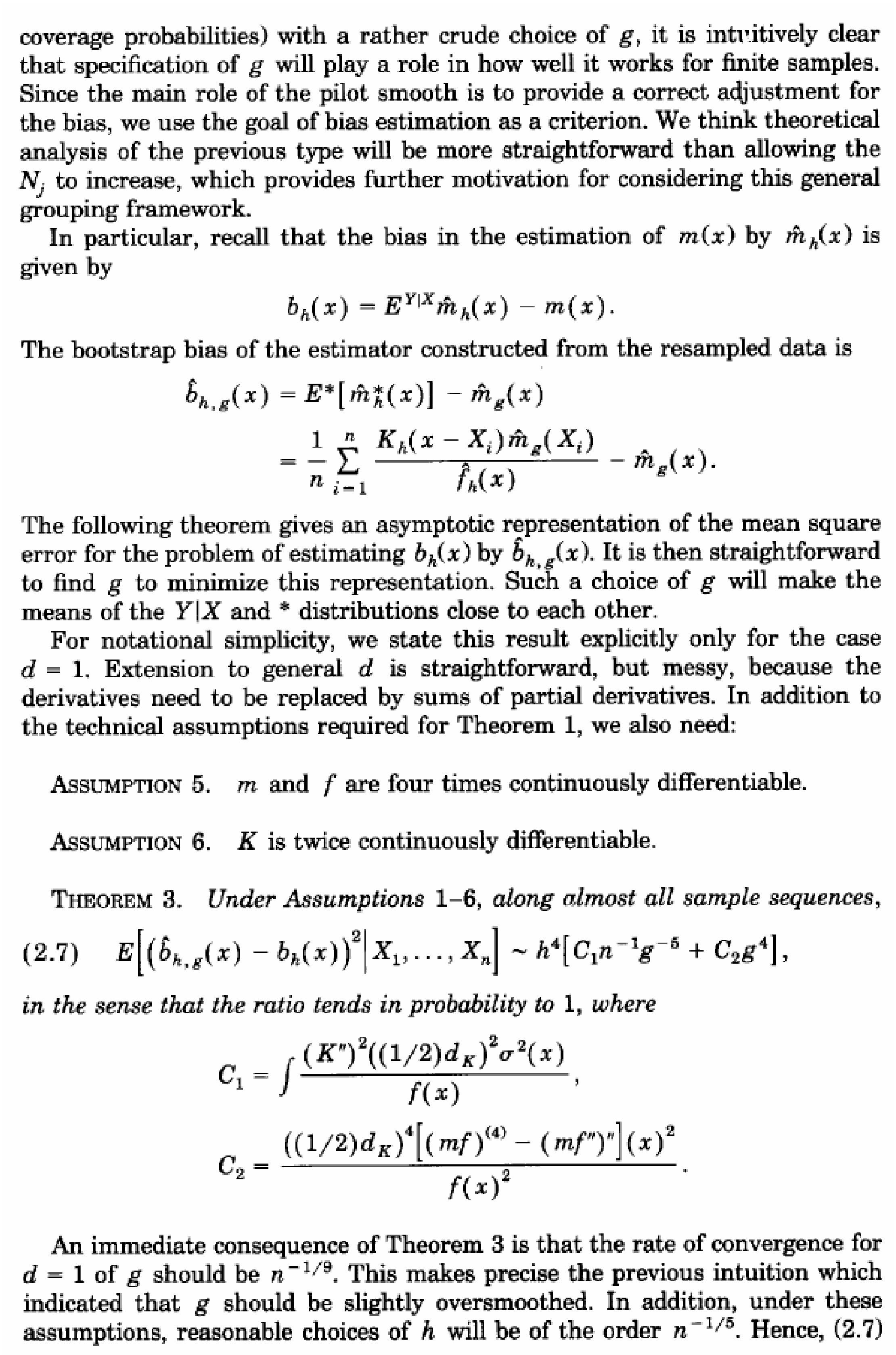
$$\lim_{n\to\infty} P\{m(x_{j,k})\in I_{j,k}(\beta_j)-\hat{m}_g(x_{j,k})+\hat{m}_h(x_{j,k}),\quad k=1,\dots,N_j,\ j=1,\dots,M\}\ge 1-\alpha. \tag{2.6}$$

As a practical method for finding β_j for each group j, we suggest the following halving approach (also called a bisection search). In particular, first try $\beta=\alpha/2M$ and calculate α_β. If the result is more than α/M, then try $\beta=\alpha/4M$, otherwise next try $\beta=3\alpha/4M$. Continue this halving approach unit neighboring (since only finitely many bootstrap replications are made, there is only a finite grid of possible β's available) values β_* and β^* are found so that $\alpha_{\beta_*}<\alpha/M<\alpha_{\beta^*}$.. Finally, take a weighted average of the β_* and the β^* intervals where the weights are $(\alpha_{\beta^*}-\alpha/M)/(\alpha_{\beta^*}-\alpha_{\beta_*})$ and $(\alpha/M-\alpha_{\beta_*})/(\alpha_{\beta^*}-\alpha_{\beta_*})$, respectively.

Note that Theorem 2 contains, as a special case, the asymptotic validity of both the Bonferroni and the direct simultaneous error bars. Bonferroni is the special case $N_1=\cdots=N_M=1$ and the direct method is where $M=1$.

The previous theorems require that M, the number of neighborhoods, remain constant with respect to n. The reason is that otherwise, the Bonferroni method of combining across neighborhoods, will require the significance level for each neighborhood to tend to zero. This means we could no longer apply Theorem 1, because it is formulated in terms of fixed $\underset{\sim}{z}$. An interesting direction for further work would be to investigate a suitable analogue of Theorem 1, which would allow M to grow. The neighborhood approach should be very useful here because only M need grow, not N.

The next issue is how to fine tune the choice of the pilot bandwidth g. While it is true that the bootstrap works (in the sense of giving asymptotically correct



coverage probabilities) with a rather crude choice of g, it is intuitively clear that specification of g will play a role in how well it works for finite samples. Since the main role of the pilot smooth is to provide a correct adjustment for the bias, we use the goal of bias estimation as a criterion. We think theoretical analysis of the previous type will be more straightforward than allowing the N_j to increase, which provides further motivation for considering this general grouping framework.

In particular, recall that the bias in the estimation of $m(x)$ by $\hat{m}_h(x)$ is given by

$$b_h(x) = E^{Y|X}\hat{m}_h(x) - m(x).$$

The bootstrap bias of the estimator constructed from the resampled data is

$$\begin{aligned}\hat{b}_{h,g}(x) &= E^*[\hat{m}_h^*(x)] - \hat{m}_g(x)\\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{K_h(x - X_i)\hat{m}_g(X_i)}{\hat{f}_h(x)} - \hat{m}_g(x).\end{aligned}$$

The following theorem gives an asymptotic representation of the mean square error for the problem of estimating $b_h(x)$ by $\hat{b}_{h,g}(x)$. It is then straightforward to find g to minimize this representation. Such a choice of g will make the means of the $Y|X$ and * distributions close to each other.

For notational simplicity, we state this result explicitly only for the case $d = 1$. Extension to general d is straightforward, but messy, because the derivatives need to be replaced by sums of partial derivatives. In addition to the technical assumptions required for Theorem 1, we also need:

Assumption 5. m and f are four times continuously differentiable.

Assumption 6. K is twice continuously differentiable.

Theorem 3. *Under Assumptions* 1–6, *along almost all sample sequences*,

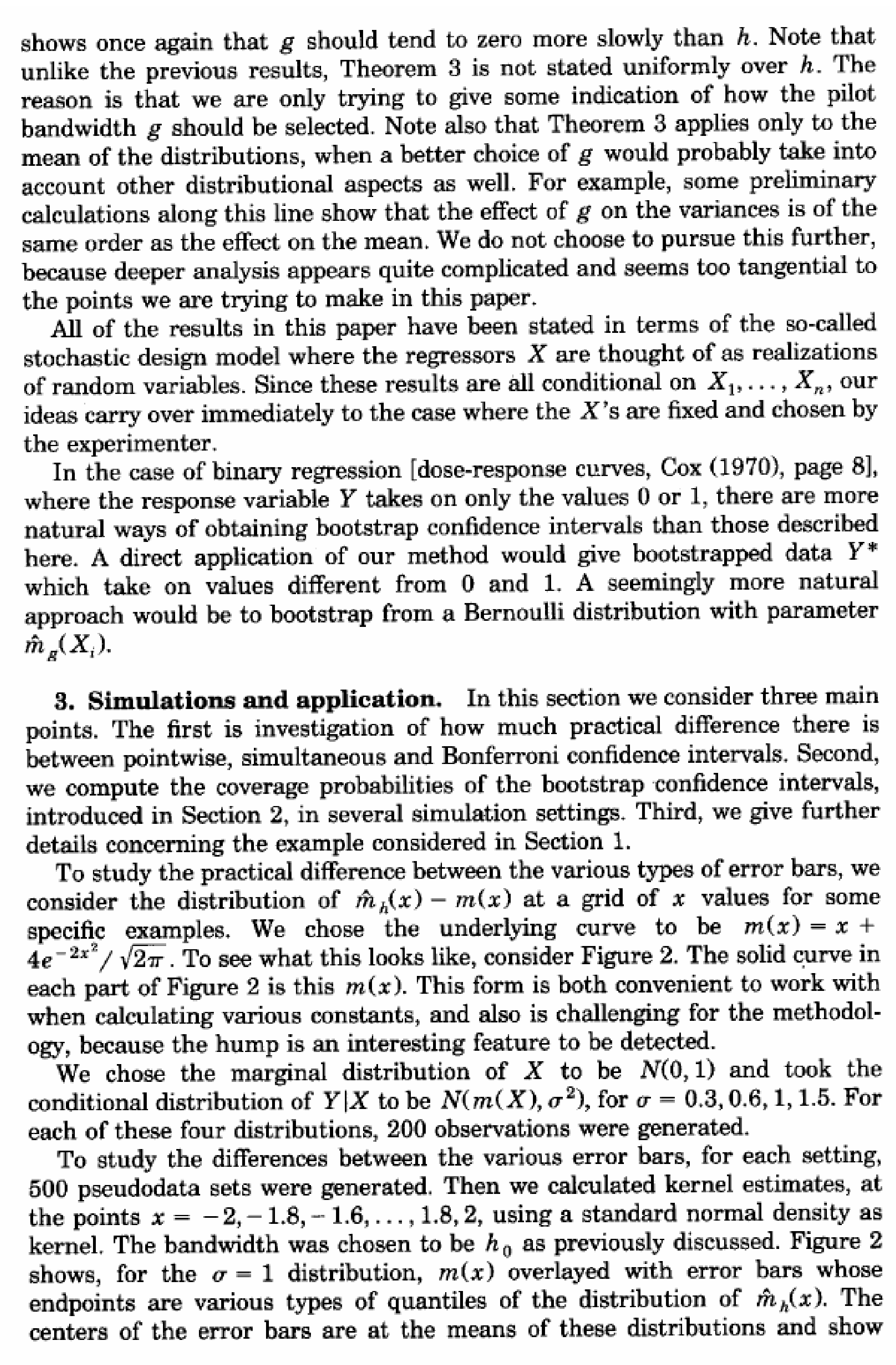
$$E\left[\left(\hat{b}_{h,g}(x) - b_h(x)\right)^2 \middle| X_1, \dots, X_n\right] \sim h^4\left[C_1 n^{-1} g^{-5} + C_2 g^4\right], \tag{2.7}$$

in the sense that the ratio tends in probability to 1, *where*

$$C_1 = \int \frac{(K'')^2((1/2)d_K)^2\sigma^2(x)}{f(x)},$$

$$C_2 = \frac{((1/2)d_K)^4\left[(mf)^{(4)} - (mf'')''\right](x)^2}{f(x)^2}.$$

An immediate consequence of Theorem 3 is that the rate of convergence for $d = 1$ of g should be $n^{-1/9}$. This makes precise the previous intuition which indicated that g should be slightly oversmoothed. In addition, under these assumptions, reasonable choices of h will be of the order $n^{-1/5}$. Hence, (2.7)



shows once again that g should tend to zero more slowly than h. Note that unlike the previous results, Theorem 3 is not stated uniformly over h. The reason is that we are only trying to give some indication of how the pilot bandwidth g should be selected. Note also that Theorem 3 applies only to the mean of the distributions, when a better choice of g would probably take into account other distributional aspects as well. For example, some preliminary calculations along this line show that the effect of g on the variances is of the same order as the effect on the mean. We do not choose to pursue this further, because deeper analysis appears quite complicated and seems too tangential to the points we are trying to make in this paper.

All of the results in this paper have been stated in terms of the so-called stochastic design model where the regressors X are thought of as realizations of random variables. Since these results are all conditional on $X_1, \dots, X_n$, our ideas carry over immediately to the case where the X's are fixed and chosen by the experimenter.

In the case of binary regression [dose-response curves, Cox (1970), page 8], where the response variable Y takes on only the values 0 or 1, there are more natural ways of obtaining bootstrap confidence intervals than those described here. A direct application of our method would give bootstrapped data Y^* which take on values different from 0 and 1. A seemingly more natural approach would be to bootstrap from a Bernoulli distribution with parameter $\hat{m}_g(X_i)$.

3. Simulations and application. In this section we consider three main points. The first is investigation of how much practical difference there is between pointwise, simultaneous and Bonferroni confidence intervals. Second, we compute the coverage probabilities of the bootstrap confidence intervals, introduced in Section 2, in several simulation settings. Third, we give further details concerning the example considered in Section 1.

To study the practical difference between the various types of error bars, we consider the distribution of $\hat{m}_h(x) - m(x)$ at a grid of x values for some specific examples. We chose the underlying curve to be $m(x) = x + 4e^{-2x^2}/\sqrt{2\pi}$. To see what this looks like, consider Figure 2. The solid curve in each part of Figure 2 is this $m(x)$. This form is both convenient to work with when calculating various constants, and also is challenging for the methodology, because the hump is an interesting feature to be detected.

We chose the marginal distribution of X to be $N(0,1)$ and took the conditional distribution of $Y|X$ to be $N(m(X), \sigma^2)$, for $\sigma = 0.3, 0.6, 1, 1.5$. For each of these four distributions, 200 observations were generated.

To study the differences between the various error bars, for each setting, 500 pseudodata sets were generated. Then we calculated kernel estimates, at the points $x = -2, -1.8, -1.6, \dots, 1.8, 2$, using a standard normal density as kernel. The bandwidth was chosen to be h_0 as previously discussed. Figure 2 shows, for the $\sigma = 1$ distribution, $m(x)$ overlayed with error bars whose endpoints are various types of quantiles of the distribution of $\hat{m}_h(x)$. The centers of the error bars are at the means of these distributions and show

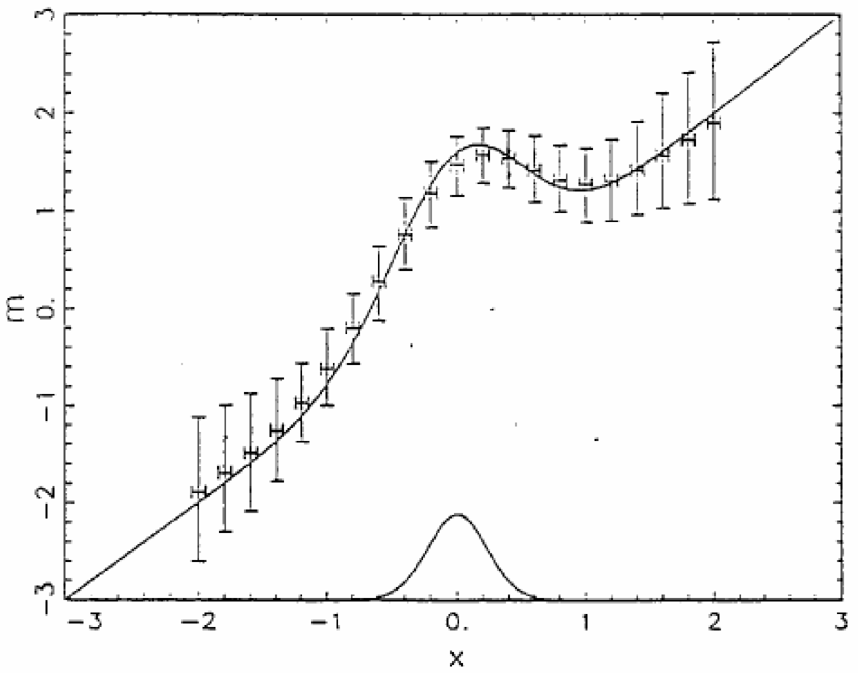


FIG. 2. *Overlay of $m(x)$ with empirical (from 500 simulation runs) quantiles of $\hat{m}_{h_0}(x)$ distribution. Centers of bars are means of distributions. Error bars are 80% simultaneous.*

clearly the bias that is inherent to nonparametric regression estimation. Note in particular how substantial bias is caused by both the curvature of $m(x)$ near the hump and by the curvature of $f(x)$, near $x = -2, 2$. The bars in Figure 2 are simultaneous bars.

For easy comparison of the lengths of these intervals with the other types, consider Figure 3. This shows, for the same x values, the lengths of the four types of bars. Of course these bars are all shorter near the center, which reflects the fact that there is more data there, so the estimates are more accurate. As expected, the lengths increase from pointwise, to actual simultaneous, to neighborhood, to Bonferroni. Also note that, as stated in Section 2, the difference between the actual simultaneous bars and the neighborhood simultaneous bars is really quite small, while the pointwise are a lot narrower. The one perhaps surprising feature is that the Bonferroni bars are not very much wider than the neighborhood bars.

To see how the bootstrap methodology proposed in Section 2 performed for the simulation settings considered here, we calculated estimates of the simultaneous coverage probabilities for 21 equally spaced error bars on $[-1, 1]$. These estimates were calculated by applying the methodology to 500 psuedo-data sets, for each of the various settings. For each data set we used 500 bootstrap replications. The pilot bandwidth g was taken to minimize a global version of the asymptotic representation given in (2.7), where the quantities that depend on x were replaced by their integral over $[-1, 1]$. The bootstrap

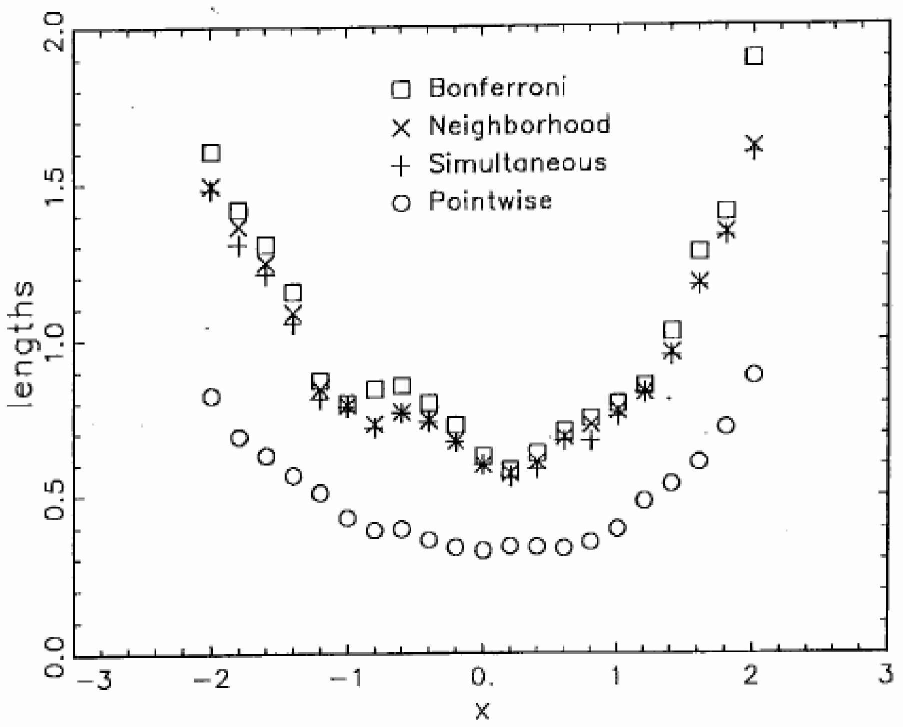


FIG. 3. *Lengths of the bars in Figure 3, x locations are the same.*

distribution was then used to derive the four types of error bars: pointwise, actual simultaneous, neighborhood simultaneous and Bonferroni. Then for each type of bar, the estimated simultaneous coverage probability is the proportion of times that the 500 bars cover the true curve $m(x)$ at each x value. The estimates are given in Table 1. To give an idea of the Monte Carlo variability in these estimates, also included are the radii of approximate 95% confidence intervals, of the form $1.96\sqrt{\hat{p}(1-\hat{p})}\,/\,\sqrt{500}$, where $\hat{p}$ is the estimated probability. Such confidence intervals are of course rather poor for $\hat{p}$

TABLE 1
Estimated (from 500 simulation runs) coverage probabilities for bootstrap error bars

	Pointwise	Simultaneous	Neighborhood	Bonferroni
$\sigma = 0.3,\ h = h_0$	0.03 ± 0.02	0.52 ± 0.04	0.55 ± 0.04	0.65 ± 0.04
$\sigma = 0.6,\ h = h_0$	0.09 ± 0.02	0.55 ± 0.04	0.59 ± 0.04	0.69 ± 0.04
$\sigma = 1.0,\ h = h_0$	0.10 ± 0.03	0.59 ± 0.04	0.63 ± 0.04	0.74 ± 0.04
$\sigma = 1.5,\ h = h_0$	0.16 ± 0.03	0.56 ± 0.04	0.65 ± 0.04	0.79 ± 0.04
$\sigma = 1.0,\ h = h_0/2$	0.04 ± 0.02	0.57 ± 0.04	0.60 ± 0.04	0.65 ± 0.04
$\sigma = 1.0,\ h = h_0$	0.10 ± 0.03	0.59 ± 0.04	0.63 ± 0.04	0.74 ± 0.04
$\sigma = 1.0,\ h = 2*h_0$	0.01 ± 0.01	0.10 ± 0.03	0.16 ± 0.03	0.33 ± 0.04

close to 0, but in most cases suffice to give a decent idea of the variability involved.

This table looks somewhat disappointing since the observed coverage probabilities are all significantly below the desired value of 80%. Careful investigation revealed that this was due to problems with the estimated bias. More precisely it was caused by a systematic underadjustment in our bias correction (i.e., bias in the estimated bias adjustment). In Figure 4 the difference between the solid curve $m(x)$ and the dashed curve $E\hat{m}_h(x)$ is the true bias for our simulation setting in the case $\sigma = 0.3$, $h = h_0$. This bias is estimated for each data set by the difference between $\hat{m}_g(x)$ and $E^*\hat{m}_h^*(x)$. The bias in this estimation process is then the difference between the curve made of dots and dashes $E\hat{m}_g(x)$ and the dotted curve $E(E^*\hat{m}_h^*(x))$. Observe that because $E\hat{m}_g(x)$ has less curvature than $m(x)$, the estimated bias will typically be smaller than the actual bias. The effect does not look very large, but simultaneous coverage turns out to be a very sensitive quantity. Note that this also explains why the $h = 2 * h_0$ line of Table 1 has much smaller coverage probabilities than the others, since such a large h value means more bias than in the other settings. Of course this bias effect goes away asymptotically, but in the example considered here, Figure 4 shows that it is not negligible (and we believe this problem will exist quite often). Experiments with different values

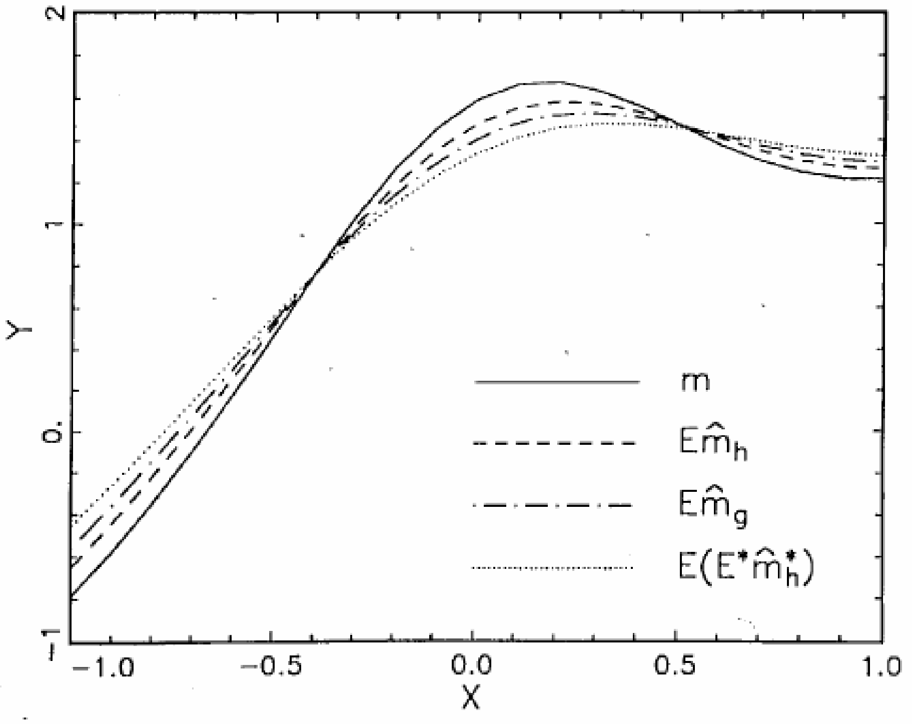


FIG. 4. *Comparison of true bias* ($E\hat{m}_h - m$) *with expected estimated bias* ($E(E^*\hat{m}_g^*) - E\hat{m}_g$) *for* $\sigma = 0.3$, $h = h_0$.

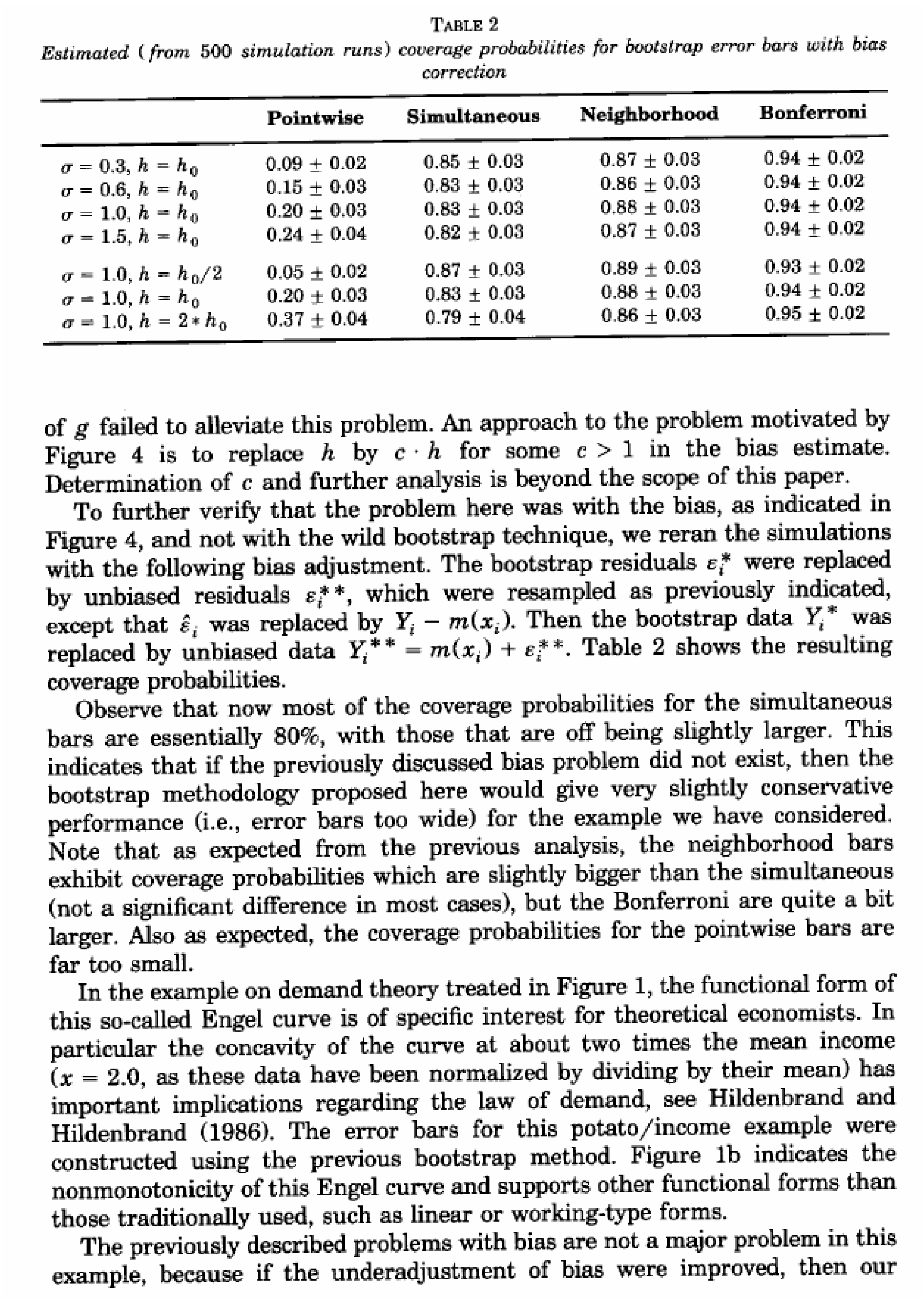


TABLE 2

Estimated (from 500 simulation runs) coverage probabilities for bootstrap error bars with bias correction

	Pointwise	**Simultaneous**	**Neighborhood**	**Bonferroni**
$\sigma = 0.3$, $h = h_0$	0.09 ± 0.02	0.85 ± 0.03	0.87 ± 0.03	0.94 ± 0.02
$\sigma = 0.6$, $h = h_0$	0.15 ± 0.03	0.83 ± 0.03	0.86 ± 0.03	0.94 ± 0.02
$\sigma = 1.0$, $h = h_0$	0.20 ± 0.03	0.83 ± 0.03	0.88 ± 0.03	0.94 ± 0.02
$\sigma = 1.5$, $h = h_0$	0.24 ± 0.04	0.82 ± 0.03	0.87 ± 0.03	0.94 ± 0.02
$\sigma = 1.0$, $h = h_0/2$	0.05 ± 0.02	0.87 ± 0.03	0.89 ± 0.03	0.93 ± 0.02
$\sigma = 1.0$, $h = h_0$	0.20 ± 0.03	0.83 ± 0.03	0.88 ± 0.03	0.94 ± 0.02
$\sigma = 1.0$, $h = 2 * h_0$	0.37 ± 0.04	0.79 ± 0.04	0.86 ± 0.03	0.95 ± 0.02

of g failed to alleviate this problem. An approach to the problem motivated by Figure 4 is to replace h by $c \cdot h$ for some $c > 1$ in the bias estimate. Determination of c and further analysis is beyond the scope of this paper.

To further verify that the problem here was with the bias, as indicated in Figure 4, and not with the wild bootstrap technique, we reran the simulations with the following bias adjustment. The bootstrap residuals ε_i^* were replaced by unbiased residuals ε_i^{**}, which were resampled as previously indicated, except that $\hat{\varepsilon}_i$ was replaced by $Y_i - m(x_i)$. Then the bootstrap data Y_i^* was replaced by unbiased data $Y_i^{**} = m(x_i) + \varepsilon_i^{**}$. Table 2 shows the resulting coverage probabilities.

Observe that now most of the coverage probabilities for the simultaneous bars are essentially 80%, with those that are off being slightly larger. This indicates that if the previously discussed bias problem did not exist, then the bootstrap methodology proposed here would give very slightly conservative performance (i.e., error bars too wide) for the example we have considered. Note that as expected from the previous analysis, the neighborhood bars exhibit coverage probabilities which are slightly bigger than the simultaneous (not a significant difference in most cases), but the Bonferroni are quite a bit larger. Also as expected, the coverage probabilities for the pointwise bars are far too small.

In the example on demand theory treated in Figure 1, the functional form of this so-called Engel curve is of specific interest for theoretical economists. In particular the concavity of the curve at about two times the mean income ($x = 2.0$, as these data have been normalized by dividing by their mean) has important implications regarding the law of demand, see Hildenbrand and Hildenbrand (1986). The error bars for this potato/income example were constructed using the previous bootstrap method. Figure 1b indicates the nonmonotonicity of this Engel curve and supports other functional forms than those traditionally used, such as linear or working-type forms.

The previously described problems with bias are not a major problem in this example, because if the underadjustment of bias were improved, then our conclusion of concavity near $x = 2.0$ is in fact strengthened. Also as the sample size in much larger now, it seems reasonable to hope that the asymptotic negligibility of the bias problem is closer to being realized.

4. Proofs.

PROOF OF THEOREM 1. For notational simplicity, the proof is given explicitly only for the case $d = 1$. The theorem is an immediate consequence of the following lemmas.

LEMMA 1. *Along almost all sample sequences,*

$$\sqrt{nh}\,[\hat{m}_h(\underset{\sim}{x}) - m(\underset{\sim}{x})] \to N(B, V),$$

uniformly in h and g, in the sense that for all $\underset{\sim}{z} \in \mathbb{R}^N$,

$$\sup_{h \in H_n} \sup_{g \in G_n} \left| P^{Y|X}\{\sqrt{nh}\,[\hat{m}_h(\underset{\sim}{x}) - m(\underset{\sim}{x})] < \underset{\sim}{z}\} - \Phi_{B,V}(\underset{\sim}{z}) \right| \to 0,$$

where $\Phi_{B,V}$ denotes the normal cumulative distribution with mean B and covariance V and where

$$B = d_K\left\{ m''(\underset{\sim}{x}) + 2m'(\underset{\sim}{x})\frac{f'(\underset{\sim}{x})}{f(\underset{\sim}{x})} \right\},$$

$$V = (v_{kl}), \qquad v_{kl} = \frac{K^{(2)}(c_k - c_l)\sigma^2(x_0)}{f(x_0)}$$

for $K^{(2)}$ the convolution of K with itself.

LEMMA 2. *Along almost all sample sequences,*

$$\sqrt{nh}\,[\hat{m}_h^*(\underset{\sim}{x}) - \hat{m}_g(\underset{\sim}{x})] \to N(B, V),$$

*uniformly in h and g, in the same sense as in Lemma 1 (except that the $Y|X$ distribution is replaced by the * distribution).*

PROOF OF LEMMA 1. The Cramér–Wold device is used in this proof. We will show that for all $\underset{\sim}{t} \in \mathbb{R}^N$ and all $z \in \mathbb{R}$,

$$\begin{aligned} (4.1) \qquad & \Big| P^{Y|X}\{\underset{\sim}{t}^T(\sqrt{nh}\,[\hat{m}_h(\underset{\sim}{x}) - m(\underset{\sim}{x})]) < z\} \\ & \qquad - \Phi\Big((z - \underset{\sim}{t}^T B)\Big/\sqrt{\underset{\sim}{t}^T V \underset{\sim}{t}}\Big) \Big| \to 0, \end{aligned}$$

uniformly over $h \in H_n$, where Φ denotes the univariate standard normal c.d.f. To obtain uniformity over h requires some modification of the Cramér–Wold device. In particular, Theorems 7.6 and 7.7 of Billingsley (1968) need to be extended in a straightforward fashion. To establish this, following Härdle and Marron (1985), we first make the linear approximation

$$\sqrt{nh}\,[\hat{m}_h(\underset{\sim}{x}) - m(\underset{\sim}{x})] = L_n + o_p(L_n), \tag{4.2}$$

where

$$L_n = \sqrt{nh}\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{K_h(\underset{\sim}{x} - X_i)[Y_i - m(\underset{\sim}{x})]}{f(\underset{\sim}{x})}\right\}.$$

The term $o_p(L_n)$ is of lower order uniformly over H_n by (5.1) of Härdle and Marron (1985) and by Lemma 1 of that paper. Now write

$$L_n = V_n + B_n,$$

where

$$V_n = \sqrt{nh}\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{K_h(\underset{\sim}{x} - X_i)\varepsilon_i}{f(\underset{\sim}{x})}\right\}$$

and $\varepsilon_i = Y_i - m(X_i)$,

$$B_n = \sqrt{nh}\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{K_h(\underset{\sim}{x} - X_i)[m(X_i) - m(\underset{\sim}{x})]}{f(\underset{\sim}{x})}\right\}.$$

The proof of Lemma 1 follows from

$$\underset{\sim}{t}^T V_n \to N(0, \underset{\sim}{t}^T V \underset{\sim}{t}), \tag{4.3}$$

$$\underset{\sim}{t}^T B_n \to \underset{\sim}{t}^T B, \tag{4.4}$$

uniformly over $h \in H_n$.

To prove (4.1), we use Esseen's inequality for arbitrary independent random variables given, for example, on page 111 of Petrov (1975). For this purpose define $W_{hi}(\underset{\sim}{x}) = n^{-1/2}h^{1/2}K_h(\underset{\sim}{x} - X_i)/f(\underset{\sim}{x})$,

$$S_{2n} = \sum_{i=1}^{n}\operatorname{Var}(\underset{\sim}{t}TW_{hi}(\underset{\sim}{x})\varepsilon_i | X_1, \ldots, X_n)$$

and

$$S_{3n} = \sum_{i=1}^{n}E\left(|\underset{\sim}{t}^T W_{hi}(\underset{\sim}{x})\varepsilon_i|^3 | X_1, \ldots, X_n\right).$$

The Esseen inequality completes the verification of (4.3), when we show that

$$\sup_h S_{3n}/S_{2n}^{3/2} = o(1) \quad \text{a.s.}$$

To evaluate S_{2n}, note that $E^X S_{2n} = \underset{\sim}{t}^T V_{1n}\underset{\sim}{t}$, where the (k,l) element of V_{1n} is by the assumption on x_k

$$\int K_h(x_k - u)K_h(x_l - u)f(u)\sigma^2(u)\,du/(f(x_k)f(x_l))$$

$$= h^{-1}K^{(2)}(c_l - c_k)\sigma^2(x_0)/f(x_0) + O(h^2).$$

Since $S_{2n} \to ES_{2n}$, a.s. by Theorem 1 of Feller [(1970), page 238] we have that

$$S_{2n} = h^{-1}K^{(2)}(c_l - c_k)\sigma^2(x_0)/f(x_0) + o(1) \quad \text{a.s.}$$

Uniformity over h is obtained by a suitable strengthening of the previous theorem. In the same manner the term S_{3n} can be evaluated to see that

$$\sup_h n^{1/2}h^{1/2}S_{3n} = O(1) \quad \text{a.s.}$$

Thus the statement (4.3) follows.

For the proof of (4.4), see the bias evaluation in Collomb (1981) or Härdle (1989). □

The proof of Proof of Lemma 2 is similar in spirit to that of Lemma 1, but is slightly more complicated because more terms arise.

Proof of Theorem 3. The proof of (2.7) uses methods related to those in the proof of Theorem 1, so only the main steps are explicitly given. The first step is to decompose into variance and squared bias components,

$$E\Big[\big(\hat{b}_{h,g}(x) - b_h(x)\big)^2\Big|X_1,\dots,X_n\Big] = \mathscr{V}_n + \mathscr{B}_n^2, \tag{4.5}$$

where

$$\mathscr{V}_n = \operatorname{Var}\big(\hat{b}_{h,g}(x)\big|X_1,\dots,X_n\big),$$

$$\mathscr{B}_n = E\big(\hat{b}_{h,g}(x) - b_h(x)\big|X_1,\dots,X_n\big).$$

Using the same linearization technique as at (4.2) together with

$$\mathscr{B}_n = \mathscr{B}_{n1} + o(\mathscr{B}_{n1}),$$

where

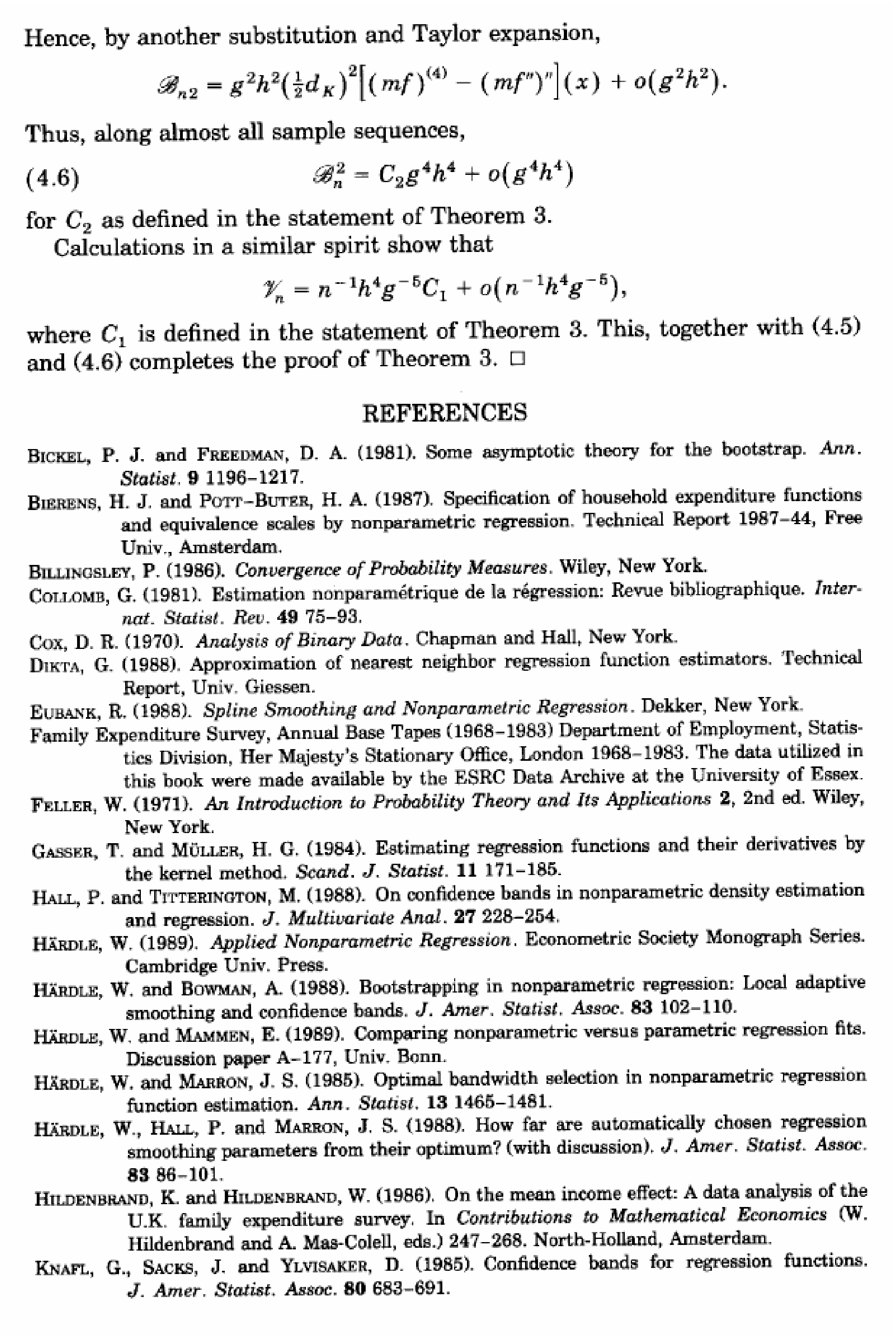
$$\mathscr{B}_{n1} = \Big[\int K_g(x - t)\mathscr{U}_h(t)\,dt - \mathscr{U}_h(x)\Big]\Big/f(x)$$

for

$$\mathscr{U}_h(x) = \int K_h(x - s)[m(s) - m(x)]f(s)\,ds.$$

Now by first integrating by substitution, then differentiating and finally Taylor expanding and collecting terms,

$$\mathscr{U}_h'(x) = h^2(\tfrac{1}{2}d_K)\big[(mf)^{(4)} - (mf'')''\big](x) + o(h^2).$$



Hence, by another substitution and Taylor expansion,

$$\mathscr{B}_{n2} = g^2h^2\left(\tfrac{1}{2}d_K\right)^2\left[(mf)^{(4)} - (mf'')''\right](x) + o(g^2h^2).$$

Thus, along almost all sample sequences,

$$\mathscr{B}_n^2 = C_2 g^4 h^4 + o(g^4 h^4) \tag{4.6}$$

for C_2 as defined in the statement of Theorem 3.

Calculations in a similar spirit show that

$$\mathscr{V}_n = n^{-1}h^4 g^{-5} C_1 + o(n^{-1}h^4 g^{-5}),$$

where C_1 is defined in the statement of Theorem 3. This, together with (4.5) and (4.6) completes the proof of Theorem 3. □

REFERENCES

BICKEL, P. J. and FREEDMAN, D. A. (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.* **9** 1196–1217.

BIERENS, H. J. and POTT-BUTER, H. A. (1987). Specification of household expenditure functions and equivalence scales by nonparametric regression. Technical Report 1987–44, Free Univ., Amsterdam.

BILLINGSLEY, P. (1986). *Convergence of Probability Measures*. Wiley, New York.

COLLOMB, G. (1981). Estimation nonparamétrique de la régression: Revue bibliographique. *Internat. Statist. Rev.* **49** 75–93.

COX, D. R. (1970). *Analysis of Binary Data*. Chapman and Hall, New York.

DIKTA, G. (1988). Approximation of nearest neighbor regression function estimators. Technical Report, Univ. Giessen.

EUBANK, R. (1988). *Spline Smoothing and Nonparametric Regression*. Dekker, New York.

Family Expenditure Survey, Annual Base Tapes (1968–1983) Department of Employment, Statistics Division, Her Majesty's Stationary Office, London 1968–1983. The data utilized in this book were made available by the ESRC Data Archive at the University of Essex.

FELLER, W. (1971). *An Introduction to Probability Theory and Its Applications* **2**, 2nd ed. Wiley, New York.

GASSER, T. and MÜLLER, H. G. (1984). Estimating regression functions and their derivatives by the kernel method. *Scand. J. Statist.* **11** 171–185.

HALL, P. and TITTERINGTON, M. (1988). On confidence bands in nonparametric density estimation and regression. *J. Multivariate Anal.* **27** 228–254.

HÄRDLE, W. (1989). *Applied Nonparametric Regression*. Econometric Society Monograph Series. Cambridge Univ. Press.

HÄRDLE, W. and BOWMAN, A. (1988). Bootstrapping in nonparametric regression: Local adaptive smoothing and confidence bands. *J. Amer. Statist. Assoc.* **83** 102–110.

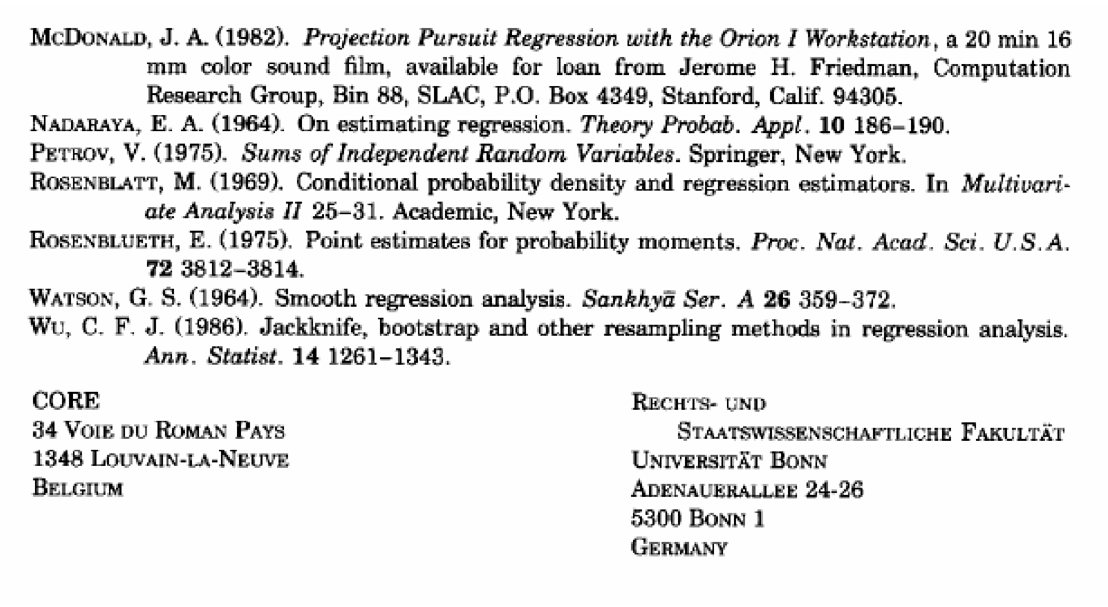
HÄRDLE, W. and MAMMEN, E. (1989). Comparing nonparametric versus parametric regression fits. Discussion paper A–177, Univ. Bonn.

HÄRDLE, W. and MARRON, J. S. (1985). Optimal bandwidth selection in nonparametric regression function estimation. *Ann. Statist.* **13** 1465–1481.

HÄRDLE, W., HALL, P. and MARRON, J. S. (1988). How far are automatically chosen regression smoothing parameters from their optimum? (with discussion). *J. Amer. Statist. Assoc.* **83** 86–101.

HILDENBRAND, K. and HILDENBRAND, W. (1986). On the mean income effect: A data analysis of the U.K. family expenditure survey. In *Contributions to Mathematical Economics* (W. Hildenbrand and A. Mas-Colell, eds.) 247–268. North-Holland, Amsterdam.

KNAFL, G., SACKS, J. and YLVISAKER, D. (1985). Confidence bands for regression functions. *J. Amer. Statist. Assoc.* **80** 683–691.



McDonald, J. A. (1982). *Projection Pursuit Regression with the Orion I Workstation*, a 20 min 16 mm color sound film, available for loan from Jerome H. Friedman, Computation Research Group, Bin 88, SLAC, P.O. Box 4349, Stanford, Calif. 94305.

Nadaraya, E. A. (1964). On estimating regression. *Theory Probab. Appl.* **10** 186–190.

Petrov, V. (1975). *Sums of Independent Random Variables*. Springer, New York.

Rosenblatt, M. (1969). Conditional probability density and regression estimators. In *Multivariate Analysis II* 25–31. Academic, New York.

Rosenblueth, E. (1975). Point estimates for probability moments. *Proc. Nat. Acad. Sci. U.S.A.* **72** 3812–3814.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā Ser. A* **26** 359–372.

Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *Ann. Statist.* **14** 1261–1343.

CORE
34 Voie du Roman Pays
1348 Louvain-la-Neuve
Belgium

Rechts- und
Staatswissenschaftliche Fakultät
Universität Bonn
Adenauerallee 24-26
5300 Bonn 1
Germany

Econometrica, Vol. 59, No. 6 (November, 1991), 1525–1549

EMPIRICAL EVIDENCE ON THE LAW OF DEMAND

BY WOLFGANG HÄRDLE, WERNER HILDENBRAND,
AND MICHAEL JERISON[1]

A sufficient condition for market demand to satisfy the Law of Demand is that the mean of all households' income effect matrices be positive definite. We show how this mean income effect matrix can be estimated from cross section data under metonymy, an assumption about the distribution of households' characteristics. The estimation procedure uses the nonparametric method of average derivatives. Income effect matrices estimated this way from U.K. family expenditure data are in fact positive definite. This result can be explained by a special form of heteroskedasticity in the data: households' demands are more dispersed at higher income levels.

KEYWORDS: Law of demand, income effect, average derivatives, nonparametric estimation, metonymy.

1. INTRODUCTION

WHEN GENERAL EQUILIBRIUM MODELS are used to make comparative static predictions they cease to be general. This is necessarily so. Without a specific structure of the demand and supply system one cannot expect any definite comparative static results. However, in most analyses, conclusions depend upon structure imposed either by aggregating consumers into a single representative, or by assuming restrictive forms for utility or production functions. Such analyses therefore deal with special cases. The present paper considers an alternative way of imposing structure on a general equilibrium model. It considers sufficient conditions for the multimarket version of the "Law of Demand" in a consumption sector; cf. Hicks (1956). The sufficient conditions are a hybrid, combining standard theoretical restrictions with restrictions that do not come from a theoretical model. The latter restrictions can, under certain conditions, be tested and we provide such a test using U.K. family expenditure data.

The Law of Demand concerns effects of price changes when households' budgets (total expenditures) are fixed. It is a condition referring to a counterfactual, asking how mean demand would differ if prices were different. As such it cannot generally be tested using time series data. If the observation period were long enough to reveal significant price variation, it would probably also show changes in households' budgets, preferences, and demographic characteristics. Our analysis describes a way of relating the Law of Demand to cross section data.

[1] Supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 303 der Universität Bonn. We thank Kurt Hildenbrand, Rosa Matzkin, and Thomas Stoker for stimulating discussion, Sigbert Klinke and Berwin Turlach for computing assistance. We also thank Robert Porter and three referees for comments on earlier versions of the paper. We especially thank Whitney Newey for formulating a statistical test of the metonymy assumption.

The Law of Demand is essentially equivalent to negative definiteness of the Jacobian matrix of price derivatives of mean demand. Note that this is much stronger than the requirement that demand for a good be downward sloping with respect to its own price. The Jacobian matrix can be decomposed into a mean of individual Slutsky substitution matrices and a mean of income effect matrices. Standard theory implies that the former matrix is negative semidefinite, but says nothing about the latter. A sufficient condition for the Law of Demand is positive definiteness of the mean income effect matrix. However, for a single consumer, the income effect matrix cannot be positive definite. It can be positive semidefinite, but only in the restrictive case of homothetic preferences. Hildenbrand (1983) and Hildenbrand and Hildenbrand (1986) have shown that when households have identical demand functions, dispersion in the income distribution contributes to the positive definiteness of the mean income effect matrix. In this paper we show that dispersion in tastes can also help. In particular, if the Engel curves of different consumers spread out at higher income levels, the income effect matrix is likely to be positive definite. This type of spreading of demands, a special form of heteroskedasticity is well supported by the expenditure data examined below. Our cross section estimate of the mean income effect matrix is indeed positive definite.

Our estimation procedure is nonparametric. Such nonparametric estimates are ordinarily less efficient than parametric ones when the parametric forms are known. However, the functional forms of the households' demands are in fact not known and cannot be accurately estimated from our data given that they differ across households. The potential advantage of parametric estimation is likely to turn out to be a disadvantage if the hypothesized parametric family is misspecified. More important, even this *potential* advantage is illusory in our framework. We estimate a matrix of derivatives averaged over households, and for these average derivatives, nonparametric estimates achieve the same rate of convergence as parametric ones; c.f. Stoker (1986) and Härdle and Stoker (1989).

There is another subtler reason for avoiding assumptions about functional form. Suppose we assume that households of a particular type have identical demand functions with a form commonly used in empirical analysis. The Engel curves for such demands are quite smooth, i.e. do not wiggle much. It can be shown that if the distribution of the households' budgets is sufficiently dispersed, then the mean income effect matrix is positive semidefinite; c.f. Chiappori (1985) and Grodal and Hildenbrand (1989). The sufficient degree of budget dispersion depends on the form of the Engel curves but for most commonly used forms it is not large, and the dispersion in our data is larger. Thus by assuming one of the standard functional forms for household Engel curves one effectively obtains the Law of Demand by assumption (with no further restrictions on households' demands). Among the standard forms we have in mind are polynomials of degree less than 5 or the forms proposed by Leser (1963). The nonparametric approach permits us to relax an assumption that is clearly too strong since it implies the conclusion we are investigating.

The paper proceeds as follows. In Section 2 we present a model of a large consumption sector. We define the Law of Demand and the mean income effect matrix and show how a closely related matrix can be estimated using cross section data. In Section 3 we discuss the latter matrix, estimated using the method of average derivatives. The estimation procedure is described in the Appendix.

2. THE LAW OF DEMAND AND THE MEAN INCOME EFFECT MATRIX

2.1. *A Sufficient Condition for the Law of Demand*

We consider a group (population) of households. Each household spends its exogenously given budget (total expenditure), $b \geqslant 0$, on the demand for l consumption goods. The consumption behavior of a household is described by an *individual demand function* $f: (p, b) \mapsto f(p, b) \in \mathbb{R}^l_+$, where $p \in \mathbb{R}^l_{++}$ denotes the vector of prices of the l commodities. By definition we have $p \cdot f(p, b) = b$ for all price vectors p. In empirical literature, demand is commonly treated as a function of current budget and prices although household decisions during the period of observation depend on expectations about conditions after the period. The above formulation is appropriate if the household has preferences for goods during the period that are separable from later consumption, or alternatively if the household faces a binding constraint on borrowing and the budget is fixed in advance. More generally, the household could face a longer term budget constraint, and price changes could affect the total expenditure b during the observation period. The Law of Demand concerns the effect of price changes with b held fixed, and this effect can be induced by changing prices *and the long run budget* by the proper amount. Then long run optimization does not imply the usual Slutsky conditions for the short run demand function f, but as noted below, we will not need to assume that all households' demands satisfy the Slutsky conditions.

Typically, different households may have different individual demand functions f and different budgets b. The class of all admissible individual demand functions f is denoted by $\mathcal{F}$. For example, $\mathcal{F}$ might be the class of demand functions which are generated by all (or a certain subset of) strictly convex and continuous (or smooth) preference relations on $\mathbb{R}^l_+$ or, more generally, the class of all demand functions which satisfy the Weak Axiom of Revealed Preference. It will be convenient in the following to label the demand functions in $\mathcal{F}$ by an index α (we then write $f^\alpha(p, b)$) with $f^\alpha(\cdot, \cdot) \neq f^{\alpha'}(\cdot, \cdot)$ if $\alpha \neq \alpha'$. The index set $\mathcal{A}$ may be a finite set, any subset of Euclidian space or, more generally, any metric space. We shall assume that $f^\alpha(p, b)$ depends continuously on the index α. (This representation of $\mathcal{F}$ entails no loss of generality since we can always choose $\mathcal{F}$ itself as an index set.)

With this notation every household i is described by a pair $(b_i, \alpha_i) \in \mathbb{R}_+ \times \mathcal{A}$, that is to say, by its budget b_i and its demand function f^{α_i}. A population of households is described by a *joint distribution* of budgets b and individual demand functions f. Let μ be any probability measure on the space of consumption characteristics $\mathbb{R}_+ \times \mathscr{A}$. The *mean demand* F of a consumption sector described by the distribution μ is then defined by

$$p \mapsto F(p) = \int_{\mathbb{R}_+ \times \mathscr{A}} f^\alpha(p,b)\, d\mu \in \mathbb{R}^l_+.$$

We say that the *Law of Demand* holds in the consumption sector μ if the mean demand function F is *monotone*, i.e.,

$$(p-q)\cdot(F(p)-F(q)) < 0$$

for every $p, q \in \mathbb{R}^l_{++}$ with $p \neq q$. This says that for any two different price vectors p and q, the vector $(p-q)$ of price changes and the vector $(F(p)-F(q))$ of corresponding demand changes point in opposite directions. Thus, in particular, every partial demand curve is downward sloping. There is no need here to emphasize the importance and the implications of the Law of Demand (see, for example, Hicks (1956, p. 59)).

The Law of Demand holds trivially if all individual demand functions f are monotone in p for every given budget b. The standard example for this case is the set of demand functions which are derived from homothetic preferences. For a general characterization of utilities or preferences which lead to monotone demand functions we refer to Mitjuschin and Polterovich (1978) or Kannai (1989). Another case where one obtains the Law of Demand quite easily is given by a consumption sector with a decreasing density of budgets and a common demand function which satisfies the Weak Axiom of Revealed Preference (Hildenbrand (1983)). These cases, however, are examples; they cannot be considered satisfactory foundations for the Law of Demand.

In this paper we shall proceed as follows; in a first step we derive, under suitable assumptions on the individual demand functions, a *sufficient condition* for the monotonicity of the mean demand function F. There is no reason to suppose that this sufficient condition is implied by any reasonable restriction on the individual consumption characteristics and/or assumptions on the distribution μ. Then, in a second step, we develop for this sufficient condition, under suitable assumptions on the distribution μ, an empirical test based on cross-section data.

We assume from now on that the individual demand functions in $\mathscr{F}$ are continuously differentiable in prices and budget. It is well-known that the differentiable mean demand function F is monotone if the Jacobian matrix

$$\partial F(p) = \left(\partial_{p_j} F_k(p)\right)_{j,k=1,\dots,l}$$

is negative definite for every $p \in \mathbb{R}^l_{++}$. Define the Slutsky (substitution) matrix of the demand function $f^\alpha(p,b)$ by

$$S(p,b,\alpha) = \partial_p f^\alpha(p,b) + \partial_b f^\alpha(p,b) f^\alpha(p,b)^T$$

where $f^\alpha(p,b)$ and $\partial_b f^\alpha(p,b)$ are column vectors and the superscript T

denotes the transpose. For the Jacobian matrix of the mean demand function F we then obtain

$$\partial F(p) = \bar{S}(p) - \bar{M}(p),$$

where

$$\bar{S}(p) = \int_{\mathbb{R}_+ \times \mathscr{A}} S(p,b,\alpha)\, d\mu \qquad \text{(mean Slutsky matrix)}$$

and

$$\bar{M}(p) = \int_{\mathbb{R}_+ \times \mathscr{A}} \partial_b f^\alpha(p,b) f^\alpha(p,b)^T\, d\mu$$

(mean income effect matrix).

Consequently, a sufficient condition for the monotonicity of the mean demand function F is that the mean Slutsky matrix $\bar{S}$ is negative semidefinite and the mean income effect matrix $\bar{M}$ is positive definite. If one is willing to accept the hypothesis that individual demand functions $f(p,b)$ are either derived from preference maximization or, more generally, satisfy the Weak Axiom of Revealed Preference, then it is well-known that every individual Slutsky matrix $S(p,b,\alpha)$, and hence the mean Slutsky matrix $\bar{S}(p)$, is negative semidefinite.

Of course such hypotheses are made throughout the theoretical and empirical literature. As noted above, they could be problematic when the consumers' time horizon is longer then the observation period. There is little empirical evidence concerning whether individual demands satisfy the revealed preference axioms. Battalio, et. al. (1973) describe individual consumer expenditure data in which violations of the Strong Axiom are fairly common but are small in a well-defined sense. Even if some consumers violate the Weak Axiom slightly, their effect on the Slutsky matrix $\bar{S}$ can be counterbalanced by other consumers who satisfy the axiom.

In conclusion, assuming that the mean Slutsky matrix $\bar{S}(p)$ is negative semidefinite, a sufficient condition for monotonicity of F is that the mean income effect matrix $\bar{M}(p)$ is positive definite. This property does not follow from an assumption on "rational" individual behavior. Our goal is to develop a better understanding of the class of consumption sectors μ that lead to a positive definite mean income effect matrix $\bar{M}(p)$. For the remainder of the paper we fix the price vector p and omit it as an argument.

2.2. *The Mean Income Effect Matrix for Metonymic Consumption Sectors*

The mean income effect matrix $\bar{M}$ cannot be estimated directly. In this section we describe a closely related matrix A, that can be estimated from cross section data. Note that the matrix $\bar{M}$ is positive definite if and only if the symmetrized matrix

$$M = \bar{M} + \bar{M}^T$$

has this property. The matrix M is given by

$$M = \left(\int_{\mathbb{R}_+ \times \mathscr{A}} \partial_b \left(f_j^\alpha(b) \cdot f_k^\alpha(b) \right) d\mu \right)_{j,k=1,\ldots,l}.$$

To simplify notation, let $g_{jk}(b,\alpha) = f_j^\alpha(b) \cdot f_k^\alpha(b)$. We call the matrix $G(b,\alpha) = (g_{jk}(b,\alpha))$ the product matrix of the demand function f^α at expenditure level b. Thus, in matrix notation,

$$M = \int_{\mathbb{R}_+ \times \mathscr{A}} \partial_b G(b,\alpha)\, d\mu.$$

In order to define a matrix A which will be shown to be related to the matrix M and which can be estimated from cross section data we need the following properties of the distribution μ on $\mathbb{R}_+ \times \mathscr{A}$.

(i) The marginal distribution of budgets is absolutely continuous, i.e., there exists a density for the budget distribution, which we denote by ρ. In addition we shall assume that the density ρ is smooth.

(ii) Let $\mu|b$ denote the conditional distribution of α given the budget level b and consider the functions

$$\bar{f}_j(b) = \int_{\mathscr{A}} f_j^\alpha(b)\, d\mu|b \qquad (j = 1,\ldots,l)$$

and

$$\bar{g}_{jk}(b) = \int_{\mathscr{A}} f_j^\alpha(b) \cdot f_k^\alpha(b)\, d\mu|b \qquad (j,k = 1,\ldots,l).$$

We shall assume that the statistical Engel curve $\bar{f}_j(\cdot)$ and the conditional mean product function $\bar{g}_{jk}$ are continuously differentiable.

Let $\overline{G}(b)$ be the matrix with components $\bar{g}_{jk}$ and define the matrix A by

$$A = \int_{\mathbb{R}_+} (\partial_b \overline{G}(b)) \rho(b)\, db.$$

This matrix can be estimated from cross section data since the element a_{jk} of A is the *average derivative* of the regression function $b \mapsto \int_{\mathscr{A}} g_{jk}(b,\alpha)\, d\mu|b$. For details we refer to the Appendix.

The matrices M and A are closely related. Indeed, since

$$M = \int_{\mathbb{R}_+} \left[\int_{\mathscr{A}} \partial_b G(b,\alpha)\, d\mu|b \right] \rho(b)\, db,$$

they are in fact identical, if for every b,

$$(*) \qquad \int_{\mathscr{A}} \partial_b G(b,\alpha)\, d\mu|b = \partial_b \int_{\mathscr{A}} G(b,\alpha)\, d\mu|b,$$

i.e., the $\mu|b$ conditional mean of the derivatives of $f_j(b,\alpha) \cdot f_k(b,\alpha)$ is equal to

Econometrica, 59, 1525-1550

the derivative of the conditional mean $\int_{\mathscr{A}} f_j(b,\alpha) f_k(b,\alpha)\, d\mu|b$. Thus, in particular, if the conditional distribution $\mu|b$ of individual demand functions does not depend on the budget level b (i.e. μ is a product measure), then $M = A$.

The case in which $M = A$ is particularly interesting since it permits the estimation of the symmetric mean income effect matrix M from cross section data. This motivates the following definition.

Definition: A distribution μ of households' characteristics (b, α) with properties (i) and (ii) is called *metonymic* if $M = A$, which is impled by $(*)$.

To obtain a better understanding of the metonymy assumption we shall now clarify the general relationship between the two matrices M and A. For this it is helpful to imagine a Gedanken experiment in which the initially given household budgets are perturbed. Households with initial budget b will be called b-households. The derivative $\partial_b G(b,\alpha)$ in the expression for M is determined by comparing the product matrix of b-households to their product matrix when their budgets change. The derivative $\partial_b \bar{G}(b)$ in the definition of A is determined by comparing the mean product matrix for a different set of households. Define

$$\tilde{G}(b,\beta) = \int_{\mathscr{A}} G(\beta,\alpha)\, d\mu|b,$$

the mean product matrix that b-households would have if their budgets were changed to β. Then we obtain

$$M = A - U$$

where

$$U = \int \left[\partial_1 \tilde{G}(b,b)\right] \rho(b)\, db.$$

($\partial_1 \tilde{G}$ denotes the partial derivative of $\tilde{G}$ with respect to the first argument.) Metonymy requires that the matrix U vanish. The left-hand side of $(*)$ is $\partial_2 \tilde{G}(b,b)$ and the right-hand side is $\partial_1 \tilde{G}(b,b) + \partial_2 \tilde{G}(b,b)$. Thus the equality $(*)$, which is equivalent to $\partial_1 \tilde{G}(b,b) = 0$, implies that $U = 0$. Note that for a product measure μ the mapping $\tilde{G}(b,\beta)$ is constant in its first argument, hence the matrix U vanishes. The property $(*)$ is weaker since it only requires that the partial derivative of $\tilde{G}$ with respect to the first argument is zero on the diagonal $b = \beta$. Metonymy is weaker still, requiring only that the integral U be zero.

Roughly speaking, under the condition $(*)$ the distribution of demands by β-households can be used to represent what the corresponding distribution for b-households would look like if their budgets changed to β, for β near b. We will make this more precise. Define

$$\tilde{f}(b,\beta) = \int_{\mathscr{A}} f^{\alpha}(\beta)\, d\mu|b$$

and let v be the corresponding unit length eigenvector. Consider the composite commodity formed by weighting the commodities by the components of v. Mean demand for this composite commodity at the price vector p is $v \cdot F(p)$. When prices change in the direction v, the directional derivative of demand for the composite derivative is

$$\begin{aligned} v \cdot \partial F(p) v &= v \cdot \bar{S} v - v \cdot \overline{M} v \\ &\leqslant -v \cdot \overline{M} v = -\tfrac{1}{2} v \cdot M v, \end{aligned}$$

and this last term under metonymy is $-\lambda/2$. For a discrete price change, say from q to $p = q + tv$, the effect on demand is $F(p) - F(q) \approx t\partial F(q)v$, so the effect on demand for the composite commodity is

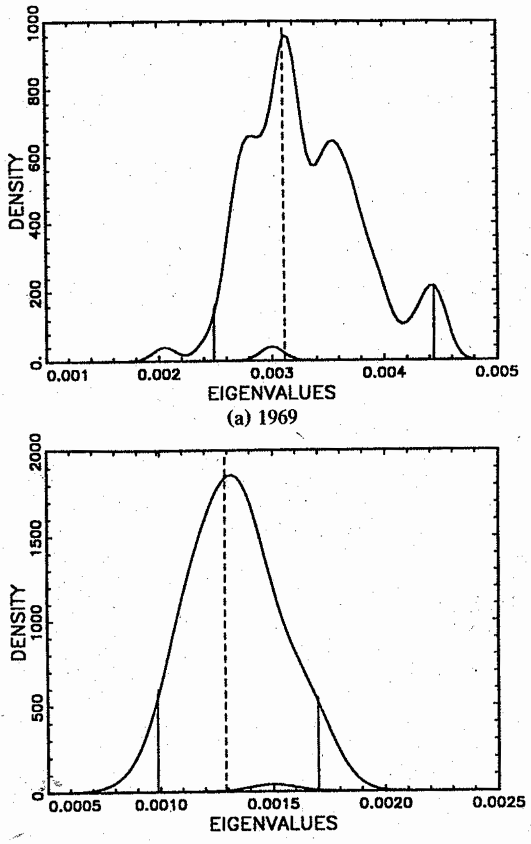
$$(p-q)(F(p)-F(q)) \approx tv \cdot \partial F(q) v = -t\frac{\lambda}{2}.$$

Table I shows that in each year the maximal eigenvalue λ is near 0.2. This implies that if prices change from q to p in the direction of the eigenvector corresponding to λ, then the term $(p-q)(F(p)-F(q))$ is bounded above by $-(.1)|p-q|$.

3.3. *Sensitivity of Estimates*

Computation of the estimate of A involves estimating ρ, the density of households' budgets, using a kernel estimator. The smoothness of this estimator is controlled by a "bandwidth" parameter. A second parameter is used to delete observations at which the estimate of ρ is very small. (See Härdle and Stoker (1989) for discussion of these parameters.)

The estimated components and eigenvalues of A are not very sensitive to the choice of bandwidth and cut-off parameters. Variations in these parameters never overturn the positive definiteness of the estimated $\hat{A}$. Concerning sampling variation, there is to our knowledge no theory of the distribution of eigenvalues of a matrix with correlated random components. However, one gets an idea of the distribution of the estimated minimum eigenvalue of $\hat{A}$ by considering the sample distribution of minimum eigenvalues computed from bootstrap estimates of $\hat{A}$. One selects randomly (with replacement) n observations from the original sample, and estimates A using the constructed bootstrap sample. Figure 1a, b shows smoothed kernel density functions for the smallest eigenvalues of the matrices estimated in this way from 100 bootstraps of the 1969 and 1983 samples. All the eigenvalues computed from the bootstrap samples were strictly positive. The Appendix contains an argument relating the bootstrap distributions to the sampling distribution of minimum eigenvalues. An elaborated theory can be found in Härdle and Hart (1989).



(a) 1969

(b) 1983

FIGURE 1.—Estimated smallest eigenvalue kernel density functions from bootstrapping.

3.4. *Subpopulations*

The metonymy condition is more plausible the more "homogeneous" the population. For this reason we tested the positive definiteness of the matrix A for subgroups of the population, considering stratifications by age and occupation of the household head, and household composition. Table IV lists the smallest eigenvalues of the estimates of $A \times 100$ for each age group. Nearly all of the estimated matrices are positive definite and most of the others belong to the age group 80–89 with the smallest sample size.

The sum of the A matrices for the subgroups, weighted by the sample size provides an alternative estimate for M, and the minimum eigenvalue of this estimate is bounded below by the sum of the eigenvalues for the subgroups, weighted by sample size. These weighted sums are strictly positive for all years.

and

$$\tilde{C}(b,\beta)=\tilde{G}(b,\beta)-\tilde{f}(b,\beta)\tilde{f}(b,\beta)^T,$$

respectively, the mean demand and the covariance matrix of the demands by b-households whose budgets are changed to β. By the budget identity we have $G(\beta,\alpha)p=\beta f^\alpha(\beta)$, so $(*)$ implies

$$0=\partial_1\tilde{G}(b,\beta)p=\partial_b\int_{\mathscr{A}}\beta f^\alpha(\beta)\,d\mu|b$$

where the derivatives are evaluated at $b=\beta$. Thus $(*)$ implies

$(*.1)\qquad \partial_1\tilde{f}(b,b)=0,$

and by definition of $\tilde{C}$,

$(*.2)\qquad \partial_1\tilde{C}(b,b)=0.$

These conditions say that the mean demand and the covariance matrix of demands by $(b+\Delta b)$-households are essentially equal respectively to what the mean demand and covariance for the b-households would be if their budgets expanded by Δb. Conditions $(*.1)$ and $(*.2)$ together imply $(*)$ and hence are equivalent to $(*)$. Thus a distribution μ satisfying $(*)$ looks locally like a product measure at least in so far as its first and second conditional moments are concerned. In fact, if the individual demand functions are homogeneous of degree zero then $\tilde{f}(b,\beta)$ is independent of b.

In summary: Let the individual demand functions in $\mathscr{F}$ be *continuously differentiable* and satisfy the *Weak Axiom of Revealed Preference*. If μ is a *metonymic distribution* on $\mathbb{R}_+\times\mathscr{A}$, then a sufficient condition for the mean demand

$$F(p)=\int_{\mathbb{R}_+\times\mathscr{A}}f^\alpha(p,b)\,d\mu$$

to be monotone is that the matrix A be positive definite.

Given the importance of the metonymy assumption it is worthwhile considering an example in which it is violated. Let the consumption sector have a finite number of household types. All households of the same type α are assumed to have the same demand function f^α. The types of households might be identified by demographic characteristics such as the number of household members, their ages, etc. Among the households with budget b, the fraction that are of type f^α will be denoted by $\nu_\alpha(b)$. If μ is a product measure, then the functions $\nu_\alpha(\cdot)$ are constant. On the other hand for certain demographic characteristics these functions cannot be assumed constant. In our example we obtain for the matrix U:

$$U=\sum_\alpha\int\left(f^\alpha(b)f^\alpha(b)^T\right)\nu'_\alpha(b)\rho(b)\,db.$$

The matrix U may be positive or negative definite or indefinite. The example shows that it might well happen that metonymy is not satisfied for the whole population but that after appropriate stratification the subpopulations satisfy it.

The violation of metonymy poses no problem in the above example. If the household types can be identified, then the mean income effect matrix for the entire population can be calculated from the corresponding matrices of the various household types. More generally, we can consider the case in which the population is partitioned into subgroups that each satisfy metonymy. The mean income effect matrix is then a weighted average of the average derivative A matrices of the subgroups. To be more precise, let v_i be the fraction of the population in subgroup i and let μ_i be the (conditional) distribution of household characteristics within that subgroup. The average derivative matrix for subgroup i is

$$A_i = \int_{\mathbb{R}_+} (\partial_b \overline{G}_i(b)) \rho_i(b)\, db$$

where $\overline{G}_i(b)$ has jk component

$$\int_{\mathscr{A}} f_j^{\alpha}(b) \cdot f_k^{\alpha}(b)\, d\mu_i|b$$

and where $\rho_i(b) = \int_{\mathscr{A}} d\mu_i|b$. Metonymy for subgroup i implies that the matrix A_i equals the subgroup's symmetrized mean income effect matrix

$$M_i = \int_{\mathbb{R}_+ \times \mathscr{A}} \partial_b G(b, \alpha)\, d\mu_i.$$

Since $\mu = \Sigma_i v_i \mu_i$, the symmetrized mean income effect matrix for the entire population is $M = \Sigma v_i M_i = \Sigma v_i A_i$. So the matrix M can be estimated by estimating the average derivative matrices A_i for all the subgroups. In this case, metonymy for the entire population can be tested by comparing A to $\Sigma v_i A_i$. If they are not equal, the population or some subgroup must violate metonymy. A statistical test based on estimates of A and A_i is described and carried out in the Appendix. Whitney Newey has pointed out that average derivatives can be computed conditioning on any covariates of the households' demands. The tests based on stratification are simply special cases of such conditioning.

We conclude this section with a brief discussion of the matrix A. In order to isolate the factors that contribute to its positive definiteness, it is useful to compare A to the income effect matrix estimated by Hildenbrand and Hildenbrand (1986). In a consumption sector described by the distribution μ on $\mathbb{R}_+ \times \mathscr{A}$, the *statistical Engel curve* is defined by the function

$$b \mapsto \int_{\mathscr{A}} f^{\alpha}(p, b)\, d\mu|b = \bar{f}(p, b).$$

Hildenbrand and Hildenbrand (1986) estimate the symmetrized mean income effect matrix of $\bar{f}$, i.e., the matrix

$$B = \int \partial_b\left(\bar{f}(p,b)\bar{f}(p,b)^T\right)\rho(b)\,db.$$

This matrix turns out to be "approximately" positive definite. More precisely, the matrix B is typically ill-conditioned; some eigenvalues are very small in magnitude (positive or negative), however the larger eigenvalues are always positive. It is easy to imagine consumption sectors for which the matrix B is singular. For example, if ρ is the uniform distribution on the interval $[0, \beta]$, then $B = \bar{f}(\beta)\bar{f}(\beta)^T$, which is a positive semidefinite matrix of rank one. Under appropriate assumptions on the form of the statistical Engel curves one can show, as mentioned above, that the matrix B is always positive semidefinite provided the variance of the budget distribution is sufficiently large (for details see Chiappori (1985) and Grodal and Hildenbrand (1989)).

The matrix B differs from the above matrix A by the average derivative of a conditional covariance matrix. To see this, we note that the jk component of the conditional covariance matrix $C(b)$ of the demands of b-households is

$$\operatorname{cov}_{\mu|b}\left(f_j^{\alpha}(b), f_k^{\alpha}(b)\right) = \int_{\mathscr{A}} f_j^{\alpha}(b) f_k^{\alpha}(b)\,d\mu|b - \bar{f}_j(b)\bar{f}_k(b).$$

Hence we obtain

$$A = B + V$$

where

$$V = \int \partial_b C(b)\rho(b)\,db$$

is the average derivative of the conditional covariance matrix $C(b)$. Note that $C(b)p = 0$ and hence, $Vp = 0$, so V is singular.

The jth diagonal component of $C(b)$ is the variance of the demands for good j by b-households. The magnitude of the jth diagonal component of V measures the heteroskedasticity of the households' demands for good j since it is an average derivative with respect to b of the conditional variances of demands for good j. In a typical cross-section, demand for each good is heteroskedastic (variance increases with total expenditure b), so the diagonal components of V are strictly positive.

Positive semidefiniteness of the matrix V means roughly that on average the dispersion in consumer demands rises with the size of the budget b. A closely related type of increasing dispersion was shown by Jerison (1982) to be the weakest Engel curve restriction ensuring that mean demand satisfies the Weak Axiom (see also Freixas and Mas-Colell (1987)). Increasing dispersion has a simple geometric representation. Given a budget b, the dispersion of the b-households' demands for, say, the first m goods is measured by the principal minor matrix $\hat{C}(b)$ formed from $C(b)$ by deleting its last $l - m$ rows and columns. When $\hat{C}(b)$ is nonsingular, this demand dispersion can be represented geometrically. There is a unique ellipsoid (called the *ellipsoid of concentration*) centered at the origin in $\mathbb{R}^m$ such that a uniform distribution over the ellipsoid has the variance-covariance matrix $\hat{C}(b)$. The ellipsoid consists of the set of x satisfying

$$x \cdot \hat{C}(b)^{-1} x = m + 2;$$

cf. Cramér (1946, Ch. 22). The ellipsoid gives a simple description of the form of the dispersion of the b-households' demands for the m goods. Larger variances correspond to a larger ellipsoid. A strong form of increasing dispersion can be represented by nested ellipsoids, with the ellipsoid at budget b contained in the one at $\beta > b$. The formal requirement for this is that $x \cdot \hat{C}(\beta)^{-1} x \leqslant lm + 2$ for each x with $x \cdot \hat{C}(b)^{-1} x \leqslant lm + 2$. This is equivalent to the positive semidefiniteness of $\hat{C}(b)^{-1} - \hat{C}(\beta)^{-1}$, which is equivalent to positive semidefiniteness of $\hat{C}(\beta) - \hat{C}(b)$, c.f. Dhrymes (1984, Prop. 65, p. 76). This last condition implies that the matrix of derivatives $\partial_b C(b)$ is positive semidefinite, so the corresponding principal minor matrix of V is also positive semidefinite. Note that the matrix $\hat{C}(b)$ cannot be taken to be $C(b)$ in the argument above since the latter matrix is singular with $C(b)p = 0$. However if $C(b)$ has maximal rank $l - 1$ then $\hat{C}(b)$ can be taken to be its leading principal minor matrix of order $l - 1$. This principal minor is positive definite and hence nonsingular. If the ellipsoids of concentration for the first $l - 1$ goods are nested, then as above $\hat{C}(\beta) - \hat{C}(b)$ is positive for $\beta > b$. But this implies that $C(\beta) - C(b)$ is positive semidefinite and hence V also. (To see this, note that any l-vector x can be written as $v + \lambda p$, where λ is a scalar and the last component of v is 0. Then $x \cdot [C(\beta) - C(b)]x = u \cdot [\hat{C}(\beta) - \hat{C}(b)]u \geqslant 0$, where u is obtained from v by removing its last component.) Thus, for V to be positive semidefinite it is sufficient but not necessary that the ellipsoids of concentration for the first $l - 1$ goods be nested, expanding with the budget level. Sections of estimated ellipsoids projected on the plane are illustrated in Figure 4 below.

3. EMPIRICAL EVIDENCE

In this section we present estimates of the matrix A for various populations, along with other empirical evidence that will help in interpreting the results.

3.1. *The Variables and Data*

We consider expenditures on nine commodity aggregates:

1. Housing (HOU)
2. Fuel, light and power (FUE)
3. Food (FOO)
4. Clothing and footwear (CLO)
5. Durable household goods (DUR)
6. Services (SER)
7. Transport (TRA)
8. Other goods, and miscellaneous (OGM)
9. Alcohol and tobacco (ATO)

by each sampled household in the U.K. Family Expenditure Surveys (FES) from

1969 to 1983. Each year the expenditures of approximately 7000 households are reported. For details concerning the samples and commodity classification, see Family Expenditure Survey (1968–1983), Kemsley, Redpath, and Holmes (1980), and Schmidt (1989). In order to interpret the results, it is convenient to normalize the mean budget and the price indices of all the commodity aggregates to equal 1. This is legitimate since the estimation of a given A matrix involves observations from a single period. The demand for a good by a particular household is therefore the household's expenditure on the good divided by the mean budget for the whole population.

3.2. *Estimates of A*

The procedure for estimating A by the method of average derivatives is described in the Appendix. The estimate $\hat{A} = (\hat{a}_{jk})$ is symmetric, and is positive definite if all of its eigenvalues are strictly positive. Table I contains the smallest and largest eigenvalues of $\hat{A}$ estimated from the entire FES sample in each of the years 1969–1983. These eigenvalues are all strictly positive, so the matrices are positive definite.

The ratio of the largest to the smallest eigenvalue in Table I is never greater than 200. So the estimated matrices are well conditioned and their positive definiteness cannot be attributed to numerical (rounding) errors. In order to interpret the magnitudes of the eigenvalues in Table I it is helpful to consider the components of $\hat{A}$. Tables IIa and IIb show the components of the 1969 and 1983 $\hat{A}$ matrices multiplied by 100.

The diagonal components of $\hat{A}$ yield estimated bounds on the own price elasticities of demand. To see this, recall that $\partial_p F = \bar{S} - \bar{M}$. Under the assumption that the mean substitution matrix $\bar{S}$ is negative semidefinite, the own price effect $\partial F_j/\partial p_j$ is bounded above by the jth diagonal component of $-\bar{M}$. Under

TABLE I

MINIMAL AND MAXIMAL EIGENVALUES OF $\hat{A}$.

Year	Sample Size	$\lambda_{\min} \times 100$	$\lambda_{\max} \times 100$
1969	7007	0.31	25
1970	6391	0.24	25
1971	7238	0.31	25
1972	7017	0.28	25
1973	7125	0.26	24
1974	6694	0.29	24
1975	7201	0.33	24
1976	7203	0.29	24
1977	7198	0.26	24
1978	7001	0.20	24
1979	6777	0.14	23
1980	6943	0.28	24
1981	7525	0.18	23
1982	7428	0.20	24
1983	6973	0.13	23

TABLE IIA

$\hat{A} \times 100$ FOR 1969.

HOU	FUE	FOO	CLO	DUR	TRA	SER	OGM	ATO
2.91	0.86	3.84	1.75	1.40	2.91	1.65	1.33	1.47
	0.74	2.03	0.92	0.60	1.50	0.80	0.67	0.85
		10.03	4.24	2.73	6.54	3.56	3.13	4.10
			3.53	1.27	2.60	1.58	1.41	1.71
				4.10	1.64	0.96	0.94	1.10
					8.84	2.56	2.11	2.56
						3.74	1.26	1.39
							1.75	1.22
								3.00

TABLE IIB

$\hat{A} \times 100$ FOR 1983.

HOU	FUE	FOO	CLO	DUR	TRA	SER	OGM	ATO
5.12	1.18	4.21	1.76	1.87	4.13	2.63	2.01	1.67
	0.52	1.53	0.67	0.68	1.45	0.94	0.72	0.63
		6.48	2.66	2.40	5.29	3.31	2.77	2.56
			2.34	1.03	2.24	1.47	1.28	1.05
				4.23	2.04	1.29	1.16	0.94
					8.86	3.23	2.36	2.12
						5.62	1.53	1.33
							2.42	1.10
								1.89

metonymy, $A = \overline{M} + \overline{M}^T$, so this diagonal component is $a_{jj}/2$, and the own price elasticity ε_j of demand for good j satisfies

$$\varepsilon_j(p) \equiv \left| \frac{p_j}{F_j(p)} \frac{\partial F_j(p)}{\partial p_j} \right| \geqslant \frac{a_{jj}}{2} \frac{p_j}{F_j(p)}.$$

Since we normalized prices to equal 1 and divided each household's demand by the mean budget, the mean demand $F_j(p)$ equals the budget share for good j for the entire consumption sector. The estimate of $a_{jj}/2F_j(p)$ is an estimated lower bound on the magnitude of the jth own price elasticity, the bound due to income effects. The set of estimated bounds is given in Table III for 1969 and 1983.

The eigenvalues of A yield similar bounds for the effects of price changes on the demand for certain composite commodities. Let λ be an eigenvalue of A

TABLE III

LOWER BOUNDS FOR OWN PRICE ELASTICITIES, 1969 AND 1983.

Year	HOU	FUE	FOO	CLO	DUR	TRA	SER	OGM	ATO
1969	0.12	0.06	0.19	0.20	0.35	0.33	0.22	0.11	0.16
1983	0.15	0.04	0.15	0.17	0.32	0.31	0.27	0.15	0.12

Econometrica, 59, 1525-1550

TABLE IV

MINIMAL EIGENVALUES OF $\hat{A}$ FOR THE STRATA "AGE."

	20–29		30–39		40–49		50–59		60–69		70–79		80–89	
Year	n	λ_{min} ×100	n	λ_{min} ×100	n	λ_{min} ×100	n	λ_{min} ×100	n	λ_{min} ×100	n	λ_{min} ×100	n	λ_{min} ×100
1969	825	0.30	1275	0.25	1380	0.34	1292	0.29	1310	0.20	706	0.29	198	−0.30
1970	874	−0.49	1106	0.19	1216	0.20	1125	0.21	1192	0.47	659	0.44	190	−0.23
1971	980	0.24	1245	0.26	1336	0.31	1307	0.36	1309	0.30	820	0.13	209	0.85
1972	998	0.11	1244	0.15	1268	0.17	1299	0.42	1239	0.38	750	0.55	186	0.11
1973	1003	0.14	1180	0.30	1167	−0.29	1309	0.47	1354	0.68	844	0.50	229	0.12
1974	912	0.16	1211	0.25	1109	0.28	1179	0.91	1248	0.16	775	0.35	227	−0.76
1975	1034	0.49	1296	0.75	1173	0.37	1217	0.20	1348	0.19	828	−0.20	264	−0.37
1976	1026	0.17	1270	0.16	1140	0.24	1244	0.16	1332	0.36	905	0.83	249	−0.96
1977	991	0.14	1361	0.29	1174	0.19	1216	0.15	1282	0.16	888	0.29	246	0.29
1978	940	−0.13	339	0.78	1103	0.87	1268	0.15	1220	0.32	832	0.57	252	−0.91
1979	957	0.75	1313	0.10	1079	−0.15	1143	0.18	1078	0.13	903	0.11	260	−0.90
1980	912	0.62	1416	0.69	1107	0.74	1170	0.16	1169	0.62	851	0.61	285	−0.16
1981	918	0.13	1594	0.10	1212	0.20	1229	0.27	1290	0.19	973	0.22	271	0.34
1982	987	0.45	1533	0.56	1201	0.85	1225	0.70	1194	0.42	939	0.63	295	−0.19
1983	898	0.78	1451	0.75	1147	0.44	1089	0.14	1170	0.33	927	0.50	254	−0.11

Thus the minimal eigenvalues of the weighted sum of the subpopulation matrices are positive also. The weighted sums of these subpopulation matrices are statistically different from the A matrix estimated from the entire population. However, this difference is not large in magnitude; see the Appendix.

Similar results obtain for the stratifications by occupation in Table V and by household composition in Table VI. The categories for the latter stratification

TABLE V

MINIMAL AND MAXIMAL EIGENVALUES OF $\hat{A}$ FOR THE STRATA "PROFESSION"

	Pensioneer			Worker			Self-employed			Others		
Year	n	λ_{min} ×100	λ_{max} ×100	n	λ_{min} ×100	λ_{max} ×100	n	λ_{min} ×100	λ_{max} ×100	n	λ_{min} ×100	λ_{max} ×100
1969	1200	0.19	26	3193	0.33	25	529	0.13	23	2085	0.38	25
1970	1127	0.49	25	2899	0.16	26	486	0.15	24	1879	0.25	25
1971	1332	0.13	24	3102	0.39	25	580	0.27	24	2224	0.32	26
1972	1282	0.41	25	3065	0.20	26	468	0.14	22	2202	0.26	25
1973	1422	0.33	24	3010	0.26	25	492	0.09	21	2201	0.34	24
1974	1343	0.39	24	2735	0.11	25	561	0.50	23	2055	0.21	24
1975	1521	0.33	25	2901	0.35	25	497	0.11	24	2282	0.44	23
1976	1568	0.70	25	2951	0.22	25	454	−2.12	23	2230	0.32	25
1977	1567	0.34	25	2884	0.24	25	506	0.14	22	2241	0.23	24
1978	1529	0.62	25	2764	0.15	26	434	0.27	23	2274	0.14	24
1979	1565	0.13	24	2567	0.11	23	429	−0.90	23	2216	0.18	24
1980	1584	0.46	26	2571	0.46	26	462	0.12	23	2326	0.16	25
1981	1774	0.16	24	2659	0.15	24	564	0.09	20	2528	0.22	24
1982	1725	0.52	26	2474	0.02	25	491	−0.16	22	2737	0.14	24
1983	1719	0.46	24	1982	0.04	24	509	0.24	22	2763	0.10	24

(1991) Härdle, W., Hildenbrand, W. and Jerison, M.
Empirical Evidence for the Law of Demand

TABLE VI

MINIMAL EIGENVALUES OF $\hat{A}$ FOR THE STRATA "HOUSEHOLD TYPE"

	1M		1F		1A + 1		2A		2A + 1		2A + 2		2A + 3		2A + + 3	
Year	n	$\lambda_{\min} \times 100$	n	$\lambda_{\min} \times 100$	n	$\lambda_{\min} \times 100$	n	$\lambda_{\min} \times 100$	n	$\lambda_{\min} \times 100$	n	$\lambda_{\min} \times 100$	n	$\lambda_{\min} \times 100$	n	$\lambda_{\min} \times 100$
1969	334	0.15	777	0.27	101	0.00	2120	0.22	723	0.31	839	0.12	322	0.30	206	0.27
1970	307	0.13	752	0.12	132	0.02	1909	0.22	621	0.07	787	0.12	339	0.04	168	0.07
1971	365	0.15	863	0.24	157	0.18	2209	0.22	695	0.12	832	0.08	359	0.10	194	0.07
1972	373	0.06	820	0.05	143	−0.01	2118	0.28	735	0.14	831	0.11	362	0.18	189	−0.01
1973	410	0.40	909	0.20	175	0.00	2196	0.14	796	0.20	858	0.15	410	0.23	212	0.14
1974	368	0.19	881	0.03	200	0.02	2075	0.38	664	0.06	872	0.31	392	0.21	203	0.21
1975	400	0.02	1020	0.45	185	0.04	2139	0.29	668	0.19	1025	0.35	373	0.15	204	0.17
1976	476	0.11	985	0.05	240	0.13	2277	0.42	668	0.27	961	0.25	354	0.06	168	0.15

are:

1 male (1 M)
1 female (1 F)
1 adult + 1 child (1A + 1)
2 adults (2A)
2 adults + 1 child (2A + 1)
2 adults + 2 children (2A + 2)
2 adults + 3 children (2A + 3)
2 adults + more than 3 children (2A + + 3)

For all stratifications, the only negative eigenvalues occur in small subpopulations.

3.5. *Further Evidence*

The estimates presented above support the hypothesis that the cross section matrix A is positive definite. Rather than present a theory consistent with such a result we will discuss further evidence that makes the above estimates more understandable. The jk component of A was shown in Section 2 to be the average derivative of the regression function $\bar{g}_{jk}$ that associates with each budget level b the average of the products of demands for goods j and k by households with budget b.

The larger the diagonal components of A the more likely is the matrix positive definite. Kernel estimates of the functions $\bar{g}_{jj}$ for 1969 are shown in Figure 2, where the index j runs over the commodity aggregates food, fuel, and transport. Estimates of $\bar{g}_{jk}$ for cross products of the same commodities $(j \neq k)$ are shown in Figure 3. The household budgets and demands have been normalized, so the unit on the horizontal axis is the mean budget.

All the curves have positive slopes. What is important is that the slopes of the cross product curves are sufficiently small compared with the slopes of the corresponding (own) product curves. For example, consider the curves for food and fuel in Figures 2 and 3. The distribution of household budgets is concentrated on the interval from 0 to twice the mean budget and we can see that the

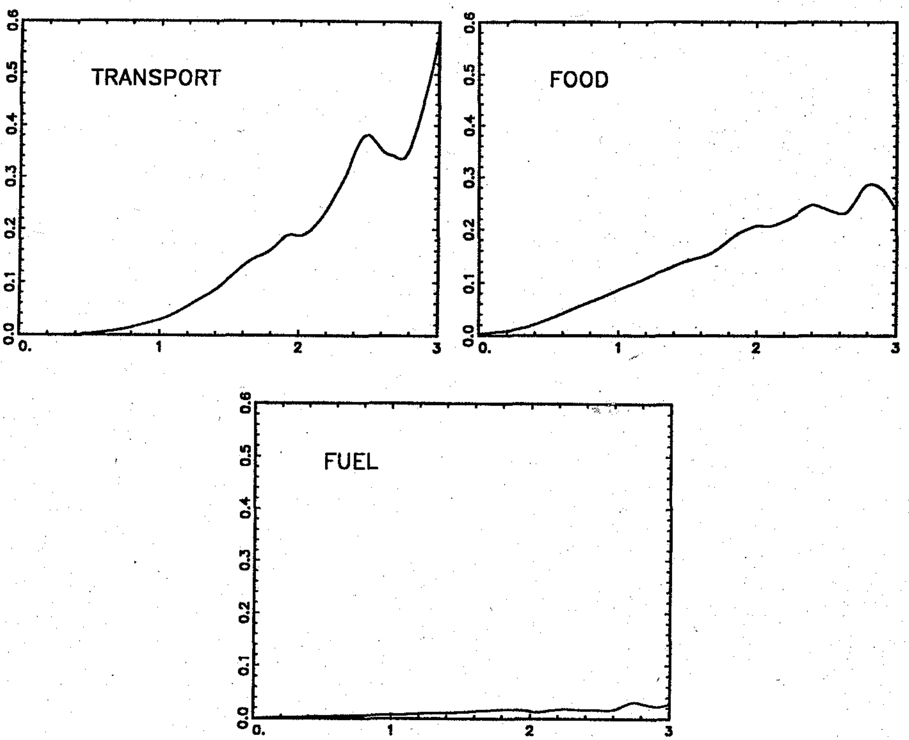
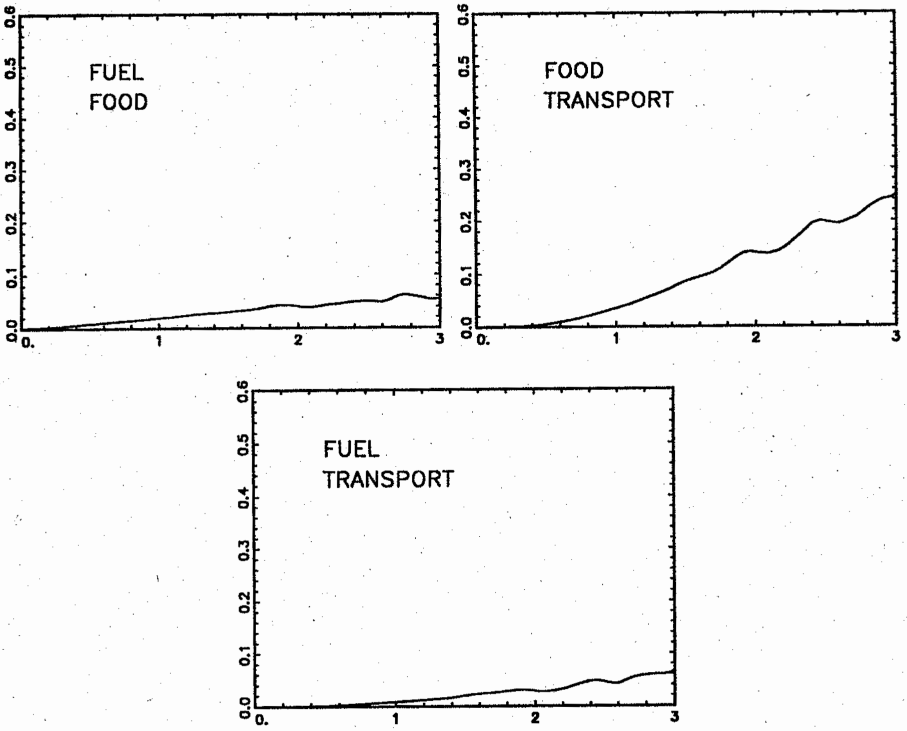


FIGURE 2.—Mean product functions $\bar{g}_{ij}$ for 1969. The unit on the horizontal axis is total expenditure divided by its mean.

slopes of the food, fuel, and food-fuel cross product curves are approximately .1, .01, and .02 respectively. These are essentially the values appearing in the 2 × 2 minor matrix for food and fuel in Table IIA, and this minor matrix is positive semidefinite. The graphs of $\bar{g}$ for other commodity aggregates have shapes and slopes similar to the ones shown here.

As discussed in Section 2, the positive semidefiniteness of $\hat{A}$ can be better understood by comparing it to the matrix of income effects of the cross section (statistical) Engel curve estimated by Hildenbrand and Hildenbrand (1986). The difference between these two matrices is the matrix V, the average derivative of the conditional covariance matrix. The V matrices estimated from the entire sample for the years 1969–83 are all positive semidefinite. By construction $V_p = 0$ so V cannot be positive definite. However all the estimated matrices V are positive definite on the space orthogonal to p. Unlike the product matrices, they are nearly dominant diagonal. The matrix estimates for 1969 and 1983 are shown in Table VIIa, b.


FUEL FOOD

FOOD TRANSPORT

FUEL TRANSPORT

Figure 3.—Mean cross product functions $\bar{g}_{jk}$ for 1969. The unit on the horizontal axis is total expenditure divided by its mean.

The matrices for all the years are quite similar. Since the matrices are symmetric by definition, they have 45 components which can vary independently. All the components remain of the same order of magnitude during the sample period, and only two change sign. The spectrum of eigenvalues is also quite stable over time. For example, the eigenvalues vary by less than 30 percent. The strong positive definiteness of the estimates of V on the orthogonal component of p can be explained along lines suggested in Section 2. Positivity of the diagonal components follows from the heteroskedasticity of the households' demand for each good. This is sufficient to make V nearly dominant diagonal because the conditional correlations of households' demands for pairs of goods are rather small (generally below .2 in magnitude) and do not vary systematically with total expenditure.

Kernel estimates of the conditional covariance matrices $C(b)$ for budget levels of 0.5, 1, 1.5, and 2 times the mean budget have been computed using 1983 data and a bandwidth equal to 0.2 (see Appendix). As discussed in Section

Econometrica, 59, 1525-1550

TABLE VII

a. ENTRIES OF V FOR 1969.

HOU	FUE	FOO	CLO	DUR	TRA	SER	OGM	ATO
1.09	0.00	−0.15	−0.22	−0.10	−0.25	−0.06	−0.07	−0.25
	0.31	0.01	−0.05	−0.08	−0.11	−0.04	−0.02	−0.04
		0.98	0.02	−0.34	−0.52	−0.22	−0.03	0.18
			1.62	−0.26	−0.71	−0.24	−0.03	−0.08
				2.39	−0.70	−0.46	−0.16	−0.29
					3.66	−0.54	−0.33	−0.44
						1.94	−0.13	−0.24
							0.86	−0.14
								1.30

b. ENTRIES OF V FOR 1983.

HOU	FUE	FOO	CLO	DUR	TRA	SER	OGM	ATO
1.42	0.06	−0.14	−0.24	−0.24	−0.37	−0.14	−0.16	−0.21
	0.14	0.02	−0.03	−0.05	−0.08	−0.03	−0.02	−0.02
		0.82	0.12	−0.30	−0.35	−0.36	0.05	0.14
			1.19	−0.23	−0.40	−0.35	0.00	−0.04
				2.76	−0.75	−0.76	−0.21	−0.18
					3.43	−0.76	−0.35	−0.29
						2.94	−0.25	−0.23
							1.02	−0.05
								0.89

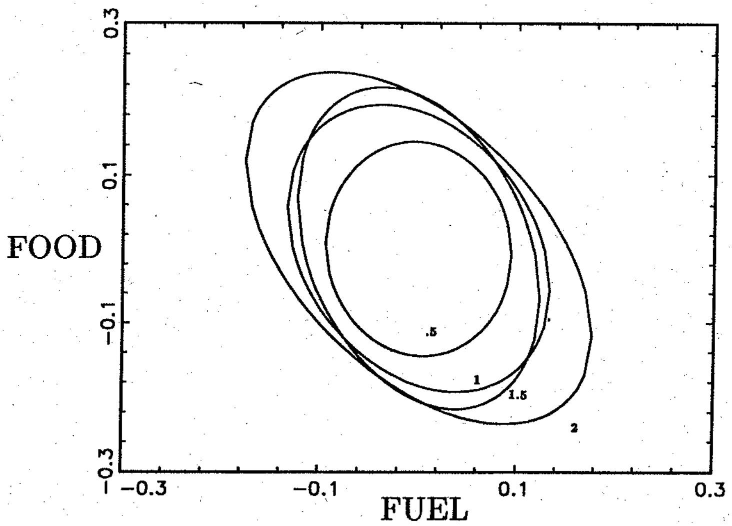


FIGURE 4.—Ellipses of concentration for 1983 at budget levels 0.5, 1.0, 1.5, 2.0 times the mean budget.

(1991) Härdle, W., Hildenbrand, W. and Jerison, M.
Empirical Evidence for the Law of Demand

Econometrica, 59, 1525-1550

2, these matrices determine ellipses of concentration for each pair of goods. (The coordinates of the ellipsoid that correspond to the other goods are set equal to zero.) These ellipses are not always nested, but are nearly so. Figure 4 shows the ellipses for food and fuel. The conditional variances of demands for nearly all goods are larger for β-households than for b-households when $\beta > b$. The only exception is for fuel with $b = 1$ and $\beta = 1.5$. On average, the dispersion of the consumers' demands clearly increases with the budget level.

CORE, Université Catholique de Louvain, 1348 Louvain-la-Neuve, Belgium,

Department of Economics, Universitat Bonn, Adenauerallee 24-26, W-5300 Bonn 1, Germany,

and

Department of Economics, State University of New York, Albany, NY 12222, USA

Manuscript received September, 1988; final revision received February, 1991.

APPENDIX

Estimation of A

In this section, we describe the procedure used to estimate the matrix

$$A = \int_{\mathbb{R}_+} (\partial_b \bar{G}(b))\rho(b)\,db.$$

The data consist of households' expenditures on each of the 9 commodity aggregates during a given period.

We normalize the prices of all commodity aggregates to be 1. A household's demand for a good is then equal to its expenditure on the good. The characteristics (b_i, α_i) of a randomly sampled household i have the distribution μ. The mean budget in the sample is denoted $\bar{b}$. We consider a fixed pair of goods j and k, and define $X_i = b_i/\bar{b}$ and $Y_i = f_j^{\alpha_i}(p, b_i) f_k^{\alpha_i}(p, b_i)/(\bar{b})^2$. Then we can interpret X_i as the budget of household i and Y_i as the jk component of the household's product matrix when the mean budget is normalized to 1. Since $\bar{b}$ is a sample mean, the pairs (X_i, Y_i) are correlated for different households. However, since the sample is large, the correlation is slight, and we will ignore it, treating the (X_i, Y_i) as i.i.d. These random variables have a distribution induced by μ, and the regression function is denoted $m(x) = E(Y_i | X_i = x)$. The jk component of A is then $\delta \bar{b}$, where

$$\delta = E_X m'(X)$$
$$= \int m'(x)\rho(x)\,dx$$

is the average derivative of m. By construction, the sum of the components of $f^{\alpha_i}(p, b_i)$ is b_i, and the b_i variables are distributed with compact support. Thus the distribution of (X_i, Y_i) has compact support.

Our approach to estimation of the average derivative δ is based on the simple observation that if ρ vanishes at the boundary of its support, then partial integration gives

$$\delta = \int m(x) L(x) \rho(x)\,dx$$

with

$$L = -d \log \rho / dx = -\rho'/\rho. \tag{4.1}$$

Since $L(\cdot)$ is not known we have to estimate it. We use the kernel technique and estimate the

(1991) Härdle, W., Hildenbrand, W. and Jerison, M.
Empirical Evidence for the Law of Demand

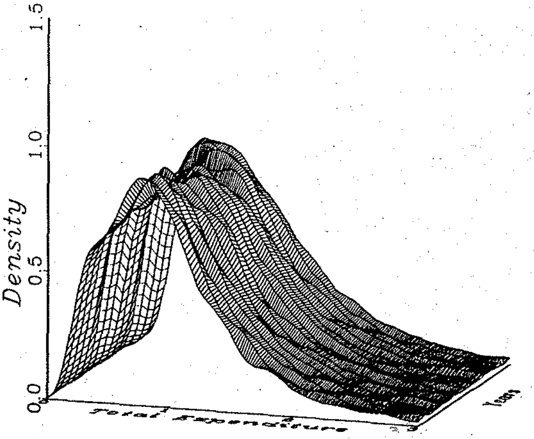


FIGURE 5.—The estimated densities of total expenditure $\hat{\rho}_h(x)$, 1968–1983.

density function $\rho(x)$ by a Rosenblatt-Parzen kernel density estimator

$$\hat{\rho}_h(x) = n^{-1} \sum_{i=1}^{n} K_h(x - X_i) \tag{4.2}$$

where $K_h(\cdot) = h^{-1}K(\cdot/h)$ is a kernel function with bandwidth h. We use a quartic kernel, $K(u) = (15/16)(1 - u^2)^2$ for $|u| \leqslant 1$; see Härdle (1990). Figure 5 shows the estimated density functions $\hat{\rho}_h$ for the entire sample period.

From the estimates $\hat{\rho}_h(x)$ we obtain as an approximation to $L(x)$ the ratio $\hat{L}_h(x) - \hat{\rho}_h'(x)/\hat{\rho}_h(x)$. (To avoid a zero denominator in low density regions we compute this only for budgets in the interval from 0.1 to 3 times the mean budget.) The Average Derivative Estimator $\hat{\delta}$ is then defined as

$$\hat{\delta} = n^{-1} \sum_{i=1}^{n} Y_i \hat{L}_h(X_i). \tag{4.3}$$

The argument in Härdle and Stoker (1989) yields the following theorem.

AVERAGE DERIVATIVE ESTIMATION THEOREM: *There exists a sequence of bandwidths $h_n \to 0$ with corresponding average derivative estimator $\hat{\delta}$, defined in* (4.3) *such that $\sqrt{n}(\hat{\delta} - \delta)$ has a limiting Normal distribution with mean* 0 *and variance σ^2, where*

$$\sigma^2 = \operatorname{var}\left[m'(X) + (Y - m(X))L(X)\right]. \tag{4.4}$$

This version of the theorem can be proved by modifying the proof of Härdle and Stoker (1989) slightly to allow for nonnegative kernels. The $\sqrt{n}$ rate of convergence is remarkable in that all the components of δ are nonparametrically estimated without any structural assumptions on ρ and m. Thus, although nonparametric estimation typically exhibits slower rates of convergence, the specific structure of the average derivative functional makes it possible to achieve the rate of convergence that is typical for parametric problems.

The computations for the A-matrix have been performed with a variety of values for the bandwidth h. All of the results reported in Section 3 use $h = 0.2$ (i.e. two tenths of the mean budget). This is the optimal value of h minimizing the mean square error (MSE) of (4.3). Härdle, Hart, Marron, and Tsybakov (1991) analyzed this mean square error and showed that there exist constants C_1 and C_2 such that $\text{MSE} = \sigma^2 n^{-1} + C_1 n^{-2} h^{-3} + C_2 h^4$. From this expression a "plug-in" estimate for the optimal h can be derived. The optimization of the kernel function for Average Derivative Estimation has been considered by Mammitzsch (1989) who showed that the Quartic kernel used in our studies is optimal.

In order to estimate the variability of the average derivative estimates we used the sample based terms given in Härdle and Stoker (1989, formula (3.6)),

$$\hat{r}_{hi} = \hat{L}_h(X_i)Y_i + n^{-1}\sum_{j=1}^{n}\left[K_h'(X_i - X_j) - K_h(X_i - X_j)\hat{L}_h(X_j)\right]\frac{Y_j}{\hat{\rho}_h(X_j)}. \tag{4.5}$$

The sample variance of these terms approximates the variance given in (4.4). The formula (4.5) is based on a linearization of the average derivative estimator in (4.3). The fact however that we used a fixed smoothing parameter for the whole range of income created high estimated variances for the entries of the $\hat{A}$ matrices. This becomes evident from Figure 5 which shows the estimated densities of total expenditure over time: at the far end (near the value of total expenditure 3.0) the estimate $\hat{\rho}_h(x)$ is very small. Therefore the score function L, although we used the cutoff technique described in Härdle and Stoker (1989), must become rather unstable. To overcome this difficulty we could, of course, use a varying bandwidth $h = h(x)$ but this is still an open problem.

An alternative method of measuring the standard error of the average derivatives is to compute the interquartile range (or F-spread) of the terms $\hat{r}_{hi}$ in (4.5). The F-spreads (times 100) for the diagonal elements of the A matrix of 1983 for instance are

$$(2.3, 0.6, 3.8, 1.9, 2.2, 7.4, 2.6, 1.0, 1.1).$$

The variances (times 100) of the terms $\hat{r}_{hi}$ for these diagonal elements are

$$(19.4, 2.14, 11.9, 11.0, 29.5, 32.4, 36.7, 7.1, 6.9).$$

The variances are much larger than the F-spreads because the distributions are highly skewed. For normal data the standard deviation is 1.39 times the F-spread.

Using these measures of variation we can consider the question of metonymy of the full population and each subclass defined by stratification. As an example we consider the age strata. Metonymy requires that A equal the weighted average of the A_i matrices estimated from the strata; see Section 2. For simplicity we consider the comparison of the diagonal elements. The weighted average matrix had the following diagonal elements in 1983:

$$(4.82, 0.46, 4.13, 1.81, 3.23, 7.29, 5.17, 1.41, 1.08).$$

As a first step one could treat these diagonal elements as given and apply a t test for each element. However, this procedure is inadequate because the two matrices are computed from the same data. The resulting correlation is accounted for in the following test suggested by Whitney Newey. Let ξ denote the vector of elements of A. Then

$$T = n(\hat{\xi}_1 - \hat{\xi}_2)\cdot\hat{\Sigma}^{-1}(\hat{\xi}_1 - \hat{\xi}_2)$$

is an asymptotic chi-square statistic for the difference between the stratified and unstratified estimates of A. Here $\hat{\xi}_1$ is the vector of components of $\hat{A}$, $\hat{\xi}_2$ is the vector of components of the weighted average of $\hat{A}_i$ estimates from the strata and $\hat{\Sigma}$ denotes a consistent variance estimator for the difference. Formula (4.5) can be used to calculate T:

$$\sqrt{n}\left(\hat{\xi}_j - \xi\right) \approx \sum_{i=1}^{n} r_{hi}^{(j)}/\sqrt{n}, \qquad j = 1, 2,$$

where $r_{hi}^{(j)}$ denotes the vector of terms in (4.5) for the stratified and unstratified case. The covariance matrix of the difference can be estimated by

$$\hat{\Sigma} = n^{-1}\sum_{i=1}^{n}\left(r_{hi}^{(1)} - r_{hi}^{(2)}\right)\left(r_{hi}^{(1)} - r_{hi}^{(2)}\right)^T.$$

We performed this test for the diagonal of A and obtained the value of $T = 0.046$ for the year 1983. The other years had T values in the range 0.03 to 0.1. So the hypothesis that the matrices are equal cannot be rejected.

Bootstrapping the Distribution of the Smallest Eigenvalue of $\hat{A}$

The distribution of the smallest eigenvalue of $\hat{A}$ is asymptotically normal, as is seen below in Theorem A. In the context of estimating covariance matrices similar asymptotic normality have been

Econometrica, 59, 1525-1550

derived. To our knowledge such a result for general random matrices is not available. In the following presentation we follow the paper by Härdle and Hart (1991). A column vector of 0's and a $k \times k$ identity matrix will be denoted, respectively, 0 and I. The eigenvalues of A are $\lambda_1 < \lambda_2 < \cdots < \lambda_k$, while those of $\hat{A}$ are $\hat{\lambda}_1 < \hat{\lambda}_2 < \cdots < \hat{\lambda}_k$. $C = [c_{ij}]$ will denote a $k \times k$ matrix with typical element c_{ij}. For any $k \times k$ symmetric matrix C, $u\,\mathrm{vec}(C)$ is the $k(k+1)/2$ component column vector $(c_{11}, \ldots, c_{1k}, c_{22}, \ldots, c_{2k}, \ldots, c_{kk})'$. Let V denote the asymptotic covariance matrix of $u\,\mathrm{vec}(\hat{A})$.

THEOREM A: *Define $A_{ij}(\lambda_1)$ to be the cofactor of the ijth element of $A - \lambda_1 I$. Let $B = 2[A_{ij}(\lambda_1)] - \mathrm{diag}(A_{11}(\lambda_1), \ldots, A_{kk}(\lambda_1))$, and let $D(x) = |A - xI|$. Then*

$$\sqrt{n}\left(\hat{\lambda}_1 - \lambda_1\right) \xrightarrow{D} N\left(0, \sigma_1^2\right),$$

where

$$\sigma_1^2 = \frac{u\,\mathrm{vec}(B)' V u\,\mathrm{vec}(B)}{\left(D'(\lambda_1)\right)^2}.$$

Although an estimator $\hat{V}$ of V can be constructed to use this result for testing $\lambda_1 > 0$ the procedure for doing so will be quite complicated. Therefore a bootstrap approximation to the distribution of $\sqrt{n}(\hat{\lambda}_1 - \lambda_1)$ seems to be an attractive alternative. The bootstrap we used resamples from the data $\{(b_i, f_j^{\alpha_i}(p, b_i))\}_{i=1}^n$ for a given year. More precisely n new observations are sampled with replacement. The bootstrap sample determines for each pair of goods a pair (X_i^*, Y_i^*) defined the same way as (X_i, Y_i).

To define the bootstrap distribution P^* of the smallest eigenvalue we have to compute A^*, the matrix $\hat{A}$ computed from a bootstrap sample (X_i^*, Y_i^*). Now calculate λ_1^*, the smallest eigenvalue of A^*. Repeated sampling allows one to approximate the bootstrap distribution P^* of $(\lambda_1^* - \hat{\lambda}_1)$ and then to conduct a test of the relevant hypothesis. Theorem B in Härdle and Hart (1991) shows, in fact, that the bootstrap distribution of $\sqrt{n}(\lambda_1^* - \hat{\lambda}_1)$ is asymptotically close to that of $\sqrt{n}(\hat{\lambda}_1 - \lambda_1)$.

A bootstrap test can now be conducted as follows. One determines an interval $[-B^*, C^*]$ from the bootstrap distribution of $\lambda_1^* - \hat{\lambda}_1$ which has probability, say, .95. Then one computes a confidence interval for λ_1 as $[\hat{\lambda}_1 - C^*, \hat{\lambda}_1 + B^*]$. The hypothesis of positive definiteness is rejected if $\hat{\lambda}_1 - C^* > 0$. (Of course, the nominal level of this one-sided test is .025.)

REFERENCES

BATTALIO, R. G. ET AL. (1973): A Test of Consumer Demand Theory Using Observations of Individual Consumer Purchases. *Western Economic Journal*, 11, 411–428.

CHIAPPORI, P. A. (1985): "Distribution of Income and the 'Law of Demand'," *Econometrica*, 53, 109–127.

CRAMÉR, H. (1946): *Mathematical Methods of Statistics*. Princeton: Princeton University Press.

DHRYMES, P. J. (1984): *Mathematics for Econometrics*. New York: Springer-Verlag.

FAMILY EXPENDITURE SURVEY, ANNUAL BASE TAPES (1968–1983): Department of Employment, Statistics Division, Her Majesty's Stationary Office, London, 1968–1983. The data utilized in this paper were made available by the ESRC Data Archive at the University of Essex.

FREIXAS, X., AND A. MAS-COLELL (1987): "Engel Curves Leading to the Weak Axiom in the Aggregate," *Econometrica*, 55, 515–531.

GRODAL, B., AND W. HILDENBRAND (1989): "Statistical Engelcurves, Income Distribution and the Law of Demand," SFB 303, Universität Bonn DP No. A-108. To appear in *Aggregation, Consumption and Trade: Essays in Honor of H. S. Houthakker*, ed. by L. Phlips and L. D. Taylor. Dordrecht: Kluwer Academic Publishers, 1992.

HÄRDLE, W. (1990): *Applied Nonparametric Regression*. Econometric Society Monograph Series 19. Cambridge: Cambridge University Press.

HÄRDLE, W., AND J. HART (1991): "A Bootstrap Test for Positive Definiteness of Income Effect Matrices," to appear in *Econometric Theory*.

HÄRDLE, W., J. HART, J. S. MARRON, AND A. B. TSYBAKOV (1991): "Choice of Smoothing Parameters for Average Derivative Estimation," to appear in the *Journal of the American Statistical Association*.

(1991) Härdle, W., Hildenbrand, W. and Jerison, M.
Empirical Evidence for the Law of Demand

HÄRDLE, W., AND T. M. STOKER (1989): "Investigating Smooth Multiple Regression by the Method of Average Derivatives," *Journal of the American Statistical Association*, 84, 986–995.

HICKS, J. R. (1956): *A Revision of Demand Theory*. Oxford: Clarendon Press.

HILDENBRAND, K., AND W. HILDENBRAND (1986): *On the Mean Income Effect: A Data Analysis of the U.K. Family Expenditure Survey*. Contributions to Mathematical Economics, in Honor of Gérard Debreu, ed. by W. Hildenbrand and A. Mas-Colell. Amsterdam: North Holland, 247–268.

HILDENBRAND, W. (1983): "On the Law of Demand," *Econometrica*, 51, 997–1019.

JERISON, M. (1982): "The Representative Consumer and the Weak Axiom when the Distribution of Income is Fixed," SUNY Albany, DP 150.

KANNAI, Y. (1989): "A Characterization of Monotone Individual Demand Functions," *Journal of Mathematical Economics*, 18, 87–94.

KEMSLEY, W. F., R. D. REDPATH, AND M. HOLMES (1980): *Family Expenditure Survey Handbook*. London: Her Majesty's Stationary Office.

LESER, C. E. (1963): "Forms of Engel Functions," *Econometrica*, 31, 594–703.

MAMMITZSCH, V. (1989): "Asymptotically Optimal Kernels for Average Derivative Estimation," manuscript, also given as an IMS Lecture, Davis, California, June, 1989.

MITJUSCHIN, L. G., AND J. POLTEROVICH (1978): "Criteria for Monotonicity of Demand Functions" (in Russian), *Ekonomika i Matematicheski Metody*, 14, 122–128.

SCHMIDT, H. (1989): "Family Expenditure Survey—Methodology, and Data Used in Microeconomic Demand Analysis," SFB 303, Universitat Bonn, DP No. A231.

STOKER, T. M. (1986): "Consistent Estimation of Scaled Coefficients," *Econometrica*, 54, 1461–1481.

Cross Section Engel Curves over Time

Wolfgang HÄRDLE[*]
CORE
Michael JERISON
Department of Economics, SUNY

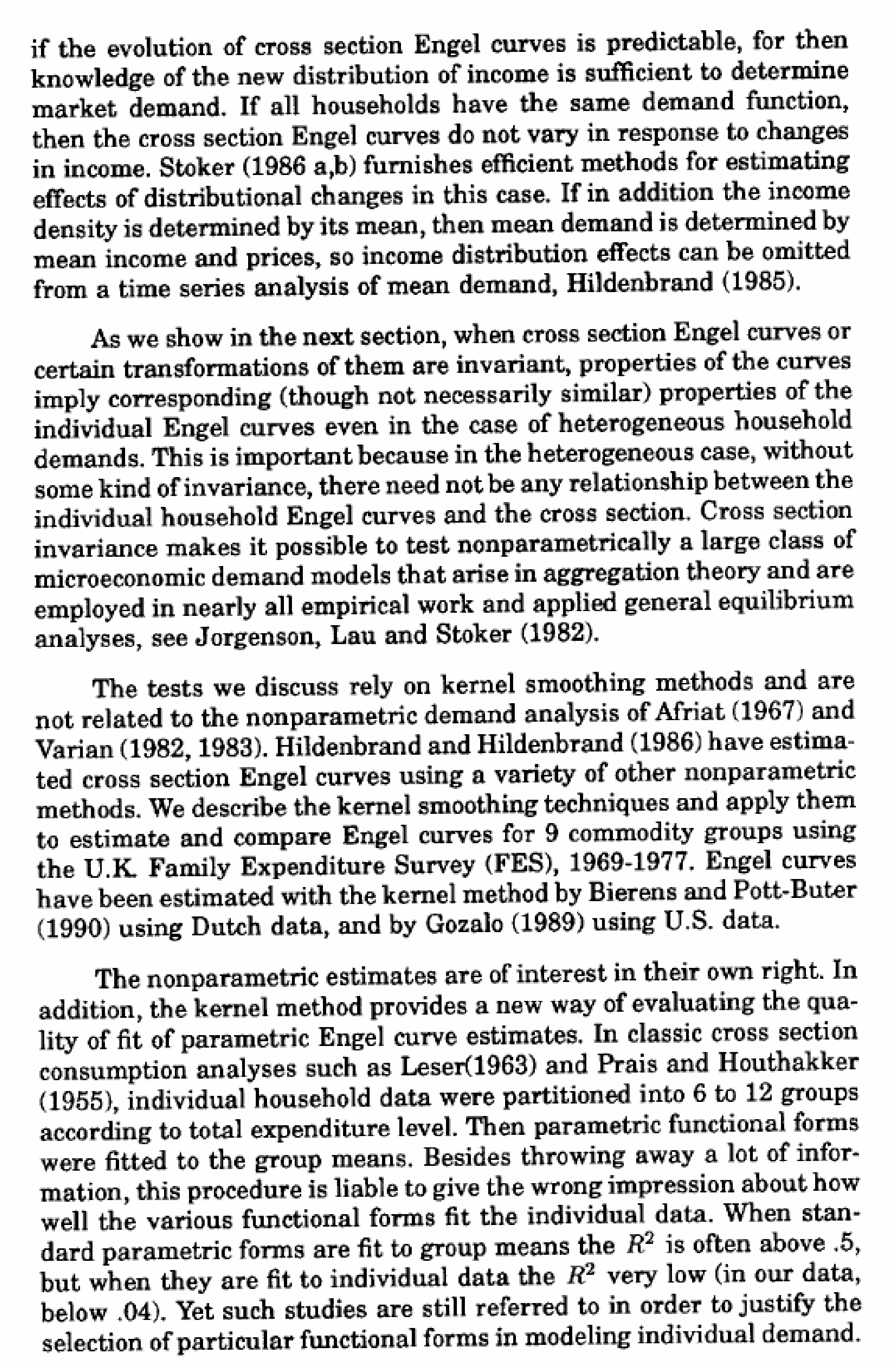
1 Introduction

The shapes of estimated cross section Engel curves are similar for many different countries and subgroups within countries. These regularities are sometimes referred to as empirical laws of demand behaviour. However, they might simply be consequences of the limited functional forms used in estimation. The present paper describes a framework for comparing cross section Engel curves using nonparametric techniques. These techniques permit flexible estimation and statistical comparison of curves without *a priori* assumptions about their functional form. We use the techniques to study Engel curves for a given population over time.

The theoretical and applied demand literature has devoted little attention to the evolution of cross section Engel curves. This is surprising since comparison of Engel curves from different periods can provide information about sensible interpretations and uses for cross section expenditure data. For example, cross section Engel curves are commonly used to classify goods as necessities or luxuries and to predict which income classes are most likely to benefit from small changes in commodity taxes. However, if the Engel curves change substantially over time —even during periods of stable prices and incomes— commodity classifications are likely to vary as well. This would cast doubt on the corresponding inferences about tax incidence.

Cross section data are also used to estimate income elasticities in order to predict effects of changing incomes. This approach is useful

[*] This paper was written at the University of Bonn. We would like to express our gratitude to the Deutsche Forschungsgemeinschaft,SFB 303 for financial support. We are also grateful for numerous discussions with Kurt Hildenbrand,Werner Hildenbrand, Heinz-Peter Schmitz and Tom Stoker. We thank Angus Deaton, Arthur Lewbel and three anonymous referees for comments on earlier versions of the paper.The data were made available by the ESRC Data Archive at the University of Essex.



if the evolution of cross section Engel curves is predictable, for then knowledge of the new distribution of income is sufficient to determine market demand. If all households have the same demand function, then the cross section Engel curves do not vary in response to changes in income. Stoker (1986 a,b) furnishes efficient methods for estimating effects of distributional changes in this case. If in addition the income density is determined by its mean, then mean demand is determined by mean income and prices, so income distribution effects can be omitted from a time series analysis of mean demand, Hildenbrand (1985).

As we show in the next section, when cross section Engel curves or certain transformations of them are invariant, properties of the curves imply corresponding (though not necessarily similar) properties of the individual Engel curves even in the case of heterogeneous household demands. This is important because in the heterogeneous case, without some kind of invariance, there need not be any relationship between the individual household Engel curves and the cross section. Cross section invariance makes it possible to test nonparametrically a large class of microeconomic demand models that arise in aggregation theory and are employed in nearly all empirical work and applied general equilibrium analyses, see Jorgenson, Lau and Stoker (1982).

The tests we discuss rely on kernel smoothing methods and are not related to the nonparametric demand analysis of Afriat (1967) and Varian (1982, 1983). Hildenbrand and Hildenbrand (1986) have estimated cross section Engel curves using a variety of other nonparametric methods. We describe the kernel smoothing techniques and apply them to estimate and compare Engel curves for 9 commodity groups using the U.K. Family Expenditure Survey (FES), 1969-1977. Engel curves have been estimated with the kernel method by Bierens and Pott-Buter (1990) using Dutch data, and by Gozalo (1989) using U.S. data.

The nonparametric estimates are of interest in their own right. In addition, the kernel method provides a new way of evaluating the quality of fit of parametric Engel curve estimates. In classic cross section consumption analyses such as Leser(1963) and Prais and Houthakker (1955), individual household data were partitioned into 6 to 12 groups according to total expenditure level. Then parametric functional forms were fitted to the group means. Besides throwing away a lot of information, this procedure is liable to give the wrong impression about how well the various functional forms fit the individual data. When standard parametric forms are fit to group means the R^2 is often above .5, but when they are fit to individual data the R^2 very low (in our data, below .04). Yet such studies are still referred to in order to justify the selection of particular functional forms in modeling individual demand.

The kernel method generates a uniform confidence region for the estimated Engel curve. The size of the region describes the precision of the estimate of the conditional mean. The region can be used to test hypotheses concerning the form of the underlying regression function.

The way Engel curves vary over time clearly depends on the way they are defined. For example, the *nominal* Engel curve for food, obtained by plotting mean food expenditure by households at each level of total expenditure, would probably shift up if food prices rose rapidly while other prices and incomes were fixed. We focus instead on the *real* Engel curve, obtained by plotting a quantity index of demand for the good at each level of *real* total expenditure in the cross section. If all prices and all households' total expenditures rose by the same percentage from one period to the next, homogeneity of individual demands would require the quantities demanded to remain fixed. This prediction could be tested by comparing real Engel curves from the two periods. One would expect real Engel curves to change in response to changes in relative prices. In fact we find that the estimated real Engel curve for a good generally moves in the direction opposite the change in the relative price of the good. Comparison of entire Engel curves allows one to observe the change in demand throughout the income distribution.

As noted above, cross section data can be used to test restrictions on microeconomic models if cross section Engel curves or transformations of them are invariant. We define *mean normalized* Engel curves which turn out to be more nearly invariant than the real Engel curves for our sample. Mean normalized curves are obtained from the nominal Engel curves by rescaling the axes, dividing x and y variables by their sample means. Invariance of the mean normalized curve for a particular good implies that households whose total expenditure on all commodities is a given fraction of economy-wide total expenditure spend a fixed fraction of economy-wide expenditure on the given good.

In the next section, we will describe some theoretical implications of invariance of real or mean normalized Engel curves when relative prices and household shares of total expenditure are fixed. These implications concern microeconomic models in which individual Engel curves lie in low dimensional subspaces of commodity space. Such models arise in the theory of income, preference and commodity aggregation, cf. Lau (1982), Jerison (1984), Lewbel (1991). The subspaces spanned by the Engel curves of different households need not be the same, so in general there is no connection between the individual and cross section Engel curves. However, if the real or mean normalized cross section curves are invariant when total expenditures of all households change by the same proportion, then the dimensions of the individual and cross section Engel curves are related. Because of this relationship one can learn from analyses of the dimension of cross section Engel curves (e.g. Hausman, Newey and Powell, 1988, and Lewbel, 1991) even when demand functions differ across households.

The model and results on dimension are described in section 2. In section 3 we discuss the kernel estimation procedure. The application to U.K. expenditure data is presented in section 4. Section 5 suggests some ways the analysis could be extended.

2 Some Consequences of Engel Curve Invariance

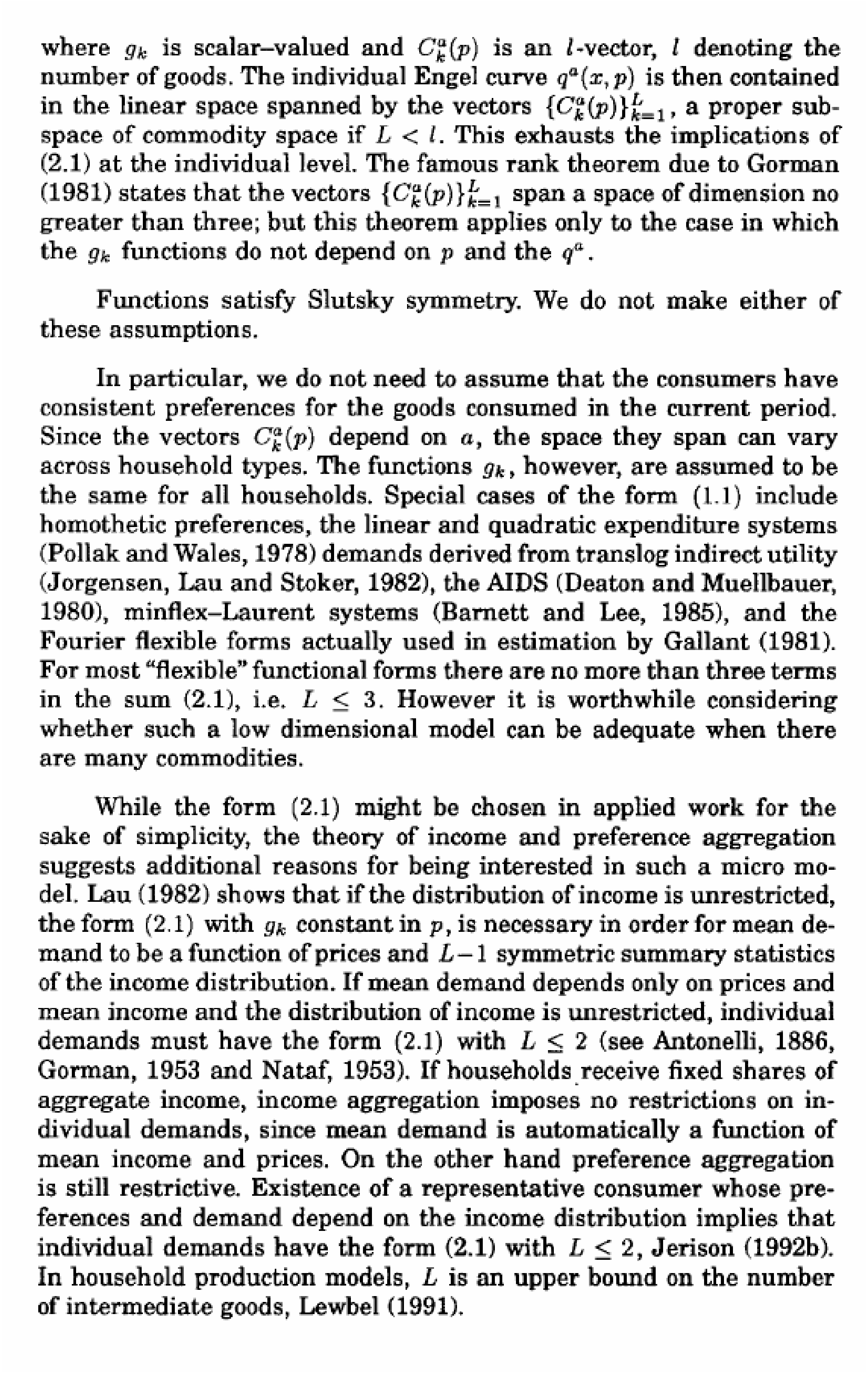
This section provides some theoretical motivation for studying Engel curve invariance. In particular, we show how standard microeconomic demand models can be related to cross section Engel curves if the latter do not vary during a period when relative prices and household shares of aggregate expenditure are fixed. Let $q^{aj}(x,p)$ be the demand for good j by a household of type a with total expenditure x at prices p. Although q^{aj} is a single valued function, this formulation can allow for random household demands by supposing that the type of a particular household is a random variable. The economy at time t is represented by a joint density function $z_t(a,x|p)$ of household types a and total expenditures x given the prices p. Let Y_t^j and X_t be random variables on the space of households at time t representing respectively expenditure on good j and total household expenditure. Then the cross section Engel curve for good j at time t is $q_t^j(\cdot,p_t)$ where

$$\begin{aligned} q_t^j(x,p_t) &= E_t(Y_t^j|X_t=x)/p_t^j \\ &= \int q^{aj}(x,p_t)z_t(a|x,p_t)da, \end{aligned}$$

with p_t^j the price of good j. Speaking in statistical terms the cross section Engel curve is the graph of the regression function for conditional mean demand with argument total expenditure x. Note that unless the households are identical there is no reason to expect any relationship between the individual and cross section Engel curves.

In order to investigate some of the theoretical consequences of invariance we assume that household demand is modeled in the following additive form used in virtually all empirical work and all applied (numerical) general equilibrium models:

$$q^a(x,p) = \sum_{k=1}^{L} g_k(x,p)C_k^a(p) \tag{2.1}$$

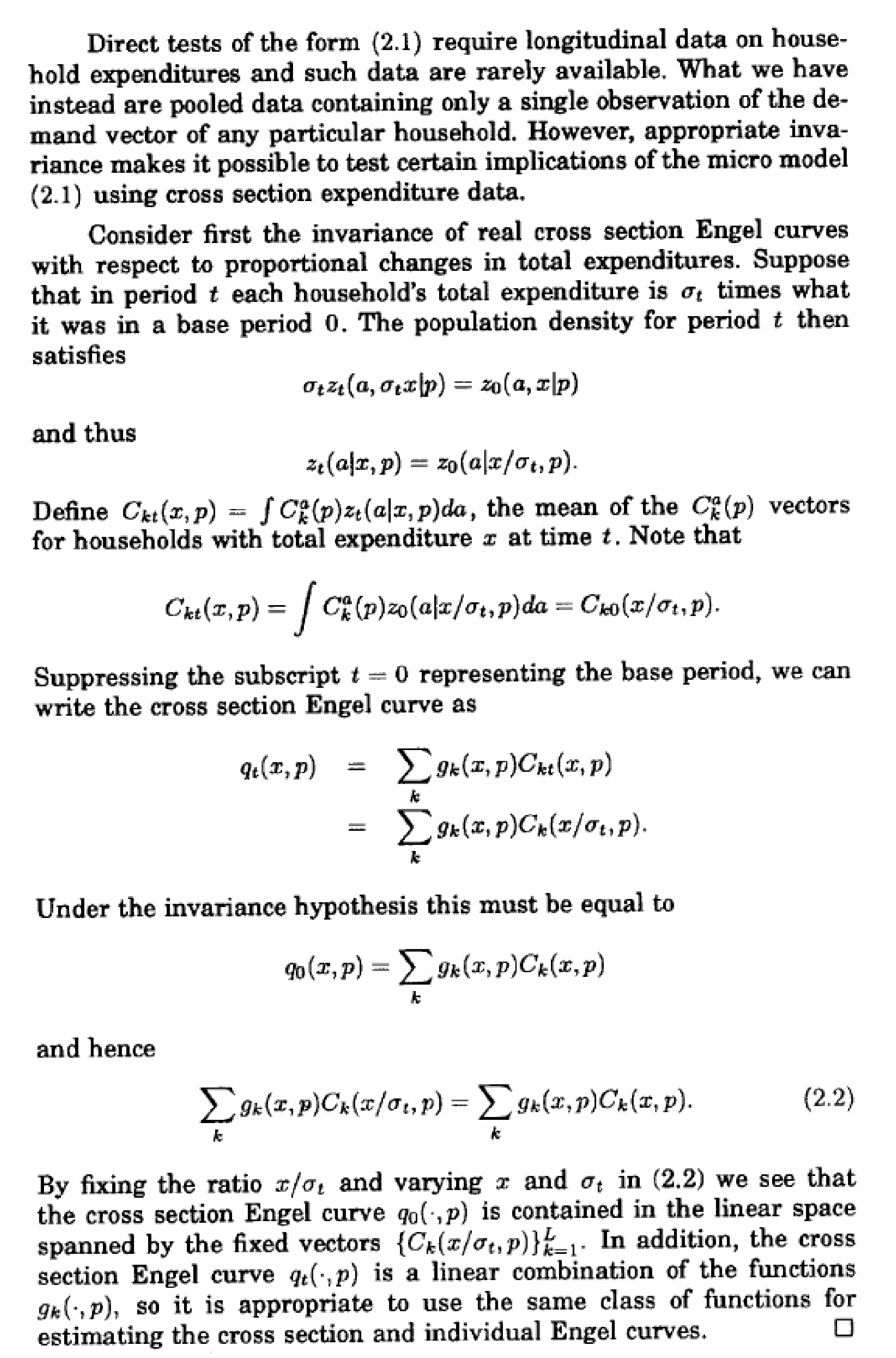


where g_k is scalar–valued and $C_k^a(p)$ is an l-vector, l denoting the number of goods. The individual Engel curve $q^a(x,p)$ is then contained in the linear space spanned by the vectors $\{C_k^a(p)\}_{k=1}^L$, a proper subspace of commodity space if $L < l$. This exhausts the implications of (2.1) at the individual level. The famous rank theorem due to Gorman (1981) states that the vectors $\{C_k^a(p)\}_{k=1}^L$ span a space of dimension no greater than three; but this theorem applies only to the case in which the g_k functions do not depend on p and the q^a.

Functions satisfy Slutsky symmetry. We do not make either of these assumptions.

In particular, we do not need to assume that the consumers have consistent preferences for the goods consumed in the current period. Since the vectors $C_k^a(p)$ depend on a, the space they span can vary across household types. The functions g_k, however, are assumed to be the same for all households. Special cases of the form (1.1) include homothetic preferences, the linear and quadratic expenditure systems (Pollak and Wales, 1978) demands derived from translog indirect utility (Jorgensen, Lau and Stoker, 1982), the AIDS (Deaton and Muellbauer, 1980), minflex–Laurent systems (Barnett and Lee, 1985), and the Fourier flexible forms actually used in estimation by Gallant (1981). For most “flexible” functional forms there are no more than three terms in the sum (2.1), i.e. $L \leq 3$. However it is worthwhile considering whether such a low dimensional model can be adequate when there are many commodities.

While the form (2.1) might be chosen in applied work for the sake of simplicity, the theory of income and preference aggregation suggests additional reasons for being interested in such a micro model. Lau (1982) shows that if the distribution of income is unrestricted, the form (2.1) with g_k constant in p, is necessary in order for mean demand to be a function of prices and $L-1$ symmetric summary statistics of the income distribution. If mean demand depends only on prices and mean income and the distribution of income is unrestricted, individual demands must have the form (2.1) with $L \leq 2$ (see Antonelli, 1886, Gorman, 1953 and Nataf, 1953). If households receive fixed shares of aggregate income, income aggregation imposes no restrictions on individual demands, since mean demand is automatically a function of mean income and prices. On the other hand preference aggregation is still restrictive. Existence of a representative consumer whose preferences and demand depend on the income distribution implies that individual demands have the form (2.1) with $L \leq 2$, Jerison (1992b). In household production models, L is an upper bound on the number of intermediate goods, Lewbel (1991).



Direct tests of the form (2.1) require longitudinal data on household expenditures and such data are rarely available. What we have instead are pooled data containing only a single observation of the demand vector of any particular household. However, appropriate invariance makes it possible to test certain implications of the micro model (2.1) using cross section expenditure data.

Consider first the invariance of real cross section Engel curves with respect to proportional changes in total expenditures. Suppose that in period t each household's total expenditure is σ_t times what it was in a base period 0. The population density for period t then satisfies

$$\sigma_t z_t(a, \sigma_t x | p) = z_0(a, x|p)$$

and thus

$$z_t(a|x,p) = z_0(a|x/\sigma_t, p).$$

Define $C_{kt}(x,p) = \int C_k^a(p) z_t(a|x,p)da$, the mean of the $C_k^a(p)$ vectors for households with total expenditure x at time t. Note that

$$C_{kt}(x,p) = \int C_k^a(p) z_0(a|x/\sigma_t, p)da = C_{k0}(x/\sigma_t, p).$$

Suppressing the subscript $t = 0$ representing the base period, we can write the cross section Engel curve as

$$\begin{aligned} q_t(x,p) &= \sum_k g_k(x,p) C_{kt}(x,p) \\ &= \sum_k g_k(x,p) C_k(x/\sigma_t, p). \end{aligned}$$

Under the invariance hypothesis this must be equal to

$$q_0(x,p) = \sum_k g_k(x,p) C_k(x,p)$$

and hence

$$\sum_k g_k(x,p) C_k(x/\sigma_t, p) = \sum_k g_k(x,p) C_k(x,p). \qquad (2.2)$$

By fixing the ratio x/σ_t and varying x and σ_t in (2.2) we see that the cross section Engel curve $q_0(\cdot, p)$ is contained in the linear space spanned by the fixed vectors $\{C_k(x/\sigma_t, p)\}_{k=1}^L$. In addition, the cross section Engel curve $q_t(\cdot, p)$ is a linear combination of the functions $g_k(\cdot, p)$, so it is appropriate to use the same class of functions for estimating the cross section and individual Engel curves. □

Theorem 2.1 *If individual Engel curves have the additive form* (2.1) *and if the cross section Engel curve does not change when all households' total expenditures change by the same proportion, then the cross section Engel curve is contained in an L dimensional subspace of commodity space and has the form (2.1) with the superscript a removed.*

Theorem 2.1 provides a nonparametric test of the model (2.1) for individual demands. If the cross–section Engel curve does not change when households' total expenditures change by the same proportion then the dimension of the span of the cross–section Engel curve is a lower bound on the number of terms L in the sum (2.1). This test is nonparametric since it does not require specifying the forms of the functions g_k and C_k^a in (2.1).

Consider now the "macro" effect of the proportional increase in total expenditures by all households. Mean demand in period t is

$$\int \sum g_k(\sigma_t x, p) C_k^a(p) z_0(a, x|p) da\ dx \ ,$$

a function of σ_t that is also a linear combination of the functions g_k. This function is in general different from the cross section Engel function $q_0(\cdot, p)$ but the images of the two functions lie in the same L dimensional space.

The hypothesis that the total expenditure of each household changes by the same percentage is restrictive, but of course the invariance assumption would be even stronger if it were to hold with respect to a larger class of changes in household budgets. The case of proportional changes is important since the distribution of total expenditures can in fact nearly be parametrized by its mean (see Hildenbrand, 1985). If households' budgets do not move in fixed proportion the effects on mean demand would be as described above if the deviations from proportionality were uncorrelated with the households' marginal propensities to consume.

Note that the test applies even if prices are not constant. Constant *relative* prices are all that is needed. (See section 4.b.) Since the data we analyze in the next section exhibit significant relative price changes, they do not provide sharp tests of (2.1). The data appear to be more compatible with invariance of transformed Engel curves than they are with invariance of the real Engel curves themselves. For this reason we consider the consequences of a second invariance concept. We say that Engel curves exhibit *mean normalized invariance* with respect to some change in the population density if the Engel curves with x and y variables divided by their sample means do not change. For the density change considered above (proportional changes in all households' total expenditure) mean normalized invariance is equivalent to the requirement that for each good j,

$$q_t^j(\sigma_t x, p)/Q_t^j(p)$$

is independent of t, where

$$Q_t^j(p) = \int q^a(x,p) z_t(a, x|p) da \, dx$$

is economy-wide mean demand for good j at time t with prices p. Jerison (1992a) proves that if individual demands satisfy homogeneity and the budget identity and have the form (2.1) with g_k constant in p, then under mean normalized invariance plus a weak regularity condition the cross section Engel curve lies in an $l - L + 1$ dimensional subspace of commodity space $\mathbb{R}_+^l$. We will prove this result for the special case when the functions g_k are power functions (see Appendix).

Theorem 2.2 *Suppose all households' demands have the form*

$$q^a(x,p) = \sum_{k=1}^{L} x^{\gamma_k} C_k^a(p) \qquad (2.3)$$

and satisfy the budget identity. If the cross section Engel curve for each good satisfies mean normalized invariance with respect to proportional changes in households' total expenditure then the cross section Engel curve at prices p lies in a linear subspace of dimension $l - L^* + 1$ *where* L^* *is the rank of the moments matrix with k-th column* $E_0(x^{\gamma_k} C_k(x,p))$, *where the expectation is with respect to x in the base period.*

By considering perturbations of economies with demands of the form (2.3) we can say that generically the moments matrix has full rank. This implies that generically $L^* = L$, so the cross section Engel curve lies in an $l - L + 1$ dimensional subspace. As shown above, the "macro" function associating mean demand with mean total expenditure is a linear combination of the functions $g_k(\cdot, p)$ and hence its image lies in an L dimensional space. Theorem 2.2 thus shows that this macro function and the cross section Engel curve are in a sense complementary to each other. When one has high dimensional span the other's span must be of low dimension. This illustrates the importance of the evolution of the cross section when one tries to interpret the form of the Engel curves.

Gorman (1981) showed that if a demand function of the form (2.3) satisfies the budget identity and Slutsky symmetry then at each p the rank of the matrix with columns $C_k^a(p)$ is at most three. Of course the rank cannot be greater than L, but this does not prevent L from being greater than three. Under the invariance hypothesis of Theorem 2.2, if for some good the individual households' Engel curves are high degree polynomials then the span of the cross section Engel curve must be *low* dimensional.

Finally, there is a large literature concerning the treatment of zero expenditures in the estimation of Engel curves, cf. Keen (1986), Pudney (1987). Household s have a variety of reasons for not purchasing a commodity during a particular period. They might never purchase the commodity, or they might purchase it rarely but in large quantities. It is possible that such variation in purchasing behaviour could lead to violations of the invariances defined above. However, if either invariance holds, we have a theorem relating the observed cross section En gel curves to the Engel curves of individual households. Thus, if the hypothesized invariance is satisfied, no adjustment is needed to take account of zero expenditures.

3 Kernel Smoothers and Confidence Bands

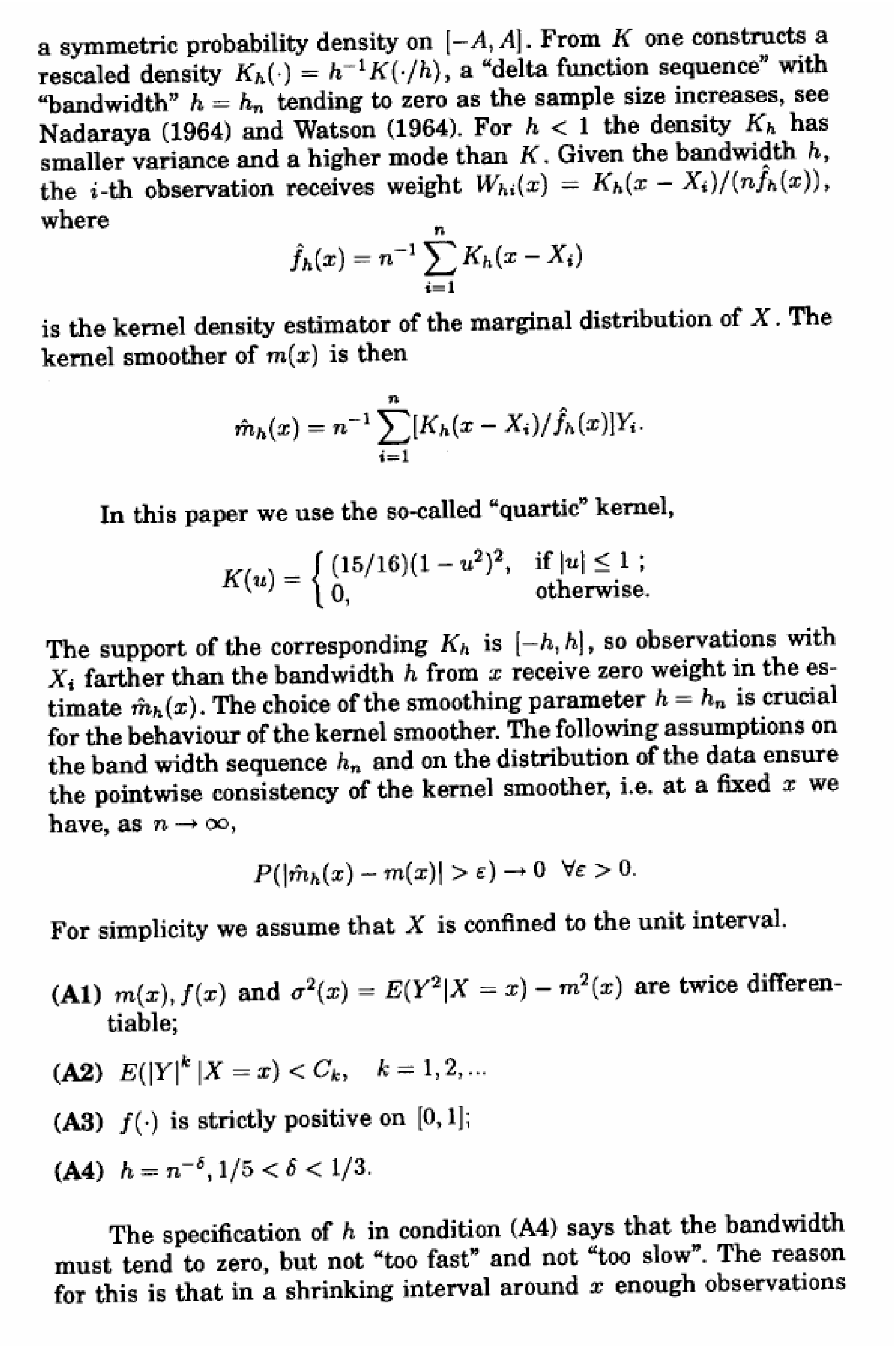
In this section we give some background on the statistical theory of nonparametric smoothing techniques that will be used to compare Engel curves. The basic framework is to treat the joint observations of household expenditure Y on a particular good and total expenditure X as a sample $\{(X_i, Y_i)\}_{i=1}^n$ of independent identically distributed random variables with density $g(x, y)$. Given this sample the aim is to estimate the mean expenditure given income, i.e. the conditional expectation

$$m(x) = E(Y|X = x) = \int yg(x, y)dy/f(x)$$

where

$$f(x) = \int g(x, y)dy$$

denotes the marginal density of X. The graph of m is the nominal Engel curve for the good during the given period. In this paper we use kernel smoothers for the sake of statistical and methodological as well as numerical simplicity. Kernel smoothers provide an estimate of $m(x)$ as a weighted average of the observations Y_i. The weights also depend on the sample size. As the sample grows the number of observations X_i within a certain distance from a given x increases, so in estimating $m(x)$, increasingly heavy weight can be placed on observations with X_i's nearer to x. The weights are constructed via a ***kernel*** function K,



a symmetric probability density on $[-A, A]$. From K one constructs a rescaled density $K_h(\cdot) = h^{-1}K(\cdot/h)$, a "delta function sequence" with "bandwidth" $h = h_n$ tending to zero as the sample size increases, see Nadaraya (1964) and Watson (1964). For $h < 1$ the density K_h has smaller variance and a higher mode than K. Given the bandwidth h, the i-th observation receives weight $W_{hi}(x) = K_h(x - X_i)/(n\hat{f}_h(x))$, where

$$\hat{f}_h(x) = n^{-1}\sum_{i=1}^{n} K_h(x - X_i)$$

is the kernel density estimator of the marginal distribution of X. The kernel smoother of $m(x)$ is then

$$\hat{m}_h(x) = n^{-1}\sum_{i=1}^{n}[K_h(x - X_i)/\hat{f}_h(x)]Y_i.$$

In this paper we use the so-called "quartic" kernel,

$$K(u) = \begin{cases} (15/16)(1-u^2)^2, & \text{if } |u| \le 1 \,; \\ 0, & \text{otherwise.} \end{cases}$$

The support of the corresponding K_h is $[-h, h]$, so observations with X_i farther than the bandwidth h from x receive zero weight in the estimate $\hat{m}_h(x)$. The choice of the smoothing parameter $h = h_n$ is crucial for the behaviour of the kernel smoother. The following assumptions on the band width sequence h_n and on the distribution of the data ensure the pointwise consistency of the kernel smoother, i.e. at a fixed x we have, as $n \to \infty$,

$$P(|\hat{m}_h(x) - m(x)| > \varepsilon) \to 0 \quad \forall \varepsilon > 0.$$

For simplicity we assume that X is confined to the unit interval.

(A1) $m(x), f(x)$ and $\sigma^2(x) = E(Y^2|X = x) - m^2(x)$ are twice differentiable;

(A2) $E(|Y|^k\,|X = x) < C_k, \quad k = 1, 2, \ldots$

(A3) $f(\cdot)$ is strictly positive on $[0, 1]$;

(A4) $h = n^{-\delta}, 1/5 < \delta < 1/3.$

The specification of h in condition (A4) says that the bandwidth must tend to zero, but not "too fast" and not "too slow". The reason for this is that in a shrinking interval around x enough observations must be collected to ensure that the variance of the estimator tends to zero. A proof of consistency is given in Härdle (1990, Proposition 3.1). Pointwise confidence intervals may be constructed from the following result on asymptotic normality of the estimator, see Schuster (1972).

Theorem 3.1 *Suppose that (A1-5) hold and let $x_1, x_2, \ldots, x_k$ be distinct points in $(0,1)$. Define*

$$Z_n(x) = (nh)^{1/2} \left\{ \frac{(\hat{m}_h(x) - m(x))}{(\sigma^2(x) \int K^2/f(x))^{1/2}} \right\} .$$

Then for the kernel smoother $\hat{m}_h(\cdot)$ the vector $(Z_n(x_1), \ldots, Z_n(x_k))$ converges in distribution to a multivariate normal random vector with zero mean vector and identity covariance matrix.

The theorem depends on the sample being random. Adjustments can be made for stratified samples, but we will not discuss them here, cf. Bierens and Pott-Buter (1990). A crucial step in applying the kernel smoothing technique is to find a reasonable bandwidth h for the given data set. Qualitatively speaking, the kernel smoother $\hat{m}_h(\cdot)$ tends to "follow the observations" closely if this smoothing parameter is chosen too small. On the other hand if this parameter is chosen too big, the kernel smoother will "flatten out" local fluctuations in the regression curve. The former choice, with a bandwidth that is too small, leads to an increase in variance of the estimator but a decrease in bias, whereas the latter choice with too large a bandwidth will result in an increased bias and decreased variance. A way to deal with the trade-off between these two effects, which are typical for nonparametric smoothing methods, is to choose the bandwidth by cross validation, see Härdle and Marron (1985). One selects the smoothing parameter such that the cross validation function

$$CV(h) = n^{-1} \sum_{i=1}^{n} (Y_i - \hat{m}_{h,i}(X_i))^2$$

is minimized. Here $\hat{m}_{h,i}(\cdot)$ denotes the kernel smoother computed from the subsample $\{(X_j, Y_j)\}_{j \neq i}$, leaving out the i-th observation. The method of cross validation automatically selects the best smoothing parameter possible (in a squared error sense) as was shown by Härdle and Marron (1985). It turns out that in an asymptotic sense the optimal bandwidth has the speed $n^{-1/5}$ which is just at the border of the range of bandwidths h allowed in Theorem 3.1. The graph of the cross validation function for the data set (X = total expenditure/mean total expenditure, Y = housing expenditure) is depicted in figure 1.

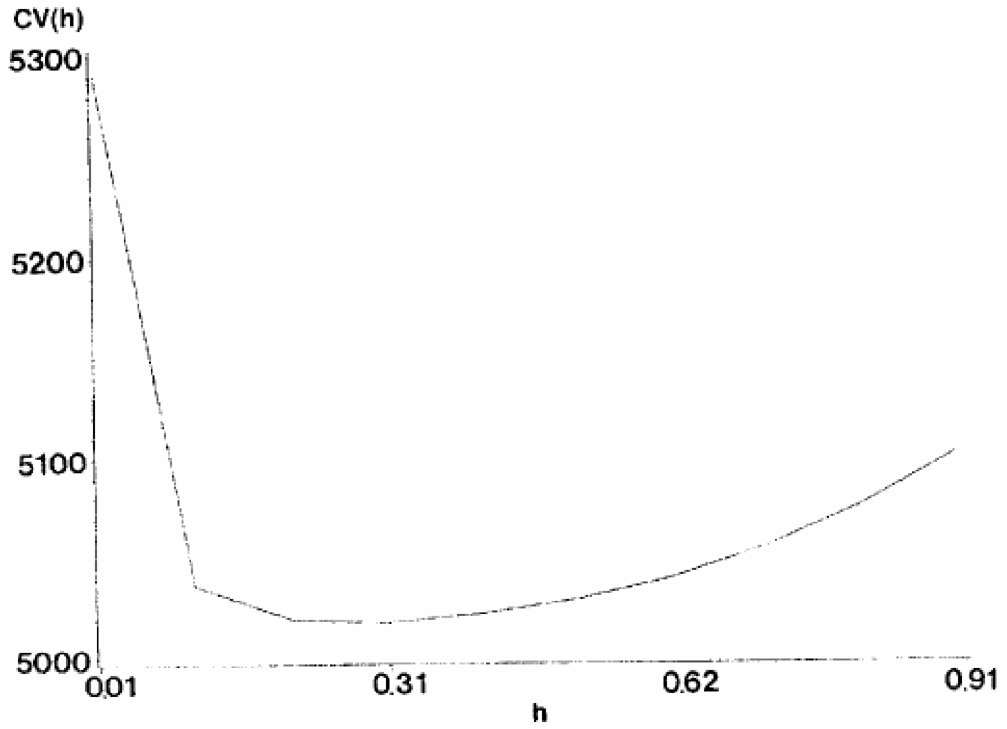
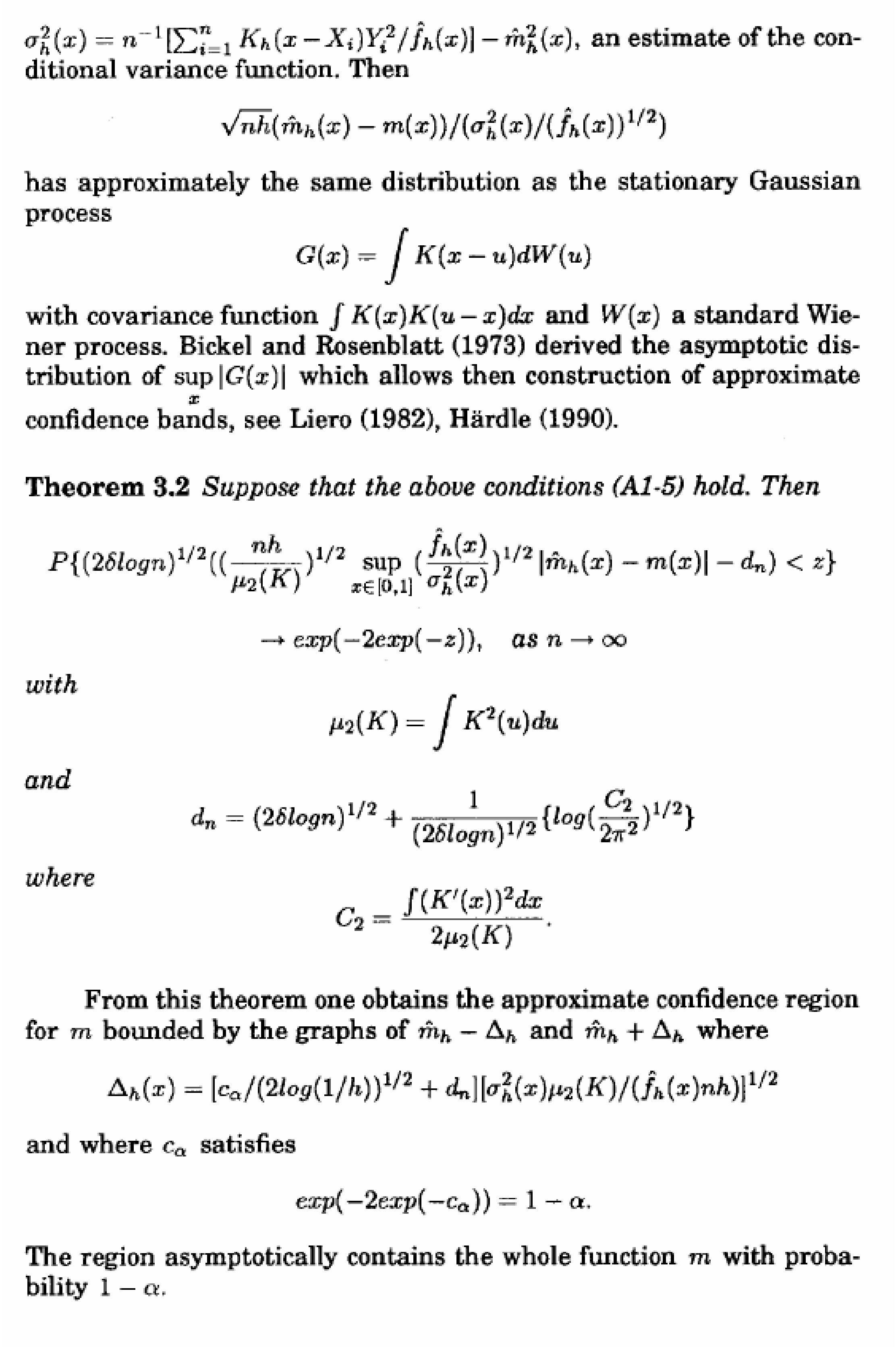


Figure 1
The cross validation function for the housing Engel curve. UK Family Expenditure Survey,1973.

The cross validation function above indicates that one should choose the bandwidth roughly equal to 0.3 times the mean of total expenditure to optimize the tradeoff between bias and variance. However, the bandwidth chosen in this way is sample dependent, which makes Theorem 3.1 inapplicable. Therefore in selecting the bandwidth for a given sample, we use c ross validation scores from the samples drawn in adjacent years. This procedure is justified by the invariance of the cross validation functions over time. In fact the optimal bandwidths for all goods and all years fall in the range $[0.2, 0.3]$ (times the mean of total expenditure) so we set $h = 0.25$ (times the mean of total expenditure) as an overall reasonable smoothing parameter.

The construction of uniform confidence bands

The idea in constructing uniform confidence bands is to approximate the suitably standardized process $\sqrt{nh}(\hat{m}_h(x) - m(x))$ by a certain Gaussian process. Consider the above kernel weights. Define



$\sigma_h^2(x) = n^{-1}[\sum_{i=1}^{n} K_h(x-X_i)Y_i^2/\hat{f}_h(x)] - \hat{m}_h^2(x)$, an estimate of the conditional variance function. Then

$$\sqrt{nh}(\hat{m}_h(x) - m(x))/(\sigma_h^2(x)/(\hat{f}_h(x))^{1/2})$$

has approximately the same distribution as the stationary Gaussian process

$$G(x) = \int K(x-u)dW(u)$$

with covariance function $\int K(x)K(u-x)dx$ and $W(x)$ a standard Wiener process. Bickel and Rosenblatt (1973) derived the asymptotic distribution of $\sup_x |G(x)|$ which allows then construction of approximate confidence bands, see Liero (1982), Härdle (1990).

Theorem 3.2 *Suppose that the above conditions (A1-5) hold. Then*

$$P\{(2\delta log n)^{1/2}((\frac{nh}{\mu_2(K)})^{1/2} \sup_{x\in[0,1]} (\frac{\hat{f}_h(x)}{\sigma_h^2(x)})^{1/2} |\hat{m}_h(x) - m(x)| - d_n) < z\}$$

$$\to exp(-2exp(-z)), \quad as\ n \to \infty$$

with

$$\mu_2(K) = \int K^2(u)du$$

and

$$d_n = (2\delta log n)^{1/2} + \frac{1}{(2\delta log n)^{1/2}}\{log(\frac{C_2}{2\pi^2})^{1/2}\}$$

where

$$C_2 = \frac{\int (K'(x))^2 dx}{2\mu_2(K)}.$$

From this theorem one obtains the approximate confidence region for m bounded by the graphs of $\hat{m}_h - \Delta_h$ and $\hat{m}_h + \Delta_h$ where

$$\Delta_h(x) = [c_\alpha/(2log(1/h))^{1/2} + d_n][\sigma_h^2(x)\mu_2(K)/(\hat{f}_h(x)nh)]^{1/2}$$

and where c_α satisfies

$$exp(-2exp(-c_\alpha)) = 1 - \alpha.$$

The region asymptotically contains the whole function m with probability $1-\alpha$.

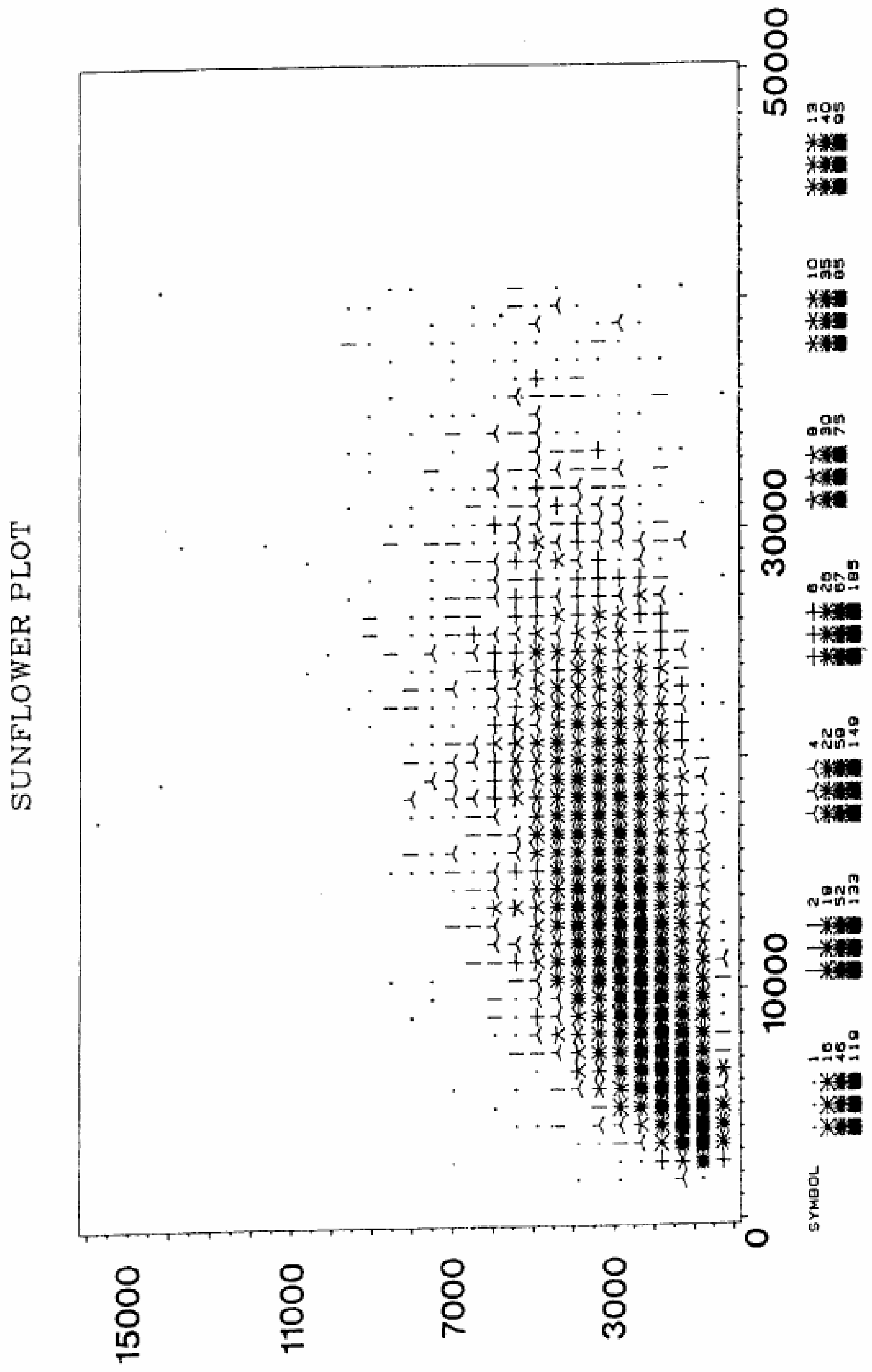


Figure 2
Sunflower Plot of Food Expenditure vs. Total Expenditure, unit: pence per week, 1981, FES

Note that for a given sample size n and bandwidth h, the size of the confidence band $2\Delta_h(x)$ depends only on the density estimate $\hat{f}_h(x)$ and the conditional variance estimate $\sigma_h^2(x)$ of expenditure Y on the good. It is proportional to the standard deviation $\sigma_h(x)$ and inversely proportional to the square root of the density.

4 Empirical Results

In this section we apply the techniques described above to U.K. Family Expenditure Survey (FES) data from 1969 to 1977. Cross section Engel curves for nine commodity aggregates have been estimated on the entire sample of households in each odd numbered year. The commodity aggregates are housing, fuel, food, clothing, durables, transport, services, alcohol and tobacco, and "miscellaneous and other goods". The curve for each aggregate j was estimated using the quartic kernel, the Working (1943) parametric model

$$\hat{q}^j(x) = (\beta_j + \alpha_j ln x)x \tag{4.1}$$

and a cubic polynomial in x, where x is total expenditure. We compare the parametric and nonparametric estimates for the same data from a given year. We then compare transformed versions of the parametrically and nonparametrically estimated Engel curves from different years, using transformations appropriate for testing the invariance discussed in section 2.

Figure 2 is a sunflower plot of the 1981 distribution of households' total expenditures and expenditures on food. The density of observations in a region is represented by the number of petals on the flowers there. Expenditures are measured here and below (unless otherwise noted) in pence per week. The dispersion in Figure 2 is striking. Households in the same column, i.e. with weekly total expenditure within a few pence of each other differ by a factor of fifteen in their food expenditures. This is undoubtedly related to the fact that the expenditures are observed during a period of only two weeks.

Figure 3 shows the 1971 nominal cross section Engel curve along with the uniform 95% confidence band described in section 3. Although the scatter plot for the year 1971 is as dispersed as in Figure 2, the conditional mean food expenditure is estimated very precisely at total expenditurelevels near the sample mean of 3099. The confidence bands spread out at higher total expenditure levels since there are few observations above twice the mean. Thus one cannot hope for precise Engel curve estimates over this high range of expenditure levels.

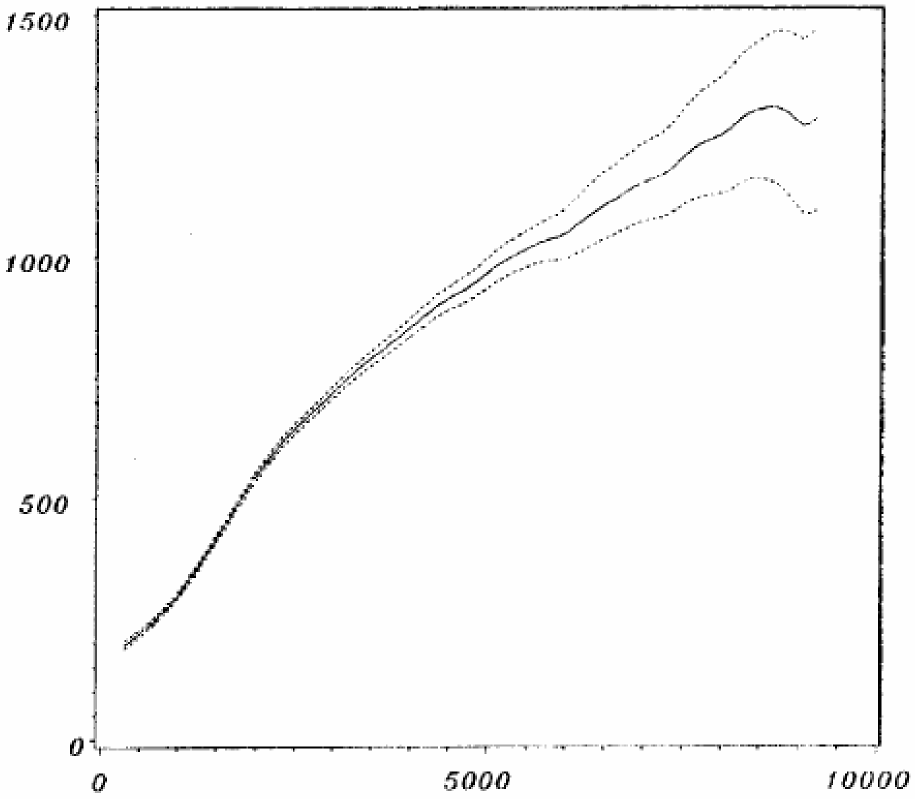


Figure 3
Cross section nominal Engel curve for food with confidence bands, year 1971

4.1 Comparison of Parametric and Nonparametric Estimates

The confidence bands defined in the last section can be used to test the hypothesis that data (X_i, Y_i) from a given year were randomly drawn from a distribution on the positive orthant with a given curve as its conditional expectation function $m(x) = E(Y|X = x)$. The test is simply whether the given curve lies within the confidence bands. In this way we see that the parametric Working (1943) form of Engel function (4.1) is nearly always rejected at a 95% confidence level. The main exception is for the durables aggregate, which has the largest coefficient of variation of expenditure at most levels of total expenditure x. This implies that its confidence bands are wider than for other goods (relative to the level of expenditure on the commodity). Examples of the above tests are given in Figures 4a,b and c which illustrate kernel estimators and their confidence bands along with Working parametric estimators of the 1971 Engel curves for food, housing and fuel.

For many goods and years the rejection of the Working fit is small in the sense that the parametric estimate lies outside the confidence bands only over a small interval of total expenditures, and its distance from the confidence band is relatively small. For example, the rejection for transport occurs only at low levels of total expenditure. In other cases, e.g. food and housing in Figures 4a and 4b, the parametric fit leaves the confidence bands over a larger set of total expenditures, but the distance from the bands remains relatively small. The comparison of parametric and nonparametric estimates yields a new test of the quality of fit of parametric estimates. As the dispersion in Figure 2 suggests, the R^2 for the Working Engel curve estimates is extremely low (often below 0.03). However, what matters is the difference between the estimated Engel function $\hat{q}^j$ in (4.1) and the population Engel function q_t^j for the period. The comparison of nonparametric and parametric estimates gives an idea of the distance between the parametric estimate and the conditional mean at each level of the independent variable x. If parametric forms are needed for theoretical or computable economic models, the analyst has a basis for trading off simplicity of functional form against goodness of fit.

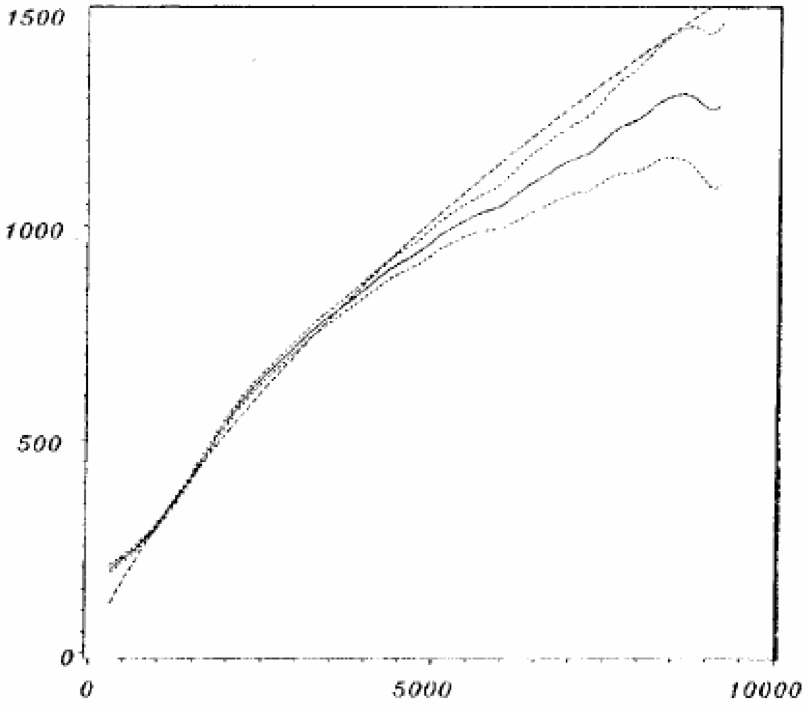


Figure 4.a
Nonparametric nominal Engel curve for food, and Working fit,1971, FES

The Working model is generally thought to provide a good fit for food (cf. Leser, 1963), yet in Figure 4a it appears to underestimate

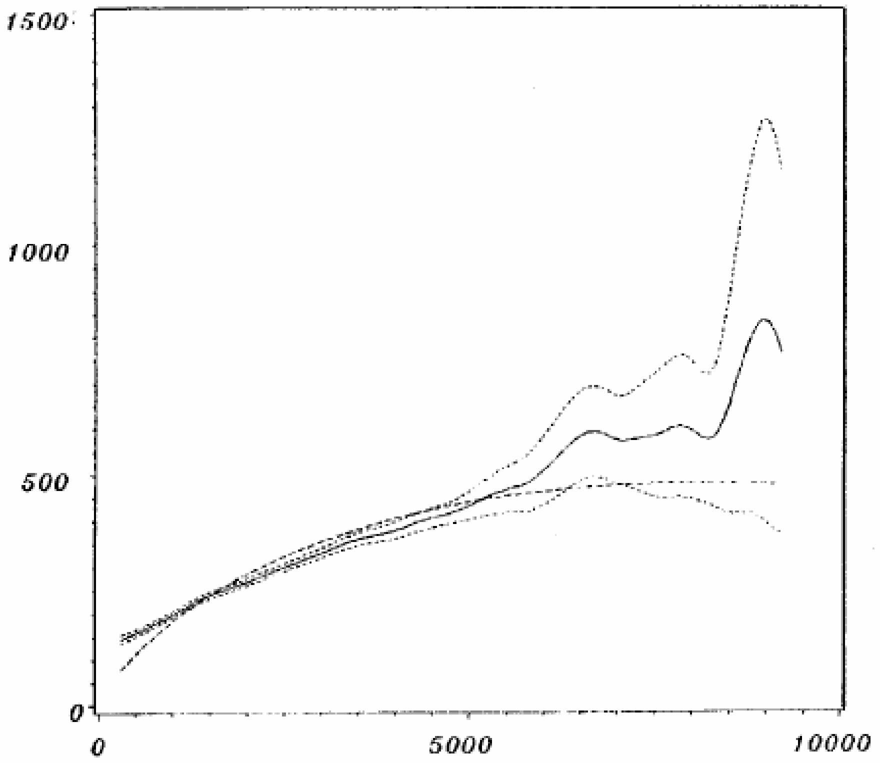


Figure 4.b: Nominal Engel curve for housing, and Working fit, 1971, FES

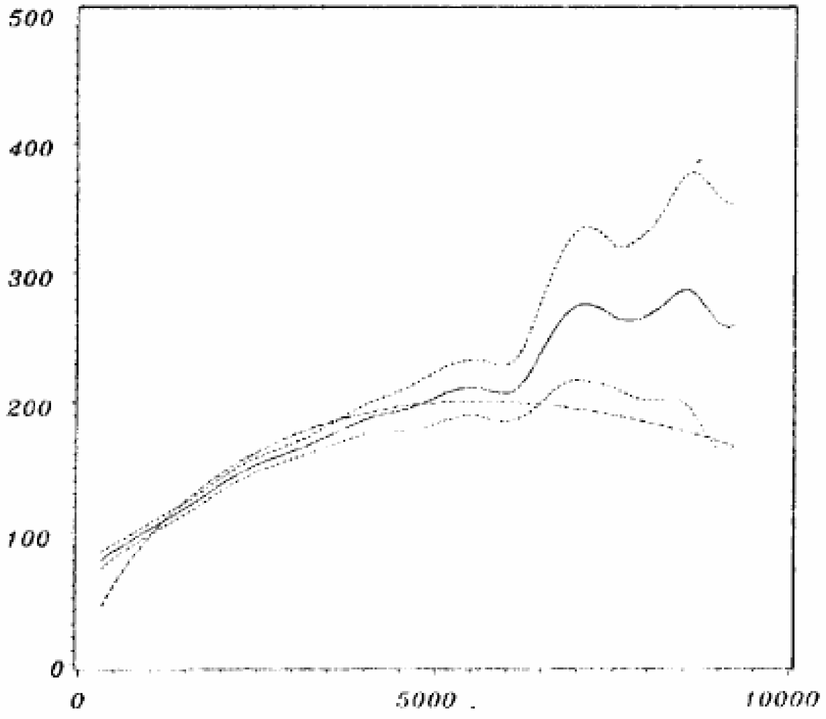


Figure 4.c: Nominal Engel curve for fuel, and Working fit, 1971, FES

the concave curvature of the regression function in the range from .5 to 3 times the mean total expenditure, and it is too rigid to permit a convex portion as appears in the kernel estimate below .5 times mean x. The Working regression (4.1) is strictly convex [concave] depending on whether $\alpha_j >$ [<]0.

For housing in Figure 4b, the parametric estimate is too concave to fit the confidence bands since the nonparametric estimate is convex over an interval beginning slightly above the mean total expenditure. The difference between the parametric and nonparametric estimates for fuel in Figure 4c is even more striking. The Working estimate is downward sloping above 1.5 times mean total expenditure, whereas the average slope of the nonparametric estimate is positive and large in magnitude over this interval. The curvature of the parametric estimates is quite sensitive to outliers at high and low levels of total expenditure. It is possible that the Working model provides a more adequate mod el of the cross section Engel curves for more homogeneous strata of the population. For narrowly defined strata the Working model often cannot be rejected. However this follows from the imprecision of the estimated mean in small samples and cannot be viewed as offering support for the parametric model.

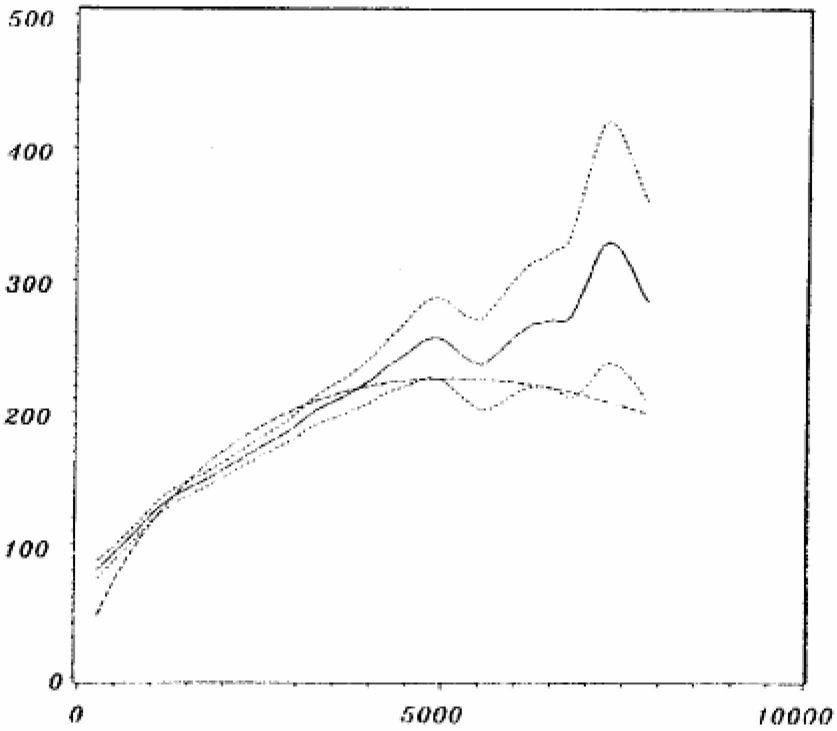


Figure 4.d: Nominal Engel curve for fuel, and Working fit, 1969, FES

We have also compared the nonparametric estimates to parametric estimates from the family of cubic polynomials. The comparisons were made for all commodity aggregates during odd years from 1969 to 1979. The cubic polynomials nearly always fit better than the Working model, which is not surprising since they have an additional parameter. On the other hand cubic polynomial fits still fall outside the 95% nonparametric confidence bands in several cases. Kneip (1991) has shown that the generalization of the Working model in which the budget shares are fourth degree polynomials in the logarithm of total expenditure cannot be rejected at standard significance levels.

The qualitative differences between the parametric and nonparametric estimates persist over time, as we see for example by comparing the fuel estimate for 1969 in Figure 4d with that for 1971. We consider the evolution of the estimated Engel curves more systematically below.

4.2 Invariance and Nonparametric Estimates Over Time

In this section we describe the evolution of cross section Engel curves estimated with the quartic kernel. The main conclusion is that the real Engel curves shift over time while the mean normalized curves (with x and y coordinates divided by their sample means) are much more stable.

We use the terms "budget" and "total expenditure" interchangeably. In order to test the hypothesis that Engel curves do not shift over time one must specify more precisely which curves are to be compared. It seems natural to decompose the effects of changes in budgets and relative prices (represented by changes in the population density z) into effects of proportional changes in prices and budgets, and effects of deviations from proportionality. The effect of a proportional change in prices and budgets can be formalized by letting $z_t(a|x,p)$ be homogeneous of degree zero in x and p. This assumption is required in order for the analysis to be consistent with the standard model of a private ownership economy in which income from profits and initial endowments rises by the factor γ if all prices rise by that factor. Homogeneity of $z_t(a|\cdot)$ and of the individual demands q^a implies that the cross section function

$$q_t(x,p) = \int q^a(x,p) z_t(a|x,p)da$$

is homogeneous of degree zero. The meaning of Engel curve invariance with respect to changes in the distribution of income is then clear as long as relative prices are fixed. For example, if the price index for every good in period t is λ_t using period 0 as base, then $p_t = \lambda_t p_0$ and the cross section Engel function of period t evaluated at expenditure level $\lambda_t x$ is $q_t(\lambda_t x, p_t) = q_t(x, p_0)$. Invariance of the real cross section Engel curve with respect to the actual changes in income requires that the right-hand side equal $q_0(x, p_0)$. Thus when relative prices are constant, such invariance implies

$$q_t(\lambda_t x, p_t) = q_0(x, p_0), \tag{3.1}$$

and it can be tested by comparing the two sides of this equation. If relative prices or income shares change, the invariance hypothesis of Theorem 2.1 no longer implies (3.1). However one can still ask whether the difference between the left and right-hand sides of (3.1) can plausibly be explained by the changes in relative prices and income shares. We will refer to (3.1) as expressing *real* invariance.

Kernel estimates of the real Engel curves $q_t^j(\lambda_t x, p_t)$ for food, clothing, housing and fuel, 1969-77, are plotted in Figures 5-8. There is considerable variation over time, however relative prices also changed during the period. Table 1 shows the values of relative price indices p_t^j/λ_t for odd numbered years, where λ_t is a consumer price index with base year 1969. The second column contains an index of the mean real household budget: $M_t/(\lambda_t M_0)$ where M_t is the mean budget in year t, and $t = 0$ denotes the base year. The price indices were taken from Employment Gazette(1982). The direct impact of the 1973-74 rise in real energy prices is reflected mainly in the price indices for fuel and transport. It is notable however that of all the aggregates, food experienced the most rapid price inflation during the period.

Table 1

Mean real total expenditure and relative price indices
(1969, 71, 73, 75, 77, FES)

		Good						
Year	$M_t/(\lambda_t M_0)$	**HOU**	**FUE**	**FOO**	**CLO**	**DUR**	**TRA**	**SER**
1969	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
1971	1.0092	1.0088	1.0032	1.0205	0.9650	0.9833	1.0207	1.0226
1973	1.0983	1.0650	0.9506	1.0930	0.9681	0.9235	0.9784	1.0435
1975	1.0543	0.9793	1.0280	1.1239	0.9068	0.8945	1.0359	1.0307
1977	1.0277	0.9352	1.0915	1.1884	0.8409	0.8423	1.0146	0.9764

The real Engel curve for food in Figure 5 generally shifts down over time. It is tempting to interpret the shift as the result of the rapid food price inflation. However, the percentage difference between the curves of different years is very close to the corresponding percentage

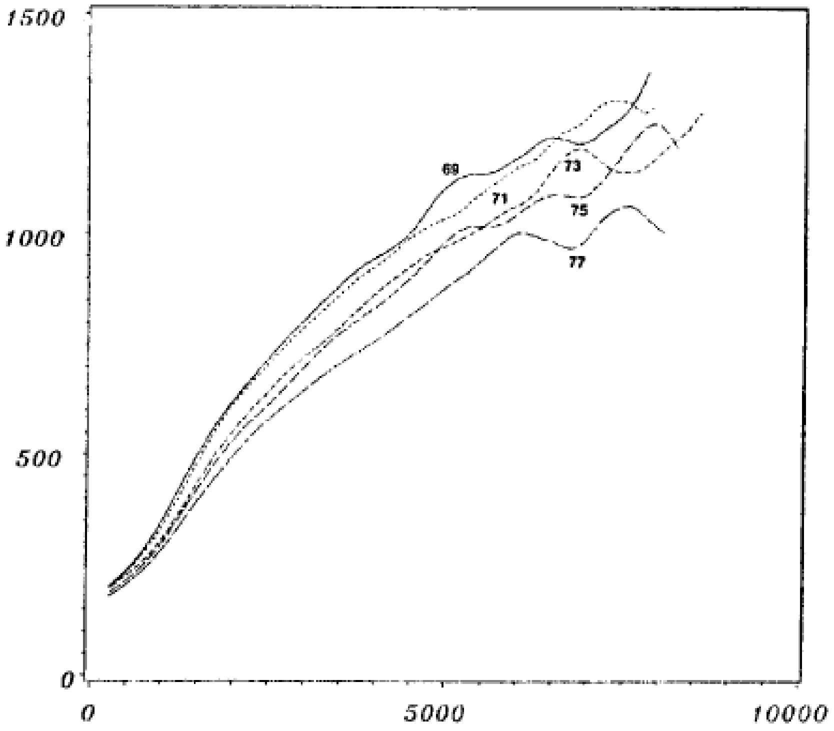


Figure 5: Real Engel curves for food 1969, 71, 73, 75, 77, FES

change in the relative food price. Thus if relative prices are responsible for the differences among the curves and if cross price effects on the demand for food are small or cancel each other out, then the own-price elasticity for food is approximately -1. This seems quite large. It suggests that implausibly large price and income effects might be required for the data to be consistent with the real invariance hypothesis in Theorem 2.1.

Changes in the real Engel curve for a good cannot be explained by contemporaneous changes in the relative price of that good alone. For example the relative price of clothing hardly changed from 1971 to 1973. Yet in Figure 6 we see that the real Engel curve for clothing dropped significantly (by over 17%) at the real total expenditure level 3500, roughly 1.3 times the 1969 mean total expenditure.

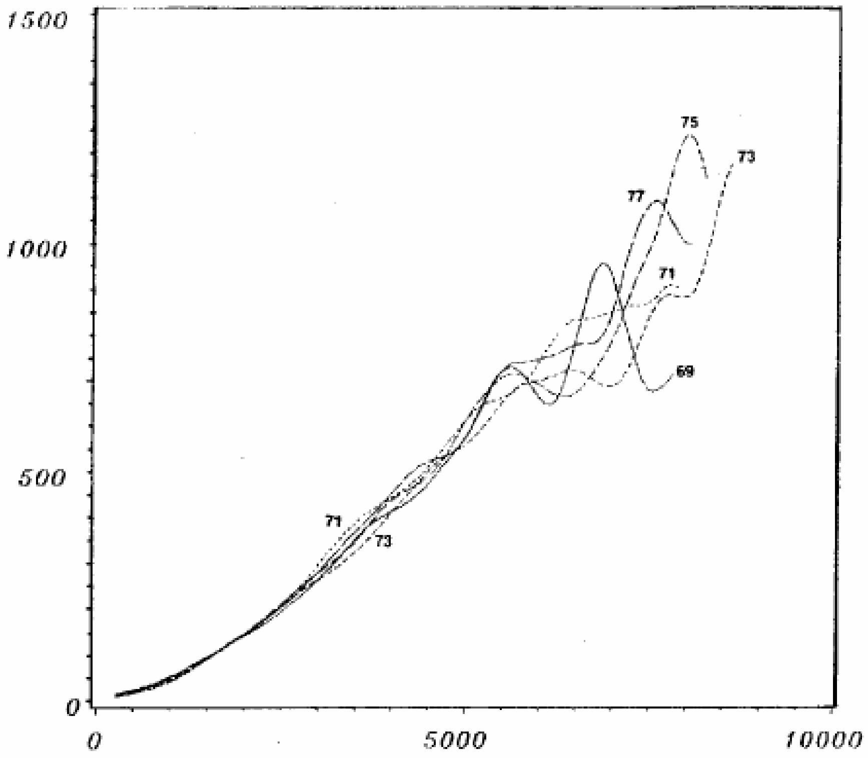


Figure 6: Real Engel curves for clothing 1969, 71, 73, 75, 77, FES

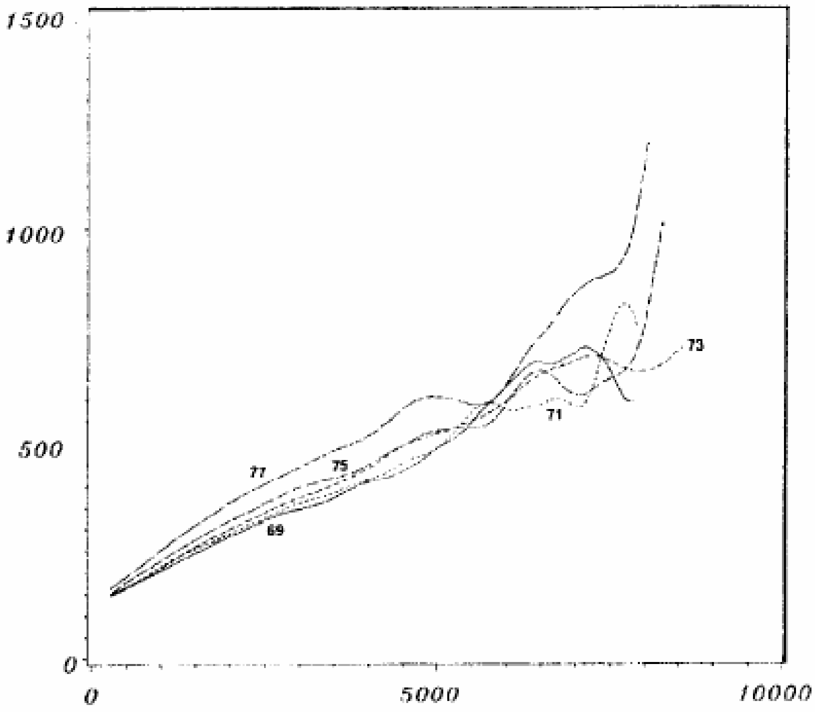


Figure 7: Figure 7. Real Engel curves for housing 1969, 71, 73, 75, 77, FES

The real Engel curve for housing (Figure 7) shifted steadily upward during the period, whereas the relative cost of housing rose from 1969 to 1973 then fell from 1973 to 1977. The real Engel curve for fuel (Figure 8) shifted down from 1973 to 1975 as the relative price rose. But from 1975 to 1977 the curve shifted upward over much of its range while the relative fuel price rose by another 6 percent.

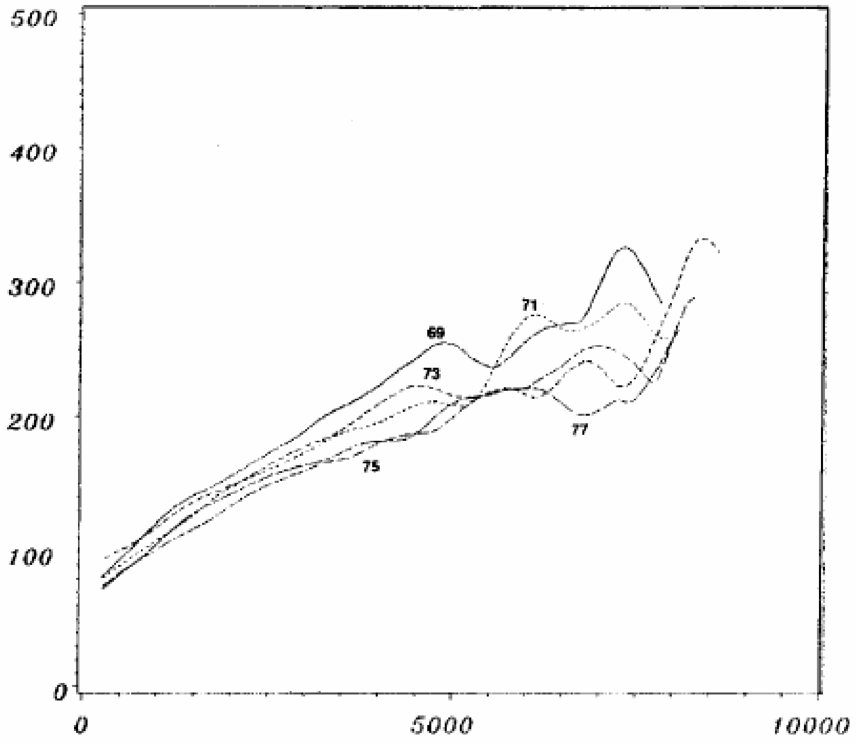


Figure 8: Real Engel curves for fuel 1969, 71, 73, 75, 77, FES

Examination of the evolution of Engel curves in principle permits a much richer analysis of price and income effects than is possible in the traditional time series modelling of mean demand. In particular, it is possible to study the way price changes affect different income classes. Such a study provides a middle ground between aggregate time series analysis, in which distributional issues cannot be raised, and a completely disaggregated study in which it can be difficult to develop intuition about the way conclusions depend on the choice of parametric models (see for example Blundell, Pashardes and Weber, 1988). Figures 5-8 show clearly that real Engel curves shift during a period of changing relative prices. It is remarkable, however that the shifts are nearly monotonic in the sense that there are relatively few crossings of curves. A second point is that the shifts are small from 1969 to 1971 when relative price changes were small. This might not seem surprising. However, since the sample populations for the two years are different many factors could have accounted for larger shifts.

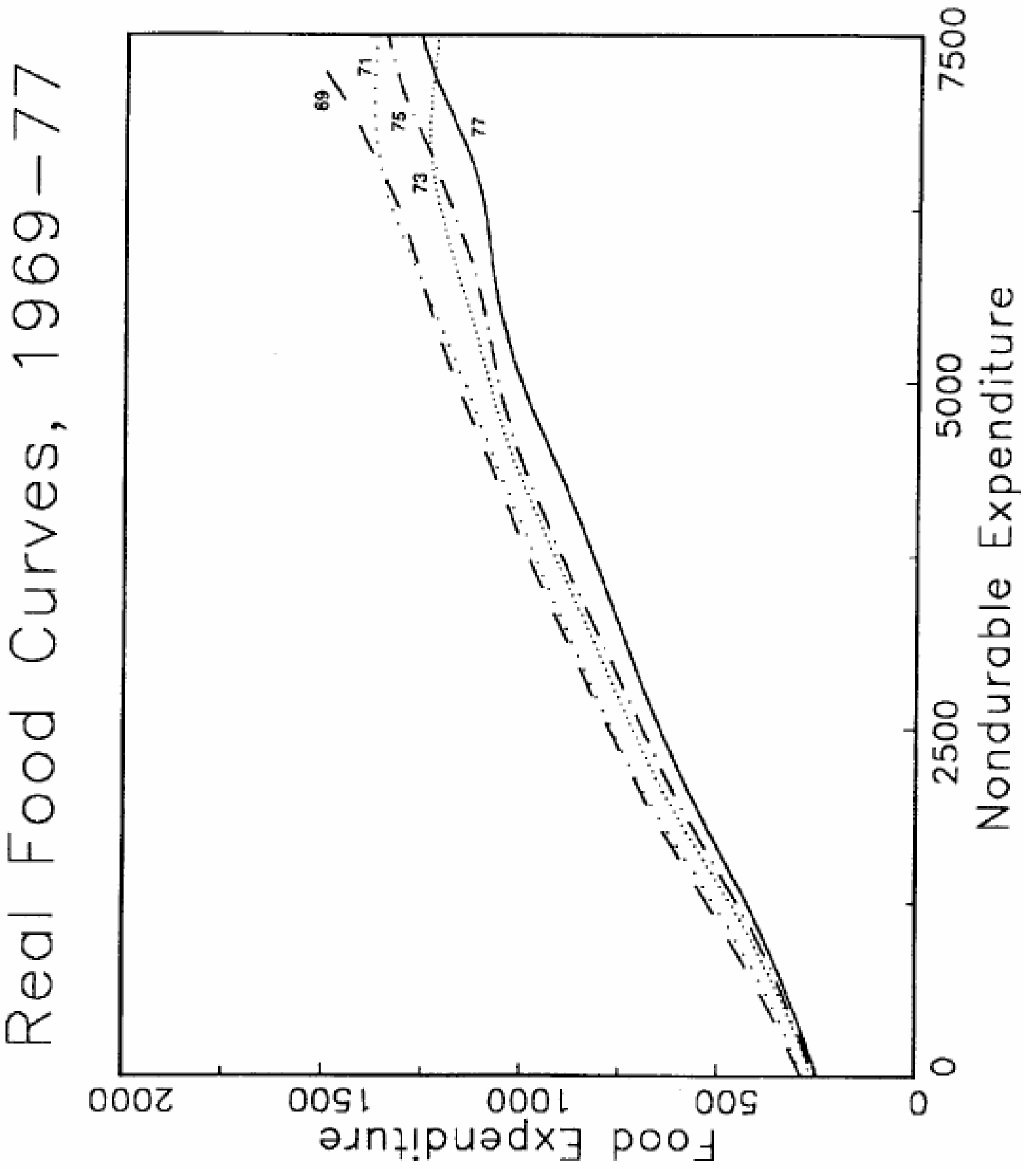


Figure 9: Real food Engel curves, 1969, 71, 73, 75, 77, FES

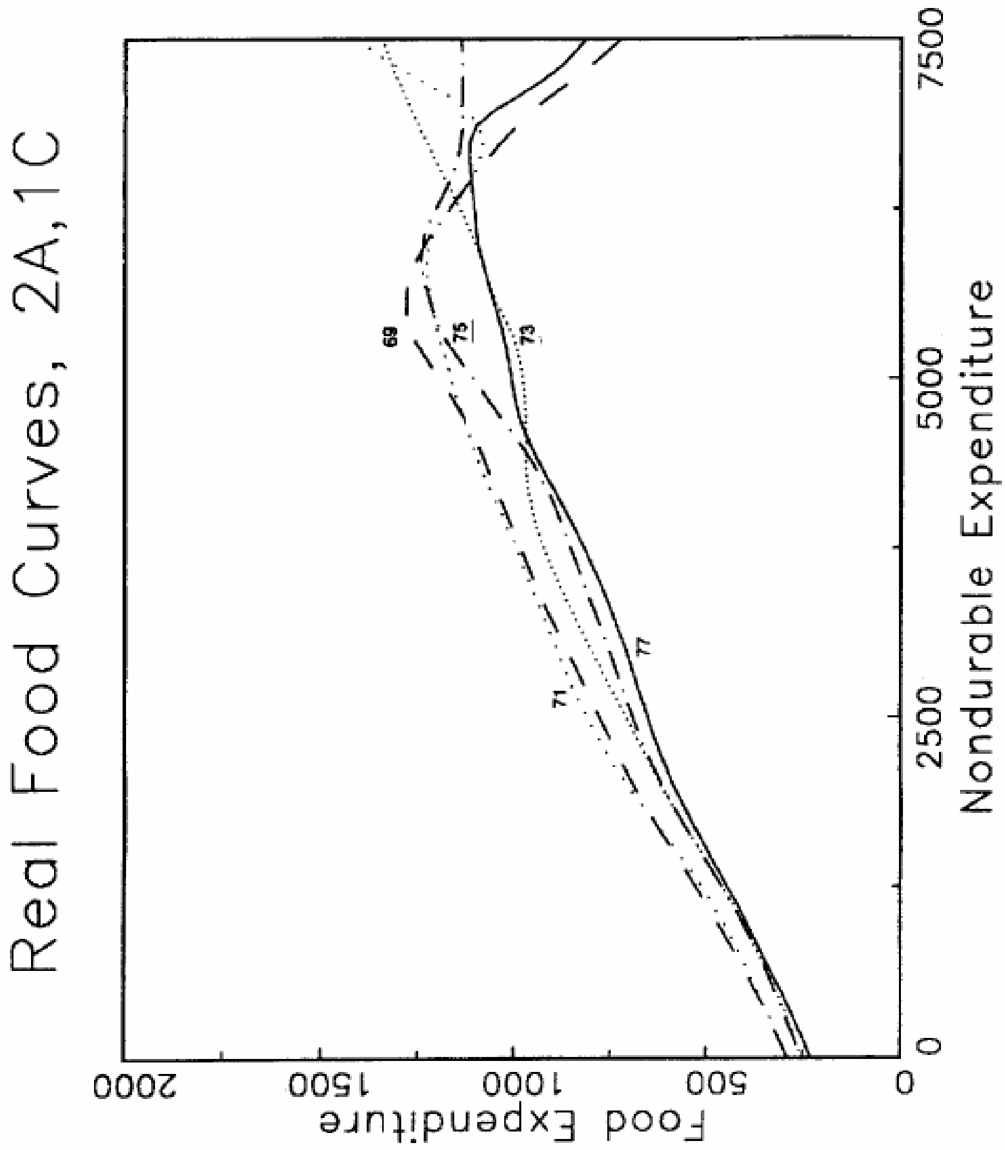


Figure 10: Real food Engel curves, 2 Adults, 1 child, FES

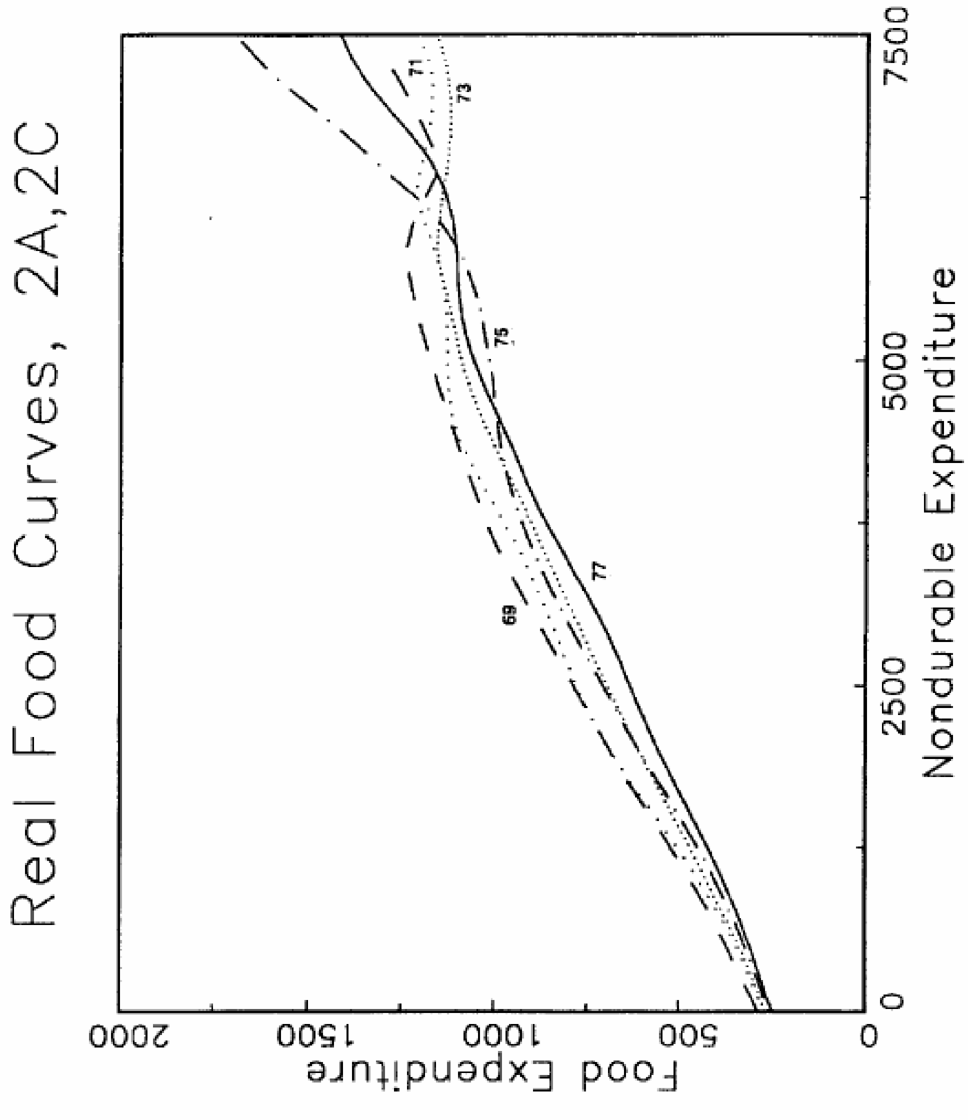
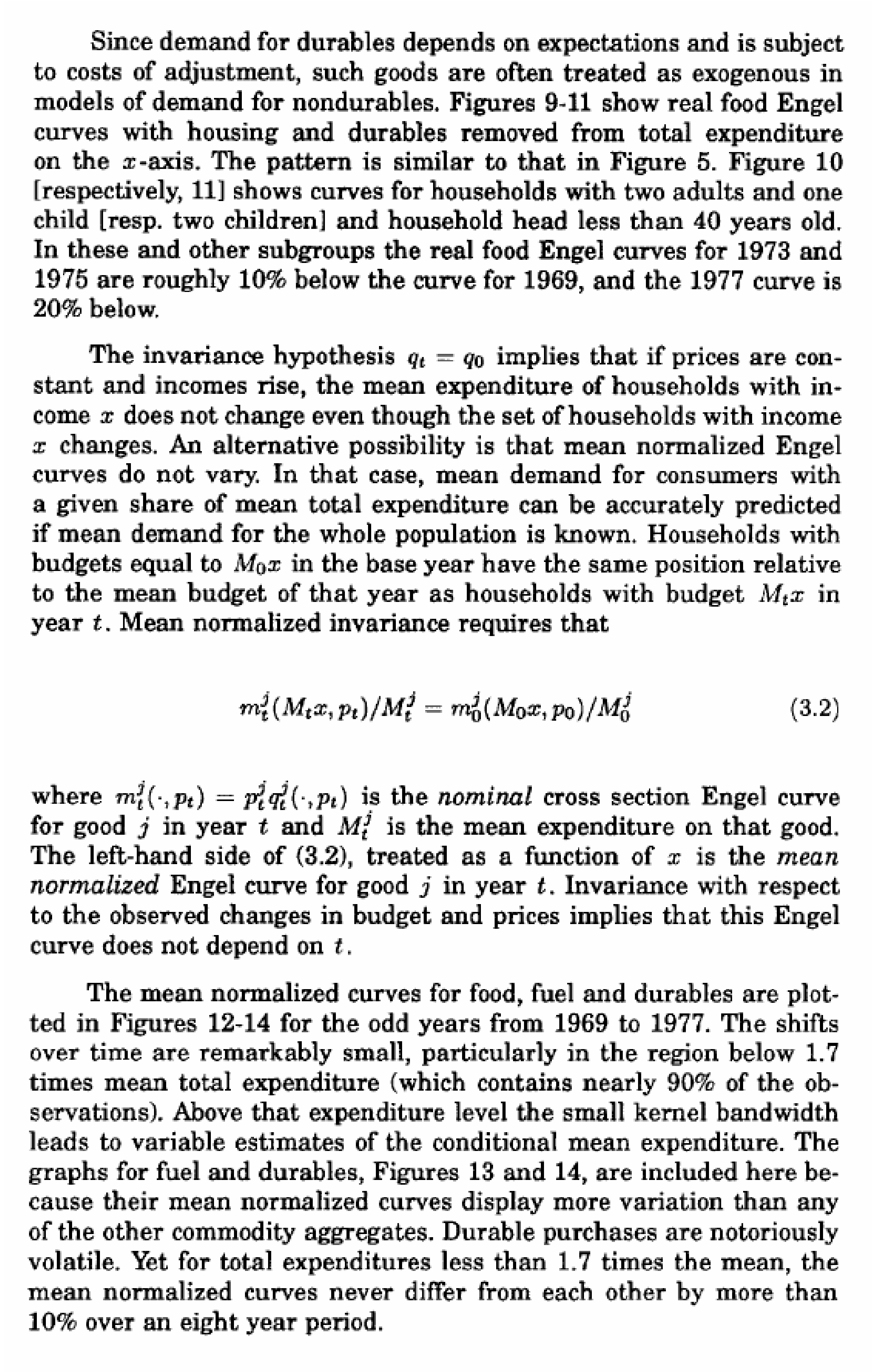


Figure 11: Real food Engel curves, 2 Adults, 2 children, FES



Since demand for durables depends on expectations and is subject to costs of adjustment, such goods are often treated as exogenous in models of demand for nondurables. Figures 9-11 show real food Engel curves with housing and durables removed from total expenditure on the x-axis. The pattern is similar to that in Figure 5. Figure 10 [respectively, 11] shows curves for households with two adults and one child [resp. two children] and household head less than 40 years old. In these and other subgroups the real food Engel curves for 1973 and 1975 are roughly 10% below the curve for 1969, and the 1977 curve is 20% below.

The invariance hypothesis $q_t = q_0$ implies that if prices are constant and incomes rise, the mean expenditure of households with income x does not change even though the set of households with income x changes. An alternative possibility is that mean normalized Engel curves do not vary. In that case, mean demand for consumers with a given share of mean total expenditure can be accurately predicted if mean demand for the whole population is known. Households with budgets equal to $M_0 x$ in the base year have the same position relative to the mean budget of that year as households with budget $M_t x$ in year t. Mean normalized invariance requires that

$$m_t^j(M_t x, p_t)/M_t^j = m_0^j(M_0 x, p_0)/M_0^j \tag{3.2}$$

where $m_t^j(\cdot, p_t) = p_t^j q_t^j(\cdot, p_t)$ is the *nominal* cross section Engel curve for good j in year t and M_t^j is the mean expenditure on that good. The left-hand side of (3.2), treated as a function of x is the *mean normalized* Engel curve for good j in year t. Invariance with respect to the observed changes in budget and prices implies that this Engel curve does not depend on t.

The mean normalized curves for food, fuel and durables are plotted in Figures 12-14 for the odd years from 1969 to 1977. The shifts over time are remarkably small, particularly in the region below 1.7 times mean total expenditure (which contains nearly 90% of the observations). Above that expenditure level the small kernel bandwidth leads to variable estimates of the conditional mean expenditure. The graphs for fuel and durables, Figures 13 and 14, are included here because their mean normalized curves display more variation than any of the other commodity aggregates. Durable purchases are notoriously volatile. Yet for total expenditures less than 1.7 times the mean, the mean normalized curves never differ from each other by more than 10% over an eight year period.

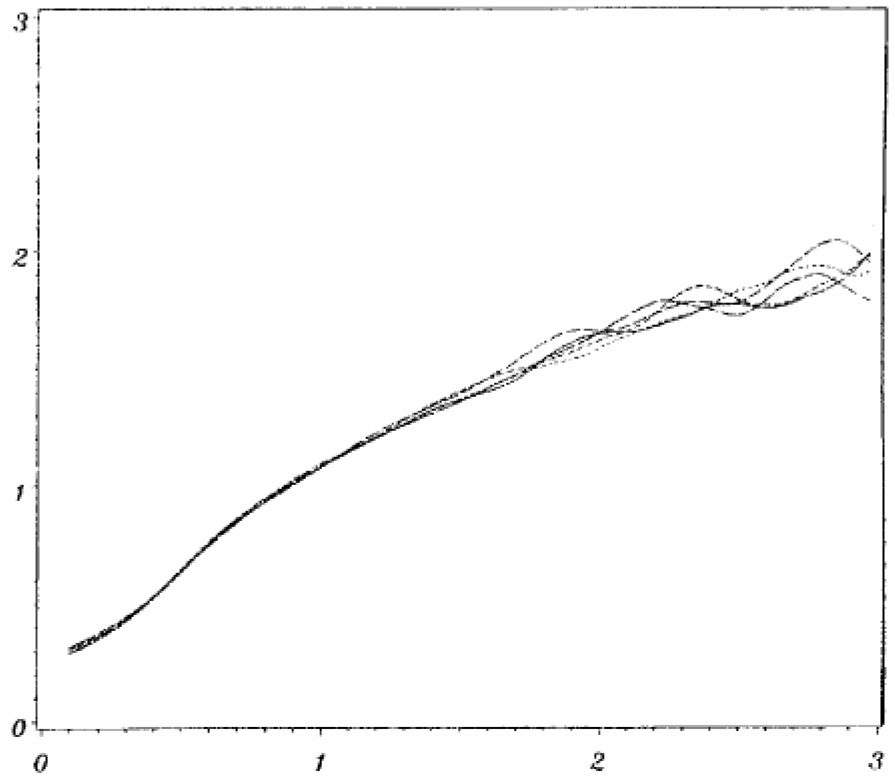


Figure 12: Mean normalized Engel curves for food 1969, 71, 73, 75, 77, FES

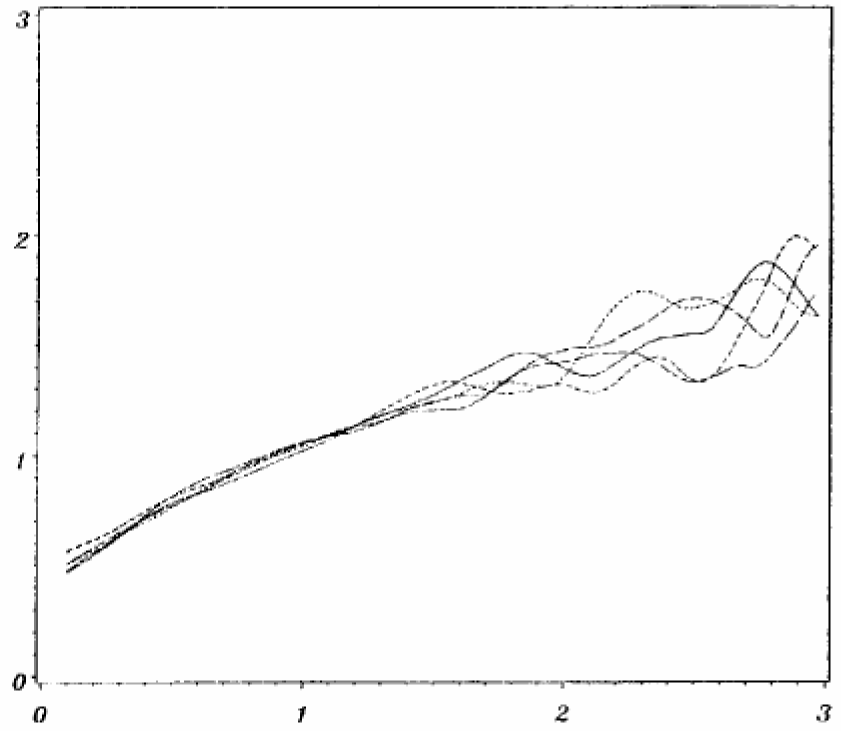


Figure 13: Mean normalized Engel curves for fuel 1969, 71, 73, 75, 77, FES

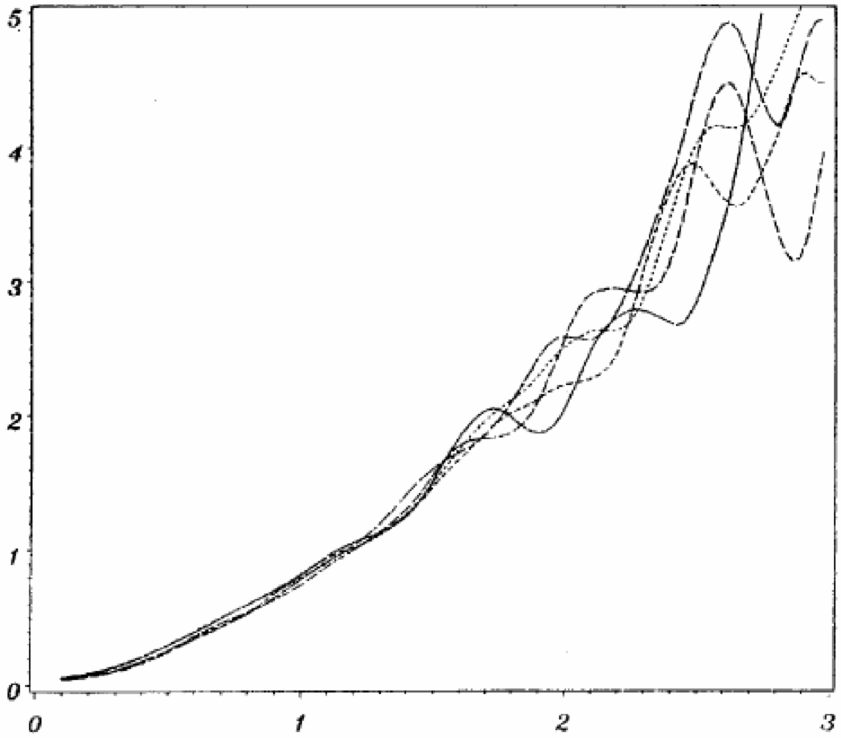


Figure 14
Mean normalized Engel curves for durables 1969, 71, 73, 75, 77, FES

The approximate invariance of the mean normalized Engel curves shows that the real Engel curve shapes are very stable during the sample period even though their positions change. This is an important result if the curves are to be used to classify goods as necessities or luxuries for different income classes. If household budget levels were nearly constant throughout the sample period and relative prices changed, then mean normalized invariance would imply that the conditional mean price elasticity of demand for households with a given budget level x is independent of x, i.e. that average price elasticities are the same throughout the budget distribution (see Jerison, 1992a).

Many other economically plausible invariance hypotheses can be formulated. For example the mean budget share of a group of households for a good might depend on their total expenditure relative to the mean total expenditure in the population. More precisely, the mean budget share $m_t^j(x_t, p_t)/x_t$ of households with total expenditure x_t in year t would equal the mean budget share for households with total expenditure x_0 in year 0, where $x_t/M_t = x_0/M_0$. This *budget share* invariance implies $\mu_t(x) \equiv m_t^j(M_t x, p_t)/M_t = m_0^j(M_0 x, p_0)/M_0$, and can be tested by comparing the functions μ_t for different years.

In the diagrams discussed so far, the competing invariance hypotheses cannot be compared since the axes are not scaled uniformly. Uniform scaling can be achieved by comparing the nominal Engel curve in year t with the base year Engel curve transformed according to an invariance hypothesis. Figures 15-17 show food Engel curves for 1971, 1973 and 1975 along with their confidence bands and the following transformations of the 1969 curve:

(a) $q_0^j(x/\lambda_t)$,

(b) $(M_t p_0^j/M_0 p_t^j)q_0^j(M_0 x/M_t)$,

(c) $(M_t^j p_0^j/M_0^j p_t^j)q_0^j(M_0 x/M_t)$,

where M_t^j and M_t are respectively mean expenditure on good j and mean total expenditure in year t. The curve (a) must be close to the year t Engel curve if the real Engel curve did not shift from the base period 0 (taken to be 1969) to period t, i.e. under real invariance. Under budget share invariance [resp. mean normalized invariance]. The curve (b) [resp. (c)] must be close to the year t Engel curve. Since the mean value of the (c) curve (weighted by the population density) is mean expenditure on good j, the same as the mean of the year t Engel curve, the (c) curve cannot be everywhere above or everywhere below the Engel curve. On the other hand, there is no further theoretical restriction implying that these curves must be close to each other. For example they could cross each other many times and differ greatly under the sup norm.

The 1969 Engel curve transformed according to the budget share (b) and mean normalized (c) invariance hypotheses fits the quantity food Engel curves of other years more closely than do the real 1969 Engel curves transformed according to (a). In 1971 the (c) curve lies inside the confidence bands, so invariance with respect to the observed changes in prices and incomes cannot be rejected. Similarly, the hypothesis is not rejected in 1973. In 1975 the (c) curve falls slightly below the confidence band at a total expenditure level of 9,000 pence per week. Note however that the (c) curve never diverges by more than 5% from the kernel estimate of the true Engel curve at total expenditure levels below 2.5 times the mean. The curves can be expected to diverge at high levels of total expenditure where the population density is low. In this region the bandwidth, which is constant over the entire domain, is lower than optimal, and the variance of the estimated conditional mean is higher than optimal. The transformation (b), based on budget share invariance, fits the nominal 1971 Engel curve better than (a) but not as well as (c).

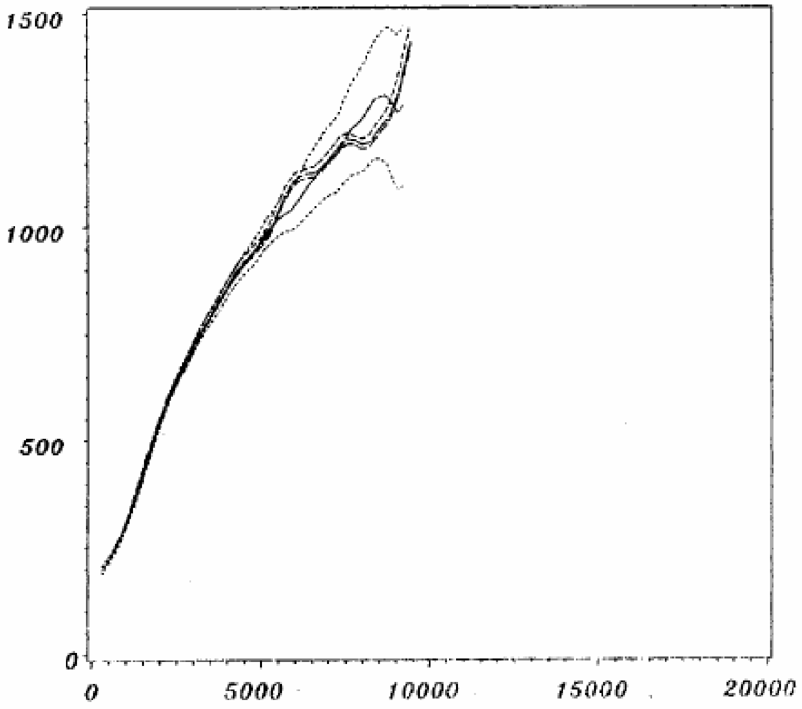


Figure 15: Nominal 1971 food Engel curve and transformed 1969 curves, FES

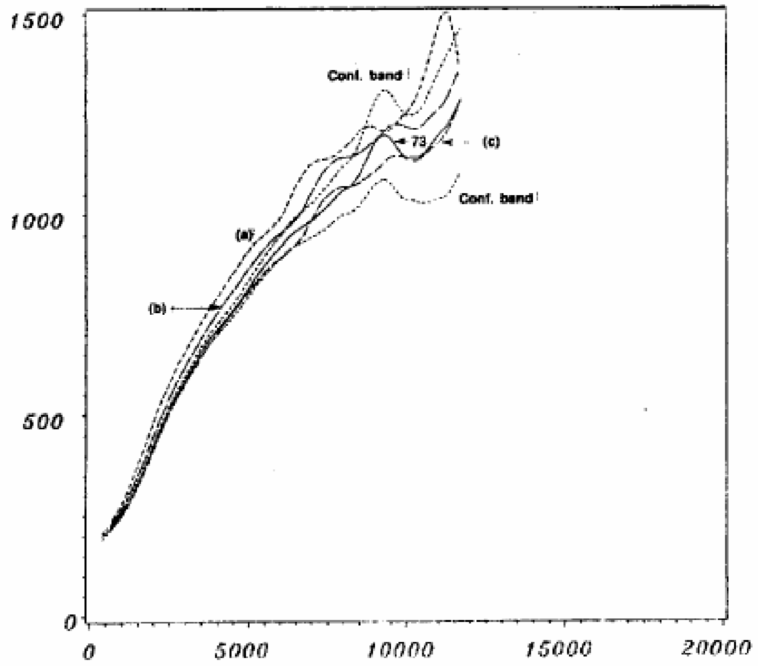


Figure 16: Nominal 1973 food Engel curve and transformed 1969 curves, FES

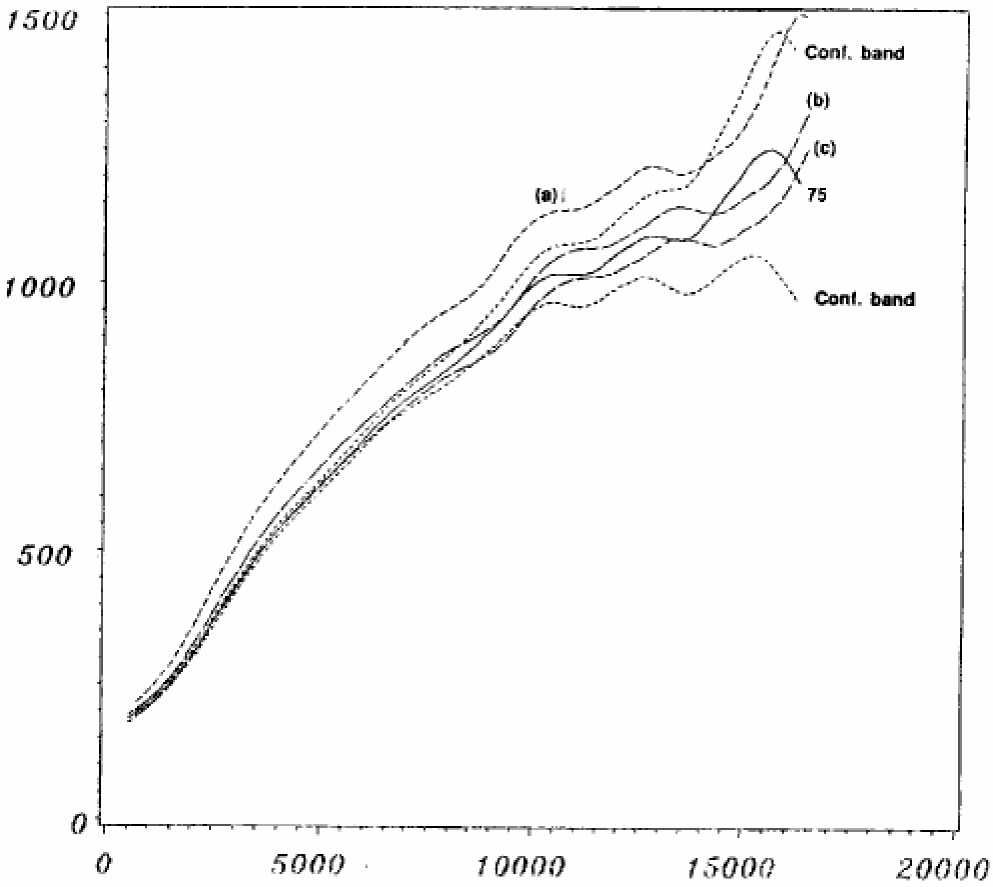


Figure 17

Figure 17. Nominal 1975 food Engel curve and transformed 1969 curves, FES

The confidence bands in Figures 15-17 are 95% bands around the untransformed Engel curves. However they must be interpreted as 87% bands in a comparison of the conditional means of two curves. This is because both curves are estimated. Since the distributions from the two years are independent but the densities of the real budgets are essentially the same, the asymptotic variance of the difference between the conditional means is twice that of a single mean.

Figures 18 and 19 show similar comparisons for housing and fuel, where transformed 1969 Engel curves are again graphed with the quantity Engel curve of 1971. In each case mean normalized invariance leads to a better prediction of the 1971 curve than does the real Engel curve invariance hypothesis. The real Engel curve transformation (a) leaves the confidence bands over a large region for each commodity aggregate.

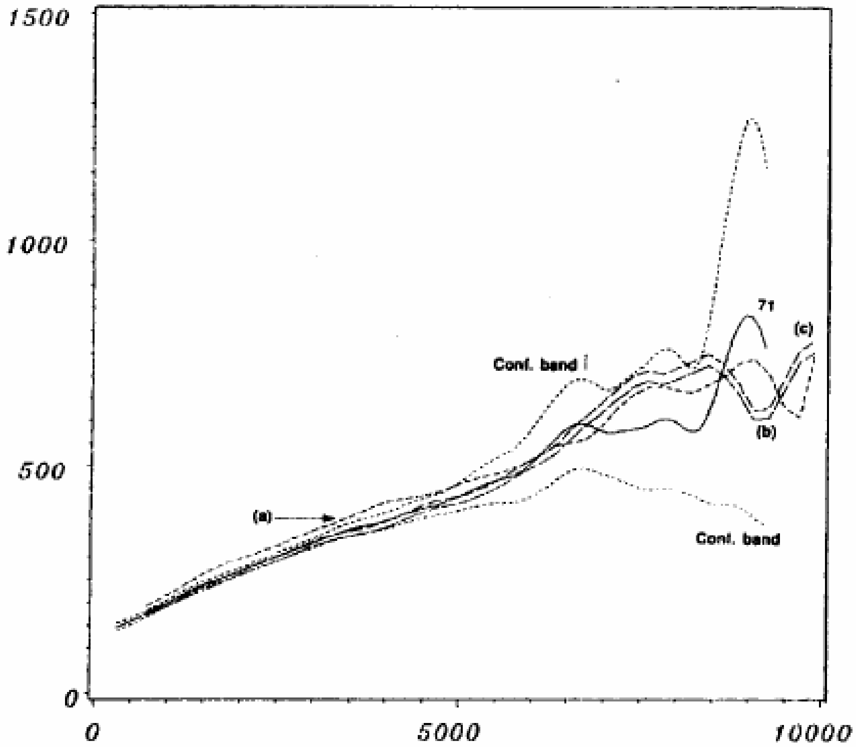


Figure 18
Nominal 1971 Engel curve for housing and transformed 1969 curves, FES

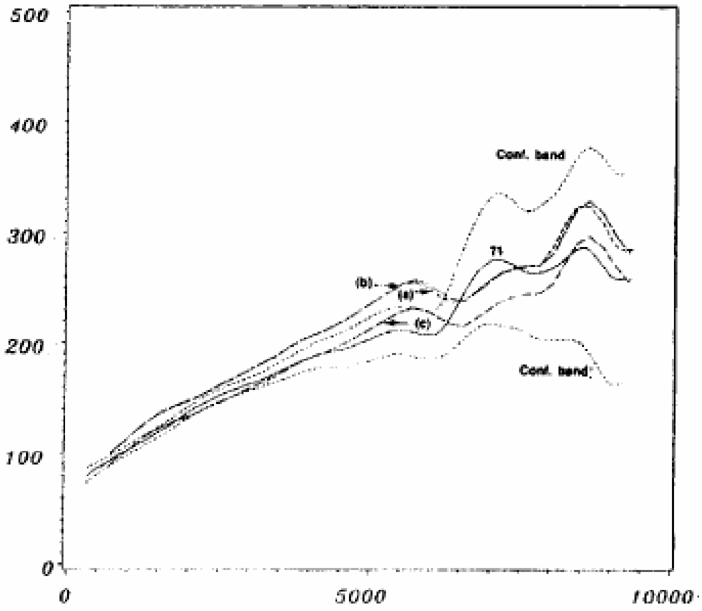


Figure 19
Nominal 1971 Engel curve for fuel and transformed 1969 curves, FES

4.3 The Working Model over Time

Finally we have compared Working model estimates of real Engel curves from different periods. The Working model for year t relates real expenditure Y_t^j/p_t^j on good j to real total expenditure X_t/λ_t according to the relation

$$(Y_t^j/p_t^j)/(X_t/\lambda_t) = \alpha_t ln(X_t/\lambda_t) + \beta_t + \varepsilon_t \tag{4.2}$$

where Y_t^j and X_t are the random variables described in Section 3 and ε_t is a random variable with zero mean, uncorrelated with X_t. We wish to test the hypothesis that the real Engel curve does not shift from period 0 to period t. Since the samples for different periods are independent we can use a standard F test of the hypothesis $\alpha_0 = \alpha_t$ and $\beta_0 = \beta_t$ (see, e.g. Neter and Wasserman, 1974, Section 5.6). Let S_τ denote the sum of squared residuals from the OLS estimate of the model (4.2) for period τ with n_τ observations and $(n_\tau - 2)$ degrees of freedom. The combined sample from periods 0 and t is then used to estimate the model (4.2) restricted by the null hypothesis. This regression has $n \equiv n_0 + n_t - 2$ degrees of freedom and the resulting sum of squared errors is denoted S. The statistic

$$F_t^* = \frac{S - (S_0 + S_t)}{n - (n_0 + n_t - 4)} / (\frac{S_0 + S_t}{n_0 + n_t - 4})$$

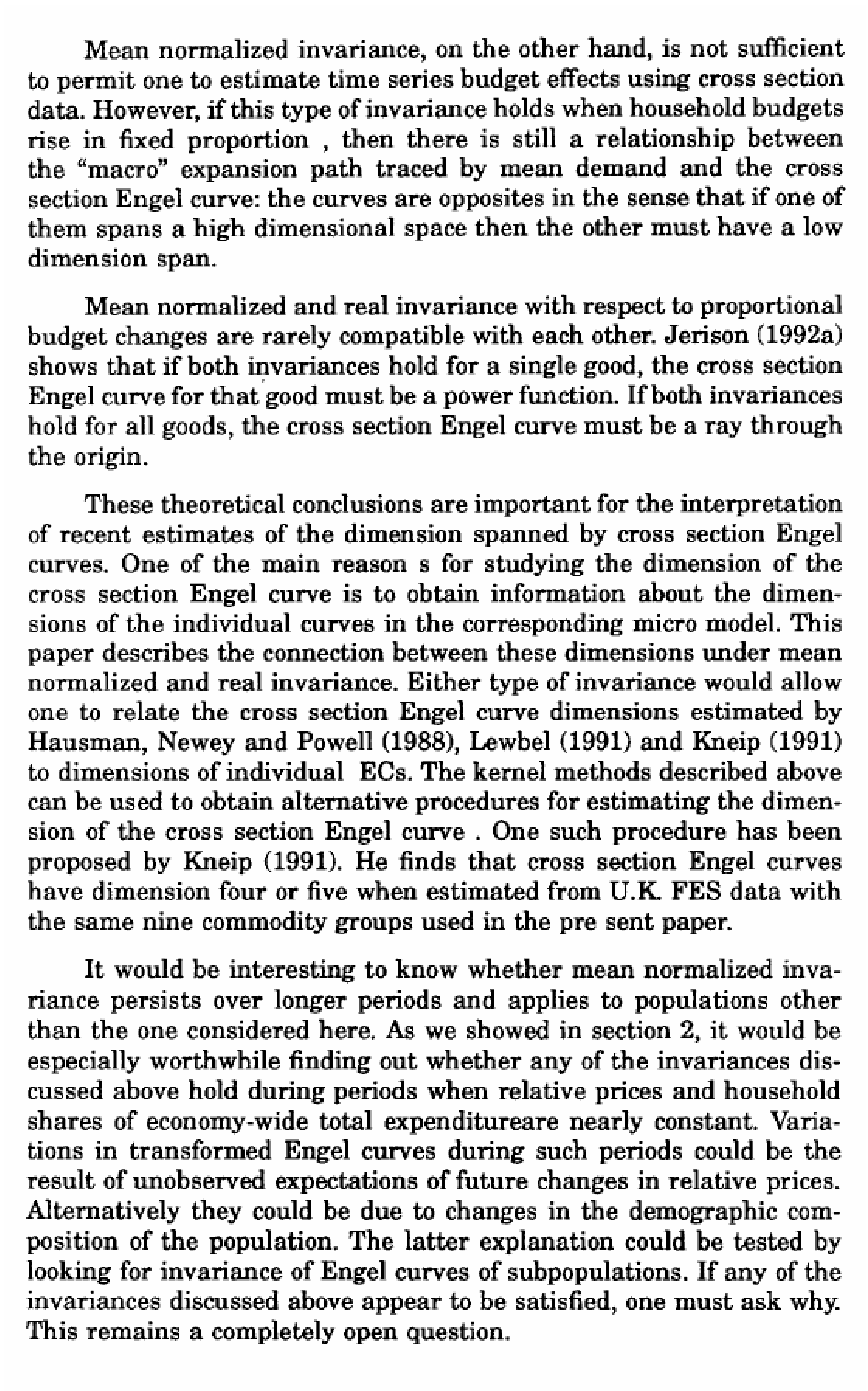
has an $F(2, n_0 + n_t - 4)$ distribution, and we reject the null hypothesis of Engel curve invariance at significance level α if $F_t^* > F(1-\alpha; 2, n_0 + n_t - 4)$. Since the sample sizes are close to 7000 in each period, the left hand side is approximately $F(1 - \alpha; 2, \infty)$ which equals 2.3 and 3.0 at confidence levels $1 - \alpha = .90$ and .95 respectively. The conclusions from these tests turn out to be generally similar to those based on comparison of the nonparametric estimates. For example, for $t = 1971$ and base period 1969, F_t^* takes the values 12.3, 3.1, 25.8 for food, housing and fuel respectively. This means that the hypothesis of invariance of the real Engel curves estimated by (4.2) is rejected at the 95% level in all three cases. The same conclusion holds for the nonparametric estimates since the 1969 curves transformed according to (a) leave the confidence bands in Figures 15, 18 and 19.

5 Conclusion

We have described nonparametric techniques for estimating and comparing cross section Engel curves. In applying the techniques to U.K. expenditure data we found that real Engel curves (with real total expenditure and quantity demanded on x and y axes) do change over time. The Engel curve for a commodity aggregate typically shifts in the direction opposite the change in the relative price index for that aggregate. However, this is not always the case, and the magnitude of the shift is sometimes too large to be interpreted as a contemporaneous "own price" effect. More work needs to be done to understand the effects of price and income changes on household demands. For example, commonly used parametric forms of price dependence imply invariance of the entire set of Engel curves under alternative transformations. Thus, invariance tests similar to the ones considered here could be applied to test these for standard forms of parametric price dependence. The advantage of the approach presented above is that it allows one to observe effects of price and in come changes on the mean demands of households throughout the income distribution.

The shapes of the nonparametrically estimated Engel curves are remarkably stable over time. This is reflected in the approximate mean normalized invariance found in section 4. It follows that in our sample the commodity aggregates can be classified as necessities or luxuries over intervals in the relative distribution of household budgets, and that the classifications are stable over time. Moreover, this result does not depend on restrictions on functional form. Mean normalized invariance requires that from one period to the next, the percentage change in the mean demand of households with a given relative share of economy-wide expenditure is independent of that share. Such independence could be treated as a standard for comparison. This would call attention to interesting situations in which the invariance fails. For example, when the relative price of a particular good rises, a temporary change in the shape of the mean normalized Engel curve for that good could indicate that households at different budget levels have different speeds of adjustment to the price change.

If real cross section Engel curves are invariant when the households' budgets change, then the time series effects of those changes can be estimated from cross section data. Furthermore the functional forms that fit the cross section Engel curve also fit the curves of the individual households and fit the "macro" expansion path that is followed by mean demand when the household budgets rise. The dimensions of the cross section and individual Engel curves and of the macro expansion path are the same, and this holds even when the individual Engel curves differ across households and the households' demand vectors collectively span the entire consumption space. In this case, the diversity of household behaviour does not enter the analysis of the effects of budget changes on mean demand. It is only the mean demand at each budget level (i.e. the cross section Engel curve) that matters, just as if the household s' demand functions were identical.

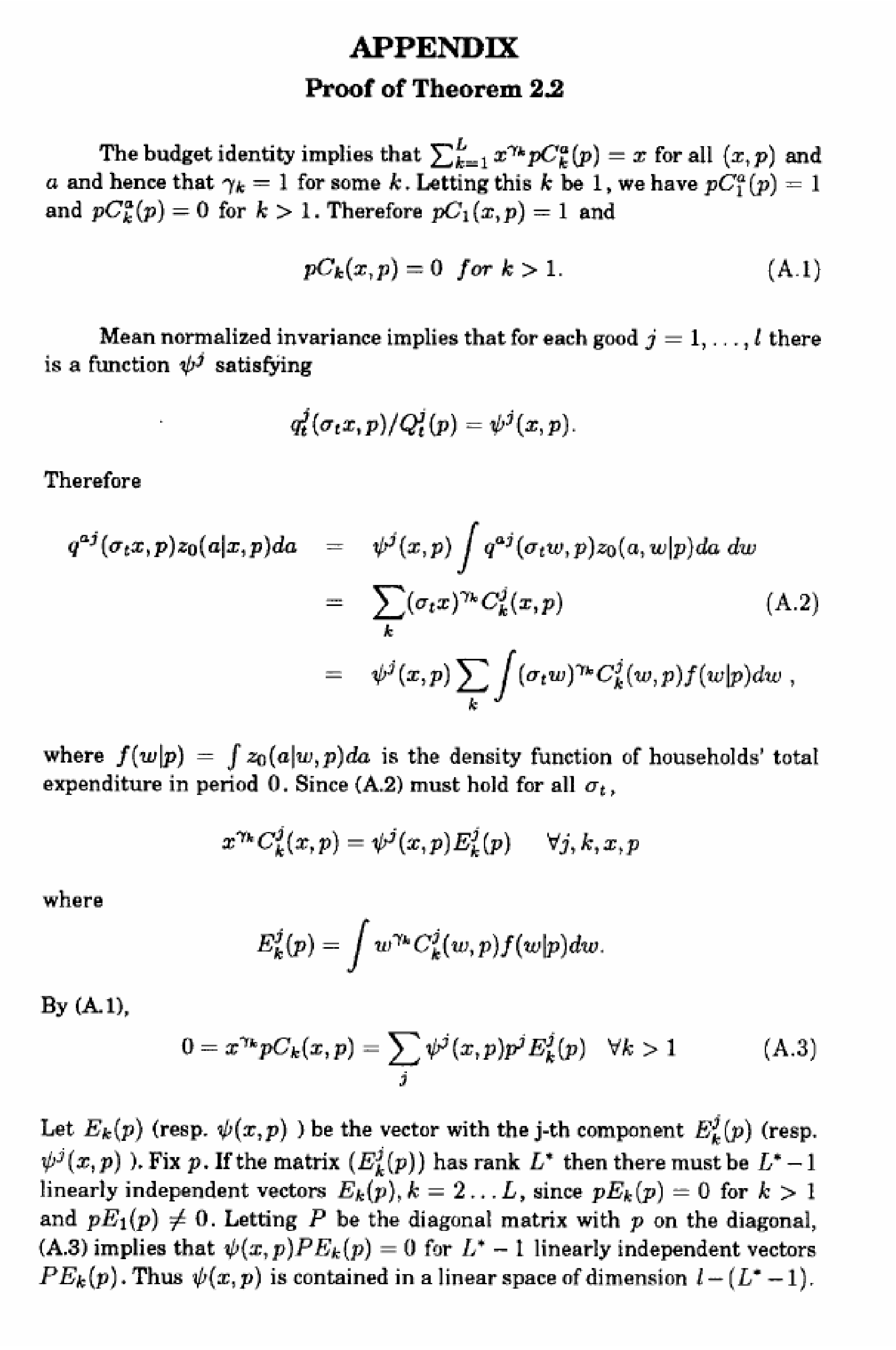


Mean normalized invariance, on the other hand, is not sufficient to permit one to estimate time series budget effects using cross section data. However, if this type of invariance holds when household budgets rise in fixed proportion , then there is still a relationship between the "macro" expansion path traced by mean demand and the cross section Engel curve: the curves are opposites in the sense that if one of them spans a high dimensional space then the other must have a low dimension span.

Mean normalized and real invariance with respect to proportional budget changes are rarely compatible with each other. Jerison (1992a) shows that if both invariances hold for a single good, the cross section Engel curve for that good must be a power function. If both invariances hold for all goods, the cross section Engel curve must be a ray through the origin.

These theoretical conclusions are important for the interpretation of recent estimates of the dimension spanned by cross section Engel curves. One of the main reason s for studying the dimension of the cross section Engel curve is to obtain information about the dimensions of the individual curves in the corresponding micro model. This paper describes the connection between these dimensions under mean normalized and real invariance. Either type of invariance would allow one to relate the cross section Engel curve dimensions estimated by Hausman, Newey and Powell (1988), Lewbel (1991) and Kneip (1991) to dimensions of individual ECs. The kernel methods described above can be used to obtain alternative procedures for estimating the dimension of the cross section Engel curve . One such procedure has been proposed by Kneip (1991). He finds that cross section Engel curves have dimension four or five when estimated from U.K. FES data with the same nine commodity groups used in the pre sent paper.

It would be interesting to know whether mean normalized invariance persists over longer periods and applies to populations other than the one considered here. As we showed in section 2, it would be especially worthwhile finding out whether any of the invariances discussed above hold during periods when relative prices and household shares of economy-wide total expenditureare nearly constant. Variations in transformed Engel curves during such periods could be the result of unobserved expectations of future changes in relative prices. Alternatively they could be due to changes in the demographic composition of the population. The latter explanation could be tested by looking for invariance of Engel curves of subpopulations. If any of the invariances discussed above appear to be satisfied, one must ask why. This remains a completely open question.



APPENDIX

Proof of Theorem 2.2

The budget identity implies that $\sum_{k=1}^{L} x^{\gamma_k} pC_k^a(p) = x$ for all (x,p) and a and hence that $\gamma_k = 1$ for some k. Letting this k be 1, we have $pC_1^a(p) = 1$ and $pC_k^a(p) = 0$ for $k > 1$. Therefore $pC_1(x,p) = 1$ and

$$pC_k(x,p) = 0 \quad for \; k > 1. \tag{A.1}$$

Mean normalized invariance implies that for each good $j = 1, \ldots, l$ there is a function ψ^j satisfying

$$q_t^j(\sigma_t x, p)/Q_t^j(p) = \psi^j(x,p).$$

Therefore

$$\begin{aligned} q^{aj}(\sigma_t x, p) z_0(a|x,p) da &= \psi^j(x,p) \int q^{aj}(\sigma_t w, p) z_0(a, w|p) da \, dw \\ &= \sum_k (\sigma_t x)^{\gamma_k} C_k^j(x,p) \qquad\qquad (A.2) \\ &= \psi^j(x,p) \sum_k \int (\sigma_t w)^{\gamma_k} C_k^j(w,p) f(w|p) dw \, , \end{aligned}$$

where $f(w|p) = \int z_0(a|w,p)da$ is the density function of households' total expenditure in period 0. Since (A.2) must hold for all σ_t,

$$x^{\gamma_k} C_k^j(x,p) = \psi^j(x,p) E_k^j(p) \qquad \forall j, k, x, p$$

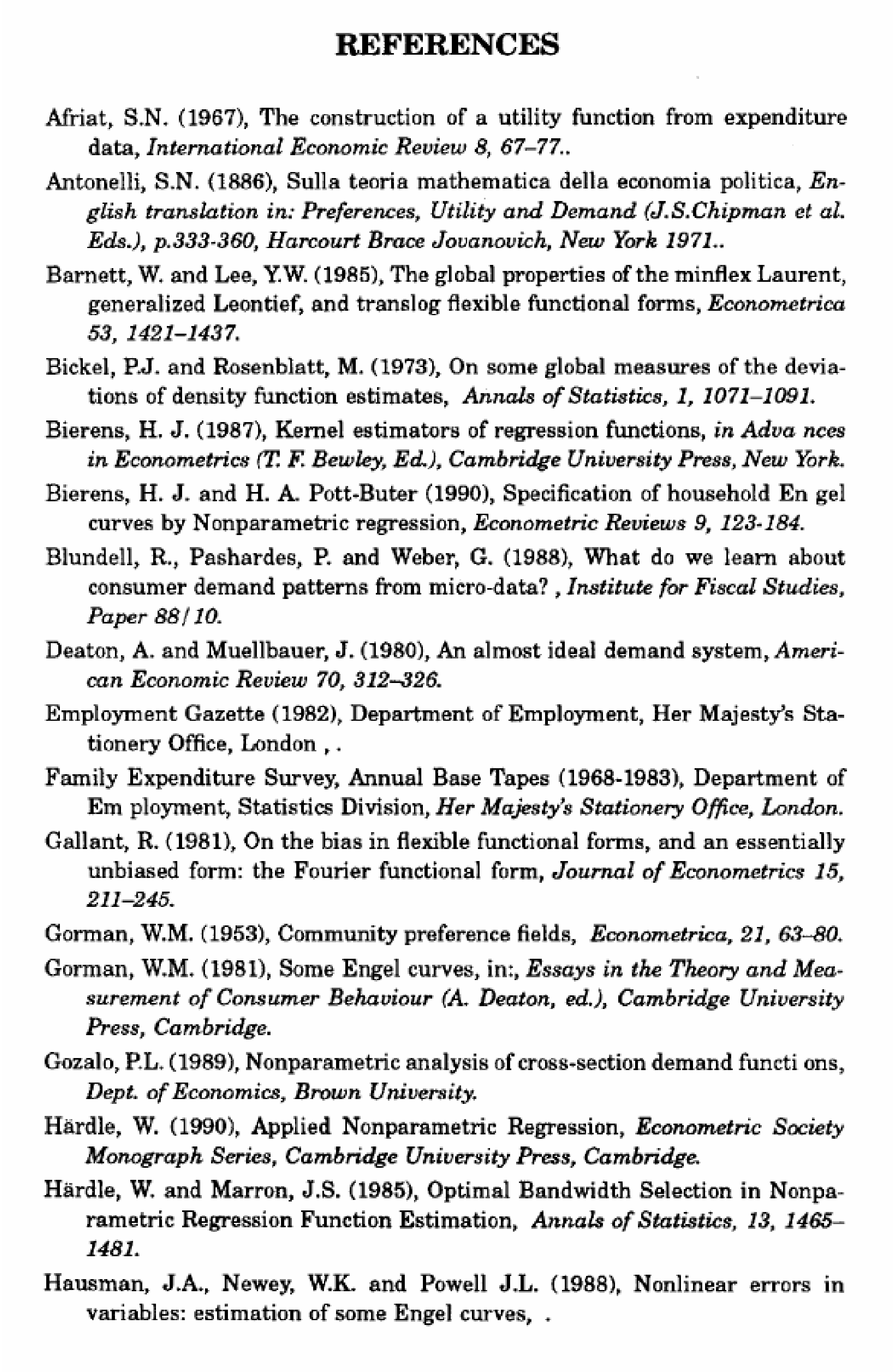
where

$$E_k^j(p) = \int w^{\gamma_k} C_k^j(w,p) f(w|p) dw.$$

By (A.1),

$$0 = x^{\gamma_k} pC_k(x,p) = \sum_j \psi^j(x,p) p^j E_k^j(p) \quad \forall k > 1 \tag{A.3}$$

Let $E_k(p)$ (resp. $\psi(x,p)$) be the vector with the j-th component $E_k^j(p)$ (resp. $\psi^j(x,p)$). Fix p. If the matrix $(E_k^j(p))$ has rank L^* then there must be $L^* - 1$ linearly independent vectors $E_k(p), k = 2 \ldots L$, since $pE_k(p) = 0$ for $k > 1$ and $pE_1(p) \neq 0$. Letting P be the diagonal matrix with p on the diagonal, (A.3) implies that $\psi(x,p) PE_k(p) = 0$ for $L^* - 1$ linearly independent vectors $PE_k(p)$. Thus $\psi(x,p)$ is contained in a linear space of dimension $l - (L^* - 1)$.



REFERENCES

Afriat, S.N. (1967), The construction of a utility function from expenditure data, *International Economic Review 8, 67–77..*

Antonelli, S.N. (1886), Sulla teoria mathematica della economia politica, *English translation in: Preferences, Utility and Demand (J.S.Chipman et al. Eds.), p.333-360, Harcourt Brace Jovanovich, New York 1971..*

Barnett, W. and Lee, Y.W. (1985), The global properties of the minflex Laurent, generalized Leontief, and translog flexible functional forms, *Econometrica 53, 1421–1437.*

Bickel, P.J. and Rosenblatt, M. (1973), On some global measures of the deviations of density function estimates, *Annals of Statistics, 1, 1071–1091.*

Bierens, H. J. (1987), Kernel estimators of regression functions, *in Adva nces in Econometrics (T. F. Bewley, Ed.), Cambridge University Press, New York.*

Bierens, H. J. and H. A. Pott-Buter (1990), Specification of household En gel curves by Nonparametric regression, *Econometric Reviews 9, 123-184.*

Blundell, R., Pashardes, P. and Weber, G. (1988), What do we learn about consumer demand patterns from micro-data? , *Institute for Fiscal Studies, Paper 88/10.*

Deaton, A. and Muellbauer, J. (1980), An almost ideal demand system, *American Economic Review 70, 312–326.*

Employment Gazette (1982), Department of Employment, Her Majesty's Stationery Office, London , .

Family Expenditure Survey, Annual Base Tapes (1968-1983), Department of Em ployment, Statistics Division, *Her Majesty's Stationery Office, London.*

Gallant, R. (1981), On the bias in flexible functional forms, and an essentially unbiased form: the Fourier functional form, *Journal of Econometrics 15, 211–245.*

Gorman, W.M. (1953), Community preference fields, *Econometrica, 21, 63–80.*

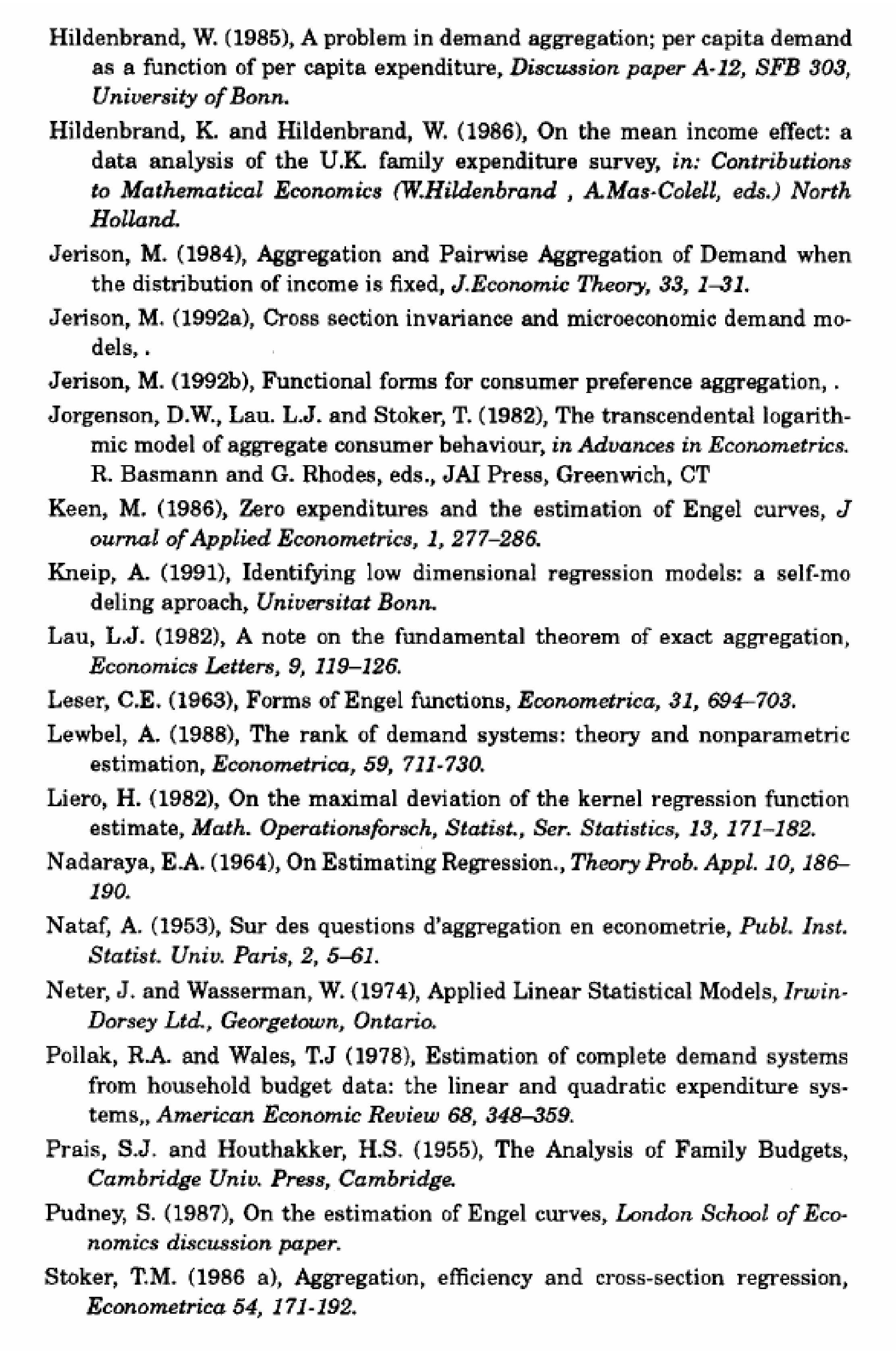
Gorman, W.M. (1981), Some Engel curves, in:, *Essays in the Theory and Measurement of Consumer Behaviour (A. Deaton, ed.), Cambridge University Press, Cambridge.*

Gozalo, P.L. (1989), Nonparametric analysis of cross-section demand functi ons, *Dept. of Economics, Brown University.*

Härdle, W. (1990), Applied Nonparametric Regression, *Econometric Society Monograph Series, Cambridge University Press, Cambridge.*

Härdle, W. and Marron, J.S. (1985), Optimal Bandwidth Selection in Nonparametric Regression Function Estimation, *Annals of Statistics, 13, 1465–1481.*

Hausman, J.A., Newey, W.K. and Powell J.L. (1988), Nonlinear errors in variables: estimation of some Engel curves, .



Hildenbrand, W. (1985), A problem in demand aggregation; per capita demand as a function of per capita expenditure, *Discussion paper A-12, SFB 303, University of Bonn.*

Hildenbrand, K. and Hildenbrand, W. (1986), On the mean income effect: a data analysis of the U.K. family expenditure survey, *in: Contributions to Mathematical Economics (W.Hildenbrand , A.Mas-Colell, eds.) North Holland.*

Jerison, M. (1984), Aggregation and Pairwise Aggregation of Demand when the distribution of income is fixed, *J.Economic Theory, 33, 1–31.*

Jerison, M. (1992a), Cross section invariance and microeconomic demand models, .

Jerison, M. (1992b), Functional forms for consumer preference aggregation, .

Jorgenson, D.W., Lau. L.J. and Stoker, T. (1982), The transcendental logarithmic model of aggregate consumer behaviour, *in Advances in Econometrics.* R. Basmann and G. Rhodes, eds., JAI Press, Greenwich, CT

Keen, M. (1986), Zero expenditures and the estimation of Engel curves, *Journal of Applied Econometrics, 1, 277–286.*

Kneip, A. (1991), Identifying low dimensional regression models: a self-modeling aproach, *Universitat Bonn.*

Lau, L.J. (1982), A note on the fundamental theorem of exact aggregation, *Economics Letters, 9, 119–126.*

Leser, C.E. (1963), Forms of Engel functions, *Econometrica, 31, 694–703.*

Lewbel, A. (1988), The rank of demand systems: theory and nonparametric estimation, *Econometrica, 59, 711-730.*

Liero, H. (1982), On the maximal deviation of the kernel regression function estimate, *Math. Operationsforsch, Statist., Ser. Statistics, 13, 171–182.*

Nadaraya, E.A. (1964), On Estimating Regression., *Theory Prob. Appl. 10, 186–190.*

Nataf, A. (1953), Sur des questions d'aggregation en econometrie, *Publ. Inst. Statist. Univ. Paris, 2, 5–61.*

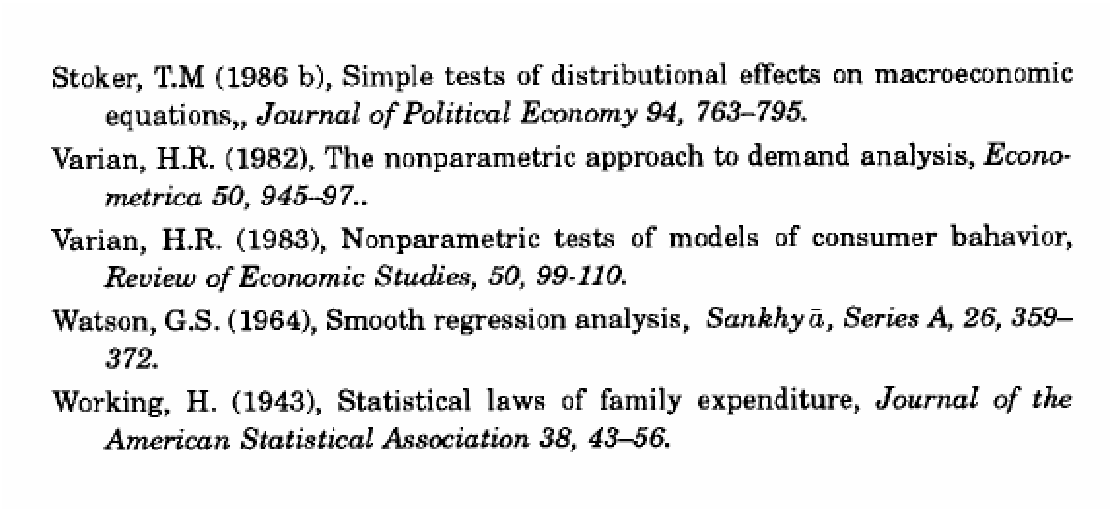
Neter, J. and Wasserman, W. (1974), Applied Linear Statistical Models, *Irwin-Dorsey Ltd., Georgetown, Ontario.*

Pollak, R.A. and Wales, T.J (1978), Estimation of complete demand systems from household budget data: the linear and quadratic expenditure systems,, *American Economic Review 68, 348–359.*

Prais, S.J. and Houthakker, H.S. (1955), The Analysis of Family Budgets, *Cambridge Univ. Press, Cambridge.*

Pudney, S. (1987), On the estimation of Engel curves, *London School of Economics discussion paper.*

Stoker, T.M. (1986 a), Aggregation, efficiency and cross-section regression, *Econometrica 54, 171-192.*



Stoker, T.M (1986 b), Simple tests of distributional effects on macroeconomic equations,, *Journal of Political Economy 94, 763–795.*

Varian, H.R. (1982), The nonparametric approach to demand analysis, *Econometrica 50, 945–97..*

Varian, H.R. (1983), Nonparametric tests of models of consumer bahavior, *Review of Economic Studies, 50, 99-110.*

Watson, G.S. (1964), Smooth regression analysis, *Sankhy ā, Series A, 26, 359–372.*

Working, H. (1943), Statistical laws of family expenditure, *Journal of the American Statistical Association 38, 43–56.*

KERNEL REGRESSION SMOOTHING OF TIME SERIES

By Wolfgang Härdle and Philippe Vieu

Université Catholique de Louvain and Université Paul Sabatier

First version received March 1990

Abstract. A class of non-parametric regression smoothers for times series is defined by the kernel method. The kernel approach allows flexible modelling of a time series without reference to a specific parametric class. The technique is applicable to detection of non-linear dependences in time series and to prediction in smooth regression models with serially correlated observations.

In practice these estimators are to be tuned by a *smoothing parameter*. A data-driven selector for this smoothing parameter is presented that asymptotically minimizes a squared error measure. We prove asymptotic optimality of this selector. We illustrate the technique with a simulated example and by constructing a smooth prediction curve for the variation of gold prices. In both cases the non-parametric method proves to be useful in uncovering non-linear structure.

Keywords. Prediction; time series analysis; autoregressive processes; data-driven bandwidth, kernel, α-mixing.

1 INTRODUCTION

Prediction of observations yet to be made is an important task in the statistical analysis of economic time series. By far the most common technique is to model the dependence of future observations on the past by a parametric class of functions. A typical example of such an approach is the Gaussian linear autoregressive scheme. The parametric model typically describes the whole distribution of the data, or the conditional distribution given exogenous variables. However, it may be able to describe only partial features of the time series. Research interest in recent years has therefore concentrated on non-parametric or semi-parametric analysis. Robinson (1988) surveyed semi-parametric methods. A striking example of a non-linear time series is given by Freedman *et al*. (1988, p. 14).

In the present paper we investigate a model of the form $Y = r(X) + \varepsilon$ where $r(X)$ is an unspecified function not restricted in its functional form. In the context of prediction X may denote a lag 1 observation of Y and a future observation would be predicted by $r(x)$. The function r can be estimated by a variety of techniques, e.g. running medians, splines or orthogonal polynomials. In this paper, we concentrate on the conceptually simple kernel smoothers (Robinson, 1983; Collomb, 1984; Bierens, 1985).

For a bivariate time series $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n), \ldots \in \mathbb{R}^2$, a *kernel smoother* is defined as (assuming that $0/0 = 0$)

0143-9782/92/03 209–24 \$02.50/0
JOURNAL OF TIME SERIES ANALYSIS Vol. 13, No. 3

$$\hat{r}_{h_n}(x) = n^{-1}h_n^{-1}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right) Y_i \Big/ n^{-1}h_n^{-1}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right) \tag{1.1}$$

where K is the *kernel* and h_n is a sequence of positive real numbers called the *bandwidth* sequence. This estimator was proposed by Nadaraya (1964) and Watson (1964); for an introduction into basic statistical properties of this estimator for independent and identically distributed (i.i.d.) observations, see Collomb (1981) or Härdle (1990, Chapter 3).

Convergence properties of this kernel smoother have been considered under certain dependence concepts for the process (X_i, Y_i) (see, for example, Collomb, 1984; Roussas, 1989; Truong and Stone, 1989; Györfi *et al.*, 1990). In this paper we assume that the observations are generated by a stationary α-mixing process (see Section 2), thus relaxing the mixing conditions usually made by Collomb. The concept of α-mixing extends the independence assumptions by allowing dependence among neighbouring observations with 'vanishing memory'.

The basic problem of applying this estimator to a given time series is the choice of the bandwidth h_n which we shall abbreviate from now on as h. The selection of this smoothing parameter decides the form of the predictor function $\hat{r}_h(x)$. A value of h which is too small will give predictions with too high a variance, i.e. it will result in undersmoothing. A value of h which is too large will lead to an oversmooth function with high bias. A mathematical quantification of these effects can be obtained by considering the average squared error

$$d_A(h) = n^{-1}\sum_{i=1}^{n} \{\hat{r}_h(X_i) - r(X_i)\}^2 w(X_i) \tag{1.2}$$

where $w(.)$ denotes a weight function. An estimator $\hat{r}_h$ that balances the trade-off between the squared bias and the variance component of $d_A(h)$ is certainly desirable. The main result of this paper is the construction of a data-driven bandwidth $\hat{h}$ that asymptotically minimizes the above measure of accuracy. More precisely, we construct a bandwidth selector that is *asymptotically optimal*, i.e.

$$\frac{d_A(\hat{h})}{\inf_{h\in H_n} d_A(h)} \xrightarrow{p} 1,$$

where H_n is a set of possible smoothing parameters. (For related results in the setting of independent data, see Shibata (1981) and Härdle and Marron (1985).)

The present α-mixing concept applies to the prediction of univariate time series $\{Z_i, i \geqslant 1\}$. If R denotes the autoregression function, so that

$$Z_n = R(Z_{n-1}) + \varepsilon_n, \tag{1.3}$$

the adaptive predictor $\hat{r}_h$ then provides smooth estimates of R.

Extensions of this technique to more than lag 1 predictions are possible but require more tedious calculations and run into the problem of sparsity of data. Therefore we restrict ourselves here to one-term prediction and delay multi-term prediction (using additive structure) to a future paper.

In Section 2 we discuss the dependence structure and define the bandwidth selector by cross-validation. Applications of the adaptive predictor are presented in Section 3. In particular, we examine the optimization method on a simulated data set and construct the adapted kernel smoother for a time series concerning the variation of gold prices (for a different approach see Frank and Stengos (1987)). In Section 4 we give conditions that ensure the asymptotic optimality of the algorithm. Section 5 contains the proofs. In the Appendix we give the detailed calculations that are used in the proofs.

2. ADAPTIVE SMOOTHING FOR TIME SERIES

Let $\{(X_i, Y_i): i \geqslant 1\} \subset \mathbb{R} \times \mathbb{R}$ be a two-dimensional time series. We assume throughout that this process is α-mixing. Rosenblatt (1956) defined this mixing condition as follows:

$$|P(A \cap B) - P(A)P(B)| \leqslant \alpha(k) \tag{C.1}$$

holds for any $n \in \mathbb{N}$ ($k \in \mathbb{N}$) and any set A (or B) which is $\sigma\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ (or $\sigma\{(X_{n+k}, Y_{n+k}), \ldots\}$) measurable, with the sequence of positive numbers $\{\alpha(n)\}$ being such that $\lim_{n\to\infty} \alpha(n) = 0$. This dependence structure is one of the least restrictive of the different mixing conditions discussed in the literature (see, for example, Bradley 1985; Györfi *et al*., 1990, Chapter II). If the process is stationary the best predictor (in a quadratic sense) for Y given $X = x$ is the conditional expectation

$$r(x) = E(Y|X = x).$$

Our aim is to estimate $r(.)$ from data $\{(X_i, Y_i)\}_{i=1}^{n}$. What does this estimation technique look like for $\{Z_i: i \geqslant 1\}$, a real-valued process, given that we are interested in predicting Z_{n+s} from Z_n for some $s > 0$? The predictor is provided by the autoregression function

$$R(z) = E(Z_{n+s}|Z_n = z) \qquad (\forall n \geqslant 1). \tag{2.1}$$

The autoregression function R can then be interpreted as a regression curve of Y on X if we define $X_i = Z_i$, $Y_i = Z_{i+s}$ ($\forall i \geqslant 1$). Clearly, $\{(X_i, Y_i), i \geqslant 1\}$ is α-mixing when $\{Z_i, i \geqslant 1\}$ has this property.

For examples of processes satisfying this α-mixing condition we refer to Györfi *et al*. (1990), Chapters II.2 and III.4. For instance, any Markov process satisfying Doeblin's conditions is α-mixing with coefficients that verify (C.1) above. Also, linear processes of the form

$$Z_n = \sum_{i=0}^{\infty} \gamma_i T_{n-i},$$

where $(T_j)_{j\in\mathbb{N}}$ is a sequence of i.i.d. variables, can be shown to be α-mixing under appropriate summability conditions on (γ_i) (see Chanda, 1974; Gorodetskii, 1977).

A measure of the closeness of $\hat{r}_h(x)$ to the curve $r(x)$ is provided by the average squared error $d_A(h)$ (see (1.2)). We shall state in Lemma 8 that, under suitable assumptions on h,

$$d_A(h) = n^{-1}h^{-1}C_1 + h^4C_2 + o_p(n^{-1}h^{-1} + h^4). \tag{2.2}$$

Here, C_1 and C_2 denote constants depending on the kernel, conditional moments and higher-order derivatives of r. Formula (2.2) specifies that the variance of $\hat{r}_h$ asymptotically tends to zero proportional to $n^{-1}h^{-1}$ and that the squared bias of $\hat{r}_h$ is a multiple of h^4. This fact is interesting from the point of view of the asymptotic behaviour of d_A and h. It tells us that we should select a bandwidth $h = (n^{-1}C_1/4C_2)^{1/5}$ in order to minimize the leading term in (2.2). Unfortunately, this optimization procedure has the drawback that it involves functionals of the underlying distribution.

There seem to be two ways out of this dilemma: first, we could estimate the unknown quantities; second, we could estimate the distance $d_A(h)$ (probably up to a constant) directly from the data. The first approach is usually called the 'plug-in' estimation procedure and is discussed in the setting of independent observations (e.g. Scott and Terrel, 1987). In this paper we follow the second route. In order to use the 'plug-in' recipe it is necessary to estimate higher derivatives of r which creates yet another smoothing parameter selection problem. This problem is avoided in our approach. To give some insight into how the adaptation works, decompose the average squared error

$$d_A(h) = n^{-1}\left\{\sum_{i=1}^{n} \hat{r}_h^2(X_i)w(X_i) - 2\sum_{i=1}^{n} r(X_i)\hat{r}_h(X_i)w(X_i) + \sum_{i=1}^{n} r^2(X_i)w(X_i)\right\}$$

and note that the final term is independent of h. The problem of minimizing d_A over a set of bandwidths is thus the same as that of minimizing the first two terms. The first term can be computed from the data, but the second one needs to be estimated since it involves the unknown regression function r. A first attempt could be to plug in Y_i for $r(X_i)$, but it is not hard to see that this estimate of the cross-term is of the same order as the variance term of $Ed_A(h)$. Therefore the following *leave-out* estimate of the cross-term will be considered:

$$n^{-1}\sum_{i=1}^{n} Y_i\hat{r}_{h,i}(X_i)w(X_i)$$

where

$$\hat{r}_{h,i}(x) = (n-1)^{-1}\sum_{j\neq i}\frac{Y_jK_h(x-X_j)}{\hat{f}_{h,i}(x)}$$

$$\hat{f}_{h,i}(x) = (n-1)^{-1}\sum_{j\neq i}K_h(x-X_j)$$

$$K_h(.) = h^{-1}K(./h).$$

We call $\hat{r}_{h,i}$ the leave-out estimator of r. The leave-out technique ensures that the estimate of the cross-term is asymptotically unbiased. Note that adding the random variable $n^{-1}\sum_{i=1}^{n} Y_i^2 w(X_i)$ does not change the minimization problem. We therefore consider

$$\mathrm{CV}(h) = n^{-1} \sum_{i=1}^{n} \{Y_i - \hat{r}_{h,i}(X_i)\}^2 w(X_i),$$

which is commonly called the cross-validation function.

The adaptive estimation then works as follows:

(i) compute the leave-out estimator $\hat{r}_{h,i}(X_i)$ for any h;
(ii) find the $\hat{h}$ that minimizes $\mathrm{CV}(h)$;
(iii) predict $r(x)$ using $\hat{r}_{\hat{h}}(x)$.

The optimality result stated in this paper ensures that $\hat{r}_{\hat{h}}(x)$ has the asymptotically smallest distance to $r(x)$. From now on we shall abbreviate $\hat{r}_{h,i}(.)$ as $\hat{r}_i(.)$ and $\hat{f}_{h,i}$ as $\hat{f}_i(.)$.

3. ADAPTIVE SMOOTHING IN PRACTICE

We simulated an autoregressive process as in (1.3) with

$$R(x) = \frac{x}{1 + x^2} \qquad (-1 \leqslant x \leqslant 1)$$

where the innovations were uniformly distributed over the interval $(-\frac{1}{2}, \frac{1}{2})$. Such a process is α-mixing with geometrically decreasing $\alpha(n)$ as shown by Doukhan and Ghindès (1980) and Gÿorfi *et al.* (1990, Chapter III.4.4). The sample size investigated was $n = 500$. The quartic kernel function

$$K(u) = \begin{cases} (15/16)(1 - u^2)^2 & |u| \leqslant 1 \\ 0 & |u| > 1 \end{cases}$$

was used.

All computations were done in GAUSS 2.0. A plot of the generated time series (Z_0 uniform in $(-\frac{1}{2}, \frac{1}{2})$) is given in Figure 1 as a function of the time index. We are interested in finding the dependence structure between Z_{n-1} and Z_n. When we plot Z_{n-1} versus Z_n we obtain Figure 2. The (uniform) error structure becomes quite visible here, but the shape of $R(x)$ can be guessed to be linear from this figure. Only at the far ends do we seem to see a curved structure of this point cloud.

We now apply the smoothing parameter selection technique described in this paper. Since this is a simulated example we can also compute the distance $d_{\mathrm{A}}(h)$. The cross-validation function $\mathrm{CV}(h)$ was computed using the discretization technique of Härdle (1990, Chapter 3). Both functions are shown in Figure 3. The minimum $\mathrm{CV}(h)$ is $\hat{h} = 0.18$, and the optimum of $d_{\mathrm{A}}(h)$ is at

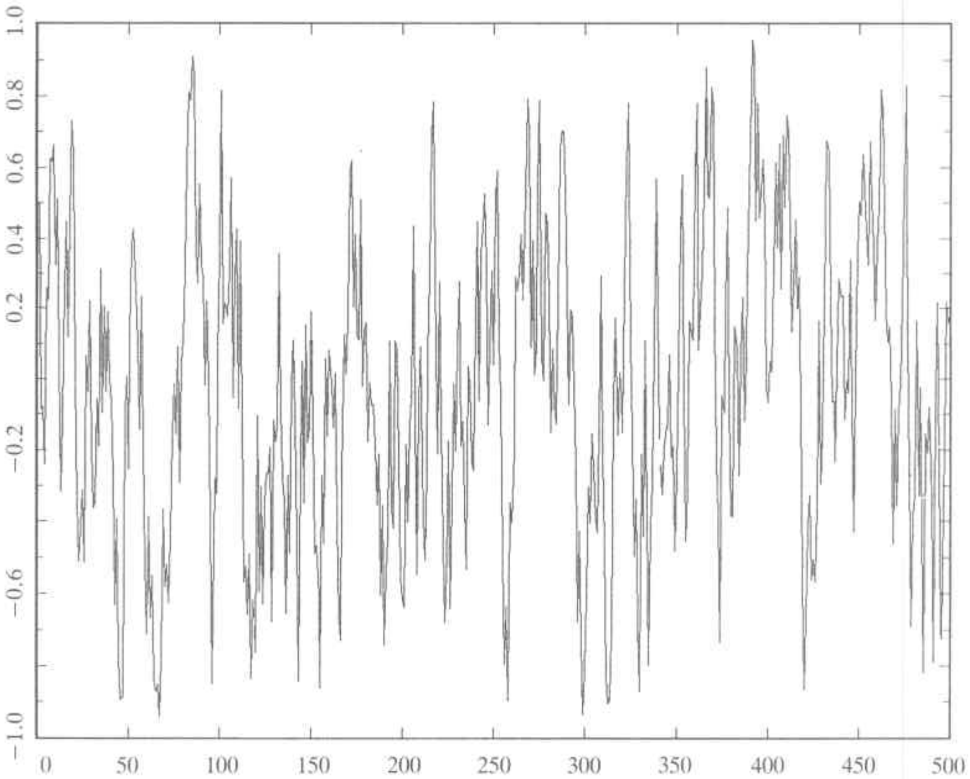


FIGURE 1. The simulated time series with $R(x) = x/(1 + x^2)$, $\varepsilon \sim U(-\frac{1}{2} - \frac{1}{2})$.

0.17. The curve $d_A(h)$ is very flat for this example since we recall that there is almost no bias. (For this display $d_A(h)$ has been shifted by an amount $\min \mathrm{CV}(h)$.) The comparison of the optimally estimated $\hat{r}_h$ with the time regression function gives an impression of how well the smoothing method works. This comparison is displayed in Figure 4 where we find good coincidence with the time regression curve.

It might be reasonable to leave out more than just one observation, especially when the time series is strongly correlated. Such a leave-out estimator, where, in fact, we sum over indexes $|i - j| > \rho_n$ for a slowly increasing sequence ρ_n, is also covered by our theory (see Remark 1). This 'leave-out-more' technique is sometimes also appealing in the independent setting (see the discussion of Härdle *et al.* (1988)). The examples treated by Hart and Vieu (1990), in the setting of density estimation, also discuss this point.

We also applied the algorithm to an economic time series: gold prices from 1978 to 1985. This series was kindly provided by D. Sondermann, University of Bonn. The data set consists of daily (log) gold prices in (in Deutschmarks) from 1978 to May 1986. The sample size n was 2041.

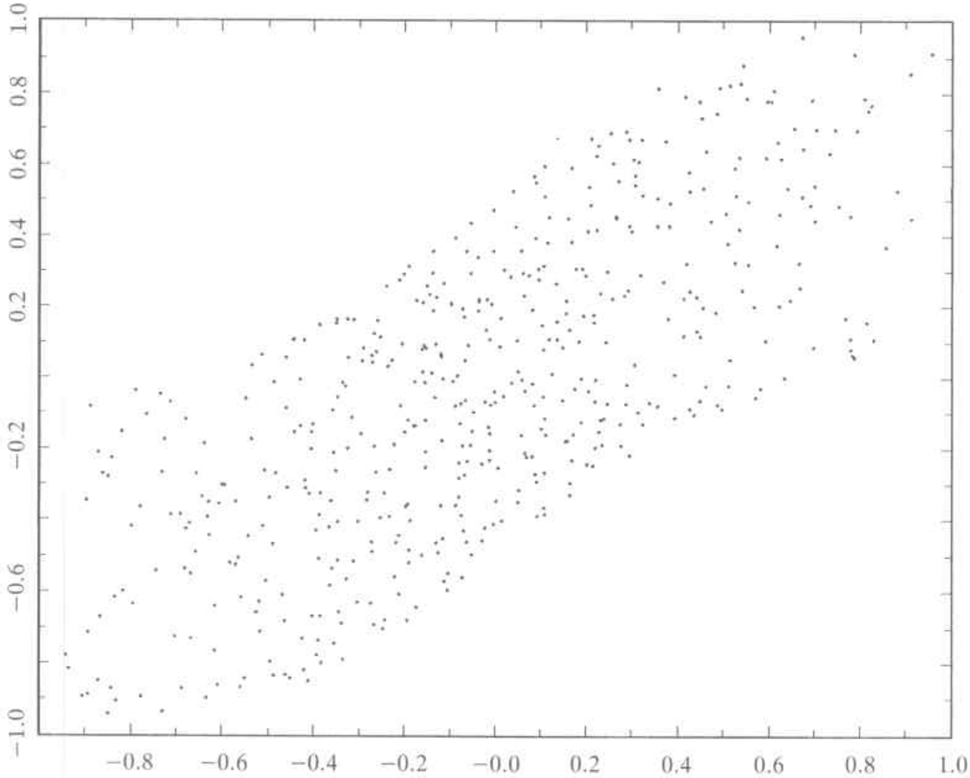


FIGURE 2. The simulated series from Figure 1 plotted as Z_{n-1} versus Z_n.

Let Z_i denote the price at time i. Then an interesting quantity for prediction is the elasticity in this series which is defined by

$$Y_i = \frac{Z_{i+1} - Z_i}{Z_i} = \frac{Z_{i+1}}{Z_i} - 1.$$

Thus the series Y_i indicates whether the price at the next time point will be relative to $X_i = Z_i$ below or above X_i. Certainly for high X_i we expect a tendency for Y_i to fall below zero and the other way around for smaller values of X. This tendency is hard to see in Figure 5 where we show a scatter plot of $\{(X_i, Y_i)\}_{i=1}^{2040}$. However, it will become clearer when we compute $\hat{r}_{\hat{h}}(x)$.

The cross-validation function CV(h) for this example had a minimum at $\hat{h} = 0.45$ and the corresponding optimal regression smoother $\hat{r}_{\hat{h}}$ is shown in Figure 6. As predicted, this curve is downward sloping but shows some non-linearities at the ends. It is interesting to analyze the residual structure in this example. We constructed the estimated residuals $\hat{\varepsilon}_i = Y_i - \hat{r}_{\hat{h}}(X_i)$ and regressed $\hat{\varepsilon}_i^2$ on X_i to obtain an estimate for the conditional variance

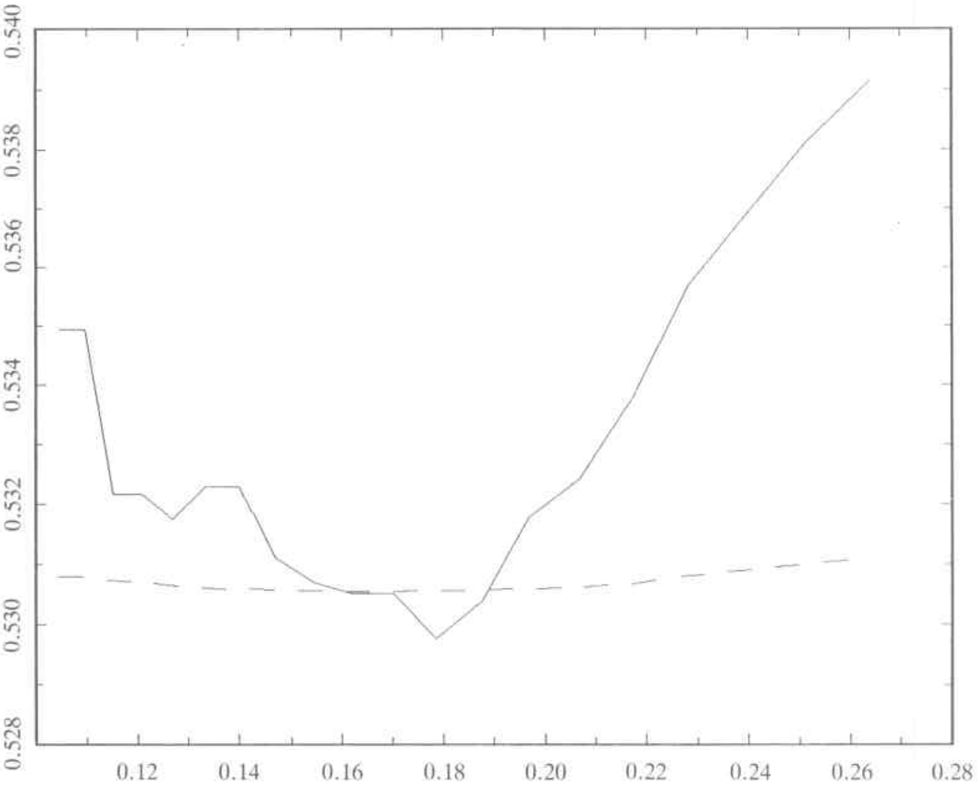


FIGURE 3. The functions $d_A(h)$ (broken curve) and CV(h) (full curve) for the simulated example.

function. The cross-validation function CV(h) for this problem is shown in Figure 7.

Using the parameter that minimized $CV(h)$ we obtained an optimal estimate for the variance function. The variance function is shown in Figure 8. It is apparent that it has a pronounced heteroscedasticity in the medium region of the observations.

4. ASYMPTOTIC OPTIMALITY OF THE ADAPTIVE ESTIMATOR

In view of (2.2), a reasonable candidate for a smoothing parameter h should be proportional to $n^{-1/5}$. The bandwidths are therefore selected in

$$H_n = [an^{-1/5-\varepsilon},\ bn^{-1/5+\varepsilon}],\ \text{for some } 0 < a < b < \infty \text{ and } 0 < \varepsilon < 1/10. \quad (C.2)$$

Assume that the kernel function satisfies

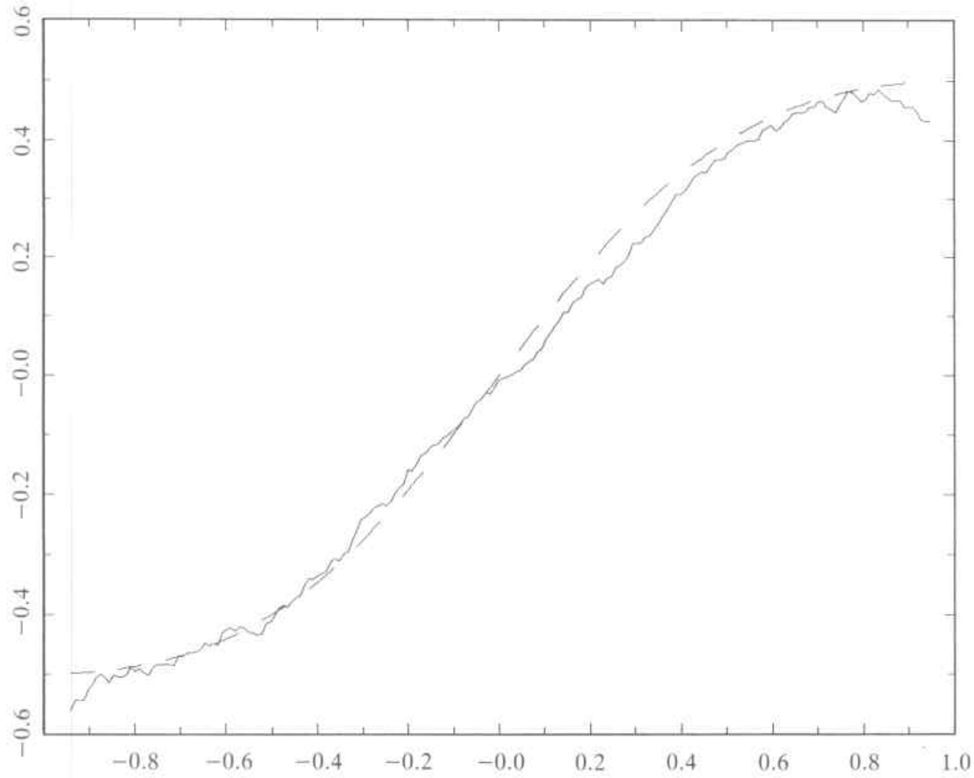


FIGURE 4. The time regression function $R(x) = x/(1 + x^2)$ for the simulated example (full curve) and the asymptotically optimal kernel smoothers (broken curve).

$$K \text{ is symmetric, Lipschitz continuous and has an absolutely integrable Fourier transform} \tag{C.3}$$

$$\int K(u)\,du = 1 \qquad K(.) \geqslant 0 \qquad \int u^2 K(u)\,du < \infty. \tag{C.4}$$

Assume that the weight function w is bounded and that its support S is compact. Let f denote the marginal density of X. Make the following assumptions:

$$f \text{ has a compact support containing } S; \tag{C.5}$$

$$r \text{ and } f \text{ have two continuous derivatives on the interior of } S; \tag{C.6}$$

$$E|Y|^k < \infty \ (k \geqslant 1). \tag{C.7}$$

To make our proofs shorter and clearer we have assumed that

$$\exists s \in]0, +\infty[, \qquad \exists t \in]0, 1[, \qquad \alpha(n) = st^n \qquad (\forall n \geqslant 1). \tag{C.8}$$

This assumption of a geometrically decaying mixing coefficient is quite

Journal Time Series Analysis, 13, 209-232

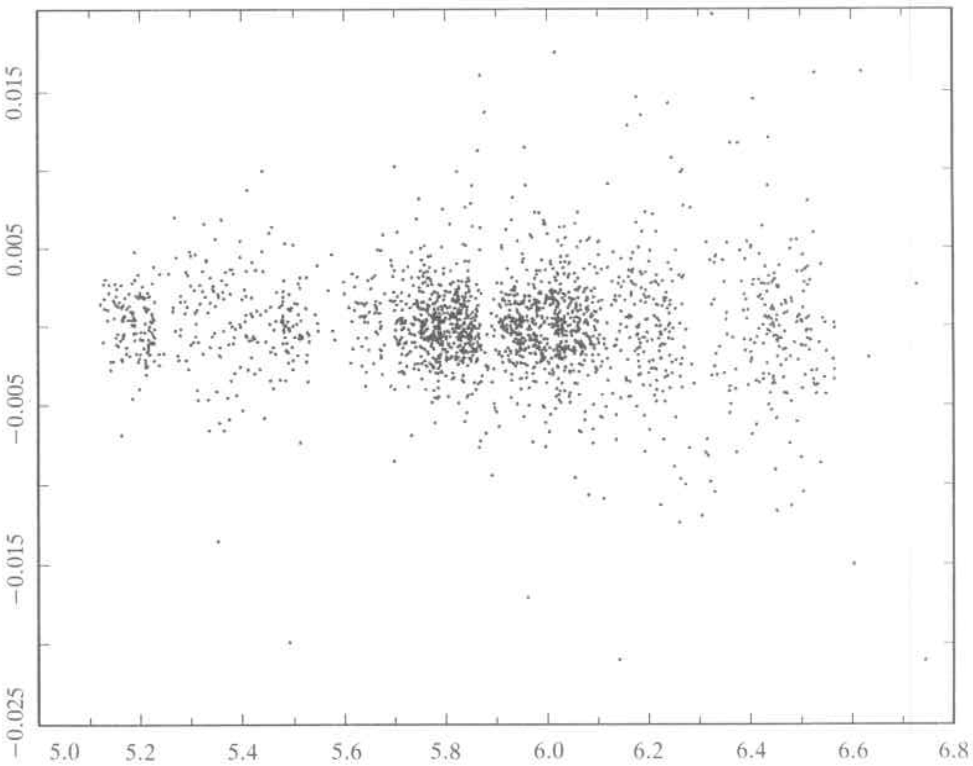


FIGURE 5. Values of X_i versus Y_i ($1 \leqslant i \leqslant 2040$) for the gold price data.

common in such problems (e.g. Truong and Stone, 1989). However, it would be possible, but with more tedious proofs, to obtain Theorem 1 under less restrictive assumptions that include some algebraically decaying rates. These assumptions would be similar to (L.1) and (L.2) in Hart and Vieu (1990).

THEOREM 1. *Under conditions (C.1)–(C.8) the adaptive non-parametric prediction algorithm is asymptotically optimal in the sense that*

$$\frac{d_A(\hat{h})}{d_A(h_0)} \xrightarrow{p} 1 \tag{4.1}$$

where

$$h_0 = \arg\min_{h \in H_n} d_A(h)$$

and

$$\hat{h} = \arg\min_{h \in H_n} \mathrm{CV}(h).$$

(1992) Härdle, W. and Vieu, P.
Kernel Regression Smoothing of Time Series.

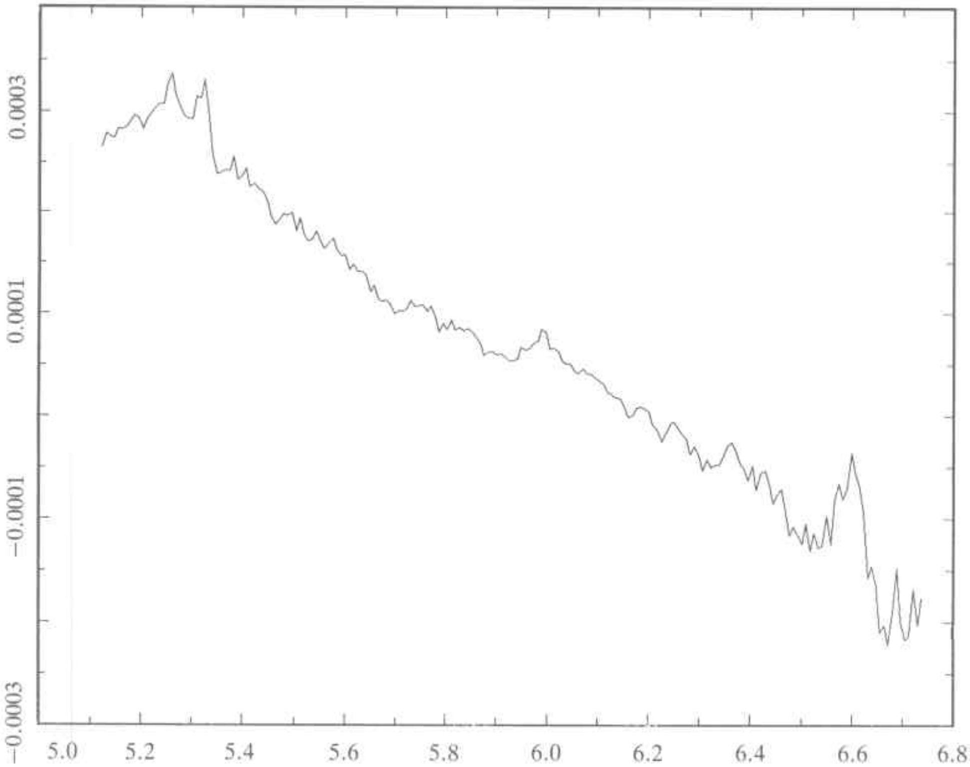


FIGURE 6. The optimal kernel smoother $\hat{r}_h(x)$ for the gold price data.

REMARK 1. In the proof of this theorem we use the more general leave-one-out estimator already mentioned in Section 3. More precisely we use a sequence ρ_n such that

$$1 \leqslant \rho_n \leqslant \rho_n^*, \text{ where } \rho_n^* = n^\tau \text{ for some } 0 < \tau < 1/15 - \varepsilon/3. \qquad \text{(C.9)}$$

It thus follows from our proof that Theorem 1 is also valid for the adaptive estimator when we leave out more than one observation.

4.1 *Application to non-parametric prediction of time series*

The kernel estimator of the s-step predictor of Z_{n+s} given $\{Z_i\}_{i=1}^n$ is defined by

$$\hat{R}_h(x) = \frac{\sum_{i=1}^{n-s} Z_{i+s} K_h(x - Z_i)}{\sum_{i=1}^{n-s} K_h(x - Z_i)}$$

and h is selected to minimize

$$\mathrm{CV}(h) = (n - s)^{-1} \sum_{i=1}^{n-s} \{Z_{i+s} - \hat{R}_i(Z_i)\}^2 w(Z_i),$$

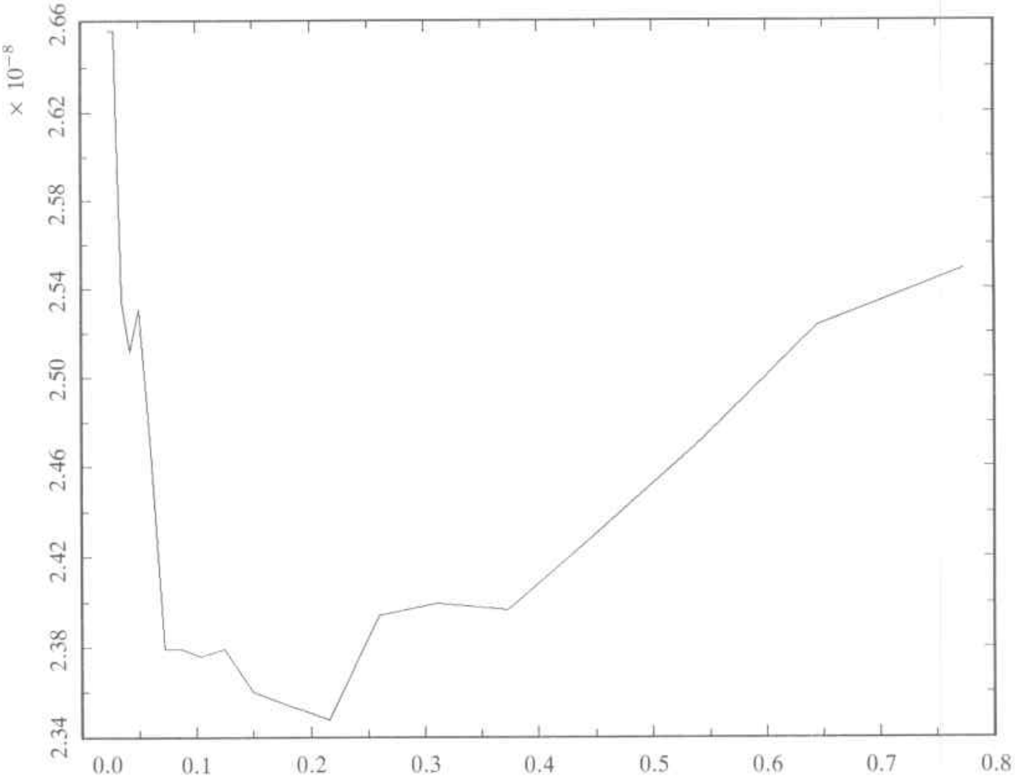


FIGURE 7. The cross-validation function CV(h) for the residual pattern.

where

$$\hat{R}_i(x) = \frac{\sum_{j \neq i}^{n-s} Z_{j+s} K_h(x - Z_j)}{\sum_{j \neq i}^{n-s} K_h(x - Z_j)}.$$

From Theorem 1 we have the following theorem.

THEOREM 2. *If $\{Z_n, n \geqslant 1\}$ is a stationary process, and if (C.1)–(C.8) are satisfied by*

$$X_i = Z_i \qquad Y_i = Z_{i+s} \qquad r = R,$$

$\hat{h}$ is optimally selected to estimate R, in the sense that (4.1) is satisfied for $\hat{r}_h = \hat{R}_h$.

5. PROOF OF THEOREM 1

We prove the asymptotic optimality property (4.1) in the more general case when we do not necessarily leave out just one point. In this case, the

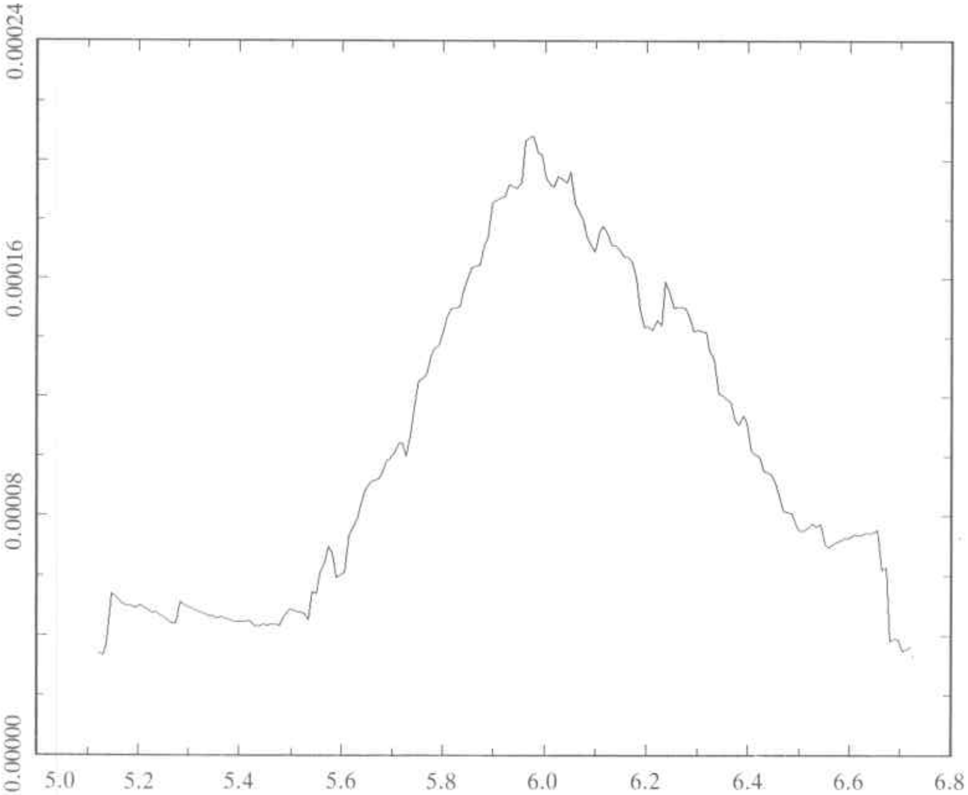


FIGURE 8. The variance function for the gold price data.

leave-out estimator $\hat{r}_i$ is defined by

$$\hat{r}_i(x) = n_i^{-1} \sum_{|j-i|>\rho_n} \frac{Y_j K_h(X_j - X_i)}{\hat{f}_i(x)}$$

where

$$\hat{f}_i(x) = n_i^{-1} \sum_{|j-i|>\rho_n} K_h(X_j - X_i)$$

$$n_i = \#\{j \in \{1, \ldots, n\}, |j - i| > \rho_n\}$$

and where ρ_n is defined in (C.9).

To check (4.1) it is enough to show that

$$\sup_{h,h' \in H_n} \frac{|d_{\mathrm{A}}(h) - d_{\mathrm{A}}(h') - \{\mathrm{CV}(h) - \mathrm{CV}(h')\}|}{d_{\mathrm{A}}(h)} = \mathrm{o}_p(1). \tag{5.1}$$

NOTATION. The average squared error based on the leave-out estimator is defined as

$$\bar{d}_{\mathrm{A}}(h) = n^{-1} \sum_{j=1}^{n} \{\hat{r}_j(X_j) - r(X_j)\}^2 w(X_j).$$

In the following we shall denote by C any finite real constant and we shall write ε_j in place of $Y_j - r(X_j)$.

LEMMA 1. *Under the conditions of Theorem 1 we have for any compact subset G of $\mathbb{R}$*

$$\sup_{x \in G} \sup_{h \in H_n} |\hat{f}_h(x) - f(x)| = \mathrm{o}_p(n^{-1/5})$$

and

$$\sup_{x \in G} \sup_{h \in H_n} |\hat{r}_h(x) - f(x)| = \mathrm{o}_p(n^{-1/5}).$$

LEMMA 2. *Under the conditions of Theorem 1 we have*

$$\sup_{h \in H_n} \frac{|\bar{d}_{\mathrm{A}}(h) - d_{\mathrm{A}}(h)|}{d_{\mathrm{A}}(h)} = \mathrm{o}_p(1).$$

Statement (5.1) now follows from Lemma 2 and

$$\sup_{h, h' \in H_n} \frac{|\bar{d}_{\mathrm{A}}(h) - \bar{d}_{\mathrm{A}}(h') - \{\mathrm{CV}(h) - \mathrm{CV}(h')\}|}{d_{\mathrm{A}}(h)} = \mathrm{o}_p(1). \tag{5.2}$$

Decompose

$$\bar{d}_{\mathrm{A}}(h) + n^{-1} \sum_{i=1}^{n} \varepsilon_i^2 w(X_i) = \mathrm{CV}(h) + 2C_n(h)$$

where

$$C_n(h) = n^{-1} \sum_{i=1}^{n} \varepsilon_i \{\hat{r}_i(X_i) - r(X_i)\} w(X_i).$$

Then (5.2) is true if we show that

$$\sup_{h \in H_n} \frac{|C_n(h)|}{d_{\mathrm{A}}(h)} = \mathrm{o}_p(1). \tag{5.3}$$

Let us define

$$\bar{C}_n(h) = n^{-1} \sum_{i=1}^{n} \varepsilon_i \{\hat{r}_i(X_i) - r(X_i)\} \frac{\hat{f}_i(X_i)}{f(X_i)} w(X_i)$$

and decompose

$$\bar{C}_n(h) = \bar{C}_{n,1}(h) + \bar{C}_{n,2}(h),$$

where

$$\bar{C}_{n,1}(h) = n^{-1} \sum_{i=1}^{n} n_i^{-1} \sum_{|j-i| > \rho_n} \varepsilon_i \varepsilon_j K_h(X_i - X_j) w(X_i) f^{-1}(X_i)$$

and

$$\bar{C}_{n,2}(h) = n^{-1} \sum_{i=1}^{n} n_i^{-1} \sum_{|j-i|>\rho_n} \varepsilon_j \{r(X_i) - r(X_j)\} K_h(X_i - X_j) w(X_i) f^{-1}(X_i).$$

Define H'_n to be a finite subset of H_n consisting of equally spaced elements and such that

$$\# H'_n = n^{\tau_1} \text{ for some } \tau_1 > 8/5 + 2\varepsilon.$$

Then (5.3) follows from the following lemmas.

LEMMA 3. *Under conditions (C.1)–(C.9) and if $\rho_n = \rho_n^*$ we have*

$$\sup_{h \in H'_n} \frac{|\bar{C}_{n,1}(h)|}{d_{\mathrm{A}}(h)} = \mathrm{o}_p(1).$$

LEMMA 4. *Under conditions (C.1)–(C.9) and if $\rho_n = \rho_n^*$, we have*

$$\sup_{h \in H'_n} \frac{|\bar{C}_{n,2}(h)|}{d_{\mathrm{A}}(h)} = \mathrm{o}_p(1).$$

LEMMA 5. *Under conditions (C.1)–(C.9) and if $\rho_n = \rho_n^*$, we have*

$$\sup_{h \in H'_n} \frac{|\bar{C}_n(h) - C_n(h)|}{d_{\mathrm{A}}(h)} = \mathrm{o}_p(1).$$

LEMMA 6. *Under conditions (C.1)–(C.9) and if $\rho_n = \rho_n^*$, we have*

$$\sup_{h \in H'_n} \frac{|C_n(h) - C_n(h^*)|}{d_{\mathrm{A}}(h)} = \mathrm{o}_p(1).$$

where, for any $h \in H_n$, h^* is defined to be the element of H'_n that is closest to h.

LEMMA 7. *Under conditions (C.1)–(C.9) we have*

$$\sup_{h \in H'_n} \frac{|C_n(h) - C_n^*(h)|}{d_{\mathrm{A}}(h)} = \mathrm{o}_p(1)$$

where C_n^* denotes the quantity C_n which applies when $\rho_n = \rho_n^*$.

An important tool which will be used throughout the proofs of these lemmas is the following variance-squared bias decomposition of the error d_{A}.

LEMMA 8. *Under the conditions of Theorem 1, we have*

$$d_{\mathrm{A}}(h) = C_1(nh)^{-1} + C_2 h^4 + \mathrm{o}_p\{d_{\mathrm{A}}(h)\}$$

where C_1 and C_2 are real positive constants. Similarly, we have for some real positive constants C_1' and C_2'

$$n^{-1}\sum_{i=1}^{n}\{\hat{f}_i(X_i)-f(X_i)\}^2 w(X_i) - C_1'(nh)^{-1} + C_2'h^4 + \mathrm{o}_p\{d_{\mathrm{A}}(h)\}.$$

APPENDIX

Proof of Lemma 1. Lemma 1 is in fact a weaker version of Theorems 3.3.6 and 5.3.3 of Gÿorfi *et al.* (1990), observing that the bounds given in these theorems are independent of $h \in H_n'$. See also Roussas (1989) for similar results. ■

Proof of Lemma 2. Consider

$$\bar{d}_{\mathrm{A}}(h) - d_{\mathrm{A}}(h) = -2n^{-1}\sum_{i=1}^{n}\{\hat{r}_h(X_i)-r(X_i)\}\{\hat{r}_h(X_i)-\hat{r}_i(X_i)\}w(X_i)$$
$$+\, n^{-1}\sum_{i=1}^{n}\{\hat{r}_i(X_i)-\hat{r}_h(X_i)\}^2 w(X_i).$$

By the definitions of $\hat{r}$ and $\hat{r}_i$ we obtain

$$\hat{r}_i(X_i) = \frac{\hat{f}_h(X_i)}{\hat{f}_i(X_i)}\left\{\frac{n}{n_i}\,\hat{r}_h(X_i)\right\} - \frac{A_i}{\hat{f}_i(X_i)}$$

where

$$A_i = n_i^{-1}\sum_{|i-j|\le\rho_n} K_h(X_i - X_j)Y_j.$$

Then it follows that

$$\bar{d}_{\mathrm{A}}(h) - d_{\mathrm{A}}(h) = 2T_1(h) + T_2(h),$$

with

$$T_1(h) = n^{-1}\sum_{i=1}^{n}\{\hat{r}_h(X_i)-r(X_i)\}\left[\hat{r}_h(X_i)\left\{\frac{\hat{f}_h(X_i)}{\hat{f}_i(X_i)}\frac{n}{n_i}-1\right\} - \frac{A_i}{\hat{f}_i(X_i)}\right]w(X_i)$$

and

$$T_2(h) = n^{-1}\sum_{i=1}^{n}\left[\hat{r}_h(X_i)\left\{\frac{\hat{f}_h(X_i)}{\hat{f}_i(X_i)}\frac{n}{n_i}-1\right\} - \frac{A_i}{\hat{f}_i(X_i)}\right]^2 w(X_i).$$

It is enough to show that

$$\sup_{h\in H_n'}\frac{T_1(h)}{d_{\mathrm{A}}(h)} \xrightarrow[n\to\infty]{p} 0 \tag{A.1}$$

and

$$\sup_{h\in H_n'}\frac{T_2(h)}{d_{\mathrm{A}}(h)} \xrightarrow[n\to\infty]{p} 0. \tag{A.2}$$

Using Lemma 1 we have

$$\sup_{x,h}|\hat{r}_h(x) - r(x)| = \mathrm{o}_p(n^{1/5}). \tag{A.3}$$

We also have

$$|A_i| \leq n_i^{-1} 2\rho_n h^{-1} \sup_j |Y_j| \sup_u K(u)$$

which gives

$$\sup_{h,i} \frac{A_i}{\hat{f}_i(X_i)} = \mathrm{o}_p(\rho_n n^{-1} h^{-1}). \tag{A.4}$$

$\hat{f}_i(X_i)$ is bounded below from zero with probability approaching 1 in view of (C.5) and Lemma 1. Similarly we have

$$|n\hat{f}_h(X_i) - n_i \hat{f}_i(X_i)| \leq 2\rho_n h^{-1} \sup_u K(u)$$

and so it follows that

$$\sup_{h,i} \left| \frac{\hat{f}_h(X_i) n}{\hat{f}_i(X_i) n_i} - 1 \right| = \mathrm{o}_p(\rho_n n^{-1} h^{-1}). \tag{A.5}$$

Then (A.1) (or (A.2)) follows from (A.3), (A.4), (A.5) and Lemma 8 (or (A.4), (A.5) and Lemma 8).

Proof of Lemma 3. It suffices to show for some $b_1 > 0$ that we have for any integer k

$$\# H'_n \sup_{h \in \bar{H}'_n} E \left\{ \frac{\bar{C}_{n,1}(h)}{d_{\mathrm{A}}(h)} \right\}^{2k} = \mathrm{O}(n^{-kb_1}). \tag{A.6}$$

Let us define

$$C^+(h) = n^{-2} \sum_{i=1}^{n} \sum_{j > i + \rho_n^*} U(i, j)$$

and

$$C^-(h) = n^{-2} \sum_{i=1}^{n} \sum_{j < i - \rho_n^*} U(i, j)$$

where

$$U(i, j) = K_h(X_i - X_j) \varepsilon_i \varepsilon_j \frac{w(X_i)}{f(X_i)}.$$

Since, for any i, n_i is of the same order of magnitude as n, (A.6) will follow from

$$\# H'_n \sup_{h \in \bar{H}_n} E \left\{ \frac{C^*(h)}{d_{\mathrm{A}}(h)} \right\}^{2k} = \mathrm{O}(n^{-kb_1}) \tag{A.7}$$

where C^* is either C^+ or C^-. We show (A.7) for $C^* = C^+$. The analysis of C^- is similar. Defining $U(i, j)$ to be equal to zero when (i, j) is not in the set $\{(i, j), i + \rho_n^* < j \leq n\}$, we can write $C^+(h)$ as

$$C^+(h)$$
$$= n^{-2} \sum_{s=0}^{1} \sum_{t=0}^{1} \sum_{j_1=1}^{\rho_n^*} \sum_{j_2=1}^{\rho_n^*} \sum_{q=1}^{n_1} \sum_{m=1}^{q} U\{j_2 + 2(m-1)\rho_n^* + s\rho_n^*, j_1 + (2q-1)\rho_n^* + t\rho_n^*\}$$

where n_1 is the greatest integer less than or equal to $n/2\rho_n^*$. Therefore, in fact, we have to prove that

$$(nh)^{2k} \# H'_n (n^{-1} \rho_n^*)^{4k} \sup_{j_1, j_2, s, t} E \left\{ \sum_{q=1}^{n_1} \sum_{m=1}^{q} U(m', q') \right\}^{2k} = \mathrm{O}(n^{-kb_1}) \tag{A.8}$$

where, for j_1, j_2, s and t fixed, we use the notation

$$m' = j_2 + 2(m-1)\rho_n^* + s\rho_n^* \qquad q' = j_1 + (2q-1)\rho_n^* + t\rho_n^*$$

and where we bounded $d_A(h)^{-1}$ by (nh) because of Lemma 8.

Define now

$$\mathcal{J} = \{J = (m_1, q_1, \ldots, m_{2k}, q_{2k});\ 1 \leqslant m_i \leqslant q_i \leqslant n_i,\ \forall i = 1, \ldots, 2k\}$$

and decompose it into $\mathcal{J} = \mathcal{J}_1 \cup \mathcal{J}_2$ where

$$\mathcal{J}_1 = \{J = (m_1, \ldots, q_{2k}) \in \mathcal{J},\ \exists u \in \{m_1, \ldots, q_{2k}\},\ \forall v \in \{m_1, \ldots, q_{2k}\},\ |u - v| \geqslant 2\}$$

and $\mathcal{J}_2 = \mathcal{J} - \mathcal{J}_1$.

Therefore what we have to deal with in fact is

$$E\left\{\sum_{q=1}^{n_1} \sum_{m=1}^{q} U(m', q')\right\}^{2k} = \sum_{J\in\mathcal{J}_1} E\Psi(J) + \sum_{J\in\mathcal{J}_2} E\Psi(J) \tag{A.9}$$

where, for $J \in \mathcal{J}$, $\Psi(J)$ is defined by

$$\Psi(J) = \prod_{i=1}^{2k} U(m_i', q_i').$$

Let us first consider the case when $J \in \mathcal{J}_1$. Let m_0 be an element of $\{m_1, \ldots, q_{2k}\}$ which differs from all others by at least 2. (The proof would be the same if the index was some q_0.) By definition of the m_i' and q_i', we have that, for any $u \in \{m_1, \ldots, q_{2k}\} - \{m_0\}$, $|u' - m_0| \geqslant \rho_n^*$, and so by applying Proposition 1 of Hart and Vieu (1990) we obtain

$$|E\Psi(J)| \leqslant \left| \int \left\{ \int \Psi(J) dP_{(X_{m_0'}, Y_{m_0'})} \right\} dP_{X_u', Y_u', u\in\{m_1,\ldots,q_{2k}\}-\{m_0\}} \right| + \mathrm{o}\{h^{-2k}\alpha(\rho_n^*)\}.$$

Conditioning now with respect to $X_{m_0'}$ and using the fact that the ε_j have mean zero, we obtain

$$\int \Psi(J)\, dP_{(X_{m_0'}, Y_{m_0'})} = 0.$$

Finally, since $\#\mathcal{J}_1$ is of order n^{4k} we have

$$\left| \sum_{J\in J_1} E\Psi(J) \right| = \mathrm{O}\{n^{4k}(\rho_n^*)^{-4k} h^{-2k} \alpha(\rho_n^*)\}. \tag{A.10}$$

Let us consider the case when $J \in \mathcal{J}_2$. For this we write

$$\mathcal{J}_2 = \bigcup_{l=1}^{4k} \mathcal{J}_2^l \qquad .$$

where

$$\mathcal{J}_2^l = \{J \in \mathcal{J}_2,\ \#\{m_1', \ldots, q_{2k}'\} = l\}.$$

Again applying Proposition 1 of Hart and Vieu (1990), we have

$$|E\Psi(J)| = \left| \int \Psi(J) \prod_{u\in\{m_1',\ldots,q_{2k}'\}} dP_{(X_u, Y_u)} \right| + \mathrm{O}\{h^{-2k}\alpha(\rho_n^*)\}.$$

Integration by substitution (see Marron and Härdle, 1986, formula (3.4)) leads to

$$\left| \int \Psi(J) \prod_{u\in\{m_1',\ldots,q_{2k}'\}} dP_{(X_u, Y_u)} \right| = \mathrm{O}\{h^{-2k+l/2}\} \qquad \text{for } J \in \mathcal{J}_2^l.$$

Finally, noting that $\#\mathcal{J}_2^l = O(n_1^{\inf(l,2k)})$, we have

$$\left|\sum_{J\in\mathcal{J}_2} E\Psi(J)\right| = O\{n^{2k}(\rho_n^*)^{-2k}\alpha(\rho_n^*)h^{-2k}\} + O\left\{\sum_{l=2k}^{4k} n^{2k}(\rho_n^*)^{-2k}h^{-2k+l/2}\right\} + O\left\{\sum_{l=1}^{2k} n^{l}(\rho_n^*)^{-l}h^{-2k+l/2}\right\}.$$

Then we have

$$\left|\sum_{J\in\mathcal{J}_2} E\Psi(J)\right| = O\{n^{2k}(\rho_n^*)^{-2k}\alpha(\rho_n^*)h^{-2k}\} + O\{n^{2k}(\rho_n^*)^{-2k}h^{-k}\}. \qquad (A.11)$$

It follows from (A.9), (A.10) and (A.11) that

$$E\left\{\sum_{q=1}^{n_1}\sum_{m=1}^{q} U(m', q')\right\}^{2k} = O\{h^{-2k}n^{4k}(\rho_n^*)^{-4k}\alpha(\rho_n^*)\} + O\{n^{2k}(\rho_n^*)^{-2k}h^{-k}\},$$

and (A.8) follows by using (C.2), (C.8) and (C.9). This completes the proof of Lemma 3.

PROOF OF LEMMA 4. The main body of this proof is the same as that of Lemma 3. Proceeding similarly we have to show that

$$(nh)^{2k}\#H_n'(n^{-1}\rho_n^*)^{4k}\sup_{j_1,j_2,s,t} E\left\{\sum_{q=1}^{n_1}\sum_{m=1}^{q} V(m', q')\right\}^{2k} = O(n^{-b_1k}) \qquad (A.12)$$

for some $b_1 > 0$, where

$$V(i, j) = K_h(X_i, X_j)\{r(X_i) - r(X_j)\}\varepsilon_j \frac{w(X_i)}{f(X_i)}$$

and where n_1, m' and q' are defined as in Lemma 3. The set $\mathcal{J}$ is defined as before, and we decompose it in the following way:

$$\mathcal{J} = \mathcal{J}_3 \cup \left(\bigcup_{l=1}^{2k} \mathcal{J}_4^l\right)$$

where

$$\begin{aligned}\mathcal{J}_3 = \{J = \{&m_1, \ldots, q_{2k}\} \in \mathcal{J},\\ &\exists i \in \{1, \ldots, 2k\},\ q_i \neq q_j,\ \forall j \in \{1, \ldots, 2k\} - \{i\},\\ &|q_i - m_j| > 1,\ \forall j \in \{1, \ldots, 2k\}\}\end{aligned}$$

$$\mathcal{J}_4 = \mathcal{J} - \mathcal{J}_3$$

$$\mathcal{J}_4^l = \{J = \{m_1, \ldots, q_{2k}\} \in \mathcal{J}_4,\ \#\{m_1, \ldots, m_{2k}\} = l\}.$$

Similarly to the proof of (A.9), application of Proposition 1 in Hart and Vieu (1990) leads to

$$\sum_{J\in\mathcal{J}_3} E\left\{\prod_{i=1}^{2k} V(m_i', q_i')\right\} = O(n^{4k}(\rho_n^*)^{-4k}h^{-2k}\alpha(\rho_n^*)). \qquad (A.13)$$

Now let J be an element of $\mathcal{J}_4^l$. Because K is compactly supported we have that

$$\forall h \in H_n', \qquad K_h(X_i - X_j)\{r(X_i) - r(X_j)\} = o_p(1), \qquad (A.14)$$

and we also have, by well-known bias expansions (Parzen, 1962), that

$$\forall h \in H_n', \qquad \sup_{t\in\mathbb{R}} E|K_h(X_k - t)\{r(X_k) - r(t)\}| = O(h^2). \qquad (A.15)$$

In the product $\prod_{i=1}^{2k} V(m_i', q_i')$, we first bound all the terms

$$K_h(X_{m_i'} - X_{q_i'})\{r(X_{m_i'} - r(X_{q_i'})\}$$

for which m_i' is not in the set

$$A = \{u \in (m_1, \ldots, m_{2k}\}, \forall v \in \{m_1, \ldots, m_{2k}\}, u \neq v\}$$

by using (A.14). Then, we apply Proposition 1 of Hart and Vieu (1990) and obtain for some $0 < C < +\infty$

$$E\left\{\prod_{i=1}^{2k} V(m_i', q_i')\right\} \leq C\left\{\sup_t E|K_h(X - t)\{r(X) - r(t)\}|\right\}^l + \mathrm{O}\{h^{-2k}\alpha(\rho_n^*)\} \tag{A.16}$$

and finally we have by (A.15)

$$\forall J \in \mathcal{J}_4^l, \qquad E\left\{\prod_{i=1}^{2k} V(m_i', q_i')\right\} = \mathrm{O}\{h^{-2k}\alpha(\rho_n^*)\} + \mathrm{O}(h^{2l}). \tag{A.17}$$

Noting that $\#\mathcal{J}_3 = \mathrm{O}(n_1^{4k})$ and $\#\mathcal{J}_4^l = \mathrm{O}(n_1^{k+l})$, we finally obtain from (A.13) and (A.17)

$$\begin{aligned} E\left\{\sum_{q=1}^{n_1} \sum_{m=1}^{q} V(m', q')\right\}^{2k} &= \sum_{J \in \mathcal{J}_3} E\left\{\prod_{i=1}^{2k} V(m_i', q_i')\right\} + \sum_{l=1}^{2k} \sum_{J \in \mathcal{J}_4^l} E\left\{\prod_{i=1}^{2k} V(m_i', q_i')\right\} \\ &= \mathrm{O}\{n^{4k}(\rho_n^*)^{-4k} h^{-2k}\alpha(\rho_n^*)\} + \mathrm{O}(n^{3k}\rho_n^{-3}h^{4k}). \end{aligned}$$

As for Lemma 3, we complete the proof by observing that (A.12) follows from this equality together with (C.2), (C.8) and (C.9).

Proof of Lemma 5. We have

$$C_n(h) - \bar{C}_n(h) = D_1(h) + D_2(h) \tag{A.18}$$

where

$$D_1(h) = n^{-1}\sum_{i=1}^{n} \varepsilon_i\{\hat{r}_i(X_i) - r(X_i)\}w(X_i)\,\frac{\hat{f}_i(X_i)}{f(X_i)}\left\{\frac{f(X_i) - \hat{f}_i(X_i)}{f(X_i)}\right\}$$

and

$$D_2(h) = n^{-1}\sum_{i=1}^{n} \varepsilon_i\{\hat{r}_i(X_i) - r(X_i)\}w(X_i)\left\{\frac{\hat{f}(X_i) - f(X_i)}{f(X_i)}\right\}^2.$$

If follows from Lemmas 1 and 8 that

$$\sup_{h \in H_n'} \frac{|D_2(h)|}{d_{\mathrm{A}}(h)} = \mathrm{o}_p(1). \tag{A.19}$$

Note now that D_1 has roughly the same structure as $\bar{C}_n$. Therefore we can write D_1 as

$$D_1(h) = D_{11}(h) + D_{12}(h) \tag{A.20}$$

where, using the same notation as in Lemmas 3 and 4,

$$D_{11}(h) = n^{-1}\sum_{i=1}^{n} \sum_{|j-i| > \rho_n^*} n_i^{-1}U(i, j)\,\frac{f(X_i) - \hat{f}_i(X_i)}{f(X_i)}$$

and

$$D_{12}(h) = n^{-1}\sum_{i=1}^{n} \sum_{|j-i| > \rho_n^*} n_i^{-1}V(i, j)\,\frac{f(X_i) - \hat{f}_i(X_i)}{f(X_i)}$$

Proceeding as in Lemmas 3 and 4 we can show that

$$\sup_{h\in H_n} \frac{|D_{11}(h)|}{d_{\mathrm{A}}(h)} = \mathrm{o}_p(1) \tag{A.21}$$

and

$$\sup_{h\in H_n} \frac{|D_{12}(h)|}{d_{\mathrm{A}}(h)} = \mathrm{o}_p(1). \tag{A.22}$$

The proof of Lemma 5 is now complete by (A.18)–(A.22).

Proof of Lemma 6. Write $C_n(h)$ as

$$\begin{aligned} C_n(h) = n^{-1}/ &\sum_{|j-i|>\rho_n^*}\sum n_i^{-1}K_h(X_i - X_j)w(X_i)\hat{f}_i^{-1}(X_i)\varepsilon_i\varepsilon_j \\ &+ n^{-1} \sum_{|j-i|>\rho_n^*}\sum n_i^{-1}K_h(X_i - X_j)w(X_i)\hat{f}_i^{-1}(X_i)\varepsilon_j\{r(X_i) - r(X_j)\}. \end{aligned}$$

We have for some $0 < C < +\infty$

$$\begin{aligned} |C_n(h) - C_n(h^*)| &\leqslant C\{K_h(X_i - X_j) - K_{h^*}(X_i - X_j)\} \\ &= C\left\{\left(\frac{1}{h} - \frac{1}{h^*}\right) K\left(\frac{X_i - X_j}{h}\right) \right. \\ &\qquad \left. + \frac{1}{h^*}\left\{K\left(\frac{X_i - X_j}{h}\right) - K\left(\frac{X_i - X_j}{h^*}\right)\right\}\right\}. \end{aligned}$$

Since the points in H'_n are equally spaced, we have by (C.2)

$$\left|\frac{1}{h} - \frac{1}{h^*}\right| = \mathrm{O}\{(\# H'_n)^{-1} n^{-2/5+2\varepsilon} h^{-1}\},$$

and because K is Lipschitz continuous and compactly supported

$$\left|K\left(\frac{X_i - X_j}{h}\right) - K\left(\frac{X_i - X_j}{h^*}\right)\right| = \mathrm{O}\left\{h^*\left(\frac{1}{h} - \frac{1}{h^*}\right)\right\}.$$

Finally, we have for any h in H_n

$$|C_n(h) - C_n(h^*)| = \mathrm{O}\{(\# H'_n)^{-1} n^{-2/5+2\varepsilon} h^{-1}\},$$

and by Lemma 8 we obtain

$$\sup_{h\in H_n} \frac{|C_n(h) - C_n(h^*)|}{d_{\mathrm{A}}(h)} = \mathrm{O}(n^{-\tau_1+3/5+2\varepsilon}) = \mathrm{o}(n^{-1}).$$

This completes the proof of this lemma.

Proof of Lemma 7. Let us define

$$\hat{g}_i(X_i) = \frac{1}{n_i} \sum_{|j-i|>\rho_n} Y_j K_h(X_j - X_i)$$

and denote by $\hat{g}_i^*$, $\hat{f}_i^*$, n_i^* the quantities $\hat{g}_i$, $\hat{f}_i$, n_i that apply when $\rho_n = \rho_n^*$.

We have, for some $0 < C < \infty$,

$$\begin{aligned} |C_n(h) - C_n^*(h)| &\leqslant C \sup_{i=1,\ldots,n} |\hat{n}_i(X_i) - \hat{r}_i^*(X_i)| \\ &= \mathrm{O}(\sup_i |\hat{g}_i(X_i) - \hat{g}_i^*(X_i)| + |\hat{f}_i(X_i) - \hat{f}_i^*(X_i)|). \end{aligned}$$

We have

$$\begin{aligned}|\hat{g}_i(X_i) - \hat{g}_i^*(X_i)| &= \left|\left(\frac{1}{n_i^*} - \frac{1}{n_i}\right) \sum_{|j-i|>\rho_n^*} Y_j K_h(X_j - X_i)\right| \\ &\quad + \left|\frac{1}{n_i} \sum_{\rho_n<|j-i|\leq\rho_n^*} Y_j K_h(X_j - X_i)\right| \\ &= \mathrm{O}\left(\frac{n_i - n_i^*}{n_i}\right) = \mathrm{O}\left(\frac{\rho_n^*}{n}\right).\end{aligned}$$

The same maximization holds for $|\hat{f}_i(X_i) - \hat{f}_i^*(X_i)|$, and so we finally have, by Lemma 8,

$$\frac{|C_n(h) - C_n^*(h)|}{d_{\mathrm{A}}(h)} = \mathrm{O}(h\rho_n^*).$$

Using (C.9) the proof of Lemma 7 is complete.

Proof of Lemma 8. We prove the first part of this lemma. The second part is shown similarly. Let us define

$$\begin{aligned}d_{\mathrm{A}}^*(h) &= n^{-1} \sum_{i=1}^{n} \{r(X_i) - \hat{r}_h(X_i)\}^2 \left\{\frac{\hat{f}_h(X_i)}{f(X_i)}\right\}^2 w(X_i) \\ d_{\mathrm{I}}(h) &= \int \{r(u) - \hat{r}_h(u)\}^2 f(u) w(u)\, du \\ d_{\mathrm{I}}^*(h) &= \int \{r(u) - \hat{r}_h(u)\}^2 \frac{\hat{f}_h^2(u)}{f(u)} w(u)\, du \\ d_{\mathrm{M}}^*(h) &= E d_{\mathrm{I}}^*(h).\end{aligned}$$

We have to prove that

$$d_{\mathrm{M}}^*(h) = \frac{C_1}{nh} + C_2 h^4 + \mathrm{o}\{d_{\mathrm{A}}(h)\} \qquad \text{(A.23)}$$

and that we have for any d and d' among d_{A}, d_{I}, d_{A}^*, d_{I}^* and d_{M}^*

$$\sup_{h\in H_n} \frac{|d(h) - d'(h)|}{d(h)} = \mathrm{o}_p(1). \qquad \text{(A.24)}$$

Proof of (A.23). Let us denote by E' the expectation that would apply if the variables were independent, and by $d_{\mathrm{M}}^{*\prime}$ the quantity $d^*{}_{\mathrm{M}}$ that would apply in such a case. From Proposition 1 of Hart and Vieu (1990) we obtain

$$|d_{\mathrm{M}}^*(h) - d_{\mathrm{M}}^{*\prime}(h)| = \mathrm{O}\left\{(nh)^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha(|i-j|)\right\}.$$

We have

$$\sum_{i,j} \alpha(|i-j|) \leq 2 \sum_{i\leq j} \alpha(j-i) = 2 \sum_{k=0}^{n-1} k\alpha(k).$$

The last sum is bounded because of (C.8) and therefore we have

$$d^*{}_{\mathrm{M}}(h) = d_{\mathrm{M}}^{*}{}'(h) + \mathrm{o}\{(nh)^{-1}\}. \qquad \text{(A.25)}$$

Regression Smoothing Parameters That Are Not Far From Their Optimum

W. HÄRDLE, P. HALL, and J. S. MARRON*

It is well known that data-driven regression smoothing parameters $\hat{h}$ based on cross-validation and related methods exhibit a slow rate of convergence to their optimum. In an earlier article we showed that this rate can be as slow as $n^{-1/10}$; that is, for a bandwidth $\hat{h}_0$ optimizing the averaged squared error, $n^{1/10}(\hat{h} - \hat{h}_0)/\hat{h}_0$ tends to an asymptotic normal distribution. In this article we consider mean averaged squared error optimal bandwidths h_0. This (nonrandom) smoothing parameter can be approximated much faster. We use the technique of double smoothing to show that there is an $\hat{h}$ such that, under certain conditions, $n^{1/2}(\hat{h} - h_0)/h_0$ tends to an asymptotic normal distribution.

KEY WORDS: Automatic smoothing; Double smoothing; Kernel estimation; Nonparametric regression.

Data-driven smoothing methods are a necessary tool for a variety of statistical procedures. For a data analyst, a look at smoothed data often provides useful insight into features of the data. Many examples of this approach are given in Tukey (1947), Eubank (1988), Müller (1988) and Härdle (1990). For such applications an automated choice of the amount of smoothing is useful. For certain procedures, this automation is essential: Projection pursuit and additive model approaches to the analysis of high-dimensional data require repeated application of effective one-dimensional smoothing. Intensive use of one-dimensional smoothing is made by the *backfitting algorithm* for generalized additive models [see Hastie and Tibshirani (1986)]. In an earlier article (Härdle, Hall, and Marron 1988), we addressed the issue of how far an automated data driven smoothing parameter is away from its optimum. We showed there that smoothing parameters optimizing the averaged squared error can be approximated with the rate of $n^{-1/10}$. In the present work we improve on this by showing that the mean averaged squared error optimal smoothing parameter can be approximated with the much better rate of $n^{-1/2}$.

A nonparametric regression model with given *design* is formulated as

$$Y_i = m(x_i) + \epsilon_i, \qquad 1 \le i \le n,$$

where each $x_i \in (0, 1)$ and the errors are iid with mean zero and variance σ^2. Our goal is to estimate the curve $m(\cdot)$ from these n observations. In this article we use the Nadaraya–Watson kernel smoother

$$\hat{m}_h(x) = \frac{\Sigma_j Y_j K[(x - x_j)/h]}{\Sigma_j K[(x - x_j)/h]},$$

with a kernel function K. It is well known that the statistical precision of this estimator crucially depends on the *bandwidth h*. A common measure of accuracy for studying the influence of varying h on how close $\hat{m}_h$ is to m is the Mean Averaged Squared Error (MASE); see Härdle, Hall, and Marron (1988). It is the aim of a practitioner to find a good bandwidth h when using the preceding kernel smoother for approximating $m(\cdot)$.

A matter that inevitably arises in this context is that of apportioning the smoothness assumptions between the curve estimation part of the problem and the bandwidth estimation part. In theory this is a perplexing issue since there are no clear empirical guidelines for resolving it. In practice, however, the order of the kernel used for the curve estimation part would usually be determined by prior preferences (e.g., a disinclination to use high-order kernels having negative side lobes) and by the need to ensure good performance in problems involving small samples or high error variances (since high-order kernels exacerbate the problem of variability). Thus, we contend, practitioners would very often take kernels to be of second-order kernel, even in the fact of evidence suggesting that the mean function has considerably more than two derivatives.

Widely studied methods for automatic smoothing parameter selection include cross-validation, Generalized Cross Validation (GCV), Akaike's Information Criterion (AIC), and C_p; see Härdle, Hall, and Marron (1988) for definitions and references. In that article it was demonstrated that each of these methods is subject to an unacceptably large amount of across-sample variability, which is probably why the methods have not become widely used data-analytic tools. In this article the technique of double smoothing is used to provide a data-driven smoothing parameter that has much better stability properties than these earlier methods. The main features of double smoothing and theoretical properties will be discussed in Section 1, and the results of numerical trials of double smoothing will be described in Section 2.

1. DOUBLE SMOOTHING

For the technique of double smoothing we need two different kernel smoothers of the form

$$\hat{m}(x) = \hat{m}_h(x) = \frac{\Sigma_j Y_j K[(x - x_j)/h]}{\Sigma_j K[(x - x_j)/h]} = \frac{n^{-1}\Sigma_j Y_j K_h(x - x_j)}{\hat{f}_h(x; K)}.$$

* W. Härdle is Professor, CORE, Université Catholique de Louvain, B-1348 Louvain-la-Neuve, Belgium. P. Hall is Professor, Department of Statistics, Australian National University, Canberra ACT 2601, Australia. J. S. Marron is Professor, Department of Statistics, University of North Carolina, Chapel Hill, NC 27514, U.S.A. This research supported by Deutsche Forschungsgemeinschaft, SFB 303, and by CORE.

Journal of the American Statistical Association
March 1992, Vol. 87, No. 417, Theory and Methods

Here $\hat{f}_h(x; K)$ denotes a density estimate with kernel K, and $K_h(\cdot) = h^{-1}K(\cdot/h)$. The double smoothing operation is performed with a second kernel smoother with different kernel L and different bandwidth g:

$$\hat{m}_g(x) = \frac{\Sigma_j Y_j L[(x - x_j)/g]}{\Sigma_j L[(x - x_j)/g]} = \frac{n^{-1}\Sigma_j Y_j L_g(x - x_j)}{\hat{f}_g(x; L)}.$$

Note that we abbreviated the kernel smoother $\hat{m}_h(x)$ as $\hat{m}(x)$. We assume that the kernels are of orders r and s respectively, that is, the kernel constants

$$\kappa_r = (-1)^r(r!)^{-1}\int u^r K(u)\,du$$

and

$$\lambda_s = (-1)^s(s!)^{-1}\int u^s L(u)\,du$$

are not equal to zero and moments between orders 1 and r or s, respectively, vanish.

We consider the MASE as a distance between $\hat{m}(x)$ and $m(x)$,

$$M = M(h) = n^{-1}\sum_i^{*} E[\hat{m}(x_i) - m(x_i)]^2,$$

where Σ_i^* denotes summation over indices i such that $c < x_i < d$, where $0 < c < d < 1$. (The reason for this restricted range of summation is the occurrence of enlarged bias near the boundary.) It is well known that the MASE splits up into a stochastic and a squared bias part. At each point x, the bias is given by

$$b(x) = E[\hat{m}(x)] - m(x). \quad (1.1)$$

The idea of double smoothing is to estimate this bias using the second kernel smooth $\hat{m}_g$. More precisely, we are going to estimate $b(x)$ by

$$\hat{b}(x) = \frac{\Sigma_j \hat{m}_g(x_j)K[(x - x_j)/h]}{\Sigma_j K[(x - x_j)/h]} - \hat{m}_g(x). \quad (1.2)$$

Note that this bias estimate can be thought of as an iterated smoothing algorithm. The pilot smooth $\hat{m}_g(x)$ (with kernel L and bandwidth g) is resmoothed with kernel K and bandwidth h. A similar bias estimate has been employed in Härdle and Marron (1991) in a bootstrap technique for constructing confidence bands.

The stochastic part of $M(h)$ is defined by

$$V = V(h) = n^{-1}\sum_i^{*} \mathrm{var}\,[\hat{m}(x_i)]$$
$$= \sigma^2 n^{-1}\sum_i^{*} n^{-2}\sum_j K_h^2(x_i - x_j)\{\hat{f}_h(x_i; K)\}^{-2}.$$

The systematic bias part is $B = B(h) = n^{-1}\,\Sigma_i^*\, b(x_i)^2$. To simplify the formula for the estimated bias $\hat{b}$, put

$$A_j(x) = n\sum_k K[(x - x_k)/h]\Bigg\{L[(x_k - x_j)/g] \times \Big\{\sum_l L[(x_k - x_l)/g]\Big\}^{-1} - L[(x - x_j)/g]\Big\{\sum_l L[(x_k - x_l)/g]\Big\}^{-1}\Bigg\} \times \Big\{\sum_l K[x - x_l)/h]\Big\}^{-1}.$$

Then the estimated bias can be written more simply as

$$\hat{b}(x) = n^{-1}\sum_j Y_j A_j(x).$$

The quantity

$$\hat{B} = \hat{B}(h) = n^{-1}\sum_i^{*} \hat{b}(x_i)^2$$

estimates B. It will turn out later that there is a variance term in this estimate, $n^{-3}\sigma^2\,\Sigma_i^*\,\Sigma_j\, A_j(x_i)^2$, that must be allowed for. Therefore, by subtracting this variance twice, we estimate B by

$$\hat{B} - n^{-3}\hat{\sigma}^2\sum_i^{*}\sum_j A_j(x_i)^2,$$

where $\hat{\sigma}^2$ is an estimator of σ^2. The variance V is estimated by

$$\hat{V} = \hat{\sigma}^2 n^{-1}\sum_i^{*}\sum_j K[(x_i - x_j)/h]^2\Big\{\sum_k K[(x_i - x_k)/h]\Big\}^{-2}.$$

Hence our final estimate of M is

$$\hat{M} = \hat{V} + \hat{B} - n^{-3}\hat{\sigma}^2\sum_i^{*}\sum_j A_j(x_i)^2 = \hat{V}_1 + \hat{B},$$

where

$$\hat{V}_1 = \hat{\sigma}^2 n^{-1}\sum_i^{*}\sum_j \times \Bigg\{K[(x_i - x_j)/h]^2\Big[\sum_k K[(x_i - x_k)/h]\Big]^{-2} - n^{-2}A_j(x_i)^2\Bigg\}.$$

Let h_0 denote the minimizer of M. It is our goal to estimate this h_0 as well as possible. Our proposal is to use the minimizer of $\hat{M}(h)$ an estimate of $M(h)$. We call this data-driven bandwidth $\hat{h}$. Before discussing asymptotic properties of $\hat{h}$, we state some assumptions.

Assumption 1. K and L are compactly supported kernels of orders r and s, respectively; K' and $L^{(r+1)}$ are bounded.

Assumption 2. Let $r' = \max\,(r, s)$. Assume that $m^{(r+r')}$ is continuous on (0, 1).

Assumption 3. Assume $\hat{\sigma}^2$ is $\sqrt{n}$ consistent for σ^2, that is, $\hat{\sigma}^2 = \sigma^2 + O_p(n^{-1/2})$.

The availability of a $\sqrt{n}$-consistent estimator of σ^2 is well known; see, for example, Hall and Marron (1989a) or Gasser, Sroka, and Jennen-Steinmetz (1986).

It can be shown [as in Härdle, Hall, and Marron (1988)]

(1992) Härdle, W., Hall, P. and Marron, J.S.
Regression smoothing parameters that are not far from their optimum.

that, under Assumptions 1–3, there exist positive constants c_1 and c_2 such that

$$M''(h_0) \approx c_1(nh_0^3)^{-1} \approx c_2 h_0^{2r-2}.$$

Define

$$\gamma_1 = c_1^{-1}(d-c)\int K^2,$$

$$\gamma_2 = 2c_2^{-2}(d-c)r^2\kappa_r^4\sigma^4 \int\left[\int L^{(r)}(y)L^{(r)}(y+z)\,dy\right]^2 dz,$$

$$\gamma_3 = 4c_2^{-2}r^2\kappa_r^4\sigma^2\int_c^d (m^{(2r)})^2,$$

and

$$\gamma_4 = -4c_2^{-1}r\kappa_r^2\lambda_s\int_c^d m^{(r+s)}m^{(r)}.$$

The following result gives an expression of $\hat h$ in terms of $\hat\sigma$ and then bandwidths h and g. The proof is deferred to the Appendix.

Theorem 1. Under the assumptions above,

$$(\hat h - h_0)/h_0 = \gamma_1(\hat\sigma^2-\sigma^2) + (\gamma_2 n^{-2}g^{-(4r+1)} + \gamma_3 n^{-1})^{1/2}Z_n + \gamma_4 g^s + o(g^s), \qquad (1.3)$$

where Z_n is asymptotically normal $N(0, 1)$. If $g^{-(4r+1)} = o(n)$, then we may replace the term $(\gamma_2 n^{-2}g^{-(4r+1)} + \gamma_3 n^{-1})^{1/2}Z_n$ by $(-1)^{r+1}(\gamma_3 n^{-1})^{1/2}Z_n^*$, where

$$Z_n^* = \left[n\sigma^2\int_c^d (m^{(2r)})^2\right]^{-1/2} \sum_j^{*} m^{(2r)}(x_j)\epsilon_j$$

and is asymptotically normal $N(0, 1)$.

Remark 1. The last part of the theorem enables us to compute the asymptotic variance of $(\hat h - h_0)/h_0$ in cases of $\sqrt{n}$ consistency as follows. Suppose that

$$n^{-2}g^{-(4r+1)} + g^{2s} = o(n^{-1}) \qquad (1.4)$$

(see Remark 2 for a discussion of the circumstances under which this is attainable), and assume for the sake of definiteness that $\hat\sigma^2 = (1/2n)\sum_{j=2}^n (Y_j - Y_{j-1})^2$. Then

$$\hat\sigma^2 - \sigma^2 = \frac{1}{2n}\sum_{j=2}^n[(\epsilon_j-\epsilon_{j-1})^2 - 2\sigma^2] + o_p(n^{-1/2)}$$

$$= \frac{1}{n}\sum_{j=1}^n(\epsilon_j^2-\sigma^2) - \frac{1}{n}\sum_{j=2}^n \epsilon_j\epsilon_{j-1} + o_p(n^{-1/2}).$$

Hence

$$(\hat h - h_0)/h_0 = \frac{1}{n}\sum_{j=1}^n[\gamma_1(\epsilon_j^2-\sigma^2) + (-1)^{r+1}2c_2^{-1}r\kappa_r^2 m^{(2r)}(x_j)I(c<x_j<d)\epsilon_j] - \frac{1}{n}\gamma_1\sum_{j=1}^n\epsilon_j\epsilon_{j-1} + o_p(n^{-1/2}),$$

whence it follows that $n^{1/2}(\hat h - h_0)/h_0$ is asymptotically normal with zero mean and variance

$$\gamma_1^2E(\epsilon^4) + \gamma_3 + \gamma_5,$$

where

$$\gamma_5 = (-1)^{r+1}4\gamma_1E(\epsilon^3)C_2^{-1}r\kappa_r^2[m^{(2r-1)}(d) - m^{(2r-1)}(c)].$$

Remark 2. Condition (2.4) holds if $s \ge 2r+1$ and g is of larger order than $n^{-1/(4r+1)}$ but of smaller order than $n^{-1/2s}$. In this circumstance we may deduce immediately from (2.3) that $(\hat h - h_0)/h_0 = O_p(n^{-1/2})$.

Remark 3. Here we discuss optimal choice of g and the convergence rate of $\hat h$ in the case $r = s$. There, the terms involving g in (1.3) are balanced when $(n^{-2}g^{-(4r+1)})^{1/2}$ is of the same size as g^r. That demands that g be of size $n^{-2/(6r+1)}$. Then both $(n^{-2}g^{-(4r+1)})^{1/2}$ and g^r are of size $n^{-2r/(6r+1)}$, which is of larger order than $n^{-1/2}$. Hence, by (1.3),

$$(\hat h - h_0)/h_0 \simeq (\gamma_2 n^{-2}g^{-(4r+1)})^{1/2}Z_n + \gamma_4 g^r.$$

The asymptotic mean square error of the right side equals

$$\gamma_2^2 n^{-2}g^{-(4r+1)} + \gamma_4^2 g^{2r}$$

and is minimized by taking $g = Cn^{-2/(6r+1)}$, where

$$C = [(4r+1)\gamma_2^2/(2r\gamma_4^2)]^{1/(6r+1)}.$$

For this choice of g,

$$(\hat h - h_0)/h_0 = [(\gamma_2 C^{-(4r+1)})^{1/2}Z_n' + \gamma_4C^r]n^{-2r/(6r+1)},$$

where Z_n' is asymptotically normal $N(0, 1)$. In particular, the convergence rate of $(\hat h - h_0)/h_0$ is $n^{-2r/(6r+1)}$, which reduces to $n^{-4/13}$ when $r = 2$.

Remark 4. A version of double smoothing in the context of nonparametric density estimation has been discussed by Hall and Marron (1989b). There the technique may be viewed as a smoother form of cross-validation. However, the links between double smoothing and cross-validation are much more tenuous in the case of nonparametric regression.

2. THE METHOD IN PRACTICE

The proposed method of double smoothing has been investigated in a simulation study. As the regression function we have chosen the one also used by McDonald and Owen (1986),

$$m(x) = \sin^3(2\pi x^3), \qquad 0 < x < 1.$$

Observations Y_i were taken at $x_i = (i - 1/2)/n$, for $n = 50$, with ϵ_i normal $(0, \sigma^2)$ and with different σ. As the kernel K we selected the quartic kernel, which is of order (0, 2, 2) in the terminology of Müller (1988, table 5.7). In our notation, K is of order $r = 2$. As can be seen from the preceding theory, the kernel L should be chosen of higher order s. From table 5.7 of Müller, we have selected the kernel

$$L(u) = (105/64)(1 - 5u^2 + 7u^4 - 3u^6)I(|u| \le 1),$$

which is of order (0, 4, 2); in our notation, $s = 4$. For

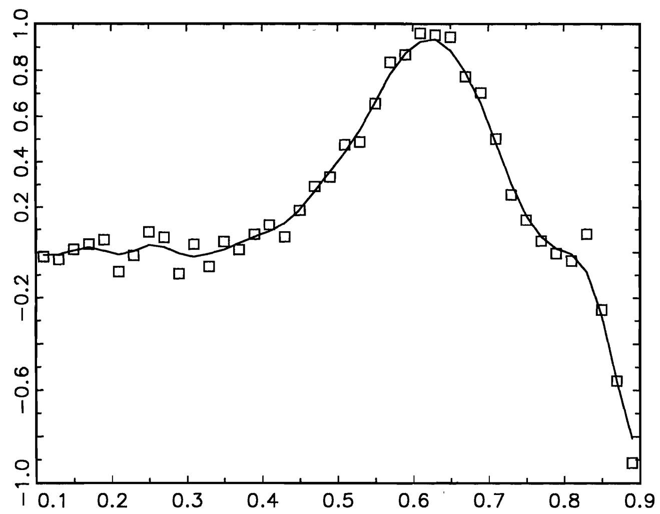


Figure 1. The Curve $\hat{m}_{h_0}(x)$ With MASE Optimal h_0, the Observations (x_i, Y_i), and Normal Errors With $\sigma = .05$.

practical reasons, we also investigate the choice of $K = L$. To give some insight into the shape of this curve, consider Figure 1. It shows the observations for one of the data sets together with the estimated curve evaluated at the points x_i between c and d. One sees that the estimated curve is able to model the hardly visible inflection point at $x = .8$.

The distance measure $M(h)$ for this setting is shown in Figure 2. We actually estimated $M(h)$ by Monte Carlo integration; that is, we averaged the averaged squared error $n^{-1} \Sigma_j^* (\hat{m}(x_j) - m(x_j))^2$ over repeated $(N = 50)$ simulated samples. The cutoff points were chosen as $c = .1$, $d = .9$. The variability of this (stochastic) measure of accuracy is made visible through the dotted lines in Figure 2, denoting two standard deviations (2 * STD) bands around $M(h)$. The bandwidth minimizing $M(h)$ is $h_0 = .054$. We could have computed M directly, by integration, but chose not to, since the error bands in Figure 2 provide an illuminating comparison with those for $\hat{M}$. See Figures 3 and 4.

For the sake of economical presentation, it is convenient, in our figures, to fix a relationship between g and h. There is no "canonical" way of relating these two bandwidths, since they are for quite different kernels with quite different scales. See Marron and Nolan (1988). Empirically, we found that for the particular distribution and sample size under investigation, the ratio $g/h \doteq 1/2$ gave reasonable results, and so we present that case.

For each of the $N = 50$ Monte Carlo samples, we estimated the curve M by $\hat{M}$. Figure 3 shows the variation of the $N = 50$ $\hat{M}$ curves around its means for $g = h/2$ when g is varied with h.

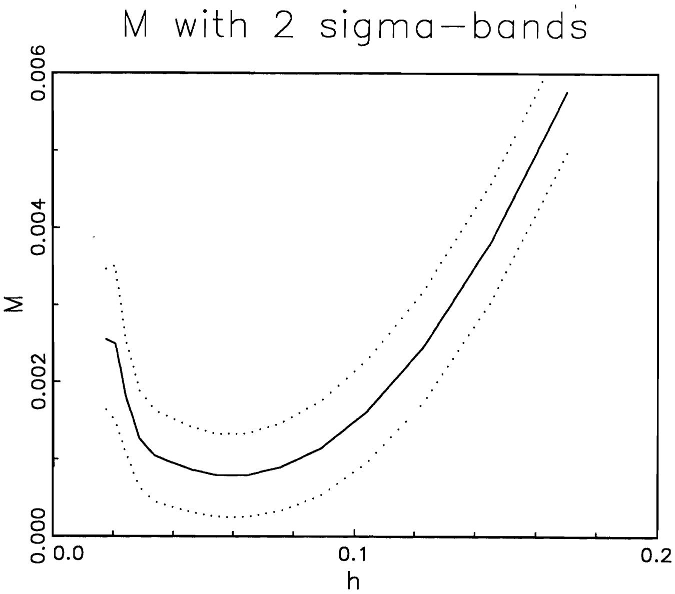


*Figure 2. The Mean Averaged Squared Error M(h) as a Function of h. The dotted lines denote 2 * STD Bands.*

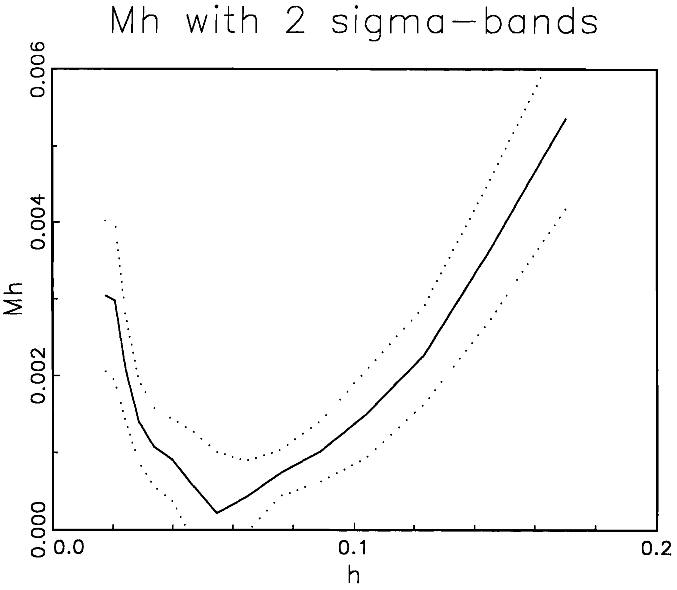


*Figure 3. The Variation of the $\hat{M}$ Curves Around Their Means. The solid line is the (Monte Carlo) average of $\hat{M}$, and the dotted curves are the 2 * STD bands around it.*

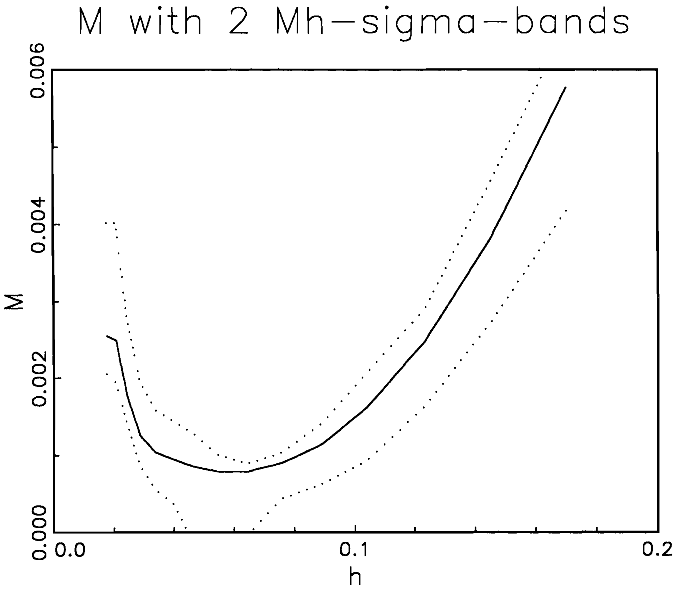


*Figure 4. The Curve M(h) and the 2 * STD Bands from Figure 3*

(1992) Härdle, W., Hall, P. and Marron, J.S.
Regression smoothing parameters that are not far from their optimum.

Journal of the American Statistical Assoication, 87, 227-233

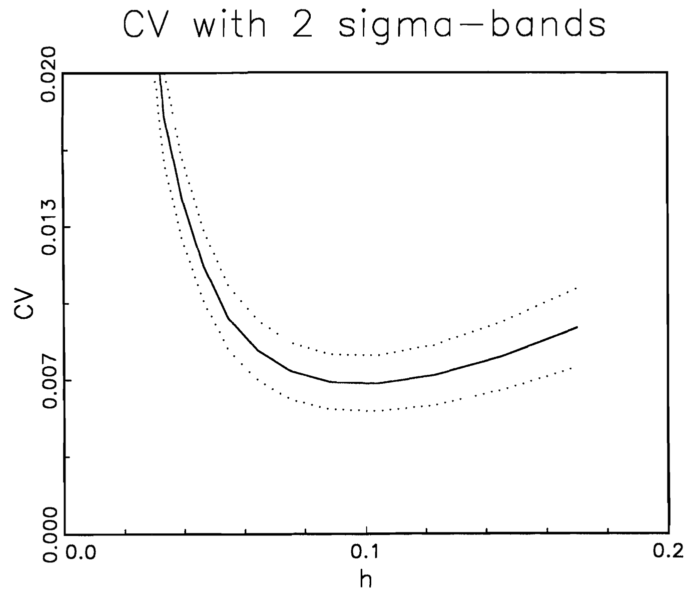


*Figure 5. The Average CV Curve and 2 * STD Bands.*

One sees that the averaged $\hat{M}$ curve has a clear minimum at the point where M is also minimized. Of course, these are only the averaged $\hat{M}$ curves. To give an impression of how $\hat{M}$ varies across the Monte Carlo setting, we added 2 * STD bands to this curve. These same bands overlayed with the mean averaged squared error are shown in Figure 4. One sees again that $\hat{M}$ approaches M.

The often used cross validation (CV) method is shown in Figure 5. The same display techniques as for the other figures is used. The minimum of the averaged $CV(h) = n^{-1} \Sigma_i^* (Y_i - \hat{m}_{h,i}(x_i))^2$ ($\hat{m}_{h,i}$ is a leave-one-out smoother) is at $h = 0.085$, quite far away from the optimal h_0. This behavior of $CV(h)$ confirms our experience, which was discussed in our earlier article, Härdle, Hall, and Marron (1988). From looking at Figure 1 we can imagine why $CV(h)$ will oversmooth the data. The small shoulder at the right end will, with the leave-one-out-method, be treated as part of the noise structure.

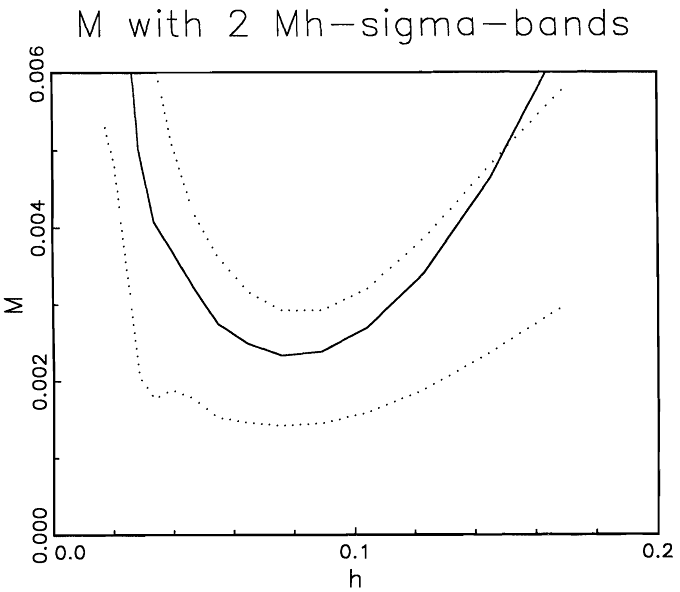


*Figure 6. The Curve M(h), $\sigma^2 = (.1)^2$, and 2 * STD Bands.*

Table 1. Results of the Monte Carlo Simulation for N = 100

Interval for $M(\hat{h})/M(h_0)$	*1*	*1.33*	*1.66*	*2*	*2.33*
1.00	43	41	31	5	1
1.00–1.01	8	8	5	3	1
1.01–1.02	6	5	2	3	2
1.02–1.03	8	6	3	1	1
1.03–1.04	8	7	4	7	5
1.04–1.05	1	1	1	1	1
1.05–1.10	9	7	10	13	6
1.10–1.20	12	16	24	33	19
1.20–1.30	1	3	10	19	15
1.30–1.40	1	2	4	7	5
1.40–1.50	3	4	5	7	9
1.50–2	0	0	1	1	21
>2	0	0	1	1	14

The effect of increasing σ^2 and g is shown in Figure 6. This plot is constructed in the same way as Figure 4, but now with $\sigma = .1$ and $g = h$. Note that this setting has two effects. First, the 2 * STD bands are enlarged, and, second, the too-large bandwidth g causes the right branch of $\hat{M}(h)$ to be flatter in Figure 4. This has also been observed in our experiments with $\sigma = .05$, but we do not report this here. Of course, the minimum of $M(h)$ lies more to the right than in Figure 4 since σ^2 is increased.

In practical studies one would probably choose $K = L$, for the sake of convenience and because K and L would often both be selected as symmetric density functions. Moreover, one would leave g at a fixed value. We have done this with the quartic kernel by varying (over α) the bandwidth $g = \alpha h_0$; $h_0 = 0.054$ (α independently of h). The results are presented in Table 1. The figures in this table show how many times out of the $N = 100$ Monte Carlo simulations the ratio $M(h)/M(h_0)$ was in the interval indicated in the first column.

It is obvious that choosing a too-high g, leading to an extremely oversmoothed bias estimate, has the effect of shifting the data driven $\hat{h}$ to the right.

A criticism that can reasonably be leveled at our procedure is that it requires selection of the initial smoothing parameter g. This difficulty arises with all other $\sqrt{n}$-consistent methods of which we are aware, and there seems to be no entirely satisfactory way of removing it. However, the arbitrariness can be eliminated by specifying g by a formula such as $g = n^{-c}$, for an appropriate constant c. Such a choice allows good asymptotic performance but is not necessarily appropriate for real, finite data sets. In practice, there appears to be no substitute for trying a small number of different g's.

APPENDIX: PROOF OF THEOREM 1

Assume that K and L vanish outside $(-C, C)$. In the case of fixed design, if $x = m/n$ for an integer m and for $c < x < d$, then $K[(x - x_k)/h] = 0$ unless $c - Ch < x_k < c + Ch$. And, if x_k does satisfy this constraint, then $L[(x_k - x_l)/g] = 0$ unless c

(1992) Härdle, W., Hall, P. and Marron, J.S.
Regression smoothing parameters that are not far from their optimum.

$-C(h+g) < x_l < d + C(h+g)$. It follows that, for $x = m/n$, $c < x < d$, and large n, the terms

$$nh_1 = \sum_{l=1}^{n} K[(x - x_l)/h]$$

and

$$ng_1 = \sum_{l=1}^{n} L[(x_k - x_l)/g] = \sum_{l=1}^{n} L[(x - x_l)/g]$$

appearing in the definition of $A_j(x)$ do not depend on x or x_k. Therefore,

$$A_j(x) = (nh_1g_1)^{-1} \sum_{k=1}^{n} K[(x - x_k)/h] \times \{L[(x_k - x_j)/g] - L[(x - x_j)/g]\}.$$

Observe that

$$M(h) = n^{-1} \sum_{i}^{*} E[\hat{m}(x_i) - m(x_i)]^2$$

$$= n^{-1}(nh_1)^{-2}\sigma^2 \sum_{i}^{*} \sum_{j=1}^{n} K[(x_i - x_j)/h]^2 + B(h),$$

and so

$$\hat{M}(h) - M(h) = (\hat{\sigma}^2 - \sigma^2)D_1(h) + D_2(h), \qquad (A.1)$$

where

$$D_1(h) = n^{-1}(nh_1)^{-2} \sum_{i}^{*} \sum_{j=1}^{n} K[(x_i - x_j)/h]^2 - n^{-3} \sum_{i}^{*} \sum_{j=1}^{n} A_j(x_i)^2$$

and

$$D_2(h) = \hat{B}(h) - B(h) - n^{-3}\sigma^2 \sum_{i}^{*} \sum_{j=1}^{n} A_j(x_i)^2.$$

Since $Y_i = m(x_i) + \epsilon_i$, then

$$D_2 = n^{-3} \sum_{i}^{*} \left\{ \sum_{j=1}^{n} [m(x_j) + \epsilon_j]A_j(x_i) \right\}^2 - n^{-1} \sum_{i}^{*} b(x_i)^2 - n^{-3}\sigma^2 \sum_{i}^{*} \sum_{j=1}^{n} A_j(x_i)^2$$

$$= T_1 + T_2 + T_3 + 2T_4, \qquad (A.2)$$

where

$$T_1 = n^{-1} \sum_{i}^{*} \left\{ \left[n^{-1} \sum_{j=1}^{n} m(x_j)A_j(x_i) \right]^2 - b(x_i)^2 \right\},$$

$$T_2 = n^{-3} \sum_{i}^{*} \sum_{j=1}^{n} (\epsilon_j^2 - \sigma^2)A_j(x_i)^2,$$

$$T_3 = n^{-3} \sum_{i}^{*} \sum_{j \neq k} \sum \epsilon_j \epsilon_k A_j(x_i)A_k(x_i),$$

and

$$T_4 = n^{-3} \sum_{i}^{*} \sum_{j=1}^{n} \sum_{k=1}^{n} \epsilon_j m(x_k)A_j(x_i)A_k(x_i).$$

From this point we only sketch the proof, with the aim to explain to the reader why the main terms admit the asymptotic formulas that we claim for them. Our argument is readily made rigorous, although at the expense of considerable additional algebra. Since $\hat{M}'(\hat{h}) = M'(h_0) = 0$, then

$$0 = (\hat{M} - M)'(\hat{h}) + M'(\hat{h})$$

$$= (\hat{M} - M)'(h_0) + (\hat{h} - h_0)M''(h_0) + o[(\hat{h} - h_0)/h_0].$$

Hence

$$\hat{h} - h_0 \simeq -[\hat{M}'(h_0) - M'(h_0)]/M''(h_0), \qquad (A.3)$$

and so we must investigate the asymptotic properties of $\hat{M}' - M'$. In view of (A.1) and (A.2),

$$\hat{M}' - M' = (\hat{\sigma}^2 - \sigma^2)D_1' + T_1' + T_2' + T_3' + 2T_4'. \qquad (A.4)$$

As a prelude to examining the terms on the right side, we next develop approximate formulas for a number of series. Note that

$$h_1 \simeq \int K[(x - y)/h]\, dy = h,$$

$$g_1 \simeq \int L[(x - y)/g]\, dy = g,$$

and

$$A_j(x) \simeq \bar{A}_j(x) \equiv g^{-1} \int K(y) \times \{L[g^{-1}(x - x_j) - hg^{-1}y] - L[g^{-1}(x - x_j)]\}\, dy.$$

Put

$$\alpha_j = n^{-1} \sum_{i}^{*} A_j(x_i)^2, \qquad \bar{\alpha}_j = n^{-1} \sum_{i}^{*} \bar{A}_j(x_i)^2,$$

$$\alpha_{jk} = n^{-1} \sum_{i}^{*} A_j(x_i)A_k(x_i), \qquad \bar{\alpha}_{jk} = n^{-1} \sum_{i}^{*} \bar{A}_j(x_i)\bar{A}_k(x_i),$$

$$\beta(x) = n^{-1} \sum_{k=1}^{n} m(x_k)A_k(x), \qquad \bar{\beta}(x) = n^{-1} \sum_{k=1}^{n} m(x_k)\bar{A}_k(x),$$

$$\beta_j = n^{-2} \sum_{i}^{*} \sum_{k=1}^{n} m(x_k)A_j(x_i)A_k(x_i),$$

and

$$\bar{\beta}_j = n^{-2} \sum_{i}^{*} \sum_{k=1}^{n} m(x_k)\bar{A}_j(x_i)\bar{A}_k(x_i).$$

We first develop approximations to $\bar{\alpha}_j$, $\bar{\alpha}_{jk}$, ..., and use those results to approximate α_j, α_{jk}, ... Since

$$\bar{A}_j(x) \simeq \kappa_r g^{-1}(h/g)^r L^{(r)}[g^{-1}(x - x_j)],$$

then

$$\bar{\alpha}_j \simeq \kappa_r^2 g^{-1}(h/g)^{2r} \int (L^{(r)})^2 I(c < x_j < d),$$

$$\bar{\alpha}_{jk} \simeq \int_c^d [\kappa_r g^{-1}(h/g)^r]^2 L^{(r)}[g^{-1}(x - x_j)]L^{(r)}[g^{-1}(x - x_k)]\, dx$$

$$\simeq \kappa_r^2 g^{-1}(h/g)^{2r} \int L^{(r)}(y)L^{(r)}[y + g^{-1}(x_j - x_k)]\, dy \times I(c < x_j, x_k < d),$$

$$\bar{\beta}(x) \simeq g^{-1} \iint m(z)K(y)\{L[g^{-1}(x - z) - hg^{-1}y] - L[g^{-1}(x - z)]\}\, dy\, dz$$

$$= \iint [m(x - gu - hy) - m(x - gu)]K(y)L(u)\, dy\, du$$

$$\simeq \kappa_r h^r m^{(r)}(x),$$

$$\bar{\beta}_j = n^{-1} \sum_{i}^{*} \bar{A}_j(x_i)\bar{\beta}(x_i)$$

$$\simeq \int_c^d \bar{A}_j(x)\bar{\beta}(x)\, dx$$

$$\simeq (-1)^r \kappa_r^2 h^{2r} m^{(2r)}(x_j) I(c < x_j < d)s,$$

$$\bar{\alpha}_j'(h) \simeq 2r\kappa_r^2 h^{2r-1} g^{-(2r+1)} \int (L^{(r)})^2 I(c < x_j < d), \qquad (A.5)$$

$$n^{-2}\sum_{j\neq k}\sum \bar{\alpha}'_{jk}(h)^2 \simeq 4(d-c)r^2\kappa_r^4h^{4r-2}g^{-(4r+1)}$$
$$\times \int\left[\int L^{(r)}(y)L^{(r)}(y+z)\,dy\right]^2 dz, \qquad (A.6)$$

$$n^{-1}\sum_j \bar{\beta}'_j(h)^2 \simeq 4r^2\kappa_r^4h^{4r-2}\int_c^d (m^{(2r)})^2, \qquad (A.7)$$

$$\bar{\beta}(x)-\beta(x) \simeq \kappa_r\lambda_s h^r g^s m^{(r+s)}(x),$$

$$n^{-1}\sum_i^{*} \bar{\beta}(x_i)^2 - n^{-1}\sum_i^{*} b(x_i)^2 \simeq 2\kappa_r\lambda_s h^{2r}g^s\int_c^d m^{(r+s)}m^{(r)}, \qquad (A.8)$$

$$n^{-1}\sum_i^{*}\bar{\alpha}_j \simeq \kappa_r^2(d-c)h^{2r}g^{-(2r+1)}\int (L^{(r)})^2,$$

and

$$D_1(h) \simeq (nh)^{-1}(d-c)\left[\int K^2 - \kappa_r^2\left(\frac{h}{g}\right)^{2r+1}\int (L^{(r)})^2\right]. \qquad (A.9)$$

Define

$$C_{01} = (d-c)\int K^2, \qquad C_{02} = 2r(d-c)\kappa_r^2\int (L^{(r)})^2,$$

$$C_1 \simeq 4r\kappa_r^2\lambda_s\int_c^d m^{(r+s)}m^{(r)},$$

$$C_3 \simeq 2(d-c)r^2\kappa_r^4\sigma^4\int\left[\int L^{(r)}(y)L^{(r)}(y+z)\,dy\right]^2 dz,$$

and

$$C_4 \simeq 4r^2\kappa_r^4\sigma^2\int_c^d (m^{(2r)})^2.$$

Noting (A.8), we have

$$\frac{\partial T_1}{\partial h} = \frac{\partial}{\partial h}\left[n^{-1}\sum_i^{*}\beta(x_i)^2 - n^{-1}\sum_i^{*} b(x_i)^2\right]$$
$$\simeq \frac{\partial}{\partial h}\left[n^{-1}\sum_i^{*}\bar{\beta}(x_i)^2 - n^{-1}\sum_i^{*} b(x_i)^2\right]$$
$$\simeq C_1h^{2r-1}g^s; \qquad (A.10)$$

noting (A.9),

$$\partial/\partial hD_1(h) = -(nh^2)^{-1}[C_{01} + C_{02}(h/g)^{2r+1}]; \qquad (A.11)$$

noting (A.5),

$$\operatorname{var}(\partial T_2/\partial h) = O\left[n^{-4}\sum_{j=1}^n \alpha'_j(h)^2\right]$$
$$= O\left[n^{-4}\sum_{j=1}^n \bar{\alpha}'_j(h)^2\right]$$
$$= O(n^{-3}h^{4r-2}g^{-(4r+2)}); \qquad (A.12)$$

noting (A.6),

$$\operatorname{var}(\partial T_3/\partial h) = 4\sigma^4n^{-4}\sum_{j<k}\sum \alpha'_{jk}(h)^2 \simeq 4\sigma^4n^{-4}\sum_{j<k}\sum \bar{\alpha}'_{jk}(h)^2$$
$$\simeq C_3n^{-2}h^{4r-2}g^{-(4r+1)}; \qquad (A.13)$$

and, noting (A.7),

$$\operatorname{var}(\partial T_4/\partial h) = \sigma^2n^{-2}\sum_{j=1}^n \beta'_j(h)^2 \simeq \sigma^2n^{-2}\sum_{j=1}^n \bar{\beta}'_j(h)^2$$
$$\simeq C_4n^{-1}h^{4r-2}. \qquad (A.14)$$

Since $ng \to \infty$, then, by (A.12) and (A.13),

$$\operatorname{var}(\partial T_2/\partial h) = o[\operatorname{var}(\partial T_3/\partial h)].$$

The variables $\partial T_3/\partial h$ and $\partial T_4/\partial h$ are symptotically independent and normally distributed with zero means and their respective variances, and so, by (A.4), (A.10), (A.11), (A.12), and (A.13),

$$\hat{M}' - m' \simeq (\hat{\sigma}^2 - \sigma^2)(nh^2)^{-1}[C_{01} + C_{02}(h/g)^{2r+1}]$$
$$+ h^{2r-1}[c_1g^s + (C_3n^{-2}g^{-(4r+1)} + C_4n^{-1})^{1/2}Z_n],$$

where Z_n is asymptotically normal $N(0, 1)$.

Finally we return to formula (A.3), take $h = h_0$, and observe that since $h/g \to 0$ and $M''(h_0) \simeq c_1(nh_0^3)^{-1} \simeq c_2h_0^{2r-2}$,

$$(\hat{h} - h_0)/h_0 \simeq (\hat{\sigma}^2 - \sigma^2)c_1^{-1}C_{01} - c_2^{-1}$$
$$\times [c_1g^s + (C_3n^{-2}g^{-(4r+1)} + C_4n^{-1})^{1/2}Z_n].$$

The theorem follows from this formula. Should $g^{-(4r+1)} = o(n)$, then the term $C_3n^{-2}g^{-(4r+1)}$ is asymptotically negligible relative to C_4n^{-1} and so may be dropped. Then, $-c_2^{-1}(C_4n^{-1})^{1/2}Z_n$ derives from

$$-[M''(h_0)h_0]^{-1}T'_4(h) = -c_2^{-1}h_0^{-2r+1}n^{-1}\sum_{j=1}^n \beta'_j(h)\epsilon_j$$
$$\simeq c_2^{-1}(-1)^{r+1}2r\kappa_r^2n^{-1}\sum_i^{*} m^{(2r)}(x_i)\epsilon_i$$
$$= (-1)^{r+1}(\gamma_4n^{-1})^{1/2}Z_n^*.$$

[*Received May 1990. Revised January 1991.*]

REFERENCES

Eubank, R. (1988), *Spline Smoothing and Nonparametric Regression,* New York: Marcel Dekker.

Gasser, T., Sroka, L., and Jennen-Steinmetz, C. (1986), "Residual Variance and Residual Pattern in Nonlinear Regression," *Biometrika,* 73, 625–633.

Hall, P., and Marron, J. S. (1989a), "On Variance Estimation and Nonparametric Regression," manuscript, University of North Carolina, Chapel Hill.

——— (1989b), "Smoothed Cross-Validation," manuscript, University of North Carolina, Chapel Hill.

Härdle, W. (1990), *Applied Nonparametric Regression,* Cambridge, MA: Cambridge University Press.

Härdle, W., and Marron, J. S. (1991), "Bootstrap Simultaneous Error Bars for Nonparametric Regression," *The Annals of Statistics,* 19, 778–796.

Härdle, W., Hall, P., and Marron, J. S. (1988), "How Far Are Automatically Chosen Regression Smoothing Parameters From Their Optimum?" (with discussion), *Journal of the American Statistical Association,* 83, 86–99.

Hastie, T., and Tibshirani, R. (1986), "Generalized Additive Models" (with discussion), *Statistical Science,* 1, 297–318.

Marron, J. S., and Nolan, D. (1988), "Canonical Kernels for Density Estimation," *Statistics and Probability Letters,* 7, 195–199.

McDonald, J. A., and Owen, A. (1986), "Smoothing With Split Linear Fits," *Technometrics,* 28, 195–208.

Müller, H. G., (1988), *Nonparametric Regression Analysis of Longitudinal Data* (Springer Lecture Notes in Statistics, 46), Heidelberg: Springer-Verlag.

Tukey, J. W. (1947), "Nonparametric Estimation II. Statistically Equivalent Blocks and Tolerance Regions. The Continuous Case," *The Annals of Mathematical Statistics,* 18, 529–539.

The Annals of Statistics
1992, Vol. 20, No. 1, 121–145

ON BOOTSTRAPPING KERNEL SPECTRAL ESTIMATES

BY J. FRANKE AND W. HÄRDLE

University of Kaiserslautern and Université Catholique de Louvain

An approach to bootstrapping kernel spectral density estimates is described which is based on resampling from the periodogram of the original data. We show that it is asymptotically valid under suitable conditions, and we illustrate its performance for a medium-sized time series sample with a small simulation study.

1. Introduction. During the last years, Efron's (1979) bootstrap has been recognized as a powerful tool for approximating certain characteristics, that is, variance or confidence limits, of statistics, which cannot at all or only with undue effort be calculated by analytical means. In time series analysis, due to the complicated data structure, this kind of difficulty quite often crops up, particularly if one is not willing to assume Gaussianity of the data. In spite of the need for an improved evaluation of the performance of spectrum or parameter estimates for stationary processes, the bootstrap has only recently been applied to problems from time series analysis. Most authors, like Freedman (1984), Efron and Tibshirani (1986), Swanepoel and van Wyk (1986) and Kreiss and Franke (1989), consider resampling the estimated innovations of parametric time series models, whereas Künsch (1989) discusses resampling blocks of data from a stationary process. In this paper, we discuss an intuitive approach to bootstrapping kernel spectral estimates based on resampling from the periodogram of the data, an idea which has been pursued independently in a quite different manner by Hartigan (1990). We prove a theorem asserting that our procedure works provided we take care of the bias in a particular manner. This result is related to similar observations of Romano (1988) for bootstrapping kernel probability density estimates. Some simulations illustrate that our procedure works for moderate sample sizes.

2. Kernel estimates for spectral densities. Let $X_1, \dots, X_T$ be a sample from a strictly stationary real-valued process $\{X_n, -\infty < n < \infty\}$ with mean 0, finite variance and spectral density $f(\omega)$. Let

$$I_T(\omega) = \frac{1}{T}\left|\sum_{k=1}^{T} X_k e^{ik\omega}\right|^2, \qquad -\pi \le \omega \le \pi,$$

denote the periodogram of the sample. Let N denote the largest integer less than or equal to $T/2$. Let the discrete frequencies ω_k be given by

Received June 1987; revised October 1990.

AMS 1980 *subject classifications.* Primary 62M15; secondary 62G05.

Key words and phrases. Bootstrap, time series, kernel spectral estimate, periodogram, bandwidth selection.

Annals of Statistics, 20, 121-145

$2\pi k/T$, $-N \le k \le N$. We consider estimation of $f(\omega)$ by a kernel spectral estimate of the form

$$(1) \qquad \hat{f}(\omega; h) = \frac{1}{Th} \sum_{k=-N}^{N} K\left(\frac{\omega - \omega_k}{h}\right) I_T(\omega_k),$$

where the kernel $K(\theta)$ is a given symmetric, nonnegative function on the real line. We stress the dependency of $\hat{f}$ on the bandwidth h, as the performance of the estimate essentially depends on this smoothing parameter. As the functional measuring the local performance we consider the mean-square percentage error (MSPE), originally proposed by Parzen (1957),

$$\mathrm{MSPE}(\omega; h) = E\left\{\frac{\hat{f}(\omega; h) - f(\omega)}{f(\omega)}\right\}^2.$$

Here, we have taken into account that $f(\omega)$ is a scale parameter of the asymptotic distribution of $I_T(\omega)$. Under suitable assumptions on the process $\{X_n\}$ and on the kernel K,

$$(2) \quad \mathrm{MSPE}(\omega; h) = \left\{\frac{1}{2} h^2 \frac{f''(\omega)}{f(\omega)}\right\}^2 + \frac{1}{2\pi} \int_{-\infty}^{\infty} K^2(\theta)\, d\theta \, \frac{1}{Th} + o\left(\frac{1}{Th}\right),$$

and $T^{-1/5}$ is the rate at which h has to go to 0 if we want to minimize MSPE$(\omega; h)$ asymptotically [compare Priestley (1981), Chapter 7.2]. In this paper, we direct our attention to this most common situation in kernel spectrum estimation.

Härdle and Bowman (1988) apply the bootstrap to kernel estimates for regression curves, and Romano (1988) discusses the related problem of bootstrapping kernel estimates for probability densities. We use the familiar device of interpreting the spectral estimation problem as an approximate multiplicative regression problem, starting from

$$(3) \qquad I_T(\omega_k) = f(\omega_k)\varepsilon_k, \qquad k = 1, \dots, N.$$

The residuals are approximately independent and identically distributed for large T. There are several precise formulations of this vague statement which differ with respect to the—always finitely numbered—frequencies at which the periodogram is considered and with respect to the assumptions on the process $\{X_n\}$ [compare, e.g., Brillinger (1981), Chapters 4 and 5].

3. The bootstrap procedure. In this section, we apply the bootstrap approach of Härdle and Bowman (1988) to (3) by pretending that $\varepsilon_1, \dots, \varepsilon_N$ are really i.i.d. As we want to resample from the residuals, we need an initial estimate of $f(\omega)$. For this purpose, we consider a kernel estimate $\hat{f}(\omega; h_i)$ of the form (1) with an arbitrary initial bandwidth h_ι. In the resampling step, we use another kernel spectrum estimate $\hat{f}(\omega; g)$ of the form (1) to get the bootstrap approximation of the law of $\hat{f}(\omega; h)$. The bandwidth h, which we want to use in spectrum estimation, the resampling bandwidth g and the

(1992) Franke, J. and Härdle, W.
On Bootstrapping kernel spectral estimates.

initial bandwidth h_i may all be different subject to some conditions which we shall discuss later. We now consider the following procedure for getting a bootstrap approximation for $\hat{f}(\omega; h)$.

STEP 1. We choose an initial global bandwidth $h_i > 0$ which does not depend on ω. We estimate the residuals ε_k, $k = 1, \dots, N$, of (3) as

$$\hat{\varepsilon}_k = \frac{I_T(\omega_k)}{\hat{f}(\omega_k; h_i)}, \qquad k = 1, \dots, N.$$

We rescale these empirical residuals and consider

$$\tilde{\varepsilon}_k = \frac{\hat{\varepsilon}_k}{\hat{\varepsilon}.}, \qquad k = 1, \dots, N, \quad \text{where } \hat{\varepsilon}. = \frac{1}{N} \sum_{j=1}^{N} \hat{\varepsilon}_j.$$

STEP 2. We draw independent bootstrap residuals $\varepsilon_1^*, \dots, \varepsilon_N^*$ from the empirical distribution of $\tilde{\varepsilon}_1, \dots, \tilde{\varepsilon}_N$, that is, for all $j = 1, \dots, N$,

$$\operatorname{pr}\{\varepsilon_j^* = \tilde{\varepsilon}_k\} = \frac{1}{N}, \qquad k = 1, \dots, N.$$

Keeping (3) in mind, we define bootstrap periodogram values as

$$I_T^*(\omega_k) = I_T^*(-\omega_k) = \hat{f}(\omega_k; g)\varepsilon_k^*, \qquad k = 1, \dots, N,$$

with some resampling bandwidth g. For convenience, we set $I_T^*(0) = 0$, which corresponds to the periodogram value at 0 taken from a mean-corrected sample. Finally, we get a bootstrap spectral estimate as

$$\hat{f}^*(\omega; h, g) = \frac{1}{Th} \sum_{k=-N}^{N} K\left(\frac{\omega - \omega_k}{h}\right) I_T^*(\omega_k).$$

The rescaled empirical residual ε_j^* has mean 1 with respect to the empirical distribution of $\tilde{\varepsilon}_1, \dots, \tilde{\varepsilon}_N$. This is asymptotically the correct value as the true residual ε_j is asymptotically distributed as an exponential variable with parameter 1. Like recentering in additive regression models [Freedman (1981)], rescaling avoids an additional bias at the resampling stage. Apart from this appealing property, we need this device also from a theoretical point of view. Without rescaling, a proof of the validity of the bootstrap procedure would require more detailed information about the asymptotic properties of $\varepsilon_1, \dots, \varepsilon_N$ than given by Chen and Hannan (1980), and, presumably, Theorem 1 would not even be true in general for resampling directly from $\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_N$.

Resampling from the periodogram is considered independently by Hartigan (1990). He appeals to the fact that the $I_T(\omega_j)$ asymptotically are independent exponential variables and derives resampling estimates for the variance of linear combinations of the periodogram ordinates by systematically perturbing the $I_T(\omega_j)$. However, his procedure has bias problems for non-Gaussian data.

Exploiting our knowledge about the asymptotic distribution of the ε_k, we can modify the preceding bootstrap procedure by replacing $\varepsilon_1^*, \dots, \varepsilon_N^*$ with independent exponential variables $\chi_1, \dots, \chi_N$ with parameter 1. As in Step 2, we get modified bootstrap periodogram values

$$I_T^+(\omega_k) = I_T^+(-\omega_k) = \hat{f}(\omega_k; g)\chi_k, \qquad k = 1, \dots, N,\ I_T^+(0) = 0,$$

and a modified bootstrap spectral estimate

$$\hat{f}^+(\omega; h, g) = \frac{1}{Th} \sum_{k=-N}^{N} K\left(\frac{\omega - \omega_k}{h}\right) I_T^+(\omega_k).$$

As we see in the next section, the bootstrap principle holds for $\hat{f}^+$ as well as for $\hat{f}^*$. Higher-order asymptotics and/or elaborate Monte Carlo studies would be needed to detect differences between both methods. Up to now, some scant simulation results support the intuition that $\hat{f}^*$ is to be preferred for not too large samples and, in particular, for non-Gaussian time series.

4. The bootstrap principle holds. The basic idea of bootstrapping, as applied to the spectral estimation context, is to infer properties of the distribution of the estimate $\hat{f}(\omega; h)$ from the conditional distribution of its bootstrap approximation $\hat{f}^*(\omega; h, g)$, given the original data. To prove the theoretical validity of this bootstrap principle, we follow Bickel and Freedman (1981) and consider the Mallows distance between the pivotal quantity $\sqrt{Th}\{\hat{f}(\omega; h) - f(\omega)\}/f(\omega)$ and its bootstrap approximation $\sqrt{Th}\{\hat{f}^*(\omega; h, g) - \hat{f}(\omega; g)\}/\hat{f}(\omega; g)$. Here, the Mallows distance between distributions F and G is defined as

$$d_2(F, G) = \inf\left\{E(X - Y)^2\right\}^{1/2},$$

where the infimum is taken over all pairs of random variables X and Y having marginal distributions F and G, respectively. We adopt the convention that where random variables appear as arguments of d_2 these represent the corresponding distributions. In particular, bootstrap quantities represent their conditional distribution given the original data $X_1, \dots, X_T$.

For our main result, we need the process generating the data to show sufficiently weak dependence between observations taken at time points far apart. To make this statement precise, we restrict our attention to linear processes, and we assume that the coefficients of the infinite moving average representation decrease sufficiently fast. Furthermore, we consider only the most common situation in kernel spectrum estimation by assuming that the spectral density $f(\omega)$ which we want to estimate is twice continuously differentiable and by choosing a kernel K for which $T^{-1/5}$ is the optimal rate of decrease for the bandwidth h if one is interested in a small mean-square percentage error. This fact is guaranteed by condition (C4) [compare, e.g., Priestley (1981), page 511].

If we want the bootstrap principle to hold in the simple form described in Section 3, we have to make the crucial assumption that the resampling bandwidth g, which we use for defining the bootstrap spectral estimate, converges to 0 a bit slower than $T^{-1/5}$. The reference estimate $\hat{f}(\omega; g)$,

Annals of Statistics, 20, 121-145

therefore, is a bit smoother than an optimal estimate of $f(\omega)$. However, this should not worry us as we do not use $\hat{f}(\omega; g)$ for estimating $f(\omega)$ but only for inferring information about the distribution of $\hat{f}(\omega; h)$, which itself is a kernel estimate with bandwidth h decreasing to 0 with optimal rate $T^{-1/5}$.

We make use of the following notational convention: $h \sim a_T$ if and only if there are constants c, c' such that $0 < c \le h/a_T \le c' < \infty$ for all T large enough.

THEOREM 1. *Let $\{X_n, -\infty < n < \infty\}$ be a real-valued linear process*:

$$X_n = \sum_{k=-\infty}^{\infty} b_k \xi_{n-k}, \qquad -\infty < n < \infty,$$

where ξ_n, $-\infty < n < \infty$, are independent identically distributed random variables satisfying

(C1) $E\xi_n = 0$, $E\xi_n^2 = 1$, $E|\xi_n|^5 < \infty$, *the characteristic function* $q(u)$ *of* ξ_j *satisfies* $\sup\{|q(u)|;\ |u| \ge \delta\} < 1$ *for all* $\delta > 0$.

Assume that the spectral density f of $\{X_n\}$ is nonvanishing and twice continuously differentiable on $[-\pi, \pi]$, and

$$\text{(C2)} \qquad \sum_{k=-\infty}^{\infty} |kb_k| < \infty.$$

Let K be a symmetric, nonnegative kernel on $(-\infty, \infty)$ satisfying

$$\text{(C3)} \qquad \frac{1}{2\pi}\int_{-\infty}^{\infty} K(\theta)\, d\theta = 1, \qquad \frac{1}{2\pi}\int_{-\infty}^{\infty} \theta^2 K(\theta)\, d\theta = 1,$$

where K has compact support $[-\kappa, \kappa]$ and K is uniformly Lipschitz with constant L_K.

Let $k(u)$ denote the Fourier transform of $K(\theta)$, and assume that it is locally quadratic around 0:

$$\text{(C4)} \qquad \lim_{u \to 0} \frac{k(0) - k(u)}{u^2} \quad \textit{exists, is finite and not } 0.$$

For $T \to \infty$, let the bandwidth h of the estimate of interest, the initial bandwidth h_i and the resampling bandwidth g satisfy

$h \sim T^{-1/5}$, $h_i \to 0$ *such that* $(Th_i^4)^{-1} = O(1)$, $g \to 0$ *such that* $h/g \to 0$.

Then, using the preceding definitions, the bootstrap principle holds:

$$\text{(i)} \quad d_2\left[\sqrt{Th}\,\frac{\hat{f}(\omega; h) - f(\omega)}{f(\omega)};\ \sqrt{Th}\,\frac{\hat{f}^*(\omega; h, g) - \hat{f}(\omega; g)}{\hat{f}(\omega; g)}\right] \to 0$$

in probability,

$$\text{(ii)} \quad d_2\left[\sqrt{Th}\,\frac{\hat{f}(\omega; h) - f(\omega)}{f(\omega)};\ \sqrt{Th}\,\frac{\hat{f}^+(\omega; h, g) - \hat{f}(\omega; g)}{\hat{f}(\omega; g)}\right] \to 0$$

in probability.

(1992) Franke, J. and Härdle, W.
On Bootstrapping kernel spectral estimates.

The proof of the theorem, which is deferred to the Appendix, crucially depends on a theorem of Chen and Hannan (1980) which states the almost sure uniform convergence of the empirical distribution function F_N of the sample $I_T(\omega_j)/f(\omega_j)$, $j = 1, \dots, N$, to the distribution function $1 - \exp(-x)$ of the exponential distribution with parameter 1. To make use of this theorem, we need finiteness of the fifth moment and the condition on the characteristic function in (C1). Theorem 1 is presumably correct, assuming $E\xi_n^4 < \infty$ only, because, for our purposes, the weaker convergence in probability of $\sup|F_N(x) - (1 - e^{-x})|$ suffices. We do not try to prove this assertion, as (C1) does not appear excessively restrictive.

To cope with the bias part of the Mallows distance in (i) and (ii), the kernel K must decrease sufficiently fast to 0. For simplicity, we even assume that its support is compact. Some of the kernels which are frequently used in applied spectral analysis satisfy this assumption, for example, the Bartlett–Priestley window [compare Priestley (1981), Chapters 6.2 and 7.5], and restricting attention to kernels with compact support gives rise to considerable simplification of already quite technical proofs.

In the literature, rescaled kernel estimates of the form

$$\tilde{f}(\omega; h) = \frac{\hat{f}(\omega; h)}{S_T(\omega)}, \qquad S_T(\omega) = \frac{1}{Th} \sum_{k=-N}^{N} K\left(\frac{\omega - \omega_j}{h}\right)$$

sometimes are considered. As we shall show in the appendix, $\frac{1}{4} \le S_T(\omega) \le 2$ for all ω, if T is large enough. Therefore, the results of this paper hold for $\tilde{f}(\omega; h)$ too if they are appropriately rephrased.

As already mentioned, Härdle and Bowman (1988) propose a similar procedure for bootstrapping kernel regression estimates. In contrast to our Theorem 1, they consider resampling regression function estimates with bandwidth $g \sim T^{-1/5}$ only. In this case, the bootstrap principle does not hold in the straightforward form of Theorem 1 as the bias of the bootstrap approximation does not approach the bias of the kernel estimate fast enough. However, it is possible to handle this difficulty by essentially bootstrapping only the variance part of the bootstrap approximation and by introducing the bias part by means of an explicit estimate of $f''(\omega)$, remembering the asymptotic relation (2). The same idea works in the spectral estimation context too, and we formulate the result as Theorem 2. We do not give the proof, as its larger part is identical and the rest is quite similar to the proof of Theorem 1. Details can be found in a technical report [Franke (1987)].

THEOREM 2. *Let $\{X_n, -\infty < n < \infty\}$ be a real-valued linear process satisfying the assumptions of Theorem* 1. *Let the kernel K satisfy the assumptions of Theorem* 1, *too. Let $\hat{f}''(\omega)$ be a weakly consistent estimate of $f''(\omega)$. Let*

$$\hat{f}_c(\omega; h, g) = E^* \hat{f}^*(\omega; h, g) = \frac{1}{Th} \sum_{k=-N}^{N} K\left(\frac{\omega - \omega_j}{h}\right) \hat{f}(\omega_j; g)$$

be the conditional expectation of $\hat{f}^*(\omega; h, g)$ and of $\hat{f}^+(\omega; h, g)$ given the *original data. Then, for* $T \to \infty$, $h \sim T^{-1/5}$, $g \sim T^{-1/5}$, $h_i \to 0$ *such that* $(Th_i^4)^{-1} = O(1)$, *we have*

(i)
$$d_2\left[\sqrt{Th}\,\frac{\hat{f}(\omega;h) - f(\omega)}{f(\omega)};\right.$$
$$\left.\sqrt{Th}\,\frac{\hat{f}^*(\omega;h,g) - \hat{f}_c(\omega;h,g) + (h^2/2)\hat{f}''(\omega)}{\hat{f}(\omega;g)}\right] \to 0$$

in probability,

(ii)
$$d_2\left[\sqrt{Th}\,\frac{\hat{f}(\omega;h) - f(\omega)}{f(\omega)};\right.$$
$$\left.\sqrt{Th}\,\frac{\hat{f}^+(\omega;h,g) - \hat{f}_c(\omega;h,g) + (h^2/2)\hat{f}''(\omega)}{\hat{f}(\omega;g)}\right] \to 0$$

in probability.

Here and in the following, E^* denotes the expectation with respect to the empirical distribution of $\tilde{\varepsilon}_1, \dots, \tilde{\varepsilon}_N$.

As a consistent estimate for $f''(\omega)$, we can choose, for example, a kernel estimate of the simple form

$$\hat{f}''(\omega;h_2) = \frac{1}{Th_2^3}\sum_{k=-N}^{N} W\left(\frac{\omega - \omega_k}{h_2}\right) I_T(\omega_k), \tag{4}$$

where W is a kernel of order $(2,4)$ as defined by Gasser, Müller, Köhler, Molinari and Prader (1984).

5. Simulations. In this section, a small simulation study illustrates the performance of our bootstrap approach for a medium sample size $T = 256$. We consider data from an autoregressive process of order 5:

$$X_t = 0.5X_{t-1} - 0.6X_{t-2} + 0.3X_{t-3} - 0.4X_{t-4} + 0.2X_{t-5} + \varepsilon_t,$$

where the ε_t, $-\infty < t < \infty$, are independent standard normal variables. The process parameters have been chosen such that the spectral density has a specified shape: one major peak, one minor peak, and local minima between the peaks, at 0 and at π. We consider estimating the spectral density at the discrete frequencies $\omega_k = 2\pi k/256$ for $k = 42$, 84 (approximately at the two peaks), for $k = 30$, 54 (at the left and right slope of the major peak) and for $k = 67$ (approximately at the trough between both peaks). For those ω_k, we consider the density and skewness of the law of the asymptotic pivot $\sqrt{Th}\{\hat{f}(\omega_k;h) - f(\omega_k)\}/f(\omega_k)$, or, to be precise, a kernel probability density estimate $p_{k,h}$ with Gaussian kernel and bandwidth $b = 0.4$, chosen by a cross-validatory argument, and the sample skewness $s_{k,h}$, both based on 500 simulated data sets. For the spectral estimate, we used the parabolic Bartlett–Priestley kernel [Priestley (1981), Chapters 6.2 and 7.5], scaled such that condition (C3) of Theorem 1 is satisfied. Inspection of various spectrum

Annals of Statistics, 20, 121-145

estimates showed that a good global bandwidth selection lies somewhere between 0.10 and 0.15.

We compare five approximations of $p_{k,h}$ and $s_{k,h}$, three of them derived from the bootstrap principle and the other two from asymptotic normality. All are based on *one* particular sample, $X_1, \ldots, X_{256}$. To get something like a representative data set, we chose that one out of nine independent samples for which the average mean-square percentage error of $\hat{f}(\omega; 0.1)$ assumed its median value. The three bootstrap approximations are provided by the conditional laws of $\sqrt{Th}\{\hat{f}^*(\omega; h, g) - \hat{f}(\omega; g)\}/\hat{f}(\omega; g)$ for bootstrap bandwidths $g = 0.2, 0.3, 0.4$ and initial bandwidths $h_i = g$ in all three cases. Based on 500 resamples, we calculated kernel probability density estimates $p^*_{k,h,g}$ with, again, Gaussian kernel and bandwidth $b = 0.4$ as bootstrap approximations of $p_{k,h}$, and sample skewnesses $s^*_{k,h,g}$ as approximations of $s_{k,h}$.

Using asymptotic normality of $\hat{f}(\omega; h)$, as in Proposition A2, and the asymptotic bias expansion, contained in (2), we know that $\sqrt{Th}\{\hat{f}(\omega_k; h) - f(\omega_k)\}/f(\omega_k)$ is also approximately normally distributed with mean $\mu_{k,h}$ and variance σ^2 given by

$$\mu_{k,h} = 0.5\sqrt{Th^5}\, f''(\omega_k)/f(\omega_k), \qquad \sigma^2 = \int K^2(\theta)\, d\theta/(2\pi).$$

To cope with the additional smoothing introduced by kernel probability density estimation, we have to compare $p_{k,h}$ with the normal density $\varphi_{k,h}$ with mean $\mu_{k,h}$, but with larger variance $\sigma^2 + b^2$. The normal approximation to the skewness $s_{k,h}$ is, of course, 0.

To get $\varphi_{k,h}$, we have to know f and f''. As a realistic competitor for the bootstrap, we therefore consider $\hat{\varphi}_{k,h}$, a plug-in normal approximation with mean

$$\hat{\mu}_{k,h} = 0.5\sqrt{Th^5}\, \hat{f}''(\omega; h_2)/\hat{f}(\omega; h_1)$$

and variance $\sigma^2 + b^2$; $\hat{f}(\omega; h_1)$ denotes again a spectral estimate, given by (1) with Bartlett–Priestley kernel K and bandwidth $h_1 = 0.15$; $\hat{f}''(\omega; h_2)$ denotes a kernel estimate of $f''(\omega)$, as in (4), where the kernel W has the same support as K and, there, equals $\{c_1 \cos^4(c_2 u)\}''$ with suitable constants c_1, c_2. The bandwidths h_1, h_2 are chosen to give a visually good correspondence between the true functions and their estimates.

Figures 1 and 2 show plots of $p_{k,h}$ and its approximations for $k = 42$ (peak) and $k = 30$ (slope) and bandwidths $h = 0.05$ and $h = 0.10$, respectively. Among all these selections of ω_k and h which we have considered, Figure 1c is typical for the majority of those situations: Visually, the bootstrap provides a better fit to the true density than its competitor, the plug-in normal approximation. In a few cases, for which Figure 2 is an example, the bootstrap approximation is not better than the plug-in normal approximation, but it never was considerably worse. A bit surprising was the observation that the bootstrap densities $p^*_{k,h,g}$ did not depend as much on the chosen bootstrap bandwidth g as we originally expected, as can be seen from Figures 1b and 2.

(1992) Franke, J. and Härdle, W.
On Bootstrapping kernel spectral estimates.

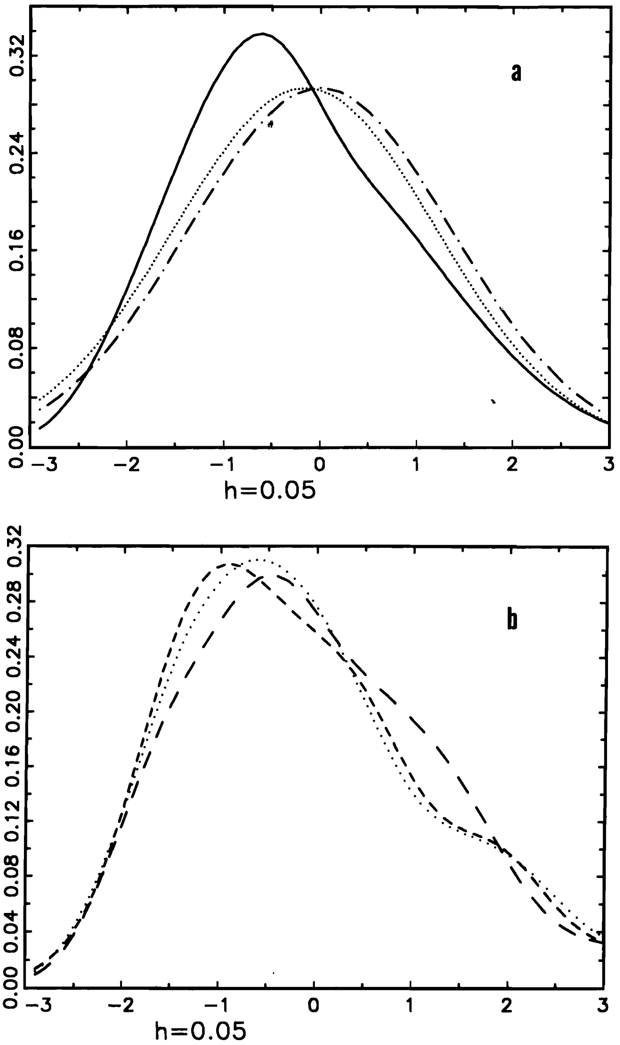


FIG. 1. (*a*) *Probability density* $p_{k,h}$ (*solid line*) *of the asymptotic pivot and its normal approximations* $\varphi_{k,h}$ (*dotted line*) *and* $\hat{\varphi}_{k,h}$ (*dots and dashes*) *for* $k = 42$ *and* $h = 0.05$. (*b*) *Bootstrap approximation* $p^*_{k;h,g}$ *for* $k = 42$ *and* $h = 0.05$ *and* $g = 0.2$ (*long dashes*), $g = 0.3$ (*short dashes*) *and* $g = 0.4$ (*dots*).

In some cases, only the heavily oversmoothed reference spectral estimate ($g = 0.4$) deviated considerably from the $p^*_{k,h,g}$ for smaller $g = 0.2$ and 0.3.

Table 1 compares the skewness $s_{k,h}$ of the asymptotic pivot and its bootstrap approximations $s^*_{k,h,g}$ for $g = 0.2$, 0.3 and 0.4. The bootstrap manages to reproduce the skewness of the distribution, which we want to approximate, quite well.

We also have repeated the simulation study with innovations ε_t drawn from a centered and scaled χ^2_4 distribution. Qualitatively, the results are the same

Annals of Statistics, 20, 121-145

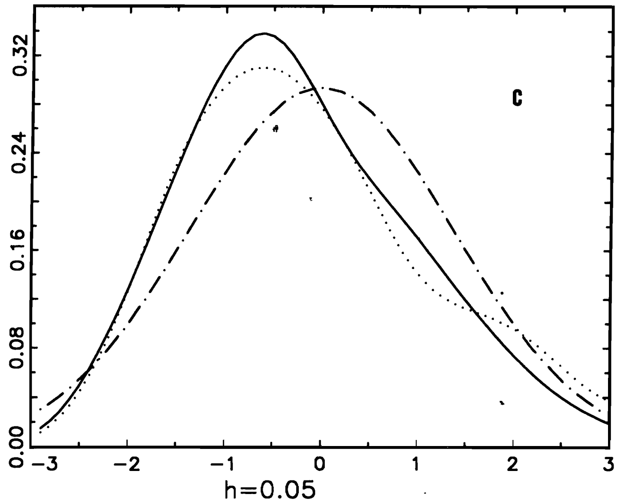


FIG. 1. (*c*) *Probability density* $p_{k,h}$ (*solid line*), *its plug-in normal approximation* $\hat{\varphi}_{k,h}$ (*dots and dashes*) *and a bootstrap approximation* $p^*_{k,h,0.4}$ (*dots*) *for* $k = 42$ *and* $h = 0.05$.

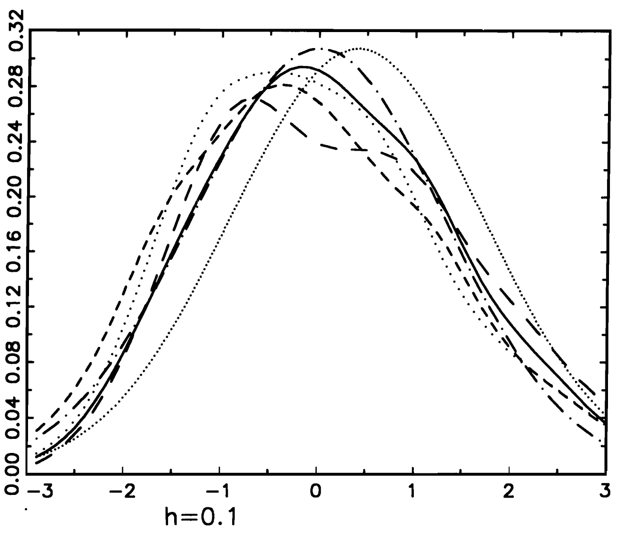


FIG. 2. *Probability density* $p_{k,h}$ (*solid line*) *of the asymptotic pivot, its normal approximations* $\varphi_{k,h}$ (*narrowly spaced dots*) *and* $\hat{\varphi}_{k,h}$ (*dots and dashes*) *and its bootstrap approximation* $p^*_{k,h,g}$ *for* $k = 30$ *and* $h = 0.10$ *and for* $g = 0.2$ (*long dashes*), $g = 0.3$ (*short dashes*) *and* $g = 0.4$ (*widely spaced dots*).

(1992) Franke, J. and Härdle, W.
On Bootstrapping kernel spectral estimates.

Annals of Statistics, 20, 121-145

TABLE 1
Skewness of asymptotic pivot for various bandwidths h and its bootstrap approximations for several values of g taken from one representative sample

		k				
		30	42	54	67	84
$h = 0.05$	$s_{k,h}$	0.800	0.606	0.866	1.055	0.674
	$s^*_{k,h,0.2}$	0.852	0.795	0.963	0.610	0.352
	$s^*_{k,h,0.3}$	0.930	0.759	0.726	0.721	0.702
	$s^*_{k,h,0.4}$	0.719	0.895	0.665	0.869	0.978
$h = 0.10$	$s_{k,h}$	0.426	0.613	0.560	0.542	0.287
	$s^*_{k,h,0.2}$	0.557	0.642	0.630	0.511	0.491
	$s^*_{k,h,0.3}$	0.417	0.546	0.532	0.583	0.372
	$s^*_{k,h,0.4}$	0.330	0.701	0.625	0.494	0.428
$h = 0.15$	$s_{k,h}$	0.414	0.489	0.467	0.554	0.330
	$s^*_{k,h,0.2}$	0.426	0.507	0.309	0.351	0.337
	$s^*_{k,h,0.3}$	0.454	0.566	0.323	0.539	0.499
	$s^*_{k,h,0.4}$	0.495	0.425	0.457	0.399	0.470

as in the Gaussian case, that is, the bootstrap outperforms the plug-in normal approximation in approximating the probability density and the skewness of the law of interest.

6. Confidence intervals and bandwidth selection. Once we know that the bootstrap principle holds for spectral density estimation we can apply it in the usual manner to get estimates for statistical quantities of interest. For the sake of illustration, we have a look at the problem of getting a confidence interval for $f(\omega)$ and of selecting a local bandwidth $h = h(\omega)$ of the kernel estimate $\hat{f}(\omega; h)$ at a given frequency ω. In this entire section, we implicitly assume that the conditions of Theorem 1 are satisfied.

Let c_α be characterized by

$$\mathrm{pr}\left[\sqrt{Th}\,\frac{\hat{f}(\omega;h) - f(\omega)}{f(\omega)} \le c_\alpha\right] = \alpha,$$

that is, $\{1 + c_\alpha(Th)^{-1/2}\}\hat{f}(\omega; h)$ is the upper bound of a $(1 - 2\alpha)$-confidence interval for $f(\omega)$. A bootstrap approximation for the generally unknown quantity c_α is given as c^*_α, defined by

$$\mathrm{pr}^*\left[\sqrt{Th}\,\frac{\hat{f}^*(\omega;h,g) - \hat{f}(\omega;g)}{\hat{f}(\omega;g)} \le c^*_\alpha\right] = \alpha,$$

where the bootstrap distribution pr* corresponds to drawing the bootstrap residuals $\varepsilon^*_1, \dots, \varepsilon^*_N$ from the empirical distribution of the rescaled residuals as described in Section 3. From Theorem 1, we know that $c^*_\alpha \to c_\alpha$ in probability if $T \to \infty$. Explicit calculation of c^*_α will be quite difficult, and, therefore, we propose to estimate it by the familiar Monte Carlo algorithm, as described, for

(1992) Franke, J. and Härdle, W.
On Bootstrapping kernel spectral estimates.

example, by Efron and Tibshirani (1986), which usually is associated with bootstrap procedures. Analogously, we can get a bootstrap approximation for the lower bound of a confidence interval for $f(\omega)$.

A major problem with kernel spectral estimates is the choice of bandwidth h. Until quite recently, the literature contains only rough guidelines for choosing h which often depend on some prior information on the shape of f. An extensive discussion of this problem has been given by Priestley [(1981), Chapter 7]. In a recent paper, Beltrão and Bloomfield (1987) have investigated the problem of selecting a global bandwidth which minimizes the average mean-square percentage error,

$$\mathrm{AMSPE}(h) = \frac{1}{N}\sum_{j=1}^{N} \mathrm{MSPE}(\omega_j; h),$$

where MSPE$(\omega; h)$ is defined as in Section 2. They have proposed a cross-validatory choice of bandwidth, and they have shown that their procedure produces a bandwidth which approximately minimizes AMSPE(h).

We consider the problem of selecting a good local bandwidth $h = h(\omega)$ which minimizes approximately the mean-square percentage error MSPE$(\omega; h)$ for a given frequency ω. Following Rice (1984) who considered bandwidth choice for the related nonparametric regression estimates, we restrict the minimization to an interval $B_T = [aT^{-1/5}, bT^{-1/5}]$ of bandwidths which shrinks to 0 at the optimal rate. Here, $0 < a < b < \infty$ are suitable constants. Let h_0, depending on the sample size, be defined by

$$\mathrm{MSPE}(\omega; h_0) = \min_{h \in B_T} \mathrm{MSPE}(\omega; h).$$

As we shall discuss in proving Theorem 3,

$$(5) \qquad T^{1/5}h_0 \to z_\infty = \left[\frac{1}{2\pi}\int_{-\infty}^{\infty} K^2(\lambda)\,d\lambda \left\{\frac{f(\omega)}{f''(\omega)}\right\}^2\right]^{1/5}, \quad \text{for } T \to \infty,$$

provided $f''(\omega) \neq 0$ and a, b are chosen such that $a < z_\infty < b$. Notice that $z_\infty T^{-1/5}$ minimizes the dominating part of the asymptotic formula (2) for MSPE$(\omega; h)$ considered as a function of h. As MSPE$(\omega; h)$ depends on the unknown spectral density f, we cannot calculate h_0. Therefore, we propose to estimate MSPE$(\omega; h)$ by its bootstrap approximation,

$$\mathrm{MSPE}^*(\omega; h) = E^*\left\{\frac{\hat{f}^*(\omega; h, g) - \hat{f}(\omega; g)}{\hat{f}(\omega; g)}\right\}^2,$$

and then to choose the bandwidth h_0^* which minimizes MSPE$^*(\omega; h)$,

$$\mathrm{MSPE}^*(\omega; h_0^*) = \min_{h \in B_T} \mathrm{MSPE}^*(\omega; h).$$

The calculation of h_0^* can be accomplished easily, as MSPE$^*(\omega; h)$ can be given explicitly. We do not have to resort to Monte Carlo methods in this case. A straightforward calculation, using the independence of the bootstrap residuals

Annals of Statistics, 20, 121-145

ε_j^* and $E^*\varepsilon_j^* = 1$, shows

$$\begin{aligned}\hat{f}^2(\omega; g)\mathrm{MSPE}^*(\omega; h) &= \frac{\mathrm{var}^*(\varepsilon_1^*)}{T^2h^2}\sum_{j=-N}^{N} K^2\left(\frac{\omega-\omega_j}{h}\right)\hat{f}^2(\omega_j; g) \\ &\quad + \left\{\frac{1}{Th}\sum_{j=-N}^{N} K\left(\frac{\omega-\omega_j}{h}\right)\hat{f}(\omega_j; g) - \hat{f}(\omega; g)\right\}^2,\end{aligned} \tag{6}$$

where, by Proposition A1 of the Appendix,

$$\mathrm{var}^*(\varepsilon_1^*) = E^*(\varepsilon_1^* - 1)^2 = \frac{1}{N}\sum_{k=1}^{N}\tilde{\varepsilon}_k^2 - 1 \to 1 \quad \text{in probability.}$$

Restricting minimization to a finite subset of B_T which is allowed to increase with the sample size at a certain rate, Härdle and Bowman (1988) have shown that the analogous bootstrap selection of the bandwidth of a kernel regression estimate is asymptotically optimal in the sense that the ratio of the minimum of the bootstrap error estimate and the minimum of the true error converges to 1 in probability. Using the explicit formula (6), we are able to prove the same result without restrictions to B_T and, furthermore, to prove consistency of h_0^* in the sense that $T^{1/5}(h_0^* - h_0) \to 0$ in probability.

THEOREM 3. *If the conditions of Theorem* 1 *are satisfied and if, additionally,* $f''(\omega) \neq 0$ *and* $0 < a < z_\infty < b < \infty$, *then, for* h_0, h_0^* *defined as before,*

(i) $\quad T^{1/5}(h_0^* - h_0) \to 0$ *in probability for* $T \to \infty$,

(ii) $\quad \dfrac{\mathrm{MSPE}^*(\omega; h_0^*)}{\mathrm{MSPE}(\omega; h_0)} \to 1$ *in probability for* $T \to \infty$.

The proof of the theorem is again postponed to the Appendix.

7. Concluding remarks. We have shown that a rather straightforward approach to bootstrapping kernel spectrum estimates works. Our procedure is quite similar to the bootstrap for both parametric and nonparametric regression with fixed design. Some care has to be taken if the bootstrap principle is to hold. Either one has to restrict the bootstrap essentially to $\hat{f}(\omega; h) - E\hat{f}(\omega; h)$, estimating the bias $E\hat{f}(\omega; h) - f(\omega)$ explicitly as in Theorem 2, or one has to choose a preliminary estimate $\hat{f}(\omega; g)$ which is asymptotically smoother than an optimal kernel spectrum estimate. If h/g does not converge to 0, then the assertion of Theorem 1 does not hold. As can be seen from a careful look at the proof, the critical quantity is

$$\sqrt{Th}\left(E^*\hat{f}^*(\omega; h, g) - E\hat{f}^*(\omega; h, g)\right), \tag{7}$$

which dominates the left-hand side of (A7) of the Appendix, and which

(1992) Franke, J. and Härdle, W.
On Bootstrapping kernel spectral estimates.

converges to 0 in probability if $h/g \to 0$. Now, remark that

$$E^*\hat{f}^*(\omega; h, g) = \frac{1}{Tg}\sum_{j=-N}^{N}\left\{\frac{1}{Th}\sum_{k=-N}^{N} K\left(\frac{\omega-\omega_j}{h}\right)K\left(\frac{\omega_j-\omega_k}{g}\right)\right\}I_T(\omega_j).$$

By the compactness of the support of K and by the asymptotic properties of the periodogram, exhibited in (A6), $E^*\hat{f}^*(\omega; h, g)$ behaves asymptotically like the mean of Tg independent random variables with uniformly bounded variance. Therefore, (7) converges to 0 only if the scaling factor $(Th)^{1/2}$ converges to ∞ slower than $(Tg)^{1/2}$. The necessity to oversmooth in resampling kernel type function estimates is not a particular feature of spectrum estimation. Similar results have been found by Romano (1988) for probability density estimates and by Härdle and Bowman (1988) for regression function estimates.

Finally, let us remark that our results do not strongly depend on the particular assumptions on the stationary process. Essentially, we need asymptotic normality of $\hat{f}(\omega; h)$, as stated in Proposition A2 of the Appendix, and the empirical distribution function of the $I_T(\omega_j)/f(\omega_j)$, $j = 1, \dots, N$, must converge uniformly to $1 - e^{-x}$ in probability.

APPENDIX

Some auxiliary results and proofs of the theorems. For real numbers a_T and random variables Z_T, we write $Z_T = o_p(a_T)$ for $T \to \infty$ $[Z_T = O_p(a_T)$ for $T \to \infty]$ if $Z_T/a_T \to 0$ in probability $[Z_T/b_T \to 0$ in probability for all sequences b_T such that $a_T = o(b_T)]$. For analyzing the bias of the kernel spectrum estimate, we repeatedly consider

$$S_T(\omega) = \frac{1}{Th}\sum_{j=-N}^{N} K\left(\frac{\omega-\omega_j}{h}\right).$$

If the kernel K satisfies the assumptions of Theorem A1, we have

$$\text{(A1)}\qquad \begin{aligned} |S_T(\omega) - 1| &= \left|S_T(\omega) - \frac{1}{2\pi}\int_{\omega-\pi}^{\omega+\pi}\frac{1}{h}K\left(\frac{\theta}{h}\right)d\theta\right| \\ &\le \frac{2\pi L_K}{Th}, \quad \text{if } |\omega| \le \pi - \kappa h, \end{aligned}$$

where $[-\kappa, \kappa]$ contains the support of K, and, for any bounded function ψ,

$$\text{(A2)}\qquad \left|\frac{1}{Th}\sum_{j=-N}^{N} K^m\left(\frac{\omega-\omega_j}{h}\right)\psi(\omega_j)\right| \le c_m^* \sup_\theta |\psi(\theta)|, \qquad m \ge 1,$$

with a suitable constant c_m^*, because only about $2\kappa Th$ summands do not vanish.

Annals of Statistics, 20, 121-145

THEOREM A1. *Let* $\{X_n, -\infty < n < \infty\}$ *be a linear process,*

$$X_n = \sum_{k=-\infty}^{\infty} b_k \xi_{n-k}, \qquad -\infty < n < \infty,$$

satisfying assumptions (C1) *and* (C2) *of Theorem* 1. *Let the spectral density* f *of* $\{X_n\}$ *be nonvanishing and satisfying a uniform Lipschitz condition. Let* K *be a symmetric, nonnegative kernel satisfying assumption* (C3) *of Theorem* 1. *Let* $I_T(\omega)$ *and* $\hat{f}(\omega; h)$ *denote the periodogram and the kernel spectrum estimate based on* $X_1, \dots, X_T$, *as in Section* 2.

If, for $T \to \infty$, *we have* $h \to 0$, $(Th^4)^{-1} = O(1)$, *then*

$$\sup_{-\pi \le \omega \le \pi} \left| \hat{f}(\omega; h) - f(\omega) S_T(\omega) \right| = O_p(h^{-1}T^{-1/3}) + O_p(h).$$

PROOF. The theorem is related to Theorem 2.1 of Woodroofe and van Ness (1967) who, under assumptions on K which are too restrictive for our purposes, give an exact rate for the convergence in probability of $\sup|\hat{f}(\omega; h) - E\hat{f}(\omega; h)|/f(\omega)$. Referring to the similarity of arguments, we only sketch the proof of our theorem. Let J_T, $\hat{\varphi}$ denote the periodogram and spectral estimate of $\xi_1, \dots, \xi_T$:

$$J_T(\omega) = \frac{1}{T}\left| \sum_{k=1}^{T} \xi_k e^{ik\omega} \right|^2, \qquad \hat{\varphi}(\omega; h) = \frac{1}{Th} \sum_{j=-N}^{N} K\left(\frac{\omega - \omega_j}{h} \right) J_T(\omega_j).$$

Because f is bounded, it suffices to show that the assertion of the theorem holds for the independent ξ_j and that the supremum of $|\hat{f}(\omega; h) - \hat{\varphi}(\omega; h) f(\omega)|$ is of the order $O_p(h)$, for

$$\begin{aligned} &\left| \hat{f}(\omega; h) - f(\omega) S_T(\omega) \right| \\ &\quad \le \left| \hat{f}(\omega; h) - \hat{\varphi}(\omega; h) f(\omega) \right| + \left| \hat{\varphi}(\omega; h) - S_T(\omega) \right| f(\omega). \end{aligned}$$

(i) We split $\sup|\hat{\varphi}(\omega; h) - S_T(\omega)|$ into two parts and show that both of them converge in probability to 0 with the desired speed. Let $a_T = hT^{-1/3}$, $m_T = [a_T^{-1}]$, and $\theta_k = \pi k / m_T$ for $-m_T \le k \le m_T$:

$$\begin{aligned} &\sup_{\omega} \left| \hat{\varphi}(\omega; h) - S_T(\omega) \right| \\ &\quad = \sup_{|k| \le m_T} \sup_{|\omega - \theta_k| \le \pi a_T} \left| \hat{\varphi}(\omega; h) - S_T(\omega) \right| \\ &\quad \le \sup_{|k| \le m_T} \left| \hat{\varphi}(\theta_k; h) - S_T(\theta_k) \right| \\ &\qquad + \sup_{|\omega - \theta| \le \pi a_T} \left| \hat{\varphi}(\theta; h) - S_T(\theta) - \hat{\varphi}(\omega; h) + S_T(\omega) \right|. \end{aligned}$$

Both terms on the right-hand side are of order $O_p(h^{-1}T^{-1/3})$. For the second

(1992) Franke, J. and Härdle, W.
On Bootstrapping kernel spectral estimates.

Annals of Statistics, 20, 121-145

term, using Lipschitz continuity of K, we have

$$hT^{1/3} \sup_{|\omega-\delta|\le\pi a_T} \left|\hat{\varphi}(\theta;h) - S_T(\theta) - \hat{\varphi}(\omega;h) + S_T(\omega)\right|$$

$$\le \pi L_K\left\{\frac{1}{Th}\sum_{j=-N}^{N} J_T(\omega_j) + 1\right\} \to 2\pi L_K \quad \text{a.s.},$$

for $T \to \infty$, by results of Chen and Hannan (1980) on the empirical distribution of the $J_T(\omega_j)$, $j = 1, \dots, N$. For the first term, we have by Chebyshev's inequality, for all $\delta > 0$,

$$\operatorname{pr}\left\{hT^{1/3} \sup_{|k|\le m_T} \left|\hat{\varphi}(\theta_k;h) - S_T(\theta_k)\right| \ge \delta\right\}$$

$$\le \sum_{|k|\le m_T} \frac{h^2T^{2/3}}{\delta^2} E\{\hat{\varphi}(\theta_k;h) - S_T(\theta_k)\}^2$$

$$\le (2m_T + 1)\frac{hT^{1/3}}{\delta^2},$$

and the right-hand side is bounded for $T \to \infty$. We have used

$$E\{\hat{\varphi}(\omega;h) - S_T(\omega)\}^2$$

$$\le \frac{c}{Th} \quad \text{for all } \omega \in [-\pi,\pi] \text{ and suitable constant } c > 0,$$

which follows from independence of the ξ_j and then from using (A2).

(ii) From (A2) and part (i), we know that $\hat{\varphi}(\omega;h)$ is $O_p(h)$ uniformly in ω. Using this result and Lipschitz continuity of f, we can show that

$$\hat{f}(\omega;h) - \hat{\varphi}(\omega;h)f(\omega) = \frac{1}{Th}\sum_{j=-N}^{N} K\left(\frac{\omega-\omega_j}{h}\right)\{I_T(\omega_j) - J_T(\omega_j)f(\omega)\}$$

is $O_p(h)$ uniformly in ω. For this purpose, we use the approximation of the discrete Fourier transform of the X_k by the discrete Fourier transform of the ξ_k as given by Hannan [(1970), page 246]. □

Theorem A1 and (A1) immediately imply Corollary A1, from which, together with (A2) and the compactness of the support of K, Corollary A2 follows.

COROLLARY A1. *Under the assumptions of Theorem* A1

$$\sup_{|\omega|\le\pi-\kappa h} \left|\hat{f}(\omega;h) - f(\omega)\right| = O_p(h^{-1}T^{-1/3}) + O_p(h).$$

(1992) Franke, J. and Härdle, W.
On Bootstrapping kernel spectral estimates.

Corollary A2. *Let the assumptions of Theorem* A1 *be satisfied. If for* $T \to \infty$, $g \to 0$ *and* $(Tg^4)^{-1} = O(1)$, *we have, for* $\omega \neq \pm\pi$,

$$\frac{1}{Th}\sum_{j=-N}^{N} K^2\left(\frac{\omega-\omega_j}{h}\right)\left\{\hat{f}(\omega_j, g) - f(\omega_j)\right\}^2 = O_p(g^{-2}T^{-2/3}) + O_p(g^2).$$

Relation (5) of Chen and Hannan (1980) on the empirical distribution function of the $I_T(\omega_j)/f(\omega_j)$, $j = 1, \dots, N$, implies for $N = [T/2] \to \infty$

$$\frac{1}{N}\sum_{j=1}^{N}\frac{I_T(\omega_j)}{f(\omega_j)} \to 1, \qquad \frac{1}{N}\sum_{j=1}^{N}\left[\frac{I_T(\omega_j)}{f(\omega_j)}\right]^2 \to 2 \quad \text{a.s.}$$

under the assumptions of Theorem 1. (A1), (A2) and Theorem A1 allow us to replace $f(\omega)$ by its estimate $\hat{f}(\omega; h)$ if we settle for convergence in probability.

Proposition A1. *Under the assumptions of Theorem* 1, *we have, for* $N = [T/2] \to \infty$ *and* $h \to 0$ *such that* $(Th^4)^{-1} = O(1)$,

$$\frac{1}{N}\sum_{j=1}^{N}\frac{I_T(\omega_j)}{\hat{f}(\omega_j; h)} \to 1, \qquad \frac{1}{N}\sum_{j=1}^{N}\left\{\frac{I_T(\omega_j)}{\hat{f}(\omega_j; h)}\right\}^2 \to 2,$$

$$\frac{1}{N}\sum_{j=1}^{N}\left\{\frac{I_T(\omega_j)}{f(\omega_j)} - \frac{I_T(\omega_j)}{\hat{f}(\omega_j; h)}\right\}^2 \to 0 \quad \textit{in probability}.$$

Proposition A2. *Let* $\{X_n,\ -\infty < n < \infty\}$ *be a linear process*,

$$X_n = \sum_{k=-\infty}^{\infty} b_k \xi_{n-k}, \qquad -\infty < n < \infty,$$

satisfying the assumptions of Theorem 1, *and let* $\hat{f}(\omega; h)$ *denote a kernel spectral estimate with a nonnegative symmetric kernel K satisfying assumption* (C3) *of Theorem* 1. *If, for* $T \to \infty$, *we have* $h \to 0$ *and* $Th^2 \to \infty$, *then, for* $|\omega| < \pi$:

$$\text{(i)}\quad Th\,\mathrm{var}\big(\hat{f}(\omega; h)\big) \to \sigma^2 = f^2(\omega)\frac{1}{2\pi}\int_{-\infty}^{\infty} K^2(\theta)\,d\theta, \quad \textit{for } T \to \infty,$$

$$\text{(ii)}\qquad \sqrt{Th}\left\{\hat{f}(\omega; h) - E\hat{f}(\omega; h)\right\} \to Z \quad \textit{in distribution},$$

where Z is a Gaussian random variable with mean 0 *and variance* σ^2.

Proof. Using the compactness of the support of K and the asymptotic properties of the periodogram $I_T(\omega_j)$, $j = 1, \dots, N = [T/2]$, as given in Theorem 6.2.3 of Priestley (1981), we have, for a suitable constant C,

$$Th\,\mathrm{var}\left\{\hat{f}(\omega; h)\right\} \le \frac{1}{Th}\sum_{j=-N}^{N} K^2\left(\frac{\omega-\omega_j}{h}\right) f^2(\omega_j) + CS_T^2(\omega)h \quad \text{for } \omega \ge \kappa h.$$

As, by (A2), $S_T(\omega)$ is bounded, (i) follows.

Annals of Statistics, 20, 121-145

Part (ii) can be shown by the same methods used in the proof of Theorem V.11 of Hannan (1970), which states a stronger result for a slightly different, but asymptotically equivalent spectral estimate. □

LEMMA A1. (i) *Let K satisfy the assumptions of Theorem* A1, *and, for* $|\omega| < \pi - \kappa h$, *let p be twice continuously differentiable on* $[\omega - \kappa h, \omega + \kappa h]$. *Then, for* $T \to \infty$, $h \to 0$ *such that* $Th \to \infty$,

$$\left|\frac{1}{Th}\sum_{j=-N}^{N} K\left(\frac{\omega - \omega_j}{h}\right)p(\omega_j) - p(\omega) - \frac{h^2}{2}p''(\omega)\right|$$
$$\le \frac{c}{Th}\left\{\sup_\theta |p(\theta)| + h \sup_\theta |p'(\theta)|\right\}$$
$$+ \frac{h^2}{2}\sup_\theta |p''(\theta) - p''(\omega)|,$$

where c is a suitable constant and the suprema are taken over the interval $[\omega - \kappa h, \omega + \kappa h]$.

(ii) *Let the assumptions of Theorem* 1 *be satisfied. Then, for* $T \to \infty$, $h \to 0$ *such that* $(Th^4)^{-1} = O(1)$, *the bias of* $\hat{f}(\omega; h)$ *satisfies*

$$E\hat{f}(\omega; h) - f(\omega) = \frac{h^2}{2}f''(\omega) + o(h^2) + O\left(\frac{\log T}{T}\right)$$

uniformly in $|\omega| \le \pi - \kappa h$.

PROOF. (i) The compactness of the support of K, its Lipschitz continuity and the differentiability of p imply, uniformly in $|\omega| < \pi - \kappa h$,

$$\left|\frac{1}{Th}\sum_{j=-N}^{N} K\left(\frac{\omega - \omega_j}{h}\right)p(\omega_j) - \frac{1}{2\pi}\int_{-\infty}^{\infty} K(\theta)p(\omega + \theta h)\,d\theta\right|$$
$$\le \frac{c}{Th}\left\{\sup_\theta |p(\theta)| + h \sup_\theta |p'(\theta)|\right\}.$$

The assertion follows from the Taylor expansion of $p(\omega + \theta h)$, using that $K(\theta)/(2\pi)$ and $\theta^2 K(\theta)/(2\pi)$ integrate to 1 and $\theta K(\theta)$ integrates to 0.

(ii) Replacing p by f in (i) and noting that, under our assumptions,

$$EI_T(\omega_j) = f(\omega_j) + O\left[\frac{\log T}{T}\right], \qquad j = 1, \dots, N, \tag{A3}$$

uniformly in j [Priestley (1981), page 418], the second assertion follows. □

PROOF OF THEOREM 1. (a) To prove (i), we use Lemma 8.8 of Bickel and Freedman (1981) and split the squared Mallows metric into a variance part

(1992) Franke, J. and Härdle, W.
On Bootstrapping kernel spectral estimates.

and a squared bias part,

$$V_T^2 = d_2^2\left[\sqrt{Th}\,\frac{\hat{f}(\omega;h) - E\hat{f}(\omega;h)}{f(\omega)},\ \sqrt{Th}\,\frac{\hat{f}^*(\omega;h,g) - E^*\hat{f}^*(\omega;h,g)}{\hat{f}(\omega;g)}\right]$$

$$B_T^2 = Th|b_T(\omega) - b_T^*(\omega)|^2,$$

where

$$b_T(\omega) = \frac{\{E\hat{f}(\omega;h) - f(\omega)\}}{f(\omega)} \quad \text{and} \quad b_T^*(\omega) = \frac{\{E^*\hat{f}^*(\omega;h,g) - \hat{f}(\omega;g)\}}{\hat{f}(\omega;g)}.$$

Throughout the proof, we use the abbreviations $I_{T,j} = I_T(\omega_j)$, $I_{T,j}^* = I_T^*(\omega_j)$ and

$$\alpha_j(\omega;h) = \frac{1}{Th}K\left(\frac{\omega - \omega_j}{h}\right),$$

$$\gamma_j(\omega;h,g) = \sum_{j=-N}^{N} \alpha_k(\omega,h)\alpha_j(\omega_k;g) - \alpha_j(\omega;g).$$

(b) We first prove that $V_T \to 0$ in probability. For this purpose, let χ_j, $|j| \geq 1$, be independent, exponentially distributed variables with parameter 1, and let $\chi_0 \equiv 0$. We remark that $I_T(\omega)/f(\omega)$ converges to χ_1 in distribution. We define

$$f^0(\omega;h) = \sum_{j=-N}^{N} \alpha_j(\omega;h)f(\omega_j)\chi_j, \qquad D^0 = \sqrt{Th}\{f^0(\omega;h) - Ef^0(\omega;h)\},$$

$$D = \sqrt{Th}\{\hat{f}(\omega;h) - E\hat{f}(\omega;h)\},$$

$$D^* = \sqrt{Th}\{f^*(\omega;h,g) - E^*\hat{f}^*(\omega;h,g)\}.$$

We use that d_2 is a metric, and we get

$$V_T \leq \frac{d_2(D,D^0)}{f(\omega)} + d_2\left(\frac{D^0}{f(\omega)}, \frac{D^0}{\hat{f}(\omega;g)}\right) + \frac{d_2(D^0,D^*)}{\hat{f}(\omega;g)}.$$

To prove $d_2(D,D^0) \to 0$ in probability, consider a zero-mean Gaussian variable Z with variance σ^2 given in Proposition A2. By this proposition, D converges to Z in distribution, and $ED^2 \to EZ^2$. Exactly as in proving the first part of Proposition A2, $E(D^0)^2 \to EZ^2$ follows. Using boundedness of f and the regularity conditions on K, it is easy to show that D^0 satisfies Liapounov's condition [Shiryayev (1984), page 331] and, therefore, converges to Z in distribution, too. Now

$$d_2(D,D^0) \leq d_2(D,Z) + d_2(Z,D^0) \to 0,$$

where the convergence holds by Lemma 8.3 of Bickel and Freedman (1981).

Theorem 8.1 of Major (1978) provides an explicit formula for the Mallows metric of real-valued random variables which implies

$$d_2^2\left[\frac{D^0}{f(\omega)}, \frac{D^0}{\hat{f}(\omega;g)}\right] = \left[\frac{1}{f(\omega)} - \frac{1}{\hat{f}(\omega;g)}\right]^2 E(D^0)^2 \to 0 \quad \text{in probability,}$$

as, for example, by Theorem A1 and (A1), $\hat{f}(\omega; g) \to f(\omega) > 0$ in probability, and as, by (A2) and the boundedness of f, $E(D^0)^2$ is bounded.

(c) To finish the proof that $V_T \to 0$ in probability, it suffices to show that $d_2(D^0, D^*) \to 0$ in probability, as $\hat{f}(\omega; g) \to f(\omega) > 0$. As D^0, D^* are sums of independent random variables (conditional on the original data), we have, by a slight modification of Lemma 8.7 of Bickel and Freedman (1981),

$$\text{(A4)} \qquad d_2^2(D^0, D^*) \le Th \sum_{j=-N}^{N} \alpha_j^2(\omega; h) d_2^2\left[f(\omega_j)\{\chi_j - 1\}, I^*_{T,j} - E^* I^*_{T,j}\right].$$

As the distributions of χ_j, ε_j^* do not depend on j, we have, using the definition of $I^*_{T,j}$,

$$\begin{aligned} &d_2^2\left[f(\omega_j)\{\chi_j - 1\}, I^*_{T,j} - E^* I^*_{T,j}\right] \\ &\quad \le 2 d_2^2\left[f(\omega_j)\{\chi_j - 1\}, \hat{f}(\omega_j;g)\{\chi_j - 1\}\right] + 2\hat{f}^2(\omega_j;g) d_2^2(\chi_j - 1, \varepsilon_j^* - 1) \\ &\quad = 2\left|f(\omega_j) - \hat{f}(\omega_j;g)\right|^2 E(\chi_1 - 1)^2 + 2\hat{f}^2(\omega_j;g) d_2^2(\chi_1, \varepsilon_1^*). \end{aligned}$$

Therefore, using Corollary A2 and (A2), we conclude that the right-hand side of (A4) converges to 0 in probability if $d_2(\chi_1, \varepsilon_1^*) \to 0$ in probability. To prove the latter convergence, we use

$$d_2(\chi_1, \varepsilon_1^*) \le d_2(\chi_1, \varepsilon_1^0) + d_2(\varepsilon_1^0, \hat{\varepsilon}_1^*) + d_2(\hat{\varepsilon}_1^*, \varepsilon_1^*),$$

where the distributions of ε_1^0 and $\hat{\varepsilon}_1^*$ are the empirical distributions of the true residuals $\varepsilon_1, \dots, \varepsilon_N$ and of the unscaled empirical residuals $\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_N$, respectively.

Theorem 1 and relation (5) of Chen and Hannan (1980) imply that the distribution function of ε_1^0 converges to the distribution function of χ_1 uniformly a.s. for $N \to \infty$ and that

$$E^0(\varepsilon_1^0)^2 = \frac{1}{N}\sum_{j=1}^{N} \varepsilon_j^2 = \frac{1}{N}\sum_{j=1}^{N}\left[\frac{I_T(\omega_j)}{f(\omega_j)}\right]^2 \to E\chi_1^2 \quad \text{a.s.}$$

Therefore, $d_2(\chi_1, \varepsilon_1^0) \to 0$ a.s. by Lemma 8.3 of Bickel and Freedman (1981).

To get an upper bound for $d_2(\varepsilon_1^0, \hat{\varepsilon}_1^*)$, we choose the joint distribution of $(\varepsilon_1^0, \hat{\varepsilon}_1^*)$ such that it assumes the value $(\varepsilon_j, \hat{\varepsilon}_j)$ with probability $1/N$, $j = 1, \dots, N$. Then, by Proposition A1,

$$\begin{aligned} d_2^2(\varepsilon_1^0, \hat{\varepsilon}_1^*) &\le \frac{1}{N} \sum_{k=1}^{N} (\varepsilon_k - \hat{\varepsilon}_k)^2 \\ &= \frac{1}{N} \sum_{k=1}^{N} \left[\frac{1}{f(\omega_k)} - \frac{1}{\hat{f}(\omega_k; h_i)} \right]^2 I_{T,k}^2 \to 0 \quad \text{in probability.} \end{aligned}$$

By exactly the same argument, we also get

$$\begin{aligned} d_2^2(\hat{\varepsilon}_1^*, \varepsilon_1^*) &\le \frac{1}{N} \sum_{k=1}^{N} (\hat{\varepsilon}_k - \tilde{\varepsilon}_k)^2 \\ &= \left[\frac{1}{N} \sum_{k=1}^{N} \hat{\varepsilon}_k^2 \right] \left[1 - \left[\frac{1}{N} \sum_{k=1}^{N} \hat{\varepsilon}_k \right]^{-1} \right]^2 \to 0 \quad \text{in probability,} \end{aligned}$$

by Proposition A1, using $\hat{\varepsilon}_k = I_T(\omega_k)/\hat{f}(\omega_k; h_i)$.

(d) We now start to discuss the bias part B_T. First, we remark that we may neglect the denominators of $b_T(\omega)$ and $b_T^*(\omega)$, as $\hat{f}(\omega; g) \to f(\omega) > 0$ in probability, and as, by Lemma A1 and Proposition A2,

$$\begin{aligned} &\sqrt{Th} \left\{ 1 - \frac{f(\omega)}{\hat{f}(\omega; g)} \right\} b_T(\omega) \\ \text{(A5)} \quad &= \sqrt{Th} \frac{\hat{f}(\omega; g) - E\hat{f}(\omega; g) + E\hat{f}(\omega; g) - f(\omega)}{f(\omega) \hat{f}(\omega; g)} \left[E\hat{f}(\omega; h) - f(\omega) \right] \\ &= O_p(g^2). \end{aligned}$$

By Theorem 6.2.3 of Priestley (1981) we have, with $\Gamma_T(j, k)$ uniformly bounded in j, k, T and with $\delta_{j,k}^* = 1$ for $j = \pm k$ and $\delta_{j,k}^* = 0$ otherwise,

$$\text{(A6)} \qquad \operatorname{cov}(I_{T,j}, I_{T,k}) = \delta_{j,k}^* f^2(\omega_j) + \frac{1}{T} \Gamma_T(j, k) \quad \text{for all } 1 \le |j|, |k| \le N.$$

Using this relation, (A3), Lemma A2 and the compactness of the support of K, a straightforward calculation shows

$$\text{(A7)} \qquad ThE \left\{ \sum_{j=-N}^{N} \gamma_j(\omega; h, g) \left[I_{T,j} - f(\omega_j) \right] \right\}^2 \le \frac{c^* h^3}{g^3} \to 0$$

for suitable $c^* > 0$. As $E^* I_{T,j}^* = \hat{f}(\omega_j; g)$, we have

$$b_T(\omega) - b_T^*(\omega) = \frac{1}{f(\omega)} \left\{ E\hat{f}(\omega; h) - f(\omega) \right\} - \frac{1}{\hat{f}(\omega; g)} \sum_{j=-N}^{N} \gamma_j(\omega; h, g) I_{T,j}.$$

Therefore, using (A5), (A7) and $\hat{f}(\omega; g) \to f(\omega) > 0$ in probability, we finally get $B_T = \sqrt{Th}\,(b_T(\omega) - b_T^*(\omega)) \to 0$ in probability by proving $\sqrt{Th}\,a_T(\omega) \to 0$ with

$$a_T(\omega) = \sum_{j=-N}^{N} \alpha_j(\omega; h) f(\omega_j) - \sum_{j=-N}^{N} \gamma_j(\omega; h, g) f(\omega_j).$$

For this purpose, we split $a_T(\omega)$ into three parts and show that the first and third parts are of order $O(1/(Tg))$ and the second is of order $o(h^2)$:

$$\begin{aligned} a_T(\omega) = {} & \sum_{j=-N}^{N} \alpha_j(\omega; g) f(\omega_j) - \frac{1}{2\pi}\int_{-\infty}^{\infty} K(\theta) f(\omega + \theta g)\, d\theta \\ & + p(\omega; g) - \sum_{j=-N}^{N} \alpha_j(\omega; h) p(\omega_j; g) \\ & + \sum_{j=-N}^{N} \alpha_j(\omega; h)\left\{\frac{1}{2\pi}\int_{-\infty}^{\infty} K(\theta) f(\omega_j + \theta g)\, d\theta \right. \\ & \left. \qquad - \sum_{j=-N}^{N} \alpha_k(\omega_j; g) f(\omega_k)\right\}, \end{aligned} \tag{A8}$$

where

$$p(\omega; g) = \frac{1}{2\pi}\int_{-\infty}^{\infty} K(\theta) f(\omega + \theta g)\, d\theta - f(\omega).$$

As in the proof of Lemma A1, the compactness of the support of K and the Lipschitz continuity of K and f imply that the first part of (A8) is bounded by a constant multiple of $1/(Tg)$ for T large enough. This upper bound is uniform in $|\omega| \le \pi - \kappa g$. Therefore, the third line of (A8) is asymptotically of order $1/(Tg)$ too, because only summands with $|\omega - \omega_j| \le \kappa h$ do not vanish.

Because f is twice continuously differentiable and K is bounded, $p(\omega; g)$ is twice continuously differentiable on $[-\pi + \kappa g,\ \pi - \kappa g]$. Using Lebesgue's theorem on dominated convergence, we conclude that $p(\omega; g)$, $p'(\omega; g)$ and $p''(\omega; g)$ converge to 0 uniformly on $[\omega - \delta,\ \omega + \delta]$ for all $\delta < \pi - |\omega|$. Applying the first part of Lemma A1, we get that the second line of (A8) is asymptotically $o(h^2)$.

(e) The proof of (ii) follows exactly the same lines as the proof of (i), but is easier. In particular, defining D^+ as D^* with $\hat{f}^+$ replacing $\hat{f}^*$, (A4) would be replaced by

$$d_2^2(D^0, D^+) \le Th \sum_{j=-N}^{N} \alpha_j^2(\omega; h) d_2^2\big[f(\omega_j)\{\chi_j - 1\},\ \hat{f}(\omega_j; g)\{\chi_j - 1\}\big] \to 0$$

in probability by Corollary A2, and the rest of part (c) of the proof is not necessary.

Lemma A2. *Let K, h, g satisfy the assumptions of Theorem 1. Then $\gamma_j(\omega; h, g) = O(h/(Tg^2))$ uniformly in $|\omega| \le \pi - \kappa h$. In particular, $\gamma_j(\omega; h, g) = 0$ if $|\omega - \omega_j| > (h + g)\kappa$.*

Proof. By definition of γ_j,

$$|\gamma_j(\omega; h, g)| \le \frac{1}{T^2 hg} \sum_{k=-N}^{N} K\left(\frac{\omega - \omega_k}{h}\right)\left|K\left(\frac{\omega_j - \omega_k}{g}\right) - K\left(\frac{\omega_j - \omega}{g}\right)\right|$$

$$+ \frac{1}{Tg} K\left(\frac{\omega_j - \omega}{g}\right)|S_T(\omega) - 1|.$$

The first term on the right-hand side is of order $h/(Tg^2)$ by Lipschitz continuity of K and (A2). The second term is of order $1/(T^2 gh)$ by (A1). The compactness of the support of K implies $\gamma_j(\omega; h, g) = 0$ for $|\omega - \omega_j| > (h + g)\kappa$. □

Proof of Theorem 3. The proof is a combination of arguments given by Rice (1984) and of results which we have already obtained in the course of proving Theorem 1. We, therefore, only give a sketchy outline of the arguments. We use the notation

$$l(z) = \frac{1}{2\pi}\int_{-\infty}^{\infty} K^2(\theta)\, d\theta \frac{1}{z} + \left[\frac{z^2 f''(\omega)}{2f(\omega)}\right]^2$$

which is the asymptotically dominating part of $T^{4/5}$ MSPE$(\omega; h)$ for $h = zT^{-1/5}$ [compare (2) of Section 2]. Using (A7) and a Taylor expansion argument as in the proof of Lemma A1,

$$\text{(A9)} \qquad \sup_{a \le z \le b} \left|T^{4/5}\, \text{MSPE}(\omega; zT^{-1/5}) - l(z)\right| \to 0 \quad \text{for } T \to \infty$$

for arbitrary $0 < a < b < \infty$. A calculation of derivatives shows that $l(z)$ is strictly convex and infinitely often differentiable on $(0, \infty)$, and that it has z_∞ of (5) as a unique minimum, provided $f''(\omega) \ne 0$. These properties and (A9) imply $T^{1/5}h_0 \to z_\infty$ for $T \to \infty$, provided $a < z_\infty < b$. As the next step, we prove

$$\text{(A10)} \qquad \sup_{h \in B_T} T^{4/5}|\text{MSPE}^*(\omega; h) - \text{MSPE}(\omega; h)| \to 0 \qquad \text{in probability for } T \to \infty.$$

This convergence is shown separately for the variance part and for the bias part of the mean-square percentage error, noticing also that we can forget about the denominators as $\hat{f}(\omega; g) \to f(\omega) > 0$ in probability. The convergence of the variance part of (A10) follows rather easily from Corollary A1, using (A2) and the asymptotic properties (A7) of the periodogram. To prove that the difference of the bootstrap bias and the bias of $\hat{f}(\omega; h)$ itself converges to 0 faster than $T^{-2/5}$, one has to repeat the arguments of part (d) of the proof of Theorem 1, remarking that all of them hold uniformly in $h \in B_T$.

Now (A9), (A10), (5) and the regularity of $l(z)$ imply $T^{1/5}(h_0^* - h_0) \to 0$ in probability, using exactly the same arguments as by Rice (1984) in the proof of his Corollary 2.2. By the first part of Theorem 3, we immediately conclude the second part of Theorem 3 because $l(z)$ is continuous and because, by (A9) and (A10), MSPE*$(\omega; h)$ and MSPE$(\omega; h)$ can both be approximated by $T^{-4/5}l(hT^{1/5})$ uniformly in $h \in B_T$. □

Acknowledgments. This research has been supported by Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 123 Stochastische Mathematische Modelle. A substantial part of this work has been done while the first author, as a visiting fellow, enjoyed the hospitality of the Department of Statistics, IAS, Australian National University, Canberra. We thank Don Poskitt and Enno Mammen for some stimulating remarks on earlier versions of this paper.

REFERENCES

BELTRÃO, K. I. and BLOOMFIELD, P. (1987). Determining the bandwidth of a kernel spectrum estimate. *J. Time Ser. Anal.* **8** 21–38.

BICKEL, P. J. and FREEDMAN, D. A. (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.* **9** 1196–1217.

BRILLINGER, D. R. (1981). *Time Series—Data Analysis and Theory*. Holden-Day, San Francisco.

CHEN, Z.-G. and HANNAN, E. J. (1980). The distribution of periodogram ordinates. *J. Time Ser. Anal.* **1** 73–82.

EFRON, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7** 1–26.

EFRON, B. and TIBSHIRANI, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy (with discussion). *Statist. Sci.* **1** 54–77.

FRANKE, J. (1987). On the choice of local bandwidth for kernel spectral estimates using the bootstrap. Preprint 409, SFB 123 Stochastische Mathematische Modelle, Univ. Heidelberg.

FREEDMAN, D. A. (1981). Bootstrapping regression models. *Ann. Statist.* **9** 1218–1228.

FREEDMAN, D. A. (1984). On bootstrapping two-stage least-squares estimates in stationary linear models. *Ann. Statist.* **12** 827–842.

GASSER, T., MÜLLER, H. G., KÖHLER, W., MOLINARI, L. and PRADER, A. (1984). Nonparametric regression analysis of growth curves. *Ann. Statist.* **12** 210–229.

HANNAN, E. J. (1970). *Multiple Time Series*. Wiley, New York.

HÄRDLE, W. and BOWMAN, A. (1988). Bootstrapping in nonparametric regression: Local adaptive smoothing and confidence bands. *J. Amer. Statist. Assoc.* **83** 100–110.

HARTIGAN, J. A. (1990). Perturbed periodogram estimates of variance. *Internat. Statist. Rev.* **58** 1–7.

KREISS, J. P. and FRANKE, J. (1989). Bootstrapping stationary ARMA-models. *J. Time Ser. Anal.* To appear

KÜNSCH, H.-R. (1989). The jackknife and the bootstrap for general stationary observations. *Ann. Statist.* **17** 1217–1241.

MAJOR, P. (1978). On the invariance principle for sums of independent, identically distributed random variables. *J. Multivariate Anal.* **8** 487–501.

PARZEN, E. (1957). On choosing an estimate of the spectral density function of a stationary time series. *Ann. Math. Statist.* **28** 921–932.

PRIESTLEY, M. B. (1981). *Spectral Analysis and Time Series* **1**. Academic, New York.

RICE, J. (1984). Bandwidth choice for nonparametric regression. *Ann. Statist.* **12** 1215–1230.

Annals of Statistics, 20, 121-145

ROMANO, J. P. (1988). Bootstrapping the mode. *Ann. Inst. Statist. Math.* **40** 565–586.

SHIBATA, R. (1981). An optimal autoregressive spectral estimate. *Ann. Statist.* **9** 300–306.

SHIRYAYEV, A. N. (1984). *Probability*. Springer, Berlin.

SWANEPOEL, J. W. H. and VAN WYK, J. W. Y. (1986). The bootstrap applied to power spectral density function estimation. *Biometrika* **73** 135–141.

WOODROOFE, M. B. and VAN NESS, J. W. (1967). The maximum deviation of sample spectral densities. *Ann. Math. Statist.* **38** 1559–1569.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF KAISERSLAUTERN
P.O. BOX 3049
D-6750 KAISERSLAUTERN
GERMANY

C.O.R.E.
UNIVERSITÉ CATHOLIQUE DE LOUVAIN
34, VOIE DU ROMAN PAYS
B-1348 LOUVAIN-LA-NEUVE
BELGIUM

(1992) Franke, J. and Härdle, W.
On Bootstrapping kernel spectral estimates.

Bandwidth Choice for Average Derivative Estimation

W. HÄRDLE, J. HART, J. S. MARRON, and A. B. TSYBAKOV*

The average derivative is the expected value of the derivative of a regression function. Kernel methods have been proposed as a means of estimating this quantity. The problem of bandwidth selection for these kernel estimators is addressed here. Asymptotic representations are found for the variance and squared bias. These are compared with each other to find an insightful representation for a bandwidth optimizing terms of lower order than n^{-1}. It is interesting that, for dimensions greater than 1, negative kernels have to be used to prevent domination of bias terms in the asymptotic expression of the mean squared error. The extent to which the theoretical conclusions apply in practice is investigated in an economical example related to the so-called "law of demand."

KEY WORDS: Bandwidth optimization; Kernel estimators.

1. AVERAGE DERIVATIVES

The average derivative is the mean slope of a regression curve. A non-parametric formulation of this problem is to use $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathbb{R}^{d+1}$ independent identically distributed, with regression function

$$m(x) = E(Y \mid X = x)\colon \mathbb{R}^d \to \mathbb{R},$$

to estimate

$$\delta = E_X[m'(X)],$$

where

$$m'(x) = [\partial m/\partial x_1, \ldots, \partial m/\partial x_d](x).$$

The average derivative provides useful generalizations of binary response models, as discussed in Manski and McFadden (1981), because it allows modeling the link function in a nonparametric fashion. One such generalization is of one-term projection pursuit type, as defined in Friedman and Stuetzle (1981). This models the regression curve as a function of the form $m(x) = g(x^T\beta)$ for some parameter vector β (identifiable up to scale). If g is nontrivial, then the average derivative is a projection vector in the same direction as β.

In an econometric context this model is called a single index model. For another setting in economic modeling, which can be effectively analyzed by the average derivative technique, we refer the reader to Powell (1986). Average derivatives occur also in the empirical verification of the "law of demand." The law of demand is a condition for the uniqueness of economic equilibria. Uniqueness of equilibria of economic situations is vital for so called comparative statics, where one compares two economies with different price systems. A sufficient condition for the law of demand to hold is that some random matrix (related to "income effects," see Section 3) is positive definite. The elements of this random matrix are average derivatives; for details, see Hildenbrand (1989).

These and other applications were also presented in Härdle and Stoker (1989), where the method is called average derivative estimation (ADE). There it was also shown that δ can be estimated at the rate $n^{-1/2}$. Equipped with this "parametric" rate of convergence of ADE, one sees that the additive model just given allows a one-dimensional rate of convergence for estimation of m. The variance in the asymptotic distribution of the ADE is the best obtainable, as shown by Samarov (1990). However, although the first-order rate for ADE is independent of smoothing parameters, these have to be properly chosen from the data in practice.

The average derivative is a functional of the joint distribution of X and Y. If full information about the regression function $m(x)$ were available to the experimenter, an obvious estimate of δ is a sample average of $m'(x)$ over the X values. However, in general, it is necessary to estimate $m(x)$ or some other nonparametric component of δ as well. In this article we base estimators of the nonparametric components of the average derivative on the kernel method. We use the kernel technique because it is straightforward to implement, easily understood on an intuitive level, and mathematically tractable to analyze. Other possibilities include spline and orthogonal series methods. With any nonparametric method there is a smoothing parameter to be selected, called the *bandwidth* in the kernel case.

The main point of this article is an analysis of how this should be done in the ADE case. Empirical motivation for our theory in a slightly different setting was provided by Hsieh and Manski (1987, p. 551) who stated that "the performance of (adaptive semiparametric) estimates has been shown to be rather sensitive to one's choice of smoothing parameter."

An interesting feature of our results is that the best choice of bandwidth for ADE is substantially smaller (undersmoothed) than the typical bandwidth for curve estimation. This is due to the fact that our goal is estimation of a functional, not the curve itself [see Hall and Marron (1987) or Carroll and Härdle (1989) for the same phenomenon]. Moreover, unlike the curve estimation problem, we will see

* W. Härdle is Professor, C.O.R.E., Université Catholique de Louvain, 34, Voie du Roman Pays, Louvain-la-Neuve, Belgium. J. Hart is Professor, Department of Statistics, Texas A&M University, College Station, TX 77843. J. S. Marron is Professor, Department of Statistics, University of North Carolina, Chapel Hill, NC 27451. A. B. Tsybakov is Senior Researcher, Academy of Sciences of the U.S.S.R., Institute for Problems of Information Transmission, Moscow, U.S.S.R. J. S. Marron's research was partially supported by National Science Foundation Grant DMS-8701201. This research was also supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 303 and CORE. We thank the associate editor and the referees for careful readings and useful suggestions concerning the presentation of this article.

Journal of the American Statistical Association
March 1992, Vol. 87, No. 417, Theory and Methods

Journal of the American Statistical Association, 87, 218-226

that $n^{-1/2}$ rates of convergence are obtainable for estimation of the average derivative functional.

We explicitly state our results in terms of the one-dimensional case, $d = 1$. Generalization to higher dimensional cases is straightforward but involves more refined arguments. One part of this extension is that, to obtain an $n^{-1/2}$ rate of convergence, one must use a higher order kernel. More precisely, the second-order term in the mean squared error expansion for $d = 1$ and a $p = 3$ times differentiable marginal density f of X is $n^{-8/7} = n^{(-4p+4)/(2p+d)}$. Thus one sees that, for $d > 1$, only for $p > (d + 4)/2$ is the next expansion term indeed of lower order. When $d > 1$, the dominant term in a mean squared expansion converges at a slower rate than n^{-1}, unless one uses a higher order kernel, that is, one that takes on negative values.

Section 2 contains a mathematical formulation of the estimator and a statement of the theorem that provides an asymptotic analysis of the bandwidth selection problem, together with a discussion of the practical implications. It is seen that, under common technical assumptions, the rate of decrease of the best bandwidth optimizing second-order terms is of the order $n^{-2/7}$, which results in a mean squared error (MSE) rate of convergence of n^{-1}. Section 3 offers an application to some economic data. The proof of the theorem in Section 2 is given in the Appendix.

2. CHOICE OF BANDWIDTH FOR ADE

If the marginal density $f(x)$ of X vanishes at the boundary, and if we apply partial integration, we can then write the average derivative as

$$\delta = E[m'(X)] = E[Yl(X)],$$

where $l(x) = -f'(x)/f(x)$. If the score function l were known, the average derivative could be estimated by a sample average over $Y_i l(X_i)$. In general, the score function is not available to the experimenter, and, therefore, it is necessary to estimate it from the data as well.

The kernel estimator of the marginal density $f(x)$ is given by

$$\hat{f}_h(x) = n^{-1} \sum_{j=1}^{n} K_h(x - X_j),$$

where $K_h(\cdot) = K(\cdot/h)/h$ for K a kernel function, which will be taken to be a bounded symmetric probability density, and where the scale factor h is called the bandwidth. The subscript of h on the estimator is used because choice of h is crucial to the efficiency of the estimator; see, for example, section 3.4 of Silverman (1986). In the multidimensional case $d > 1$, a product kernel is to be used in the preceding formula for the density estimate. The gradient $f'(x) = (\partial f/\partial x_1, \ldots, \partial f/\partial x_d)$ would then be estimated componentwise by

$$(\hat{f}'_h(x))_k = n^{-1} \sum_{i=1}^{n} \prod_{j \neq k} K_h(x_j - X_{ij}) h^{-2} K'\left(\frac{x_k - X_{ik}}{h}\right).$$

Rates of convergence and asymptotic limiting behavior of multivariate density estimators are well known; for an access to the literature, we refer to Silverman (1986).

The estimate of the derivative $f'(x)$ is, in fact, obtained by differentiating $\hat{f}_h(x)$ with respect to x. We thus form the estimate

$$\hat{l}_h(x) = -\hat{f}'_h(x)/\hat{f}_h(x).$$

The average derivative can then be estimated by

$$\hat{\delta}^*_h = n^{-1} \sum_{i=1}^{n} Y_i \hat{l}_h(X_i).$$

A different approach could be based on a samle average of kernel estimates of $m'(\cdot)$, the derivative of the regression function. It is not hard to see that a sample average of a kernel regression estimator leads to a very similar expression. Indeed, this approach leads to the same variance expressions as has been shown by Stoker (1989). The preceding representation was, historically, developed first and is slightly more tractable since it contains less terms to analyze.

It seems likely that our estimator could be improved by using different bandwidths for $\hat{f}_h$ and $\hat{f}'_h$. A drawback to this approach is that then there are two bandwidths to be selected. For the sake of simplicity in this analysis, we choose to work only with a common bandwidth for the two estimators. It will be apparent from the proof that, after linearization of $l(\cdot)$, only the bandwidth for estimating f' is of interest.

Note that, in the construction of δ_h, the quantities $\hat{f}_h$ and $\hat{f}'_h$ are evaluated only at the points $X_1, \ldots, X_n$. In each instance, this results in one term of the form $K_h(0)$ in the denominator of $\hat{l}_h(X_i)$ (in the numerator such terms vanish since $K'_0(0) = 0$ for symmetric kernels K). While these terms will be asymptotically negligible, as discussed (in a related problem) by Hall and Marron (1987), there can be a small sample difference that makes it desirable to eliminate these terms. Hence define the leave-one-out estimators,

$$\hat{f}_{h,i}(x) = (n-1)^{-1} \sum_{j \neq i} K_h(x - X_j), \qquad \hat{l}_{h,i}(x) = -\frac{\hat{f}'_{h,i}(x)}{\hat{f}_{h,i}(x)}.$$

A modified estimator of δ is given by

$$\hat{\delta}_h = n^{-1} \sum_{i=1}^{n} Y_i \hat{l}_{h,i}(X_i).$$

Inspection of the proofs shows that $\hat{\delta}_h$ is also easier to work with mathematically than $\hat{\delta}^*_h$ because the "diagonal terms" of the form $K_h(0)$ in $\hat{\delta}^*_h$ need to be handled separately.

As with many related estimators, $\hat{\delta}_h$ is technically tricky to handle because of the random denominator appearing in $\hat{l}_{h,i}(x)$. The approach to this problem taken here is similar to the linearization method used in chapter 3 of Härdle (1990). It will become apparent from the proof of the next lemma that $\hat{\delta}_h$ may, for purposes of analysis, be replaced by the "linearized" version

$$\tilde{\delta}_h = n^{-1} \sum_{i=1}^{n} Y_i L_{hi}(X_i),$$

where $L_{hi}(x) = \hat{f}'_{h,i}(x)(\hat{f}_{h,i}(x) - 2f(x))/f(x)^2$. Technical assumptions used here are:

(1992) Härdle, W., Hart, J., Marron. J.S. and Tsybakov, A.B.
Bandwidth Choice for Average Derivative Estimation.

A1. The kernel K is bounded, continuously differentiable, symmetric, and compactly supported.

A2. $\int K(u)\, du = 1$.

A3. There exists $k, k' > 0$, such that $K^2(u) \geq k'I(|u| \leq 1/k)$.

A4. $f(x)$ has three continuous derivatives on its support, and support $(f) = (a, b)$, for $-\infty < a < b < \infty$.

A5. $f''(a) \neq 0$, and $f''(b) \neq 0$.

A6. support$_x[f^{-1}(x)E(|Y| \mid X = x)] < \infty$.

A7. $h_n = h_0 n^{-2/7}$, where h_0 is some positive number.

The first four conditions are common conditions in the setting of kernel smoothing ensuring regularity of both K and f. In the multidimensional case, $d > 1$, assumption A4 has to be replaced by a cube, for example, $(a, b)^d$. Assumption A5 is introduced to control the curvature at the boundary; again it can be modified for the multidimensional case; see formula (A.1.1) in the Appendix. Assumption A6 is a growth condition necessary to control the random denominators. Assumption A7 is a condition on the rate of h already predefining the optimal range of h. It could be modified to a slightly larger range at the expense of more complicated mathematics.

The following Lemma guarantees that the replacement of $\hat{l}_{h,i}$ by L_{hi} is possible.

Linearization Lemma 1. Under assumptions A1–A7,

$$\sqrt{n}(\hat{\delta}_h - \tilde{\delta}_h) = o_p(n^{-1/14}), \qquad n \to \infty.$$

In the multidimensional case the rate $n^{-1/14}$ has to be replaced by $n^{-1/[2(2p+d)]}$.

That this bound is enough to enable replacement of δ_h by $\tilde{\delta}_h$, and that the bandwidth speed given in A7 is reasonable, are a consequence of the following theorem, which is the main result of this article. Additional technical assumptions are:

B1. $m(x)$ is three times continuously differentiable on $\mathbb{R}$, and $m(x)l(x)$ is Lipschitz.

B2. The conditional variance $\sigma^2(x) = m_2(x) - m^2(x)$ and the function $m(x)/f(x)$ are continuous, and the integrals $\int_a^b m_2(x)[(f'(x))^2/f(x)]dx$ and $\int_a^b [m_2(x)/f^3(x)]dx$ are finite. Here $m_2(x) = E(Y^2 \mid X = x)$.

Theorem 1. Under assumptions A1, A2, A4, A6, A7, B1, and B2,

$$E(\tilde{\delta}_h - \delta)^2 = Q_1 n^{-1} + Q_2 n^{-2} h_n^{-3} + Q_3 h_n^4 + o(n^{-2} h_n^{-3} + h_n^4),$$

where

$$Q_1 = \operatorname{var}(m'(X)) + E(\sigma^2(X) l^2(X)),$$

$$Q_2 = \int \sigma^2(x)dx \int (K'(t))^2\, dt,$$

and

$$Q_3 = \left(\int \frac{m(x)}{f(x)} (f'(x)f''(x) - f(x)f'''(x))\, dx\right)^2 \left(\frac{1}{2} \int t^2 K(t)dt\right)^2.$$

Corollary. The asymptotically optimal h_n is given by

$$h_n = h_0^* n^{-2/7}, \qquad h_0^* = (3Q_2/4Q_3)^{1/7}.$$

Under this choice of h_n, the first two terms of the asymptotic expansion are

$$Q_1 n^{-1} + Q_2^{4/7} Q_3^{3/7} \left(\left(\frac{4}{3}\right)^{3/7} + \left(\frac{3}{4}\right)^{4/7}\right) n^{-8/7}. \qquad (2.1)$$

The theorem and the corollary generalize to dimension $d > 1$, as explained in the Appendix.

Note that the second term is not particularly small in comparison to the first one, since their ratio is of order $n^{-1/7}$. Therefore, recalling the preceding observation by Hsieh and Manski (1987), while the choice of h is asymptotically negligible, extremely large n will be required before its influence disappears in a practical sense. The constants in (2.1) can be optimized.

Optimization of Q_1

The n^{-1} term with constant Q_1 is the leading term in the MSE expansion of the ADE. This constant cannot be improved upon in a minimax sense due to Levit (1974). Samarov (1990) proved that this first-order term Q_1 is the smallest achievable for any possible estimate of the average derivative.

Optimization of K

The second-order terms in (2.1) involve the kernel K. So, it is natural to ask whether the factor $Q_2^{4/7}\, Q_3^{3/7}$ can be optimized over the choice of kernel. Note that this is the same as seeking to minimize

$$T(K) = \left(\int (K')^2\right)^4 \left(\int u^2 K\right)^3.$$

Mammitzsch (1989) has solved this problem by showing that K' is of order (1, 3), in the terminology of Gasser, Müller, and Mammitzsch (1985). Integrating K' leads to the *quartic kernel* $K(u) = (15/16)(1 - u^2)^2 I(|u| \leq 1)$ as the kernel optimizing $T(K)$.

Optimization of h for $d > 1$

The proof of the Theorem can be extended to the case of higher dimensional X variables. The rate in the stochastic term will be, as known from other semiparametric problems, of the order $n^{-2}h^{-d-2}$. It is interesting that the bias for three-times differentiable f would be of the order as in the one-dimensional case, namely, $h^{2(p-1)}$, where $p = 3$ denotes the degree of differentiability of f. Observe now that $p = 3$ as a degree of smoothness of f is no longer feasible for $d > 1$. To speed up the rate of convergence for the bias term, we have to assume that more derivatives exist, and we have to use higher order kernels (Gasser, Müller, and Mammitzsch 1985) to obtain a rate $h^{2(p-1)}$ faster than n^{-1}. If, for example, we are in a $d = 4$ dimensional setting, we should use a kernel of, say, order $p = 6 > 4$, since then the bias term is of order $h^{2(p-1)} = h^{10}$, yielding a rate of $n^{-4p+4/(2p+d)}$. The h optimizing the second-order terms would be, in this setting, $h \sim n^{-1/8}$ More generally, for $p > (d + 4)/2$, the best bandwidth is given by $h \sim n^{-2/(2p+d)}$, yielding a rate of $n^{-(4p+4)/(2p+d)}$.

Journal of the American Statistical Association, 87, 218-226

3. THE METHOD IN PRACTICE

Empirical verification of the so-called law of demand, see Hicks (1956), provides one motivation for the average derivative estimation method. The law of demand concerns effects of price changes when a household's budget is fixed. A sufficient condition for the law of demand to hold is positive definiteness of the matrix of mean income effects. The (k, l) component of this matrix is the demand for good l multiplied by the derivative of demand for good k, with respect to income, all averaged over the population. Hildenbrand (1989) used nine goods and showed that this is equivalent to checking positive definiteness of the matrix

$$a_{kl} = \int \frac{d}{dx} E(Z_k Z_l \mid X = x) f(x)\, dx, \qquad k, l = 1, \ldots, 9.$$

Here Z_k and Z_l denote demands for goods k and l and $f(x)$ is the density of the income distribution. The matrix elements a_{kl} are thus average derivatives computed for $Y = Z_k Z_l$.

To give some insight into this data structure, consider Figure 1. It shows an estimate of $E(Z_k Z_l \mid X = x)$ for $k =$ FOOD and $l =$ TRANSPORT. The data are from the Family Expenditure Survey from the Department of Employment, Statistics Division (1968–1983) for the year 1973. The average derivative for this example, with a quartic kernel and $h = .2$, was $\hat{\delta} = .06$. For a picture of the income density $f(x)$, see Härdle (1990, chap. 1).

The bandwidth selection procedure was performed for a variety of those matrix elements for different years. We give an example from the year 1973 based on plugging in estimates $\hat{Q}_2$ and $\hat{Q}_3$ of Q_2 and Q_3 in Theorem 1. To estimate the constants in Theorem 1 most conveniently, we used kernel estimates with the Gaussian kernel $K(u) = (1/\sqrt{2\pi}) \exp(-u^2/2)$.

The reason for choosing the Gaussian kernel was that derivative estimates as occurring in Q_3 can be easily computed without referring to other special derivative kernels.

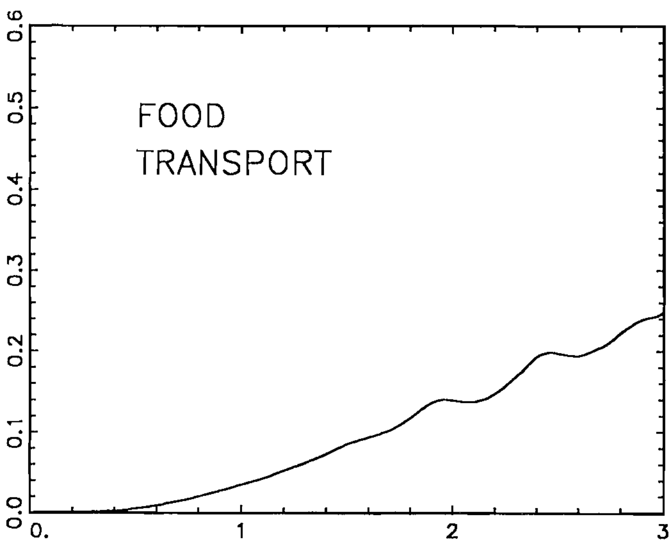


Figure 1. The Estimated Mean Product Function $E(Z_k Z_l \mid Z = x)$ for Food and Transport for 1973 (X and Z are normalized by their mean). From Hildenbrand (1989).

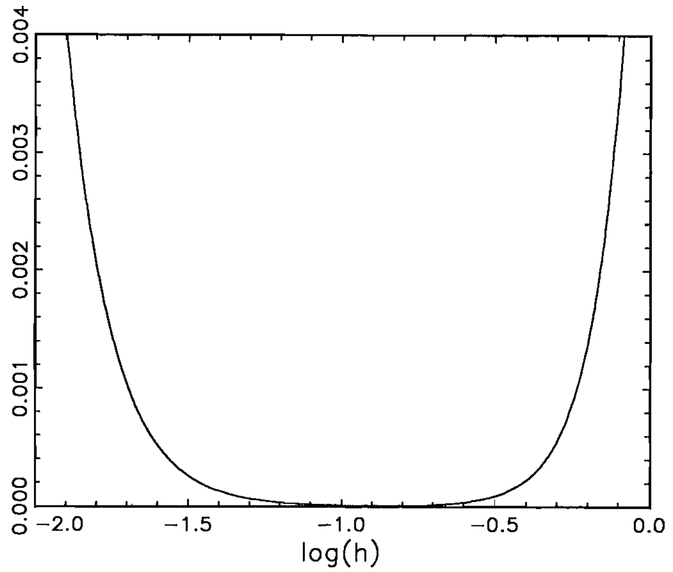


Figure 2. The Bandwidth Selection Function $\hat{Q}_2 n^{-2} h^{-3} + \hat{Q}_3 h^4$ for the Food and Transport Example.

We used numerical quadrature methods to compute the integral $(\int [m(x)/f(x)] (f'(x)f''(x) - f(x)f'''(x))dx)^2$, which is part of Q_3. The bandwidths for estimating the unknown curves in this constant Q_3 were chosen by cross-validation, using the techniques of Härdle, Marron, and Wand (1989). We are aware of the fact that this introduces further noise in a third-order level, but we varied the bandwidths for estimating Q_2 and Q_3 and found little difference in the estimated optimal h. One might suggest that optimizing a choice of bandwidths for estimating f', etc., as in Härdle, Marron, and Wand (1989), is not reasonable, given the perspective of this article. (It does not solve the very complicated problem of finding a "best" bandwidth for estimating Q_3). But note that this is yet another theory for optimizing estimation of Q_3. The selected bandwidths for estimating Q_2 and Q_3 were around .2.

Figure 2 shows the curve $\hat{Q}_2 n^{-2} h^{-3} + \hat{Q}_3 h^4$ for the Food and Transport example ($n = 7{,}123$). For this curve we also used the Gaussian kernel to compute $\int (K')^2$ and $\int u^2 K$ most easily. The curve has its minimum around $h \approx .1$. This bandwidth of $h = .1$ does not, of course, correspond directly to the bandwidth used for Figure 1, where we used the quartic kernel. To obtain an interpretable value of this bandwidth in the scale of the quartic kernel used previously, we refer the reader to the canonical kernel technique, as described in Härdle (1990, chap. 4.5).

APPENDIX: PROOFS OF THE THEOREM AND LEMMA

A.1 Proof of Theorem 1

Write the estimate $\tilde{\delta}$ as $\tilde{\delta} = (1/n) \sum_{i=1}^{n} (m(X_i) + \epsilon_i) L_{hi}(X_i)$, where $\epsilon_i = Y_i - m(X_i)$. Since $E(\epsilon_i \mid X_i) = 0$,

$$\begin{aligned} E(\tilde{\delta} - \delta)^2 &= E\left(\frac{1}{n} \sum_{i=1}^{n} \epsilon_i L_{hi}(X_i)\right)^2 \\ &\quad + E\left(\frac{1}{n} \sum_{i=1}^{n} m(X_i) L_{hi}(X_i) - \delta\right)^2 \\ &= V_1 + V_2 - 2\delta E(\tilde{\delta}) + \delta^2, \end{aligned} \tag{A.1.1}$$

(1992) Härdle, W., Hart, J., Marron. J.S. and Tsybakov, A.B.
Bandwidth Choice for Average Derivative Estimation.

where $V_1 = E((1/n)\ \Sigma_{i=1}^{n}\ \epsilon_i L_{hi}(X_i))^2$ and $V_2 = E((1/n)\ \Sigma_{i=1}^{n}\ m(X_i)L_{hi}(X_i))^2$. Note that

$$L_{hi}(X_i) = f^{-2}(X_i)\frac{1}{(n-1)^2}\sum_{\substack{k,j=1\\k,j\neq i}}^{n} R_{kj}(X_i),$$

where $R_{kj}(x) = K_h'(x - X_k)(K_h(x - X_j) - 2f(x))$, for $k, j = 1, \dots, n$. Hence

$$\begin{aligned} V_1 &= \frac{1}{n}\int \sigma^2(X_1)E(L_{h1}^2(X_1))f(X_1)\,dX_1 \\ &= \frac{1}{n}\int \sigma^2(x)f^{-3}(x)\left[\frac{1}{(n-1)^4}E\left(\sum_{i,j,i',j'=2}^{n} R_{ij}(x)R_{i',j'}(x)\right)\right]dx. \end{aligned} \tag{A.1.2}$$

The sum $\Sigma_{i,j,i',j'=2}^{n}$ can be represented as

$$\sum_{i,j,i',j'=2}^{n} = I + II + III + IV,$$

where $I = \Sigma_{i=1}^{n}\ R_{ii}^2(x)$ corresponds to the case $i = j = i' = j'$, II is the double sum over i, j that contains the products $R_{ii}R_{ij}$, $R_{ii}R_{ji}$, R_{ij}^2, $R_{ii}R_{jj}$, III is the triple sum over i, j, i' that contains the products $R_{ij}R_{ii'}$, $R_{ij}R_{i'j,}R_{ij}R_{ji'}$, $R_{ij}R_{i',i'}$, and IV is the sum over the quadruples (i, j, i', j'), with i, j, i', j' pairwise different from each other. (A.3.2)–(A.3.5) imply that

$$\begin{aligned} &E\left[\frac{1}{(n-1)^4}\sum_{i,j,i',j'=2}^{n} R_{ij}(x)R_{i'j'}(x)\right] \\ &= (f'(x)f(x))^2 + n^{-1}h^{-3}f^3(x)\int (K'(t))^2\,dt + o(n^{-1}h^{-3}). \end{aligned} \tag{A.1.3}$$

Substitution of (A.1.3) into (A.1.2) yields

$$\begin{aligned} V_1 &= \frac{1}{n}\int \sigma^2(x)\frac{(f'(x))^2}{f(x)}\,dx + n^{-2}h^{-3}\int \sigma^2(x)\,dx \\ &\quad \times \int (K'(t))^2\,dt + o(n^{-2}h^{-3}). \end{aligned} \tag{A.1.4}$$

Next,

$$\begin{aligned} V_2 &= E\left[\frac{1}{n}\sum_{i=1}^{n} m(X_i)L_{hi}(X_i)\right]^2 \\ &= n^{-2}\sum_{i,j=1;i\neq j}^{n} E[m(X_i)L_{hi}(X_i)m(X_j)L_{hj}(X_j)] \\ &\quad + n^{-2}\sum_{i=1}^{n} E[m(X_i)L_{hi}(X_i)]^2 = B_1 + B_2. \end{aligned} \tag{A.1.5}$$

Equation (A.1.3) implies that

$$\begin{aligned} B_2 &= \frac{1}{n}E(m^2(X_1)L_{h1}^2(X_1)) \\ &= \frac{1}{n}\int m^2(x)f^{-3}(x)\frac{1}{(n-1)^4}E\left[\sum_{i,j,i',j'=2}^{n} R_{ij}(x)R_{i'j'}(x)\right]dx \\ &= \frac{1}{n}\left(W_1 + \frac{1}{n}W_2\right) + o(n^{-2}h^{-3}), \end{aligned} \tag{A.1.6}$$

where $W_1 = \int m^2(x)[(f'(x))^2/f(x)]\,dx$ and $W_2 = h^{-3}\int m^2(x)\,dx \int (K')^2$. Using (A.3.1) we obtain

$$\begin{aligned} E(m(X_1)L_{h1}(X_1)) &= E(\tilde{\delta}) \\ &= \int m(x)f^{-1}(x)\frac{1}{(n-1)^2}E\left[\sum_{i,j=2}^{n} R_{ij}(x)\right]dx \\ &= \frac{n-2}{n-1}\delta + \frac{n-2}{n-1}\sqrt{Q_3}h^2 + O(n^{-1}h^{-1}) \\ &= \delta(1 - n^{-1}) + \sqrt{Q_3}h^2 + O(n^{-1}h^{-1}). \end{aligned} \tag{A.1.7}$$

Rewrite B_1 as

$$\begin{aligned} B_1 &= \frac{n-1}{n}\frac{1}{(n-1)^4}E\left[S_1S_n\sum_{s,l=2}^{n} R_{sl}(X_1)\sum_{r,p=1}^{n-1} R_{rp}(X_n)\right] \\ &= \frac{1}{n(n-1)^3}(I_1 + I_2 + I_3), \end{aligned} \tag{A.1.8}$$

where $S_i = [m(X_i)/f^2(X_i)]$ and

$$I_1 = E\left[S_1S_n\sum_{s,l=2}^{n-1} R_{sl}(X_1)\sum_{r,p=1}^{n-1} R_{rp}(X_n)\right],$$

$$I_2 = 2E\left\{S_1S_n\sum_{s,l=2}^{n-1} R_{sl}(X_1)\left[\sum_{r=1}^{n-1} R_{r1}(X_n) + \sum_{p=2}^{n-1} R_{1p}(X_n)\right]\right\},$$

and

$$\begin{aligned} I_3 = E\Bigg[S_1S_n\Bigg(&\sum_{s=2}^{n} R_{sn}(X_1) \\ &+ \sum_{l=2}^{n-1} R_{nl}(X_1)\Bigg)\left(\sum_{r=1}^{n-1} R_{r1}(X_n) + \sum_{p=2}^{n-1} R_{1p}(X_n)\right)\Bigg]. \end{aligned}$$

Denote

$$\begin{aligned} U_{sl} &= E_{X_1}(S_1R_{sl}(X_1)) \\ &= \int \frac{m(x)}{f(x)}K_h'(x - X_s)K_h(x - X_l)\,dx + 2m'(X_s) + O(h^2). \end{aligned}$$

Then $I_1 = E(\Sigma_{s,l,r,p=2}^{n-1}\ U_{sl}U_{rp}) = \Sigma_{q=1}^{5}\ I_{1q}$, where

$$I_{11} = E\left[\sum_{r\neq s,r\neq l,p\neq s,p\neq l} U_{sl}U_{rp}\right] = \sum_{r\neq s,r\neq l,p\neq s,p\neq l} E(U_{sl})E(U_{rp}),$$

$$I_{12} = E\left[\sum_{s,l=2}^{n-1} U_{sl}\sum_{p=2}^{n-1} U_{sp}\right], \qquad I_{13} = E\left[\sum_{s,l=2,l\neq s}^{n-1} U_{sl}\sum_{p=2}^{n-1} U_{lp}\right],$$

$$I_{14} = E\left[\sum_{s,l=2}^{n-1} U_{sl}\sum_{r=2,r\neq s,r\neq l}^{n-1} U_{rs}\right] \quad \text{and}$$

$$I_{15} = E\left[\sum_{s,l=2,l\neq s}^{n-1} U_{sl1}\sum_{r=2,r\neq s,r\neq l}^{n-1} U_{rl}\right].$$

By Lemma 2.1

$$\begin{aligned} \tilde{U} &= E(U_{sl}) = E_{X_1}(S_1E(R_{sl} \mid X_1) \mid X_0)) \\ &= \delta + \sqrt{Q_3}h^2 + o(h^2),\ s \neq l, \end{aligned}$$

and

$$\overline{U} = E(U_{ss}) = O(h^{-1}).$$

It can be easily seen that

$$I_{11} = (n-2)^4\left[\left(\tilde{U} + \frac{1}{n-2}\overline{U}\right)^2 - \frac{6\tilde{U}^2}{n-2} + O\left(\frac{1}{n^2h^2}\right)\right],$$

and hence

$$\frac{1}{n(n-1)^3} I_{11} = \left[1 - \frac{5}{n} + O\left(\frac{1}{n^2}\right)\right] \times \left[\left(\tilde{U} + \frac{\bar{U}}{n-2}\right)^2 - \frac{6\tilde{U}^2}{n-2} + O\left(\frac{1}{n^2h^2}\right)\right]$$
$$= \left[\delta + \sqrt{Q_3}h^2 + o(h^2) + O\left(\frac{1}{nh}\right)\right]^2 - \frac{11\delta^2}{n} + O\left(\frac{h^2}{n} + \frac{1}{n^2h^2}\right). \qquad \text{(A.1.9)}$$

The main term of l_{12} is that with $s \neq l$ and $l \neq p$. Direct calculation shows that

$$E(U_{sl}U_{sp}) = \int (m'(x))^2 f(x)\, dx + O(h), \qquad l \neq p, s \neq l,$$
$$E(U_{sl}^2) = W_3(1 + o(1)), \qquad s \neq l,$$
$$W_3 = h^{-3} \int m^2(u)\, du \left[\int K'(t)K(v + t)\, dt\right]^2 dv,$$

and

$$\frac{1}{n(n-1)^3} I_{12} = \frac{1}{n}\left[\int (m'(x))^2 f(x)\, dx + \frac{1}{n} W_3\right] + o\left(\frac{1}{n^2h^3}\right). \qquad \text{(A.1.10)}$$

Similarly, the main term of I_{13} contains the summands with $s \neq l$ and $s \neq p$:

$$E(U_{sl}U_{lp}) = W_4 + 2\delta^2 + O(h),$$

where $W_4 = \int m'(x)m(x)f'(x)\, dx$, $E(U_{sl}U_{ls}) = -W_3(1 + o(1))$, for $s \neq l$, and

$$\frac{1}{n(n-1)^3} I_{13} = \frac{1}{n}\left(W_4 - \frac{1}{n} W_3\right) + \frac{2\delta^2}{n} + o\left(\frac{1}{n^2h^3}\right). \qquad \text{(A.1.11)}$$

Note that I_{14} is just the part of I_{13} corresponding to the case $s \neq p$. Therefore,

$$\frac{1}{n(n-1)^3} I_{14} = \frac{1}{n} W_4 + \frac{2\delta^2}{n} + o\left(\frac{1}{n^2h^3}\right). \qquad \text{(A.1.12)}$$

Moreover,

$$\frac{1}{n(n-1)^3} I_{15} = \frac{1}{n} W_1 + o\left(\frac{1}{n^2h^3}\right), \qquad \text{(A.1.13)}$$

since $E(U_{sl}U_{rl}) = W_1 + O(h)$, for $s \neq l$, $s \neq r$, and $r \neq l$. To evaluate I_2 we split it into three terms: $I_2 = I_{21} + I_{22} + I_{23}$, where

$$I_{21} = 2E\left[S_1S_n \sum_{s,l,r=2}^{n-1} R_{sl}(X_1)R_{r1}(X_n)\right],$$
$$I_{22} = 2E\left[S_1S_n \sum_{s,l,p=2}^{n-1} R_{sl}(X_1)R_{1p}(X_n)\right],$$

and

$$I_{23} = 2E\left[S_1S_n \sum_{s,l=2}^{n-1} R_{sl}(X_1)R_{11}(X_n)\right].$$

Using Lemma 2.1 and the Lipschitz condition on $m(x)l(x)$, one obtains

$$E(S_1R_{sl}(X_1)U_{rl}) = -W_1 + 2\delta^2 + O(h), \qquad s \neq l, l \neq r, r \neq s.$$

and

$$E(S_1R_{sl}(X_1)U_{sl}) = -W_5,$$

where

$$W_5 = -\frac{1}{h^3}\left[\int m^2(x)\, dx \iint K'(w)K' \times (w - t)K(t)\, dw\, dt + o(1)\right], \qquad s \neq l,$$

and

$$\frac{1}{n(n-1)^3} I_{21} = -\frac{2}{n^2} W_1 - \frac{2}{n^2} W_5 + \frac{4\delta^2}{n} + o\left(\frac{1}{n^2h^3}\right). \qquad \text{(A.1.14)}$$

Next, the main term of $I_{22}/2$ corresponds to $s \neq l$, $l \neq p$, and $s \neq p$, and its summands are $E(S_1R_{sl}(X_1)U_{1p}) = -W_4 + O(h)$ and $E(S_1R_{sl}(X_1)U_{1s}) = W_5$.

Therefore,

$$\frac{1}{n(n-1)^3} I_{22} = -\frac{2}{n} W_4 + \frac{2}{n^2} W_5 + o(n^{-2}h^{-3}). \qquad \text{(A.1.15)}$$

Finally, using Lemma 2.1 and the fact that $U_{11} = O(h^{-2})$, we obtain

$$\frac{1}{n(n-1)^3} I_{23} = o(n^{-2}h^{-3}). \qquad \text{(A.1.16)}$$

Considering I_3, we see that the nonnegligible part of it is

$$E\left(S_1S_n \sum_{\substack{l,p=2 \\ l \neq p}}^{n=1} R_{nl}(X_1)R_{1p}(X_n)\right) = (n-2)(n-3)(-W_2 + O(1)),$$

and thus

$$\frac{1}{n(n-1)^3} I_3 = -\frac{1}{n^2} W_2 + o(n^{-2}h^{-3}). \qquad \text{(A.1.17)}$$

Summing up (A.1.6), (A.1.9)–(A.1.17), and using (A.1.8), we have

$$V_2 = B_1 + B_2$$
$$= \left(\delta + \sqrt{Q_3}h^2 + o(h^2) + O\left(\frac{1}{nh}\right)\right)^2 - \frac{3\delta^2}{n} + o(n^{-2}h^{-3}). \qquad \text{(A.1.18)}$$

(Note that all W_j, $j = 1, 2, 3, 4$, cancel out.) Finally, substitute (A.1.4), (A.1.7), and (A.1.18) into (A.1.1). This proves the theorem.

A.2. Proof of Lemma 1

Introduce the following notation,

$$\xi(x) = \int K^2(u)f(x + uh_n)\, du, \qquad K_{\max} = \max_x |K(x)|,$$
$$D = \operatorname{diam} \operatorname{supp} K, \quad \text{and} \quad d = \max_x f(x),$$

and put without loss of generality $h_0^* = 1$. In the following, C_i denote positive constants and h_n is abbreviated as h.

Note that Assumption (A5) entails that there exist A, B, and $\Delta > 0$ such that

$$A|z - a|^2 \leq f(z) \leq B|z - a|^2, \qquad a \leq z \leq a + \Delta,$$

and

$$A|z - b|^2 \leq f(z) \leq B|z - b|^2, \qquad b - \Delta \leq z \leq b. \qquad \text{(A.2.1)}$$

To prove Lemma 1 we need some preliminary steps.

Lemma 1.1. Under (A4) and (A5)

$$\limsup_{\tau \downarrow 0} P\{f(X) \leq \tau\}/\tau^{3/2} < \infty.$$

Proof. Assume that τ is small enough so that

$$\{x : f(x) \le \tau\} \subseteq \{a \le x \le a + \Delta\} \cup \{b - \Delta \le x \le b\}.$$

Then

$$P\{f(X) \le \tau\} \le \int_a^{a+\Delta} I\{f(x) \le \tau\} f(x)\, dx + \int_{b-\Delta}^{b} I\{f(x) \le \tau\} f(x)\, dx.$$

Next, apply (A.2.1).

Lemma 1.2. Assume (A1) to (A5). Then

$$\xi(x) \ge C_1 h^2. \qquad \text{(A.2.2)}$$

Proof. Let L be the Lipschitz constant for f. Then

$$\xi(x) \ge f(x) \int K^2(u)\, du - Lh \int |u| K^2(u)\, du. \qquad \text{(A.2.3)}$$

If $f(x) \ge [2L \int |u|K^2(u)du / \int K^2(u)du]\ h = C_2 h$, then $\xi(x) \ge (C_2/2)h$, so that (A.2.2) holds.

Now suppose that $f(x) < C_2 h$. If n is large enough, then

$$\{f(x) < C_2 h\} \subset \{a \le x \le a + \Delta\} \cup \{b - \Delta \le x \le b\}$$

and we can apply (A.2.1). Suppose, as before, that we are on the set $\{a \le x \le a + \Delta\}$. Here again we have two cases: (1) $x - a \le C_3 h/k$, and (2) $x - a > C_3 h/k$, where

$$C_3 = 1 + \left(\frac{A}{2B}\right)^{1/3}$$

First estimate ξ in the case (1). In view of Assumption A3 and (A.2.1), one obtains

$$\xi(x) \ge k' \int_{-1/k}^{1/k} f(x + uh)\, du$$
$$= \frac{k'}{h}\left[\int_a^{x+h/k} f(t)\, dt - \int_a^{x-h/k} f(t)\, dt\right]$$
$$\ge \frac{k'}{h}\left[A \int_a^{a+h/k} (t - a)^2\, dt - B \int_a^{a+(C_3-1)h/k} (t - a)^2\, dt\right].$$

Computing the integrals and using the definition of C_3, we obtain the assertion of the Lemma in the case (1). If case (2) is true, then

$$\int_{x-h/k}^{x+h/k} f(t)\, dt \ge A \int_{x-h/k}^{x+h/k} (t - a)^2\, dt$$
$$\ge \frac{2A}{k^3}(C_3 - 1)^3 h^3$$

Lemma 1.3. Under Assumptions (A1–A7) we have

$$P\left\{(\hat{f}_{h,i}(X_i) - f(X_i))^2/\xi(X_i) \ge \eta \frac{\log n}{nh}\right\}$$
$$\le 2 \exp(-C_4 \sqrt{\eta} \log n), \qquad \text{(A.2.4)}$$

for all $\eta > 0$ large enough.

Proof. Set $\epsilon = \eta \log n/(nh)$. Then

$$P\{(\hat{f}_{h,i}(X_i) - f(X_i))^2/\xi(X_i) \ge \epsilon\}$$
$$= E_{X_i}(P\{(\hat{f}_{h,i}(X_i) - f(X_i))^2/\xi(X_i) \ge \epsilon \mid X_i\}).$$

Fix some i and denote for brevity $x = X_i$ and $\hat{f}_{h,i}(\mathbf{X_i}) = \tilde{f}_h(x)$. Now it is sufficient to prove that the right side of (A.2.4) bounds, from above, the probability

$$p_n = P_{X_j, j \ne i}\{(\tilde{f}_h(x) - f(x))^2/\xi(x) \ge \epsilon\}.$$

Here

$$(\tilde{f}_h(x) - f(x))^2 \le 2(\tilde{f}_h(x) - E(\tilde{f}_h(x)))^2 + 2(E(\tilde{f}_h(x)) - f(x))^2. \qquad \text{(A.2.5)}$$

Now, by Assumptions A1, A2, and A3,

$$(E(\tilde{f}_h(x)) - f(x))^2$$
$$\le \left(\frac{d}{n} + \int K(u)(f(x + uh) - f(x))\, du\right)^2$$
$$\le 2\left(\frac{d^2}{n^2} + C_5 h^4\right) \le C_6 n^{-8/7}. \qquad \text{(A.2.6)}$$

Note that, by Lemma 1.2,

$$\frac{\xi(x)\epsilon}{2} = \frac{\xi(x)}{2} \eta \frac{\log n}{nh}$$
$$\ge C_1 \frac{\eta}{2} (\log n) n^{-8/7} > 2C_6 n^{-8/7}$$

if n is large enough. Using this (A.2.5), (A.2.6), and applying the Bernstein inequality, (Serfling 1980, p. 95) we obtain

$$p_n \le P\left\{(\tilde{F}_h(x) - E(\tilde{f}_h(x)))^2 \ge \frac{\epsilon\xi(x)}{2} - C_6 n^{-8/7}\right\}$$
$$\le P\left\{(\tilde{f}_h(x) - E(\tilde{f}_h(x)))^2 \ge \frac{\epsilon\xi(x)}{4}\right\}$$
$$\le 2 \exp\left(-\frac{(n - 1)(\sqrt{\epsilon\xi(x)}/2)^2}{2\sigma_n^2 + [K_{\max}\sqrt{\epsilon\xi(x)}/3h_n]}\right), \qquad \text{(A.2.7)}$$

for n large enough. Here $\sigma_n^2 = E\{(1/h_n^2\, K^2[(x - X_j)/h_n]\} = (1/h_n)\xi(x)$. By Lemma 1.3 the last expression in (A.2.7) does not exceed

$$2 \exp\left(-\frac{(n - 1)h_n\epsilon/4}{2 + K_{\max}\sqrt{\epsilon C_1^{-1} h_n^{-2}}}\right) \le 2 \exp(-C_4 \sqrt{\eta} \log n)\Big)$$

for n and η large enough.

Lemma 1.4. If η is large enough, then $P\{\mathcal{B}\} = o(1)$, $n \to \infty$, where

$$\mathcal{B} = \left\{\max_{i=1,\ldots,n} (\hat{f}_{h,i}(X_i) - f(X_i))^2/\xi(X_i) \ge \eta \frac{\log n}{nh}\right\}.$$

Proof. Using Lemma 1.3 we have

$$P\{\mathcal{B}\} \le \sum_{i=1}^{n} P\left\{(\hat{f}_{h,i}(X_i) - f(X_i))^2/\xi(X_i) \ge \eta \frac{\log n}{nh}\right\}$$
$$\le 2n \exp(-C_4 \sqrt{\eta} \log n) = o(1),$$

for η large enough.

Lemma 1.5. Under Assumptions (A1–A7),

$$\max_{i=1\ldots n} |\hat{f}'_{h,i}(X_i)| = O_p(1), \qquad n \to \infty.$$

This is proved by standard techniques of nonparametric estimation (see, for example, Stone 1982).

Lemma 1.6. Let Assumptions (A4), (A5), and (A7) hold. Then $P(\mathcal{A}) = o(1)$, $n \to \infty$, where

$$\mathcal{A} = \{f(X_i) \le C_7 \log n/(nh) \text{ for some } i\}.$$

Proof. Use the Bonferroni inequality and Lemma 1.1. Then

$$P(\mathcal{A}) \le nP\{f(X) \le C_7 \log n/(nh)\}$$
$$\le C_8 n \left(\frac{\log n}{nh}\right)^{3/2} = o(1), \qquad n \to \infty.$$

Proof of Lemma 1 (Linearization Lemma)

Suppressing dependence on x, h, and i for notational simplicity, observe that

$$\begin{aligned}\hat{l} - L &= -\hat{f}'/\hat{f} - \hat{f}'(\hat{f} - 2f)/f^2 \\ &= -\hat{f}'(\hat{f} - f)^2/(\hat{f}f^2).\end{aligned}$$

Hence we have to prove that

$$P\{|J_n| \geq n^{-1/14}\} \to 0, \qquad n \to \infty,$$

where

$$J_n = n^{-1/2} \sum_{i=1}^{n} \frac{-\hat{f}'_{h,i}(X_i)(\hat{f}_{h,i}(X_i) - f(X_i))^2}{\hat{f}_{h,i}(X_i) f^2(X_i)} Y_i.$$

Now

$$\begin{aligned}P\{|J_n| \geq n^{-1/14}\} \leq{}& P\{\mathscr{A}\} + P\{\mathscr{B}\} \\ &+ P\{\{|J_n| \geq n^{-1/14}\} \cap \overline{\mathscr{A}} \cap \overline{\mathscr{B}}\}.\end{aligned}$$

It follows from Lemmas 1.4 and 1.6 that the first and the second terms in the right side of this inequality tend to zero. Therefore, it suffices to prove that the third term also vanishes.

Define the slices $U_r = \{x : D/2^r \leq f(x) \leq D/2^{r-1}\}$. Then

$$J_n = n^{-1/2} \sum_{r=0}^{\infty} \sum_{\{i:X_i \in U_r\}} \left[\frac{-\hat{f}'_{h,i}(X_i)(\hat{f}_{h,i}(X_i) - f(X_i))^2}{\hat{f}_{h,i}(X_i) f^2(X_i)} Y_i\right].$$

On $\overline{\mathscr{B}}$ we have

$$\hat{f}_{h,i}(X_i) > f(X_i) - \sqrt{\eta \xi(X_i) \log n/(nh)}, \qquad i = 1, \ldots, n.$$

If $X_i \in U_r$, then

$$\hat{f}_{h,i}(X_i) \geq D/2^r - \sqrt{\eta \log n/(nh)} \times \sqrt{C_9(D/2^{r-1} + h)}, \tag{A.2.8}$$

since

$$\xi(X_i) \leq \int K^2(u)\, du \left[\frac{D}{2^{r-1}} + LDh\right]. \tag{A.2.9}$$

Define

$$r^* = \max\{r = 1, 2, \ldots : (D/2^r) \geq C_{10}[(\log n)/nh],$$

where $C_{10} = 10C_9\eta$. It can be easily seen that

$$\hat{f}_{h,i}(X_i) > D/2^{r+1}, \qquad X_i \in U_r,\ r \leq r^*,$$

if n is large enough and $\overline{\mathscr{B}}$ holds.

Note that

$$r^* \leq \log_2 \left[\frac{D}{C_{10}} \frac{nh}{\log n}\right]. \tag{A.2.10}$$

For $X_i \in U_r$, $r > r^*$, by definition of r^*,

$$f(X_i) \leq D/2^{r^*-1} < 4C_{10}[(\log n)/nh]. \tag{A.2.11}$$

Set $C_7 = 5C_{10}$. Then (A.2.11) is impossible on $\overline{\mathscr{A}}$. Therefore, the sets $\{i : X_i \in U_r\}$, $r > r^*$, are empty. Hence we have to bound

$$P\left\{\left\{n^{-1/2} \sum_{r=0}^{r^*} \sum_{\{i:X_i \in U_r\}} \left|\frac{\hat{f}'_{h,i}(X_i)(\hat{f}_{h,i}(X_i) - f(X_i))^2}{\hat{f}_{h,i}(X_i) f^2(X_i)} Y_i\right| \geq n^{-1/14}\right\} \cap \overline{\mathscr{A}} \cap \overline{\mathscr{B}}\right\}.$$

Since we are on the set $\overline{\mathscr{B}}$, this probability is smaller than

$$P\left\{\left\{n^{-1/2} \sum_{r=0}^{r^*} \sum_{\{i:X_i \in U_r\}} \frac{\eta \log n}{nh} \frac{|Y_i|\, |\hat{f}'_{h,i}(X_i)| \xi(X_i)}{\hat{f}_{h,i}(X_i) f^2(X_i)} \geq n^{-1/14}\right\} \cap \overline{\mathscr{A}} \cap \overline{\mathscr{B}}\right\}.$$

Substitute (A.2.11) into the preceding expressions and use Lemma 1.5. Then

$$P\left\{\left\{n^{-1/2} \sum_{r=0}^{r^*} \sum_{\{i:X_i \in U_r\}} \frac{\eta \log n}{nh} \frac{|Y_i|\, |\hat{f}'_{h,i}(X_i)| 2^{r+1} \xi(X_i)}{df^2(X_i)} \geq n^{-1/14}\right\} \cap \overline{\mathscr{A}} \cap \overline{\mathscr{B}}\right\}$$

$$\leq P\left\{\frac{C_{11}\eta \log n}{n^{3/2}h} \sum_{r=0}^{r^*} \sum_{\{i:X_i \in U_r\}} 2^{r+1} \left(\frac{\xi(X_i)|Y_i|}{f^2(X_i)}\right) \geq n^{-1/14}\right\} + o(1).$$

Now it remains to prove that

$$n^{1/14} \frac{\log n}{n^{3/2}h} \sum_{r=0}^{r^*} E\left[\sum_{\{i:X_i \in U_r\}} 2^{r+1} \frac{\xi(X_i)|Y_i|}{f^2(X_i)}\right] \tag{A.2.12}$$

tends to zero. Using Assumption A6 one obtains

$$\begin{aligned}E\left(\sum_{\{i:X_i \in U_r\}} \frac{\xi(X_i)|Y_i|}{f^2(X_i)}\right) &= n \int_a^b I\{x \in U_r\} \frac{\xi(x)E(|Y|\ |X = x)}{f(x)}\, dx \\ &\leq C_{12}n \int_a^b I\{x \in U_r\}\xi(x)\, dx.\end{aligned} \tag{A.2.13}$$

Together with (A.2.9) this entails that the left side of (A.2.13) is bounded by

$$C_{13}n\left(\frac{d}{2^{r-1}} + LDh\right) \int_a^b I\{x \in U_r\}\, dx \leq C_{14}n(2^{-3r/2} + h2^{-r/2}).$$

Hence (A.2.12) does not exceed

$$C_{15}n^{1/14} \frac{\log n}{n^{1/2}h} \sum_{r=0}^{r^*} (2^{-r/2} + 2^{r/2}h) \leq C_{16}n^{1/14} \frac{\log n}{n^{1/2}} 2^{r^*/2},$$

which tends to 0, by the definition of r^*.

A.3. Auxiliary Results

Define $K'_h(u) = h^{-2}K'(u/h)$, $d_K = \int u^2K(u)\, du$, $c_K = \int K^2(u)\, du$, and $c_{K'} = \int (K'(u))^2\, du$. Also define

$$A_1 = \int (K_h(x - u) - 2f(x))f(u)\, du,$$

$$A_2 = \int K'_h(x - u)f(u)\, du,$$

$$A_3 = \int (K_h(x - u) - 2f(x))^2f(u)\, du,$$

$$A_4 = \int (K'_h(x - u))^2f(u)\, du,$$

and

$$A_5 = \int K'_h(x - u)(K_h(x - u) - 2f(x))f(u)\, du.$$

We have

$$\begin{aligned}A_1 &= \int K(t)f(x - th)\, dt - 2f(x) \\ &= -f(x) + f''(x)\frac{h^2}{2} d_K + o(h^2),\end{aligned} \tag{A.3.1}$$

and, by similar techniques and partial integration,

$$A_2 = f'(x) + f'''(x)\frac{h^2}{2} d_K + o(h^2), \tag{A.3.2}$$

$$A_3 = O(h^{-1}), \tag{A.3.3}$$

$$A_4 = h^{-3}\left[f(x)\int (K'(t))^2\,dt + o(1)\right], \qquad \text{(A.3.4)}$$

and

$$A_5 = O(h^{-1}). \qquad \text{(A.3.5)}$$

In the multidimensional case the asymptotic expressions for A_1–A_5 get more complicated. Recall the multidimensional kernel density estimator, as defined in Section 2. The second term in (A.3.1) would thus change to $(d_K \sum_{j=1}^{d} \partial f^2/\partial x_j^2)\, h^2/2$. If kernels of order p are used, as in Härdle and Stoker (1989), this changes to a multiple of h^p the constant depending on pth partial derivatives of f. The term A_2 changes in a similar fashion, since it is the expected value of a kernel estimator for the gradient of f.

[*Received March 1989. Revised January 1991.*]

REFERENCES

Carroll, R., and Härdle, W. (1989), "A Note on Second-Order Effects in a Semiparametric Context," *Statistics,* 20, 179–186.

Department of Employment, Statistics Division (1968–1983), "Family Expenditure Survey, Annual Base Tapes," London: Her Majesty's Stationery Office. The data used in this article were made available by the ESRC Data Archive at the University of Essex.

Friedman, J., and Stuetzle, W. (1981), "Project Pursuit Regression," *Journal of the American Statistical Association,* 76, 817–823.

Gasser, T., Müller, H. G., and Mammitzsch, V. (1985), "Kernels for Nonparametric Curve Estimation," *Journal of the Royal Statistical Society,* Ser. B, 47, 238–252.

Hall, P., and Marron, J. S. (1987), "Estimation of Integrated Squared Density Derivatives," *Statistics and Probability Letters,* 6, 109–115.

Härdle, W. (1990), *Applied Nonparametric Regression* (Econometric Society Monograph Series 19), Cambridge, U.K.: Cambridge University Press.

Härdle, W., and Stoker, T. (1989), "Investigating Smooth Multiple Regression by the Method of Average Derivatives," *Journal of the American Statistical Association,* 84, 986–995.

Härdle, W., Marron, J. S., and Wand, M. (1989), "Bandwidth Choice for Density Derivatives," *Journal of the Royal Statistical Society,* Ser. B, 52, 223–232.

Levit, B. Y. (1974), "On Optimality of Some Statistical Estimates," in *Proceedings of the Prague Symposium on Asymptotic Statistics, II,* ed. I. Hajek, 215–238.

Hicks, J. R. (1956), *A Revision of Demand Theory,* Oxford, U.K.: Clarendon Press.

Hildenbrand, W. (1989), "Facts and Ideas in Micoeconomic Theory," *European Economic Review,* 33, 251–276.

Hsieh, D. A., and Manski, C. F. (1987), "Monte Carlo Evidence on Adaptive Maximum Likelihood Estimation," *The Annals of Statistics,* 15, 541–551.

Mammitzsch, V. (1989), "Asymptotically Optimal Kernels for Average Derivative Estimation," talk given at the American Mathematical Society conference, Davis, CA.

Manski, C. F., and McFadden, D. (1981), *Structural Analysis of Discrete Data with Econometric Applications,* Cambridge, MA: MIT Press.

Powell, J. L. (1986), "Symmetrically Trimmed Least Squares Estimation for Tobit Models," *Econometrica,* 54, 1435–1460.

Samarov, A. (1990), "On Asymptotic Efficiency of Average Derivative Estimates," unpublished manuscript, Massachusetts Institute of Technology, Dept. of Mathematics.

Serfling, R. J. (1980), *Approximation theorems of mathematical statistics*. New York, Wiley.

Silverman, B. W. (1986), *Dênsity Estimation for Statistics and Data Analysis,* London: Chapman and Hall.

Stoker, T. M. (1989), "Equivalence of Direct and Indirect Estimators of Average Derivatives," manuscript, Massachusetts Institute of Technology, Sloan School of Management.

Stone, C. J. (1982), "Optimal Global Rates of Convergence for Nonparametric Regression," *The Annals of Statistics,* 10, 1040–1053.

Computational Statistics (1992) 7:97-128

Smoothing by Weighted Averaging of Rounded Points

W. K. Härdle[1] and D. W. Scott[2]

[1]C.O.R.E., Université Catholique de Louvain, 34 Voie du Roman Pays, B-1348
Louvain-la-Neuve, Belgium
[2]Department of Statistics, Rice University, Houston, Texas 77251-1892, USA

Key Words: Smoothing, Nonparametric Density Estimation, Nonparametric Regression, Binning, WARPing, Generalized Additive models.

Abstract

Nonparametric smoothing techniques are generating much interest not only among theoretical statisticians but among applied workers in biostatistics, economics, and engineering. The benefits of this more flexible method come at the cost of greater computation. In higher dimensions, the computational burden can also be enormous when resampling methods for confidence intervals are used. One idea for reduction of computational cost is to do a data compression. In the case of multivariate density estimation, for example, the averaged shifted histogram is such an algorithm with significantly reduced computational effort. The ideas of the averaged shifted histogram algorithm can be extended to other nonparametric estimation problems such as regression and also to algorithms for additive modeling of high dimensional surfaces. In this paper the common framework for the so-called *Weighted Averaging of Rounded Points (WARPing)* is presented in these situations and examples are given with real data from LANDSAT observations and from a study of binomial response variables. The reduction of computational cost is discussed versus the loss in statistical efficiency.

1. The need for computationally efficient smoothing algorithms

Smoothing of data is a method of re-expressing the data points in a form that is easier to understand than the raw point cloud itself. In a *regression smoothing* problem a $(d+1)$-dimensional point cloud is observed consisting of observations $\{Y_i\}_{i=1}^n$ at values $\{X_i\}_{i=1}^n$ of the predictor variable, where $X_i \in \mathbb{R}^d$. It is assumed that with observation errors $\{\varepsilon_i\}_{i=1}^n$,

$$Y_i = m(X_i) + \varepsilon_i. \tag{1.1}$$

0723-712X/92/2/97-128 $ 2.50 © 1992 Physica-Verlag, Heidelberg

The goal of regression smoothing is to re-express the point cloud by approximating the function m. In a *density smoothing* problem a d-dimensional point cloud is observed and one is interested to understand the structure of the data by estimating the unknown density $f(x)$ from observations $\{X_i\}_{i=1}^n$.

Nonparametric techniques for regression and density smoothing are known for their flexibility and their ability to detect structures deviating from a postulated parametric model. For an overview of nonparametric techniques, see the recent monographs of Eubank (1988), Härdle (1990), Müller (1988), Silverman (1986), and Wahba (1990). A drawback, however, of nonparametric smoothing techniques is that they face rapidly increasing computational complexity as the dimension increases beyond one dimension, or when the sample size is bigger than some threshold, say several thousand points in one dimension. This computational burden is especially handicapping when resampling techniques are used to determine the smoothing parameter or to construct error bars.

In this paper nonparametric smoothing of data points by *Weighted Averaging of Rounded Points* will be discussed. This approach is based on discretizing the data which allows efficient computing in low and high dimensions. The approach is entirely natural for raw data automatically collected in rounded form. For example, each datum collected by the LANDSAT remote sensing satellite is stored in just eight bits; see section 4. We consider smoothing in "low and high" dimensions. A dimension will be called "low" if it is less than three or four since the smoothing operation can be interactively visualized on modern computing systems. At present, it seems reasonable to call dimensions beyond 10 "high" since our experience indicates that one rarely has enough observations to explore spaces of that dimension. The dimensions from 4 – 10 are the subject of intense study, and it is not clear exactly where "high" begins; see Scott and Wand (1990).

Smoothing in high dimensional spaces results in an inherent lack of sufficient statistical precision. Huber (1985) has discussed the so-called "curse of dimensionality." Practically speaking, observations in higher-dimensional spaces are very sparse. This sparseness limits the application of smoothing methods since they are all basically constructed by local averages over sample points. To illustrate this surprising sparseness, we recall the small example by Friedman and Stuetzle (1981) who considered a uniform distribution on a 10-dimensional unit cube. How sparse are the points in this cube?

If the dimensions of the neighborhood are chosen to cover 10 percent of the range of each coordinate, then it will (on the average) contain only $(.1)^{10}$ *of the sample, and thus will nearly always be empty. If, on the other hand, one adjusts the neighborhood to contain 10 percent of the sample, it will cover (on the average)* $(.1)^{1/10} \simeq 80$ *percent of the range of each coordinate. This problem of sparsity basically limits the success of direct d-dimensional local averaging.*

Should we give up now that we know that smoothing in high dimensions is almost impossible unless we have billions of data points that we can't analyze effectively? No, we could still try to pursue the goal to extract the most interesting low dimensional feature. One very attractive class of such models that greatly improve statistical efficiency are the *additive* ones (Stone, 1985). A simple additive model is, for example, one where the regression function m is decomposed into a sum of simpler one-dimensional functions, i.e.,

$$m(\underset{\sim}{x}) = \sum_{j=1}^{d} g_j(x_j),$$

see Hastie and Tibshirani (1987). A more general class of additive models are those that are based on projecting the $\underset{\sim}{x} \in \mathbb{R}^d$ onto the real line, i.e.,

$$m(\underset{\sim}{x}) = \sum_{j=1}^{N} g_j(\beta^T \underset{\sim}{x}),$$

see Friedman and Stuetzle (1981). A similar model has been recently studied by Duan and Li (1990). They use Sliced Inverse Regression (SIR), a method of discreting the response variable in order to find interesting projections.

The assumption of additivity significantly improves statistical convergence properties but does not eliminate the complexity of computation. The additive model algorithms proposed in the literature are heavily based on efficient smoothing algorithms since most of them are iterative or optimize the smoothing parameter. Friedman's (1984) *supersmoother*, for example, used in the alternating conditional expectation (ACE) algorithm (Breiman and Friedman, 1985) needs three pilot symmetrized k–NN "smooths" (the tweeter, woofer and midrange).

Current algorithmic techniques in typical smoothing scenarios in low or high dimensions involve stepping through some of the following operations.

SMOOTHING. A nonparametric "smooth" is first computed for a particular smoothing parameter. This smooth may be a one–dimensional building block for a higher dimensional additive model or the smooth may be a simple possibly multivariate exploratory computation.

OPTIMIZATION. Often a functional of "interestingness" is optimized over a set of parameters. Huber (1985, Chapter III) presents several projection indices for projection pursuit. In the specific projection pursuit technique the optimization is performed simultaneously over linear combinations of the (transformed) data and a nonparametric regression smooth.

ITERATION. The process of extracting interesting features often involves changing the interestingness functional, or running an optimization technique from residuals or transformed statistics in order to find good approximations to additive models. For instance, the "local scoring" algorithm of Tibshirani and Hastie (1987, Section 2) to fit generalized additive models (GAM) by "backfitting" is highly iterative. It loops over the elements of the predictor variable and uses the Aitken-weighted least squares technique in each element.

CALIBRATION. A necessary operation in nonparametric smoothing is to calibrate the smooth by finding a good smoothing parameter. There are several concepts for defining an optimal smoothing parameter but all have the goal of construction of narrow confidence bands for comparison with alternative (possibly parametric) models. Cross-validation or related resampling methods have been used both for regression and density smoothing, see Härdle, Hall and Marron (1988), Silverman (1986), or Scott and Terrell (1987). A nonasymptotic way to set up confidence bands is to use some sort of resampling scheme. Härdle and Bowman (1988) and Härdle and Marron (1989) use the bootstrap from estimated residuals to calibrate the smoothing parameter and construct variability bands.

Most of these operations involve computations that are typically a function of the squared sample size, $O(n^2)$, if one uses straightforward implementations of the formulae. Improved computation may be obtained either by reducing the number of arithmetic operations or by using novel approaches that eliminate one or more of the above four basic operations. For example, equation (2.11) below illustrates a method that avoids an optimization step. Our plan is to demonstrate how the Weighted Averaging of Rounded Points (WARPing) method achieves the goal of improved computational efficiency. We have intentionally chosen this abbreviation because of the connotation of speed, although it might lead the reader to the conclusion that we distort the data. Later, we will show that the asymptotics of this method are not unreasonable.

WARPing consists of first rounding the data points, i.e., to reduce the complexity in a step that is linear in the sample size, $O(n)$, and then performing a weighted average of these rounded points. A significant benefit of this prebinning step is that the resulting array of estimates, which falls on a uniform grid, is precisely in the form required for surface visualization algorithms such as marching cubes (Lorensen and Cline, 1987). Thus the decision to prebin is directly related to the final visualization desired (Scott and Hall, 1989) and enhances the speed of interpretation and analysis of data.

Other authors have used prebinning for one-dimensional smoothing problems. Silverman (1982) has given an algorithm based upon the Fast Fourier Transform (FFT) for computing a Gaussian kernel density estimate by rounding data to a mesh with 2^M points. At least three examples of prebinning from regression exist. Cleveland's (1979) LOWESS algorithm bins the x data in order to take advantage of a clever updating scheme for computing a sliding series of local linear regressions. O'Sullivan's unpublished BART smoothing spline implementation uses the same trick. Härdle (1987) prebins the x and y data to apply the FFT for one step M-smoothing.

Our purpose in this paper is to illustrate the effectiveness of prebinning. The usefulness of multivariate prebinning is not widely acknowledged. An alternative approach is to use any of the efficient algorithms that exist for computing k–th nearest neighbor (k–NN) estimate, which is a particular example of an adaptive kernel estimate. We hope to expand the use of the simple but significant trick of prebinning, which Marron has suggested "could make nonparametric methods accessible to a PC."

2. Weighted Averaging of Rounded Points (WARPing)

2.1 Density and regression smoothing

It is easiest to demonstrate our approach in the univariate density smoothing context. The simplest density estimator for data $\{X_i\}_{i=1}^n$ is the histogram with bin width h,

$$HG_h(x;x_0) = \frac{1}{nh}\sum_{i=1}^{n} I(X_i \in B(x;x_0,h)) \tag{2.1}$$

where $B(x;x_0,h)$ denotes the unique bin containing x of the form $[x_0 + kh, x_0 + (k+1)h)$ for some integer k. Note that the histogram is a function of the origin x_0. It is well known that different choices of x_0 may result in quite different shapes of density estimates. Figure 1 shows several histograms of the Buffalo snowfall data (Parzen, 1979). Depending on the origin, the histogram $HG_h(x;x_0)$ has one, two, or three modes and looks skewed to either the right or left. A quite drastic difference can be observed between the estimates with bin origin 2.5 (secondary mode at left) and bin origin 10.0 (secondary mode at right).

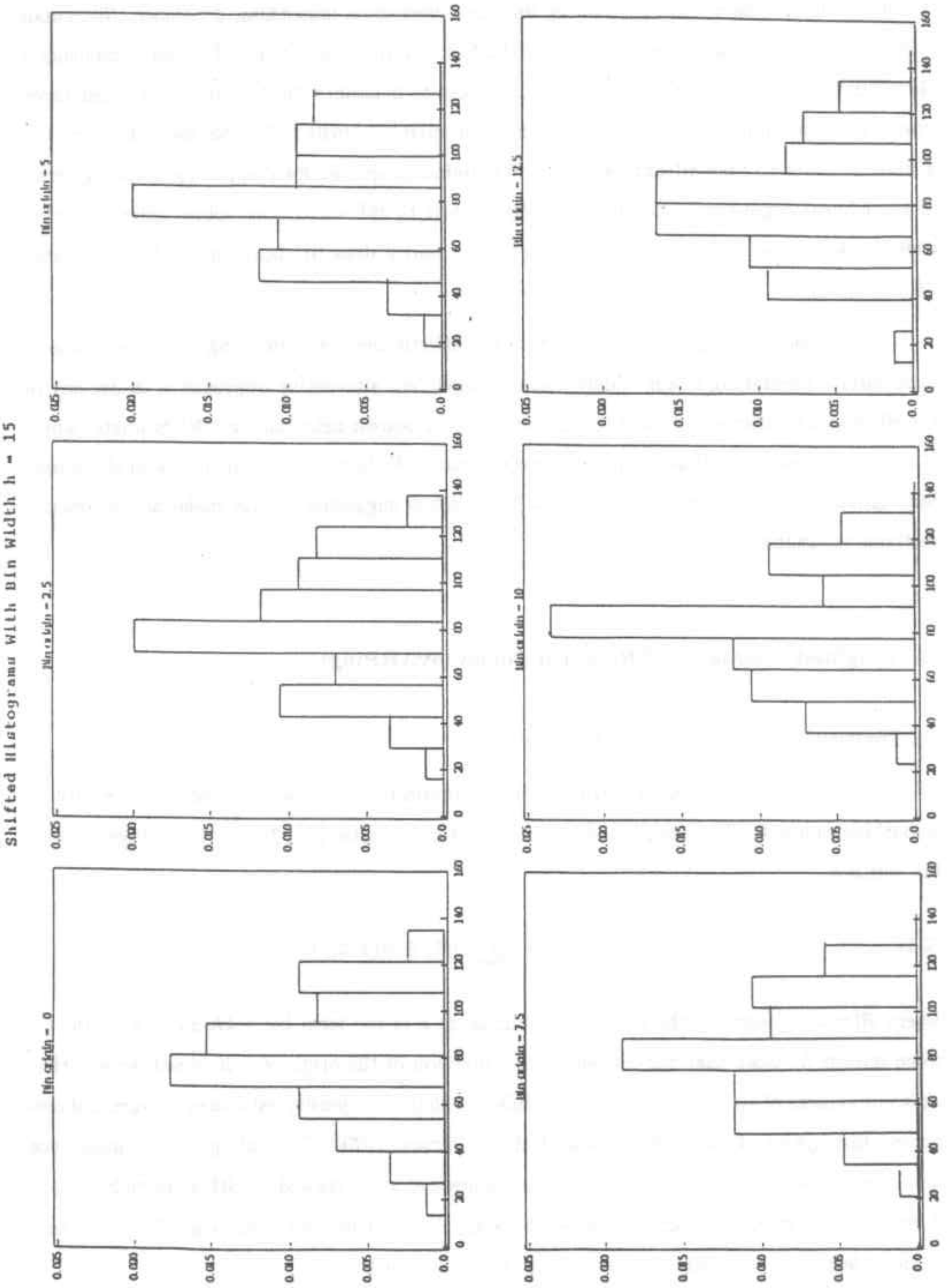


Figure 1. Six histograms with bin width h=13.5 and bin origins x_0=0,2.5,5,7.5,10,12.5 for the Buffalo snowfall data.

A natural way of eliminating this nuisance parameter is to construct an average of histograms, each with bin width h, over a collection of "origin" choices (Scott, 1985a). In particular, consider the set $\{x_{0,\ell} = \ell h/M, \ell = 0, ..., M-1\}$. Then our estimate is simply

$$\hat{f}_M(x) = \frac{1}{M} \sum_{\ell=0}^{M-1} HG_h(x, x_{0,\ell}). \tag{2.2}$$

It is straightforward to see that formula (2.2) may be written as

$$\begin{aligned} \hat{f}_M(x) &= \frac{1}{M} \sum_{\ell=0}^{M-1} \frac{1}{nh} \#\{i : X_i \in B(x; x_{0,\ell},\ h)\} \\ &= \frac{1}{nh} \sum_{\ell=1-M}^{M-1} \frac{(M - |\ell|)}{M} \#\{i : X_i \in B(x; x_{0,\ell},\ h/M)\} \end{aligned} \tag{2.3}$$

Another way to view formula (2.3) is to say we are weighting points aggregated or rounded into the smaller bins of width $\delta = h/M$. In formula (2.3), observe that the data enter only in the form of frequencies in the bin $B(\cdot; x_{0,k}, \delta)$.

The averaged shifted histogram (ASH) in (2.3) is a special case of WARPing. As $M \to \infty$, formula (2.3) can be written in the familiar form of a kernel estimate

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{x - x_i}{h}), \tag{2.4}$$

where $K(u) = (1 - |u|)I(|u| \leq 1)$.

The triangular weights $(1 - |\ell|/M) \equiv w_M(\ell)$ in formula (2.3) can be generalized in an obvious way to other kernel weights. A simple way to generate weight sequences is to discretize a continuous weight function that is defined on the interval $[-1, 1]$. For example, $K(u) = \frac{15}{16}(1 - u^2)^2 I(|u| \leq 1)$, the so-called biweight or quartic kernel, corresponds to the weights $w_M(\ell) = \frac{15}{16}(1 - \ell^2/M^2)^2$ for $|\ell| < M$. In practice we would normalize this weight sequence by a constant $C_{M,K}$ so that $M^{-1} \sum_\ell w_M(\ell) = 1$. This technique has been applied to the Buffalo snowfall data (Parzen, 1979). The sequence of WARP density estimates as the number of averages increases is shown in Figure 2.

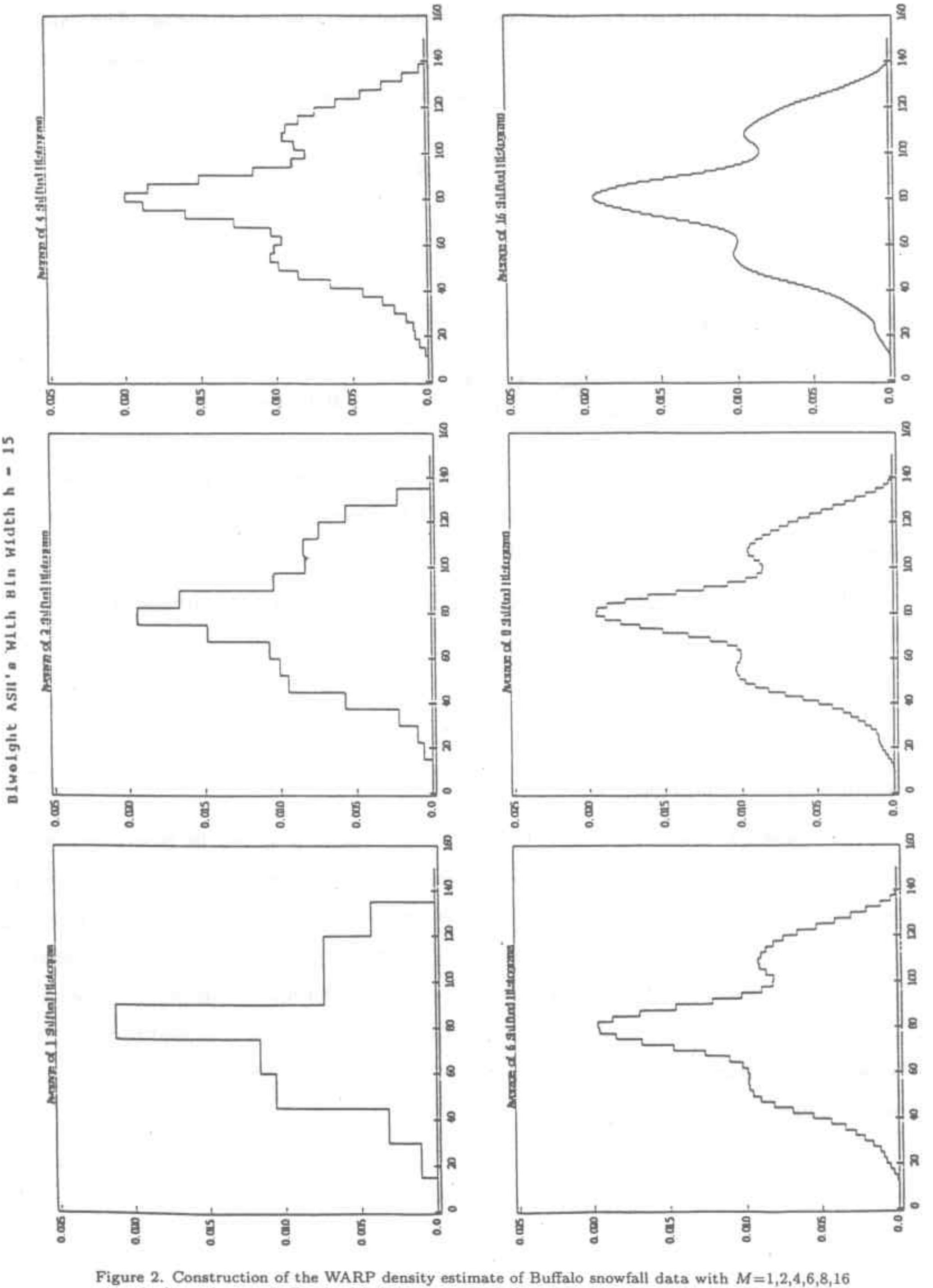


Figure 2. Construction of the WARP density estimate of Buffalo snowfall data with M=1,2,4,6,8,16 using the biweight kernel. Bandwidth h fixed.

Using this generalization we can rewrite formula (2.3) in the general form

$$(WARP) \qquad \hat{g}(x) = \frac{1}{M} \sum_{|\ell|<M} w_M(\ell) RP_{\iota(x)+\ell} \, ,$$

where $\iota(x)$ is the bin in which x falls and where, in the above case of density smoothing, RP_j is the frequency of rounded points ($\equiv RP$) in the j-th bin. We will show that the above notion is not restricted to estimating the density itself, but also related functionals such as a regression smooth. By allowing the weight sequence to take on negative values, we can estimate other quantities such as the derivative of the density. For the biweight kernel, $K'(u) = \frac{15}{4}u(u^2-1)I(|u| \leq 1)$ so that the effective WARP weight $w_M(\ell)$ for derivatives is $\frac{15}{4}\ell/M(\ell^2/M^2 - 1)$. Further generalizations to multivariate data are accomplished by rounding along each coordinate axis. For example, the multivariate version of the ASH in equation (2.3) converges to a multivariate product triangle kernel estimate as $M \to \infty$.

Let us give an overview and summary of this technique for this introductory example of density smoothing.

- One has a small bin width δ defining a sequence of bins

  $$B_j = [(j - 1/2)\delta, (j + 1/2)\delta], \quad j \in \mathbb{Z}$$

  and the *index function* $\iota(x) = j \Leftrightarrow x \in B_j$.
- The approximation parameter $M = h/\delta$ determines how many adjacent bins enter into the averaging process.
- The WARPed density estimate is

  $$\hat{f}_M(x) = \frac{C_{M,K}}{nh} \sum_{i=1}^{n} K\left(\frac{\iota(x) - \iota(X_i)}{h}\right),$$

  where the factor $C_{M,K}$ is introduced to guarantee that $\hat{f}_M$ integrates to one. The constant $C_{M,K}$ is given by

  $$C_{M,K} = M / \sum_{\ell=1-M}^{M-1} K\left(\frac{\ell}{M}\right).$$

  The discretized quartic kernel

  $$K(u) = \frac{15}{16}(1-u^2)^2 I(|u| \leq 1),$$

  for example, has the effective weight

  $$C_{M,K}\, K\left(\frac{\ell}{M}\right) = \frac{15M^4}{16M^4-1}\left(1 - \left(\frac{\ell}{M}\right)^2\right)^2 .$$

Computational Statistics, 1, 97-128

For regression smoothing, we wish to approximate $m(x)$ by a weighted average of the responses for which the predictor values are in a neighborhood of x (Stone 1977). Using the WARP approach this amounts first to rounding the predictor values to the closest bin center and computing the average response in each bin; then second, appropriately weighting these bin averages in a neighborhood of x to estimate $m(x)$. For the regression case we consider two different estimators, a step function approximation

$$\hat{m}_M^{(S)}(x) = \frac{\sum_{i=1}^{n} K\left(\frac{\iota(x)-\iota(X_i)}{M}\right) Y_i}{\sum_{i=1}^{n} K\left(\frac{\iota(x)-\iota(X_i)}{M}\right)} \tag{2.5}$$

and a polygon through the midpoints of the steps of $\hat{m}_M^{(S)}(x)$. For the latter, define $d_x = \iota(x) - x/\delta$ the distance between x and the bias center $\iota(x)\delta$. The polygon approximation is defined through

$$\begin{aligned} \hat{m}_M^{(P)}(x) &= (1-|d_x|)\ \hat{m}_M^{(S)}(\iota(x)\delta) \\ &\quad + |d_x|\big(I(x \in [(\iota(x)-1/2)\delta), \iota(x)\delta))\ \hat{m}_M^{(S)}((\iota(x)-1)\delta) \\ &\quad + I(x \in [\iota(x)\delta, (\iota(x)+1/2)\delta))\ \hat{m}_M^{(S)}((\iota(x)+1)\delta)\big). \end{aligned} \tag{2.6}$$

To give some insight note that by construction

$$K\left(\frac{\iota(x)-\iota(X_i)}{h}\right) = \sum_{\ell \in \mathbb{Z}} K(\ell/M) \times I(X_i \in B_{\iota(x)+\ell})$$

and therefore

$$\hat{m}_M^{(S)}(x) = \frac{\sum_{\ell=1-M}^{M-1} K(\ell/M) Y_{\bullet,\iota(x)+\ell}}{\sum_{\ell=1-M}^{M-1} K(\ell/M) n_{\iota(x)+\ell}}, \tag{2.7}$$

where

$$Y_{\bullet,j} = \sum_{i=1}^{n} Y_i\ I(X_i \in B_j),$$

$$n_j = \sum_{i=1}^{n} I(X_i \in B_j).$$

The rounded points RP_j are here the sum of the response variables and the frequency of the predictor variable in bin B_j. In principle, the Y_i values can be binned, but there is no computational advantage in doing so.

(1992) Härdle, W. and Scott, D.W. Smoothing in Low and High Dimensions by Weighted Averaging Using Rounded Points.

Figure 3 shows $\hat{m}_M^{(S)}(x)$, $\hat{m}_M^{(P)}(x)$, and the Nadaraya–Watson kernel estimator

$$\hat{m}_h(x) = \frac{\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right) Y_i}{\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)}.$$

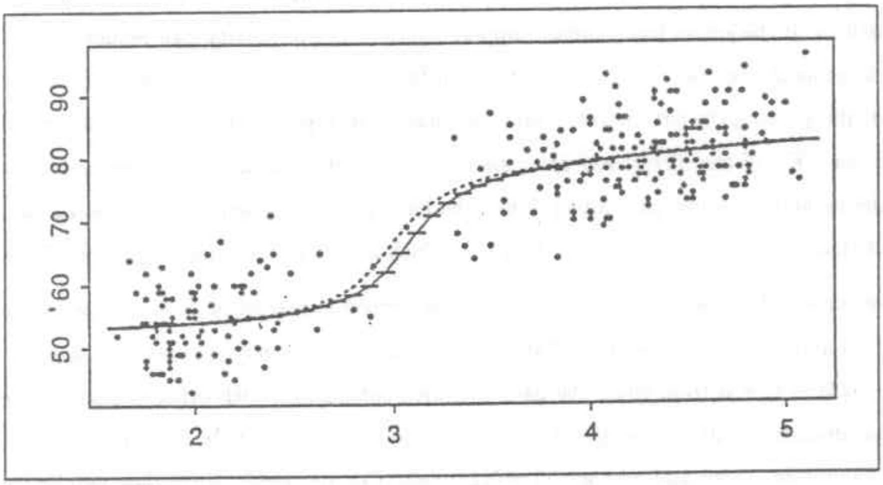


Figure 3. WARPing approximations (solid lines) and the Nadaraya–Watson kernel estimator (dashed line) for the Geyser data-set. Bandwidth h=0.8, M=15.

Qualitatively, the general WARP formula says that we first discretize the data and then smooth; compare equations (2.5) and (2.7), for example. Thus the computational advantage of the WARP method results from the partial decoupling of the sample size and the smoothing computation. The first step of binning is a computational burden of order $O(n)$ that produces a data compression into $1/\delta$ bin counts on a mesh that can be easily maintained. The smoothing is then applied to this smaller bin object taking advantage of the regular spacing of bins which in particular pays off if we use a kernel with compact support. Such a kernel allows fast computation since we know from the regularly spaced bin points where the weighting scheme gives weight zero outside the support of the kernel. Alternative degrees of smoothing can be tried without having to rebin the data.

2.2 Comparison of computational costs

To have some insight into the computational advantage of the WARP method, consider a comparison of three possible procedures for estimating a density. Assume that the density is to be computed at δ^{-1} equidistant points. The brute force kernel method amounts to averaging the rescaled kernel functions $K((x - X_i)/h)$ over the whole sample at δ^{-1} points resulting in $\delta^{-1}n$ operations. If the kernel has compact support and $d = 1$, a presorting can reduce this to $2nh/\delta$ but, in general, the cost is n/δ. The method of Silverman (1982) and of Härdle (1987) consists of first discretizing the data into δ^{-1} bins (assumed to be a power of 2) and then using the FFT. After the FFT of the discretized data, smoothing is performed in the frequency domain (δ^{-1} multiplications for a low-pass filter). Then the back FFT is applied. The discretization cost is n and that of the two FFT's is $2\delta^{-1} log_2 \delta^{-1}$, which results in the number given in Table 1.

As explained above, the WARP method is performed in two steps: first n steps to discretize the data and, then, in a window of $2M - 1$ the rounded points $RP_{\iota(x)+\ell}$ are averaged, resulting in $\delta^{-1}(2M - 1) + n$ total cost. We have excluded comparison with other smoothing methods such as discrete maximum penalized likelihood estimators (DMPLE or DiMPLE) (Scott, Tapia, and Thompson, 1980) and orthogonal polynomials (Cencov, 1962), since they require iterative solution or are equivalent to a particular kernel method, respectively. The k–NN estimator is also relatively efficient, $O(n\,log\,n + \delta^{-1}\,log\,n)$, but is deficient for many purposes since it is often too bumpy and visually rough (Silverman, 1986). Observe that the $O(n\,log\,n)$ work for sorting is replaced by prebinning, which is essentially a sorting operation.

Method	Operation Count
Kernel	$\delta^{-1}n$
WARP	$\delta^{-1}(2M - 1) + n$
FFT	$\delta^{-1}(1 + 2\,log_2\,\delta^{-1}) + n$

Table 1. Operation counts for some univariate nonparametric density estimators.

Of course, the costs of the WARP and FFT are quite similar since they are both based on discretization ideas. For example, when we discretized into 1000 bins the factor of δ^{-1} is 21 for the FFT method and 19 for $M = 10$. Such an M is already rather big, we work mostly with $M = 5$ in practice. The differences become much more drastic if we consider high dimensional smoothing problems or data with many empty bins. While the FFT technique of Silverman (1982) extends to bivariate (x, y) regression smoothing (see Härdle, 1987), it does not extend so easily to higher dimensions. By keeping the pointers to nonempty bins the discretization can be performed only for the nonempty bins so that the computation count for WARPing is in fact smaller, namely, dependent upon the number of nonempty bins, $NB(\delta, n)$.

Method	Operation Count
Kernel	$2(\delta^{-1})^d n$
WARP	$2NB(\delta,n)^d(2M-1)^d + nd$

Table 2. Comparison of operation counts between brute force kernel smoothing in d dimensions and the WARP method.

Of course, the number of nonempty bins $NB(\delta,n)$ is in $\{1,\ldots,n\}$, but in most cases the number is much smaller than n. To see this in the case of $d=1$, denote by $R = X_{(n)} - X_{(1)}$ the span of the data and $[R/\delta]$ the next integer larger than R/δ. Of course, $NB(\delta,n) \leq min(n,[R/\delta])$, but let us look more closely at the random variable $NB(\delta,n)$ for the case of uniform X-variables, i.e., $P(X_i \in B_j) \approx 1/[R/\delta]$. This assumption is rather conservative in the sense that other X-distributions will produce a smaller $NB(\delta,n)$.

Proposition 2.1 *The number of nonempty bins $NB(\delta,n)$ can be represented as a homogeneous Markov chain with transition matrix*

$$P = \frac{1}{[R/\delta]} \begin{pmatrix} 1 & [R/\delta]-1 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & 2 & [R/\delta]-2 & 0 & \cdots & \cdots & 0 \\ & & & \cdots & & & \\ 0 & \cdots & k & [R/\delta]-k & 0 & \cdots & 0 \\ & & & \cdots & & & \\ 0 & \cdots & \cdots & \cdots & 0 & [R/\delta]-1 & 1 \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & [R/\delta]. \end{pmatrix}$$

The starting distribution is given by

$$p_1^T = (1\ 0\ \cdots\ 0).$$

From this proposition we can compute the expectation

$$E(NB(\delta,n)) = p_1^T P^{n-1}(1\ 2\ 3\ \cdots\ [R/\delta])^T.$$

Figure 4 shows the increase of computational cost for the kernel smoother as a function of n. Figure 5 shows the computational cost for WARPing. One clearly sees that the cost stays relatively constant as n increases.

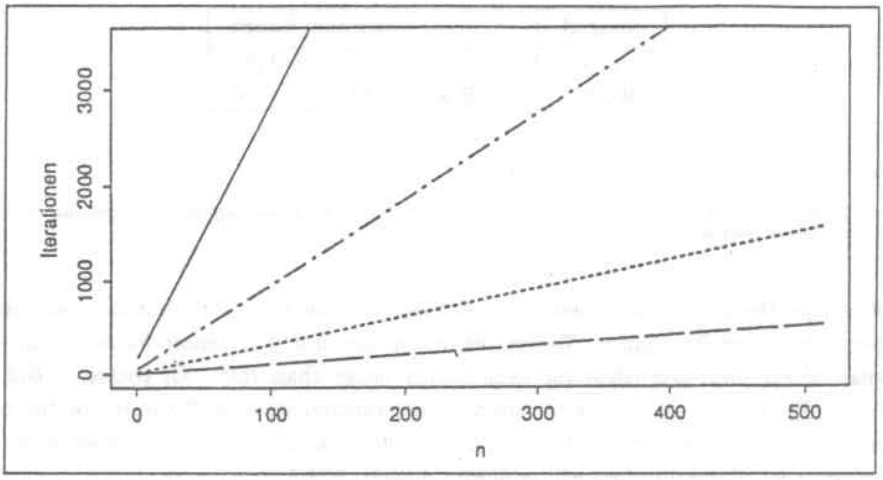


Figure 4. Computational cost of the kernel smoother as a function of n for R=1, h=0.27, M=1 (long dashed line), M=3 (short dashed line), M=9 (short-long dashed line), M=27 (solid line). M is here the number of grid points per h.

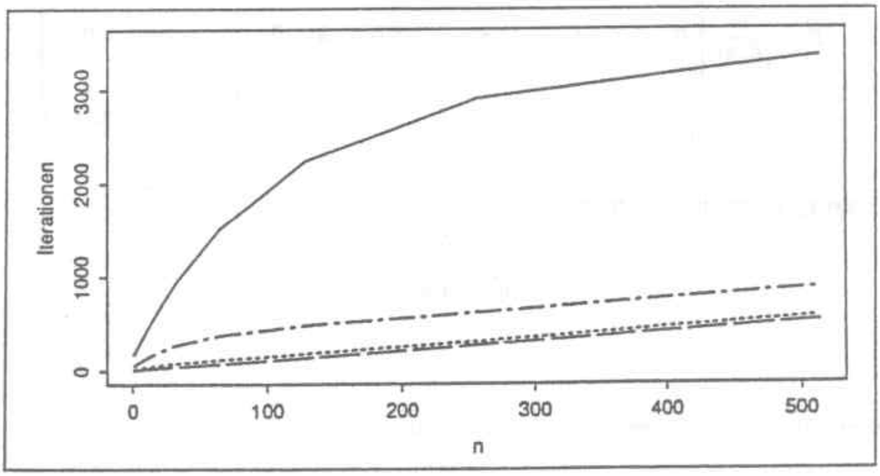


Figure 5. Computational cost for WARPing as a function of n for R=1, h=0.27, M=1 (long dashed line), M=3 (short dashed line), M=9 (short-long dashed line), M=27 (solid line). M is here the number of grid points per h.

The operation counts for density or regression estimation for one-dimensional x's will be slightly high, since empty bins require no operations. For bivariate x's, usually a substantial number of bivariate bins are empty, so that an operation count estimate of δ^{-1} may in fact be only $0.3\delta^{-1}$. For three-dimensional x's, it is possible to proceed exactly as in the previous two cases, accepting a large number of empty bins. However, if n is not too large, then it may be more attractive to keep an unarrayed list of bin counts, specifically, three integers pointing to a trivariate bin plus the integer bin count (with an additional real variable containing the sum of the bin responses in the regression case).

When the dimension $d \geq 4$, it is almost always sensible to store the data in such an intermediate compressed form, since only three-dimensional slices will be computed and displayed. It is possible to maintain $30^4 < 10^6$ bins in memory in the quadravariate case, which may be necessary when using animation techniques (Scott, 1986). Only rarely would it be sensible to estimate at one instant the density on a full mesh in R^d for $d > 5$. In such cases graphical display is not the goal but rather estimates at the sample points may be desired.

It is interesting to mention the symmetrized k–NN smoother that is used in *supersmoother*, Friedman's (1984) variable span smoother. This smoother is only defined for one-dimensional scatterplot smoothing and requires pre-sorting which results in $n(1+log\,n)$ operations altogether. By contrast, the WARP kernel smoother requires $n + \delta^{-1}M$ operations to produce a smooth curve. In fact, Friedman has tried substituting a one-dimensional WARP estimator for a k–NN smoother not for any computational reason but because the smoothness of the WARPing approach greatly improved the iteration in one version of the ACE algorithm.

As we will see in Section 3 below, the discretization step does not affect, in an asymptotic sense, the statistical accuracy of the WARP smoother. The decoupling of the sample size by introducing the rounded points makes the optimization of the nonparametric smoothers much more efficient. This becomes apparent when regression and density smoothers of any kind have to be calibrated by a smoothing parameter. In the WARP–estimator above this smoothing parameter is the number of bins M over which the average is performed. The difference in computational cost becomes drastic if cross-validation or related scores are to be computed for optimization of smoothing parameters. Scott and Schmidt (1988) in an analysis of British income data compared the least-squares and biased cross-validation techniques to optimize the smoothing parameter. Using the direct formulae used up as much as 8 CPU hours on an IBM 3081 mainframe for $n = 7123$. Using the WARPing approach brought the computation of these scores down to 30 seconds on an IBM AT (Processor 80286, 10MHz).

For regression smoothing, ordinary cross-validation requires multiple computation of the smoother at all the sample points of the cross-validation score at several bandwidths, although efficient algorithms exist for certain linear smoothers associated with spline smoothers (Eubank, 1988). The WARP estimator has the advantage that once the discretization step is performed, all further smoothing operations are linear in δ^{-1}.

2.3 Smoothing in high dimensions by means of additive models

In this section we illustrate the use of WARPing in high dimensions by an additive model example. The process of modeling high dimensional point clouds by additive models typically involves iterative approximations in order to minimize some score function. Adopting the additive model (Breiman and Friedman, 1985), construct nonparametric functions ψ^*, $\{\varphi_j^*\}_{j=1}^d$ such that

$$e^2(\psi, \varphi_1, ..., \varphi_d) = \frac{E\{[\psi(Y) - \sum_{j=1}^d \varphi_j(X_j)]^2\}}{E\psi^2(Y)} \tag{2.8}$$

is minimized. Alternating conditional expectations have to be computed based on an efficient one-dimensional scatterplot smoother. In this context, the WARP kernel smoother has to perform $2\delta^{-1}M + n$ operations for this task. Depending on δ and M it can be made highly efficient compared to a kernel or symmetrized k-nearest neighbor smoother (Carroll and Härdle, 1988).

The projection pursuit regression (PPR) algorithm of Friedman and Stuetzle (1981) requires an efficient one-dimensional scatterplot smoother as well. The PPR algorithm searches first for the best pair (β, g) such that with a suitably normalized direction β the residual sum of squares

$$\sum_{i=1}^n (Y_i - g(X_i^T\beta))^2 \tag{2.9}$$

is minimized. This task is done by iterating over several smoothing operations. In a further step residuals are fitted and the same procedure is run for the set of estimated residuals. Then another set of residuals is computed and this iteration is continued until a convergence criterion is met.

The direction β of a one-step PPR can also be estimated without optimization or iteration by proving that β is proportional to a certain expectation:

$$\eta \equiv E_X[\nabla m(X)], \tag{2.10}$$

where $\nabla m(x)$ denotes the gradient vector of partial derivatives. The so-called average derivative, η, can be estimated by

$$\hat{\eta} = \frac{-1}{n} \sum_{i=1}^n Y_i \frac{\hat{f}_h'(X_i)}{\hat{f}_h(X_i)}, \tag{2.11}$$

where $\hat{f}'_h(\cdot)$ denotes the vector of partial derivatives of the kernel estimate with bandwidth h. This can be done effectively with WARPing since averaging has to be performed only over nonempty bins. Once the estimate $\hat{\eta}$ is obtained, the one-dimensional function g is estimated by smoothing the bivariate scatterplot $\{(\hat{\eta}^T X_i, Y_i)\}_{i=1}^n$. The estimator $\hat{\eta}$ from (2.11) has desirable root–n convergence properties, and the additivity of the models enables a rate of convergence for m that is typical for one-dimensional problems; for details, see Härdle and Stoker (1989). The average derivative is also helpful in estimating parameters in the partial linear model

$$Y = \beta^T x + m(z) + \varepsilon, \qquad x \in \mathbb{R}^d, \quad z \in \mathbb{R}^k \ ;$$

see Spiegelman (1976), Rice (1986), and Heckman (1986). The first d components of the average derivative are equal to β.

Generalized additive modeling involves, similarly to ACE, finding functions $\{g_j\}_{j=1}^d$ such that

$$m(\underset{\sim}{x}) = G\left(\sum_{j=1}^{d} g_j(x_j)\right)$$

with a nonlinear (inverse) link function $G(\cdot)$. The functions g_j are determined by a combination of the *backfitting* and the *local scoring algorithm*. For a definition of this algorithm, see Hastie and Tibshirani (1987). We applied this algorithm to a data-set of simulated side-impacts with Opel Kadetts. The response variable was $Y = 0$ or 1 for survival or nonsurvival, respectively, as determined by examining the occupants after impact. The predictor variables for survival were X_1 = AGE, X_2 = VELocity, X_3 = ROSYM, a measure of acceleration at the chest. A draftsman's graphic of this four-dimensional data-set is presented in Figure 6. We will return to these data in Section 4.

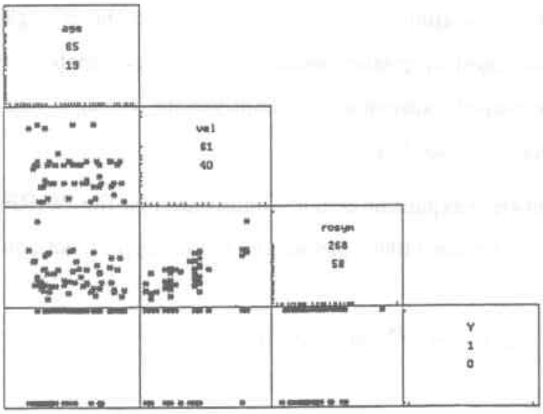


Figure 6. A draftsman's plot of the four side-impact variables. Made with XploRe (1990). X_1= AGE, X_2= VEL, X_3= ROSYM, Y= response.

3. Asymptotics of WARPing

3.1 How precise is WARPing?

Since WARP algorithms are identical to kernel smoothing when the fine bin width $\delta = h/M \to 0$, it is not surprising that the asymptotic bias and variance expressions of WARP and kernel estimators are quite similar. For example, with a univariate WARP density estimator with asymptotic triangular kernel $K(u) = (1-|u|)I(|u| \leq 1)$, the mean integrated squared error is given by

$$MISE \simeq \frac{2}{3nh}(1+\frac{1}{2M^2}) + \frac{1}{12}\delta^2 \int (f')^2 + \frac{1}{144}(1-\frac{2}{M^2}+\frac{3}{5M^4}) \int (f'')^2. \tag{3.1}$$

The second and third terms are bias terms, with the second being a histogram-like bias term. Clearly, the fine bin width δ should be chosen sufficiently small (or, equivalently, M sufficient large since $h = M\delta$ with h conceptually fixed near its optimal value) so that the second term is negligible compared to the third term. In practice, the data are often presented with finite precision so that δ (and hence h) cannot be made arbitrarily small. When δ can be chosen arbitrarily small and assuming a reasonable choice of h is known, then a (δ, h) pair should be specified so that $M \geq 5$ but not so large to eliminate the benefit of binning. Scott (1985a) also showed that the histogram-like bias term could be eliminated by the additional work of constructing a piecewise linear interpolant of the WARP estimates at the bin centers, for which

$$MISE \simeq \frac{2}{3nh} + \frac{1}{144}h^4(1+\frac{1}{M^2}+\frac{9}{20M^4}) \int (f'')^2.$$

This expression is much easier to understand and for relatively small $M \geq 3$, the WARP estimate is both theoretically and graphically virtually identical to the triangle kernel estimate. This particular issue is discussed in greater detail in Scott and Sheather (1985) and Jones (1989). Hjort (1986) has developed expressions for multivariate averaged shifted histograms using a linear blend interpolation procedure.

The mean squared error expansion is more complicated for the WARPed regression smoothers. Let $\sigma^2(x) = \text{var}(\varepsilon \mid x)$. Breuer (1990) proved the following propositions.

Proposition 3.1 *As $h \sim n^{-1/5}$, $M \sim n^\beta$, $\beta \geq 1/5$,*

$$MSE[\hat{m}_M^{(S)}(x)]$$
$$\simeq (nh)^{-1}\frac{\sigma^2(x)}{f(x)} \parallel K_M \parallel_2^2 + \frac{h^2}{M^2}d_x^2(m'(x))^2$$

$$+\frac{h^3}{M}d_x\left(\mu_2(K_M)+\frac{12d_x^2+1}{12M^2}\right)m'(x)\left\{m''(x)+2m'(x)\frac{f'(x)}{f(x)}\right\}$$
$$+\frac{h^4}{4}\left(\mu_2(K_M)+\frac{12d_x^2+1}{12M^2}\right)\left\{m''(x)+2m'(x)\frac{f'(x)}{f(x)}\right\}^2$$

where

$$\|K_M\|_2^2=\frac{C_{M,K}^2}{M}\sum_{\ell}K^2\left(\frac{\ell}{M}\right)$$

and

$$\mu_2(K_M)=\frac{C_{M,K}^2}{M^3}\sum_{\ell}\ell^2K\left(\frac{\ell}{M}\right).$$

Proposition 3.2 *Let $h\sim n^{-1/5}$, $M\sim n^{\beta}$, $\beta>0$,*

$$\begin{aligned}&MSE[\hat{m}_M^{(P)}(x)]\\&\simeq(nh)^{-1}\frac{\sigma^2(x)}{f(x)}\Big(\|K_M\|_2^2-2|d_x|(1-|d_x|)(\|K_M\|_2^2-\gamma(K_M))\Big)\\&\quad+\frac{h^4}{4}\Big[m''(x)\Big(\mu_2(K_M)+\frac{12|d_x|(1-|d_x|)+1}{12M^2}\Big)\\&\quad+2m'(x)\frac{f'(x)}{f(x)}\Big(\mu_2(K_M)+\frac{1}{12M^2}\Big)\Big]^2\end{aligned}$$

where

$$\gamma(K_M)=\frac{C_{M,K}^2}{M}\sum_{\ell=1-M}^{M-1}K\left(\frac{\ell}{M}\right)K\left(\frac{\ell+1}{M}\right).$$

These two propositions are remarkable. They show the basic difference between the step function $\hat{m}_M^{(S)}$ and the continuous polygon approximation $\hat{m}_M^{(P)}$: One sees that if binning is done so that $M\sim n^{1/5}$ the terms depending on M have the same speed as those depending solely on h. Thus if $M\sim n^{1/5}$ the MSE of WARPing is different from that of the kernel smoother $\hat{m}_h$. By contrast the estimator $\hat{m}_M^{(P)}$ will reflect the correct MSE properties of $\hat{m}_h$ provided $M\sim n^{\beta}$, $\beta>0$.

3.2 The price of discretization

An example of the MSE of $\hat{m}_M^{(S)}$ is plotted as a function of x in Figure 7. One sees especially in the right half of the interval the increased bias of $\hat{m}_M^{(S)}$ for the fact of not being directly in a bin center. The effect is, of course, here made very drastic since we are averaging over a very large bandwidth and use only $M=5$ bins for this h.

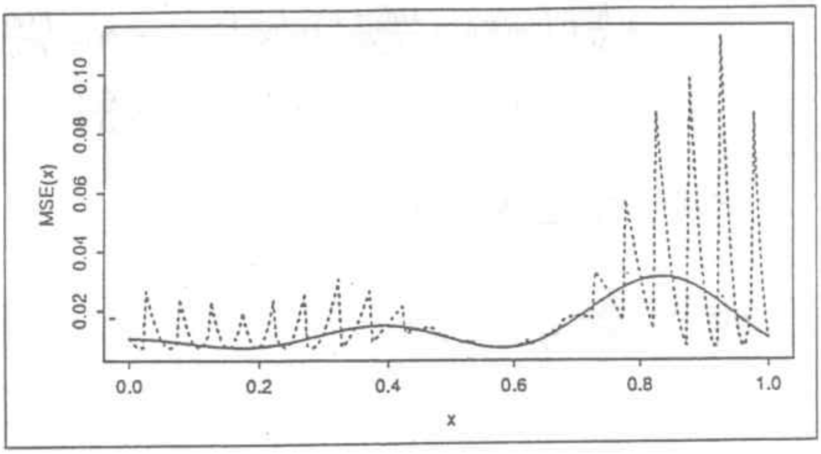


Figure 7. $MSE(x)$ for $\hat{m}_h(x)$ (solid line) and $\hat{m}_M^{(S)}$ (dashed line) for a simulated example $m(x)= x\,sin(2\pi x)$, $x \in U(0,1)$, $\varepsilon \sim N(0,\sigma^2)$, $\sigma^2=0.25$, $h=0.25$, $M=5$, $n=100$, $K=$ Quartic.

Figure 8 shows the MSE as a function of x for the estimator $\hat{m}_M^{(P)}$. Note that the scale is different here. We can see how much better $\hat{m}_M^{(P)}$ is compared to $\hat{m}_M^{(S)}$. For more details, in particular how to bound the difference between the MSE curves, see Breuer (1990).

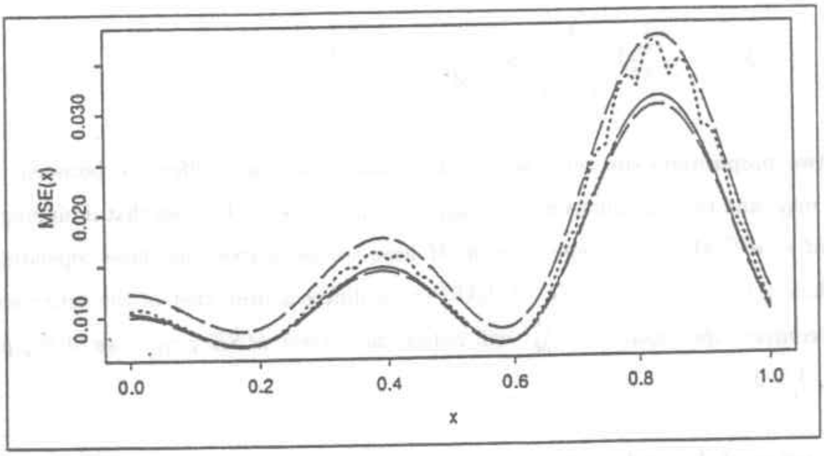


Figure 8. $MSE(x)$ for $\hat{m}_h(x)$ (solid line) and $\hat{m}_M^{(P)}$ (short dashed line) together with bounds on the difference between the two MSE curves. $m(x)=x\,sin(2\pi x)$, $x \in U(0,1)$, $\varepsilon \sim N(0,\sigma^2)$, $\sigma^2=0.25$, $h=0.25$, $M=5$, $n=100$, $K=$ Quartic.

4. WARPing in practice with visualization

The WARP technique was applied to an analysis of laboratory-simulated car side-impacts, see the data in Section 2. Figure 9 shows the WARP density estimate of three biomechanical variables for 29 dummies. The visualization technique has been discussed in Scott and Thompson (1983). The "flying duck" shape of the distribution of the dummy's variables is substantially different from the "frozen duck" shape in Figure 10 that was computed for a different group of experiments. Interestingly there are two clusters at the "drumsticks" of the frozen duck. Both density contours are plotted on the same scale and show very different shapes due to the fact that the frozen duck has higher z-values and lower x-values than the flying duck. Note that the contours would be spherical if the data were independent and normal.

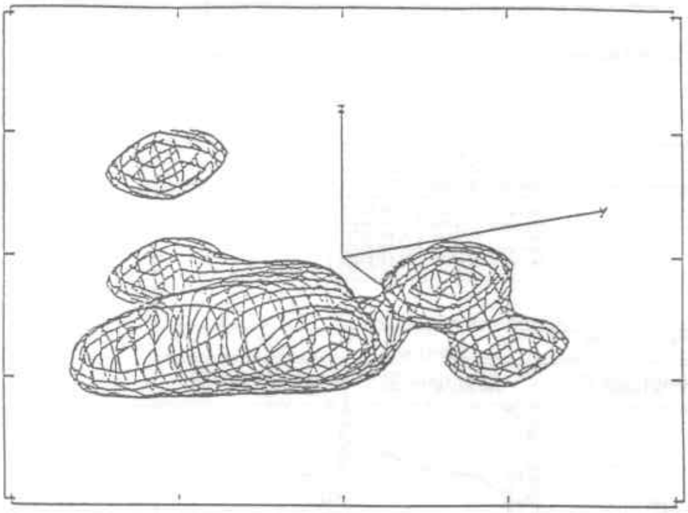


Figure 9. WARP dummy estimates (biweight kernel) for the variables $(ROSYM, RUSYM, T12RM) = (x, y, z)$ giving maximal acceleration (in g) at different regions of the thorax of the dummy's bodies. The contour is at a level 5% of the modal level.

To investigate these data further the GAM/WARP approach has been used. Note that this model generalizes related concepts like logistic regression models where the link function G is known. Figure 11 shows the fifth iteration of the local scoring algorithm (done with XploRe (1990)) using the logistic link function $G(u) = exp(u)/(1 + exp(u))$. It is interesting that the lower left curve shows a typical nonlinear structure. Figure 12 finally shows the functions $\hat{g}_j(x_j)$, $j = 1, 2, 3$. These functions are not linear and not all monotone as might be assumed. However with such small samples, such conclusions can only be tentative.

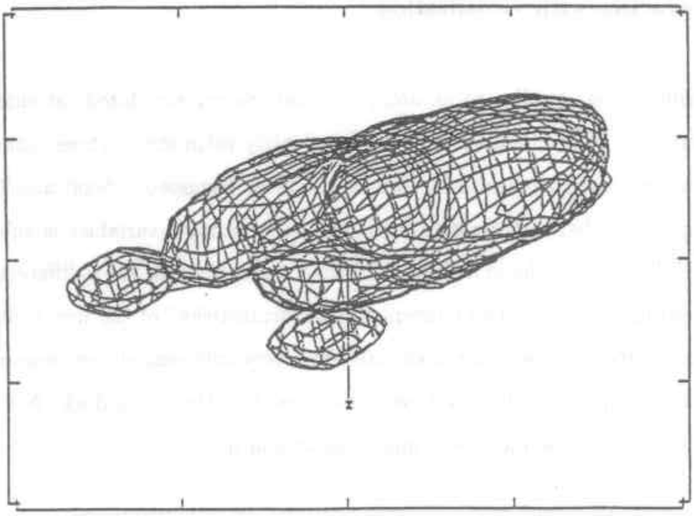
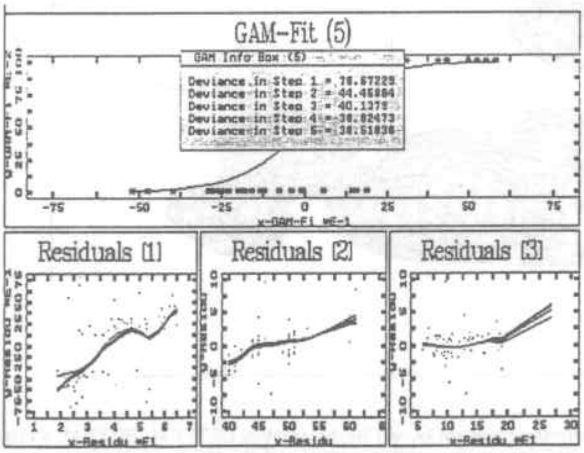


Figure 10. WARP estimates (as in Figure 9) for a different group.



```
Deviance in Step 1 = 76.67229
Deviance in Step 2 = 44.45884
Deviance in Step 3 = 40.1379
Deviance in Step 4 = 38.82473
Deviance in Step 5 = 38.51838
```

Figure 11. Fifth step of the local scoring algorithm applied to the side-impact data. The info box shows the decrease in deviance as the iteration proceeds. Behind the info box one sees the current estimate $\sum_{j=1}^{d} \hat{g}_j(x_j)$ plotted versus Y with the link function G.

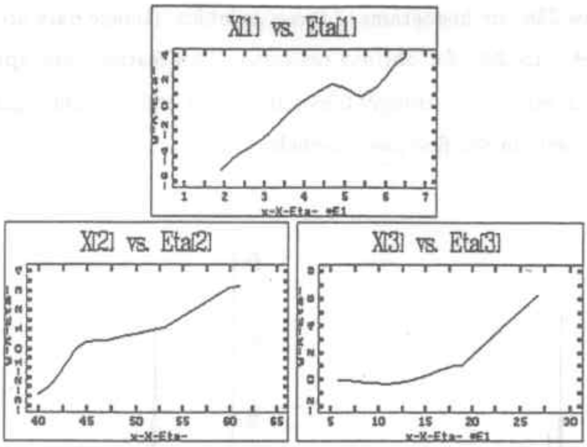


Figure 12. The WARP estimated function, $\hat{g}_j(x_j)$, j=1,2,3 for the side-impact example.

The WARP technique also applies directly to four–dimensional data, but one has to look at a sequence of contour slices with one variable fixed (Scott 1986). We re–examined the classical Iris data with variables (x_1, x_2, x_3, x_4) given by sepal length, petal width, petal length, and sepal width, respectively. The contours of $\hat{f}(x_1, x_2, x_3, x_4 = 3.14\,cm)$ (not shown) clearly reveal three separate spherical contours representing the three clusters of Iris flower varieties present. Other analyses show that the data for two of these varieties are not well-separated. This example is provocative since the sample size is so small.

The final example illustrates the application of WARPing to large data-sets. The simpler visualization techniques used previously can be replaced by more sophisticated surface rendering methods, since large-sample estimates warrant closer inspection (Scott and Hall, 1989). These data were derived from LANDSAT IV measurements taken in summer 1977 on segment 1663, a 5 by 6 nautical mile region in North Dakota in primarily agricultural use (Scott, 1983; Scott and Jee, 1984; and Scott, 1985b). Each segment contains 117 scan lines each with 196 picture elements (pixels), for a total of 22,932 1.1 acre pixels. Five acquisitions were obtained by the 4-channel satellite, which measure light reflectance intensities in fairly narrow bands. These 20 dimensions of data (ignoring any spatial information) were projected into 3 dimensions by the nonlinear model of Badhwar (1980). For agricultural pixels, the three transformed variables have the following interpretations; x, the time of peak "greenness"; y, the ripening period; and z, the peak greenness level.

Figure 13 shows raw 256–bin histograms of these variables. (Image data are usually digitized into 8 bits, numbered 0 to 255, for obvious reasons.) Two features are apparent from these histograms: first, the x-variable is strongly bimodal; second, much of the "bandwidth" in the 8 bits is wasted, particularly in the first two channels.

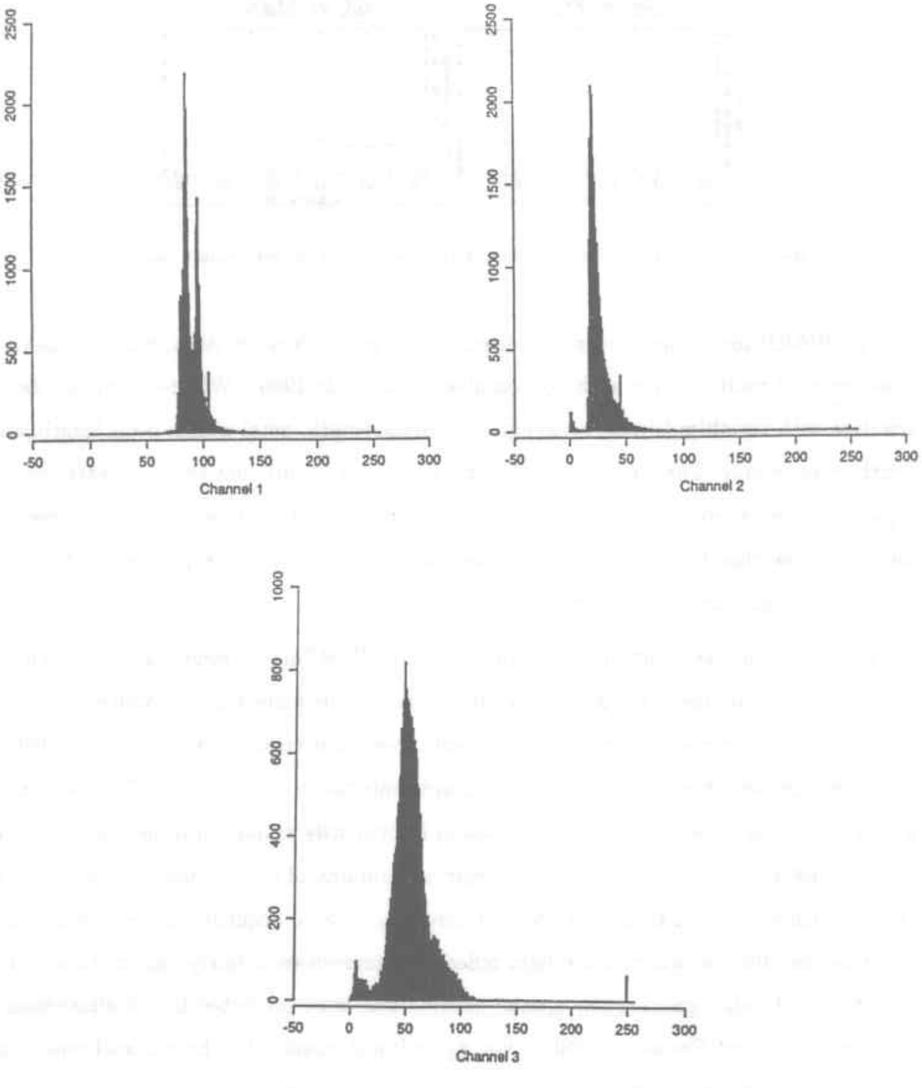


Figure 13. Histograms of the 3 channels of transformed LANDSAT data.

Figure 14 shows the pairwise scatter diagrams of these data. To improve the visualization, the data have been "blurred" by adding uniform noise to the integer values. More multivariate structure is apparent in this diagram than in Figure 13, but it is well to remember that the features away from the central cluster represent less that 10% of the data

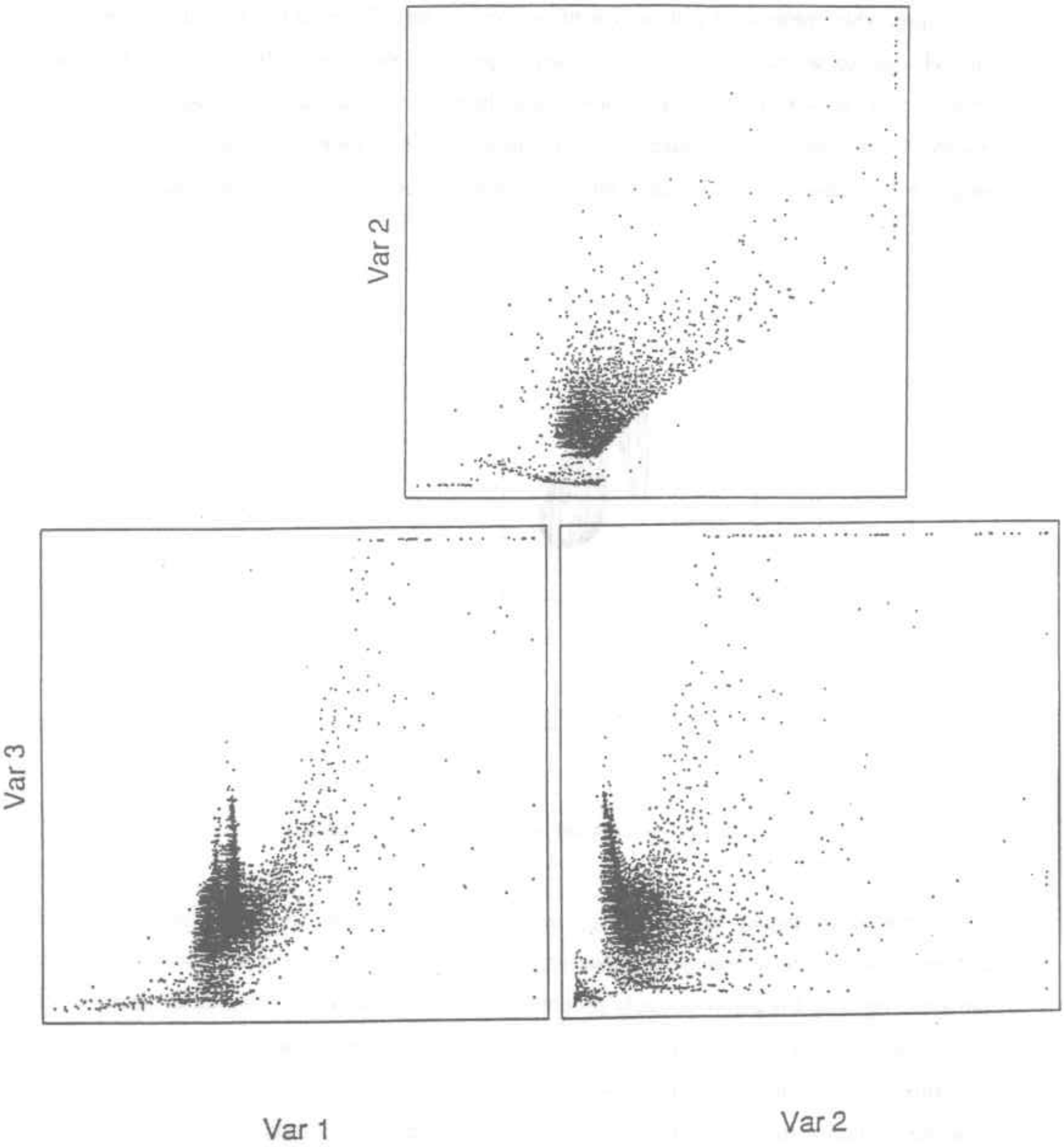


Figure 14. Pairwise scatter diagrams of the 3 channels of transformed LANDSAT data.

A WARP density estimate was computed over the region $[65, 130] \times [0, 75] \times [0, 113]$, which excluded 474 points ($< 2.1\%$). Each axis was divided into 40 bins and $M_i = 3$ was chosen subjectively. Figure 15 displays the $\alpha = 20\%$ density contour, which is clearly bimodal. The left ellipse, which represents primarily sunflower crops, actually leans towards the viewer, while the right ellipse, which represents small grain crops, leans away and is further from the viewer. Observe that the WARP bin structure and triangulation of the marching cube algorithm are both readily visible. Some further surface shading algorithms may be contemplated, but smoothing away the bin edges and local noise information is not to be recommended generally.

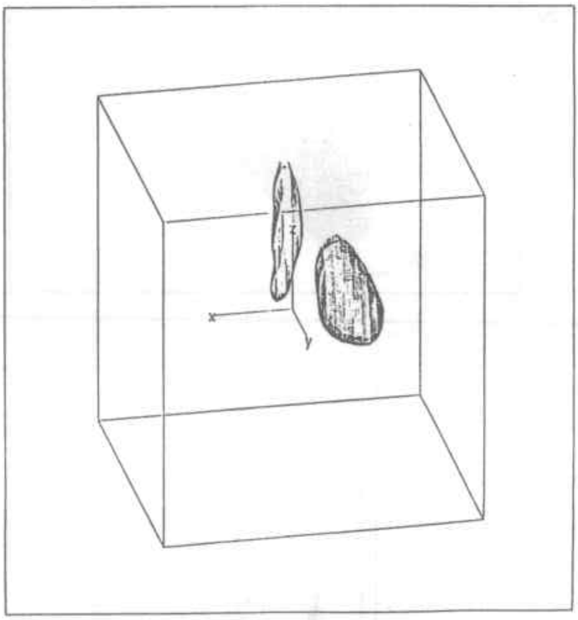


Figure 15. Density shell $S_{20\%}$ for the WARP estimate of the LANDSAT data.

Figures 16–18 display the $S_{10\%}$, $S_{5\%}$, and $S_{2\%}$ contour shells. (Conceptually, these shells are nested, and a transparency visualization algorithm could be employed.) The bump towards the lower right in Figure 17 actually protrudes more than the bump towards the bottom left, as is clear from a stereo or animated representation. Figure 18 displays not only the beautiful structure arising from the complex mixing of different pure crop data, but also begins to show some small bumps in the tails of the data. Given the very large sample, the detailed structure is most likely real and not an artifact.

We believe these surface renderings will become commonplace in nonparametric density and regression applications and that WARPing is essential to implement interactive versions of such applications.

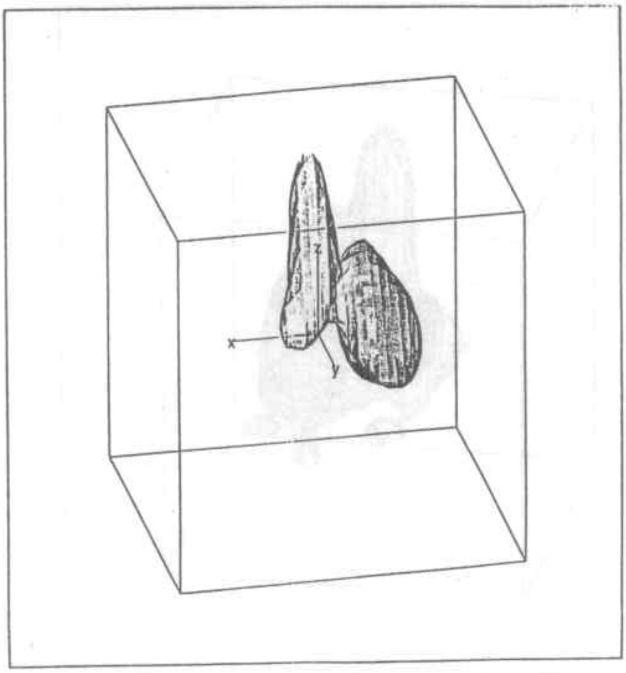


Figure 16. Density shell $S_{10\%}$ for the WARP estimate of the LANDSAT data.

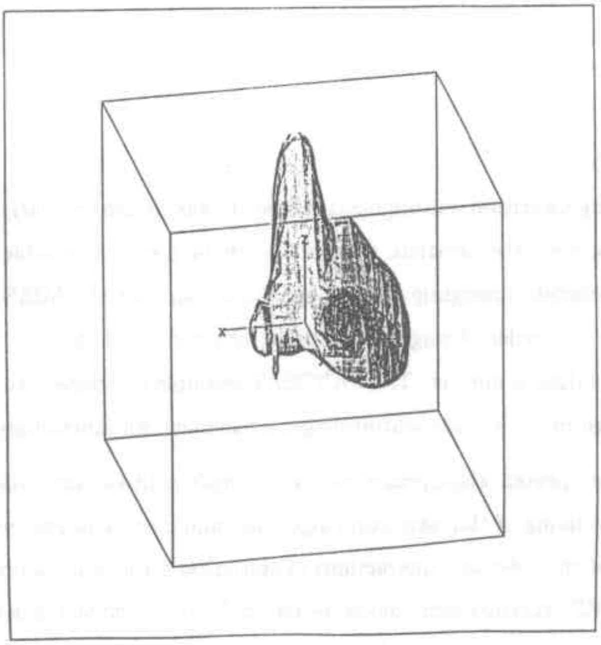


Figure 17. Density shell $S_{5\%}$ for the WARP estimate of the LANDSAT data.

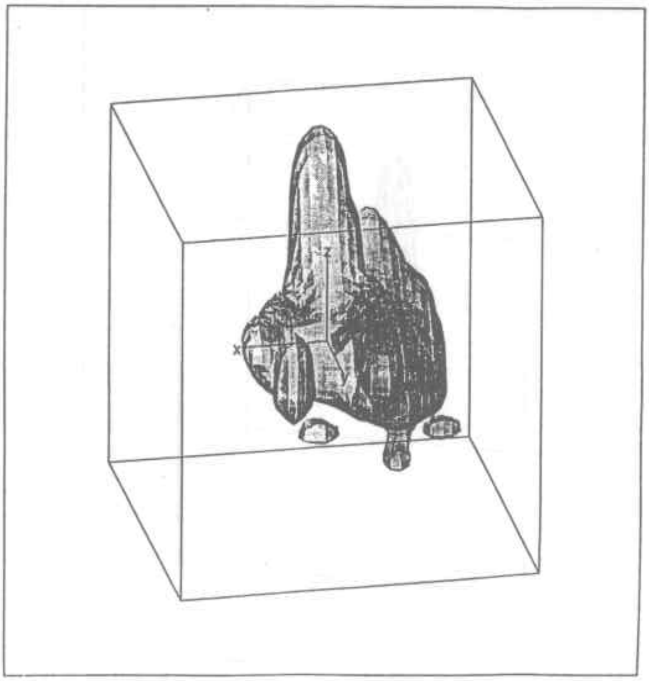


Figure 18. Density shell $S_{2\%}$ for the WARP estimate of the LANDSAT data.

5. Discussion

The WARPing algorithm was originally devised to handle the very large trivariate LANDSAT data. Exploring even the bivariate data in Figure 14 proved infeasible with ordinary kernel estimates. A bivariate histogram was a partial solution, but the MISE of the corresponding kernel estimate is an order of magnitude smaller for a sample of this size. The MISE advantage for the trivariate data is similar. The LANDSAT example illustrates the substantial difference in exploring large raw trivariate scatter diagrams and smooth functionals of those data.

WARPing has proven appropriate even with small multivariate data-sets such as the Iris example. The coupling of the estimation algorithm and the visualization scheme simplifies the understanding of the roles and interactions of statistical errors and (numerical) approximation errors. The WARP estimates carry along the interpolation machinery so no further consideration of numerical approximation errors of visualization is required.

Computational efficiency of the algorithms is acutely felt when hundreds of repetitions are required for bootstrapping, animations, or cross-validation. In the latter case, numerous convolutions are often required. WARPing simply substitutes discrete convolutions on rounded data for the exact calculations. Such an approach has been commonplace in time series analysis for many years. [Many spline-based estimators have exact representations for solutions with exactly n basis functions. Using rounded points could reduce the solution sizes as well; see Wahba (1990).]

For data in more than four dimensions, a two-stage analysis is often most appropriate. As in the LANDSAT example, the high dimensional data are projected to four or fewer dimensions. The projected data are then analyzed using a WARP estimator. In many cases, WARPing can be used in the projection stage.

Our goal is to encourage the use of the growing array of nonparametric ideas by stimulating the development of computationally feasible algorithms such as those provided by the WARPing framework.

References.

Badhwar, G.G. (1980). Crop Emergence Data Determination From Spectral Data. *Photogram. Eng. Remote Sens., 46, 369–377.*

Becker, R.A. and Chambers, J.M. (1984). S: An Interactive Environment for Data Analysis and Graphics. *Wadsworth, Belmont, California.*

Breiman, L. and Friedman, J.H. (1985). Estimating Optimal Transformations for Multiple Regression and Correlation. *Journal of the American Statistical Association, 80, 580–619.*

Breuer, K. (1990). Approximation von Kernglättern durch die WARPing-Methode. *Diploma Thesis.*

Broich, Th., Härdle, W. and Krause, A. (1990). XploRe — A computing environment for eXploRatory Regression. *Springer-Verlag, Berlin.*

Carroll, J. and Härdle, W. (1989). Symmetrized nearest neighbor regression estimates. *Statistics and Probability Letters , 7, 315–318.*

Cencov, N.N (1962). Evaluation of an Unknown Distribution Density from Observations. *Soviet Mathematics, 3, 1559–1562.*

Cleveland, W.S. (1979). Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association, 74, 829–836.*

Duan, N. and Li, K.-C. (1990). Slicing Regression: A Link-Free Regression Method *Annals of Statistics, in press.*

Eubank, R. (1988). Spline Smoothing and Nonparametric Regression. *Marcel Dekker, New York.*

Friedman, J.H. (1984). A Variable Span Smoother. *Tech Rep LCS5, Dept. of Statistics, Stanford.*

Friedman, J.H. and Stuetzle, W. (1981). Project Pursuit Regression. *Journal of the American Statistical Association, 76, 817–823.*

Härdle, W. (1987). Resistant Smoothing using the Fast Fourier Transform. *Appl. Statistics, AS 222, 36, 104–111.*

Härdle, W. (1990). Applied Nonparametric Regression. *Econometric Society Monograph Series, Cambridge University Press, in press*

Härdle, W. and Bowman, A. (1988). Bootstrapping in Nonparametric Regression: Local Adaptive Smoothing and Confidence Bands. *Journal of the American Statistical Association, 83, 102–110.*

Härdle, W., Hall, P., and Marron, J.S. (1988). How Far are Automatically Chosen Regression Smoothing Parameters from Their Optimum? (with discussion). *Journal of the American Statistical Association, 83, 86–101.*

Härdle, W. and Marron, J.S. (1990). Bootstrap Simultaneous Error Bars for Nonparametric Regression. *CORE D.P. 8923, Universite Catholque de Louvain, Belgium.*

Härdle, W. and Stoker, T. (1989). Investigating Smooth Multiple Regression by the Method of Average Derivatives. *Journal of the American Statistical Association, 84, 986–995.*

Hastie, T. and Tibshirani, R. (1987). Generalized Additive Models: Some Applications *Journal of the American Statistical Association, 82, 371-386.*

Heckman, N. (1986). Spline Smoothing in a Partly Linear Model *Journal of the Royal Statistical Society, Series B, 48, 244–248.*

Hjort, N. (1986). On Frequency Polygons and Average Shifted Histograms in Higher Dimensions. *TR 22, Dept. of Statistics, Stanford.*

Huber, P.J. (1985). Projection Pursuit. *Annals of Statistics, 13, 435–475.*

Jones, M.C. (1989). Discretised and Interpolated Kernel Density Estimates. *Journal of the American Statistical Association, 84, 733-741.*

Kallieris, D., Mattern, R., and Härdle W. (1986). Belastbarkeitsgrenze und Verletzungsmechanik des angegurteten PKW-Insassen beim Seitenaufprall. Phase II: Ansätze zur Verletzungsprädiktion. *FAT Schriftenreihe 60, Forschungsvereinigung Automobiltechnik e.V. (FAT).*

Lorensen, W.E. and Cline, H.E. (1987). Marching Cubes: A High Resolution 3D Surface Reconstruction Algorithm. *ACM Computer Graphics, 21, 163–169.*

Müller, H.-G. (1988). Nonparametric Regression Analysis of Longitudinal Data *Springer-Verlag, New York.*

Parzen, E. (1979). Nonparametric Statistical Data Modeling. *Journal of the American Statistical Association, 74, 105–131.*

Priestley, M.B. and Chao, M.T. (1972). Non-parametric Function Fitting. *Journal of the Royal Statistical Society, Series B, 34, 385–392.*

Rice, J.A. (1986). Convergence Rates for Partially Splined Models. *Statistics and Probability Letters, 4, 203–208.*

Scott, D.W. (1983). Nonparametric Probability Density Estimation for Data Analysis in Several Dimensions. *Proceedings of the Twenty-Eighth Conference on the Design of Experiments in Army Research Development and Testing, pp. 387–397.*

Scott, D.W. (1985a). Averaged Shifted Histograms: Effective Nonparametric Density Estimators in Several Dimensions. *The Annals of Statistics, 13, 1024–1040.*

Scott, D.W. (1985b). Classification Using Multivariate Nonparametric Density Estimation. *Proceedings of the Sixth Annual National Computer Graphics Association Conference, Volume III, 715–718.*

Scott, D.W. (1986). Data Analysis in 3 and 4 Dimensions With Nonparametric Density Estimation *in Statistical Image Processing, E.J. Wegman and D. DePriest, Eds., Marcel Dekker, New York, pp. 291–305.*

Scott, D.W. (1988). Software for Cross-Validation of Density Estimates. *Rice Technical Report 88–8–311.*

Scott, D.W. and Hall, M.R. (1989). Interactive Multivariate Density Estimation in the S Language. *Proceedings of the 20th Interface of Computer Science and Statistics, American Statistical Association, Alexandria, Virginia, pp. 241–245.*

Scott, D.W. and Jee, R. (1984). Nonparametric Analysis of Minnesota Spruce and Aspen Tree Data and Landsat Data *Proceedings of the Second Symposium on Mathematical Pattern Recognition and Image Analysis, Dept. of Math, Texas A&M University, pp. 27–49*

Scott, D.W. and Schmidt, H.-P. (1988). Calibrating Histograms with Applications to Economic Data. *Empirical Economics, 13, 155–168.*

Scott, D.W. and Sheather, S.J. (1985). Kernel Density Estimation with Binned Data. *Communications in Statistics, 14, 1353–1359.*

Scott, D.W., Tapia, R.A., and Thompson, J.R. (1980). Nonparametric Probability Density Estimation by Discrete Maximum Penalized-Likelihood Criteria. *Annals of Statistics, 8, 820–832.*

Scott, D.W. and Terrell, G.R. (1987). Biased and Unbiased Cross-Validation in Density Estimation. *Journal of the American Statistical Association, 82, 1131–1146.*

Scott, D.W. and Thompson, J.R. (1983). Probability Density Estimation in Higher Dimensions. *Proceedings of the 15th Symposium on the Interface of Computer Science and Statistics, J.E. Gentle, Ed., North-Holland, Amsterdam, pp. 173–179.*

Scott, D.W. and Wand, M.P. (1990). Feasibility of Multivariate Density Estimates. *To appear, Biometrika.*

Silverman, B.W. (1982). Kernel Density Estimation Using the Fast Fourier Transformation. *Applied Statistics, 31, 93–97.*

Silverman, B.W. (1986). Density Estimation for Statistics and Data Analysis. *Chapman and Hall, London.*

Spiegelman, C.H. (1976). Two Techniques for Estimating Treatment Effects in the Presence of Hidden Variables: Adaptive Regression and a Solution of Reiersols's Problem *Unpublished Ph.D. Thesis, Northwestern University, Dept. Mathematics.*

Stone, C.J. (1977). Consistent Non-parametric Regression. *Annals of Statistics, 5, 595–645.*

Stone, C.J. (1982). Optimal Global Rates of Convergence for Nonparametric Regression. *Annals of Statistics, 10, 1040–1053.*

Stone, C.J. (1985). Additive Regression and other Nonparametric Models. *Annals of Statistics, 13, 689–705.*

Wahba, G. (1990). Spline Models for Observational Data *SIAM, Philadephia*

XploRe (1990). See Broich, Härdle and Krause, 1990.