
Kernel estimators of regression functions

Herman J. Bierens

Abstract: This chapter reviews the asymptotic properties of the Nadaraya-Watson kernel estimator of an unknown (multivariate) regression function. Conditions are set forth for pointwise asymptotic normality and uniform weak consistency. These conditions cover the standard i.i.d. case with continuously distributed regressors, as well as the cases that the distribution of all, or some, regressors is discrete and/or the data are generated by a class of strictly stationary time series processes. Moreover, attention is paid to the problem of how the kernel and the window width should be specified. Furthermore, the estimation procedure under review is illustrated by a numerical example.

1 Introduction

A large extent of applied econometric research involves the specification and estimation of regression models, where in most of the cases the linear regression model is used. The most crucial assumption underlying these models is that they represent the mathematical expectation of the dependent variable conditional on the regressors, which implies that the expectation of the error term conditional on the regressors equals zero with probability 1. If the dependent variable has finite absolute first moment this conditional expectation always exists. Compare Chung (1974, Theorem 9.1.1). Therefore, regression models are either true or false in the sense that they represent conditional expectations given the regressors, or not.

This point of view is at odds with the common belief among econometricians that econometric models are always convenient but imperfect ap-

Paper presented at the Invited Symposium on Nonparametric and Robust Estimation in the Fifth World Congress of the Econometric Society, August 1985, Cambridge, MA. The assistance of Gerard Phann in making the figures and the useful comments of A. Ronald Gallant, Alexander Georgiev, Pedro Gozalo, and Charles Manski are gratefully acknowledged.

proximations of the empirical phenomena under review and that therefore econometric models are by definition false. In particular, one usually distinguishes admissible and nonadmissible models, where the admissibility of a model depends on its purpose. Given the dependent variable and the regressors, however, there is only one regression model that obeys the usual conditions on the error term, namely the conditional expectation mentioned above. Therefore, as long as these conditions on the error term are adopted, the only admissible model is the true model, in the above sense. As a consequence, the specification of the functional form of an econometric regression model is basically a statistical problem rather than an economic problem. Only the choice of the dependent and the explanatory variables themselves is an economic problem. As soon as these variables have been selected, the corresponding regression model has implicitly been selected too. In other words, the model is uniquely determined by the data. Thus, adding or deleting variables from the list of explanatory variables may change the functional form of the model. See Bierens (1984, p. 326). This argument, however, applies only to regression models. The conditional expectation is not the correct location "parameter" for every econometric problem. For example, in nonlinear structural modeling the conditional expectation is not necessarily an aid to recovering either the structural model or its reduced form.

The usual practice in constructing an econometric regression model is to specify a parametric family for the response function. Obviously, the most popular parametric family is the linear model. One could consider this as choosing a parametric functional form from a continuum of possible functional forms, analogous to sampling from a continuous distribution. Therefore, the probability that we pick the true functional form in this way is zero or at least very close to zero.

The only way to avoid model misspecification is to specify no functional form at all. But then the problem arises how information about the functional form of the model can be derived from the data. Before Gallant's (1981) paper on the Fourier flexible form, this problem has not been rigorously addressed in the econometric literature. Of course, econometricians have dealt with flexible functional forms before, but mainly in the framework of the well-known Box-Cox (1964) transformation. See Zarembka (1968), White (1972), Leech (1975), and Spitzer (1976), among others, for empirical applications of the Box-Cox transformation. As is well known, the linear and log-linear models are members of the class of functional forms covered by the Box-Cox transformation. It is also obvious, however, that there is no guarantee that the true model belongs to this class. So the Box-Cox transformation is by no means a watertight protection against model specification errors.

Gallant's method is based on the fact that under mild conditions a real function on a bounded domain can be expanded in an infinite Fourier series. Estimating the coefficients of this Fourier series after suitable truncation by, say, least squares gives a direct estimate of the functional form of the model without need to specify a parametric family. This approach is reminiscent of the orthogonal series expansion approach in estimating probability density functions of unknown form, proposed by Cencov (1962) and others. See Fryer (1977) and Tapia and Thompson (1978) for a review of this and other methods for nonparametric density estimation. That brings us to the statistical literature on nonparametric estimation of unknown density and regression functions. The problem of nonparametric estimation of a density function has received extensive attention in the statistical literature (see the above-mentioned reviews), and since the pioneering papers of Nadaraya (1964) and Watson (1964) there is now a growing extent of literature on the related problem of nonparametric estimation of unknown regression functions. See Collomb (1981, 1985b) for a review. Most of the literature on nonparametric regression function estimation deals with the kernel method and its variants.

In this chapter we shall discuss the kernel method. The reason for emphasizing the kernel method is twofold. First, due to recent generalizations of this method to time series data and discontinuously distributed regressors, it has now become relevant for econometrics. Second, the kernel method is easier to apply to multivariate regression problems than other methods (perhaps apart from Gallant's approach), although the performance of this method is negatively related to the number of regressors.

The character of this chapter is mainly that of an introduction. Our main aim is to explain the theory to the general reader in an accessible way, although some new results will be presented. Nevertheless, accessibility might be enhanced by first reading McFadden (1985), who considers the model specification problem from various points of view, including the nonparametric point of view. A possible exception to mathematical accessibility might be the section on time series applications, where we employ martingale and mixing concepts. Rather than going into mathematical sloppiness, we have sought accessibility in not employing the utmost general conditions (though general enough for econometric applications). Moreover, we shall discuss some practical matters in applying these methods, together with a numerical example.

The plan of this chapter is as follows. In Section 2 we consider pointwise asymptotic normality and uniform weak consistency in the case of an i.i.d. data-generating process with continuously distributed regressors. Section 3 deals with the i.i.d. case where all, or some, regressors are discrete, thus allowing for qualitative explanatory variables. Again point-

wise asymptotic normality and uniform weak consistency are shown. In Section 4 we set forth conditions such that the asymptotic results in the i.i.d. case go through for a class of strictly stationary time series processes. Section 5 is devoted to practical matters, including a numerical example.

2 Kernel estimators of regression functions: the continuous i.i.d. case

2.1 Introduction

For a start, let us consider the easiest data-generating process, namely the case where we have an i.i.d. sample $\{(y_1, x_1), \dots, (y_n, x_n)\}$ from an absolutely continuous $k+1$ -variate distribution with density $f(y, x)$, where $y \in R$ and $x \in R^k$. In this data the y_j 's are the dependent variables and the x_j 's are k -component vectors of regressors. If $E|y_j| < \infty$, then the conditional expectation of y_j given x_j exists and takes the form

$$E(y_j | x_j) = g(x_j) \quad (2.1.1)$$

with $g(\cdot)$ a Borel measure real function on R^k . Compare Chung (1974, Theorems 9.1.1 and 9.1.2). Denoting

$$u_j = y_j - g(x_j) \quad (2.1.2)$$

we then get the regression model

$$y_j = g(x_j) + u_j \quad (2.1.3)$$

where, by construction, the error term u_j satisfies the usual condition that its expectation conditional on the vector of regressors equals zero with probability 1 (w.p.1), that is,

$$E(u_j | x_j) = 0 \quad \text{w.p.1} \quad (2.1.4)$$

The model (2.1.3) is therefore purely tautological, that is, its setup is merely a matter a definition. Now our aim is to estimate the regression function $g(\cdot)$ without making explicit assumptions about its functional form.

As said before, the regression model is completely determined by the data-generating process. For the data-generating process under review the regression function $g(\cdot)$ takes the well-known form

$$g(x) = \int y f(y, x) dy / h(x) \quad \text{if } h(x) > 0 \quad (2.1.5)$$

where $h(x)$ is the marginal density of $f(y, x)$; that is,

$$h(x) = \int f(y, x) dy \quad (2.1.6)$$

This suggests estimation of the function $g(x)$ via estimating the densities f and h .

A convenient method for estimating unknown multivariate density functions is the kernel density estimation method proposed by Rosenblatt (1956). Important contributions to the asymptotic theory of this class of estimators have been made by Parzen (1962) for the univariate case and Cacoullos (1966) for the multivariate case. A kernel estimator of the density $h(x)$ is a random function of the form

$$\hat{h}(x) = \frac{1}{n} \sum_{j=1}^n \frac{K[(x-x_j)/\gamma_n]}{\gamma_n^k} \quad (2.1.7)$$

where $K(\cdot)$ is an a priori chosen real function on R^k , called the *kernel*, satisfying

$$\int |K(x)| dx < \infty \quad \int K(x) dx = 1 \quad (2.1.8)$$

and (γ_n) is an a priori chosen sequence of positive numbers, called *window width* parameters, satisfying

$$\lim_{n \rightarrow \infty} \gamma_n = 0 \quad \lim_{n \rightarrow \infty} n\gamma_n^k = \infty \quad (2.1.9)$$

Under conditions (2.1.8) and (2.1.9), the estimator $\hat{h}(x)$ is pointwise consistent in every continuity point of $h(x)$, provided

$$\sup_x h(x) < \infty \quad (2.1.10)$$

The proof of this proposition is simple but instructive. First, the asymptotic unbiasedness follows from

$$\begin{aligned} E\hat{h}(x) &= \int \gamma_n^{-k} K[(x-z)/\gamma_n] h(z) dz = \int h(x - \gamma_n z) K(z) dz \\ &\rightarrow h(x) \int K(z) dz = h(x) \end{aligned} \quad (2.1.11)$$

by bounded convergence. Second, the variance vanishes at order $O(1/n\gamma_n^k)$, as

$$\begin{aligned} n\gamma_n^k \text{var}[\hat{h}(x)] &= n\gamma_n^k \frac{1}{n^2} \sum_{j=1}^n \text{var} \left\{ \frac{K[(x-x_j)/\gamma_n]}{\gamma_n^k} \right\} \\ &= E\gamma_n^{-k} K[(x-x_j)/\gamma_n]^2 - \gamma_n^k \{E\gamma_n^{-k} K[(x-x_j)/\gamma_n]\}^2 \\ &= \int h(x - \gamma_n z) K(z)^2 dz - \gamma_n^k \left[\int h(x - \gamma_n z) K(z) dz \right]^2 \\ &\rightarrow h(x) \int K(z)^2 dz \end{aligned} \quad (2.1.12)$$

by bounded convergence. This completes the pointwise consistency proof.

We shall now construct a kernel density estimator $\hat{f}(y, x)$ of the joint density $f(y, x)$ such that $\hat{h}(x)$ is the marginal density of $\hat{f}(y, x)$ and the integral $\int y \hat{f}(y, x) dy$ yields an expression involving the same kernel K as in (2.1.7). This kernel estimator of $f(y, x)$ is of the form

$$\hat{f}(y, x) = \frac{1}{n} \sum_{j=1}^n K_*[(y-y_j)/\gamma_n, (x-x_j)/\gamma_n] \gamma_n^{-k-1} \quad (2.1.13)$$

where the kernel K_* satisfies

$$\int y K_*(y, x) dy = 0 \quad \int K_*(y, x) dy = K(x) \quad (2.1.14)$$

Then $\hat{h}(x)$ is the marginal density of $\hat{f}(y, x)$, and moreover,

$$\int y \hat{f}(y, x) dy = \frac{1}{n} \sum_{j=1}^n y_j K[(x-x_j)/\gamma_n] \gamma_n^{-k} \quad (2.1.15)$$

and hence the corresponding regression function estimator of (2.1.5) is

$$\hat{g}(x) = \frac{\sum_{j=1}^n y_j K[(x-x_j)/\gamma_n]}{\sum_{j=1}^n K[(x-x_j)/\gamma_n]} \quad (2.1.16)$$

This is the so-called Nadaraya-Watson kernel regression function estimator, named after Nadaraya (1964) and Watson (1964). Note that this kernel regression function estimate is a weighted mean of the dependent variables y_j , where the weights sum to 1. In particular, if the kernel is chosen to be a unimodal density function with zero mode (e.g., let the kernel be the density of the k -variate standard normal distribution) then the closer x is to x_j , the more weight is put on y_j .

Similarly to (2.1.11) and (2.1.12), it can now be shown that

$$E \hat{g}(x) \hat{h}(x) = \int g(x - \gamma_n z) h(x - \gamma_n z) K(z) dz \rightarrow g(x) h(x) \quad (2.1.17)$$

and

$$\begin{aligned} n \gamma_n^k \text{var} [\hat{g}(x) \hat{h}(x)] &= \int \sigma_y^2(x - \gamma_n z) h(x - \gamma_n z) K(z)^2 dz \\ &\quad - \gamma_n^k \left[\int g(x - \gamma_n z) h(x - \gamma_n z) K(z) dz \right]^2 \\ &= O(1) \end{aligned} \quad (2.1.18)$$

where

$$\sigma_y^2(x) = E(y_j^2 | x_j = x) \quad \text{for } h(x) > 0 \quad (2.1.19)$$

provided

$$\sup_x |g(x)| h(x) < \infty \quad \sup_x \sigma_y^2(x) h(x) < \infty \quad (2.1.20)$$

Now it is easy to verify from (2.1.11), (2.1.12), (2.1.17), and (2.1.18) that

$$p \lim_{n \rightarrow \infty} \hat{h}(x) = h(x) \quad p \lim_{n \rightarrow \infty} \hat{g}(x) \hat{h}(x) = g(x) h(x) \quad (2.1.21)$$

and hence

$$p \lim_{n \rightarrow \infty} \hat{g}(x) = g(x) \quad (2.1.22)$$

in every continuity point x of $h(x)$ and $g(x)h(x)$ for which $h(x) > 0$.

The weak consistency of the kernel regression function estimator is not limited to the case that x_j is continuously distributed, as is shown by Devroye (1978) and Bierens (1983a). We shall consider the (partly) discrete case in Section 3. Uniform weak consistency will also be considered in the sequel of this chapter. Also, strong consistency results are available in the literature (e.g., Nadaraya 1965, 1970; Noda 1976), but these results will not be discussed. Moreover, the L^p convergence concept, that is,

$$\lim_{n \rightarrow \infty} \int |\hat{g}(x) - g(x)|^p dH(x) = 0 \quad \text{for some } p > 0 \quad (2.1.23)$$

with H the distribution function of the x_j , will also not be considered. For that we refer to Konakov (1977) and Devroye and Wagner (1980a).

If we replace the window width γ_n in (2.1.16) by γ_j , we get the recursive kernel regression function estimator. This type of estimator has been studied by Ahmad and Lin (1976), Greblicki and Krzyzak (1980), and Devroye and Wagner (1980b). Finally, we note that other variants of the kernel regression approach can be found in Priestley and Chao (1972), Benedetti (1977), Clark (1977, 1979, 1980), Stone (1977), Schuster and Yakowitz (1979), Spiegelman and Sacks (1980), Devroye and Wagner (1980b), Devroye and Wise (1980), Cheng and Lin (1981a, b), Bierens (1983b), and Georgiev (1984b-e). All these variants are out of the scope of this chapter. Thus, we only consider the basic properties of the "classical" Nadaraya-Watson kernel regression estimator (2.1.16).

2.2 Asymptotic normality

The kernel regression estimation approach distinguishes itself from other nonparametric regression methods in that asymptotic distribution theory is fairly well established. In particular, the asymptotic normality of the kernel regression function estimator under the conditions under review has been proved by Schuster (1972) for the univariate case ($k = 1$). Here we shall derive asymptotic normality in a somewhat different, but much easier, way for the general case $k \geq 1$.

Observe from (2.1.3), (2.1.7), and (2.1.16) that

$$\begin{aligned}
[\hat{g}(x) - g(x)]\hat{h}(x) &= \frac{1}{n} \sum_{j=1}^n u_j K\left[\frac{x-x_j}{\gamma_n}\right] \gamma_n^{-k} \\
&\quad + \frac{1}{n} \sum_{j=1}^n \left\{ [g(x_j) - g(x)] K\left[\frac{x-x_j}{\gamma_n}\right] \gamma_n^{-k} \right. \\
&\quad \quad \left. - E[g(x_j) - g(x)] K\left[\frac{x-x_j}{\gamma_n}\right] \gamma_n^{-k} \right\} \\
&\quad + \frac{1}{n} \sum_{j=1}^n E[g(x_j) - g(x)] K\left[\frac{x-x_j}{\gamma_n}\right] \gamma_n^{-k} \\
&= \hat{q}_1(x) + \hat{q}_2(x) + \hat{q}_3(x) \tag{2.2.1}
\end{aligned}$$

say. We shall now set forth conditions such that first

$$\sqrt{n\gamma_n^k} \hat{q}_1(x) \rightarrow N\left(0, \sigma_u^2(x) h(x) \int K(z)^2 dz\right) \tag{2.2.2}$$

in distribution, where, for $h(x) > 0$,

$$\sigma_u^2(x) = E(u_j^2 | x_j = x) \tag{2.2.3}$$

is the conditional variance of the u_j . Second, we show that

$$\lim_{n \rightarrow \infty} E[\sqrt{n\gamma_n^k} \hat{q}_2(x)]^2 = 0 \tag{2.2.4}$$

and finally we set forth conditions such that

$$\lim_{n \rightarrow \infty} \gamma_n^{-2} \hat{q}_3(x) \tag{2.2.5}$$

exists. The conditions we need are the following. For $p > 0$ let

$$\sigma_u^p(x) = E(|u_j|^p | x_j = x) \tag{2.2.6}$$

provided $E|u_j|^p < \infty$ and $h(x) > 0$.

Assumption 2.2.1. There exists a $\delta > 0$ such that $\sigma_u^{2+\delta}(x)h(x)$ is uniformly bounded. The functions $g(x)^2h(x)$ and $\sigma_u^2(x)h(x)$ are continuous and uniformly bounded. The functions $h(x)$ and $g(x)h(x)$ and their first and second partial derivatives are continuous and uniformly bounded.

First we prove (2.2.2). Denote

$$v_{n,j}(x) = u_j \frac{K[(x-x_j)/\gamma_n]}{\sqrt{\gamma_n^k}} \tag{2.2.7}$$

Since $\sqrt{n\gamma_n^k} \hat{q}_1(x) = (1/\sqrt{n}) \sum_{j=1}^n v_{n,j}(x)$, it suffices to show that the double array $(v_{n,j}(x))$ satisfies the conditions of Lyapunov's central limit theorem (cf. Chung 1974, p. 209). Thus, the results

$$E v_{n,j}(x) = 0 \tag{2.2.8}$$

$$E v_{n,j}(x)^2 = E u_j^2 K[(x-x_j)/\gamma_n]^2 \gamma_n^{-k} = \int \sigma_u^2(x - \gamma_n z) K(z)^2 dz \rightarrow$$

$$\rightarrow \sigma_u^2(x) h(x) \int K^2(z) dz \quad (2.2.9)$$

and

$$\begin{aligned} \sum_{j=1}^n E|v_{n,j}(x)/\sqrt{n}|^{2+\delta} &= \left(\frac{1}{\sqrt{n\gamma_n^k}}\right)^\delta E|u_j|^{2+\delta} \left|K\left[\frac{x-x_j}{\gamma_n}\right]\right|^{2+\delta} \gamma_n^{-k} \\ &= \left(\frac{1}{\sqrt{n\gamma_n^k}}\right)^\delta \int \sigma_u^{2+\delta}(x-\gamma_n z) h(x-\gamma_n z) |K(z)|^{2+\delta} dz \\ &= O\left(\frac{1}{\sqrt{n\gamma_n^k}}\right)^\delta \rightarrow 0 \quad \text{for some } \delta > 0 \end{aligned} \quad (2.2.10)$$

imply

$$\frac{1}{\sqrt{n}} \sum_{j=1}^n v_{n,j}(x) \rightarrow N\left(0, \sigma_u^2(x) h(x) \int K(z)^2 dz\right) \quad (2.2.11)$$

in distribution. This proves (2.2.2).

Next, observe that similarly to (2.1.12),

$$\begin{aligned} E[\sqrt{n\gamma_n^k} \hat{q}_2(x)]^2 &= \int [g(x-\gamma_n z) - g(x)]^2 h(x-\gamma_n z) K(z)^2 dz \\ &\quad - \gamma_n^k \left\{ \int [g(x-\gamma_n z) - g(x)] h(x-\gamma_n z) K(z) dz \right\}^2 \rightarrow 0 \end{aligned} \quad (2.2.12)$$

by bounded convergence. This proves (2.2.4).

Finally, observe that similarly to (2.2.11),

$$\begin{aligned} \hat{q}_3(x) &= \int [g(x-\gamma_n z) - g(x)] h(x-\gamma_n z) K(z) dz \\ &= \int [g(x-\gamma_n z) h(x-\gamma_n z) - g(x) h(x)] K(z) dz \\ &\quad - g(x) \int [h(x-\gamma_n z) - h(x)] K(z) dz \\ &= -\gamma_n \int z' \frac{\partial}{\partial x'} [g(x) h(x)] K(z) dz \\ &\quad + \frac{1}{2} \gamma_n^2 \int z' \frac{\partial}{\partial x} \frac{\partial}{\partial x'} \{g[x-\gamma_n \lambda_n(x, z)] h[x-\gamma_n \lambda_n(x, z)]\} z K(z) dz \\ &\quad + \gamma_n g(x) \int z' \frac{\partial}{\partial x'} h(x) K(z) dz \\ &\quad - \frac{1}{2} \gamma_n^2 g(x) \int z' \frac{\partial}{\partial x} \frac{\partial}{\partial x'} h[x-\gamma_n \lambda_n(x, z)] z K(z) dz \end{aligned} \quad (2.2.13)$$

where $0 \leq \lambda_n(x, z) \leq 1$. The last equality in (2.2.13) follows from Taylor's theorem. Thus, if we choose K such that

$$\int xK(x) dx = 0 \quad \int xx'K(x) dx = \Omega \quad (2.2.14)$$

is finite, then by bounded convergence

$$\lim_{n \rightarrow \infty} \gamma_n^{-2} \hat{q}_3(x) = b(x) \quad (2.2.15)$$

where

$$b(x) = \frac{1}{2} \operatorname{tr} \left\{ \Omega \frac{\partial}{\partial x'} \frac{\partial}{\partial x'} [g(x)h(x)] \right\} - \frac{1}{2} g(x) \operatorname{tr} \left[\Omega \frac{\partial}{\partial x} \frac{\partial}{\partial x'} h(x) \right] \quad (2.2.16)$$

This proves (2.2.5).

From (2.2.1), (2.2.2), (2.2.4), and (2.2.5) the following theorem easily follows.

Theorem 2.2.1. *Let Assumption 2.2.1 and condition (2.2.14) hold and let $h(x) > 0$. If*

$$\lim_{n \rightarrow \infty} \gamma_n^2 \sqrt{n} \gamma_n^k = \lambda \quad \text{with } 0 \leq \lambda < \infty \quad (2.2.17)$$

then

$$\sqrt{n} \gamma_n^k [\hat{g}(x) - g(x)] \rightarrow N \left(\frac{\lambda b(x)}{h(x)}, \frac{\sigma_u^2(x)}{h(x)} \int K(z)^2 dz \right) \quad (2.2.18)$$

in distribution. If

$$\lim_{n \rightarrow \infty} \gamma_n^2 \sqrt{n} \gamma_n^k = \infty \quad (2.2.19)$$

then

$$p \lim_{n \rightarrow \infty} \gamma_n^{-2} [\hat{g}(x) - g(x)] = \frac{b(x)}{h(x)} \quad (2.2.20)$$

Note that the latter result may be considered as convergence in distribution to a degenerated limiting distribution.

At first sight it looks attractive to choose the window width γ_n such that $\lambda = 0$, as then the asymptotic bias vanishes. This is done by Singh and Ullah (1985). However, in that case the asymptotic rate of convergence in distribution is lower than in the case $\lambda > 0$, as (2.2.17) implies

$$\frac{\sqrt{n} \gamma_n^k}{n^{2/(k+4)}} \rightarrow \lambda^{k/(k+4)} \quad \text{as } n \rightarrow \infty \quad (2.2.21)$$

This corresponds to the fact that minimizing the integrated mean square error $E \int [\hat{g}(x)h(x) - g(x)h(x)]^2 dx$ yields an optimal window width of

the form (2.2.17) with $\lambda > 0$. Thus, the window width γ_n that gives the maximum rate of convergence in distribution is

$$\gamma_n = cn^{-1/(k+4)} \quad (2.2.22)$$

where $c > 0$ is a constant. Since $\lambda = c^{(k+4)/2}$, we have the following corollary.

Corollary 2.2.1. *Let the conditions of Theorem 2.2.1 hold. With the window width (2.2.22) we have*

$$n^{2/(k+4)}[\hat{g}(x) - g(x)] \rightarrow N\left(\frac{c^2 b(x)}{h(x)}, c^{-k} \frac{\sigma_u^2(x)}{h(x)} \int K(z)^2 dz\right) \quad (2.2.23)$$

Note that the asymptotic rate of convergence in distribution is negatively related to the number of regressors. This is typical for nonparametric regression, for the more regressors we have, the more information we ask from the data and thus the more observations we need to get a useful answer.

The result (2.2.23) has practical significance only if it is possible to estimate the mean and the variance of the limiting normal distribution involved. As far as the variance is concerned, consistent estimation will appear to be feasible. Regarding the mean, however, the estimation problem is too hard. Inspecting the function $b(x)$ [cf. (2.2.16)] reveals that estimating this function is awkward, as $b(x)$ is a function of the second derivatives of the unknown functions $h(x)$ and $g(x)h(x)$. It would therefore be preferable to get rid of the mean of the limiting normal distribution. We already mentioned a way to do that, namely, to choose the window width such that the limit λ in (2.2.17) is zero, but then we also sacrifice some of the speed of convergence. There is, however, another way to get rid of the asymptotic bias while maintaining the maximal rate of convergence in distribution of order $n^{2/(k+4)}$, namely, by combining the results (2.2.18) and (2.2.20). The idea is to use (2.2.20) for estimating the mean of the limiting normal distribution in (2.2.18) by subtracting the random function at the left-hand side of (2.2.20) times λ from the left-hand side of (2.2.18).

Corollary 2.2.2. *Let the conditions of Theorem 2.2.1 hold. Let $\hat{g}_1(x)$ be the kernel regression estimator with window width*

$$\gamma_n = cn^{-1/(k+4)}$$

and let $\hat{g}_2(x)$ be the kernel regression estimator with window width

$$\gamma_n = cn^{-\delta/(k+4)} \quad \text{with } \delta \in (0, 1)$$

Denote

$$\hat{g}(x) = \frac{\hat{g}_1(x) - n^{-2(1-\delta)/(k+4)} \hat{g}_2(x)}{1 - n^{-2(1-\delta)/(k+4)}} \quad (2.2.24)$$

Then

$$n^{2/(k+4)}[\hat{g}(x) - g(x)] \rightarrow N\left(0, c^{-k} \frac{\sigma_u^2(x)}{h(x)} \int K(z)^2 dz\right) \quad (2.2.25)$$

in distribution.

Note that for the estimator $g_1(x)$ the result (2.2.18) holds with $\lambda > 0$, whereas for $\hat{g}_2(x)$ the result (2.2.20) holds. The proof of this corollary follows therefore straightforwardly from the fact that, by (2.2.20),

$$n^{2/(k+4)}[\hat{g}_1(x) - g(x)] - c^2(cn^{-\delta/(k+4)})^{-2}[\hat{g}_2(x) - g(x)]$$

is asymptotically distributed as

$$n^{2/(k+4)}[\hat{g}_1(x) - g(x)] - \frac{c^2 b(x)}{h(x)}$$

This easy result, however, is not a standard result in the literature but one of the innovations in this chapter.

The rate of convergence in distribution is determined by the rate of convergence of the expectation $\hat{q}_3(x)$. If we would choose the kernel K such that $\int xK(x) dx = 0$ and $\int xx'K(x) dx = 0$, then it can be shown that instead of (2.2.15), $\lim_{n \rightarrow \infty} \gamma_n^{-3} \hat{q}_3(x)$ exists and is finite. The asymptotic rate of convergence in distribution then becomes $n^{3/(k+6)}$ instead of $n^{2/(k+4)}$. Thus, a way to improve the convergence in distribution is to choose a kernel with zero moments up to a particular order m . More precisely, following Singh (1981), we define the class $\mathcal{K}_{k,m}$ of these kernels as follows.

Definition 2.2.1. Let $\mathcal{K}_{k,m}$ be the class of all Borel measurable bounded real-valued functions $K(\cdot)$ on R^k such that, with $z = (z_1, \dots, z_k)'$, $z_i \in R$,

$$\int z_1^{i_1} z_2^{i_2} \cdots z_k^{i_k} K(z_1, z_2, \dots, z_k) dz_1 \cdots dz_k = \begin{cases} 1 & \text{if } i_1 = i_2 = \cdots = i_k = 0 \\ 0 & \text{if } 0 < i_1 + i_2 + \cdots + i_k < m \end{cases} \quad (2.2.26)$$

$$\int |z|^i |K(z)| dz < \infty \quad \text{for } i=0 \text{ and } i=m \quad \int K(z) dz = 1$$

With $K \in \mathcal{K}_{k,m}$ we then have

$$\lim_{n \rightarrow \infty} \gamma_n^{-m} \hat{q}_3(x) = b^*(x) \quad (2.2.27)$$

say, provided $h(x)$ and $g(x)h(x)$ belong to the class $\mathcal{D}_{k,m}$.

Definition 2.2.2. Let $\mathcal{D}_{k,m}$ be the class of all continuous real functions f on R^k such that the derivatives

$$\left(\frac{\partial}{\partial z_1}\right)^{i_1} \left(\frac{\partial}{\partial z_2}\right)^{i_2} \cdots \left(\frac{\partial}{\partial z_k}\right)^{i_k} f(z_1, \dots, z_k) \quad i_j \geq 0, \quad j = 1, \dots, k$$

are continuous and uniformly bounded for $0 \leq i_1 + i_2 + \cdots + i_k \leq m$.

Thus, similarly to Theorem 2.2.1, we have:

Theorem 2.2.2. Let Assumption 2.2.1 and the additional conditions $h(x) \in \mathcal{D}_{k,m}$, $g(x)h(x) \in \mathcal{D}_{k,m}$, $K \in \mathcal{K}_{k,m}$ hold, where m is an integer greater than 2. Let $h(x) > 0$. There exists a real function $b^*(x)$ on R^k such that

$$\lim_{n \rightarrow \infty} \gamma_n^m \sqrt{n\gamma_n^k} = \lambda \quad \text{with } 0 \leq \lambda < \infty \quad (2.2.28)$$

implies

$$\sqrt{n\gamma_n^k} [\hat{g}(x) - g(x)] \rightarrow N\left(\frac{\lambda b^*(x)}{h(x)}, \frac{\sigma_u^2(x)}{h(x)} \int K(z)^2 dz\right) \quad (2.2.29)$$

in distribution, and

$$\lim_{n \rightarrow \infty} \gamma_n^m \sqrt{n\gamma_n^k} = \infty \quad (2.2.30)$$

implies

$$p \lim_{n \rightarrow \infty} \gamma_n^{-m} [\hat{g}(x) - g(x)] = \frac{b^*(x)}{h(x)} \quad (2.2.31)$$

The optimal rate of convergence in distribution is now $n^{m/(2m+k)}$, and the corresponding window width is

$$\gamma_n = cn^{-1/(2m+k)} \quad (2.2.32)$$

with $c > 0$ a constant. Moreover, similarly to Corollaries 2.2.1 and 2.2.2, we now have:

Corollary 2.2.3. Let the conditions of Theorem 2.2.2 hold. With window width (2.2.32) we have:

$$n^{m/(2m+k)}[\hat{g}(x) - g(x)] \rightarrow N\left(\frac{c^m b^*(x)}{h(x)}, c^{-k} \frac{\sigma_u^2(x)}{h(x)} \int K(z)^2 dz\right) \quad (2.2.33)$$

in distribution.

Corollary 2.2.4. *Let the conditions of Theorem 2.2.2 hold. Let $\hat{g}_1(x)$ be the kernel regression estimator with window width*

$$\gamma_n = cn^{-1/(2m+k)}$$

and let $\hat{g}_2(x)$ be the kernel regression estimator with window width

$$\gamma_n = cn^{-\delta/(2m+k)} \quad \text{with } \delta \in (0, 1)$$

Denote

$$\hat{\hat{g}}(x) = \frac{\hat{g}_1(x) - n^{-(1-\delta)m/(2m+k)} \hat{g}_2(x)}{1 - n^{-(1-\delta)m/(2m+k)}} \quad (2.2.34)$$

Then

$$n^{m/(2m+k)}[\hat{\hat{g}}(x) - g(x)] \rightarrow N\left(0, c^{-k} \frac{\sigma_u^2(x)}{h(x)} \int K(z)^2 dz\right) \quad (2.2.35)$$

As is well known, the usual asymptotic normality results in parametric regression analysis hold with a rate of convergence in distribution equal to the square root of the number of observations. Now we see that in the nonparametric regression case this rate can be approached arbitrarily close by increasing m .

In Singh (1981), examples are given of members of the class $\mathcal{K}_{1,m}$ for $m = 3, 4, 5, 6$. However, a simple way to construct kernels in $\mathcal{K}_{k,m}$ for arbitrary $k \geq 1$ and even $m \geq 2$ is the following. For $x \in R^k$ and $N > 1$ let

$$K(x) = \sum_{j=1}^N \frac{\theta_j \exp(-\frac{1}{2}x' \Omega^{-1}x / \sigma_j^2)}{(\sqrt{2\pi})^k |\sigma_j|^k \sqrt{\det(\Omega)}} \quad (2.2.36)$$

where Ω is a positive definite matrix and the θ_j and σ_j are such that

$$\sum_{j=1}^N \theta_j = 1 \quad (2.2.37)$$

$$\sum_{j=1}^N \theta_j \sigma_j^{2\ell} = 0 \quad \text{for } \ell = 1, 2, \dots, N-1 \quad (2.2.38)$$

Then it is not hard to verify that $K \in \mathcal{K}_{k,m}$ with $m = 2N$.

The choice of the θ_j and σ_j affects the asymptotic variance of the estimator \hat{g} via the quantity

$$\int K(x)^2 dx = \sum_{i=1}^N \sum_{j=1}^N \theta_i \theta_j \sqrt{\sigma_i^2 + \sigma_j^2} (\sqrt{2\pi})^k \sqrt{\det(\Omega)} \quad (2.2.39)$$

Thus, at first sight, one might think of choosing the θ_j and σ_j so as to minimize (2.2.39) given Ω . However, (2.2.39) can be made arbitrarily small, for if $\theta_1, \dots, \theta_N, \sigma_1, \dots, \sigma_N$ is a solution of (2.2.37) and (2.2.38), then so is $\theta_1, \dots, \theta_N, \lambda\sigma_1, \dots, \lambda\sigma_N$ for any $\lambda > 0$. Then (2.2.39) is proportional to λ^{-1} . This indicates that the choice of the θ_j and the σ_j is not crucial, as the constant of the window width (2.2.32) may take over the role of this λ .

Finally, we consider the multivariate limiting distribution of the kernel regression estimator in distinct points. Thus, let $x^{(1)}$ and $x^{(2)}$ be distinct points in R^k such that $h(x^{(1)}) > 0, h(x^{(2)}) > 0$. Then, similarly to (2.2.9),

$$\begin{aligned} & \text{cov}\{\sqrt{n\gamma_n^k} \hat{q}_1(x^{(1)}), \sqrt{n\gamma_n^k} \hat{q}_1(x^{(2)})\} \\ &= Eu_j^2 K[(x^{(1)} - x_j)/\gamma_n] K[(x^{(2)} - x_j)/\gamma_n] \gamma_n^{-k} \\ &= \int \sigma_u^2 (x^{(1)} - \gamma_n z) h(x^{(1)} - \gamma_n z) K(z) K\{[(x^{(2)} - x^{(1)})/\gamma_n] + z\} dz \rightarrow 0 \end{aligned} \quad (2.2.40)$$

by bounded convergence, for $K\{[(x^{(1)} - x^{(2)})/\gamma] + z\} \rightarrow 0$ as $\gamma \downarrow 0$. Using this result, it is easy to show that

$$\begin{aligned} & \sqrt{n\gamma_n^k} \begin{pmatrix} \hat{q}_1(x^{(1)}) \\ \hat{q}_1(x^{(2)}) \end{pmatrix} \\ & \rightarrow N_2 \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} [\sigma_u^2(x^{(1)})/h(x^{(1)})] \int K(z)^2 dz & 0 \\ 0 & [\sigma_u^2(x^{(2)})/h(x^{(2)})] \int K(z)^2 dz \end{pmatrix} \right] \end{aligned} \quad (2.2.41)$$

in distribution. Thus,

$$\sqrt{n\gamma_n^k} \hat{q}_1(x^{(1)}) \quad \text{and} \quad \sqrt{n\gamma_n^k} \hat{q}_1(x^{(2)})$$

are asymptotically independent, and so are

$$\sqrt{n\gamma_n^k} [\hat{g}(x^{(1)}) - g(x^{(1)})] \quad \text{and} \quad \sqrt{n\gamma_n^k} [\hat{g}(x^{(2)}) - g(x^{(2)})]$$

More generally we have:

Theorem 2.2.3. *Let the conditions of Theorem 2.2.1 or Theorem 2.2.2 be satisfied and let $x^{(1)}, \dots, x^{(M)}$ be distinct points in R^k with $h(x^{(\ell)}) > 0$ for $\ell = 1, 2, \dots, M$. Then the sequence*

$$\{\sqrt{n\gamma_n^k} [\hat{g}(x^{(\ell)}) - g(x^{(\ell)})]\} \quad \ell = 1, \dots, M$$

is asymptotically independent, and so is

$$\{\sqrt{n\gamma_n^k} [\hat{\hat{g}}(x^{(\ell)}) - g(x^{(\ell)})]\} \quad \ell = 1, \dots, M$$

2.3 Uniform consistency

The uniform consistency of the kernel regression estimator is proved by Nadaraya (1965, 1970), Devroye (1978), Schuster and Yakowitz (1979), and Bierens (1983a). The approach in the latter two papers is based on an idea of Parzen (1962), namely to use the Fourier transform of the kernel. Suppose that the kernel has an absolutely integrable Fourier transform; that is,

$$\int |\psi(t)| dt < \infty \quad \text{with } \psi(t) = \int \exp(it'x) K(x) dx \quad (2.3.1)$$

[Note that this condition is satisfied for kernels of the type (2.2.36).] Then by the well-known inversion formula for Fourier transforms the kernel can be written as

$$K(x) = \left(\frac{1}{2\pi}\right)^k \int \exp(-it'x) \psi(t) dt \quad (2.3.2)$$

Consequently, $\hat{g}(x)\hat{h}(x)$ can be written as

$$\begin{aligned} \hat{g}(x)\hat{h}(x) &= \frac{1}{n} \sum_{j=1}^n y_j \gamma_n^{-k} \left(\frac{1}{2\pi}\right)^k \int \exp\left(\frac{-it'(x-x_j)}{\gamma_n}\right) \psi(t) dt \\ &= \left(\frac{1}{2\pi}\right)^k \int \left[\frac{1}{n} \sum_{j=1}^n y_j \exp(it'x_j) \right] \exp(-it'x) \psi(\gamma_n t) dt \end{aligned} \quad (2.3.3)$$

Hence

$$\begin{aligned} E \sup_x |\hat{g}(x)\hat{h}(x) - E\hat{g}(x)\hat{h}(x)| \\ \leq \left(\frac{1}{2\pi}\right)^k \int E \left| \frac{1}{n} \sum_{j=1}^n [y_j \exp(it'x_j) - Ey_j \exp(it'x_j)] \right| |\psi(\gamma_n t)| dt \end{aligned} \quad (2.3.4)$$

Moreover, using the well-known equality $\exp(ia) = \cos(a) + i \sin(a)$ we see that, uniformly in t ,

$$\begin{aligned} E \left| \frac{1}{n} \sum_{j=1}^n [y_j \exp(it'x_j) - Ey_j \exp(it'x_j)] \right| \\ \leq \left\{ \text{var} \left[\frac{1}{n} \sum_{j=1}^n y_j \cos(t'x_j) \right] + \text{var} \left[\frac{1}{n} \sum_{j=1}^n y_j \sin(t'x_j) \right] \right\}^{1/2} \\ \leq \frac{\sqrt{Ey_j^2}}{\sqrt{n}} \end{aligned} \quad (2.3.5)$$

Combining (2.3.4) and (2.3.5) yields

$$\begin{aligned} E \sup_x |\hat{g}(x)\hat{h}(x) - E\hat{g}(x)\hat{h}(x)| &\leq \sqrt{Ey_j^2} \frac{1}{\sqrt{n}} \left(\frac{1}{2\pi}\right)^k \int |\psi(\gamma_n t)| dt \\ &= O\left(\frac{1}{\gamma_n^k \sqrt{n}}\right) \end{aligned} \quad (2.3.6)$$

Furthermore, if $h(x)$ and $g(x)h(x)$ belong to the class $\mathfrak{D}_{k,m}$ and K belongs to the class $\mathfrak{K}_{k,m}$, then it can be shown, similarly to the proof of Lemma 2 of Bierens (1983a), that

$$\limsup_{n \rightarrow \infty} \sup_x \gamma_n^{-m} |E\hat{g}(x)\hat{h}(x) - g(x)h(x)| < \infty \quad (2.3.7)$$

Changing y_j to 1 we see that similar results hold for $\hat{h}(x)$. It is now easy to verify:

Theorem 2.3.1. *Let Assumption 2.2.1 and the additional conditions (2.3.1), $h(x) \in \mathfrak{D}_{k,m}$, $g(x)h(x) \in \mathfrak{D}_{k,m}$, $K \in \mathfrak{K}_{k,m}$, hold, where $m \geq 2$. Let*

$$\delta \in (0, \sup_x h(x)] \quad (2.3.8)$$

be arbitrary. Then

$$\min(\gamma_n^k \sqrt{n}, \gamma_n^{-m}) \sup_{x \in \{x \in R^k: h(x) \geq \delta\}} |\hat{g}(x) - g(x)| \quad (2.3.9)$$

is stochastically bounded.

Clearly, the best uniform consistency rate is obtained for γ_n such that $\min(\gamma_n^k \sqrt{n}, \gamma_n^{-m})$ is maximal. This is the case if

$$\gamma_n \propto n^{-1/(2m+2k)} \quad (2.3.10)$$

We then have

$$\min(\gamma_n^k \sqrt{n}, \gamma_n^{-m}) \propto n^{m/(2m+2k)} \quad (2.3.11)$$

It should be noted that the rate (2.3.11) is not the maximum obtainable rate, as is shown by Silverman (1978) for the density case and Révész (1979), Schuster and Yakowitz (1979), Liero (1982), and Cheng (1983) for the regression case. The present conservative approach has been chosen for its simplicity by which it can easily be extended to the case with partly discrete data and/or time series data. Compare Bierens (1983a).

3 The i.i.d. discrete case and the mixed continuous-discrete case

3.1 The discrete case

Economic data quite often contain qualitative variables. A typical feature of such variables is that they take a countable number of values and can usually be rescaled to integer-valued variables. Therefore, we consider now the case that all the components of x_j are of qualitative nature.

In the next section we show what happens in the mixed continuous-discrete case.

The following assumption formalizes the discrete nature of x_j .

Assumption 3.1.1. There exists a countable subset X of R^k such that

- (I) $x \in X$ implies $p(x) = P(x_j = x) > 0$;
- (II) $\sum_{x \in X} p(x) = 1$; and
- (III) every bounded subset of X is finite.

Part (III) of this assumption excludes limit points in X . It ensures that, for every $x \in X$,

$$\inf_{z \in X - \{x\}} |z - x| = \mu(x) > 0 \quad (3.1.1)$$

Now let the kernel and the window width be such that

$$K(0) = 1, \gamma_n \downarrow 0, \sqrt{n} \sup_{|z| > \lambda/\gamma_n} |K(z)| \rightarrow 0 \text{ for every } \lambda > 0 \quad (3.1.2)$$

This condition holds for kernels of the type (2.2.36) and window widths of the type $\gamma_n \propto n^{-\tau}$ with $\tau > 0$.

Since now

$$\begin{aligned} & \left| \frac{1}{\sqrt{n}} \sum_{j=1}^n y_j K \left[\frac{x - x_j}{\gamma_n} \right] - \frac{1}{\sqrt{n}} \sum_j y_j I(x_j = x) \right| \\ & \leq \frac{1}{\sqrt{n}} \sum_{j=1}^n \left| y_j K \left[\frac{x - x_j}{\gamma_n} \right] \right| I(x_j \neq x) \\ & \leq \frac{1}{n} \sum_{j=1}^n |y_j| \sqrt{n} \sup_{|z| > \mu(x)/\gamma_n} |K(z)| \rightarrow 0 \end{aligned} \quad (3.1.3)$$

in probability, where $I(\cdot)$ is the indicator function, and similarly

$$\left| \frac{1}{\sqrt{n}} \sum_{j=1}^n K \left[\frac{x - x_j}{\gamma_n} \right] - \frac{1}{\sqrt{n}} \sum_{j=1}^n I(x_j = x) \right| \rightarrow 0 \quad (3.1.4)$$

in probability, it is easy to verify that, for every $x \in X$,

$$p \lim_{n \rightarrow \infty} \sqrt{n} [\hat{g}(x) - \hat{g}^*(x)] = 0 \quad (3.1.5)$$

where

$$\begin{aligned} \hat{g}^*(x) &= \frac{\sum_{j=1}^n y_j I(x_j = x)}{\sum_{j=1}^n I(x_j = x)} \\ &= \left[\frac{\sum_{j=1}^n u_j I(x_j = x)}{\sum_{j=1}^n I(x_j = x)} \right] + g(x) \end{aligned} \quad (3.1.6)$$

It follows straightforwardly from the law of large numbers that

$$\hat{p}^*(x) = \frac{1}{n} \sum_{j=1}^n I(x_j = x) \rightarrow p(x) \quad (3.1.7)$$

in probability, whereas by the central limit theorem

$$\begin{aligned} \sqrt{n}[\hat{g}^*(x)\hat{p}^*(x) - g(x)\hat{p}^*(x)] &= \frac{1}{\sqrt{n}} \sum_{j=1}^n u_j I(x_j = x) \\ &\rightarrow N(0, \sigma_u^2(x)p(x)) \end{aligned} \quad (3.1.8)$$

in distribution. Combining (3.1.5), (3.1.7), and (3.1.8) yields:

Theorem 3.3.1. *Under Assumption 3.1.1 and condition (3.1.2) we have*

$$\sqrt{n}[\hat{g}(x) - g(x)] \rightarrow N(0, \sigma_u^2(x)/p(x)) \quad (3.1.9)$$

in distribution.

Also, similarly to Theorem 2.2.3, we have:

Theorem 3.3.2. *Let $x^{(1)}, \dots, x^{(M)}$ be distinct points in X . Under the conditions of Theorem 3.3.1 the sequence*

$$\{\sqrt{n}[\hat{g}(x^{(\ell)}) - g(x^{(\ell)})]\} \quad \ell = 1, \dots, M$$

is asymptotically independent.

Note that the discrete case differs from the continuous case in that hardly any restrictions are placed on the window width, whereas, nevertheless, the asymptotic normal distribution has zero mean. Moreover, the asymptotic rate of convergence in distribution is now the same as for the usual parametric models. Furthermore, since every bounded subset X_* of X is finite, Theorem 3.3.1 implies that

$$\max_{x \in X_*} |\sqrt{n}[\hat{g}(x) - g(x)]| \quad (3.1.10)$$

is stochastically bounded.

3.2 The mixed continuous-discrete case

We now consider the case where the first k_1 components of x_j are continuous and the remaining k_2 components are discrete. Of course, this case is only relevant if $k = k_1 + k_2 \geq 2$.

Assumption 3.2.1. Let $x_j = (x_j^{(1)}, x_j^{(2)})' \in X_1 \times X_2$, where X_1 is a k_1 -dimensional real space and X_2 is a subset of a k_2 -dimensional real space. The set X_2 is such that

- (I) $x^{(2)} \in X_2$ implies $p(x^{(2)}) = P(x_j^{(2)} = x^{(2)}) > 0$;
 (II) $\sum_{x^{(2)} \in X_2} p(x^{(2)}) = 1$; and
 (III) every bounded subset of X_2 is finite.

Let $x = (x^{(1)}, x^{(2)})' \in X_1 \times X_2$ and let $h(x^{(1)} | x^{(2)})$ be the density of the conditional distribution of $x_j^{(1)}$ given the event $x_j^{(2)} = x^{(2)}$. For every fixed $x^{(2)} \in X_2$ the following hold:

- (IV) $h(x^{(1)} | x^{(2)})$ and $g(x^{(1)}, x^{(2)})h(x^{(1)} | x^{(2)})$ belong to the class $\mathfrak{D}_{k_1, m}$ with $m \geq 2$;
 (V) there exists a $\delta > 0$ such that $\sigma_u^{2+\delta}(x^{(1)}, x^{(2)})h(x^{(1)} | x^{(2)})$ is uniformly bounded on X_1 ; and
 (VI) the functions
 $g(x^{(1)}, x^{(2)})^2 h(x^{(1)} | x^{(2)})$ and $\sigma_u^2(x^{(1)}, x^{(2)})h(x^{(1)} | x^{(2)})$
 are continuous and uniformly bounded on X_1 .

Moreover, we now choose the kernel $K(x^{(1)}, x^{(2)})$ and the window width γ_n such that, with $(z_1, z_2)' \in X_1 \times X_2$ and for $n \rightarrow \infty$,

$$\begin{aligned} \gamma_n \downarrow 0 \quad \sqrt{n} \sup_{|z_2| > \lambda/\gamma_n} \int |K(z_1, z_2)| dz_1 &\rightarrow 0 \quad \text{for every } \lambda > 0 \\ n\gamma_n^{k_1} \rightarrow \infty \quad K(z_1, 0) \in \mathfrak{K}_{k_1, m} \quad \text{with } m \geq 2 \\ \int |K(z_1, 0)| dz_1 < \infty \quad \int K(z_1, 0) dz_1 = 1 \end{aligned} \quad (3.2.1)$$

Denoting

$$h(x) = h(x^{(1)}, x^{(2)}) = h(x^{(1)} | x^{(2)})p(x^{(2)}) \quad (3.2.2)$$

we now have:

Theorem 3.2.1. *Under Assumption 3.2.1 and condition (3.2.1) the conclusions of Theorems 2.2.2, 2.2.3, and 2.3.1 and Corollaries 2.2.3 and 2.3.3 carry over with k replaced by k_1 and $\int K(z)^2 dz$ replaced by $\int K(z_1, 0)^2 dz_1$.*

This theorem can be proved by combining the arguments in Sections 2 and 3.1. The proof is somewhat cumbersome but does not involve insurmountable difficulties. It is therefore left to the reader.

4 Time series

4.1 Preliminaries

Recently the kernel regression approach has been extended to time series. Robinson (1983) and Singh and Ullah (1985) show strong consistency and

asymptotic normality using the α -mixing concept. Bierens (1983a) proves uniform consistency under ν stability in L^2 with respect to a ϕ -mixing base. Collomb (1985a) proves uniform strong consistency under the ϕ -mixing condition. Georgiev (1984a) proves consistency in the case of a Markov data-generating process. The conditions in the first four papers are reminiscent of those in Bierens (1982a, 1984) and White and Domowitz (1984) for nonlinear parametric time series regressions.

In this section we shall extend the results of Bierens (1983a) to asymptotic normality under similar conditions. Thus, we shall employ the ν -stability and mixing concepts. These concepts are defined as follows.

Let (z_t) be a stochastic process in R^k with the structure

$$z_t = \Psi_t(w_t, w_{t-1}, w_{t-2}, \dots) \quad (4.1.1)$$

where (w_t) is a stochastic process in a Euclidean space W and the Ψ_t are Borel-measurable mappings from the space of one-sided infinite sequences in W into R^k . Let $E|z_t|^r < \infty$ for some $r > 0$, and for $m = 0, 1, 2, \dots$, let

$$\nu(m) = \sup_t E|z_t - E(z_t | w_t, w_{t-1}, w_{t-2}, \dots, w_{t-m})|^r \quad (4.1.2)$$

Definition 4.1.1. The process (z_t) is said to be ν stable in L^r with respect to the base (w_t) if $\lim_{m \rightarrow \infty} \nu(m) = 0$.

Next, let $\mathcal{F}_{-\infty, t}$ be the Borel field generated by $w_t, w_{t-1}, w_{t-2}, \dots$, and let $\mathcal{F}_{t, \infty}$ be the Borel field generated by $w_t, w_{t+1}, w_{t+2}, \dots$. Denote, for $m = 0, 1, 2, 3, \dots$,

$$\alpha(m) = \sup_t \sup_{E_1 \in \mathcal{F}_{-\infty, t-m}, E_2 \in \mathcal{F}_{t, \infty}} |P(E_1 \text{ and } E_2) - P(E_1)P(E_2)| \quad (4.1.3)$$

$$\phi(m) = \sup_t \sup_{E_1 \in \mathcal{F}_{-\infty, t-m}, E_2 \in \mathcal{F}_{t, \infty}, P(E_1) > 0} |P(E_2 | E_1) - P(E_2)| \quad (4.1.4)$$

Definition 4.1.2. The process (w_t) is said to be α -mixing if $\lim_{m \rightarrow \infty} \alpha(m) = 0$ and ϕ -mixing if $\lim_{m \rightarrow \infty} \phi(m) = 0$.

A further discussion of these concepts can be found in Bierens (1983a, 1984) and White and Domowitz (1984). See also Bierens (1981, Chapter 5, 1982a, 1984) for a discussion of the related stochastic stability concept.

We now assume:

Assumption 4.1.1. The data-generating process $\{(y_t, x_t)\}$ is a strictly stationary ν -stable process in L^2 with respect to a strictly stationary ϕ -mixing base (w_t) , where

$$\nu(m) = O[\exp(-cm)] \quad \text{for some } c > 0 \quad (4.1.5)$$

$$\sum_{m=0}^{\infty} \phi(m)^{1/2} < \infty \quad (4.1.6)$$

This assumption holds, for example, if the data-generating process is an ARMA process with the usual properties. Compare Bierens (1983a, 1984).

Moreover, according to Bierens (1984), we assume that $g(x_t)$ represents the conditional expectation of y_t given the *entire* past of the data-generating process:

Assumption 4.1.2. Let

$$g(x_t) = E(y_t | x_t, x_{t-1}, x_{t-2}, \dots, y_{t-1}, y_{t-2}, \dots) \quad \text{w.p.1} \quad (4.1.7)$$

Moreover, the error $u_t = y_t - g(x_t)$ is one of the components of w_t .

The latter condition in Assumption 4.1.2 is not strictly necessary but eases the argument. Moreover, we note that Robinson (1983) assumes only $E(y_t | x_t) = g(x_t)$, which is weaker than (4.1.7). The advantage of assumption (4.1.7) is, however, that the errors u_t are now martingale differences, by which the following Lyapunov version of the martingale central limit theorem is applicable.

Lemma 4.1.1. *Let $(v_{n,j})$, $j = 1, \dots, n$, $n = 1, 2, \dots$, be a double array of martingale differences; that is,*

$$E(v_{n,j} | v_{n,j-1}, v_{n,j-2}, \dots, v_{n,1}) = 0 \quad \text{w.p.1} \quad (4.1.8)$$

for $j = 2, 3, \dots, n$ and $n = 2, 3, \dots$. If

$$p \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{j=1}^n v_{n,j}^2 = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{j=1}^n E v_{n,j}^2 = \sigma^2 \in (0, \infty) \quad (4.1.9)$$

and

$$\lim_{n \rightarrow \infty} \sum_{j=1}^n E \left| \frac{v_{n,j}}{\sqrt{n}} \right|^{2+\delta} = 0 \quad \text{for some } \delta > 0 \quad (4.1.10)$$

then

$$\frac{1}{\sqrt{n}} \sum_{j=1}^n v_{n,j} \rightarrow N(0, \sigma^2) \quad (4.1.11)$$

in distribution.

Proof. The lemma follows straightforwardly from Theorem 2.3 of McLeish (1974). ■

Furthermore, we also need the following lemma.

Lemma 4.1.2. Let $\{(z_j, x_j)\}$ be a strictly stationary stochastic process in $R \times R^k$, with

$$Ez_j^4 < \infty \quad \text{and} \quad E|x_j|^2 < \infty \quad (4.1.12)$$

Let this process be ν stable in L^2 with respect to a strictly stationary ϕ -mixing base, where ν satisfies condition (4.1.5) and ϕ satisfies condition (4.1.6). Let K be a Borel-measurable real function on R^k such that, for $\ell = 1, 2$,

$$\int |K(x)|^\ell dx < \infty \quad \int |t\psi_\ell(t)| dt < \infty \quad (4.1.13)$$

where $\psi_\ell(t) = \int \exp(it'x) |K(x)|^\ell dx$

Denote, for $x \in R^k$,

$$d_n(x) = \text{var} \left\{ \frac{1}{n} \sum_{j=1}^n z_j K \left[\frac{x-x_j}{\gamma_n} \right] \right\} - \frac{1}{n^2} \sum_{j=1}^n \text{var} \left\{ z_j K \left[\frac{x-x_j}{\gamma_n} \right] \right\} \quad (4.1.14)$$

where $\gamma_n > 0$, $\gamma_n \downarrow 0$ as $n \rightarrow \infty$. Then

$$d_n(x) = O \left[\left(\ln \frac{n}{\gamma_n} + \ln \frac{1}{Ez_0^2 K[(x-x_0)/\gamma_n]^2} \right) \times \frac{Ez_0^2 K[(x-x_0)/\gamma_n]^2}{n} \right] \quad (4.1.15)$$

Proof. See Appendix. ■

Finally, we need the following additional assumption.

Assumption 4.1.3

- (I) If Assumption 2.2.1 holds, then in addition:
- (a) $\sigma_u^4(x)h(x)$ is uniformly bounded;
 - (b) $g(x)^2h(x)$ has continuous and bounded second derivatives.
- (II) If Assumption 3.2.1 holds, then for every fixed $x^{(2)} \in X_2$:
- (a) $\sigma_u^4(x^{(1)}, x^{(2)})h(x^{(1)} | x^{(2)})$ is uniformly bounded on X_1 ;
 - (b) $g(x^{(1)}, x^{(2)})^2h(x^{(1)} | x^{(2)})$ has continuous and bounded second derivatives with respect to the components of $x^{(1)}$.

4.2 Results

Using Lemmas 4.1.1 and 4.1.2, the previous asymptotic results can be extended to the time series case.

Theorem 4.2.1

- (I) With Assumption 4.1.1 the previous uniform consistency results go through.
 (II) Let

$$Eu_j^8 < \infty \quad \text{and} \quad Eg(x_j)^4 < \infty \quad (4.2.1)$$

and let the kernel K be such that, for $\ell = 1, 2, 3, 4$,

$$\int |\psi_\ell(t)| dt < \infty \quad \text{where} \quad \psi_\ell(t) = \int \exp(it'x) |K(x)|^\ell dx \quad (4.2.2)$$

Moreover,

let $\int zz'K(z)^2 dz$ be finite in the continuous case,
 and let $\int z_1 z_1' K(z_1, 0)^2 dz$ be finite in the mixed
 continuous-discrete case, respectively. (4.2.3)

With Assumptions 4.1.1, 4.1.2, and 4.1.3 and conditions (4.2.1), (4.2.2), and (4.2.3), the previous asymptotic normality results go through.

Note that conditions (4.2.2) and (4.2.3) hold for kernels of the type (2.2.36).

To save space, we shall prove only part II of Theorem 4.2.1 for the continuous case. The reader is invited to prove the discrete and mixed continuous-discrete cases along the same lines. For the proof of part I we refer to Bierens (1983a). So we now have to show that (2.2.2) and (2.2.4) go through and that $\hat{h}(x)$ remains pointwise consistent for the time series under review, as only in these steps was the independence assumption involved.

Proof of (2.2.2). Since now the $v_{n,j}(x)$ defined by (2.2.7) are martingale differences, it suffices to show that

$$p \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{j=1}^n [v_{n,j}(x)^2 - Ev_{n,j}(x)^2] = 0 \quad (4.2.4)$$

as then (2.2.2) follows from Lemma 4.1.1. Thus, consider Lemma 4.1.2 with $z_j = u_j^2$ and $K(x)$ replaced by $K(x)^2$. Since

$$\begin{aligned} Eu_j^4 K \left[\frac{x - x_j}{\gamma_n} \right]^4 &= \gamma_n^k \int \sigma_u^4(x - \gamma_n z) h(x - \gamma_n z) K(z)^4 dz \\ &= O(\gamma_n^k) \end{aligned} \quad (4.2.5)$$

it follows from Lemma 4.1.2 that

$$\begin{aligned}
& \text{var} \left(\frac{1}{n} \sum_{j=1}^n \frac{u_j^2 K[(x-x_j)/\gamma_n]^2}{\gamma_n^k} \right) \\
&= \frac{1}{n^2} \sum_{j=1}^n \text{var} \left(\frac{u_j^2 K[(x-x_j)/\gamma_n]^2}{\gamma_n^k} \right) + \gamma_n^{-2k} d_n(x) \\
&= O \left(\frac{1}{n\gamma_n^k} \right) + O \left(\frac{1}{n\gamma_n^k} \ln \frac{n}{\gamma_n^{k+1}} \right) \rightarrow 0 \quad \text{as } n \rightarrow \infty \quad (4.2.6)
\end{aligned}$$

provided $\gamma_n \propto n^{-\tau}$ with $\tau < 1/k$. Therefore, (4.2.4) follows from (4.2.6) and Chebishev's inequality. ■

Proof of (2.2.4). Let z_j in Lemma 4.1.2 be

$$z_j = y_j - u_j - g(x) = g(x_j) - g(x) \quad (4.2.7)$$

Then

$$\begin{aligned}
& \text{var} \left(\sqrt{n\gamma_n^k} \frac{1}{n} \sum_{j=1}^n \frac{[g(x_j) - g(x)] K[(x-x_j)/\gamma_n]}{\gamma_n^k} \right) \\
&= \frac{1}{n^2} \sum_{j=1}^n \text{var} \left(\sqrt{n\gamma_n^k} \frac{[g(x_j) - g(x)] K[(x-x_j)/\gamma_n]}{\gamma_n^k} \right) \\
&= \frac{n}{\gamma_n^k} d_n(x) = O \left(\gamma_n^2 \ln \frac{n}{\gamma_n^{k+3}} \right) + o(1) \quad (4.2.8)
\end{aligned}$$

where the last conclusion follows from the fact that by Assumption 4.1.3 and Taylor's theorem

$$\begin{aligned}
& E[g(x_j) - g(x)]^2 K \left[\frac{x-x_j}{\gamma_n} \right]^2 \\
&= \gamma_n^k \int [g(x - \gamma_n z) - g(x)]^2 h(x - \gamma_n z) K(z)^2 dz \\
&\cong \gamma_n^{k+2} \int \left\{ z' \frac{\partial}{\partial x'} [g(x)^2 h(x)] \right\}^2 K(z)^2 dz = O(\gamma_n^{k+2}) \quad (4.2.9)
\end{aligned}$$

Since $\gamma_n^2 \ln(n/\gamma_n^{k+3}) \rightarrow 0$ as $n \rightarrow \infty$ if $\gamma_n \propto n^{-\tau}$ with $\tau > 0$, (2.2.4) follows. ■

The pointwise consistency of $\hat{h}(x)$ under the conditions under review follows from Bierens (1983a). So the proof of part II of Theorem 4.2.1 for the continuous case is now complete.

Finally, we note that the ϕ -mixing condition on the base can be relaxed to a similar α -mixing condition but at the expense of stronger conditions on the moments of y_j and x_j . This follows from Lemma 2.1 of McLeish (1975). The present approach has been chosen in order to keep our argument in tune with the approach in Bierens (1983a).

5 How to choose the kernel and the window width

5.1 The choice of the kernel

In the literature on kernel density and regression function estimation the problem of how the kernel should be specified has mainly been considered from an asymptotic point of view. In the case of density estimation Epanechnikov (1969) has shown that the kernel that minimizes the integrated mean squared error

$$\int [\hat{h}(x) - h(x)]^2 dx \quad (5.1.1)$$

over the class of product kernels

$$K(x) = K(x^{(1)}, x^{(2)}, \dots, x^{(k)}) = \prod_{i=1}^k K_0(x^{(i)}) \quad x^{(i)} \in R \quad (5.1.2)$$

with

$$\begin{aligned} K_0(v) = K_0(-v) \geq 0 & \quad \int K_0(v) dv = \int v^2 K_0(v) dv = 1 \\ \int v K_0(v) dv = 0 & \end{aligned} \quad (5.1.3)$$

is a product kernel with

$$K_0(v) = \begin{cases} 3/4\sqrt{5} - 3v^2/20\sqrt{5} & \text{if } |v| \leq \sqrt{5} \\ 0 & \text{if } |v| > \sqrt{5} \end{cases} \quad (5.1.4)$$

Note that Epanechnikov's kernel K_0 is the solution of the problem:

$$\text{minimize } \int K_0(v)^2 dv \quad \text{subject to (5.1.3)} \quad (5.1.5)$$

Greblicki and Krzyzak (1980) have confirmed this result for the regression case. Epanechnikov also shows there are various kernels that are nearly optimal. For example, the standard normal density satisfies conditions (5.1.3) and is almost optimal, as

$$\int \left[\frac{\exp(-\frac{1}{2}v^2)}{\sqrt{2\pi}} \right]^2 dv = 1.051 \int K_0(v)^2 dv \quad (5.1.6)$$

A disadvantage of Epanechnikov's kernel is that its Fourier transform is not absolutely integrable, a condition quite often employed in this chapter. Also, the nonnegativity of K_0 implies that the kernel (5.1.2) with (5.1.4) merely satisfies $K \in \mathcal{K}_{k,2}$, whereas higher rates of convergence in distribution than $n^{2/(k+4)}$ require $K \in \mathcal{K}_{k,m}$ with $m > 2$. Compare Theorem 2.2.2.

Since kernels of the type (2.2.36) have all the required properties, are almost arbitrarily flexible, and can easily be constructed, we advocate the use of that type of kernel. However, the question now arises how the matrix Ω should be specified. A heuristic approach to solve this problem is to specify Ω such that certain properties of the true regression function carry over to the estimate \hat{g} . The property we shall consider is the *linear translation invariance principle*. Suppose we apply a linear translation to x and the x_j 's:

$$x^* = Px + q \quad x_j^* = Px_j + q \quad (5.1.7)$$

where P is a nonsingular $k \times k$ matrix and q is a k -component vector. Then

$$g(x) = E(y_j | x_j = x) = E(y_j | x_j^* = x^*) = g^*(x^*) \quad (5.1.8)$$

say. However, if we replace the x_j and x in (2.1.16) by x_j^* and x^* , respectively, and if we leave the kernel K unchanged, then the resulting kernel regression estimator $\hat{g}^*(x^*)$, say, will in general be unequal to $\hat{g}(x)$, for

$$\hat{g}^*(x^*) = \sum_{j=1}^n y_j K \left[\frac{P(x-x_j)}{\gamma_n} \right] \bigg/ \sum_{j=1}^n K \left[\frac{P(x-x_j)}{\gamma_n} \right] \neq \hat{g}(x) \quad \text{if } P \neq I \quad (5.1.9)$$

The only way to accomplish $\hat{g}^*(x^*) = \hat{g}(x)$ in all cases (5.1.7) is to let the kernel be of the form

$$\hat{K}(x) = \eta(x' \hat{V}^{-1} x) \quad (5.1.10)$$

where η is a real function on R and \hat{V} is the sample variance matrix; that is,

$$\hat{V} = \frac{1}{n} \sum_{j=1}^n (x_j - \bar{x})(x_j - \bar{x})' \quad \text{with } \bar{x} = \frac{1}{n} \sum_{j=1}^n x_j \quad (5.1.11)$$

In particular, if we use kernels of the form (2.2.36), then we should specify $\Omega = \hat{V}$. Thus, for $m = 2, 4, 6, \dots$, we let

$$\hat{K}_m(x) = \sum_{j=1}^{m/2} \frac{\theta_j \exp[-\frac{1}{2} x' \hat{V}^{-1} x / \sigma_j^2]}{(\sqrt{2\pi})^k |\sigma_j|^k \sqrt{\det(\hat{V})}} \quad (5.1.12a)$$

in the continuous case,

$$\hat{K}_m(x) = \sum_{j=1}^{m/2} \frac{\theta_j \exp[-\frac{1}{2} x' \hat{V}^{-1} x / \sigma_j^2]}{(\sqrt{2\pi})^{k_1} |\sigma_j|^{k_1} \sqrt{\det[(\hat{V}^{(1)})^{-1}]}} \quad (5.1.12b)$$

in the mixed continuous-discrete case (with the first k_1 components of x_j continuously distributed), and

$$\hat{K}_m(x) = \hat{K}_2(x) = \exp(-\frac{1}{2} x' \hat{V}^{-1} x) \quad (5.1.12c)$$

in the discrete case, where $\hat{V}^{(1)}$ is the upper left $k_1 \times k_1$ submatrix of \hat{V}^{-1} , and

$$\sum_{j=1}^{m/2} \theta_j \sigma_j^{2\ell} = \begin{cases} 1 & \text{if } \ell = 0 \\ 0 & \text{if } \ell = 1, 2, \dots, (m/2) - 1 \end{cases} \quad (5.1.13)$$

The question now arises whether the previous asymptotic results go through for kernel regression estimators with this kernel. The answer is yes, provided the following additional conditions hold.

Assumption 5.1.1. Let $E|x_j|^4 < \infty$ and let the matrix

$$V = Ex_j x_j' - (Ex_j)(Ex_j)' \quad (5.1.14)$$

be nonsingular. Moreover, let (x_j) be ν_* stable in L^4 with respect to the base (w_j) considered in Assumption 4.1.1, with

$$\nu_*(m) = O(\exp(-c_* m)) \quad \text{for some } c_* > 0 \quad (5.1.15)$$

Denoting

$$K_m(x) = \sum_{j=1}^{m/2} \frac{\theta_j \exp[-\frac{1}{2}x'V^{-1}x/\sigma_j^2]}{(\sqrt{2\pi})^k |\sigma_j|^k \sqrt{\det(V)}} \quad (5.1.16a)$$

in the continuous case,

$$K_m(x) = \sum_{j=1}^{m/2} \frac{\theta_j \exp[-\frac{1}{2}x'V^{-1}x/\sigma_j^2]}{(\sqrt{2\pi})^{k_1} |\sigma_j|^{k_1} \sqrt{\det[(V^{(1)})^{-1}]}} \quad (5.1.16b)$$

in the mixed continuous-discrete case, where $V^{(1)}$ is the upper left $k_1 \times k_1$ submatrix of V^{-1} and the θ_j and the σ_j are the same as before, and

$$K_m(x) = K_2(x) = \exp(-\frac{1}{2}x'V^{-1}x) \quad (5.1.16c)$$

in the discrete case, we can now state:

Theorem 5.1.1. *With Assumption 5.1.1 the kernel regression estimator with kernel (5.1.12) has the same asymptotic properties (as previously considered) as the kernel regression estimator with kernel (5.1.16).*

Proof. See Appendix. ■

Remark. The approach in this section is reminiscent of that in Bierens (1983b). In that paper the regression function estimate with kernel of the type \hat{K}_2 is derived from moment conditions on the kernel density estimator of the density of the regressors x_j and the density of the observations (y_j, x_j) . In other words, these moment conditions imply the linear translation invariance principle.

5.2 The choice of the window width

From the preceding results it is clear that the asymptotic performance of the kernel regression function estimator heavily depends on the choice of the window width γ_n . In particular, the asymptotic normality results in Corollaries 2.2.2 and 2.2.4 show that the variance of the limiting normal distribution of \hat{g} shrinks down to zero if we let the constant c of the window width parameters approach infinity. But that will destroy the small-sample performance of the kernel regression estimator. If we choose too large a γ_n , the Nadaraya-Watson regression function estimate will become too flat, for

$$\hat{g}(x) \rightarrow \bar{y} = \frac{1}{n} \sum_{j=1}^n y_j \quad \text{if } \gamma_n \rightarrow \infty \quad (5.2.1)$$

Similarly, $\hat{g}(x) \rightarrow \bar{y}$ if $c \rightarrow \infty$. On the other hand, if we choose too small a γ_n , the estimate \hat{g} will go wild. For example, if we employ in (2.1.16) the kernel \hat{K}_2 and if we let $\gamma_n \downarrow 0$, then

$$\hat{g}(x) \rightarrow \sum_{j=1}^n y_j I\{(x-x_j)' \hat{V}^{-1}(x-x_j) = \min_{\ell} (x-x_{\ell})' \hat{V}^{-1}(x-x_{\ell})\} \\ \left/ \sum_{j=1}^n I\{(x-x_j)' \hat{V}^{-1}(x-x_j) = \min_{\ell} (x-x_{\ell})' \hat{V}^{-1}(x-x_{\ell})\} \right. \\ (\ell = 1, \dots, n) \quad (5.2.2)$$

where $I(\cdot)$ is the indicator function. Thus, $\hat{g}(x)$ converges to the y_j for which $(x-x_j)' \hat{V}^{-1}(x-x_j)$ is minimal, so that then the estimate \hat{g} degenerates to an inconsistent nearest-neighbor estimate. Compare Stone (1977). Again, a similar result holds for \hat{g} if we let $c \rightarrow 0$.

A somewhat heuristic but well-working trick to optimizing the window width is the cross-validation approach introduced by Stone (1974), Geisser (1975), and Wahba and Wold (1975). See also Wong (1983). The basic idea is to split the data set in two parts. Then the first part is used for calculating the estimate and the second part is used for optimizing the fit of the estimate by minimizing the mean squared error. A variant used by Bierens (1983a) is to consider various partitions and to minimize the mean of the mean squared errors. In particular, let

$$\hat{g}_*^{(\ell)}(x | \gamma) = \sum_{j=1}^n y_j I(j \neq \ell) \hat{K}_m \left[\frac{x-x_j}{\gamma} \right] \left/ \sum_{j=1}^n I(j \neq \ell) \hat{K}_m \left[\frac{x-x_j}{\gamma} \right] \right. \quad (5.2.3)$$

and denote, similarly to (2.2.34),

$$\hat{g}_1^{(\ell)}(x | c) = \hat{g}_*^{(\ell)}(x | cn^{-1/(2m+k)}) \quad (5.2.4)$$

$$\hat{g}_2^{(\ell)}(x | c) = \hat{g}_*^{(\ell)}(x | cn^{-\delta/(2m+k)}) \quad (5.2.5)$$

$$\hat{g}^{(\ell)}(x | c) = \frac{\hat{g}_1^{(\ell)}(x | c) - n^{-(1-\delta)m/(2m+k)} \hat{g}_2^{(\ell)}(x | c)}{1 - n^{-(1-\delta)m/(2m+k)}} \quad (5.2.6)$$

Then $\hat{g}^{(\ell)}(x | c)$ is the regression function estimator of the type (2.2.34) with kernel \hat{K}_m , based on the data set leaving the observation with index ℓ out. We now propose to optimize c by minimizing

$$\hat{Q}(c) = \sum_{j=1}^n \{y_j - \hat{g}^{(j)}(x_j | c)\}^2 \quad (5.2.7)$$

to c in an interval $[c_1, c_2]$ with $0 < c_1 < c_2 < \infty$. Denoting the resulting optimal c by \hat{c} , that is,

$$\hat{Q}(\hat{c}) = \inf\{\hat{Q}(c) | c \in [c_1, c_2]\} \quad (5.2.8)$$

we then propose to use

$$\hat{g}(x | \hat{c}) = \hat{g}^{(0)}(x | \hat{c}) \quad (5.2.9)$$

as the cross-validated kernel regression function estimator.

Although this approach works well in practice, it has the disadvantage that we lose control over some of the asymptotic properties of kernel estimators. From Bierens (1983a) it follows that the cross-validated kernel regression estimator remains (uniformly) consistent, but it is not clear whether asymptotic normality goes through. We can regain some control over the asymptotic behavior of $\hat{g}(x | \hat{c})$ if instead of (5.2.8) we optimize c by finite grid search; that is,

$$\hat{Q}(\hat{c}) = \min\{\hat{Q}(c^{(\ell)}) | \ell = 1, 2, \dots, M\} \quad (5.2.10)$$

where $c_1 = c^{(1)} < c^{(2)} < \dots < c^{(M)} = c_2$ are grid points. It is not hard to show that in the continuous case the M -variate limiting distribution of

$$n^{m/(2m+k)}(\hat{g}(x | c^{(1)}) - g(x), \dots, \hat{g}(x | c^{(M)}) - g(x)) \quad (5.2.11)$$

is M -variate normal with zero mean vector; hence, for x with $h(x) > 0$ we have at least

$$n^{m/(2m+k)}[\hat{g}(x | \hat{c}) - g(x)] \quad (5.2.12)$$

is stochastically bounded. A similar result holds for the mixed continuous-discrete case. However, if for this \hat{c} , $p \lim_{n \rightarrow \infty} \hat{c} = c$, then asymptotic normality goes through as if $\hat{c} = c$. Moreover, in the discrete case the cross-validated regression estimator has the same properties as before without additional conditions.

5.3 A numerical example

We shall now demonstrate the performance of the kernel regression estimator on an artificial data set of size 100. This data set, $\{(y_1, x_1), \dots, (y_{100}, x_{100})\}$ with $(y_j, x_j) \in R \times R$, has been generated according to

$$y_j = x_j^2 + u_j \quad u_j \equiv \text{NID}(0, 1) \quad x_j \equiv \text{i.i.d. } \frac{1}{2}N(-2, 1) + \frac{1}{2}N(2, 1) \quad (5.3.1)$$

Thus, the x_j 's have been generated from a mixture of the normal distributions $N(-2, 1)$ and $N(2, 1)$ with equal weights, and the errors have been generated independently from the standard normal distribution. So the regression function g and the density h are

$$g(x) = x^2 \quad h(x) = \frac{\frac{1}{2} \exp[-\frac{1}{2}(x+2)^2] + \frac{1}{2} \exp[-\frac{1}{2}(x-2)^2]}{\sqrt{2\pi}} \quad x \in R \quad (5.3.2)$$

In Figures 3.1 and 3.2 the cross-validated kernel regression estimator $\hat{g}(x | \hat{c})$ has been fitted by grid search on the 20 grid points $0.1, 0.2, \dots, 2$, where $\delta = \frac{1}{2}$ and $m = 2$ (Figure 3.1) or $m = 4$ (Figure 3.2). Compare (5.2.6). In Figures 3.3 and 3.4 are displayed the estimate $\hat{g}(x | c)$ for $m = 2$ and $m = 4$, respectively, and c fixed on the value 1. In the case $m = 4$ the σ_j in (5.1.12) have been chosen $\sigma_j = \sqrt{j}$. Also, 95% confidence bands are calculated on the basis of the following consistent estimator of the asymptotic variance of $\hat{g}(x | c)$:

$$\hat{\sigma}^2(x | c) = c^{-k} \frac{(1/n) \sum_{j=1}^n [y_j - \hat{g}(x | c)]^2 \hat{K}_m[(x - x_j)/\gamma_n(c)]^2 / \gamma_n(c)^k}{\{(1/n) \sum_{j=1}^n \hat{K}_m[(x - x_j)/\gamma_n(c)] / \gamma_n(c)^k\}^2} \quad (5.3.3)$$

with

$$\gamma_n(c) = cn^{-1/(2m+k)} \quad (5.3.4)$$

It is not too hard to show that

$$p \lim_{n \rightarrow \infty} \hat{\sigma}^2(x | c) = c^{-k} \frac{\sigma_u^2(x)}{h(x)} \int K_m(x)^2 dz \quad (5.3.5)$$

[cf. (2.2.35)]; hence

$$n^{m/(2m+k)} \frac{\hat{g}(x | c) - g(x)}{\hat{\sigma}(x | c)} \rightarrow N(0, 1) \quad (5.3.6)$$

in distribution, provided c is fixed. In Figures 3.1 and 3.2 the c is not fixed but is determined by cross-validation, so strictly speaking, the 95% confidence bands involved are not valid. Only if the cross-validated estimate \hat{c} converges in probability to a constant c are these confidence bands

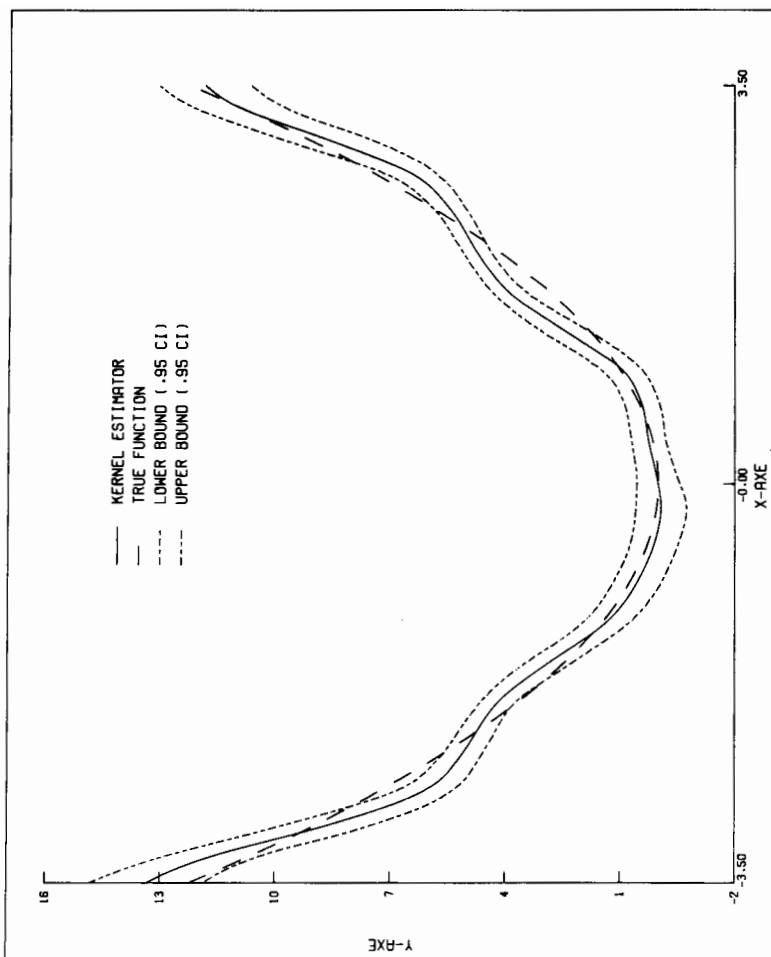


Figure 3.1

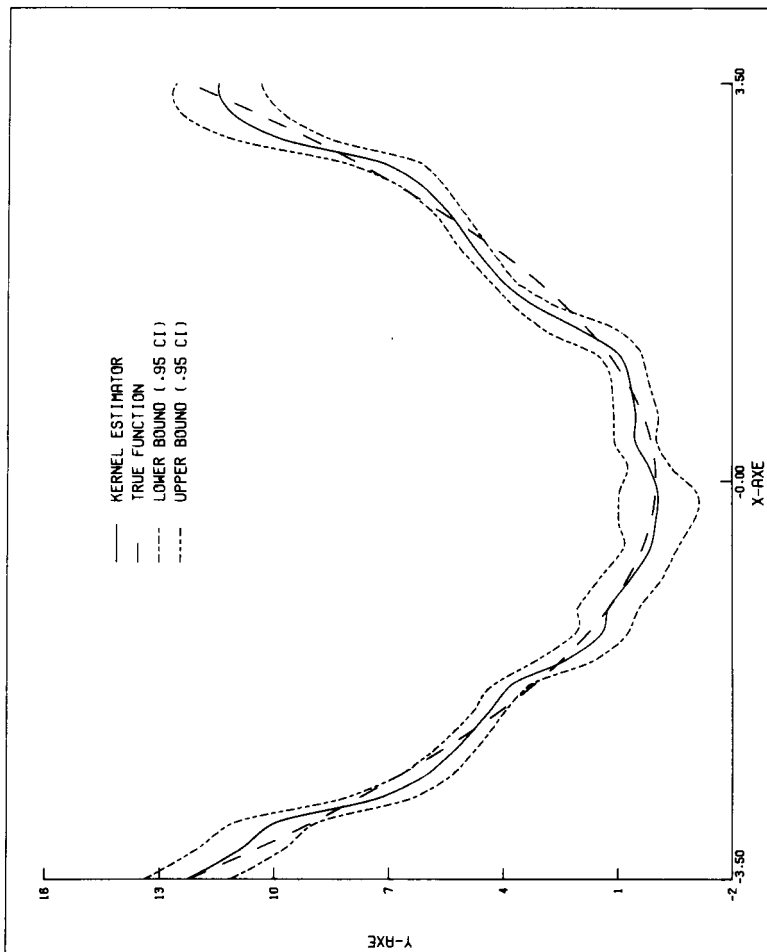


Figure 3.2

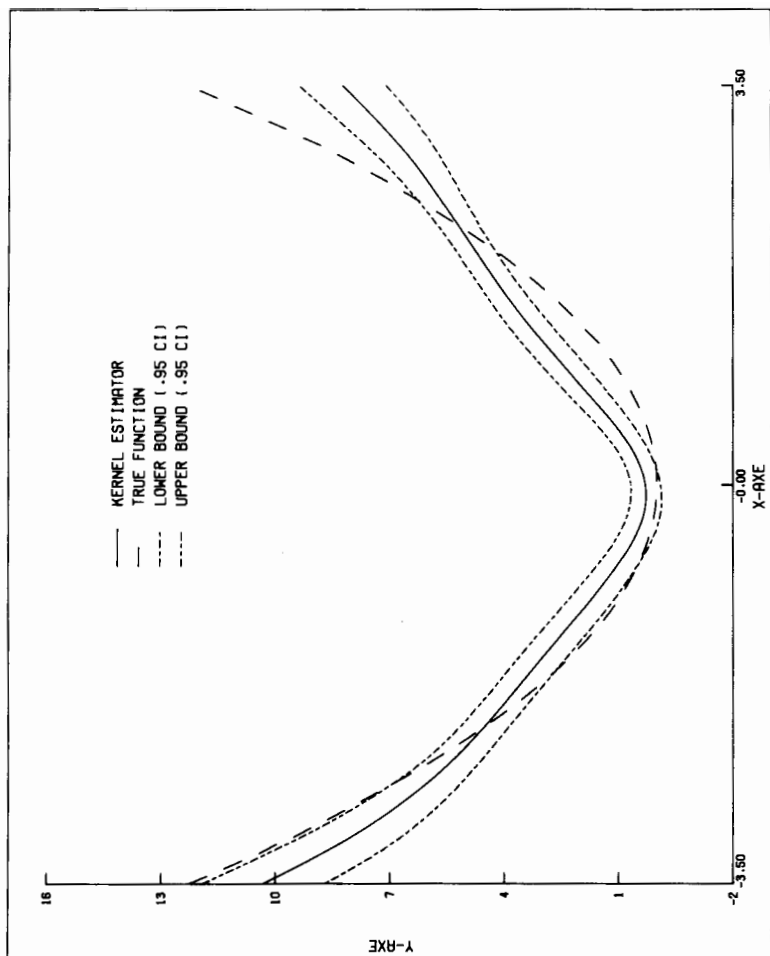


Figure 3.3

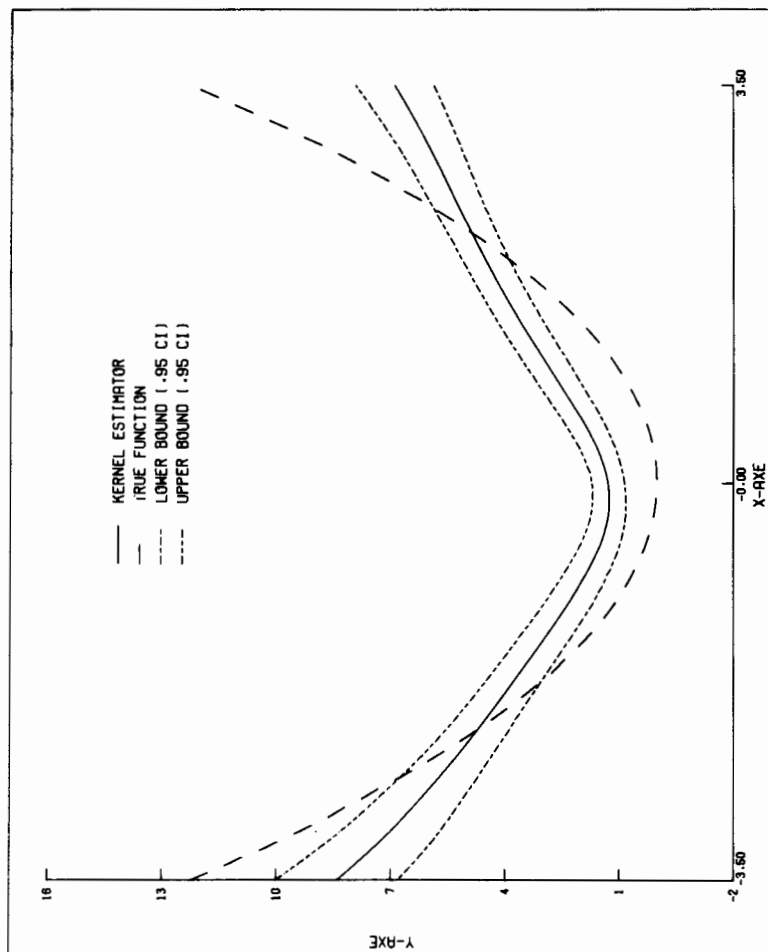


Figure 3.4

asymptotically correct. We emphasize that the confidence limits shown in the figures are pointwise limits and not uniform limits.

The cross-validated kernel estimates in Figures 3.1 and 3.2 are much closer to the true regression function than the kernel estimates with fixed $c = 1$ in Figures 3.3. and 3.4. In particular, the fact that in Figures 3.3 and 3.4 the true regression function lies too often outside the 95% confidence band indicates that with this small sample and fixed c the asymptotic normal approximation is of limited practical use. Therefore, cross-validation seems to be an essential step in kernel regression estimation. Also, the choice of the kernel type does not seem critical as long as cross-validation is applied. The estimates in Figures 3.1 and 3.2 look quite similar, and the estimate in Figure 3.1 is even somewhat smoother than in Figure 3.2, although the latter estimate is preferable from an asymptotic point of view. Figures 3.1 and 3.2 show that one should be careful to not interpret every bump on the estimated regression curve as structural, as the estimation errors manifest themselves as bumps on the estimated regression curve.

In the case of one or two regressors this sort of plot may serve as a tool in specifying an appropriate parametric functional form. Subsequently, model specification tests may be applied to check the correctness of the specified parametric functional form. In particular, the tests of Bierens (1982b, 1984) seem appropriate for that.

For models with more than two regressors the interpretation of the estimation results become difficult. Admittedly, this is a serious drawback of the kernel regression approach and any other nonparametric estimation method. Nevertheless, also for higher-dimensional models, the kernel regression approach may be useful, namely as an out-of-sample forecasting scheme, as for this application visual inspection of the estimation results is not needed.

Appendix

Proof of Lemma 4.1.2. We can write

$$\begin{aligned}
 d_n(x) &= 2 \frac{1}{n^2} \sum_{\ell=1}^{n-1} \sum_{j=1}^{n-\ell} \text{cov} \left\{ z_0 K \left[\frac{x-x_0}{\gamma_n} \right], z_j K \left[\frac{x-x_j}{\gamma_n} \right] \right\} \\
 &= 2 \frac{1}{n^2} \sum_{\ell=1}^{n-1} \sum_{j=1}^{n-\ell} E \left(z_0 z_j K \left[\frac{x-x_0}{\gamma_n} \right] K \left[\frac{x-x_j}{\gamma_n} \right] \right) \\
 &\quad - 2 \frac{1}{n^2} \sum_{\ell=1}^{n-1} \sum_{j=1}^{n-\ell} E \left(z_0 K \left[\frac{x-x_0}{\gamma_n} \right] \right)^2
 \end{aligned} \tag{A.1}$$

Similarly, let

$$\begin{aligned} d_n^{(m)}(x) &= 2 \frac{1}{n^2} \sum_{\ell=1}^{n-1} \sum_{j=1}^{n-\ell} \text{cov} \left\{ z_0 K \left[\frac{x-x_0^{(m)}}{\gamma_n} \right], z_j K \left[\frac{x-x_j^{(m)}}{\gamma_n} \right] \right\} \\ &= 2 \frac{1}{n^2} \sum_{\ell=1}^{n-1} \sum_{j=1}^{n-\ell} E \left(z_0 z_j K \left[\frac{x-x_0^{(m)}}{\gamma_n} \right] K \left[\frac{x-x_j^{(m)}}{\gamma_n} \right] \right) \\ &\quad - 2 \frac{1}{n^2} \sum_{\ell=1}^{n-1} \sum_{j=1}^{n-\ell} E \left(z_0 K \left[\frac{x-x_0^{(m)}}{\gamma_n} \right] \right)^2 \end{aligned} \quad (\text{A.2})$$

where

$$\begin{aligned} z_j^{(m)} &= E(z_j | w_j, w_{j-1}, w_{j-2}, \dots, w_{j-m}) \\ x_j^{(m)} &= E(x_j | w_j, w_{j-1}, w_{j-2}, \dots, w_{j-m}) \end{aligned} \quad (\text{A.3})$$

We shall prove the lemma in four steps:

Step 1:

$$|d_n^{(m)}(x)| \leq 4n^{-1} \left(m+1 + \sum_{\ell=0}^{\infty} \phi(\ell)^{1/2} \right) E \left(z_0^{(m)} K \left[\frac{x-x_0^{(m)}}{\gamma_n} \right] \right)^2$$

Step 2:

$$\begin{aligned} & \left| E \left(z_0 K \left[\frac{x-x_0}{\gamma_n} \right] \right)^2 - E \left(z_0^{(m)} K \left[\frac{x-x_0^{(m)}}{\gamma_n} \right] \right)^2 \right| \\ & \leq c_1 \gamma_n^{-1} \nu(m)^{1/2} + c_2 \gamma_n^{-2} \nu(m) \end{aligned}$$

if n is sufficiently large;

Step 3:

$$\begin{aligned} & \left| E \left(z_0 K \left[\frac{x-x_0}{\gamma_n} \right] \right) \left(z_j K \left[\frac{x-x_j}{\gamma_n} \right] \right) \right. \\ & \quad \left. - E \left(z_0^{(m)} K \left[\frac{x-x_0^{(m)}}{\gamma_n} \right] \right) \left(z_j^{(m)} K \left[\frac{x-x_j^{(m)}}{\gamma_n} \right] \right) \right| \leq c_3 \gamma_n^{-1} \nu(m)^{1/2} \end{aligned}$$

if n is sufficiently large;

Step 4:

$$\left| E \left(z_0 K \left[\frac{x-x_0}{\gamma_n} \right] \right)^2 - E \left(z_0^{(m)} K \left[\frac{x-x_0^{(m)}}{\gamma_n} \right] \right)^2 \right| \leq c_3 \gamma_n^{-1} \nu(m)^{1/2}$$

if n is sufficiently large. We first show that the results of these four steps imply the lemma.

Let (m_n) be an arbitrary sequence of positive integers converging to infinity and let n be so large that

$$m_n \geq 1 + \sum_{\ell=0}^{\infty} \phi(\ell)^{1/2} \quad (\text{A.4})$$

Then the results of the above four steps imply

$$d_n(x) = O[\rho_n(x)] \quad (\text{A.5})$$

where

$$\rho_n(x) = \max \left\{ \frac{m_n}{n} \gamma_n^{-1} \nu(m_n)^{1/2}, \frac{m_n}{n} E z_0^2 K \left[\frac{x-x_0}{\gamma_n} \right]^2, \gamma_n^{-1} \nu(m_n)^{1/2}, \gamma_n^{-2} \nu(m_n) \right\} \quad (\text{A.6})$$

Without loss of generality, we may assume

$$\gamma_n^{-2} \nu(m_n) \leq 1 \quad \gamma_n^{-1} \nu(m_n)^{1/2} \leq E \{ z_0 K[(x-x_0)/\gamma_n] \}^2 \quad (\text{A.7})$$

(as will appear). Then

$$\begin{aligned} \rho_n(x) &= \max \left\{ \frac{m_n}{n} E z_0^2 K \left[\frac{x-x_0}{\gamma_n} \right]^2, \gamma_n^{-1} \nu(m_n)^{1/2} \right\} \\ &\leq \frac{m_n}{n} E z_0^2 K \left[\frac{x-x_0}{\gamma_n} \right]^2 + \gamma_n^{-1} \exp(-\frac{1}{2} c m_n) \end{aligned} \quad (\text{A.8})$$

Minimizing the right-hand side of (A.8) to m_n yields

$$\nu(m_n)^{1/2} = \exp(-\frac{1}{2} c m_n) = [2\gamma_n/(cn)] \{ E z_0^2 K[(x-x_0)/\gamma_n] \}^{-1} \quad (\text{A.9})$$

with

$$m_n = \frac{2}{c} \ln \frac{2}{c} + \frac{2}{c} \ln \frac{n}{\gamma_n} + \frac{2}{c} \ln \left[E z_0^2 K \left[\frac{x-x_0}{\gamma_n} \right]^2 \right]^{-1} \quad (\text{A.10})$$

[thus observe that indeed (A.7) holds]. Hence

$$\begin{aligned} \rho_n(x) &\leq \frac{2}{c} \left| 1 + \ln \frac{2}{c} + \ln \frac{n}{\gamma_n} + \ln \left\{ \left[E z_0^2 K \left[\frac{x-x_0}{\gamma_n} \right]^2 \right]^{-1} \right\} \right| \\ &\quad \times \frac{E z_0^2 K[(x-x_0)/\gamma_n]^2}{n} \\ &= O \left(\left(\ln \frac{n}{\gamma_n} + \ln \left[E z_0^2 K \left(\frac{x-x_0}{\gamma_n} \right)^2 \right]^{-1} \right) \frac{E z_0^2 K[(x-x_0)/\gamma_n]^2}{n} \right) \end{aligned} \quad (\text{A.11})$$

as will be shown.

Proof of step 1. Since $\{z_0^{(m)} K[(x-x_0^{(m)})/\gamma_n]\}$ is a ϕ^* -mixing sequence, where

$$\phi^*(j) = \begin{cases} 1 & \text{if } j < m+1 \\ \phi(j-m) & \text{if } j \geq m+1 \end{cases} \quad (\text{A.12})$$

it follows from Lemma 1 of Billingsley (1968, p. 170) that

$$\begin{aligned} &\left| \text{cov} \left\{ z_0^{(m)} K \left[\frac{x-x_0^{(m)}}{\gamma_n} \right], z_j^{(m)} K \left[\frac{x-x_j^{(m)}}{\gamma_n} \right] \right\} \right| \\ &\leq 2\phi^*(j)^{1/2} E \left(z_0^{(m)} K \left[\frac{x-x_0^{(m)}}{\gamma_n} \right] \right)^2 \end{aligned} \quad (\text{A.13})$$

Step 1 now follows from the fact that

$$\sum_{j=0}^{\infty} \phi^*(j)^{1/2} \leq m+1 + \sum_{j=0}^{\infty} \phi(j)^{1/2} \quad (\text{A.14})$$

Proof of step 2. From the inversion formula for Fourier transforms we have

$$K(x)^\ell = \left(\frac{1}{2\pi}\right)^k \int \exp(-it'x) \psi_\ell(t) dt \quad \ell=1, 2 \quad (\text{A.15})$$

Hence,

$$\begin{aligned} & \left| Ez_0 K\left[\frac{x-x_0}{\gamma_n}\right] - Ez_0^{(m)} K\left[\frac{x-x_0^{(m)}}{\gamma_n}\right] \right| \\ & \leq \left(\frac{1}{2\pi}\right)^k \int E|z_0 \exp(it'x_0) - z_0^{(m)} \exp(it'x_0^{(m)})| |\psi_1(\gamma_n t)| dt \quad (\text{A.16}) \end{aligned}$$

Moreover,

$$\begin{aligned} E|z_0 \exp(it'x_0) - z_0^{(m)} \exp(it'x_0^{(m)})| & \leq |t| E|z_0| |x_0 - x_0^{(m)}| + E|z_0 - z_0^{(m)}| \\ & \leq \{\sqrt{Ez_0^2} |t| + 1\} \nu(m)^{1/2} \quad (\text{A.17}) \end{aligned}$$

Combining (A.16) and (A.17), it easily follows that

$$|Ez_0 K[(x-x_0)/\gamma_n] - Ez_0^{(m)} K[(x-x_0^{(m)})/\gamma_n]| \leq c_* \gamma_n^{-1} \nu(m)^{1/2} \quad (\text{A.18})$$

for some $c_* > 0$. Moreover, using the trivial inequality $|a^2 - b^2| \leq 2|a||a-b| + |a-b|^2$, it easily follows from (A.18) that

$$\begin{aligned} & |Ez_0 K[(x-x_0)/\gamma_n] - Ez_0^{(m)} K[(x-x_0^{(m)})/\gamma_n]| \\ & \leq 2|Ez_0 K[(x-x_0)/\gamma_n]| c_* \gamma_n^{-1} \nu(m)^{1/2} + c_*^2 \gamma_n^{-1} \nu(m) \quad (\text{A.19}) \end{aligned}$$

Realizing that, by (A.15),

$$\left| Ez_0 K\left[\frac{x-x_0}{\gamma_n}\right] \right| \leq E|z_0| \left(\frac{1}{2\pi}\right)^k \int |\psi_1(t)| dt < \infty \quad (\text{A.20})$$

step 2 follows from (A.19) and (A.20).

Proof of step 3. From (A.15) it follows that

$$\begin{aligned} & \left| E_0 z_j K\left[\frac{x-x_0}{\gamma_n}\right] K\left[\frac{x-x_j}{\gamma_n}\right] - E_0^{(m)} z_j^{(m)} K\left[\frac{x-x_0^{(m)}}{\gamma_n}\right] K\left[\frac{x-x_j^{(m)}}{\gamma_n}\right] \right| \\ & \leq \left(\frac{1}{2\pi}\right)^{2k} \gamma_n^{-2k} \\ & \quad \times \iint E|z_0 z_j \exp(it'_1 x_0 + it'_2 x_j) - z_0^{(m)} z_j^{(m)} \exp(it'_1 x_0^{(m)} + it'_2 x_j^{(m)})| \\ & \quad \times |\psi_1(\gamma_n t_1) \psi_1(\gamma_n t_2)| dt_1 dt_2 \quad (\text{A.21}) \end{aligned}$$

Moreover, similarly to (A.17), we have

$$\begin{aligned}
 E|z_0 z_j \exp(it'_1 x_0 + it'_2 x_j) - z_0^{(m)} z_j^{(m)} \exp(it'_1 x_0^{(m)} + it'_2 x_j^{(m)})| \\
 \leq \sqrt{E(z_0 z_j)^2} |(t_1, t_2)| \sqrt{E|(x_0, x_j) - (x_0^{(m)}, x_j^{(m)})|^2} + E|z_0 z_j - z_0^{(m)} z_j^{(m)}| \\
 \leq \{\sqrt{2} \sqrt{Ez_0^4} |(t_1, t_2)| + 2 \sqrt{Ez_0^2}\} \nu(m)^{1/2} + \nu(m) \\
 \leq \{c_*^{(1)} |(t_1, t_2)| + c_*^{(2)}\} \nu(m)^{1/2} \quad \text{for some } c_*^{(1)} > 0, c_*^{(2)} > 0 \quad (\text{A.22})
 \end{aligned}$$

Combining (A.21) and (A.22), step 3 easily follows.

Proof of step 4. Similarly to (A.21) we have

$$\begin{aligned}
 \left| Ez_0^2 K \left[\frac{x - x_0}{\gamma_n} \right]^2 - E(z_0^{(m)})^2 K \left[\frac{x - x_0^{(m)}}{\gamma_n} \right]^2 \right| \\
 \leq \left(\frac{1}{2\pi} \right)^k \gamma_n^{-k} \int E|z_0^2 \exp(it'x_0) - (z_0^{(m)})^2 \exp(it'x_0^{(m)})| |\psi_2(\gamma_n t)| dt \quad (\text{A.23})
 \end{aligned}$$

and similarly to (A.22) we have

$$\begin{aligned}
 E|z_0^2 \exp(it'x_0) - (z_0^{(m)})^2 \exp(it'x_0^{(m)})| \\
 \leq \{c_{**}^{(1)} |t| + c_{**}^{(2)}\} \nu(m)^{1/2} \quad \text{for some } c_{**}^{(1)} > 0, c_{**}^{(2)} > 0 \quad (\text{A.24})
 \end{aligned}$$

Step 4 easily follows from (A.23) and (A.24). This completes the proof. ■

Proof of Theorem 5.1.1. We shall prove this theorem only for the continuous case with kernel $\hat{K}_2(x)$. The proof for the other cases is similar and therefore left to the reader.

Let

$$\hat{s}(x) = \frac{1}{n} \sum_{j=1}^n \frac{y_j \exp[-\frac{1}{2}(x - x_j)' \hat{V}^{-1}(x - x_j)/\gamma_n^2]}{(\sqrt{2\pi})^k \gamma_n^k \sqrt{\det(\hat{V})}} \quad (\text{A.25})$$

$$\tilde{s}(x) = \frac{1}{n} \sum_{j=1}^n \frac{y_j \exp[-\frac{1}{2}(x - x_j)' V^{-1}(x - x_j)/\gamma_n^2]}{(\sqrt{2\pi})^k \gamma_n^k \sqrt{\det(\hat{V})}} \quad (\text{A.26})$$

$$\bar{s}(x) = \frac{1}{n} \sum_{j=1}^n \frac{y_j \exp[-\frac{1}{2}(x - x_j)' V^{-1}(x - x_j)/\gamma_n^2]}{(\sqrt{2\pi})^k \gamma_n^k \sqrt{\det(V)}} \quad (\text{A.27})$$

Moreover, let

$$\hat{M} = \max_{i,j} |\hat{v}^{(i,j)} - v^{(i,j)}| \quad (\text{A.28})$$

where $\hat{v}^{(i,j)}$ is the typical element of \hat{V}^{-1} and $v^{(i,j)}$ is the corresponding typical element of V^{-1} . For every $t = (t_1, \dots, t_k)' \in R^k$ we have

$$|t' \hat{V}^{-1} t - t' V^{-1} t| \leq \hat{M} \sum_{i,j} |t_i, t_j| \leq k \hat{M} t' t \leq \rho \hat{M} t' V^{-1} t \quad (\text{A.29})$$

where ρ is the maximum eigenvalue of V times k . Using inequality (A.29) and the mean value theorem, it is not too hard to verify that

$$\begin{aligned} |\hat{s}(x) - \bar{s}(x)| &\leq \frac{\rho \hat{M}}{n \gamma_n^k} \sum_{j=1}^n |y_j| \frac{(x-x_j)' V^{-1} (x-x_j)}{\gamma_n^2} \\ &\quad \times \exp\left[\frac{1}{2} \rho \hat{M} (x-x_j)' V^{-1} (x-x_j) / \gamma_n^2\right] \\ &\quad \times \frac{\exp\left[-\frac{1}{2} (x-x_j)' V^{-1} (x-x_j) / \gamma_n^2\right]}{(\sqrt{2\pi})^k \sqrt{\det(\hat{V})}} \end{aligned} \quad (\text{A.30})$$

Now suppose for the moment that the x_j 's are independent. Then, for every $\epsilon > 0$,

$$n^{1/2-\epsilon} (\hat{V} - V) \rightarrow 0 \quad (\text{A.31})$$

in probability, as is easy to verify. Since the elements of an inverse matrix are continuously differentiable functions of the elements of the inverted matrix, provided the inverted matrix is nonsingular, it follows that (A.31) implies

$$n^{1/2-\epsilon} (\hat{V}^{-1} - V^{-1}) \rightarrow 0 \quad (\text{A.32})$$

in probability, and consequently

$$p \lim_{n \rightarrow \infty} n^{1/2-\epsilon} \hat{M} = 0 \quad \text{for every } \epsilon > 0 \quad (\text{A.33})$$

Thus also

$$\lim_{n \rightarrow \infty} P(\rho \hat{M} < \frac{1}{4}) = 1 \quad (\text{A.34})$$

Now (A.30) and (A.34) imply that the inequality

$$|\hat{s}(x) - \bar{s}(x)| \leq \frac{2\rho k \hat{M}}{\gamma_n^k} \frac{\sqrt{\det(V)}}{\sqrt{\det(\hat{V})}} (\sqrt{2})^k \frac{1}{n} \sum_{j=1}^n |y_j| K_* \left[\frac{x-x_j}{\gamma_n} \right] \quad (\text{A.35})$$

with

$$K_*(x) = (x' V^{-1} x / 2k) \exp(-\frac{1}{4} x' V^{-1} x) / [(\sqrt{2\pi})^k (\sqrt{2})^k \sqrt{\det(V)}] \quad (\text{A.36})$$

holds with probability converging to 1. Since

$$\begin{aligned} E \frac{1}{n} \sum_{j=1}^n |y_j| K_* \left[\frac{x-x_j}{\gamma_n} \right] \gamma_n^{-k} \\ &\leq E(1+y_0^2) K_* \left[\frac{x-x_0}{\gamma_n} \right] \gamma_n^{-k} \\ &= \int [1 + \sigma_u^2(x - \gamma_n z) + g(x - \gamma_n z)^2] h(x - \gamma_n z) K_*(z) dz \\ &\rightarrow [1 + \sigma_u^2(x) + g(x)^2] h(x) \quad \text{as } n \rightarrow \infty \end{aligned} \quad (\text{A.37})$$

and since (A.33) implies

$$p \lim_{n \rightarrow \infty} \sqrt{n\gamma_n^k} \tilde{M} = 0 \quad (\text{A.38})$$

it now follows that, pointwise in x ,

$$p \lim_{n \rightarrow \infty} \sqrt{n\gamma_n^k} |\hat{s}(x) - \bar{s}(x)| = 0 \quad (\text{A.39})$$

Next, observe that

$$p \lim_{n \rightarrow \infty} \sqrt{n\gamma_n^k} |\bar{s}(x) - \bar{s}(x)| = 0 \quad (\text{A.40})$$

for (A.31) implies that

$$p \lim_{n \rightarrow \infty} \sqrt{n\gamma_n^k} [\det(\hat{V}) - \det(V)] = 0 \quad (\text{A.41})$$

Thus,

$$p \lim_{n \rightarrow \infty} \sqrt{n\gamma_n^k} [\hat{s}(x) - \bar{s}(x)] = 0 \quad (\text{A.42})$$

From this result it follows straightforwardly that under the i.i.d. assumption the asymptotic normality results go through. The proof that the uniform consistency results go through is analogous to Bierens (1983a). So the proof of the theorem under review is complete if we show that (A.31) goes through for time series. This follows from Lemma A.

Lemma A. Let (z_j) be a strictly stationary stochastic process in R satisfying Ez_j^4 . Let (z_j) be ν stable in L^4 with respect to a strictly stationary ϕ -mixing base, where

$$\nu(m) = O[\exp(-cm)] \quad \text{for some } c > 0 \quad (\text{A.43})$$

$$\sum_{\ell=0}^{\infty} \phi(\ell)^{1/2} < \infty \quad (\text{A.44})$$

Then for every $\epsilon > 0$,

$$p \lim_{n \rightarrow \infty} n^{1/2-\epsilon} \frac{1}{n} \sum_{j=1}^n (z_j - Ez_j) = 0 \quad (\text{A.45})$$

and

$$p \lim_{n \rightarrow \infty} n^{1/2-\epsilon} \frac{1}{n} \sum_{j=1}^n (z_j^2 - Ez_j^2) = 0 \quad (\text{A.46})$$

Proof. Let (w_j) be the base and let

$$z_j^{(m)} = E(z_j | w_j, w_{j-1}, w_{j-2}, \dots, w_{j-m}) \quad (\text{A.47})$$

Denote, similarly to (4.1.14), (A.1), and (A.2),

$$d_n = \text{var} \left(\frac{1}{n} \sum_{j=1}^n z_j^2 \right) - \frac{1}{n^2} \sum_{j=1}^n \text{var}(z_j^2) \quad (\text{A.48})$$

$$d_n^{(m)} = \text{var} \left(\frac{1}{n} \sum_{j=1}^n (z_j^{(m)})^2 \right) - \frac{1}{n^2} \sum_{j=1}^n \text{var}[(z_j^{(m)})^2] \quad (\text{A.49})$$

Then, similarly to the proof of Lemma 4.1.2, it follows that

$$\begin{aligned} |d_n^{(m)}| &\leq 4 \left(\frac{m+1 + \sum_{\ell=0}^{\infty} \phi(\ell)^{1/2}}{n} \right) E(z_j^{(m)})^4 \\ &\leq 8 \left(\frac{m+1 + \sum_{\ell=0}^{\infty} \phi(\ell)^{1/2}}{n} \right) [Ez_0^4 + E(z_0 - z_0^{(m)})^4] \\ &\leq 16 \left(\frac{m+1 + \sum_{\ell=0}^{\infty} \phi(\ell)^{1/2}}{n} \right) Ez_0^4 \\ &\leq 32 \frac{m}{n} Ez_0^4 \\ &= \frac{c_1 m}{n} \end{aligned} \quad (\text{A.50})$$

if $\nu(m) \leq Ez_0^4$ and $m \geq 1 + \sum_{\ell=0}^{\infty} \phi(\ell)^{1/2}$, and moreover,

$$|d_n - d_n^{(m)}| \leq c_2 \nu(m)^{1/4} \quad \text{for some } c_2 > 0 \quad (\text{A.51})$$

Thus, we have, for sufficiently large $m > 0$,

$$|d_n| \leq c_1(m/n) + c_2 \exp(-\frac{1}{4}cm) \quad (\text{A.52})$$

Now choose $m = n^\epsilon$. Then

$$\begin{aligned} n^{1-2\epsilon} \text{var} \left(\frac{1}{n} \sum_{j=1}^n z_j^2 \right) &\leq n^{1-2\epsilon} [c_1 n^{\epsilon-1} + c_2 \exp(-\frac{1}{4}cn^\epsilon)] + n^{-2\epsilon} Ez_0^4 \\ &\rightarrow 0 \quad \text{as } n \rightarrow \infty \end{aligned} \quad (\text{A.53})$$

This proves (A.46). The proof of (A.45) is nearly the same. \blacksquare

References

- Ahmad, I. A., and P. E. Lin, 1976, Non-parametric sequential estimation of a multiple regression function, *Bull. Math. Statist.* **17**, 63-75.
 Benedetti, J. K., 1977, On the non-parametric estimation of regression functions, *J. Roy. Statist. Soc. B* **39**, 248-53.
 Bierens, H. J., 1981, *Robust methods and asymptotic theory in nonlinear econometrics*, Lecture notes in economics and mathematical systems, Vol. 192, Springer-Verlag, New York.

- 1982a, A uniform weak law of large numbers under ϕ -mixing with application to nonlinear least squares estimation, *Statist. Neerland.* **36**, 81-6.
- 1982b, Consistent model specification tests, *J. Econometr.* **20**, 105-34.
- 1983a, Uniform consistency of kernel estimators of a regression function under generalized conditions, *JASA* **77**, 699-707.
- 1983b, Sample moments integrating normal kernel estimators of multivariate density and regression functions, *Sankhya B* **45**, 160-92.
- 1984, Model specification testing of time series regressions, *J. Econometr.* **26**, 323-53.
- Billingsley, P., 1968, *Convergence of probability measures*, Wiley, New York.
- Box, G. E. P., and D. R. Cox, 1964, An analysis of transformations, *J. Roy. Statist. Soc. B* **26**, 211-43.
- Cacoullos, T., 1966, Estimation of a multivariate density, *Ann. Inst. Statist. Math.* **18**, 179-89.
- Cencov, N. N., 1962, Evaluation of an unknown distribution density from observations, *Sov. Math.* **3**, 1559-62.
- Cheng, K. F., 1983, Strong convergence in nonparametric estimation of regression functions, *Period. Math. Hung.* **14**, 177-87.
- Cheng, K. F., and P. E. Lin, 1981a, Nonparametric estimation of a regression function, *Z. Wahrscheinlichkeitstheorie verw. Gebiete* **57**, 223-33.
- 1981b, Nonparametric estimation of a regression function: limiting distribution, *Austr. J. Statist.* **23**, 186-95.
- Chung, K. L., 1974, *A course in probability theory*, Academic, New York.
- Clark, R. M., 1977, Non-parametric estimation of a smooth regression function, *J. Roy. Statist. Soc. B* **39**, 107-13.
- 1979, Calibration, cross-validation and carbon-14, I, *J. Roy. Statist. Soc. A* **142**, 47-62.
- 1980, Calibration, cross-validation and carbon-14, II, *J. Roy. Statist. Soc. A* **143**, 177-94.
- Collomb, G., 1981, Estimation non-paramétrique de la régression: revue bibliographique, *Int. Statist. Rev.* **49**, 75-93.
- 1985a, Non-parametric time series analysis and prediction: uniform almost sure convergence of the window and K-NN autoregression estimates, *Statistics* **16**, 297-307.
- 1985b, Nonparametric regression: an up-to-date bibliography, *Statistics* **16**, 309-24.
- Devroye, L. P., 1978, The uniform convergence of the Nadaraya-Watson regression function estimate, *Can. J. Statist.* **6**, 179-91.
- Devroye, L. P., and T. J. Wagner, 1980a, On the L_1 convergence of kernel estimators of regression functions with applications in discrimination, *Z. Wahrscheinlichkeitstheorie verw. Gebiete* **51**, 15-25.
- 1980b, Distribution-free consistency results in non-parametric discrimination and regression function estimation, *Ann. Statist.* **8**, 231-9.
- Devroye, L. P., and G. L. Wise, 1980, Consistency of a recursive nearest neighborhood regression function estimate, *J. Mult. Anal.* **10**, 539-50.
- Epanechnikov, V. A., 1969, Non-parametric estimation of a multivariate probability density, *Theory Prob. Applic.* **14**, 153-8.
- Fryer, M. J., 1977, A review of some non-parametric methods of density estimation, *J. Inst. Math. Applic.* **20**, 335-54.

- Gallant, A. R., 1981, On the bias in flexible functional forms and an essentially unbiased form: the Fourier flexible form, *J. Econometr.* **15**, 211-45.
- Geisser, S., 1975, The predictive sample reuse method with applications, *JASA* **70**, 320-8.
- Georgiev, A. A., 1984a, Nonparametric system identification by kernel methods, *IEEE Trans. Aut. Cont.* **29**, 356-8.
- 1984b, Nonparametric mathematical model for individual human growth curve, *Cyb. Syst. Res.* **2**, 277-9.
- 1984c, A nonparametric algorithm for identification of linear dynamic SISO systems of unknown order, *Syst. Contr. Lett.* **4**, 273-80.
- 1984d, Speed of convergence in nonparametric kernel estimation of a regression function and its derivatives, *Ann. Inst. Statist. Math.* **36**, 455-62.
- 1984e, Kernel estimates of functions and their derivatives with applications, *Statist. Prob. Lett.* **2**, 45-50.
- Greblicki, W., and A. Krzyzak, 1980, Asymptotic properties of kernel estimates of a regression function, *J. Statist. Plan. Inf.* **4**, 81-90.
- Konakov, V. D., 1977, On a global measure of deviation for an estimate of the regression line, *Theory Prob. Applic.* **22**, 858-68.
- Leech, D. 1975, Testing the error specification in nonlinear regression, *Econometrica* **43**, 719-25.
- Liero, H., 1982, On the maximal deviation of the kernel regression function estimate, *Math. Operationsforsch. Statist., Ser. Statist.* **13**, 171-82.
- McFadden, D., 1985, Specification of econometric models, Presidential Address, Fifth World Congress of the Econometric Society.
- McLeish, D. L., 1974, Dependent central limit theorems and invariance principles, *Ann. Prob.* **2**, 620-8.
- 1975, A maximal inequality and dependent strong laws, *Ann. Prob.* **3**, 829-39.
- Nadaraya, E. A., 1964, On estimating regression, *Theory Prob. Applic.* **9**, 141-2.
- 1965, On non-parametric estimation of density functions and regression curves, *Theory Prob. Applic.* **10**, 186-90.
- 1970, Remarks on non-parametric estimates for density functions and regression curves, *Theory Prob. Applic.* **15**, 134-7.
- Noda, K., 1976, Estimation of a regression function by the Parzen kernel-type density estimators, *Ann. Inst. Statist. Math.* **28**, 221-34.
- Parzen, E., 1962, On estimation of a probability density function and mode, *Ann. Math. Statist.* **33**, 1065-76.
- Priestley, M. B., and M. T. Chao, 1972, Non-parametric function fitting, *J. Roy. Statist. Soc. B.* **34**, 385-95.
- Révész, P., 1979, On the nonparametric estimation of the regression function, *Probl. Contr. Inf. Theory* **8**, 297-302.
- Robinson, P. M., 1983, Nonparametric estimators for time series, *J. Time Ser. Anal.* **4**, 185-207.
- Rosenblatt, M., 1956, Remarks on some non-parametric estimates of a density function, *Ann. Math. Statist.* **27**, 832-7.
- Schuster, E. F., 1972, Joint asymptotic distribution of the estimated regression function at a finite number of distinct points, *Ann. Math. Statist.* **43**, 84-8.
- Schuster, E. F., and S. Yakowitz, 1979, Contributions to the theory of nonparametric regression, with application to system identification, *Ann. Statist.* **7**, 139-49.

- Silverman, B. W., 1978, Weak and strong uniform consistency of the kernel estimate of a density and its derivatives, *Ann. Statist.* **6**, 177-84.
- Singh, R. S., 1981, Speed of convergence in non-parametric estimation of a multivariate μ -density and its mixed partial derivatives, *J. Statist. Plan. Inf.* **5**, 287-98.
- Singh, R. S., and A. Ullah, 1985, Nonparametric time series estimation of joint DGP, conditional DGP and vector autoregression, *Econometric Theory* **1**, 27-52.
- Spiegelman, C., and J. Sacks, 1980, Consistent window estimation in nonparametric regression, *Ann. Statist.* **8**, 240-6.
- Spitzer, J. J., 1976, The demand for money, the liquidity trap, and functional forms, *Int. Econ. Rev.* **17**, 220-7.
- Stone, C., 1977, Consistent nonparametric regression (with discussion), *Ann. Statist.* **5**, 595-645.
- Stone, M., 1974, Cross-validated choice and assessment of statistical predictions (with discussion), *J. Roy. Statist. Soc. B* **36**, 111-47.
- Tapia, R. A., and J. R. Thompson, 1978, *Nonparametric probability density estimation*, Johns Hopkins University Press, Baltimore and London.
- Wahba, G., and S. Wold, 1975, A completely automatic French curve: fitting spline functions by cross-validation, *Commun. Statist.* **4**, 1-17.
- Watson, G. S., 1964, Smooth regression analysis, *Sankhya A* **26**, 359-72.
- White, H., and I. Domowitz, 1984, Nonlinear regression with dependent observations, *Econometrica* **52**, 143-61.
- White, K. J., 1972, Estimation of the liquidity trap with a generalized functional form, *Econometrica* **40**, 193-9.
- Wong, W. H., 1983, On the consistency of cross-validation in kernel nonparametric regression, *Ann. Statist.* **11**, 1136-41.
- Zarembka, P., 1968, Functional form in the demand for money, *JASA* **18**, 502-11.